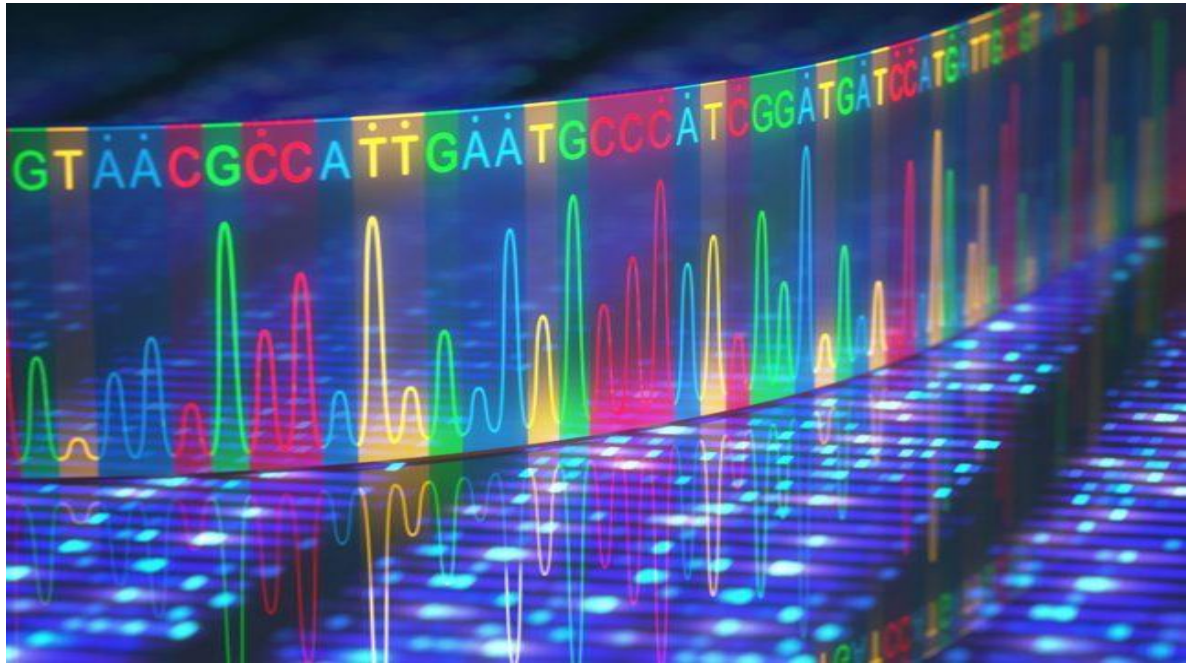
Benchmarking Cell Typing Methods in Spatial Transcriptomics

- Howard Baek
- Eliza Chai
- Alexis Harris
- Ingrid Luo
- Makayla Tang



This Photo by Unknown author is licensed under CC BY-NC-ND.

UNIVERSITY of WASHINGTON

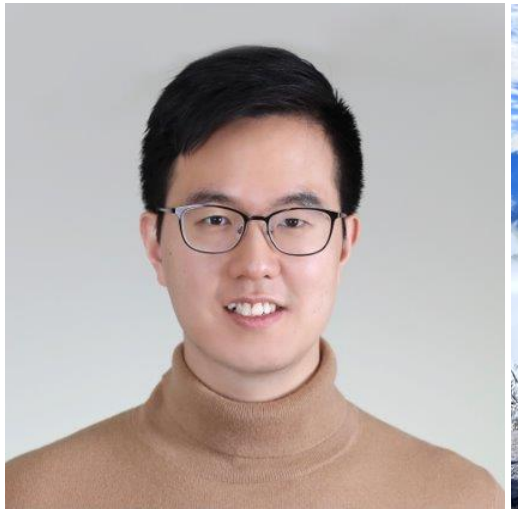
BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH

Meet the Team

Alexis Harris

Eliza Chai

Ingrid Luo



Howard Baek

Makayla Tang

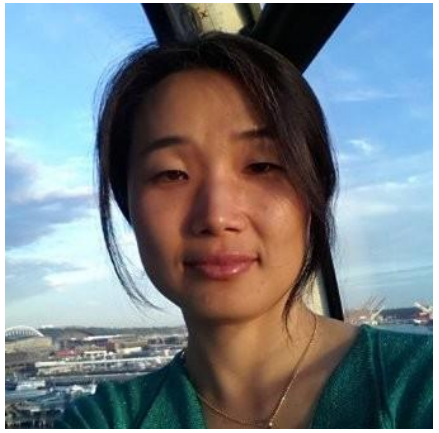
UNIVERSITY of WASHINGTON

Acknowledgements

Patrick Danaher

Edward Zhao

Ronalyn Vitancol



Sangsoon Woo

Lloyd Mancl

Patrick Heagerty

UNIVERSITY *of* WASHINGTON

Spatial Transcriptomics and Cell Typing

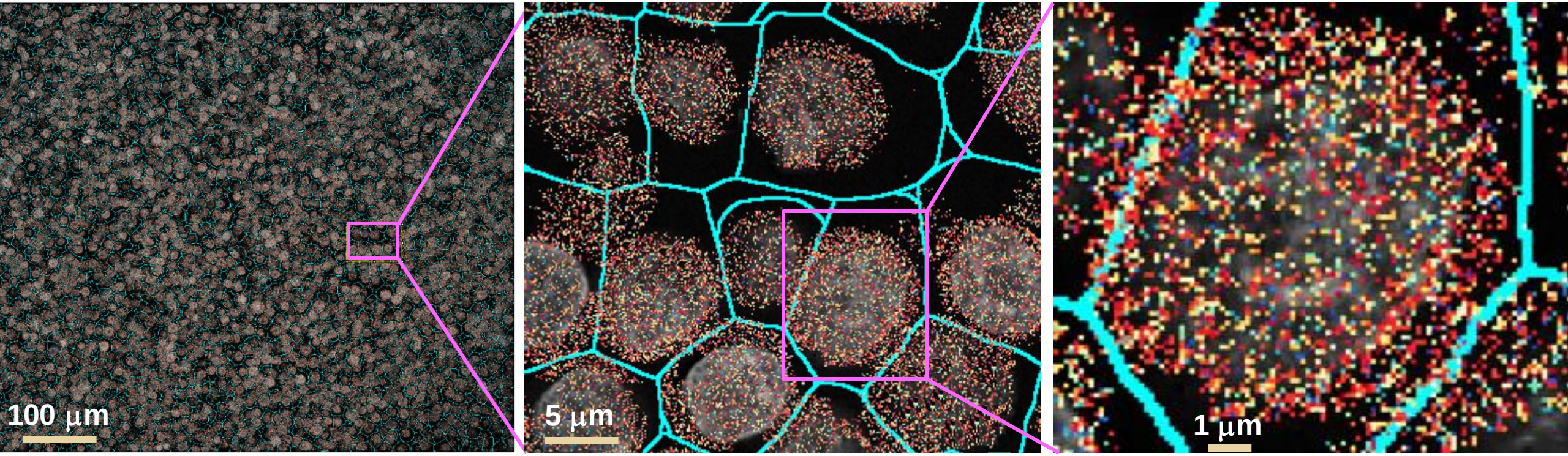
Spatial transcriptomics technologies measure single cell gene expression while recording cells' positions within a tissue

- detect individual RNA molecules where they lie on tissue slides
- produce datasets of single-cell gene expression data while preserving cells' location data

Cell typing is a classification used to identify cells that share features

- supervised and unsupervised methods

Single Cell Spatial Imager Localizes RNA at Subcellular Scale



Thousands of RNA transcripts
(represented as different colored single pixel in the image)
detected per single cell

UNIVERSITY of WASHINGTON

InSituType Cell Typing

- *in situ* - "in the original place"
- Likelihood model for spatial transcriptomics data
- Supervised cell typing from reference datasets via a Bayes classifier
- Unsupervised and semi-supervised cell typing via an Expectation Maximization (EM) algorithm

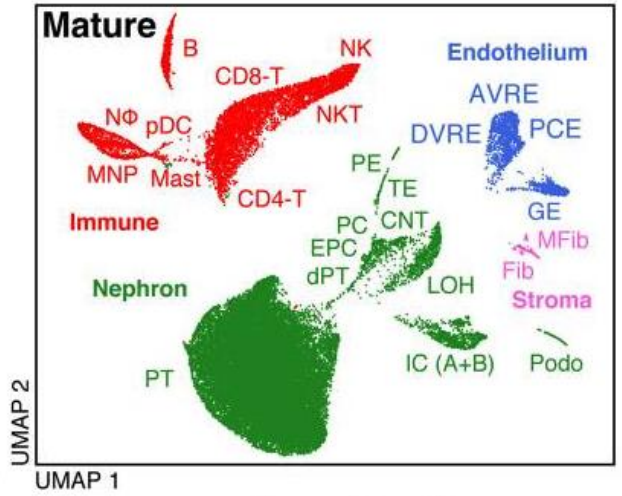
Dataset - Supervised

Training set: Kidney Single Cell Atlas.

