HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

Nat Biotechnol. 2022 April ; 40(4): 517–526. doi:10.1038/s41587-021-00830-w.

Robust decomposition of cell type mixtures in spatial transcriptomics

Dylan M. Cable[1,2,3], **Evan Murray**[2], **Luli S. Zou**[2,3,4], **Aleksandrina Goeva**[2], **Evan Z. Macosko**[2,5], **Fei Chen**[2,6,*], **Rafael A. Irizarry**[3,4,*]

[1]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139

[2]Broad Institute of Harvard and MIT, Cambridge, MA, 02142

[3]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, 02215

[4]Department of Biostatistics, Harvard University, Boston, MA, 02115

[5]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, 02114

[6]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge MA 02138

Abstract

A limitation of spatial transcriptomics technologies is that individual measurements may contain contributions from multiple cells, hindering the discovery of cell type-specific spatial patterns of localization and expression. Here, we develop Robust Cell Type Decomposition (RCTD), a computational method that leverages cell type profiles learned from single-cell RNA-seq to decompose cell type mixtures, while correcting for differences across sequencing technologies. We demonstrate RCTD's ability to detect mixtures and identify cell types on simulated datasets. Furthermore, RCTD accurately reproduces known cell type and subtype localization patterns in Slide-seq and Visium datasets of the mouse brain. Finally, we show how RCTD's recovery of cell type localization enables the discovery of genes within a cell type whose expression depends on spatial environment. Spatial mapping of cell types with RCTD enables defining spatial components of cellular identity, uncovering new principles of cellular organization in biological tissue. RCTD is publicly available as an open source R package at https://github.com/dmcable/RCTD.

Editorial summary:

Correspondence to: rafa@ds.dfci.harvard.edu, chenf@broadinstitute.org.
*These authors contributed equally

Author Contributions

D.M.C., F.C., R.A.I, and E.Z.M. conceived the study; F.C., E.M., and E.Z.M. designed the Slide-seq experiment; E.M. generated the Slide-seq data; D.M.C., R.A.I., and F.C. developed the statistical methods; D.M.C., F.C., R.A.I, and E.Z.M. designed the analysis; D.M.C., R.A.I., F.C, A.G., and L.S.Z. analyzed the data; D.M.C., F.C., R.A.I., E.Z.M., and L.S.Z. wrote the manuscript; all authors read and approved the final manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

Code Availability Statement

RCTD is implemented in the open-source R package RCTD, with source code freely available at https://github.com/dmcable/RCTD. Additional code used for analysis in this paper is available at https://github.com/dmcable/RCTD/tree/dev/AnalysisPaper.

Cell-type mapping in spatial transcriptomics is enabled by accounting for compositional mixtures and differences in sequencing technologies.

Introduction

Tissues are composed of diverse cell types and states whose spatial organization governs interaction and function. Recent advances in spatial transcriptomics technologies [1–3] have enabled high throughput collection of RNA-sequencing coupled with spatial information in biological tissues. Using such technologies to spatially map cell types is fundamental to our understanding of tissue structure. In particular, knowledge of spatial localization of specific cellular subtypes remains incomplete and laborious to obtain [4,5].

Spatial transcriptomics technologies have the potential to elucidate interactions between cellular environment and gene expression, augmenting our knowledge of healthy functions and disease states of tissues. Spatial transcriptomics data is composed of gene expression counts for each of the spatial measurement locations, here referred to as *pixels*, that tile a two dimensional surface. A common task of interest is identifying genes with expression varying across space. Current computational methods search for spatial patterns in gene expression without stratifying by cell type [6–8]. However, much of the variation detected by these methods may be driven by varying cell type composition across the spatial landscape, since single-cell RNA sequencing (scRNA-seq) studies have revealed that cell type can explain a majority of the variation within a population of cells [9,10]. It is therefore necessary to consider cell type information when searching for spatial gene expression patterns.

Assignment of cell types is analytically challenging, even for high-resolution approaches such as Slide-seq, due to the fact that although pixel resolution can approach the size of mammalian cells (e.g. Slide-seq, 10 microns) [11], fixed pixel locations may overlap with multiple cells. As a result, gene expression measurements at a single pixel may be the result of a mixture of multiple cell types. Currently, the most widely used approach to identifying cell types relies on unsupervised clustering [12]; however, this approach does not allow for the possibility of cell type mixtures. A fundamental challenge is thus to correctly identify these mixture pixels as a combination of multiple cell types, permitting a more complete characterization of the spatial localization of cell types in spatial transcriptomics.

Several recent methods have used scRNA-seq references to predict cell types on spatial transcriptomics data [11,13]; however, some do not statistically model platform effects, Poisson sampling, and overdispersed counts [11], despite recent evidence that methods in scRNA-seq data analysis accounting for these aspects of gene expression count data outperform those assuming normality [14,15]. Others have not yet been demonstrated to scale to large datasets, such as those obtained by Slide-seq, and they perform cell type enrichment testing at the level of regions rather than individual pixels [13].

Here, we introduce Robust Cell Type Decomposition (RCTD), a supervised learning approach to decompose RNA sequencing mixtures into single cell types, enabling assignment of cell types to spatial transcriptomic pixels. Specifically, we leverage annotated

scRNA-seq data to define cell type-specific profiles for the cell types expected to be present in the spatial transcriptomics data. Several supervised cell type assignment methods have achieved high accuracy in scRNA-seq [12,16], but are not designed for mixtures of multiple cell types. RCTD fits a statistical model that estimates mixtures of cell types at each pixel.

A pertinent challenge for supervised cell type learning is what we term *platform effects*: the effects of technology-dependent library preparation on the capture rate of individual genes between sequencing platforms. We show that if these platform effects are not accounted for, supervised methods are unlikely to succeed since systematic technical variability dominates relevant biological signals [17]. These effects have been previously found in comparisons between single-cell and single-nucleus RNA-seq on the same biological sample [18], where it has been shown that e.g. nucleus-localized genes are enriched in single-nucleus RNA-seq. Here, we demonstrate that platform effects between the scRNA-seq reference and spatial transcriptomics target present a challenge when transferring cell type knowledge to spatial transcriptomics. To enable cross-platform learning in RCTD, we have developed and validated a platform effect normalization procedure.

We demonstrate that RCTD can accurately discover localization of cell types in both simulated and real spatial transcriptomic data. Furthermore, we show that RCTD can detect subtle transcriptomic differences to spatially map cellular subtypes. Finally, we use RCTD to compute expected cell type-specific gene expression, which enables detection of changes in gene expression based on the spatial environment of a cell. Below, we demonstrate how RCTD learns mixtures of cell types in spatial transcriptomics data, facilitating quantification of the effect of spatial position and local cellular environment on gene expression within a cell type.

Results

Challenges in Spatial transcriptomics: cell type mixtures and platform effects

Spatial transcriptomics pixels source RNA from multiple, rather than single, cells creating a challenge for cell type learning. In Slide-seq cerebellum data, we found that the most widely used approach for scRNA-seq cell type identification, unsupervised clustering [12], incorrectly classifies cell types that colocalize spatially but are not similar transcriptionally. For example, Bergmann and Purkinje cells spatially colocalize to the same layer, resulting in a population of pixels that possess marker genes from both cell types (Figure 1a). The most likely explanation for this observation is that these pixels contain two or more cells of different types, but unsupervised clustering assigns these *doublet* pixels to just one cell type. Moreover, this approach predicts granule cells not exclusively in the granular layer, with many cells incorrectly predicted inside the molecular layer and oligodendrocyte layer and possessing low granule marker expression (Figure 1b–c, Supplementary Figure 1–2).

An additional challenge, platform effects, arises in applying supervised learning, in which scRNA-seq cell type profiles are leveraged to classify spatial transcriptomic cell types. For instance, a standard supervised learning approach trained on an assessment single-nucleus RNA-seq cerebellum dataset with known cell types obtained much higher accuracy in the training platform than the testing platform, a single-cell RNA-seq cerebellum dataset (Figure

1d–e). This difference is explained by the presence of *platform effects*, which can cause gene expression to change multiplicatively between single-nucleus and single-cell RNA-seq (Figure 1f). NMFreg, a supervised cell-type mixture assignment algorithm previously developed for Slide-seq, also does not account for platform effects. Testing on the Slide-seq cerebellum dataset, NMFreg assigned a minority (24.8% out of $n = 11626$) of pixels confidently to cell types and mislocalized broad cell type classes (Supplementary Figure 3). Likewise, DWLS, a method designed for bulk RNA-seq deconvolution [19], does not account for platform effects and performed better at within-reference cell type classification than cross-platform cell type classification (Supplementary Figure 4).

Robust Cell Type Decomposition enables cross-platform detection of cell type mixtures

To address these challenges, RCTD accounts for platform effects while using a scRNA-seq reference to decompose each spatial transcriptomics pixel into a mixture of individual cell types. RCTD first calculates the mean gene expression profile of each cell type within the annotated scRNA-seq reference (Figure 2a). Next, RCTD creates a spatial map of cell types by fitting each spatial transcriptomics pixel as a linear combination of individual cell types. RCTD takes as input RNA-sequencing counts for each pixel and assumes an unknown mixture of multiple cells (Figure 2a). Each cell type contributes an unobserved proportion of counts to each gene. RCTD estimates the proportion of each cell type for each pixel by fitting a statistical model where, for each pixel i and gene j, the observed gene counts $Y_{i,j}$ are assumed to be Poisson-distributed. The rate parameter is determined by the pixel's total transcript count, N_i, and $\lambda_{i,j}$, a mixture of K cell type expression profiles:

$$Y_{i,j} \mid \lambda_{i,j} \sim \text{Poisson}\left(N_i \lambda_{i,j}\right).$$

To account for platform effects and other sources of natural variability, such as spatial variability, we assume $\lambda_{i,j}$ is a random variable defined by

$$\log\left(\lambda_{i,j}\right) = \alpha_i + \log\left(\sum_{k=1}^{K} \beta_{i,k} \mu_{k,j}\right) + \gamma_j + \epsilon_{i,j},$$

with $\mu_{k,j}$ the mean gene expression profile for cell type k, α_i a fixed pixel-specific effect, γ_j a gene-specific platform random effect and $\epsilon_{i,j}$ a random effect to account for gene-specific overdispersion.

