Supplements to 'SCDC: Bulk Gene Expression Deconvolution by Multiple Single-Cell RNA Sequencing References'

Meichen Dong Aatish Thennavan Eugene Urrutia
Yun Li Charles M. Perou Fei Zou Yuchao Jiang

Evaluation Measurement

The metrics we used for method evaluation include mean absolute deviation (mAD), Pearson correlation, and Spearman correlation. Given a parameter z and its estimator $\hat{z}$, these metrics can be defined as:

$$\begin{aligned} \text{mAD} &= \mathsf{mean}\{\|z - \hat{z}\|_1\} \\ \text{Pearson/Spearman R} &= \mathit{corr}(z, \hat{z}) \end{aligned}$$

$\|\cdot\|_1$ denotes the $L1$ norm.

Inverse Sum of Squared Errors

In the spirit of meta analysis (Borenstein et al. [2011], DerSimonian and Laird [2015]), we propose an alternative weighting scheme, where we assign reference-specific weights that are proportional to the inverse sum of squared errors (SSE) between Y and $\hat{Y}$. That is, for the rth single-cell reference, we assign its weights as

$$\hat{w}_r = iSSE_r = \frac{1/SSE_r}{\sum_{r'} 1/SSE_{r'}}, r = 1, ..., R; \quad SSE_r = ||Y - \hat{Y}_r||_2^2,$$

where $\|\cdot\|_2$ denotes the $L2$ norm. This score can be used as a criteria while the grid search/regression-based methods show any discrepancy. To reduce

the effects of the genes with extreme expression values that could bias the weight selection procedure, SCDC opts to further filter out genes whose gene expressions in bulk samples are within the upper 5% quantile or the lower 15% quantile.

Quality Control and Clustering of scRNA-seq Data

For single cells from the three cell-line experiment, cells with a high percentage of mitochondrial gene expressions were filtered out. Genes with lengths greater than 200kb, ribosomal genes, and genes with undetectable expressions were filtered out. Seurat was applied for single cell clustering: genes detected in at least three cells were kept; cells with less than 200 genes detected were filtered out; the number of genes detected and the number of UMIs were regressed out in the scaling procedure; 'FindClusters' was applied using the first twenty principal components, with resolution parameter set from 0.6 to 1. Finally, cell types were annotated according to previously reported marker genes.

For the mouse mammary gland data, single cell clustering was performed within each subject separately. In addition to the Seurat clustering procedures described above, the percentage of cell-cycle gene expressions was also regressed out when scaling the gene expression matrix. Epithelial cells were first identified as a major cluster and were further subgrouped into luminal and basal cells. 'FindMarkers' function was applied to each pair of cell types, and the number of marker genes from each pair was used to determine whether or not to combine the two clusters.

Two-Level Deconvolution

Similar to MuSiC (Wang et al. [2019]), for cases where closely related cell types are present in the data, SCDC adopts a two-step approach, which first separates remotely connected cell types and, in the second step, dissociates cell types that share high similarities. However, there is no consensus on how to determine the order of deconvolution, especially when multiple scRNA-seq datasets are available. To solve this, we employ MNN (Haghverdi et al. [2018]) to correct for batch effect and to calculate a basis matrix from the ad-

justed data. Hierarchical clustering is applied to determine the relationship between the cell types of interest. The hierarchical structure is further used to guide the two-step approach for deconvolution. For the mouse mammary gland dataset, the first-round deconvolution separates cluster 1 = {immune cells} from cluster 2 ={endothelial, fibroblast, basal, luminal cells} and the second-round deconvolution further separates the cell types in cluster 2 (Figure S4A). Within each level of deconvolution, differentially expressed genes are first identified by Wilcoxon rank-sum test with multiple testing correction and then used as input.

