De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution

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

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

[2] Future Health Laboratory, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China

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

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

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

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

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

[8] Department of Pharmaceutical Sciences, Beijing Institute of Radiation Medicine, Beijing, 100850, China

[9] Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, 310024, China.

Correspondence should be addressed to X.F. (e-mail: fanxh@zju.edu.cn), H.C. (e-mail: huajunsir@zju.edu.cn) and Y.G. (e-mail: gaoyue@bmi.ac.cn).

[†] These authors contributed equally to this work.

Abstract

Uncovering the tissue molecular architecture at single-cell resolution could help better understand organisms' biological and pathological processes. However, bulk RNA-seq can only measure gene expression in cell mixtures, without revealing the transcriptional 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 disclose the spatial and cellular heterogeneity of bulk RNA-seq data using existing single-cell and spatial transcriptomics references. The use of bulk transcriptomics to validate Bulk2Space unveils, 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 is utilized to perform spatial deconvolution analysis on bulk transcriptome data from two different mouse brain regions derived from our in-house developed sequencing approach termed Spatial-seq. We have not only reconstructed the hierarchical structure of the mouse isocortex but also further annotated cell types that were not identified by original methods in the mouse hypothalamus.

Introduction

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

55 two decades, RNA-seq has been extensively applied in transcriptome analysis[18], with
56 many large projects having been carried out, such as the Encyclopedia of DNA
57 Elements (ENCODE)[19], The Cancer Genome Atlas (TCGA)[20], and projects of the
58 International Cancer Genome Consortium (ICGC)[21]. A wealth of bulk RNA-seq data has
59 become a legacy for biological and clinical research[22]. Thus, reanalysis of the enormous
60 amount of bulk data to explain both cellular diversity and spatial expression patterns
61 is a challenging but consequential task.

62 In silico methods have great potential to predict spatial heterogeneity from bulk
63 RNA-seq data at single-cell resolution by integrating cutting-edge technologies. Several
64 approaches such as CPM[23], CIBERSORT[24], and MuSiC[25] can only extrapolate
65 proportions of cell types from bulk RNA-seq data and have failed to further generate
66 single-cell data, let alone map them to tissue coordinates. The emergence of an
67 approach that can efficiently decompose bulk RNA-seq data into spatially resolved
68 single-cell expression profiles is expected to reveal the cell diversity of complex tissue
69 and the spatial expression variation simultaneously.

70 Herein, we introduce Bulk2Space, a spatial deconvolution algorithm based on deep
71 learning frameworks, which generates spatially resolved single-cell expression profiles
72 from bulk transcriptomes using existing high-quality scRNA-seq data and spatial
73 transcriptomics as references. We hypothesize that the process of scRNA-seq is similar
74 to sampling cells from bulk tissue, and each selected cell is labeled with a unique
75 barcode. Consequently, bulk transcriptome data can be used as a weighted collection
76 of single-cell expression data in a defined clustering space of cells. Bulk2Space first
77 generates single-cell transcriptomic data within the clustering space to find a set of
78 cells whose aggregated data is proximate to the bulk data. Next, the generated single
79 cells were allocated to optimal spatial locations using a spatial transcriptome reference.
80 For this step, we chose as a spatial reference one of the two most commonly used
81 spatially resolved transcriptomics technologies. One is an image-based method with
82 limited target genes[7-9], and the other is a spatial barcoding method without single-cell
83 resolution[11-14]. Taken all, Bulk2Space showed a robust performance across multiple
84 datasets and conditions and is an open-access algorithm on GitHub
85 (https://github.com/ZJUFanLab/bulk2space).

Results

Design concept of Bulk2Space

The overall design of the Bulk2Space algorithm is illustrated in Fig. 1, which is divided into two steps, deconvolution, and spatial mapping. Although bulk RNA-seq data are obtained by sequencing a mixture of cells, whereas scRNA-seq labels individual cells in advance, both methods share comparable cell types and states[23-26]. Thus, we assume that bulk transcriptomics data can be decomposed into single-cell transcriptomics data by a well-designed deconvolution algorithm (Step 1). Subsequently, in the second step, the above-mentioned two prevailing spatially resolved transcriptomics methods, though, failed in either achieving single-cell resolution or delineating the whole transcriptome, can provide reference locations for single cells generated in the first step of Bulk2Space based on the similarity of their expression profiles (Step 2)[27, 28]. As a consequence, it could be rationally conceived 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 (Fig. 1a).

For the deconvolution of bulk transcriptomics data, we hypothesize that the process of single-cell sequencing identifies with the process of sampling from the tissue, while the ratio between different cell types is related to the zone of the tissue, the ways of single-cell capturing, and the size of the cell itself. However, even if the ratio between cell types changes, the state of each cell type still fluctuates within a relatively stable high-dimensional space, which is namely the clustering space of cell types. This hypothesis has been supported by several studies[29-31]. As shown in Fig. 1b, a single-cell reference can be utilized to characterize the clustering space and average the gene expression for each cell type[31, 32]. The expression vector of the bulk transcriptome is taken as the input, which is equal to a product of the average gene expression matrix of cell types and their abundance vector. By solving this nonlinear equation, the proportion of each cell type is determined. Then, the solved proportion of each cell type is taken as a control parameter to generate the corresponding number of single cells. Specifically, a deep generative model, termed beta variational autoencoder (β-VAE)[33], is employed to simulate a given number of single cells within the characterized clustering space of each cell type. The simulation stops when the training loss no

118 longer reduces, thus deconvolves the heterogeneous bulk transcriptome into single-
119 cell transcriptome data. Instead of only calculating the abundance values of cell types,
120 as performed by other methods, such as Cell Population Mapping (CPM)[23],
121 CIBERSORT[24], and ImmuCC[26], Bulk2Space can generate biologically feasible single-cell
122 expression profiles.

123 For spatial mapping, the single-cell profiles generated in the first step are assigned
124 to the reference tissue section based on an optimization strategy. In the second step,
125 we consider two of the most commonly used spatially resolved transcriptomics
126 approaches as the spatial reference, namely, spatial barcoding-based RNA-seq
127 methods and image-based targeted methods.

128 Spatial barcoding-based methods employ arrayed reverse transcription primers
129 with unique positional barcodes to preserve the spatial information of mRNA via in
130 situ complementary DNA (cDNA) synthesis. Such methods could provide the whole
131 transcriptome as well as the location of each sequenced spot after decoding the spatial
132 barcodes. However, these methods are unable to achieve single-cell resolution,
133 because the spots are customized in shape and size, such as ST[14] (100 µm in diameter),
134 Visium (55 µm in diameter), Slide-seq[13] (10 µm in diameter), etc. As shown in Fig. 1c,
135 for spatial barcoding-based references, each barcoded spot is regarded as a mixture of
136 several cells. Similar to existing decomposition methods such as RCTD[34], SpatialDWLS[35],
137 stereoscope[36], and SPOTlight[37], we first calculate the cell-type composition of each
138 spot. The difference is that we then map the generated single cells into the
139 recommended spots based on the similarity of their cell expression profiles and ensure
140 that the proportions of cell types in each spot are consistent with the calculated results.
141 Since Bulk2Space generates spatially resolved single-cell transcriptomics data from
142 bulk RNA-seq, the spatial heterogeneity of individual cells can be analyzed, which
143 cannot be realized by RCTD, SpatialDWLS, stereoscope, and SPOTlight.

144 As illustrated in Fig. 1d, another strategy is to map the single-cell data generated in
145 the first step onto image-based targeted references capable of measuring hundreds to
146 thousands of RNA species[38], such as MERFISH[7], STARmap[9], seqFISH[10], etc. The pairwise
147 similarity between cells is calculated based on genes shared by both datasets, and
148 eventually, each generated single cell is robustly mapped to an optimized coordinate
149 of the target cell in the spatial reference. The resulting spatially resolved single-cell
150 RNA-seq data can provide unbiased transcriptomes of individual cells and improve the

151 gene coverage.

152 Taken all, unlike other computational methods that focus only on either of the two
153 spatially resolved transcriptomics approaches, Bulk2Space takes into account both
154 techniques to greatly increase the scalability, applicability, and reference data of the
155 algorithm.

156

157 Performance evaluation of Bulk2Space using simulated and biological 158 datasets

159 To demonstrate the robustness of the deconvolution step of Bulk2Space, a benchmark
160 test was performed using 30 paired simulations (both bulk and single-cell data were
161 generated from the same datasets) from 10 different high-quality single-cell datasets
162 across human blood, brain, kidney, liver, and lung, and mice brain, kidney, lung,
163 pancreas, and testis (Fig. S1a), and 12 unpaired simulations (both bulk and single-cell
164 data were generated from different datasets of mice pancreas) from 8 single-cell RNA-
165 seq data of human pancreas (Fig. S1b). All datasets used in this study were listed in
166 Supplementary Data 1.

167 First, for 30 paired simulations, each single-cell transcriptome dataset was divided
168 into two parts, one for reference and the other for the synthesis of bulk data. The
169 procedure for bulk data synthesis is shown in Fig. S1c. For each single-cell dataset, we
170 changed the proportion of cell types and synthesized 3 corresponding bulk
171 transcriptomics data with different cell compositions. Three generative deep learning
172 models, namely, the beta variational autoencoder (β-VAE)[33], generative adversarial
173 networks (GAN)[39], and conditional GAN (CGAN)[40], were introduced here for the
174 benchmarking of different methods using 30 paired single-cell datasets. The gene
175 expression correlation between the generated single-cell data and input bulk data was
176 calculated to evaluate and compare the three candidate algorithms. As illustrated in
177 Fig. 2a, β-VAE performed better than the other algorithms in both Pearson and
178 Spearman correlations with a lower root mean squared error (RMSE). The Pearson
179 correlation, Spearman correlation, and RMSE were calculated by combining the gene
180 expression across all cell types and averaging these metrics for each cell type. Ten
181 examples of different species and tissues are shown in Fig. S2. The results

182 demonstrated that the clustering space of the generated single-cell data was similar
183 to that of the test data but varied in the proportions of different cell types (Fig. S2a
184 and Fig. S2b). Furthermore, the correlation analysis of the expression of marker genes
185 for each cell type between the generated and test data was performed for each
186 method. The comparison of the correlation heatmaps between Bulk2Space, GAN, and
187 CGAN showed that Bulk2Space had higher correlations across different simulations
188 (Fig. S2c).

189 We then benchmarked the deconvolution step of Bulk2Space using 12 unpaired
190 simulation data of mice pancreas. Because other deconvolution methods, such as
191 CPM[23], CIBERSORT[24], and ImmuCC[26], can only predict cell-type proportions instead of
192 gene expression of generated data, we compared Bulk2Space with GAN, CGAN, and a
193 Bayesian deconvolution method termed bMIND[41]. As illustrated in Fig. 2b and Fig. S3a,
194 Bulk2Space outperformed the other three methods with a higher Pearson correlation
195 of gene expression and lower gene expression variation (RMSE). Although CGAN had
196 a comparable performance as Bulk2Space in the single-cell generation, its computing
197 speed was significantly lower than Bulk2Space. The detailed comparison of the cell-
198 type-specific marker gene expression correlations of generated single-cell data and the
199 ground truth between the four candidate methods were shown in Fig. S3b.

200 To further investigate the robustness of the single-cell generation of Bulk2Space,
201 we introduced two noise mechanisms to test the performance of the algorithm (Fig.
202 S4). One mechanism changes the expression values of certain genes in randomly
203 selected cells (Fig. S4a), and the other alters the cell type labels of selected cells (Fig.
204 S4b). Bulk2space showed robust performance and could effectively avoid the over-
205 fitting phenomenon (Fig. S4c and Fig. S4d). We also demonstrated that Bulk2Space can
206 deconvolve bulk RNA-seq data with an annotation-free single-cell reference in the
207 supplementary information Fig. S5. Since the core generation model of Bulk2Space is
208 β-VAE, a perturbation analysis was conducted to verify the robustness of single-cell
209 generation by Bulk2Space. The perturbation test confirmed that better generation
210 results can be obtained by assuming that the latent vectors follow the Gaussian
211 distribution (Fig. S6). Finally, we validated Bulk2Space with paired biological bulk and
212 single-cell data derived from different mice liver[42]. As shown in Fig. S7, the
213 deconvolution results indicated that Bulk2Space could be well applied in biological
214 scenarios.

215 In the second step, Bulk2Space assigns the generated single cells to spatial
216 coordinates based on a spatially resolved transcriptome reference (Fig. 1c and Fig. 1d).
217 There are two types of most common spatially resolved transcriptomic techniques.
218 One is the spatial barcoding-based RNA-seq technology, which includes 'spatial
219 transcriptomics' (ST)[14], high-definition spatial transcriptomics (HDST)[11], and Slide-
220 seq[12, 13], 10 Visium, etc. These methods are based on tissue microregions with certain
221 regular shapes (spots), and cannot achieve single-cell resolution. The other is image-
222 based in situ transcriptomics, including MERFISH[7], seqFISH[8, 10], and STARmap[9]. These
223 methods can only measure the expression of target genes but cannot cover the whole
224 transcriptome. To overcome the bottlenecks of the two approaches, we designed
225 corresponding spatial assignment strategies.

226 The first strategy was designed to map single cells to spots in spatially resolved
227 transcriptomics based on spatial barcoding. In this step, we first used paired synthetic
228 data to test the performance of Bulk2Space. The data synthesis procedure is illustrated
229 in Fig. S1c. Similar to the commonly used spatially resolved transcriptomics
230 technologies, such as ST, HDST, 10X Visium, and Slide-seq, we randomly chose 10 cells
231 from each scRNA-seq data and aggregated their gene expression profiles as a spot of
232 pseudo spatial transcriptomics data. A comprehensive comparison between
233 Bulk2Space (deep forest), six machine learning or deep learning approaches, and four
234 published methods, RCTD[34], SpatialDWLS[35], stereoscope[36], and SPOTlight[37] was then
235 conducted using 50 paired simulation datasets. These methods were benchmarked for
236 the optimization of cell-type composition and spatial gene expression patterns. As
237 shown in Fig. 2c, compared with these methods, Bulk2Space showed higher
238 correlations in gene expression and lower RMSEs than other methods. The cell
239 composition and spatial cell-type distribution predicted by Bulk2Space were highly
240 correlated to the ground truth (Fig. S8). Next, we benchmarked the spatial mapping
241 step of Bulk2Space using unpaired simulation data of mice pancreas. As shown in Fig.
242 2d, Bulk2Space showed the highest correlations in gene expression and the lowest
243 RMSEs. The comparison of the cell-type composition for each spot between
244 Bulk2Space and the spatial reference was shown in Fig. S9.

