Deconvolving spatial transcriptomic data using heterogeneous single-cell datasets

Joshua Sodicoff[1,2], Joshua Welch[1]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[2]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI.

Introduction

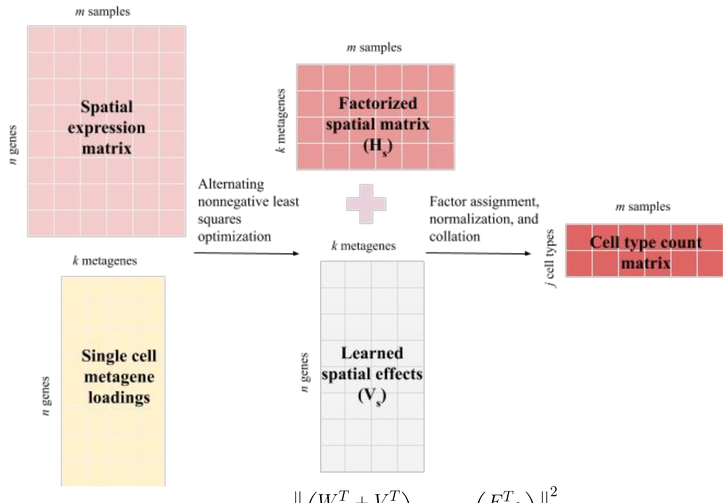
- Single-cell genomic technologies can measure gene expression or epigenomic state of millions of cells within a tissue sample but require tissue dissociation
- Spatial transcriptomic technologies, including those based on fluorescent in situ hybridization or spatially barcoded oligonucleotides, measure expression in space but often at a lower resolution than single cell
- Spatially resolved and dissociated measurements employ different sequencing protocols, creating platform effects that complicate the deconvolution problem.
- The integration of these modalities would allow for the imputation of properties at the single cell level, such as cell state and signalling

Methods

- Integrative nonnegative matrix factorization (iNMF) can jointly decompose multiple expression matrices of dimensions m genes by n_i samples into a k by m shared metagene loading matrix (W), k by m dataset specific metagene loading matrices (V_i) and n_i by k factor loading matrices, which represent the expression in a k-dimensional latent space.

$$\arg\min_{H\geq 0, W\geq 0, V\geq 0} \sum_i^d \left\|E_i - H_i(W+V_i)\right\|_F^2 + \lambda \sum_i^d \left\|H_i V_i\right\|_F^2$$

- We use the learned metagene loadings to iteratively optimize the projection of the spatial data into the k-dimensional latent space (H_S) and a matrix of spatial-specific metagene loadings (V_S).
- To generate cell state proportion predictions, factors are normalized, assigned to specific cell states or removed, and combined to yield a matrix representing cell type proportions or counts per spatial sample.



$$H_s = \arg\min_{H\geq 0} \left\| \begin{pmatrix} W^T + V^T \\ \sqrt{\lambda} V^T \end{pmatrix} H^T - \begin{pmatrix} {E^T}_1 \\ 0_{g\times n} \end{pmatrix} \right\|_F^2$$

$$V_s = \arg\min_{V\geq 0} \left\| \begin{pmatrix} H \\ \sqrt{\lambda} H \end{pmatrix} V - \begin{pmatrix} E - HW \\ 0_{g\times n} \end{pmatrix} \right\|_F^2$$

Results

- To demonstrate the capabilities of ANLS deconvolution, an implementation was applied to the deconvolution of spatial transcriptomic measured with multiple modalities from the mouse brain.
- For each analysis, a metagene loading matrix (W) with k = 40 was generated by running LIGER, an R implementation of iNMF for single cell data, on a reference single cell transcriptomic dataset matched from the same brain region.
- The alternating nonnegative least squares deconvolution was run with 5 random starts with $\lambda = 10$

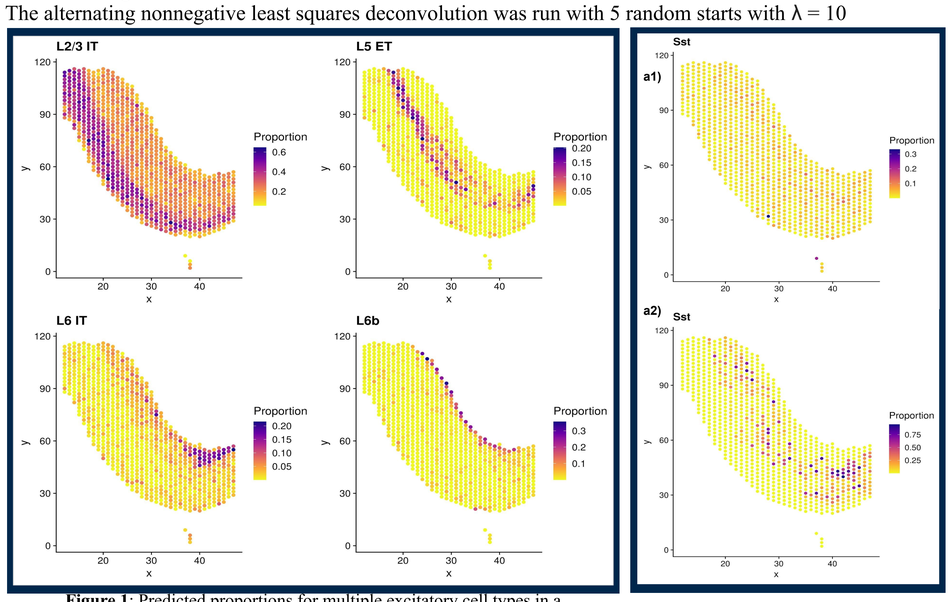


Figure 1: Predicted proportions for multiple excitatory cell types in a sagittal section of mouse frontal cortex measured with the 10X Visium spatial sequencing platform. The method correctly describes the laminar nature of the types with only limited loss of specificity on account of overlap in cell type markers.

