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

Bulk RNA sequencing technology provides the opportunity to understand biology at the whole transcriptome level without the prohibitive cost of single cell profiling. Advances in spatial transcriptomics enable to dissect tissue organization and function by genome-wide gene expressions. However, the readout of both technologies is the overall gene expression across potentially many cell types without directly providing the information of cell type constitution. Although several in-silico approaches have been proposed to deconvolute RNA-Seq data composed of multiple cell types, many suffer a deterioration of performance in complex tissues. Here we present AdRoit, an accurate and robust method infer the cell composition from transcriptome data comprised of multiple cell types. AdRoit uses gene expression profile obtained from single cell RNA sequencing as a reference. It employs an adaptive learning approach to correct the sequencing technique difference between the single cell data and the bulk or spatial transcriptome data, enabling cross-platform readout comparability. Our

systematic benchmarking and applications, which include deconvoluting complex mixtures that encompass 30 cell types, demonstrate its superior sensitivity and specificity compared to other existing methods as well as its utilities. In addition, AdRoit is computationally efficient and runs orders of magnitude faster than many existing methods.

Introduction

RNA sequencing is a powerful tool to understand the biology of normal and disease tissues at the whole transcriptome level. It helps to elucidate disease mechanism and identify potential treatment targets[1]. Due to the presence of heterogeneous cell populations, the bulk tissue transcriptome only characterizes the overall gene expression across multiple cell types. The cell type identifies and their prevalence remain unknown in the bulk data. However, knowledge of the cell type composition is often critical to understand the biology. For instance, the constitution of stromal and immune cells sculpts the tumor microenvironment, which is essential in cancer progression or response to immune check point inhibition[2–6]. Excessive expression of cytokines in particular leukocyte types underlines the etiology of many chronic inflammatory diseases[7–11]. Such information cannot be directly read out from the bulk RNA-Seq.

Recent breakthroughs in spatial transcriptomics methods enable characterizing whole transcriptome-wise gene expressions at spatially resolved locations in a tissue section[12]. However, it remains challenging to reach a single cell resolution while measuring tens of thousands of genes transcriptome-wise. Some widely used technologies can achieve a resolution of 50-100 µm, equivalent to 3–30 cells depending on the tissue type[12,13]. The

48 transcripts therein may originate from more than one cell type, resulting in another form of
49 RNA-Seq data of multiple cell types. Unlike the bulk RNA-Seq, the profiling data at each spot
50 contains substantial dropouts as merely a few cells are sequenced, imposing additional
51 challenges to demystify the cell type content. We refer to bulk RNA-Seq data and spatial
52 transcriptome data at the multi-cell resolution as compound RNA-Seq data hereafter.
53
54 The rapid development of single-cell RNA-Seq (scRNA-Seq) technologies has allowed for cell-
55 type specific transcriptome profiling[14]. It provides the information missing from the RNA-Seq
56 data of tissues comprised of multiple cell types (e.g., bulk RNA-Seq). Nevertheless, the
57 technologies have low sensitivity and substantial noise due to the high dropout rate and the
58 cell-to-cell variability. Consequently, scRNA-Seq technologies require a large number of cells
59 (thousands to tens of thousands) to ensure statistical significance in the results. In addition, the
60 cells must remain viable during the capture. These requirements render the scRNA-Seq
61 technologies costly, prohibiting their application in clinical studies that involve many subjects or
62 cannot allow real time tissue dissociation and cell capture. Furthermore, scRNA-Seq
63 technologies may not be well suited to characterizing cell-type proportions in solid tissues
64 because the dissociation and capture steps may not have the same efficiency for different cell
65 types [15–17].
66
67 Because sequencing at the single cell level is not always feasible and also for the purpose of
68 better interpreting many highly valuable existing bulk RNA-Seq or spatial transcriptomic data
69 sets, in silico approaches have been developed to infer cell type proportions from compound

70 RNA-Seq data[18–24]. The most common strategy is to conduct a statistical inference through the
71 maximum likelihood estimation (MLE)[25] or the maximum a posterior probability estimation
72 (MAP)[26] on a constrained linear regression framework, wherein the unobserved mixing
73 proportion of a finite number of cell types are part of the latent variables to be optimized[19,21–
74 24]. The deconvolution methods are often applied to dissect the immune cell compositions in
75 blood samples[27–31]. However, their performance in more complex tissues, such as the nervous,
76 ocular, respiratory or gastrointestinal organs, remains unclear. These tissues often contain
77 many cell types (10-10^2) and the difference among related cells can be subtle, rendering the
78 deconvolution a challenging task. For example, a recent study on the mouse nervous system
79 using scRNA-Seq found more than 200 cell clusters and many are highly similar neuronal
80 subtypes[32].
81
82 Earlier works often utilized the transcriptome profiling of the purified cell populations to
83 estimate the gene expressions of 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 Although it provides higher throughput by measuring multiple cell types in one experiment,
86 profiling at the single cell level has a high noise floor. The accuracy of deconvolution using
87 scRNA-Seq data as the reference may be affected by the data noise 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 does not hold, as the impact of changing platforms
101 is different for each gene. Though learning a uniform scaling factor would correct the difference
102 in most genes, a few genes that remain significantly biased can easily confound the estimation,
103 especially under a linear model framework. Thus, a gene-wise correction should be considered.
104
105 In this work, we present a new deconvolution method, AdRoit. A unified framework that jointly
106 models the gene-wise technology bias, as well as the cell type specificity and cross-sample
107 variability of genes. The method estimates the cell type constitution in the compound RNA-Seq
108 samples using relevant single cell data as a training source. Genes used for deconvolution were
109 automatically selected from the single cell data based on their information richness. Uniquely,
110 it uses an adaptively learning approach to estimate gene-wise scaling factors, addressing the
111 issue that each platform impact genes differently. AdRoit further makes use of regularization to
112 avoid collinearity among closely related cell subtypes that are common in complex tissues. Over

113 a set of comprehensive benchmarking data with a varying cell composition complexity, AdRoit
114 showed superior sensitivity and specificity to other existing methods. Applications to real RNA-
115 Seq bulk data and spatial transcriptomics data revealed strong and expected biologically
116 relevant information. We believe AdRoit offers an accurate and robust tool for cell type
117 deconvolution and will enhance the value of bulk RNA-Seq and the spatial transcriptome
118 profiling.
119

120 ## Results

121 **Overview of the AdRoit framework**

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

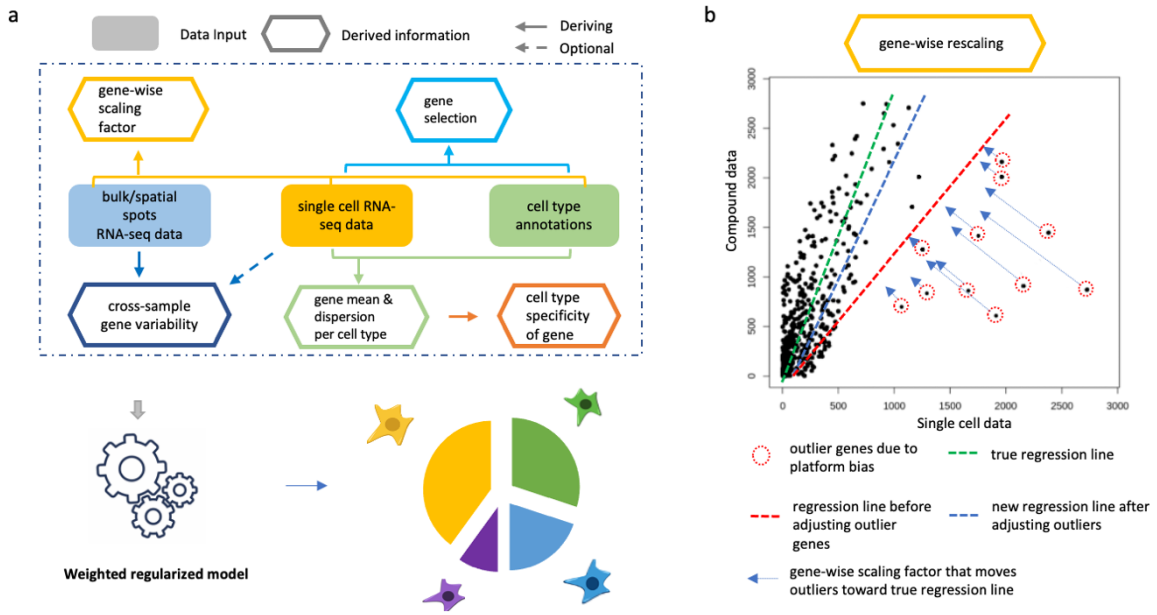

Fig. 1

AdRoit considers a comprehensive set of other relevant factors including the cross-sample variability and the cell type specificity of genes, as well as the collinearity of expression profiles among closely related cell types. The cross-sample variability of a gene is the variation of its expression in the same cell type across multiple bulk samples. It shall be distinguished from the expression change in the gene over different cell types. AdRoit decreases the weight of genes with high cross-sample variability whilst increasing the weight of those with an expression highly specific to certain cell types. The definition of cross-sample variability and cell type specificity also accounts for the overdispersed nature of counts data. Lastly, AdRoit adopts a linear model to ensure the interpretability of the coefficients. At the same time, AdRoit includes a regularization term to minimize the impact of statistical collinearity. Each of the factors

contributes an indispensable part to AdRoit, leading to an accurate and robust deconvolution method for inferring complex cell compositions.

To evaluate the performance, we compared AdRoit with NNLS[18,36], MuSiC[22], Bisque[37], and SPOTlight[38] for bulk data deconvolution, and Stereoscope[23], Cell2location[39], and SPOTlight[38] for spatial transcriptomics data deconvolution. When evaluating the algorithms, a common practice is to pool the single cell data to synthesize a “bulk” sample with the known ground truth of the cell composition. We measured the performance by comparing the estimated cell proportions with true proportions using four metrics: mean absolute difference (mAD), root mean square deviation (RMSD), and two correlation statistics (i.e., Pearson and Spearman). While Pearson reflects linearity, Spearman measures whether the estimated results and the ground truth are monotonically related, even if their relationship is not linear, which avoids the artificial high linearity scores driven by outliers when the majority of estimates are small. Thus, both correlations statistics were included. Good estimations feature low mAD and RMSD along with high correlation statistics. When estimating cell proportions for a synthetic sample, cells from this sample were excluded from the input single cell reference (i.e., leave-one-out) to avoid overfitting. We further applied AdRoit to real bulk RNA-Seq data and validated the results by available RNA fluorescence *in-situ* hybridization (RNA-FISH) data. The estimates were further confirmed by the biological knowledge of human pancreatic islets. We also used AdRoit to map cell types within spatial spots, and the accuracy was verified by *in-situ* hybridization (ISH) images from Allen mouse brain atlas[40].

166 **AdRoit excels in both simple and complex cell constitutions**

167 We started with a simple human pancreatic islets data set that contains 1492 cells and four
168 distinct endocrine cell types (Alpha, Beta, Delta, and PP cells)[41] (Supplementary Fig. 1a;
169 Supplementary Table 1). The synthesized bulk data were constructed by mixing the single cell
170 data at known proportions. All methods except SPOTlight achieved satisfactory performance
171 according to the evaluation metrics, AdRoit performed the best as reflected by scatterplots of
172 estimated proportion vs. true proportion (Supplementary Fig. 1b, Supplementary Table 2). It
173 had the lowest mAD and RMSD, and the highest correlation values (MuSiC's correlations are
174 comparable) among the five methods tested (Supplementary Fig. 1c). Given that MuSiC is a
175 weighted NNLS and consistently showed an improved accuracy here and in the original
176 publication[22], we excluded NNLS from further evaluation hereafter.