245 The second spatial mapping strategy uses image-based targeted methods. A
246 scRNA-seq data ("GSE113576
247 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113576]") and MERFISH
248 data (image-based reference) of the mouse hypothalamus tissue from the same

249 experiment[43] were used as paired datasets to test the performance of the Bulk2Space
250 algorithm. Another scRNA-seq data[44] derived from a separated experiment was used
251 as unpaired single-cell data. We wanted to find the shortest path that could shift single
252 cells to optimal locations. First, a nonparametric empirical Bayes network was used to
253 eliminate the expression differences between the single-cell data and spatially
254 resolved transcriptome data. As illustrated in Fig. 2e, the clustering spaces of the two
255 datasets were very similar, with most cell types sharing the distribution in both
256 datasets. The two discrete datasets could be well merged in distribution, and the gene
257 expression of each cell type had a correlation of over 0.8 between the two datasets
258 (Fig. 2f). Of the 150 target genes identified in MERFISH, a leave-one-out test was
259 conducted for the evaluation of the spatial mapping step of Bulk2Space. Specifically,
260 149 were used as reference genes, and the rest one was used for validation. The results
261 showed that spatially, the gene expression pattern of the matched cells was highly
262 correlated with that of the reference, and the predicted spatial expression patterns
263 across all 150 targeted genes are strongly correlated with the ground truth with
264 averaged Pearson correlation coefficients (PCCs) over 0.9 for both paired and unpaired
265 datasets (Fig. 2g and Fig. S10).

266 We further verified Bulk2Space's spatial mapping using five-fold cross-validation
267 for image-based methods (Fig. S11). specifically, the 150 targeted genes in MERFISH
268 were split into five folds using R package caret (version 6.0-92), and 80% of genes were
269 used as the reference and the remaining was used for validation of predicted spatial
270 expression. As shown in Fig. S11, the spatial expressions of the top 25 genes predicted
271 by Bulk2Space for the paired and unpaired single-cell data were compared with
272 MERFISH data.

273 To validate the spatial mapping step of Bulk2Space in the real situation, two
274 biological datasets, a spatial reference sequenced by Slide-seq v2[12], and a single-cell
275 data[45] from separated experiments were used to reconstruct the structure of the
276 mouse hippocampus (Fig. S12). The spatial distribution of single cells predicted by
277 Bulk2Space was consistent with the real pattern of cell types in that region. Notably,
278 the refined subregions in the mouse hippocampus were successfully reconstructed by
279 Bulk2Space, which was confirmed by the spatial expression of cell-type-specific
280 marker genes (Fig. S12e).

281 Since Bulk2Space employed β-VAE to generate single-cell profiles within the

clustering space of cell types, we then investigated whether the randomness of β-VAE could affect the generation of single-cell data and the spatial mapping results. We conducted 100 repetitions on the deconvolution step of Bulk2Space to evaluate the robustness of β-VAE using simulated bulk and single-cell reference[46]. The results demonstrated that the single-cell generation remained highly robust across 100 repetitions (Fig. S13).

We then used simulated PDAC[47] and melanoma[48, 49] bulk data to evaluate the robustness of the spatial mapping step of Bulk2Space using three repetitions of bulk data deconvolution and spatial mapping. As shown in Fig. S14 and Fig. S15, the spatial deconvolution results suggested a robust performance of Bulk2Space with 3 repetitions for PDAC and melanoma bulk data, respectively.

In conclusion, although the single-cell data generated by β-VAE were slightly different each time, the overall prediction results showed robust performance in the spatial distribution of cell types, the cell-type composition and proportion in spots, and the spatial patterns of gene expression.

Validation of Bulk2Space using biological data

To further verify the performance of the Bulk2Space, two consecutive slices[49] from the same melanoma tissue were used to demonstrate the spatial deconvolution result. As shown in Fig. S16, one slice was used as the spatial reference and the other was used to synthesize the bulk data. The expression of cell-type-specific marker genes between the generated data and the reference were highly correlated and the spatial distribution pattern of generated single cells was consistent with the histological annotations of different regions. Moreover, we found a spatial heterogeneity of generated B cells from different tissue regions, which was associated with the biological functions in the lymph node area and the melanoma region. Because Bulk2Space can predict spatially resolved single-cell transcriptomics data from bulk RNA-seq or scRNA-seq, spatial heterogeneity of the same cell type can be discovered. This was difficult for other spatial deconvolution algorithms to achieve such as RCTD[34], SpatialDWLS[35], stereoscope[36], and SPOTlight[37].

Similarly, the same spatial deconvolution procedure of Bulk2Space was performed for another two discrete slices from different PDAC[47] tissues to reveal the spatial heterogeneity of bulk transcriptomics data. One slice was used as the spatial reference and the other was used to synthesize the bulk data. As shown in Fig. S17, similar results were retrieved as the validation of melanoma data. The expression of cell-type-specific marker genes between the generated data and the reference were highly correlated and the spatial distribution pattern of generated single cells was consistent with the histological feature of tissue regions.

Bulk2Space integrates spatial gene expression and histomorphology in PDAC

Based on the success in the spatial deconvolution of PDAC using simulated data and discrete slices, we further verified the performance of the Bulk2Space algorithm using biological bulk RNA-seq data. As shown in Fig. 3a, two bulk RNA-seq data[50] of the pancreatic adjacent (PA) and the PDAC tissues, one scRNA-seq data[47], and one spatial transcriptomics data[47] were derived from two individual experiments and three different technologies. The scRNA-seq data sequenced by inDrop-seq[51] were used as the single-cell reference for the deconvolution of PA tissue and PDAC bulk data, and the spatial barcoding-based ST data were used as the spatial reference.

The PA and PDAC bulk data were deconvolved into single-cell data with different cell-type proportions by Bulk2Space. We compared the cell-type proportions of single cells generated from PA and PDAC bulk data with Bulk2Space. As shown in Fig. 3b, the proportion of acinar cells in PA was significantly higher than that in PDAC, where the proportion of two cancer clone subtypes was higher. The clustering space of single cells generated from PDAC and PA tissue was exhibited in Fig. 3c. The detailed information for the deconvolution of PA tissue was shown in Fig. S19.

Here, we focused on the PDAC bulk data. The expression of cell-type-specific marker genes between generated single-cell profiles by Bulk2Sapce and the reference data was highly correlated (Fig. 3d). Then, each cell was mapped to coordinates based on the spatial reference (Fig. 3e). To further correlate the spatial expression pattern of the tissue with its histological feature, we investigated the molecular architecture of the annotated regions in the spatial reference (Fig. 3f). As shown, the spatial distribution of two cancer clones, acinar cells, and ductal cells predicted by Bulk2Space (Fig. 3e) was consistent with the histological annotation of cancer cells and desmoplasia, normal pancreatic tissues, and duct epithelium, as well as the spatially resolved transcriptomics of the cancer region, pancreatic tissue and stroma, and the duct epithelium, respectively. The spatial cell-type proportions predicted by Bulk2Space were illustrated in Fig. 3g and Fig. S19c. The spatial expression of cell-type-specific marker genes at spot (left) and single-cell (right) resolution for different cell types were shown in Fig. 3h and Fig. S19d. Compared with the spatial distribution of single cells generated by Bulk2Space, the spatial expression of cell-type-specific marker genes

353 exhibited consistent patterns.

354 Bulk2Space reconstructs the hierarchical structure of the mouse 355 isocortex region sequenced by Spatial-seq

356 In addition to linking the histomorphology and transcriptomics in pathological tissues,
357 another application scenario of Bulk2Space is to reconstruct the structure of tissues
358 with spatial patterns. For instance, the spatial organization of the isocortex region of
359 the mouse brain exhibits a layered pattern. Therefore, Bulk2Space was applied to
360 reconstruct the hierarchical structure of the mouse isocortex region through spatial
361 deconvolution of the bulk transcriptomics data. The bulk transcriptomics data of
362 mouse isocortex were sequenced by our in-house developed multiplexed RNA-seq
363 approach, termed Spatial-seq. The detailed information for Spatial-seq was described
364 in the 'Methods' section and illustrated in Fig. S20a. In short, the laser capture
365 microdissection (LCM) was used to isolate regions of interest from the tissue sections
366 and each isolated tissue was collected independently. Then, barcoded beads were
367 used to capture and label the mRNA from the collected samples. Finally, the captured
368 mRNA was pooled together for RNA-seq. The transcriptome data of each sample can
369 be obtained by identifying the sample-specific barcode sequence. Using Spatial-seq,
370 we isolated and sequenced 13 main brain regions from coronal and sagittal sections
371 across the entire mouse brain (Fig. S20b).

372 As shown in Fig. 4a, the bulk transcriptomics data of the mouse isocortex was
373 obtained by Spatial-seq. The mouse primary visual cortex regions in different coronal
374 sections from anterior to posterior were collected and sequenced by SMART-seq2[52].
375 The result scRNA-seq data were used as the single-cell reference. A sagittal section of
376 the mouse brain was divided into two parts and sequenced using 10X Visium to obtain
377 spatial transcriptomics data of the isocortex region. The spatial reference data were
378 downloaded from 10X datasets. The detailed information of datasets used in this study
379 was summarized in Supplementary Data 1.

380 The bulk data were first deconvolved into single-cell RNA-seq data (Fig. 4b).
381 Subsequently, the generated single cells were mapped to spatial locations by
382 Bulk2Space and the spatial distribution of cell types showed a distinct layered pattern
383 (Fig. 4c). The expression of cell-type-specific marker genes was highly correlated

384 between Bulk2Space results and the single-cell reference data (Fig. 4d). The spatial
385 distribution of the cell-type proportion in each spot predicted by Bulk2Space was
386 illustrated in Fig. 4e, which corresponded well with the hierarchical structure of the
387 mouse isocortex. The spatial distributions of cell-type proportions predicted by
388 Bulk2Space for seven cell types including Astro, L2/3 IT, L4, L5 IT, L5 PT, L6 CT, and L6 IT
389 cells were shown in Fig. 4f. The results supported the ability of Bulk2Space to
390 reconstruct the spatial organization of tissues. Moreover, the spatial expression
391 patterns of cell-type-specific marker genes predicted by Bulk2Space at single-cell
392 resolution were consistent with the proportion distributions of the corresponding cell
393 types (Fig. 4g).

394 Different from traditional spatial deconvolution algorithms such as RCTD[34],
395 SpatialDWLS[35], stereoscope[36], and SPOTlight[37], Bulk2Space can generate spatially
396 resolved single-cell transcriptomics data from bulk RNA-seq. Besides, Bulk2Space
397 allowed us to analyze the spatial heterogeneity of individual cells from the spatially
398 resolved transcriptomics data without single-cell resolution.

399 Bulk2Space re-annotates ambiguous cells in the mouse hypothalamus

400 The Bulk2Space algorithm was also used to spatially deconvolve bulk transcriptome
401 data derived from the hypothalamus region of the mouse brain using Spatial-seq, to
402 explain the spatial distribution of single cells and gene expression. The resources of
403 the bulk, single-cell, and spatial data was illustrated in Fig. 5a. The single-cell reference
404 was obtained by profiling ~31000 cells using Drop-seq and the MERFISH data of the
405 mouse hypothalamus were used as the spatial reference[43]. More details were
406 summarized in Supplementary Data 1.

407 After deconvolution of the bulk RNA-seq data of the mouse hypothalamus by
408 Bulk2Space, the clustering space of the generated single cells was close to that of the
409 MERFISH data (Fig. 5b). The generated single cells were then mapped to spatial
410 coordinates based on the image-based spatial reference (Fig. 5c). As shown in Fig. 5c,
411 the cell-type distributions in MERFISH data and Bulk2Space results were comparable
412 at single-cell resolution. Among the 10 cell types, immature oligodendrocytes were
413 unique to MERFISH, and macrophages were predicted by Bulk2Space, but absent in
414 MERFISH data. The heatmap suggested that 7 of the 8 shared cell types were strongly
415 correlated in both MERFISH and Bulk2Space results with an average pairwise

416 correlation over 0.9. The poor correlation of mural cells between MERFISH data and
417 Bulk2Space results may be related to the few numbers of pericytes. The spatial
418 patterns of the MERFISH targeted genes were consistent with that of Bulk2Space
419 results (Fig. 5d).

420 Meanwhile, owing to the limited number of measured target genes, some
421 ambiguous cells in MERFISH data could not be further identified according to their
422 expression profiles. However, using Bulk2Space, the generated single cells with
423 unbiased gene expression were assigned to the spatial context, thus further clustering
424 these ambiguous cells using novel genes that were absent in MERFISH data (Fig. 5e).
425 Meanwhile, the spatial distribution of novel genes beyond the targeted RNA species
426 was predicted by Bulk2Space and was consistent with the spatial pattern of
427 corresponding cell types as illustrated in Fig. 5f and Fig. S21.

428 Bulk2Space uncovers spatial gene expression dynamics in different 429 stages of the inflammation-induced prostate cancer

430 Tumor development is linked to chronic infection, dietary factors, obesity, inhaled
431 pollutants, tobacco use, and autoimmunity. The unifying principle underlying these
432 processes is inflammation, which is an aberrantly prolonged form of a protective
433 response to a loss of tissue homeostasis[53]. To explore whether Bulk2Space could
434 identify the cellular and molecular pathways that coordinate the tumor-promoting
435 effects in inflammation-induced cancer, we deconvolved the bulk transcriptome data
436 of prostate cancer and mapped single cells generated by Bulk2Space to the
437 corresponding spatial references, including normal glands, inflammatory tissue, and
438 tumor sites. The resources of the bulk (from TCGA), single-cell[54], and spatial data[55]
439 were listed in Supplementary Data 1.

440 These tissue regions were isolated from the prostate, and three of the sections
441 were subjected to RNA-seq based on spatial barcoding (Fig. 6a). Next, Bulk2Space was
442 used to deconvolve the bulk data into single-cell profiles (Fig. 6b) and mapped the
443 generated single cells to spatial coordinates (Fig. 6c). Notably, the composition and
444 distribution of cells in different tissue regions were quite different and closely related
445 to the state of the tissue (Fig. 6d and 6e). A comparison of cell distribution between
446 normal, inflammatory, and cancerous tissues (Fig. 6f to 6h) showed that in normal

447 glands, cancer-associated fibroblasts (CAFs) were the most abundant, but in
448 inflammatory and cancer tissues, the proportions of CAFs were significantly lower
449 (52.4% in normal, 20.5% in inflammation, and 27.5% in cancer). In the areas of
450 inflammation, cells involved in the inflammatory response, such as immunocompetent
451 B cells, endothelial cells, monocytes, and NK cells (Fig. 6g), were more abundant than
452 in normal or tumor tissues (13.4% in inflammation, 2.0% in normal, and 1.7% in cancer).
453 In cancer with a Gleason score (Gs) of 3+4, the proportion of cancer epithelial cells was
454 significantly higher (53.0% in cancer, 6.3% in normal, and 4.3% in inflammation) than
455 that in normal and inflammatory areas (Fig. 6h), while the proportion of the
456 periventricular layer cells was significantly lower (17.9% in cancer, 39.3% in normal,
457 and 44.1% in inflammation).

