Uncovering Regional, Cellular, and Biological Patterns of Gene Expression in the Mouse Brain Using Semi-Supervised NMF

Kyla Woyshner[1], Yi Wang[2], Kasper Hansen[1,2], Genevieve Stein-O'Brien[1,3,4,5], and Loyal Goff[1,3,4]

[1]Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.,
[3]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, USA. [4]Kavli Neurodiscovery Institute, Johns Hopkins University, Baltimore, USA.
[5]Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, USA.

Introduction

The mammalian brain is a highly complex organ composed of various cell types and functionally distinct regions. Understanding the intricate orchestration of gene expression across cells within the brain is crucial for comprehending its function, development, and pathology. High-throughput spatial transcriptomic technologies enable us to obtain gene expression data from individual cells within brain tissue. However, dealing with the vast amount of transcriptomic data can be difficult.

Non-negative matrix factorization (NMF) is a dimensionality reduction technique that has been successfully applied in genomic research to identify interpretable patterns in gene expression data[1,2,3]. However, NMF is generally unsupervised, and thus limited in its ability to integrate prior biological knowledge. In this study ***we employ semi-supervised NMF*** to analyze gene expression data from the adult mouse brain, incorporating prior knowledge of functionally distinct anatomical regions in the brain. ***This allows us to discover features beyond existing annotations as well as learn new gene correlations to known regions.*** This study serves as a proof of concept for the utility of semi-supervised NMF in deciphering complex biology, with the ultimate goal of creating an efficient, scalable semi-supervised NMF method for a more comprehensive understanding of brain function.