177

178 We then tested the methods on a complex tissue - human trabecular meshwork (TM). We
179 acquired published single cell data that contains 8758 cells and 12 cell types from 8 donors[42].
180 The data include 3 similar types of endothelial cells, 2 types of Schwann cells, and 2 types of TM
181 cells (Supplementary Fig. 2; Supplementary Table 3). Cells from each donor were pooled as
182 individual synthetic bulk samples. The cell type proportions vary from <1% to 43%. These
183 proportions were the ground truth cell composition and were compared head-to-head with the
184 estimated proportions inferred by four methods. For each synthetic bulk sample, estimations
185 were performed using a reference built from cells of remaining donors (i.e., leaving-one-out). In
186 each of the 8 samples, the estimates made by AdRoit best approximated the true proportions.
187 In particular, AdRoit had the lowest mAD (0.016) and RMSD (0.025), and the highest correlation

188 values (Pearson = 0.97; Spearman = 0.94) among the methods (Fig. 2a). We further assessed the
189 deviation of the estimates from the true proportions for each cell type. AdRoit consistently had
190 the lowest deviations from the true proportions for all cell types, as well as the lowest variation
191 among 8 samples (Fig. 2b, blue dots), indicating robustness over various cell types and samples
192 (Supplementary Fig. 3, Supplementary Table 4).

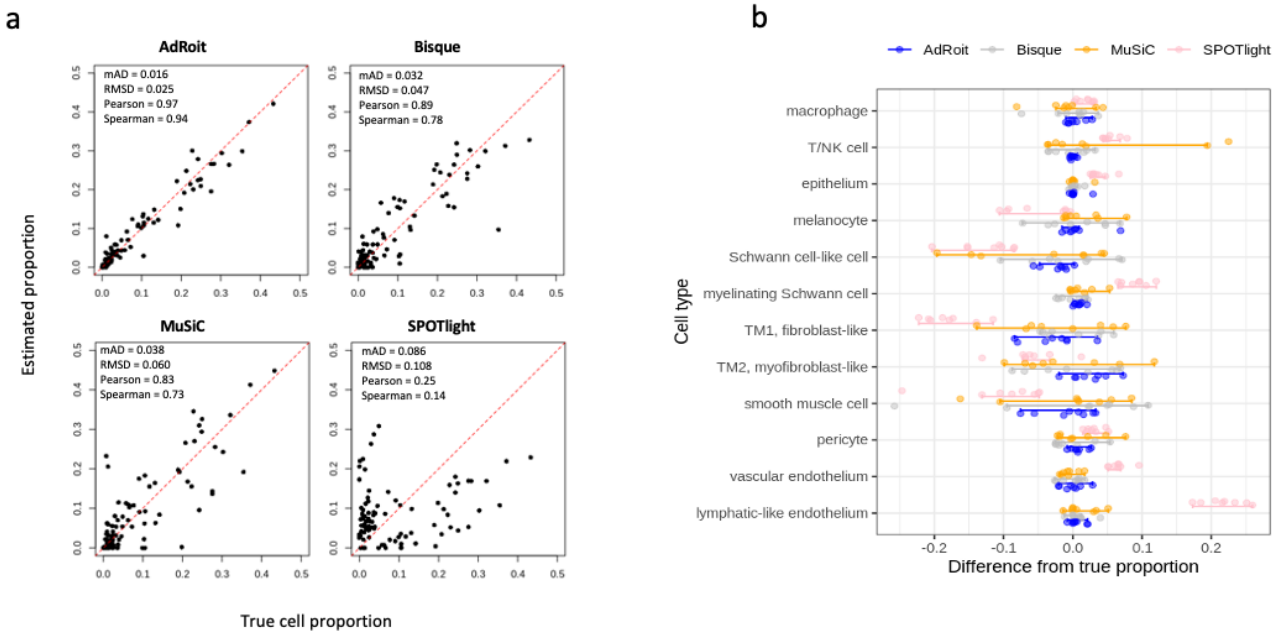

Fig. 2

193

194 **AdRoit has better sensitivity and specificity**

195 To assess the level of granularity that AdRoit can achieve when deconvoluting similar cell types,
196 we used well-studied, closely related immune cells from myeloid lineage and lymphoid lineage.
197 We obtained B cells, naïve $CD4^+$ T cells, memory $CD4^+$ T cells, $CD8^+$ T cells, natural killer (NK)
198 cells, dendritic cells (DC), $CD14^+$ monocytes, $FCGR3A^+$ monocytes, and platelets from a public
199 release of the human peripheral blood mononuclear cell (PBMC) single cells by 10x Genomics
200 (see also Data Availability). We simulated bulk samples that each contained a mixture of

201 myeloid lineage or lymphoid lineage cell types. Such a sample was created by randomly
202 selecting cells from a given set of cell types and mixing them according to their predefined
203 percentages. We repeated this procedure 100 times for a series of mixing proportions among
204 myeloid or T lymphoid cells. The bulk mixtures of $CD14^+$ monocytes, $FCGR3A^+$ monocytes and
205 dendritic cells (DC) followed three schemes of proportions: 0.33:0.33:0.33 (mix0), 0.1:0.45:0.45
206 (mix1) and 0.1:0.3:0.6 (mix2). The same ratios were applied to the mixtures of naïve $CD4^+$ T,
207 memory $CD4^+$ T and $CD8^+$ cells. All the nine cell types in the PBMC data were used to build the
208 reference for the deconvolution. AdRoit accurately estimated the known percentages in all
209 cases (Fig. 3a, Supplementary Table 5). In addition, there were no noticeable false positive
210 predictions of any excluded cell types. These results highlight that AdRoit can capture the fine
211 difference among similar cell types and quantitatively distinguish them.

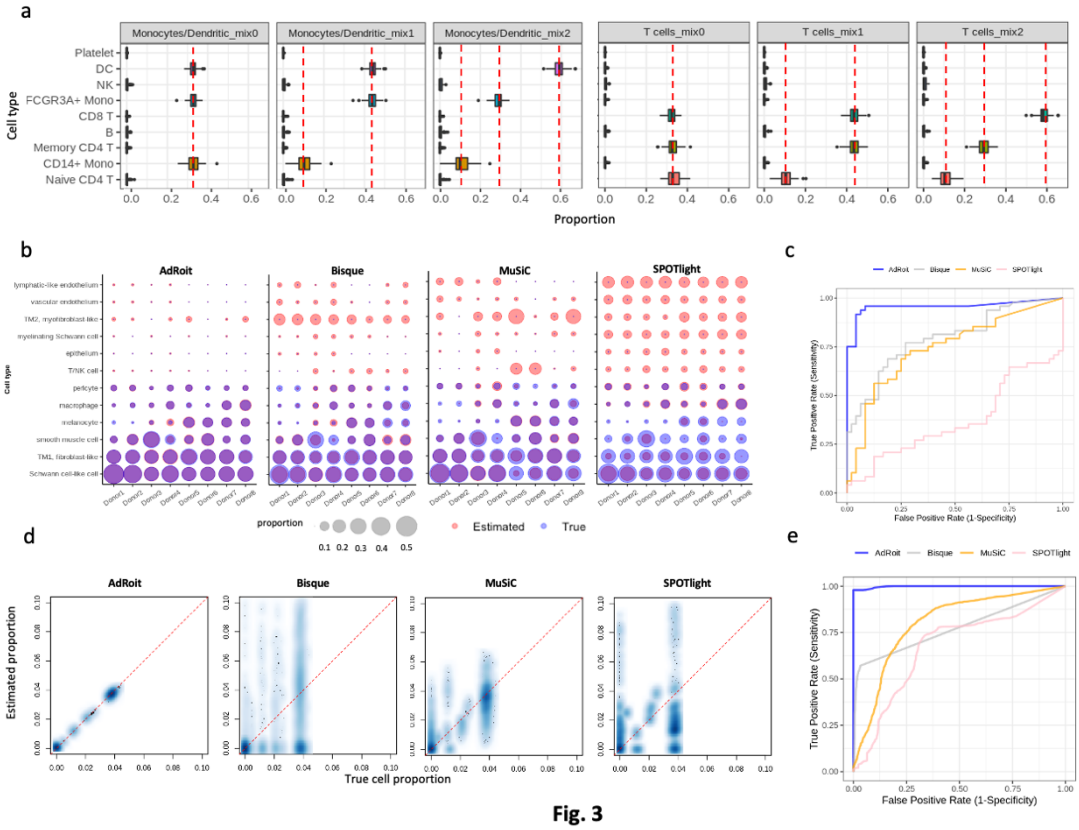

Fig. 3

We next systematically benchmarked the sensitivity and specificity of each of the algorithms. In the context of the cell type deconvolution, a false negative occurs when the proportion of an existing cell type is predicted to be zero (or below a given threshold). Conversely, a non-zero prediction (or above a given threshold) of an absent cell type results in a false positive. False negatives and false positives measure the sensitivity and specificity of a deconvolution algorithm, respectively. Both quantities are crucial to establish the utility of the algorithm. Particularly, in real-world applications, it is often difficult to know *a prior* what cell types exist in a bulk sample, users may inform the algorithm to consider more possible cell types than what actually exist in the sample. False positive predictions in this situation would make the algorithm unusable.

223

224 We designed a simulation to test sensitivity and specificity. we selected 6 out of the 12 human
225 trabecular meshwork cell types, i.e., Schwann-cell like cell, TM1, smooth muscle cell,
226 melanocyte, macrophage, and pericyte, from each donor sample and pooled them within that
227 sample to synthesize 8 new bulk samples. The unselected 6 cell types were considered absent
228 in the bulk samples. Some cell types selected highly resembled those in absence, challenging
229 the programs to pinpoint the right cell type present in the bulk among similar candidates. We
230 provided the full list of 12 single cell types as the reference to the programs to estimate the cell
231 type proportions. Consistently across 8 samples, AdRoit had the most accurate estimates for
232 the 6 present cell types, and 0 or close to 0 estimated values for the non-existing cell types in
233 the simulated bulk data. MuSiC, Bisque and SPOTlight were notably less accurate on the 6
234 selected cell types, as well as they had many non-negligible values for the 6 cell types excluded
235 in the 8 synthetic samples (Fig. 3b, Supplementary Table 6). For example, smooth muscle cells
236 accounted for ~14% in donor 4 but were largely missed (~0.03%) by MuSiC. We noted that TM2
237 had false non-zero estimates from all methods though not included. This is because TM2 is
238 easily mistaken as TM1 due to their high similarity[42]. Nonetheless, AdRoit's estimates of TM2
239 were consistently low across samples (<1% for 44 out of 48 estimates), while the other
240 methods had significantly larger estimates of TM2 that occasionally even exceeded the TM1
241 estimates. For a systematic comparison, we constructed the receiver operating characteristic
242 (ROC) curve by varying the threshold of detection (i.e. a cutoff below which the cell type was
243 deemed undetected) (Fig. 3c). AdRoit had a significantly higher area under the curve (AUC) than

244 the other methods (AdRoit: 0.95, Bisque: 0.79, MuSiC: 0.74, SPOTlight: 0.37), implying both
245 better sensitivity and specificity.
246
247 As complex tissues often contain several tens of different cell types, we continued the
248 evaluation of the sensitivity and specificity in more complicated cell mixtures. We utilized the
249 published mouse brain single cell atlas by Zeisel et al that contains a comprehensive set of
250 neuronal and supporting cells[32]. Without loss of generality, we consolidated the original
251 annotation into 46 major cell types (Methods, Supplementary Table 7). To synthesize a bulk
252 mixture, we randomly selected 30 cell types and pooled all their associated cells. This
253 procedure was repeated independently 100 times for thorough coverage of different cell type
254 combinations. The deconvolution was performed using all 46 cell types as the reference. As
255 illustrated in Fig. 3d, AdRoit was able to estimate the proportions closely consistent with the
256 ground truth while the other three tools made less satisfactory predictions (see also
257 Supplementary Table 8). The ROC curve further indicates that AdRoit has the best sensitivity
258 and specificity given its highest AUC (Fig. 3e; AdRoit: 0.99, Bisque: 0.77, MuSiC: 0.80, SPOTlight:
259 0.67). These observations convince us that AdRoit offers a leading performance on highly
260 complex cell constitutions.
261
262 **AdRoit outperforms in deconvoluting closely related cell subtypes**
263 To further evaluate AdRoit when multiple highly homologous subtypes of cells are present in a
264 complex tissue, we performed scRNA-Seq experiment on mouse lumbar dorsal root ganglion
265 (DRG) from five mice. Following the standard analysis pipeline (Methods), we obtained 3352

266 single cells after quality control procedures. After clustering and annotation, we discovered 14
267 cell types including multiple subtypes of neuronal cells (Fig. 4a, Supplementary Table 9). The
268 heatmap of the top marker genes showed distinct patterns of the major cell types as well as
269 similar patterns of the subtypes (Supplementary Fig. 4a), with the cell type proportions varying
270 from 0.5% to 33.71% (Supplementary Fig. 4b). These 14 cell types include 3 subtypes of
271 neurofilament containing neurons (i.e., NF_Calb1, NF_Pvalb, NF_Ntrk2.Necab2), 3 subtypes of
272 non-peptidergic neurons (i.e., NP_Nts, NP_Mrgpra3, NP_Mrgprd), and 5 subtypes of
273 peptidergic neurons (i.e., PEP1_Dcn, PEP1_S100a11.Tagln2, PEP1_Slc7a3.Sstr2,
274 PEP2_Htr3a.Sema5a, PEP3_Trpm8). Also discovered were tyrosine hydroxylase containing
275 neurons (Th), satellite glia, and endothelial cells. Such complex compositions formed a
276 challenging testing ground for evaluating the ability to distinguish closely related cell types. We
277 again performed the leave-one-out deconvolution on five synthesized bulk samples.

