iIMPACT: integrating image and molecular profiles for spatial transcriptomics analysis

Xi Jiang[1,2], Shidan Wang[1], Lei Guo[1], Bencong Zhu[3,4], Zhuoyu Wen[1], Liwei Jia[5], Lin Xu[1*], Guanghua Xiao[1*] and Qiwei Li[4*]

*Correspondence:
Lin.Xu@utsouthwestern.edu;
Guanghua.Xiao@utsouthwestern.
edu; Qiwei.Li@utdallas.edu

[1] Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX, USA
[2] Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, USA
[3] Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China
[4] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX, USA
[5] Department of Pathology, The University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Current clustering analysis of spatial transcriptomics data primarily relies on molecular information and fails to fully exploit the morphological features present in histology images, leading to compromised accuracy and interpretability. To overcome these limitations, we have developed a multi-stage statistical method called iIMPACT. It identifies and defines histology-based spatial domains based on AI-reconstructed histology images and spatial context of gene expression measurements, and detects domain-specific differentially expressed genes. Through multiple case studies, we demonstrate iIMPACT outperforms existing methods in accuracy and interpretability and provides insights into the cellular spatial organization and landscape of functional genes within spatial transcriptomics data.

Keywords: Spatially resolved transcriptomics, AI-reconstructed histology image, Markov random field, Spatial clustering, Spatially variable gene

Background

Spatially resolved transcriptomics (SRT), a new generation of RNA profiling techniques, provides biological information at the cellular level while preserving the organization of the tissue and cellular microenvironment [1–4]. One category of SRT methods builds upon next-generation sequencing (NGS)-based SRT techniques, including spatial transcriptomics (ST) [5], 10x Visium (an improved ST platform), Slide-seq [6], Slide-seqV2 [7], and high-definition spatial transcriptomics (HDST) [8]. These techniques capture RNA molecules via spatially arrayed barcoded probes. The barcoded areas, namely spots, cover a group of cells and are usually arrayed on a two-dimensional grid. Another category of SRT platforms is based on imaging techniques, such as seqFISH [9], MERFISH [10], and STARmap [11]. They measure the expression level for hundreds to thousands of genes at the single-cell resolution with detailed spatial organization information. With these advancements, SRT techniques have been widely applied to facilitate discoveries of novel insights in biomedical studies.

Jiang *et al. Genome Biology* (2024) 25:147

Page 2 of 25

A central challenge for SRT data analysis is to define clinically or biologically meaningful spatial domains by partitioning regions with similar molecular and/or histological characteristics, because the spatial domain identification serves as the foundation for several important downstream analyses, including but not limited to the domain-based differential expression analysis, trajectory analysis, and functional pathway analysis [12, 13]. However, current state-of-the-art methods typically focus on achieving this goal solely by analyzing SRT molecular profiles, such as gene expression, while neglecting the valuable morphological or biological information present in the associated histology images. For example, the Seurat package, the most prevalent single-cell RNA sequencing data analysis pipeline [14, 15], utilizes only the high-throughput gene expression of each spot for clustering analysis but omits spatial context. On the other hand, several recently developed methods, such as the hidden-Markov random field model [16], BayesSpace [17], and BASS [18], integrate spatial information using Bayesian frameworks, but do not leverage any information from the paired histology images. Meanwhile, a series of deep-learning-based methods are designed to integrate features extracted from the histology image to enhance SRT data clustering analysis. For example, SpaGCN [19] relies on the RGB channel data from areas surrounding spots for histological insights, whereas stLearn [20], MUSE [21], and SiGra [22] achieve image feature extraction via various deep neural network models. However, those features do not explicitly reveal detailed morphological information (e.g., cell locations and types) and, thus, have limited ability to directly provide biologically or clinically relevant insights.

Different from molecular information, histology images characterize cellular structures and tissue microenvironments, which have been proven valuable in clinical diagnosis and prognosis [23, 24]. Computer vision algorithms have enabled us to automatically segment cell nuclei from digital histology images at a large scale [25]. Recent developments in deep convolutional neural networks (e.g., H-DenseUNet [26], Micro-Net [27], Hover-Net [28], and HD-Staining model [23]) have further integrated the automatic identification, classification, and feature extraction of each observed nucleus in a histology image. In practice, a histology-based spatial domain (e.g., tissue) is defined as a group of cells with similar morphological and molecular context as a unit. Thus, we hypothesize that integrating spot-level molecular profiles and cellular-level image profiles from AI-reconstructed histology images—digitally processed tissue samples with AI-identified and classified nuclei, primarily using deep learning—could enhance the spatial domain identification in terms of both accuracy and interpretability.

Another challenge for SRT data analysis is to identify spatial domain-specific differentially expressed genes (spaDEGs), which are defined as genes enriched in a given spatial domain. Recently developed methods, such as SpatialDE [29], SPARK [30], BOOST-GP [31], and BOOST-MI [32] focus on identifying spatially variable genes (SVGs), which represent genes with spatially correlated expression patterns [30, 31]. They characterize the global spatial dependency of a gene in the whole domain while ignoring the spatial pattern heterogeneity due to cellular organization, which could be observed in AI-reconstructed histology images. SpaGCN [19] proposed domain-guided differential expression analysis to detect spaDEGs without a rigorous statistical framework. Therefore, there is an urgent need to develop a reliable statistical method to detect spaDEGs.

This paper proposes a two-stage statistical approach by integrating Image and Molecular Profiles to Analyze and Cluster spatial Transcriptomics data, or iIMPACT for short. The first stage is implementing a Bayesian finite mixture model to allocate all spots into mutually exclusive clusters, namely histology-based spatial domains. We decompose each mixture component into two sub-components to integrate image and molecular profiles. In particular, a multinomial sub-component is employed to model cell type abundance available in histology images. Following BayesSpace [17], we use a normal sub-component to model the low-dimensional representation of normalized gene expression from the matching SRT molecular profile. The Bayesian model also adopts a Markov random field prior (MRF) to encourage neighboring spots to be clustered in the same histology-based spatial domain. The spots' neighborhood structure can be straightforwardly defined from the NGS-based SRT geospatial profile, as spots are usually arrayed on square or triangular lattices. Through the resulting posterior inference, we obtain histology-based spatial domains and their interactive zones, while characterizing each identified histology-based spatial domain by inferring its underlying domain-specific relative abundance of cell types. The second stage is implementing a negative binomial (NB) regression model to search for spaDEGs, which are differentially expressed between a given histology-based spatial domain identified in the first stage and all others. This approach directly models the numbers of read counts (used as a proxy for gene expression) in the SRT molecular profile to achieve minimum information loss. iIMPACT could also be extended to analyze imaging-based SRT data via some special handling. Compared with existing state-of-the-art methods, iIMPACT is able to fully leverage information from the nuclei segmentation procedure on the histology images for clustering analysis and has strong biological interpretability. Applying iIMPACT on multiple datasets from different SRT platforms (summarized in Additional file 1: Table S1), we confirmed that iIMPACT performed better on both spatial domain identification and spaDEG detection than state-of-the-art methods. We further demonstrated that iIMPACT could capture biological features at both the spatial domain level and gene level. Therefore, by integrating image and molecular information, iIMPACT facilitates the discovery of new biological insights from SRT datasets.

Results

Overview of iIMPACT

iIMPACT is a two-stage statistical method to analyze SRT data, with its workflow shown in Fig. 1. It includes two stages—histology-based spatial domain identification by a Bayesian normal-multinomial mixture model and spaDEG detection by an NB regression model.

To achieve the above goals, iIMPACT utilizes the morphological context of histology images and the spatial context of gene expression measurements, referring to the image and molecular profiles in Fig. 1A and throughout the paper. In particular, the molecular profile refers to the low-dimensional representation of normalized gene expression values at the spot level (denoted by Y), which is obtained by a pre-specified dimension reduction technique, such as principal component analysis (PCA). The accompanying SRT geospatial profile that records all spots' locations is processed as an adjacent matrix (denoted by G) representing the spots' neighborhood structure. iIMPACT requires the

Jiang *et al. Genome Biology* (2024) 25:147

Page 4 of 25
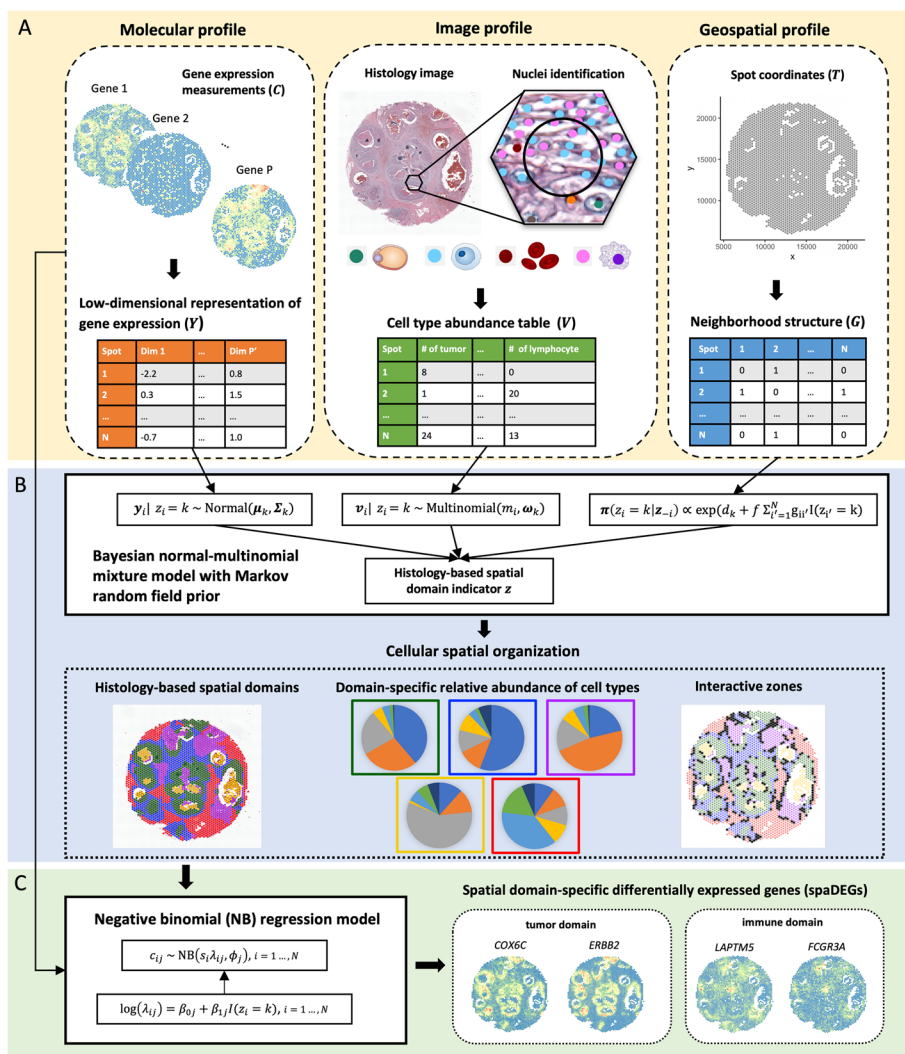
Fig. 1 Workflow of iIMPACT: **A** iIMPACT starts by combining and processing image profile from AI-reconstructed histology images, and geospatial and molecular profiles from SRT data (circled by dashed lines) to conduct the histology-based spatial domain identification. **B** A Bayesian normal-multinomial mixture model with the Markov random field (circled by solid lines) is fitted for histology-based spatial domain identification. Based on the spatial domain identification results, biologically important cellular spatial organization can be characterized, including the domain-specific relative abundance of cell types and interactive zones (circled by dotted lines). **C** Domain-specific spaDEGs are identified by a negative binomial (NB) regression model

locations and types of all cell nuclei in the matching histology image. Combining with the geospatial profile, we can generate the image profile (denoted by V), which indicates the spot-level cell type abundance, i.e., the number of different cell types within a spot and its expanded area.

