Bioinformatics Program

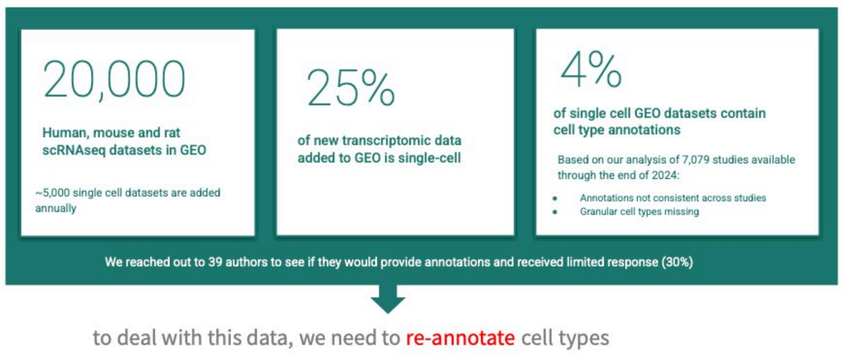
Automated Cell Type Re-Annotation for Enhanced Reproducibility and Meta-analysis in Single-Cell Brain Studies

Rachel Schwartz; Paul Pavlidis

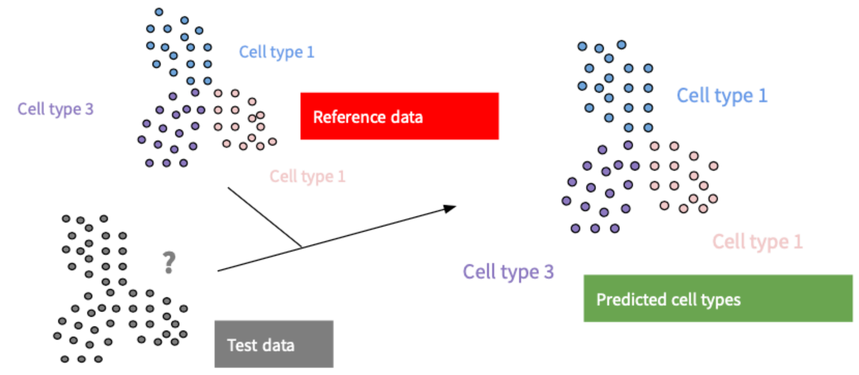
Introduction

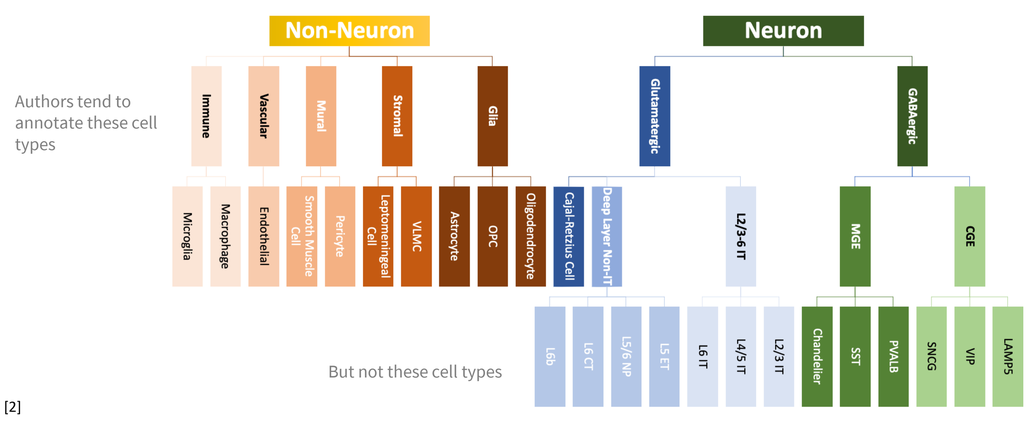
•**Challenge:** Inconsistency and lack of cell-type annotations hinder meta-analysis of single-cell data and limit reproducibility across studies.
•**Approach:** We a pipeline for automated annotation of public single-cell transcriptomes, focusing on human and mouse neocortex under a unified cell type taxonomy.
• **Methods:** We benchmarked several approaches to automated re-annotation, focusing on the impact of method, reference dataset, confidence thresholding, and experimental factors.
•**Impact:** Re-annotation enables robust identification of overlooked cell types, exposing them to meta-analysis in the context of brain disease.

