Spatially resolved single-cell deconvolution of bulk RNA-seq using Bulk2Space

Jie Liao[1,4,†], Jingyang Qian[1,4,†], Yin Fang[2,3,†], Zhuo Chen[2,3,†], Xiang Zhuang[2,3,†], Ningyu Zhang[2,3], Xin Shao[1,4], Haihong Yang[2,3], Penghui Yang[1,4], Junyun Cheng[1,4], Yang Hu[1,4], Yining Hu[1], Lingqi Yu[1,4], Jinlu Zhang[1,2], Xiaoyan Lu[1,4], Li Shao[8], Dan Wu[9], Yue Gao[10,*], Huajun Chen[2,3,6,7,*], Xiaohui Fan[1,4,5*]

[1]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

[3]Hangzhou Innovation Center, Zhejiang University, Hangzhou 310058, China

[4]Innovation Center in Zhejiang University, State Key Laboratory of Component-Based Chinese Medicine, Hangzhou 310058, China

[5]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310058, China

[6]Alibaba-ZJU Frontier Technology Research Center, Hangzhou 310023, China

[7]The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310003, China

[8]Institute of Translational Medicine, The Affiliated Hospital of Hangzhou Normal University, Hangzhou Normal University, Hangzhou 310015, China

[9]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310013, China

[10]Department of Pharmaceutical Sciences, Beijing Institute of Radiation Medicine, Beijing, 100852, China.

*Corresponding Authors: Prof. Xiaohui Fan, e-mail: fanxh@zju.edu.cn, Prof. Huajun Chen, e-mail: huajunsir@zju.edu.cn, and Prof. Yue Gao, e-mail: gaoyue@bmi.ac.cn.

[†]These authors contribute equally to this work.

Abstract

Uncovering the tissue molecular architecture at single-cell resolution could help better understand the biological and pathological processes in organisms. However, bulk RNA-seq can only measure gene expression in cell mixtures, without revealing molecular heterogeneity and spatial patterns of single cells. Herein, we introduce Bulk2Space (https://github.com/ZJUFanLab/bulk2space), a deep learning framework-based spatial deconvolution algorithm that can simultaneously reveal the spatial and cellular heterogeneity of bulk RNA-seq data. The use of bulk transcriptomics to validate Bulk2Space revealed, in particular, the spatial variance of immune cells in different tumor regions, the molecular and spatial heterogeneity of tissues during inflammation-induced tumorigenesis, and spatial patterns of novel genes in different cell types. Moreover, Bulk2Space further annotated cell types that cannot be identified by original methods when it was used to spatially deconvolve the bulk data derived from our in-house developed sequencing approach Spatial-seq. Bulk2Space is distributed via an open-source Python software package.

Keywords

RNA-seq, Spatially resolved transcriptomics, Spatial deconvolution, Spatial-seq, scRNA-seq, Deep learning.

Main

Tissue complexity is portrayed by the spatial diversity and heterogeneity of cells[1]. Advances in spatially resolved transcriptomics[2-4] have made it possible to understand the cell composition, molecular architecture, and functional details of tissues at unanticipated spatial levels[5, 6]. State-of-the-art experimental technologies, including image-based methods[7-10], spatial barcoding RNA-seq methods[11-14], and laser capture microdissection-based methods[15, 16], have been developed to address either high throughput measuring of cells, unbiased detection of mRNA species, or single-cell resolution. To investigate the molecular variation during biological and pathological processes at a higher resolution, each sample is encouraged to be analyzed by spatially resolved single-cell transcriptomics, which is not yet fully achieved and is time-consuming, costly, and difficult to scale up[17]. Meanwhile, with investment for almost two decades, RNA-seq has been extensively applied in transcriptome analysis[18], with many large projects having been carried out, such as the Encyclopedia of DNA Elements (ENCODE)[19], The Cancer Genome Atlas (TCGA)[20], and projects of the International Cancer Genome Consortium (ICGC)[21]. A wealth of bulk RNA-seq data has become a legacy for biological and clinical research[22]. Thus, reanalysis of the enormous amount of bulk data to explain both cellular diversity and spatial expression patterns is a challenging but consequential task.

The *in silico* method has great potential to predict spatial heterogeneity from bulk RNA-seq data at single-cell resolution by integrating cutting-edge technologies. Several approaches such as CPM[23], CIBERSORT[24], and MuSiC[25] stopped short of extrapolating the proportion of cell types from bulk RNA-seq data yet failed to obtain single-cell data, let alone map them to tissue coordinates. The emerging of an approach that can efficiently convert bulk RNA-seq data into spatially resolved single-cell expression profiles is expected to reveal the cell diversity of complex tissue and the spatial expression variation simultaneously.

Herein, we introduce Bulk2Space, a spatial deconvolution method based on deep learning frameworks, which converts bulk transcriptomes into spatially resolved single-cell expression profiles using existing high-quality scRNA-seq data and spatial transcriptomics as references. We hypothesize that the process of scRNA-seq is similar to sampling cells from bulk tissue, and each

selected cell is labeled with a unique barcode. Consequently, bulk transcriptome data can be used as a weighted collection of single-cell expression data in a defined clustering space of cells. Bulk2Space first generates single-cell transcriptomic data within the clustering space to find a set of cells whose aggregated data are close to the bulk data. Next, the generated single cells were allocated to optimal spatial locations using a spatial transcriptome reference. For this step, we chose as a spatial reference one of the two most commonly used spatially resolved transcriptomics technologies. One is an image-based method with limited target genes[7-9], and the other is a spatial barcoding method without single-cell resolution[11-14]. Bulk2Space showed a robust performance across multiple datasets and conditions and is an open-access algorithm in GitHub (https://github.com/ZJUFanLab/bulk2space).

Results

Design of Bulk2Space

The overall design of the Bulk2Space algorithm is illustrated in Fig. 1, which is divided into deconvolution and spatial mapping. Although bulk data are obtained by sequencing a mixture of cells, whereas scRNA-seq labels individual cells, both methods share comparable cell types and states[23, 24, 26]. Meanwhile, the prevailing spatially resolved transcriptomics methods can provide reference coordinates for single cells based on their expression profiles[27, 28]. As a consequence, it could rationally conceive that the combination of single-cell profiles and spatial transcriptomics is expected to overcome technical bottlenecks and accomplish the spatially resolved single-cell deconvolution of bulk transcriptomes. We hypothesize that the process of single-cell sequencing is the process of sampling from tissue and that the ratio between different cell types is related to the zone of tissue, the ways of single-cell capture, and the size of the cell itself. However, even if the ratio of cell types changes, the state of each cell still fluctuates within a relatively stable high-dimensional space, which is namely the clustering space of cells. Several studies have backed up this hypothesis[29-31].

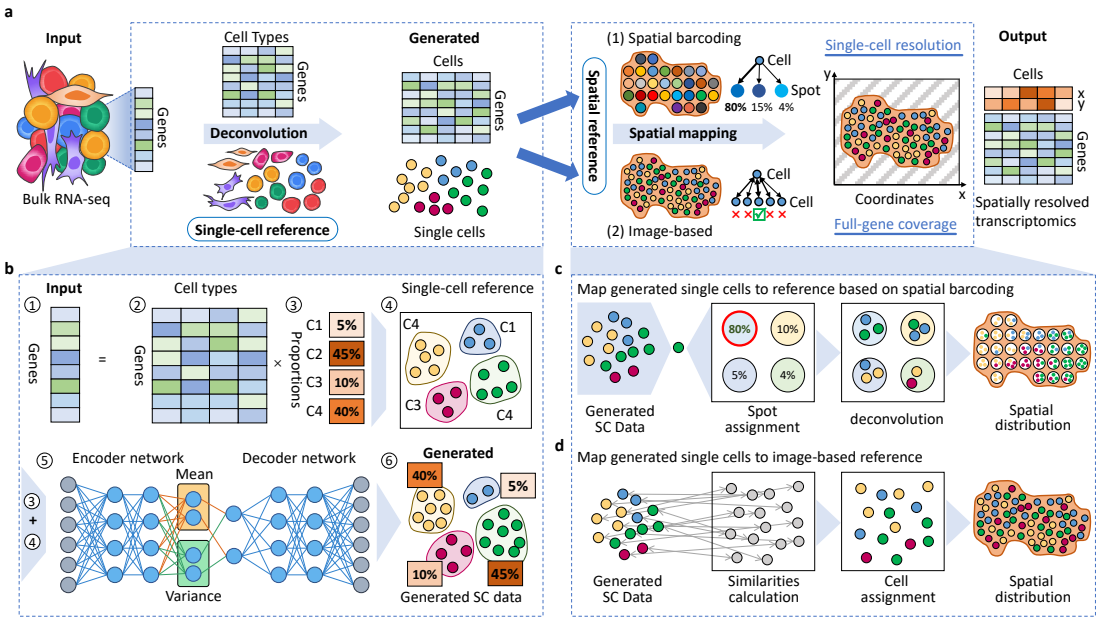

Fig. 1 Workflow of Bulk2Space. (a) Overview of the design concept of Bulk2Space. Bulk transcriptome data is taken as input, and a single-cell profile is used as the reference for characterizing the clustering space of the heterogeneous tissue. After deconvolution, input bulk data is transformed into single-cell transcriptomics. Then, either of the two spatially resolved transcriptomics, spatial barcoding RNA-seq or image-based in-situ hybridization, is used as the spatial reference. Generated single cells are assigned to the corresponding coordinates based on the spatial reference. The output is a set of deconvolved single-cell profiles with specified x and y spatial coordinates. **(b)** Detailed deconvolution procedure. The input vector of bulk tissue is a production of the expression matrix of cell types and the proportion vector of each cell type. The calculated proportion of all cell types is employed for the subsequent single cells generation. The single-cell reference is used to characterize the clustering space of the tissue, and a deep learning model generates single-cell profiles within the clustering space of each cell type. **(c)** and **(d)** demonstrate the strategies for spatial mapping based on the two mainly used spatially resolved transcriptomics approaches. **(c)** For spatial barcoding reference, each generated single cell is assigned to the spot with the highest gene expression correlation until the aggregation of cells within the spot is close enough to the exact expression value. **(d)** For image-based reference, each generated cell is assigned to the location where the cell on tissue has the highest similarity with the given cell.