- scRNA study that makes an analysis of all the cell types
- 40,268 mature kidney cells, consisting of 33 cell types

Test set: Kidney core biopsy from a lupus nephritis patient profiled using the CosMx™ Spatial Molecular Imager

- scRNA dataset with 61,073 mature kidney cells
- No cell type data
- Have a priori about whether the cells are located within the glomerulus or not



Dataset - Unsupervised

- Unsupervised method can train on any tissue
- Cell pellet array (CPA) profiled using the CosMx™ Spatial Molecular Imager
- Dataset of 52,518 cells
- 13 known clusters
- Benchmarking is performed on the full dataset and two downsampled datasets containing a random sample of 50% of the genes and the 10% most highly variable genes

Objectives

Objective 1 – Supervised Cell Typing

Benchmark **InSituType supervised** against established cell typing methods: SingleR, CHETAH, SeuratV3, SingleCellNet, scPred and SVM

- Evaluate the accuracy of the results using known kidney biology
- Compare running time

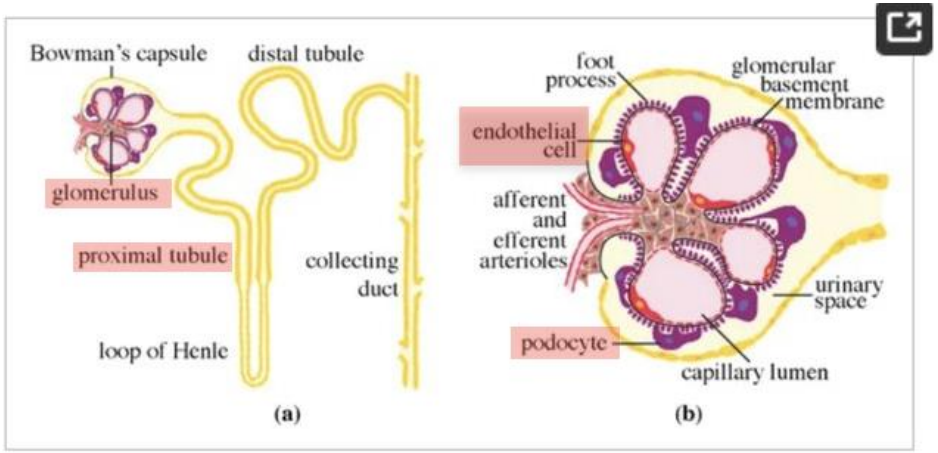
Objective 2 – Unsupervised Cell Typing

Optimize the performance of **InSituType unsupervised** method by tuning parameters

Methods

Supervised Cell Typing

- Score cell typing results based on the distribution of distinct cell types in substructures called "glomeruli"
- Record the intensity of specific marker proteins (PanCK and CD45) that are highly informative of cell types (cytokeratin and immune cells)



Unsupervised Cell Typing

Modify tuning parameters (# of clusters, iterations, Subsample size, number of cohorts)

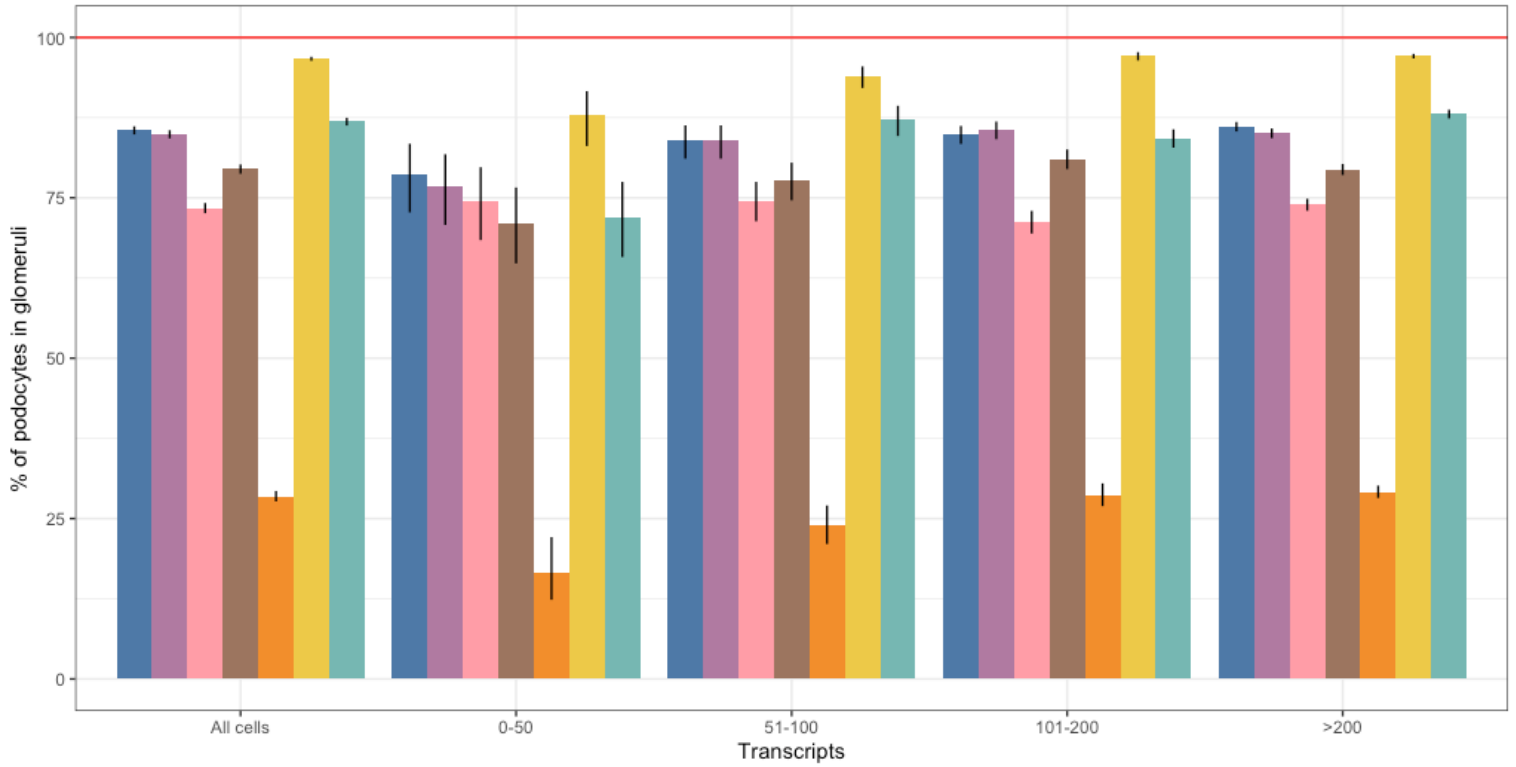
Metrics: Bayesian information criteria (BIC), Adjusted rand index (ARI), Computation time

