OBSERVED COSINE SIMILARITY WITH GCODE PREDICTIONS VIA TRANSFORMS

David Banh
Data Science
AskExplain
St Lucia, Brisbane, Australia
david.b@askexplain.com

ABSTRACT

Previously, Pearson correlation was used as a measure of accuracy - however, this lead to a poor representative metric for similarity between observed and predicted values due to centering of values by the expectation in the correlation formulae. Here, cosine similarity is used to compare between gene vectors and sample vectors of the same label or ID between observed and predicted. Metrics improve by a large margin for gene similarity measure, whereas sample similarity measures are maintained.

1 Introduction

Four spatial transcriptomic datasets are used on different tissues: human adipose tissue [Bäckdahl et al., 2021], human breast tumour tissue [Ståhl et al., 2016], human developing heart tissue [Asp et al., 2019] and mouse spinal tissue [Maniatis et al., 2019].

For each tissue, the gene expression spot, and the pixel spot cropped as a square are taken and assigned the same ID in separate matrices (the coordinate files were used to extract an appropriate sized pixel spot). Generative Encoding is used to learn a transformation between gene expression and image by learning a set of orthogonal components that are representative of both tissue image and gene levels. From the learned transformation, a test set of pixel spots are transformed to the predicted gene expression and compared to the with-held observed gene expression values.

Previously Pearson correlations were used and while this resulted in a high correlation between genes for individual sample spots, it did not score highly the correlation between spots for individual genes. The main reason for this is because the formulae expression for correlation centers the genes by the expectation, which scores the non-linear "relation" between spots for individual genes poorly. An alternative measure of similarity that is more flexible at representing similarity of differing magnitudes is cosine similarity, used in a method to optimise for gene alignment to spatial tissue [Biancalani et al., 2021]. Cosine and correlation metrics are quite similar - the main difference is the lack of centering by the expectation in the cosine similarity measure that appears in correlation.

2 Methods

2.1 Baseline

The model for Generative Encoding is learned on a train set, and the validation is done on a test set of pixels - transformed to predictions of gene expression and compared with a with-held test set of observed gene expression.

This is to measure the baseline accuracy of a direct transformation from spot to gene expression.

2.2 Spatially displaced perturbation

The spatial tissue spot window is moved from the center to capture a displaced spatial tissue spot representing the four quadrants of a square window that overlap the border and other areas of the histology outside of the observed spot window. With cosine similarity, the predicted gene expression from these spatially displaced perturbations are compared to the observed gene expression, originally measured by the Spatial Transcriptomics platform device experimentally for the corresponding spot.

This is to measure the robustness of the accuracy of a spatially displaced spot that still overlaps with the original spot, as well as, other histology imaging tissue not within the original spot.

