AdRoit: an accurate and robust method to infer complex transcriptome composition

Tao Yang[1], Nicole Alessandri-Haber[1], Wen Fury[1], Michael Schaner[1], Robert Breese[1], Michael LaCroix-Fralish[2], Jinrang Kim[1], Christina Adler[1], Lynn E. Macdonald[1], Gurinder S. Atwal[1], Yu Bai[1, *]

Affiliations

1. Regeneron Pharmaceuticals, Inc., Tarrytown NY 10591
2. Cellular Longevity, Inc., San Francisco, CA 94103

*Corresponding author

Abstract

RNA sequencing technology promises an unprecedented opportunity in learning disease mechanisms and discovering new treatment targets. Recent spatial transcriptomics methods further enable the transcriptome profiling at spatially resolved spots in a tissue section. In controlled experiments, it is often of immense importance to know the cell composition in different samples. Understanding the cell type content in each tissue spot is also crucial to the spatial transcriptome data interpretation. Though single cell RNA-seq has the power to reveal cell type composition and expression heterogeneity in different cells, it remains costly and sometimes infeasible when live cells cannot be obtained or sufficiently dissociated. To computationally resolve the cell composition in RNA-seq data of mixed cells, we present AdRoit, an accurate and robust method to infer transcriptome composition. The method estimates the proportions of each cell type in the compound RNA-seq data using known single cell data of relevant cell types. It uniquely uses an adaptive learning approach to correct the bias gene-wise

26 due to the difference in sequencing techniques. AdRoit also utilizes cell type specific genes
27 while control their cross-sample variability. Our systematic benchmarking, spanning from
28 simple to complex tissues, shows that AdRoit has superior sensitivity and specificity compared
29 to other existing methods. Its performance holds for multiple single cell and compound RNA-
30 seq platforms. In addition, AdRoit is computationally efficient and runs one to two orders of
31 magnitude faster than some of the state-of-the-art methods.
32

33 ## Introduction

34 RNA sequencing is a powerful tool to address the transcriptomic perturbations in disease
35 tissues and help understand the underlying mechanism to develop treatments[1]. Due to the
36 presence of heterogeneous cell populations, bulk tissue transcriptome only characterizes the
37 averaged expression of genes over a mixture of different types of cells. The identity of
38 individual cell types and their prevalence remain unelucidated in the bulk data. However,
39 knowledge of the cell type composition and gene expression perturbation at the cell type level
40 is often critical to identifying disease-manifesting cells and designing targeted therapies. For
41 instance, the constitution of stromal and immune cells sculpts the tumor microenvironment
42 that is essential in cancer progression and control[2–6]. Excessive expression of cytokines in
43 particular leukocyte types underlines the etiology of many chronic inflammatory diseases [7–11].
44 Such information cannot be directly read out from the bulk RNA-Seq.
45

46 Recent breakthroughs in spatial transcriptomics methods enable characterizing whole
47 transcriptome-wise gene expressions at spatially resolved locations in a tissue section[12].

48 However, it remains challenging to reach a single cell resolution while measuring tens of
49 thousands of genes transcriptome-wise. Some widely used technologies can achieve a
50 resolution of 50-100 µm, equivalent to 3–30 cells depending on the tissue type[12,13]. The
51 transcripts therein may originate from one or more cell types. Unlike the bulk RNA-seq, the
52 profiling data at each spot contains substantial dropouts as merely a few cells are sequenced,
53 imposing additional challenges to demystify the cell type content. We refer to bulk RNA-seq
54 and spatial transcriptomics data at the multi-cell resolution as compound RNA-seq data
55 hereafter.
56
57 The rapid development of single-cell RNA-seq (scRNA-seq) technologies has allowed for cell-
58 type specific transcriptome profiling[14]. It provides the information missing from the compound
59 RNA-seq data. Nevertheless, the technologies have low sensitivity and substantial noise due to
60 the high dropout rate and the cell-to-cell variability. Consequently, scRNA-seq technologies
61 require a large number of cells (thousands to tens of thousands) to ensure statistical
62 significance in the results. In addition, the cells must remain viable during capture. These
63 requirements render the scRNA-seq technologies costly, prohibiting their application in clinical
64 studies that involve many subjects or cannot allow real time tissue dissociation and cell capture.
65 Furthermore, scRNA-seq technologies may not be well suited to characterizing cell-type
66 proportions in solid tissues because the dissociation and capture steps can be ineffective to
67 certain cell types [15–17].
68

69 As sequencing at the single cell level is not always feasible, in silico approaches have been
70 developed to infer cell type proportions from compound RNA-seq data[18–24]. The most common
71 strategy is to conduct a statistical inference through the maximum likelihood estimation
72 (MLE)[25] or the maximum a posterior estimation (MAP)[26] on a constrained linear regression
73 framework, wherein the unobserved mixing proportion of a finite number of cell types are part
74 of the latent variables to be optimized. [1921–24]The deconvolution methods are often applied to
75 dissect the immune cell compositions in blood samples[27–31]. However, their performance in
76 more complex tissues, such as the nervous, ocular, respiratory and gastrointestinal organs,
77 remains unclear. These tissues often contain many cell types (10-10^2) and the difference among
78 related cells can be subtle, rendering the deconvolution a challenging task. For example, a
79 recent study on the mouse nervous system contains more than 200 cell clusters and many are
80 highly similar neuronal subtypes[32].
81
82 Earlier works often utilized the transcriptome profiling of the purified cell populations to
83 estimate the gene expressions per cell type (e.g. Cibersort)[19]. More recently, acquiring cell type
84 specific expression from the scRNA-seq data was shown to be an intriguing alternative[21–24].
85 Though it provides higher throughput by measuring multiple cell types in one experiment,
86 profiling at single cell level is substantially noisy. Deconvolution using scRNA-seq data as
87 reference can be biased by noise non-relevant to cell identities if not treated properly.
88 Moreover, the platform difference between the compound data and the single cell data cannot
89 be ignored.
90

91 To overcome these challenges, additional information from the data may be considered. A
92 recent method that weighs genes according to their expression variances across samples
93 greatly improved the accuracy[22], highlighting the importance of gene variability in inferring cell
94 type composition. Some other methods and applications have pointed out the importance of
95 cell type specific genes[24,28,31,33]. In these works, the cell type specific expression was only used
96 to select the input genes (e.g., markers). Nonetheless, it measures how informative a gene is in
97 distinguishing cell types and thus can be incorporated as a part of the model. To address the
98 platform difference between the compound data and the single cell data it is usually assumed
99 there exists a single scaling factor or a linearly scaled bias for all genes that can be learned and
100 corrected accordingly[22,23]. This assumption is hardly held because the impact of the platform
101 difference to each gene is different. Though learning a uniform scaling factor would correct the
102 difference in the majority of genes, a few genes that remain significantly biased can easily
103 confound the estimation, especially under a linear model framework. Thus, a gene-wise
104 correction should be considered.
105
106 In this work, we presented a new deconvolution method, AdRoit, a unified framework that
107 jointly models the gene-wise technology bias, genes' cell type specificity and cross-sample
108 variability. The method estimated the cell type constitution in the compound RNA-seq samples
109 using relevant single cell data as a training source. Genes used for deconvolution were
110 automatically selected from the single cell data based on their information richness. Uniquely,
111 it uses an adaptively learning approach to estimate gene-wise scaling factors, addressing the
112 issue that different platforms impact genes differently. The model of AdRoit is further

113 regularized to avoid collinearity among closely related cell subtypes that are common in
114 complex tissues. Over a comprehensive benchmarking data sets with a varying cell composition
115 complexity, AdRoit showed superior sensitivity and specificity to other existing methods.
116 Applications to real RNA-seq bulk data and spatial transcriptomics data revealed strong and
117 expected biologically relevant information. We believe AdRoit offers an accurate and robust
118 tool for cell type deconvolution and will promote the value of the bulk RNA-seq and the spatial
119 transcriptomics profiling.
120

121 ## Results

122 **Overview of the AdRoit framework**

123 AdRoit estimates the proportions of cell types from compound transcriptome data including but
124 not limited to bulk RNA-seq and spatial transcriptome. It directly models the raw reads without
125 normalization, preserving the difference in total amounts of RNA transcript in different cell
126 types. The method utilizes as reference the relevant pre-existing single cell RNA-seq data with
127 cell identity annotation. It selects informative genes, estimates the mean and dispersion of the
128 expression of selected genes per cell type, and constructs a weighted regularized linear model
129 to infer percent combinations (Fig. 1a). Because sequencing platform bias impacts genes
130 differently[15,34,35], a uniform scaling factor for all genes does not sufficiently eliminate such bias.
131 A key innovation of AdRoit is that it uniquely adopts an adaptive learning approach, where the
132 bias was first estimated for each gene, then adjusted such that more biased gene is corrected
133 with a larger scaling factor (Fig. 1b).
134

135 We also attribute the success of AdRoit to the consideration of a comprehensive set of other
136 relevant factors including genes' cross-sample variability, cell type specificity and collinearity of
137 expression profiles among closely related cell types. The cross-sample variability of a gene
138 confounds its biological expression variability due to the variety of cell types. The latter is
139 referred as the cell type specific expression that helps identify the cell type. AdRoit weighs
140 down genes with high cross-sample variability whilst weighs up those with an expression highly
141 specific to certain cell types. The definition of cross-sample variability and cell type specificity
142 also accounts for the overdispersion nature in counts data. Lastly, AdRoit adopted a linear
143 model to ensure the interpretability of the coefficients. At the same time, AdRoit included a
144 regularization term to minimize the impact of the statistical collinearity. Each of the factors
145 contributes an indispensable part to AdRoit, leading to an accurate and robust deconvolution
146 method for inferring complex cell compositions.
147
148 To evaluate the performance, we compared AdRoit with MuSiC[22] and NNLS[18,36] for bulk data
149 deconvolution, and stereoscope[23] for spatial transcriptomics data deconvolution. When
150 evaluating the algorithms, a common practice is to pool the single cell data to synthesize a
151 "bulk" sample with the known ground truth of the cell composition. We measured the
152 performance by comparing the estimated cell proportions with true proportions using four
153 metrics: mean absolution difference (mAD), rooted mean squared deviation (RMSD) and two
154 correlation statistics (i.e., Pearson and Spearman). Both correlations are included because
155 Pearson reflects linearity, while Spearman avoids the artificial high scores driven by outliers
156 when majority of estimates are tiny. Good estimations feature low mAD and RMSD along with

157 high correlations. When estimating cell proportions for a synthetic sample, cells from this
158 sample are excluded from the input single cell reference (i.e., leave-one-out) to avoid
159 overfitting. We further applied AdRoit to real bulk RNA-seq data and validated the results by
160 available RNA fluorescence *in-situ* hybridization (RNA-FISH) data. The estimates were further
161 confirmed by relevant biology knowledge of human pancreatic islets. We also used AdRoit to
162 map cell types on spatial spots, and the accuracy was verified by *in-situ* hybridization (ISH)
163 images from Allen mouse brain atlas[37].
164
165 **AdRoit excels in datasets with both simple and complex cell constitutions**
166 We started with a simple human pancreatic islets dataset that contains 1492 cells and four
167 distinct endocrine cell types (i.e., Alpha, Beta, Delta, and PP cells)[38] (Extended Data Fig. 1a;
168 Supplementary Table 1). The synthesized bulk data were constructed by mixing the single cell
169 data at known proportions. Though all three methods achieved satisfactory performance
170 according to the evaluation metrics, AdRoit has slightly better performance as reflected by
171 scatterplots of estimated proportion vs. true proportion (Extended Data Fig. 1b, Supplementary
172 Table 2). It has moderately lower mAD (0.029 vs. 0.031 for MuSiC and 0.066 for NNLS), and
173 RMSD (0.039 vs. 0.046 for MuSiC and 0.095 for NNLS) and comparable correlations (Pearson:
174 0.99 vs 0.98 for MuSiC and 0.93 for NNLS; Spearman: 0.97 vs 0.98 for MuSiC and 0.91 for NNLS)
175 (Extended Data Fig. 1c). This performance was expected because there were only four cell types
176 with very distinct transcriptome profiles. Deconvoluting such data was a relatively easy task for
177 all three methods.
178

179 We then tested the methods on a couple of complex tissues that are more challenging to
180 deconvolute. One is the human trabecular meshwork (TM) tissue. We acquired published single
181 cell data that contains 8758 cells and 12 cell types from 8 donors[39]. The data include 3 similar
182 types of endothelial cells, 2 types of Schwann cells and 2 types TM cells (Supplementary Fig. 1;
183 Supplementary Table 3). Cells from each donor were pooled as a synthetic bulk sample. The cell
184 type proportions vary from <1% to 43%. These proportions were the ground truth cell
185 composition and were compared head-to-head with the estimated proportions inferred by
186 AdRoit, MuSiC and NNLS. For each synthetic bulk sample, estimations were performed using a
187 reference built from cells of other donors (i.e., leaving-one-out). In each of the 8 samples, the
188 estimates made by AdRoit best approximated the true proportions. In particular, AdRoit had
189 significantly lower mAD (0.016) and RMSD (0.025), and higher correlations (Pearson = 0.97;
190 Spearman = 0.94), comparing to MuSiC (mAD = 0.038; RMSD = 0.06; Pearson = 0.83; Spearman
191 = 0.73) and NNLS (mAD = 0.06; RMSD = 0.088; Pearson = 0.69; Spearman = 0.63) (Fig. 2a). We
192 further assessed the deviation of the estimates from the true proportions for each cell type.
193 AdRoit consistently had the lowest deviations from the true proportions for all cell types, as
194 well as the lowest variation among 8 samples (Fig. 2b, blue dots), indicating a higher robustness
195 over various cell types and samples. Notably, AdRoit only missed one rare cell type (true
196 proportion = 0.3%) out of 12 cell types in one sample, while MuSiC missed 1 to 5 cell types in 6
197 of the 8 samples, and NNLS missed 3 to 7 cell types in all 8 samples (Supplementary Fig. 2,
198 Supplementary Table 4).
199
200 **AdRoit has better sensitivity and specificity**