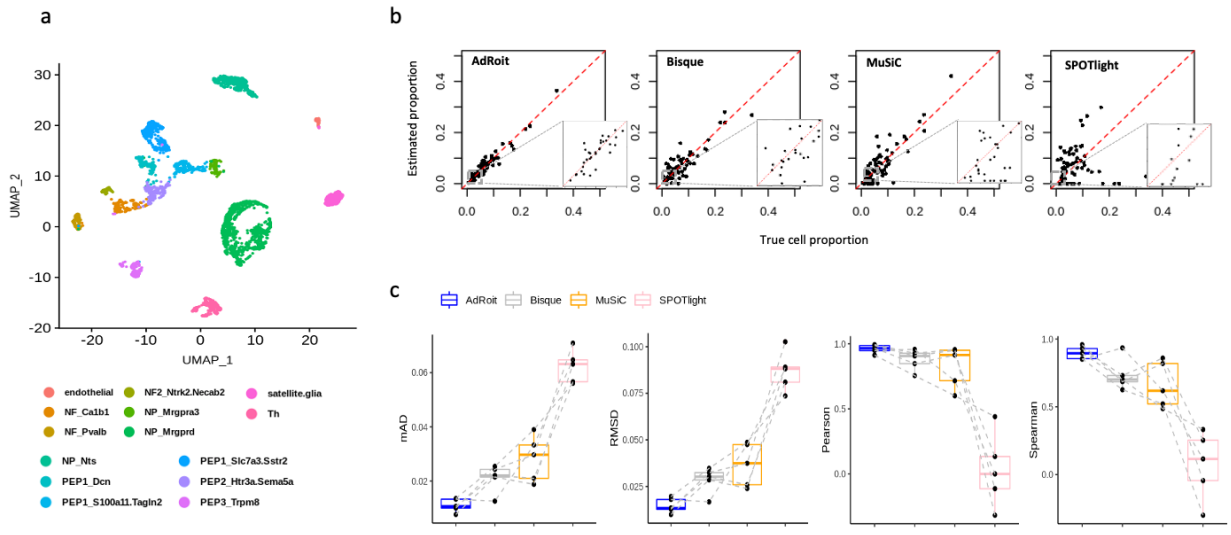

Fig. 4

278

279 AdRoit had highly accurate estimations on all cell subtypes across samples (Fig. 4b). Worth
280 noting, for the rare cell types that account for less than 5%, AdRoit still had a good estimation
281 that is close to the true proportions and never missed a single cell type, showing that AdRoit is
282 very robust on rare cell types. For example, 0.51% endothelial cells were predicted to be 0.35%,
283 and 1.05% NF2_Ntrk2.Necab2 cells were predicted to be 0.85% (Supplementary Fig. 5,
284 Supplementary Table 10). Conversely, Bisque, MuSiC, and SPOTlight were notably less accurate,
285 especially for the cell types less than 5%, and all three missed multiple cell types including some
286 large clusters accounting for ~10% of the cells (PEP1_Slc7a3.Sstr2 cells of Sample5). We
287 continued to examine how much the variability of the estimates was in each sample. We
288 computed the same 4 metrics used previously to evaluate the performance on each of the 5
289 synthetic samples and compared them head-to-head among the algorithms. This fine
290 comparison showed AdRoit significantly outperformed on nearly every sample (Fig. 4c).
291 Moreover, the performance metrics of AdRoit were highly consistent across samples with the
292 lowest variability among all four methods.
293
294 **AdRoit excels on spatial transcriptomics data**
295 Given the promising performance on complex tissues, we moved forward to test AdRoit's
296 applicability to spatial transcriptome data. Spatial transcriptomics data differs from bulk RNA-
297 Seq data in that each spot only contains transcripts from a handful of cells (3-30)[12]. Some of the
298 spots contain multiple cells of the same type, while others may have heterogeneous cell types
299 at varying mixing percentages (e.g., spatial spots at the boundary of different cell types). As the
300 spot is a pool of only a few cells, the variations across spatial spots are expected to be greater

301 than the changes in bulk samples, imposing an additional challenge for deconvolution. We
302 simulated a large number of spatial spots (3200 in total) by using sampled cells from the DRG
303 single cell data above (Methods), then compared AdRoit with Stereoscope, Cell2location, and
304 SPOTlight over a range of simulation scenarios.
305
306 We first tested whether the methods could correctly infer a single cell type when the spots
307 contain cells from that same type. For each of the 14 cell types from DRG, we sampled 10 cells
308 and pooled them to form a spatial spot. We repeated the simulation 100 times for robust
309 testing, then used the full set of 14 cell types as reference to deconvolute the 1400 simulated
310 spots. All methods except SPOTlight were able to identify the correct cell types with high
311 accuracy on the target cell types (i.e., percentages close to 1) and comparably low estimated
312 values (i.e., percentages close to 0) for other cell types excluded from the
313 simulation(Supplementary Fig. 6, Supplementary Table 11).
314
315 We then proceeded to a more difficult scenario where we sampled cells from the 5 PEP neuron
316 subtypes and mixed them. We created three simulation schemes for a comprehensive
317 evaluation: 1) 5 PEP subtypes had the same percent of 0.2; 2) PEP1_Dcn was 0.1 and the other
318 4 were 0.225; 3) PEP1_S100a11.Tagln2 and PEPE1_Dcn were 0.1, PEP2_Htr3a.Sema5a and
319 PEP1_Slc7a3.Sstr2 were 0.2, and PEP3_Trpm8 was 0.4. Again, each simulation scheme was
320 repeated 100 times. Under each scheme, the estimates by AdRoit consistently centered around
321 the true proportions while the predictions of irrelevant cell types remained close to 0 (Fig. 5a,
322 Supplementary Table 11). In comparison, Cell2location systematically overestimated the

PEP1_Slc7a3.Sstr2 by about 2.5%. Stereoscope more notably overestimated PEP2_Htr3a.Sema5a and underestimated the PEP3_Trpm8. SPLOTlight had high noise for some cell types (e.g., endothelial and satellite glia) and generally deviated from the truth for multiple cell types under all simulation schemes.

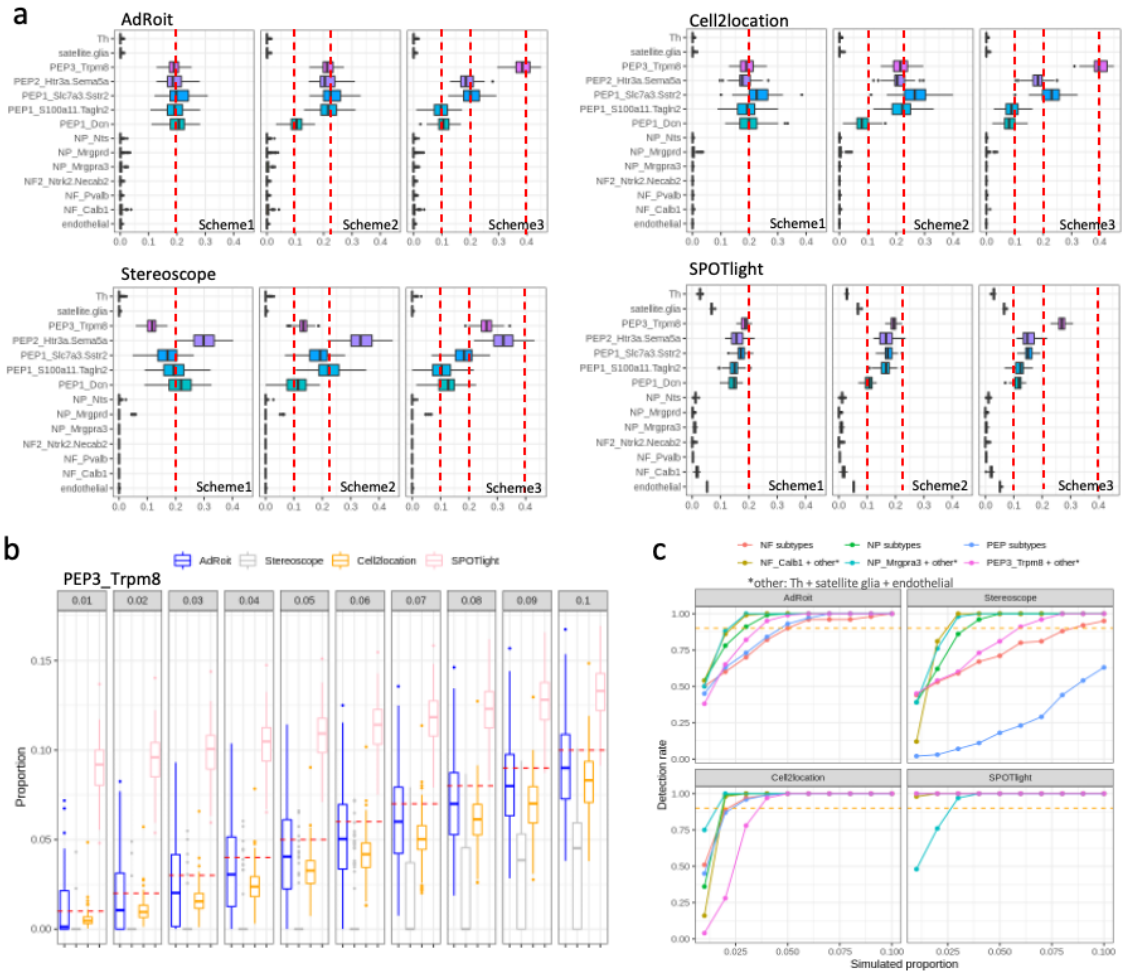

Fig. 5

We expanded the simulated spatial spots to the mixture of 3 NP cell subtypes and the mixture of 3 NF subtypes. In addition, we blended either NP_Mrgpra3 cells, NF_Calb1 cells, or PEP3_Trpm8 cells with three other cell types (i.e., Th, satellite glia and endothelial). For all these simulated spatial spots, AdRoit's estimates consistently centered around the true

332 proportions, whereas the other methods deviated by different degrees in most of the simulated
333 schemes (Supplementary Fig. 7, Supplementary Table 11). We speculate one important reason
334 for the underperformance of some methods such as Stereoscope is that they normalize the
335 total UMI counts to the same value for all cells. However, in real world, different cell types
336 unlikely have the same amount of RNA transcripts (e.g. immune cells have about 10 fold fewer
337 total UMIs than the neurons). Our simulation pooled the cells by adding up the raw UMI counts
338 per gene, which we believe best mimics the real data. Normalizing the data by total UMI not
339 only eliminates this factual difference between cell types but may distort the data from the
340 assumed negative binomial distribution. AdRoit avoids this problem by model the raw UMI
341 counts.
342
343 Next, we asked how sensitive the methods are in detecting rare cell populations. We simulated
344 mixtures of 3 PEP subtypes (i.e., PEP1_Slc7a3.Sstr2, PEP2_Htr3a.Sema5a, and PEP3_Trpm8)
345 wherein the percentage of PEP3_Trpm8 cells varied from 0.01 to 0.1 by 0.01, and the other two
346 cell types sharing the remaining percentage equally (Methods). At each given mixing ratio, the
347 simulation was repeated 100 times. We then checked how accurately the concentration of
348 PEP3_Trpm8 cells was estimated. The medians of AdRoit's estimates were closest to the true
349 proportions (Fig. 5b, red dashed lines) among all 4 methods, followed by Cell2location.
350 Stereoscope's estimates were systematically lower than the true values and failed to detect the
351 PEP3_Trpm8 cells when the simulated proportion was below 0.06. On the other hand,
352 SPOTlight constantly overestimated the presence of PEP3_Trpm8 cells. The median of its
353 predictions remained around 9% even when the true population was at 1%. Such observations

