RESCUE: recovery of unattributed expression patterns in spatial transcriptomics

Young Joo Lee[1], Seokjin Yeo[2], Alex W. Schrader[3], Ian M. Traniello[4], Amy Cash Ahmed[5], Gene E. Robinson[5,6,7], Hee-Sun Han[2,3,5,6*], Sihai Dave Zhao[1,5*]

[1]Department of Statistics, University of Illinois Urbana-Champaign

[2]Department of Bioengineering, University of Illinois Urbana-Champaign

[3]Department of Chemistry, University of Illinois Urbana-Champaign

[4]Lewis-Sigler Institute for Integrative Genomics, Princeton University

[5]Carl R. Woese Institute of Genomic Biology, University of Illinois Urbana-Champaign

[6]Neuroscience Program, University of Illinois Urbana-Champaign

[7]Department of Entomology, University of Illinois Urbana-Champaign

* Corresponding authors: sdzhao@illinois.edu (S.D.Z), hshan@illinois.edu (H.-S.H)

ABSTRACT

Spatial transcriptomics (ST) enables gene expression profiling while preserving the spatial architecture of intact tissue. Analyzing ST data often proceeds by first extracting cell-level information, typically through cell segmentation or cell-type deconvolution. However, a critical oversight has been that a substantial portion of molecular expression is systematically lost or unannotated by these methods. This lost expression can arise from diverse and biologically important sources like fragile or underrepresented cell types, subcellular structures like neurites, and extracellular expression. These omissions can result in biased analyses and incorrect or incomplete biological interpretations. We describe a new computational method, RESCUE, that can recover the unattributed spatial expression patterns missed by existing ST analysis methods and enable robust inference even when reference is incomplete. We validate RESCUE using MERFISH data from the honey bee brain and apply it to multiple ST datasets to demonstrate how it can reveal novel insights into complex tissue biology.

INTRODUCTION

Profiling cellular expression within its spatial context has become possible with the recent development of spatial transcriptomics (ST) technologies, which measures gene expression in intact tissue sections[1,2]. Analyzing ST data typically proceeds by extracting cell-level information from the spatial expression data, as cells are natural units of analysis. With high-resolution platforms such as MERFISH[3], 10x Visium HD[4], or 10x Xenium[5], this is usually done using cell segmentation[6–8]. With low-resolution platforms like 10x Visium[9] and Slide-seq[10,11], where each spatial location can contain multiple cells, it is common to deconvolve the observed expression into separate contributions from different cell types, based on reference single-cell/nuclei RNA-seq (sc/snRNA-seq) data[12–16].

However, a major problem with these existing ST analysis methods is that they can fail to capture important expression patterns. First, they may omit entire cells or cell types. Cell segmentation algorithms for high-resolution data have difficulty with cells with complex morphologies, such as fibroblasts, or cells lacking nuclei, like erythrocytes[17,18]. Deconvolution methods for low-resolution data cannot account for cell types that are absent from the reference sc/snRNA-seq data because they are too fragile to survive tissue dissociation[19–21]. Second, existing ST approaches can overlook expression patterns unique to cellular substructures. Segmentation methods simply drop transcripts located in substructures that are difficult to segment, such as the axons and dendrites of neurons. These fragile structures are also typically destroyed during tissue dissociation and therefore are not represented in reference single-cell data used in deconvolution. Indeed, it has been estimated that over 40% of the total transcriptome of an adult mouse brain may be missed by sc/snRNA-seq[22]. Similarly, reference snRNA-seq data, by

55 definition, ignores cytoplasmic expression. Finally, existing methods do not capture
56 extracellular tissue expression patterns, which cannot be studied by segmentation or
57 deconvolution.

58 These difficult-to-capture expression patterns can be critically important. First, they
59 can bias analyses that do not take them into account. An example with dramatic
60 implications is that deconvolution estimates of cell-type proportions can be invalid if the
61 reference is missing any cell types[23], as we illustrate later in this work. Second, there may
62 be important biological processes that are unique to the overlooked expression patterns.
63 For example, important missed cell types include adipocytes[5], which play crucial roles in
64 metabolic disease and cancer[24], and neutrophils[25], which are core components of the
65 immune system. Important missed substructural expression includes axonal or dendritic
66 transcripts that are central to synaptic development and plasticity[26–30] and diseases such
67 as amyotrophic lateral sclerosis[31,32], as well as cytoplasmic transcripts that regulate key
68 processes such as metabolism, signaling, and stress responses[33–35] and are involved in
69 Alzheimer's disease[36,37]. Important missed extracellular expression patterns include
70 extracellular RNAs that mediate intercellular communication in cancer[38–40].

71 We describe a new computational method, RESCUE, that can recover the spatial
72 expression patterns missed by single-cell inference on ST data. Our approach is based
73 on making a fundamental correction to the class of statistical models currently used for
74 integrating ST data and reference data, such as sc/snRNA or cell segmentation results.
75 Briefly, our method partitions observed gene expression into one component explainable
76 by the reference data, which we refer to as “canonical” expression, and another
77 component that is not, which we refer to as “unattributed” expression. Our purified

canonical ST expression component can be used to correct existing cell-type deconvolution methods. Our recovered unattributed ST expression component gives us the new ability to specifically target biology that is distinct from our reference data, as discussed above. The unattributed expression recovery step in RESCUE is analogous to negative selection in experimental biology[41–44], where we identify and remove canonical expression. We demonstrate that RESCUE accurately recovers unattributed expression, enables robust deconvolution of ST data despite incomplete reference data, and reveals biologically meaningful patterns overlooked by existing methods. While we focus here on ST data, we note that RESCUE can also be directly applied to non-spatial modalities, such as bulk RNA-seq data.

RESULTS

Overview of RESCUE

RESCUE is predicated on a new model that we propose for integrating ST and reference single-cell data, and can be applied to diverse ST platforms, including MERFISH and Visium, across different organisms. Most existing models[12–16] posit that ST expression at each location is composed of a linear combination of expression patterns present in the reference, where the coefficients of this linear combination are interpreted as the proportions of each reference cell type in that location. However, as described in the introduction, there are numerous sources of ST expression that are left out of this model; we will refer to these as unattributed expression patterns, because they are not attributed to canonical expression patterns reflected in the reference.

To address this issue, we propose a corrected model (**Fig. 1**) that allows the ST profile to have both canonical and unattributed expression components. While it appears minor, this change has important implications. Under this model, results from the statistics literatures[45–47] imply that under certain conditions it is possible to simultaneously recover unattributed expression and estimate the linear combination of reference patterns that constitute canonical expression. The conditions are that only relatively few genes show non-zero, or “sparse” expression, in the unattributed component, which is reasonable given the nature of unattributed expression and the read depth/detection efficiency of ST technologies (**Methods**).

Leveraging this insight, we developed RESCUE, an algorithm that simultaneously deconvolves ST data while rescuing spatial unattributed expression. RESCUE solves a regularized robust multivariate regression problem and can be applied to both low- and high-resolution ST data. Unlike existing deconvolution methods[12–16], it is robust to

113 incomplete references, such as snRNA-seq data which lack cytoplasmic expression
114 profiles. We emphasize that the interpretation of the unattributed component depends on
115 both the biological context of the tissue and the choice of the reference dataset
116 (**Methods**). As a result, the unattributed component may comprise multiple biological
117 sources, including expression from cell types missing in the reference, substructural
118 expression, and extracellular RNA. RESCUE is designed to recover these signals jointly
119 rather than to assign them to predefined categories. The rescued unattributed expression
120 can then be further decomposed using standard unsupervised approaches, such as
121 clustering and non-negative matrix factorization (NMF), to identify distinct expression
122 patterns within the unattributed component. In addition, differentially expressed gene
123 (DEG) and spatially variable gene (SVG) analyses can be applied to characterize the
124 biological and spatial features of the recovered unattributed signals. Careful consideration
125 of biological structures and tissue organization is therefore essential when interpreting
126 the unattributed expression. We illustrate this in a broad range of examples below.
127

128 **Simulations**

129 To evaluate the performance of RESCUE, we conducted simulation studies on synthetic
130 data designed to mimic real ST measurements with varying degrees of unattributed
131 expression. We simulated ST data using real scRNA-seq data from the honey bee brain[48]
132 by introducing synthetic unattributed expression of a small number of genes (**Methods**).
133 Unlike other tissues where cell bodies and non-cell-body structures are spatially
134 intermingled, honey bee brains provide a unique benefit for simulating unattributed
135 expression due to the existence of the neuropil region, where axons and dendrites are

136 spatially segregated from cell bodies. Molecular expression in these regions is largely
137 absent in single-cell sequencing data due to the structural fragility of neurites. During the
138 tissue dissociation step, a necessary step in single-cell sequencing, neurites are mostly
139 sheared off and lost. In contrast, ST measurements performed on intact tissue sections
140 preserve these structures and therefore capture molecular expression within neuropils.

141 To guide the choice of sparsity levels used in our simulations, we examined
142 empirical expression patterns in the neuropil region of our in-house MERFISH data. We
143 found that neuropil regions exhibit highly sparse gene expression, with an average
144 sparsity level of 0.92 across spatial locations (**Fig. 2a**). Here, sparsity refers to the
145 structure of the unattributed component (i.e., that only a subset of genes is present at a
146 given spatial location), rather than to low expression levels (**Methods**). Based on this
147 empirical distribution, we simulated unattributed expression across a range of sparsity
148 levels by varying the proportion of zero entries (ranging from 0.8 to 0.95). In addition, we
149 varied the proportion of unattributed expression (ranging from 0 to 0.8) to cover different
150 levels of the overall magnitude of unattributed expression. Importantly, these parameter
151 ranges fall within those observed in real biological samples and are not artificially chosen
152 to favor RESCUE.

153 We first compared the deconvolution performance of RESCUE with existing
154 methods. **Fig. 2b** summarizes the accuracy of cell-type proportion estimates across
155 methods, measured using the average Lin's concordance correlation coefficient[49] (CCC),
156 root mean squared error (RMSE), and Jensen-Shannon divergence (JSD) (**Methods**). As
157 expected, the performance of standard deconvolution methods deteriorates sharply as
158 the level of unattributed expression increases. In contrast, RESCUE maintains high

accuracy across all conditions, demonstrating robustness to both high unattributed content and reduced sparsity. These results indicate that failing to explicitly model unattributed expression can lead to systematic biases in cell-type proportion estimates.

We also evaluated the accuracy of unattributed expression recovery (**Fig. 2c**) using the relative Frobenius error between the true and estimated unattributed expression matrices (**Methods**). Overall, RESCUE achieves the lowest recovery error even under high unattributed expression and low sparsity levels, whereas other methods deviate substantially from the ground truth, especially when unattributed signals become strong. These results indicate that explicitly modeling unattributed expression enables accurate recovery of hidden transcriptional signals across a wide range of sparsity and signal-strength regimes.

To assess robustness and generalizability, we expanded the simulation study by varying noise levels via library size variability across spots and varying the effective grid resolution via changes in the average total UMI count per spot (**Supplementary Fig. 1**). As expected, overall performance tends to decrease with increasing library size variability and increase with higher average total UMIs. Across these regimes, RESCUE remains among the top-performing methods, demonstrating robustness across a range of noise levels and grid sizes.

RESCUE accurately recovers the unattributed transcriptome in real tissues

Using a custom MERFISH platform[50,51], we generated single molecule-resolution ST data from coronal sections of a honey bee brain, targeting 130 genes relevant to aggression (**Supplementary Table S1**)[52]. Transcripts in axons and dendrites, which are collectively

182 termed as neurites, have been difficult to profile due to segmentation challenges. Cell
183 segmentation methods, such as Cellpose[6] and Baysor[7], can only identify the boundaries
184 of neuronal cell bodies, also called somata. Although individual axons and dendrites can
185 be stained in 2D cell cultures[53], in native tissues they extend across multiple imaging
186 planes and therefore cannot be reliably segmented as individual cellular units. As a result,
187 neurite-associated expression has traditionally been ignored or incorrectly merged into
188 nearby cytosolic expression.