201 We next systematically addressed the sensitivity and specificity of these algorithms. In the
202 context of the cell type deconvolution, a false negative occurs when the proportion of an
203 existing cell type is predicted to be zero (or below a given threshold). Conversely, a non-zero
204 prediction (or above a given threshold) of an absent cell type results in a false positive. False
205 negatives and false positives measure the sensitivity and specificity of a deconvolution
206 algorithm, respectively. Both quantities are crucial to establish the utility of the algorithm.
207 Particularly, in real world applications, it is often difficult to know *a prior* what cell types exist in
208 a bulk sample, users may inform the algorithm to consider more possible cell types than what
209 are actually in the sample. False positive predictions in this situation would make the algorithm
210 unusable.
211
212 We designed a simulation to test the sensitivity and specificity. we selected 6 out of the 12 cell
213 types, i.e., Schwann-cell like cell, TM1, smooth muscle cell, melanocyte, macrophage and
214 pericyte, from each donor sample and pooled them within that sample to synthesize 8 new bulk
215 samples. The unselected 6 cell types are considered absent in the bulk samples. Some cell types
216 in presence are highly similar to those in absence, challenging the programs to pinpoint the
217 right cell type present in the bulk among similar candidates. We provided the full list of 12
218 single cell types as reference to the programs to estimate the cell type proportions. NNLS was
219 excluded from this evaluation due to its low benchmarking performance observed earlier (Fig.
220 2a, b).
221

222 Consistently across 8 samples, AdRoit had very accurate estimates for the 6 present cell types,
223 and zero or close-to-zero estimated values for the non-existing cell types in the synthetic bulk
224 data. MuSiC was notably less accurate on the 6 selected cell types, meanwhile it had many non-
225 negligible values (>1% for 26 out 48 estimates) of the 6 cell types excluded in the 8 synthetic
226 samples (Fig. 2c, Supplementary Table 5). For example, smooth muscle cells accounted for
227 ~14% in donor 4 but was largely missed (~0.03%) by MuSiC. We noted that TM2 had false non-
228 zero estimates from both methods though not included. This is because TM2 is easily mistaken
229 as TM1 due to their high similarity[39]. Nonetheless, AdRoit's estimates of TM2 were consistently
230 small across samples (<1% for 44 out of 48 estimates), while MuSiC had significantly larger
231 estimates of TM2 that occasionally even exceeded the TM1 estimates (donors 5 and 8 in Fig. 2c
232 right). For a systematic comparison, we constructed the receiver operating characteristic (ROC)
233 curve by varying the threshold of detection (i.e., a cutoff below which the cell type was deemed
234 undetected) (Fig. 2d). AdRoit had significantly higher area under the curve (AUC) than MuSiC
235 (0.95 vs. 0.74), implying a dominantly better sensitivity and specificity.
236
237 **AdRoit outperforms in deconvoluting closely related subtypes**
238 To further evaluate AdRoit when multiple cell subtypes present in a complex tissue, we
239 performed scRNA-seq experiment on mouse lumbar dorsal root ganglion (DRG) from five mice.
240 Following the standard analysis pipeline (Methods), we obtained 3352 single cells after quality
241 control procedures. After clustering and annotation, we discovered 14 cell types including
242 multiple subtypes of neuronal cells (Fig. 3a, Supplementary Table 6). The heatmap of the top
243 marker genes showed distinct patterns of the major cell types as well as similar patterns of the

244 subtypes (Extended Data Fig. 2a), and the cell type proportions varied from 0.5% to 33.71%
245 (Extended Data Fig. 2b). These 14 cell types include 3 subtypes of neurofilament containing
246 neurons (i.e., NF_Calb1, NF_Pvalb, NF_Ntrk2.Necab2), 3 subtypes of non-peptidergic neurons
247 (i.e., NP_Nts, NP_Mrgpra3, NP_Mrgprd), and 5 subtypes of peptidergic neurons (i.e., PEP1_Dcn,
248 PEP1_S100a11.Tagln2, PEP1_Slc7a3.Sstr2, PEP2_Htr3a.Sema5a, PEP3_Trpm8). Also discovered
249 were tyrosine hydroxylase containing neurons (Th), satellite glia and endothelial cells. Such
250 complex compositions formed a challenging testing ground for evaluating the ability to
251 distinguish closely related cell types. We again did the leave-one-out deconvolution on five
252 synthesized bulk samples.
253
254 AdRoit had highly accurate estimations on all cell types across samples (Fig. 3b). It is worth to
255 mention that, for the rare cell types that account for less than 5%, AdRoit still had a good
256 estimation that is fairly close to the true proportions and never missed a single cell type,
257 showing that AdRoit is very robust on rare cell types. For example, 0.51% endothelial cells were
258 predicted to be 0.35%, and 1.05% NF2_Ntrk2.Necab2 cells were predicted to be 0.85%
259 (Supplementary Fig. 3, Supplementary Table 7). On the contrary, MuSiC and NNLS were notably
260 less accurate, especially for the cell types less than 5%, and missed multiple cell types including
261 some large cell clusters taking account of ~10% (PEP1_Slc7a3.Sstr2 cells of Sample5). We
262 further examined how much the variability of the estimates was in each individual sample. We
263 computed the 4 metrics to evaluate the performance on each of the 5 synthetic samples and
264 compared them head-to-head among the algorithms. This fine comparison showed AdRoit
265 significantly outperformed MuSiC and NNLS on every sample (Fig. 3c). Further, the performance

266 metrics of AdRoit were highly consistent across samples with the lowest variability among the
267 three methods.
268
269 **AdRoit excels on simulated spatial transcriptomics data**
270 Given the promising performance on complex tissues, we continued to test AdRoit's
271 applicability to spatial transcriptomics data. Spatial transcriptomics data differs from bulk RNA-
272 seq data in that each spot only contains transcripts from a handful of cells (3-30)[12]. Some of the
273 spots contain multiple cells of the same type, while others may have mixtures of cell types at
274 varying mixing percentages (e.g., spatial spots at the boundary of different cell types). Also,
275 because the mixture is a pool of only a few cells, the variations across spatial spots are
276 expected to be greater than in bulk samples. We simulated a large number of spatial spots
277 (3200 in total) by using sampled cells from the DRG single cell data above (Methods), then
278 compared AdRoit with Stereoscope over a range of simulation scenarios.
279
280 We first tested whether the methods could correctly infer a single cell type when the spots
281 contain cells from that same type. For each of the 14 cell types from DRG, we sampled 10 cells
282 and pooled them to form a spatial spot. We repeated the simulation for 100 times for a robust
283 testing, then used the full set of 14 cell types as reference to deconvolute the 1400 simulated
284 spots. Both methods were able to identify the correct cell types with indistinguishable accuracy
285 on the simulated cell types (i.e., estimates close to 1) and comparably low estimated values
286 (i.e., estimates close to zero) for other cell types not included when simulating the spots
287 (Extended Data Fig. 3).

288

289 We then continued a difficult scenario where we sampled cells from the 5 PEP subtypes and
290 mixed them. We created three simulation schemes for a comprehensive evaluation: 1) 5 PEP
291 subtypes had same percent of 0.2; 2) PEP1_Dcn was 0.1 and the other 4 were 0.225; 3)
292 PEP1_S100a11.Tagln2 and PEPE1_Dcn were 0.1, PEP2_Htr3a.Sema5a and PEP1_Slc7a3.Sstr2
293 were 0.2, and PEP3_Trpm8 was 0.4. Again, each simulation scheme was repeated 100 times.
294 Under each scheme, the estimates by AdRoit consistently centered around true proportions
295 and the other cell types had very low estimated values (close to zero) (Fig. 4a, Supplementary
296 Table 8). In comparison, though the estimates for the other cell types were also generally close
297 to zero, the estimates of the PEP cells by Stereoscope systematically deviated from the true
298 proportions for all three simulated schemes except for PEP1_S100a11.Tagln2.

299

300 We further expanded the simulated spatial spots to the mixture of 3 NP cell types and mixture
301 of 3 NF cell types. In addition, we sampled NP_Mrgpra3 cells and mixed them with other
302 distinct cell types (i.e., Th, satellite glia and endothelial), as well as NF_Calb1 cells mixed with
303 other distinct cell types, and PEP3_Trpm8 mixed with other distinct cell types. For all these
304 simulated spatial spots, AdRoit's estimates were consistently centered at true proportions,
305 whereas Stereoscope's estimates deviated in almost all simulated schemes (Extended Data Fig.
306 4, Supplementary Table 8). We speculate the main reason Stereoscope underperformed at
307 these simulated spots is that it normalizes the total UMI counts to the same number for all
308 cells. In real world, a spatial spot is unlikely to be a pool of cells that have the same total RNA
309 transcripts sampled, especially when a spot contains different cell types (e.g., immune cells

310 have about 10-fold less total UMIs than the neuronal cells or subtypes of neuronal cells). Our
311 simulation pooled the sampled cells by adding up the raw UMI counts per gene, which we
312 believe best mimics the real data.
313
314 Next, we asked how sensitive the methods are in detecting rare cell populations. We simulated
315 mixtures of 3 PEP subtypes (i.e., PEP1_Slc7a3.Sstr2, PEP2_Htr3a.Sema5a, PEP3_Trpm8) with a
316 series of low percent PEP3_Trpm8 (from 0.01 to 0.1 by 0.01), and the other two cell types
317 sharing the rest percentage equally (Methods). At each given percent, the simulation was
318 repeated 100 times. We then checked how accurately the percent of PEP3_Trpm8 cells was
319 estimated. The medians of AdRoit's estimates were always close to the true proportions (Fig.
320 4b, red lines), whereas that of Stereoscope's estimates were largely lower than true
321 proportions. Stereoscope also missed the majority of PEP3_Trpm8 cell type when the simulated
322 proportion was below 0.06. This comparison implied AdRoit is more advantageous in detecting
323 low percent cells. For a complete comparison, we also simulated 5 other types of cell mixtures
324 in the same way. At each given low percent, we computed how many times out of 100 the low
325 percent cell component was detected (estimates > 0.005). AdRoit had systematically higher
326 detection rates, as well as higher consistency across different cell mixtures (Fig. 4c,
327 Supplementary Table 9). Notably, at a simulated percent of 5%, AdRoit achieved >90% of
328 detention rate, making it a powerful tool in detecting rare cells.
329
330 Though MuSiC was not designed for deconvoluting spatial spots, theoretically it also can be
331 applied to spatial transcriptomics data. We thus also compared AdRoit to MuSiC on the same

332 sets of simulation data above. We observed AdRoit was also significantly more accurate over all
333 simulation scenarios of spatial spots (Fig. 4a, Extended Data Fig. 3 and 4, Supplementary Fig. 4),
334 and more sensitive when detecting low percent cells (Fig. 4b, c, Supplementary Fig. 5).
335
336 **Application to real bulk RNA-seq data of human pancreatic islets**
337 Though using synthetic bulk data based on mixing of single cells is a useful benchmarking
338 strategy, the bulk and single cell RNA-seq often use distinct RNA library preparation and
339 sequencing protocols. The capability of a method to deconvolute real bulk samples shall be
340 addressed to ensure it is useful in the real-world applications. We acquired 70 real human
341 pancreatic islets bulk samples from published studies[38,40,41] (Supplementary Table 10) and used
342 single cell data of the same tissue[38] as reference to infer the percentages of 4 endocrine cell
343 types (i.e., Alpha, Beta, Delta, PP). The 70 bulk samples were collected from 39 distinct donors,
344 including 26 healthy donors, and 13 donors with type 2 diabetes (T2D). Each donor contributed
345 1 to 5 replicated bulk RNA samples.
346
347 Replicates from the same donor are expected to have similar compositions and thus were used
348 to assess the reproducibility of the estimates from AdRoit. For all cell types, AdRoit had highly
349 consistent estimates for the same donors (Fig. 5a, Supplementary Table 11). The average
350 standard deviations did not exceed 1% for all 4 cell types (i.e., Alpha: 0.010; Beta: 0.008; Delta:
351 0.004; PP: 0.002). To seek an independent validation, we obtained cell sorting results by RNA-
352 FISH for 4 of the 39 donors[38] (Supplementary Table 12). The estimated cell proportions of the 4
353 were highly consistent with the percentages measured by RNA-FISH (Fig. 5b), and the

354 consistency held for both major cells (Alpha and Beta) and the minor cells (Delta and PP).
355 Reproducibility and independent validation showed AdRoit is reliable in deconvoluting real bulk
356 RNA-seq data.
357
358 We then asked if AdRoit can detect known biological differences between healthy and T2D
359 donors. Loss of functional insulin-producing Beta cells is a prominent characteristic of T2D[42–44],
360 typically reflected by elevated level of hemoglobin A1c (HbA1c)[45,46]. Among the healthy donors,
361 the majority of Beta cell proportions estimated by AdRoit ranged from 50% to 75% (Fig. 5c),
362 agreed with the known percent range of Beta cells in human islets tissue[47,48]. A significant
363 decreasing of the estimated Beta cell proportions was seen in T2D patients (P value = 4.1e-6).
364 Further, a linear regression of estimated Beta cell proportions on HbA1c levels showed a
365 statistically significant negative association (P value = 1.8e-6). AdRoit adequately reflected the
366 cell composition difference between healthy donors and T2D patients.
367
368 **Application to mouse brain spatial transcriptomics**
369 We lastly demonstrated an application to the real spatial transcriptomics data. Given the
370 molecular architecture of brain tissue has been well studied, we chose mouse brain spatial
371 transcriptomics data generated by 10x genomics, containing 2703 spatial spots (Methods). The
372 reference single cell data were acquired from an independent study which contains a
373 comprehensive set of nervous cell types in brain[32]. We curated the cell types by merging highly
374 similar clusters and came down to a consolidated set of 46 distinct brain cell types (Methods,
375 Supplementary Table 13).