Lack of consistent cell-type annotations challenges meta-analysis of single-cell data



Leveraging BICCN cell type taxonomy for re-annotation of human and mouse neocortex



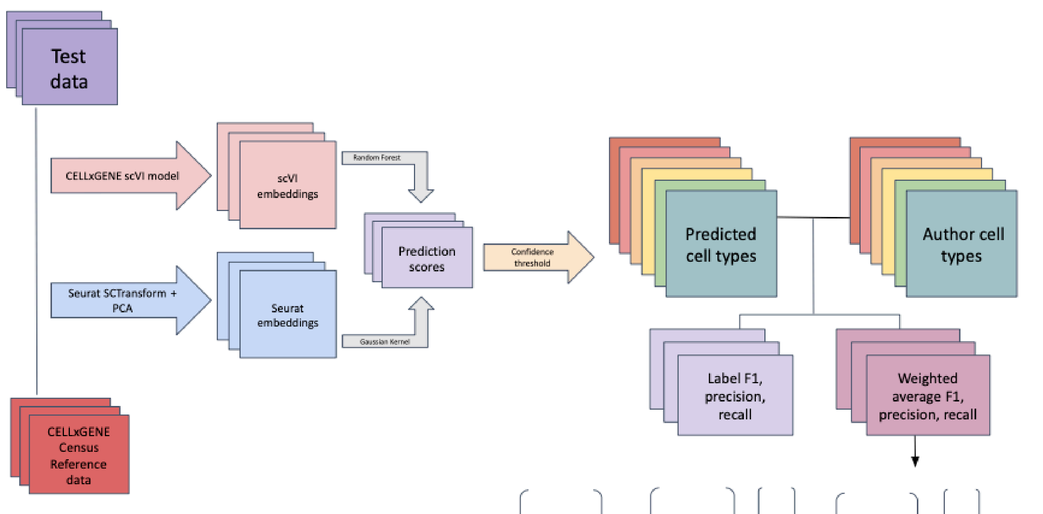


Authors tend to annotate these cell types

But not these cell types

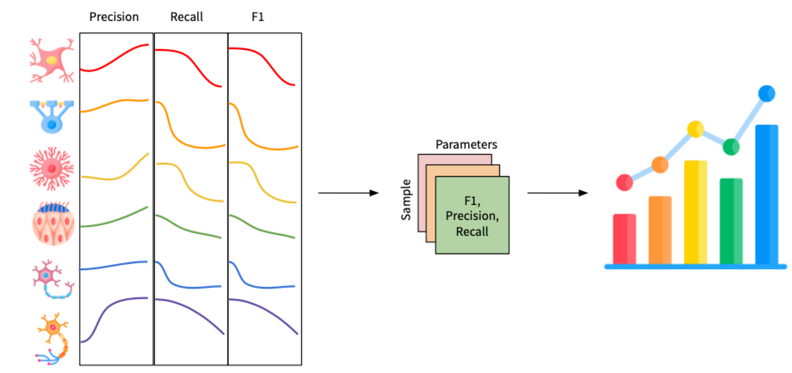
[2]

Benchmarking of label transfer methods from BICCN references in diverse public datasets