Supervised

Results – Accuracy of glomerular cell assignments

All genes

Note that scPred has 98% unassigned cells;
CHETAH has 50% unassigned cells



Best: SVM

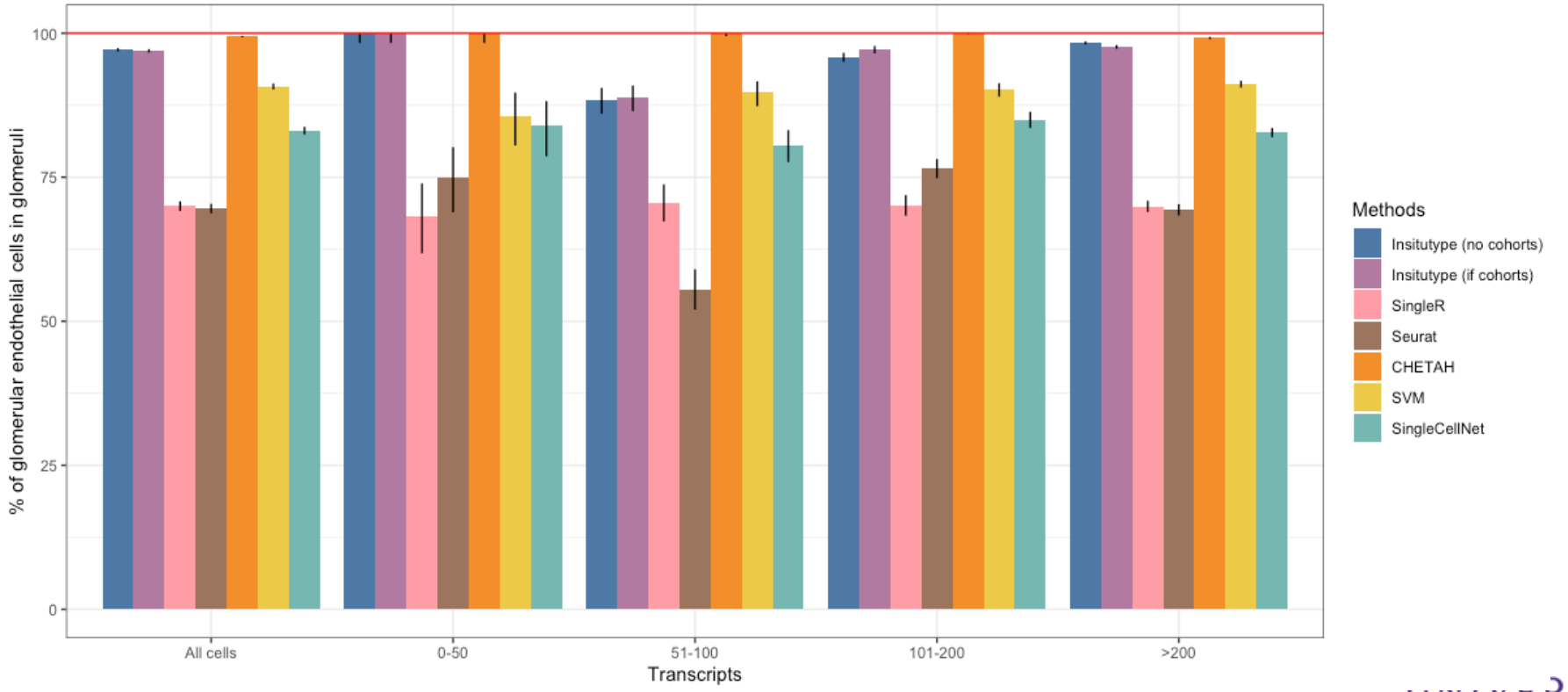
Worst: CHETAH

UNIVERSITY of WASHINGTON

Supervised

Results – Accuracy of glomerular cell assignments

All genes



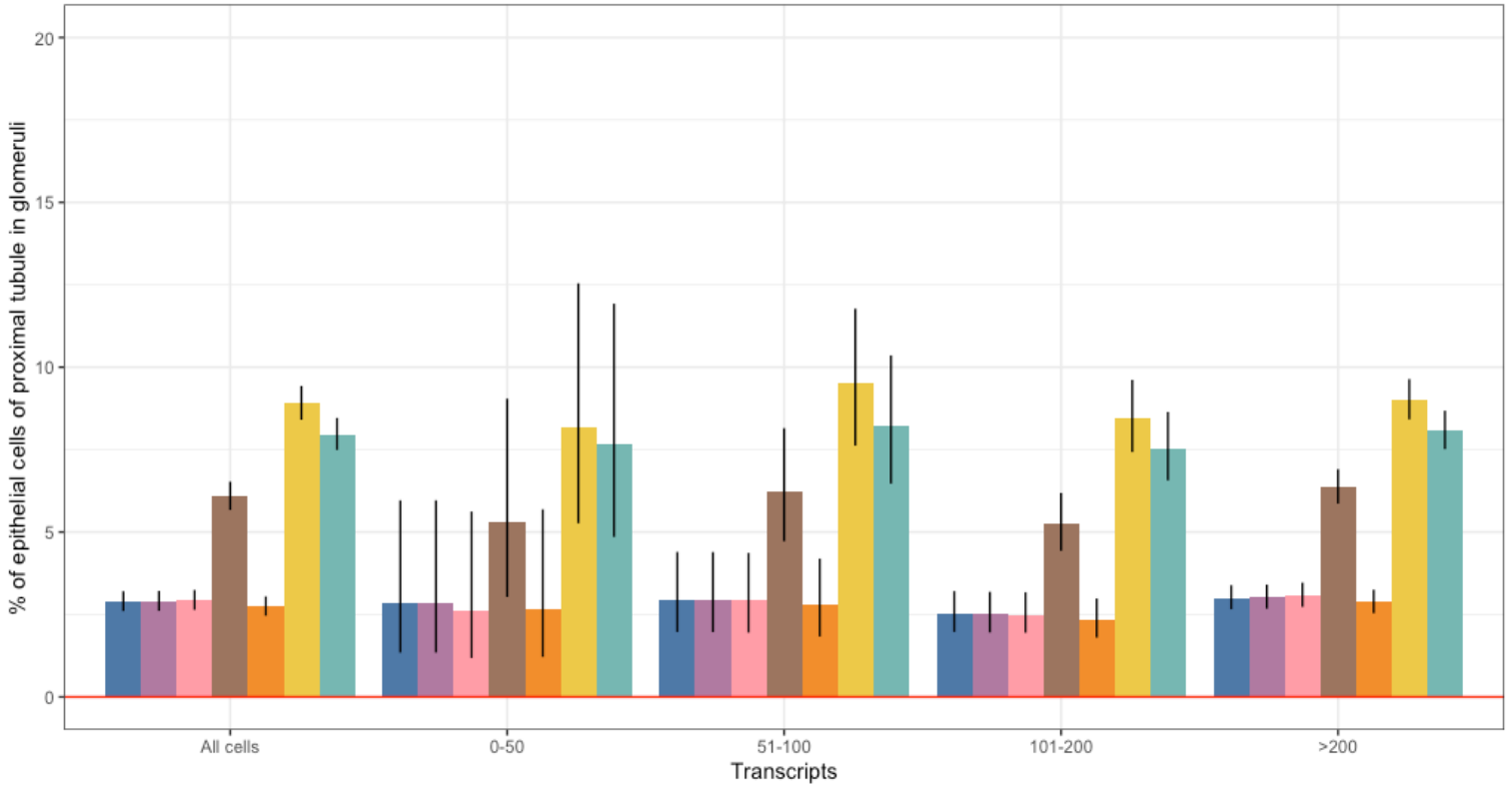
Best: CHETAH

Worst: Seurat

Supervised

Results – Accuracy of glomerular cell assignments

All genes



Best: /

Worst:

SVM

SingleCellNet

Seurat

UNIVERSITY of WASHINGTON

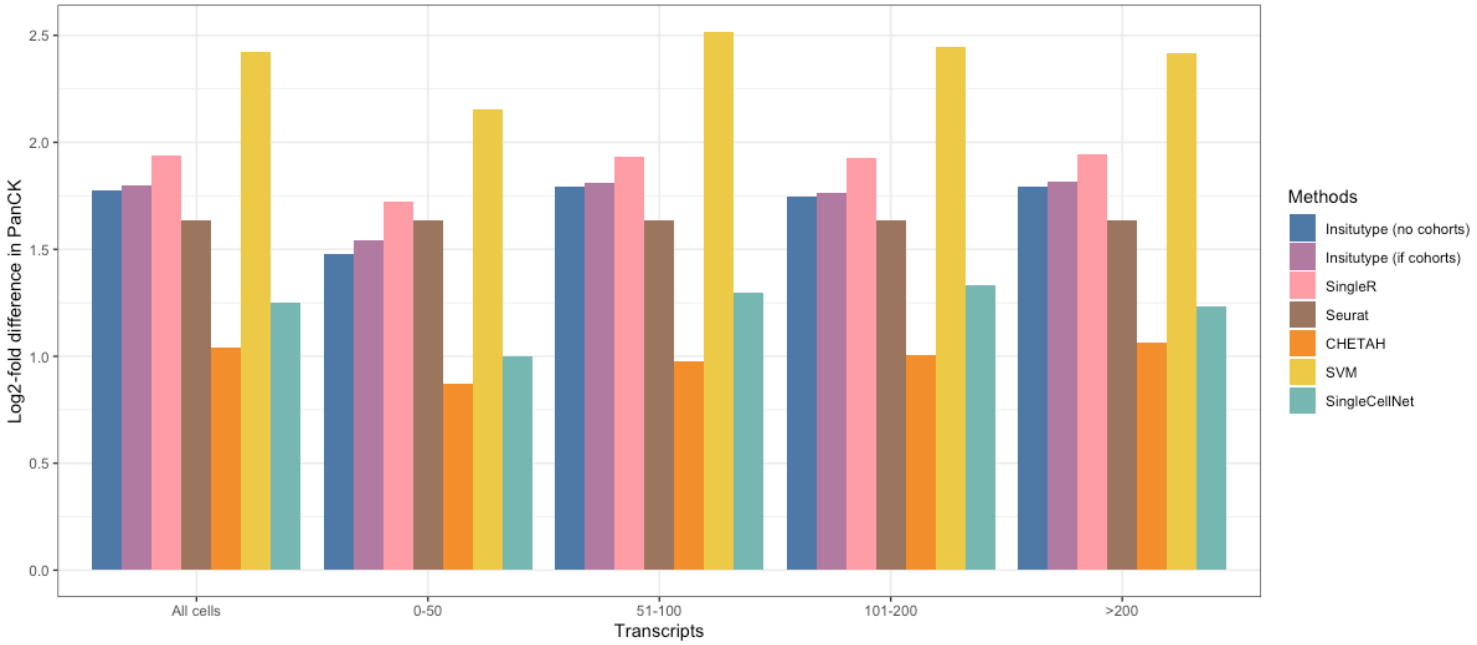
Supervised

Results – Difference in PanCK intensity

- Cytokeratin+ cells show strong PanCK staining; Cytokeratin- cells do not show PanCK staining
- Significant difference in PanCK intensity between cytokeratin+ and cytokeratin-cells

  -> cell typing is accurate

All genes



Best: SVM

Worst: CHETAH

Cytokeratin + cell types:

epithelial cell of proximal tubule,
kidney connecting tubule **epithelial** cell,
kidney loop of Henle thick ascending limb **epithelial** cell,
kidney **epithelial** cell,
renal **alpha-intercalated** cell,
renal **beta-intercalated** cell,
renal **intercalated** cell, renal principal cell

UNIVERSITY of WASHINGTON

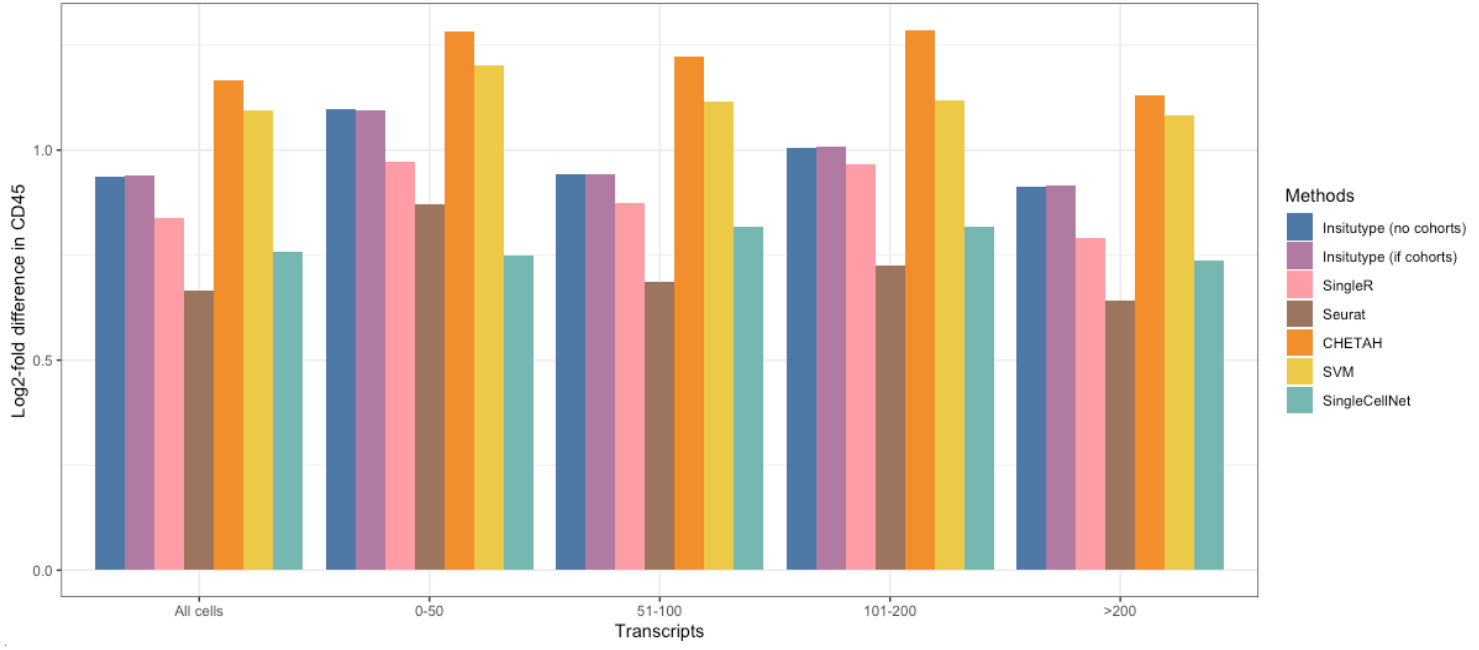
Supervised

Results – Difference in CD45 intensity

- Immune cells express high levels of CD45; Non-immune cells express low levels of CD45
- Significant difference in CD45 intensity between immune and non-immune cells

  -> cell typing is accurate

All genes



Best: **CHETAH**

Worst: **Seurat**

Immune cell types:

B cell, CD4-positive alpha-beta T cell, CD8-positive alpha-beta T cell, classical monocyte, dendritic cell, kidney resident macrophage, mature NK T cell, natural killer cell, neutrophil, non-classical monocyte, plasmacytoid dendritic cell