354 are consistent with the frequent false positives made by SPOTlight when deconvoluting the
355 simulated bulk samples (Fig. 3b). This comparison implied AdRoit is more advantageous in
356 detecting low percent cells. For a complete evaluation, we replicated the comparison using 5
357 additional sets of cell mixtures: NF_Calb1 with NF_Pvalb and NF2_Ntrk2.Necab2 (NP subtypes);
358 NP_Mrgpra3 with NP_Mrgprd and NP_Nts (NP subtypes); NF_Calb1 with Th, satellite glia and
359 endothelial (NF_Calb1 + others); NP_Mrgpra3 with Th, satellite glia and endothelial
360 (NP_Mrgpra3 + others); and PEP_Trpm8 with Th, satellite glia and endothelial (PEP_Trpm8 +
361 others). In each set, the first cell type mentioned was considered rare with its percent varied
362 from 0.01 to 0.1. The rest cell types shared the remaining proportion evenly. At each given
363 percent of the rare cell type, we computed how many times out of 100 the rare cell component
364 was detected (estimation > 0.005). AdRoit had systematically high detection rates especially
365 when the percent is below 3%. (Fig. 5c, Supplementary Table 12). Note that the apparent high
366 detection rates in SPOTlight were merely the result of its high false positive estimates (Fig. 3b
367 and Fig. 5b). The detection sensitivity of Cell2location was comparable to that of AdRoit.
368 Notably, given a rare cell population of 5%, both AdRoit and Cell2location achieved a detection
369 rate >90% under all simulation schemes, making them powerful tools to uncover rare cells.
370

371 **Application to real bulk RNA-Seq data of human pancreatic islets**

372 Though using synthetic bulk data based on the mixing of single cells is a useful benchmarking
373 strategy, the bulk and single cell RNA-Seq often use distinct RNA library preparations and
374 sequencing protocols. The capability of a method to deconvolute real bulk samples shall be
375 addressed to ensure it is useful in real-world applications. We acquired 70 real human

376 pancreatic islets bulk samples from published studies[41,43,44] (Supplementary Table 13) and used
377 single cell data of the same tissue[41] as the reference to infer the percentages of 4 endocrine cell
378 types (i.e., Alpha, Beta, Delta, PP). The 70 bulk samples were collected from 39 distinct donors,
379 including 26 healthy donors, and 13 donors with type 2 diabetes (T2D). Each donor contributed
380 1 to 5 bulk RNA sample repeats.
381
382 Replicates from the same donor are expected to have similar compositions and thus were used
383 to assess the reproducibility of the estimates from AdRoit. For all cell types, AdRoit had
384 consistent estimates for the same donors (Fig. 6a, Supplementary Table 14). The average
385 standard deviations did not exceed 1% for all 4 cell types (i.e., Alpha: 0.010; Beta: 0.008; Delta:
386 0.004; PP: 0.002). To seek an independent validation, we obtained cell sorting results by RNA-
387 FISH for 4 of the 39 donors[41] (Supplementary Table 15). The estimated cell proportions in the 4
388 samples agreed well with the percentages measured by RNA-FISH (Fig. 6b). The consistency
389 held for both major cells (Alpha and Beta) and the minor cells (Delta and PP). The observed
390 reproducibility and the success of the independent validation showed AdRoit is reliable in
391 deconvoluting real bulk RNA-Seq data.

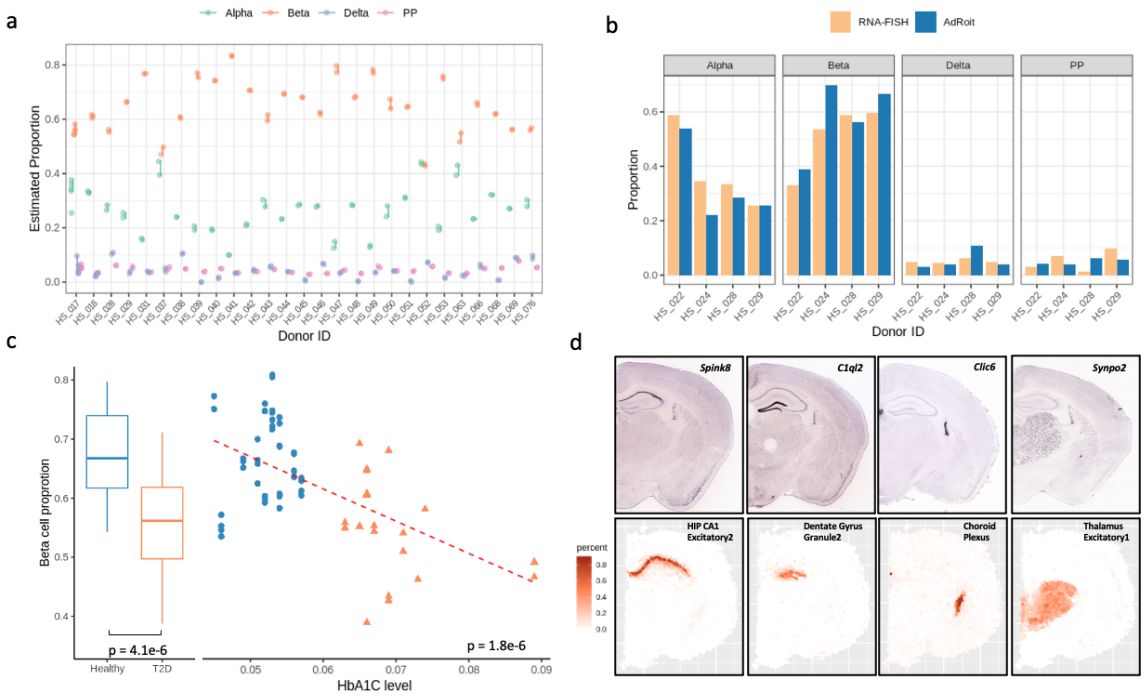

Fig. 6

We then asked if AdRoit can detect known biological differences between healthy and T2D donors. Loss of functional insulin-producing Beta cells is a prominent characteristic of T2D[45–47], typically reflected by an elevated level of hemoglobin A1c (HbA1c)[48,49]. Among the healthy donors, most Beta cell proportions estimated by AdRoit ranged from 50% to 75% (Fig. 6c), agreeing with the known percent range of Beta cells in human islets tissue[50,51]. A significant decrease in the estimated Beta cell proportions was seen in T2D patients (p value = 4.1e-6 by two-sided t-test). Further, a linear regression of estimated Beta cell proportions on HbA1c levels showed a statistically significant negative association (p value = 1.8e-6), showing that AdRoit adequately reflected the cell composition difference between healthy donors and T2D patients.

404 **Application to mouse brain spatial transcriptomics**

405 We lastly demonstrated an application to the real spatial transcriptomics data. Given that the
406 molecular architecture of brain tissues has been well studied, for this evaluation, we chose a set
407 of mouse brain spatial transcriptomics data generated by 10x genomics, containing 2703 spatial
408 spots. The consolidated version of the aforementioned mouse brain single cell atlas [32] was used
409 as the reference for deconvolution (Methods).

410

411 The cell contents inferred by AdRoit per spot appear to accurately match the expected cell
412 types at that location (Supplementary Fig. 8, Supplementary Table 16). For example, the three
413 subtypes of cortex excitatory neurons each occupied a sub-area in the cerebral cortex region.
414 As another example, the shape of the hippocampal region was delineated by the estimated
415 percentages of dentate gyrus granule/excitatory neurons. As an independent validation, we
416 checked the consistency between the estimated cell types with the *in-situ* hybridization (ISH)
417 images from Allen mouse brain atlas[52]. We chose 4 genes highly expressed in 4 brain regions
418 respectively, i.e., Spink8 for the hippocampal field CA1; C1ql2 for the dentate gyrus; Clic6 for
419 the choroid plexus; and Synpo2 for the thalamus[32]. The spots enriched with the 4 indicative cell
420 types, hippocampal CA1 excitatory neuron type 2, dentate gyrus granule neuron type 2, choroid
421 plexus cell and thalamus excitatory neuron type 1, as mapped by AdRoit, precisely co-localized
422 with the regions that showed a strong signal of the 4 marker genes on the ISH images
423 respectively (Fig. 6d). This agreement confirmed that the spatial mapping of cell types by AdRoit
424 is reliable.

425

426 **Computational efficiency**

427 Besides the accuracy and robustness, another major advantage of AdRoit is its magnitude
428 higher computational efficiency. AdRoit uses a two-step procedure to do the inference. The first
429 step prepares the reference using the single cell data where the per-gene means and
430 dispersions are estimated, and cell type specificity is subsequently computed. The built
431 reference can be saved and reused. We tested the running time on the reference construction
432 using the mouse brain single cell dataset discussed earlier that contains ~15,000 cells. It took
433 about 4.5 minutes on a CPU that has 24 cores (23 used for parallel computing). The second step
434 inputs the built reference and the compound RNA-Seq data and does the estimation.
435 Deconvoluting ~2700 RNA-Seq samples of mixed cell populations took around 5 minutes.
436 Therefore, AdRoit in total took less than 10 minutes and ~3Gb memory usage on a regular CPU.
437 As a comparison, MuSiC and Bisque took about 1 hour and 30 minutes on the same data using
438 the same CPU setup. SPOTlight took about 2 hours with the default parameters (cl_n = 100, hvg
439 = 3000, ntop = 200, transf = "uv", method = "nsNMF", min_cont = 0). Stereoscope ran about 24
440 hours continuously with the published parameter setting (-scb 256 -sce 75000 -topn_genes
441 5000 -ste 75000 -lr 0.01 -stb 100 -scb 100) on a powerful V100 GPU with 80 cores and 16G
442 memory. The efficiency of Cell2location was comparable to Stereoscope, taking about 18 hours
443 on the V100 GPU (posterior n_sample = 1000, n_iteration = 10000, selection_specificity =
444 0.45). These methods can be prohibitive when seeking a quick turnaround.

445

446 ## Discussion

447 In this work we have demonstrated that AdRoit is capable of deconvoluting the cell
448 compositions from the compound RNA-Seq data with a leading accuracy, measured by the
449 consistency between the true and predicted cell proportions. Its advantage over the existing
450 state-of-the-art methods was verified over a wide range of use cases. In particular, AdRoit
451 excelled in complex tissues composed of more than ten different cell types with a wide range of
452 cell proportions (e.g., trabecular meshwork, mouse brain and dorsal root ganglion). In these
453 cases, AdRoit performed significantly better than the comparators on deconvoluting bulk RNA-
454 Seq data. AdRoit is also more accurate and sensitive than tools specifically designed for
455 deconvoluting spatial transcriptomics spots, especially in detecting low percent rare cells.
456 Previous benchmarking often assumed the types of cells in the synthetic bulk data are not more
457 or less than the cell types collected in the reference, and thus the only unknown was the
458 proportion of each cell type. This assumption may not hold. Missing existing cell types or false
459 predictions of non-existing ones can hinder the utility of an algorithm. Thus, besides the overall
460 accuracy, we also examined the sensitivity and specificity of the algorithms. We observed a
461 superior sensitivity and specificity in AdRoit, an important leverage for its usage in practice.
462
463 The reference single cell data used by AdRoit came from different platforms, such as the 10x
464 Genomics Chromium Instrument (the mouse dorsal root ganglion), and the Fluidigm C1 system
465 (the human pancreatic islets data). AdRoit consistently exhibited excellent performance across
466 all benchmarking datasets independent of their single cell sequencing technology platforms.
467 More importantly, this statement holds not only for deconvoluting the synthesized bulk data,
468 but also for the real bulk RNA-Seq data. The latter typically does not apply the unique molecular

469 barcoding and requires a significantly different cDNA amplification procedure from what is used
470 in the single cell RNA-Seq (Methods). Moreover, the sequencing depth, read mapping and gene
471 expression quantification are dissimilar as well. The fact that AdRoit accurately dissected the
472 cell compositions in the real bulk samples based on the single cell reference data further
473 supports its cross-platform applicability.
474
475 We attribute the power of AdRoit to its comprehensive modeling of relevant factors. Firstly, we
476 think a common rescaling factor is insufficient to correct the platform difference between
477 single cells and the compound data. Rather, the impact of platform difference on genes is quite
478 different and not linearly scaled. Correcting such differences entails rescaling factors specifically
479 tailored to each gene. AdRoit uses an adaptive learning approach to estimate such gene-wise
480 correcting factors and does the correction in a unified model. In addition, the contribution of a
481 gene in a cell type to the loss function is jointly weighted by its specificity and variability in a cell
482 type, where specificity and variability are defined in a way accounting for the overdispersion
483 property of counts data. Our observations over the multiple benchmarking data sets also show
484 that the coexistence of similar cell types may have induced a collinearity condition that
485 negatively impacted the regression-based methods developed by others. Being able to alleviate
486 this problem gives AdRoit an edge. All these factors help AdRoit to distinguish similar cell
487 clusters while remaining sensitive enough to separate rare cell types.
488
489 Technically, the input profiles of individual cell types to AdRoit need not come from the single
490 cell RNA-Seq. Bulk RNA-Seq profiles of individual isolated cell types can be used as well.