For single-cell deconvolution, single-cell reference profiles can be used to characterize the clustering space and the average gene expression in each cell type[31, 32]. The expression vector of bulk transcriptome is taken as an input, which is a product of the average gene expression matrix

121 of each cell type and its abundance vector. By solving this nonlinear equation, the proportion of
122 each cell type is determined. Using the percentage of each cell type as a control parameter to
123 generate the number of single cells, a deep learning model, termed the constrained variational
124 autoencoder (β-VAE)[33], is utilized to randomly simulate a given number of single cells within the
125 clustering space of each cell type, thus transforming the heterogeneous bulk transcriptome into
126 single-cell transcriptome data (Fig. 1b).

127 For spatial mapping, the generated single-cell profiles are assigned to the reference tissue
128 section map based on either of the two most commonly used spatially resolved transcriptomics
129 approaches, namely, spatial barcoding or image-based targeted method. As shown in Fig. 1c, for
130 spatial barcoding references, each barcoded spot was regarded as a mixture of several cells. We
131 calculated the probability of each generated single cell in each spot based on its gene expression
132 and assigned this cell to the most likely spot until the recommended spot was saturated. Finally,
133 the cell composition of each spot was determined. For image-based references, simulated single-
134 cell data were mapped to the coordinates of the target cell according to the similarity in their gene
135 expression (Fig. 1d). Instead of only calculating the abundance values of cell types, as performed
136 by other methods, such as Cell Population Mapping (CPM)[23], CIBERSORT[24], and ImmuCC[26],
137 Bulk2Space can generate biologically feasible single-cell expression profiles and assign them
138 robustly to tissue coordinates, thus achieving spatially resolved single-cell deconvolution of the
139 bulk transcriptomics.

140

141 **Bulk2Space robustly deconvolves bulk transcriptomes**

142 To demonstrate the robustness of the Bulk2Space algorithm, a benchmark test was performed
143 using 30 single-cell transcriptome datasets (Supplementary Table 1) obtained from different tissues.
144 First, each single-cell transcriptome dataset was divided into two parts, one for reference and the
145 other for the synthesis of bulk data. The procedure for bulk data synthesis is shown in Fig. 2a. The
146 composition ratio of each cell type was first assumed, and then single-cell data were randomly
147 selected from the given dataset according to this ratio. This newly created set of cells can be
148 employed as the ground truth for subsequent evaluation and is aggregated to synthesize the input

149 bulk transcriptome data. Three generative deep learning models, namely, the constrained
150 variational autoencoder (β-VAE)[33], generative adversarial networks (GAN)[34], and conditional GAN
151 (CGAN)[35], were introduced here for the benchmarking of the 30 single-cell datasets. The gene
152 expression correlation between the generated single-cell data and input bulk data was calculated
153 to evaluate and compare the three candidate algorithms. As illustrated in Fig. 2b, β-VAE performed
154 better than the other algorithms in both Pearson and Spearman correlations with a lower root
155 mean squared error (RMSE). Two examples are shown in Fig. 2c. The first includes single-cell data
156 for human peripheral blood[36], which consists of four cell types. Bulk2Space was used to deconvolve
157 the synthetic data, and the results demonstrated that the clustering space of the generated single-
158 cell data was similar to that of the test data but varied in the proportions of different cell types.
159 Furthermore, the correlation analysis of the expression of marker genes for each cell type showed
160 that the generated and test data had a correlation over 0.9. Similar results were obtained for single-
161 cell data collected generated from the human lung[37] (Fig. 2c) and other datasets (Supplementary
162 Fig. 1). To further investigate the robustness of Bulk2Space, we introduced two noise mechanisms
163 to test the performance of the algorithm (Supplementary Fig. 2). One mechanism changes the
164 expression values of certain genes in randomly selected cells, and the other alters the cell type
165 labels of selected cells. Bulk2space showed robust performance and could effectively avoid the
166 over-fitting phenomenon.

167

168 **Bulk2Space maps single cells to spatial locations**

169 In the second step, Bulk2Space assigns the generated single cells to spatial coordinates based on a
170 spatially resolved transcriptome reference (Fig.2 d-i). There are two most common spatially
171 resolved transcriptomic techniques. One is the spatial barcoding technology, which includes
172 'spatial transcriptomics' (ST)[14], high-definition spatial transcriptomics (HDST)[11], and Slide-seq[12, 13].
173 These methods are based on tissue microregions with certain regular shapes (spots), and cannot
174 achieve single-cell resolution. The other is image-based *in situ* transcriptomics, including MERFISH[7],
175 seqFISH[8, 10], and STARmap[9]. These methods can only measure the expression of target genes but
176 cannot cover the whole transcriptome. To overcome the bottlenecks of the two approaches, we

177 designed corresponding spatial assignment strategies.

178 The first strategy was designed to map single cells to spots in spatially resolved transcriptomics
179 based on spatial barcoding. In this step, we used the synthesized data to test the performance of
180 Bulk2Space. The data synthesis procedure is illustrated in Fig. 2d. Similar to the commonly used
181 spatially resolved transcriptomics technologies, such as ST, HDST, 10X Visium, and Slide-seq, we
182 randomly selected certain single cells from the scRNA-seq data to constitute a spot and assigned it
183 a defined spatial coordinate in a virtual tissue with a reference system. The gene expression value
184 of a spot is the sum of the gene expression values of all cells within it. A comprehensive comparison
185 of six machine learning or deep learning approaches was then conducted using 40 datasets. These
186 methods were benchmarked for the optimization of cell type composition and spatial gene
187 expression patterns. As shown in Fig. 2e, compared with these methods, Bulk2Space showed a
188 higher correlation with gene expression and a lower RMSE. The cell composition and spatial
189 expression patterns predicted by Bulk2Space were highly correlated to the ground truth (Fig. 2f).

190 A scRNA-seq data (GSE113576) and MERFISH data (image-based reference) from the same
191 tissue[38] were used to test the performance of the Bulk2Space algorithm for the second spatial
192 mapping strategy. We wanted to find the shortest path that could shift single cells to optimal
193 locations. First, a nonparametric empirical Bayes network was used to eliminate the expression
194 differences between the single-cell data and spatially resolved transcriptome data. As illustrated in
195 Fig. 2g, the clustering spaces of the two datasets were very similar. Most cell types were matched
196 in both datasets. The two discrete datasets could be well merged in distribution, and the gene
197 expression of each cell type had a strong correlation between the two datasets (Fig. 2h). Of the
198 150 target genes identified in MERFISH, 149 were used as reference genes, and one was used for
199 validation. The results showed that spatially, the gene expression pattern of the matched cells was
200 highly correlated with that of the reference (Fig. 2i), and the predicted spatial expression patterns
201 across all 150 targeted genes are strongly correlated with the ground truth (Supplementary Fig. 3).

