Reference-free cell-type deconvolution of pixel-resolution spatially resolved transcriptomics data

Brendan F. Miller[1,2], Lyla Atta[1,2,3], Arpan Sahoo[1,4], Feiyang Huang[1,2], Jean Fan[1,2,4,*]

[1]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21211, USA

[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

[3]Medical Scientist Training Program, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[4]Department of Computer Science, Johns Hopkins University, Baltimore MD 21218, USA

[*]To whom correspondence should be addressed

Correspondence should be addressed to:

Jean Fan (jeanfan@jhu.edu)

Key words:

Bioinformatics, Computational biology, Gene expression, Single Cell, Deconvolution, Spatial Transcriptomics, Topic Modeling

Abstract

Recent technological advancements have enabled spatially resolved transcriptomic profiling but at multi-cellular pixel resolution, thereby hindering the identification of cell-type spatial co-localization patterns. We developed STdeconvolve as an unsupervised approach to deconvolve underlying cell-types comprising such multi-cellular pixel resolution spatially resolved transcriptomics datasets. We show that STdeconvolve effectively recovers the putative transcriptomic profiles of cell-types and their proportional representation within spatially resolved pixels without reliance on external single-cell transcriptomics references.

Main

Delineating the spatial organization of transcriptionally distinct cell-types within tissues is critical for understanding the cellular basis of tissue function[1]. Recent technologies have enabled spatially resolved transcriptome (ST) profiling within tissues at multi-cellular pixel-resolution[2]. As such, these ST measurements represent cell mixtures that may comprise multiple cell-types. This lack of single-cell resolution hinders the characterization of cell-type specific spatial organization. To address this challenge, supervised deconvolution approaches such as SPOTlight[3] and RCTD[4] have recently been developed to predict the proportion of cell-types within ST pixels. However, these supervised deconvolution approaches rely on the availability of a suitable single-cell reference, which may present limitations if such a reference does not exist due to budgetary, technical[5], or biological limitations[6]. Here, we developed STdeconvolve (available at https://github.com/JEFworks-Lab/STdeconvolve and as Supplementary Software) as an unsupervised, reference-free approach for deconvolving multi-cellular pixel resolution ST data (Figure 1). Given a counts matrix of ST data, STdeconvolve infers the putative

44 transcriptomic profiles of distinct cell-types and their proportional representation within each
45 multi-cellular spatially resolved ST pixel (Methods). Briefly, STdeconvolve first feature selects
46 for genes likely to be informative of transcriptionally distinct cell-types. STdeconvolve then
47 builds on Latent Dirichlet Allocation (LDA)[7] to estimate the number of transcriptionally distinct
48 cell-types, *K*, and deconvolves these *K* cell-types across ST pixels. STdeconvolve leverages
49 several unique features of ST data that make this application of LDA particularly amenable
50 (Supplementary Note 1, 2).

51 As a proof of concept, we first evaluated the performance of STdeconvolve in recovering
52 the proportional representations of cell-types and their transcriptomic profiles using simulated
53 ST data. This was done by aggregating single-cell resolution multiplex error-robust fluorescence
54 *in situ* hybridization (MERFISH) data of the mouse medial pre-optic area (MPOA)[8] into 100
55 μm^2 pixels (Figure 2A, Supplementary Figure S1A-B, S2, Methods). Applying STdeconvolve,
56 we identified *K*=9 transcriptionally distinct cell-types and deconvolved their transcriptomic
57 profiles and proportional representation in each simulated pixel (Figure 2B, Supplementary
58 Figure S1C, S3A). To infer the identities of the deconvolved cell-types for benchmarking
59 purposes, we matched their deconvolved transcriptional profiles with the transcriptional profiles
60 of ground truth cell-types (Methods) (Supplementary Figure S3B-C). We observed strong
61 correlations between the transcriptomic profiles of each deconvolved cell-type and matched
62 ground truth cell-type across genes (Figure 2C) and, likewise, between the proportions of each
63 deconvolved cell-type and matched ground truth cell-type across simulated pixels (Figure 2D).
64 Some deconvolved cell-types such as cell-types X2 and X8 both matched to excitatory neurons
65 while cell-types X4 and X7 both matched to inhibitory neurons. Further partitioning the ground-
66 truth excitatory and inhibitory cell-types into additional sub-types (76 total) based on previous

67 annotations[8] found that these deconvolved cell-types correlated with specific combinations of
68 neuronal sub-types (Supplementary Figure S3D). When we further expanded the number of
69 deconvolved cell-types to K=76, we were able to identify deconvolved cell-types that are highly
70 correlated in terms of both transcriptional profiles and pixel proportions to finer neuronal
71 subtypes as well as rare cell-types such as pericytes and microglia (Supplementary Figure 3E-F,
72 Supplementary Note 2, 3). In addition, as current ST technologies allow for spatial
73 transcriptomic profiling at varying resolutions[2], we further simulated another ST dataset at 20
74 μm^2 resolution and observed similarly strong correlations between the deconvolved cell-type
75 transcriptomic profiles and proportions with the ground truth by STdeconvolve (Supplementary
76 Figure S4).

77 Using our simulated 100 μm^2 resolution ST data of the MPOA, we also compared
78 STdeconvolve to existing supervised deconvolution approaches SPOTlight and RCTD. For an
79 ideal single cell transcriptomics reference, we used the original single-cell MERFISH data. We
80 evaluated the performance of each approach using the root-mean-square-error (RMSE) of the
81 deconvolved cell-type proportions compared to ground truth across simulated pixels (Methods).
82 In general, we find the performance of STdeconvolve to be comparable to SPOTlight and RCTD
83 (Supplementary Fig. S5A). One potential limitation of such existing supervised deconvolution
84 approaches is their reliance on a suitable single-cell reference. We thus sought to evaluate their
85 performance when a suitable single-cell reference is not available by removing neuronal cell-
86 types from the MERFISH single-cell reference. Reevaluating performance resulted in an increase
87 in RMSE for both SPOTlight and RCTD (Supplementary Fig. S5B).