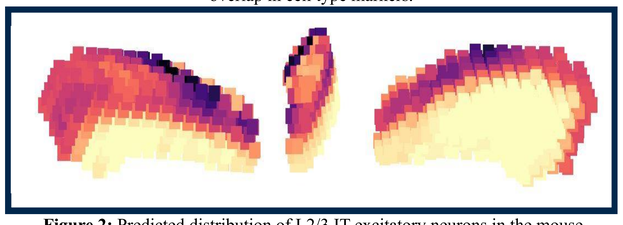


Figure 2: Predicted distribution of L2/3 IT excitatory neurons in the mouse primary motor cortex, from quantified and spatially registered in situ hybridization data from the Allen Mouse Brain Atlas. While the modality differs from other spatial transcriptomic protocols in its dimensionality, distribution and resolution, ANLS deconvolution still correctly identifies the correct layer distribution

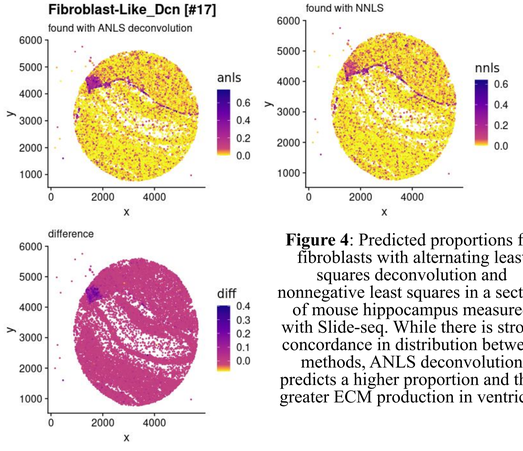


Figure 4: Predicted proportions for fibroblasts with alternating least squares deconvolution and nonnegative least squares in a section of mouse hippocampus measured with Slide-seq. While there is strong concordance in distribution between methods, ANLS deconvolution predicts a higher proportion and thus greater ECM production in ventricle.

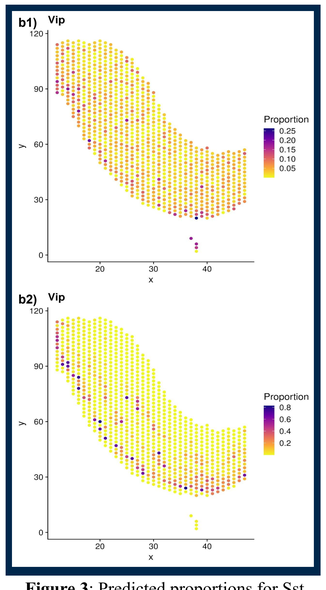


Figure 3: Predicted proportions for Sst and Vip interneurons from the same dataset as figure x, generated with our method (a1 and b1) and Seurat (a2 and b2). ANLS deconvolution again correctly predicts locations of the types while retaining the expected diffusion. Seurat's predictions match only in the most dense regions and do not account for regions with low but non-negligible expression.

Discussion

- ANLS deconvolution is capable of correctly predicting the location of neuronal cell types across
  - Multiple regions with a myriad of laminar structures and diversities of cell types
  - Multiple modalities with different dimensionalities, resolutions, number of genes, and measurement techniques
- While the relative distribution is accurate, the true proportion is likely inaccurate
  - Cell size may be important in normalizing proportions, but is confounded by preprocessing and gene selection
  - Determining ground truth values in a tissue can be difficult on account of dissociation bias and limits on spatial resolution associated with single cell data
- Choice of hyperparameters, especially k and those associated with the factor assignment, can significantly impact the results
- The current method for factor assignment, based on a one-tailed t-test of factor loadings between cells within and outside of the cluster, does not always result in deconvolution of all cell types in the reference dataset.

Conclusions and Direction

- ANLS deconvolution provides several benefits over other existing deconvolution techniques, including
  - Limited assumptions as to the underlying distribution of expression in both the spatial and reference single cell data, allowing for application to a wider array of modalities
  - The ability to discern the presence of small populations of cells
  - Integration with LIGER, allowing for imputation of high resolution single cell data across spatial data
- Several elements of the method must still be studied, namely
  - How the generation of W (from a single reference dataset, multiple reference datasets, across multiple modalities) impacts the quality of the deconvolution
  - The significance of the learned spatial effects
    - Can they be used to correct batch effects in bulk RNAseq of the tissue?
    - Can they be used to iteratively modify W to remove batch effects from the single cell data?
  - How the similarity of cell states present in the single cell data impacts the quality of the deconvolution

Acknowledgement

Special thanks to

- Dr. Joshua Welch
- Kinsey Van Deynze
- NIH Brain Initiative Cell Census Network

References

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Irizarry, R. A. (2020). Robust decomposition of cell type mixtures in spatial transcriptomics. *bioRxiv*.

Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, *10*(1), 1-9.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, *177*(7), 1873-1887.

Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R. S., Aldrige, A. I., ... & Bertagnolli, D. (2020). An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *Biorxiv*.