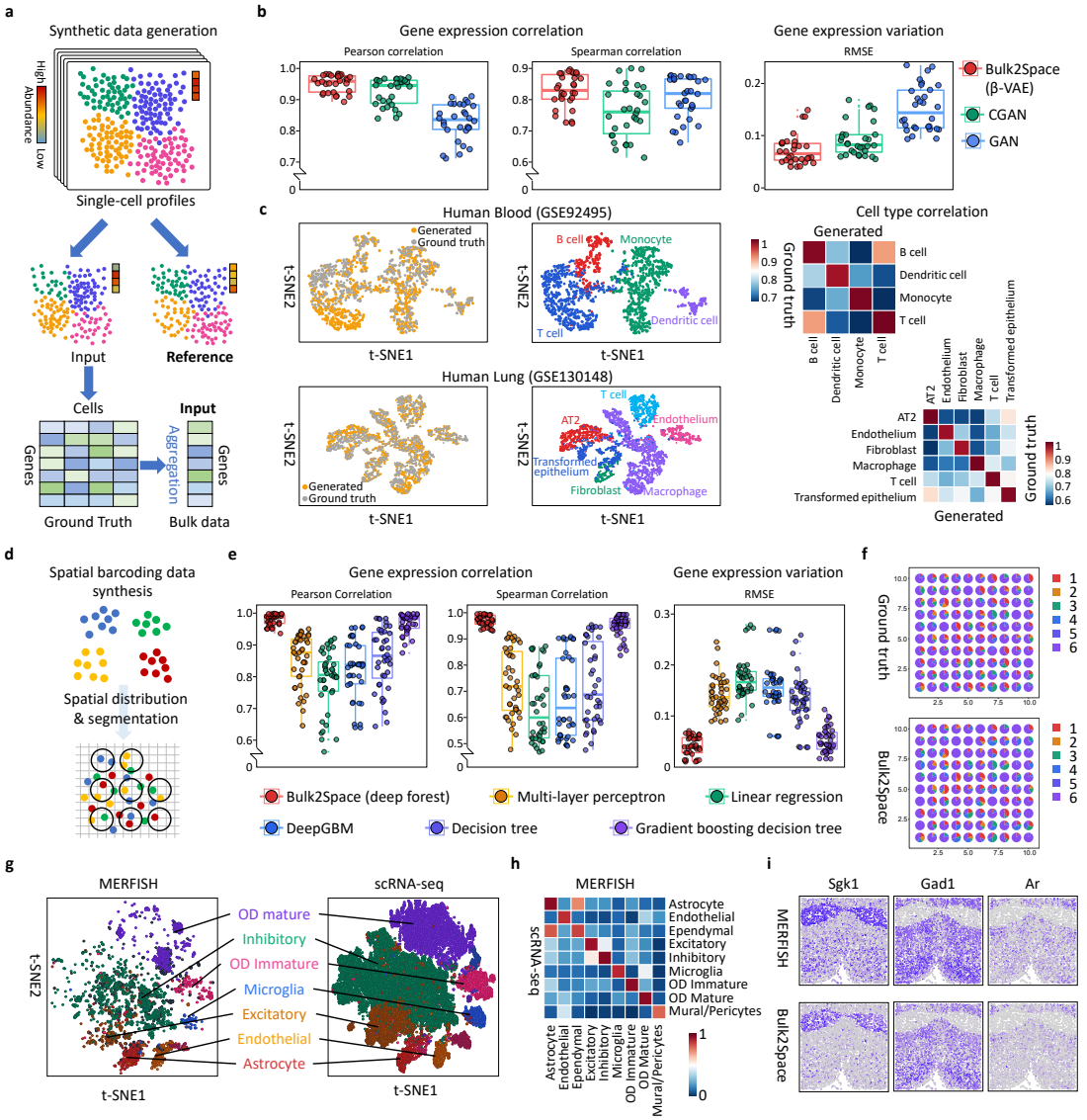

Fig. 2. Benchmark test for Bulk2Space. (a) Synthesis of the bulk transcriptome data. The color bar donated the abundance of distinct cell types. For a given single-cell profile and the specified composition proportions of cell types, single cells were randomly selected from each cell type to form a new dataset. The input bulk transcriptomics was retrieved by aggregating the generated single-cell profiles, which can be used as the ground truth for the performance evaluation of various methods. **(b)** Gene expression correlation (Pearson correlations, left, and Spearman correlations, middle) and root mean squared error (RMSE, right) of three methods, namely, Bulk2Space (β-VAE), GAN, and CGAN. Each point represented a data set. **(c)** Illustration of the deconvolution examples by Bulk2Space on two datasets, human peripheral blood (GSE92495) on the top and human lung (GSE130148) on the bottom. Left, distribution of generated (yellow) and test (grey) data. Middle, t-SNE layout of the datasets. Different colors represented distinct cell types. Right, Pearson expression correlation heatmap of different cell types for generated and test data. **(d)** Synthesis of the spatially resolved transcriptome for spatial barcoding methods. Single cells were randomly distributed on the virtual tissue section. Arrayed regions were uniformly segmented from the virtual tissue at a certain diameter (regarding the spatial resolution of

218 different technologies) to simulate existing spatial barcoding-based RNA-seq data, e.g., ST, HDST,
219 slide-seq, and 10X Visium, etc. Therefore, the specific cell composition of the generated spatial
220 reference was known. **(e)** Benchmark test for 6 spatial mapping methods (red, Bulk2Space (deep
221 forest); yellow, multi-layer perceptron; green, linear regression; blue, DeepGBM; indigo, decision
222 tree; and purple, gradient boosting decision tree). Left, box plot of gene expression correlation and
223 RMSE across 40 datasets for each method. **(f)** The layout of the cell compositions that were
224 predicted by Bulk2Space and the ground truth. Each pie chart exhibited the cell type proportions
225 in each spot. **(g)** The t-SNE layout of the MERFISH data (image-based spatially resolved
226 transcriptomics) and the single-cell data. Share cell types were colored with corresponding colors.
227 **(h)** Expression correlation heatmap of the marker genes of all shared cell types in MERFISH data
228 and scRNA-seq data. **(i)** Spatial expression patterns of genes in the ground truth and predicted by
229 Bulk2Space.

230

231 **Bulk2Space integrates spatial gene expression and histomorphology in pancreatic**
232 **ductal adenocarcinoma (PDAC)**

233 To further verify the performance of the Bulk2Space algorithm, it was applied to reveal the spatial
234 heterogeneity in bulk PDAC data. Moncada[39] et al. performed scRNA-seq and ST on PDAC tissues.
235 As shown in Fig. 3a, the scRNA-seq data were used as a deconvolution reference, and one piece of
236 a tissue section was used as a spatial reference (Slice 1). An additional slice (Slice 2) was used to
237 synthesize bulk data. After deconvolution, the bulk data were transformed into single-cell data.
238 There was a good correlation between the generated single-cell profiles and Slice 2 on the
239 expression of marker genes corresponding to each cell type (Fig. 3b). The deconvolved single cells
240 were then mapped to Slice 1 by Bulk2Space (Fig. 3c). To further correlate the spatial expression
241 pattern of the tissue with its histological feature, we investigated the molecular architecture of the
242 annotated regions in Slice 1 (Fig. 3d). As shown, the spatial distribution of cancer, endothelial, and
243 ductal cells predicted by Bulk2Space (Fig. 3e) was consistent with the histological annotation and
244 the spatially resolved transcriptomics of cancer, interstitium, and duct epithelium regions. This was
245 confirmed by the spatial expression of marker genes in these three histologically associated cell
246 types (Fig. 3f). The gene expression correlation of all six cell types between the predicted spatial
247 data and Slice 1 was exhibited in Fig. 3g.

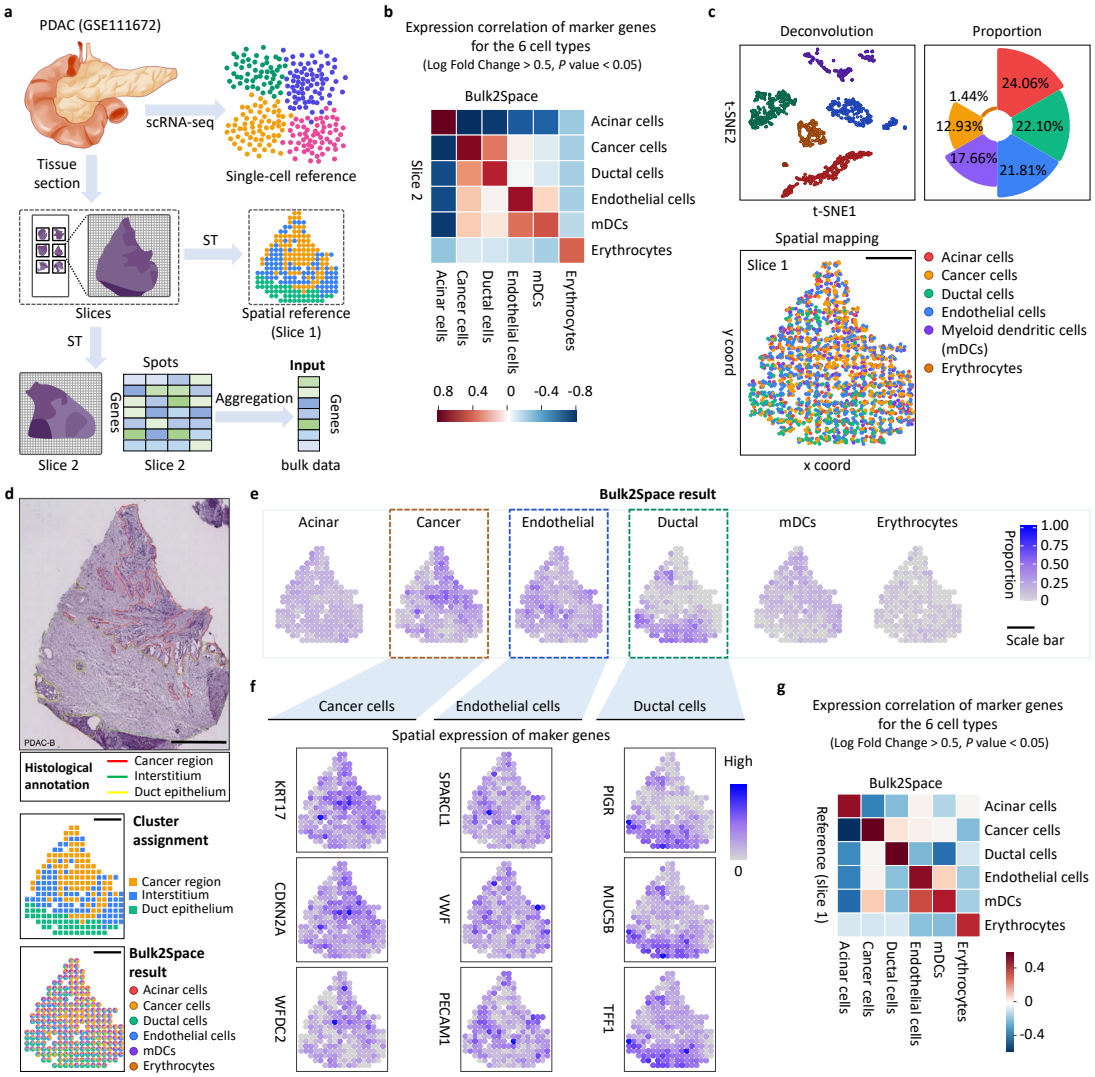

Fig. 3. Spatially resolved analysis of PDAC by Bulk2Space. (a) Synthesis of the bulk transcriptome data. The scRNA-seq data is used as the deconvolution reference, one slice of the sectioned tissue is employed as the spatial reference (Slice 1), and another slice (Slice 2) is used to synthesize the input bulk data. **(b)** The expression correlation of the marker genes in six cell types between the Bulk2Space result and Slice 2. **(c)** Single-cell deconvolution of the bulk data. Top, t-SNE plot showed the clustering of generated single cells and the rose chart showed the proportion of each predicted cell type. Bottom, distribution of predicted single cells. Different colors represented distinct cell types. **(d)** The top, Histological annotation for cancer region (red), interstitium (green), and duct epithelium (yellow). Middle, clusters of cancer region (yellow), interstitium (blue), and duct epithelium (green) from the ST data. Bottom, single-cell decomposition result of each spot in the ST data by Bulk2Space. **(e)** The spatial abundance of different cell types in each spot on the tissue section is predicted by Bulk2Space. **(f)** The spatial expression of the marker genes in cancer (left), endothelial (middle), and ductal (right) cells in predicted data. **(g)** The expression correlation of the marker genes in six cell types between the Bulk2Space result and Slice 1.

