PROOF
Centre of | Centre d'
EXCELLENCE

Biomarqueurs – Solutions en soins de santé.

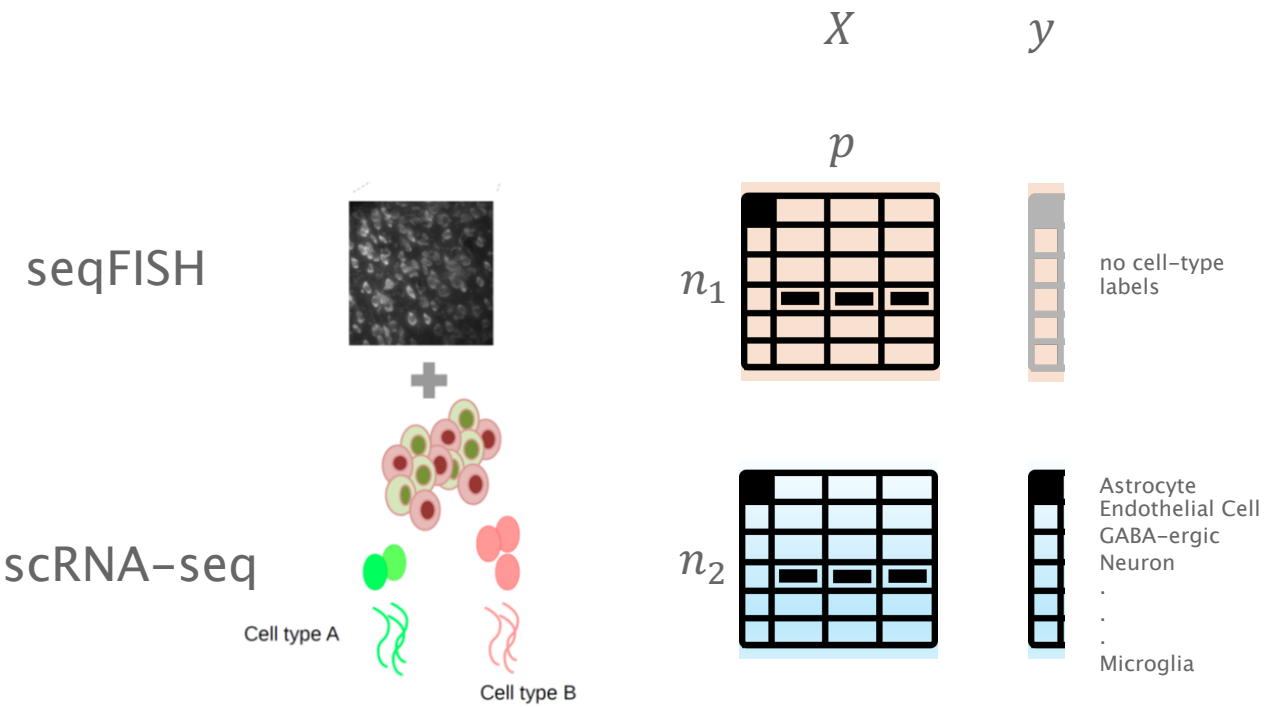
Biomarker solutions for health care.

CORTEX seqFISH: integration with scRNA-seq data by self-training an elastic net classifier

Amrit Singh, PhD
Postdoc – University of British Columbia
https://amritsingh.ca

Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Types (Online)
June 15, 2020