In the first stage, we employ a Bayesian normal-multinomial mixture model with the MRF prior [33, 34] to identify the histology-based spatial domains (Fig. 1B) and interactive zones, corresponding to those spots with less confidence to be allocated to any histology-based spatial domains. Through model parameter estimation, iIMPACT can infer the underlying relative abundance of cell types at each histology-based spatial

Jiang *et al. Genome Biology* (2024) 25:147

Page 5 of 25

domain to provide a reference to distinguish their histological types. In the second stage, an NB regression model is fitted for each gene and each histology-based spatial domain of interest, where spaDEGs can be defined (Fig. 1C).

Application to human breast cancer dataset

We applied iIMPACT to analyze an SRT dataset from a human breast cancer study. This dataset includes 2518 spots and 17,651 genes. The gene expression was measured on a section of human breast with invasive ductal carcinoma via the 10x Visium platform, along with annotation from pathologists that was used to evaluate the accuracy of spatial domain detection (H&E-stained image with five annotated tissue regions in Fig. 2A). After applying HD-Staining [23] to the histology image of breast cancer tissue, we identified 156,235 cells within seven categories: macrophage, ductal epithelium, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (detailed information in Additional file 1: Fig. S1).

Firstly, we compared the five spatial domains identified by iIMPACT, SpaGCN [19], BayesSpace [17], BASS [18], stLearn [20], and MUSE [21], with manually annotated domains by pathologists. We quantified the clustering performance via the widely used adjusted Rand index (ARI). It generally ranges from 0 to 1, with higher values indicating greater consistency between the identified spatial domain pattern and the manual annotation, as illustrated in Additional file 1: Fig. S2. We found that iIMPACT achieved the highest consistency with the manual annotation (see Fig. 2B. ARI$=0.634$). stLearn (ARI$=0.527$) and SpaGCN (ARI$=0.520$) took the image-extracted features or image RGB values, respectively, instead of detailed histology information, which might contribute to their less satisfactory segmentation of non-tumor regions. However, they outperformed BASS (ARI$=0.496$) and BayesSpace (ARI$=0.419$). Notably, none of the methods performed well in separating the fat region (in blue) from the fibrous tissue (in red) per the manual annotation. Detailed comparisons of spatial domains across different numbers of domains K and the corresponding ARIs are presented in Additional file 1: Fig. S3 and S4, respectively. In summary, the better performance of iIMPACT suggests the advantage of integrating both molecular and image profiles in the clustering analysis of SRT data.

Secondly, iIMPACT is able to define each individual histology-based spatial domain simultaneously by inferring the latent spatial domain-specific relative abundance of cell types parametrized by the Bayesian multinomial-normal mixture model (Fig. 2C). In contrast, other methods, despite their good capabilities in identifying spatial domains, currently lack the ability to effectively integrate cell type information and directly interpret the identified domains in a biologically meaningful way. For example, as detailed in Fig. 2C, the proportion of tumor cells is higher in domain 1 (green spots in Fig. 2B) than in other domains, indicating that domain 1 is the tumor region. This inference is consistent with tumor regions in the manual annotation. Domain 2 (blue) and domain 5 (red) have a similar proportion of stromal cells, while the proportion of lymphocytes in domain 2 is higher than in domain 5. The difference in the relative abundance of cell types may indicate the functional difference between these two domains. These examples confirm that iIMPACT is able to provide biological interpretation of spatial domains.

Jiang *et al. Genome Biology* (2024) 25:147

Page 6 of 25
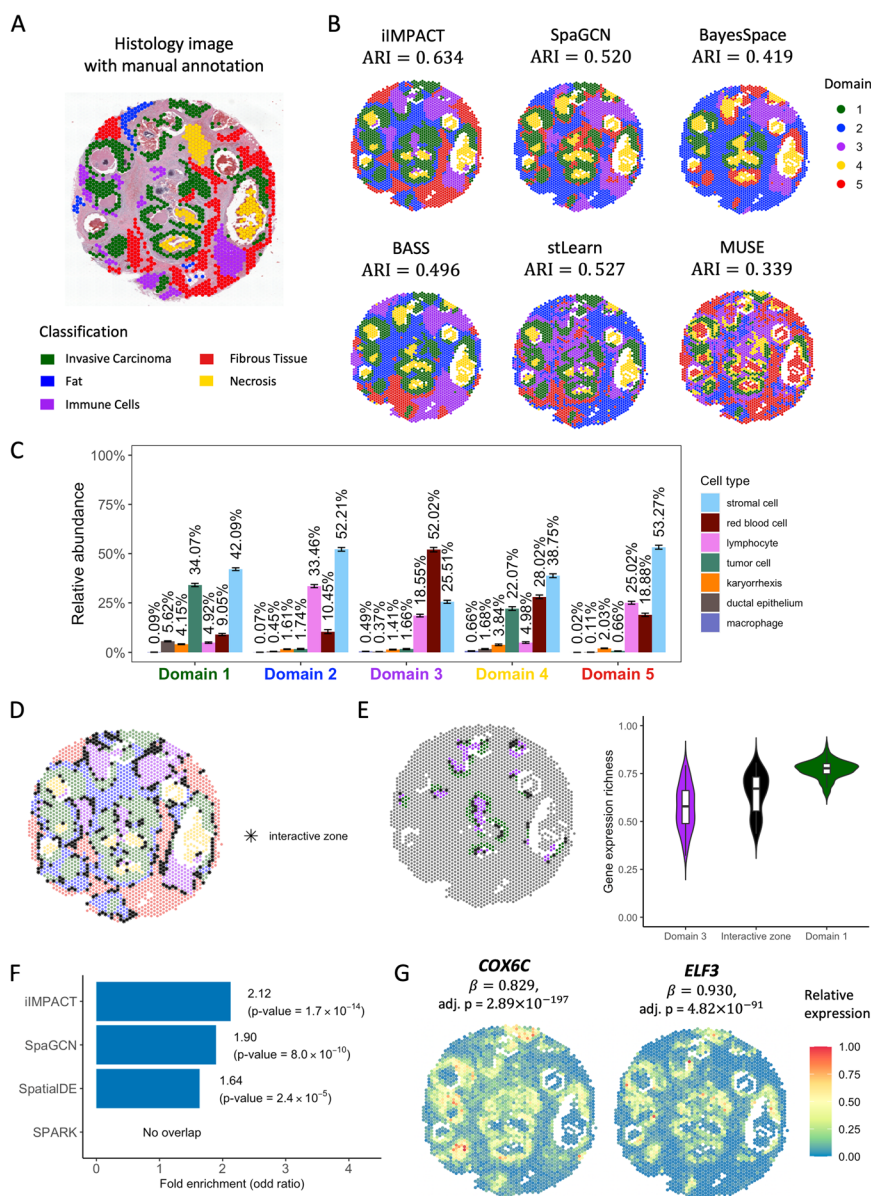
Fig. 2 Human breast cancer dataset: **A** H&E-stained image of the tissue section with spot-level manual annotation from pathologists. **B** Spatial domains detected by iIMPACT, SpaGCN, BayesSpace, BASS, stLearn, and MUSE, with the number of clusters to be five. **C** Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the seven cell types observed in the AI-reconstructed histology image. **D** Interactive zones (black asterisk spots) defined by iIMPACT. **E** Identified interactive zones (black asterisk spots) and other boundary areas of tumor domain and its adjacent domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. **F** Gene enrichment analysis between genes detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and known breast cancer genes from the COSMIC database. **G** Spatial expression patterns of two example spaDEGs, *COX6C* and *ELF3*, that were only detected by iIMPACT

Thirdly, iIMPACT can identify the interactive zones among histology-based spatial domains (Fig. 2D). Interactive zones are distinguished from the identified spatial domain. It is defined as spots with higher uncertainty on domain allocation, which potentially have higher diversity in cell type abundance and heterogeneity in gene

Jiang *et al. Genome Biology* (2024) 25:147

Page 7 of 25

expression compared with neighboring spots with unambiguous domain definition. We calculated the gene expression richness at each spot within the tumor boundary, immune boundary, and interactive zone, defining it as the percentage of genes exhibiting non-zero read counts. Note that this measure was taken at the spot level rather than the single-cell level. We observed statistically significant differences among these comparisons (Fig. 2E), implying that the identified zones are connected areas between tumor and immune domains with a high level of heterogeneity in gene expression and complex cellular interactions. By further comparing the gene expressions for these groups, we found several known cancer or immune genes with high expression in the interactive zones (e.g., *GREM1* [35]*),* suggesting the possible tumor-immune interactions in these zones.