88 We next evaluated the performance of STdeconvolve by analyzing 100 μm^2 resolution
89 ST data of the mouse main olfactory bulb (MOB)[9]. The MOB consists of multiple bilaterally

90 symmetric and transcriptionally distinct cell layers due to topographically organized sensory
91 inputs[10]. While previous clustering analysis of MOB ST data revealed coarse spatial organization
92 of coarse cell layers, finer structures such as the rostral migratory stream (RMS) could not be
93 readily observed (Supplementary Figure S6A-B). We applied STdeconvolve to identify K=12
94 transcriptionally distinct cell-types (Figure 2E, Supplementary Figure S6C) that either
95 overlapped with or further split coarse cell layers previously identified from clustering analysis
96 (Supplementary Fig S6D). In particular, while deconvolved cell-type X7 overlapped with the
97 granule cell layer previously identified from clustering analysis, it was spatially placed where the
98 RMS is expected[11] (Figure 2F). Upregulated genes in its deconvolved transcriptional profile,
99 including *Nrep*, *Sox11,* and *Dcx*, are known to be associated with neuronal differentiation or
100 upregulated in neuronal precursor cells within the RMS[12] and mark the expected spatial
101 organization based on ISH staining[13] (Figure 2G, Supplementary Figure S6E). This suggests that
102 deconvolved cell-type X7 corresponds to the neuronal precursor cell-type within the RMS
103 unidentified from clustering analysis. Again, we compared STdeconvolve to SPOTlight and
104 RCTD using an appropriate MOB scRNA-seq reference[14] and found a high degree of
105 correspondence among all evaluated methods (Supplementary Figure S7).
106 We again compared the performance of such supervised deconvolution approaches when
107 a suitable single-cell reference is lacking by removing olfactory ensheathing cells (OECs) from
108 the MOB scRNA-seq reference. OECs were initially predicted to be enriched in the olfactory
109 nerve layer by all evaluated methods (Supplementary Fig S8A-C). However, given this new
110 reference without OECs, SPOTlight and RCTD erroneously predicted N2 cells to be enriched in
111 the olfactory nerve layer and highly abundant (Supplementary Figure S8A, S8D) even though N2
112 cells were initially predicted to be rare by all methods. In addition, we trained SPOTlight and

113 RCTD on a scRNA-seq reference from the mouse cortex[13] resulting in the vascular
114 leptomeningeal (VLMC) cell cluster of the cortex reference to be erroneously predicted as highly
115 enriched in the olfactory nerve layer (Supplementary Figure S9). As such, supervised
116 deconvolution approaches may be sensitive to the single-cell transcriptomics reference used.
117 Finally, to demonstrate the potential of an unsupervised, reference-free approach, we
118 applied STdeconvolve to 100 μm^2 resolution ST data of 4 breast cancer sections[15]. Here, a
119 matched scRNA-seq reference was not available and using a scRNA-seq reference from another
120 breast cancer sample may be inappropriate due to potential inter-tumoral heterogeneity[16].
121 Transcriptional clustering of the ST pixels previously identified 3 transcriptionally distinct
122 clusters that corresponded to 3 histological regions of the tissue: ductal carcinoma *in situ*,
123 invasive carcinoma, and non-malignant[15] (Supplementary Figure S10A-B). However, the tumor
124 microenvironment is a complex milieu of many additional cell-types[17]. We applied
125 STdeconvolve to identify additional cell-types and interrogate their spatial organization,
126 resulting in K=15 identified cell-types (Figure 3A, Supplementary Figure S10C). Of these,
127 deconvolved cell-types X3, X13, and X15 corresponded spatially and had deconvolved
128 transcriptional profiles consistent with the non-malignant, ductal in situ carcinoma, and invasive
129 carcinoma annotations, respectively (Supplementary Figure S10D, S11). However, deconvolved
130 cell-type X15 was spatially enriched at the interface of the cancerous and non-malignant regions
131 of the tissue (Figure 3B). Highly expressed genes for deconvolved cell-type X15 included
132 immune genes such as *CD74* and *CXCL10* and gene set enrichment analysis suggested
133 significant enrichment in immune processes (Figure 3C-E, Supplementary Figure S12,
134 Supplementary Table S1), suggesting that deconvolved cell-type 15 may correspond to immune
135 cells. Such spatial organization of immune cells along a tumor boundary has been previously

136 implicated to play a role in breast cancer prognosis[18]. Therefore, STdeconvolve may be able to
137 assist in deconvolving transcriptionally distinct cell-types in heterogeneous tissues to discover
138 potentially clinically relevant spatial organization.

139 In conclusion, we have developed STdeconvolve as a tool for analyzing ST data to
140 recover cell-type proportion and transcriptional profiles without reliance on single-cell
141 transcriptomics references. We show that STdeconvolve can recapitulate expected biology and
142 provide competitive performance to existing supervised methods when suitable single-cell
143 references are available, as well as potentially superior performance when suitable single-cell
144 references are not available. In general, we anticipate that STdeconvolve will help interrogate the
145 spatial relationships between transcriptionally distinct cell-types in complex heterogeneous
146 tissues.

147

148 **Methods**

149

150 **STdeconvolve Overview**

151 STdeconvolve uses latent Dirichlet allocation (LDA)[7], a generative probabilistic model, to
152 deconvolve the latent cell-types contained within multi-cellular pixels of spatially resolved
153 transcriptome (ST) measurements. In this context, each pixel is defined as a mixture of K cell-
154 types represented as a multinomial distribution of cell-type probabilities (θ), and each cell-type
155 is defined as a probability distribution over the genes (β) present in the ST dataset.

156

157 *LDA Modeling*

158 The ST dataset is represented as a $D \times V$ matrix of discrete gene counts for each pixel d and

159 gene v. The number of total gene counts in a given pixel d is N_d.

160

161 As a generative probabilistic model, the LDA model generates a set of new pixels as follows:

162

163 For each pixel d:

164 a. draw a cell-type distribution $\theta_d \sim Dir(\alpha)$, where θ_d is a multinomial distribution of

165 length K drawn from a uniform Dirichlet distribution with scaling parameter α.

166 b. for each gene count n in N_d:

167 i. draw cell-type $k_{d,n} \sim mult(\theta_d)$

168 ii. draw gene $v_{d,n} \sim mult(\beta_{k_{d,n}})$

169

170 The central goal is to identify the posterior distribution of the latent parameters given the input

171 data, where for each pixel d:

172 $$p(\theta_d, \boldsymbol{k} \mid \boldsymbol{v}, \alpha, \beta) = \frac{p(\theta_d, \boldsymbol{k}, \boldsymbol{v} \mid \alpha, \beta)}{p(\boldsymbol{v} \mid \alpha, \beta)}$$

173 where $\boldsymbol{k}$ is a vector of N_d cell-types associated with each gene in pixel d, and $\boldsymbol{v}$ is the vector of

174 N_d genes for pixel d. A variational expectation-maximization approach is used to estimate the

175 values of the latent parameters[7,19]. By default, β is initialized with 0 for all cell-types and genes,

176 and α as 50/K.