Bulk2Space reveals spatial, molecular, and functional heterogeneity of B cells in melanoma

Similarly, the same spatial deconvolution procedure of Bulk2Space was performed for another dataset deriving from Tirosh[40] et al., the melanoma bulk transcriptome data. The scRNA-seq data were used as a single-cell reference, one slice of tissue section was used as a spatial reference (Slice 1), and another (Slice 2) was used for bulk data synthesis (Fig. 4a). The expression correlation of marker genes for five cell types between Bulk2Space result and Slice 2 was shown in Fig. 4b. After deconvolution, single cells were assigned to Slice 1 by Bulk2Space (Fig. 4c). As illustrated in Fig. 4d, the composition of cell types in each spot of the tissue showed a strong spatial pattern, which was consistent with the histological annotation of the tissue section. The spatial distribution of each cell type is shown in Fig. 4e. The predicted results for B cells, cancer-associated fibroblasts (CAFs), macrophages, and malignant cells showed distinct spatial heterogeneity. Next, we compared the expression of marker genes in both predicted and reference tissues (Fig. 4f) and found that the spatial pattern of the marker genes in the predicted tissue was highly correlated with that in the spatial reference (Fig. 4g). Surprisingly, B cells assigned to the melanoma area were quite different from those assigned to the lymph node area (Fig. 4h), which suggested that there was heterogeneity in B cells from different tissue regions. However, this spatial difference cannot be detected by other deconvolution methods that only predict the composition of cell types in each spot. The differential expression analysis (Fig. 4i) and pathway enrichment (Fig. 4j) were carried out to further investigate the molecular and functional heterogeneity of B cells. The result indicated that B cells in the lymph node area were related to immunoregulation such as lymphocyte activation, T cell activation, and immune effector process, etc. However, in the melanoma region, many pathways, such as regulation of growth, cell-substrate junction, cell-substrate adhesion, and other pathways associated with cancer cell growth, adhesion, and diffusion were enriched.

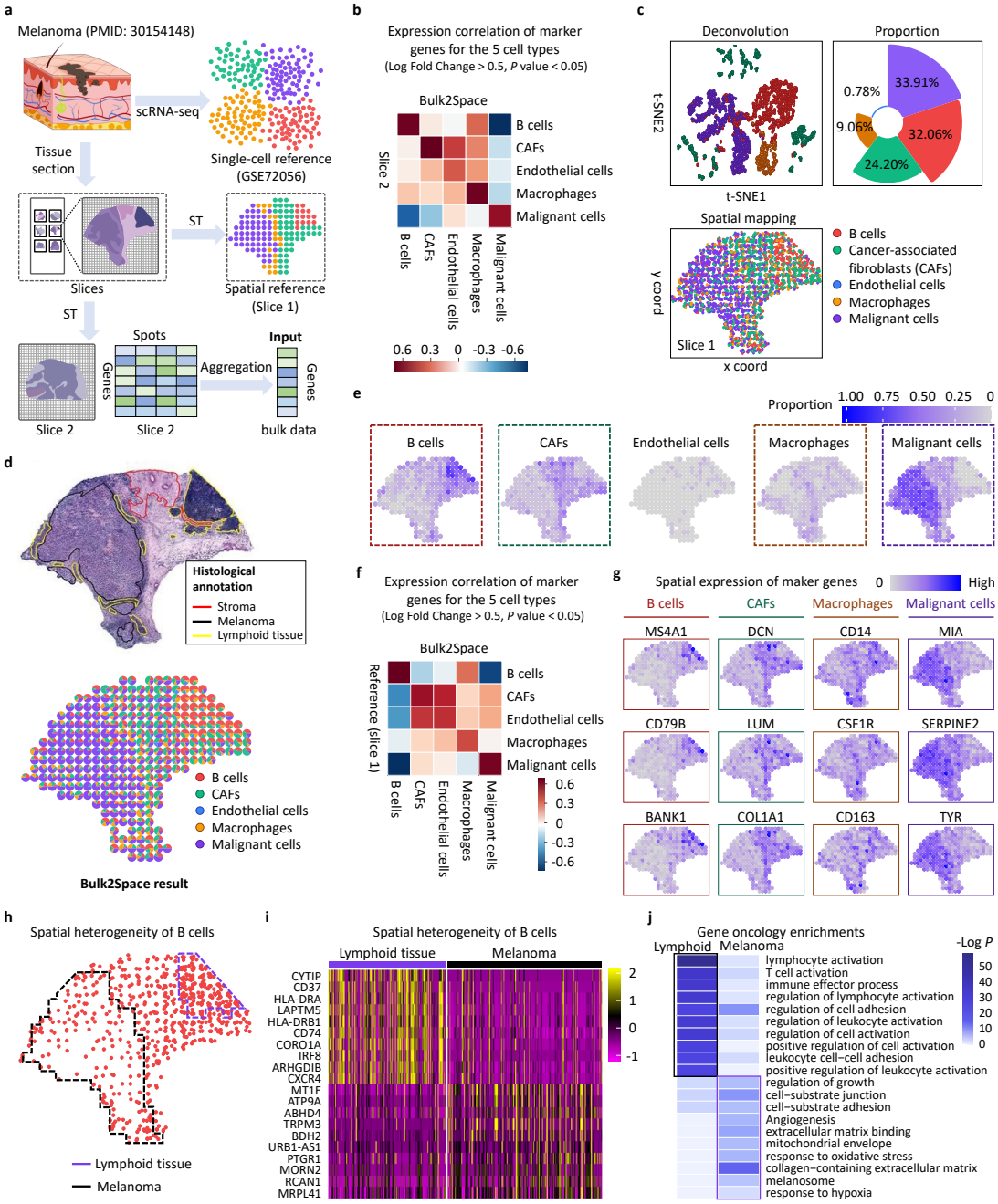

Fig. 4. Spatially resolved analysis of Melanoma by Bulk2Space. (a) Synthesis of the bulk transcriptome data. The scRNA-seq data is used as deconvolution reference, one slice of the sectioned tissue is employed as a spatial reference, and another slice is used to synthesize the input bulk data. **(b)** Expression correlation of marker genes for five predicted cell types between Bulk2Space result and Slice 2. **(c)** Single-cell deconvolution of the bulk data. Top, t-SNE plot showed the clustering of generated single cells and the rose chart showed the proportion of each predicted cell type. Bottom, distribution of predicted single cells. Different colors represented distinct cell types. **(d)** The top, Histological annotation for stroma (red), melanoma (black), and lymphoid region (yellow). Bottom, single-cell decomposition result of each spot in the ST data by Bulk2Space. **(e)** The spatial abundance of different cell types in each spot on the tissue section is predicted by Bulk2Space. **(f)** The expression correlation of the marker genes in six cell types between the

301 Bulk2Space result and Slice 1. **(g)** The spatial expression of the marker genes in B cells, CAFs,
302 macrophages, and malignant cells in predicted data, respectively. **(h)** Spatial distribution of B cells.
303 Segmented area, lymphoid tissue (purple), and melanoma region (black). **(i)** Differentially
304 expressed genes for lymphoid and melanoma tissue. **(j)** Pathway enrichment analysis for lymphoid
305 and melanoma tissue.

306

307 **Bulk2Space re-annotates ambiguous cells in the mouse hypothalamus**

308 The Bulk2Space algorithm was also used to spatially deconvolve bulk transcriptome data derived
309 from the hypothalamus region of the mouse brain using our in-house developed multiplexed
310 barcoding RNA-seq method, termed Spatial-seq, to explain the spatial distribution of single cells
311 and gene expression. The Spatial-seq workflow is shown in Fig. 5a, laser capture microdissection
312 (LCM) was used to isolate brain regions and a barcoding strategy was utilized for multiplexed RNA-
313 seq (see 'methods' section). After deconvolution by Bulk2Space, the resulting single cells had a
314 clustering space close to the reference scRNA-seq data and MERFISH data (Fig. 5b). The generated
315 single cells were mapped to spatial coordinates based on a spatial reference (Figure 5c). As shown
316 in Fig. 5c, the MERFISH and Bulk2Space results were very close in single-cell spatial distribution.
317 Among the 10 cell types, immature oligodendrocytes were unique to MERFISH, and macrophages
318 were predicted by Bulk2Space, but absent in MERFISH. The heatmap suggested 7 of the 8 shared
319 cell types were strongly correlated in both MERFISH and Bulk2Space results. The low correlation
320 of mural cells may be related to the low number of pericytes. The spatial patterns of the MERFISH
321 targeted genes were quite similar to that of Bulk2Space results (Fig. 5d). Meanwhile, owing to the
322 limited number of measured target genes, some ambiguous cells in MERFISH could not be defined
323 according to their expression profiles. Using Bulk2Space, the resulting single cells were assigned to
324 the spatial context, thus further clustering these ambiguous cells (Fig. 5e). Meanwhile, the spatial
325 distribution of novel genes beyond the targeted RNA species was predicted by Bulk2Space and was
326 highly consistent with the spatial pattern of corresponding cell types. (Fig. 5f).

