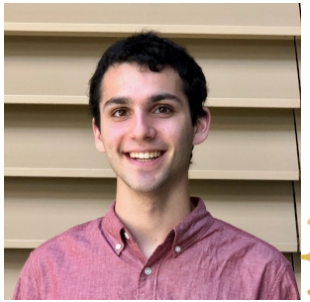
Robust decomposition of cell type (RCTD) mixtures in spatial transcriptomics

Dylan Cable 11-18-20



Dylan Cable, PhD student in Computer Science, MIT

Supported by:

NSF the Hertz FOUNDATION freedom to innovate

Advised By:

Rafael Irizarry, Dana Farber/ Harvard Biostatistics

Fei Chen, Broad Institute, Harvard

In Collaboration With:

Evan Macosko, MGH and Broad

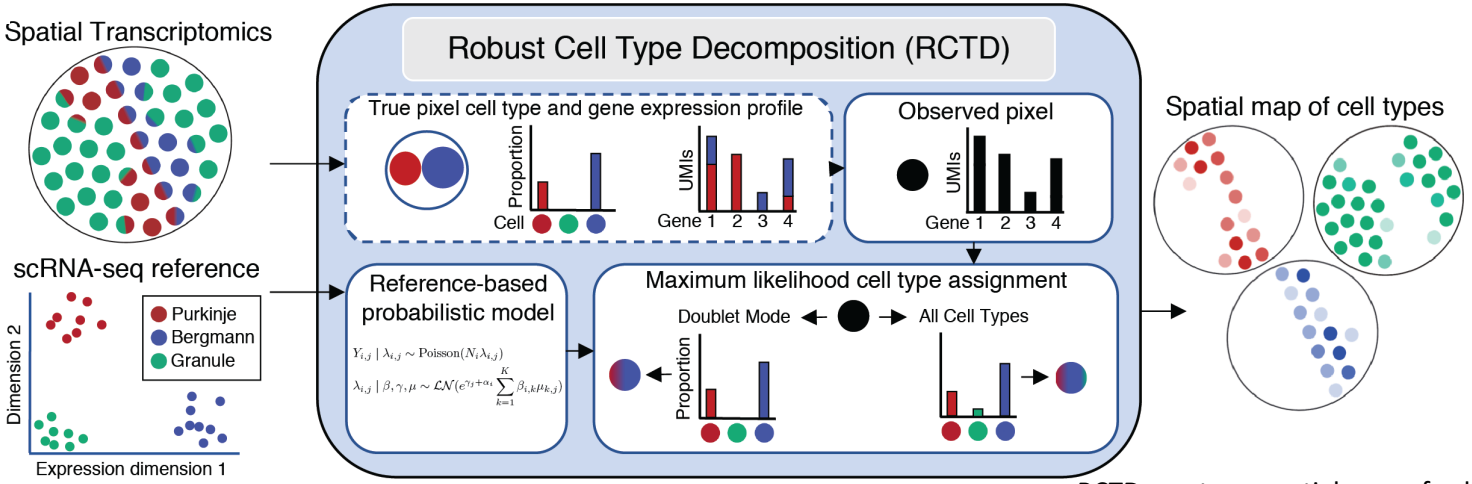
Evan Murray
Luli Zou
Aleks Goeva

Our project: RCTD uses a labeled single-cell reference to assign cell type assignments and proportions to spatial transcriptomics pixels (or spots), which can contain mixtures of multiple cell types.

RCTD is publicly available as an R package (https://github.com/dmcable/RCTD), and has been tested on Slide-seq, Visium, and MERFISH datasets.

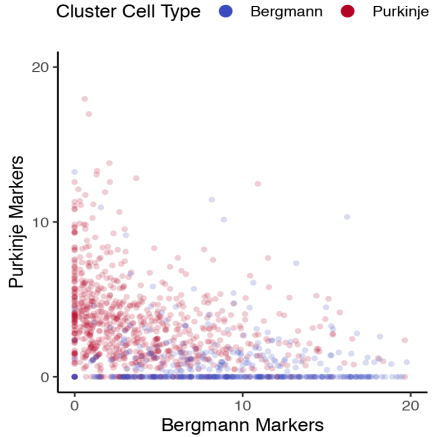
RCTD is recently accepted for publication at Nature Biotechnology.