Finally, we asked whether the spaDEGs defined by iIMPACT are more consistent with biological knowledge than those from other algorithms, which is an independent evaluation step frequently used for validating the clustering approaches on single-cell and spatial profiling data [19, 29, 30]. We focused on the tumor-domain specific spaDEGs defined by iIMPACT and SpaGCN [19], and SVGs by SpatialDE [29] and SPARK [30], respectively, and performed the enrichment analysis by comparing tumor-domain spaDEGs or SVGs defined by these four methods with the known breast cancer gene set defined in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. The number of genes identified by each method, along with their overlaps with the referenced gene set, is detailed in Additional file 1: Table S2. As summarized in Fig. 2F, the tumor-domain spaDEGs detected by iIMPACT showed higher overlap with the known breast cancer gene set than the genes detected by SpaGCN, SpatialDE, and SPARK, respectively, including two example genes that can only be detected by iIMPACT (Fig. 2G): *COX6C*, a known biomarker for the identification of hormone-responsive breast cancer [36], and *ELF3*, an epithelial-specific gene that is a novel therapeutic target of breast cancer and has been amplified in early breast cancer [37]. To have an additional diagnosis of the spatial signals of those detected genes, we employed Moran's *I* [38] statistic to quantify the degree of spatial autocorrelation of gene expression (details in Additional file 1: Section S1). Results are shown in Additional file 1: Fig. S5. Genes detected by iIMPACT exhibit a notably higher average Moran's *I* than those detected by SpaGCN and SPARK across various selection thresholds. Additionally, these genes demonstrate a higher average Moran's I than the top 1000 SpatialDE-identified SVGs. These results confirm that iIMPACT-defined spaDEGs are more closely aligned with established biological knowledge and display a more pronounced spatial expression pattern.

Application to human prostate cancer dataset

To evaluate the performance of iIMPACT in different tissue types, we studied another SRT dataset from a human prostate cancer study, which includes 4371 spots and 17,651 genes. The gene expression was measured on a section from invasive carcinoma of the human prostate via the 10x Visium platform. We applied HD-Staining to analyze the histology image of this tissue (Fig. 3A). 259,257 cells were segmented and classified into six categories: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (detailed information in Additional file 1: Fig. S6).

Jiang *et al. Genome Biology* (2024) 25:147

Page 8 of 25
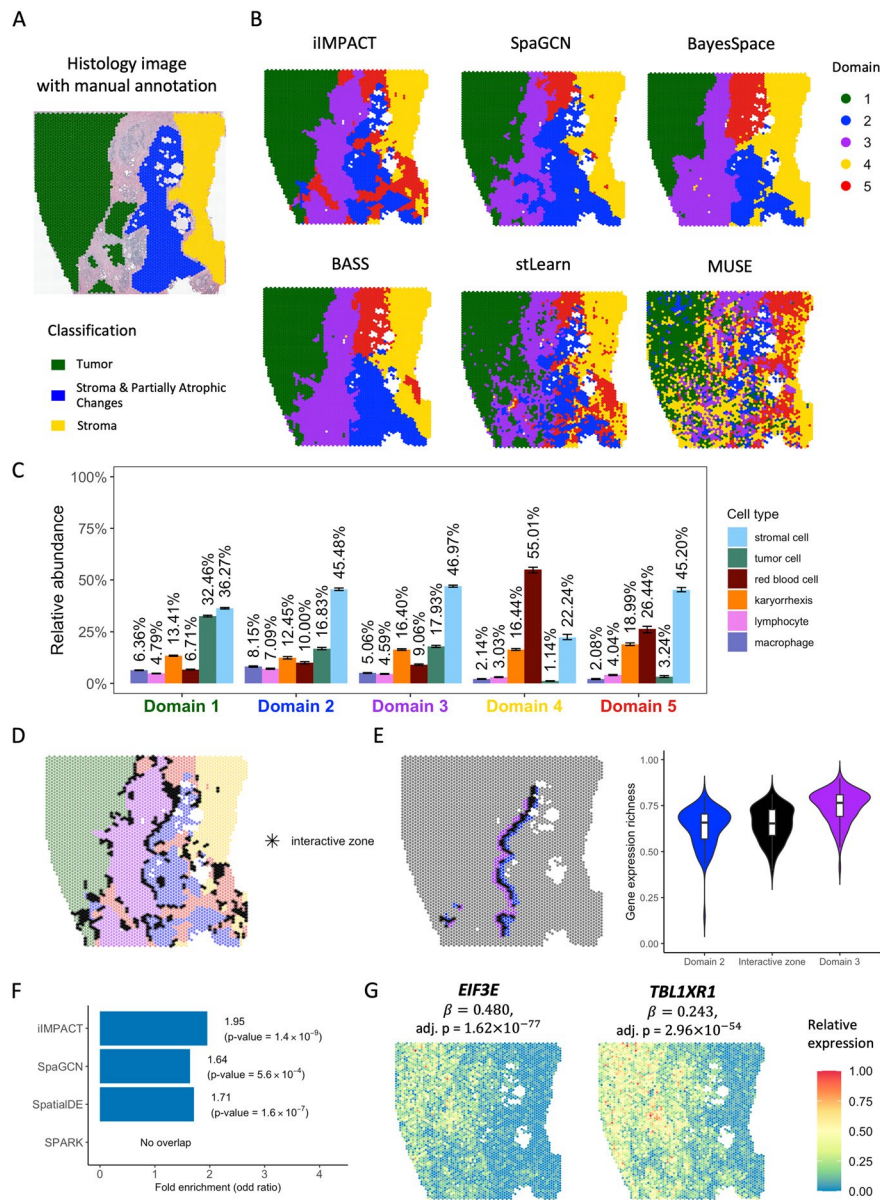
Fig. 3 Human prostate cancer dataset: **A** H&E-stained image of the tissue section with spot-level manual annotation from pathologists. **B** Spatial domains detected by iIMPACT, SpaGCN, BayesSpace, BASS, stLearn, and MUSE, setting the number of clusters to be five. **C** Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. **D** Interactive zones (black asterisk spots) defined by iIMPACT. **E** Identified interactive zones (black asterisk spots) and other boundary areas of domain 2 and domain 3, and boxplots of gene expression richness for spots in the interactive zone and other boundaries. **F** Gene enrichment analysis between genes detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known prostate cancer genes from the COSMIC database. **G** Spatial expression patterns of two example spaDEGs, *EIF3E* and *TBL1XR1*, that were only detected by iIMPACT

iIMPACT identified spatial domains that align more closely with the manual annotation than other methods. As shown in Fig. 3B, when the number of spatial domains *K* set to 5 suggested by the integrated completed likelihood (ICL) plot (Additional file 1: Fig. S7), these six methods could identify the domain (marked in green) with

Jiang *et al. Genome Biology* (2024) 25:147

Page 9 of 25

a high proportion of tumor cells, compared with the spatial distribution of tumor cells (Additional file 1: Fig. S6C). Interestingly, iIMPACT could distinguish histology-based spatial domains with different red blood cell proportions (Fig. 3B, yellow region vs. red region). We confirmed that iIMPACT outperformed other methods in spatial domain identification, given that there are three morphologically distinguished spatial domains: tumor, stroma and partially atrophic changes, and stroma (Additional file 1: Fig. S8). We observed that iIMPACT achieved the highest consistency with the manual annotation (ARI = 0.659). Additional file 1: Fig. S9 displays the spatial domains identified across the settings of the number of domains K from 2 to 8, and Additional file 1: Fig. S4 shows the corresponding ARI comparisons among different methods.

To demonstrate the interpretability of iIMPACT, we characterized the domain-specific relative abundance of cell types in Fig. 3C. We observed that domain 1 has a higher proportion of tumor cells than other domains, indicating that it is probably the tumor domain. Comparing domain 2 with domain 3, we observed that they have a similar proportion of tumor cells, but domain 2 has a higher proportion of immune cells (i.e., lymphocyte and macrophage), implying the heterogeneity of immune composition within tumors.

In addition, interactive zones can also be defined by iIMPACT (Fig. 3D). By checking the interactive zones of domains 2 and 3 and calculating the gene expression richness, we observed a clear trend between the interactive zones and the surrounding boundaries, indicating the unique characteristics of interactive zones (Fig. 3E). We further found that gene *DNAJC5* [39] expressed higher on the identified interactive zones, implying its potential relationship with the intermediate areas of immune cell distribution.

We also compared iIMPACT, SpaGCN [19], SpatialDE [29], and SPARK [30] in detecting biologically meaningful genes in this prostate cancer dataset. We confirmed that, for tumor-domain (domain 1) specific spaDEGs, iIMPACT outperformed SpaGCN, SpatialDE, and SPARK in detecting known prostate cancer genes from the COSMIC database (Fig. 3F and Additional file 1: Table S2), illustrating that iIMPACT could detect spaDEGs that are biologically relevant. These iIMPACT-defined spaDEGs in tumor domains have experimental evidence to support their functional relevance to the development of prostate cancer. For example, as shown in Fig. 3G, *EIF3E*, which is associated with increased cell cycle progression and motility in prostate cancer [40], and *TBL1XR1*, which displays an oncogene role for prostate cancer cell proliferation [41]. Based on the calculation of Moran's I (Additional file 1: Fig. S5), genes detected by iIMPACT have strong spatial correlation, similar with SpaGCN, and higher than those detected by SpatialDE and SPARK.

Application to human ovarian cancer dataset

The third NGS-based SRT dataset is from a section of human ovarian tumor tissue. This dataset includes 3455 spots and 17,651 genes. The gene expression was measured on a section of serous papillary carcinoma from human ovarian via the 10x Visium platform, with the H&E-stained image shown in Fig. 4A, HD-Staining model segmented and classified 211,746 cells in six categories: macrophage, karyorrhexis, tumor cell, lymphocyte, red blood cell, and stromal cell (Detailed information in Additional file 1: Fig. S10). By utilizing the cell type abundance information from

