

Assessing Normalization Methods in Spatial Transcriptomics

Benton Anderson, Stat 877, 5/4/2022

Introduction

Spatial transcriptomics is an important development in understanding complex biological samples using RNA sequencing. Whereas bulk RNAseq and single cell RNAseq can provide an averaged or single-cell level view of a complex tissue sample, spatial RNAseq has unique utility in that it retains the spatial relationship between spots and allows inspection of the tissue sample in 2 or even 3 dimensions, while still providing similar information as bulk tissue sequencing or single cell experimentation.

As with all RNAseq workflows, it is essential to perform an appropriate pre-processing of the raw data, accounting for various factors that confound the true biological signal behind technical noise. Normalization is an important step in raw data pre-processing, as appropriate normalization whose assumptions match the true confounders of the data collection process will boost biological signal and minimize variance from within-sample or between-sample variance¹.

To study the effects of normalization on spatial transcriptomics data, it is convenient to utilize a set of samples studied and analyzed with high-quality methods in sample preparation, sequencing and data analysis. To that end, we have chosen to use the analysis from Maynard et al.² that uses dorsolateral prefrontal cortex samples acquired from 3 human brains, which were processed using the 10x Genomics Visium method and then subjected to numerous data methods. Their efforts yielded data with distinct cellular layers of the brain in that region annotated and validated to provide a ground truth assignment of the dominant cell type within each spot. Additionally, by testing many different clustering methods, both unsupervised and semi-supervised, the authors have explored the landscape of possible clustering techniques that show meaningful patterns in cell types and differential expression of genes within cell layers. Notably, the authors only tested one normalization method, and therefore the dataset is ripe for reanalysis with a focus on varying only the normalization used.

Background

In the Spatial LIBD paper, the authors used the logNormCounts normalization in scater³ which computes a size factor based on total library size for each spot, then log2 transforms the counts after adding a pseudocount. Given the lack of normalization methods specific to spatial transcriptomic data, this method can be considered appropriate given that it is unknown whether the assumptions present in other, more advanced single-cell specific normalization methods will hold in this case. Researchers have questioned the utility of single cell specific methods when used in spatial transcriptomics due to questions about whether the counts obtained per spot reflects its true cellular contents versus technical variance in transcript amplification⁴.

Other popular transformations include SCTransform⁵ which uses a regression model with a negative binomial error. This model makes the key assumption that most of the genes in a sample do not display biological variation, which seems plausible to hold for spatial samples given that any cells that fall within the spot also fall close to each other, and will therefore likely be of the same cell type and have similar gene expression. However, this assumption can be compromised by tissue types known to have high cell diversity within a single spot, such as the kidney⁶. Interestingly, the authors comparing kidney samples in different batches show that because using batch as a variable in SCTransform only slightly improves gene expression alignment of housekeeping genes, this suggests that most of the technical variation in our samples can be modeled by sequencing depth alone.

Methods

Package and language versions:

Python 3.7 and R 4.2 with Bioconductor version 3.15 was used to access packages. Python scripts are located within Jupyter notebooks.

Data sourcing:

Raw data for the Spatial LIBD 10x Genomics Visium experiments was accessed in two ways. In R, data were accessed via the spatialLIBD package. In Python, raw data were read in from the filtered .h5 files located at spatial.libd.org/spatialLIBD. Note that only the filtered data were used, as they corresponded to the data accessed in the R package. For the gene and spot metadata, dataframes for each were saved as .csv files from R using the spatialLIBD functions colData and rowData. Only 6 total samples were used, corresponding to the two 0 μ m depth replicates from each of the 3 brains.

Gene filtering:

To minimize problems with computation time and memory constraints, the number of genes was filtered to include only genes with summed raw expression across all 12 samples greater than 10,000 counts, leaving 2,831 genes.

Normalization:

All normalizations were applied to each replicate sample individually after spot filtering for non-tissue spots, and gene filtering. The 6 data matrices used in normalization have size approximately 4,000 by 2,831.

SCTransform⁵ and Dino⁷ were accessed using Bioconductor/R packages Seurat and Dino. The normalization functions were applied with all default parameters.

In addition, Python scanpy.preprocessing.normalize_total function was used to perform count-based normalization, with all default parameters except `exclude_highly_expressed=True`.

Additionally, a non-normalized dataset was used in the following steps as negative control.