We use maximum likelihood estimation to infer the cell type proportions, $\beta_{i,k}$, indicating which cell types are present in each pixel (see Methods for details). RCTD may be used without constraining the number of cell types per pixel or with what we refer to as *doublet mode*, which searches for the best fitting one or two cell types per pixel (see Methods for details). In particular, we refer to pixels as *singlets* if they contain only one cell type and *doublets* if they contain two cell types. Doublet mode may mitigate overfitting if mixtures of three or more cell types are expected to be rare, as we found in Slide-seq (Supplementary Figure 5). We have also extended doublet mode to optionally fit more than two cell types per pixel (Methods).

Because gene-specific platform effects are not observable from the raw data, we developed a procedure to estimate platform effects between sequencing platforms with RCTD (Methods, Supplementary Table 1). Training RCTD on the single-nucleus RNA-seq cerebellum reference and testing on the single-cell RNA-seq cerebellum dataset, we validated that our approach is able to reliably recover the platform effects (R^2 = 0.90) (Figure 2b). After normalizing cell type profiles for platform effects, RCTD achieved high cross-platform single-cell classification accuracy (89.5% of n = 3960 cells) (Figure 2c). Transcriptomically similar cell types, e.g. oligodendrocytes/polydendrocytes, accounted for most of the remaining errors (91.8% of n = 415 errors).

Because ground-truth cell type identities are not known in spatial transcriptomics datasets, we benchmarked RCTD's performance on single-nucleus and single-cell RNA-sequencing datasets with ground-truth cell types obtained from previous studies [20,21]. To evaluate RCTD's ability to detect and decompose mixtures in spatial transcriptomics data in the presence of platform effects, we trained RCTD on the single-nucleus RNA-Seq (snRNA-seq) cerebellum reference (Supplementary Figure 6), and tested on a dataset of doublets simulated as computational mixtures of single cells with known cell types in the scRNA-seq dataset (See Methods for details). By varying the true underlying cell type proportion, we observed that RCTD correctly classified singlets (89.2% ± 0.5% s.e.) and doublets (81.1% ± 0.3% s.e.) with high accuracy (Figure 3a, Supplementary Figure 7). A large proportion of doublet misclassifications came from transcriptionally similar cell types appearing on the same doublet pixel, which RCTD often misclassified (87.0% ± 1.2% s.e.) as singlets (Supplementary Figure 7). Additionally, RCTD identified each cell class present on each doublet with 98.2% accuracy on confident calls (Methods, ± 2.8% s.d. across 66 cell type pairs) (Figure 3b, Supplementary Figure 8–9). Finally, RCTD accurately estimated the proportion of each cell type on the sample with 12.8% RMSE (± 6.9% s.d. across 66 cell type pairs) (Figure 3c–d). These technical validations show that RCTD can accurately learn cell type information in a dataset with mixtures of single cells.

Next, we extended our validation of RCTD on simulated data to additional contexts including varying unique molecular identifier (UMI) counts per pixel, more than two cell types per pixel, and missing cell types in the reference (Supplementary Figure 10–13). We found that additional UMIs per pixel led to an increased confidence rate, and RCTD achieved high classification accuracy on pixels containing ≥ 100 UMIs (Supplementary Figure 10). Moreover, we additionally found that RCTD was able to accurately predict cell class proportions on pixels containing three or four cell types, a typical regime in lower resolution spatial transcriptomics (e.g. Visium, Supplementary Figure 12–13). When cell types in the simulated spatial data were missing from the reference, RCTD classified pixels as the most transcriptionally similar cell type in the reference, if available (Supplementary Figure 10). When no closest cell type was available in the reference, RCTD predicted cell types with reduced confidence rates (Supplementary Figure 10), but often misclassified such pixels (Supplementary Figure 11).

RCTD localizes cell types in spatial transcriptomics data

We next applied RCTD to assign and decompose cell types in spatial transcriptomics data. We first applied RCTD to localize cell types in the mouse cerebellum, using a single-nucleus RNA-seq (snRNA-Seq) reference for training, and a Slide-seqV2 dataset collected on the adult mouse cerebellum as the target. RCTD confidently classified a majority (86.9%, out of $n = 11626$) of pixels, and the resulting cell type calls are consistent with the spatial architecture of the cerebellum (Figure 4a) [22]. Since ground truth cell type labels do not exist for spatial transcriptomic data, to assess the accuracy of RCTD, we used multiple validation strategies including comparison to marker genes and prior knowledge of spatial organization. While presence of marker genes should be expected to roughly correspond to cell type presence, we do not expect a perfect relationship and consequently look for marker gene presence in conjunction with prior biological knowledge. We first considered Purkinje/Bergman cells, two cell types which are spatially co-localized in the cerebellum. We found that RCTD's singlet pixels assigned to Purkinje or Bergmann cell types do not possess markers of the other cell type (Figure 4b). Moreover, pixels predicted as doublets contained marker genes of both Bergmann and Purkinje cells, with estimated cell type proportion correlating with marker gene ratio (Figure 4c). We next observed that RCTD correctly localized molecular layer interneurons to the molecular layer [20], granule cells to the granular layer, and oligodendrocytes to the white matter layer [22], predictions further supported by the spatial correspondence between RCTD's assignments and the marker genes of each cell type (Figure 4d, Supplementary Figure 14). Next, to validate RCTD's ability to correctly localize doublets, we leveraged the layered organization of the cerebellum (Figure 4e) [22]. RCTD finds doublets within a layer and between adjacent layers, but rarely between spatially separated layers (Figure 4f).

To test if RCTD achieved consistent results when trained on multiple datasets, we additionally trained RCTD on the single-cell RNA-seq cerebellum dataset and tested on Slide-seq cerebellum (Supplementary Figure 15). On confidently classified pixels, RCTD, trained on two different references, agreed on 95.7% of cell type predictions (Supplementary Figure 16). We additionally validated RCTD's ability to reproduce the layered structure of cortical cell types in a Slide-seqV2 dataset of the mouse somatosensory cortex, training on a Smart-seq2 reference (Methods) (Supplementary Figure 17–18) [1,23]. We found that cortical neuron subtypes appeared in appropriate layers, with L2/3 intratelencephalic (IT), followed by L4, followed by L5 IT and L5 pyramidal tract (PT), followed by L6 corticothalamic (CT) and L6 IT, followed by L6b, consistent with the results of additional studies [24].

RCTD discovers spatial localization of cellular subtypes

Next, we tested the ability of RCTD to profile the spatial localization of cellular subtypes, recently defined by large-scale transcriptomic analyses [21], for which there is limited knowledge of spatial position in their resident tissues. To this end, we validated RCTD's ability to classify previously defined [21] subtypes of interneurons in the hippocampus (Methods). We first used RCTD to spatially annotate cell types in Slide-seq data of the mouse hippocampus (Figure 5a), training on a scRNA-seq hippocampus dataset [21]. We found that RCTD correctly localizes hippocampal cell types (Supplementary Figure 19–20).

We also validated RCTD's ability to localize hippocampal cell types in a Visium spatial transcriptomics dataset (Supplementary Figure 20–21) [2] and found qualitative agreement between predicted cell type localization patterns on Slide-seq and Visium (Supplementary Figure 22). We then observed spatial clustering of pixels assigned to the broad class of interneurons (Figure 5b), which we inferred to be derived from large, single interneuron cells [4], an inference supported by histological examination [25] (Supplementary Figure 23). Consequently, we tested RCTD's performance in assigning pixels within a cluster to the same interneuron subclass and found high agreement (97.1% ± 0.09% s.e.) of coarse subclass classification between confident pixels within the same spatial cluster (Figure 5c, Methods). Additionally, we found that the spatial localization of the Basket/OLM subclass coincides with expression of *Sst*, a differentially expressed gene for this subclass (Figure 5d). Finally, we used RCTD to assign each spatial cluster to one of 27 transcriptomically defined interneuron subtypes, confidently classifying the majority of interneuron pixels (Figure 5e, Supplementary Figure 24). Localizations of known subtypes, such as CA1-Lacunosum, which appears in the stratum lacunosum-moleculare (SLM) layer of the CA1 [26], and OLM, which appears primarily in the stratum oriens (SO) [27], agree with known anatomy. We conclude that RCTD enables the identification of spatial locations of cellular subtypes in spatial transcriptomics data.

RCTD enables detection of spatially variable genes within cell type

Previous computational methods search for spatially variable genes without incorporating cell type information [6–8]. However, because cell types are not evenly distributed in space, and different cell types have different expression profiles, this approach will likely lead to confusing cell type marker genes with spatially variable genes. For example, we found that the 20 genes with the highest spatial autocorrelation in the Slide-seq hippocampus (Methods) were primarily expressed in only a few cell types, indicating that their spatial variation is partially driven by cell type composition (Figure 6a). After conditioning on cell type, a majority of these genes exhibited small remaining spatial variation (Figure 6b). For example, *Ptk2b* is differentially expressed in excitatory neurons, but does not exhibit any spatial variation that is unexplained by cell type alone (Figure 6c).

Instead, RCTD enables estimation of spatial gene expression patterns within each cell type. After identifying cell types, we used RCTD to compute the expected cell type-specific gene expression for each cell type within each pixel (see Methods for details). Using this cell type-specific expected gene expression, we detected genes with large spatial variation within CA3 pyramidal neurons (Figure 6b, $p \leq 0.01$, permutation F-test, Supplementary Table 2). For these genes, we recovered smooth patterns of gene expression over space with locally weighted regression (Figure 6d, see Methods for details). In addition to spatially variable genes, RCTD can be used to detect the effect of cellular environment on gene expression. In the hippocampus, RCTD detected astrocyte doublets with many cell types in distinct spatial regions (Figure 6e); we hypothesized that astrocytic transcriptomes could vary based on their cellular environment. We detected genes whose expression within astrocytes depended on co-localization with another cell type (Figure 6f–g, Methods, Supplementary Table 2). For instance, we found that *Entpd2* was enriched in astrocytes colocalizing with dentate neurons ($p = .025$, two-tailed z-test). This is consistent with a prior

study that detected a population of astrocyte-like progenitor cells in the dentate expressing *Entpd2* [28]. Moreover, *Slc6a11*, which enables uptake of the GABA neurotransmitter and likely modulates inhibitory synapses [29], was differentially expressed in astrocytes around excitatory neurons ($p < 10^{-6}$, two-tailed z-test) [30]. Thus, RCTD enables measurement of the effect of the cellular environment and space on gene expression.