UNIVERSITY of WASHINGTON

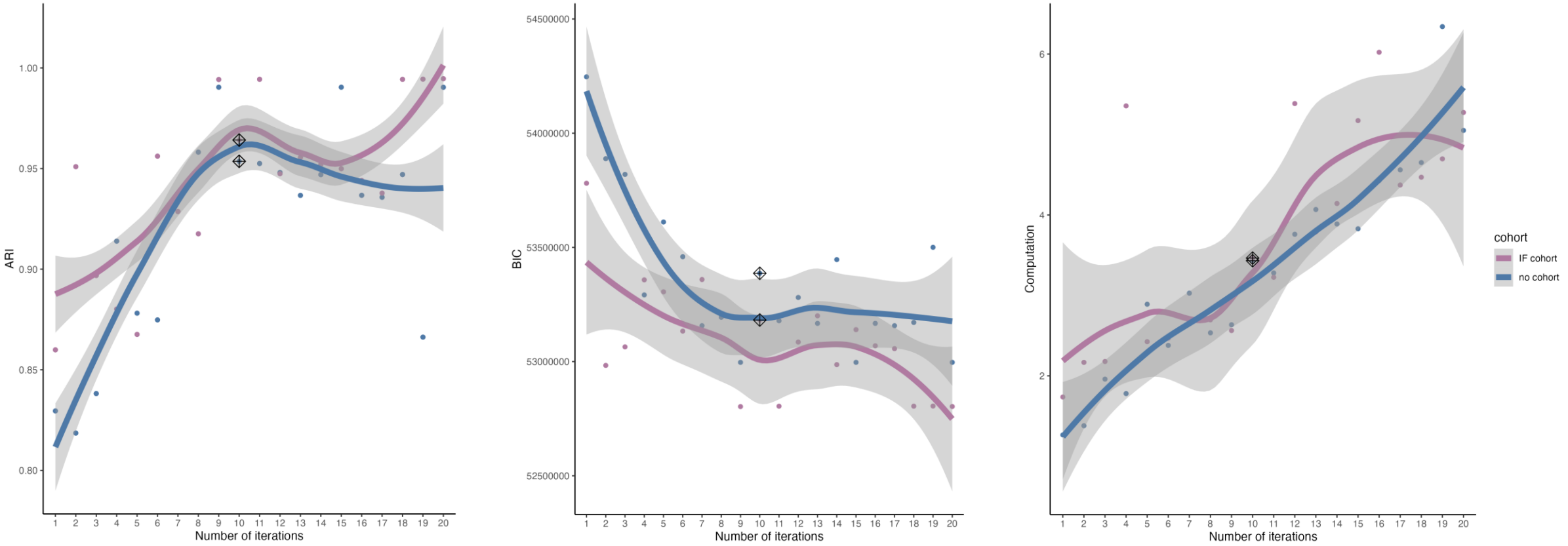
Results – Summary

- No method that outperforms all others in every aspect
- **InSituType** performs relatively well
- Consistent accuracy rates across cells with different transcript numbers
- Inconsistent accuracy rates across all genes, half genes, and 10% highly variable genes (not displayed in the slides, but will be included in the final report)
- Running time: SingleR and InSituType are the fastest (<1 min), SingleCellNet is the slowest (~1.5 hrs)