Unsupervised clustering:

All 3 normalized and 1 non-normalized dataset were subjected to PCA with 50 PCs using `scanpy.preprocessing.pca`. KMeans clustering was performed using `scikit-learn` using `n_clusters=7` or `5`, depending on the number of layers found in the `spatialLIBD` column in the spot metadata table, where the 151507 and 151673 brains had 7 layers (L1 – L6 and WM), but the 151669 brain was annotated with only 5 layers.

Measuring clustering performance:

To assess performance, both `accuracy_score` and `balanced_accuracy_score` from `sklearn.metrics` were used to compare the ground truth assigned label in the `spatialLIBD` metadata against the KMeans cluster assignment. Because the clustering was unsupervised, the KMeans cluster label is assigned arbitrarily, and therefore each of the permutations of the 7 or 5 KMeans cluster labels was tested against the `spatialLIBD` cluster assignment to find the permutation with the maximum accuracy score. The maximum score is given in the results.

Code sharing:

Project files can be found within GitHub repository [benton-anderson/877](https://github.com/benton-anderson/877), with all files located in the `/project` folder.

Results

Figure 1 Results from 0 um replicates 1 and 2 of all 3 brain samples. Ground truth represents labels from the 'spatialLIBD' column of metadata in published dataset. HVG spatial PCA is the best unsupervised clustering algorithm according to the adjusted rand index metric as used in the paper, corresponding to the column 'HVG_PCA_spatial' in the metadata. The following 4 columns give the result and accuracy of each normalization method for each sample.