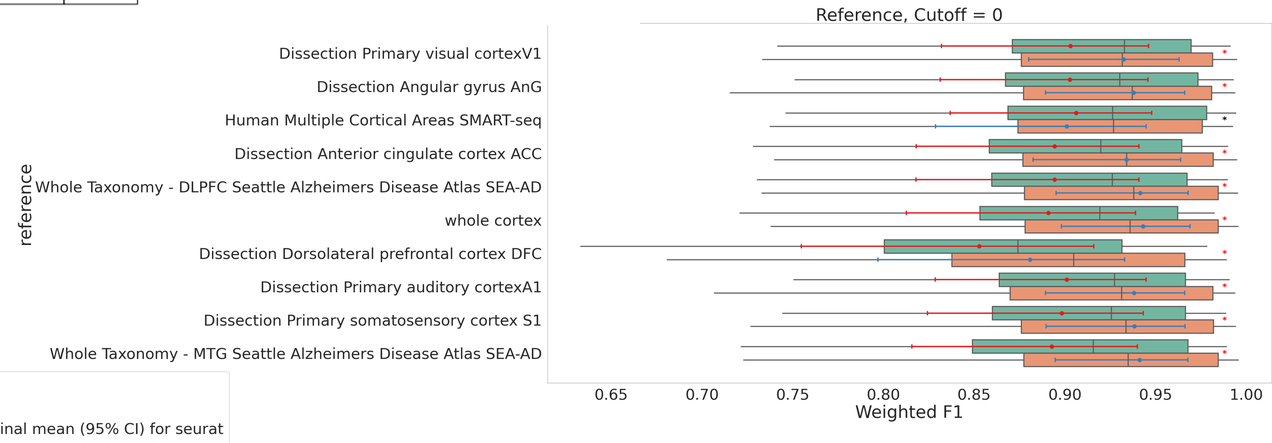


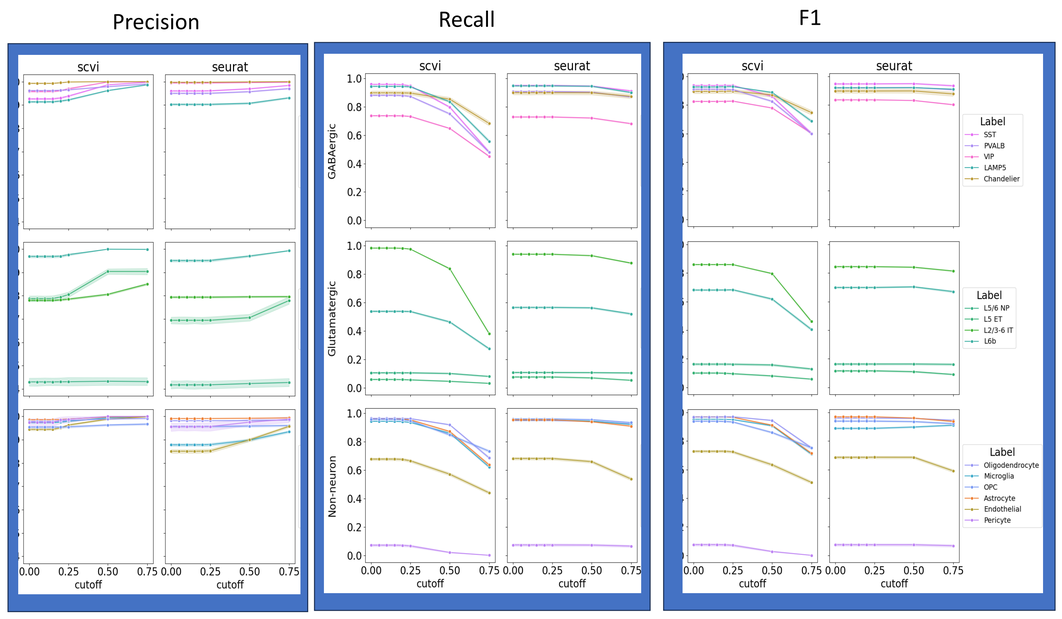
[3][5][6]

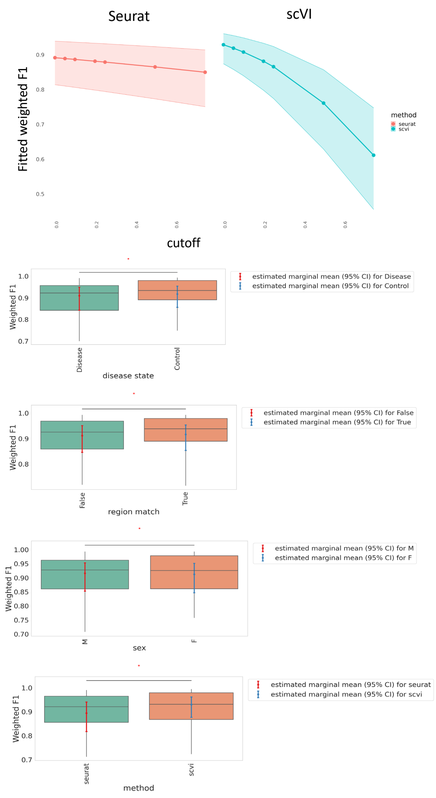
Evaluation of label transfer across 15 cell types in 198 human neocortex samples



factor	contrast	OR	SE	Tukey's p-value
method	seurat / scvi	0.63	0.004	<0.001
sex	F / M	0.95	0.004	<0.001
region match	False / True	0.94	0.008	<0.001
disease state	Control / Disease	1.11	0.005	<0.001