3 Results

3.1 Human Adipose Tissue (Backdahl et al, 2021)

3.1.1 Baseline cosine similarity

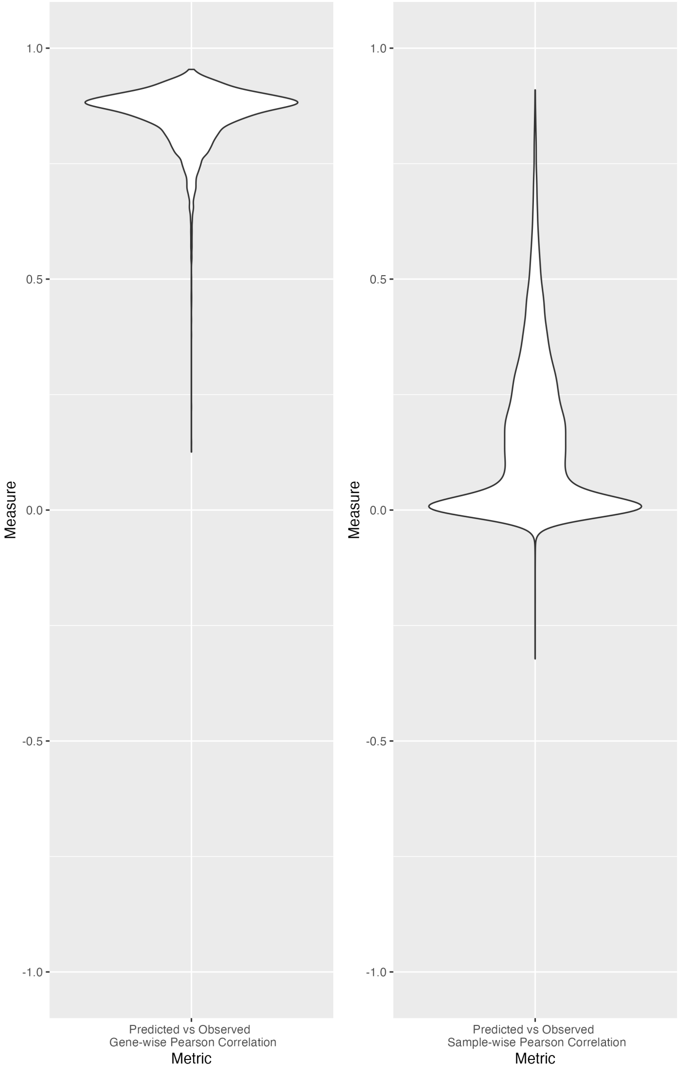

Figure 1: Baseline correlation accuracy for human adipose tissue [Bäckdahl et al., 2021]

3.1.2 Spatially displaced perturbed cosine similarity

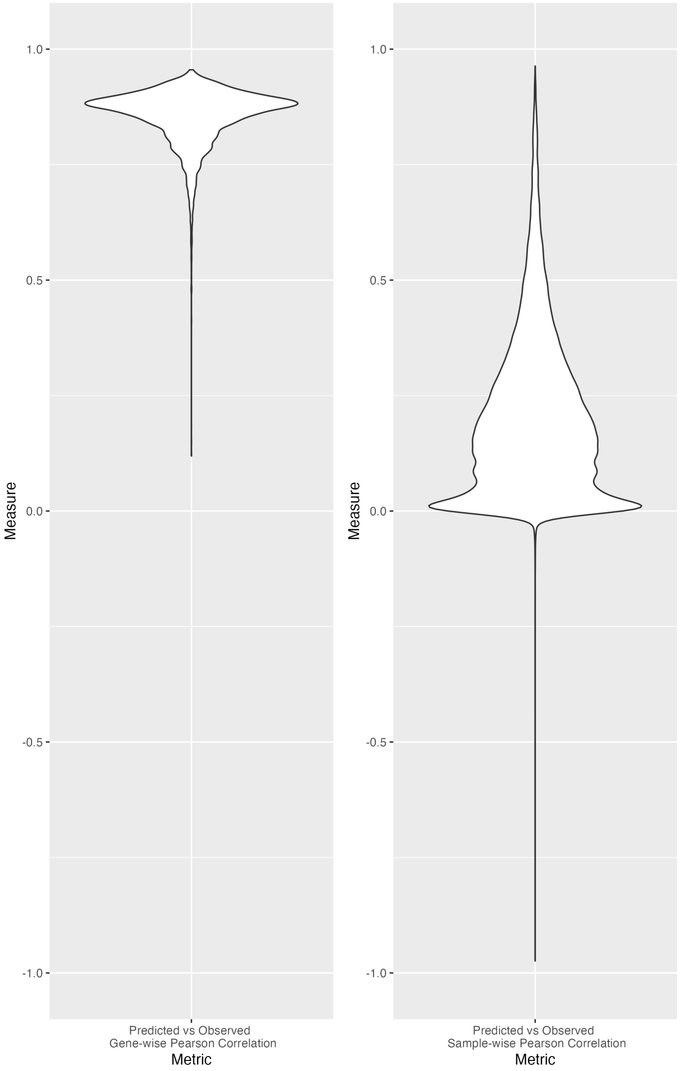

Figure 2: Spatially displaced perturbation correlation accuracy for human adipose tissue [Bäckdahl et al., 2021]