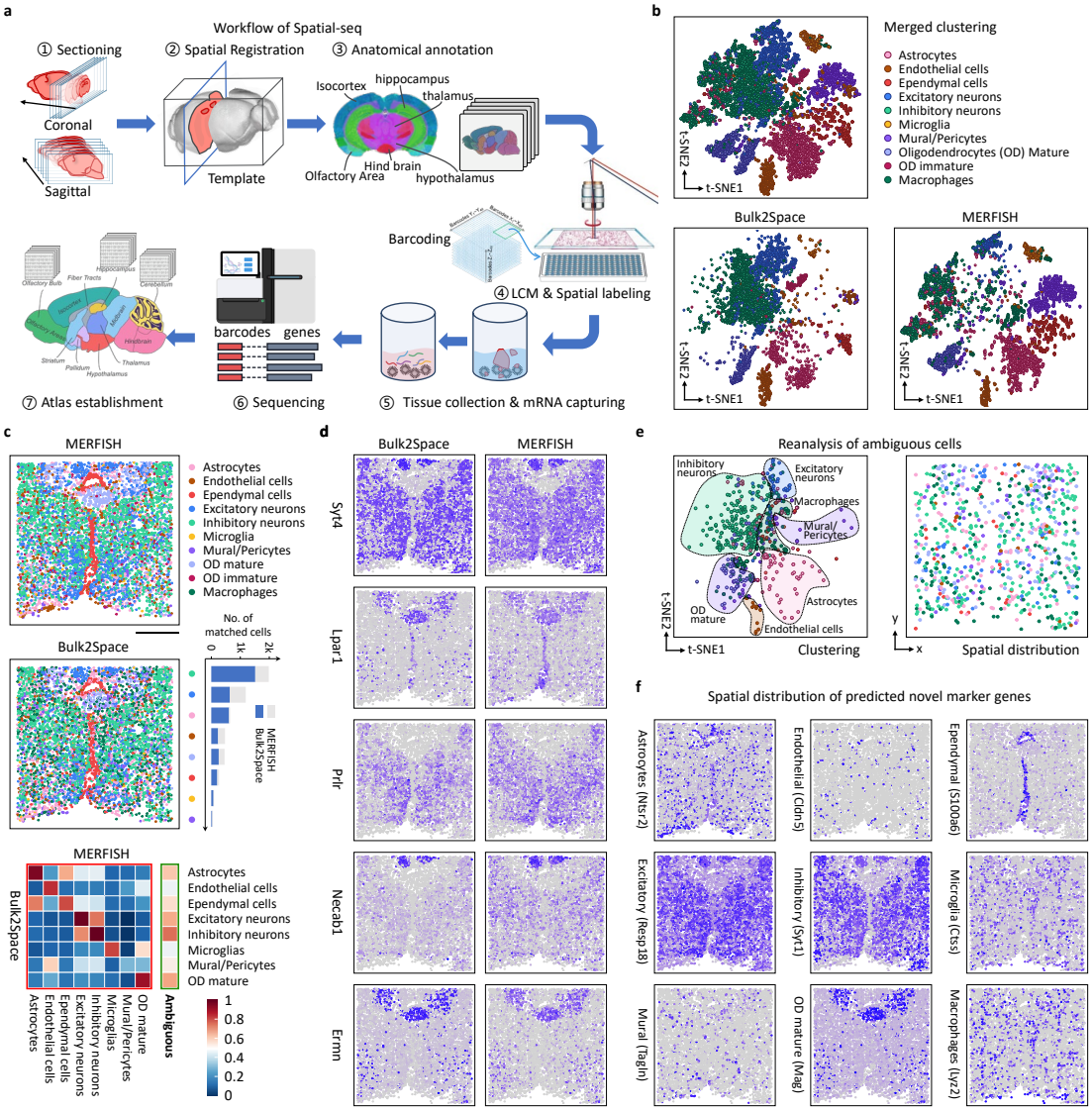

Fig. 5. Spatially resolved analysis of mouse hypothalamus by Bulk2Space. (a) Spatial-seq workflow. The mouse brain was sectioned continuously from coronal and sagittal directions. For each slice, spatial registration was carried out to annotate each brain region to be isolated. Brain regions are dissected by LCM and labeled with a unique barcode and then followed with RNA-seq. **(b)** The t-SNE layout of single-cell profiles in MERFISH data and generated from bulk tissue by Bulk2Space. The top, clustering of all cells in both datasets. Two heterogeneous sets of data were aligned through a joint analysis. Each cell type was represented by a unique color. **(c)** The spatial distribution of distinct cells in MERFISH and Bulk2Space results. The stacked bar chart showed the cell number of each cell type (grey) and the number of cells predicted by Bulk2Space to the corresponding cell type at each coordinate (blue). Bottom, Pearson correlation of gene expression for different cell types in MERFISH and Bulk2Space data. Correlation for shared cell types was in the red border and the expression correlation between MERFISH ambiguous cells and Bulk2Space predicted cell types was in the green border. **(d)**The spatial expression of the Bulk2Space predicted genes and the ground truth (MERFISH). **(e)** Re-annotation of ambiguous cells in MERFISH using Bulk2Space results. Left, clustering of the ambiguous cells. Right, spatial assignment of cells that

343 were colored according to their predicted cell types. **(f)** Predicted spatial expression patterns of
344 novel marker genes for different cell types.

345

346 Bulk2Space uncovers spatial gene expression dynamics in inflammation-induced 347 prostate cancer

348 Tumor development is linked to chronic infection, dietary factors, obesity, inhaled pollutants,
349 tobacco use, and autoimmunity. The unifying principle underlying these processes is inflammation,
350 which is an aberrantly prolonged form of a protective response to a loss of tissue homeostasis[41-43].
351 To explore whether Bulk2Space could identify the cellular and molecular pathways that coordinate
352 the tumor-promoting effects in inflammation-induced cancer, we deconvolved the bulk
353 transcriptome data of prostate cancer[44] and mapped them to the corresponding spatial references,
354 including normal glands, inflammatory tissue, and tumor sites. These tissue regions were isolated
355 from the prostate, and three of the sections were subjected to RNA-seq based on spatial barcoding
356 (Fig. 6a). Next, Bulk2Space was used to deconvolve the bulk data into single-cell profiles (Fig. 6b)
357 and mapped the generated single cells to spatial coordinates (Fig. 6c). Interestingly, the
358 composition and distribution of cells in different tissue regions were quite different and closely
359 related to the state of the tissue (Fig. 6d and 6e). Comparison of cell distribution among normal,
360 inflammatory, and cancerous tissues (Fig. 6f to 6h) showed that in normal glands, cancer-
361 associated fibroblasts (CAFs) were the most abundant, but in inflammatory and cancer tissues, the
362 proportions of CAFs were significantly lower (52.4% in normal, 20.5% in inflammation, and 27.5%
363 in cancer). In the areas of inflammation, cells involved in the inflammatory response, such as
364 immunocompetent B cells, endothelial cells, monocytes, and NK cells (Fig. 6g), were more
365 abundant than in normal or tumor tissues (13.4% in inflammation, 2.0% in normal, and 1.7% in
366 cancer). In cancer with a Gleason score (Gs) of 3+4, the proportion of cancer epithelial cells was
367 significantly higher (53.0% in cancer, 6.3% in normal, and 4.3% in inflammation) than that in normal
368 and inflammatory areas (Fig. 6h), while the proportion of the periventricular layer cells was
369 significantly lower (17.9% in cancer, 39.3% in normal, and 44.1% in inflammation). The Bulk2Space
370 results indicated that, during the progress of inflammation-induced prostate cancer, CAFs
371 accumulated in normal glands at the early stage, resulting in high expression of inflammatory

factors and promoting the occurrence of local inflammation, a fundamental innate immune response to perturbed tissue homeostasis, thus leading to tissue cancerization.

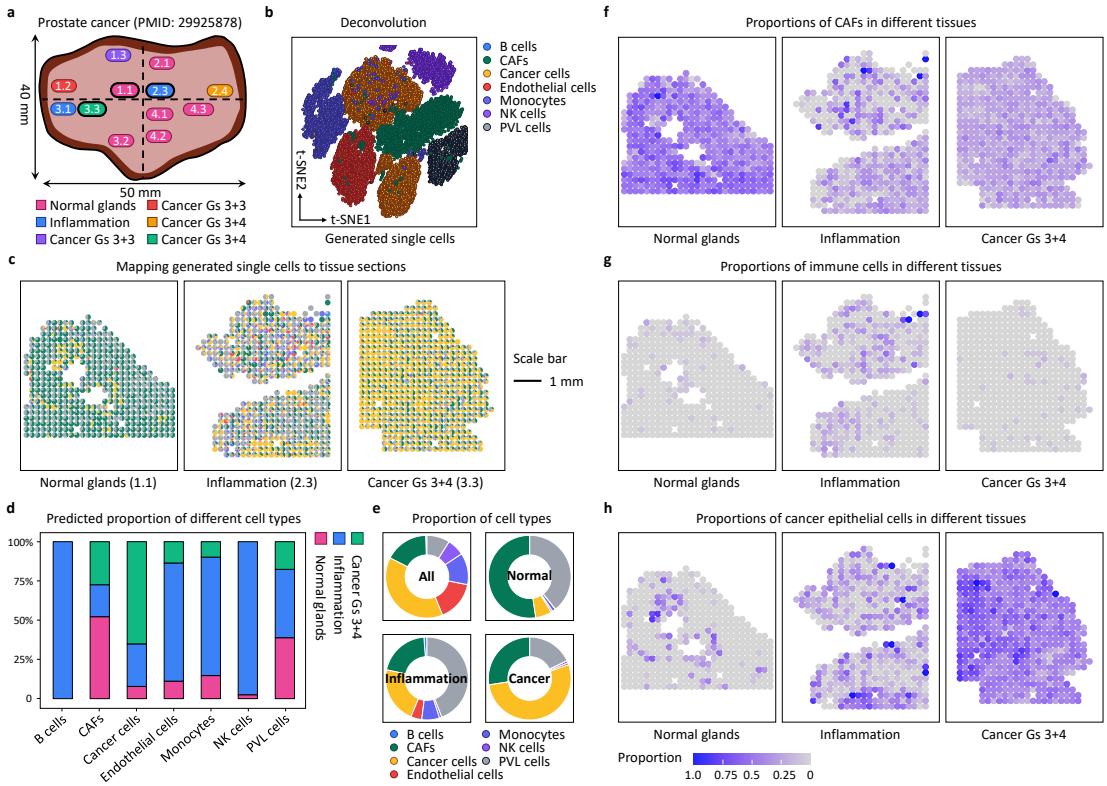

Fig. 6. Spatial gene expression dynamics analysis of inflammation-induced prostate cancer by Bulk2Space. (a) Experimental design of the bulk transcriptome data. The tissue was segmented into distinct regions, namely, normal glands (pink), inflammation (blue), and several cancer regions (red, green, purple, and yellow). Three tissue regions (1.1 normal, 2.3 inflammation, and 3.3 cancer region) were isolated to conduct spatially resolved transcriptomics. **(b)** The t-SNE layout of the deconvolved single-cell profiles by Bulk2Space. Each cell type was represented by a unique color. **(c)** The proportions of generated single cells that were mapped to normal, inflammation, and cancer region by Bulk2Space, respectively. The color of each cell type was consistent with (b). **(d)** The proportion of cells assigned to normal (pink), inflammation (blue), and cancer (green) regions for each cell type. **(e)** The predicted proportion of distinct cell types for each tissue region by Bulk2Space. **(f)** The proportions of CAFs in normal (left), inflammation (middle), and cancer regions (right). **(g)** The proportions of immune cells in normal, inflammation, and cancer regions. **(h)** The proportions of cancer epithelial cells in normal, inflammation, and cancer regions.