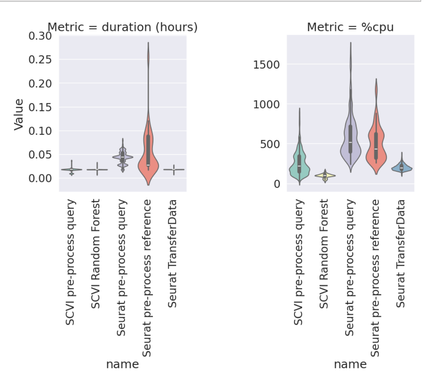




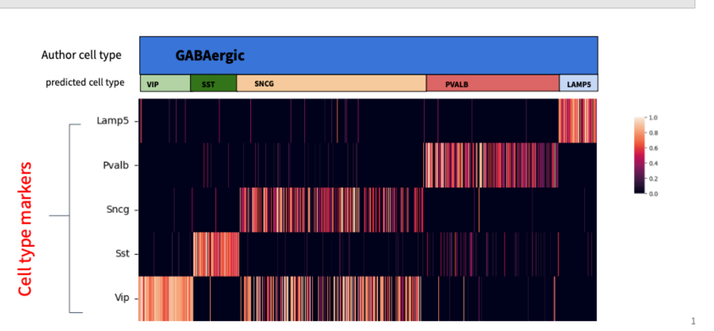


Precision, recall, and F1 scores for 15 cell subclasses across 8 confidence thresholds (cutoff) were generated for 7 studies comprising 198 individual samples. Weighted average F1 scores for each sample were fit to a beta-logit generalized linear model treating study as a random effect. Marginal means for parameters (reference, method, disease state, and region matching) were estimated from the model and compared with Tukey's post-hoc tests.

scVI is faster and more efficient than Seurat



Re-annotation exposes granular cell types



Re-annotation of GABAergic cell types in [1].

Conclusions

- Using an scVI-trained classifier without confidence thresholding achieves highest overall performance for human data; Seurat classifier is more robust to thresholds
- Cortical region, sex, disease state have little overall effect on performance when accounting for the effects of individual studies
- We implemented a Nextflow pipeline informed by benchmarking results to re-annotate experiments using an aggregate "whole cortex" reference [4]
- So far, we have re-annotated 79 scRNAseq experiments in human and mouse with harmonized cell types in the Pavlidis lab's Gemma database [7]
- Future work will expose harmonized cell types to meta-analysis across human and mouse studies of neurodegenerative and psychiatric disease

References
1. Bhattacherjee, A., Djekidel, M. N., Chen, R., Chen, W., Tuesta, L. M., & Zhang, Y. (2019). Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nature Communications*, *10*(1), 4169. https://doi.org/10.1038/s41467-019-12054-3
2. BRAIN Initiative Cell Census Network (BICCN), BRAIN Initiative Cell Census Network (BICCN) Corresponding authors, Callaway, E. M., Dong, H.-W., Ecker, J. R., Hawrylycz, M. J., Huang, Z. J., Lein, E. S., Ngai, J., Osten, P., Ren, B., Tolias, A. S., White, O., Zeng, H., Zhuang, X., BICCN contributing principal investigators, Ascoli, G. A., Behrens, M. M., Chun, J., … Sunkin, S. (2021). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature, 598(7879), 86–102. https://doi.org/10.1038/s41586-021-03950-0
3. CZI Single-Cell Biology Program, Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Michael Cherry, J., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., … Carr, A. (2023). *CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data*. https://doi.org/10.1101/2023.10.30.563174
4. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820
5. Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, *42*(2), 293–304. https://doi.org/10.1038/s41587-023-01767-y
6. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2
7. Lim, N., Tesar, S., Belmadani, M., Poirier-Morency, G., Mancarci, B. O., Sicherman, J., Jacobson, M., Leong, J., Tan, P., & Pavlidis, P. (2021). Curation of over 10 000 transcriptomic studies to enable data reuse. *Database*, *2021*, baab006. https://doi.org/10.1093/database/baab006

THE UNIVERSITY OF BRITISH COLUMBIA
Michael Smith Laboratories
UBC DEPARTMENT of Psychiatry
INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
NSERC CRSNG
NIMH National Institute of Mental Health
Graduate Program in BIOINFORMATICS