458 During the progress of inflammation-induced prostate cancer, CAFs accumulated
459 in normal glands at the early stage, resulting in high expression of inflammatory factors
460 and promoting the occurrence of local inflammation, a fundamental innate immune
461 response to perturbed tissue homeostasis, thus leading to tissue cancerization[53].
462 Bulk2Space results were consistent with this phenomenon.

463 Discussion

464 Understanding the transcriptional heterogeneity within the tissue from the
465 perspective of single-cell spatial resolution and full gene coverage is an essential
466 direction for the development in the field of biological sciences[1, 56]. The spatial
467 locations of cells may determine their identity and how they interplay with each other
468 in the microenvironment[57, 58]. However, there are many technical challenges, and thus,
469 such a technology has not been fully realized yet. Traditional bulk transcriptome data
470 can provide gene expression patterns and reflect the overall status of tissue, which is
471 widely used in studies on the occurrence and development of diseases. Nevertheless,
472 disease-inherent cell composition and spatial distribution are difficult to analyze.
473 Although several attempts can infer the cell-type proportions from traditional RNA-
474 seq data, understanding bulk tissue at single-cell spatial resolution remains a pressing
475 need in the field. Cell identity can be characterized by the clustering space of single-
476 cell gene expression profiles[59], and expression features can be stably retained across
477 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 deconvolving the bulk transcriptomics into spatially resolved single-cell transcriptomics data.

As thoroughly considered the merits and demerits of the widespread use of 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 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 providing a spatially resolved single-cell deconvolution strategy for these spatial data. 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 predicting the molecular mechanism underlying the progression of the disease.

Methods

Datasets

All scRNA-seq, spatially resolved transcriptomics, and bulk RNA-seq datasets used in this study were collected from high-quality publications, Gene Expression Omnibus (GEO), and The Cancer Genome Atlas (TCGA), wherein unannotated, ambiguous, or low quality-cells were excluded. The detailed description of each dataset was summarized in Supplementary Data 1.

Computing environment

The Bulk2Space was developed on two workstations, as listed below.

Workstation 1: Dell Precision Tower 7820 Workstation, CPU (Intel Xeon Gold 5118, 2.3 GHz × 2), RAM (64 GB, 16 GB × 4, DDR4, 2933 MHz), Hard Drive (SSD, SATA Class 20, 512 GB; HDD, 7200 rpm, SATA), Graphics Card (NVIDIA, Quadro P4000, 8 GB), Operating System (Ubuntu 16.04), Running Environment (CUDA 11.6, Torch 1.12.1, Python 3.8.5, deep-forest 0.1.5, easydict 1.9, numpy 1.19.2, pandas 1.1.3, scanpy 1.8.1, scikit-learn 1.0.1, scipy 1.5.2, tqdm 4.50.2, Unidecode 1.3.0).

Workstation 2: Dell Precision Tower 7920 Workstation, CPU (Intel Xeon Gold 6230, 2.1 GHz × 2), RAM (192 GB, 16 GB × 12, DDR4, 2933 MHz), Hard Drive (SSD, SATA Class 20, 512 GB; HDD, 7200 rpm, SATA), Graphics Card (NVIDIA, RTX2080Ti VIDEO CARD V2 × 2, 22 GB), Operating System (Ubuntu 18.04), Running Environment (CUDA 11.0, Torch 1.7.1, Python 3.8.5, deep-forest 0.1.5, easydict 1.9, numpy 1.19.2, pandas 1.1.3, scanpy 1.8.1, scikit-learn 1.0.1, scipy 1.5.2, tqdm 4.50.2, Unidecode 1.3.0).

Data processing

All data were preprocessed using R (version 4.1.1). For mouse brain scRNA-seq data by Moffitt[43] et al., we filtered 841 ambiguous cells and 88 unstable cells. For human prostate cancer scRNA-seq data by Wu[54] et al., we filtered 183 unassigned cells. All cells of other scRNA-seq as well as spatially resolved transcriptomics datasets were retained. For mouse hypothalamus MERFISH data, we filtered 5 “blank” barcodes as well as *the Fos* gene, whose expression value in all cells was ‘NA’. For human prostate cancer RNA-seq data, we used biomaRt (version 2.48.3) to transform ensemble id to gene name. For all scRNA-seq and spatially resolved transcriptomics datasets, the raw counts were normalized using the global-scaling normalization method “LogNormalize” by Seurat (version 4.0.4)[60]. For the 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:

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

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.

$$\min_{\boldsymbol{p}} \|\mathbf{x} - \mathbf{C}\boldsymbol{p}\|_2 \quad (2)$$

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.

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:

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

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

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

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

$$= -\mathcal{L}(\theta,\phi;\mathbf{x}) + \log\big(p_\theta(x)\big) \quad (3)$$

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

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

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

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

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(p_\theta(\mathbf{x}|\mathbf{z})\right)\right] - \mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right) \quad (4)$$

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:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] \ \textit{subject to}\ \mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right) < \epsilon \quad (5)$$

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

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{x}|\mathbf{z}))] - \beta\left(\mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right) - \epsilon\right) \quad (6)$$

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:

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

$$\beta\left(\mathcal{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right)\right) \quad (7)$$

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 learned 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 learned 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.

Model Configuration. In step1, we simulate single-cell by β-VAE. Both the encoder and decoder apply a four-layer perceptron, where each layer is followed by a RELU activation except the last layer of the encoder. For the encoder, the number of neurons in each layer is 2048, 1024, 512 and 512 respectively. While for the decoder, the number of neutrons in each layer is 512, 1024, 2048 and k respectively, where k represents the number of genes in the dataset. The dimensionality of the latent space learned by the beta-VAE is 256. The relative weighting of the reconstruction loss and regularization loss is 1:4. We apply the Adam with weight decay (AdamW) optimizer with an initial learning rate as 1e-4, the decoupled weight decay as 5e-4 and Adam's β parameters as 0.9 and 0.999. Moreover, the default running epoch is fixed to 3000, but we use early stopping during the training phase, with which we stop training when the training loss no longer reduces for 50 epochs.

Here we introduce the implementation details of our benchmark approaches. For GAN and cGAN, the generator consists of two fully connected layers, followed by LeakyRELU and RELU activations respectively. The discriminator is also made up of two fully connected layers, where the first layer is followed by a LeakyRELU activation. We use the Adam optimizer with the initial learning rate of 1e-4 and the betas parameters of 0.5 and 0.999. We train util the loss in the generative and discriminative phases is no longer reduced for 50 epochs, and use the model obtained at this time as the final model for prediction.

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 100 trees. We concatenated the output of these two forests as the transformed feature

671 vectors, which also acted as the input of cascade forest. The final model was a cascade
672 of cascades, where each cascade consisted of multiple levels each corresponding to a
673 grain of scanning. Each level consisted of 4 completely-random tree forests, each
674 containing 100 trees. In other words, the transformed feature vectors were
675 augmented with the class vector generated by the previous grade, and then were used
676 to train the current grade of cascade forests.

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

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

686 *Image-based in situ hybridization*. We developed a simple method for predicting the
687 spatial distributions of genes not measured in spatial transcriptomic data which
688 produced by in situ RNA imaging-based technologies. To do this, Bulk2Space used the
689 spatial transcriptomic data as the reference, and generated scRNA-seq data were
690 mapped to the tissue space corresponding to this reference. To eliminate the
691 differences between scRNA-seq data and spatial transcriptomic data caused by
692 different experiment types, an empirical Bayes framework was used to remove the
693 batch effect, so that both sets of data are at the same scale level. Subsequently, the
694 cross-dataset k-nearest neighbor graph of spatial transcriptomic data in scRNA-seq
695 data was computed in the aligned space, and then the predicted whole gene
696 expression profile of each cell of spatial transcriptomic data could be calculated to the
697 mean of its k nearest neighbors in scRNA-seq data.

698 *Model Configuration.* Deep forest is applied for spot recommendation. Specifically, the
699 number of samples used to construct feature discrete bins is set to 200000. If the size
700 of training set is smaller than it, then all training samples will be used. The type of
701 binner used to bin feature values into integer-valued bins is “percentile”, which means
702 each bin will have approximately the same number of distinct feature values. We set

the maximum number of cascade layers in the deep forest as 20, and apply 2 estimator in each cascade layer. Gini impurity is used to measure the quality of a split. We have no constraints on the maximum depth of each tree. The training process terminates when the validation performance on the training set does not improve compared against the best validation performance achieved so far to 2 tolerant rounds. And the counting on tolerant rounds is triggered if the performance of a fitted cascade layer does not improve by 1e-5 compared against the best validation performance achieved so far.

We also introduce the implementation details of our benchmark approaches. (1) For Logistic Regression (LR), we implement L2 regularization as the additional penalty term to solve the problem of overfitting. We use L-BFGS algorithm as the solver, which uses the Hessian matrix to iteratively optimize the loss function. We fit the model according to the given training data and return the probability estimates on testing set. (2) For Decision Tree (DT), we use the Gini impurity to measure the quality of a split and choose the best split at each node. We don't set any maximum depth of the tree, so nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples. We build the decision tree classifier from the training set, and then predict class probabilities of the input samples. (3) For Gradient Boosting Decision Tree(GBDT), we set the learning rate as 0.1 and the number of boosting stages to perform as 100. The loss function to be optimized is log loss function, and we choose Friedman MSE to measure the quality of a split. The maximum depth of the individual regression estimators is set to 3. We fit the gradient boosting model on the training set and use the trained model to make predictions on testing set. (4) For Multilayer Perceptron(MLP), we apply a two-layer perceptron, where each layer is followed by a batch normalization, a RELU activation, and a dropout layer with the probability of element zeroing as 0.1. We use the Adam optimizer with the initial learning rate of 1e-4 and the betas parameters of 0.9 and 0.999. We train until the loss is no longer reduced for 30 epochs, and use the model obtained at this time as the final model for prediction. (5) For DeepGBM, we set the number of tree groups to 100. The dimension of leaf embedding for a tree group is set to 20. The structure of the distilled NN model is a fully connected network with "100-100-100-50" hidden layers. We adopt the feature selection in each tree group, where we first sort the features according to the information gain, and the top 128 of them are selected as the inputs of distilled NN model. We use the Adam with weight decay (AdamW) optimizer with an initial learning rate as 2e-3, the decoupled weight decay as 1e-6, and Adam's betas parameters as 0.9

and 0.999. We train until the model reaches the highest ROC-AUC score, and use it for prediction.

Performance evaluation of Bulk2Space (Simulated datasets)

Benchmarking the deconvolution step of Bulk2Space on simulated datasets

1. Simulated bulk data and the single-cell reference were generated from same datasets (paired simulations)

Datasets. Ten scRNA-seq datasets were applied to benchmark our method, including human and mouse primary tissues. Five human scRNA-seq datasets including peripheral blood[46] ("GSE92495 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495]"), brain[61] ("GSE103723 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103723]"), kidney[62] ("GSE121862 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121862]"), liver[63] ("GSE124395 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395]"), and lung[64] ("GSE130148 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130148]"). Five mice scRNA-seq datasets including brain[65] ("GSE60361 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361]"), kidney[66] ("GSE119531 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119531]"), lung[67] ("GSE127465 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127465]"), pancreas[68] ("GSE84133 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133]"), and testis[69] ("GSE112393 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112393]").

Data simulations. The simulated data were derived from the above 10 scRNA-seq datasets. We randomly divided the scRNA-seq data into two parts, one was treated as the single-cell reference (data_2) and the other (data_1) for the construction of bulk transcriptome data via aggregating all the single-cell gene expression profiles. Considering the cell composition of bulk RNA-seq data varies greatly in the natural situation, for each data_1, we further changed the cell-type proportion and

synthesized 3 corresponding bulk transcriptome data with different cell compositions. In total, 30 paired simulated data were synthesized in this study. The detailed description of the experimental design was summarized in Supplementary Data 2.

Comparied methods. **a**, Generative adversarial networks (GAN), **b**, conditional generative adversarial networks (CGAN). The generator consists of two fully connected layers, followed by LeakyRELU and RELU activations respectively. The discriminator is also made up of two fully connected layers, where the first layer is followed by a LeakyRELU activation. We use the Adam optimizer with the initial learning rate of 1e-4 and the betas parameters of 0.5 and 0.999. We train until the loss in the generative and discriminative phases is no longer reduced for 50 epochs, and use the model obtained at this time as the final model for prediction.

Benchmark metrics. Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC), and root mean squared error (RMSE) were used to assess the similarity of the gene expression profile between scRNA-seq data reference and synthetic bulk transcriptome data.

2. Simulated bulk data and the single-cell reference were generated from different datasets (unpaired simulations)

Datasets. Eight human pancreas scRNA-seq datasets from different resources were also applied to benchmark our method, including 1 CelSeq[70] ("GSE81076 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81076]"), 1 CelSeq2[71] ("GSE85241 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85241]"), 1 Fluidigm C1[72] ("GSE86469 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86469]"), 4 inDrops[68] ("GSE84133 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133]"), and 1 SMART-Seq2[73] (E-MTAB-5061).

Data simulations. The simulated data were derived from the above 8 scRNA-seq datasets. We randomly selected one as the single-cell reference and another dataset from different resources as bulk transcriptome data. The detailed description of the experimental design was summarized in Supplementary Data 3.

Comparied methods. **a**, Generative adversarial networks (GAN). **b**, conditional

generative adversarial networks (CGAN). The model details were described above. **c**, bMIND[41]. We followed the guidelines on the bMIND GitHub repository: https://github.com/randel/MIND. We set the cell type proportion parameter "frac" to the output value of Bulk2Space to ensure the generated gene expression profile comparability.

Benchmark metrics. The benchmark metrics were described above.

Evaluating the robustness of the deconvolution step of Bulk2Space

we evaluated the robustness of Bulk2space by introducing two kinds of noise: cell type noise and gene expression noise. For cell type noise, we randomly altered the 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 expression noise, we firstly constructed a noise expression profile $\boldsymbol{C}$,

$$\boldsymbol{C} \sim \mathrm{U}(-x, x) \quad (8)$$

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

We next constructed synthetic expression matrix by combining noise expression profile with scRNA-seq data reference, and the ratio of gene expression-altered cells remains 0.01, 0.02, 0.04, 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 matrix and constructed bulk RNA-seq data were used to evaluate performance.