Methods

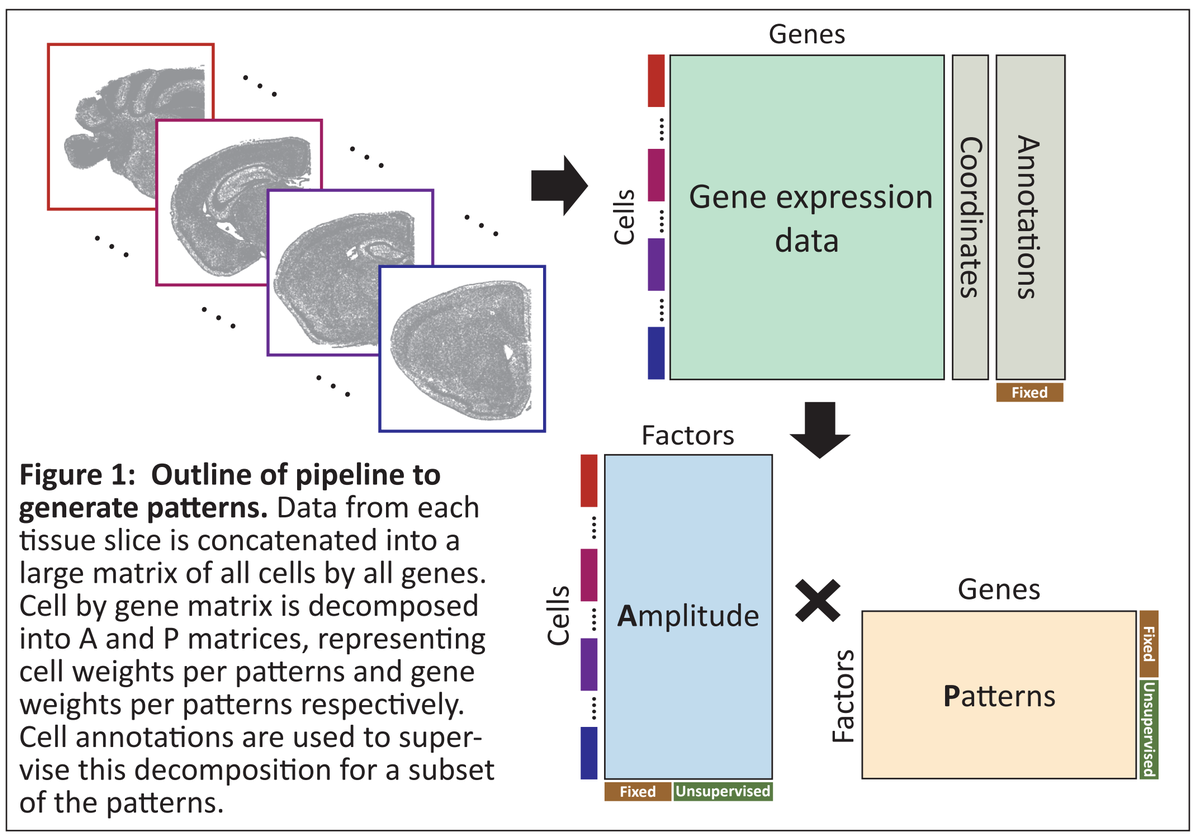


Figure 1: Outline of pipeline to generate patterns. Data from each tissue slice is concatenated into a large matrix of all cells by all genes. Cell by gene matrix is decomposed into A and P matrices, representing cell weights per patterns and gene weights per patterns respectively. Cell annotations are used to supervise this decomposition for a subset of the patterns.

Spatial transcriptomic dataset

- Data obtained from public STARmap PLUS study[4]
- 1,022 genes were profiled for each sample[4]
- Cell annotations were provided for cell type and tissue region[4]
- Selected 12 coronal slices from the same mouse for analysis

Semi-Supervised Non-Negative Matrix Factorization (NMF)

- Per-cell Allen Brain Atlas CCF based annotations were used to fix patterns for high-level anatomical regions
- Fixed 15 patterns and learned an additional 35 unsupervised patterns
- Non-negative least squares used to learn fixed patterns[5]
- Non-negative matrix factorization used for unknown patterns[5]

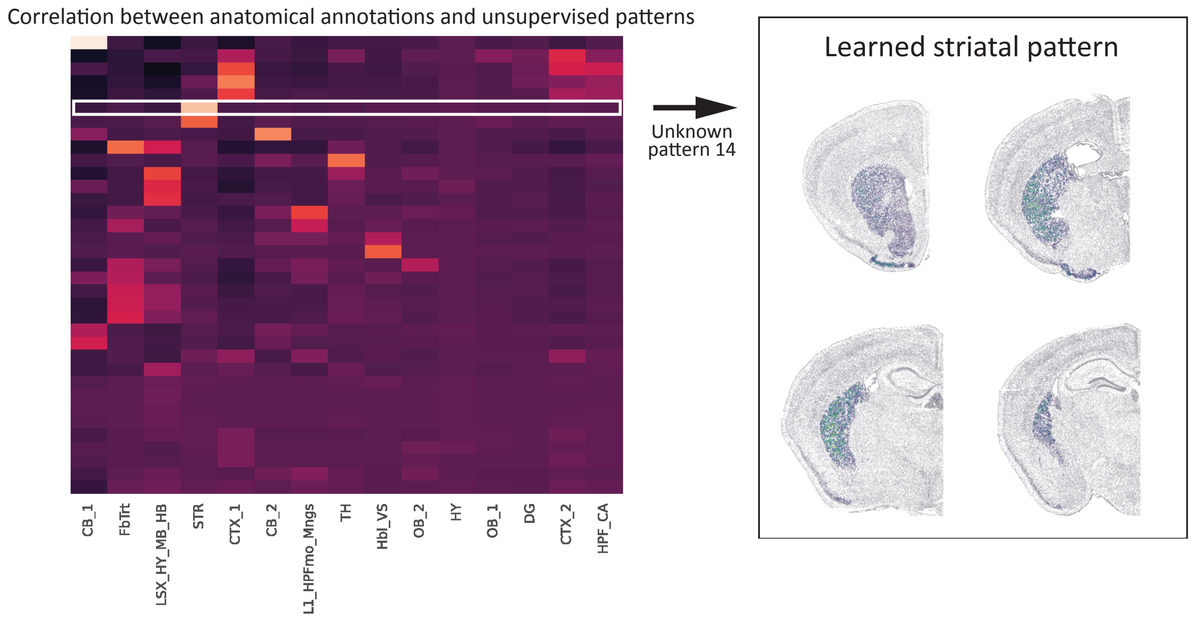
Interpretation of patterns

- Pearson correlation calculated between the Amplitude matrix and the one-hot encoded annotations included in the dataset
- Gene set enrichment analysis using GSEApy