189 We first experimentally validated that RESCUE can accurately recover gene
190 expression in the neuropil region of the honey bee brain (**Fig. 3a**). This validation is
191 possible because the physical separation of neurites and somata in the brain makes it
192 simple for existing algorithms to separate somatic from unattributed expression. Cell
193 soma were segmented using Cellpose[6], revealing that only 73.5% of transcripts were
194 localized within cells, with the remaining 26.5% comprising the unattributed expression.
195 Using honey bee brain scRNA-seq data[48] as a reference, we then applied RESCUE and
196 competing methods to our MERFISH data after dividing the brain into grids of size
197 $20 \times 20\ \mu m^2$ (**Methods**). For each grid, we estimated the proportion of canonical
198 expression relative to total expression and obtained a ground-truth estimate by identifying
199 transcripts localized within cell bodies via cell segmentation. RESCUE's estimates closely
200 matched the ground truth (**Fig. 3b**) and were more accurate than those from other
201 methods, as quantified by Lin's CCC and RMSE (**Methods**). Notably, RESCUE relies
202 only on reference scRNA-seq data and does not use cell segmentation information.

203 To further clarify the source of discrepancies among methods, we evaluated
204 canonical and unattributed expression separately in anatomically well-defined regions.

205 Specifically, we examined grids corresponding to pure cell bodies and compared the
206 average canonical expression of each gene against the ground truth derived from
207 segmentation (**Fig. 3c**). RESCUE closely matched the true somatic expression levels,
208 whereas competing methods systematically underestimated canonical expression. This
209 behavior reflects the fact that existing deconvolution approaches are designed to estimate
210 cell-type proportions rather than expression levels, and under the subtraction-based
211 framework (**Methods**), expression that is not well explained by the reference profiles is
212 left as residual rather than attributed to the canonical component. Conversely, in grids
213 containing no identifiable cell bodies, RESCUE accurately recovered unattributed
214 expression in neuropil-only regions (**Fig. 3d**), while competing methods were significantly
215 less accurate. Together, these results demonstrate that RESCUE effectively partitions
216 the full transcriptome into canonical and unattributed components in real tissue, capturing
217 valuable expression signals that would otherwise be lost using conventional approaches.
218 Importantly, transcripts outside segmented somata reflect genuine biological
219 signals rather than technical artifacts: in our MERFISH sample (**Fig. 3a**), neuropil-
220 associated transcripts form structured spatial patterns that are clearly segregated from
221 DAPI-stained nuclei (**Supplementary Fig. 2**). Moreover, the MERFISH protocol includes
222 gel embedding, expansion, and clearing steps that reduce autofluorescence and
223 suppress nonspecific signal[54,55]. These signals, which would have been ignored by
224 standard cell segmentation approaches, highlight the importance of RESCUE in capturing
225 overlooked transcriptomic information.
226 Finally, to investigate whether the unattributed expression recovered by RESCUE
227 reflects structured biological organization rather than unstructured noise, we performed

228 clustering analysis of the unattributed expression profiles across spatial locations.
229 Clustering revealed anatomically distinct transcriptional niches within neuropil regions
230 (**Fig. 3e**), indicating that the recovered unattributed component exhibits coherent spatial
231 patterns. Specifically, distinct clusters localize to known neuropil structures, including the
232 mushroom body calyces, antennal lobes, optic lobes, and regions proximal to the central
233 complex. In contrast, spatial locations corresponding to somatic regions collapsed into a
234 single cluster, consistent with the fact that canonical expression is explicitly removed, and
235 only minimal unattributed signal remains in these regions. Examination of representative
236 genes within the neuropil-associated clusters showed spatially structured expression
237 patterns that are largely absent from the canonical component (**Fig. 3f**). These findings
238 further support that the unattributed expression recovered by RESCUE captures
239 biologically meaningful and spatially organized signals, revealing hidden tissue structure
240 that is not accessible through conventional analyses.

241

242 **RESCUE enables accurate deconvolution of ST data with incomplete reference**

243 Here, we illustrate how RESCUE's estimated canonical expression component can be
244 used to achieve accurate deconvolution of ST data when key cell types are missing from
245 the reference, a challenge for all existing methods that has been noted in a recent study[23].
246 We analyzed paired single-cell FFPE-seq (scFFPE-seq) and 10x Visium ST data from
247 human breast cancer tissue[5]. In this ST dataset, adipocytes are manually annotated in
248 approximately 10% of spatial spots (**Fig. 4a**). However, adipocytes are absent from the
249 scFFPE-seq reference, likely due to their high lipid content interfering with fixation and
250 processing. These cells are prone to floating, rupturing, or sticking to plastic surfaces

251 during tissue dissociation, which hinders their recovery in single-cell sequencing
252 protocols[5,17]. As a result, reference-based methods that assume the observed expression
253 is fully explained by the available reference cell types tend to misattribute adipocyte-
254 derived transcripts to other cell types.

255 In adipocyte-rich regions, existing deconvolution methods fail to account for the
256 missing adipocyte transcriptome and instead misattribute the signal to other reference
257 cell types, such as stromal and endothelial cells (**Fig. 4b**). This systematic misassignment
258 likely arises from transcriptional similarity, as shown by the strong correlation between
259 adipocyte and stromal/endothelia profiles in pseudo-bulk measurements from Visium and
260 scFFPE-seq (**Fig. 4c**). In contrast, RESCUE explicitly models unattributed expression
261 and thus avoids forcing unassigned signal into existing reference cell types, which results
262 in more accurate and biologically plausible estimates. As shown in **Fig. 4d-e**, RESCUE's
263 estimated proportions closely align with the scFFPE-seq composition, while other
264 methods exhibit significant distortions. Together, these results demonstrate that RESCUE
265 improves deconvolution accuracy in settings with incomplete references by preventing
266 the systematic misassignment of unattributed transcriptomes to incorrect cell types.

267 To further interpret the biological content of the unattributed expression recovered
268 by RESCUE, we applied NMF to the unattributed component. This analysis revealed
269 multiple latent factors capturing structured expression patterns that are not explained by
270 the canonical component (**Fig. 4f**). Notably, the NMF Factor 1 (Unattributed) captures the
271 spatial distribution of adipocyte-rich regions, effectively recovering a missing cell type
272 absent from the reference. Genes with the highest loadings in this factor include well-
273 established adipocyte markers such as *FABP4*, *CFD*, and *G0S2*[5,56] (**Fig. 4g**). In addition,

274 the NMF Factor 2 (Unattributed) is enriched for genes such as *DCN*, *COL14A1*, and *C1R*,
275 which have been implicated in extracellular matrix (ECM) domains[57–59]. Together, these
276 findings indicate that the unattributed component captured by RESCUE is biologically
277 meaningful, thereby supporting that explicitly modeling it leads to improved deconvolution
278 in the presence of incomplete references.

279 Finally, to assess whether these deconvolution improvements generalize beyond
280 the human breast cancer example, we benchmarked RESCUE using additional datasets,
281 including human dorsolateral prefrontal cortex (DLPFC)[60] and cichlid fish telencephalon[61]
282 Visium samples. In both datasets, reference profiles were derived from snRNA-seq, which
283 inherently underrepresent cytoplasmic and process-localized RNA, leading standard
284 deconvolution methods to misattribute unexplained ST signals. Across these datasets,
285 RESCUE consistently outperformed competing methods (**Supplementary Fig. 3-4**),
286 indicating that the benefits of explicitly modeling unattributed expression extend across
287 tissues, species, and reference modalities.

288

289 **RESCUE reveals unattributed biological processes in diverse tissues**

290 In many neuroscience experiments, cell-level information is often characterized using
291 snRNA-seq[62,63], which can leave a substantial fraction of biologically relevant expression
292 unattributed. Here, we illustrate how RESCUE can uncover unique biological processes
293 within the unattributed expression of the human DLPFC and cichlid fish telencephalon,
294 using snRNA-seq as a reference.

295

296 **Human DLPFC.** Maynard et al.[60] profiled the DLPFC of neurotypical adult donors using
297 paired snRNA-seq and 10x Visium ST. We applied RESCUE to eight samples in which
298 all six cortical layers and the white matter (WM) were annotated, to recover unattributed
299 expression patterns not represented in the snRNA-seq reference (**Fig 5a**). As shown in
300 **Fig. 5b**, RESCUE reveals that Layer 1 and the WM contain substantially higher levels of
301 unattributed expression compared to other cortical layers. This is consistent with the fact
302 that Layer 1 has sparser cell densities and contains neuropil-rich regions[57], and that WM
303 contains a substantial amount of axons and myelin[64], whose transcriptomic content is
304 mostly cytoplasmic and thus is absent from the reference snRNA-seq data.

305 RESCUE enables us to study biology specific to the unattributed transcriptome of
306 Layer 1. We first identified DEGs by comparing Layer 1 to all other layers using limma[65],
307 based on either the raw observed ST expression or the unattributed expression recovered
308 by RESCUE. These two analyses gave substantially different results (**Fig. 5c**). For
309 example, the top five genes most enriched in Layer 1 based on unattributed expression
310 had either fold-changes in the opposite direction (*MT-ND1*, *MT-ND2*, and *MT-CO1*) or
311 negligible fold-changes (*MT3* and *CST3*) in the differential expression analysis using raw
312 data. We also identified genes such as *CLU*, *GFAP*, and *MALAT1* that exhibit clear spatial
313 enrichment within the unattributed component of Layer 1 (**Fig. 5d**). These genes are
314 associated with local translation and neurite outgrowth, and their enrichment outside
315 nucleus is consistent with known roles of astrocytic processes and non-somatic cellular
316 structures in Layer 1 of the cortex[66,67].

317 To explore the underlying biological processes, we performed gene set enrichment
318 analysis separately for DEGs derived from the unattributed expression and from the raw

ST data (**Fig. 5e**). Notably, gene ontology (GO) terms enriched among the unattributed DEGs include mitochondrial processes: “oxidative phosphorylation”, “energy derivation by oxidation of organic compounds”, and “generation of precursor metabolites and energy”. In contrast, no mitochondria-related pathways reached statistical significance in the analysis based on observed ST expression; only the broad term “response to stress” passed the FDR < 0.05 threshold. These results align with the elevated metabolic demands of a neurite-enriched Layer 1, which is densely populated with synapses and reliant on local energy production to support axonal and dendritic activity[68,69]. Collectively, these findings demonstrate that RESCUE can uncover previously obscured yet functionally important gene expression programs in Layer 1 of the human DLPFC.

Cichlid Fish Telencephalon. Johnson et al.[61] collected paired snRNA-seq and Visium ST data from the dorsal telencephalon of individual cichlid fishes exhibiting distinct social behaviors. As shown in **Fig. 6a**, the dorsal telencephalon is characterized by prominent radial glial (RG) processes that originate in the ventricular zone and extend long projections toward superficial pallial regions[70,71]. These RG projections span nearly the entire depth of the pallium and may play a role in shaping spatial gene expression profiles across distinct layers and compartments, but they do not survive tissue dissociation and thus their transcriptomes are not captured by snRNA-seq.

We applied RESCUE to this dataset to recover unattributed transcriptomic signals using the snRNA-seq data as a reference. To assess spatial heterogeneity, we performed SVG analysis using SPARK-X[72] on both the observed ST expression and the unattributed expression. As shown in **Fig. 6b**, a substantial number of SVGs are uniquely identified in

342 the unattributed transcriptome. In **Fig. 6c**, we show the expression maps of those SVGs
343 that were undetectable in the observed ST expression yet exhibited clear spatial structure
344 once nuclear signals were removed. Among these, *app*, *cntn1*, *dner*, and *eno1* mark
345 distinct transcriptional programs that may operate along RG projections. These genes
346 highlight transcriptional signals that are likely to reflect localized molecular activities along
347 RG scaffolds: *app* encodes a transmembrane protein that promotes synapse formation
348 and plasticity[73,74], *cntn1* (Contactin-1) is a GPI-anchored adhesion molecule essential for
349 axon guidance and neurite outgrowth[75], *dner* mediates neuron-glia interactions that are
350 critical for astrocyte differentiation[76], and *eno1* (α-enolase) promotes neuritogenesis and
351 contributes to structural remodeling in the developing nervous system[77]. Indeed, Johnson
352 et al.[61] found that bower-building behavior in cichlid fishes may be related to neuronal
353 rebalancing driven by a subpopulation of RG cells and their processes, highlighting the
354 importance of studying transcriptional processes in RG projections, which RESCUE
355 enables.