CORTEX seqFISH: integration with scRNA-seq data



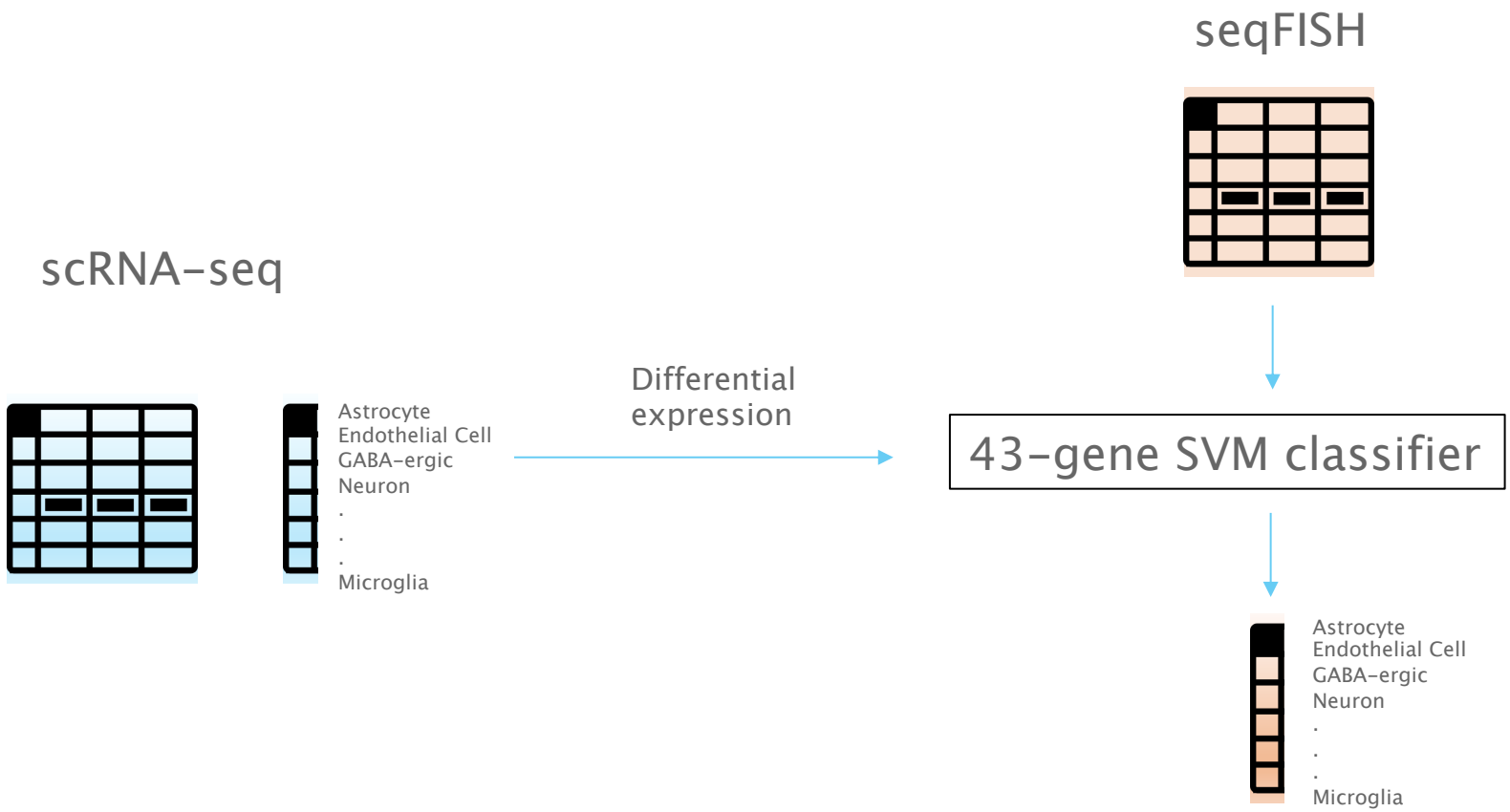
1) Can scRNA-seq data be overlaid onto seqFISH for resolution enhancement?

2) What is the minimal number of genes needed for data integration?