Jiang *et al. Genome Biology* (2024) 25:147

Page 10 of 25
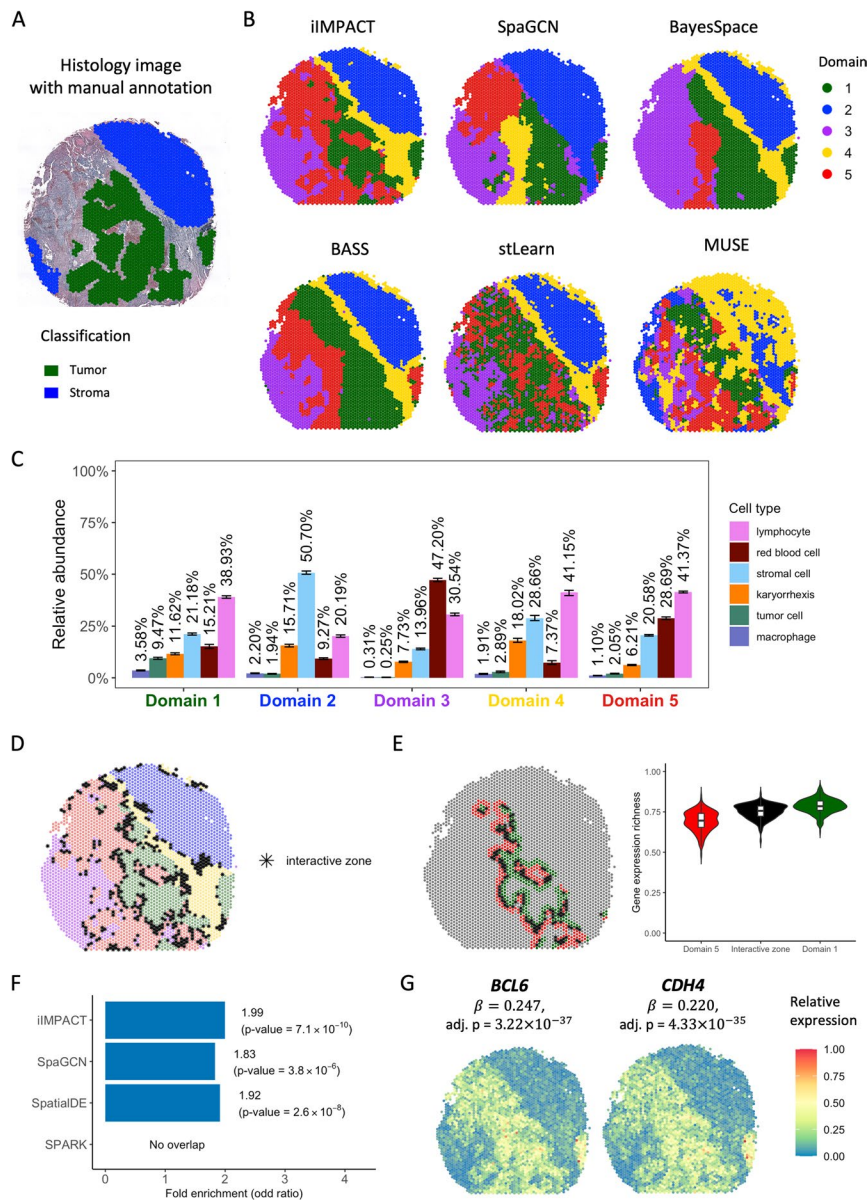
Fig. 4 Human ovarian cancer dataset: **A** H&E-stained image of the tissue section with spot-level manual annotation from pathologists. **B** Spatial domains detected by iIMPACT, SpaGCN, BayesSpace, BASS, stLearn, and MUSE, setting the number of clusters to be five. **C** Estimates (posterior means and credible intervals) of domain-specific relative abundance of cell types for the six cell types observed in the AI-reconstructed histology image. **D** Interactive zones (black asterisk spots) defined by iIMPACT. **E** Identified interactive zones (black asterisk spots) and other boundary areas of tumor domain and its adjacent domain 5, and boxplots of gene expression richness for spots in interactive zone and other boundaries. **F** Gene enrichment analysis between genes detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and the known ovarian cancer genes from the COSMIC database. **G** Spatial expression patterns of two example spaDEGs, *BCL6* and *CHD4*, that were only detected by iIMPACT

the histology image, we observed that iIMPACT had better performance on spatial domain identification. Setting the number of spatial domains to be 5, as suggested by the ICL plot (Additional file 1: Fig. S7), iIMPACT could identify the domain (Fig. 4B, domain marked in green) with a high proportion of tumor cells, which has a high

consistency with the tumor region annotated by the pathologist (Additional file 1: Fig. S11) and the region with a high amount of tumor cells (Additional file 1: Fig. S10C). By comparing the clustering results of six methods (iIMPACT, SpaGCN, BayesSpace, BASS, stLearn, and MUSE) with the annotated tumor and benign domains for this SRT dataset, we observed a remarkable concordance between the clustering results obtained from iIMPACT and the pathologist's annotations (ARI = 0.967, see Additional file 1: Fig. S11). Additional file 1: Fig. S12 shows the spatial domains identified across the settings of the number of domains K from 2 to 8.

iIMPACT could also distinguish domains with different red blood cell proportions. Figure 4C shows the estimation of the relative abundance of cell types for the five histology-based spatial domains. Domain 1 has a higher proportion of tumor cells than other domains, indicating that it is likely to be the tumor domain. We further examined the interactive zones (Fig. 4D) and compared the interactive zone between domains 1 and 5 with other boundary spots (Fig. 4E). A significant difference in gene expression richness between boundary spots and the interactive zone was observed. Furthermore, we found that gene *TTLL5* [42] and *CLEC12A* [43] have a higher expression on the interactive zone between domains 1 and 5, which may infer their potential relationship with the tumor-immune interaction.

We further detected spaDEGs using iIMPACT and then queried tumor-region spaDEGs with the known ovarian cancer gene set defined by the COSMIC database. We observed that iIMPACT-defined ovarian cancer spaDEGs showed a higher overlap with the known ovarian cancer gene set than that of SpaGCN, SpatialDE, and SPARK (Fig. 4F and Additional file 1: Table S2). Moreover, we explored these ovarian cancer spaDEGs only defined by iIMPACT and found that many of them possess compelling experimental evidence substantiating their functional relevance to ovarian cancer. For example, our list included *BCL6*, which displays pro-oncogenic activity in ovarian cancer [44], and *CHD4*, which is associated with apoptosis mediated by cisplatin in ovarian cancer cells [45] (Fig. 4G). Additional file 1: Fig. S5 illustrates that the top 1,000 spaDEGs identified by iIMPACT exhibit greater average Moran's I values than those identified by other methods, indicating a stronger spatial correlation in their expression patterns.

Application to mouse visual cortex STARmap dataset

To demonstrate that iIMPACT is also able to analyze data from imaging-based SRT platforms, we applied iIMPACT to a STARmap dataset [11]. This dataset was generated from mouse visual cortex, including the hippocampus, corpus callosum, and the neocortical layers. In total, 1020 genes were measured in 1207 cells with 15 cell types. The layer structure and cell type distribution of the tissue section provided by the original study are displayed in Fig. 5A.

As shown in Fig. 5A, iIMPACT displayed the second accurate clustering results with the known layer structure (ARI = 0.592). BASS is designed for single-cell-resolution SRT data, thus it had the best performance (ARI = 0.666). We also noticed that implementing iIMPACT on a lower resolution level (grids in Fig. 5A) might reduce the influence of noise, thus making the clustering result more robust. We also leveraged iIMPACT to identify the interactive zones (Fig. 5B). The majority areas of identified interacting areas were boundaries between two adjacent layers.

Jiang *et al. Genome Biology* (2024) 25:147

Page 12 of 25
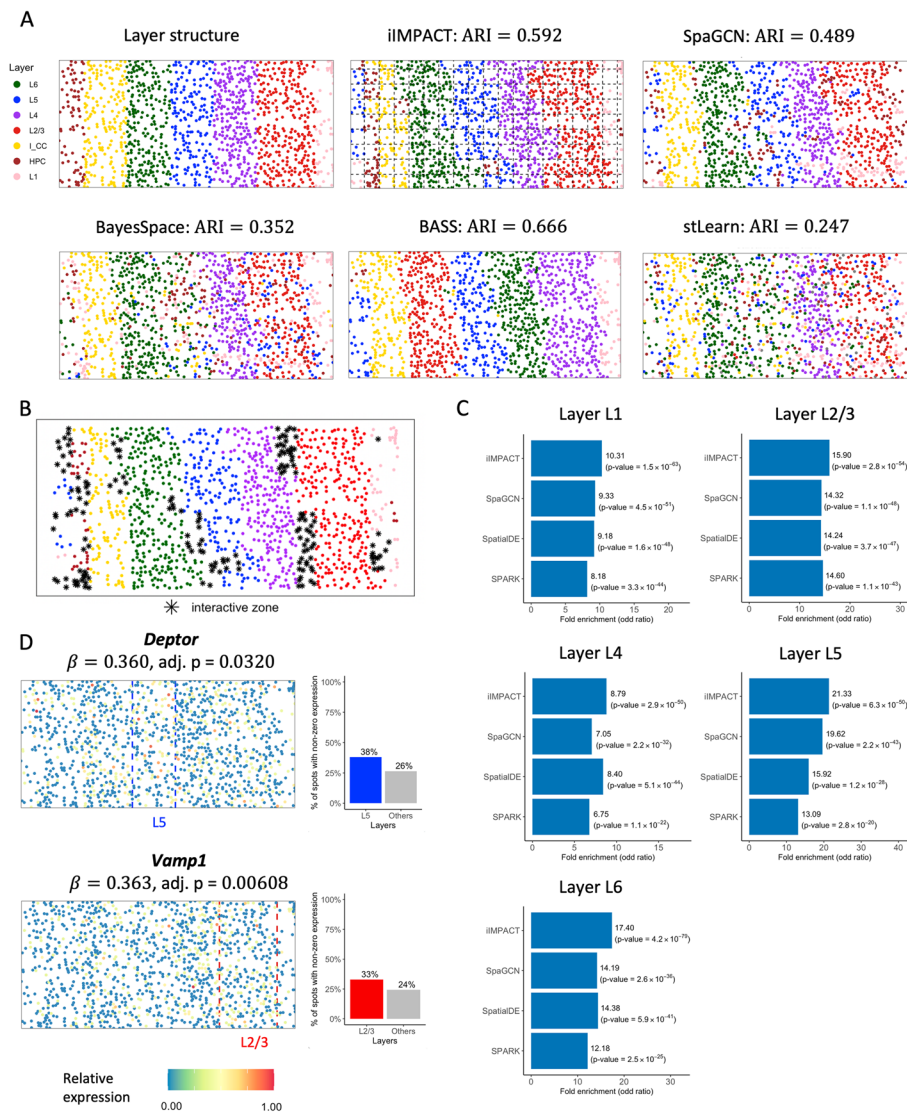
Fig. 5 Mouse visual cortex STARmap data: **A** Layer structure of the tissue section from the original study. Spatial domains detected by iIMPACT, SpaGCN, BayesSpace, BASS, and stLearn, setting the number of clusters to seven (the number of layers). Manually added square lattice grid when fitting iIMPACT is displayed with dashed lines. **B** Interactive zones (black asterisk spots) defined by iIMPACT. **C** Gene enrichment analysis between genes detected by iIMPACT, SpaGCN, SpatialDE, and SPARK, and genes functionally relevant to visual cortex for five layers. **D** Spatial expression patterns and barplots of proportion of non-zero expression of two example spaDEGs, *Deptor* and *Vamp3*, that were only detected by iIMPACT

We found these iIMPACT-defined spaDEGs are frequently functionally relevant to the visual cortex (Fig. 5C and Additional file 1: Table S2). For example, we observed *Deptor*, which is highly expressed and functions in a significant portion of corticostriatal and callosal neurons, located in the middle and superficial portions of layer 5 (L5) [46], and *Vamp1*, which is ubiquitously expressed and functioned in layer III pyramidal neurons in higher-order areas [47] (Fig. 5D). These two genes were detected by iIMPACT only.