The RCTD model uses a scRNA-seq dataset to assign cell types to spatial transcriptomics mixtures.

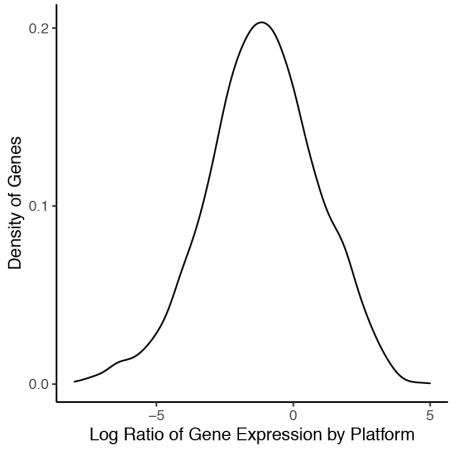


Challenges

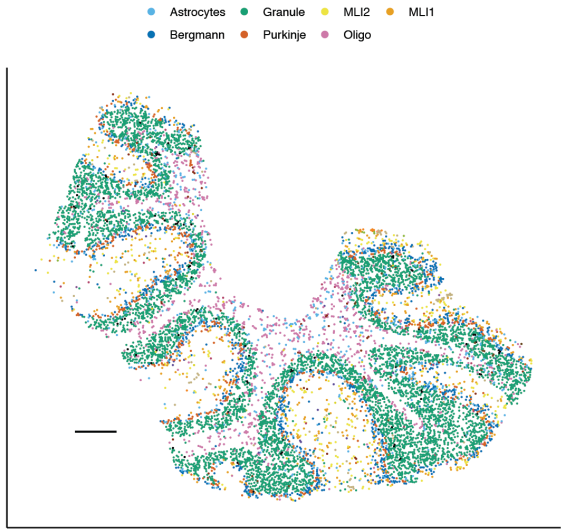
Spatial Transcriptomics pixels can contain genes from multiple nearby cell types, as seen from visualization in marker gene space. Unsupervised clustering may not correctly label the mixtures.



Single-cell RNA-seq vs Spatial Transcriptomics can have large technical effects, platform effects, effecting the rate of capture of genes.



RCTD creates a spatial map of cell types



Robust Cell Type Decomposition (RCTD) Model

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

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

$$\gamma_j \sim \text{Normal}(0, \sigma_\gamma^2), \quad \varepsilon_{i,j} \sim \text{Normal}(0, \sigma_\varepsilon^2)$$

Index of variables: We use i to index pixels, j for genes, and k for cell types. $Y_{i,j}$ is the observed count of gene j at pixel i. N_i is the total UMI counts at pixel i. $\mu_{k,j}$ is the mean gene expression of cell type k for gene j in the single-cell reference, and w is a normalized version of β. γ_j is the platform effect of gene j. $\beta_{i,k}$ is the proportion of cell type k on pixel i. $\varepsilon_{i,j}$ is cell-specific biological variation. $\lambda_{i,j}$ is the gene expression level of gene j in pixel i. S_j is the pseudobulk measured counts of gene j, and W_j is the pseudobulk proportion of cell type k. $\mathcal{L}$ represents the log-likelihood of the model on an individual pixel.

Platform Effect Normalization:
Estimation of Platform Effects

$$S_j \equiv \sum_{i=1}^{I} Y_{i,j} \sim \text{Poisson}(\sum_{i=1}^{I} N_i \lambda_{i,j})$$

$$S_j|\gamma_j \sim \text{Poisson}\left(I\bar{N}e^{\gamma_j}\sum_{k=1}^{K}\mu_{k,j}W_k\right), \quad \gamma_j \sim \text{Normal}(0, \sigma_\gamma^2)$$

$$\bar{S}_j \approx e^{\gamma_i}\sum_{k=1}^{K}\mu_{k,j}W_k \implies \gamma_j \mid \hat{W} \approx \log(\bar{S}_j) - \log\left(\sum_{k=1}^{K}\mu_{k,j}\hat{W}_k\right) \equiv \hat{\gamma}_j$$