Unsupervised

Results – Number of iterations

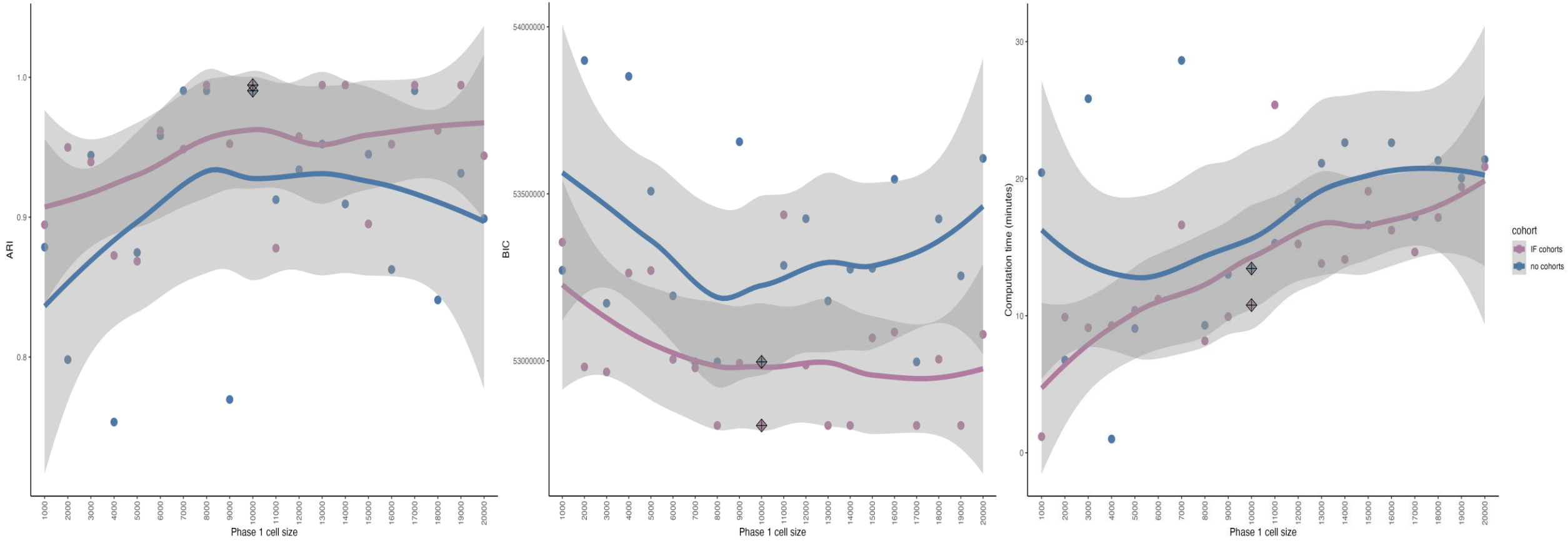
All genes



Unsupervised

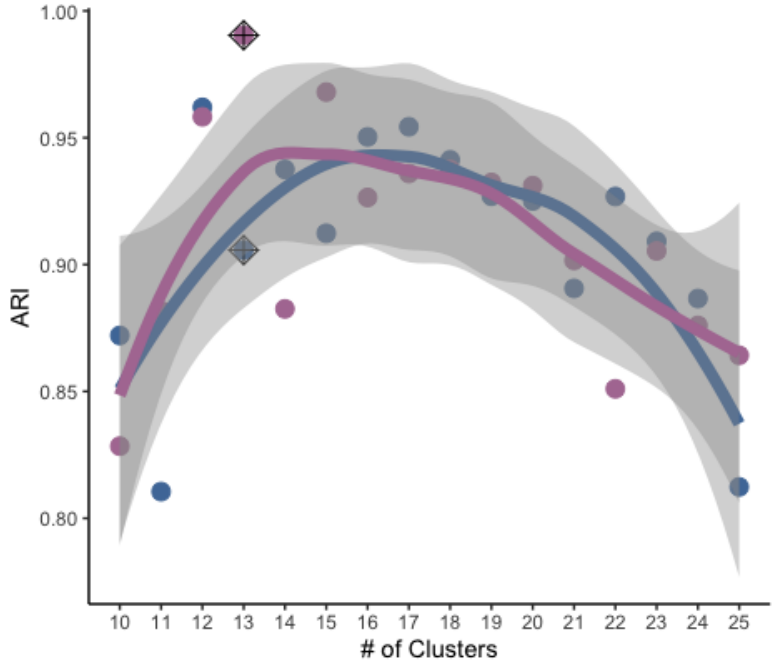
Results – Number of cells in phase 1

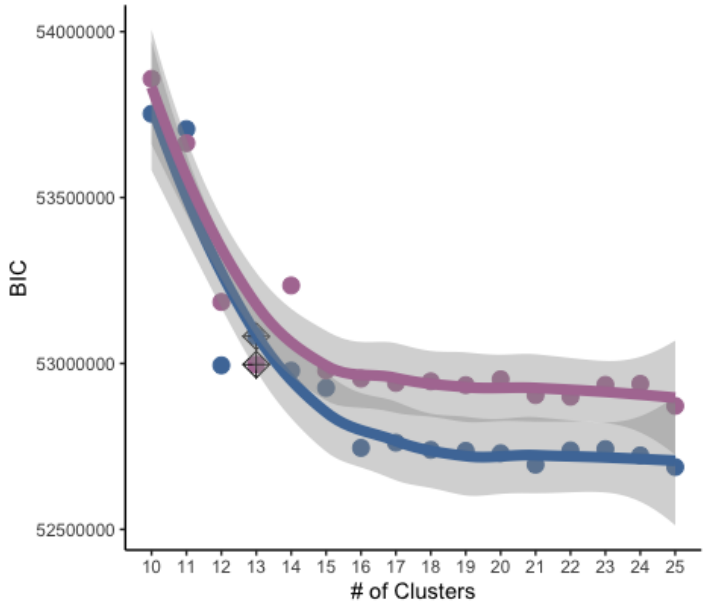
All genes

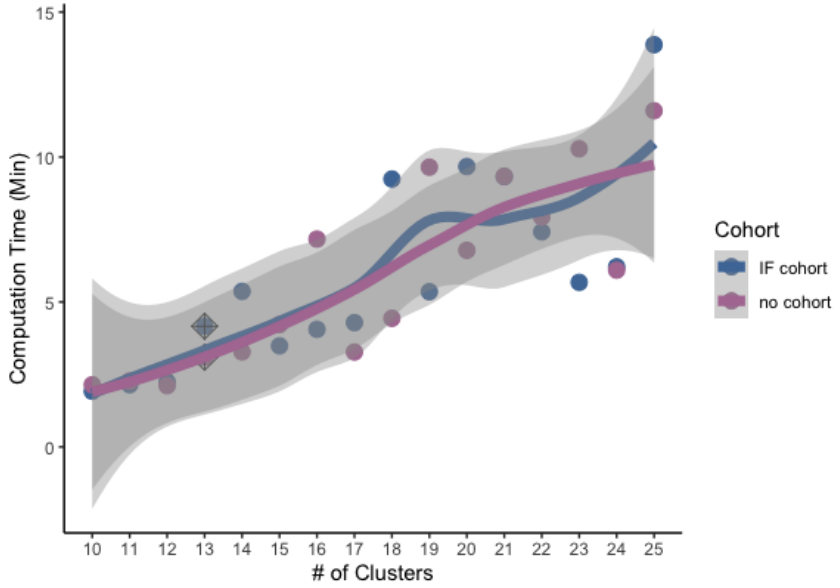


Results – Number of clusters

All genes







Results – Summary

- **Default parameters** perform reasonably well regarding ARI, BIC, and computation time
  - Variability in accuracy for cells in phase 1 parameter
- Computation time increases relatively linear as parameter values increase

Future Work

Supervised Methods

- Experiment with different tuning parameters for SingleCellNet

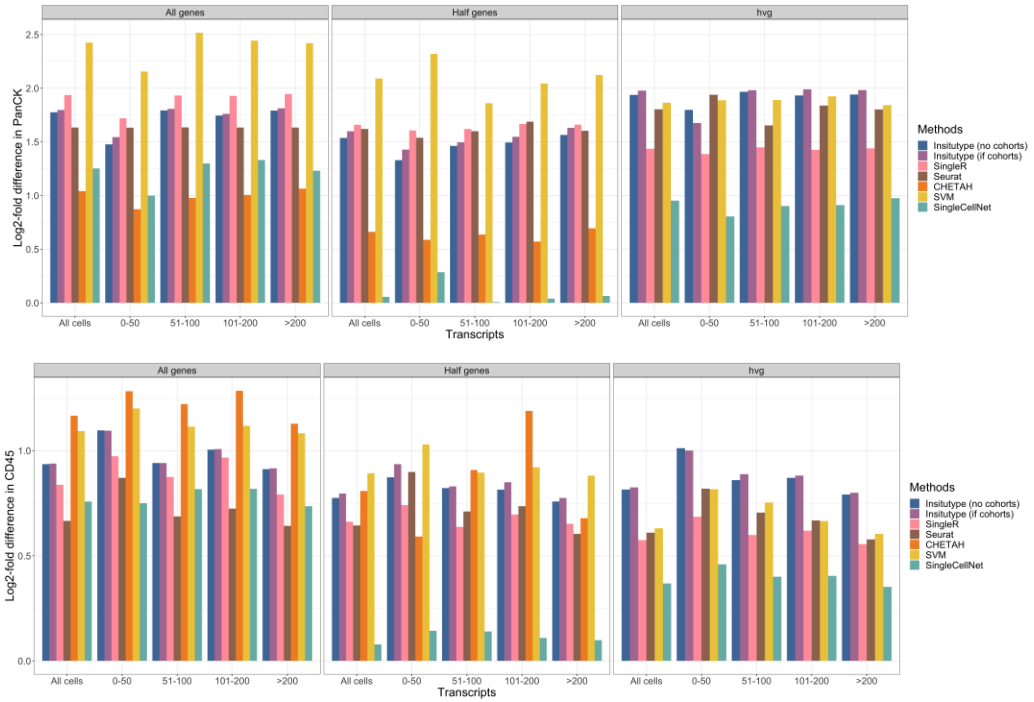
Unsupervised Methods

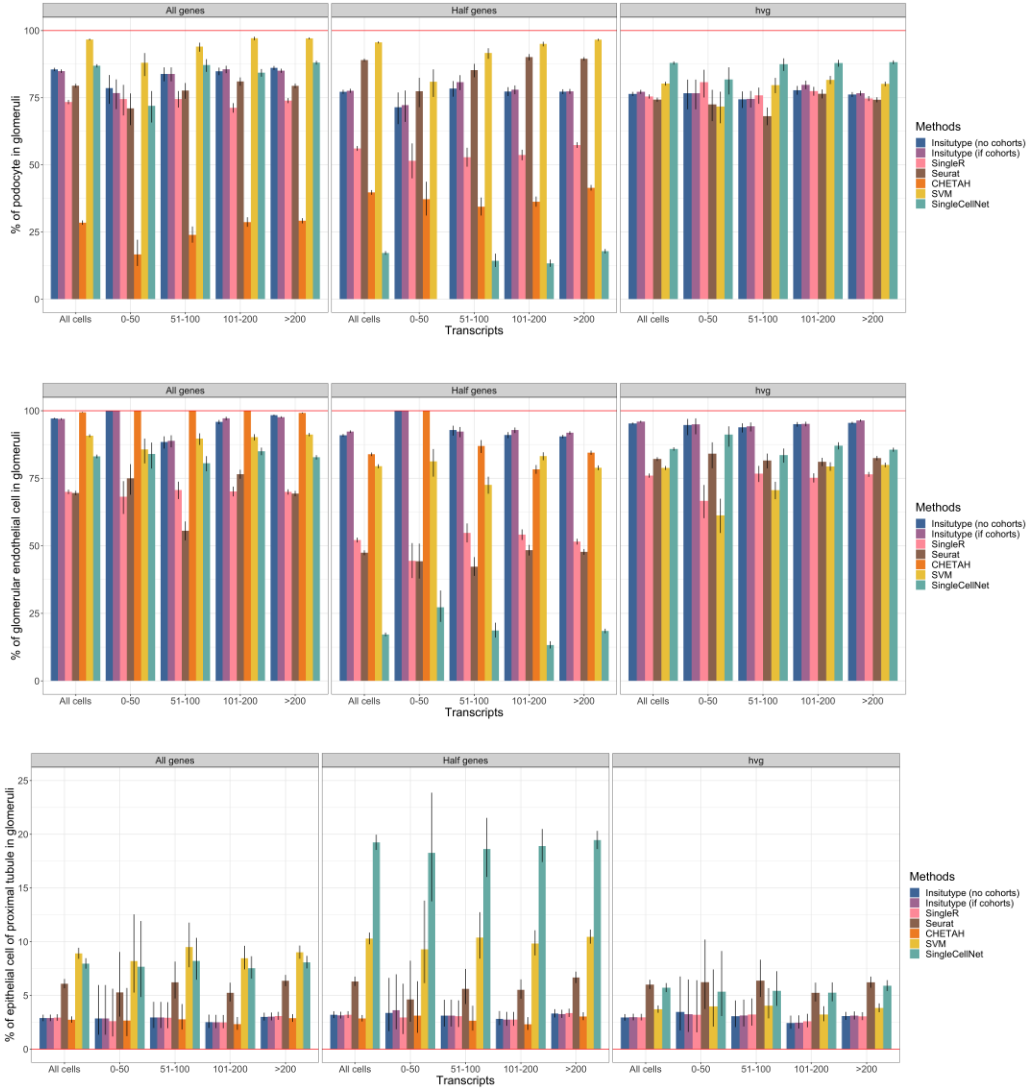
- Benchmark different combinations of parameters
- Utilize parallel computation and run all models on the same computing environment such as Amazon EC2

Thank you!

Any Questions?