Discussion

Accurate spatial mapping of cell types and detection cell type-specific spatial patterns of gene expression is critical for understanding tissue organization and function. Here, we introduce RCTD, a computational method for accurate decomposition of spatial transcriptomic pixels into mixtures of cell types, using a single-cell RNA-seq reference normalized for platform effects. RCTD takes as input RNA sequencing counts at each pixel containing an unknown mixture of multiple cells, and predicts the proportion of each cell type on each pixel. RCTD accurately maps cell types, as demonstrated on both a dataset of simulated doublets as well as cerebellum and hippocampus spatial transcriptomics datasets. We additionally demonstrated RCTD's ability to correctly localize subtypes in a Visium hippocampus spatial transcriptomics dataset, showing that RCTD can be applied broadly to different platforms. We further showed RCTD can spatially localize transcriptomically-defined cellular subtypes of interneurons of the hippocampus. Lastly, we demonstrated that RCTD enables discovery of spatially varying gene expression within cell types in the hippocampus.

As the cost of sequencing diminishes, scRNA-seq datasets are becoming more prevalent and easier to generate [31]. Individual scRNA-seq methods can be more or less similar to a spatial transcriptomics dataset in their platform effects, which can be measured by RCTD. For example, relative to Slide-seq, we found a lower magnitude of platform effects for the single-cell hippocampus reference than for the single-nucleus cerebellum reference. While spatial platform effects are hard to measure a priori, we have demonstrated our platform effect normalization procedure to be robust to the choice of reference (scRNA-seq, snRNA-seq, SMART-seq). We thus anticipate it to be compatible with future scRNA-seq modalities. Furthermore, our method is flexible to the choice of target platform. For example, our procedure for estimating platform effects depends only on merging all pixels into one *pseudo-bulk* measurement. Our method can consequently be applied to estimate platform effects from a scRNA-seq reference to any other sequencing technology, including bulk RNA sequencing, providing a generally-applicable normalization procedure for RNA sequencing. Although motivated by spatial transcriptomics, we expect that RCTD can learn cell types on other non-spatial datasets with single cells or mixtures of multiple cell types [32].

A limitation of RCTD is that it relies on an assumption that platform effects are shared among cell types. This is a general problem with reference-based cell type learning, and it will be important to explore learning cell type-specific platform effects in future work. We additionally found that a challenging problem for RCTD is cell types missing from the reference but present in the spatial data. This issue may be mitigated by cropping the spatial data to exclude regions known a priori to primarily contain cell types not present in the

reference. Future work includes improving our method to identify pixels with cell types out of reference.

When fine spatial resolution causes localization of three or more cell types to one pixel to be uncommon (e.g. Slide-seq [11]), we recommend using doublet mode of RCTD, which constrains at most two cell types per pixel. Otherwise, RCTD can be used to decompose any number of cell types per pixel (e.g. Visium). Similar in principle to AIC model selection methods [33], doublet mode reduces overfitting by penalizing the number of cell types used, improving RCTD's statistical power. This concept can be readily extended to triplets and beyond in future work.

A major goal of spatial transcriptomics is understanding the contributions of cell type and cellular environment on cell state. RCTD facilitates the discovery of these effects by computing expected cell type-specific gene expression for each spatial transcriptomics pixel. For instance, we analyzed gene expression within astrocytes to detect astrocytic genes influenced by local cellular environment. There are many drivers of a gene's dependence on cellular environment: cell-to-cell interactions, regional signalling factors, or cellular history during development. The ability of RCTD to localize cell types uniquely enables high-throughput generation of biologically-relevant hypotheses concerning the effects of space and environment on gene expression. As more spatial transcriptomics datasets are generated, we expect that RCTD will facilitate the discovery of new principles of cellular organization in biological tissue.

Methods

Statistical model

Here, we describe the statistical model used to perform Robust Cell Type Decomposition (RCTD) to identify mixtures of cell types. For each pixel $i = 1, \ldots, I$ in the spatial transcriptomics dataset, we denote the observed gene expression counts as $Y_{i,j}$ for each gene $j = 1, \ldots, J$. We model these counts with the following hierarchical model,

$$\begin{aligned} Y_{i,j} \mid \lambda_{i,j} &\sim \text{Poisson}\left(N_i \lambda_{i,j}\right) \\ \log\left(\lambda_{i,j}\right) &= \alpha_i + \log\left(\sum_{k=1}^{K} \beta_{i,k} \mu_{k,j}\right) + \gamma_j + \varepsilon_{i,j}, \end{aligned} \tag{1}$$

with N_i the total transcript count or number of unique molecular identifies (UMIs) for pixel i, K the number of cell types present in our dataset, α_i a fixed pixel-specific effect, $\mu_{k,j}$ the mean gene expression profile for cell type k and gene j, $\beta_{i,k}$ the proportion of the contribution of cell type k to pixel i, γ_j a gene-specific platform random effect and $\varepsilon_{i,j}$ a random effect to account for other sources of variation, such as spatial effects. By modeling Poisson noise, RCTD can account for sampling noise including when overall UMI counts are low ($\approx 100 - 1000$), such as in Slide-seq. Data exploration (Figure 1f) supported a Poisson-lognormal mixture, used previously for count data [34]. Thus, we assume γ_j and $\varepsilon_{i,j}$ both follow normal distributions with mean 0 and standard deviation σ_γ and σ_ε, respectively. We note that in practice we additionally modify the random effects distributions to include a heavier tail that is robust to outliers (using an approximation to a Cauchy-Gaussian mixture

distribution [35]; see supplementary methods for details). The main goal of our analysis is to estimate the $\beta_{i,k}$'s, which represent the cell type or cell types present in each pixel i, constrained so that $\sum_{k=1}^{K} \beta_{i,k} = 1$ and each $\beta_{i,k} \geq 0$.

Fitting the model

Model (1) is a complex model with thousands of parameters (many, $K \times J$, of these parameters are introduced by the cell type-specific gene expression profiles). We overcome this challenge by fitting our model using a stepwise approach that includes a supervised learning step for estimating these expression profiles, $\mu_{k,j}$. The steps of our estimation approach are as follows:

1. Supervised estimation of cell type profiles: We use a *reference* dataset, refered to as the *training* dataset, to obtain estimates for the mean gene expression profiles $\mu_{k,j}$. We refer to these estimates as $\hat{\mu}_{k,j}$, which are then considered fixed in the next steps.
2. Gene filtering: We use the estimated cell type profiles $\hat{\mu}_{k,j}$ to filter out genes that are unlikely to be informative. We do this by selecting genes that show differential expression across cell types.
3. Platform Effect Normalization: The random effects γ_j account for the unwanted technical variation resulting from gene expression profiles varying across different sequencing platforms. The next step is therefore to estimate σ_γ and predict γ_j for each gene j. We denote the prediction of the random effects as $\hat{\gamma}_j$, which are then considered fixed in the next step.
4. Robust Cell Type Decomposition: We use the plugin estimates $\hat{\mu}_{k,j}$ and $\hat{\gamma}_j$ and assume they are fixed. Conditional on these estimates, for each sample i and treating $\varepsilon_{i,j}$ as a random effects, we can compute the maximum likelihood estimate (MLE) for $\beta_{i,k}$, α_i, and σ_ε.

Next we describe each of these steps in detail.

Supervised estimation of cell type profiles

First, we obtain a single-cell RNA-seq reference, which has been previously annotated with cell types. We estimate $\hat{\mu}_{k,j}$ as the average normalized expression of gene j within all cells of cell type k.

Gene filtering

Using the estimated cell type expression profiles $\hat{\mu}_{k,j}$, we select differentially expressed genes that will be informative when estimating cell type proportions. For each cell type in the scRNA-seq reference, we select genes with minimum average expression above .0625 counts per 500 and at least 0.5 log-fold-change compared to the average expression across all cell types. Typically, this results in about 5,000 genes for the platform effect normalization step. These parameters are further increased for the Robust Cell Type Decomposition step, to reduce the set to about 3,000 genes for computational efficiency.

Platform effect normalization

Estimating the $\beta_{i,k}$ in the presence of the unobserved platform effects γ_j is challenging. However, γ_j can be reliably predicted independently from the other parameters by summarizing the spatial transcriptomics data as a single *pseudo-bulk* measurement $S_j \equiv \sum_{i=1}^{I} Y_{i,j}$. Notice that, conditioned on the rates $\lambda_{i,j}$, S_j is Poisson distributed with the average $\bar{Y}_j = \frac{1}{I} S_j$ having expectation:

$$\begin{aligned}\log\left\{\mathbb{E}\left(\bar{Y}_j \middle| \lambda_{1,j}, \ldots, \lambda_{I,j}\right)\right\} &= \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i \lambda_{i,j}\right) \\ &= \gamma_j + \log\left(\bar{N}\sum_{k=1}^{K} \mu_{k,j} B_{k,j}\right) \\ &\approx \gamma_j + \log\left(\bar{N}\sum_{k=1}^{K} \mu_{k,j}\beta_k\right) + \log(\beta_0)\end{aligned}$$

with

$$\beta_0 \text{ a scaling factor constant, } \bar{N} = \frac{1}{I}\sum_{i=1}^{I} N_i \text{ and } B_{k,j} = \frac{1}{I}\sum_{i=1}^{I} \frac{N_i}{\bar{N}} \beta_{k,i} \exp\left(\alpha_i + \varepsilon_{i,j}\right)$$

a random variable that is approximately proportional to $\beta_k = \frac{1}{I}\sum_{i=1}^{I} \frac{N_i}{\bar{N}} \beta_{i,k}\alpha_i$, the proportion of cell type k in our target dataset:

$$B_{k,j} \approx \beta_k \beta_0 .$$

This follows from the fact that $\mathbb{E}\left(B_{k,j}\right) = \beta_k\beta_0$, and $\mathrm{Var}(B_{k,j})$ converges to 0 when I is large (see supplementary methods for details). By plugging in the $\hat{\mu}_{i,j}$ obtained in the first step and treating them as known, we can then obtain the maximum likelihood estimator (MLE) for β_0, the β_k's, and σ_γ and subsequently estimate the platform effects γ_j as $\hat{\gamma}_j$.