Deconvolution Using Single Cells from One Subject

To accommodate experimental designs using single cells from only one subject, we adapt the W-NNLS framework to calculate the gene-specific weights by within-subject variation only. Denote the cell-type proportion vector for bulk sample d as $\mathbf{P}_d = (P_{1d}, P_{2d}, ..., P_{Kd})^T$ and the normalized bulk gene expression as $\mathbf{Y}_d = (Y_{1d}, Y_{2d}, ..., Y_{Gd})^T$. The gene-specific expression can be formalized as

$$Y_{dg} - \mathbf{B}_g \mathbf{P}_d = \epsilon_{dg} \sim F\left(\mu_g, \delta_g^2\right),$$

where $\mathbf{B}_g$ is the g^{th} row in the basis matrix $\mathbf{B}$; the residual term ϵ_{dg} follows a certain distribution F with mean μ_g and variance δ_g^2. Adjusting for the variance of residuals, we derive:

$$\frac{Y_{dg}}{\delta_g} - \frac{\mathbf{B}_g \mathbf{P}_d}{\delta_g} = \frac{\epsilon_{dg}}{\delta_g} \sim F\left(\frac{\mu_g}{\delta_g}, 1\right).$$

We can iteratively estimate the proportion vector $\mathbf{P}_d$ and derive the residual vector in the meantime. If two consecutive estimated proportion vectors $\hat{\mathbf{P}}_d$ and $\hat{\mathbf{P}}'_d$ are equal, then we derive a consistent estimation result. That is, if $\|\hat{\mathbf{P}}_d - \hat{\mathbf{P}}'_d\| < a \to 0^+$ and $\hat{\mathbf{P}}_d \approx \hat{\mathbf{P}}'_d$, then

$$\frac{1}{\hat{\delta}_g}\left(Y_{dg} - \epsilon_{dg}\right) - \frac{1}{\hat{\delta}'_g}\left(Y_{dg} - \epsilon_{dg}\right) = \mathbf{B}_g\left(\frac{\hat{\mathbf{P}}_d}{\hat{\delta}_g} - \frac{\hat{\mathbf{P}}'_d}{\hat{\delta}'_g}\right) \approx \mathbf{B}_g \hat{\mathbf{P}}_d \left(\frac{1}{\hat{\delta}_g} - \frac{1}{\hat{\delta}'_g}\right).$$

Hence, as the proportion estimates converge, we derive a final deconvolution result:

$$Y_{dg} - \epsilon_{dg} \approx \mathbf{B}_g \hat{\mathbf{P}}_d.$$

Supplemental Figure Titles and Legends

Figure S1. Empirical results via simulations show that the metrics on gene expression levels $\mathbf{Y}$ are good proxies for the metrics on cell-type proportions $\mathbf{P}$. **A-C**: Prediction errors $\|\mathbf{Y}-\hat{\mathbf{Y}}\|_1$ against Pearson correlation between cell-type proportions $\mathbf{P}$ and $\hat{\mathbf{P}}$ for pseudo-bulk samples constructed using single cells from Åsa Segerstolpe et al. [2016], Baron et al. [2016], and Xin et al. [2016], respectively. **D-F**: Prediction errors $\|\mathbf{Y}-\hat{\mathbf{Y}}\|_1$ against $\|\mathbf{P}-\hat{\mathbf{P}}\|_1$ for pseudo-bulk samples constructed from Åsa Segerstolpe et al. [2016], Baron et al. [2016], and Xin et al. [2016], respectively.