Appendix





UNIVERSITY of WASHINGTON

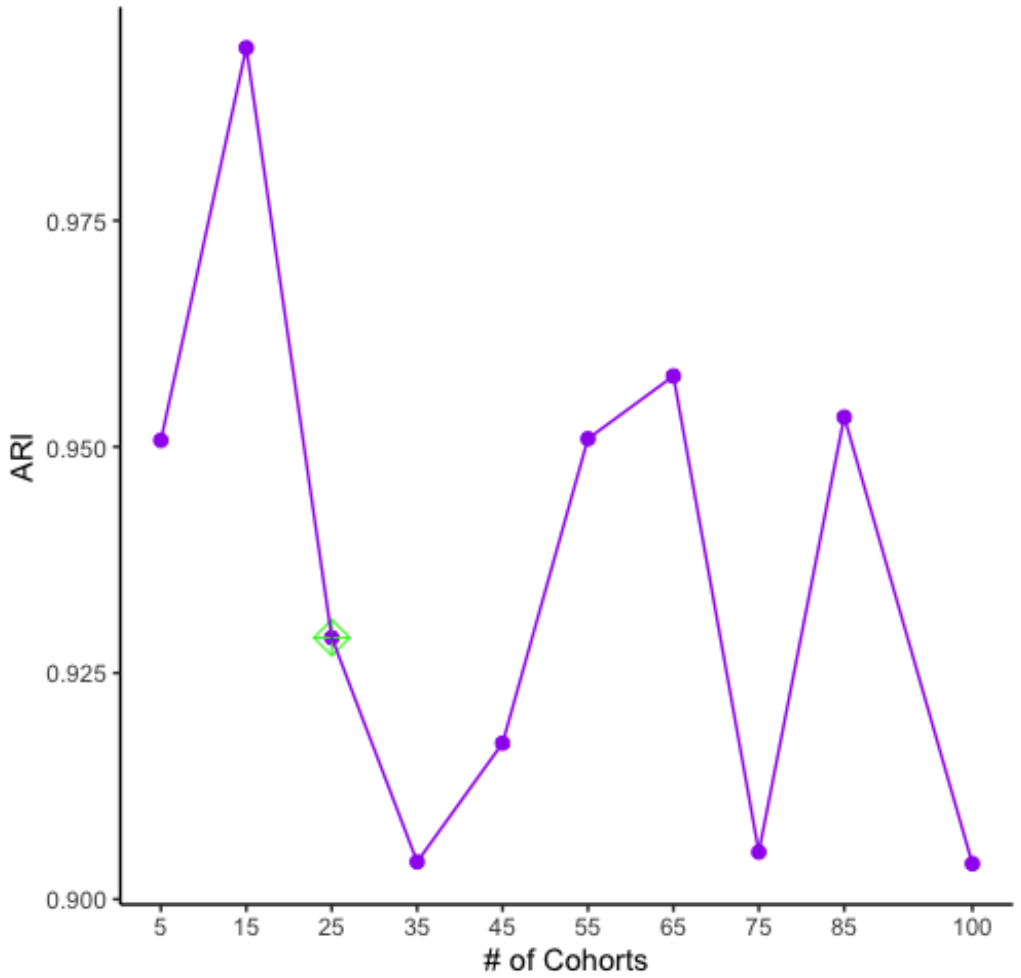
Appendix

	InSituType (no cohort)	InSituType (if cohort)	SingleR	Seurat	CHETAH	SVM	SingleCellNet
All genes	31.25 sec	33.0 sec	1.18 min	2.99 min	26.11 min	6.48 min	~1.5 hr
Half genes	15.65 sec	15.73 sec	44.07 sec	3.99 min	34.27 min	3.55 min	64.13 min
10% hvg	5.24 sec	5.06 sec	3.93 sec	5.93 min	NA	1.79 min	39.58 min

Appendix-

unsupervised cell typing performance (number of cohorts in `fastCohorting()`)