491 Nevertheless, using single cell RNA-Seq data as the reference has a few key advantages. It is a
492 high throughput approach wherein multiple cell types can be interrogated simultaneously. Prior
493 knowledge of the cell types in presence as well as their specific gene markers are not required,
494 which allows novel cell types to be identified. Although detection of lowly expressed genes has
495 been a challenge for the single cell RNA-Seq, significant enhancements have been
496 demonstrated. For example, the number of detectable genes currently can reach an order of
497 10,000 per cell and keeps improving[53]. As AdRoit focuses on the informative genes whose
498 expressions are generally high, the detection limit of the single cell RNA-Seq does not impose a
499 significant drawback. Indeed, given the single cell reference profiles, AdRoit successfully
500 deconvoluted the real bulk RNA-Seq data and spatial transcriptomics data. The results suggest
501 that, besides enriching our understanding of the bulk transcriptome data, AdRoit can leverage
502 the usage of the vast amount and continuously growing single cell data as well.
503
504 AdRoit is a reference-based deconvolution algorithm. A comprehensive collection of the
505 possible cell components is important. However, completeness may not always be guaranteed.
506 Even with the single cell acquisition that is independent of prior knowledge, rare and/or fragile
507 cell types may not survive through the capture procedure and hence are excluded. It is also
508 difficult to generate a solid reference profile for cells that are versatile from sample to sample
509 (e.g., tumor cells). Currently AdRoit deals implicitly with the components unknown to the
510 reference. If an unknown cell type reassembles one of the referenced ones, it may be
511 considered as part of the known cell type and their joint population is predicted. Such an
512 outcome is acceptable as treating two similar cell types as one is still biologically meaningful

513 although the resolution of the system may be compromised. If the unknown component is
514 dissimilar to all the known ones, it will be ignored by AdRoit because its representative markers
515 are unlikely among the top weighted genes associated with the known components. At the
516 same time, the distinct component is expected to have a unique gene expression pattern and
517 thus unlikely to interfere significantly with the gene expressions from the known cell types.
518 Therefore, AdRoit essentially deconvolutes the relative populations among the known cell
519 components. For example, AdRoit was able to correctly uncover the populations of 4 endocrine
520 cell types from the human islet bulk data despite the absence of many other cell types such as
521 macrophages, Schwann cells and endothelial cells in the input single cell reference[20]. Although
522 under such a circumstance, the absolute percentages of the cells remain obscure, we expect
523 their relative proportions can be studied and valuable. A future improvement is to explicitly
524 model the unknown cell types and estimate their percentages upon the signals in the
525 compound data that cannot be explained by the contribution from the known components.
526 AdRoit may also be coupled with other programs[54] to output the deconvoluted cell type specific
527 transcriptome profiles. The inferred cell type proportions by AdRoit can serve as the input to
528 the downstream program and benefit the outcome with the reliability and accuracy
529 demonstrated in this work.
530

531 ## Methods

532 **Gene selection**

533 AdRoit selects genes that provide information about cell type identity, excluding non-
534 informative genes that potentially introduce noise. There are two ways for selecting such

genes: 1) union of the genes whose expression is enriched in one or more cell types in the single cell UMI count matrix. These genes are referred to as marker genes; 2) union of the genes that vary the most across all the cells in the single cell UMI count matrix, referred to as the highly variable genes. For marker genes, we recommend selecting the top ~200 genes, ranked by fold change, from each cell type for resolving complex compound transcriptome data. Considering some genes may be enriched in more than one cell type, we further require selected markers presenting in no more than 5 cell types to ensure specificity. We also suggest selecting a minimum of 1000 total unique marker genes for an accurate estimation. If not satisfied, one may consider expanding the number of top genes and/or loosen the statistical significance cutoff.

AdRoit also offers the option to use highly variable genes. To avoid the selected highly variable genes being dominated by large cell clusters, AdRoit first balances the cell types in the single cell UMI count matrix by finding the median size (i.e., the number of cells) of all clusters, then samples cells from each cluster to make its size equal to the median. Next, AdRoit computes the variance of each gene across all the cells in the balanced single cell UMI matrix. The variance-stabilizing transformation (VST)[55] is applied to normalize the data prior to the variance calculation. Genes with the top 2000 largest variances are then selected.

In both ways, mitochondria genes are excluded as their expression does not have information of cell identity. The results shown in the current paper were based on the marker genes derived

556 as described above. We also demonstrated that using the balanced highly variable genes
557 yielded comparably accurate estimations (Supplementary Fig. 9).
558
559 **Estimate gene-wise mean and variance per cell type**
560 Modeling single cell RNA-Seq data is challenging due to cell heterogeneity, technical sensitivity,
561 and noise. Some genes may not be detected by chance, while others may be found to be highly
562 dispersed. These factors can lead to excessive variability even within the same cell type. AdRoit
563 combats high noise and computational complexity by building models with estimated mean and
564 dispersion per gene per cell type. This strategy reduces the data complexity while preserves the
565 cell type specific information.
566
567 Although typical analyses of RNA-Seq data start with normalization, AdRoit does not do
568 normalization before the mean estimation. Performing a normalization across all cell types
569 forces every cell type to have the same amount of RNA transcripts, measured by the total
570 unique molecular identifier (UMI) counts per cell. However, different cell types can have
571 dramatically different amounts of transcripts. For example, the amount of RNA transcripts in
572 neuronal cells is about 5 times higher than that in glial cells. Thus, normalization can falsely
573 alter the relative abundance of cell types, misleading the estimation of cell type percentages.
574 To avoid this problem, AdRoit models the means using the raw UMI counts.
575
576 Studies have shown that UMI counts follow a negative binomial distribution[56,57]. We therefore
577 fit negative binomial distributions to genes of each cell type and build the model based on the

578 estimated means and dispersions from the selected genes. More specifically, let X_{ik} be the set

579 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 of

580 selected genes, and K denotes the number of cell types in the single cell count matrix. The

581 distribution of X_{ik} follows a negative binomial distribution,

582 $$X_{ik} \sim NB(\lambda_{ik}, p_{ik}), \quad (1)$$

583 where λ_{ik} is the dispersion parameter of the gene i in cell type k, and p_{ik} is the success

584 probability, i.e., the probability of one observed UMI belonging to gene i in cell type k. The two

585 parameters are estimated by MLE. The likelihood function is

586 $$LH(\lambda_{ik}, p_{ik}|X_{ik}) = \prod_{i=1}^{n_k} f(X_{ik}|\lambda_{ik}, p_{ik}), \quad (2)$$

587 where n_k is the number of cells in cell type k, and f is the probability mass function of the

588 negative binomial distribution. The MLE estimates are then given by

589 $$(\widehat{\lambda_{ik}}, \widehat{p_{ik}}) = \underset{\lambda_{ik}, p_{ik}}{arg\max}\, LH(\lambda_{ik}, p_{ik}|X_{ik}). \quad (3)$$

590 Once the success probability and the dispersion are estimated, the mean and the variance of

591 the negative binomial distribution can be computed numerically,

592 $$\mu_{ik} = \frac{\widehat{\lambda_{ik}} \cdot \widehat{p_{ik}}}{1-\widehat{p_{ik}}}, \quad (4)$$

593 $$\sigma_{ik}^2 = \frac{\widehat{\lambda_{ik}} \cdot \widehat{p_{ik}}}{(1-\widehat{p_{ik}})^2}. \quad (5)$$

594 Model fitting using MLE has been readily coded in many R packages. We choose the 'fitdist'

595 function from the 'fitdistrplus' package[58] for its fast computation speed and flexibility in

596 selecting distributions. The mean and variance estimations are done for each selected gene in

597 each cell type, resulting in a $I \times K$ matrix each.

598

599 **Cell type specificity of genes**

600 Genes with a cell-type specific expression pattern better distinguish cell types, thus are more
601 important for resolving cell type composition. In line with this property, AdRoit weighs more
602 the genes with a higher cell type specificity. Highly specific genes usually have consistently high
603 expression and thus relatively low variance among cells within a cell type. To compute the cell
604 type specificity of a gene, we first identify the cell type in which the gene has the highest
605 expression (i.e., most expressed cell type), then defines the specificity of this gene as the mean-
606 to-variance ratio within that cell type. A high ratio renders a high weight to the gene in the
607 model. Here the mean and the variance refer to the estimated values from the negative
608 binomial model (μ_{ik} and σ_{ik}^2 in eq. 4 and 5). Let k' be the index of the cell type that has the
609 highest mean expression of gene i,

610 $$k' = \underset{k}{arg\max}\{\mu_{ik} |\ k \epsilon\ 1 \dots K\}, \qquad (6)$$

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

612 $$w_i^S = \frac{\mu_{ik'}}{\sigma_{ik'}^2}, \qquad (7)$$

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

614

615 **Cross-sample gene variability**

616 The variability of a gene contrasts with how stable a gene is across samples. The idea of
617 weighting genes based on their variability across samples is first explored by Wang et al[22],
618 wherein the variability was defined as the cross-sample variance. By weighting down the high
619 variability genes, the authors achieved a great advantage over the traditional unweighted
620 method. Genes with a low cross-sample variability better represent the sample population,

621 hence are more trustworthy for learning the cell composition. AdRoit incorporates the same
622 notion to weigh the importance of genes, however, it defines the variability more
623 comprehensively. AdRoit acknowledges the dependency between the gene expression level
624 and its variance and thus computes a variance-to-mean ratio (VMR) as the cross-sample
625 variability. Here the mean and the variance are computed across samples. The distribution of
626 the VMR values is less skewed than that of the variance alone. The VMR also circumvents
627 underweighting genes with a low expression, or overweighting genes highly dispersed. Let Y_{ij}
628 denote the counts for gene i in sample $j \in 1, \ldots, J$, then

629 $$Y_{ij} \sim NB(\lambda_{ij}, p_{ij}), \quad (8)$$

630 where λ_{ij} is the dispersion parameter of the gene i in sample j, and p_{ij} is the success
631 probability. Again, we use MLE to estimate $\widehat{\lambda_{ij}}$ and $\widehat{p_{ij}}$, following which the cross-sample mean
632 and variance can be numerically computed:

633 $$\mu_i^S = \frac{\widehat{\lambda_{ij}} \cdot \widehat{p_{ij}}}{1 - \widehat{p_{ij}}}, \quad (9)$$

634 $$(\sigma_i^2)^S = \frac{\widehat{\lambda_{ij}} \cdot \widehat{p_{ij}}}{(1 - \widehat{p_{ij}})^2}, \quad (10)$$

635 and the cross-sample variability VMR for gene i is then defined as

636 $$VMR_i = \frac{(\sigma_i^2)^S}{\mu_i^S} = \frac{1}{w_i^C}, \quad (11)$$

637 where w_i^C is later used in the model. The cross-sample variability weight is computed for each
638 gene in the set of selected genes.
639 Typically, the RNA-Seq data of mixed cells to be deconvoluted include multiple samples to
640 control for biological variability. In this case, the VMR is computed as described above using the
641 cross-sample mean and variance of the gene. In case the RNA-Seq data to be deconvoluted lack

