Expanding Gene Expression profile in Targeted Spatial Transcriptomics through scRNA-seq data integration

Authors: Sumanth Reddy N[1], Riya Podder[1], Rahul Tyagi[1]
[1] Elucidata Inc.

Abstract

We present an improved methodology for **gene expression imputation (GEI)** in Xenium, a targeted spatial transcriptomics platform. Xenium provides single-cell resolution but measures only a limited set of genes, restricting downstream analyses. Our approach of aligning spatial and reference scRNA-seq data through "Symphony" in a common embedding space, integrating them using single-cell neighborhood information improved recovery of genome-wide expression profiles from targeted panels, enabling more accurate downstream analysis.

Introduction

Targeted spatial transcriptomics technologies such as Xenium, CosMx, and MERFISH provide single-cell resolution but are limited to a fixed panel of pre-selected genes. "Gene expression imputation (GEI)", the task of inferring genome-wide expression profiles from these targeted panels is therefore critical for enabling more nuanced downstream analysis such as cell phenotyping, cell-cell communication analysis and profiling of tissue microenvironment. Here, we explored multiple imputation tools that leverage paired or reference single-cell RNA-seq (scRNA-seq)datasets. While existing methods can project scRNA-seq information into the Xenium space, we observed that their predictions are poorly correlated with ground truth values. To address this, we explored different methods and different benchmarking metrics to improve the imputation quality and result in the imputation more aligned with scRNA-seq data. We found that "Symphony" integration followed by expression aggregation over the common embedding-space that leverages single-cell neighbor relationships for each spatial cell yielded improved results.

Methods

We systematically compared imputation methods leveraging paired scRNA-seq reference data in kidney tissue, through latent-space alignment approaches. Our pipeline involves Preprocessing of Xenium data and matched scRNA-seq datasets. Followed by Integration via Symphony to align latent embeddings across modalities. The **GEI** is performed by weighted average of the top 50 nearest neighbors of single cells corresponding to each spatial cell in the common embedding space. We compared the integration performance for various GEI tools - Tangram, ENVI, GIMVI, Seurat CCA, Seurat RPCA, Harmony and Symphony using UMAP visualization. To evaluate the imputation performance, we used Pearson correlation coefficient (PCC) scores between ground truth and imputed values, visual inspection of tissue structural regions and gene expression UMAPs for three mutually exclusive marker gene pairs. For the purpose of PCC calculation, we smoothened both ground truth and imputed gene expression for each xenium cell by taking the average of expression of the 10 nearest xenium neighbor cells in the integrated symphony UMAP coordinate space and the cell's expression itself.

Results & Discussion

The Symphony integration achieved the most effective alignment of the two data modalities, visually confirmed by the superior overlap in the UMAP embedding. Leveraging this strong integration, we found that Symphony-based imputation of mutually exclusive marker gene pairs robustly enhanced gene expression signals compared to ground-truth levels. Furthermore, the imputation successfully preserved critical biological patterns, maintaining both mutual exclusivity at the cluster level and the spatial integrity of the genes. The high accuracy of this approach was validated by strong PCC scores (0.75–0.98) between the ground-truth and imputed data, underscoring its potential as a highly efficient method for GEI.

Conclusion & Future Work

Our findings highlight the value of finding embedding-based integration methods that perform a gene expression imputation in targeted spatial platforms. Future extensions will explore adapting the method to different tissue types to validate the robustness and exploring the performance with various hyperparameter variations.

References

1. Kang, J.B., Nathan, A., Weinand, K. *et al.* Efficient and precise single-cell reference atlas mapping with Symphony. *Nat Commun* 12, 5890 (2021). https://doi.org/10.1038/s41467-021-25957-x
2. Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, Marcel J T Reinders, SpaGE: Spatial Gene Enhancement using scRNA-seq, *Nucleic Acids Research*, Volume 48, Issue 18, 09 October 2020, Page e107, https://doi.org/10.1093/nar/gkaa740
3. Biancalani, T., Scalia, G., Buffoni, L. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods 18, 1352–1362 (2021). https://doi.org/10.1038/s41592-021-01264-7
4. Haviv, D., Remšík, J., Gatie, M. *et al.* The covariance environment defines cellular niches for spatial inference. *Nat Biotechnol* 43, 269–280 (2025). https://doi.org/10.1038/s41587-024-02193-4
5. Butler, A., Hoffman, P., Smibert, P. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018). https://doi.org/10.1038/nbt.4096
6. Lopez, R., Nazaret, A., Langevin, M. et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv (2019). https://arxiv.org/pdf/1905.02269v1