Robust Cell Type Decomposition

With $\hat{\mu}_{k,j}$ and $\hat{\gamma}_j$ in place, we plug them into equation (1) which we can rewrite as,

$$Y_{i,j} | \varepsilon_{i,j} \sim \mathrm{Poisson}\left\{N_i \exp\left[\alpha_i + \log\left(\sum_{k=1}^{K} \beta_{i,k}\hat{\mu}_{k,j}\right) + \hat{\gamma}_j + \varepsilon_{i,j}\right]\right\} \quad (2)$$

$$\varepsilon_{i,j} \sim \mathrm{Normal}\left(0, \sigma_\varepsilon^2\right), \quad (3)$$

and we obtain the MLE α_i, $\beta_{i,k}$ and σ_ε. The algorithm implemented to find the MLE is in the supplementary methods, and we have validated its ability to find the MLE (Supplementary Figure 25).

Cell type identification by model selection

Notice that in the procedure described above, $\hat{\beta}_{i,k} > 0$ for as many as K cell types, implying that pixel i is a mixture of several cell types. However, for many spatial transcriptomics technologies, we do not expect more than two cell types per pixel. We therefore implemented a version of our model and estimation procedure that constrains the number of k's for which $\beta_{i,k} > 0$ to two. We refer to this version of method as *doublet* mode. In doublet mode, cell type identification is accomplished using a model selection framework, where we compare likelihoods and penalize the inclusion of an additional features. In this version of our method, we refer to the two possible outcomes as *singlet* and *doublet*. The maximum number of cell types per pixel can be optionally increased (e.g. to 3 or 4, Supplementary Methods), or RCTD can be run without constraining the number of cell types per pixel.

Specifically, for each cell type k, we compute $\mathscr{L}(k)$ as the log-likelihood of the model fit with only cell type k, and $\mathscr{L}(k, \ell)$ as the log-likelihood of the model fit with only cell types k and ℓ. For each pixel i we then define

$$\hat{k} = \arg\max_k \mathscr{L}(k) \text{ and } \hat{\ell} = \arg\max_{l \neq k} \mathscr{L}(\hat{k}, \ell).$$

Because we expect many pixels to represent only one cell type, we then used a penalized approach similar to AIC [33] to decide between the two models, using only one cell $\hat{k}$ or two $\hat{k}, \hat{\ell}$. Specifically, we select the model $\mathscr{M}$ maximizing,

$$\text{AIC}(\mathscr{M}) \equiv \mathscr{L}(\mathscr{M}) - V p(\mathscr{M}),$$

with p the number of parameters (cell types) and V a penalty weight. In the results presented here, we selected $V = 25$ based on simulation studies.

We then use an ad-hoc approach to classify our selections into either *confident* or *unconfident* in the following way:

1. Consider pairs of cell types (k, ℓ) such that $\left|\mathscr{L}(\hat{k}, \hat{\ell}) - \mathscr{L}(k, \ell)\right| < \delta$. If there exists one such pair such that $\hat{k} \notin \{k, \ell\}$ and another (possibly identical) pair where $\hat{\ell} \notin \{k, \ell\}$, then we assume that we do not have enough information to predict cell types and call this pixel *unconfident*. If this condition does not hold, then we will be *confident* of at least one cell type, $\hat{k}$ and/or $\hat{\ell}$, that appears in all such pairs.
2. If condition 1 does not hold, and we select the singlet model, then we call this a *confident singlet*.
3. If condition 1 does not hold, and we select the doublet model, then if there exists a cell type pair $\{k, \ell\}$ distinct from $\{\hat{k}, \hat{\ell}\}$ for which $|\mathscr{L}(\hat{k}, \hat{\ell}) - \mathscr{L}(k, \ell)| < \delta$, we call this a *unconfident doublet*, otherwise we call this a *confident doublet*.

For the work in this paper, we set $\delta = 10$ based on simulation studies. Although this value of δ was used for RCTD's accurate results across all datasets in this study, users can decide to increase δ from the default, which will reduce the number of confident pixels and potentially achieve more accurate results on fewer confident pixels.

Classification of cellular subtypes

We apply the RCTD procedure described above to detect major cell types. But, as mentioned in the results section, recently characterized cellular subtypes have been identified and defined by large-scale transcriptomic analyses [21]. After selecting pixels in which RCTD was confident of the presence of the cell type of interest, we re-ran RCTD on these pixels using a larger set of cellular subtype profiles defined by the reference. During the subtype step of RCTD, we constrained the major cell types appearing on each pixel so be the same as originally detected by RCTD.

For interneurons, we used 27 previously defined [21] interneuron subtypes and hierarchically clustered the log average expression vectors of these subtypes into 3 major subclasses (Supplementary Figure 26). In order to define spatial clusters of Slide-seq interneurons, we hierarchically clustered the points in space and manually split doublets. To classify a set of pixels presumed to comprise the same cell, we selected the subtype maximizing the joint density of these pixels by summing the log-likelihoods.

Expected cell type-specific gene expression

Once β has been estimated by RCTD, we can compute the expected cell type-specific gene expression at each pixel. Specifically, we compute the conditional expectation of $Y_{i,k,j}$, the expression of gene j on pixel i from cell type k (see supplementary methods for derivation):

$$\mathbb{E}\left[Y_{i,k,j} \middle| \beta, Y_{i,j}\right] = \frac{Y_{i,j}\beta_{k,i}\hat{\mu}_{k,j}}{\sum_{k'=1}^{K}\beta_{k',i}\hat{\mu}_{k',j}} \quad (4)$$

Intuitively, the expected expression of a cell type is proportional to the proportion of the cell type on the pixel and the probability of observing the gene in each cell type. We note that we are only computing the conditional expectation $\mathbb{E}\left[Y_{i,k,j} \middle| \beta, Y_{i,j}\right]$, but $Y_{i,k,j} \mid \beta$, $Y_{i,j}$ may have large variance for a single pixel, due to sampling noise. Furthermore, this estimate is based on a strong assumption of the model that random effects of gene expression $\varepsilon_{i,j}$ are shared across cell types.

Collection and processing of scRNA-seq and spatial transcriptomics data

We used publicly available single-cell RNA-seq datasets, which have previously been annotated by cell type using clustering. While clustering itself is an imperfect annotation of cellular identity, for the purposes of our study we assumed that these annotations were sufficiently accurate. For running RCTD on cerebellum, we trained on a single-nucleus RNA-seq dataset [20]. For training RCTD on hippocampus, and testing (cross-platform) RCTD in cerebellum, we used the DropViz single-cell RNA-seq dataset [21]. This single-cell cerebellum dataset was also used as training data to predict on Slide-seq cerebellum

(Supplementary Figure 15). The DropViz hippocampus dataset also contained annotations for interneuron subtypes. Before training RCTD on Smart-seq2 (to predict on Slide-seq somatosensory cortex), we normalized read counts by gene length, following the approach of transcripts per million (TPM). For marker gene plots, we define a *metagene* for each cell type as the sum of genes that are over-expressed with a log-fold-change above 3. In Supplementary Figure 3 concerning the Slide-seq cerbellum, for the cell type MLI2, the log-fold-change threshold was increased to 4 to achieve additional specificity.

Slide-seq mouse cerebellum and somatosensory cortex data was collected using the Slide-seqV2 protocol, developed and described recently (see supplementary methods for details) [1]. Slide-seqV2 hippocampus and Visium hippocampus data were used from previous studies [1,2]. Data pre-processing occurred using the Slide-seq tools pipeline [1]. The region of interest (ROI) was cropped prior to running RCTD, and spatial transcriptomic spots were filtered to have a minimum of 100 UMIs. We used prior anatomical knowledge to crop the ROI from an image of the total UMI counts per pixel across space, which in many cases allows one to observe overall anatomical features. For example, in Slide-seq hippocampus, the somatosensory cortex was cropped out prior to analysis. In Slide-seq cerebellum, the granule region was defined as pixels that are within 40 microns of at least 6 pixels expressing granule markers at the level of 5 counts per 500.

Validation with simulated doublets dataset

We trained RCTD on the cerebellum single-nucleus RNA-seq reference, and tested the model on a dataset of doublets simulated from the single-cell RNA-seq cerebellum dataset. We restricted to 12 cell types that appeared both in the single-nucleus and single-cell reference. In order to simulate a doublet, we randomly chose a cell from each cell type, and sampled a predefined number of UMIs from each cell (total 1, 000). In order to assess the performance of RCTD on predicting more than 1 or 2 cell types, we developed simulated spatial transcriptomics datasets with either three or four cell types per pixel. Cell type proportions per pixel were determined by choosing three or four cell types at random, and drawing the true cell type proportion from the uniform distribution. To model the conditions of Visium, we sampled 10, 000 UMIs per pixel. If a single cell did not have enough UMIs for its cell type, we continued augmenting the pixel with additional cells from that cell type.

We defined a doublet as containing 25–75% of UMIs for each of the two cell types, whereas a singlet contained 0% or 100%. We defined *doublet classification rate* (Figure 3a) as the ratio of number of predicted doublets to total predicted singlets or doublets. Cell type proportion estimation (Figure 3b, 3c) was measured with RCTD fit using the two cell types present on the simulated doublet. We defined coarser *classes* of cell types (used for e.g. Figure 3d) based on a previously defined dendrogram [20]. This resulted in pairing of MLI1/MLI2, Astrocytes/Bergmann, Oligoden./Polyden., and Endothelial/Fibroblast. Cell class identification rate (Figure 3d, top) was calculated on the subset of confidently called cell types.