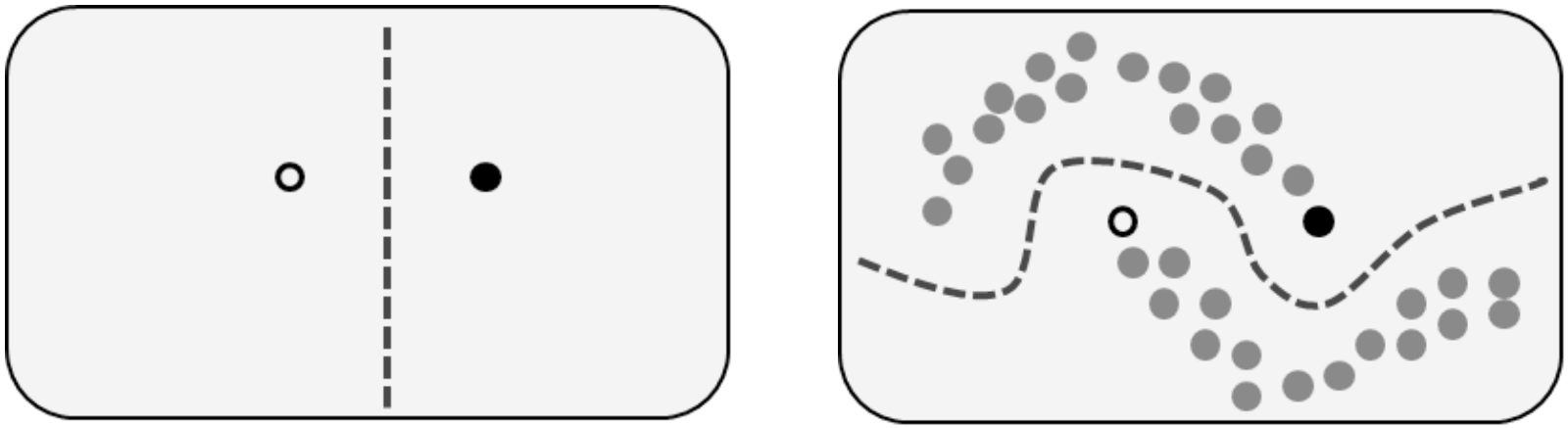
Purpose: Identify a gene signature predictive of cell-types in the mouse visual cortex by integrating seqFISH+scRNASeq data.

Figure modified from Zhu *et al. Nature Biotechnology* 36, 1183–1190 (2018)

How did Zhu *et al.* do it?