The cell contents inferred by AdRoit per spot appear to accurately match the expected cell types at that location (Extended Data Fig. 5, Supplementary Table 14). For example, the three subtypes of cortex excitatory neurons each occupied a sub-area in the cerebral cortex region. As another example, the shape of hippocampal region was delineated by the estimated percentages of dentate gyrus granule/excitatory neurons. For an independent validation, we checked the consistency between estimated cell types with the *in-situ* hybridization (ISH) images from Allen mouse brain atlas[49]. We chose 4 genes highly expressed in 4 brain regions respectively, i.e., Spink8 for hippocampal field CA1, C1ql2 for dentate gyrus, Clic6 for choroid plexus, and Synpo2 for thalamus[32]. The spots enriched with the 4 cell types (i.e., hippocampal CA1 excitatory neuron type 2, dentate gyrus granule neuron type 2, choroid plexus cell, thalamus excitatory neuron type 1), as mapped by AdRoit, precisely co-localized with the strong signals of the 4 marker genes on the ISH images respectively (Fig. 5d). This agreement confirmed that the spatial mapping of cell types by AdRoit is reliable.

Computational efficiency

Besides the accuracy and robustness, another major advantage of AdRoit is its magnitude higher computational efficiency. AdRoit uses a two-step procedure to do the inference. The first step prepares the reference on single cell data where per-gene means and dispersions are estimated, and cell type specificity is subsequently computed. The built reference can be saved and reused. We tested the running time on the reference building using the aforementioned mouse brain single cell dataset containing ~15,000 cells. It took about 4.5 minutes on a CPU

398 that has 24 cores (23 used for parallel computing). The second step inputs the built reference
399 and target compound data and does the estimation. Deconvoluting ~2700 compound RNA-seq
400 samples took around 5 minutes. Therefore, AdRoit in total took less than 10 minutes and ~3Gb
401 memory usage on a regular CPU. As a comparison, MuSiC took about 1 hour and 37 minutes on
402 the same data using the same CPU. Stereoscope ran about 24 hours continuously with the
403 published parameter setting (-scb 256 -sce 75000 -topn_genes 5000 -ste 75000 -lr 0.01 -stb 100
404 -scb 100) on a powerful V100 GPU with 80 cores and 16G memory, which is prohibitive for
405 seeking a quick turnaround.
406

407 **Discussion**

408 In this work we have demonstrated that AdRoit is capable of deconvoluting the cell
409 compositions from the compound RNA-seq data with a leading accuracy, measured by the
410 consistency between the true and predicted cell proportions. Its advantage over the existing
411 state-of-the-art methods was verified over a wide range of use cases. In particular, AdRoit
412 excelled in complex tissues composed of more than ten different cell types with wide range of
413 cell proportions (e.g., trabecular meshwork, dorsal root ganglion). In both cases, AdRoit
414 performed significantly better than the comparators MuSiC and NNLS on deconvoluting bulk
415 RNA-seq data. AdRoit is also more accurate and sensitive than Stereoscope in demystifying
416 spatial transcriptomics spots, especially in detecting low percent cells. Previous benchmarking
417 often assumed the types of cells in the synthetic bulk data are not more or less than the cell
418 types collected in the reference, and thus the only unknown was the proportion of each cell
419 type. This assumption may not hold. Missing existing cell types or false predictions of non-

420 existing ones can hinder the utility of an algorithm. Thus, besides the overall accuracy, we also
421 examined the sensitivity and specificity of the algorithms. We observed a superior sensitivity
422 and specificity in AdRoit, an important leverage for its usage in practice.
423
424 The reference single cell data used by AdRoit came from different platforms, such as the 10x
425 Genomics Chromium Instrument (the mouse dorsal root ganglion), and the Fluidigm C1 system
426 (the human pancreatic islets data). AdRoit consistently exhibited excellent performance across
427 all benchmarking datasets independent of their single cell sequencing technology platforms.
428 More importantly, this statement holds not only for deconvoluting the synthesized bulk data,
429 but also for the real bulk RNA-seq data. The latter typically does not apply the unique molecular
430 barcoding and requires a significantly different cDNA amplification procedure from what is used
431 in the single cell RNA-seq (Methods). Besides, the sequencing depth, read mapping and gene
432 expression quantification are dissimilar as well. The fact that AdRoit accurately dissected the
433 cell compositions in the real bulk samples based on the single cell reference data further
434 supports its cross-platform applicability.
435
436 We attribute the power of AdRoit to its comprehensive modeling of relevant factors. Firstly, we
437 think a common rescaling factor is not sufficient to correct the platform difference between
438 single cells and the compound data. Rather, the impact of platform difference to genes is quite
439 different and hardly is linearly scaled. Correcting such differences entails rescaling factors
440 specifically tailored to each gene. AdRoit uses an adaptive learning approach to estimate such
441 gene-wise correcting factor and does the correction in a unified model. In addition, the

442 contribution of a gene in a cell type to the loss function is jointly weighted by its specificity and
443 variability in a cell type, where specificity and variability are defined in a way accounting for the
444 overdispersion property of counts data. Our observations over the multiple benchmarking
445 dataset also show that the coexistence of similar cell types may have induced a collinearity
446 condition that negatively impacted the regression-based methods developed by others. Being
447 able to alleviate this problem gives AdRoit an edge to outperform. All these factors help AdRoit
448 to distinguish similar cell clusters while sensitive enough to separate rare cell types.
449
450 Technically, the input profiles of individual cell types to AdRoit does not necessarily come from
451 the single cell RNA-seq. Bulk RNA-seq profiles of individual isolated cell types can be used as
452 well. Nevertheless, using single cell RNA-seq data as the reference has a few key advantages. It
453 is a high throughput approach wherein multiple cell types can be interrogated simultaneously.
454 Prior knowledge of the cell types in presence as well as their specific gene markers are not
455 required, which allows novel cell types to be identified. Although detection of lowly expressing
456 genes has been a challenge for the single cell RNA-seq, significant enhancements have been
457 demonstrated. For example, the number of detectable genes currently can reach an order of
458 10,000 per cell and keeps improving[50]. As AdRoit focuses on the informative genes whose
459 expressions are generally high, the detection limit of the single cell RNA-seq does not impose a
460 significant drawback. Indeed, given the single cell reference profiles, AdRoit successfully
461 deconvoluted the real bulk RNA-seq data and spatial transcriptomics data. The results suggest
462 that, besides enriching our understanding of the bulk transcriptome data, AdRoit can leverage
463 the usage of the vast amount and continuously growing single cell data as well.

464

465 AdRoit is a reference-based deconvolution algorithm. A comprehensive collection of the
466 possible cell components is important. However, completeness may not always be guaranteed.
467 Even with the single cell acquisition that is independent of prior knowledge, rare and/or fragile
468 cell types may not survive through the capture procedure and hence are excluded. It is also
469 difficult to generate a solid reference profile for cells that are versatile from sample to sample
470 (e.g., tumor cells). Currently AdRoit deals implicitly with the components unknown to the
471 reference. If an unknown cell type reassembles one of the referenced ones, it may be
472 considered as part of the known cell type and their joint population is predicted. Such an
473 outcome is acceptable as treating two similar cell types as one is still biologically meaningful
474 although the resolution of the system may be compromised. If the unknown component is
475 dissimilar to all the known ones, it will be ignored by AdRoit because its representative markers
476 are unlikely among the top weighted genes associated with the known components. At the
477 same time, the distinct component is expected to have a unique gene expression pattern and
478 thus unlikely interferes significantly with the gene expressions from the known cell types.
479 Therefore, AdRoit essentially deconvolutes the relative populations among the known cell
480 components. For example, AdRoit was able to correctly uncover the populations of 4 endocrine
481 cell types from the human islet bulk data despite the absence of many other cell types such as
482 macrophages, Schwann cells and endothelial cells in the input single cell reference[20]. Although
483 under such a circumstance, the absolute percentages of the cells remain obscure, we expect
484 their relative proportions can be studied and valuable. A future improvement is to explicitly

485 model the unknown cell types and estimate their percentages upon the signals in the
486 compound data that cannot be explained by the contribution from the known components.
487

488 ## Methods

489 **Gene selection**

490 AdRoit selects genes that contain information about cell type identity, excluding non-
491 informative genes that potentially introduce noise. There are two ways for selecting such
492 genes: 1) union of the genes whose expression is enriched in one or more cell types in the
493 single cell UMI count matrix. These genes are referred as marker genes; 2) union of the genes
494 that vary the most across all the cells in the single cell UMI count matrix, referred as the highly
495 variable genes. For marker genes, we recommend selecting top ~200 genes (P value < 0.05),
496 ranked by fold change, from each cell type for resolving complex compound transcriptome
497 data. Considering some genes may mark more than one cell types, we further require selected
498 markers presenting in no more than 5 cell types to ensure specificity. We also suggest select a
499 minimal of 1000 total number unique genes for an accurate estimation. If not satisfied, one
500 may consider expand the number of top genes and/or loose the P value cutoff.
501

502 AdRoit also offer the option to use highly variable genes. To avoid the selected highly variable
503 genes being dominated by large cell clusters whilst underrepresents small clusters, AdRoit first
504 balances the cell types in the single cell UMI count matrix by finding the median size among all
505 cell clusters, then sample cells from each cluster to make them equal to this size. Next, AdRoit
506 computes the variance of each gene across the cells in the balanced single cell UMI matrix. Due

507 to the well-known dispersion effect in RNA-seq data, directly computing variances from count
508 matrix can results in overestimation. We thus compute variances on the normalized data done
509 by variance-stabilizing transformation (VST)[51]. Genes with top 2000 large variances are then
510 selected.
511
512 In both ways, mitochondria genes were excluded as their expression do not have information of
513 cell identity. The results shown in current paper were based the marker genes as described
514 above. But we also demonstrated that using the balanced highly variable genes yields
515 comparably accurate estimations (Supplementary Fig. 6).
516
517 **Estimate gene mean and dispersion per cell type**
518 Modeling single cell RNA-seq data is challenging due to the cellular heterogeneity, technical
519 sensitivity, and noise. While the expression of some genes can be not detected by chance, other
520 genes may be found to be highly dispersed. These factors can lead to excessive variability even
521 within the same cell type. AdRoit combats high noise and computational complexity by building
522 models with estimated mean and dispersion per cell type. This strategy reduced the data
523 complexity while preserve the cell type specific information.
524
525 Although typical analyses of RNA-seq data starts with normalization, Adroit does not do
526 normalization prior to the mean estimation. Performing a normalization across all cell types
527 forces every cell type to have the same amount of RNA transcripts, measured by the total
528 unique molecular identifier (UMI) counts per cell. However, different cell types can have

529 dramatically different amounts of transcripts. For example, the amount of RNA transcripts in
530 neuronal cells is about 10 times fold of that in glial cells. Thus, normalization can falsely alter
531 the relative abundance of cell types, misleading the estimation of cell type percentages. To
532 avoid this problem, AdRoit models the means using the raw UMI counts.
533
534 Studies have shown that UMI counts follows negative binomial distribution[52,53], we therefore fit
535 negative binomial distributions to single cells of each cell type and build the model based on
536 the estimated means and dispersions from the selected genes. More specifically, let X_{ik} be the
537 set of single cell UMI counts of gene $i \in 1,..,I$ for all cells in cell type $k \in 1,...,K$. I is the number
538 of selected genes, and K denotes number of cell types in the single cell reference. The
539 distribution of X_{ik} follows negative binomial distribution,

540 $$X_{ik} \sim NB(\lambda_{ik}, p_{ik}), \quad (1)$$

541 where λ_{ik} is the dispersion parameter of the gene i in cell type k, and p_{ik} is the success
542 probability, i.e., the probability of gene i in cell type k getting one UMI. The two parameters are
543 estimated by MLE. The likelihood function is

544 $$LH(\lambda_{ik}, p_{ik}|X_{ik}) = \prod_{i=1}^{n_k} f(X_{ik}|\lambda_{ik}, p_{ik}), \quad (2)$$

545 where n_k is the number of cells in cell type k, and f is the probability mass function of negative
546 binomial distribution. The MLE estimates are then given by

547 $$(\widehat{\lambda_{ik}}, \widehat{p_{ik}}) = \underset{\lambda_{ik}, p_{ik}}{arg\max}\, LH(\lambda_{ik}, p_{ik}|X_{ik}). \quad (3)$$

548 Once success probability and dispersion are estimated, the mean estimates can be computed
549 numerically according to the property of negative binomial distribution,

$$\mu_{ik} = \frac{\widehat{\lambda_{ik}} \cdot \widehat{p_{ik}}}{1-\widehat{p_{ik}}}, \quad (4)$$

$$\sigma_{ik}^2 = \frac{\widehat{\lambda_{ik}} \cdot \widehat{p_{ik}}}{(1-\widehat{p_{ik}})^2}. \quad (5)$$

Estimation using MLE has been readily coded in many R packages. We choose 'fitdist' function from 'fitdistrplus' package[54] for its fast computation speed and flexibility in selecting distributions. Estimations are done for each selected gene in each cell type, resulting in a $I \times K$ matrix of cell type means.

Cell type specificity of genes

Genes with cell-type specific expression patterns better represent cell types, thus are more important when be used for resolving cell type composition. In line with this property, AdRoit weights genes with high specificity more than less specific ones. Highly specific genes usually have consistently high expression and thus relatively low variance among cells within a cell type. To compute cell type specificity of a gene, we first identify the cell type in which the gene has the highest expression (i.e., most specifically expressed cell type), then defines the specificity of this gene as the mean-to-variance ratio within the cell type. A high ratio renders high weight to the gene in the model. We use the estimated means and variances from negative binomial fitting (μ_{ik} and σ_{ik}^2 in eq. 4 and 5). Let k' be the index of cell type that has the highest mean expression of gene i,

$$k' = \underset{k}{argmax}\{\mu_{ik} |\ k \in 1 \dots K\}, \quad (6)$$

then the cell type specificity weight for gene i, denoting w_i^S, is given by,

$$w_i^S = \frac{\mu_{ik'}}{\sigma_{ik'}^2}, \quad (7)$$

571 and it is computed for each gene in the set of selected genes.