177

178 The resulting estimated θ and β matrices represent the deconvolved proportions of cell-types in

179 each pixel and the gene expression profiles for each cell-type, scaled to a library size of 1. β

180 represents a $K \times V$ gene-probability matrix for each cell-type k and each gene v with each row
181 summing to 1. The β matrix can be multiplied by a scaling factor of one million to be more like
182 conventional counts-per-million expression values for interpretability. θ represents a $D \times K$
183 pixel-cell-type proportion matrix for each pixel d and each cell-type k. LDA modeling in
184 STdeconvolve is implemented through the `topicmodels` R package[19].
185
186 Of note, LDA inherently assumes for each pixel, there are a few cell-types present with high
187 probability. We find this assumption reasonable for ST data due to the limited number of cells
188 being captured within an ST pixel (Supplementary Fig. 1). Likewise, LDA assumes for each cell-
189 type, there is a set of genes associated with high probability. Therefore, STdeconvolve uses
190 feature selection for genes more likely to be associated with cell-types, which can improve cell-
191 type deconvolution.
192
193 *Selection of genes for LDA model*
194 To filter for genes that are more likely to be specifically expressed in particular cell-types to
195 improve cell-type deconvolution by LDA, STdeconvolve first removes genes that are not
196 detected in a sufficient number of pixels. By default, genes detected in less than 5% of pixels are
197 removed. Because LDA attempts to identify tightly occurring, and ideally non-overlapping
198 clusters of genes in the pixels, the most expressed genes in the dataset, as well as genes that are
199 expressed in many pixels may also be removed. By default, genes detected in 100% of pixels are
200 removed. STdeconvolve then selects for significantly overdispersed genes, or genes with higher-
201 than-expected expression variance across pixels, as a means to detect transcriptionally distinct
202 cell-types[20]. We assume that the proportion of cell-types will vary across pixels and thus

203 differences in their cell-type-specific transcriptional profiles manifest as overdispersed genes
204 across pixels in the dataset. Additional gene filtering or cell-type specific marker genes to
205 include in the input ST dataset may also be augmented by the user.
206
207 *Selection of LDA model with optimal number of cell-types*
208 The number of cell-types K in the LDA model must be chosen *a-priori*. To determine the
209 optimal number of cell-types K to set for an LDA model for a given dataset, we fit a set of LDA
210 models using a different value for K over a user defined range of positive integers greater than 1.
211 Additionally, users may select a held-out subset of pixels to then apply the fitted model to. We
212 then compute the perplexity of each fitted model:

213 $$Perplexity(D) = \exp\left\{-\frac{\log\big(p(D)\big)}{\sum_{d=1}^{D}\sum_{v=1}^{V} n^{(dv)}}\right\}$$

214 Where $p(D)$ is the likelihood of the dataset and $n^{(dv)}$ is the gene count of gene v in pixel d. We
215 can interpret $p(D)$ as the posterior likelihood of the dataset conditional on the cell-type
216 assignments using the final estimated θ and β. The lower the perplexity, the better the model
217 represents the real dataset. Thus, the trend between choice of K and the respective model
218 perplexity can then serve as a guide. By default, the perplexity is computed by comparing $p(D)$
219 to the entire input dataset used to estimate θ and β.
220 In addition, STdeconvolve also reports the trend between K and the number of
221 deconvolved cell-types with mean pixel proportions < 5% (as default). We chose this default
222 threshold based on the difficulty of STdeconvolve, SPOTlight, and RCTD to deconvolve cell-
223 types at low proportions, (i.e., "rare" cell-types) (Supplementary Note 2). We note that as K is
224 increased for fitted STdeconvolve models, the number of such "rare" cell-types generally
225 increases. Such rare deconvolved cell-types are often distinguished by fewer distinct

226 transcriptional patterns in the data and may represent non-relevant or spurious subdivisions of
227 primary cell-types. We can use this metric to help set an upper bound on *K*.
228 Generally, perplexity decreases and the number of "rare" deconvolved cell-types
229 increases as *K* increases. Given these model perplexities and number of "rare" deconvolved cell-
230 types for each tested *K,* the optimal *K* can then be determined by choosing the maximum *K* with
231 the lowest perplexity while minimizing number of "rare" deconvolved cell-types. To further
232 guide the choice of *K*, an inflection point ("knee") is derived from the maximum second
233 derivative of the plotted *K* versus perplexity plot and *K* versus number of "rare" deconvolved
234 cell-types. We find that the optimal *K* is stable for models fitted to similar datasets, and that the
235 deconvolved cell-types are highly similar in terms of their deconvolved transcriptional profiles
236 (Supplementary Note 4). Ultimately, the choice of *K* is left up to the user and can be chosen
237 taking into consideration prior knowledge of the biological system.
238
239 **Simulating ST data from single-cell resolution spatially resolved MERFISH data**
240 MERFISH data of the mouse medial preoptic area (MPOA) was obtained from the original
241 publication[8]. Normalized gene expression values were converted back to counts by dividing by
242 1000 and multiplying by each cell's absolute volume. Datasets for an untreated female animal
243 (FN7, datasets 171021_FN7_2_M22_M26 and 171023_FN7_1_M22_M26) containing counts
244 for 135 genes assayed by MERFISH were used. Genes with non-count expression intensities
245 assayed by sequential FISH were omitted. Counts of blank control measurements were also
246 removed. Cells were previously annotated as being one of 9 major cell-types (astrocyte,
247 endothelial, microglia, immature or mature oligodendrocyte, ependymal, pericyte, inhibitory
248 neuron, excitatory neuron). Cells originally annotated as "ambiguous" were removed from the

249 dataset to ensure the ground truth was composed of cells with known cell-types. Because certain
250 cell-types are only present in specific regions of the MPOA, we combined 12 tissue sections
251 across the anterior and posterior regions to ensure that all expected cell-types would be present in
252 the final simulated ST dataset. After filtering, the final dataset contained 59651 cells representing
253 9 total cell-types and counts for the 135 genes.