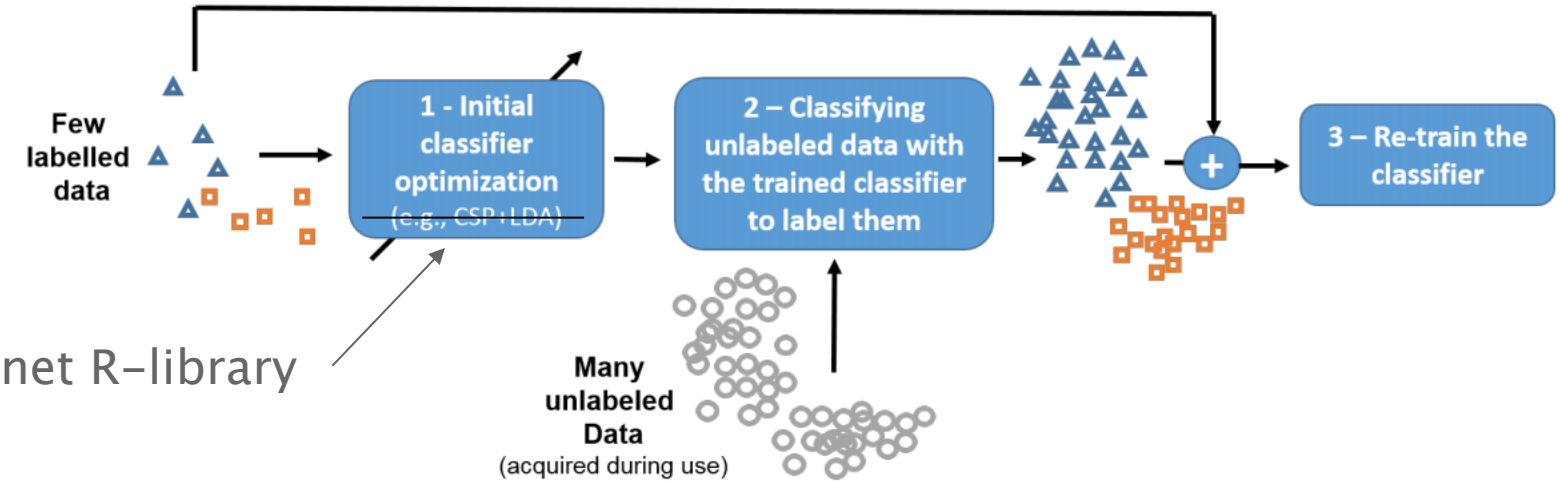
Semi-supervised learning



By Techerin - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=19514958

Self-training

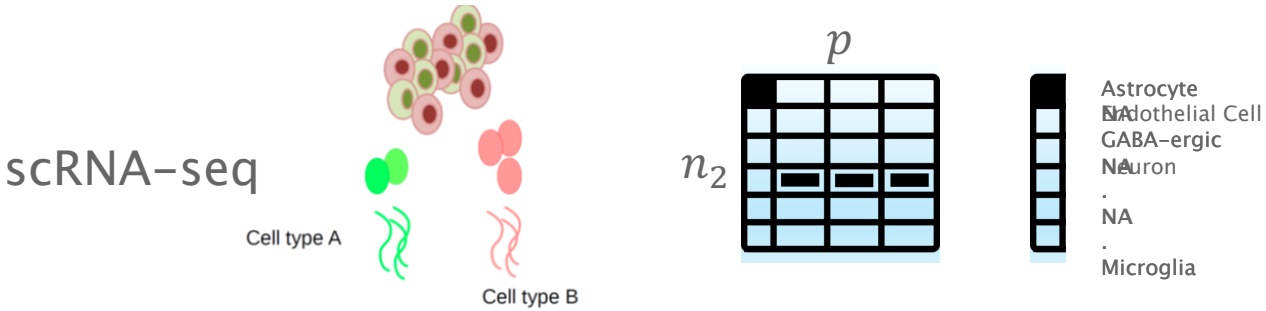
ssc R-library



glmnet R-library

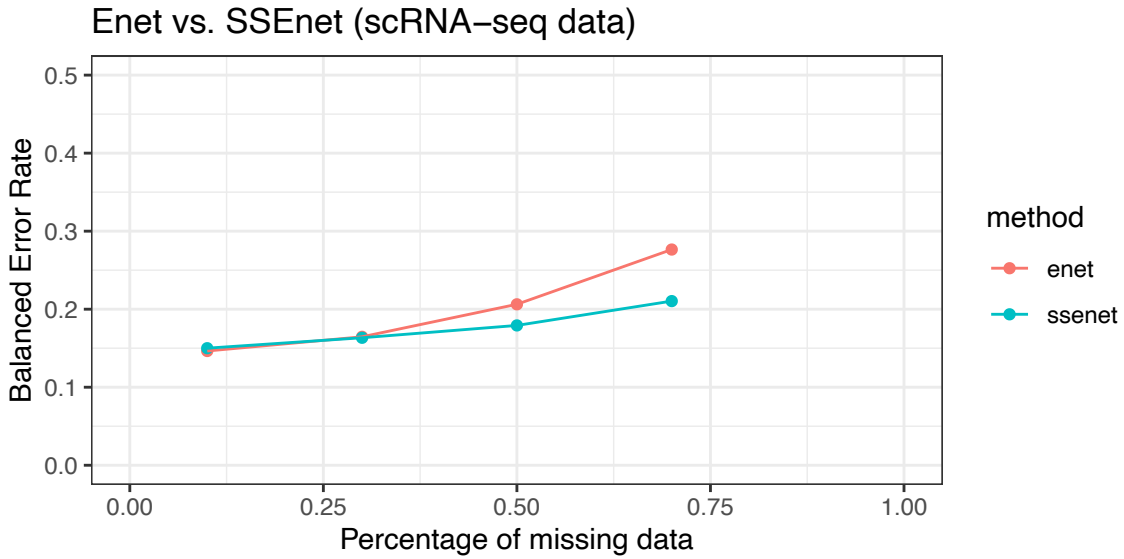
Fabien Lotte. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based Brain-Computer Interfaces. Proceedings of the IEEE, Institute of Electrical and Electronics Engineers (IEEE), 2015, 103 (6), pp.871-890.