Jiang *et al. Genome Biology* (2024) 25:147

Page 13 of 25

Discussion

In this paper, we presented iIMPACT, a two-stage statistical method that integrates histology images and molecular profiles. The first stage is a Bayesian finite normal-multinomial mixture model for identifying histology-based spatial domains. Numerous methods for spatial domain identification necessitate a dimensionality reduction step applied to the molecular profile, which compromises the clarity and direct interpretability of the identified spatial domains. However, iIMPACT fully leverages cellular-level information from histology images to improve clustering performance and increase interpretability. The cell type abundance data derived from HD-Staining encompasses comprehensive morphological details, such as cell growth pattern, cell–cell interaction, and cell interaction with the surrounding microenvironment, thereby improving the performance. On the other side, the latent spatial domain-specific relative abundance of cell types parametrized by iIMPACT offers a straightforward and user-friendly approach to define and characterize the identified spatial domains. The second stage is a NB regression model for detecting domain-specific spaDEGs. From both the simulation study (details in Additional file 1: Section S2) and real data analysis, we demonstrated that iIMPACT had higher accuracy in identifying spatial domains than published state-of-the-art methods due to the integration of histopathology images in iIMPACT. In addition, iIMPACT is versatile in analyzing both NGS-based and imaging-based SRT techniques, and therefore have broad impacts in the SRT field. Furthermore, iIMPACT has good biological interpretability to characterize histology-based spatial domains. For example, the inferred domain-specific cell-type compositions are consistent with curated annotations, and the interactive zones emphasize the areas with highly heterogeneous cell-type composition and gene expression compared with surroundings. Compared with other SVG detection methods, iIMPACT-defined spaDEGs are more enriched of known functional genes, confirming that iIMPACT could provide a better understanding of both cellular spatial organization and functional gene landscape of developmental and diseased tissues. Last but not least, compared with other methods, we also confirmed that iIMPACT is computationally efficient (Additional file 1: Table S3).

In real data applications, we assessed the performance of spatial domain identification by measuring the consistency between the identified domains and the manual annotation provided by the pathologists. While we recognize that these manual annotations might not perfectly reflect the true segmentation of domains integrating both morphological and molecular information, using them as a benchmark remains a standard and widely accepted practice in spatial domain identification, as established by precedent in foundational work [17–19]. Moreover, iIMPACT, alongside other spatial domain identification methods primarily relying on molecular profiles, exhibits limited capability in characterizing regions that are histologically distinct but have similar or low-quality gene expression. For instance, for the human breast cancer dataset, none of the methods effectively distinguished the fat region (in blue) from the fibrous tissue (in red) as per the manual annotation. The constrained number of cells in fat tissue results in a limited amount of gene expression measured, leading to low-quality molecular profiles and consequently unsatisfactory performance in the identification of those domains.

iIMPACT, BASS, and BayesSpace utilize a Bayesian mixture model with a Markov random field model for the identification of smooth spatial domains based on the SRT

Jiang *et al. Genome Biology* (2024) 25:147

Page 14 of 25

molecular profile. Unlike BASS and BayesSpace omit complementary information from the paired histology image, iIMPACT integrates cell type abundance derived from the image as an additional component. This integration significantly improves the accuracy of spatial domain identification (comparison of ARIs in Additional file 1: Table S4) and enables the biological interpretability of these domains. Moreover, iIMPACT assumes the image and molecular profiles—specifically, cell type abundance and gene expression levels—to contribute to spatial clustering, with adjustable weighting to optimize results. In contrast, BASS models cell type composition as a hidden layer within its Bayesian hierarchical model, asserting a direct probabilistic link between gene expression features and latent cell types. Notably, BASS specializes in analyzing imaging-based SRT data, which typically achieve the single-cell resolution and supports multi-sample clustering, whereas iIMPACT conducts spatial domain identification at the spot resolution, rendering it more suitable for analyzing NGS-based SRT data.

Nuclei identification methods for histology image analysis exhibit several limitations that hinder their widespread applicability and accuracy. One primary challenge is their generalizability. Most deep-learning-based algorithms require model training on high-quality labeled data, making them less adaptable to varied datasets and potentially limiting their generalizability across different tissue types and staining techniques. Besides, the performance of these nuclei identification methods may decrease when handling overlapping nuclei, where segmentation becomes intricate due to the lack of clear boundaries. To address the limitations of existing nuclei identification methods and enhance the versatility of iIMPACT, we proposed an alternative approach for the data preparation outlined in Additional file 1: Fig. S13. When implementing iIMPACT on tissue sections where precise nuclei classification proves challenging using nuclei identification methods, we leveraged the outputs from deep-learning-based or statistical nuclei segmentation methods to derive the nuclei localization, enabling us to determine the number of nuclei in each spot. Many methods [48–50] exist for isolating cell nuclei across various tissue types without relying on manual labeling data for training. Subsequently, we recommend using reference-free cell-type deconvolution methods [51] to generate the cell type abundance table. This data preparation pipeline was applied to additional SRT data, the LIBD human dorsolateral prefrontal cortex (DLPFC) data generated via 10x Visium [52]. Notably, iIMPACT demonstrated superior performance in spatial domain identification under this alternative approach to generate the image profile (Additional file 1: Fig. S14). Details are introduced in Additional file 1: Section S3. We also validated this data preparation pipeline on the human breast cancer data, and iIMPACT achieved better performance than only utilizing the molecular profile for the human breast cancer dataset and similar performance for human prostate and ovarian cancer datasets, as shown in Additional file 1: Fig. S15–S17 and Table S5. To leverage the histology image more sufficiently and enhance the interpretability of identified domains, we suggest obtaining the image profile, i.e., the cell type abundance table, by conducting nuclei identification through HD-Staining for cancer tissues. While originally designed for lung cancer, HD-Staining has proven effective for breast, prostate, and ovarian cancers, as demonstrated in our study, indicating its broader utility. The alternate data preparation pipeline should be reserved for instances where HD-Staining is less effective, such as with non-cancerous tissue sections.

There are several important future extensions for iIMPACT. First, improvement of nuclei segmentation and classification methods might further improve the performance of iIMPACT and therefore will be our focus in the near future. Second, the number of histology-based spatial domains has to be pre-specified when implementing the current version of iIMPACT. To automatically estimate the number of spatial domains, we plan to replace the proposed Bayesian finite mixture model with a Bayesian nonparametric model, such as the Dirichlet process mixture model [53] or a mixture of finite mixture model [54, 55]. Third, iIMPACT's performance in spatial domain identification was less satisfactory when dependent solely on histological image profiles, as shown in Additional file 1: Fig. S15–S17. This may be due to the extensive cell-type heterogeneity within domains, exemplified in Additional file 1: Fig. S18 for the human breast cancer dataset. Thus, integrating molecular information is crucial for effective spatial clustering. However, further investigation into better utilization of image profiles is also warranted. For instance, cell–cell interaction information can be incorporated into iIMPACT to improve the accuracy of histology-based spatial domain identification and increase the model interpretability. These future directions could potentially further boost the performance and interpretability of iIMPACT.

Conclusions

In conclusion, we have introduced iIMPACT, a multi-stage method that integrates histology image and spatial transcriptomics data to identify histology-based spatial domains and detect spatial domain-specific differentially expressed genes. Compared with existing methods, iIMPACT improves spatial domain identification accuracy and enhances biological interpretability by leveraging cellular-level information from AI-reconstructed histology images, and identifies spaDEGs enriched with known functional genes, making it a powerful tool for spatial transcriptomics analysis.

Methods

In this section, we first define the molecular and geospatial profiles from NGS-based SRT data (e.g., spatial transcriptomics and the improved $10 \times$ Visium platform) and the image profile from the matching AI-reconstructed histology image. Then we discuss how to construct the corresponding profiles from imaging-based SRT (e.g., STARmap) data. After that, we detail the statistical models used in the two stages of iIMPACT. Additional file 1: Table S6 summarizes all key notations introduced in this section.

Data preparation

Molecular profile Y

In general, the spot-level molecular profile of NGS-based SRT data can be represented by an $N \times P$ count table $\boldsymbol{C}$, where each entry $c_{ij} \in \mathbb{N}$, $i = 1, \cdots, N, j = 1, \cdots, P$ is the read count for gene j measured at spot i. To account for nuisance effects across spots, including sequencing depth, amplification and dilution efficiency, and reverse transcription efficiency, we normalize each read count c_{ij} to its relative level $\widetilde{c}_{ij} = c_{ij}/s_i$, where s_i is the total sum of counts across all genes at spot i, $s_i = \sum_{j=1}^{P} c_{ij}$, although other normalization methods are acceptable. Then, the relative gene expression $\widetilde{c}_{ij}$ are further log transformed to approximately conform to normality. Following the preprocessing steps

Jiang *et al. Genome Biology* (2024) 25:147

Page 16 of 25

in BayesSpace [17], we select the top 2000 most highly variable genes in terms of their relative expression and perform principal component analysis (PCA), or other dimension reduction techniques (e.g., t-SNE [56] or UMAP [57]), to obtain the low-dimensional representation of normalized gene expression denoted by an $N \times P'$ matrix Y, where each entry $y_{ij} \in \mathbb{R}$, $i = 1, \cdots, N, j = 1, \cdots, P'$ is the value of the j-th top principal component (PC) at spot i. We choose to model the PCs in Y rather than the raw count table C to avoid the use of complex finite mixture models with feature selection based on cumbersome multivariate distributions. Here, we recommend modeling the top three PCs ($P' = 3$) for simplicity. A sensitivity analysis on the human breast cancer data (see Additional file 1: Fig. S19) shows that larger P' only provided marginal improvements in clustering performance.

Image profile V

To integrate the image profile into iIMPACT, we applied a nuclei segmentation and identification algorithm, the histology-based digital (HD)-Staining model [23], to extract cellular features from images. The HD-Staining model is a trained deep-learning model implemented by the mask regional convolutional neural network (Mask R-CNN) architecture [58] for the tumor morphological microenvironment to segment the nuclei of different types of cells, such as immune, tumor, and stromal cells. The model was first trained using histology images from lung adenocarcinoma patients in the National Lung Screening Trial study, which has nuclei of six different cell types manually labeled by pathologists. Although the model was originally trained by lung cancer data, it has been improved and verified to be widely adapted to histology image datasets with other cancer types, such as breast cancer, head and neck cancer, ovarian cancer, prostate cancer, and other carcinomas.