Benchmarking the spatial mapping step of Bulk2Space on simulated datasets

1. Spatial barcoding-based data

Datasets. Eighteen datasets metioned above were applied to benchmark our method. 10 datasets were applied to construct single-cell data and spatial reference from same datasets, and other 8 datasets were applied to construct single-cell data and spatial reference from different datasets.

Data simulations. Starting from scRNA-seq data reference, we randomly chose 10 cells from it and aggregated their transcriptomic profiles as a spot of pseudo spatial transcriptomic data. The spot with over 25000 UMI counts would be sampled down to

825 20000 UMI counts to better meet the true situation. We also constructed pseudo
826 spatial transcriptomic data with 100, 200, 500, 1000, and 5000 spot numbers to
827 simulate true spatial transcriptomics data produced by different spatial barcoding
828 technologies. Besides, the simulation that the scRNA-seq and spatial datasets are from
829 different sources was also considered. The detailed description of the experimental
830 design was summarized in Supplementary Data 4 (paired simulation) and
831 Supplementary Data 5 (unpaired simulation).

832 *Comparied methods.* **a**, Logistic Regression (LR). We implement L2 regularization as
833 the additional penalty term to solve the problem of overfitting. We use L-BFGS
834 algorithm as the solver, which uses the Hessian matrix to iteratively optimize the loss
835 function. We fit the model according to the given training data and return the
836 probability estimates on testing set. **b**, Decision Tree (DT). We use the Gini impurity to
837 measure the quality of a split and choose the best split at each node. We don't set any
838 maximum depth of the tree, so nodes are expanded until all leaves are pure or until all
839 leaves contain less than 2 samples. We build the decision tree classifier from the
840 training set, and then predict class probabilities of the input samples. **c**, Gradient
841 Boosting Decision Tree (GBDT). We set the learning rate as 0.1 and the number of
842 boosting stages to perform as 100. The loss function to be optimized is log loss function,
843 and we choose Friedman MSE to measure the quality of a split. The maximum depth
844 of the individual regression estimators is set to 3. We fit the gradient boosting model
845 on the training set and use the trained model to make prediction on testing set. **d**,
846 Multilayer Perceptron (MLP). We apply a two-layer perceptron, where each layer is
847 followed by a batch normalization, a RELU activation, and a dropout layer with the
848 probability of element zeroing as 0.1. We use the Adam optimizer with the initial
849 learning rate of 1e-4 and the betas parameters of 0.9 and 0.999. We train until the loss
850 is no longer reduced for 30 epochs, and use the model obtained at this time as the
851 final model for prediction. **e**, DeepGBM. We set the number of tree groups to 100. The
852 dimension of leaf embeddin for a tree group is set to 20. The structure of the distilled
853 NN model is a fully connected networks with "100-100-100-50" hidden layers. We
854 adopt the feature selection in each tree group, where we first sort the features
855 according to the information gain, and the top 128 of them are selected as the inputs
856 of distilled NN model. We use the Adam with weight decay (AdamW) optimizer with
857 an initial learning rate as 2e-3, the decoupled weight decay as 1e-6, and Adam's betas
858 parameters as 0.9 and 0.999. We train until the model reach the highest ROC-AUC
859 score, and use it for prediction. **f**, RCTD[34]. We followed the guidelines on the RCTD

GitHub repository: https://github.com/dmcable/spacexr. We set the parameter doublet_mode = 'full'. **g**, spatialDWLS[35]. We followed the guidelines on the spatialDWLS GitHub repository: https://github.com/rdong08/spatialDWLS_dataset/tree/main/codes. We set the parameter n_cell = 20. **h**, stereoscope[36]. We followed the guidelines on the stereoscope GitHub repository: https://github.com/almaan/stereoscope. We set the parameter sc epochs = 10000, st epochs = 10000. **i**, SPOTlight[37]. We followed the guidelines on the SPOTlight GitHub repository: https://github.com/MarcElosua/SPOTlight. We set the parameter n_cells = 75.

Benchmark metrics. Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC), and root mean squared error (RMSE) were used to assess the similarity of the gene expression profile per spot between the predict result and ground truth. Due to RCTD, spatialDWLS, stereoscope, and SPOTlight could not obtain gene expression profiles for each spot, we calculated gene signature matrix of each cell type based on single-cell reference, and the product of the predicted proportions of each cell type and the gene signature matrix of each cell type was regarded as the gene expression profile of each spot.

2. Image-based in situ hybridization data

Datasets. A mouse hypothalamus scRNA-seq data ("GSE113576 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113576]") and MERFISH data (Bregma +0.26) (image-based reference) from the same laboratory[43] were applied to test the performance of the Bulk2Space algorithm for the second spatial mapping strategy. We also applied another mouse hypothalamus scRNA-seq data[44] ("GSE87544 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87544]") from different source as single-cell reference to confirm the robustness of Bulk2Space.

Benchmark metrics. Pearson correlation coefficient (PCC) was used to assess the similarity of the gene expression profile between the predict result and ground truth. We took two strategies, five-fold cross validation and leave-one-out cross validation, to validate performance of Bulk2Space. In five-fold cross validation, we random split 150 target genes identified in MERFISH into five folds using R package caret (version

891 6.0-92), a split of 80% for reference and 20% for validation. In leave-one-out cross
892 validation, 149 were used as reference genes and and one was used for validation.

893 Performance evaluation of Bulk2Space using biological datasets

894 *Deconvolution performance evaluation using paired bulk and single-cell datasets*

895 Three paired mouse liver bulk and single-cell datasets[74] ("GSE119340
896 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119340]") were applied to
897 evaluate the performance of the deconvolution step of Bulk2Space.

898 *Reconstruction of mouse hippocampus subregions at single-cell resolution*

899 The mouse hippocampus Slide-seq v2[12] and scRNA-seq data[45] were downloaded from
900 the Seurat website: https://satijalab.org/seurat/articles/spatial_vignette.html were
901 applied to evaluate the performance of the spatial mapping step of Bulk2Space on real
902 single-cell data and real spatial data. We down sampled Slide-seq v2 data to 5000 spots
903 to speed the spatial mapping step. For Bulk2Space, we set the parameter k = 2 and
904 top_marker_num = 100, all other parameter followed the default values. For Seurat,
905 we followed the standard analysis workflow on the Seurat website and set the
906 parameter resolution = 0.3 to perform unsupervised clustering.

907 *Robustness evaluation of Bulk2Space using repeated data*

908 We conducted 100-time repetitions on the deconvolution step of Bulk2Space to
909 evaluate the robustness of β-VAE using the human peripheral blood[46] ("GSE92495
910 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495]") scRNA-seq data.
911 The construction of the single-cell reference and bulk transcriptome data was the
912 same as mentioned above. We further evaluated the spatial mapping results using the
913 pancreatic ductal adenocarcinoma (PDAC) dataset[47] and the human melanoma
914 dataset[48, 49] with 3-time repetitions. For Bulk2Space, all parameters followed the
915 default values.

916 *Deconvolution of bulk RNA-seq data using annotation-free single-cell reference by*
917 *Bulk2Space*

918 We further evaluated the performance of Bulk2Space without providing cell type

information. We randomly divided the human peripheral blood[46] ("GSE92495 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495]") scRNA-seq data into two parts, one was treated as the single-cell reference and the other for the construction of bulk transcriptome data via aggregating all the single-cell gene expression profiles. For the single-cell reference, we followed the scRNA-seq data analysis workflow on the Seurat website: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html to get 5 "clustering spaces" using "FindClusters" function with the parameter resolution = 0.2. For Bulk2Space, all parameters followed the default values.

Revealing spatial, molecular, and functional heterogeneity of B cells in melanoma (consecutive slices)

The scRNA-seq data ("GSE72056 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056]") by Tirosh[48] et. al. was used as the single-cell reference, and two consecutive slices of the ST data by Thrane[49] et. al. were used as spatial reference (slice 1) and synthesized bulk data (slice 2) (via aggregating all the spots gene expression profiles), respectively. For Bulk2Space, we set the parameter epoch_num = 3500, k = 10, and top_marker_num = 500, all other parameters followed the default values.

We firstly evaluated the expression correlation of marker genes for five major cell types between generated single-cell expression profiles from synthesized bulk data (slice 2) and ground truth. We applied "FindAllMarkers" function of Seurat (version 4.0.4)[60] to calculate the marker genes of each cell type with the parameter logfc.threshold = 0.5. We also evaluated the expression correlation of marker genes for five major cell types between spatial mapping result of Bulk2Space and spatial reference (slice 1).

For B cells spatial heterogeneity analysis, we determined two spatial areas (lymphoid area and tumor area) based on histological annotation firstly. Then, differentially expressed genes analysis between B cells from different areas was applied by the "FindAllMarkers" function with the parameter logfc.threshold = 0.25. For these differentially expressed genes, we further performed the pathway enrichment analysis using the Metascape (https://metascape.org) to investigate the biological functions.

Linking histomorphology and transcriptomics in PDAC (disrete slices)

Bulk2Space was performed for another PDAC dataset[47]. We selected PDAC-B scRNA-seq data as the single-cell reference, and two inconsecutive slices of the ST data were used as spatial reference (PDAC-B ST1, slice 1) and synthesized bulk data (PDAC-B ST2, slice 2) (via aggregating all the spots gene expression profiles), respectively. For Bulk2Space, we set the parameter epoch_num = 3500, k = 10, and top_marker_num = 500, all other parameters followed the default values. The evaluation procedure of the marker gene expression correlation between Bulk2Space results and slices was same as it in the human melanoma dataset analysis.

Application of Bulk2Space

Bulk2Space integrates spatial gene expression and histomorphology in PDAC

The pancreatic adjacent tissues and pancreatic cancer tissues bulk data by Wu[50] et.al. were downloaded from "GSE171485 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171485]". We firstly applied the PDAC-A scRNA-seq data by Moncada[47] et. al. as the single-cell reference to deconvolute pancreatic adjacent tissues bulk data and pancreatic cancer tissues bulk data respectively. We then applied the PDAC-A ST1 by Moncada[47] et. al. as the spatial reference and performed spatial mapping step for the generated single-cell gene profiles from pancreatic cancer tissues bulk data. For Bulk2Space, we set the parameter epoch_num = 3500, k = 10, and top_marker_num = 200, all other parameters followed the default values.

Reconstruction of mouse isocortex layers at single-cell resolution using Spatial-seq data

The mouse isocortex bulk data was produced by Spatial-seq (see Methods). We applied a mouse primary visual cortex scRNA-seq data[52] and two tissue sections (anterior section 1 and posterior section 1) sequenced by 10X Visium as the single-cell reference and spatial references, respectively. For Bulk2Space, we set the parameter epoch_num = 3500 and top_marker_num = 300, all other parameters followed the default values.

980 ***Reconstruction of mouse hypothalamus structure at single-cell resolution using***
981 ***Spatial-seq data***

982 The mouse hypothalamus bulk data was produced by Spatial-seq (see Methods). We
983 applied a mouse hypothalamus scRNA-seq data ("GSE113576
984 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113576]") and MERFISH
985 data (Bregma -0.04) by Moffitt[43] et.al. as the single-cell reference and spatial reference,
986 respectively. For Bulk2Space, we set the parameter epoch_num = 3500 and
987 top_marker_num = 500, all other parameters followed the default values. For those
988 ambiguous cells in MERFISH data, we re-annotated their cell type based on the whole
989 gene expression profiles predicted by Bulk2Space.

990 ***Uncovering spatial gene expression dynamics in inflammation-induced prostate***
991 ***cancer***

992 The human prostate cancer bulk data was downloaded from TCGA
993 (https://portal.gdc.cancer.gov). We selected the human prostate cancer scRNA-seq
994 data by Wu[54] et.al. as the single-cell reference, and three ST data from different tissue
995 section location as spatial references (p1.1 Normal glands, p2.3 Inflammation, p3.3
996 Cancer Gs 3+4). For Bulk2Space, we set the parameter epoch_num = 3500, k = 10, and
997 top_marker_num = 500, all other parameters followed the default values. We then
998 compared the distribution of various cell types in these three different tissue regions.

999 Spatial-seq protocol

1000 *Multiplexed spatial barcoding.* A virtual 48×48×48 three-dimensional well array
1001 (barcoding array) was constructed in X, Y, and Z directions, with 48×48 wells in each
1002 layer. 384-well plates (24×16 wells) were selected as the basic units (2×3 plates) of
1003 each layer. Each well in the barcoding array can be labeled as barcode
1004 $(\{X_i, Y_j, Z_k\}, i, j, k \in N, [1, 48])$. Magnetic beads coated with carboxyl group (Cat. #
1005 40200, purchased from BEAVER Biomedical Engineering Co., Ltd.) were distributed into
1006 each well for barcode synthesis. There were three rounds of barcode extension
1007 reactions. In the first round, 5' amino-modified barcoded oligonucleotides (barcode
1008 X_i) were conjugated to the beads. In the second round of extension reaction, barcoded
1009 oligonucleotides (barcode Y_j) which had a sequence of bases at the 5' end
1010 complementary to the 3' end of barcode X_i were linked to the end of barcode X_i by

1011 PCR. Similarly, in the third round of reaction, barcode sequences containing a unique
1012 molecular identifier (UMI) and a polyT tail (barcode Z_k) were extended to barcode Y_j
1013 by PCR. All oligonucleotides were purchased from Sangon Biotech (Shanghai) Co., Ltd.
1014 This method only needs 144 (48×3) different barcodes to theoretically generate
1015 magnetic beads with 48^3 distinct kinds of magnetic beads with known sequences.
1016 More importantly, the entire synthesis process was serialized and multiplexed, which
1017 can be accomplished with the help of arrayed pipettes or liquid workstations, rather
1018 than synthesizing the magnetic beads well by well. The spatially barcoded beads can
1019 be used in many areas, such as multiplexed labeling of tissue samples or even single
1020 cells.

1021 *Collection of mouse brain regions.* Three wild-type adult C57BL/6J mice (SPF, male, 20-
1022 25 g) aged 8-10 weeks were used for Spatial-seq experiments. The environmental
1023 conditions in the mouse facility were: 12 h light and 12 h dark cycle (light on from 8:00
1024 a.m. to 8:00 p.m.), light intensity range of 15-20 lux, temperature range of 22–26 °C,
1025 humidity range of 40–70%, and free access to food and water. The use and care of the
1026 mice were in accordance with the guidelines of the Animal Advisory Committee of
1027 Zhejiang University and the US National Institutes of Health Guidelines for the Care
1028 and Use of Laboratory Animals. All procedures were approved by the Animal Advisory
1029 Committee of Zhejiang University. The mouse brain was sliced into 14-µm sections
1030 from the coronal and sagittal directions. Each tissue slice was registered to a reference
1031 brain template provided by the Allen Brain Atlas (https://portal.brain-map.org/). After
1032 spatial registration, the anatomical regions of the mouse brain were delineated and
1033 annotated. The outlines, dissection sequences, and collectors of the brain regions
1034 were specified from the annotated data and then imported to an LCM instrument
1035 (Laser Microdissection System equipped with a DM68 microscope, a Leica LMD6 laser
1036 cutter, and a single-cell capture collector LMT350, Leica Microsystems, Germany).LCM
1037 is a microscope-guided powerful cutting system incorporating UV light for contact- and
1038 contamination-free isolation of areas of interest from tissue sections[15]. Brain regions
1039 were dissected from the tissue by LCM according to the imported files and fell into the
1040 collector loaded with barcoded beads in advance. Spatial-seq has the potential to
1041 achieve spatially resolved whole transcriptome sequencing of a large number of single
1042 cells.