Discussion

Understanding transcriptional heterogeneity within the tissue from the perspective of single-cell spatial resolution and full gene coverage is an essential direction for the development in the field of biological sciences[1]. However, there are many technical challenges, and thus, such a technology has not been fully realized yet. Traditional bulk transcriptome data can provide gene expression patterns and reflect the overall status of tissue, which is widely used in studies on the occurrence and development of diseases. Nevertheless, disease-inherent cell composition and spatial distribution are difficult to analyze. Although several attempts can infer the cell-type proportions from traditional RNA-seq data, understanding bulk tissue at single-cell spatial resolution remains a pressing need in the field. Cell identity can be characterized by the clustering space of single-cell gene expression profiles[45], and expression features can be stably retained across different conditions, technologies, and species[29]. Based on this, we used a deep learning model, termed β-VAE, to generate single-cell profiles with biological significance within the clustering space of each cell type and map them accurately to tissue coordinates, thus transforming bulk transcriptomics into spatially resolved single-cell transcriptome data. As thoroughly considered the merits and demerits of the widespread used two spatially resolved transcriptomics technologies, we employed different spatial mapping strategies for each method. For image-based *in situ* hybridization, Bulk2Space could predict the spatial expression of novel mRNA species apart from the target genes. Moreover, Bulk2Space can help annotate cells whose types are difficult to further distinguish by targeted methods. For spatial barcoding-based RNA-seq, each sequenced tissue region contains transcripts from different cells and can thus be regarded as a mixture of cells, not allowing single-cell resolution. Bulk2Space can assign single cells to optimal spots, and these mixtures are split into collections of individual cells, thus achieving single-cell resolution. Our results showed that Bulk2Space had great biological and clinical application prospects, including linking tissue molecular characteristics with histological phenotypes, revealing the spatial specific variation of cells, discovering the spatial expression pattern of novel genes, achieving more refined cell clustering, and uncovering the molecular mechanism underlying the progression of the disease.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (81973701, 91846204, and U19B2027), the Natural Science Foundation of Zhejiang Province (LZ20H290002), the Innovation Team and Talents Cultivation Program of the National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202002), the Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine), and the Alibaba Cloud (Alibaba Cloud Computing Co. Ltd.).

Author contributions

X.F., H.C., Y. G., and J. L. conceived the study. J. L. and J. Q. collected datasets involved in this article, benchmarked all methods, and participated in the development of the Bulk2Space algorithm. Y. F., Z. C., and X. Z. implemented the code and interface to the Bulk2Space algorithm. N. Z., X. S., H. Y., P. Y., J. C., L. S., and J. Z. provided a lot of advice on algorithm implementation and biological applications. J. L., Y. H., Y. H., L. Y., and X. L. conducted the multiplexed RNA-seq experiment using Spatial-seq. D. W. provided important advice on the brain tissue isolation, spatial registration, and annotation. All authors wrote the manuscript, read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Methods

Datasets

All scRNA-seq, spatially resolved transcriptomics, and bulk RNA-seq datasets used for

440 benchmarking of Bulk2Space were retrieved from high-quality publications, Gene Expression
441 Omnibus (GEO), and The Cancer Genome Atlas (TCGA), wherein unannotated, ambiguous, or low
442 quality-cells were excluded.

443 *scRNA-seq datasets.* Ten scRNA-seq were applied to benchmark the Bulk2Space algorithm,
444 including five human datasets, peripheral blood, brain, kidney, liver, and lung tissues, and five
445 mouse datasets, brain, kidney, lung, pancreas, and testis tissues. A dataset from the hypothalamic
446 preoptic region in mouse brain containing 30370 cells of 12 cell types is used for spatial mapping
447 using image-based methods. The human pancreatic ductal adenocarcinomas (PDAC), human
448 melanoma, and human prostate cancer data were used for the validation of Bulk2Space.

449 *Spatially resolved transcriptome data*. For image-based spatial mapping, the mouse
450 hypothalamus MERFISH data were obtained from Dryad (doi:10.5061/dryad.8t8s248), and the
451 Bregma +0.26 slice of mouse #1 was chosen to further analysis. For spatial barcoding-based
452 methods, the human PDAC, human melanoma, and human prostate cancer data were used for
453 reference and validation.

454 *Bulk transcriptome data.* Two bulk RNA-seq datasets were used to illustrate the applications of
455 Bulk2Space. The mouse hypothalamus data were derived from experiments of our in-house
456 developed Spatial-seq protocols. The human prostate cancer RNA-seq data were downloaded from
457 TCGA.

458 **Data processing**

459 All data were preprocessed using R (version 4.1.1). For mouse brain scRNA-seq data by Moffitt[38] et
460 al., we filtered 841 ambiguous cells and 88 unstable cells. For human prostate cancer scRNA-seq
461 data by Wu[46] et al., we filtered 183 unassigned cells. All cells of other scRNA-seq as well as spatially
462 resolved transcriptomics datasets were retained. For mouse hypothalamus MERFISH data, we
463 filtered 5 "blank" barcodes as well as *Fos* gene, whose expression value in all cells was 'NA'. For
464 human prostate cancer RNA-seq data, we used biomaRt (version 2.48.3) to transform ensemble id
465 to gene name. For all scRNA-seq and spatially resolved transcriptomics datasets, the raw counts

were normalized using the global-scaling normalization method "LogNormalize". For MERFISH dataset, the normalized data was determined as the raw count per cell divided by the cell volume and scaled by 1000.

Step 1: Deconvolution of bulk transcriptome data.

Cell type proportions prediction. For a given bulk RNA-seq dataset, we aimed to calculate its cell type proportions firstly, and then generate single-cell gene expression profiles based on the calculated proportions. Since cell type and gene expression of all single cells can be accessed from the scRNA-seq data reference, we collected cells with the same cell type and then average their gene expression. This average vector $\mathbf{c}_i \in \mathbb{R}^N$ is defined as the gene expression of cell type i ($i \in \{1,2,\dots, C\}$, where C denotes the total number of cell types). Given the gene expression vector $\mathbf{x} \in \mathbb{R}^N$ of a bulk RNA-seq dataset, we aimed to predict the cell type proportion:

$$\sum_{i=1}^{C} p_i \mathbf{c}_i = \mathbf{x}$$

Here, p_i is the proportion of cell type i.

We applied least square estimation (LSE) to estimate p_i by minimizing the squared discrepancies between observed data.

$$\min_{\boldsymbol{p}} \|\mathbf{x} - \mathbf{C}\boldsymbol{p}\|_2$$

where $\mathbf{C} \in \mathbb{R}^{N \times C}$ denotes the cell type gene expression profile, and each column represents a different cell type. $\boldsymbol{p} = [p_1, p_2, \dots, p_C]^{\boldsymbol{T}}$ is the proportion vector to be estimated.

Notably, we only chose top 500 marker genes of each cell type in this step, to reduce the noise effect.

Single-cell simulation by β*-VAE.* Let $\mathcal{D} = \{X, V, W\}$ be a set consisting of gene expression vectors $x \in \mathbb{R}^N$ and two sets of ground truth data generative factors: conditionally independent factors $v \in \mathbb{R}^K$, where $\log p(v|x) = \sum_k \log p(v_k|x)$; and conditionally dependent factors $w \in \mathbb{R}^H$. We

assume that x are generated by a true world simulator S using the corresponding ground truth data generative factors: $p_\theta(x|v,w) = S(v,w)$, where θ is the generative model parameter.

We wanted to develop an unsupervised deep generative model that, using samples from X only, can learn the joint distribution of the data x and a set of generative latent factors z ($z \in \mathbb{R}^M$, where $M \geq K$) that can be used to generate the observed data x; i.e., $p_\theta(x|z) \approx p(x|v,w) = S(v,w)$. However, since the integral of the marginal likelihood $p_\theta(x) = \int p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})\, d\mathbf{z}$ is intractable (so we cannot evaluate or differentiate the marginal likelihood), the true posterior density $p_\theta(\mathbf{z}|x) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) / p_\theta(\mathbf{x})$ is intractable.

To solve this problem, for a given observation x, we described the inferred posterior configurations of the latent factors $\mathbf{z}$ through a probability distribution $q_\phi(\mathbf{z}|\mathbf{x})$: an approximation to the intractable true posterior $p_\theta(x|z)$. We aimed to ensure that the inferred latent factors $q_\phi(\mathbf{z}|\mathbf{x})$ capture the generative factors v in a disentangled manner. A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors. In a disentangled representation, knowledge about one factor can generalize to novel configurations of other factors. The conditionally dependent data generative factors w can remain entangled in a separate subset of $\mathbf{z}$ that is not used for representing v.