254 To simulate a multi-cellular pixel resolution ST dataset from such single-cell resolution
255 spatially resolved MERFISH data, we generated a grid of squares, each square with an area of
256 100 μm^2. Each square was considered a simulated pixel and the gene counts of cells whose x-y
257 centroid was located within the coordinates of a square pixel were summed together. A grid of
258 square pixels was generated for each of the 12 tissue sections separately and the simulated pixels
259 for all 12 tissue sections were subsequently combined into a single ST dataset. For a given tissue
260 section, the bottom edge of the grid was the lowest y-coordinate of the cell centroids and the left
261 edge of the grid was the lowest x-coordinate. Square boundaries were then drawn from each of
262 these edges in 100 μm^2 increments until the position of the farthest increment from the origin
263 was greater than the highest respective cell centroid coordinate. After generating the grid, square
264 pixels whose edges formed one of the outside edges of the grid were discarded in order to
265 remove simulated pixels, which by virtue of their placement, encompassed space outside of the
266 actual tissue sample. The retained pixels covered 49142 out of the original 59651 cells in the 12
267 tissue sections. This resulted in a simulated ST dataset with 3072 pixels by 135 genes. We used
268 the original cell-type labels of each cell to compute the ground truth proportions in each
269 simulated pixel. Likewise, to generate the ground truth transcriptional profiles of each cell-type,
270 we averaged the gene counts for cells of the same cell-type from the original 59651 cells and
271 normalized the resulting gene count matrix to sum to 1 for each cell-type. To simulate pixels of

272 20 μm^2, an identical approach was taken using the same cells except that square boundaries were
273 drawn from each edge in 20 μm^2 increments.
274
275 **Deconvolution of simulated MERFISH ST data**
276 STdeconvolve was applied to the simulated MERFISH MPOA ST dataset. We selected the
277 model with the *K* with the lowest perplexity and where the number of "rare" cell-types = 0
278 resulting in *K* = 9 detected cell-types. To compare deconvolved cell-types to the ground truth
279 cell-types in the simulated ST dataset, we computed the Pearson's correlation between every
280 combination of deconvolved cell-type and ground truth cell-type transcriptional profile.
281 Likewise, the Pearson's correlation between the pixel proportions of each deconvolved cell-type
282 and ground truth cell-type was computed. After assignment of deconvolved to ground truth cell-
283 types, the ranking of each gene based on its expression level in the transcriptional profile of the
284 deconvolved or ground truth cell-type for each assigned match was compared.
285
286 *Annotation and matching of deconvolved and ground truth cell-types*
287 Each deconvolved cell-type was first matched with the ground truth cell-type that had the highest
288 Pearson's correlation between their transcriptional profiles. This was done by computing the
289 Pearson's correlation between every combination of deconvolved and ground truth cell-type
290 transcriptional profile.
291 The assignment of deconvolved cell-types to ground truth cell-types was confirmed by
292 testing for enrichment of differentially upregulated genes of the ground truth cell-types in the
293 deconvolved cell-type transcriptional profiles. To determine the differentially upregulated genes
294 of the ground truth cell-types, ground truth transcriptional profiles were converted to counts per

295 thousand and low expressed genes, defined as those with average expression values less than 5,
296 were removed. For each ground truth cell-type, the $\log_2$ fold-change of each remaining gene with
297 respect to the average expression across the other ground-truth cell-types was computed.
298 Differentially upregulated genes were those with $\log_2$ fold-change > 1. We performed rank-based
299 gene set enrichment analysis of the ground truth upregulated gene sets in each deconvolved cell-
300 type transcriptional profile using the `liger` R package[21]. A match to a ground truth cell-type was
301 confirmed and assigned if the ground truth gene set had the lowest gene set enrichment adjusted
302 p-value that was at least < 0.05, followed by the highest positive enrichment score to break ties.
303
304 *Comparison to SPOTlight and RCTD*
305 For both SPOTlight and RCTD, a single cell transcriptomic profile reference was
306 required. To construct this reference, the matrix of gene counts for the 49142 individual cells
307 included in the simulated ST dataset and their predefined cell-type labels were input into the
308 `seurat` R package[22] (v4.0.1) as recommended in both the SPOTlight and RCTD pipelines. In
309 SPOTlight, a minimum cell-type proportion threshold is set to remove cell-types contributing
310 low amounts to pixels. To be consistent across methods, after deconvolution, cell-types in each
311 pixel whose proportions were less than the lowest ground truth pixel proportion for a cell-type
312 (2.5%) were removed, and the remaining cell-type proportions in a pixel were adjusted to sum to
313 1.
314 To compare the performance between STdeconvolve, SPOTlight, and RCTD, the root
315 mean squared error (RMSE) was computed for each pixel between the deconvolved and matched
316 ground truth cell-type proportions for each pixel in the ST dataset:

317 $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

318 where n is the number of cell-types, $\hat{y}_i$ is the predicted cell-type proportion for the cell-type i,
319 and y_i is the ground truth cell-type proportion for the cell-type i.
320
321 To compare the accuracy of each method deconvolving individual cell-types, the RMSE
322 between the predicted pixel proportions for a given cell-type and the matched ground truth cell-
323 type proportions across MERFISH ST dataset pixels was computed. Here, n is the number of
324 dataset pixels, $\hat{y}_i$ is the predicted cell-type proportion for the given cell-type in pixel i, and y_i is
325 the ground truth cell-type proportion for the given cell-type in pixel i. Because the RMSE scales
326 with abundance, the RMSE for each cell-type were divided by the standard deviation of the pixel
327 proportions for the corresponding ground truth-type to compare RMSEs across cell-types.
328
329 **Deconvolution of simulated MERFISH ST data using a reference with missing cell-types**
330 To simulate a single-cell reference with missing cell-types, cells annotated as "excitatory" and
331 "inhibitory" were removed from the previously constructed MERFISH single cell transcriptomic
332 profile reference and used to train SPOTlight and RCTD as described under `*Comparison to*
333 *SPOTlight and RCTD*`. The new trained models were then reapplied to the simulated ST
334 MERFISH dataset of 3072 pixels for deconvolution. After deconvolution, cell-types in each
335 pixel whose proportions were less than the lowest ground truth pixel proportion for a cell-type
336 (2.5%) were removed and the remaining cell-type proportions were adjusted to sum to 1, as
337 described previously. Pixel RMSEs were computed as described above based on the