Conclusions

- Semi-supervised NMF is a useful tool for dissecting complex gene expression patterns within the mouse brain
- Incorporating prior knowledge about anatomical regions in a semi-supervised manner allows us to identify biologically relevant patterns of coregulated genes
- Fixed patterns expand our understanding of gene expression in known regions
- Unsupervised patterns learn heterogeneity within the fixed brain regions, the distribution of cell types, and other shared biological processes
- Future work is needed to scale this method for larger spatial transcriptomic datasets

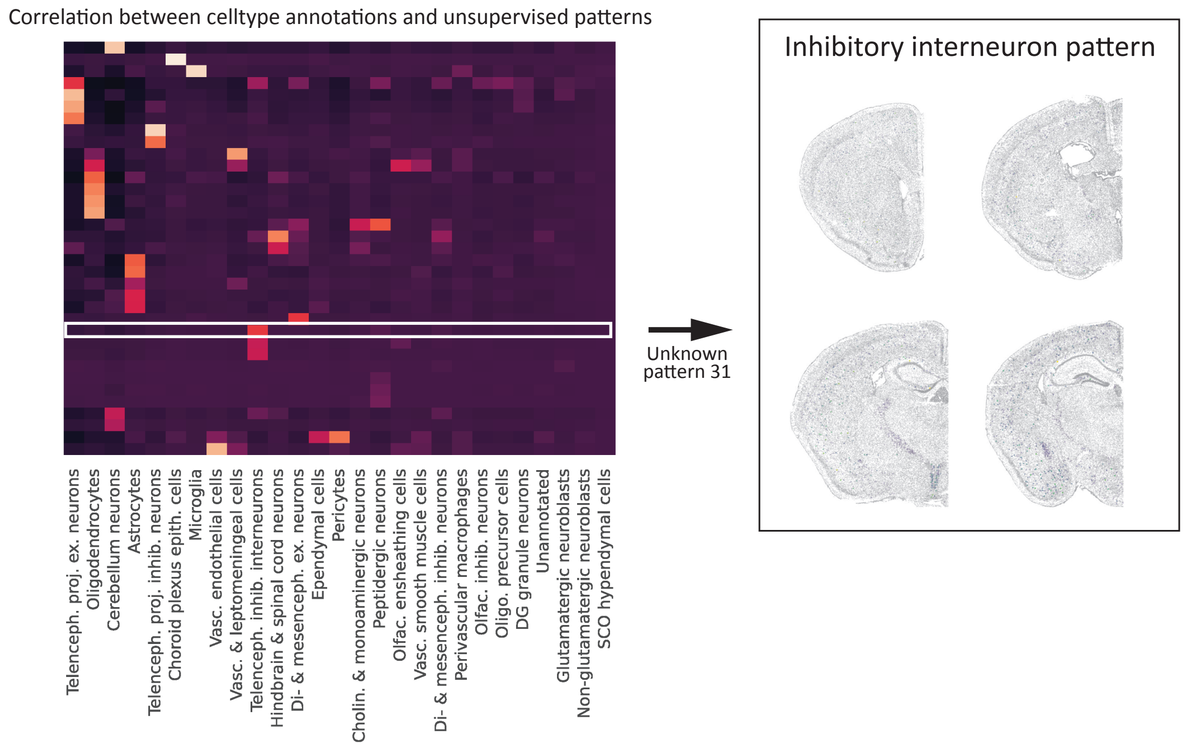
Learned patterns reveal sub-feature heterogeneity