1043 *Multiplex RNA-seq.* The tissue was lysed in the barcoded well, allowing mRNA to be
1044 captured by polyT tail on the surface of the magnetic beads. The lysis buffer was

prepared with Tris-HCl (120 µL, pH 7.5, Cat. # T1140, purchased from Beijing Solarbio Science & Technology Co., Ltd.), LiCl (80 µL, Cat. # AM9480, purchased from Invitrogen), 10% SDS solution (120 µL, RNase-free, Cat. # AM9823, purchased from Invitrogen), EDTA (16 µL, Cat. # ST066, Shanghai beyotime Biological Co.,Ltd.), 0.5 M DTT solution (16 µL, DNase, RNase & Protease free, Cat. # ST041, Shanghai beyotime Biological Co.,Ltd.), and water (852 µL, nuclease-free, Cat. # AM9930, Ambion). Each well was loaded with lysis buffer. The tissue was lysed on ice for 12 minutes. The beads were then washed and transferred for reverse transcription (RT), The captured mRNA was reverse transcribed using PrimeScript II Reverse Transcriptase (Cat. # 2690A, Takara).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 was utilized for quality control of the RNA-seq data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). 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.

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

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

$$\mathrm{RPM} = \frac{ExonMappedReads*10^6}{TotalMappedReads} \quad (11)$$

$$\mathrm{TPM} = \frac{N_i/L_i*10^6}{sum(N_1/L_1+N_2/L_2+\cdots+N_n/L_n)} \quad (12)$$

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.

Ethical statement. All experiments were approved by and conducted in accordance with the ethical guidelines of the Zhejiang University Animal Experimentation Committee (Protocol number, 14875).

Data Availability

The original data used in this paper can be accessed through the following links: (1) single-cell RNA-seq data of the human blood: GEO accession: "GSE92495 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495]"[46]; (2) single-cell RNA-seq data of the human brain: GEO accession: "GSE103723 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103723]"[61]; (3) single-cell RNA-seq data of the human kidney: GEO accession: "GSE121862 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121862]"[62]; (4) single-cell RNA-seq data of the human liver: GEO accession: "GSE124395 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395]"[63]; (5) single-cell RNA-seq data of the human lung: GEO accession: "GSE130148 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130148]"[64]; (6) single-cell RNA-seq data of the mouse brain: GEO accession: "GSE60361 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361]"[65]; (7) single-cell RNA-seq data of the mouse kidney: GEO accession: "GSE119531 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119531]"[66]; (8) single-cell RNA-seq data of the mouse lung: GEO accession: "GSE127465 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127465]"[67]; (9) single-cell RNA-seq data of the mouse pancreas: GEO accession: "GSE84133 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133]"[68]; (10) single-cell RNA-seq data of the mouse testis: GEO accession: "GSE112393 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112393]"[69]; (11) single-cell RNA-seq data of the human pancreas with 975 cells: GEO accession: "GSE81076 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81076]"[70]; (12) single-cell RNA-seq data of the human pancreas with 2133 cells: GEO accession: "GSE85241 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85241]"[71]; (13) single-cell RNA-seq data of the human pancreas with 597 cells: GEO accession: "GSE86469 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86469]"[72]; (14) four sets of single-cell RNA-seq data of the human pancreas with 1635, 1562, 3330, 1230 cells,

respectively: GEO accession: "GSE84133 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133]"[68]; (15) single-cell RNA-seq data of the human pancreas with 2288 cells[73] (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5061/); (16) single-cell RNA-seq data of the mouse hypothalamus using 10X Genomics: GEO accession: "GSE113576 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113576]"[43]; (17) single-cell RNA-seq data of the mouse hypothalamus using Drop-seq: "GSE87544 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87544]"[44]; (18) three sets of single-cell RNA-seq data of the mice liver: GEO accession: "GSE119340 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119340]"[42]; (19) single-cell RNA-seq data of the mouse hippocampus region[45] (https://www.dropbox.com/s/cs6pii5my4p3ke3/mouse_hippocampus_reference.rds?dl=0); (20) two sets of single-cell RNA-seq data of the human pancreatic ductal adenocarcinoma (PDAC): GEO accession: "GSE111672 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672]"[47]; (21) single-cell RNA-seq data of the human melanoma: GEO accession: "GSE72056 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056]"; (22) single-cell RNA-seq data of the mouse cortex region[52] (https://www.dropbox.com/s/dl/cuowvm4vrf65pvq/allen_cortex.rds); (23) single-cell RNA-seq data of the human prostate cancer[54] (https://singlecell.broadinstitute.org/single_cell/study/SCP1415); (24) MERFISH data of the mouse hypothalamic preoptic region at bregma 0.26[43] (https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248); (25) slide-seq v2 data of the mouse hippocampus region[12] (https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2); (26) three sets of spatially resolved transcriptomics data of the human PDAC using “Spatial Transcriptomics”: GEO accession: "GSE111672 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672]"[47]; (27) two sets of spatially resolved transcriptomics data of the human melanoma using “Spatial Transcriptomics”[49] (https://www.spatialresearch.org/resources-published-datasets/doi-10-1158-0008-5472-can-18-0747/); (28) two sets of 10X Visium data of the mouse anterior cortex and posterior cortex regions (https://www.10xgenomics.com/cn/resources/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-1-0); (29) two sets of spatially resolved

transcriptomics data of the human prostate cancer using “Spatial Transcriptomics”[55] (https://www.spatialresearch.org/resources-published-datasets/10-1038-s41467-018-04724-5/); (30) three sets of bulk RNA-seq data of the mice liver: GEO accession: "GSE119340 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119340]"[42]; (31) two sets of bulk RNA-seq data of the human PDAC: GEO accession: "GSE171485 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171485]"[50]; (32) bulk RNA-seq data of the human prostate cancer (https://portal.gdc.cancer.gov).

The bulk RNA-seq data of the mouse cortex and mouse hypothalamus regions reported in this manuscript using our in-house developed Spatial-seq have been deposited to the Gene Expression Omnibus under accession number "GSE192999 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE192999]".All other relevant data supporting the key findings of this study are available within the article and its Sup-plementary Information files or from the corresponding author upon reasonable request. All source data are provided with this paper.

Statistics & Reproducibility

In this study, 152 simulated datasets (the data points of paired simulations for deconvolution, unpaired simulations for deconvolution, paired simulations for spatial mapping, and unpaired simulations for spatial mapping are 30, 12, 50, and 60, respectively), 13 biological datasets, and 2 experimental datasets were used to evaluate the Bulk2Space algorithm. Unannotated, ambiguous, or low quality-cells were excluded from the analysis. Pearson correlation coefficient, Spearman's rank correlation coefficient, and root mean squared error were used to compare the performance between different methods. For paired data simulations, single cells were randomly selected to synthesize the reference and the test set. For unpaired data simulations, single-cell profiles were randomly selected to synthesize the reference and the test set. The Investigators were not blinded to allocation during experiments and outcome assessment.

Code Availability

Bulk2Space is available as a python package and the source code is deposited at the GitHub (https://github.com/ZJUFanLab/bulk2space)[75].

1170

1171 # References

1172 1. Liao, J., Lu, X., Shao, X., Zhu, L. & Fan, X. Uncovering an Organ's Molecular Architecture at Single-Cell
1173 Resolution by Spatially Resolved Transcriptomics. *Trends Biotechnol* **39**, 43-58 (2021).

1174 2. Wang, Y.S. & Guo, J. Multiplexed Single-Cell in situ RNA Profiling. *Front Mol Biosci* **8**, 775410 (2021).

1175 3. Srivatsan, S.R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111-117 (2021).

1176 4. Waylen, L.N., Nim, H.T., Martelotto, L.G. & Ramialison, M. From whole-mount to single-cell spatial
1177 assessment of gene expression in 3D. *Commun Biol* **3**, 602 (2020).

1178 5. Qian, X. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat*
1179 *Methods* **17**, 101-106 (2020).

1180 6. Larsson, L., Frisen, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new dimension to
1181 genomics. *Nat Methods* **18**, 15-18 (2021).

1182 7. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved,
1183 highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

1184 8. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by
1185 sequential hybridization. *Nat Methods* **11**, 360-361 (2014).

1186 9. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states.
1187 *Science* **361** (2018).

1188 10. Eng, C.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**,
1189 235-239 (2019).

1190 11. Vickovic, S. et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**,
1191 987-990 (2019).

1192 12. Stickels, R.R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-
1193 seqV2. *Nat Biotechnol* **39**, 313-319 (2021).

1194 13. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high
1195 spatial resolution. *Science* **363**, 1463-1467 (2019).

1196 14. Stahl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial
1197 transcriptomics. *Science* **353**, 78-82 (2016).

1198 15. Nichterwitz, S. et al. Laser capture microscopy coupled with Smart-seq2 for precise spatial
1199 transcriptomic profiling. *Nat Commun* **7**, 12139 (2016).

1200 16. Chen, J. et al. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat*
1201 *Protoc* **12**, 566-580 (2017).

1202 17. Cheng, J., Liao, J., Shao, X., Lu, X. & Fan, X. Multiplexing Methods for Simultaneous Large-Scale
1203 Transcriptomic Profiling of Samples at Single-Cell Resolution. *Adv Sci (Weinh)* **8**, e2101229 (2021).

1204 18. Liu, Y., Ai, N., Liao, J. & Fan, X. Transcriptomics: a sword to cut the Gordian knot of traditional Chinese
1205 medicine. *Biomark Med* **9**, 1201-1213 (2015).

1206 19. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

1207 20. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human
1208 glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

1209 21. International Cancer Genome, C. et al. International network of cancer genome projects. *Nature* **464**,
1210 993-998 (2010).

1211 22. Consortium, S.M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and
1212 information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903-914
1213 (2014).

1214 23. Frishberg, A. et al. Cell composition analysis of bulk genomics using single-cell data. *Nat Methods*
1215 **16**, 327-332 (2019).

1216 24. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat*
1217 *Methods* **12**, 453-457 (2015).

1218 25. Wang, X., Park, J., Susztak, K., Zhang, N.R. & Li, M. Bulk tissue cell type deconvolution with multi-
1219 subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).

1220 26. Chen, Z. et al. Inference of immune cell composition on the expression profiles of mouse tissue. *Sci*
1221 *Rep* **7**, 40508 (2017).

1222 27. Gangalum, R.K. et al. Spatial Analysis of Single Fiber Cells of the Developing Ocular Lens Reveals
1223 Regulated Heterogeneity of Gene Expression. *iScience* **10**, 66-79 (2018).

1224 28. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for
1225 Single-Cell Multi-omics. *Trends Biotechnol* **38**, 1007-1022 (2020).

1226 29. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data
1227 across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

1228 30. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

1229 31. Shao, X. et al. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA
1230 Sequencing Data. *iScience* **23**, 100882 (2020).

1231 32. Shao, X. et al. scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics
1232 using deep learning with a weighted graph neural network. *Nucleic Acids Res* **49**, e122 (2021).

1233 33. Higgins, I. et al. beta-vae: Learning basic visual concepts with a constrained variational framework.
1234 *International Conference on Learning Representations*, 1-22 (2017).

1235 34. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat*
1236 *Biotechnol* **40**, 517-526 (2022).

1237 35. Dong, R. & Yuan, G.C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome*
1238 *Biol* **22**, 145 (2021).

1239 36. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type
1240 topography. *Commun Biol* **3**, 565 (2020).

1241 37. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to
1242 deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* **49**, e50
1243 (2021).

1244 38. Li, H., Wei, W. & Xu, H. Drug discovery is an eternal challenge for the biomedical sciences. *Acta*
1245 *Materia Medica* **1**, 1-3 (2022).

1246 39. Goodfellow, I.J. et al. Generative Adversarial Networks. *arXiv*, 1-9. (2014).
1247 https://doi.org/10.48550/arXiv.1406.2661.

1248 40. Mirza, M. & Osindero, S.J.A. Conditional Generative Adversarial Nets. arXiv, 1-7. (2014).
1249 https://doi.org/10.48550/arXiv.1411.1784.

1250 41. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior
1251 derived from single-cell data. *Genome Res* **31**, 1807-1818 (2021).

1252 42. Xiong, X. et al. Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-
1253 Cell Secretome Gene Analysis. *Mol Cell* **75**, 644-660 e645 (2019).

1254 43. Moffitt, J.R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic
1255 region. *Science* **362** (2018).

1256 44. Chen, R., Wu, X., Jiang, L. & Zhang, Y. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell*
1257 *Rep* **18**, 3227-3241 (2017).

1258 45. Saunders, A. et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain.
1259 *Cell* **174**, 1015-1030 e1016 (2018).

1260 46. Gierahn, T.M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput.
1261 *Nat Methods* **14**, 395-398 (2017).

1262 47. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq
1263 reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38**, 333-342 (2020).

1264 48. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-
1265 seq. *Science* **352**, 189-196 (2016).

1266 49. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially Resolved Transcriptomics
1267 Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. *Cancer*
1268 *Res* **78**, 5970-5979 (2018).

1269 50. Wu, H. et al. Identification and functional analysis of novel oncogene DDX60L in pancreatic ductal
1270 adenocarcinoma. *BMC Genomics* **22**, 833 (2021).

1271 51. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using
1272 Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

1273 52. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat*
1274 *Neurosci* **19**, 335-346 (2016).

1275 53. Elinav, E. et al. Inflammation-induced cancer: crosstalk between tumours, immune cells and
1276 microorganisms. *Nat Rev Cancer* **13**, 759-771 (2013).

1277 54. Wu, S.Z. et al. Cryopreservation of human cancers conserves tumour heterogeneity for single-cell
1278 multi-omics analysis. *Genome Med* **13**, 81 (2021).

1279 55. Berglund, E. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape
1280 of heterogeneity. *Nat Commun* **9**, 2419 (2018).

1281 56. Shao, X., Lu, X., Liao, J., Chen, H. & Fan, X. New avenues for systematically inferring cell-cell
1282 communication: through single-cell transcriptomics data. *Protein Cell* **11**, 866-880 (2020).