338 deconvolved cell-type proportions and the ground truth dataset, which retained excitatory and
339 inhibitory neuronal cell-types.
340
341 **Deconvolution of ST data of the mouse olfactory bulb (MOB)**
342 Mouse olfactory bulb datasets were obtained from the original publication[9]. We focused on
343 MOB replicate #8, as the primary MOB ST dataset in this work. We first removed genes with
344 less than 100 reads detected across pixels and pixels with fewer than 100 total gene counts,
345 resulting in a cleaned dataset of 260 pixels and 7365 genes. Overdispersed genes were
346 determined were defined as genes with higher-than-expected observed expression variance
347 across the pixels[20]. Expression variance was modeled based on the expression magnitude using a
348 general additive model with a basis of 5. The p-value of a gene being overdispersed was
349 determined using the cumulative distribution function of the $\mathcal{X}^2$ distribution with degrees of
350 freedom equal to the number of pixels – 1. A gene was overdispersed if the multiple testing
351 adjusted p-value was < 0.05. For MOB replicate #8, we obtained 255 overdispersed genes. We
352 used STdeconvolve to fit LDA models with a range of integer *K*s from 2 to 20 and chose the
353 model with *K*=12, which was within the range of *K*'s that produced the lowest perplexity and the
354 number of "rare" cell-types with mean pixel proportion < 5% was 0 (Supplementary Figure
355 S6C). After deconvolution, cell-types in each pixel whose proportions were less than 5% were
356 removed and the remaining cell-type proportions in each pixel were adjusted to sum to 1.
357 Without a ground truth reference, this filtering threshold was based on the variable performance
358 of the different deconvolution methods to accurately deconvolve cell-types represented below
359 this pixel proportion (Supplementary Note 2).

360 For SPOTlight and RCTD, we used a previously generated scRNA-seq reference of the
361 MOB[14]. We retained only cells collected from untreated wildtype animals and the resulting
362 matrix encompassed 17709 cells representing 38 previously annotated cell-type clusters and raw
363 counts for 18560 genes. The trained models were then applied to deconvolve cell-types in the
364 cleaned MOB replicate #8 ST dataset of 260 pixels and 7365 genes. After deconvolution, cell-
365 types in each pixel whose proportions were less than 5% were removed and the remaining cell-
366 type proportions were adjusted to sum to 1.
367
368 **Deconvolution of ST data of the mouse olfactory bulb (MOB) using a reference with**
369 **missing cell-types**
370 To simulate a single-cell reference with missing cell-types, cells of the MOB scRNA-seq
371 reference that were part of "OEC" clusters 1-5 were removed. SPOTlight and RCTD were
372 trained using this new reference and the newly trained models were then reapplied to the cleaned
373 MOB replicate #8 ST dataset of 260 pixels and 7365 genes for deconvolution. After
374 deconvolution, cell-types in each pixel whose proportions were less than 5% were removed and
375 the remaining cell-type proportions were adjusted to sum to 1.
376
377 **Deconvolution of ST data of the mouse olfactory bulb (MOB) using cortex reference**
378 For the scRNA-seq reference of the mouse cortex, we used the scRNA-seq cortex dataset
379 provided by SPOTlight containing 1404 cells representing 23 transcriptionally distinct clusters
380 and 34617 genes. SPOTlight and RCTD were trained using this reference and the newly trained
381 models were then reapplied to the cleaned MOB replicate #8 ST dataset of 260 pixels and 7365

genes for deconvolution. After deconvolution, cell-types in each pixel whose proportions were less than 5% were removed and the remaining cell-type proportions were adjusted to sum to 1.

Comparison between STdeconvolve, SPOTlight, and RCTD

As neither RCTD nor SPOTlight returns transcriptional profiles, we compared methods by evaluating the Pearson's correlation between the pixel proportions of each deconvolved cell-type from any two methods. Cell-type clusters based on the MOB reference that were deconvolved by RCTD or SPOTlight were matched to STdeconvolve cell-types that had the highest Pearson's correlation.

Deconvolution of ST data of breast cancer sections

ST datasets of 4 breast cancer sections were obtained from the original publication[15]. Genes with less than 10 reads across pixels or pixels with less than 10 total reads were removed from each dataset and genes present in more than 95% of pixels for given dataset were removed. Overdispersed genes were determined for each dataset as described in `*Deconvolution of ST data of the mouse olfactory bulb (MOB)*` using the same parameters. After, we combined the 4 breast cancer datasets into a single dataset of 1029 pixels with counts for 372 genes found to be overdispersed in at least one dataset. We trained LDA models on this combined dataset with STdeconvolve using a range of K from 2 to 20 and selected K=15, which was within the range of K's that produced the lowest perplexity and the number of "rare" cell-types with mean pixel proportion < 5% was 0 (Supplementary Figure S10C).

Gene set enrichment analysis of deconvolved breast cancer cell-types

405 To interpret the transcriptional profiles of the deconvolved cell-types in ST data of breast cancer
406 sections, we used gene set enrichment analysis as implemented in the `liger` R package[21]. We
407 filtered the list of 16771 Homo sapiens Gene Ontology gene set terms[23] to include those which
408 contained at least 1 gene present in the input ST dataset corpus used with STdeconvolve,
409 resulting in 4238 terms. We then performed iterative gene set enrichment analysis on the ranked
410 expression profile of the genes as previously described in `*Annotation and matching of*
411 *deconvolved and ground truth cell-types*`
412
413 **Clustering analysis of ST pixels**
414 For clustering analysis of the MOB, using the cleaned MOB replicate #8 ST dataset of 260 pixels
415 and 7365 genes, the raw counts were normalized to counts per million and adjusted to a $\log_{10}$
416 scale with pseudo count 1. Subsequently, dimensionality reduction using PCA was performed,
417 and pixels were visualized using 2-D embedding with t-SNE on the top 5 principal components
418 and perplexity = 30. Graph-based cluster detection using Louvain clustering[24] was performed
419 using the top 5 principal components with the maximum number of nearest neighbors equal to
420 30, resulting in the assignment of pixels to 5 clusters, which were manually annotated based on
421 the physical locations of the pixel clusters on the MOB tissue section[9].
422 For clustering analysis of the breast cancer sections, we took the combined dataset of
423 1029 pixels and 372 overdispersed genes, and $\log_{10}$ transformed with pseudo count of 1, and
424 dimensionality reduction using PCA was performed. In a manner similar to the original
425 publication, pixels were clustered into 3 groups using the "Ward.D" method and the Euclidean
426 distance calculated using the first 2 principle components This resulted in the assignment of

427 pixels to 3 clusters which corresponded to the annotations of the 3 histological sections

428 previously annotated by pathologists[15].