Detection of cell type-specific gene expression patterns

After computing expected cell type-specific gene expression, we detected spatially variable genes within a cell type. Genes were filtered for minimum average expression within the scRNA-seq reference of the cell type of interest (.01 counts per 500, and at least 50% as large as average expression of other cell types). We applied 2D local regression to these genes, and calculated coefficient of variation (CV) of the estimated smooth function. We selected genes with CV ≥ 0.5 and tested the local regression variation with a permutation *F*-test ($p \leq 0.01$, 99 permutations of spatial locations).

Next, we searched for genes that changed their expression within astrocytes based on co-localization with another cell type. We classified astrocytes as co-localizing with another particular cell type if at least 25% of their neighbors within a 40 micron radius were that cell type. If at least 80% of these neighbors were other astrocytes, the cell was classified as co-localizing with other astrocytes. We filtered for genes in the scRNA-seq reference with minimum average expression within astrocytes (0.01 counts per 500, and log-fold-change of ≥ 1.6 vs. each other cell type). We looked for genes that were differentially expressed depending on the co-localized cell type, testing with a two-tailed *z*-test ($p < 0.05$). We pooled together *excitatory neuron* cell types, defined as CA1, CA3, and dentate cell types, for the analysis of several genes, including *Kcnj16*, *Slc7a10*, and *Slc6a11*, that were differentially expressed in astrocytes localized around each of CA1, CA3, and dentate cell types.

Implementation details

RCTD is publicly available as an R package (https://github.com/dmcable/RCTD). The quadratic program that arises in the RCTD optimization algorithm is solved using the quadprog package in R [36]. We used and modified code from the DWLS package to implement sequential quadratic programming for RCTD [19,37]. Non-negative least squares regression was also implemented as a quadratic program. Unsupervised clustering was performed using the Seurat package, following Seurat's spatial transcriptomics vignette [38]. Clusters were assigned by their expression of marker genes and spatial localization. Additionally, detection of globally spatially variable genes was accomplished using Seurat's implementation of Moran's I. Local regression was accomplished with the loess function. The NMFreg python notebook was used with default parameters (factors = 30) for testing NMFreg. To test DWLS on cell type classification, we used the buildSignatureMatrixUsingSeurat function to build the cell type signature matrix and used the solveDampenedWLS function to predict cell type proportions for each pixel, which were scaled to units of UMI counts. RCTD was tested on a Macintosh laptop computer with a 2.4 GHz Intel Core i9 processor, 8 cores, and 32GB of memory (we recommend at least 4GB of memory to run RCTD. For example, we timed RCTD on the Slide-seq cerebellum dataset, containing 11, 626 pixels, 19 cell types, and 3, 272 differentially expressed genes detected by RCTD. Under these conditions, RCTD ran in 14 minutes and 57 seconds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Robert Stickels for providing valuable input on the analysis. We thank members of the Chen lab, Irizarry lab, and Macosko lab for helpful discussions. D.C. was supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. This work was supported by an NIH Early Independence Award (DP5, 1DP5OD024583 to F.C.), the Burroughs Wellcome Fund (F.C.), the NHGRI (R01, R01HG010647 to E.Z.M. and F.C.), as well as the Schmidt Fellows Program at the Broad Institute and the Stanley Center for Psychiatric Research. R.A.I. was supported by NIH grants R35GM131802 and R01HG005220.

Data Availability Statement

Slide-seq V2 data generated for this study is available at the Broad Institute Single Cell Portal https://singlecell.broadinstitute.org/single_cell/study/SCP948. Additional publicly available data from other studies that was used for analysis is also included in this repository. Furthermore, the Life Sciences Reporting Summary is available.

References

[1]. Stickels RR et al. Sensitive spatial genome wide expression profiling at cellular resolution. bioRxiv (2020). https://www.biorxiv.org/content/early/2020/03/14/2020.03.12.989806.full.pdf.

[2]. 10x Genomics. 10x genomics: Visium spatial gene expression. https://www.10xgenomics.com/solutions/spatial-gene-expression/ (2020).

[3]. Vickovic S et al. High-definition spatial transcriptomics for in situ tissue profiling. Nature methods 16, 987–990 (2019). [PubMed: 31501547]

[4]. Pelkey KA et al. Hippocampal gabaergic inhibitory interneurons. Physiological reviews 97, 1619–1747 (2017). [PubMed: 28954853]

[5]. Cembrowski MS et al. The subiculum is a patchwork of discrete subregions. Elife 7, e37701 (2018). [PubMed: 30375971]

[6]. Edsgärd D, Johnsson P & Sandberg R Identification of spatial expression trends in single-cell gene expression data. Nature methods 15, 339 (2018). [PubMed: 29553578]

[7]. Sun S, Zhu J & Zhou X Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods 17, 193–200 (2020). [PubMed: 31988518]

[8]. Svensson V, Teichmann SA & Stegle O SpatialDE: identification of spatially variable genes. Nature methods 15, 343–346 (2018). [PubMed: 29553579]

[9]. Wagner A, Regev A & Yosef N Revealing the vectors of cellular identity with single-cell genomics. Nature biotechnology 34, 1145 (2016).

[10]. Regev A et al. Science forum: the human cell atlas. Elife 6, e27041 (2017). [PubMed: 29206104]

[11]. Rodriques SG et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 363, 1463–1467 (2019). [PubMed: 30923225]

[12]. Stuart T et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902 (2019). [PubMed: 31178118]

[13]. Moncada R et al. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nature Biotechnology 38, 333–342 (2020).

[14]. Townes FW, Hicks SC, Aryee MJ & Irizarry RA Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. Genome biology 20, 1–16 (2019). [PubMed: 30606230]

[15]. Hafemeister C & Satija R Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome biology 20, 1–15 (2019). [PubMed: 30606230]

[16]. Pliner HA, Shendure J & Trapnell C Supervised classification enables rapid annotation of cell atlases. Nature methods 16, 983–986 (2019). [PubMed: 31501545]

[17]. Leek JT et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11, 733–739 (2010).

[18]. Bakken TE et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PloS one 13 (2018).

[19]. Tsoucas D et al. Accurate estimation of cell-type composition from gene expression data. Nature communications 10, 1–9 (2019).

[20]. Kozareva V et al. A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. bioRxiv (2020). https://www.biorxiv.org/content/early/2020/03/05/2020.03.04.976407.full.pdf.

[21]. Saunders A et al. Molecular diversity and specializations among the cells of the adult mouse brain. Cell 174, 1015–1030 (2018). [PubMed: 30096299]

[22]. Brown AM et al. Molecular layer interneurons shape the spike activity of cerebellar purkinje cells. Scientific reports 9, 1–19 (2019). [PubMed: 30626917]

[23]. Tasic B et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature neuroscience 19, 335–346 (2016). [PubMed: 26727548]

[24]. Zhang M et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. bioRxiv (2020).

[25]. Sunkin SM et al. Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic acids research 41, D996–D1008 (2012). [PubMed: 23193282]

[26]. Capogna M Neurogliaform cells and other interneurons of stratum lacunosum-moleculare gate entorhinal–hippocampal dialogue. The Journal of physiology 589, 1875–1883 (2011). [PubMed: 21135049]

[27]. Leão RN et al. OLM interneurons differentially modulate CA3 and entorhinal inputs to hippocampal CA1 neurons. Nature neuroscience 15, 1524 (2012). [PubMed: 23042082]

[28]. Gampe K et al. Ntpdase2 and purinergic signaling control progenitor cell proliferation in neurogenic niches of the adult mouse brain. Stem Cells 33, 253–264 (2015). [PubMed: 25205248]

[29]. Dikow N et al. 3p25. 3 microdeletion of gaba transporters slc6a1 and slc6a11 results in intellectual disability, epilepsy and stereotypic behavior. American Journal of Medical Genetics Part A 164, 3061–3068 (2014).

[30]. Lee T-S et al. Gat1 and gat3 expression are differently localized in the human epileptogenic hippocampus. Acta neuropathologica 111, 351–363 (2006). [PubMed: 16456667]

[31]. Kulkarni A, Anderson AG, Merullo DP & Konopka G Beyond bulk: a review of single cell transcriptomics methodologies and applications. Current opinion in biotechnology 58, 129–136 (2019). [PubMed: 30978643]

[32]. Halpern KB et al. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nature biotechnology 36, 962–970 (2018).

[33]. Sakamoto Y, Ishiguro M & Kitagawa G Akaike information criterion statistics. Dordrecht, The Netherlands: D. Reidel 81 (1986).

[34]. Zhou M, Li L, Dunson D & Carin L Lognormal and gamma mixed negative binomial regression. In Proceedings of the… International Conference on Machine Learning. International Conference on Machine Learning, vol. 2012, 1343 (NIH Public Access, 2012). [PubMed: 25279391]

[35]. Swami A Non-gaussian mixture models for detection and estimation in heavy-tailed noise. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), vol. 6, 3802–3805 (IEEE, 2000).

[36]. Turlach BA & Weingessel A quadprog: Functions to solve quadratic programming problems. R package version 1.5–5 (2013).

[37]. Duchi J Sequential Convex Programming, notes for EE364b: Convex Optimization II, Stanford University (2018).

[38]. SatijaLab. Analysis, visualization, and integration of spatial datasets with Seurat. https://satijalab.org/seurat/v3.1/spatial_vignette.html (2020).