1283 57. Shao, X. et al. CellTalkDB: a manually curated database of ligand-receptor interactions in humans
1284 and mice. *Brief Bioinform* **22** (2021).

1285 58. Shao, X. et al. Knowledge-graph-based cell-cell communication inference for spatially resolved
1286 transcriptomic data with SpaTalk. *Nat Commun* **13**, 4429 (2022).

1287 59. Shao, L. et al. Identify differential genes and cell subclusters from time-series scRNA-seq data using
1288 scTITANS. *Comput Struct Biotechnol J* **19**, 4132-4141 (2021).

1289 60. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).

1290 61. Fan, X. et al. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-
1291 seq analysis. *Cell Res* **28**, 730-745 (2018).

1292 62. Lake, B.B. et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and
1293 pathophysiology of human kidneys. *Nat Commun* **10**, 2832 (2019).

1294 63. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*
1295 **572**, 199-204 (2019).

1296 64. Vieira Braga, F.A. et al. A cellular census of human lungs identifies novel cell states in health and in
1297 asthma. *Nat Med* **25**, 1153-1163 (2019).

1298 65. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-
1299 cell RNA-seq. *Science* **347**, 1138-1142 (2015).

1300 66. Wu, H., Kirita, Y., Donnelly, E.L. & Humphreys, B.D. Advantages of Single-Nucleus over Single-Cell
1301 RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J Am Soc*
1302 *Nephrol* **30**, 23-32 (2019).

1303 67. Zilionis, R. et al. Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved
1304 Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-1334 e1310 (2019).

1305 68. Baron, M. et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter-
1306 and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344 (2016).

1307 69. Green, C.D. et al. A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell
1308 RNA-Seq. *Dev Cell* **46**, 651-667 e610 (2018).

1309 70. Grun, D. et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell*
1310 *Stem Cell* **19**, 266-277 (2016).

1311 71. Muraro, M.J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-394
1312 e383 (2016).

1313 72. Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-
1314 specific expression changes in type 2 diabetes. *Genome Res* **27**, 208-222 (2017).

1315 73. Segerstolpe, A. et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and
1316 Type 2 Diabetes. *Cell Metab* **24**, 593-607 (2016).

1317 74. Xiong, X. et al. Mapping the molecular signatures of diet-induced NASH and its regulation by the
1318 hepatokine Tsukushi. *Mol Metab* **20**, 128-137 (2019).

1319 75. Liao, J. et al. De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution.
1320 ZJUFanLab/bulk2space: Bulk2Space v1.0.0 (v1.0.0), doi: 10.5281/zenodo.7134575 (2022).

1321

Acknowledgments

This work is supported by the National Natural Science Foundation of China (81973701, X.F.; 91846204, H.C.; and U19B2027, H.C.), the Natural Science Foundation of Zhejiang Province (LZ20H290002, X.F.), the Innovation Team and Talents Cultivation Program of the National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202002, X.F.; ZYYCXTD-D-202207, Y.G.), the Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine), and the Alibaba Cloud (Alibaba Cloud Computing Co. Ltd.).

Author Contributions Statement

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. (Yang Hu), Y.H.(Yining Hu), L.Y., and X.L. conducted the multiplexed RNA-seq experiment using Spatial-seq. D.W. provided important advice on brain tissue isolation, spatial registration, and annotation. All authors wrote the manuscript, and read and approved the final manuscript.

Competing Interests Statement

The authors declare no competing interests.

Figure Legends

Fig. 1 Workflow of Bulk2Space. a, Overview of the design concept of Bulk2Space. Bulk transcriptome data is taken as the 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 deconvolved into single-cell transcriptomics data. Then, either of the two spatially resolved transcriptomics, spatial barcoding-based 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 generated single-cell profiles with specified x and y spatial coordinates. **b,** Detailed deconvolution procedure. The input vector of bulk tissue is equal to the 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-cell 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-based 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.

Fig. 2. Benchmark test for Bulk2Space. a, Gene expression correlation (Pearson correlations, left, and Spearman correlations, middle) and gene expression variation (RMSE, right) of three methods, namely, Bulk2Space (β-VAE), GAN, and CGAN, using paired simulation data (n=30). Each point represented a simulated dataset. Data are presented as boxplots (minima, 25th percentile, median, 75th percentile, and maxima). **b**, Gene expression correlation and gene expression variation of three methods with unpaired simulation data (n=12). Data are presented as boxplots (minima, 25th percentile, median, 75th percentile, and maxima). **c**, Benchmark test for 10 spatial mapping methods (from left to right, Bulk2Space (deep forest), gradient boosting decision tree, decision tree, multi-layer perceptron, DeepGBM, linear regression, RCTD, spatialDWLS, stereoscope, and SPOTlight) using paired simulation data (n=50). Data are presented as boxplots (minima, 25th percentile, median, 75th percentile, and maxima). Left, Pearson correlation of gene expression, Middle, Spearman correlation of gene expression, Right, gene expression variation (RMSE). **d**, Benchmark test for 10 spatial mapping methods using paired simulation data (n=60). Data are presented as

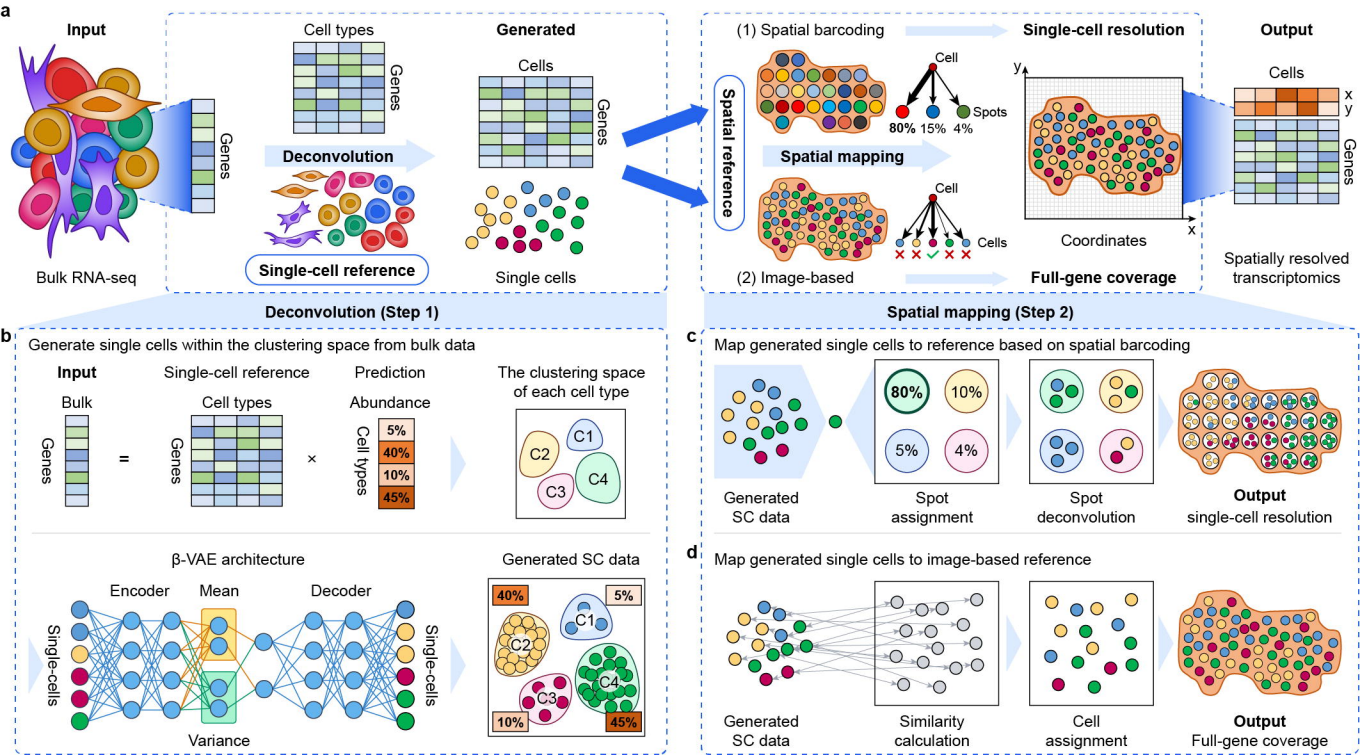
boxplots (minima, 25th percentile, median, 75th percentile, and maxima). **e**, The UMAP layout of the MERFISH data (image-based spatially resolved transcriptomics) and the single-cell data. Share cell types were colored with corresponding colors. **f**, Expression correlation heatmap of the marker genes of all shared cell types in MERFISH data and scRNA-seq data. **g**, Spatial expression of genes in the ground truth and predicted by Bulk2Space using paired and unpaired single-cell RNA-seq data. PCC for each gene was shown. Source data are provided as a Source Data file.

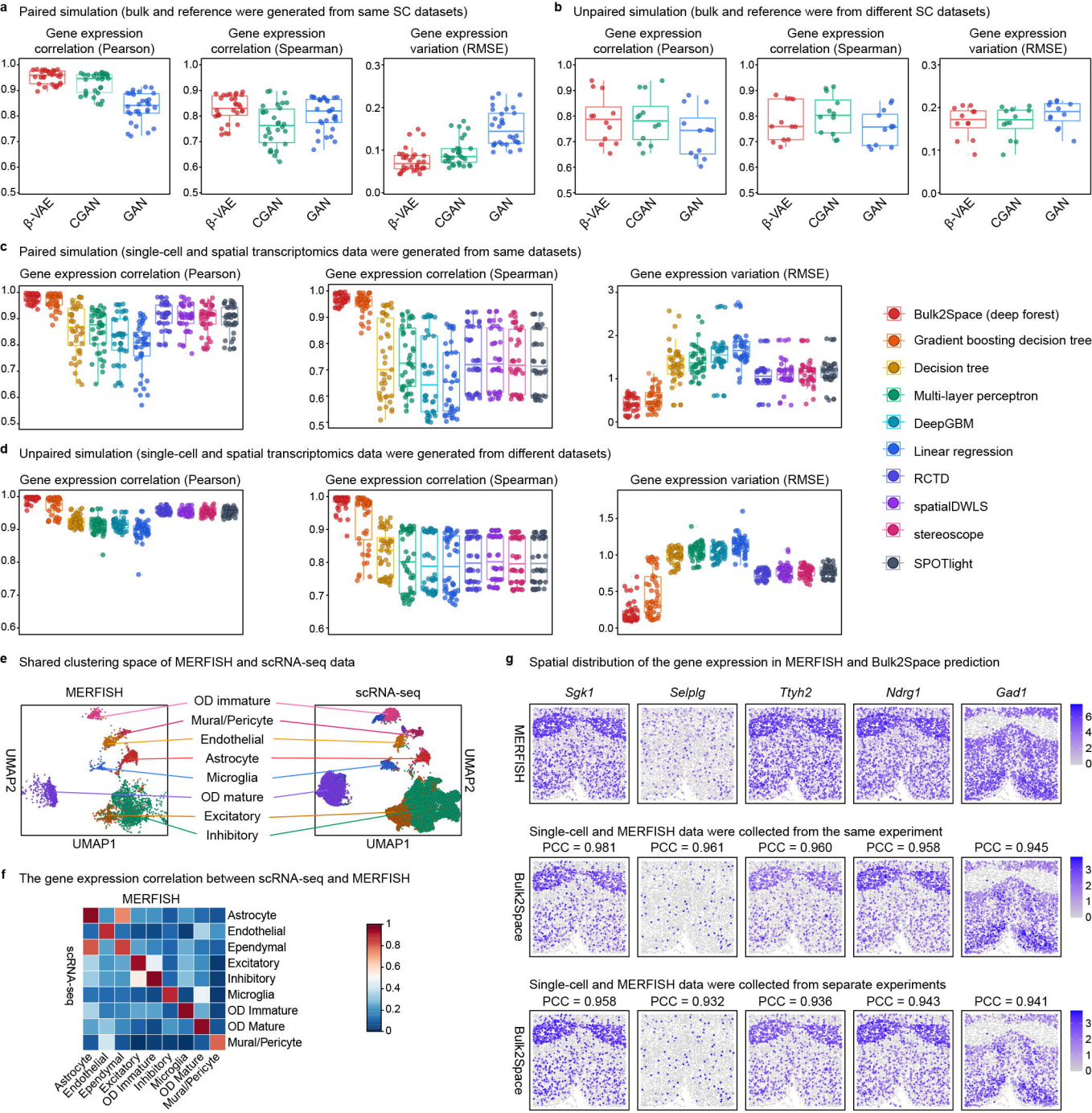
Fig. 3. Spatially resolved analysis of PDAC by Bulk2Space. a, Collection of pancreatic adjacent (PA) and PDAC bulk transcriptome data, PDAC scRNA-seq data, and PDAC spatial transcriptomics data. Two bulk data were sequenced by bulk RNA-seq. The scRNA-seq data sequenced by inDrop-seq[51] is used as the single-cell reference. One slice of the sectioned PDAC tissue sequenced by ST[14] was employed as the spatial reference. **b**, The cell-type proportions of single cells generated from PDAC and PA tissues by Bulk2Space. Purple box, the proportion of the cell type was higher in PA tissue than that in PDAC tissue. Black box, the proportion of the cell type was higher in PDAC tissue than that in PA tissue. Teal, the PDAC tissue. Red, PA tissue. **c**, The clustering space of single cells generated from the PDAC (Top) and PA (Bottom) bulk data by Bulk2Space using UMAP layout. Different colors represented distinct cell types. **d**, Pairwise expression correlation of cell-type-specific marker genes between single cells generated by Bulk2Space and the single-cell reference for PDAC. Marker genes were found by 'FindAllMarkers' function in Seurat. *P* value was calculated with the Wilcoxon Rank Sum test. **e**, Spatial mapping of single cells generated from PDAC bulk data. Scale bar, 1 mm. **f**, Top, Histological annotation for cancer region (red), duct epithelium (yellow), and normal pancreatic tissue (light blue). Middle, clusters of cancer region (yellow), pancreatic tissue (blue), duct epithelium (green), and stroma (dark grey) from the spatial transcriptomics data. Bottom, spatial deconvolution result of the bulk PDAC data by Bulk2Space with each spot displaying the composition of cell types in a pie chart. Scale bar, 1 mm. **g**, The spatial abundance of acinar cells, cancer clone A cells, cancer clone B cells, and ductal cells in each spot on the tissue section predicted by Bulk2Space. Scale bar, 1 mm. **h**, The spatial expression of the marker genes in acinar cells, cancer clone A cells, cancer clone B cells, and ductal cells predicted by Bulk2Space at spot (left) and cellular (right) resolution. Source data are provided as a Source Data file.