642 multiple samples whereas the single cell reference contains replicates, AdRoit pools all cells per
643 replicate to synthesize multiple compound samples and subsequently estimate the VMRs with
644 them. If neither input has multiple repeats, AdRoit takes a bootstrapping approach to
645 subsample and pool cells from each cell type in the single cell reference several times to
646 generate multiple compound samples.
647
648 **Gene-wise scaling factor to correct the platform bias**
649 When comparing the RNA-Seq data of cell mixtures to the single cell data, a rescaling factor is
650 often used to account for the library size and platform difference. The existing methods adopt a
651 single rescaling factor for each sample, i.e., all genes in a sample are multiplied by the same
652 factor[22,23]. This operation is based on a strong assumption that the platform difference impacts
653 every gene equally and linearly scalable among different cell types, which hardly holds. In
654 addition, because linear models are sensitive to outliers, the estimation of cell proportions can
655 be steered away from the truth by genes that are significantly affected by the platform bias.
656 Therefore, applying a uniform scaling factor to all genes is inappropriate.
657
658 To overcome this problem, AdRoit estimates a gene-wise scaling factor per sample via an
659 adaptive learning strategy and rescales each gene with its respective scaling factor. To proceed,
660 we first input the mean gene expression from the mixture samples (μ_i^S in eq. 9) and the
661 estimated means of each cell type from the single cell data (μ_{ik} in eq. 4), then apply a non-
662 negative least square regression (NNLS) to get a coarse estimation of the proportions of each
663 cell type, τ_k. For each gene i, a predicted mean expression ($\sum_k^K \widehat{\tau_k}\, \mu_{ik}$ in eq. 13) in the sample of

664 mixture cells is computed as the weighted sum of the means of this gene in each cell type

665 wherein the weights are τ_k. The regression equation is given by,

666 $$\mu_i^S = A \cdot (\sum_k^K \tau_k \, \mu_{ik} + \varepsilon), \qquad 0 < \tau_k, \ \sum_k^K \tau_k = 1 \qquad (12)$$

667 where A is a constant to ensure the sum of τ_k's is 1 and ε is the error term. We use the 'nnls'

668 function in the 'nnls' package[59] to estimate τ_k's. Next, we calculate the ratio between the mean

669 expression from the mixture samples and the predicted counterpart, and define the gene-wise

670 rescaling factor as the logarithm of the ratio plus 1,

671 $$r_i = \log\left(\frac{\mu_i^S}{\sum_k^K \widehat{\tau_k} \mu_{ik}} + 1\right). \qquad (13)$$

672 Given the dispersed nature of count data, the logarithm of the ratio is a more appropriate

673 statistic as it results in relatively stable scaling factors. The addition of 1 avoids taking logarithm

674 on zeros. By multiplying the gene-wise rescaling factor, the genes affected more strongly by the

675 platform bias (i.e., "outliers") receive a larger correction (Fig. 1b).

676

677 **Weighted and regularized loss function for training**

678 AdRoit employs a non-negative least square regression model to infer the cell compositions in a

679 RNA-Seq sample of cell mixtures. The model (eq. 14) approximates the gene expression

680 observed in the mixture sample by the weighted (the weights are the mixing percentages) sum

681 of the cell type-specific gene expressions from the single cell reference, after correcting the

682 platform difference.

683 $$y_{ij} \approx r_i \cdot \sum_k^K \beta_k \widehat{\mu_{ik}} \qquad (\beta_1, \dots, \beta_K > 0) \qquad (14)$$

684 where y_{ij} is the expression of gene i in the mixture transcriptome sample j; the coefficient β_k is

685 the mixing proportion for cell type k; $\widehat{\mu_{ik}}$ is the estimated mean expression of gene i in cell type

686 k computed according to eq. 4; r_i is the gene-wise scaling factor computed by eq. 13 to correct
687 the bias due to the technology difference. To infer the β_k coefficients, AdRoit uniquely
688 incorporates all the factors discussed in the previous sections. This is implemented in a
689 weighted sum-of-squared loss function L, where the weights consist of two components, w_i^C in
690 eq. 7 and w_i^S in eq. 11. Thus, a larger weight is given to the genes with a higher cell type
691 specificity and a lower cross-sample variability. In cases of complex tissues where many similar
692 cell subtypes exist, strong collinearity among the subtypes can render the model sensitive to
693 noise and prone to overfitting. AdRoit handles this problem by including an L2 norm of the
694 mixing proportions as the regularization component in the loss function. For a compound
695 transcriptome sample j, the loss function is given by,

696 $$L_j(\beta_1, \dots, \beta_K | y_{ij}, w_i^C, w_i^S, r_i, \widehat{\mu_{ik}}) = \sum_i^I w_i^C \cdot w_i^S \cdot (y_{ij} - r_i \cdot \sum_k^K \beta_k \widehat{\mu_{ik}})^2 + \sum_k^K \beta_k^2. \quad (15)$$

697 where i is the index of a gene among the selected genes used for the deconvolution; y_{ij}, β_k, r_i,
698 and $\widehat{\mu_{ik}}$ are the same as defined in eq. 14; w_i^S and w_i^C are computed as described above in eq. 7
699 and 11, respectively. The β_k coefficients are estimated by minimizing the loss function subject
700 to the constraint $\beta_1, \dots, \beta_K > 0$,

701 $$\widehat{\beta_1}, \dots, \widehat{\beta_K} = \underset{\beta_1,\dots,\beta_K}{\operatorname{argmax}}_{\beta_1,\dots,\beta_K > 0} L_j. \quad (16)$$

702 The minimization is done by a gradient projection method proposed by Byrd et al[60]. We derive
703 the gradient function by taking the partial derivative of the loss function with respect to β_k,

704 $$G_k = \nabla_{\beta_k} L_j = -2 \sum_i^I r_i \cdot \widehat{\mu_{ik}} \cdot w_i^C \cdot w_i^S \cdot \left(y_{ij} - r_i \cdot \sum_k^K \beta_k \widehat{\mu_{ik}}\right) + 2\beta_k. \quad (17)$$

705 AdRoit uses the function 'optim' from the R package 'stats' to perform the minimization[61],
706 providing the loss function (eq. 15) and the gradient (eq. 17). To get the final estimates of cell
707 type proportions, we rescale the coefficients β_k's to ensure a summation of 1,

708 $$\theta_k = \frac{\widehat{\beta_k}}{\sum_k^K \widehat{\beta_k}}. \quad (18)$$

709 Each mixture RNA-Seq sample j is independently estimated by the model described above.

710

711 **Evaluation statistics**

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

713 following four statistics. The mean absolute difference (mAD) and root mean square deviation

714 (RMSD) are given by,

715 $$mAD = \frac{\sum_k^K |\theta_k - \theta_k^0|}{K} \quad (19)$$

716 $$RMSD = \frac{\sum_k^K (\theta_k - \theta_k^0)^2}{K} \quad (20)$$

717 where θ_k and θ_k^0 are the estimated proportions and true proportions, respectively.

718 Pearson correlation coefficient is computed as,

719 $$\rho_p = \frac{\sum_k^K (\theta_k - \overline{\theta_k})(\theta_k^0 - \overline{\theta_k^0})}{\sqrt{\sum_k^K (\theta_k - \overline{\theta_k})^2}\sqrt{\sum_k^K (\theta_k^0 - \overline{\theta_k^0})^2}} \quad (21)$$

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

721 respectively. Spearman correlation coefficient is given by,

722 $$\rho_s = \frac{\sum_k^K (r_k - \overline{r_k})(r_k^0 - \overline{r_k^0})}{\sqrt{\sum_k^K (r_k - \overline{r_k})^2}\sqrt{\sum_k^K (r_k^0 - \overline{r_k^0})^2}} \quad (22)$$

723 where r_k is the rank of θ_k.

724

725 **Generation of synthesized bulk RNA-Seq and spatial transcriptomics data**

726 Bulk RNA-Seq data used for the benchmarking were synthesized by adding up the raw UMI

727 counts per gene from all cells within a subject. Let t_k denote a cell of cell type k, $t_k \in 1, \ldots, T_k$,

728 where T_k is the total number of cells in cell type k. Let Y_{ij}^B be the count of gene i in a

729 synthesized bulk sample associated with the subject j, and X_{ijt_k} be the UMI count of gene i in

730 cell t_k within the subject j, then

731 $$Y_{ij}^B = \sum_k^K \sum_{t_k}^{T_k} X_{ijt_k} \qquad (23)$$

732 The true proportion of cell type k is given by,

733 $$\theta_k^0 = \frac{T_k}{\sum_k^K T_k} \qquad (24)$$

734

735 To simulate the spatial transcriptomic spots, we first sampled 10 cells without replacement

736 from each cell type, then weighed them by their respective mixing proportions before adding

737 them up. For example, to simulate a spot with p_k percent of cell type k, the count Y_{ij}^S of gene i

738 in a spatial spot j is given by,

739 $$Y_{ij}^S = \sum_k^K p_k \sum_{n=1}^{10} X_{ikn} \qquad (25)$$

740 where X_{ikn} is the UMI count of gene i in a selected cell n of cell type k. For each mixing scheme,

741 the simulation was repeated 100 times.

742

743

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

745 As described previously[62], lumbar DRGs were isolated from five 18-week old adult C57BL/6

746 mice (three males and two females) and transferred to a dissociation buffer (Dulbecco's

747 modified Eagle's medium supplemented with 10% heat-inactivated Fetal Calf Serum) (Gibco; cat
748 # A38400-02). To generate a single cell suspension, DRGs were subjected to a 2 step-enzymatic
749 dissociation followed by a mechanical dissociation. In brief, DRGs were first incubated with
750 0.125% collagenase P from Clostridium histolyticum (Roche Applied Science; cat #
751 11249002001) for 90 minutes in an Eppendorf Thermomixer C (37°C; intermittent 750 rpm
752 shaking for about 10 sec every 2 minutes). Then, DRGs were transferred to a Hank's Balanced
753 Salt Solution (HBSS, Mg^{2+} and Ca^{2+} free; Invitrogen) supplemented with 0.25% Trypsin
754 (Worthington biochemical corp.; cat # LSoo3707) and 0.0025% EDTA and incubated for 10
755 minutes at 37°C in the Eppendorf Thermomixer C. Trypsin was neutralized by the addition of 2.5
756 mg/ml MgSO4 (Sigma; cat #M-3937) and DRGs were triturated with Pasteur pipettes. The
757 resulting cell suspension was passed through a 70 µm mesh filter to remove remaining chunks
758 of tissues and centrifuged for 5 minutes at 2500 rpm at room temperature. The pellet was
759 resuspended in HBSS ($Ca^{2+,}$ Mg^{2+} free; Invitrogen) and the cell suspension was run on a 30%
760 Percoll Plus gradient (Sigma GE17-5445-02) to further remove debris. Finally, cells were
761 resuspended in PBS supplemented with 0.04% BSA at a concentration of 200 cells/µl and cell
762 viability was determined using the automated cell analyzer NucleoCounter® NC-250™. The
763 suspended single cells were loaded on a Chromium Single Cell Instrument (10X Genomics) with
764 about 6000 cells per lane to minimize the presence of doublets. 2000-3000 cells per lane were
765 recovered. RNA-Seq libraries were constructed using Chromium Single Cell 3' Library, Gel Beads
766 & Multiplex Kit (10X Genomics). Single end sequencing was performed on Illumina NextSeq500.
767 Read 1 starts with a 26-bp UMI and cell barcode, followed by an 8-bp i7 sample index. Read 2
768 contains a 55-bp transcript read. Sample de-multiplexing, alignment, filtering, and UMI

769 counting were conducted using Cell Ranger Single-Cell Software Suite[63] (10X Genomics, v2.0.0).
770 Mouse mm10 Genome assembly and UCSC gene model were used for the alignment.
771
772 **Data preprocessing**
773 *DRG single cell data*
774 The UMI data output from the Cell Ranger Single-Cell Software Suite (10X Genomics, v2.0.0)
775 was analyzed using Seurat package[64] to assess the cell quality and identify cell types, similar to
776 what was described previously[42]. Cells with the number of detected genes less than 500 or over
777 15000, or with a UMI ratio of mitochondria encoded genes versus all genes over 0.1 were also
778 removed. The UMI data was normalized by the 'NormalizeData' method in Seurat with default
779 settings. To avoid potential sample-to-sample variation caused by the technical variation at
780 various experiment steps, we employed Seurat data integration method. The top 2000 variable
781 genes of each of the 5 DRG samples were identified using 'FindVariableFeatures' with
782 selection.method='vst'. Based on the union of these variable genes, the anchor cells in each
783 sample were identified by 'FindIntegrationAnchors'. All the samples were then integrated by
784 'IntegrateData'. We subsequently scaled the integrated data ('ScaleData') and performed
785 dimension reduction ('RunPCA'). Cells were then clustered based on the first 15 principal
786 components by applying 'FindNeighbors' and 'FindClusters' (resolution=0.6, algorithm=1).
787 Marker genes for each cluster were identified using 'FindAllMarkers'. Parameters were used
788 such that these genes were expressed in at least 25% of the cells in the cluster, and on average
789 2-fold higher than the rest of cells with a multiple-testing adjusted Wilcoxon test p value of less
790 than 0.01. The specificity of the canonical cell type-specific genes or cell cluster-specific genes

