Rescuing Sparse Gene Expression Data Using snRNA-seq

Andy Han[1], Kirsten Nishino[2], Adam Reeson[3]

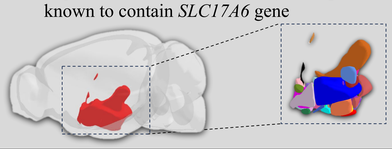
[1]Georgetown University, [2]University of Michigan, [3]Marquette University

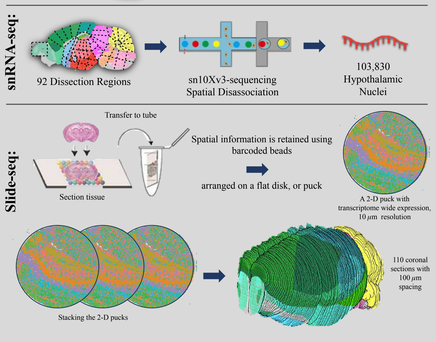
Abstract:

- Fluorescence is a common technique to study the molecular activity in a mouse brain.
- Despite the given Slide-seq dataset already containing gene expression for each bead, sparsity of the dataset makes drawing conclusions difficult.
- Our study provides a computational solution by utilizing sn10Xv3 data to decrease the sparsity of gene expression throughout the mouse hypothalamus. This not only increases the comprehensiveness of our gene expression data, but also offers an alternative to using fluorescence to study mouse brain activity.
- After creating a new gene expression matrix by incorporating sn10Xv3 data through a proposed algorithm, sparsity of the gene expression decreased.
- By testing our algorithm on *SLC17A6*, a well-studied gene, expression was more visible throughout the mouse hypothalamus in 3D plots, particularly in the suprachiasmatic nucleus (SCH) region as we would expect.

Background:

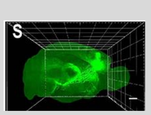
- Hypothalamus:
  - Coordinates endocrine and cardiovascular systems
  - Regulates circadian rhythms, food intake, and reproduction
  - Composed of many nuclei
  - Suprachiasmatic nucleus (SCH) subregion – known to contain *SLC17A6* gene





Motivation:

- Create a computational alternative to using fluorescence to examine molecular processes in the mouse brain.

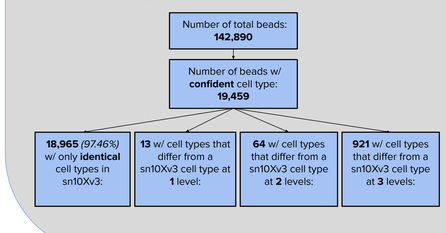


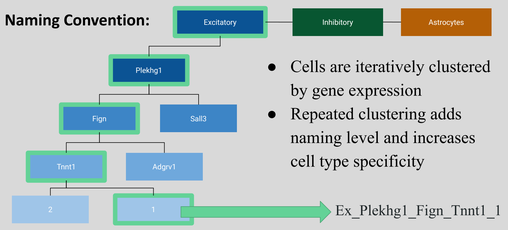
Research Question: How can sparsity in a Slide-seq gene expression matrix be overcome to better analyze the spatial distribution of specific genes throughout the mouse hypothalamus?

Methods:

Finding Common Cell Types

- Ideally, if the snRNA-seq and Slide-seq datasets had all cell types in common, we would be able to confidently interchange their molecular profiles. However, only **436** out of the **658** Slide-seq full cell type names are found in the snRNA-seq data.
- We propose **'peeling'**: sequentially removing the last cluster label for cell type names until a match is found between both datasets.
  - **658** identified Slide-seq cell types in hypothalamus:
    - **436** w/ **full** match, **4** w/ **1** different level, **24** w/ **2** different levels, **141** w/ **3** different levels
- Once a match is found, compute average gene expression across cells with that cell type name.
- After peeling at most 3 times, gene expression profiles for **605** Slide-seq cell types were obtained.

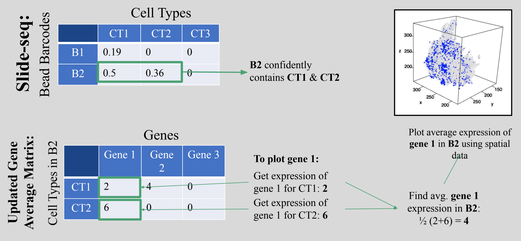




- Cells are iteratively clustered by gene expression
- Repeated clustering adds naming level and increases cell type specificity

Average Gene Expression Imputation Algorithm

- Create a list of beads that confidently contain cell types in the hypothalamus, and therefore, ideally, exhibit some level of gene expression that we can find in the sn10Xv3 dataset.
- Iterate through every bead in the Slide-seq dataset. In each bead, iterate through all the confident cell types, use the sn10Xv3 to find the average expression for every gene in the cell type, and 'peel' the last cluster labels of the cell type name when necessary to estimate gene expression. The calculated average gene expression per bead is outputted and stored in a gene new expression matrix. Repeat this process for every bead in the Slide-seq dataset.