356 To further characterize the structure of the unattributed transcriptome, we applied
357 NMF to the unattributed expression component (**Fig. 6d**). Several unattributed factors are
358 enriched for genes associated with synaptic and process-localized functions, including
359 *slc1a2*, *calm2*, and mitochondrial ATP synthase (**Fig. 6e**), consistent with transcriptional
360 activity in extended RG scaffolds and neuronal processes. In addition, one factor shows
361 strong enrichment for ribosomal protein genes, suggesting localized protein synthesis
362 within these structures[78]. Notably, these structured transcriptional programs emerge only
363 after the canonical expression is removed, indicating that the unattributed expression

364 recovered by RESCUE captures biologically meaningful signals associated with radial
365 glial and cytoplasmic processes that are not represented in snRNA-seq reference data.

DISCUSSION

RESCUE is the first method designed to acknowledge and recover the unattributed transcriptome, which encompasses gene expression patterns not captured by cell segmentation or deconvolution approaches in ST data analysis. These unattributed signals, though typically overlooked, can carry substantial functional and biological significance. While the field has developed numerous transcriptomic profiling methods, such as scRNA-seq for isolating gene expression at the single-cell level, Ribo-seq for ribosome-associated transcripts, and CLIP-seq for binding-protein interactions, each of these technologies inevitably captures only a fraction of the total transcriptome. RESCUE provides a systematic framework to quantify and analyze what is missing: the molecular signals outside the scope of reference datasets.

By formulating the problem as a robust sparse recovery framework, RESCUE decomposes observed ST data into a reference-aligned canonical component and a biologically meaningful unattributed component. RESCUE focuses on characterizing what is not explained by reference profiles and introduces a broader conceptual innovation: computational negative selection. Rather than focusing solely on what to include (as in positive selection), we define a reference component to subtract out and focus on what remains. In this work, we used scRNA-seq and snRNA-seq data as reference profiles and focused primarily on brain tissue, but the concept is general and can be applied to a wide range of other settings, for example, muscle tissue, where multinucleated cells make segmentation challenging and single-cell reference data difficult to interpret. The choice of reference therefore requires careful design and quality control.

389 One striking finding emerging from our brain tissue analyses is the substantial
390 amount of expression (26.5% of total expression) found in the unattributed transcriptome
391 of the honey bee brain MERFISH data. While localization of mRNAs to neurites is a well-
392 studied phenomenon[26–30], our results, along with the estimates for the adult mouse brain
393 from Ament and Poulpoulos[22], additionally suggest that these local expression patterns
394 may be different from those found in cell soma. RESCUE enables a reexamination of
395 existing datasets to more carefully study this apparently divergent expression. Despite
396 these promising biological insights revealed by RESCUE, reliance on computational
397 inference alone may have potential limitations, underscoring the value of experimental
398 validation when feasible. Orthogonal approaches, such as *in situ* hybridization, could
399 further strengthen and refine these findings; however, such validation is often impractical
400 in tissues where cell bodies and non-cell-body structures are spatially intermingled and
401 difficult to segment. These challenges highlight the importance of discovery-oriented
402 computational frameworks like RESCUE, which are designed to uncover biologically
403 meaningful patterns and guide interpretation in settings where experimental confirmation
404 is difficult.

405 There are several potential directions for future work. First, in cases where no
406 suitable reference is available, extensions of RESCUE to reference-free or unsupervised
407 frameworks may be possible. For example, methods based on robust principal
408 component analysis (PCA)[79,80] could be adapted to extract unattributed structure without
409 predefined reference matrices. Second, applying RESCUE to high-resolution ST data
410 requires creating a predefined grid, and further research is required to develop a
411 systematic method to determine the optimal grid size. Third, while the gene-level sparsity

assumption underlying RESCUE is biologically plausible in many settings, there may exist scenarios in which the unattributed component is not sparse but instead exhibits alternative structure, such as low-rank or low-rank plus sparse residuals. Extending RESCUE to accommodate such residual regimes represents an important direction for future research. Despite these remaining challenges, we have already shown that RESCUE can still provide meaningful analyses of unattributed expression.

Finally, RESCUE reminds us that the biology we wish to study and the measurements we are able to make are not always one and the same. For instance, we typically accept single-cell reference data as truly representing the expression of the entire single cell, but in many cases it does not, as we have illustrated above. Any measurement technology has inherent biases. As a result, recognizing, acknowledging and properly interpreting what is missed, in other words, the “unattributed expression”, is important for a comprehensive understanding of a system.

METHODS

RESCUE model

We assume that our ST data profiles G genes in J grids across the tissue section. In low-resolution ST technologies, these grids are predefined as multicellular spots, whereas in higher-resolution ST technologies, these can be constructed at any desired size, which allows inference at different biological scales. To introduce the statistical model behind RESCUE, we begin by reviewing existing linear deconvolution methods[12–16]. These methods decompose the ST expression matrix $Y \in \mathbb{R}_{\geq 0}^{G \times J}$ of G genes in J grids as

$$Y = X\alpha + E, \tag{1}$$

where $X \in \mathbb{R}_{\geq 0}^{G \times K}$ is a matrix of K cell-type expression reference profiles, $\alpha \in \mathbb{R}_{\geq 0}^{K \times J}$ is a matrix of inferred cell-type proportions within each grid, and $E \in \mathbb{R}^{G \times J}$ is assumed to be a mean-zero noise matrix. The issue is that (1) implies that expectation of ST expression is $\mathbb{E}(Y) = X\alpha$, which ignores unattributed transcriptome.

Instead, we further decompose the error E, motivated by the robust regression framework[45–47]. Specifically, we consider the multivariate mean-shift model:

$$Y = X\alpha + R + \mathcal{E}, \tag{2}$$

where $X\alpha$ represents the expected expression values of each gene in each grid due to the presence of cells whose types are represented in the reference sc/snRNA-seq data, $R \in \mathbb{R}_{\geq 0}^{G \times J}$ is an unattributed term that allows for outliers that cannot be captured by the reference data, and $\mathcal{E} \in \mathbb{R}^{G \times J}$ is a noise matrix that allows for residual variability between the ST sample and the reference data.

Following the mean-shift model in the robust regression framework, we assume that the unattributed component R is sparse. Specifically, the sparsity assumption reflects

the expectation that only a subset of genes exhibits unattributed expression within each grid. This assumption is required for identifiability of the decomposition and is biologically plausible, as we expect relatively few genes to exhibit unattributed expression within each grid. Consistent with this assumption, we observe empirically that unattributed expression in neuropil regions of the honey bee brain is highly sparse, with 92% of genes showing no unattributed expression on average. Moreover, even in extracellular environments, ECM proteins typically involve highly sparse, restricted gene sets, with only 1.3–1.5% of genes showing nonzero expression, rather than transcriptome-wide activation[81].

Optimization problem

RESCUE aims to simultaneously recover the sparse non-negative matrix R and non-negative cell-type proportions α in our model (2). Specifically, we consider the following optimization problem:

$$\min_{\alpha, R \geq 0} \|W \odot R\|_1 \ \text{ subject to } \ \|Y - R - X\alpha\|_F^2 \leq \lambda,$$

where W is a weight matrix, $\odot$ denotes the Hadamard product, $\|\cdot\|_F$ denotes the Frobenius norm, and $\lambda > 0$ is a tuning parameter. We set each entry of the weight matrix as $W_{gj} = 1/\sqrt{Var(R_{gj})}$ to account for the fact that different genes have different variances across different locations. These weights will be estimated through an iterative variance estimation procedure. The ℓ_1 objective function enforces sparsity of our estimated R, while the constraint ensures that the residual variability $\mathcal{E}$ is not too large, in other words, that the ST data do not deviate too much from the reference sc/snRNA-seq data.

In practice, we solve the Lagrangian form of this problem:

$$\min_{\alpha, R \geq 0} f(\alpha, R) = \|W \odot R\|_1 + \frac{\mu}{2}\|Y - R - X\alpha\|_F^2, \tag{3}$$

where $\mu > 0$ is a tuning parameter. We apply an efficient alternating minimization to solve two convex sub-problems in each iteration:

$$\hat{\alpha} = \underset{\alpha \geq 0}{\operatorname{argmin}} \|Y - R - X\alpha\|_F^2 , \tag{4}$$

$$\hat{R} = \underset{R \geq 0}{\operatorname{argmin}} \|W \odot R\|_1 + \frac{\mu}{2} \|Y - R - X\alpha\|_F^2 = S_+\left(Y - X\alpha, \frac{W}{\mu}\right), \tag{5}$$

where $S_+(x, \tau) = \max(x - \tau, 0)$ denotes the entry-wise non-negative soft-thresholding function.

Estimation of adaptive weights

Kong et al.[47] proposed adaptive lasso-type weights to reduce masking and swamping effects in the presence of multiple outliers. Our weighting scheme plays a similar role, but differs in that the weights are learned iteratively through variance estimation of the unattributed component. In ST deconvolution models, the observed count $\tilde{Y}_{gj}$ is often modeled under a Poisson assumption[13,16]:

$$\tilde{Y}_{gj} | \lambda_{gj} \sim \text{Poisson}\left(N_j \lambda_{gj}\right)$$

where $N_j = \sum_g \tilde{Y}_{gj}$ represents the total transcript count, or number of UMIs for grid j, and λ_{gj} denotes the expected expression rate of gene g at grid j. Rather than directly modeling $\tilde{Y}_{gj}$, we assume that it can be decomposed as $\tilde{Y}_{gj} = \tilde{L}_{gj} + \tilde{R}_{gj}$, where

$$\tilde{L}_{gj} \big| \lambda_{gj}^L \sim \text{Poisson}\left(N_j^L \lambda_{gj}^L\right) \quad \text{and} \quad \tilde{R}_{gj} \big| \lambda_{gj}^R \sim \text{Poisson}\left(N_j^R \lambda_{gj}^R\right).$$

Here, $N_j^L = \sum_g \tilde{L}_{gj}$ and $N_j^R = \sum_g \tilde{R}_{gj}$ represent the total transcript counts of the canonical and unattributed expression for grid j, respectively, and $(\lambda_{gj}^L, \lambda_{gj}^R)$ denote the expected expression rates of these components. We can write $N_j^L = (1 - \pi_j) N_j$ and $N_j^R = \pi_j N_j$, where $\pi_j \in [0, 1]$ denotes the proportion of unattributed expression for grid j.

When using counts per million (CPM) normalization, i.e., $Y_{gj} = \tilde{Y}_{gj}/N_j \times 10^6$, we can estimate the variance of unattributed component as

$$\mathrm{Var}(R_{gj}) = \mathrm{Var}\left(\frac{\tilde{R}_{gj}}{N_j} \times 10^6\right) = \left(\frac{10^6}{N_j}\right)^2 \times \mathrm{Var}(\tilde{R}_{gj}) \approx \frac{10^6}{N_j} \times \hat{R}_{gj}, \tag{6}$$

where the approximation follows from $\widehat{\mathrm{Var}}(\tilde{R}_{gj}) = \widehat{N_j^R \lambda_{gj}^R}$ and $\tilde{R}_{gj} = R_{gj}/10^6 \times N_j$. Here, the term $\widehat{N_j^R \lambda_{gj}^R}$ can be iteratively updated using the plug-in estimator $\tilde{R}_{gj} \approx N_j^R \lambda_{gj}^R$. To avoid division by zero, we set $W_{gj} = 1/\max\left(\sqrt{\frac{10^6}{N_j} \times R_{gj}}, \delta\right)$, where δ is a small positive constant. We outline the iterative algorithm for RESCUE below (**Algorithm 1**).

Algorithm 1 Iterative algorithm for RESCUE

Input: Normalized expression $Y \in \mathbb{R}_{\geq 0}^{G \times J}$, observed UMI counts $N \in \mathbb{N}^J$, reference profile $X \in \mathbb{R}_{\geq 0}^{G \times K}$, $\mu > 0$, $\delta > 0$, and $\epsilon > 0$.