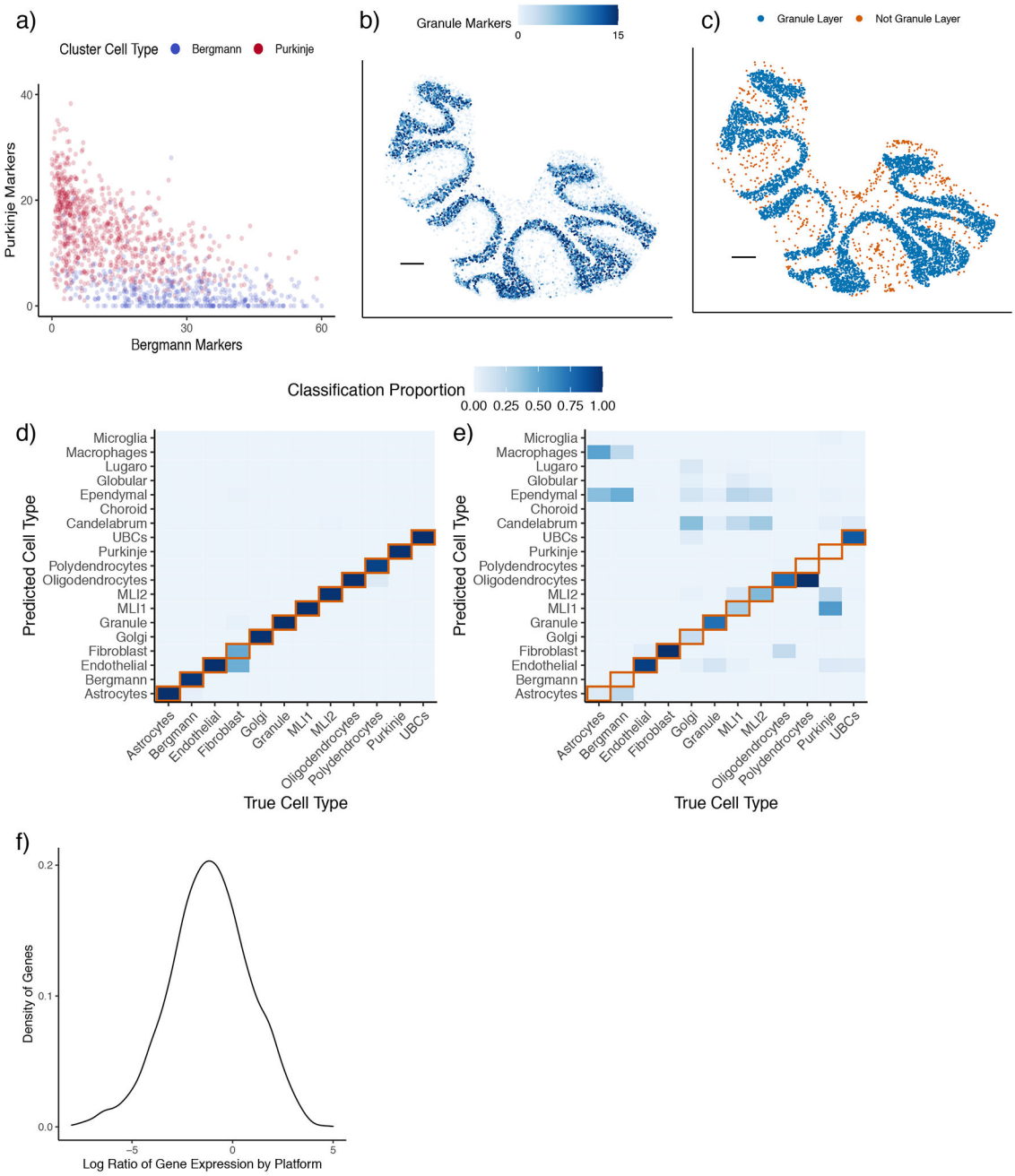

Figure 1:

Spatial transcriptomics data presents challenges for cell type learning.

a) Expression of Bergmann and Purkinje marker genes for pixels colored by unsupervised clustering cell type assignment within a Slide-seq cerebellum dataset. The e.g. Bergmann markers axis is the sum of the expression (counts per 500) of Bergmann differentially expressed genes.

b) Expression (counts per 500) of granule marker genes in Slide-seq. Scale bar: 250 microns.

c) Spatial plot of granule cells identified by unsupervised clustering. Pixels are colored by whether they spatially belong to the granule layer. Scale bar: 250 microns.
d) Confusion matrix of true vs predicted cell types within training dataset (single-nucleus RNA-seq) by non-negative least squares regression. Color represents the proportion of the cell type on the *x*-axis classified as the cell type on the *y*-axis. The diagonal representing ground truth is boxed in red.
e) Confusion matrix of cell type predictions across platforms using non-negative least squares regression trained on single-nucleus RNA-seq, tested on single-cell RNA-seq. Same color scale as (d).
f) Density plot, across genes, of measured platform effects between cerebellum single-cell RNA-seq and single-nucleus RNA-seq. The platform effect is defined as the $\log_2$ ratio of average gene expression between platforms.

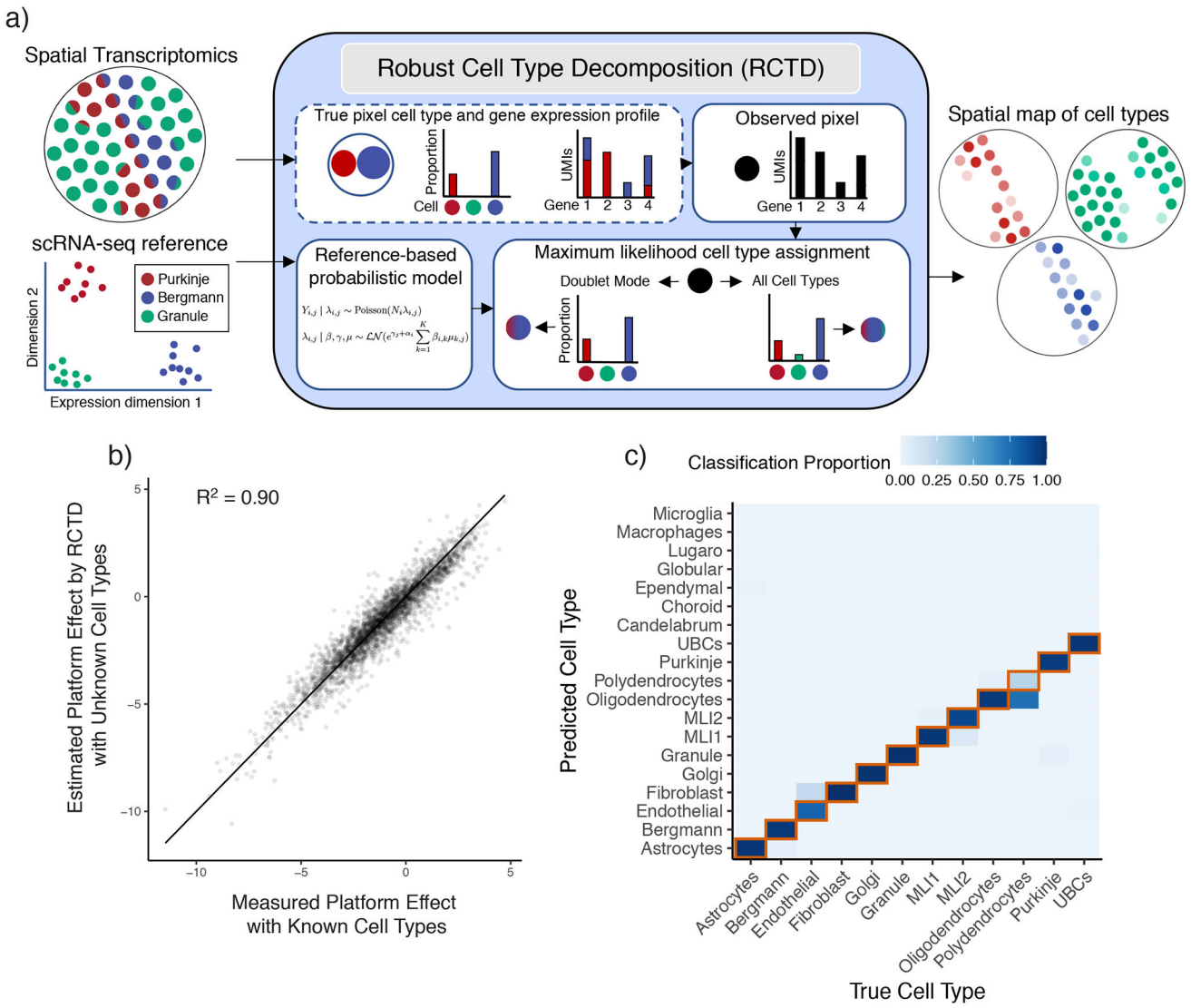

Figure 2:

Robust Cell Type Decomposition enables cross-platform learning of cell types.

a) Left: RCTD inputs: a scRNA-seq dataset, annotated by cell type, and a spatial transcriptomics dataset with unknown cell types. Middle: RCTD uses a scRNA-seq reference-based probabilistic model to predict cell types on a single pixel containing a mixture of two cell types (e.g. Bergmann/Purkinje), with unknown cell type proportions. RCTD predicts the maximum likelihood cell type proportions. In *doublet mode*, RCTD constrains each pixel to contain at most two cell types; alternatively, RCTD can estimate the best fit at a pixel using all cell types. Right: RCTD outputs a spatial map of cell types, with opacity representing the inferred cell type proportion.

b) Scatter plot of measured vs predicted platform effect (by RCTD) for each gene between the single-cell and single-nucleus cerebellum datasets. Line is the identity line. Measured platform effect is calculated as the $\log_2$ ratio of average gene expression between platforms.

c) Confusion matrix for RCTD's performance on cross-platform (trained on single-nucleus RNA-seq, tested on single-cell RNA-seq) cell type assignments for single cells. Color

represents the proportion of the cell type on the *x*-axis classified as the cell type on the *y*-axis. The diagonal representing ground truth is boxed in red.