The HD-Staining model takes a batch of high-resolution histology image patches of a tissue section as input and simultaneously segments and classifies cell nuclei on this image patch. It provides the locations and types for all identified nuclei in the whole histology image. To match the molecular information measured at spots, which only take less than half area (e.g., the area of all spots in 10x Visium platform is about 38% of the whole domain area), we count cells with different types within each spot and its expanded area (see Additional file 1: Fig. S20) so that all the cellular information can be utilized. The result is summarized into an $N \times Q$ count matrix V, namely cell abundance table, where each entry $v_{iq} \in \mathbb{N}$, $i = 1, \cdots, N, q = 1, \cdots, Q$ is the number of cells with type q observed at spot i and its expanded area. iIMPACT leverages the single-cell level histology information from the image profile to enhance spatial domain identification.

Geospatial profile G

Spots are the round area of barcoded mRNA capture probes where gene expression is measured. The spatial distribution of spots is arrayed on a square or triangular lattice. We denote the SRT geospatial profile by an $N \times 2$ matrix T, where each row $t_i = (t_{i1}, t_{i2})$ gives the x and y coordinates of the spot i on a two-dimensional Cartesian plane. ST and $10 \times$ Visium spots are arranged on square and triangular lattice grids, respectively. Thus, defining a neighborhood structure provides an alternative way to represent the geospatial profile G. In particular, G is an $N \times N$ binary adjacent matrix, where each entry $g_{ii'} = 1$

if spot i and i' are neighbors (i.e., the Euclidean distance $\sqrt{(t_{i1} - t_{i'1})^2 + (t_{i2} - t_{i'2})^2}$ between spot i and i' is less than a threshold) and $g_{ii'} = 0$ otherwise. Note that each diagonal entry $g_{ii'}$ is equal to zero. There are four and six neighbors for each non-boundary spot from the ST and $10 \times$ Visium platforms, respectively. With this neighborhood structure G as our geospatial profile, the spatial information can be easily integrated into Bayesian cluster analysis via an appropriate prior setting.

Special handling to imaging-based SRT data

Imaging-based SRT techniques usually have a higher spatial resolution than NGS-based SRT techniques, which is capable of profiling mRNA at the single-cell level. Data from some imaging-based platforms might provide the spatial distribution and types of cells on the tissue section in the original study. To fit iIMPACT to imaging-based SRT data such as STARmap [11], we manually add a square lattice grid with appropriate size to the whole domain and consider each square unit as a spot (see Fig. 5A). Note that those 'spots' fill the whole domain; thus, there is no gap between two adjacent spots. For STARmap data in the RESULTS section, the grid size was chosen to be 750×750 pixels, resulting in $N = 170$ spots. Each non-boundary spot has four neighboring spots. We define G with each entry $g_{ii'} = 1$ if spot i and i' are neighbors. To construct the molecular profile Y, we first normalize, transform, and reduce the dimension of the gene expression counts at the single-cell level, and then average the resulting values across all cells within each spot. To obtain the "image" profile V, we directly count the cells with different types in each spot.

Stage I: a Bayesian normal-multinomial mixture model for identifying histology-based spatial domains

The first stage of iIMPACT is to use a Bayesian finite mixture model to partition the whole domain into K mutually exclusive histology-based spatial domains. In general, a finite mixture model [59] generates random variables from a weighted sum of K independent distributions that belong to the same parametric family,

$$f(x_i | z, \theta_1, \cdots, \theta_K) = \sum_{k=1}^{K} \pi(z_i = k) f_k(x_i | z_i = k, \theta_k),$$

where $z = (z_1, \cdots, z_N)^{\mathrm{T}}$ denotes the latent variables specifying the identity of the mixture component f_k, characterized by θ_k, to each observation x_i. In the context of this paper, $x_i = \left\{ y_i \in \mathbb{R}^{P'}, v_i \in \mathbb{N}^Q \right\}$ represents the observed molecular and image profiling data, and $z_i = k$ indicates that spot i belongs to histology-based spatial domain k. Since there are two modalities Y and V, we decompose the mixture component f_k into two sub-components described below. In addition, we incorporate the information from the geospatial profile G into the prior placed over the auxiliary variable z, encouraging the neighboring spots to be in the same histology-based spatial domain.

Modeling the molecular profile Y

We use a multivariate normal (MN) sub-component for modeling the low-dimensional gene expression y_i at spot i:

Jiang *et al. Genome Biology* (2024) 25:147

Page 18 of 25

$$y_i | z_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \text{MN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kP'})^\text{T}, \mu_{kp} \in \mathbb{R}$ is the domain-specific mean vector and $\boldsymbol{\Sigma}_k$ is the $P' \times P'$ domain-specific variance–covariance matrix, requiring positive definiteness. For computational efficiency, we specify a normal prior for $\boldsymbol{\mu}_k$ conditional on $\boldsymbol{\Sigma}_k$, and an inverse-Wishart (IW) prior for $\boldsymbol{\Sigma}_k$, i.e., $\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim \text{MN}(\boldsymbol{v}_0, \boldsymbol{\Sigma}_k/\tau_0)$ and $\boldsymbol{\Sigma}_k \sim \text{IW}(\eta_0, \boldsymbol{\Phi}_0)$. This conjugate setting leads to analytically tractable posterior distributions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Here, $\boldsymbol{v}_0, \tau_0, \eta_0$, and Φ_0 are fixed hyperparameters. We set $\boldsymbol{v}_0$ to be the empirical mean vector over all spots and $\tau_0 = 0.01$ to provide a weak prior information so that the data itself would dominate the estimation of $\boldsymbol{\mu}_k$. We set the degree of freedom parameter $\eta_0 = P' + 1$, controlling the informative strength, and the scale matrix $\boldsymbol{\Phi}_0$ to be the identity matrix. Let $n_k = \sum_{i=1}^N \text{I}(z_i = k)$ and $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^N \text{I}(z_i = k)y_i$, the closed-form posterior distributions are $\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k, Y \sim \text{MN}(\boldsymbol{v}_k, \boldsymbol{\Sigma}_k/\tau_k)$ and $\boldsymbol{\Sigma}_k | Y \sim \text{IW}(\eta_k, \boldsymbol{\Phi}_k)$, where $\tau_k = \tau_0 + n_k, \quad \eta_k = \eta_0 + n_k, \quad \boldsymbol{v}_k = (\tau_0 \boldsymbol{v}_0 + n_k \bar{y}_k)/(n_k + \tau_0),$ $\boldsymbol{\Phi}_k = \boldsymbol{\Phi}_0 + \sum_{i=1}^N \text{I}(z_i = k)(y_i - \bar{y}_k)^\text{T}(y_i - \bar{y}_k) + \frac{n_k \tau_0}{\tau_0 + n_k}(\bar{y}_k - \boldsymbol{v}_0)^\text{T}(\bar{y}_k - \boldsymbol{v}_0).$

Suppose we choose PCA to perform an orthogonal projection of the scaled and normalized SRT molecular profiling data, we can further set all off-diagonal entries in $\boldsymbol{\Sigma}_k$ to be zero, i.e., $\sigma_{kpp'} = 0, \forall p \neq p'$. In this case, the multivariate normal model can be decomposed into a product of P' independent normal model,

$$y_i | z_i = k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \sim \prod_{p=1}^{P'} \text{N}(\mu_{kp}, \sigma_{kp}^2),$$

where $\boldsymbol{\sigma}_k^2 = (\sigma_{k1}^2, \ldots, \sigma_{kP'}^2)^\text{T}$ is the set of diagonal entries in $\boldsymbol{\Sigma}_k$. The conjugate setting for each dimension becomes a normal-inverse-gamma (IG) distribution [60], $\mu_{kp} | \sigma_{kp}^2 \sim \text{N}(0, \sigma_{kp}^2/\tau_0)$ and $\sigma_{kp}^2 \sim \text{IG}(v_0/2, \Phi_0/2)$, resulting in the closed-form posteriors $\mu_{kp} | \sigma_{kp}^2 \sim \text{N}(0, \sigma_{kp}^2/\tau_k)$ and $\sigma_{kp}^2 \sim \text{IG}(v_k/2, \Phi_k/2)$, where $\tau_k = \tau_0 + n_k, \eta_k = \eta_0 + n_k$, and $\Phi_k = \Phi_0 + \sum_{i=1}^N \text{I}(z_i = k)(y_{ip} - \bar{y}_{kp})^2 + \frac{n_k \tau_0}{\tau_0 + n_k}\bar{y}_{kp}^2$. One standard way of setting a weakly informative IG prior is to choose small values of both parameters, such as $v_0/2 = \Phi_0/2 = 0.1$.

Modeling the image profile V

We use a multinomial sub-component for modeling the number of cells with different types $\boldsymbol{v}_i$ within spot i and its expanded area:

$$\boldsymbol{v}_i | z_i = k, \boldsymbol{\omega}_k \sim \text{Multi}(m_i, \boldsymbol{\omega}_k),$$

where $m_i = \sum_{q=1}^Q v_{iq}$ is the total number of cells observed within the area and $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kQ})^\text{T}$ is defined on a Q-dimensional simplex (i.e., $\omega_{kq} > 0, \forall q$ and $\sum_{q=1}^Q \omega_{kq} = 1$), representing the underlying relative abundance of cell types in histology-based spatial domain k. Of particular note is that $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K$ are the parameters of key interest in iIMPACT, because it can be used to interpret or even define the identified histology-based spatial domains. For example, if a histology-based spatial domain is heavily dominated by cell type q, i.e., $\omega_{kq} \gg \omega_{kq'}, \forall q'$, then it could be named after cell type q. Note that cell type abundance is assumed to be homogeneous across the same

Jiang *et al. Genome Biology (2024) 25:147*

Page 19 of 25

histology-based spatial domain. For computational efficiency, we specify a Dirichlet prior setting for $\boldsymbol{\omega}_k$, i.e., $\boldsymbol{\omega}_k \sim \text{Dir}(\boldsymbol{\alpha}_0)$, where $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0Q})^\text{T}, \alpha_{0q} \in \mathbb{R}^+$ are fixed hyperparameters. This conjugate setting leads to an analytically tractable posterior distribution on $\boldsymbol{\omega}_k | V \sim \text{Dir}(\boldsymbol{\alpha}_k)$ with each entry $\boldsymbol{\alpha}_{kq} = \alpha_{0q} + \sum_{i=1}^{N} \text{I}(z_i = k)v_{iq}$. We recommend $\alpha_{01} = \cdots = \alpha_{0Q} = 1/2$ or 1 for a non or weakly informative setting.