Figure S2. Prediction errors of Y serve as a surrogate for the estimation errors of P. The simulation setups differ from those in Figure 2. **A**: Outline of simulation setup, where single cells of human pancreatic islets from Baron et al. [2016] are aggregated to generate pseudo-bulk samples, whose cell-type proportions are known. We examine the results of deconvolution via ENSEMBLE under two settings, both with and without paired single-cell reference datasets. **B**: mAD($\mathbf{P}-\hat{\mathbf{P}}$) and mAD($\mathbf{Y}-\hat{\mathbf{Y}}$) with three varying dataset-specific weights for deconvolution of bulk samples with paired scRNA-seq. The two metrics agreed on the assignment of the optimal weights: around $(\hat{w}_1, \hat{w}_2, \hat{w}_3) = (1, 0, 0)$. **C**: Spearman correlation and mAD of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, Pearson correlation and mAD of $\mathbf{P}$ and $\hat{\mathbf{P}}$ with varying dataset-specific weights for deconvolution of bulk samples without paired scRNA-seq. The two metrics are highly correlated with varying weights for the reference dataset from Åsa Segerstolpe et al. [2016]. **D**: Outline of simulation setup, where single cells of human pancreatic islets from Åsa Segerstolpe et al. [2016] are aggregated to generate pseudo-bulk samples, whose cell-type proportions are known. **E**: mAD($\mathbf{P}-\hat{\mathbf{P}}$) and mAD($\mathbf{Y}-\hat{\mathbf{Y}}$) with three varying dataset-specific weights for deconvolution of bulk samples with paired scRNA-seq. The two metrics agreed on the assignment of the optimal weights to be around $(\hat{w}_1, \hat{w}_2, \hat{w}_3) = (0, 1, 0)$. **F**: Spearman correlation and mAD of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, Pearson correlation and mAD of $\mathbf{P}$ and $\hat{\mathbf{P}}$ with varying dataset-specific weights for deconvolution of bulk samples without paired scRNA-seq. While the mAD metrics are minimized at different optimal weights, the Spearman correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, and the Pearson correlation between $\mathbf{P}$ and $\hat{\mathbf{P}}$ largely agree with each other.

Figure S3. Single-cell clustering visualization by t-SNE. **A-B**: scRNA-seq

data from the Perou Lab. **C-D**: scRNA-seq data from the Tabula Muris Consortium.

Figure S4. Deconvolution results without the tree-guided approach hardly separate closely related cell types. **A**: Pairwise correlation of cell-type-specific gene expression profiles estimated by scRNA-seq. **B**: Estimated cell-type proportions of mouse mammary gland 10X bulk samples without tree-guided approach. **C**: Estimated cell-type proportions of mouse mammary gland fresh-frozen bulk samples without tree-guided approach.

Figure S5. A first-pass SCDC run on the single-cell reference dataset removes potentially mislabeled cells and doublets. Each single cell is treated as a "bulk" sample and used as input for SCDC. The highly binary cell-type proportions indicate good data quality and reliable cell type clustering. Cells whose estimated cell-type proportions have a maximum less than a user-defined threshold (0.7 by default) are filtered out. These cells are potentially doublets, mis-classified, poorly sequenced, or from cell types not of interest. **A**: A first-pass SCDC run using cells as "bulk" samples. **B**: Unique cell identities after QC.

Figure S6. Number and percentage of single cells grouped by cell type clusters using scRNA-seq data of human pancreatic islets and mouse mammary glands. **A**: Baron et al. [2016]. **B**: Åsa Segerstolpe et al. [2016]. **C**: Xin et al. [2016]. **D**: Perou Lab. **E**: Tabula Muris.

Figure S7. Running time for SCDC with varying number of single-cell references. With less than or equal to three references, both **A** grid search and **B** LAD can finish within 200 seconds. With greater than three references, which may be rare in empirical study, grid search can take longer to run while the computing time for LAD does not change much with the increasing number of references.

References

Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and

intra-cell population structure. *Cell Systems*, 3(4):346 – 360.e4, 2016. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2016.08.011. URL `http://www.sciencedirect.com/science/article/pii/S2405471216302666`.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials revisited. *Contemporary clinical trials*, 45:139–145, 2015.

Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

Yurong Xin, Jinrang Kim, Haruka Okamoto, Min Ni, Yi Wei, Christina Adler, Andrew J Murphy, George D Yancopoulos, Calvin Lin, and Jesper Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell metabolism*, 24(4):608–615, 2016.

Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K. Bjursell, David M. Smith, Maria Kasper, Carina Ämmälä, and Rickard Sandberg. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, 24(4):593 – 607, 2016. ISSN 1550-4131. doi: https://doi.org/10.1016/j.cmet.2016.08.020. URL `http://www.sciencedirect.com/science/article/pii/S1550413116304363`.