572

573 **Cross-sample gene variability**

574 The variability of a gene contrasts how much stable a gene is across samples. The idea of
575 weighting genes based on variability across samples is first explored by Wang et al[22], where
576 variability was defined as the cross-sample variance. By weighting down the high variability
577 genes, the authors achieved a great advantage over the traditional unweighted method. Genes
578 with low cross-sample variability better represent the population, hence are more trust-worthy
579 to be used to learn the cell composition. AdRoit incorporates the same notion to weight the
580 importance of genes, however, defines the variability in a more sophisticated way. Similar as
581 we define the cell type specificity, AdRoit utilizes mean and variance, and computes variance-
582 to-mean ratio (VMR) to stand for cross-sample gene variability. But here the mean and variance
583 are computed across samples. The VMR is better scaled than the simple variance, and it can
584 avoid underweighting genes that has low expression, while circumvent overweighting genes
585 hugely dispersed.

586

587 In addition, AdRoit extends the method to fit the case where multiple samples are not
588 available. We proposed three ways to compute the VMR, depending on whether multi-sample
589 data is available. Typically, the compound transcriptome data to be deconvolved have multiple
590 samples. In bulk RNA-seq data, multiple samples are usually included to control for biological
591 variability. In spatial transcriptome data, the spatial dots can be seen as multiple samples.
592 Therefore, we first consider computing the cross-sample gene variability from compound

593 transcriptome data. In case multi-sample for compound data is not available, AdRoit utilizes the

594 single cell reference, and synthesizes compound samples by pooling all cells belonging to the

595 same sample. If multi-sample is not available for both data, AdRoit subsample single cells and

596 pool them to make pseudo samples. Let Y_{ij} denote the counts of sequences for gene i in

597 sample $j \in 1,\ldots,J$, then

598 $$Y_{ij} \sim NB(\lambda_{ij}, p_{ij}), \quad (8)$$

599 where λ_{ij} is the dispersion parameter of the gene i in sample j, and p_{ij} is the success

600 probability. Again, we use MLE to get the estimates $\widehat{\lambda_{ij}}$and $\widehat{p_{ij}}$, following which cross-sample

601 mean and variance can be numerically computed:

602 $$\mu_i^S = \frac{\widehat{\lambda_{ij}} \cdot \widehat{p_{ij}}}{1-\widehat{p_{ij}}}, \quad (9)$$

603 $$(\sigma_i^2)^S = \frac{\widehat{\lambda_{ij}} \cdot \widehat{p_{ij}}}{(1-\widehat{p_{ij}})^2}, \quad (10)$$

604 and cross-sample variability for gene i is then defined as

605 $$VMR_i = \frac{(\sigma_i^2)^S}{\mu_i^S} = \frac{1}{w_i^C}, \quad (11)$$

606 where w_i^C is later used in the model. The cross-sample variability weight is computed for each

607 gene in the set of selected genes.

608

609 **Gene-wise scaling factor to correct platform bias**

610 When linking the compound data to the single cell data, rescaling factor is often used to

611 account for the library size and platform difference. The existing methods adopt a single

612 rescaling factor for each unit of sample, i.e., all genes of a single sample are multiplied by the

613 same factor[22,23]. This operation is based on a strong assumption that the impact of platform

614 difference to every gene is the same and linearly scaled among different cell types, which is
615 hardly true. In addition, because estimates can be easily affected by outliers in linear model,
616 estimation of cell proportions can be steered away from the truth by extremely high expression
617 genes. Therefore, applying a uniform scaling factor to all gene is inappropriate.
618
619 To overcome this problem, AdRoit instead estimates gene-wise scaling factors via an adaptive
620 learning strategy and rescales each gene with its respective scaling factor. To proceed, we first
621 input the mean gene expression from the compound samples (μ_i^S in eq. 9) and the estimated
622 means of each cell type from the single cell data (μ_{ik} in eq. 4), then apply a traditional non-
623 negative least square regression (NNLS) to get a rough estimation of the proportions of each
624 cell type, denoting τ_k. For each gene, a predicted mean expression ($\sum_k^K \widehat{\tau_k}\, \mu_{ik}$ in eq. 13) is
625 computed as the weighted sum of the means of each cell type wherein the weights are the
626 roughly estimated proportions. The regression equation is given by,

627 $$\mu_i^S = A \cdot (\sum_k^K \tau_k\, \mu_{ik} + \varepsilon), \qquad 0 < \tau_k,\ \sum_k^K \tau_k = 1 \qquad (12)$$

628 where A is a constant to ensure τ_k's sum to 1 and ε is the error term. We use 'nnls' function in
629 the 'nnls' package[55] to estimate τ_k's. Next, we calculate the ratio between the mean expression
630 from compound samples and the predicted means, and define the gene-wise rescaling factor as
631 the logarithm of the ratio plus 1,

632 $$r_i = \log\left(\frac{\mu_i^S}{\sum_k^K \widehat{\tau_k}\mu_{ik}} + 1\right). \qquad (13)$$

633 Given the dispersion property of count data, the logarithm of the ratio is a more appropriate
634 statistic as it results in relatively stable scaling factors. The addition of 1 avoids taking logarithm
635 on zero. By multiplying the flexible gene-wise rescaling factor, the "outlier" genes will be

636 pushed toward the truth regression line direction, while the genes around the true regression

637 lines are less affected (Fig. 1b).

638

639 **Weighted and regularized model**

640 We next designed a model that incorporates all these factors to do the actual estimation of cell

641 type proportions. AdRoit builds upon non-negative least square regression model. It gives high

642 weights to the genes with high cell type specificity and low cross-sample variability. This was

643 done by optimizing a weighted sum of squared loss function L, where the weights consist of

644 two components (w_i^C in eq. 7, w_i^S in eq. 11). The gene-wise scaling factor tailored for each gene

645 effectively corrects the bias due to technology difference between compound sample and

646 single cell data ($r_i$in eq 13). In cases of complex tissues (e.g., neural tissues) where many highly

647 similar subtypes are common, closely related subtypes can have strong collinearity, leading to

648 overestimation of some cell types whilst underestimate or miss some others. AdRoit handles

649 this problem by including a L2 norm of the estimates as the regularization component. Denote

650 β_k as the unscaled coefficient for cell type k. For a compound transcriptome sample j, the loss

651 function is given by,

652 $$L_j(\beta_1, \dots, \beta_K | y_{ij}, w_i^C, w_i^S, r_i, \widehat{\mu_{ik}}) = \sum_i^I w_i^C \cdot w_i^S \cdot (y_{ij} - r_i \cdot \sum_k^K \beta_k \widehat{\mu_{ik}})^2 + \sum_k^K \beta_k^2. \quad (14)$$

653 Then the coefficient β_k can be estimated by minimizing the loss function with the constraint

654 $\beta_1, \dots, \beta_K > 0$,

655 $$\widehat{\beta_1}, \dots, \widehat{\beta_K} = \underset{\beta_1,\dots,\beta_K}{\operatorname{argmax}}_{\beta_1,\dots,\beta_K>0} L_j. \quad (15)$$

656 The estimation is done by a gradient projection method by Byrd et al[56]. We derive the gradient

657 function by taking partial derivative of the loss function with $w.r.t.\ \beta_k$,

$$G_k = \nabla_{\beta_k} L_j = -2\sum_i^I r_i \cdot \widehat{\mu_{ik}} \cdot w_i^C \cdot w_i^S \cdot \left(y_{ij} - r_i \cdot \sum_k^K \beta_k \widehat{\mu_{ik}}\right) + 2\beta_k. \qquad (16)$$

AdRoit uses the function 'optim' from the R package 'stats' to do the estimation[57], providing the loss function (eq. 15) and the gradient (eq. 16). To get the final estimates of cell type proportions, we rescale the coefficients β_k's to ensure a summation of 1,

$$\theta_k = \frac{\widehat{\beta_k}}{\sum_k^K \widehat{\beta_k}}. \qquad (17)$$

Each compound sample j is independently estimated by the model described above.

Simulation of bulk RNA-seq and spatial transcriptomics data

Bulk RNA-seq data used for benchmarking are synthesized by adding up the raw UMI reads per gene from all single cells of a sample regardless of cell types. Denote t_k as a cell in cell type k, and $t_k \in 1, \ldots, T_k$, where T_k is the number of cells in cell type k. Let Y_{ij}^B be the read count of gene i in a synthesized bulk sample j, and X_{ijt_k} be the UMI count of the gene, then

$$Y_{ij}^B = \sum_k^K \sum_{t_k}^{T_k} X_{ijt_k}.$$

The true proportion of cell type k is given by,

$$\theta_k^0 = \frac{T_k}{\sum_k^K T_k}.$$

To simulate spatial transcriptomic spots, we first sample 10 cells without replacement from each cell type and added them up, then mix them with designed proportions. For example, to simulate a spot with p_k percent of cell type k, the read count Y_{ij}^S of gene i in a spatial spot j is given by,

$$Y_{ij}^S = \sum_k^K p_k \sum_{n=1}^{10} X_{ikn},$$

679 where X_{iks} is UMI count of gene i in a sampled cell n of cell type k. For each mixing scheme, the

680 simulation is repeated 100 times.

681

682 **Evaluation statistics**

683 We compared the estimated cell type proportions with the ground truth by calculating 4

684 statistics. The mAD and RMSD are given by,

685 $$mAD = \frac{\sum_k^K |\theta_k - \theta_k^0|}{K},$$

686 $$RMSD = \frac{\sum_k^K (\theta_k - \theta_k^0)^2}{K}.$$

687 Pearson correlation coefficient is computed as,

688 $$\rho_p = \frac{\sum_k^K (\theta_k - \overline{\theta_k})(\theta_k^0 - \overline{\theta_k^0})}{\sqrt{\sum_k^K (\theta_k - \overline{\theta_k})^2}\sqrt{\sum_k^K (\theta_k^0 - \overline{\theta_k^0})^2}},$$

689 where $\overline{\theta_k}$ and $\overline{\theta_k^0}$ are means of the estimated proportions and true proportions, respectively.

690 Spearman correlation coefficient is given by,

691 $$\rho_s = \frac{\sum_k^K (r_k - \overline{r_k})(r_k^0 - \overline{r_k^0})}{\sqrt{\sum_k^K (r_k - \overline{r_k})^2}\sqrt{\sum_k^K (r_k^0 - \overline{r_k^0})^2}},$$

692 where r_k is the rank of θ_k.

693

694 **Single cell RNA sequencing of mouse dorsal root ganglion**

695 As described previously[58], lumbar DRGs were isolated from adult C57BL/6 mice and transferred

696 to a dissociation buffer (Dulbecco's modified Eagle's medium supplemented with 10% heat-

697 inactivated Fetal Calf Serum) (Gibco; cat # A38400-02). To generate a single cell suspension,

698 DRGs were subjected to a 2 step-enzymatic dissociation followed by a mechanical dissociation.

699 In brief, DRGs were first incubated with 0.125% collagenase P from Clostridium histolyticum
700 (Roche Applied Science; cat # 11249002001) for 90 minutes in an Eppendorf Thermomixer C
701 (37°C; intermittent 750 rpm shaking for about 10 sec every 2 minutes). Then, DRGs were
702 transferred to a Hank's Balanced Salt Solution (HBSS, Mg^{2+} and Ca^{2+} free; Invitrogen)
703 supplemented with 0.25% Trypsin (Worthington biochemical corp.; cat # LSoo3707) and
704 0.0025% EDTA and incubated for 10 minutes at 37°C in the Eppendorf Thermomixer C. Trypsin
705 was neutralized by the addition of 2.5 mg/ml MgSO4 (Sigma; cat #M-3937) and DRGs were
706 triturated with Pasteur pipettes. The resulting cell suspension was passed through a 70 µm
707 mesh filter to remove remaining chunks of tissues and centrifuged for 5 minutes at 2500 rpm at
708 room temperature. The pellet was resuspended in HBSS ($Ca^{2+,}$ Mg^{2+} free; Invitrogen) and the
709 cell suspension was run on a 30% Percoll Plus gradient (Sigma GE17-5445-02) to further remove
710 debris. Finally, cells were resuspended in PBS supplemented with 0.04% BSA at a concentration
711 of 200 cells/µl and cell viability was determined using the automated cell analyzer
712 NucleoCounter® NC-250™. The suspended single cells were loaded on a Chromium Single Cell
713 Instrument (10X Genomics) with about 6000 cells per lane to minimize the presence of
714 doublets. 2000-3000 cells per lane were recovered. RNA-seq libraries were constructed using
715 Chromium Single Cell 3' Library, Gel Beads & Multiplex Kit (10X Genomics). Single end
716 sequencing was performed on Illumina NextSeq500. Read 1 starts with a 26-bp UMI and cell
717 barcode, followed by an 8-bp i7 sample index. Read 2 contains a 55-bp transcript read. Sample
718 de-multiplexing, alignment, filtering, and UMI counting were conducted using Cell Ranger
719 Single-Cell Software Suite[59] (10X Genomics, v2.0.0). Mouse mm10 Genome assembly and UCSC
720 gene model were used for the alignment.