Incorporating the geospatial profile G

To utilize the available spatial information in the geospatial profile, we employ a Markov random field prior [33, 34] on the histology-based spatial domain indicator $\boldsymbol{z}$, encouraging neighboring spots to be clustered into the same histology-based spatial domain:

$$\pi(z_i = k | \boldsymbol{z}_{-i}) \propto \exp(d_k + f\sum_{i'=1}^{N} g_{ii'}\text{I}(z_{i'} = k))$$

where $\boldsymbol{z}_{-i}$ denotes the set of all entries in $\boldsymbol{z}$ excluding the i th one, the hyperparameters $\boldsymbol{d} = (d_1, \cdots, d_N)^\text{T}$ control the number of spots belonging to each of the K histology-based spatial domains and $f \in \mathbb{R}^+$ controls the spatial dependency or smoothness. Note that if a spot has no neighbors, the above prior distribution reduces to a multinomial distribution, $z_i = k \sim \text{Multi}\left(N, \exp(\boldsymbol{d})/\sum_{k=1}^{K}\exp(d_k)\right)$. Although the larger the f, the smoother the pattern of spatial domains, careful determination of f is required. This is because a large value of f may lead to a phase transition problem (i.e., all spots are assigned to the same histology-based spatial domain). In this paper, we choose $d_1 = \cdots = d_K = 1$ and $f = 1$ by default, as this setting performs very well in the simulation study and yields reasonable results in our real data analysis.

Posterior sampling via MCMC algorithm

iIMPACT integrates the molecular, image, and geospatial profiles to partition the whole domain into K biologically meaningful spatial domains. Because the low-dimensional molecular profile Y and AI-reconstructed image profile V are generated from different sources, they are conditionally independent of each other. Thus, we define the mixture component

$$f_k(\boldsymbol{x}_i = \{\boldsymbol{y}_i, \boldsymbol{v}_i\} | z_i = k, \boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\omega}_k\}) = \text{MN}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\text{Multi}(\boldsymbol{v}_i; m_i, \boldsymbol{\omega}_k)^w$$

(1)

where the tuning parameter $w \in [0,1]$ controls the image profile's contribution to the clustering process, with respect to that of the molecular profile. Parameterizing the data likelihood above by decreasing w will result in a flatter multinomial distribution, thus downplaying the role of the image profile. When $w = 0$, iIMPACT will not depend on any cell type abundance information. We conducted a sensitivity analysis to search for the best choice of w. Our result suggests setting $w = 0.05$ and 0.5 for $10 \times$ Visium and STARmap data, respectively (see Additional file 1: Fig. S21). Note that in addition to the SRT platform and application, we should also consider the image and molecular profiles' dimensionalities (i.e., Q and $P\prime$) to determine the value of w with some degree of caution. Finally, we give the full posterior distribution as,

$$\pi\left(\boldsymbol{z}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_k, \boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_k | V, Y\right) \propto \pi(\boldsymbol{z}) \prod_{k=1}^{K} \mathrm{MN}\left(\boldsymbol{\mu}_k; \boldsymbol{v}_0, \frac{\boldsymbol{\Sigma}_k}{\tau_0}\right)$$

$$\mathrm{IW}(\boldsymbol{\Sigma}_k; \eta_0, \boldsymbol{\Phi}_0) \mathrm{Dir}(\boldsymbol{\omega}_k; \boldsymbol{\alpha}_0)^w \times \prod_{k=1}^{K} \prod_{\{i:z_i=k\}} \mathrm{MN}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \mathrm{Multi}(\boldsymbol{v}_i; m_i, \boldsymbol{\omega}_k)^w.$$

To identify histology-based spatial domains, the posterior distribution of z_i will be of direct interest to us, given by

$$\pi(z_i = k|\bullet) \propto \mathrm{MN}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \mathrm{Multi}(\boldsymbol{v}_i; m_i, \boldsymbol{\omega}_k)^w \times \exp\left(d_k + f \sum_{i'=1}^{N} g_{ii'} \mathrm{I}(z_{i'} = k)\right).$$

The individual quantities of all possible values of z_i are first computed and then summed to find the normalization constant $e = \sum_{k=1}^{K} \pi(z_i = k|\bullet)$. A new value of z_i can be drawn from a multinomial distribution $\mathrm{Multi}\left(1, (\pi(z_i = 1|\bullet)/e, \cdots, \pi(z_i = K|\bullet)/e)^{\mathrm{T}}\right)$. For any particular domain-specific parameters, i.e., $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\omega}_k$, we only require the partial data likelihood in estimating its posterior density as detailed before. Since the posterior conditional distributions for all parameters are in closed form, it is straightforward to use a Gibbs sampler, a type of Markov chain Monte Carlo (MCMC) algorithm, to obtain a sequence of observations approximated from the multivariate distribution $\pi\left(\boldsymbol{z}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_k, \boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_k | V, Y\right)$ (details in Additional file 1: Section S4). Consequently, the posterior inference can be made by post-processing the MCMC samples, such as $\left\{\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(U)}\right\}$ and $\left\{\boldsymbol{\omega}_k^{(1)}, \cdots, \boldsymbol{\omega}_k^{(U)}\right\}$, where u indexes the MCMC iteration and U is the total number of iterations after burn-in.

In any finite mixture model, the invariance of the likelihood under permutation of the cluster labels $\boldsymbol{z}$ may result in an identifiability problem, leading to symmetric and multimodal posterior distributions with up to $K!$ copies of each genuine model. What is worse, it will also complicate inference on other parameters. To address this issue, we impose an order restriction on the posterior samples of parameters $\boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_k$ based on a given cell type q. In particular, at each iteration u, we relabel $\boldsymbol{z}$ and switch all the related domain-specific parameters of the MCMC outputs to satisfy the constraint $\omega_{kq}^{(u)} > \omega_{k'q}^{(u)}$ for cluster indicator $k < k'$. In other words, the first histology-based spatial domain has the largest proportion of cell type q, while histology-based spatial domain K has the small proportion of cell type q.

Identifying histology-based spatial domains and interactive zones

Our primary interest lies in identifying histology-based spatial domains via making inferences on the spatial domain indicator vector $\boldsymbol{z}$. Here we apply the mode estimates [61] based on the marginal probabilities $\pi(z_i = k|\cdot) \approx \frac{1}{U} \sum_{u=1}^{U} \mathrm{I}(z_i^{(u)} = k)$. The estimate of $\widehat{z}_i$ can be obtained by selecting the highest value:

$$\widehat{z}_i = \mathrm{argmax}_k \pi(z_i = k|\cdot) \approx \mathrm{argmax}_k \mathrm{I}\left(z_i^{(u)} = k\right).$$

Uncertainty quantification is one advantage of the proposed Bayesian finite mixture model. For example, if the marginal probability of assigning spot i to histology-based spatial domain k is considerably high, e.g., $\pi(z_i = k|\cdot) \geq 0.9$, then we are confident about the assignment. However, if some marginal probabilities are almost equivalent or there is no significant mode for a spot, e.g., $\pi(z_i = k|\cdot) < 0.9, \forall k$, then we tend not

Jiang *et al. Genome Biology* (2024) 25:147

Page 21 of 25

to assign the spot to any histology-based spatial domains. Instead, we define the spot as the boundary spot, and the resulting connected area as the interactive zone.

Interpreting and defining histology-based spatial domains

The domain-specific relative abundance of cell types $\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K$ are another group of parameters of interest in our model, because it can be used to interpret or even define the identified histology-based spatial domains. We use the posterior mean as the estimate,

$$\widehat{\omega}_{kq} = \mathrm{E}\left[\omega_{kq}|\cdot\right] \approx \frac{1}{U}\sum\nolimits_{u=1}^{U} \omega_{kq}^{(u)},$$

averaging over all its MCMC samples. Additionally, the credible interval for each ω_{kq} can be approximated by its post-burn-in MCMC sample quantiles. Note that the MCMC samples can also be used to approximate any other quantity of interest that analytical solution is impossible, e.g., $\pi\left(\omega_{kq} > \omega_{k'q}|\cdot\right)$ for some k, k', and q.

Choosing the number of histology-based spatial domains K.

The number of histology-based spatial domain K can be determined by prior biological knowledge when available. In the absence of this information, we could apply the integrated completed likelihood (ICL) [62] as the criterion for selecting K. The ICL is calculated using the following:

$$\mathrm{ICL}(K) = -2\log\left(\mathrm{L}\left(Y, V, \widehat{z}|\widehat{\boldsymbol{\mu}}_1, \cdots, \widehat{\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}_1, \cdots, \widehat{\boldsymbol{\Sigma}}_K, \widehat{\boldsymbol{\omega}}_1, \cdots, \widehat{\boldsymbol{\omega}}_K\right)\right) + d\log(n),$$

where $\mathrm{L}\left(\boldsymbol{Y}, \boldsymbol{V}, \widehat{\boldsymbol{z}}|\widehat{\boldsymbol{\mu}}_1, \cdots, \widehat{\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}_1, \cdots, \widehat{\boldsymbol{\Sigma}}_K, \widehat{\boldsymbol{\omega}}_1, \cdots, \widehat{\boldsymbol{\omega}}_K\right)$ is complete data likelihood, i.e., the product of Eq. (1) over i, and $d = 2KP' + K(Q-1)$ is the total number of model parameters.

Stage II: a generalized linear regression model for detecting spaDEGs

To test if each gene is differentially expressed among those identified histology-based spatial domains in Stage I of iIMPACT, we use a generalized linear regression model, where the response variable is gene expression counts, and the predictor variables are the histology-based spatial domain indicators. In particular, we assume that all read counts from a gene j across different spots indexed by i are from an NB distribution:

$$c_{ij} \sim \mathrm{NB}\left(s_i\lambda_{ij}, \psi_j\right), i = 1, \cdots, N$$

where s_i is the size factor of spot i, ψ_j is the over-dispersion parameter of gene j, and λ_{ij} is the underlying normalized expression level for gene j at spot i. We further use the canonical link,

$$\lambda_{ij} = \alpha_{j,k} + \beta_{j,k}x_{i,k},$$

which is typically used in the Poisson and NB regression models. Here, $x_{i,k} = \mathrm{I}(z_i = k)$ is a binary indicator. If spot i is assigned to histology-based spatial domain k in Stage I of iIMPACT, then $x_{i,k} = 1$; otherwise, $x_{i,k} = 0$. Thus, we can interpret the intercept $\alpha_{j,k}$ as the baseline expression level of gene j in the whole domain

Jiang *et al. Genome Biology* (2024) 25:147

Page 22 of 25

excluding histology-based spatial domain k, and the slope $\beta_{j,k}$ as the differential expression level of gene j in histology-based spatial domain k as a shift from the baseline. With this modeling framework, spaDEGs, which are differentially expressed in a given histology-based spatial domain k compared with all other domains, can be identified via testing the null hypothesis $H_0 : \beta_{j,k} = 0$ versus the alternative $H_\alpha : \beta_{j,k} \neq 0$. For those genes whose resulting adjusted p-values are less than a significance level (e.g., 0.05), we define them as domain-k-specific spatially variable genes. To control the false discovery rate, the Benjamini and Hochberg method [63] needs to be applied to adjust p-values. The above NB regression model is fitted via the function *glm.nb* in the R package MASS [64].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03289-5.