An intuitive approach was to minimize the KL divergence between the approximate and the true posterior:

$$\mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|x)\right) = -\sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{p_\theta(\mathbf{z}|x)}{q_\phi(\mathbf{z}|\mathbf{x})}\right) = -\sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{\frac{p_\theta(x,\mathbf{z})}{p_\theta(x)}}{q_\phi(\mathbf{z}|\mathbf{x})}\right)$$

$$= -\sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \left[\log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right) - \log\big(p_\theta(x)\big)\right]$$

$$= -\sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right) + \log\big(p_\theta(x)\big) = -\mathcal{L}(\theta,\phi;\mathbf{x}) + \log\big(p_\theta(x)\big)$$

Here, $\mathcal{L}(\theta,\phi;\mathbf{x})$ is called the variational lower bound, and can be written as:

$$\mathcal{L}(\theta,\phi;\mathbf{x}) = \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right) = \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right)$$

$$= \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \left[\log\big(p_\theta(\mathbf{x}|\mathbf{z})\big) + \log\left(\frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right)\right]$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log\big(p_\theta(\mathbf{x}|\mathbf{z})\big)\big] - \mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right)$$

To encourage this disentangling property in the inferred $q_\phi(\mathbf{z}|\mathbf{x})$, we introduced a constraint over it by trying to match it to a prior $p(\boldsymbol{z})$ that can both control the capacity of the latent information bottleneck, and embodies the desiderata of statistical independence mentioned above. We set the prior to be an isotropic unit Gaussian $p(z) \sim \mathcal{N}(0, I)$, then the constrained optimization problem can be written as:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] \ \textit{subject to}\ \mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right) < \epsilon$$

where ϵ specifies the strength of the applied constraint. Re-writing the above equation as a Lagrangian under the KKT conditions, we obtain:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] - \beta\left(\mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right) - \epsilon\right)$$

where the KKT multiplier β is the regularization coefficient that constrains the capacity of the latent information channel $\mathbf{z}$ and puts implicit independence pressure on the learnt posterior due to the isotropic nature of the Gaussian prior $p_\theta(\mathbf{z})$. Since $\beta, \epsilon \geq 0$, according to the complementary slackness KKT condition, the equation can be re-written as:

$$\mathcal{F}(\theta,\phi,\beta;\mathbf{x},\mathbf{z}) \geq \mathcal{L}(\theta,\phi;\mathbf{x},\mathbf{z},\beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] - \beta\left(\mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right)\right)$$

which is the β-VAE formulation with the addition of the β coefficient.

Here, different β will change the degree of applied learning pressure during training, thereby encouraging different learnt representations. We postulated that to learn disentangled representations of the conditionally independent data generative factors $\boldsymbol{v}$, it is important to set

$\beta > 1$ to impose a stronger constraint on latent bottleneck than the original VAE. These constraints limit the capacity of $\mathbf{z}$, coupled with the pressure to maximize the log likelihood of the training data $\mathbf{x}$ under the model, encouraging the model to learn the most efficient representation of the data. The additional pressure from high β values may create a trade-off between reconstruction fidelity and the quality of disentanglement within the learnt latent representations. When the appropriate balance is found between information preservation (reconstruction cost as regularization) and latent channel capacity restriction ($\beta > 1$), disentangled representations emerge.

Step2: Mapping generated single cells to spatial locations

Spatial barcoding-based RNA-seq. After obtaining the generated data, we needed to map each cell and the spot it belongs to. To improve the accuracy and reduce the complexity of this process, we divided it into two steps. First, we calculated the cell type proportion of each spot, and then predicted which cells are contained in each spot based on this proportion.

1) Calculate cell type proportions. The procedure is identical to the Step 1 deconvolution, we repeated this process to obtain the cell type proportion of each spot in the tissue.

2) Deep forest for spot recommendation. With cell type proportion, we next predicted which cells are contained in each spot. We defined this task as a binary classification problem. We designed a classifier, whose input is the gene expression vectors of both single cell $\mathbf{s}$ and spot $\mathbf{x}$. When the cell belongs to the spot, the output of the classifier is 1, otherwise the output is 0. Specifically, the input can be represented as $\mathbf{v} = [\mathbf{s};\ \mathbf{x};\ \mathbf{s} - \mathbf{x}] \in \mathbb{R}^{N\times 3}$, where semicolon means concatenate operation. $\mathbf{s} - \mathbf{x} \in \mathbb{R}^{N}$ can be regarded as an auxiliary signal, since when there is a negative number occurring in this vector, the single cell must not belong to the spot.

We applied multi-Grained Cascade Forest (gcForest), a decision tree ensemble with a cascade structure and further enhanced by multi-grained scanning, to build our classifier. Each level of cascade receives feature information processed by its preceding level and outputs its processing result to the next level. Each level is an ensemble of decision tree forests, i.e., an ensemble of ensembles. We also included completely-random tree forests) to encourage diversity. Given an

instance, each forest counts the percentage of different classes of training samples at the leaf node where the concerned instance falls, and then average across all trees in the same forest to produce an estimate of class distribution. The estimated class distribution forms a class vector, which was generated by k-fold cross validation and then concatenated with the original feature vector to be input to the next level of the cascade.

To enhance cascade forest, we introduced a procedure of multi-grained scanning. We used sliding windows to scan the raw features. Specially, we used a window with a fixed length of k to slide the input vector into k-dimensional feature vectors. We then feeded these vectors to n forests and finally produce n two-dimensional vectors. By using multiple sizes k of sliding windows, differently grained feature vectors were generated.

The we summarized the overall procedure of gcForest. For the original raw features, three window sizes $\{\lfloor N/16 \rfloor, \lfloor N/8 \rfloor, \lfloor N/4 \rfloor\}$ were used for multi-grained scanning. The generated data were used to train a completely-random tree forest containing 500 trees. We concatenated the output of these two forests as the transformed feature vectors, which also acted as the input of cascade forest. The final model was a cascade of cascades, where each cascade consisted of multiple levels each corresponding to a grain of scanning. Each level consisted of 4 completely-random tree forests, each containing 500 trees. In other words, the transformed feature vectors were augmented with the class vector generated by the previous grade, and then were used to train the current grade of cascade forests.

Given a test instance, it would obtain its corresponding transformed feature vector through the multi-grained scanning procedure, and then go through the cascade till the last level. The two-dimensional class vectors at the last level were aggregated, and the class with the largest aggregate value was selected to obtain the final prediction result.

This decision tree ensemble approach not only had much fewer hyper-parameters than deep neural networks, but also retained the interpretability of tree models. Its model complexity could be automatically determined in a data-dependent manner, which made gcForest work well even on small-scale data.

587 *Image-based in situ hybridization*. We developed a simple method for predicting the spatial
588 distributions of genes not measured in spatial transcriptomic data which produced by *in situ* RNA
589 imaging-based technologies. To do this, Bulk2space used the spatial transcriptomic data as the
590 reference, and generated scRNA-seq data were mapped to the tissue space corresponding to this
591 reference. To eliminate the differences between scRNA-seq data and spatial transcriptomic data
592 caused by different experiment types, an empirical Bayes framework was used to remove the batch
593 effect, so that both sets of data are at the same scale level. Subsequently, the cross-dataset k-
594 nearest neighbor graph of spatial transcriptomic data in scRNA-seq data was computed in the
595 aligned space, and then the predicted whole gene expression profile of each cell of spatial
596 transcriptomic data could be calculated to the mean of its k nearest neighbors in scRNA-seq data.

597

598 **Data simulations**

599 Using several simulations, we evaluated the performance of Bulk2space. The first simulation
600 compared the generation of single-cell gene expression profiles by Bulk2Space with other methods,
601 GAN and CGAN. The simulated data were derived from 10 scRNA-seq data of benchmarking usage.
602 We randomly divided the scRNA-seq data into two parts, one was treated as the single-cell
603 reference and the other for the construction of bulk transcriptome data via aggregating all the
604 single-cell gene expression profiles. Considering the cell composition of bulk RNA-seq data varies
605 greatly in the natrual situation, we further changed the cell-type proportion and synthesized three
606 bulk transcriptome data with different cell compositions from the same part of scRNA-seq data. In
607 total, 40 paired simulated data were synthesized in this study. Performance was assessed using the
608 Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC), and root
609 mean squared error (RMSE) between the gene expression profile of scRNA-seq data reference and
610 synthetic bulk transcriptome data.

611 In the second simulation, we evaluated the robustness of Bulk2space. We introduced two kinds
612 of noise: cell type noise and gene expression noise. For cell type noise, we randomly altered the
613 type of cells with the ratio of 0.01, 0.02, 0.04, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, and 0.9. For gene

614 expression noise, we firstly constructed a noise expression profile $\mathbf{C}$,

615 $$\mathbf{C} \sim \mathrm{U}(-x, x)$$

616 Here, x is the maximum gene expression value of scRNA-seq data reference.

617 We next constructed synthetic expression matrix by combining noise expression profile with
618 scRNA-seq data reference, and the ratio of gene expression-altered cells remains 0.01, 0.02, 0.04,
619 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, and 0.9. PCC, SRCC and RMSE between the synthetic expression
620 matrix and constructed bulk RNA-seq data were used to evaluate performance.

621 In the third simulation, we simulated spatial barcoding data to evaluate the spatial
622 reconstruction capacity of Bulk2space with other five classical classification models: Logistic
623 Regression (LR), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT), Multilayer Perceptron
624 (MLP) and DeepGBM. Starting from scRNA-seq data reference, we randomly chose 10 cells from it
625 and aggregated their transcriptomic profiles as a spot of pseudo spatial transcriptomic data. The
626 spot with over 25000 UMI counts would be sampled down to 20000 UMI counts to better meet
627 the true situation. We also constructed pseudo spatial transcriptomic data with 100, 200, 500 and
628 1000 spot numbers to simulate true ST data produced by different spatial barcoding technologies.
629 PCC, SRCC, and RMSE of cell type proportion and gene expression profile between model
630 prediction results and ground truth were used to evaluate performance.

631 In the fourth and fifth simulations, we systematically evaluated the performance of Bulk2space.
632 We used Bulk2space into 2 types of cancer data, pancreatic ductal adenocarcinomas (PDAC) and
633 melanoma. The bulk RNA-seq data were simulated from the biological replicates of selected ST
634 slices.