721

722 **Data preprocessing**

723 *DRG single cell data*

724 The UMI data output from Cell Ranger Single-Cell Software Suite (10X Genomics, v2.0.0) was
725 analyzed using Seurat package[60] to assess the cell quality and identify cell types, similar to what
726 described previously[39]. Cells with the number of detected genes less than 500 or over 15000, or
727 with a UMI ratio of mitochondria encoded genes versus all genes over 0.1 were also removed.
728 The UMI data was normalized by the 'NormalizeData' method in Seurat with default settings.
729 To avoid potential sample-to-sample variation caused by technical variation at various
730 experiment steps, we employed Seurat data integration method. The top 2000 variable genes
731 of each of the 5 samples were identified using 'FindVariableFeatures' with
732 selection.method='vst'. Based on the union of these variable genes, the anchor cells in each
733 sample were identified by 'FindIntegrationAnchors'. All the samples were then integrated by
734 'IntegrateData'. We subsequently scaled the integrated data ('ScaleData') and performed
735 dimension reduction ('RunPCA'). Cells were then clustered based on the first 15 principal
736 components by applying 'FindNeighbors' and 'FindClusters' (resolution=0.6, algorithm=1).
737 Marker genes for each cluster were identified using 'FindAllMarkers'. Parameters were used
738 such that these genes were expressed in at least 25% of the cells in the cluster, and on average
739 2-fold higher than the rest of cells with a multiple-testing adjusted Wilcoxon test p value of less
740 than 0.01. The specificity of the canonical cell type-specific genes or cell cluster-specific genes
741 were further examined by visualizations (Extended Data Fig. 2) and used to define the cell type
742 for each cluster. At the end, the original UMI data from 17271 genes and 3352 cells that passed

743 the quality control were organized into a matrix (genes as rows and cell identifiers as columns).
744 This matrix, together with the cell type label for each cell therein, were loaded into AdRoit as
745 reference profiles.
746
747 *Mouse brain single cell data*
748 The scRNA-seq reference data of the mouse brain were obtained from Zeisel et. al[32]. Among all
749 the available data, we only retained 96,572 cells that were acquired from the brain regions, had
750 an assigned cell type by the authors and a minimal total UMI of 1000. These cells corresponded
751 to 183 clusters at the finest taxonomy level in the original study. As many of the clusters are
752 highly similar, we decided to merge some of them to simplify the reference landscape. First, the
753 top 50 cluster enriched markers were derived using Scanpy[61] via the 'rank_genes_groups'
754 function (method='wilcoxon'), following the normalization ('normalize_per_cell'), log
755 transformation ('log1p') and regressing out ('regress_out') the variances associated with the
756 total UMI and the percentage of mitochondrial chromosome encoded genes per cell. Then, the
757 pair-wise overlapping p-values among the clusters were calculated using the top 50 marker
758 genes assuming the hypergeometric null distribution. Last, clusters with overlapping p-values
759 more significant than 1e-10 were merged and new names were assigned by combinedly
760 considering the original annotation, the molecular features and the specificity to certain brain
761 regions. A total of 46 cell types were determined that cover all the 12 brain regions and their
762 important substructures[37] (Supplementary Table 13). To make the reference dataset more
763 manageable in size and more balanced in the representation of cell types, we down sampled

764 each cluster to no more than 360 cells. A final set of 14,666 cells over 46 cell types were used
765 for the deconvolution of the mouse brain spatial transcriptome data.
766
767 *Human Islets*
768 We used the 1492 high quality human islets single cell and annotation from Xin et al[38]. The
769 RPKM expression table was directly downloaded and used as is. The RNA-FISH data was also
770 from this study[38]. For the real bulk human pancreatic islets data[38,40,41], the read counts table
771 were deconvoluted. Only data from donors with HbA1C level available were included in the
772 regression of Beta cell proportion on HbA1C level (Fig. 4c, Supplementary Table 10).
773
774 *Trabecular Meshwork*
775 We downloaded the raw sequence data and followed the same analysis procedure as in Patel et
776 al[39] for quality control and cell type identification.
777
778 *Mouse Brain Spatial transcriptomics data by 10x Visium platform*
779 The filtered cell matrix, tissue image and the spatial coordinates of a coronal section of an adult
780 C57BL/6 mouse brain from the 10x Genomics were available for download and used as is.
781
782 *Mouse Brian ISH images*
783 The ISH images were directly downloaded from Allen mouse Brain Atlas[37] by searching the gene
784 names. THE images were used with further editing except for cropping.
785

786 **Data availability**

787 DRG single cell data are deposited at NCBI GEO (accession number: GSE163252) . The bulk RNA-
788 seq and RNA-FISH data for human pancreatic islets were initially published as aggregated data
789 where the data processing and experimental procedure were described therein[38,40,41]. We
790 acquired the individual sample data from the authors and released them along with the current
791 study (Supplementary Table 10 and Supplementary Table 12). The other public data analyzed in
792 this study are available from: GEO (human pancreatic islets single cell data: GSE81608); NCBI
793 (human trabecular meshwork single cell data: PRJNA616025; mouse brain single cell data:
794 SRP135960). Mouse brain spatial transcriptomic data was downloaded from the 10x Genomics
795 website (https://support.10xgenomics.com/spatial-gene-
796 expression/datasets/1.1.0/V1_Adult_Mouse_Brain_Coronal_Section).

797

798 **Code availability**

799 AdRoit's source code is available on Github (https://github.com/TaoYang-dev/AdRoit).

800

801 **Software**

802 The statistical analyses were done with R statistical software (v3.6.0)[57] and python (v3.7.2)[62].
803 The packages used include Seurat (v3.0.1)[60], scanpy (v1.6.0)[61], dplyr (v0.8.0.1)[63], doParallel
804 (v1.0.14)[64], data.table (v1.12.4)[65], fitdistrplus (v1.1-1)[54], nnls (v1.4)[55].

805

806 **Reference**

807 1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics.
808 *Nature Reviews Genetics* (2009) doi:10.1038/nrg2484.
809 2. Chu, G. C., Kimmelman, A. C., Hezel, A. F. & DePinho, R. A. Stromal biology of pancreatic
810 cancer. *Journal of Cellular Biochemistry* (2007) doi:10.1002/jcb.21209.
811 3. Bussard, K. M., Mutkus, L., Stumpf, K., Gomez-Manzano, C. & Marini, F. C. Tumor-
812 associated stromal cells as key contributors to the tumor microenvironment. *Breast*
813 *Cancer Research* (2016) doi:10.1186/s13058-016-0740-2.
814 4. Munn, D. H. & Bronte, V. Immune suppressive mechanisms in the tumor
815 microenvironment. *Current Opinion in Immunology* (2016)
816 doi:10.1016/j.coi.2015.10.009.
817 5. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: From tumor
818 initiation to metastatic progression. *Genes and Development* (2018)
819 doi:10.1101/GAD.314617.118.
820 6. Garner, H. & de Visser, K. E. Immune crosstalk in cancer progression and metastatic
821 spread: a complex conversation. *Nature Reviews Immunology* (2020)
822 doi:10.1038/s41577-019-0271-z.
823 7. Singh, U. P. *et al.* Chemokine and cytokine levels in inflammatory bowel disease patients.
824 *Cytokine* (2016) doi:10.1016/j.cyto.2015.10.008.
825 8. Van Lint, P. & Libert, C. Chemokine and cytokine processing by matrix metalloproteinases
826 and its effect on leukocyte migration and inflammation. *J. Leukoc. Biol.* (2007)
827 doi:10.1189/jlb.0607338.
828 9. Zelová, H. & Hošek, J. TNF-α signalling and inflammation: Interactions between old

829 acquaintances. *Inflammation Research* (2013) doi:10.1007/s00011-013-0633-0.

830 10. Koelman, L., Pivovarova-Ramich, O., Pfeiffer, A. F. H., Grune, T. & Aleksandrova, K.
831 Cytokines for evaluation of chronic inflammatory status in ageing research: Reliability
832 and phenotypic characterisation. *Immun. Ageing* (2019) doi:10.1186/s12979-019-0151-1.

833 11. Landskron, G., De La Fuente, M., Thuwajit, P., Thuwajit, C. & Hermoso, M. A. Chronic
834 inflammation and cytokines in the tumor microenvironment. *Journal of Immunology*
835 *Research* (2014) doi:10.1155/2014/149185.

836 12. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial
837 transcriptomics. *Science* (2016) doi:10.1126/science.aaf2403.

838 13. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat.*
839 *Methods* (2019) doi:10.1038/s41592-019-0548-y.

840 14. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*
841 (2009) doi:10.1038/nmeth.1315.

842 15. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in
843 single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* (2020)
844 doi:10.1186/s13059-020-02048-6.

845 16. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental considerations
846 for single-cell RNA sequencing approaches. *Frontiers in Cell and Developmental Biology*
847 (2018) doi:10.3389/fcell.2018.00108.

848 17. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism.
849 *Nature* (2017) doi:10.1038/nature21350.

850 18. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of

851 blood microarray data identifies cellular activation patterns in systemic lupus
852 erythematosus. *PLoS One* (2009) doi:10.1371/journal.pone.0006098.
853 19. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
854 *Nat. Methods* (2015) doi:10.1038/nmeth.3337.
855 20. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas
856 Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* (2016)
857 doi:10.1016/j.cels.2016.08.011.
858 21. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression
859 data. *Nat. Commun.* (2019) doi:10.1038/s41467-019-10802-z.
860 22. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution
861 with multi-subject single-cell expression reference. *Nat. Commun.* (2019)
862 doi:10.1038/s41467-018-08023-x.
863 23. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference
864 of cell type topography. *Commun. Biol.* **3**, 565 (2020).
865 24. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues
866 with digital cytometry. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0114-2.
867 25. Myung, I. J. Tutorial on maximum likelihood estimation. *J. Math. Psychol.* (2003)
868 doi:10.1016/S0022-2496(02)00028-7.
869 26. Bassett, R. & Deride, J. Maximum a posteriori estimators as a limit of Bayes estimators.
870 *Math. Program.* (2019) doi:10.1007/s10107-018-1241-0.
871 27. Zhao, Y. & Simon, R. Gene expression deconvolution in clinical samples. *Genome*
872 *Medicine* (2010) doi:10.1186/gm214.

873 28. Chiu, Y. J., Hsieh, Y. H. & Huang, Y. H. Improved cell composition deconvolution method
874 of bulk gene expression profiles to quantify subsets of immune cells. *BMC Med.*
875 *Genomics* (2019) doi:10.1186/s12920-019-0613-5.
876 29. Kang, K. *et al.* CDSeq: A novel complete deconvolution method for dissecting
877 heterogeneous samples using gene expression data. *PLoS Comput. Biol.* (2019)
878 doi:10.1371/journal.pcbi.1007510.
879 30. Qiao, W. *et al.* PERT: A Method for Expression Deconvolution of Human Blood Samples
880 from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.*
881 (2012) doi:10.1371/journal.pcbi.1002838.
882 31. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of
883 cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* (2019)
884 doi:10.1038/s41467-019-09990-5.
885 32. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* (2018)
886 doi:10.1016/j.cell.2018.06.021.
887 33. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular
888 deconvolution of GTEx tissues powers discovery of disease and cell-type associated
889 regulatory variants. *Nat. Commun.* (2020) doi:10.1038/s41467-020-14561-0.
890 34. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA
891 sequencing protocols. *F1000Research* (2017) doi:10.12688/f1000research.11290.1.
892 35. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational
893 data analysis. *Frontiers in Genetics* (2019) doi:10.3389/fgene.2019.00317.
894 36. Chen, D. & Plemmons, R. J. Nonnegativity constraints in numerical analysis. in *The Birth*

895 *of Numerical Analysis* (2009). doi:10.1142/9789812836267_0008.

896 37. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature*

897 (2007) doi:10.1038/nature05453.

898 38. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes.

899 *Cell Metab.* (2016) doi:10.1016/j.cmet.2016.08.018.

900 39. Patel, G. *et al.* Molecular taxonomy of human ocular outflow tissues defined by single-

901 cell transcriptomics. *Proc. Natl. Acad. Sci.* **117**, 12856 LP – 12867 (2020).

902 40. Xin, Y. *et al.* Pseudotime ordering of single human B-cells reveals states of insulin

903 production and unfolded protein response. *Diabetes* (2018) doi:10.2337/db18-0365.

904 41. Gutierrez, G. D. *et al.* Gene signature of proliferating human pancreatic a cells.

905 *Endocrinology* (2018) doi:10.1210/en.2018-00469.

906 42. Cerf, M. E. Beta cell dysfunction and insulin resistance. *Frontiers in Endocrinology* (2013)

907 doi:10.3389/fendo.2013.00037.

908 43. Maedler, K. & Donath, M. Y. Beta-cells in type 2 diabetes: a loss of function and mass.

909 *Hormone research* (2004).

910 44. Donath, M. Y. *et al.* Mechanisms of β-cell death in type 2 diabetes. *Diabetes* (2005)

911 doi:10.2337/diabetes.54.suppl_2.S108.

912 45. Calanna, S. *et al.* Alpha- and beta-cell abnormalities in haemoglobin A1c-defined

913 prediabetes and type 2 diabetes. *Acta Diabetol.* (2014) doi:10.1007/s00592-014-0555-5.

914 46. Kanat, M. *et al.* The Relationship Between β-Cell Function and Glycated Hemoglobin.

915 *Diabetes Care* **34**, 1006 LP – 1010 (2011).

916 47. Nepton, S. Beta-Cell Function and Failure. in *Type 1 Diabetes* (2013). doi:10.5772/52153.

48. Dolenšek, J., Rupnik, M. S. & Stožer, A. Structural similarities and differences between the human and the mouse pancreas. *Islets* (2015) doi:10.1080/19382014.2015.1024405.

49. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).

50. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* (2019) doi:10.1038/s41467-019-12266-7.

51. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* (2010) doi:10.1186/gb-2010-11-10-r106.

52. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* (2019) doi:10.1186/s13059-019-1874-1.

53. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* (2020) doi:10.1038/s41587-019-0379-5.

54. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R package for fitting distributions. *J. Stat. Softw.* (2015) doi:10.18637/jss.v064.i04.

55. Mullen, Katharine M., I. H. M. van S. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). *R Packag. version 1.4* (2012).

56. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* (1995) doi:10.1137/0916069.