Additional file 1. Supplementary materials for iIMPACT, including the supplementary sections, figures, and tables.

Additional file 2. Review history.

Availability of data and materials
The authors analyzed four publicly available SRT datasets. Raw count matrices, images, and spatial data for three SRT datasets from 10xVisium are accessible on the 10x Genomics website at https://support.10xgenomics.com/spatial-gene-expression/datasets. Mouse visual cortex STARmap data can be downloaded from https://www.starmapresources.com/data. Processed data of the analyzed four SRT datasets can be downloaded from Zenodo [65]. An open-source implementation of the iIMPACT algorithm in R/C++ is available at Github (https://github.com/Xijiang1997/iIMPACT) [66] and Zenodo [65].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Jiang *et al. Genome Biology* (2024) 25:147

Page 23 of 25

References

1. Asp M, Bergenstrahle J, Lundeberg J. Spatially resolved transcriptomes-next generation tools for tissue exploration. BioEssays. 2020;42:e1900221.
2. Burgess DJ. Spatial transcriptomics coming of age. Nat Rev Genet. 2019;20:317.
3. Zhang M, Sheffield T, Zhan X, Li Q, Yang DM, Wang Y, Wang S, Xie Y, Wang T, Xiao G. Spatial molecular profiling: platforms, applications and analysis tools. Brief Bioinform. 2021;22(3):bbaa145.
4. Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. Curr Opin Biotechnol. 2017;46:126–33.
5. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.
6. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363:1463–7.
7. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, Arlotta P, Macosko EZ, Chen F. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat Biotechnol. 2021;39:313–9.
8. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, Äijö T, Bonneau R, Bergenstråhle L, Navarro JF, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods. 2019;16:987–90.
9. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. In Nat Methods. 2014;11:360–1.
10. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348:aaa6090.
11. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400):eaat5691.
12. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;19:534–46.
13. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. Cancer Res. 2018;78:5970–9.
14. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
15. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20:273–82.
16. Zhu Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat Biotechnol. 2018;36(12):1183–90.
17. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SEB, Nghiem P, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol. 2021;39:1375–84.
18. Li Z, Zhou X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. Genome Biol. 2022;23:168.
19. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18:1342–51.
20. Pham D, Tan X, Xu J, Grice LF, Lam PY, Raghubar A, Vukovic J, Ruitenberg MJ, Nguyen Q. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. bioRxiv. 2020:2020.2005.2031.125658.
21. Bao F, Deng Y, Wan S, Shen SQ, Wang B, Dai Q, Altschuler SJ, Wu LF. Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. Nat Biotechnol. 2022;40:1200–9.
22. Tang Z, Li Z, Hou T, Zhang T, Yang B, Su J, Song Q. SiGra: single-cell spatial elucidation through an image-augmented graph transformer. Nat Commun. 2023;14:5618.
23. Wang S, Rong R, Yang DM, Fujimoto J, Yan S, Cai L, Yang L, Luo D, Behrens C, Parra ER, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. Cancer Res. 2020;80:2056–66.
24. Fox H. Is H&E morphology coming to an end? J Clin Pathol. 2000;53:38–40.
25. Cui Y, Zhang G, Liu Z, Xiong Z, Hu J. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. Med Biol Eng Comput. 2019;57:2027–43.
26. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imaging. 2018;37:2663–74.
27. Raza SEA, Cheung L, Shaban M, Graham S, Epstein D, Pelengaris S, Khan M, Rajpoot NM. Micro-Net: A unified model for segmentation of various objects in microscopy images. Med Image Anal. 2019;52:160–73.
28. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, Rajpoot N. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal. 2019;58:101563.
29. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15:343–6.

Jiang *et al. Genome Biology* (2024) 25:147

Page 24 of 25

30. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17:193–200.

31. Li Q, Zhang M, Xie Y, Xiao G. Bayesian modeling of spatial molecular profiling data via gaussian process. Bioinformatics. 2021;37(22):4129–36.

32. Jiang X, Xiao G, Li Q. A Bayesian modified Ising model for identifying spatially variable genes from spatial transcriptomics data. Stat Med. 2022;41:4647–65.

33. Clifford P. Markov random fields in statistics. Disorder in physical systems: A volume in honour of John M Hammersley. 1990. p. 19–32.

34. Morris R, Descombes X, Zerubia J. Fully Bayesian image segmentation-an engineering perspective. In Proceedings of International Conference on Image Processing. IEEE; 1997. pp. 54–57.

35. Neckmann U, Wolowczyk C, Hall M, Almaas E, Ren J, Zhao S, Johannessen B, Skotheim RI, Bjørkøy G, Ten Dijke P. GREM1 is associated with metastasis and predicts poor prognosis in ER-negative breast cancer patients. Cell Commun Signal. 2019;17:1–17.

36. Wang C, Lv J, Xue C, Li J, Liu Y, Xu D, Jiang Y, Jiang S, Zhu M, Yang Y, Zhang S. Novel role of COX6c in the regulation of oxidative phosphorylation and diseases. Cell Death Discov. 2022;8:336.

37. Gajulapalli VN, Samanthapudi VS, Pulaganti M, Khumukcham SS, Malisetty VL, Guruprasad L, Chitta SK, Manavathi B. A transcriptional repressive role for epithelial-specific ETS factor ELF3 on oestrogen receptor alpha in breast cancer cells. Biochem J. 2016;473:1047–61.

38. Li H, Calder CA, Cressie N. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geogr Anal. 2007;39:357–75.

39. Moses MA, Kim YS, Rivera-Marquez GM, Oshima N, Watson MJ, Beebe KE, Wells C, Lee S, Zuehlke AD, Shao H. Targeting the Hsp40/Hsp70 chaperone axis as a novel strategy to treat castration-resistant prostate cancer. Can Res. 2018;78:4022–35.

40. Gao Y, Teng J, Hong Y, Qu F, Ren J, Li L, Pan X, Chen L, Yin L, Xu D, Cui X. The oncogenic role of EIF3D is associated with increased cell cycle progression and motility in prostate cancer. Med Oncol. 2015;32:518.

41. Daniels G, Li Y, Gellert LL, Zhou A, Melamed J, Wu X, Zhang X, Zhang D, Meruelo D, Logan SK, et al. TBLR1 as an androgen receptor (AR) coactivator selectively activates AR target genes to inhibit prostate cancer growth. Endocr Relat Cancer. 2014;21:127–42.

42. Sand M, Bechara FG, Gambichler T, Sand D, Bromba M, Hahn SA, Stockfleth E, Hessam S. Circular RNA expression in cutaneous squamous cell carcinoma. J Dermatol Sci. 2016;83:210–8.

43. Chatterjee R, Chatterji U. CLEC12A: a promise target for cancer therapy. Arch Clin Med Case Rep. 2022;6:706–14.

44. Wang YQ, Xu MD, Weng WW, Wei P, Yang YS, Du X. BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer. Am J Cancer Res. 2015;5:255–66.

45. Oyama Y, Shigeta S, Tokunaga H, Tsuji K, Ishibashi M, Shibuya Y, Shimada M, Yasuda J, Yaegashi N. CHD4 regulates platinum sensitivity through MDR1 expression in ovarian cancer: A potential role of CHD4 inhibition as a combination therapy with platinum agents. PLoS One. 2021;16:e0251079.

46. Sorensen SA, Bernard A, Menon V, Royall JJ, Glattfelder KJ, Desta T, Hirokawa K, Mortrud M, Miller JA, Zeng H, et al. Correlated gene expression and target specificity demonstrate excitatory projection neuron diversity. Cereb Cortex. 2015;25:433–49.

47. Krienen FM, Yeo BT, Ge T, Buckner RL, Sherwood CC. Transcriptional profiles of supragranular-enriched genes associate with corticocortical network architecture in the human brain. Proc Natl Acad Sci U S A. 2016;113:E469-478.

48. Hayakawa T, Prasath VS, Kawanaka H, Aronow BJ, Tsuruoka S. Computational nuclei segmentation methods in digital pathology: a survey. Arch Comput Methods Eng. 2021;28:1–13.

49. Mouroutis T, Roberts SJ, Bharath AA. Robust cell nuclei segmentation using statistical modelling. Bioimaging. 1998;6:79–91.

50. Fatakdawala H, Xu J, Basavanhally A, Bhanot G, Ganesan S, Feldman M, Tomaszewski JE, Madabhushi A. Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. IEEE Trans Biomed Eng. 2010;57:1676–89.

51. Miller BF, Huang F, Atta L, Sahoo A, Fan J. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. Nat Commun. 2022;13:2339.

52. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, Catallini JL, Tran MN, Besich Z, Tippani M. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24:425–36.

53. Ferguson TS. A Bayesian analysis of some nonparametric problems. The annals of statistics. 1973. p. 209–30.

54. Miller JW, Harrison MT. Mixture models with a prior on the number of components. J Am Stat Assoc. 2018;113:340–56.

55. Hu G, Yang HC, Xue Y. Bayesian group learning for shot selection of professional basketball players. Stat. 2021;10:e324.

56. Van der Maaten L, Hinton G. Visualizing data using t-SNE. JMLR; 2008. p. 9.

57. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37:38–44.

58. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2961–9.

59. Kontkanen P, Myllymaki P, Tirri H. Constructing Bayesian finite mixture models by the EM algorithm. In: ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT. Citeseer; 1996.

60. Alvarez I, Niemi J, Simpson M. Bayesian inference for a covariance matrix. arXiv preprint arXiv:14084050 2014.

61. Li Q, Dahl DB, Vannucci M, Hyun J, Tsai JW. Bayesian model of protein primary sequence for secondary structure prediction. PLoS One. 2014;9:e109832.

62. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell. 2000;22:719–25.

Jiang *et al. Genome Biology* (2024) 25:147

Page 25 of 25

63. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57:289–300.
64. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB. Package 'mass.' Cran r. 2013;538:113–20.
65. Jiang X, Wang S, Guo L, Zhu B, Wen Z, Jia L, Xu L, Xiao G, Li Q. iIMPACT: integrating image and molecular profiles for spatial transcriptomics analysis. Zenodo. 2024. https://doi.org/10.5281/zenodo.11117768.
66. Jiang X, Wang S, Guo L, Zhu B, Wen Z, Jia L, Xu L, Xiao G, Li Q. iIMPACT: Integrating image and molecular profiles for spatial transcriptomics analysis. Github. https://github.com/Xijiang1997/iIMPACT.

Publisher's Note