3.2 Human Breast Tumour Tissue (Stahl et al, 2016)

3.2.1 Baseline cosine similarity

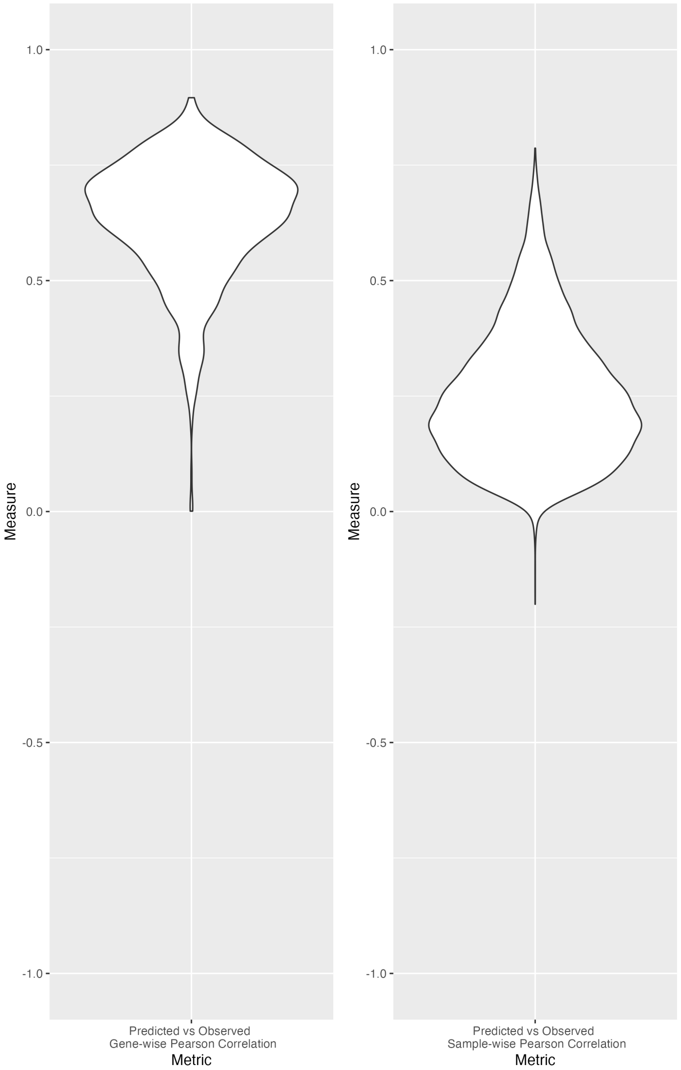

Figure 3: Baseline correlation accuracy for human breast tissue [Bäckdahl et al., 2021]