Output: Cell-type proportion $\alpha \in \mathbb{R}_{\geq 0}^{K \times J}$ and unattributed expression $R \in \mathbb{R}_{\geq 0}^{G \times J}$.

Initialize $R^{(0)} = 0$, $W_{gj}^{(0)} = 1$, and $t = 1$.
repeat
  $\alpha^{(t)} = \mathrm{NNLS}(X, Y - R^{(t-1)})$ from (4).
  $R^{(t)} = S_+(Y - X\alpha^{(t)}, W/\mu)$ from (5).
  for $g = 1$ to G and $j = 1$ to J **do**
    $W_{gj}^{(t)} = 1/\max\left(\sqrt{\frac{10^6}{N_j} \times R_{gj}^{(t)}}, \delta\right)$ from (6).
  end for
  $t = t + 1$.
until $|f(\alpha^{(t)}, R^{(t)}) - f(\alpha^{(t-1)}, R^{(t-1)})| / f(\alpha^{(t)}, R^{(t)}) < \epsilon$, where f is defined in (3).
Return: $\alpha^{(t)}$ and $R^{(t)}$.

Tuning parameter selection

RESCUE aims to recover unattributed expression by balancing the sparsity of R, which reflects the number of rejected outliers, against the level of residual variability $\mathcal{E}$. Following standard practices in robust regression and robust PCA[45,79], we set $\mu = GJ/C\|Y\|_1$ and

focus on selecting the optimal value of the tuning parameter C. However, choosing the optimal tuning parameter is challenging in the presence of outliers[80,82].

To select the optimal tuning parameter, we adopt a data-driven strategy recently proposed for related deconvolution problems in bulk RNA-seq anaysis[83]. The key idea is to generate simulated data that closely resemble the observed ST data and use this for evaluation. Specifically, we use a surrogate ST dataset with "ground truth" proportions of canonical and unattributed expression. To construct this surrogate ST dataset, we first apply NNLS to the observed ST data to obtain initial estimates for α and R:

$$\hat{R}_{NNLS} = (Y - X\hat{\alpha}_{NNLS})_+, \quad \text{where} \quad \hat{\alpha}_{NNLS} = \underset{\alpha \geq 0}{\operatorname{argmin}} \|Y - X\alpha\|_F^2, \tag{7}$$

where $(x)_+ = \max(x, 0)$ denotes the zero-thresholding function. We then simulate a surrogate ST data $\tilde{Y}_{\text{sim}}$ from a Poisson distribution with rescaled expected expression rates, preserving the overall signal-to-noise characteristics observed in the real data. RESCUE is applied to $\tilde{Y}_{\text{sim}}$ across a range of μ values, and the optimal C is selected as the value that minimizes the recovery error of the canonical and unattributed components in the surrogate dataset. We summarize the tuning procedure below (**Algorithm 2**).

Algorithm 2 Tuning parameter selection via surrogate match

Input: Normalized expression $Y \in \mathbb{R}_{\geq 0}^{G \times J}$, observed UMI counts $N \in \mathbb{N}^J$, reference profile $X \in \mathbb{R}_{\geq 0}^{G \times K}$, and $S > 0$.

Output: Optimal tuning parameter μ.

1: **for** $s = 1$ to S **do**
2: Obtain $\hat{L}_{\text{NNLS}} = X\hat{\alpha}_{\text{NNLS}}$ and $\hat{R}_{\text{NNLS}}$ from (7).
3: Simulate ST data surrogate $\tilde{Y}_{\text{sim}} \sim \text{Pois}(N_j \lambda_{gj})$ with $\lambda_{gj} = (\hat{L}_{\text{NNLS}} + \hat{R}_{\text{NNLS}})/10^6$.
4: Apply Algorithm 1 with normalized Y_{sim} to obtain $\hat{L} = X\hat{\alpha}$ and $\hat{R}$.
5: Evaluate $\eta_C^{(s)} = \max\left(\|\hat{L}_{\text{NNLS}} - \hat{L}\|_F / \|\hat{L}_{\text{NNLS}}\|_F, \|\hat{R}_{\text{NNLS}} - \hat{R}\|_F / \|\hat{R}_{\text{NNLS}}\|_F\right)$.
6: **end for**

Return: $\mu = GJ/C\|Y\|_1$, where $C = \text{argmin}_C\left(\frac{1}{S}\sum_{s=1}^{S} \eta_C^{(s)}\right)$.

---

523 **Platform effect normalization**

524 To mitigate cross-platform effects between ST data and sc/snRNA-seq reference data,
525 we apply platform effect normalization. This step is essential because differences in
526 experimental platforms can introduce systematic shifts in gene expression levels that
527 confound downstream analysis. Without proper alignment of scale, these differences may
528 lead to biased estimation of cell-type contributions or unattributed signals. We adopt the
529 normalization strategy from RCTD[13], which rescales each gene in the reference data to
530 match the overall expression level of the corresponding gene in the ST data. Specifically,
531 it first computes pseudo-bulk measurements for both ST and reference datasets and then
532 apply a gene-wise log-scale correction to harmonize them.

533

534 **Data normalization and reference matrix**

535 To account for technical variability and the effect of library size, we adopt CPM
536 normalization but fit RESCUE in the linear space. We avoid log-transformation for two
537 main reasons. First, log-transformed input tends to underestimate true signals in gene
538 expression deconvolution[84]. Second, log-transformation requires adding pseudo-counts,
539 which can introduce significant artificial distortion between zeros and non-zeros[85,86].
540 Instead, we work in the linear scale, employing an iteratively reweighted algorithm
541 (**Algorithm 1**) to stabilize variance adaptively. This approach allows the model to adjust
542 for varying data scales and minimizes the influence of outliers without artificially altering
543 data distributions. The normalized reference matrix is then used to construct average cell-
544 type expression.

545

546 **Reference dependence of unattributed expression**

547 RESCUE is a computational negative selection method which relies on the choice of
548 reference, and interpretation of the recovered unattributed expression therefore requires
549 consideration of both biological context and reference completeness. To further assess
550 the impact of reference selection within the same tissue, we extended the honey bee
551 brain analysis using snRNA-seq[87] and segmented MERFISH cells as alternative
552 reference profiles (**Supplementary Fig. 5**). Using segmented MERFISH cells produced
553 unattributed expression patterns highly consistent with those obtained using scRNA-seq,
554 as both references capture cellular transcripts beyond the nucleus and therefore explain
555 similar fractions of the ST signal. In contrast, the snRNA-seq reference yielded higher
556 levels of unattributed expression on average, reflecting its enrichment for nuclear RNA
557 and underrepresentation of cytoplasmic and neuritic transcripts. These results indicate
558 that variation in the recovered unattributed component reflects biologically meaningful
559 differences in reference coverage rather than technical artifacts, and that interpretation of
560 unattributed expression should be guided by the biological properties of both the tissue
561 and the reference dataset.

562

563 **Choice of grid resolution for honey bee brain MERFISH data**

564 For high-resolution ST platforms such as MERFISH, RESCUE operates on spatially
565 aggregated expression profiles defined over user-specified grids. The choice of grid size
566 reflects a trade-off between spatial resolution and sparsity: smaller grids preserve fine
567 spatial structure but increase sparsity, whereas larger grids reduce sparsity at the cost of
568 spatial detail. As a practical guideline, we recommend selecting a grid size such that the

average total transcript count per grid is comparable to the typical total transcript count per cell. In the honey bee MERFISH analysis, we used a grid size of $20 \times 20\ \mu m^2$, which approximates cell-scale resolution. To assess sensitivity to grid resolution, we varied grid size across a range of spatial scales and found that RESCUE produces consistent results across a broad range of grid resolutions (**Supplementary Fig. 6**).

Subtraction-based estimation of unattributed expression for existing methods

Most existing ST deconvolution methods are designed to estimate cell-type proportions, rather than to recover expression levels directly. Because these methods do not explicitly model unattributed expression, we adopted a subtraction-based framework, following a strategy used in related deconvolution settings for bulk RNA-seq data[23]. Specifically, we first estimated cell-type proportions using each deconvolution method and then recovered the unattributed expression by subtracting the estimated cellular contributions from the observed expression, followed by zero-thresholding to avoid negative values. Under this framework, the unattributed expression is defined as the portion of the observed ST expression that cannot be explained by the reference-based cellular expression implied by each method's estimated cell-type proportions.

Data generating process for simulation studies

For a total of $J = 2{,}500$ grids, we first generated the total UMI count for grid j from a negative binomial distribution $N_j \sim \mathrm{NB}(\mathrm{mean} = M, \mathrm{size} = 1/\sigma)$. We generated the raw count for each gene in each grid as $\tilde{Y}_{gj} = \tilde{L}_{gj} + \tilde{R}_{gj}$, where $\tilde{R}_{gj}$ denotes the unattributed

591 expression and $\tilde{L}_{gj}$ represents the canonical expression that could have been captured
592 by sc/snRNA-seq reference data.

593 To generate $\tilde{L}_{gj}$, we used the real scRNA-seq data from honey bee brains[48] as the
594 source of canonical expression profiles. We selected the top $G = 500$ number of highly
595 variable genes and constructed the reference matrix X using $K = 20$ number of cell types
596 annotated via Louvain clustering in Seurat[88]. Latent mixing proportions were drawn as
597 $\beta_{kj} \sim \mathrm{Beta}(1,3)$ and normalized as $\alpha_{kj} = \beta_{kj} / \sum_k \beta_{kj}$, with at most 10 cell types per spot.
598 Defining $\lambda^L_{gj} = (X\alpha)_{gj}$, we generated the cell-type-driven canonical expression as

599 $$\tilde{L}_{gj} \sim \mathrm{Multinomial}\left((1-\pi_j)N_j, P_{gj} = \frac{\lambda^L_{gj}}{\sum_g \lambda^L_{gj}}\right).$$

600 To generate sparse unattributed expression, we sampled

601 $$\tilde{R}_{gj} \sim \mathrm{Multinomial}\left(\pi_j N_j, Q_{gj} = \frac{\lambda^R_{gj}}{\sum_g \lambda^R_{gj}}\right),$$

602 where $\lambda^R_{gj} \sim \mathrm{Gamma}(1, 1/3)$ with probability $(1-s)$ and $\lambda^R_{gj} = 0$ with probability s. Here,
603 we introduced a sparsity parameter s, which makes $(s \times 100)$ percentage of intensity
604 parameter λ^R_{gj} zero-inflated. Under this setting, we can generate $\tilde{L}_{gj}$ and $\tilde{R}_{gj}$ that satisfy
605 $\sum_g(\tilde{L}_{gj} + \tilde{R}_{gj}) = N_j$. In our simulations, we fixed $M = 10{,}000$ and $\sigma = 0.5$, while varying
606 $\pi_j \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and $s \in \{0.8, 0.85, 0.9, 0.95\}$ to control the proportion of canonical
607 expression and the sparsity level of unattributed transcriptome, respectively (**Fig. 2**). We
608 then examined the robustness of RESCUE under additional simulation settings by varying
609 the average total UMI count per spot (M), as well as varying noise levels by adjusting
610 library size variability across spots (σ) (**Supplementary Fig. 1**).

611

612 **Evaluation metrics for cell-type deconvolution in simulation studies**

613 To evaluate the accuracy of cell-type deconvolution, we considered three complementary
614 metrics: Lin's CCC, RMSE, and JSD. We first measured Lin's CCC between the true and
615 estimated proportion of the kth cell type, defined as

616 $$\mathrm{CCC}(\alpha_k, \hat{\alpha}_k) = \frac{2\mathrm{Cov}(\alpha_k, \hat{\alpha}_k)}{\sigma_{\alpha_k}^2 + \sigma_{\hat{\alpha}_k}^2 + \left(\mu_{\alpha_k} - \mu_{\hat{\alpha}_k}\right)^2},$$

617 where μ_{α_k} and $\mu_{\hat{\alpha}_k}$ are the means of the two variables, and $\sigma_{\alpha_k}^2$ and $\sigma_{\hat{\alpha}_k}^2$ denote the
618 corresponding variances. We used Lin's CCC because it captures shifts in both location
619 and scale, making it robust to high leverage points[83]. The concordance increases as the
620 CCC value gets closer to 1. We report the average CCC across all cell types.