57. The R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2019).

939 58. Alessandri-Haber, N. *et al.* Hypotonicity induces TRPV4-mediated nociception in rat.
940 *Neuron* (2003) doi:10.1016/S0896-6273(03)00462-8.
941 59. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat.*
942 *Commun.* (2017) doi:10.1038/ncomms14049.
943 60. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019)
944 doi:10.1016/j.cell.2019.05.031.
945 61. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data
946 analysis. *Genome Biol.* (2018) doi:10.1186/s13059-017-1382-0.
947 62. van Rossum, G. & Drake, F. L. *Python 3 Reference Manual. Scotts Valley, CA* (2009).
948 63. Wickham, H. & Francois, R. dplyr: A Grammar of Data Manipulation. *R Packag. version*
949 *0.4.2.* (2015).
950 64. Weston, S., Calaway, R. & Tenenbaum, D. *doParallel: Foreach Parallel Adaptor for the*
951 *Parallel Package*. *Cran* (2014).
952 65. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. R Package Version
953 1.12.8. *Manual* (2019).
954

955 **Acknowledgements**

956 We thank Yurong Xin for pointing us to the relevant public data resource. We also thank Gabor
957 Halasz and Yuan Zhu for the advice to algorithm design.
958

959 **Author contributions**

960 T.Y., Y.B., W.F., N.A.-H., M.L.-F., L.E.M. and G.S.A. designed the research. T.Y., Y.B., and W.F.
961 developed the algorithm. T.Y., Y.B., W.F. and J.K. participated in the data analyzing. M.S. and
962 R.B. performed the DRG tissue collection. C.A. performed the single cell library preparation and
963 sequencing experiment. T.Y., Y.B., N.A.-H. and G.S.A. wrote the manuscript.
964

965 ## Competing interests

966 T.Y., Y.B., W.F. and G.S.A. have filed a patent application relating to the AdRoit computational
967 framework. M.L.-F. is an employee of Cellular Longevity. All other authors are employees and
968 shareholders of Regeneron Pharmaceuticals, although the manuscript's subject matter does not
969 have any relationship to any products or services of this corporation.
970

971 ## Figure legends

972 **Fig. 1: Schematic representation of AdRoit computational framework. a,** AdRoit inputs bulk or
973 spatial RNA-seq data, single cell RNA-seq data and cell type annotations. It first selects
974 informative genes and estimates their means and dispersions, based on which the cell type
975 specificity of genes is computed. Depending on multi-sample availability, cross-sample gene
976 variability is estimated from compound data, or single cell samples (dashed arrow). Lastly the
977 gene-wise scaling factors are estimated using both compound data and single cell data. These
978 computed quantities are fed to a weighted regularized model to infer the transcriptome
979 composition. **b,** A mock example to illustrate the role of gene-wise scaling factor. Ideally, an
980 accurate estimation of slop (i.e., cell proportion) would be the slope of the green line, however
981 direct fitting would result in the red line due to the impact of the outlier genes. Outlier genes

982 can be induced due to platform difference affecting genes differently. AdRoit adopts an
983 adaptive learning approach that first learns a rough estimation of the slop (red line), then
984 moves the outlier genes toward it such that the more deviated genes will be moved more
985 toward the true line (i.e., longer arrows). After the adjustment, the new estimated slop (blue
986 line) is closer to the truth (green line), thus is a more accurate estimation.
987
988 **Fig.2: Benchmark on simulated bulk data synthesized from trabecular meshwork (TM) single**
989 **cells data. a,** AdRoit has the closest estimation to the true cell proportion comparing to MuSiC
990 and NNLS. Each dot is a cell type from one donor. **b,** For each cell type in TM, AdRoit has the
991 smallest differences from the true cell type proportion and the smallest variance of estimates
992 across the 8 donors. For each cell type, a dot on the graph denotes a donor, and the bars
993 represent the $1.5 \times$ interquartile ranges. Estimation was done by using the single cell as
994 reference leaving out the donor used for synthesizing bulk. **c,** AdRoit's estimates are more
995 accurate and specific than MuSiC's estimates on synthetic bulk that contains partial cell types.
996 The synthetic bulk was simulated by using only 6 out of the 12 cell types per donor, then
997 estimated with the reference of 12 cell types. AdRoit has notably fewer false positive estimates
998 of the 6 cell types not included, and more accurate estimation of the 6 cell types used for
999 synthesizing bulk. **d,** Receiver operating characteristic (ROC) curve shows AdRoit has a
1000 significantly higher AUC than MuSiC (0.95 vs 0.74), meaning better sensitivity and specificity.
1001
1002 **Fig. 3: Benchmark on scRNA-seq data from dorsal root ganglion (DRG) where these exist many**
1003 **closely related subtypes of neuronal cells. a,** 14 cell types were identified from scRNA-seq

1004 samples of 5 mice, including multiple subtypes of neurofilaments (NF), peptidergic (PEP) and
1005 non-peptidergic (NP) neurons. **b,** Benchmarking with the synthetic data shows AdRoit's
1006 estimation of cell type proportions are highly accurate. In particular, AdRoit achieves
1007 reasonably high accuracy when the cells are rare (e.g., < 5%). Each dot represents a cell type
1008 from one sample. **c,** For each individual sample, mAD, RMSD, Pearson and Spearman
1009 correlations were computed and compared across three methods. AdRoit has the lowest mAD
1010 and RMSD, and highest Pearson and Spearman correlations. In addition, AdRoit's estimation is
1011 also the most stable across samples. Each dot on the boxplot is a sample. Estimation was done
1012 by using the single cell reference leaving out the sample used for synthesizing bulk.
1013
1014 **Fig. 4: AdRoit is more accurate and sensitive than Stereoscope on spatial spots simulated**
1015 **from real DRG cells. a,** AdRoit and Stereoscope estimations on simulated spatial spots that
1016 contains 5 PEP neuron subtypes. True mixing proportions were denoted by the red dashed
1017 lines. Three schemes were simulated: 1) the proportions of 5 PEP cell types are the same and
1018 equal to 0.2; 2) PEP1_Dcn is 0.1 and the other 4 are 0.225; 3) PEP1_Dcn and
1019 PEP1_S100a11.Tagln2 are 0.1, PEP1_Slc7a3.Sstr2 and PEP2_Htr3a.Sema5a 0.2 are 0.2, and
1020 PEP3_Trpm8 is 0.4. In all simulation schemes, AdRoit's estimates are more consistently
1021 centered around the true proportions than Stereoscope's estimates. **b,** AdRoit is more accurate
1022 in estimating rare cells in spatial spots. The spots were simulated by simulating mixtures of 3
1023 PEP cell types (i.e., PEP1_Slc7a3.Sstr2, PEP2_Htr3a.Sema5a and PEP3_Trpm8), with a series of
1024 low percent of PEP3_Trpm8 cell type from 1% to 10% and the other two cell types sharing the
1025 rest proportion equally. AdRoit's estimates are systematically closer to the true simulated

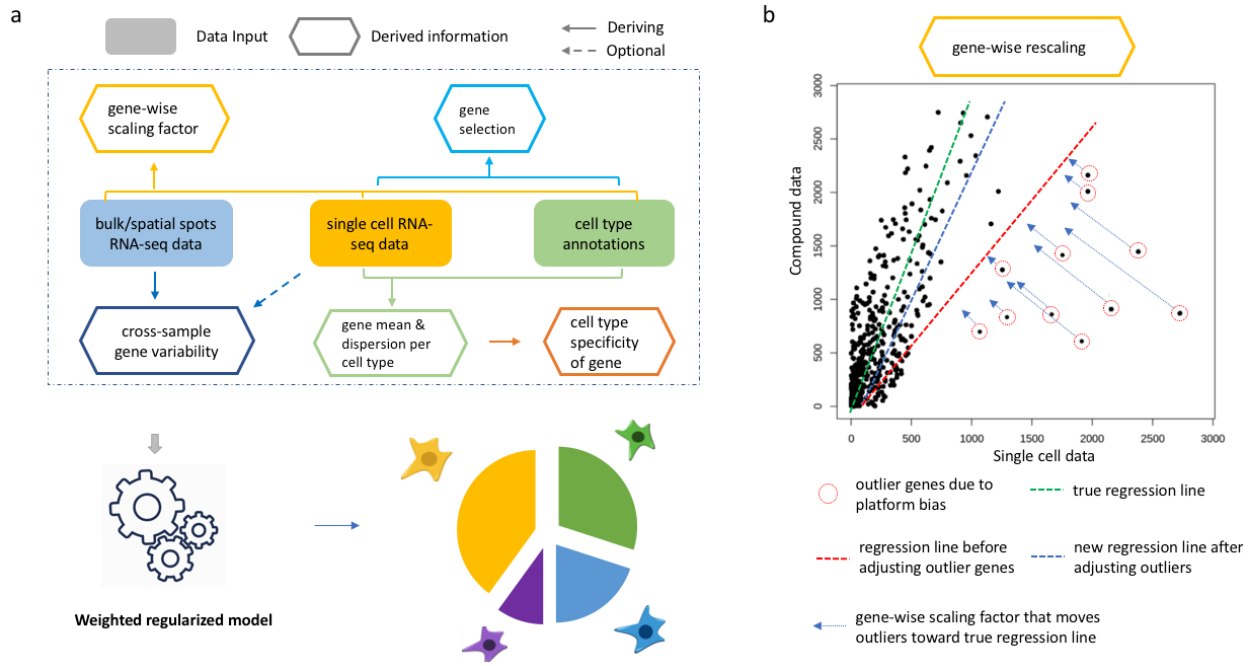
1026 proportions than Stereoscope's estimates. **c,** AdRoit is consistently more sensitive than
1027 Stereoscope in detecting low percent cells (estimates > 0.5% deemed as detected) in simulated
1028 spots of 1) low percent of NF_Calb1 mixed with NF_Pvalb and NF2_Ntrk2.Necab2, 2) low
1029 percent of NP_Mrgpra3 mixed with NP_Mrgprd and NP_Nts, 3) low percent of PEP3_Trpm8
1030 mixed with PEP1_Slc7a3.Sstr2 and PEP2_Htr3a.Sema5a, 4) low percent of NF_Calb1 mixed with
1031 Th, satellite glia and endothelial, 5) low percent of NP_Mrgpra3 mixed with Th, satellite glia and
1032 endothelial, and 6) low percent of PEP_Trpm8 mixed with Th, satellite glia and endothelial.
1033
1034 **Fig. 5: Applications to real bulk human islets RNA-seq data and mouse brain spatial**
1035 **transcriptome data. a,** AdRoit's estimates on real human Islets bulk RNA-seq data were highly
1036 reproducible for the repeated samples from same donor. **b,** AdRoit estimated cell type
1037 proportions agreed with the RNA-FISH measurements. **c,** AdRoit estimated Beta cell
1038 proportions in type 2 diabetes patients are significantly lower than that in healthy subjects. In
1039 addition, the estimated proportions have a significant negative linear association with donors'
1040 HbA1C level. **d,** The spatial mapping of 4 mouse brain cell types is consistent with the ISH
1041 images of 4 marker genes from Allen mouse brain atlas[37] respectively. The 4 genes, Spink8
1042 (marker of hippocampal field CA1), C1ql2 (marker of Dentate Gyrus), Clic6 (marker of Choroid
1043 Plexus), Synpo2 (marker of Thalamus) were identified as markers of corresponding tissues by
1044 Zeisel et al[32].
1045
1046 **Extended Data Fig. 1: Benchmark three methods on human pancreatic islets data. a,** Human
1047 islets single cell data contains 4 cell types from 18 subjects including two major cell types Alpha

1048 and Beta cells, and two minor cells PP and Delta cells[38]. The cell proportion varies across
1049 different subjects. **b, c,** AdRoit achieves leading accuracy when applied to the bulk data
1050 synthesized from the single cell data. Each dot on scatterplot is a cell type from one subject.
1051 Estimation was done by using the single cell reference leaving out the subject used to
1052 synthesize bulk.
1053
1054 **Extended Data Fig. 2: Dorsal root ganglion single cell shows 14 cell types including 3 subtypes**
1055 **of neurofilament, 3 subtypes of non-peptidergic neurons, and 5 subtypes of peptidergic**
1056 **neurons. a,** Heatmap of top markers shows distinction between cell types as well as similarity
1057 between subtypes. **b,** The proportion of each cell type varies from 0.5% to 33.71% across
1058 different samples.
1059
1060 **Extended Data Fig. 3: Comparing the performance on estimated simulated spatial spots of 14**
1061 **pure cell type respectively. a,** Estimates by AdRoit and **b,** estimates by Stereoscope are
1062 comparably accurate. Simulations were done by sampling cells from the same cell type and
1063 adding up the read counts per gene. For each of the 14 cell types of the DRG tissue, we
1064 repeated the simulation 100 times. The results shown were a summary of 100 simulations for
1065 each cell type. For both methods, the median estimates of the sampled cell type were close to
1066 1 (red lines), whereas the cell type not sampled has zero or close-to-zero values.
1067
1068 **Extended Data Fig. 4: The comparison of AdRoit and Stereoscope on the simulated spots of**
1069 **additional cell mixing schemes.** 5 more types of mixed spatial spots were simulated: 1) mixture

1070 of 3 neurofilaments (NF); 2) mixture of 3 non-peptidergic (NP) cell types; 3) NF2_Ntrk2.Necab2
1071 mixing with Th, satellite glia and endothelial; 4) NP_Nts mixing with Th, satellite glia and
1072 endothelial; and 5) PEP3_Trpm8 mixing with Th, satellite glia and endothelial. Each simulation
1073 was repeated 100 times. Consistently for all simulation schemes, AdRoit's estimates were
1074 always closer to the true simulated proportions (red lines), whereas Stereoscope's estimates
1075 largely deviated from the true proportions.
1076
1077 **Extended Data Fig. 5: Spatial mapping of 46 cell types with AdRoit quantitative depicts the**
1078 **content in each spot.** Spatial transcriptomics data was downloaded from 10x genomics
1079 (https://support.10xgenomics.com/spatial-gene-
1080 expression/datasets/1.1.0/V1_Adult_Mouse_Brain_Coronal_Section). The reference single cells
1081 were sampled from Zeisel et al[32] and curated into 46 cell types.
1082
1083
1084
1085
1086
1087
1088
1089
1090