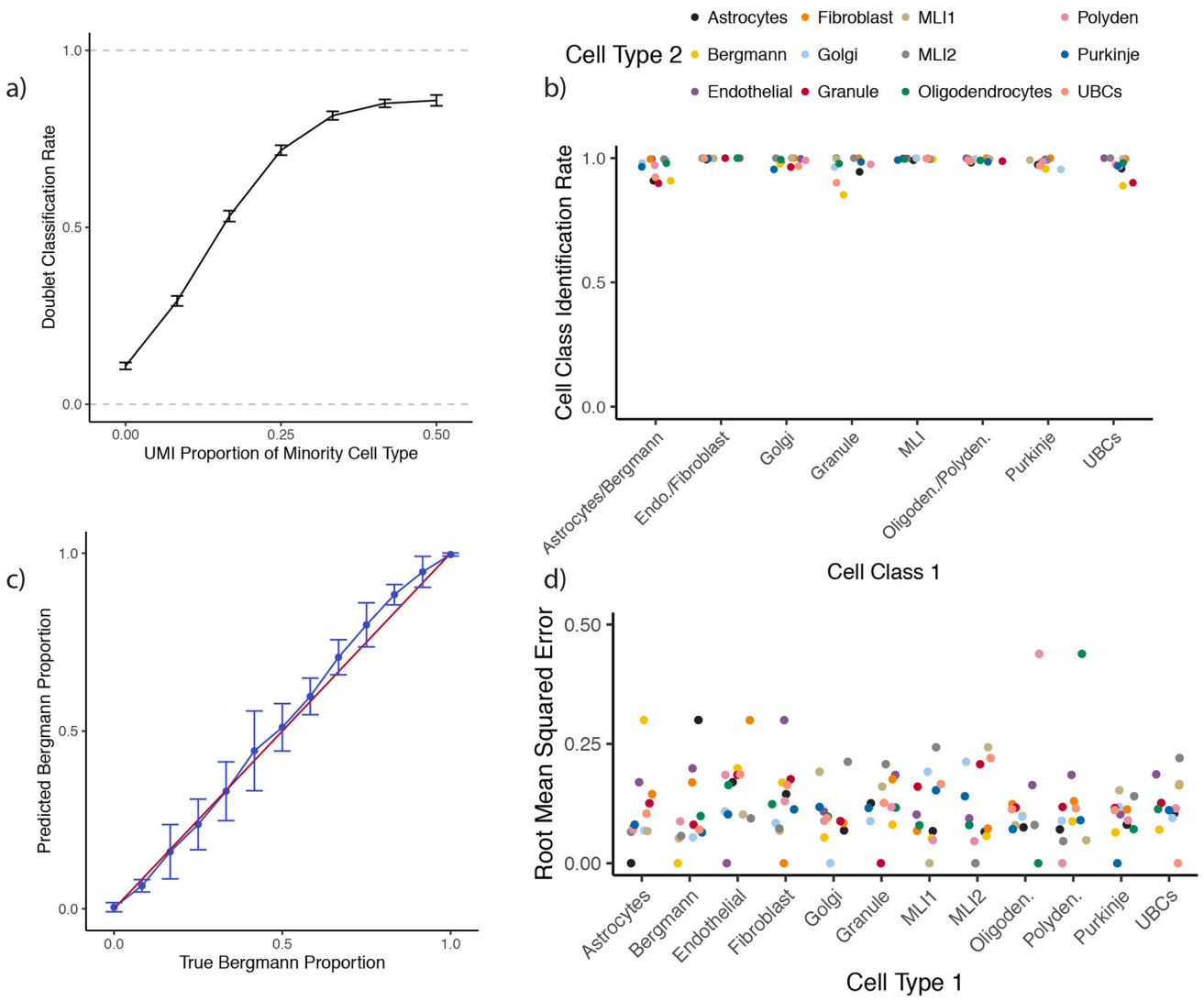

Figure 3:

RCTD performs cross-platform detection and decomposition of doublets.

All: RCTD was trained on the single-nucleus RNA-seq cerebellum dataset and tested on a dataset of simulated mixtures of single cells from a single-cell RNA-seq cerebellum dataset.

a) Rate of doublet classification by RCTD on simulated mixtures of single cells, with 95% confidence intervals. The x-axis represents the true proportion of UMIs sampled from the minority cell type, ranging from 0% (true singlet) to 50% (equal proportion doublet) ($1980 \leq n \leq 3860$ simulations per condition).

b) On simulated doublets of cell class 1 and cell type 2, the percentage of confident calls by RCTD that correctly identify the cell class, where cell classes group four pairs of transcriptionally similar cell types based on a previous dendrogram [20] (polydendrocytes/oligodendrocytes, MLI1/MLI2, Bergmann/astrocytes, endothelial/fibroblasts). Column represents cell class 1, and color represents cell type 2.

c) On simulated Bergmann-Purkinje doublets, predicted Bergmann proportions by RCTD. The *x*-axis represents the true proportion of UMIs sampled from the Bergmann cell. The

red line is the identity line, and the blue line is the average and standard deviation (n = 30 simulations per condition) of RCTD's prediction.

d) For each pair of cell types, root mean squared error (RMSE) of predicted vs true cell type proportion (as in (c)) by RCTD on simulated doublets (n = 390 simulations per cell type pair). Column represents cell type 1, and color represents cell type 2.

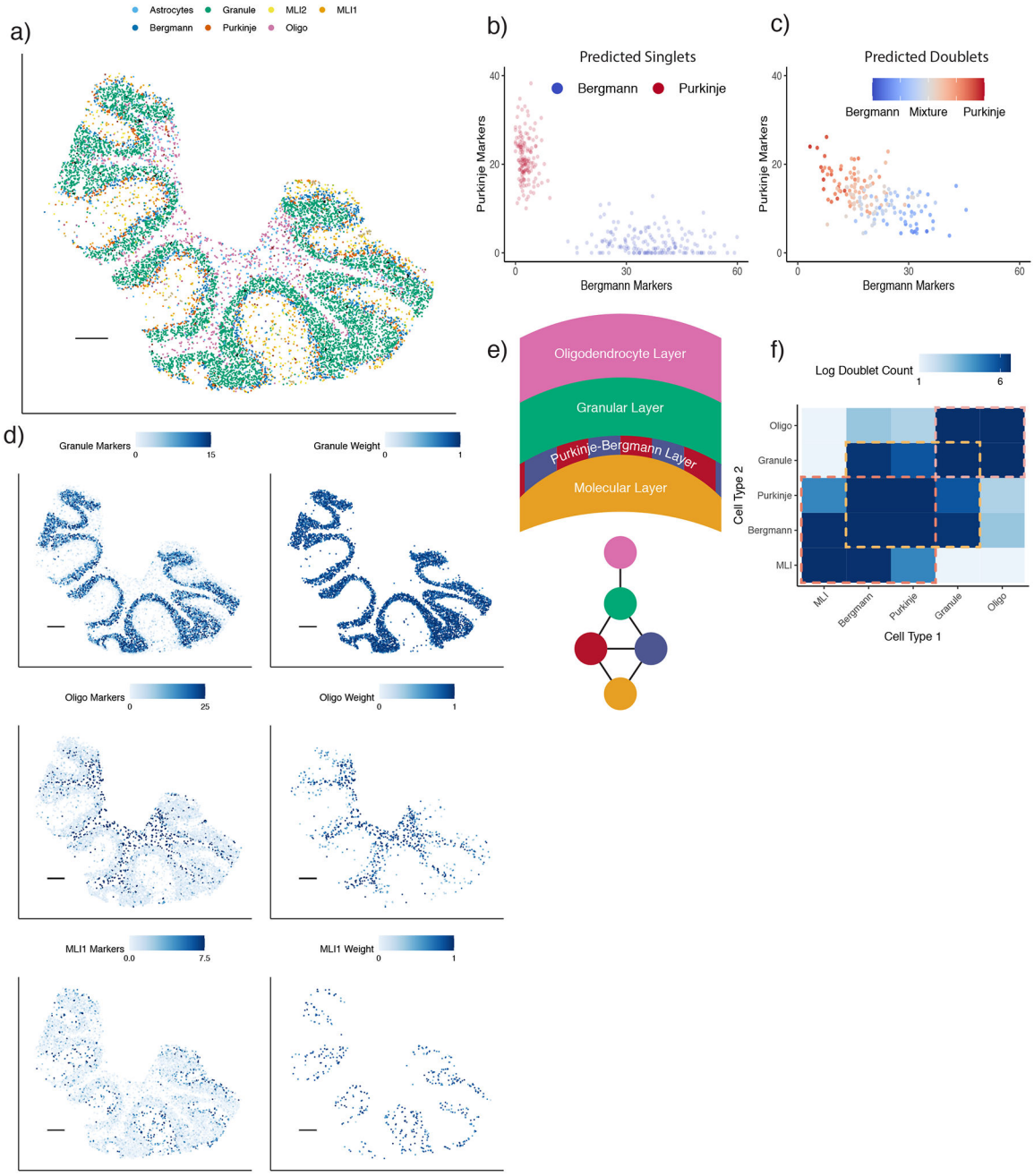

Figure 4:

RCTD applied to cell type learning in Slide-seq datasets.

a) RCTD's spatial map of cell type assignments in the cerebellum. Out of 19 cell types, the seven most common appear in the legend (individual cell types displayed in Supplementary Figure 14).

b) Analogous to (1a), expression of Bergmann and Purkinje marker genes for RCTD's predicted singlet pixels within a Slide-seq cerebellum dataset (colored by cell type assignment). The e.g. Bergmann markers axis is the sum of the expression (counts per 500) of Bergmann differentially expressed genes.

c) Expression of Bergmann and Purkinje marker genes for doublet pixels predicted by RCTD, colored by predicted cell type proportion.
d) Predicted spatial localization of cell types by RCTD for granule, oligodendrocytes, and molecular layer interneurons 1 (MLI1). Left: summed expression (counts per 500) (represented by color) of cell type-specific marker genes. Right: predicted spatial locations of each cell type, with color representing predicted cell type proportion.
e) (Top) Schematic of spatial cell type organization within the cerebellum [22]. (Bottom) Connectivity graph of cell types that are likely to spatially colocalize. Cell types are colored as in (a).
f) Frequency of doublets identified by RCTD between each pair of cell types. Color represents $\log_2$ scale counts. Dotted boxes represent communities anatomically expected to exhibit spatial co-localization. Diagonal represents prevalence of singlets. Color bar range: 2 to 100 counts.
All scale bars 250 microns.