635 Spatial-seq protocol

636 *Multiplexed spatial barcoding.* A virtual 48×48×48 three-dimensional well array (barcoding array)
637 was constructed in X, Y, and Z directions, with 48×48 wells in each layer. 384-well plates (24×16
638 wells) were selected as the basic units (2×3 plates) of each layer. Each well in the barcoding array

639 can be labeled as barcode $(\{X_i, Y_j, Z_k\}, i, j, k \in N, [1, 48])$. Magnetic beads coated with carboxyl
640 group were distributed into each well for barcode synthesis. There were three rounds of barcode
641 extension reactions. In the first round, 5' amino-modified barcoded oligonucleotides (barcode X_i)
642 were conjugated to the beads. In the second round of extension reaction, barcoded
643 oligonucleotides (barcode Y_j) which had a sequence of bases at the 5' end complementary to the
644 3' end of barcode X_i were linked to the end of barcode X_i by PCR. Similarly, in the third round
645 of reaction, barcode sequences containing a unique molecular identifier (UMI) and a polyT tail
646 (barcode Z_k) were extended to barcode Y_j by PCR. This method only needs 144 (48×3) different
647 barcodes to theoretically generate magnetic beads with 48^3 distinct kinds of magnetic beads with
648 known sequences. More importantly, the entire synthesis process was serialized and multiplexed,
649 which can be accomplished with the help of arrayed pipettes or liquid workstations, rather than
650 synthesizing the magnetic beads well by well. The spatially barcoded beads can be used in many
651 areas, such as multiplexed labeling of tissue samples or even single cells.

652 *Collection of mouse brain regions.* The mouse brain was sliced into 14-μm sections from the coronal
653 and sagittal directions. Each tissue slice was registered to a reference brain template provided by
654 the Allen Brain Atlas (https://portal.brain-map.org/). After spatial registration, the anatomical
655 regions of the mouse brain were delineated and annotated. The outlines, dissection sequences,
656 and collectors of the brain regions were specified from the annotated data and then imported to
657 an LCM instrument (Laser Microdissection System equipped with a DM68 microscope, a Leica
658 LMD6 laser cutter, and a single-cell capture collector LMT350, Leica Microsystems, Germany).LCM
659 is a microscope-guided powerful cutting system incorporating UV light for contact- and
660 contamination-free isolation of areas of interest from tissue sections[15]. Brain regions were
661 dissected from the tissue by LCM according to the imported files and fell into the collector loaded
662 with barcoded beads in advance. Spatial-seq has the potential to achieve spatially resolved whole
663 transcriptome sequencing of a large number of single cells.

664 *Multiplex RNA-seq.* The tissue was lysed in the barcoded well, allowing mRNA to be captured by
665 polyT tail on the surface of the magnetic beads. Each well was loaded with lysis buffer. The tissue
666 was lysed on ice for 12 minutes. The beads were then washed and transferred for reverse

transcription (RT). The captured mRNA was reverse transcribed to construct cDNA libraries following the guidance of Illumina Nextera XT DNA Library Preparation Kits. A paired-end sequencing was conducted to decode the spatial barcode in the 3' end and detect RNA species in the 5' end of the cDNA.

Sequence alignment. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was utilized for quality control of the RNA-seq data. The derived RNA-seq data were fragmented into lots of files based on the spatial barcodes. The digital gene expression matrix for each barcode was retrieved follow the Drop-seq sequence alignment cookbook.

Data normalization. FPKM, RPKM, RPM, and TPM were calculated from the gene expression matrix according to the following formulas.

$$\mathrm{FPKM} = \frac{ExonMappedFragments * 10^9}{TotalMappedFragments * ExonLength}$$

$$\mathrm{RPKM} = \frac{(ExonMappedRead/TotalMappedReads * 10^6)}{ExonLength} * 10^3$$

$$\mathrm{RPM} = \frac{ExonMappedReads * 10^6}{TotalMappedReads}$$

$$\mathrm{TPM} = \frac{N_i/L_i * 10^6}{sum(N_1/L_1 + N_2/L_2 + \cdots + N_n/L_n)}$$

The data were normalized using the global-scaling normalization method "LogNormalize" in Seurat. A Bayesian network-based method termed ComBat was used for batch effect removal of data derived from different cDNA libraries. The expression matrix of each sample was normalized using the "Deconvolution Normalization" algorithm to align the median gene expression.

Data availability

The original and processed mouse brain data used in this study were uploaded to GEO (GSE192999).

Code availability

Bulk2Space is available as a python package and the source code is deposited at the GitHub (https://github.com/ZJUFanLab/bulk2space).

Statistics

Python (version 3.8.5) and R (version 4.1.1) were used for the statistical analysis.

695 # References

696 1. Liao, J., Lu, X., Shao, X., Zhu, L. & Fan, X. Uncovering an Organ's Molecular Architecture at Single-Cell
697 Resolution by Spatially Resolved Transcriptomics. *Trends Biotechnol* **39**, 43-58 (2021).

698 2. Wang, Y.S. & Guo, J. Multiplexed Single-Cell in situ RNA Profiling. *Front Mol Biosci* **8**, 775410 (2021).

699 3. Srivatsan, S.R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111-117 (2021).

700 4. Waylen, L.N., Nim, H.T., Martelotto, L.G. & Ramialison, M. From whole-mount to single-cell spatial
701 assessment of gene expression in 3D. *Commun Biol* **3**, 602 (2020).

702 5. Qian, X. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat*
703 *Methods* **17**, 101-106 (2020).

704 6. Larsson, L., Frisen, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new dimension to
705 genomics. *Nat Methods* **18**, 15-18 (2021).

706 7. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved,
707 highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

708 8. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by
709 sequential hybridization. *Nat Methods* **11**, 360-361 (2014).

710 9. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states.
711 *Science* **361** (2018).

712 10. Eng, C.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**,
713 235-239 (2019).

714 11. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**,
715 987-990 (2019).

716 12. Stickels, R.R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-
717 seqV2. *Nat Biotechnol* **39**, 313-319 (2021).

718 13. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high
719 spatial resolution. *Science* **363**, 1463-1467 (2019).

720 14. Stahl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial
721 transcriptomics. *Science* **353**, 78-82 (2016).

722 15. Nichterwitz, S. et al. Laser capture microscopy coupled with Smart-seq2 for precise spatial
723 transcriptomic profiling. *Nat Commun* **7**, 12139 (2016).

724 16. Chen, J. et al. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat*
725 *Protoc* **12**, 566-580 (2017).

726 17. Cheng, J., Liao, J., Shao, X., Lu, X. & Fan, X. Multiplexing Methods for Simultaneous Large-Scale
727 Transcriptomic Profiling of Samples at Single-Cell Resolution. *Adv Sci (Weinh)* **8**, e2101229 (2021).

728 18. Liu, Y., Ai, N., Liao, J. & Fan, X. Transcriptomics: a sword to cut the Gordian knot of traditional Chinese

729 medicine. *Biomark Med* **9**, 1201-1213 (2015).

730 19. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

731 20. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human
732 glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

733 21. International Cancer Genome, C. et al. International network of cancer genome projects. *Nature* **464**,
734 993-998 (2010).

735 22. Consortium, S.M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and
736 information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903-914
737 (2014).

738 23. Frishberg, A. et al. Cell composition analysis of bulk genomics using single-cell data. *Nat Methods*
739 **16**, 327-332 (2019).

740 24. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat*
741 *Methods* **12**, 453-457 (2015).

742 25. Wang, X., Park, J., Susztak, K., Zhang, N.R. & Li, M. Bulk tissue cell type deconvolution with multi-
743 subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).

744 26. Chen, Z. et al. Inference of immune cell composition on the expression profiles of mouse tissue. *Sci*
745 *Rep* **7**, 40508 (2017).

746 27. Gangalum, R.K. et al. Spatial Analysis of Single Fiber Cells of the Developing Ocular Lens Reveals
747 Regulated Heterogeneity of Gene Expression. *iScience* **10**, 66-79 (2018).

748 28. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for
749 Single-Cell Multi-omics. *Trends Biotechnol* **38**, 1007-1022 (2020).

750 29. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data
751 across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

752 30. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

753 31. Shao, X. et al. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA
754 Sequencing Data. *iScience* **23**, 100882 (2020).

755 32. Shao, X. et al. scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics
756 using deep learning with a weighted graph neural network. *Nucleic Acids Res* **49**, e122 (2021).

757 33. Higgins, I. et al. in ICLR (2017).

758 34. Goodfellow, I.J. et al. in NIPS (2014).

759 35. Mirza, M. & Osindero, S.J.A. Conditional Generative Adversarial Nets. **abs/1411.1784** (2014).

760 36. Gierahn, T.M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput.
761 *Nat Methods* **14**, 395-398 (2017).

762 37. Vieira Braga, F.A. et al. A cellular census of human lungs identifies novel cell states in health and in
763 asthma. *Nat Med* **25**, 1153-1163 (2019).

764 38. Moffitt, J.R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic
765 region. *Science* **362** (2018).

766 39. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq
767 reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38**, 333-342 (2020).

768 40. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-
769 seq. *Science* **352**, 189-196 (2016).

770 41. Elinav, E. et al. Inflammation-induced cancer: crosstalk between tumours, immune cells and
771 microorganisms. *Nat Rev Cancer* **13**, 759-771 (2013).

772 42. Shao, X., Lu, X., Liao, J., Chen, H. & Fan, X. New avenues for systematically inferring cell-cell
773 communication: through single-cell transcriptomics data. *Protein Cell* **11**, 866-880 (2020).

774 43. Shao, X. et al. CellTalkDB: a manually curated database of ligand-receptor interactions in humans
775 and mice. *Brief Bioinform* **22** (2021).

776 44. Berglund, E. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape
777 of heterogeneity. *Nat Commun* **9**, 2419 (2018).

778 45. Shao, L. et al. Identify differential genes and cell subclusters from time-series scRNA-seq data using
779 scTITANS. *Comput Struct Biotechnol J* **19**, 4132-4141 (2021).

780 46. Wu, S.Z. et al. Cryopreservation of human cancers conserves tumour heterogeneity for single-cell
781 multi-omics analysis. *Genome Med* **13**, 81 (2021).

782