Optimization of RCTD log-likelihood via sequential quadratic programming

Log-likelihood of an individual pixel (after platform effect normalization)

$$\max_w \mathcal{L}(w) = \sum_{j=1}^{J} \log P(Y_j \mid \lambda_j(w)) \qquad w_{k,i} = \beta_{k,i}e^{\alpha_i}$$

Second-order Taylor approximation of log-likelihood

$$-\mathcal{L}(w) \approx -\mathcal{L}(w_0) + b(w_0)^T(w - w_0) + \frac{1}{2}(w - w_0)^T A(w_0)(w - w_0)$$

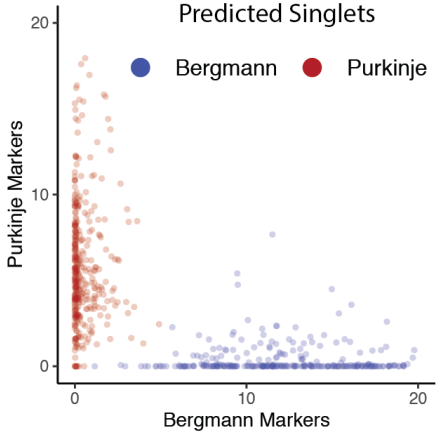
Optimization of approximate log-likelihood as a Quadratic Program (QP)

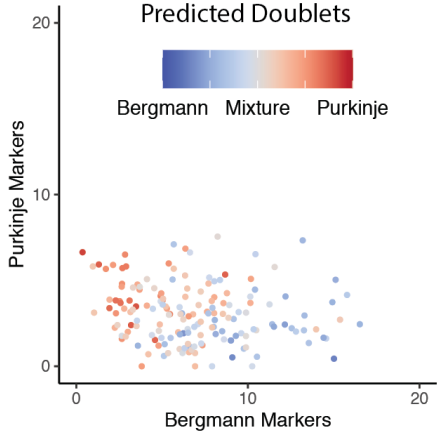
$$\begin{aligned} \min_d \quad & b(w_0)^T d + \frac{1}{2} d^T A(w_0) d \\ \text{s.t.} \quad & d + w_0 \geq 0 \end{aligned}$$

$$b(w) = -\nabla L(w) \qquad A(w) = \text{Hess}(-L(w))$$

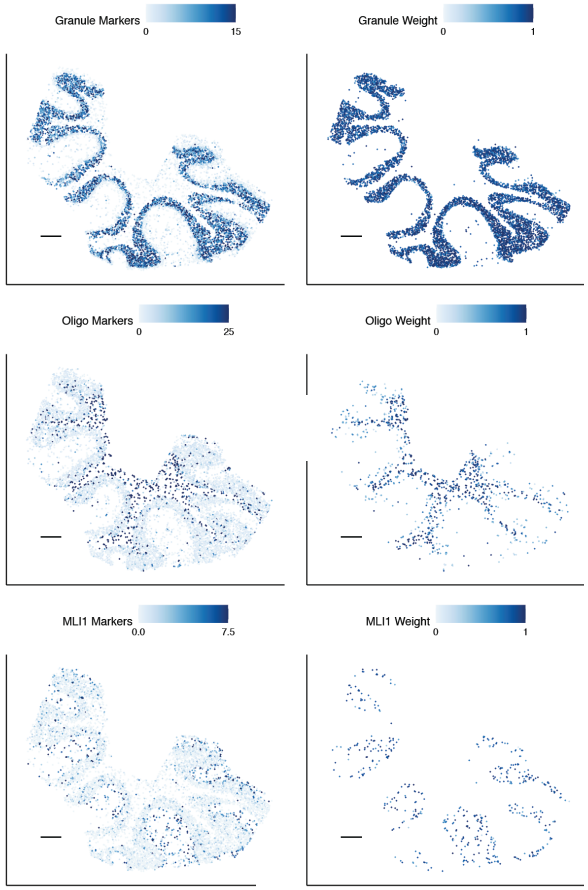
Technical Validations on Spatial Data

RCTD's predictions of singlets (single cell type per pixel) and doublets (two cell types per pixel) are consistent with visualization of pixels in marker gene space:





RCTD's predictions agree with spatial maps of marker genes:



Technical Validations on Simulated Data