621 To quantify estimation error at the spatial location level, we computed the RMSE
622 between the true and estimated cell-type proportion vectors. For spatial location j, RMSE
623 is defined as

624 $$\mathrm{RMSE}_j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\alpha_{kj} - \hat{\alpha}_{kj}\right)^2},$$

625 where K is the number of cell types, and α_{kj} and $\hat{\alpha}_{kj}$ denote the true and estimated
626 proportions of cell type k at location j, respectively. We report the average RMSE across
627 all spatial locations, with smaller values indicating better performance.

628 Finally, we computed the JSD between the true and estimated cell-type proportion
629 distributions at each spatial location. For spatial location j, JSD is defined as

630 $$\mathrm{JSD}_j = \frac{1}{2}\mathrm{KL}(\alpha_j \| m_j) + \frac{1}{2}\mathrm{KL}(\hat{\alpha}_j \| m_j), \qquad m_j = \frac{1}{2}\left(\alpha_j + \hat{\alpha}_j\right),$$

where $\mathrm{KL}(\cdot \,\|\, \cdot)$ denotes the Kullback-Leibler divergence. We report the average JSD across all spatial locations, with smaller values indicating better performance.

Evaluation metrics for unattributed expression recovery in simulation studies

To evaluate the accuracy of unattributed expression recovery, we measured the relative Frobenius error, defined as

$$(\text{Relative Frobenius Error}) = \frac{\|R - \hat{R}\|_F}{\|R\|_F},$$

where R and $\hat{R}$ denote the true and estimated unattributed expression matrices, respectively.

Evaluation metrics for honey bee brain MERFISH data

We measured Lin's CCC and the RMSE between the segmentation-based (true) and estimated unattributed expression proportions, defined as follows:

$$\mathrm{CCC}(\rho, \hat{\rho}) = \frac{2\mathrm{Cov}(\rho, \hat{\rho})}{\sigma_\rho^2 + \sigma_{\hat{\rho}}^2 + \left(\mu_\rho - \mu_{\hat{\rho}}\right)^2} \quad \text{and} \quad \mathrm{RMSE}(\rho, \hat{\rho}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\rho_j - \hat{\rho}_j\right)^2},$$

where $\rho_j = \sum_g \tilde{R}_{gj} \,/\, \sum_g \tilde{Y}_{gj}$ denotes the true unattributed expression proportion in grid j, and $\hat{\rho}_j$ denotes its estimated value. Here, μ_ρ and $\mu_{\hat{\rho}}$ are the means of two variables, and σ_ρ^2 and $\sigma_{\hat{\rho}}^2$ denote the corresponding variances.

Benchmarking computational runtime

We evaluated the computational runtime of each method across ST datasets of varying sizes (**Supplementary Fig. 7**). Across all benchmarks, RESCUE completed within

practical time limits for contemporary ST datasets, demonstrating scalability to datasets with hundreds to thousands of spatial locations and genes. All analyses were performed using CPU-only computation on a standard laptop (Apple M3 Pro with 36GB of memory), without GPUs or high-performance computing clusters.

DATA AVAILABILITY

Human breast cancer dataset[5]: 10x Visium and scFFPE-seq data are available via GEO accession GSE243280. Human DLPFC dataset[60]: 10x Visium and snRNA-seq data are available at https://github.com/LieberInstitute/spatialLIBD. Cichlid fish brain dataset[61]: 10x Visium and snRNA-seq data are available via GEO accession GSE217619. Honey bee brain MERFISH data is available upon reasonable request. Honey bee brain scRNA-seq data[48] is available at https://doi.org/10.6084/m9.figshare.16832518.

CODE AVAILABILITY

RESCUE is available as an R package at https://github.com/brunoyjlee/RESCUE.

ACKNOWLEDGEMENTS

We thank Dr. Tina Barbasch on valuable discussions on the analysis of fish brain data. Funding: This work was supported by the National Institutes of Health (R21HG013180 to Y.J.L., S.Y., H.-S.H. and S.D.Z., R35GM147420 to A.W.S and H.-S.H. and R01AT013189 to S.D.Z.).

FIGURES

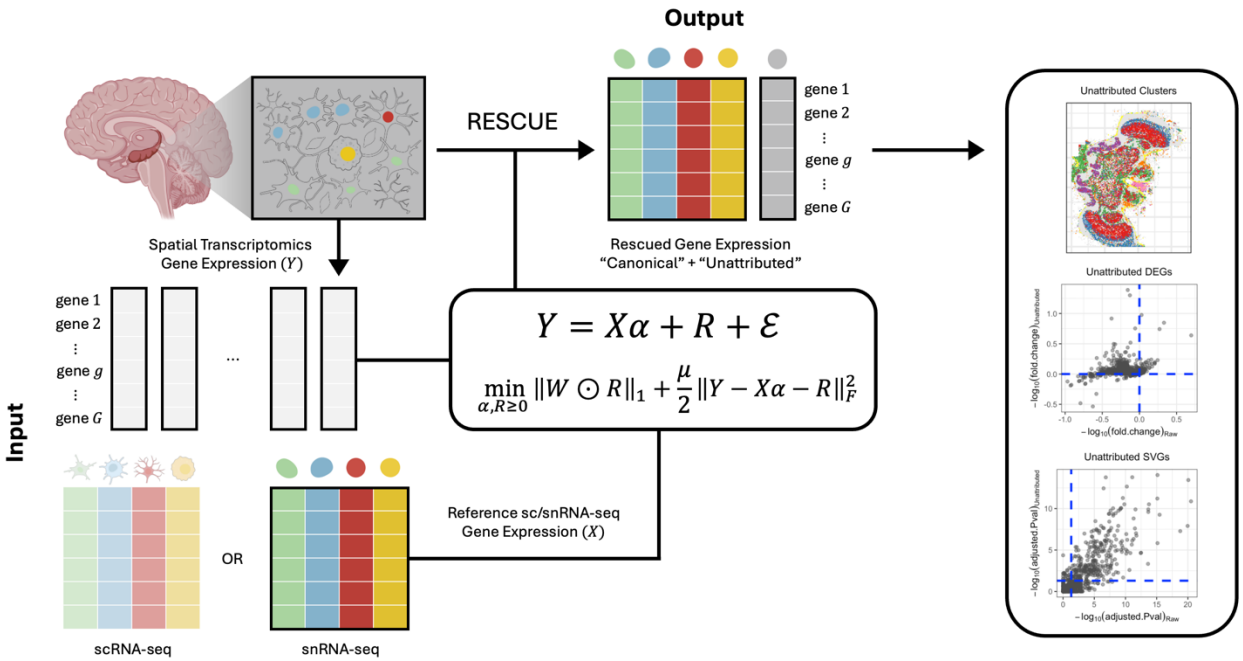


Figure 1. Schematic overview of RESCUE. RESCUE is a novel computational negative selection method designed to recover the spatial expression patterns missed by standard single-cell inference on ST data. It takes as input ST data to be decomposed, along with a reference matrix constructed from sc/snRNA-seq data from the same tissue, and performs robust sparse recovery via a regularized multivariate regression framework. RESCUE partitions observed gene expression into a reference-explainable "canonical" component and a residual "unattributed" component. As a result, the purified canonical expression enables more accurate deconvolution, while the recovered unattributed expression enables downstream analyses such as clustering, differentially expressed gene (DEG), and spatially variable gene (SVG) analyses. Created with BioRender.com.

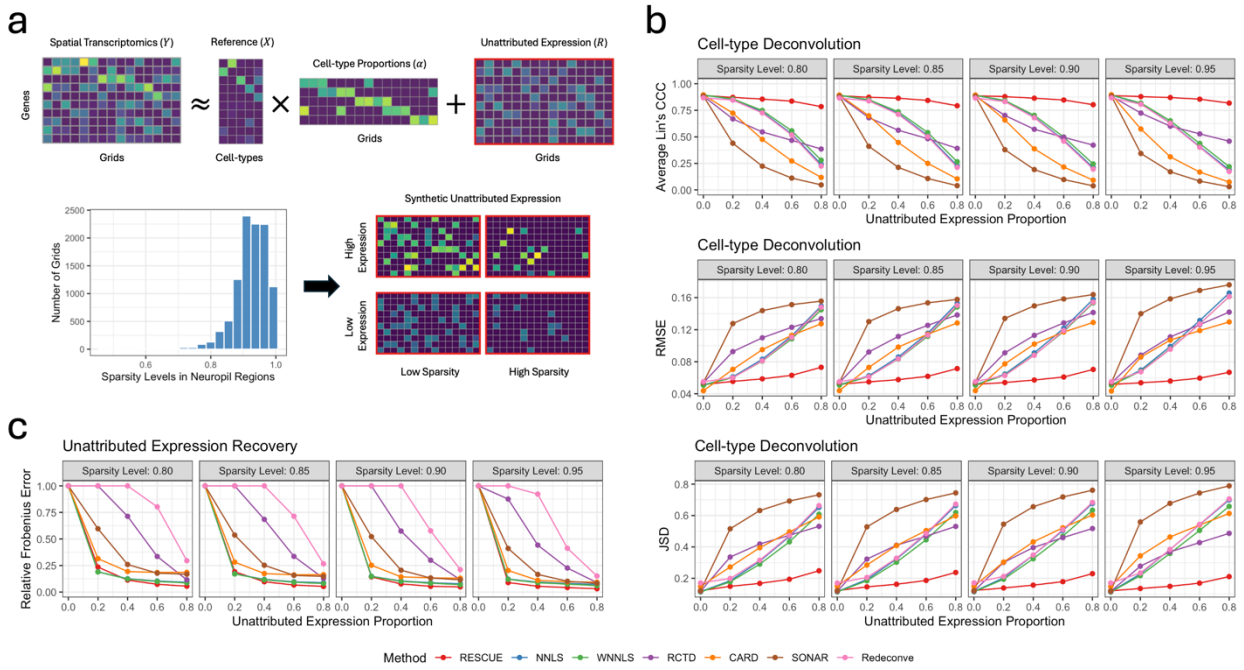
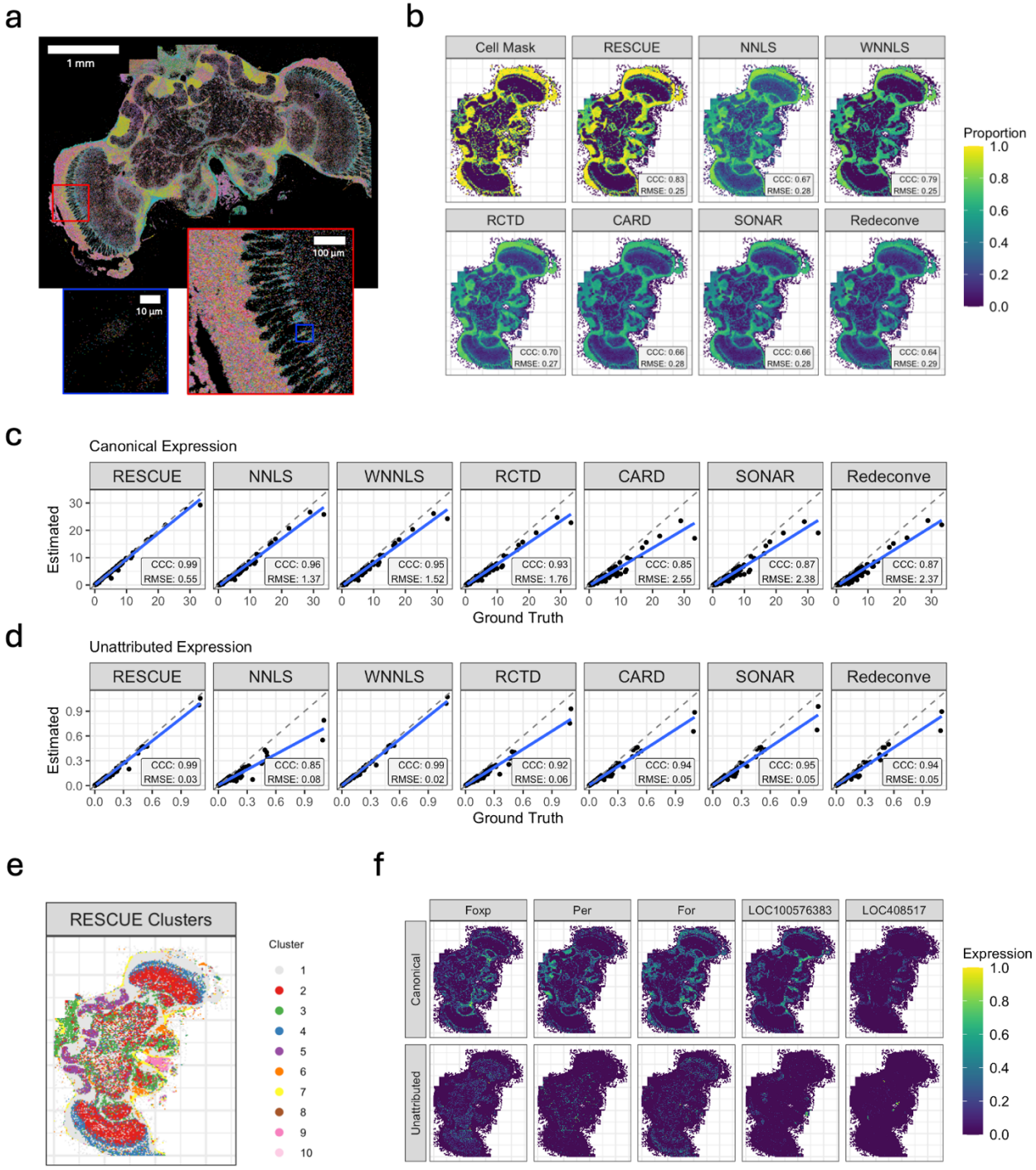