1091 **Figures**

1092 **Fig. 1**



1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

Fig. 2

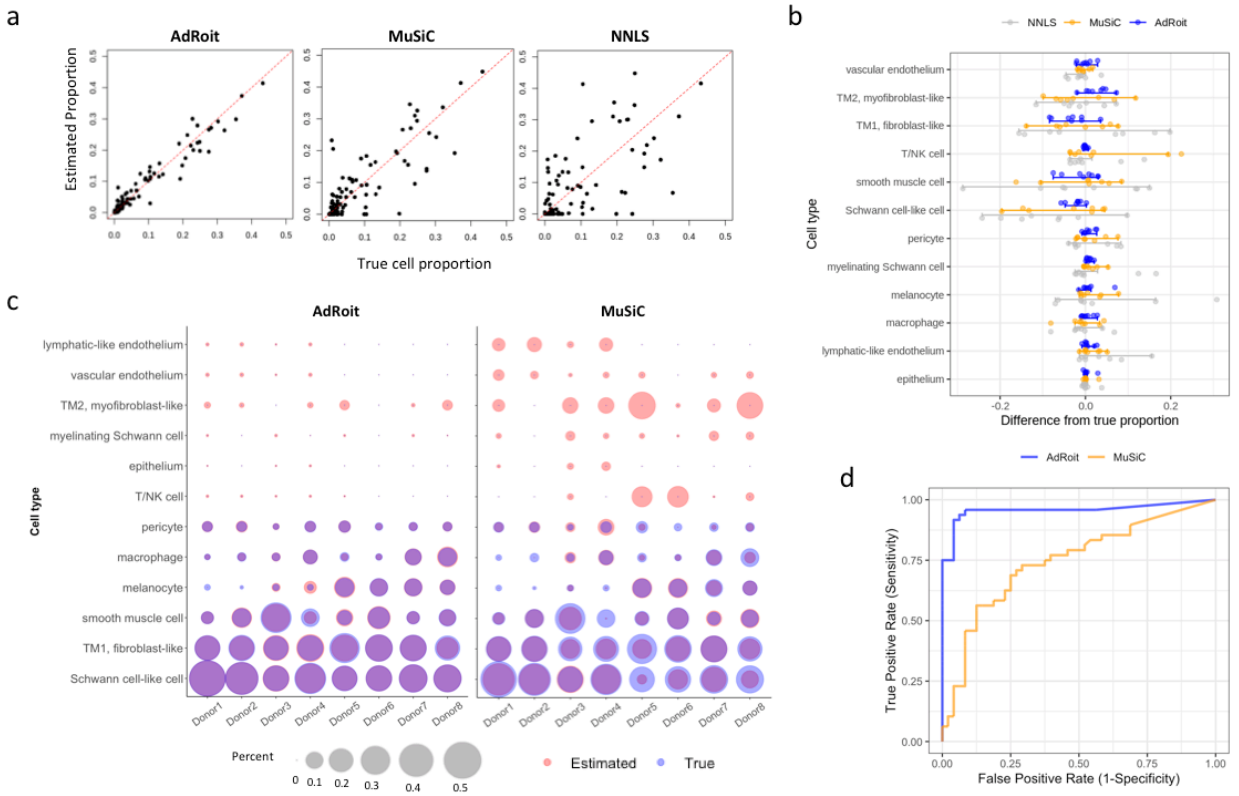

a
AdRoit
MuSiC
NNLS
Estimated Proportion
True cell proportion
b
NNLS
MuSiC
AdRoit
vascular endothelium
TM2, myofibroblast-like
TM1, fibroblast-like
T/NK cell
smooth muscle cell
Schwann cell-like cell
pericyte
myelinating Schwann cell
melanocyte
macrophage
lymphatic-like endothelium
epithelium
Cell type
Difference from true proportion
c
AdRoit
MuSiC
Cell type
Donor1
Donor2
Donor3
Donor4
Donor5
Donor6
Donor7
Donor8
Percent
Estimated
True
d
AdRoit
MuSiC
True Positive Rate (Sensitivity)
False Positive Rate (1-Specificity)

Fig. 3

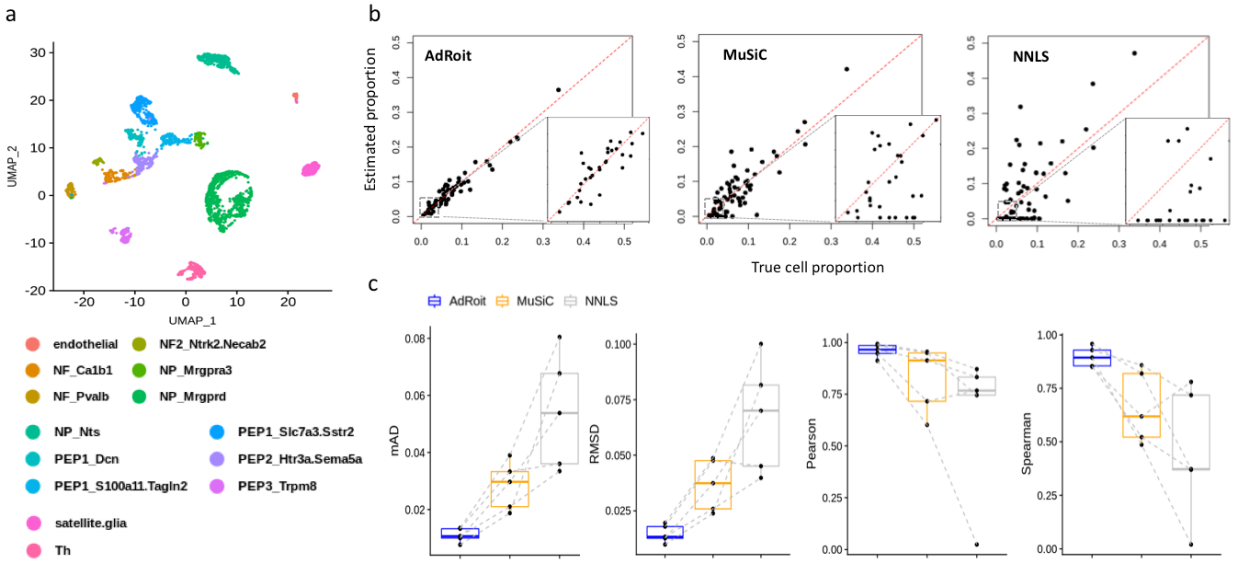

a
UMAP_2
UMAP_1
endothelial
NF2_Ntrk2.Necab2
NF_Ca1b1
NP_Mrgpra3
NF_Pvalb
NP_Mrgprd
NP_Nts
PEP1_Slc7a3.Sstr2
PEP1_Dcn
PEP2_Htr3a.Sema5a
PEP1_S100a11.Tagln2
PEP3_Trpm8
satellite.glia
Th
b
AdRoit
MuSiC
NNLS
Estimated proportion
True cell proportion
c
AdRoit
MuSiC
NNLS
mAD
RMSD
Pearson
Spearman

Fig. 4

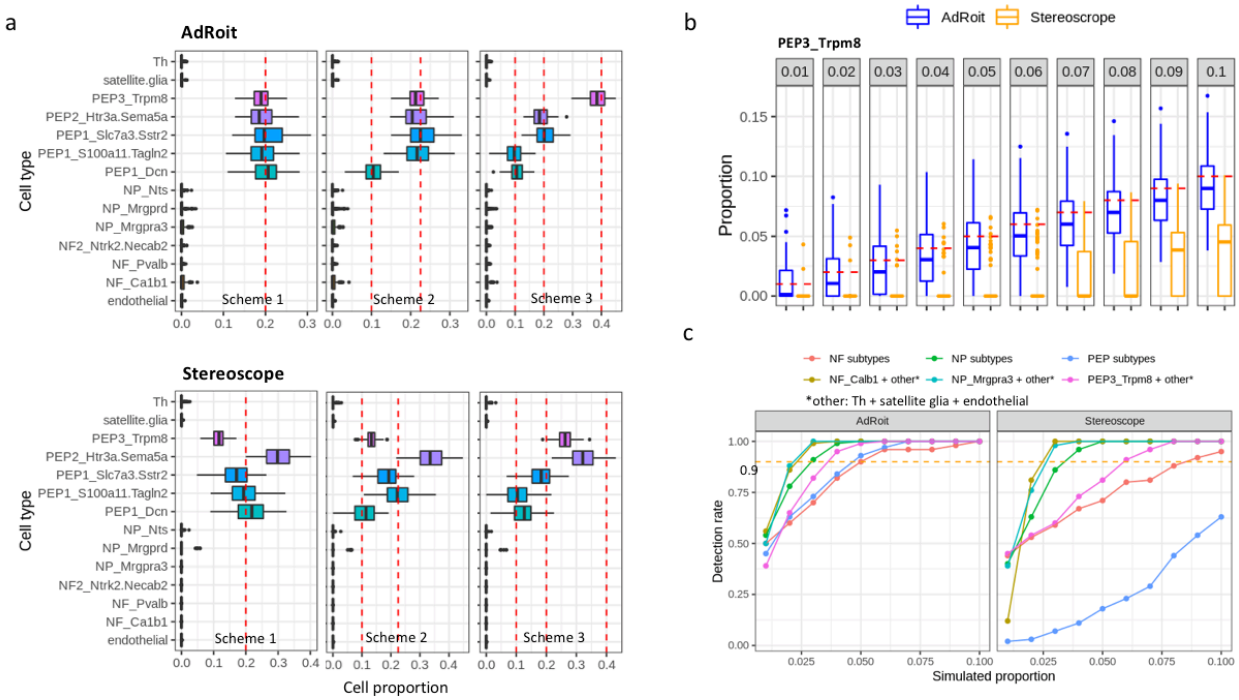
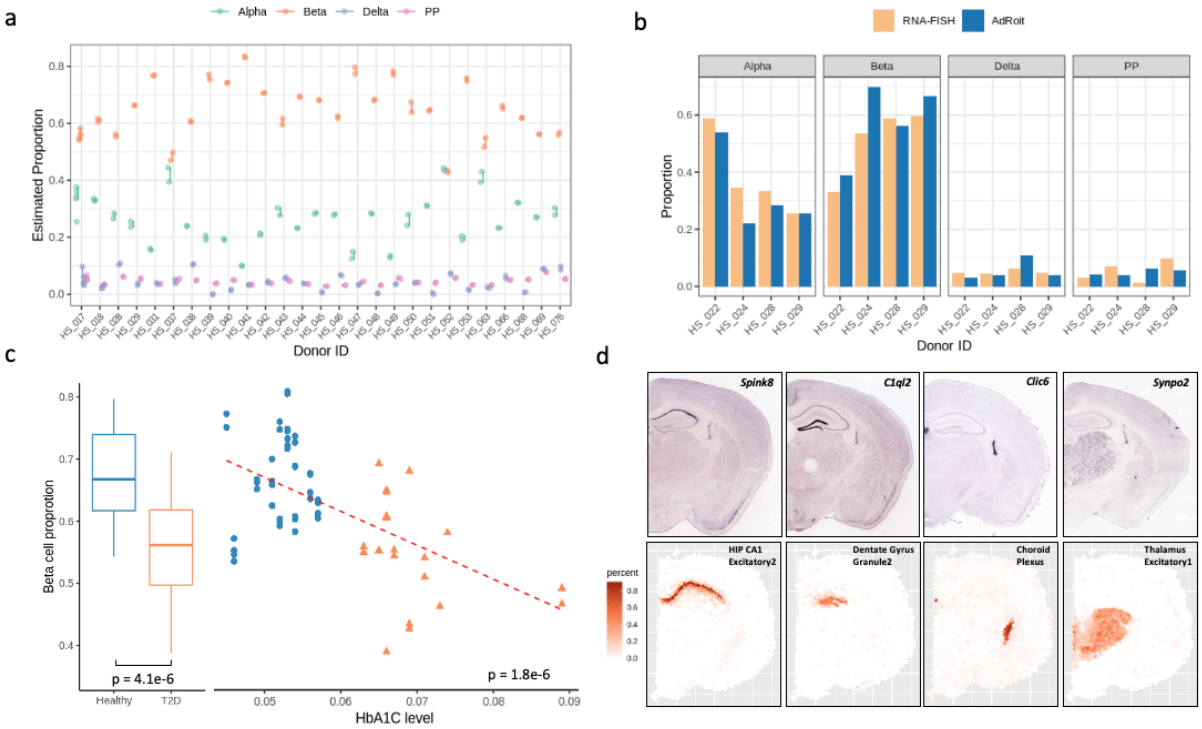
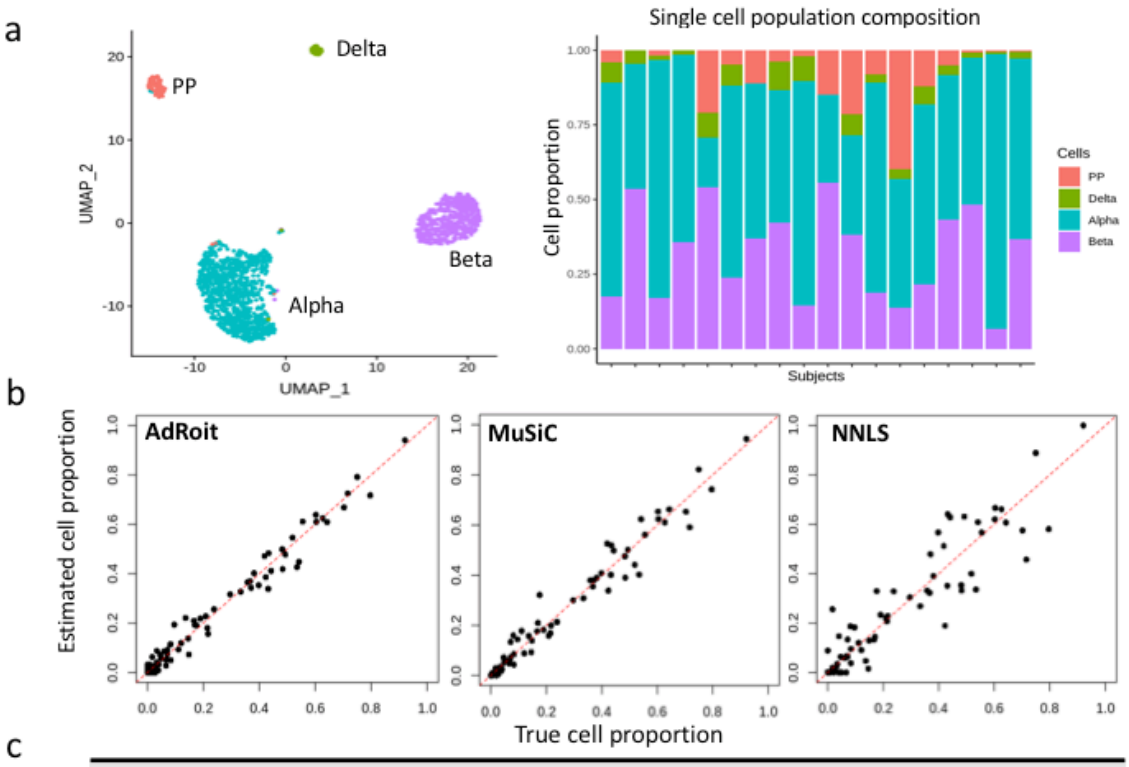

a
AdRoit
Stereoscope
Cell type
Th
satellite.glia
PEP3_Trpm8
PEP2_Htr3a.Sema5a
PEP1_Slc7a3.Sstr2
PEP1_S100a11.Tagln2
PEP1_Dcn
NP_Nts
NP_Mrgprd
NP_Mrgpra3
NF2_Ntrk2.Necab2
NF_Pvalb
NF_Ca1b1
endothelial
Scheme 1
Scheme 2
Scheme 3
Cell proportion
b
AdRoit
Stereoscrope
PEP3_Trpm8
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
Proportion
c
NF subtypes
NP subtypes
PEP subtypes
NF_Calb1 + other*
NP_Mrgpra3 + other*
PEP3_Trpm8 + other*
*other: Th + satellite glia + endothelial
AdRoit
Stereoscope
Detection rate
Simulated proportion