3.2.2 Spatially displaced perturbed cosine similarity

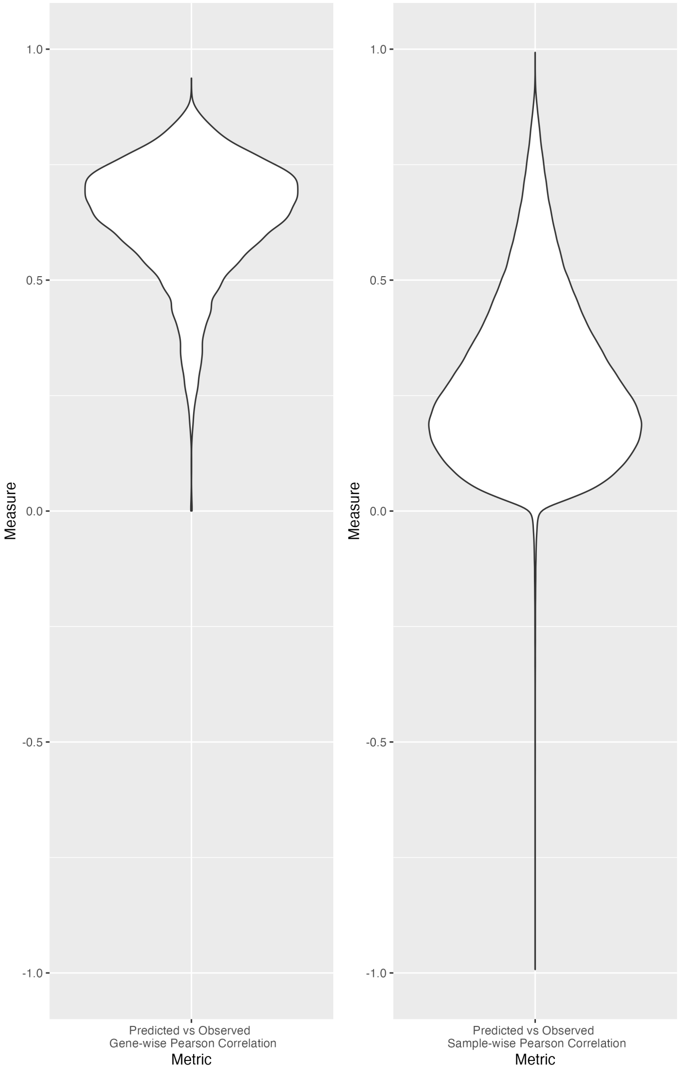

Figure 4: Spatially displaced perturbation correlation accuracy for human breast tissue [Bäckdahl et al., 2021]

3.3 Human Developing Heart Tissue (Asp et al, 2019)

3.3.1 Baseline cosine similarity

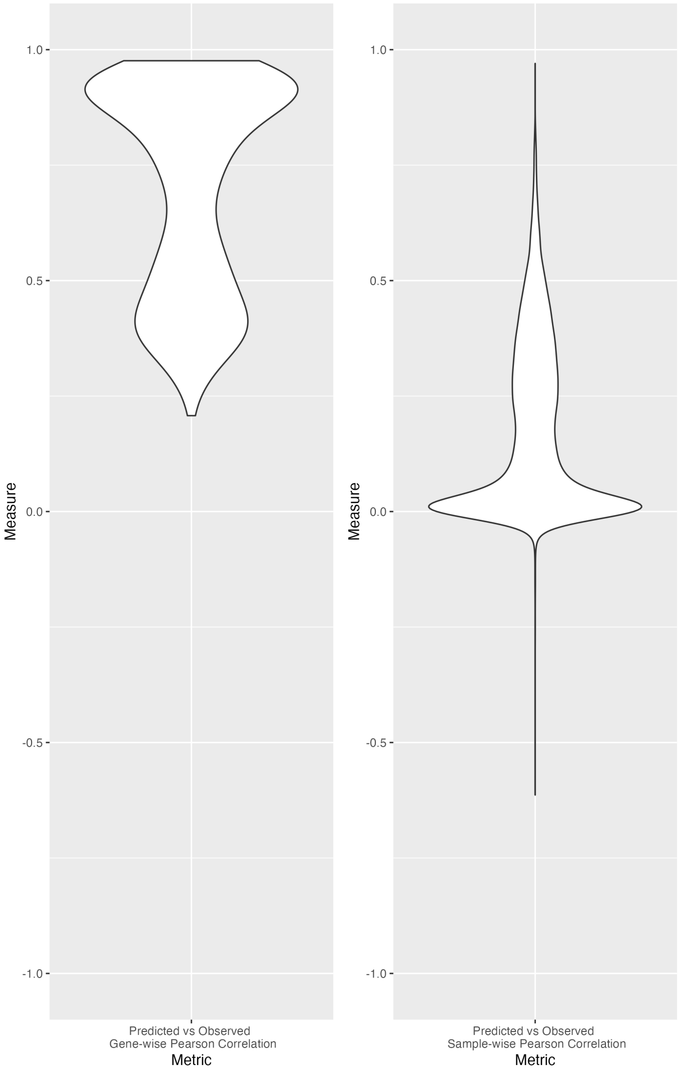

Figure 5: Baseline correlation accuracy for human fetal$_heart tissue$
[Bäckdahl et al., 2021]