Does self-training actually work?

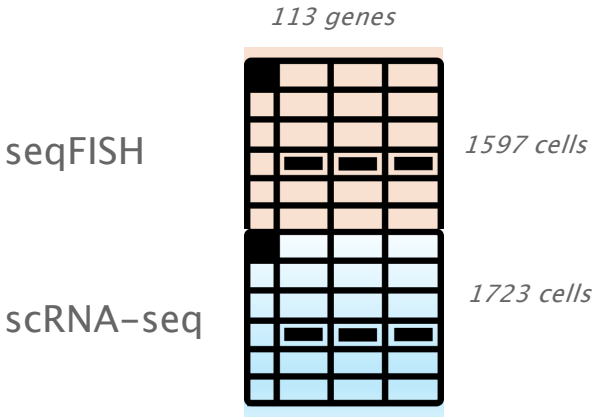


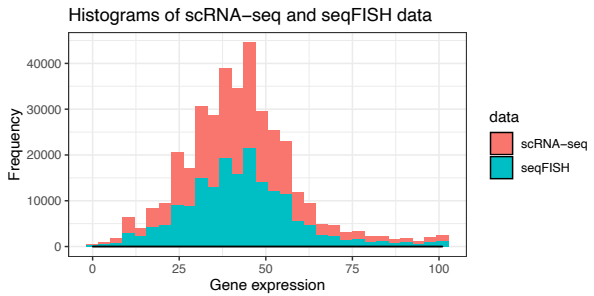
Astrocyte	Endothelial Cell	GABA-ergic Neuron	Glutamatergic Neuron	Microglia	Oligodendrocyte.1	Oligodendrocyte.2	Oligodendocyte.3
43	29	761	812	22	19	6	31

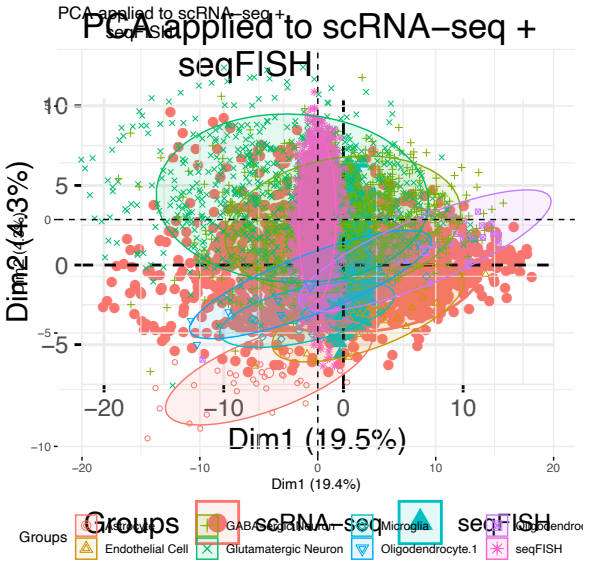
Removed: too few samples



Data normalization







"…bias-corrected, quantile-normalized seqFISH data to assign cell types…"

ComBat

(sva R-library): treat this issue as a batch correction problem

MFA: Multiple Factor Analysis

(FactoMineR R-library): treat this issue as a scaling problem:

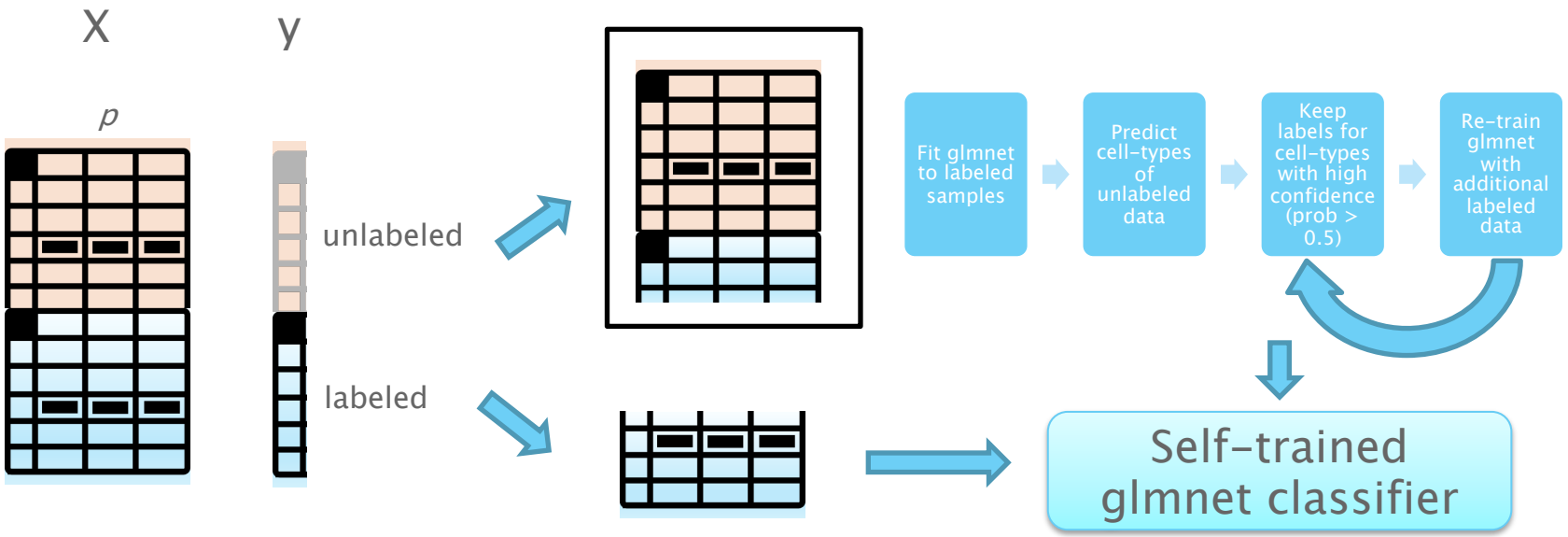
1. For each dataset apply standardization (center+scale) of each dataset.
2. Further divide all elements by the square root of the first eigenvalue (from applying PCA to the dataset).

SGCCA/DIABLO

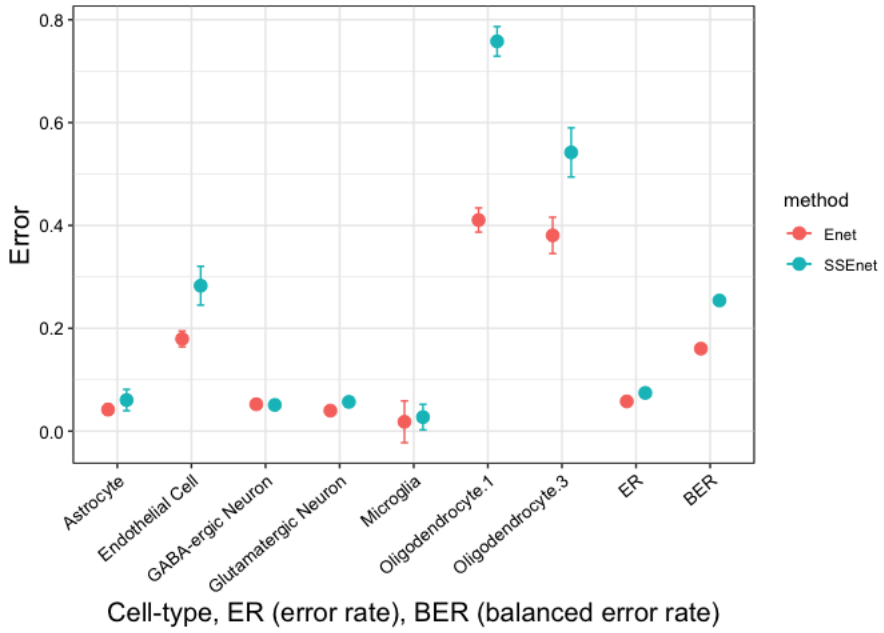
(RGCCA/mixOmics R-libraries): treat difference in the number of variables problem