Fig. 5


a
Alpha
Beta
Delta
PP
Estimated Proportion
Donor ID
b
RNA-FISH
AdRoit
Alpha
Beta
Delta
PP
Proportion
Donor ID
c
Beta cell proprotion
Healthy
T2D
p = 4.1e-6
p = 1.8e-6
HbA1C level
d
Spink8
C1ql2
Clic6
Synpo2
HIP CA1 Excitatory2
Dentate Gyrus Granule2
Choroid Plexus
Thalamus Excitatory1
percent

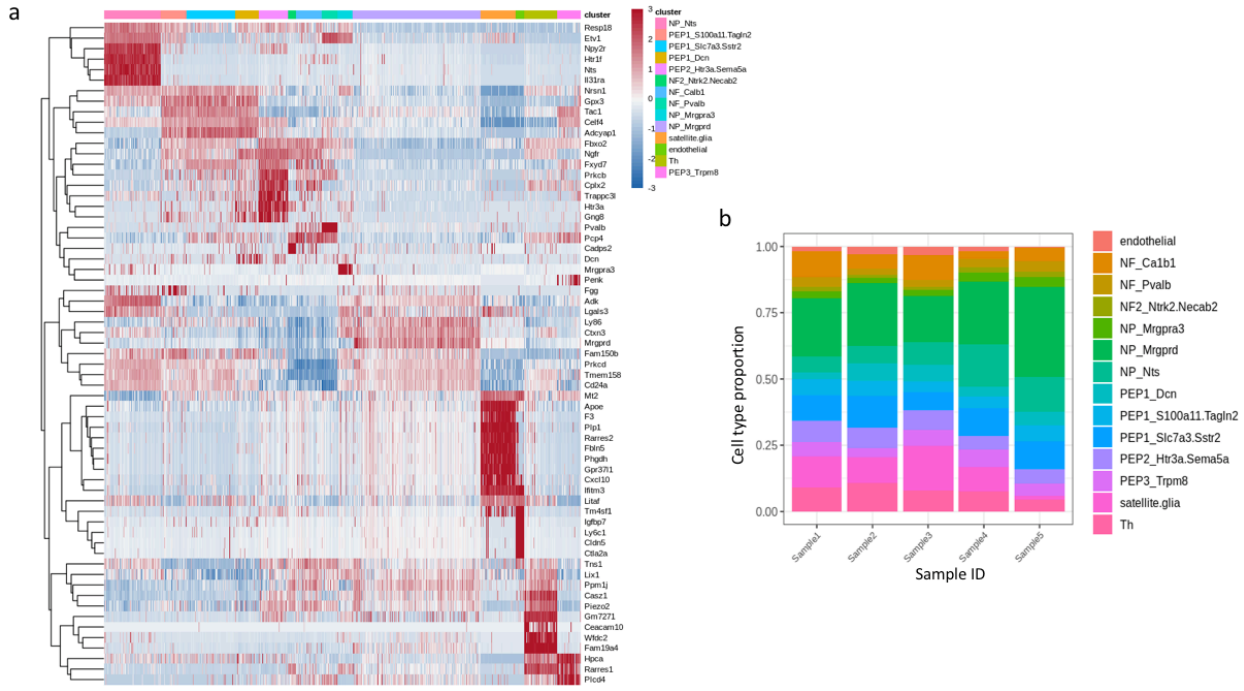
Extended Data Fig. 1



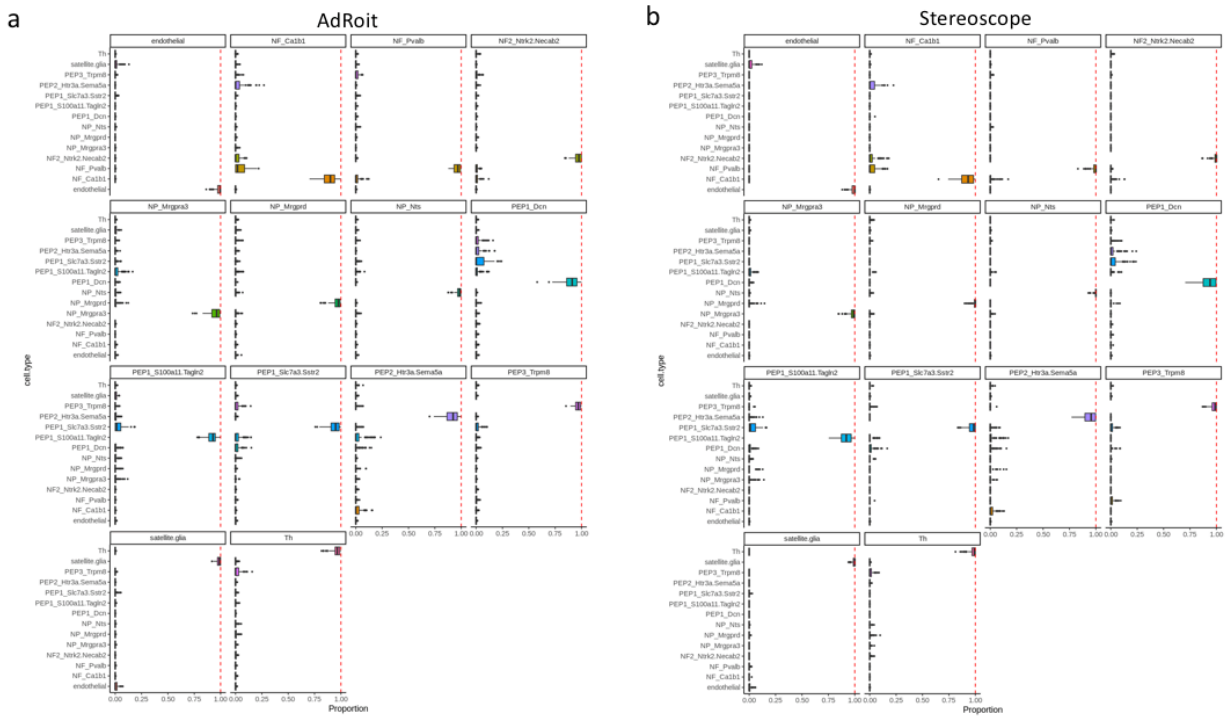
c

	AdRoit	MuSiC	NNLS
mAD	0.029	0.031	0.066
RMSD	0.039	0.046	0.095
Pearson	0.99	0.98	0.93
Spearman	0.97	0.98	0.91

Extended Data Fig. 2



1183 **Extended Data Fig. 3**



1184

1185

1186

1187

1188

1189

1190

1191

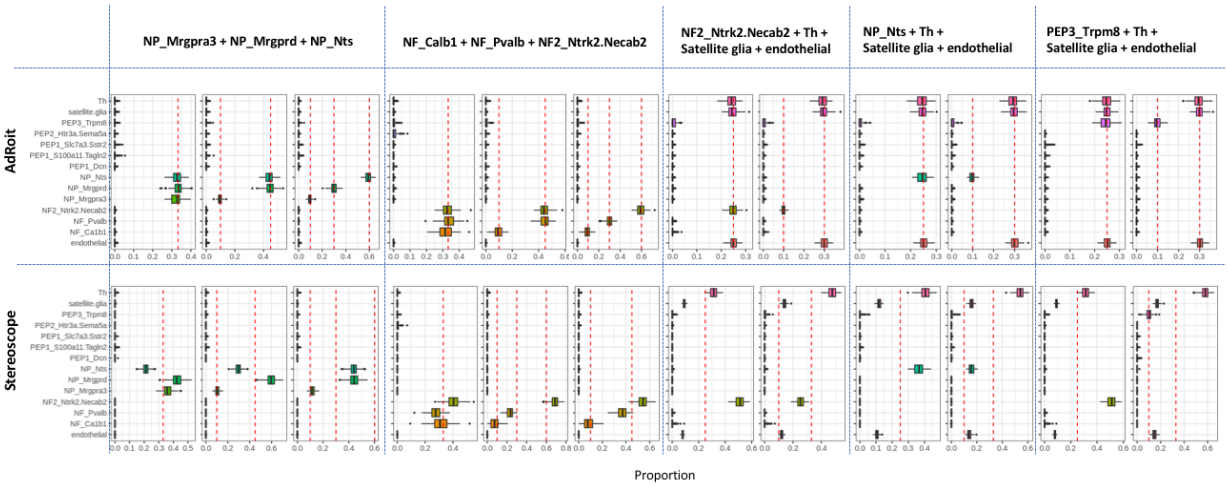
1192

1193

1194

1195

1196 **Extended Data Fig. 4**



1197

1198

1199

1200

1201

1202

1203

1204

1205

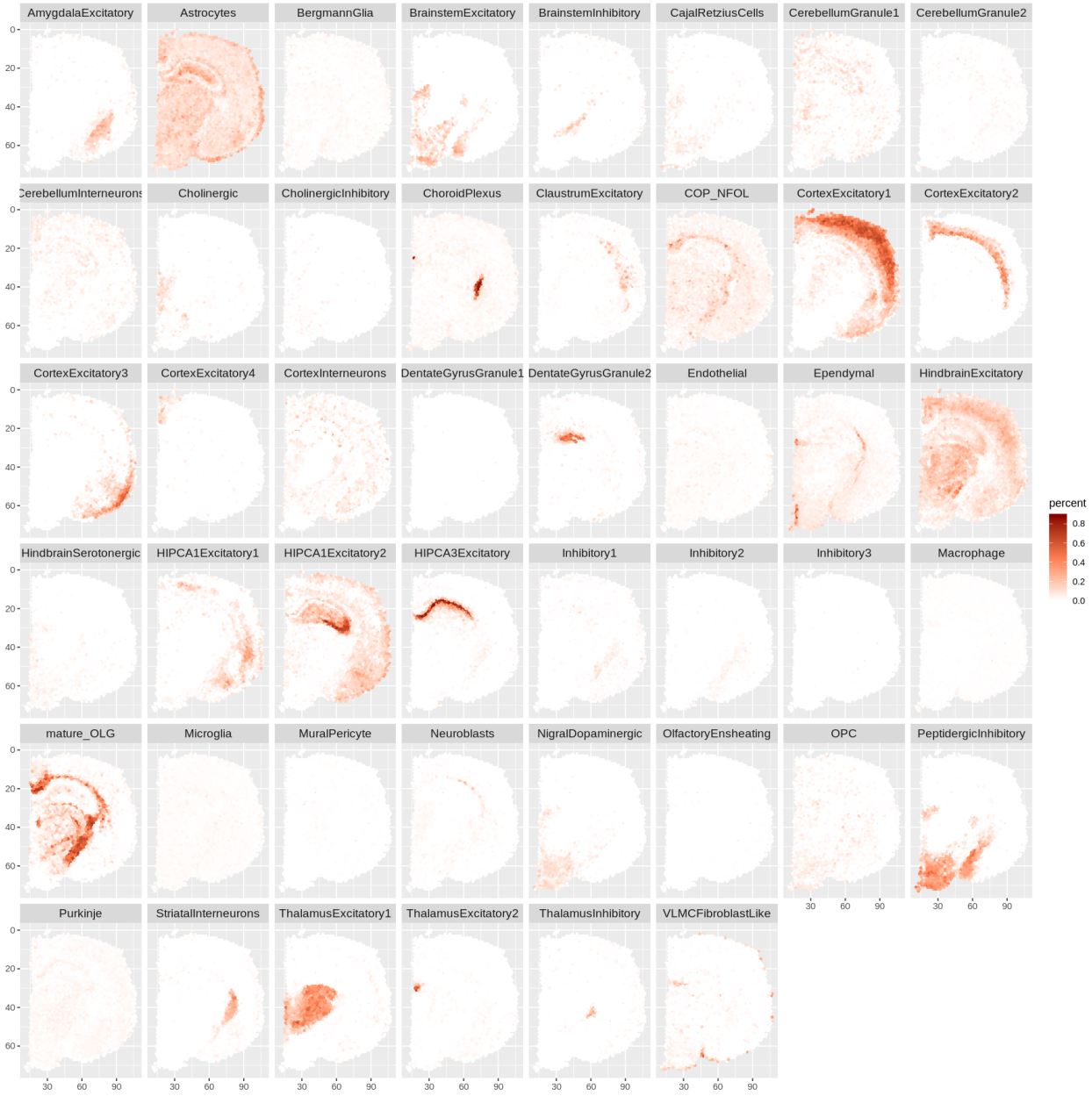
1206

1207

1208

1209

1210

1211

Extended Data Fig. 5