3.3.2 Spatially displaced perturbed cosine similarity

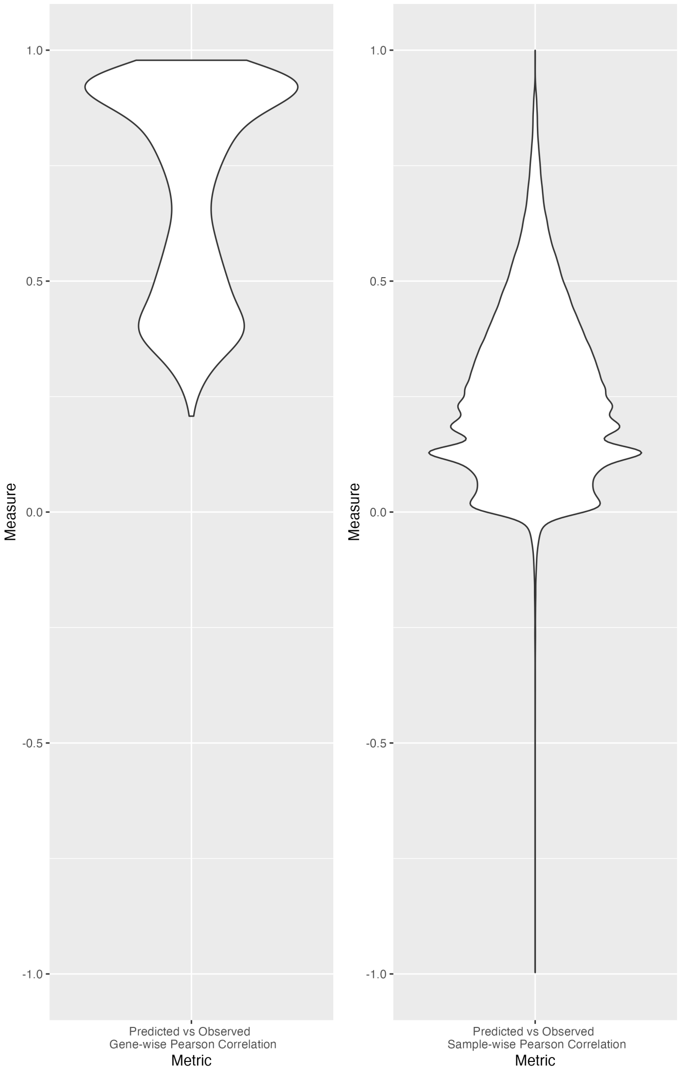

Figure 6: Spatially displaced perturbation correlation accuracy for human fetal$_{h}earttissue$[Bäckdahl et al., 2021]

3.4 Mouse Spinal Tissue (Maniatis et al, 2019)

3.4.1 Baseline cosine similarity

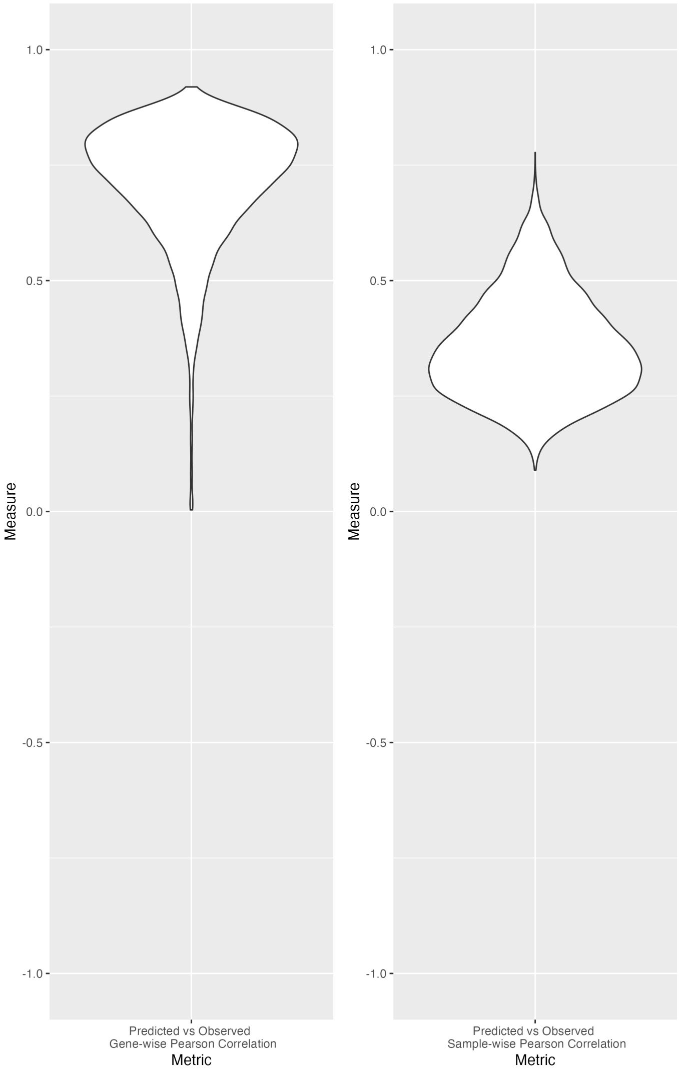

Figure 7: Baseline correlation accuracy for mouse spinal tissue [Bäckdahl et al., 2021]