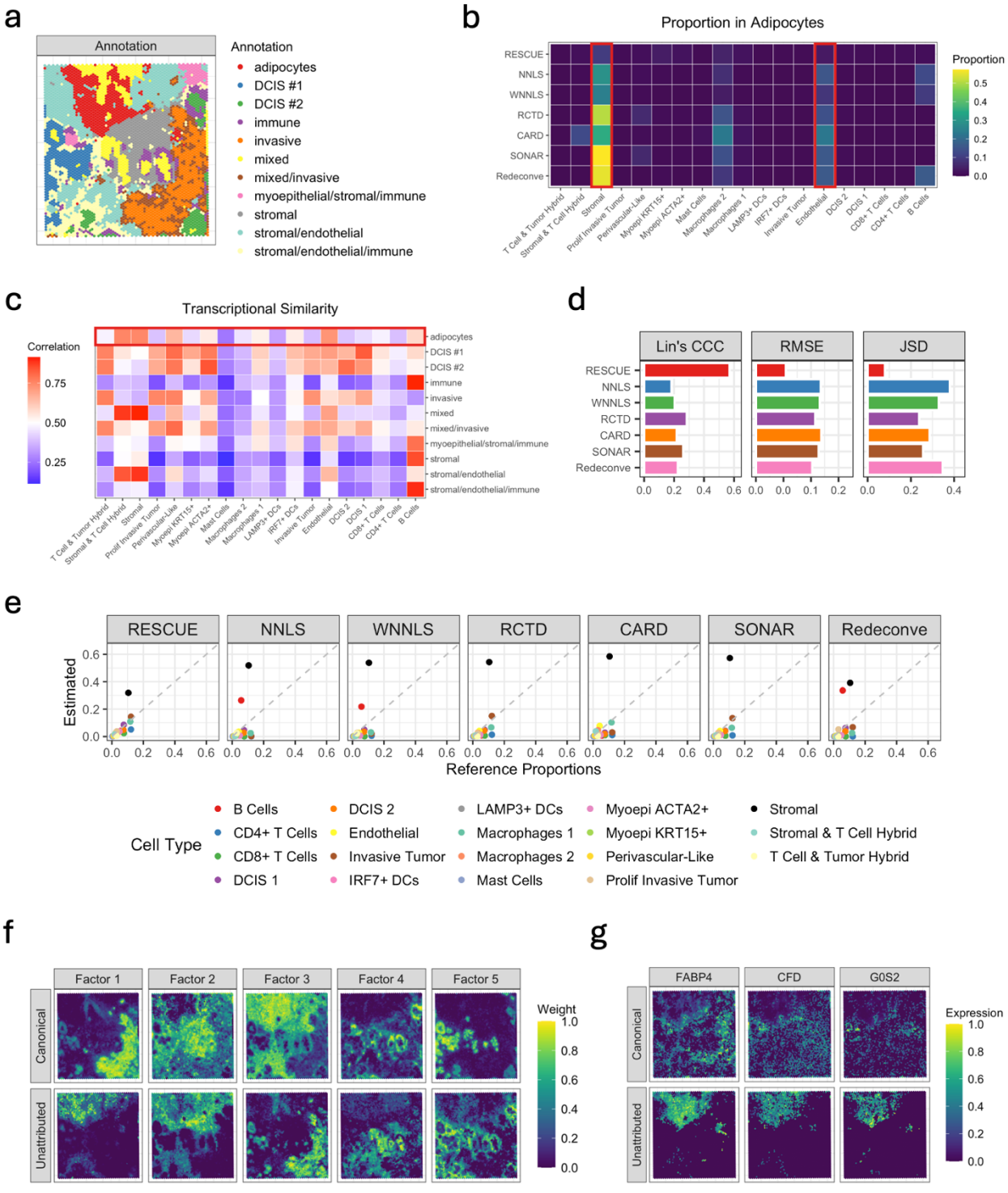
Figure 2. Simulation results. a. Schematic of the simulation framework. ST data are generated as the sum of reference-driven canonical expression and synthetic unattributed expression. The distribution of sparsity levels observed in neuropil regions of the honey bee brain motivates the simulated sparsity regimes. Representative examples of synthetic unattributed expression under low- and high-sparsity/expression levels are shown. **b.** Cell-type deconvolution performance across methods, evaluated using average Lin's concordance correlation coefficient (CCC; top), root mean squared error (RMSE; middle), and Jensen-Shannon divergence (JSD; bottom) as a function of the sparsity/proportion levels of the unattributed expression. **c.** Unattributed expression recovery performance measured by relative Frobenius error under the same simulation settings.


a
1 mm
100 µm
10 µm
b
Cell Mask
RESCUE
NNLS
WNNLS
RCTD
CARD
SONAR
Redeconve
CCC: 0.83
RMSE: 0.25
CCC: 0.67
RMSE: 0.28
CCC: 0.79
RMSE: 0.25
CCC: 0.70
RMSE: 0.27
CCC: 0.66
RMSE: 0.28
CCC: 0.66
RMSE: 0.28
CCC: 0.64
RMSE: 0.29
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
c
Canonical Expression
Estimated
Ground Truth
CCC: 0.99
RMSE: 0.55
CCC: 0.96
RMSE: 1.37
CCC: 0.95
RMSE: 1.52
CCC: 0.93
RMSE: 1.76
CCC: 0.85
RMSE: 2.55
CCC: 0.87
RMSE: 2.38
CCC: 0.87
RMSE: 2.37
d
Unattributed Expression
Estimated
Ground Truth
CCC: 0.99
RMSE: 0.03
CCC: 0.85
RMSE: 0.08
CCC: 0.99
RMSE: 0.02
CCC: 0.92
RMSE: 0.06
CCC: 0.94
RMSE: 0.05
CCC: 0.95
RMSE: 0.05
CCC: 0.94
RMSE: 0.05
e
RESCUE Clusters
Cluster
1
2
3
4
5
6
7
8
9
10
f
Foxp
Per
For
LOC100576383
LOC408517
Canonical
Unattributed
Expression
1.0
0.8
0.6
0.4
0.2
0.0

697

698

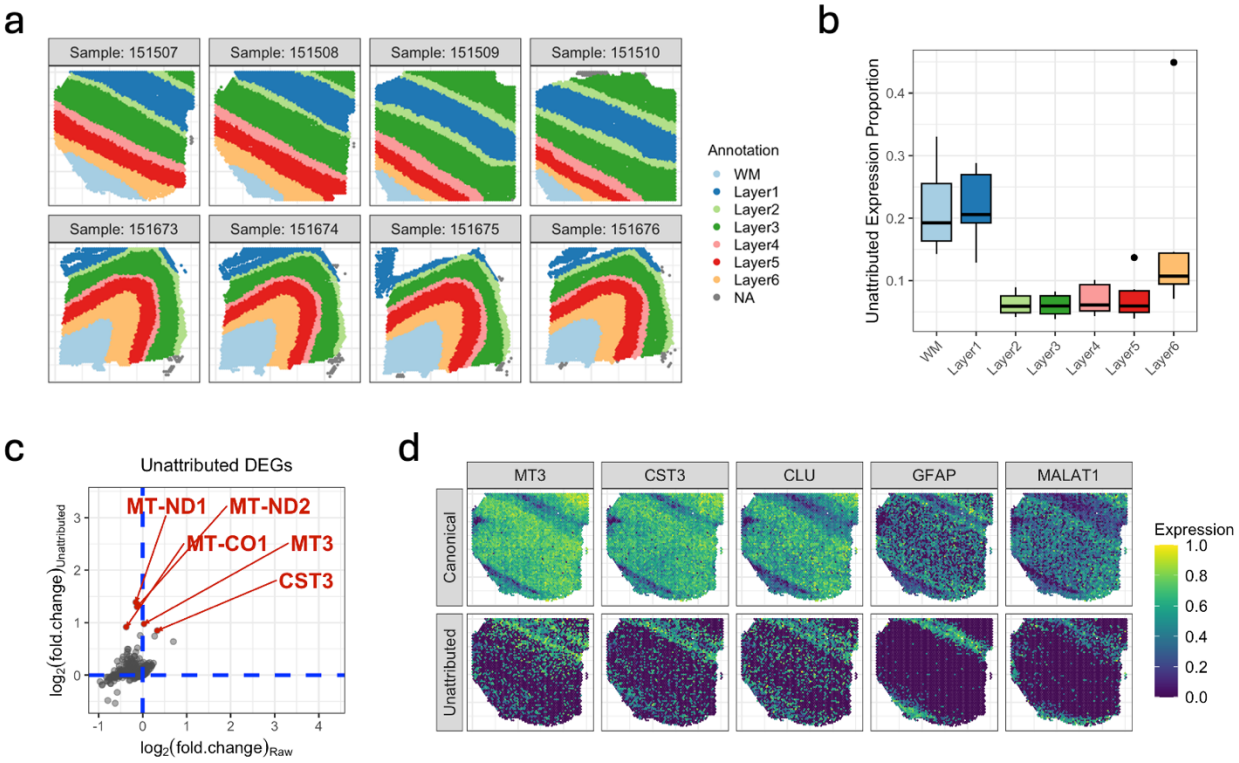
Figure 3. Application to the honey bee brain MERFISH data. a. MERFISH profiling of a coronal honey bee brain section. Each dot represents an individual transcript, with genes shown in distinct colors. A substantial fraction of transcripts localize to neurites, rather than to cell bodies. **b.** Comparison of segmentation-based ground-truth ("Cell Mask") and estimated somatic proportions. Cell mask ratio is computed as the proportion of MERFISH spots in each grid lying within a cell. Lin's concordance correlation coefficient (CCC) and root mean squared error (RMSE) between ground-truth and estimated proportions are reported. **c.** Comparison of estimated versus ground-truth canonical gene expression levels in grids corresponding to pure cell body regions. **d.** Comparison of estimated versus ground-truth unattributed gene expression levels in neuropil-only grids lacking segmented somata. For **c** and **d,** each point represents a gene averaged across grids; grey dashed lines indicate the identity line, and blue lines denote linear regression fits. Lin's CCC and RMSE between ground-truth and estimated average expression are reported. **e.** Clustering of the unattributed expression profiles inferred by RESCUE reveals distinct transcriptional niches within neuropil regions, while somatic regions collapse into a single cluster. **f.** Spatial expression patterns of representative genes, shown separately for canonical and unattributed components. Expression values are min-max normalized separately for each panel.