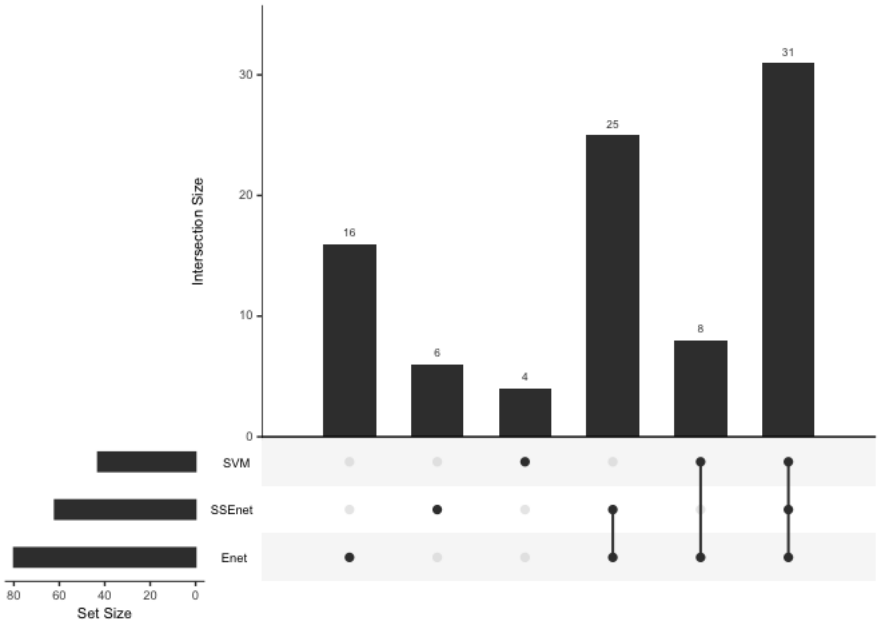
1. For each dataset apply standardization (center+scale) of each dataset.
2. Further divide by the square root of the number of variables in that dataset.

Estimate seqFISH cell-type labels using a semi-supervised elastic net classifier (ssenet)



ssenet() applied to scRNA-seq+seqFISH vs. enet() applied to scRNA-seq only





Summary

- ssenet improves classification performance by self-training on unlabeled data
- Although the performance of ssenet < enet, it may generalize better to seqFISH data

Limitations of present study:

- Data distributions of labeled (scRNA-seq) and unlabeled data (seqFISH) are different and should be mitigated using:
  - strategies to normalize between datasets: ComBat, eigenvalue, # of variables

Future directions:

- Use observational weights for imbalance class sizes
- Try different data normalization strategies

Word of caution of current implement of ssenet::ssenet()

- Don't use "singha53/ssenet" with R4.0 or glmnet4.0 at the moment

Supervisors

- Dr. Bruce M McManus
- Dr. Kim-Anh Lê Cao
- Dr. Scott Tebbutt

MICHAEL SMITH FOUNDATION FOR HEALTH RESEARCH
Discover. Connect. Engage.

MITACS

UBC