791 were further examined by visualizations (Supplementary Fig. 4) and used to define the cell type
792 for each cluster. At the end, the original UMI data from 17271 genes and 3352 cells that passed
793 the quality control were organized into a matrix (genes as rows and cell identifiers as columns).
794 This matrix, together with the cell type label for each cell therein, were loaded into AdRoit as
795 the reference profiles.
796
797 *Mouse brain single cell atlas data*
798 The scRNA-Seq reference data of the mouse brain were obtained from Zeisel et. al[32]. Among all
799 the available data, we only retained 96,572 cells that were acquired from the brain regions, had
800 an assigned cell type by the authors and a minimal total UMI of 1000. These cells corresponded
801 to 183 clusters at the finest taxonomy level in the original study. As many of the clusters are
802 highly similar, we decided to merge some of them to simplify the reference landscape. First, the
803 top 50 cluster enriched markers were derived using Scanpy[65] via the 'rank_genes_groups'
804 function (method='wilcoxon'), following the normalization ('normalize_per_cell'), log
805 transformation ('log1p') and regressing out ('regress_out') the variances associated with the
806 total UMI and the percentage of mitochondrial chromosome encoded genes per cell. Then, the
807 pair-wise overlapping p-values among the clusters were calculated using the top 50 marker
808 genes assuming a hypergeometric null distribution. Last, clusters with overlapping p-values
809 more significant than 1e-10 were merged and new names were assigned by combinedly
810 considering the original annotation, the molecular features and the specificity to certain brain
811 regions. A total of 46 cell types were determined that cover all the 12 brain regions and their
812 important substructures[40] (Supplementary Table 7). To make the reference dataset more

813 manageable in size and more balanced in the representation of cell types, we down sampled
814 each cluster to no more than 360 cells. A final set of 14,666 cells over 46 cell types was used for
815 the deconvolution of the mouse brain spatial transcriptome data.
816
817 *Human Islets single cell and bulk RNA-Seq data*
818 We used 1492 high quality human islets single cells and the associated annotations from Xin et
819 al[41]. The RPKM expression table was directly downloaded and used as-is. The RNA-FISH data
820 was also from this study[41]. The real bulk RNA-Seq data of human islets was acquired from a
821 large scale study referred in multiple publications[41,43,44]. We only included the data from donors
822 with a valid HbA1C level measurement in the regression analysis of the Beta cell proportion
823 with respect to the HbA1C level (Fig. 5c, Supplementary Table 13).
824
825 *Trabecular Meshwork single cell data*
826 We acquired the UMI counts, tSNE coordinates and the cell type annotation of 8759 high
827 quality cells from the authors of Patel et al[42]. Details of the quality control and the cell type
828 identification were presented in the original publication.
829
830 *Mouse Brain Spatial transcriptomics data by 10x Visium platform*
831 The filtered count matrix, tissue image and the spatial coordinates of a coronal section of an
832 adult C57BL/6 mouse brain were download from the 10x Genomics website (see Data
833 availability). A total of 2698 in-tissue spots were provided in the data set and used as-is.
834

835 *Mouse Brian ISH images*

836 The ISH images were directly downloaded from Allen mouse Brain Atlas[40] by searching the gene
837 names. The images were used without further editing except for cropping.

838

839 Data availability

840 The mouse DRG single cell data were deposited at NCBI GEO (accession number: GSE163252).
841 The bulk RNA-Seq and RNA-FISH data for human pancreatic islets were initially published as
842 aggregated data where the data processing and experimental procedure were described
843 therein[41,43,44]. We acquired the individual sample data from the authors and released them
844 along with the current study (Supplementary Table 10 and Supplementary Table 12). The other
845 public data analyzed in this study were obtained using GEO accession number GSE81608
846 (human pancreatic islets single cell data), NCBI SRA accession number PRJNA616025 (human
847 trabecular meshwork single cell data), and NCBI SRA accession number SRP135960 (mouse
848 brain single cell data). The 10x Genomics PBMC data was downloaded from the link
849 (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k), and
850 the Visium mouse brain spatial transcriptomic data was downloaded from the link
851 (https://support.10xgenomics.com/spatial-gene-
852 expression/datasets/1.1.0/V1_Adult_Mouse_Brain_Coronal_Section).

853

854 Code availability

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

856

857 ## Software

858 The statistical analyses were done with R statistical software (v3.6.0)[61] and python (v3.7.2)[66].
859 The packages used include Seurat (v3.0.1)[64], scanpy (v1.6.0)[65], dplyr (v0.8.0.1)[67], doParallel
860 (v1.0.14)[68], data.table (v1.12.4)[69], fitdistrplus (v1.1-1)[58], nnls (v1.4)[59], MuSiC (v0.1.1)[22],
861 BisqueRNA (v1.0.4)[21], SPOTlight (v1.0.4)[38], Cell2location (v0.05-alpha)[39], Stereoscope(v_03)[23].

862

863 ## Reference

864 1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics.
865 *Nature Reviews Genetics* (2009) doi:10.1038/nrg2484.
866 2. Chu, G. C., Kimmelman, A. C., Hezel, A. F. & DePinho, R. A. Stromal biology of pancreatic
867 cancer. *Journal of Cellular Biochemistry* (2007) doi:10.1002/jcb.21209.
868 3. Bussard, K. M., Mutkus, L., Stumpf, K., Gomez-Manzano, C. & Marini, F. C. Tumor-
869 associated stromal cells as key contributors to the tumor microenvironment. *Breast*
870 *Cancer Research* (2016) doi:10.1186/s13058-016-0740-2.
871 4. Munn, D. H. & Bronte, V. Immune suppressive mechanisms in the tumor
872 microenvironment. *Current Opinion in Immunology* (2016)
873 doi:10.1016/j.coi.2015.10.009.
874 5. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: From tumor
875 initiation to metastatic progression. *Genes and Development* (2018)
876 doi:10.1101/GAD.314617.118.
877 6. Garner, H. & de Visser, K. E. Immune crosstalk in cancer progression and metastatic
878 spread: a complex conversation. *Nature Reviews Immunology* (2020)

879 doi:10.1038/s41577-019-0271-z.

880 7. Singh, U. P. *et al.* Chemokine and cytokine levels in inflammatory bowel disease patients.
881 *Cytokine* (2016) doi:10.1016/j.cyto.2015.10.008.

882 8. Van Lint, P. & Libert, C. Chemokine and cytokine processing by matrix metalloproteinases
883 and its effect on leukocyte migration and inflammation. *J. Leukoc. Biol.* (2007)
884 doi:10.1189/jlb.0607338.

885 9. Zelová, H. & Hošek, J. TNF-α signalling and inflammation: Interactions between old
886 acquaintances. *Inflammation Research* (2013) doi:10.1007/s00011-013-0633-0.

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

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

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

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

897 14. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*
898 (2009) doi:10.1038/nmeth.1315.

899 15. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in
900 single-cell and single-nucleus RNA-Seq workflows. *Genome Biol.* (2020)

901 doi:10.1186/s13059-020-02048-6.

902 16. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental considerations
903 for single-cell RNA sequencing approaches. *Frontiers in Cell and Developmental Biology*
904 (2018) doi:10.3389/fcell.2018.00108.

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

907 18. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of
908 blood microarray data identifies cellular activation patterns in systemic lupus
909 erythematosus. *PLoS One* (2009) doi:10.1371/journal.pone.0006098.

910 19. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
911 *Nat. Methods* (2015) doi:10.1038/nmeth.3337.

912 20. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas
913 Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* (2016)
914 doi:10.1016/j.cels.2016.08.011.

915 21. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression
916 data. *Nat. Commun.* (2019) doi:10.1038/s41467-019-10802-z.

917 22. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution
918 with multi-subject single-cell expression reference. *Nat. Commun.* (2019)
919 doi:10.1038/s41467-018-08023-x.

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

922 24. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues

923 with digital cytometry. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0114-2.

924 25. Myung, I. J. Tutorial on maximum likelihood estimation. *J. Math. Psychol.* (2003)
925 doi:10.1016/S0022-2496(02)00028-7.

926 26. Bassett, R. & Deride, J. Maximum a posteriori estimators as a limit of Bayes estimators.
927 *Math. Program.* (2019) doi:10.1007/s10107-018-1241-0.

928 27. Zhao, Y. & Simon, R. Gene expression deconvolution in clinical samples. *Genome*
929 *Medicine* (2010) doi:10.1186/gm214.

930 28. Chiu, Y. J., Hsieh, Y. H. & Huang, Y. H. Improved cell composition deconvolution method
931 of bulk gene expression profiles to quantify subsets of immune cells. *BMC Med.*
932 *Genomics* (2019) doi:10.1186/s12920-019-0613-5.

933 29. Kang, K. *et al.* CDSeq: A novel complete deconvolution method for dissecting
934 heterogeneous samples using gene expression data. *PLoS Comput. Biol.* (2019)
935 doi:10.1371/journal.pcbi.1007510.

936 30. Qiao, W. *et al.* PERT: A Method for Expression Deconvolution of Human Blood Samples
937 from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.*
938 (2012) doi:10.1371/journal.pcbi.1002838.

939 31. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of
940 cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* (2019)
941 doi:10.1038/s41467-019-09990-5.

942 32. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* (2018)
943 doi:10.1016/j.cell.2018.06.021.

944 33. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular

945 deconvolution of GTEx tissues powers discovery of disease and cell-type associated
946 regulatory variants. *Nat. Commun.* (2020) doi:10.1038/s41467-020-14561-0.
947 34. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA
948 sequencing protocols. *F1000Research* (2017) doi:10.12688/f1000research.11290.1.
949 35. Chen, G., Ning, B. & Shi, T. Single-cell RNA-Seq technologies and related computational
950 data analysis. *Frontiers in Genetics* (2019) doi:10.3389/fgene.2019.00317.
951 36. Chen, D. & Plemmons, R. J. Nonnegativity constraints in numerical analysis. in *The Birth*
952 *of Numerical Analysis* (2009). doi:10.1142/9789812836267_0008.
953 37. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust
954 integration of single-cell information. *Nat. Commun.* (2020) doi:10.1038/s41467-020-
955 15816-6.
956 38. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF
957 regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.
958 *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab043.
959 39. Kleshchevnikov, V. *et al.* Comprehensive mapping of tissue cell architecture via
960 integrated single cell and spatial transcriptomics. *bioRxiv* (2020).
961 40. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature*
962 (2007) doi:10.1038/nature05453.
963 41. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes.
964 *Cell Metab.* (2016) doi:10.1016/j.cmet.2016.08.018.
965 42. Patel, G. *et al.* Molecular taxonomy of human ocular outflow tissues defined by single-
966 cell transcriptomics. *Proc. Natl. Acad. Sci.* **117**, 12856 LP – 12867 (2020).