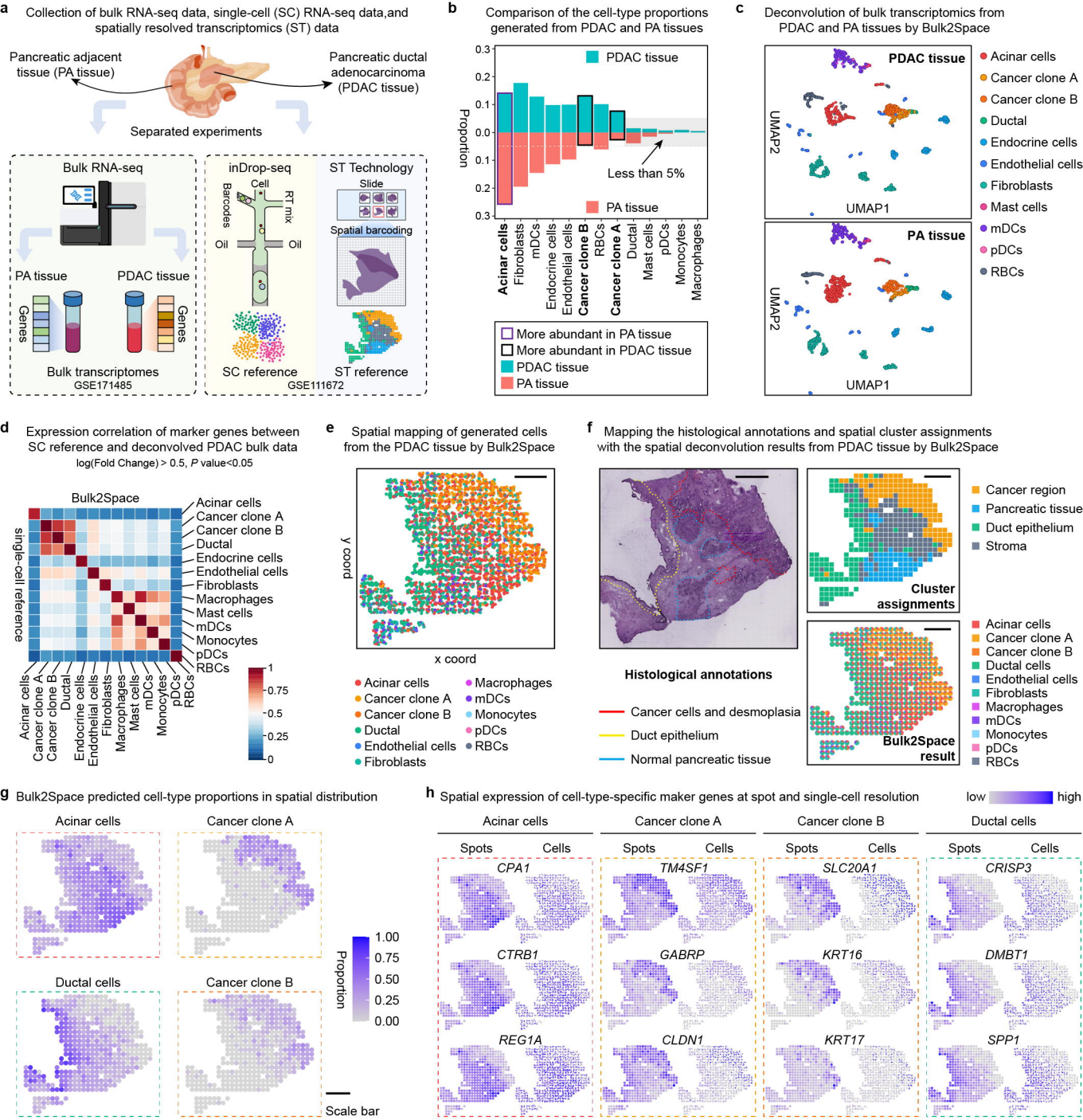
Fig. 4. Spatially resolved single-cell analysis of the mouse isocortex bulk data by Bulk2Space. a, The resources of the bulk transcriptome data, single-cell reference, and spatial reference. The bulk RNA-seq data of the mouse isocortex was sequenced using our in-house developed Spatial-seq which combines LCM and spatial barcoding strategies. The single-cell reference data is derived from scRNA-seq of a collection of consecutive mouse primary visual cortex sections from anterior to posterior. The spatial reference data is obtained from an open-access database using the 10X Visium approach. **b**, The deconvolution results of the mouse isocortex bulk data by Bulk2Space. The UMAP layout showed the clustering space of the generated single cells. **c**, The spatial mapping of the single cells generated from the mouse isocortex bulk data by Bulk2Space. Scale bar, 1 mm. **d**, The pairwise expression correlation of cell-type-specific marker genes between single cells generated by Bulk2Space and single cells in the reference data. **e**, The spatial distribution of the cell-type proportion predicted by Bulk2Space. Scale bar, 1 mm. **f**, The spatial cell-type abundance of seven cell types with layered structure predicted by Bulk2Space. Scale bar, 1 mm. **g**, The spatial expression distribution of cell-type-specific marker genes predicted by Bulk2Space at single-cell resolution. Scale bar, 1 mm. Source data are provided as a Source Data file.

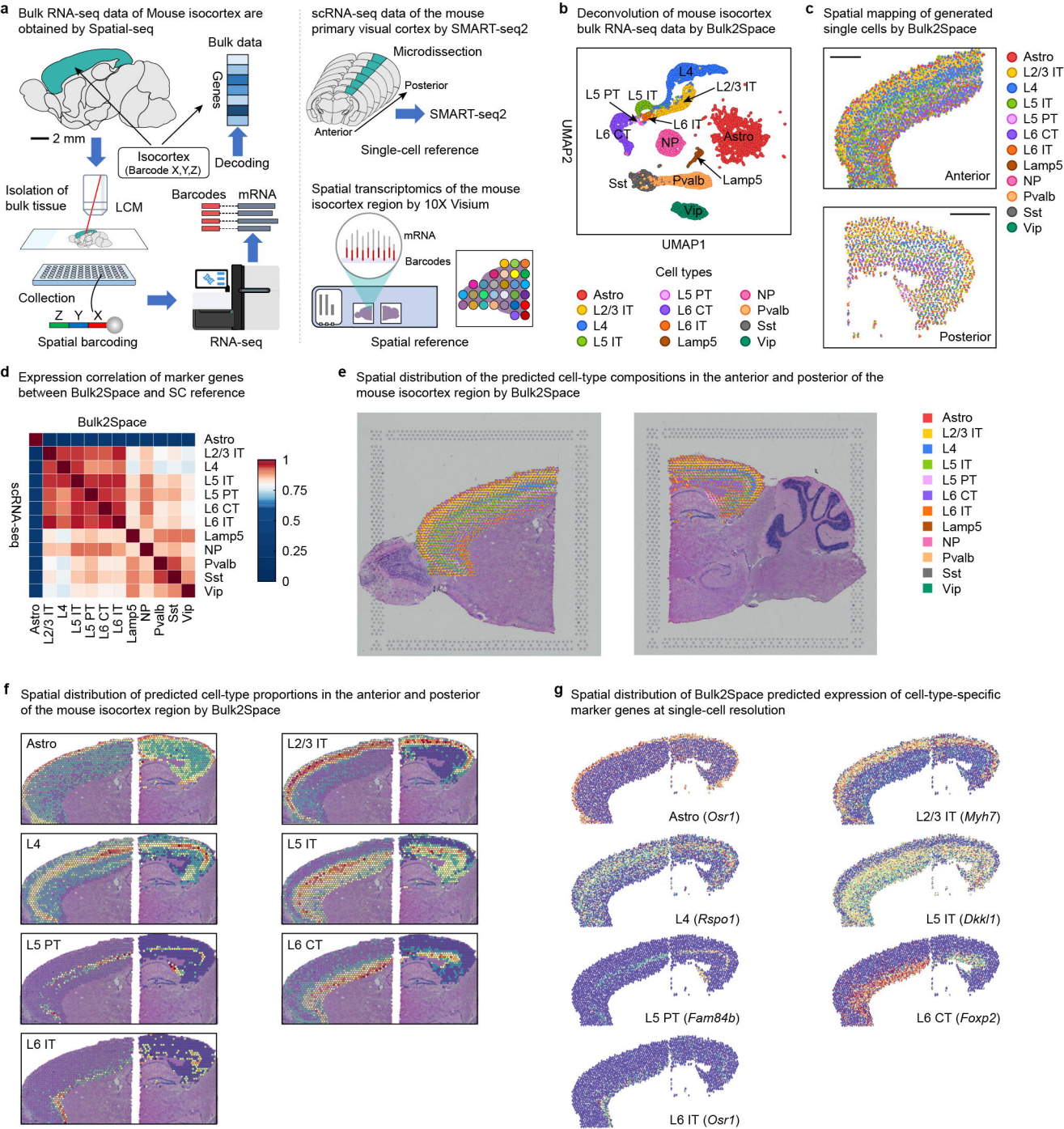
Fig. 5. Spatially resolved analysis of the mouse hypothalamus bulk data by Bulk2Space. a, The resources of the bulk transcriptome data, single-cell reference, and spatial reference. The bulk RNA-seq data of the mouse hypothalamus was profiled using our in-house developed Spatial-seq. The single-cell reference data was derived from Drop-seq. The spatial reference data is obtained from MERFISH. **b**, The UMAP layout of single-cell profiles in MERFISH data and generated from bulk tissue by Bulk2Space. 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 between cell types in MERFISH data and Bulk2Space results. The correlation of shared cell types was in red border and the expression correlation between MERFISH ambiguous cells and Bulk2Space predicted cell types was in green border. **d**, The spatial expression of genes predicted by Bulk2Space and in 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 were colored according to their predicted cell types. **f**, Predicted spatial expression patterns of novel marker genes for different cell types. Source data are provided as a Source Data file.

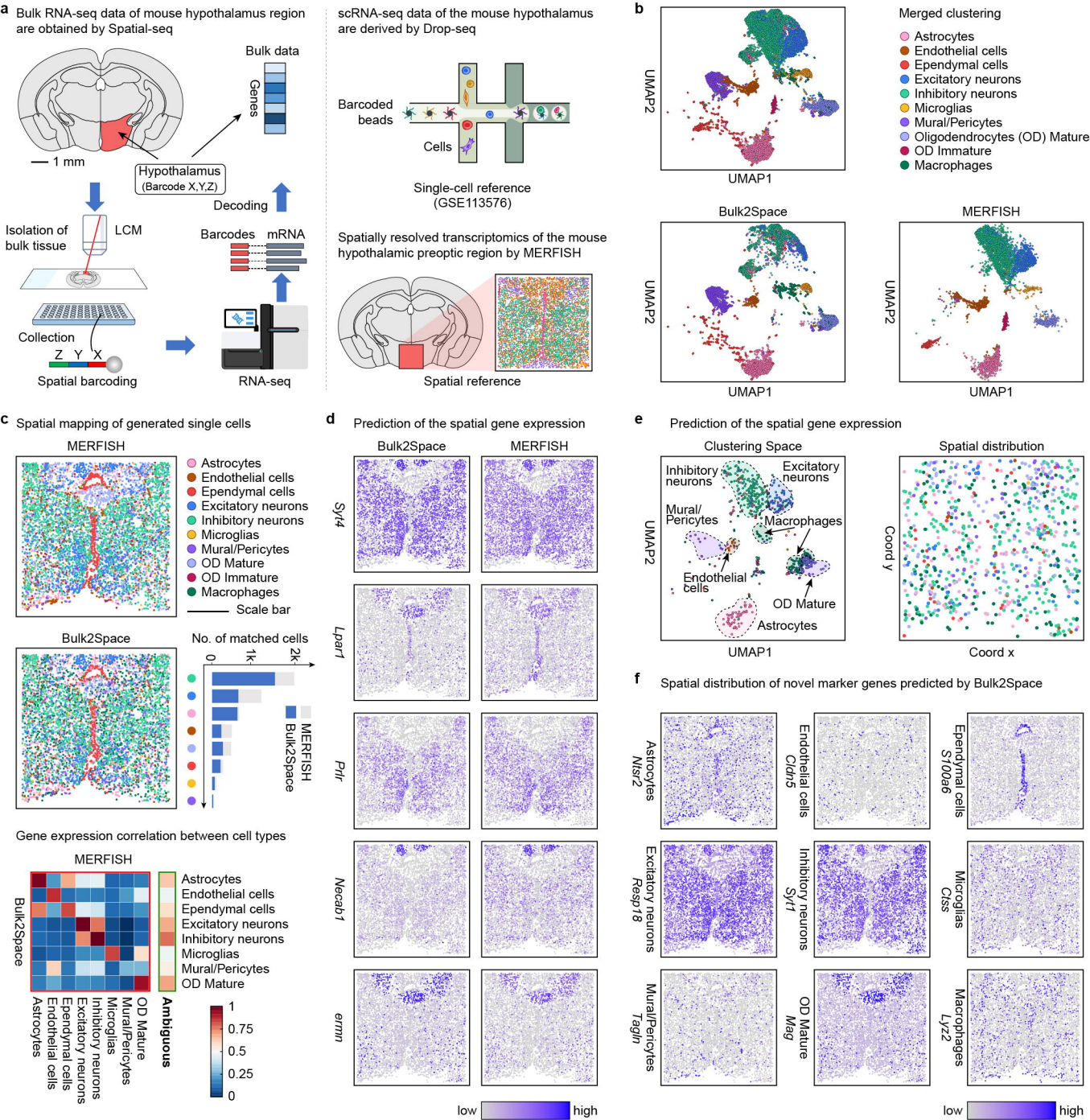
Fig. 6. Spatial gene expression dynamics analysis of inflammation-induced prostate cancer by Bulk2Space. a, Experimental design of the bulk transcriptome data. The tissue was segmented into distinct regions, namely, normal glands (pink), inflammation (blue), and several cancer regions (red, green, purple, 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 generated 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 cancer-associated fibroblasts (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. Source data are provided as a Source Data file.