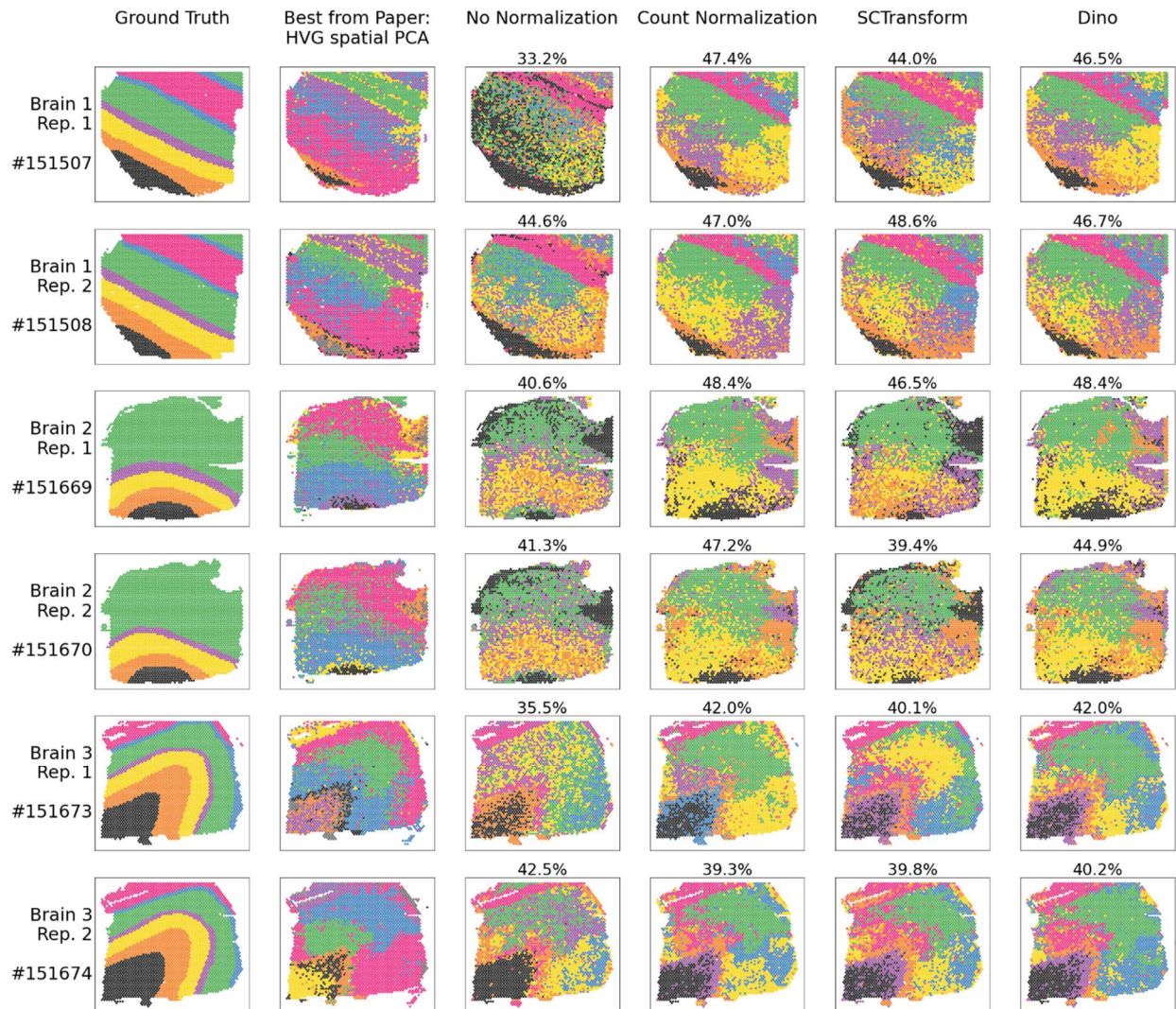


Table 1 Percent accuracy scores when comparing normalization to ground truth label.

	No norm.	TPM	SCTransform	Dino
Brain 1 rep 1	33.2	47.4	44.0	46.5
Brain 1 rep 2	44.6	47.0	48.6	46.7
Brain 2 rep 1	40.6	48.4	46.5	48.4
Brain 2 rep 2	41.3	47.2	39.4	44.9
Brain 3 rep 1	35.5	42.0	40.1	42.0
Brain 3 rep 2	42.5	39.3	39.8	40.2

Table 2 Percent balanced accuracy scores comparing normalization to ground truth label. Balanced accuracy weights each label according to its proportion in the sample.

	No norm.	TPM	SCTransform	Dino
Brain 1 rep 1	34.6	40.6	45.0	43.4
Brain 1 rep 2	46.4	42.4	47.8	44.6
Brain 2 rep 1	36.8	44.6	38.5	44.4
Brain 2 rep 2	39.7	41.0	37.4	41.7
Brain 3 rep 1	37.1	39.8	44.5	41.6
Brain 3 rep 2	41.3	36.6	37.3	36.9

Discussion

Applying any of the 3 normalization methods generally showed any improvement over using no normalization, with the exception of sample 151674, where the non-normalized control had a superior accuracy score, 42.5%, versus 39.3%, 39.8% and 40.2% for TPM, SCT and Dino respectively. The source of this discrepancy is unexpected given the opposite trend in its paired replicate and merits further study. One notable feature in the spot-labeled visual is that the black chunk of white matter in the lower left is highly homogeneous, unlike in the 3 normalized samples. This suggests that something about the white matter in this sample was particularly distinct, yet sensitive to normalization, and thus boosts the accuracy score.

Comparing accuracy to balanced accuracy in Tables 1 and 2 shows that the overall trends are approximately the same, where any of the normalizations providing on average a superior accuracy score, and the Brain 3 rep 2 still showing a decrease in accuracy with normalization. Notably, the balanced accuracy scores are overall lower than the accuracy scores, suggesting that the imbalance of labels does play a meaningful role in the boost in the regular accuracy score.

As mentioned in the presentation, the non-normalized data show higher rates of spottiness, where labels intermix among themselves more often than in the normalized data, where there tend to be more blocks of continuous identical labels. While it would be

convenient to think that this is a real feature of the data, it is not necessarily biologically true, as there could be a biological source for the spottiness in the data where close proximity does not correlate with identical gene expression. This could suggest that within the spotty areas of the brain sample, there are spots that tend to overrepresent one type of cells showing differential expression from its near neighbors.

References

1. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).
2. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
3. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
4. Saiselet, M. *et al.* Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *J. Mol. Cell Biol.* **12**, 906–908 (2020).
5. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
6. Melo Ferreira, R., Freije, B. J. & Eadon, M. T. Deconvolution Tactics and Normalization in Renal Spatial Transcriptomics. *Front. Physiol.* **12**, (2022).
7. Brown, J., Ni, Z., Mohanty, C., Bacher, R. & Kendzierski, C. Normalization by distributional resampling of high throughput single-cell RNA-sequencing data. *Bioinformatics* **37**, 4123–4128 (2021).