429

430 **Availability of Code**

431 STdeconvolve is available as an open-source R software package[25] with the source code

432 available in the Supplemental Material and on GitHub at https://github.com/JEFworks-

433 Lab/STdeconvolve. Additional documentation and tutorials are available at

434 https://jef.works/STdeconvolve/

435

436 **References**

437

438 1 Zhuang, X. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nat*
439 *Methods* **18**, 18-22, doi:10.1038/s41592-020-01037-8 (2021).

440 2 Larsson, L., Frisen, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new
441 dimension to genomics. *Nat Methods* **18**, 15-18, doi:10.1038/s41592-020-01038-7
442 (2021).

443 3 Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF
444 regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.
445 *Nucleic Acids Res*, doi:10.1093/nar/gkab043 (2021).

446 4 Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics.
447 *Nat Biotechnol*, doi:10.1038/s41587-021-00830-w (2021).

448 5 Kiemen, A. *et al.* In situ characterization of the 3D microanatomy of the pancreas and
449 pancreatic cancer at single cell resolution. *bioRxiv*, 2020.2012.2008.416909,
450 doi:10.1101/2020.12.08.416909 (2020).

451 6 Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental
452 Considerations for Single-Cell RNA Sequencing Approaches. *Front Cell Dev Biol* **6**, 108,
453 doi:10.3389/fcell.2018.00108 (2018).

454 7 Blei, D. M. a. N., Andrew Y and Jordan, Michael I. Latent dirichlet allocation. *The*
455 *Journal of Machine Learning Research* **3**, 993-1022 (2003).

456 8 Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the
457 hypothalamic preoptic region. *Science* **362**, doi:10.1126/science.aau5324 (2018).

458 9 Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by
459 spatial transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).
460 10 Nagayama, S., Homma, R. & Imamura, F. Neuronal organization of olfactory bulb
461 circuits. *Front Neural Circuits* **8**, 98, doi:10.3389/fncir.2014.00098 (2014).
462 11 Hintiryan, H. *et al.* Comprehensive connectivity of the mouse main olfactory bulb:
463 analysis and online digital atlas. *Front Neuroanat* **6**, 30, doi:10.3389/fnana.2012.00030
464 (2012).
465 12 Wang, C. *et al.* Identification and characterization of neuroblasts in the subventricular
466 zone and rostral migratory stream of the adult human brain. *Cell Res* **21**, 1534-1550,
467 doi:10.1038/cr.2011.83 (2011).
468 13 Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature*
469 **445**, 168-176, doi:10.1038/nature05453 (2007).
470 14 Tepe, B. *et al.* Single-Cell RNA-Seq of Mouse Olfactory Bulb Reveals Cellular
471 Heterogeneity and Activity-Dependent Molecular Census of Adult-Born Neurons. *Cell*
472 *Rep* **25**, 2689-2703 e2683, doi:10.1016/j.celrep.2018.11.034 (2018).
473 15 Yoosuf, N., Navarro, J. F., Salmen, F., Stahl, P. L. & Daub, C. O. Identification and
474 transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Res* **22**,
475 6, doi:10.1186/s13058-019-1242-9 (2020).
476 16 Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in
477 TNBC through single-cell RNA-seq. *Nat Commun* **9**, 3588, doi:10.1038/s41467-018-
478 06052-0 (2018).

17 Wei, R., Liu, S., Zhang, S., Min, L. & Zhu, S. Cellular and Extracellular Components in Tumor Microenvironment and Their Application in Early Diagnosis of Cancers. *Anal Cell Pathol (Amst)* **2020**, 6283796, doi:10.1155/2020/6283796 (2020).

18 Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174**, 1373-1387 e1319, doi:10.1016/j.cell.2018.08.039 (2018).

19 Grün, B. & Hornik, K. topicmodels: An R Package for Fitting Topic Models. *2011* **40**, 30, doi:10.18637/jss.v040.i13 (2011).

20 Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* **13**, 241-244, doi:10.1038/nmeth.3734 (2016).

21 Fan, J. Differential Pathway Analysis. *Methods Mol Biol* **1935**, 97-114, doi:10.1007/978-1-4939-9057-3_7 (2019).

22 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *bioRxiv*, 2020.2010.2012.335331, doi:10.1101/2020.10.12.335331 (2020).

23 Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740, doi:10.1093/bioinformatics/btr260 (2011).

24 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008, doi:10.1088/1742-5468/2008/10/p10008 (2008).

25 Team, R. C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2021).

Main Figures

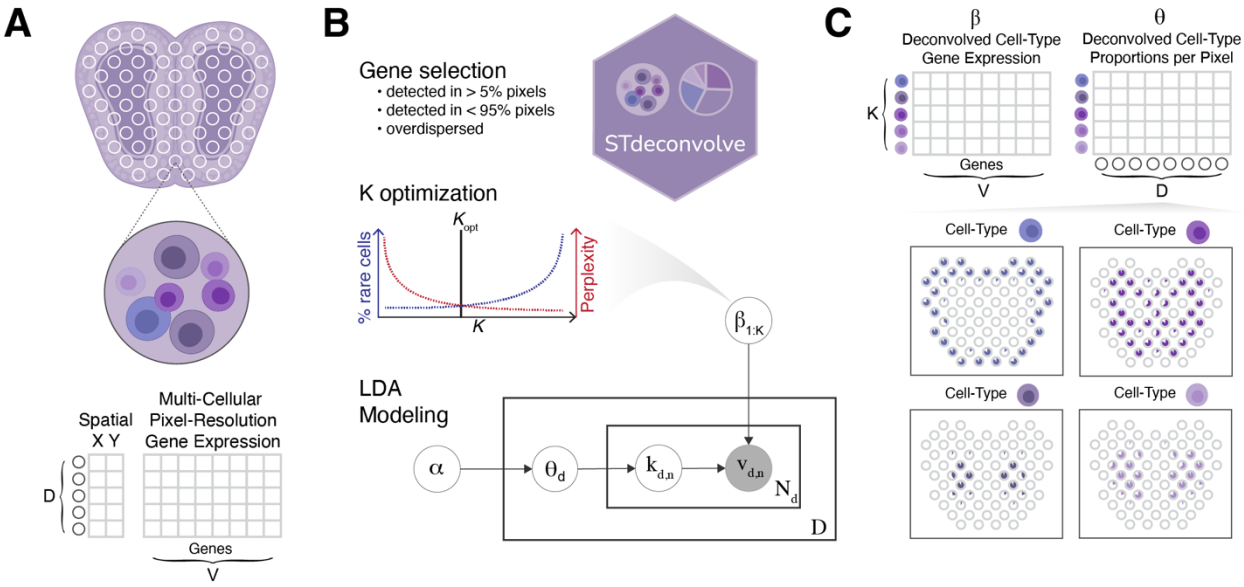

Figure 1. Overview of STdeconvolve.