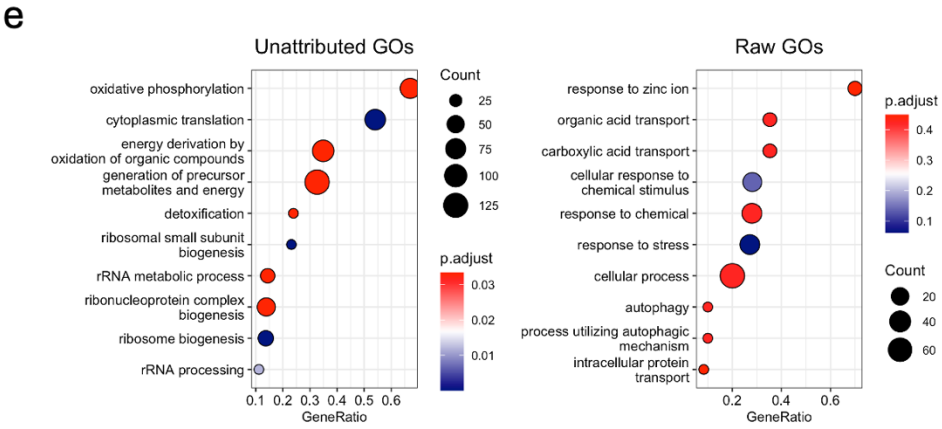

a
Annotation
Annotation
adipocytes
DCIS #1
DCIS #2
immune
invasive
mixed
mixed/invasive
myoepithelial/stromal/immune
stromal
stromal/endothelial
stromal/endothelial/immune
b
Proportion in Adipocytes
Proportion
RESCUE
NNLS
WNNLS
RCTD
CARD
SONAR
Redeconve
c
Transcriptional Similarity
Correlation
d
Lin's CCC
RMSE
JSD
e
RESCUE
NNLS
WNNLS
RCTD
CARD
SONAR
Redeconve
Estimated
Reference Proportions
Cell Type
B Cells
CD4+ T Cells
CD8+ T Cells
DCIS 1
DCIS 2
Endothelial
Invasive Tumor
IRF7+ DCs
LAMP3+ DCs
Macrophages 1
Macrophages 2
Mast Cells
Myoepi ACTA2+
Myoepi KRT15+
Perivascular-Like
Prolif Invasive Tumor
Stromal
Stromal & T Cell Hybrid
T Cell & Tumor Hybrid
f
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Canonical
Unattributed
Weight
g
FABP4
CFD
G0S2
Canonical
Unattributed
Expression

718

719

720 **Figure 4. Application to the human breast cancer Visium data. a.** Manual annotation
721 of Visium ST data from human breast cancer tissue. **b.** Estimated cell-type proportions in
722 adipocyte regions across methods. Red box highlights that existing deconvolution
723 methods misattribute adipocyte-derived signal to other reference cell types due to the
724 absence of adipocytes in the reference, whereas RESCUE mitigates such misassignment
725 by modeling unattributed component. **c.** Heatmap of correlations between pseud-bulk
726 Visium ST and scFFPE-seq reference cell types. Red box highlights high correlations
727 between adipocytes and other cell types including stromal and endothelial cells. **d.**
728 Comparison of cell-type deconvolution accuracy across methods using Lin's concordance
729 correlation coefficient (CCC), root mean squared error (RMSE), and Jensen-Shannon
730 divergence (JSD). **e.** Comparison of estimated versus reference cell-type proportions
731 across methods. Each point represents a cell type, with grey dashed lines indicating the
732 identity line. **f.** Non-negative matrix factorization (NMF) of the canonical and unattributed
733 expression component inferred by RESCUE. Weight values are min-max normalized
734 separately for each panel. **g.** Spatial expression patterns of adipocyte marker genes,
735 shown separately for canonical and unattributed components. Expression values are min-
736 max normalized separately for each panel.
737

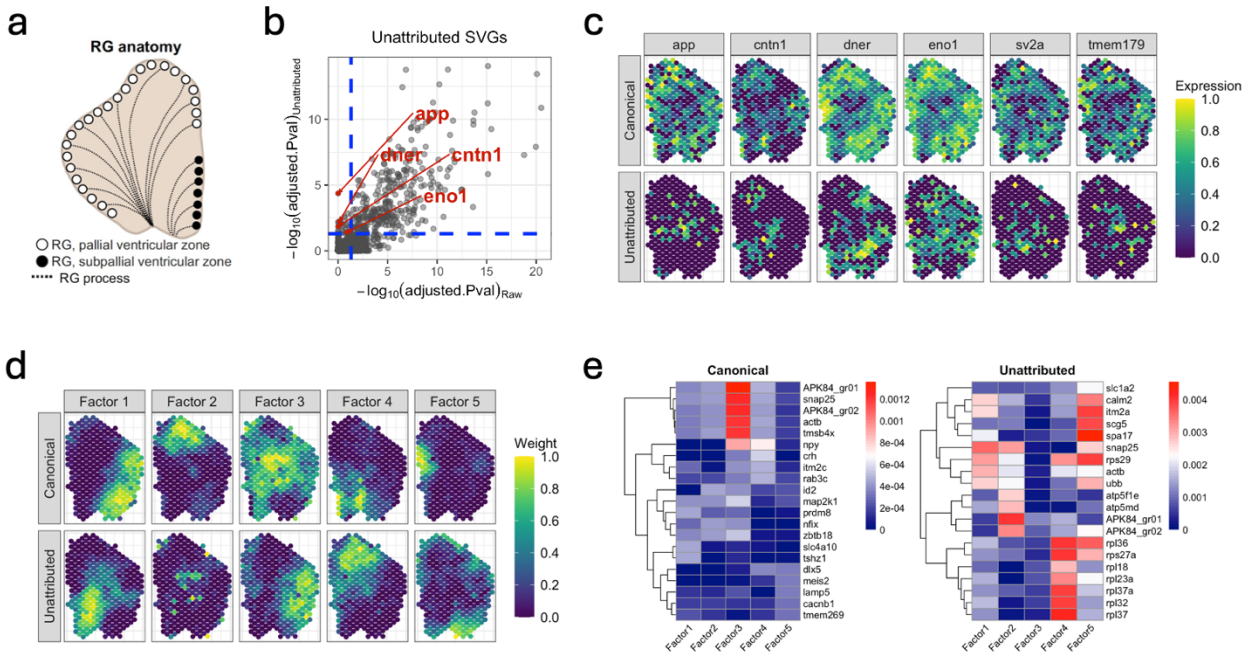

a
Sample: 151507
Sample: 151508
Sample: 151509
Sample: 151510
Sample: 151673
Sample: 151674
Sample: 151675
Sample: 151676
Annotation
WM
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
NA
b
Unattributed Expression Proportion
0.4
0.3
0.2
0.1
WM
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
c
Unattributed DEGs
MT-ND1
MT-ND2
MT-CO1
MT3
CST3
log2(fold.change)Unattributed
log2(fold.change)Raw
d
MT3
CST3
CLU
GFAP
MALAT1
Canonical
Unattributed
Expression
1.0
0.8
0.6
0.4
0.2
0.0


e
Unattributed GOs
oxidative phosphorylation
cytoplasmic translation
energy derivation by oxidation of organic compounds
generation of precursor metabolites and energy
detoxification
ribosomal small subunit biogenesis
rRNA metabolic process
ribonucleoprotein complex biogenesis
ribosome biogenesis
rRNA processing
GeneRatio
Count
25
50
75
100
125
p.adjust
0.03
0.02
0.01
Raw GOs
response to zinc ion
organic acid transport
carboxylic acid transport
cellular response to chemical stimulus
response to chemical
response to stress
cellular process
autophagy
process utilizing autophagic mechanism
intracellular protein transport
GeneRatio
p.adjust
0.4
0.3
0.2
0.1
Count
20
40
60

738

739

Figure 5. Application to the human DLPFC Visium data. a. Anatomical annotations of Visium ST samples from the DLPFC dataset, with cortical layers (Layer 1-6) and white matter (WM) labeled. **b.** Proportions of unattributed expression estimated by RESCUE across cortical layers and WM. Layer 1 and WM exhibit substantially higher unattributed expression compared to other layers. **c.** Comparison of log-fold changes for differentially expressed gene (DEG) analysis comparing Layer 1 to all other layers using limma on raw observed ST expression (x-axis) versus unattributed expression recovered by RESCUE (y-axis). Blue dashed lines indicate zero-fold change. Genes highlighted in red are top Layer 1 markers, detected from unattributed expression. **d.** Spatial expression patterns of Layer 1 marker genes derived from unattributed expression, shown separately for canonical and unattributed components. Expression values are min-max normalized separately for each panel. **e.** Gene ontology (GO) terms enriched among Layer 1-enriched DEGs based on unattributed (left) and raw (right) expression.



753

754 **Figure 6. Application to the cichlid fish telencephalon Visium data. a.** Schematic
755 illustration of radial glial (RG) anatomy[61] in the dorsal telencephalon, highlighting RG cell
756 bodies in the ventricular zone and long processes extending toward superficial pallial
757 regions. **b.** Comparison of adjusted p-values for spatially variable gene (SVG) detection
758 using SPARK-X on raw observed ST expression (x-axis) versus unattributed expression
759 recovered by RESCUE (y-axis). Blue dashed lines indicate the significance threshold at
760 adjusted p-value = 0.05. Genes labeled in red are SVGs uniquely detected from the
761 unattributed expression. Adjusted p-values are obtained using SPARK-X's default option:
762 Benjamini-Yekutieli procedure. **c.** Spatial expression patterns of unattributed-only SVGs,
763 shown separately for canonical and unattributed components. Expression values are min-
764 max normalized separately for each panel. **d.** Non-negative matrix factorization (NMF) of
765 the canonical and unattributed expression component inferred by RESCUE. Weight
766 values are min-max normalized separately for each panel. **e.** Heatmaps of genes with
767 high loadings in canonical (left) and unattributed (right) factors.

REFERENCES

1. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
2. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat Methods* **19**, 534–546 (2022).
3. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* https://doi.org/10.1126/science.aaa6090 (2015) doi:10.1126/science.aaa6090.
4. Nagendran, M. *et al.* 1457 Visium HD enables spatially resolved, single-cell scale resolution mapping of FFPE human breast cancer tissue. *J Immunother Cancer* **11**, (2023).
5. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat Commun* **14**, 8353 (2023).
6. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**, 100–106 (2021).
7. Petukhov, V. *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nat Biotechnol* **40**, 345–354 (2022).
8. Polański, K. *et al.* Bin2cell reconstructs cells from high resolution Visium HD data. *Bioinformatics* **40**, btae546 (2024).
9. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

10. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
11. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* **39**, 313–319 (2021).
12. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* **49**, e50 (2021).
13. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517–526 (2022).
14. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol* **40**, 1349–1359 (2022).
15. Zhou, Z., Zhong, Y., Zhang, Z. & Ren, X. Spatial transcriptomics deconvolution at single-cell resolution using Redeconve. *Nat Commun* **14**, 7930 (2023).
16. Liu, Z., Wu, D., Zhai, W. & Ma, L. SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. *Nat Commun* **14**, 4727 (2023).
17. Si, Y. *et al.* FICTURE: scalable segmentation-free analysis of submicron-resolution spatial transcriptomics. *Nat Methods* **21**, 1843–1854 (2024).
18. Mitchel, J., Gao, T., Cole, E., Petukhov, V. & Kharchenko, P. V. Impact of Segmentation Errors in Analysis of Spatial Transcriptomics Data. 2025.01.02.631135 Preprint at https://doi.org/10.1101/2025.01.02.631135 (2025).