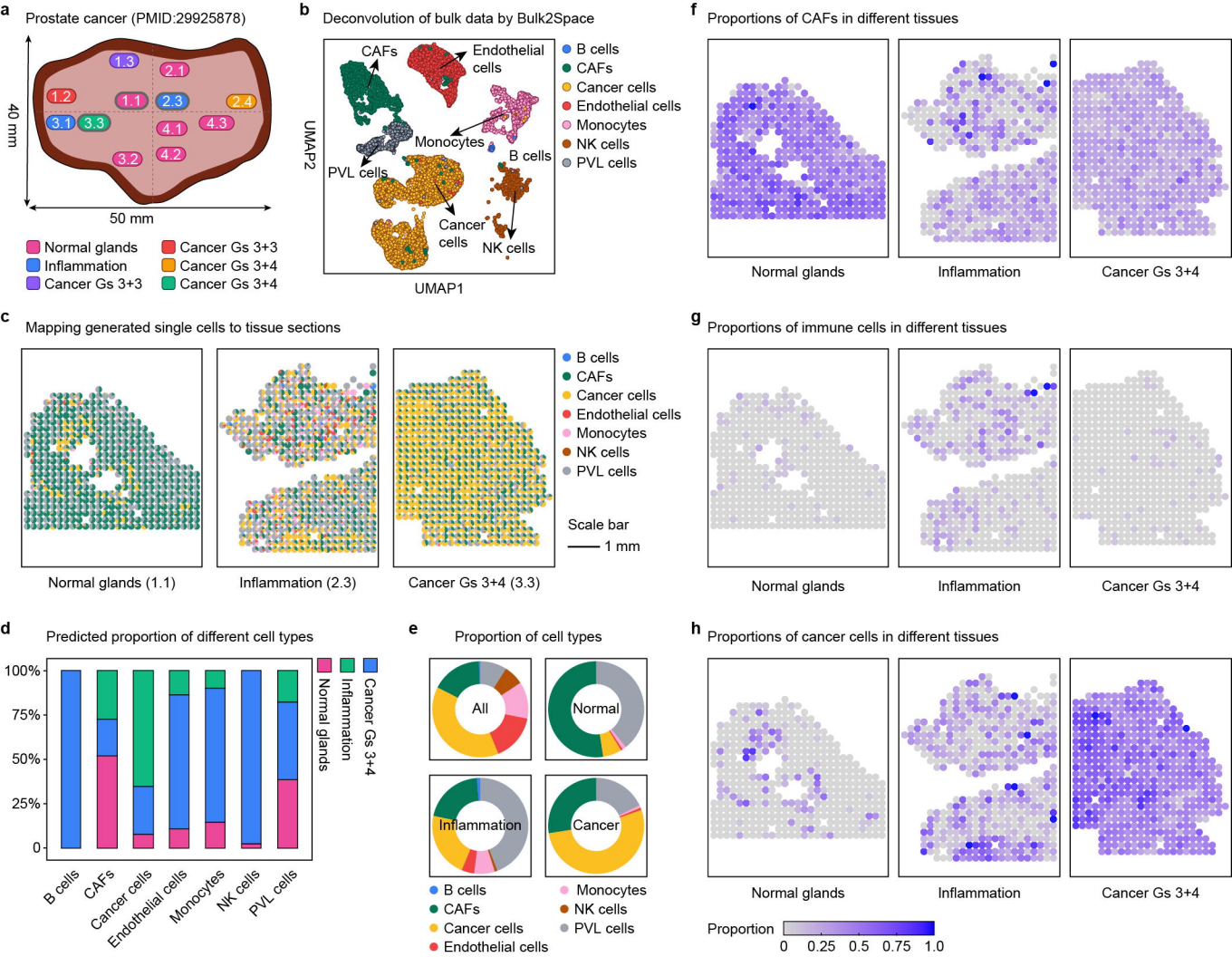

a
Input
Bulk RNA-seq
Genes
Cell types
Genes
Deconvolution
Single-cell reference
Generated
Cells
Genes
Single cells
(1) Spatial barcoding
Single-cell resolution
Output
Spatial reference
Cell
Spots
80% 15% 4%
Spatial mapping
Cell
Cells
(2) Image-based
Full-gene coverage
y
x
Coordinates
Cells
x
y
Genes
Spatially resolved transcriptomics
Deconvolution (Step 1)
Spatial mapping (Step 2)
b
Generate single cells within the clustering space from bulk data
Input
Bulk
Genes
=
Single-cell reference
Cell types
Genes
×
Prediction
Abundance
Cell types
5%
40%
10%
45%
The clustering space of each cell type
C1
C2
C3
C4
β-VAE architecture
Encoder
Mean
Decoder
Single-cells
Variance
Single-cells
Generated SC data
40%
5%
C1
C2
C3
C4
10%
45%
c
Map generated single cells to reference based on spatial barcoding
Generated SC data
80%
10%
5%
4%
Spot assignment
Spot deconvolution
Output
single-cell resolution
d
Map generated single cells to image-based reference
Generated SC data
Similarity calculation
Cell assignment
Output
Full-gene coverage


a
Paired simulation (bulk and reference were generated from same SC datasets)
Gene expression correlation (Pearson)
Gene expression correlation (Spearman)
Gene expression variation (RMSE)
β-VAE
CGAN
GAN
b
Unpaired simulation (bulk and reference were from different SC datasets)
c
Paired simulation (single-cell and spatial transcriptomics data were generated from same datasets)
d
Unpaired simulation (single-cell and spatial transcriptomics data were generated from different datasets)
Bulk2Space (deep forest)
Gradient boosting decision tree
Decision tree
Multi-layer perceptron
DeepGBM
Linear regression
RCTD
spatialDWLS
stereoscope
SPOTlight
e
Shared clustering space of MERFISH and scRNA-seq data
MERFISH
scRNA-seq
UMAP1
UMAP2
OD immature
Mural/Pericyte
Endothelial
Astrocyte
Microglia
OD mature
Excitatory
Inhibitory
f
The gene expression correlation between scRNA-seq and MERFISH
Astrocyte
Endothelial
Ependymal
Excitatory
Inhibitory
Microglia
OD Immature
OD Mature
Mural/Pericyte
g
Spatial distribution of the gene expression in MERFISH and Bulk2Space prediction
Sgk1
Selplg
Ttyh2
Ndrg1
Gad1
MERFISH
Single-cell and MERFISH data were collected from the same experiment
PCC = 0.981
PCC = 0.961
PCC = 0.960
PCC = 0.958
PCC = 0.945
Bulk2Space
Single-cell and MERFISH data were collected from separate experiments
PCC = 0.958
PCC = 0.932
PCC = 0.936
PCC = 0.943
PCC = 0.941


a
Collection of bulk RNA-seq data, single-cell (SC) RNA-seq data,and spatially resolved transcriptomics (ST) data
Pancreatic adjacent tissue (PA tissue)
Pancreatic ductal adenocarcinoma (PDAC tissue)
Separated experiments
Bulk RNA-seq
PA tissue
PDAC tissue
Genes
Genes
Bulk transcriptomes
GSE171485
inDrop-seq
Cell
Barcodes
RT mix
Oil
Oil
SC reference
ST Technology
Slide
Spatial barcoding
ST reference
GSE111672
b
Comparison of the cell-type proportions generated from PDAC and PA tissues
PDAC tissue
PA tissue
Proportion
Less than 5%
Acinar cells
Fibroblasts
mDCs
Endocrine cells
Endothelial cells
Cancer clone B
RBCs
Cancer clone A
Ductal
Mast cells
pDCs
Monocytes
Macrophages
More abundant in PA tissue
More abundant in PDAC tissue
PDAC tissue
PA tissue
c
Deconvolution of bulk transcriptomics from PDAC and PA tissues by Bulk2Space
PDAC tissue
PA tissue
UMAP2
UMAP1
Acinar cells
Cancer clone A
Cancer clone B
Ductal
Endocrine cells
Endothelial cells
Fibroblasts
Mast cells
mDCs
pDCs
RBCs
d
Expression correlation of marker genes between SC reference and deconvolved PDAC bulk data
log(Fold Change) > 0.5, P value<0.05
Bulk2Space
single-cell reference
Acinar cells
Cancer clone A
Cancer clone B
Ductal
Endocrine cells
Endothelial cells
Fibroblasts
Macrophages
Mast cells
mDCs
Monocytes
pDCs
RBCs
1
0.75
0.5
0.25
0
e
Spatial mapping of generated cells from the PDAC tissue by Bulk2Space
y coord
x coord
Acinar cells
Cancer clone A
Cancer clone B
Ductal
Endothelial cells
Fibroblasts
Macrophages
mDCs
Monocytes
pDCs
RBCs
f
Mapping the histological annotations and spatial cluster assignments with the spatial deconvolution results from PDAC tissue by Bulk2Space
Histological annotations
Cancer cells and desmoplasia
Duct epithelium
Normal pancreatic tissue
Cancer region
Pancreatic tissue
Duct epithelium
Stroma
Cluster assignments
Bulk2Space result
Acinar cells
Cancer clone A
Cancer clone B
Ductal cells
Endothelial cells
Fibroblasts
Macrophages
mDCs
Monocytes
pDCs
RBCs
g
Bulk2Space predicted cell-type proportions in spatial distribution
Acinar cells
Cancer clone A
Ductal cells
Cancer clone B
Proportion
1.00
0.75
0.50
0.25
0.00
Scale bar
h
Spatial expression of cell-type-specific maker genes at spot and single-cell resolution
low
high
Acinar cells
Cancer clone A
Cancer clone B
Ductal cells
Spots
Cells
CPA1
CTRB1
REG1A
TM4SF1
GABRP
CLDN1
SLC20A1
KRT16
KRT17
CRISP3
DMBT1
SPP1


a
Bulk RNA-seq data of Mouse isocortex are obtained by Spatial-seq
Bulk data
Genes
2 mm
Isocortex
(Barcode X,Y,Z)
Decoding
Isolation of bulk tissue
LCM
Barcodes
mRNA
Collection
Z Y X
Spatial barcoding
RNA-seq
scRNA-seq data of the mouse primary visual cortex by SMART-seq2
Microdissection
Posterior
Anterior
SMART-seq2
Single-cell reference
Spatial transcriptomics of the mouse isocortex region by 10X Visium
mRNA
Barcodes
Spatial reference
b
Deconvolution of mouse isocortex bulk RNA-seq data by Bulk2Space
UMAP2
UMAP1
L4
L2/3 IT
L5 PT
L5 IT
L6 IT
L6 CT
NP
Astro
Sst
Pvalb
Lamp5
Vip
Cell types
Astro
L2/3 IT
L4
L5 IT
L5 PT
L6 CT
L6 IT
Lamp5
NP
Pvalb
Sst
Vip
c
Spatial mapping of generated single cells by Bulk2Space
Anterior
Posterior
Astro
L2/3 IT
L4
L5 IT
L5 PT
L6 CT
L6 IT
Lamp5
NP
Pvalb
Sst
Vip
d
Expression correlation of marker genes between Bulk2Space and SC reference
Bulk2Space
scRNA-seq
Astro
L2/3 IT
L4
L5 IT
L5 PT
L6 CT
L6 IT
Lamp5
NP
Pvalb
Sst
Vip
1
0.75
0.5
0.25
0
e
Spatial distribution of the predicted cell-type compositions in the anterior and posterior of the mouse isocortex region by Bulk2Space
Astro
L2/3 IT
L4
L5 IT
L5 PT
L6 CT
L6 IT
Lamp5
NP
Pvalb
Sst
Vip
f
Spatial distribution of predicted cell-type proportions in the anterior and posterior of the mouse isocortex region by Bulk2Space
Astro
L2/3 IT
L4
L5 IT
L5 PT
L6 CT
L6 IT
g
Spatial distribution of Bulk2Space predicted expression of cell-type-specific marker genes at single-cell resolution
Astro (Osr1)
L2/3 IT (Myh7)
L4 (Rspo1)
L5 IT (Dkkl1)
L5 PT (Fam84b)
L6 CT (Foxp2)
L6 IT (Osr1)


a
Bulk RNA-seq data of mouse hypothalamus region are obtained by Spatial-seq
Bulk data
Genes
1 mm
Hypothalamus
(Barcode X,Y,Z)
Decoding
Isolation of bulk tissue
LCM
Barcodes
mRNA
Collection
Z Y X
Spatial barcoding
RNA-seq
scRNA-seq data of the mouse hypothalamus are derived by Drop-seq
Barcoded beads
Cells
Single-cell reference (GSE113576)
Spatially resolved transcriptomics of the mouse hypothalamic preoptic region by MERFISH
Spatial reference
b
Merged clustering
Astrocytes
Endothelial cells
Ependymal cells
Excitatory neurons
Inhibitory neurons
Microglias
Mural/Pericytes
Oligodendrocytes (OD) Mature
OD Immature
Macrophages
UMAP2
UMAP1
Bulk2Space
MERFISH
c
Spatial mapping of generated single cells
MERFISH
Bulk2Space
Scale bar
No. of matched cells
0
1k
2k
MERFISH
Bulk2Space
Gene expression correlation between cell types
Ambiguous
1
0.75
0.5
0.25
0
d
Prediction of the spatial gene expression
Bulk2Space
MERFISH
Syt4
Lpar1
Prlr
Necab1
ermn
low
high
e
Prediction of the spatial gene expression
Clustering Space
Spatial distribution
Inhibitory neurons
Excitatory neurons
Mural/ Pericytes
Macrophages
Endothelial cells
OD Mature
Astrocytes
Coord y
Coord x
f
Spatial distribution of novel marker genes predicted by Bulk2Space
Astrocytes Nlsr2
Endothelial cells Cldn5
Ependymal cells S100a6
Excitatory neurons Resp18
Inhibitory neurons Syt1
Microglias Ctss
Mural/Pericytes Tagln
OD Mature Mag
Macrophages Lyz2
low
high


a
Prostate cancer (PMID:29925878)
40 mm
50 mm
1.3
2.1
1.2
1.1
2.3
2.4
3.1
3.3
4.1
4.3
3.2
4.2
Normal glands
Inflammation
Cancer Gs 3+3
Cancer Gs 3+3
Cancer Gs 3+4
Cancer Gs 3+4
b
Deconvolution of bulk data by Bulk2Space
CAFs
Endothelial cells
Monocytes
B cells
PVL cells
Cancer cells
NK cells
UMAP2
UMAP1
B cells
CAFs
Cancer cells
Endothelial cells
Monocytes
NK cells
PVL cells
c
Mapping generated single cells to tissue sections
Normal glands (1.1)
Inflammation (2.3)
Cancer Gs 3+4 (3.3)
B cells
CAFs
Cancer cells
Endothelial cells
Monocytes
NK cells
PVL cells
Scale bar
1 mm
d
Predicted proportion of different cell types
100%
75%
50%
25%
0
B cells
CAFs
Cancer cells
Endothelial cells
Monocytes
NK cells
PVL cells
Normal glands
Inflammation
Cancer Gs 3+4
e
Proportion of cell types
All
Normal
Inflammation
Cancer
B cells
CAFs
Cancer cells
Endothelial cells
Monocytes
NK cells
PVL cells
f
Proportions of CAFs in different tissues
Normal glands
Inflammation
Cancer Gs 3+4
g
Proportions of immune cells in different tissues
Normal glands
Inflammation
Cancer Gs 3+4
h
Proportions of cancer cells in different tissues
Normal glands
Inflammation
Cancer Gs 3+4
Proportion
0
0.25
0.5
0.75
1.0