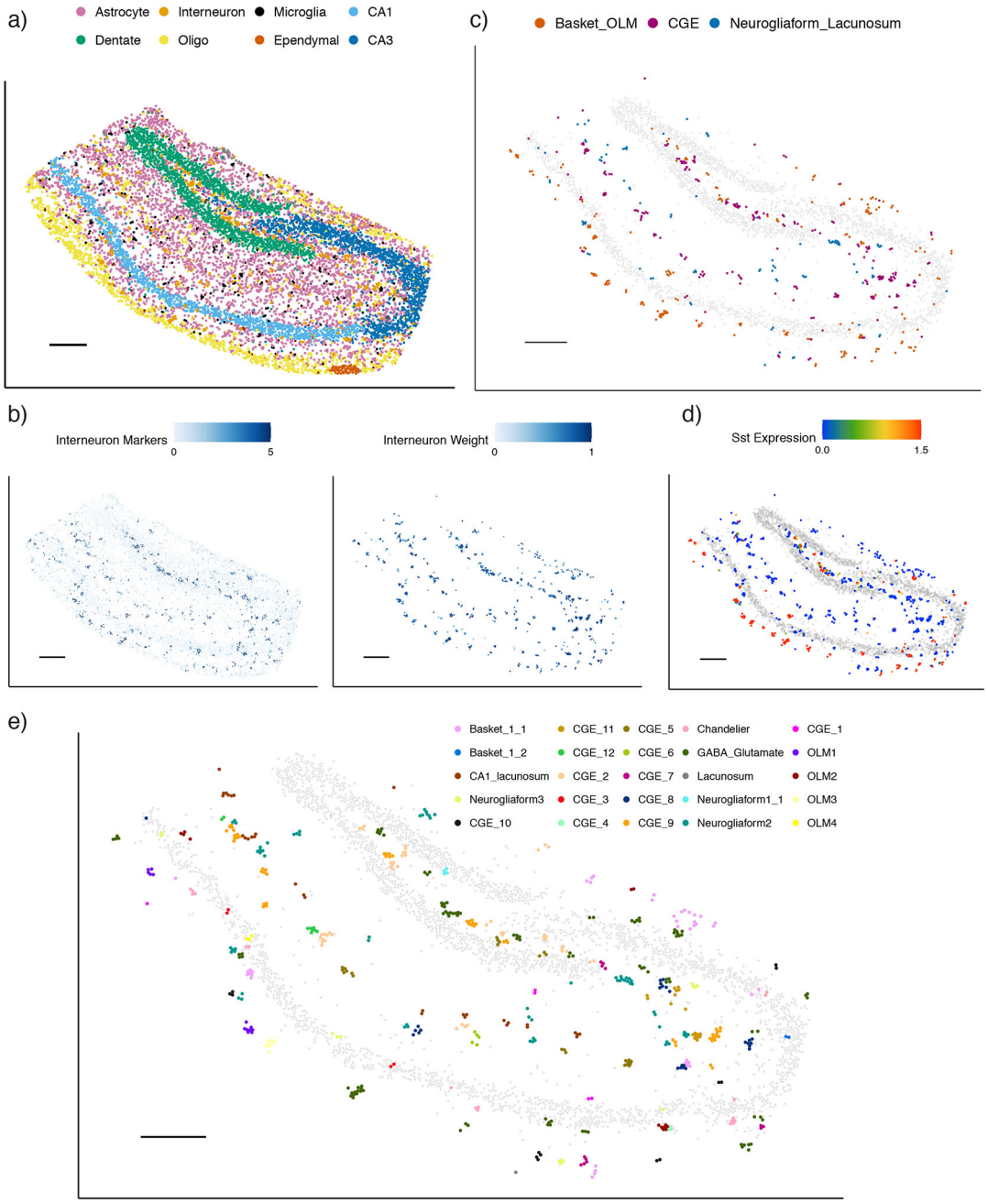

Figure 5:

RCTD maps cell types and subtypes in Slide-seq hippocampus.

a) RCTD's spatial map of predicted cell types in the hippocampus. Out of 17 cell types, the 8 most common appear in the legend (individual cell types displayed in Supplementary Figure 19).

b) Predicted spatial localization of interneuron cell types by RCTD. Left: normalized expression (represented by color, counts per 500) of marker genes. Right: predicted spatial locations of interneurons, with color representing predicted cell type proportion.

c) Predicted confident assignments of interneuron pixels by RCTD to 3 classes of interneuron subtypes, plotted in space. Color indicates predicted subclass.
d) Expression (counts per 500) of the *Sst* gene in interneurons identified by RCTD.
e) RCTD's confident assignment of spatial clusters to 27 interneuron subtypes (25/27 subtypes assigned).
All scale bars 250 microns. Grey circles represent location of CA1, CA3, and dentate gyrus excitatory neurons for reference.

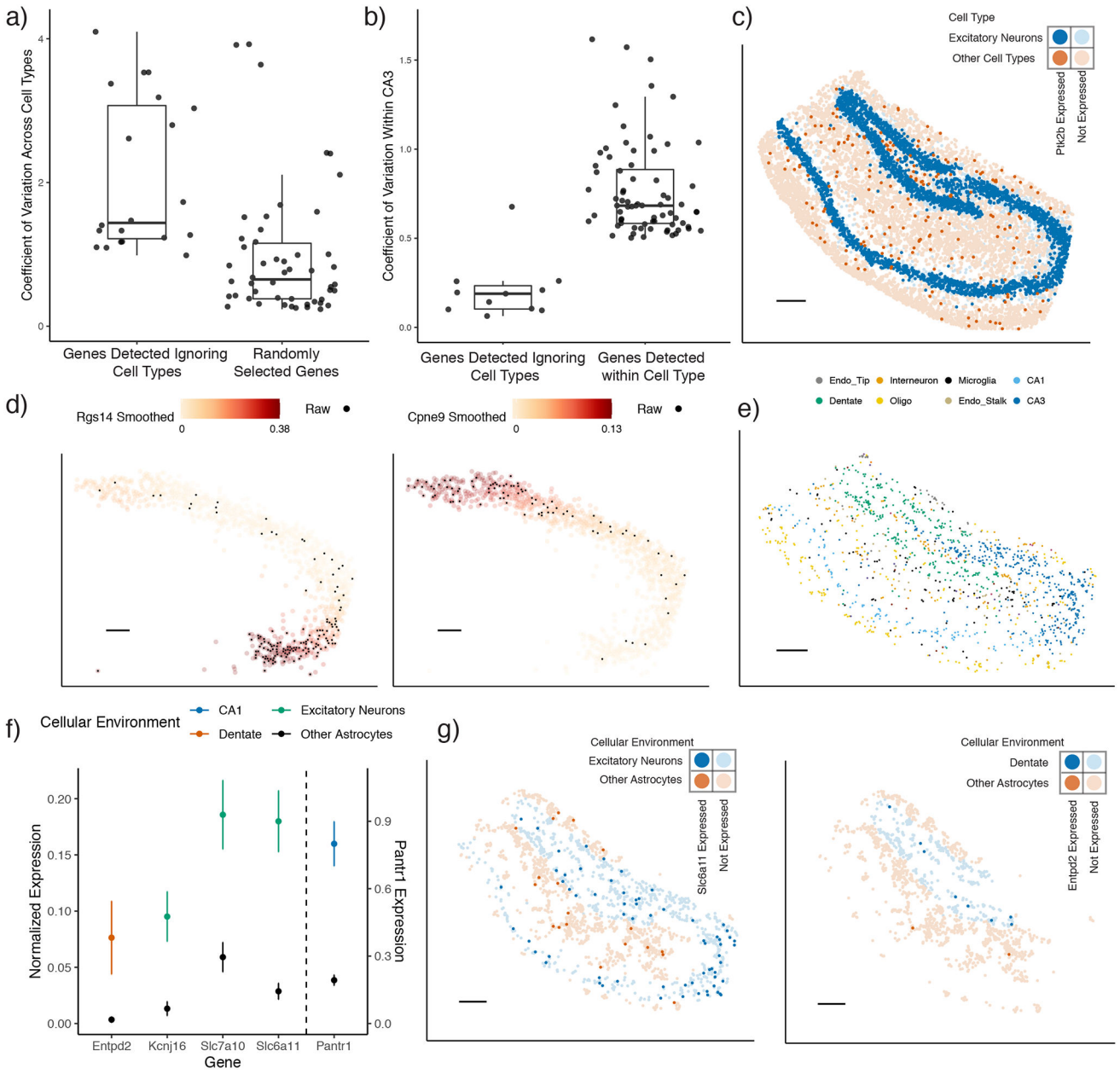

Figure 6:

RCTD enables detection of cell type-specific spatial patterns of gene expression.

a) Boxplot of coefficient of variation of genes across cell types in the hippocampus single-cell RNA-seq reference. Spatially variable genes were selected for large spatial autocorrelation in the Slide-seq hippocampus, without considering cell type. For reference, 50 randomly selected genes are shown.

b-g) Analysis on Slide-seq hippocampus data

b) Boxplot of the coefficient of variation in gene expression within CA3 cells identified by RCTD. (Left): Spatially variable genes selected for large spatial autocorrelation in the hippocampus, without considering cell type. (Right): Using RCTD's expected cell

type-specific gene expression, genes determined to be spatially variable by applying local regression within the CA3 cell type ($p \leq 0.01$, permutation F-test).
c) Bold pixels represent expression of *Ptk2b*, a gene selected to be spatially variable without considering cell type. Blue represents pixels with excitatory neurons (as detected by RCTD), whereas red represents pixels without excitatory neurons.
d) Smoothed spatial expression patterns (counts per 500), recovered by local regression, of two genes detected to have large spatial variation within RCTD's CA3 cells. Individual pixels expressing the gene are colored in black.
e) Spatial localization of astrocyte doublets in the hippocampus, detected by RCTD. Color represents the other cell type on the doublet.
f) Mean and standard error of RCTD's expected gene expression (counts per 500) within groups of astrocytes ($129 \leq n \leq 956$ cells per condition) classified by their cellular environment (color). (Scale on the right for *Pantr1*, scale on the left for other genes).
g) Spatial visualization of genes with environment-dependent expression within astrocytes. Red represents the astrocytes surrounded by other astrocytes, whereas blue represents astrocytes that are surrounded by excitatory neurons (left) or dentate gyrus cells (right). Bold points represent astrocytes expressing *Slc6a11* (left) or *Entpd2* (right).
All scale bars 250 microns. For boxplots, the median, 25th, and 75th percentile define the box, with whiskers extending the hinge by 1.5 times the inter-quartile range (IQR).