19. Nguyen, H., Nguyen, H., Tran, D., Draghici, S. & Nguyen, T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Res* **52**, 4761–4783 (2024).
20. Garmire, L. X. *et al.* Challenges and perspectives in computational deconvolution of genomics data. *Nat Methods* **21**, 391–400 (2024).
21. Gaspard-Boulinc, L. C., Gortana, L., Walter, T., Barillot, E. & Cavalli, F. M. G. Cell-type deconvolution methods for spatial transcriptomics. *Nat Rev Genet* 1–19 (2025) doi:10.1038/s41576-025-00845-y.
22. Ament, S. A. & Poulopoulos, A. The brain's dark transcriptome: Sequencing RNA in distal compartments of neurons and glia. *Current Opinion in Neurobiology* **81**, 102725 (2023).
23. Ivich, A. *et al.* Missing cell types in single-cell references impact deconvolution of bulk data but are detectable. *Genome Biology* **26**, 86 (2025).
24. Han, J. *et al.* Single-cell sequencing unveils key contributions of immune cell populations in cancer-associated adipose wasting. *Cell Discov* **8**, 122 (2022).
25. Fetit, R. *et al.* Characterizing Neutrophil Subtypes in Cancer Using scRNA Sequencing Demonstrates the Importance of IL1β/CXCR2 Axis in Generation of Metastasis-specific Neutrophils. *Cancer Research Communications* **4**, 588–606 (2024).
26. Sutton, M. A. *et al.* Miniature Neurotransmission Stabilizes Synaptic Function via Tonic Suppression of Local Dendritic Protein Synthesis. *Cell* **125**, 785–799 (2006).
27. Cajigas, I. J. *et al.* The Local Transcriptome in the Synaptic Neuropil Revealed by Deep Sequencing and High-Resolution Imaging. *Neuron* **74**, 453–466 (2012).

832 28. Hafner, A.-S., Donlin-Asp, P. G., Leitch, B., Herzog, E. & Schuman, E. M. Local protein
833 synthesis is a ubiquitous feature of neuronal pre- and postsynaptic compartments.
834 *Science* **364**, eaau3644 (2019).
835 29. Glock, C. *et al.* The translatome of neuronal cell bodies, dendrites, and axons.
836 *Proceedings of the National Academy of Sciences* **118**, e2113929118 (2021).
837 30. von Kügelgen, N. & Chekulaeva, M. Conservation of a core neurite transcriptome
838 across neuronal types and species. *WIREs RNA* **11**, e1590 (2020).
839 31. Lee, K.-W., Kim, H.-J., Sung, J.-J., Park, K.-S. & Kim, M. Defective neurite outgrowth in
840 aphidicolin/cAMP-induced motor neurons expressing mutant Cu/Zn superoxide
841 dismutase. *International Journal of Developmental Neuroscience* **20**, 521–526 (2002).
842 32. Markmiller, S. *et al.* Persistent mRNA localization defects and cell death in ALS neurons
843 caused by transient cellular stress. *Cell Reports* **36**, 109685 (2021).
844 33. Buxbaum, A. R., Haimovich, G. & Singer, R. H. In the right place at the right time:
845 visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* **16**, 95–109
846 (2015).
847 34. Gandalovičová, A., Vomastek, T., Rosel, D. & Brábek, J. Cell polarity signaling in the
848 plasticity of cancer cell invasiveness. *Oncotarget* **7**, 25022–25049 (2016).
849 35. Backlund, M. *et al.* Plasticity of nuclear and cytoplasmic stress responses of RNA-
850 binding proteins. *Nucleic Acids Res* **48**, 4725–4740 (2020).
851 36. Apicco, D. J. *et al.* Reducing the RNA binding protein TIA1 protects against tau-
852 mediated neurodegeneration in vivo. *Nat Neurosci* **21**, 72–80 (2018).

853 37. Wolozin, B. & Ivanov, P. Stress granules and neurodegeneration. *Nat Rev Neurosci* **20**,
854 649–666 (2019).
855 38. Ridder, K. *et al.* Extracellular vesicle-mediated transfer of functional RNA in the tumor
856 microenvironment. *OncoImmunology* **4**, e1008371 (2015).
857 39. Li, I. & Nabet, B. Y. Exosomes in the tumor microenvironment as mediators of cancer
858 therapy resistance. *Molecular Cancer* **18**, 32 (2019).
859 40. O'Brien, K., Breyne, K., Ughetto, S., Laurent, L. C. & Breakefield, X. O. RNA delivery by
860 extracellular vesicles in mammalian cells and its applications. *Nat Rev Mol Cell Biol*
861 **21**, 585–606 (2020).
862 41. Landry, F. J. & Findlay, S. R. Purification of human basophils by negative selection.
863 *Journal of Immunological Methods* **63**, 329–336 (1983).
864 42. Hansel, T. T. *et al.* Purification of human blood eosinophils by negative selection using
865 immunomagnetic beads. *Journal of Immunological Methods* **122**, 97–103 (1989).
866 43. Hansel, T. T. *et al.* An improved immunomagnetic procedure for the isolation of highly
867 purified human blood eosinophils. *Journal of Immunological Methods* **145**, 105–110
868 (1991).
869 44. Munoz, N. M. & Leff, A. R. Highly purified selective isolation of eosinophils from human
870 peripheral blood by negative immunomagnetic selection. *Nat Protoc* **1**, 2613–2620
871 (2006).
872 45. She, Y. & Owen, A. B. Outlier Detection Using Nonconvex Penalized Regression. *Journal*
873 *of the American Statistical Association* **106**, 626–639 (2011).

46. Lee, Y., MacEachern, S. N. & Jung, Y. Regularization of Case-Specific Parameters for Robustness and Efficiency. *Statistical Science* **27**, 350–372 (2012).

47. Kong, D., Bondell, H. D. & Wu, Y. Fully Efficient Robust Estimation, Outlier Detection and Variable Selection Via Penalized Regression. *Statistica Sinica* **28**, 1031–1052 (2018).

48. Traniello, I. M. *et al.* Single-cell dissection of aggression in honeybee colonies. *Nat Ecol Evol* **7**, 1232–1244 (2023).

49. Lin, L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **45**, 255–268 (1989).

50. Kumar, A. *et al.* Intracellular spatial transcriptomic analysis toolkit (InSTAnT). *Nat Commun* **15**, 7794 (2024).

51. Kim, Y. *et al.* SpaceExpress enables differential spatial transcriptomics with natural coordinate systems. 2024.12.19.628720 Preprint at https://doi.org/10.1101/2024.12.19.628720 (2025).

52. Yeo, S., Schrader, A. W., Lee, J., Asadian, M. & Han, H.-S. Spot-Based Global Registration for Subpixel Stitching of Single-Molecule Resolution Images for Tissue-Scale Spatial Transcriptomics. *Anal. Chem.* **96**, 6517–6522 (2024).

53. Wang, G. *et al.* Spatial organization of the transcriptome in individual neurons. 2020.12.07.414060 Preprint at https://doi.org/10.1101/2020.12.07.414060 (2020).

54. Moffitt, J. R. *et al.* High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proceedings of the National Academy of Sciences* **113**, 14456–14461 (2016).

896 55. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of
897 RNAs with MERFISH and expansion microscopy. *Sci Rep* **8**, 4847 (2018).
898 56. Khaledian, B., Thibes, L. & Shimono, Y. Adipocyte regulation of cancer stem cells.
899 *Cancer Sci* **114**, 4134–4144 (2023).
900 57. Hu, X. *et al.* Decorin-mediated suppression of tumorigenesis, invasion, and metastasis
901 in inflammatory breast cancer. *Commun Biol* **4**, 72 (2021).
902 58. Cai, R. *et al.* Primary breast tumor induced extracellular matrix remodeling in
903 premetastatic lungs. *Sci Rep* **13**, 18566 (2023).
904 59. de Freitas Oliveira-Tore, C. *et al.* Non-canonical extracellular complement pathways
905 and the complosome paradigm in cancer: a scoping review. *Front. Immunol.* **16**,
906 (2025).
907 60. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human
908 dorsolateral prefrontal cortex. *Nat Neurosci* **24**, 425–436 (2021).
909 61. Johnson, Z. V. *et al.* Cellular profiling of a recently-evolved social behavior in cichlid
910 fishes. *Nat Commun* **14**, 4891 (2023).
911 62. Lake, B. B. *et al.* A comparative strategy for single-nucleus and single-cell
912 transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA.
913 *Sci Rep* **7**, 6031 (2017).
914 63. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat*
915 *Methods* **14**, 955–958 (2017).
916 64. DeFelipe, J. *et al.* Cortical white matter: beyond the pale remarks, main conclusions
917 and discussion. *Front. Neuroanat.* **4**, 1692 (2010).

65. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

66. Sakers, K. *et al.* Astrocytes locally translate transcripts in their peripheral processes. *Proceedings of the National Academy of Sciences* **114**, E3830–E3838 (2017).

67. Jiang, T. *et al.* The lncRNA MALAT1/miR-30/Spastin Axis Regulates Hippocampal Neurite Outgrowth. *Front. Cell. Neurosci.* **14**, (2020).

68. Overly, C. C., Rieff, H. I. & Hollenbeck, P. J. Organelle motility and metabolism in axons vs dendrites of cultured hippocampal neurons. *J Cell Sci* **109**, 971–980 (1996).

69. Harris, J. J., Jolivet, R. & Attwell, D. Synaptic Energy Use and Supply. *Neuron* **75**, 762–777 (2012).

70. Than-Trong, E. & Bally-Cuif, L. Radial glia and neural progenitors in the adult zebrafish central nervous system. *Glia* **63**, 1406–1428 (2015).

71. Miranda-Negrón, Y. & García-Arrarás, J. E. Radial glia and radial glia-like cells: Their role in neurogenesis and regeneration. *Front. Neurosci.* **16**, (2022).

72. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology* **22**, 184 (2021).

73. Abramsson, A. *et al.* The zebrafish amyloid precursor protein-b is required for motor neuron guidance and synapse formation. *Developmental Biology* **381**, 377–388 (2013).

74. Newman, M., Ebrahimie, E. & Lardelli, M. Using the zebrafish model for Alzheimer's disease research. *Front. Genet.* **5**, (2014).

75. Haenisch, C. *et al*. The neuronal growth and regeneration associated Cntn1 (F3/F11/Contactin) gene is duplicated in fish: expression during development and retinal axon regeneration. *Molecular and Cellular Neuroscience* **28**, 361–374 (2005).
76. Eiraku, M. *et al.* DNER acts as a neuron-specific Notch ligand during Bergmann glial development. *Nat Neurosci* **8**, 873–880 (2005).
77. Díaz-Ramos, À., Roig-Borrellas, A., García-Melero, A. & López-Alemany, R. α-Enolase, a Multifunctional Protein: Its Role on Pathophysiological Situations. *BioMed Research International* **2012**, 156795 (2012).
78. Buxbaum, A. R., Haimovich, G. & Singer, R. H. In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* **16**, 95–109 (2015).
79. Zhou, Z., Li, X., Wright, J., Candès, E. & Ma, Y. Stable Principal Component Pursuit. in *2010 IEEE International Symposium on Information Theory* 1518–1522 (2010). doi:10.1109/ISIT.2010.5513535.
80. Mateos, G. & Giannakis, G. B. Robust PCA as Bilinear Decomposition With Outlier-Sparsity Regularization. *Trans. Sig. Proc.* **60**, 5176–5190 (2012).
81. Manabe, R. *et al*. Transcriptome-based systematic identification of extracellular matrix proteins. *Proceedings of the National Academy of Sciences* **105**, 12849–12854 (2008).
82. Giannakis, G. B., Mateos, G., Farahmand, S., Kekatos, V. & Zhu, H. USPACOR: Universal sparsity-controlling outlier rejection. in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1952–1955 (2011). doi:10.1109/ICASSP.2011.5946891.

83. Huang, P., Cai, M., Lu, X., McKennan, C. & Wang, J. Accurate estimation of rare cell-type fractions from tissue omics data via hierarchical deconvolution. *The Annals of Applied Statistics* **18**, 1178–1194 (2024).

84. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat Methods* **9**, 8–9 (2012).

85. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).

86. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**, 295 (2019).

87. Mu, X. *et al*. Single-nucleus and spatial transcriptomics identify brain landscape of gene regulatory networks associated with behavioral maturation in honeybees. *Nat Commun* **16**, 3343 (2025).

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