A. STdeconvolve takes as input a spatial transcriptomics (ST) dataset of *D* pixels (rows) and counts of *V* genes (columns). A matrix of spatial coordinates for each of the *D* pixel can also be used for visualization.

B. STdeconvolve first feature selects genes to retain in the input matrix for deconvolution, such as genes with counts in more than 5% and less than 95% of the pixels, and overdispersed across the pixels. To determine the optimal number of cell-types to be deconvolved, *K*, STdeconvolve fits multiple LDA models to the input dataset each with a different *K*. STdeconvolve computes the perplexity and number of rare deconvolved cell-types to guide the selection of the model with the optimal *K*. A graph representation of LDA modeling is shown, where $\beta_{1:K}$ is a $K \times V$ gene-probability matrix for each cell-type k and each input matrix gene v; *D* is the number of pixels in the dataset; N_d is the total gene counts in pixel *d*; α is the Dirichlet distribution scaling parameter; θ_d is the multinomial distribution of cell-type proportions in pixel *d* drawn from $Dir(\alpha)$; $k_{d,n}$ is a

drawn cell-type from $mult(\theta_d)$ for the nth gene count in pixel d; $v_{d,n}$ is a count of gene v drawn from gene-probability $mult(\beta_{k_{d,n}})$, given the cell-type $k_{d,n}$. (See Methods for details). Shaded circle indicates observed variables and clear circles indicate latent variables.

C. STdeconvolve outputs two matrices: (1) β, the deconvolved transcriptional profile matrix of K cell-types over genes V, and (2) θ, the proportions of K cell-types across the D pixels. The proportion of deconvolved cell-types can then be visualized across the pixels.

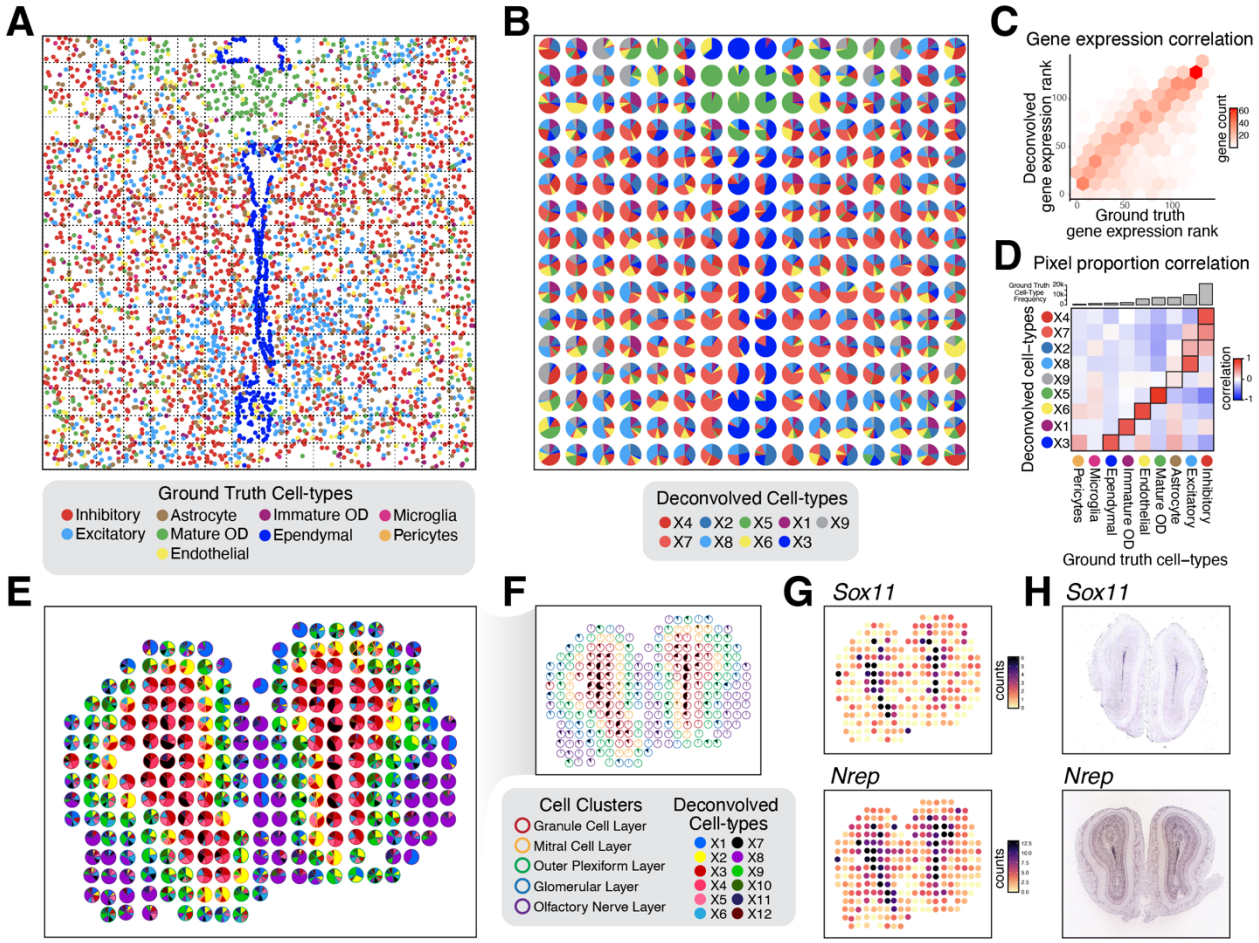

Figure 2. STdeconvolve characterizes the spatial organization of cell-types in simulated and real ST data.