Several unsupervised patterns identify anatomical brain regions

- Correlations between annotated brain regions and pattern weights reveal learned patterns that are localized to specific anatomical structures
- Unknown pattern 14 represents a ***learned striatal pattern*** and captures additional heterogeneity not captured by the fixed striatal pattern
- Top genes for *fixed* striatal pattern:
  ADORA2A, GPR88, DRD2, CPNE5, GPR83, RGS9, SCN4B, DCLK3, PTHLH, CYP26B1
- Top genes for *learned* striatal pattern:
  PPP1R1B, DRD1, SCN4B, RGS9, GPR88, CHRM4, TAC1, CLCA3A1, PDYN, FAM169B

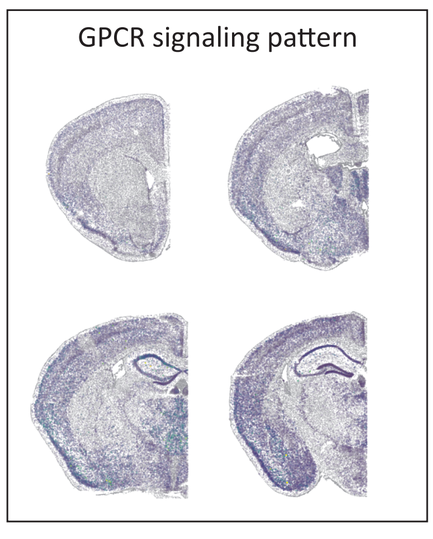
Learned patterns reflect cell type distributions



Several unsupervised patterns reflect cell types within the brain

- Learned cell type patterns identify additional genes correlated with cell types
- Cell type pattern distributions are learned across the brain
- Unknown pattern 31 represents inhibitory interneurons
- The highest weighted gene in this pattern is SST which is used as a marker for inhibitory interneurons[6]
- Top genes for learned inhibitory interneuron pattern:
  SST, BDKRB1, CRHBP, COL19A1, NPY, CCRL2, CEACAM10, RBP4, LHX6, CRCT1

Learned patterns identify biological processes



Unsupervised patterns identify shared biology within the brain

- Many biological processes are reused across multiple cell types and brain regions
- Distribution of these patterns show differential and shared usage across the brain
- Unknown pattern 6 shows enrichment for GPCR signaling (GSEA, pval = 0.004) which plays a key role in neurotransmission
- Top genes for GPCR signaling pattern:
  YJEFN3, COL19A1, HTR2A, KRT12, NPTX2, LGI2, LHX6, CORO6, LTK, RGS12, BMP3, NECAB1, PDGFRA, NPAS1, RIK1, HTR5A, DKKL1, KCNC2, GPR17, CUX2

References

1. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164-4169. doi:10.1073/pnas.0308531101
2. Stein-O'Brien, GL, Arora, R, Culhane, AC, Favorov, AV, Garmire, LX, Greene, CS, Goff, LA, Li, Y, Ngom, A, Ochs, MF, Xu, Y, & Fertig, EJ. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends in genetics : TIG, 34(10), 790–805 (2018). https://doi.org/10.1016/j.tig.2018.07.0033.
3. Stein-O'Brien GL, Clark BS, Sherman T, et al. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species [published correction appears in Cell Syst. 2021 Feb 17;12(2):203]. Cell Syst. 2019;8(5):395-411.e8. doi:10.1016/j.cels.2019.04.004
4. Shi H, He Y, Zhou Y, et al. Spatial atlas of the mouse central nervous system at molecular resolution. Nature. 2023;622(7983):552-561. doi:10.1038/s41586-023-06569-5
5. https://github.com/nloyfer/ssNMF/blob/main/README.md
6. Song YH, Yoon J, Lee SH. The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. Exp Mol Med. 2021;53(3):328-338. doi:10.1038/s12276-021-00580-4