Data Processing:

- **Read Count Normalization across cells:** divide each gene count by total # of reads for that cell, then multiply by scaling factor, i.e. $k = 10^6$

sn10Xv3: **original**

	Gene A	Gene B	…	Total
$Cell_i$	a_i	b_i	…	n_i

sn10Xv3: **normalized**

	Gene A	Gene B	…
$Cell_i$	$(a_i / n_i) * k$	$(b_i / n_i) * k$	…

"Expected count for gene A per k reads in cell i"

sn10Xv3: **normalized**

103,830 cells

Cell Type	Gene A	Gene B
CT1	x_A	x_B
CT2	y_A	y_B
CT1	z_A	z_B

21,899 genes

sn10Xv3: **gene averages by cell type**

1,125 cell types

Cell Type	μ(Gene A)	μ(Gene B)
CT1	$½(x_A + z_A)$	$½(x_B + z_B)$
CT2	y_A	y_B

21,899 genes

Results:

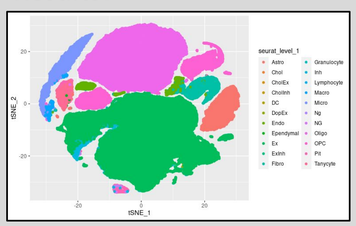


Figure 1. t-SNE plot of cells detected in mouse hypothalamus for sn10Xv3 data, labeled by first cell type level.

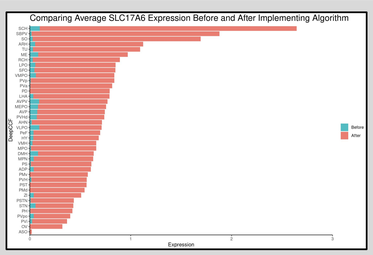


Figure 2. Histogram comparing average *SLC17A6* expression before and after implementing algorithm. 'Before' is aqua and is a direct plotting of the Slide-seq gene expression matrix. 'After' is red and utilizes our algorithm to find the gene expression per bead using the sn10Xv3 data.

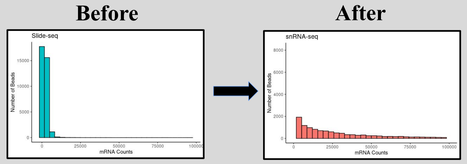
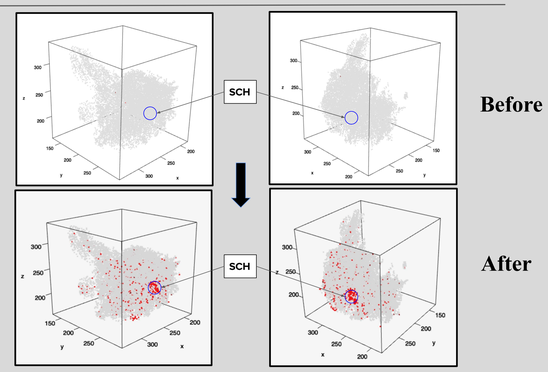


Figure 3. Histograms of total mRNA counts for all genes before and after implementing algorithm. The 'after' histogram demonstrates a more uniform distribution of reads across beads.



* Beads in upper 50% of expression range plotted in red

Figure 4. Comparing the mappings of *SLC17A6* gene expression across the mouse hypothalamus before and after implementing algorithm. Before the algorithm, **97.3%** of beads contained no reads of *SLC17A6*. After the algorithm, only **9.9%** of beads contained no reads of *SLC17A6*.

Discussion:

- We see a noticeable difference in *SLC17A6* expression between the Slide-seq dataset and the updated dataset using the sn10Xv3 data with our proposed algorithm.
- As expected, the *SLC17A6* reads are noticeably higher in the suprachiasmatic nucleus (SCH), which is known to regulate mammalian circadian rhythms.
- This strongly corresponds to the fact that the *SLC17A6* gene encodes the vGLUT2 protein, which regulates feeding patterns, glucose metabolism, and circadian rhythm.
- Limitations:
  - Proposed algorithm did not determine average gene expression for every cell type in hypothalamus (**53** hypothalamic cell types were NOT classified).
  - Did not scale average gene expression values downward for "peeled" Slide-seq cell types to account for lower accuracy.
- Future Directions:
  - How does gene expression differ between a healthy vs. diseased mouse?
    - Using GWAS, look at specific gene variants known to be associated with certain neurological disorders.
  - For calculated gene expressions for cell types that did not perfectly correspond, create some kind of confidence metric to scale the average gene expression when finding the total gene expression per bead.
  - Is the distribution of certain genes across the hypothalamus random? Or is there significance in the spatial distribution of gene expression?
  - Explore alternative methods for determining cell type confidence to include more beads in analysis.

References:

1. Langlieb, et al. 2023. The cell type composition of the adult mouse brain revealed by single cell and spatial genomics. bioRxiv 2023.03.06.531307; doi: https://doi.org/10.1101/2023.03.06.531307
2. Rodriques, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019 Mar 29;363(6434):1463-1467
3. Gao, et al. terative single-cell multi-omic integration using online learning. Nat Biotechnol 39, 1000–1007 (2021).
4. Zhang, et al. Multi-Scale Light-Sheet Fluorescence Microscopy for Fast Whole Brain Imaging. Frontiers in Neuroanatomy 15, 2021.09.24; doi: https://doi.org/10.3389/fnana.2021.732464

Acknowledgements:

- Thank you to Dr. Matt Zawistowski, April Kriebel, Dan Barker, Dr. Bhramar Mukherjee, & entire BDSI 2023 cohort.