A. Ground truth single-cell resolution MERFISH data of one section of the MPOA partitioned into 100 μm^2 pixels (black dashed squares). Each dot is a single cell colored by its ground truth cell-type label.

B. Proportions of deconvolved cell-types from STdeconvolve, represented as pie charts for each simulated pixel.

C. The ranking of each gene based on its expression level in the transcriptional profiles of the deconvolved cell-types, compared to its gene rank in the transcriptional profile of the matched ground truth cell-type.

537 D. Heatmap of Pearson's correlations between the proportions of the deconvolved cell-types
538 and ground truth cell-types across simulated pixels. Ground truth cell-types are ordered
539 by their frequencies in the ground truth dataset. Matched deconvolved and ground truth
540 cell-types are boxed.
541 E. Deconvolved cell-type proportions for ST data of the MOB from STdeconvolve,
542 represented as pie charts for each ST pixel. Pixels are outlined with colors based on the
543 pixel transcriptional cluster assignment corresponding to MOB coarse cell layers.
544 F. Highlight of deconvolved cell-type X7. Pixel proportion of deconvolved cell-type X7 are
545 indicated as black slices in pie charts. Pixels are outlined with colors as in E).
546 G. Gene counts in each pixel of the MOB ST dataset for deconvolved cell-type X7's select
547 top marker genes *Sox11* and *Nrep*.
548 H. Corresponding ISH images for deconvolved cell-type X7's top marker genes *Sox11* and
549 *Nrep* from the Allen Brain Atlas[13].
550

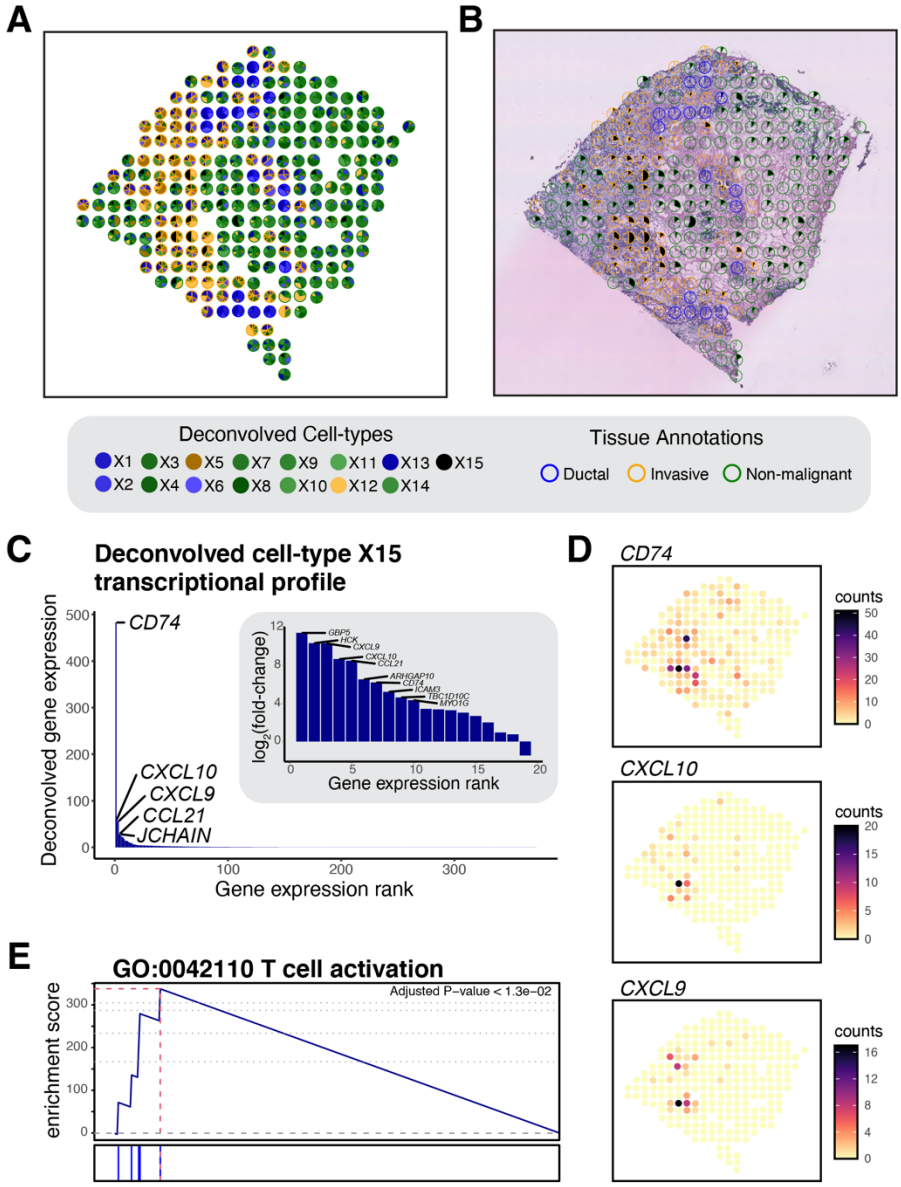

Figure 3. STdeconvolve characterizes the spatial organization of immune cells in breast cancer ST data.

A. Deconvolved cell-type pixel proportions for ST data of a breast cancer tissue section, represented as pie charts. Pixels are outlined with colors based on the pixel transcriptional cluster assignment corresponding to 3 pathological annotations.

557 B. Highlight of deconvolved cell-type X15. Pixel proportion of deconvolved cell-type X15
558 are indicated as black slices in pie charts. Pixels are outlined with colors as in A). An
559 H&E-stained image of the breast cancer tissue is shown in the background.
560 C. Barplot of the deconvolved transcriptional profile of deconvolved cell-type X15 ordered
561 by magnitude of deconvolved gene expression. Inset represents the $\log_2$ fold-change of
562 the deconvolved transcriptional profile genes with respect to the mean expression of the
563 other 14 deconvolved cell-type transcriptional profiles. Select highly expressed and high
564 fold-change genes are labeled.
565 D. Gene counts in each pixel of the breast cancer ST dataset for deconvolved cell-type
566 X15's select top marker genes.
567 E. Gene set enrichment plot for significantly enriched GO term "T cell activation" for
568 deconvolved cell-type X15's transcriptional profile.
569