967 43. Xin, Y. *et al.* Pseudotime ordering of single human B-cells reveals states of insulin
968 production and unfolded protein response. *Diabetes* (2018) doi:10.2337/db18-0365.
969 44. Gutierrez, G. D. *et al.* Gene signature of proliferating human pancreatic a cells.
970 *Endocrinology* (2018) doi:10.1210/en.2018-00469.
971 45. Cerf, M. E. Beta cell dysfunction and insulin resistance. *Frontiers in Endocrinology* (2013)
972 doi:10.3389/fendo.2013.00037.
973 46. Maedler, K. & Donath, M. Y. Beta-cells in type 2 diabetes: a loss of function and mass.
974 *Hormone research* (2004).
975 47. Donath, M. Y. *et al.* Mechanisms of β-cell death in type 2 diabetes. *Diabetes* (2005)
976 doi:10.2337/diabetes.54.suppl_2.S108.
977 48. Calanna, S. *et al.* Alpha- and beta-cell abnormalities in haemoglobin A1c-defined
978 prediabetes and type 2 diabetes. *Acta Diabetol.* (2014) doi:10.1007/s00592-014-0555-5.
979 49. Kanat, M. *et al.* The Relationship Between β-Cell Function and Glycated Hemoglobin.
980 *Diabetes Care* **34**, 1006 LP – 1010 (2011).
981 50. Nepton, S. Beta-Cell Function and Failure. in *Type 1 Diabetes* (2013). doi:10.5772/52153.
982 51. Dolenšek, J., Rupnik, M. S. & Stožer, A. Structural similarities and differences between
983 the human and the mouse pancreas. *Islets* (2015) doi:10.1080/19382014.2015.1024405.
984 52. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature*
985 **445**, 168–176 (2007).
986 53. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of
987 single cell RNA-Seq analysis pipelines. *Nat. Commun.* (2019) doi:10.1038/s41467-019-
988 12266-7.

989 54. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression
990 with prior derived from single-cell data. *Genome Res.* (2021) doi:10.1101/gr.268722.120.
991 55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome*
992 *Biol.* (2010) doi:10.1186/gb-2010-11-10-r106.
993 56. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
994 Seq data using regularized negative binomial regression. *Genome Biol.* (2019)
995 doi:10.1186/s13059-019-1874-1.
996 57. Svensson, V. Droplet scRNA-Seq is not zero-inflated. *Nature Biotechnology* (2020)
997 doi:10.1038/s41587-019-0379-5.
998 58. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R package for fitting distributions. *J.*
999 *Stat. Softw.* (2015) doi:10.18637/jss.v064.i04.
1000 59. Mullen, Katharine M., I. H. M. van S. nnls: The Lawson-Hanson algorithm for non-
1001 negative least squares (NNLS). *R Packag. version 1.4* (2012).
1002 60. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A Limited Memory Algorithm for Bound
1003 Constrained Optimization. *SIAM J. Sci. Comput.* (1995) doi:10.1137/0916069.
1004 61. The R Core Team. R: A Language and Environment for Statistical Computing. *R*
1005 *Foundation for Statistical Computing* (2019).
1006 62. Alessandri-Haber, N. *et al.* Hypotonicity induces TRPV4-mediated nociception in rat.
1007 *Neuron* (2003) doi:10.1016/S0896-6273(03)00462-8.
1008 63. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat.*
1009 *Commun.* (2017) doi:10.1038/ncomms14049.
1010 64. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019)

1011 doi:10.1016/j.cell.2019.05.031.

1012 65. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data

1013 analysis. *Genome Biol.* (2018) doi:10.1186/s13059-017-1382-0.

1014 66. van Rossum, G. & Drake, F. L. *Python 3 Reference Manual. Scotts Valley, CA* (2009).

1015 67. Wickham, H. & Francois, R. dplyr: A Grammar of Data Manipulation. *R Packag. version*

1016 *0.4.2.* (2015).

1017 68. Weston, S., Calaway, R. & Tenenbaum, D. *doParallel: Foreach Parallel Adaptor for the*

1018 *Parallel Package*. *Cran* (2014).

1019 69. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. R Package Version

1020 1.12.8. *Manual* (2019).

1021

1022 Acknowledgements

1023 We thank Yurong Xin for pointing us to the relevant public data resource. We also thank Gabor

1024 Halasz and Yuan Zhu for the helpful discussions regarding the memory usage, and Shawn

1025 Mishra for the manuscript proofreading.

1026

1027 Author contributions

1028 Y.B., T.Y., N.A.-H., M.L.-F., L.E.M. and G.S.A. designed the research. T.Y., Y.B., and W.F.

1029 developed the algorithm. T.Y., Y.B. and J.K. participated in the data analysis. M.S. and R.B.

1030 performed the DRG tissue collection. C.A. performed the single cell library preparation and

1031 sequencing experiment. Y.B., T.Y. and N.A.-H. wrote the manuscript.

1032

1033 ## Competing interests

1034 T.Y., Y.B., W.F. and G.S.A. have filed a patent application relating to the AdRoit computational
1035 framework. M.L.-F. is an employee of Cellular Longevity. All other authors are employees
1036 and/or shareholders of Regeneron Pharmaceuticals, although the manuscript's subject matter
1037 does not have any relationship to any products or services of this corporation.

1038

1039 ## Figure legends

1040 **Fig. 1: Schematic representation of AdRoit computational framework. a,** AdRoit inputs
1041 compound (bulk or spatial) RNA-Seq data, single cell RNA-Seq data and cell type annotations. It
1042 first selects informative genes and estimates their means and dispersions, then computes the
1043 cell type specificity of genes. Depending on the availability of multiple samples, cross-sample
1044 gene variability is derived from either the compound RNA-Seq, or the single cell data (dashed
1045 arrow; see also Methods). Lastly the gene-wise scaling factors are computed to correct the
1046 platform bias between the compound and the single cell RNA-Seq data. These quantities are
1047 used in a weighted regularized model to infer the cell type composition. **b,** A mock example to
1048 illustrate the role of the gene-wise scaling factor. Ideally, an accurate estimation of the cell
1049 proportions should be represented by the slope of the green line, however fitting in the
1050 presence of outlier genes would result in the red line. Outlier genes exist because the platform
1051 bias affects genes differently. AdRoit adopts an adaptive learning approach that first learns a
1052 coarse estimation of the slope (red line), from which the gene-wise scaling factors are derived
1053 and applied to the outlier genes, moving them toward the green line. The more deviated the

1054 gene, the larger the correction (i.e., longer arrows). After the adjustment, the new estimated
1055 slope (blue line) is closer to the truth (green line), thus is a more accurate estimation.
1056
1057 **Fig.2: Benchmark on simulated bulk data generated from the trabecular meshwork (TM)**
1058 **single cells. a,** AdRoit has the closest estimation to the true cell proportion comparing to
1059 Bisque, MuSiC and SPOTlight. Each dot is a cell type from a donor. **b,** For each cell type in TM,
1060 AdRoit has the smallest differences from the true cell type proportion and the smallest variance
1061 of estimates across the 8 donors. For each cell type, a dot on the graph denotes a donor, and
1062 the bars represent the $1.5 \times$ interquartile ranges. The reference used for deconvoluting each
1063 synthetic bulk sample excludes the cells used to synthesize that sample (leave-one-out).
1064
1065 **Fig. 3: AdRoit can achieve a high granularity and exhibits high sensitivity and specificity in**
1066 **complex tissues.** a. AdRoit is accurate in deconvoluting the simulated bulk samples that contain
1067 a mixture of similar cell types from myeloid or lymphoid lineage. The vertical dashed lines
1068 indicate the true mixing proportions. $CD14^+$ monocytes, $FCGR3A^+$ monocytes and dendritic cells
1069 (DC) were mixed under three schemes of proportions: 0.33:0.33:0.33 (mix0), 0.1:0.45:0.45
1070 (mix1) and 0.1:0.3:0.6 (mix2). The same ratios were applied to the mixtures of naïve $CD4^+$ T,
1071 memory $CD4^+$ T and $CD8^+$ cells. **b,** AdRoit's estimates are more accurate and specific than those
1072 from Bisque, MuSiC and SPOTlight on synthetic samples that contain only 6 out of the 12 cell
1073 types. The deconvolution was done using all 12 cell types as the reference. A pair of size-
1074 matched blue (true value) and red (estimated value) bubbles indicate an accurate prediction.
1075 Red-only and blue-only bubbles mark false positives and false negatives, respectively. **d,**

1076 Receiver operating characteristic (ROC) curve shows AdRoit has a significantly higher area
1077 under the curve (AUC) than other methods, meaning better sensitivity and specificity. **e,**
1078 Scatterplots between the ground truth and the deconvoluted cell proportions in the simulated
1079 bulk samples of high complexity (mixtures of 30 cell types). **f,** ROC curves show AdRoit has the
1080 best AUC among all methods on highly complex cell constitutions.
1081
1082 **Fig. 4: Benchmark on scRNA-Seq data from dorsal root ganglion (DRG) where many closely**
1083 **related subtypes of neuronal cells exist. a,** 14 cell types are identified from scRNA-Seq samples
1084 of 5 mice, including multiple subtypes of neurofilaments (NF), peptidergic (PEP), and non-
1085 peptidergic (NP) neurons. **b,** Benchmarking with the synthetic data shows the cell type
1086 proportions inferred by AdRoit are highly accurate. In particular, AdRoit remains a high
1087 accuracy when the cells are rare (e.g., < 5%; see also the zoom-in inserts). Each dot represents a
1088 cell type from one sample. **c,** For each sample, mAD, RMSD, Pearson and Spearman correlations
1089 are compared across four methods. AdRoit has the lowest mAD and RMSD, and the highest
1090 Pearson and Spearman correlations. In addition, AdRoit's estimation is the most stable across
1091 samples. Each dot on the boxplot is a sample. Same samples are chained by the dotted lines
1092 across the metrics. The deconvolution was done by using the leave-one-out strategy.
1093
1094 **Fig. 5: AdRoit is more accurate and sensitive on spatial spots simulated from real DRG cells. a,**
1095 Estimations from AdRoit, Cell2location,Stereoscope and SPOTlight on simulated spatial spots
1096 that contain 5 PEP neuron subtypes. True mixing proportions are denoted by the red dashed
1097 lines. Three schemes are presented: 1) the proportions of 5 PEP cell types are the same and

1098 equal to 0.2; 2) PEP1_Dcn is 0.1 and the other 4 are 0.225; 3) PEP1_Dcn and
1099 PEP1_S100a11.Tagln2 are 0.1, PEP1_Slc7a3.Sstr2 and PEP2_Htr3a.Sema5a 0.2 are 0.2, and
1100 PEP3_Trpm8 is 0.4. **b,** The performance of AdRoit, Cell2location, Stereoscope, and SPOTlight in
1101 estimating rare cell populations in the spatial spots. The spots contain a mixture of three PEP
1102 cell subtypes (i.e., PEP1_Slc7a3.Sstr2, PEP2_Htr3a.Sema5a and PEP3_Trpm8), with the percent
1103 of PEP3_Trpm8 ranging from 1% to 10% and the other two cell types sharing the remaining
1104 proportion equally. **c,** Compare the rate of detecting rare cells in simulated spots. An inferred
1105 percent greater than 0.5% is deemed as a positive detection. Six sets of cell mixtures are
1106 employed: NF_Calb1 with NF_Pvalb and NF2_Ntrk2.Necab2 (NP subtypes), NP_Mrgpra3 with
1107 NP_Mrgprd and NP_Nts (NP subtypes), PEP3_Trpm8 with PEP1_Slc7a3.Sstr2 and
1108 PEP2_Htr3a.Sema5a (PEP subtypes), NF_Calb1 with Th, satellite glia and endothelial (NF_Calb1
1109 + others), NP_Mrgpra3 with Th, satellite glia and endothelial (NP_Mrgpra3 + others), and
1110 PEP_Trpm8 with Th, satellite glia and endothelial (PEP_Trpm8 + others). In each set, the first
1111 cell type listed is the target of detection and varies its percent from 1% to 10%. The rest cell
1112 types split the remaining proportion evenly. The red dashed lines mark the detection rate of
1113 90%.
1114
1115 **Fig. 6: Applications to real-world bulk RNA-Seq data and mouse brain spatial transcriptome**
1116 **data. a,** The deconvoluted cell compositions in the real bulk RNA-Seq data of human Islets are
1117 highly reproducible for the repeated samples from the same donor. **b,** AdRoit estimation of the
1118 cell type proportions agrees with the RNA-FISH measurements. **c,** AdRoit-inferred Beta cell
1119 proportions in type 2 diabetes patients are significantly lower than those in healthy subjects. In

1120 addition, the estimated proportions have a significant negative linear association with the
1121 donor HbA1C levels. **d,** The spatial mapping of four mouse brain cell types is consistent with the
1122 locations of four region-specific markers shown on the ISH images obtained from Allen mouse
1123 brain atlas[40]. The four genes, Spink8, C1ql2, Clic6 and Synpo2, were identified by Zeisel et al[32] as
1124 markers of the hippocampal field CA1, dentate gyrus, choroid plexus and thalamus,
1125 respectively.
1126
1127
1128
1129
1130
1131
1132
1133