3.4.2 Spatially displaced perturbed cosine similarity

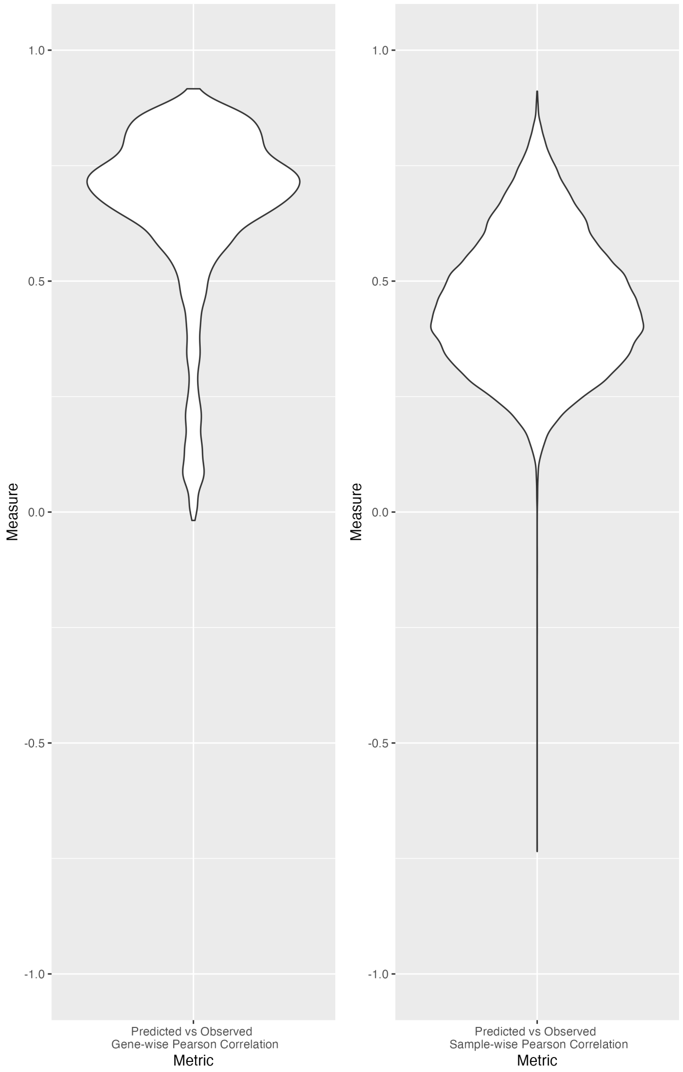

Figure 8: Spatially displaced perturbation correlation accuracy for mouse spinal tissue [Bäckdahl et al., 2021]

References

[Asp et al., 2019] Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., Österholm, C., Ståhl, P. L., Sundström, E., Åkesson, E., Bergmann, O., Bienko, M., Månsson-Broberg, A., Nilsson, M., Sylvén, C., and Lundeberg, J. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660.e19.

[Bäckdahl et al., 2021] Bäckdahl, J., Franzén, L., Massier, L., Li, Q., Jalkanen, J., Gao, H., Andersson, A., Bhalla, N., Thorell, A., Rydén, M., Ståhl, P. L., and Mejhert, N. (2021). Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metabolism*, 33(9):1869–1882.e6.

[Biancalani et al., 2021] Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C. R., Segerstolpe, Å., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N. B., Fanelli, D., Zhuang, X., Macosko, E. Z., and Regev, A. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature Methods*, 18(11):1352–1362.

[Maniatis et al., 2019] Maniatis, S., Äijö, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D., Žaneta Andrusivová, Saarenpää, S., Saiz-Castro, G., Cuevas, M., Watters, A., Lundeberg, J., Bonneau, R., and Phatnani, H. (2019). Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93.

[Ståhl et al., 2016] Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Åke Borg, Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.