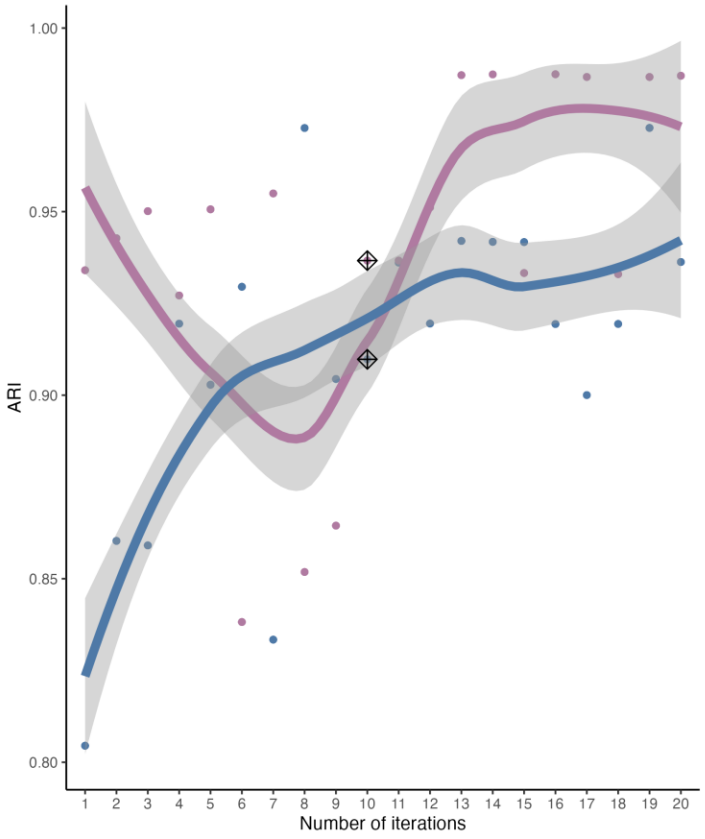
Full gene data set

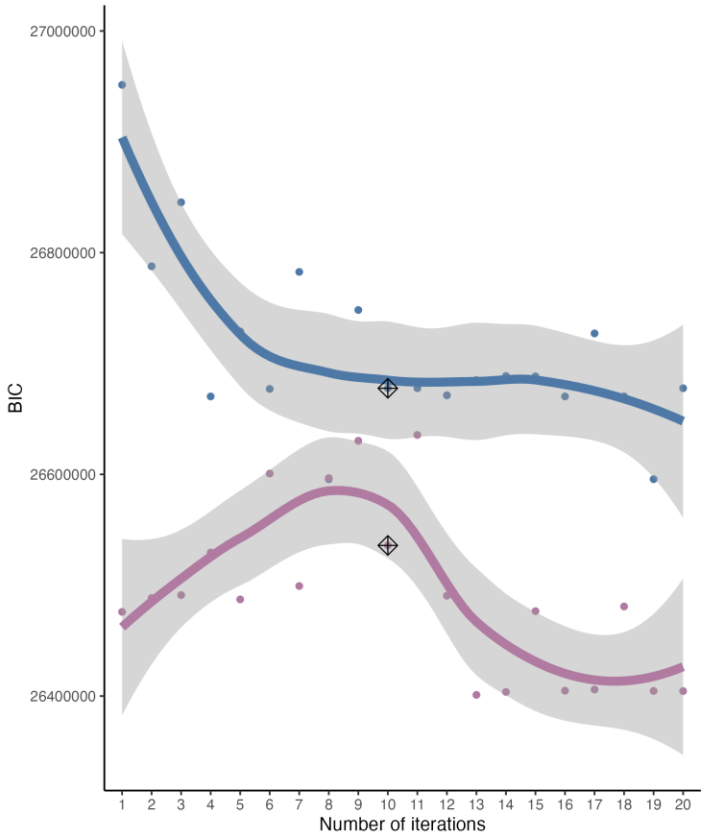


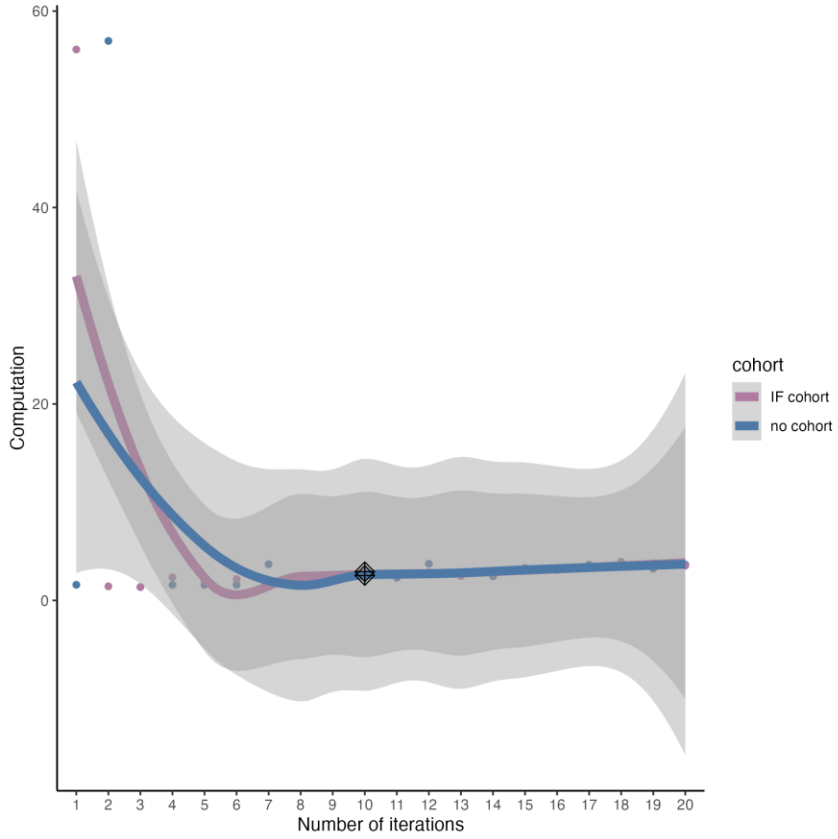
Results –

unsupervised cell typing performance (number of iterations)

Half gene data set



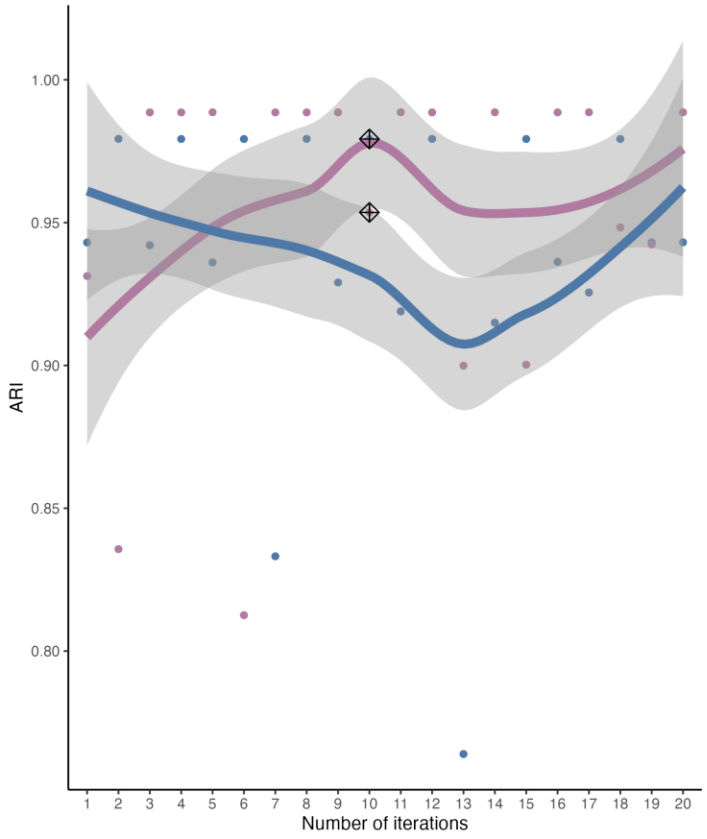


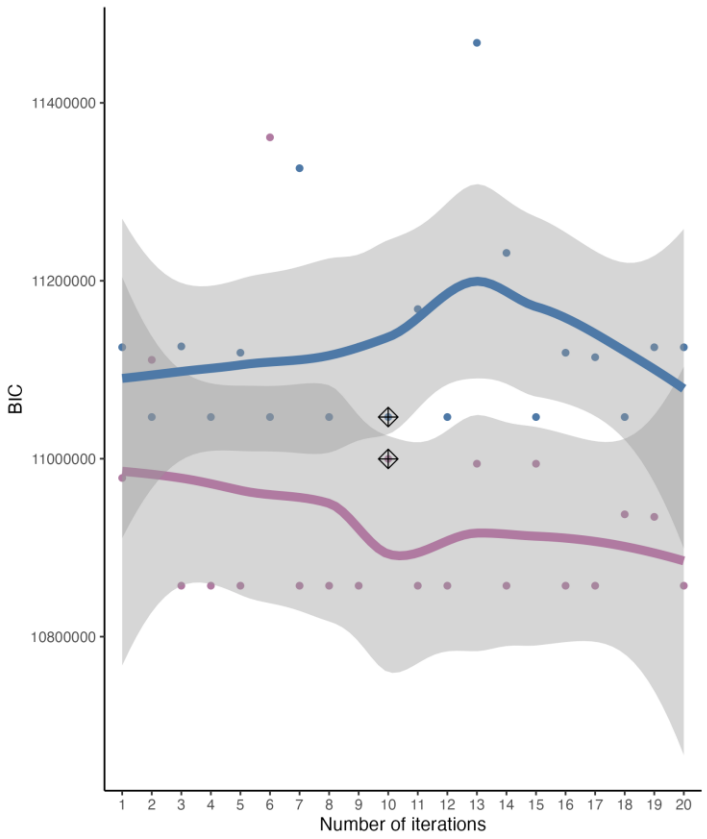


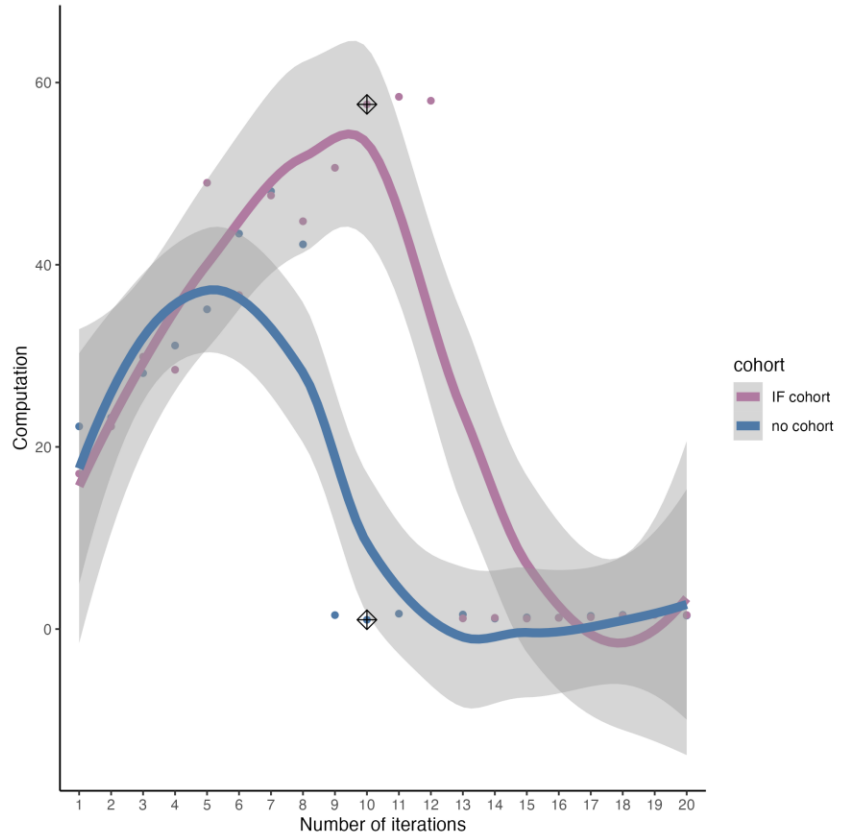
Results –

unsupervised cell typing performance (number of iterations)

10% Highly Variable Gene



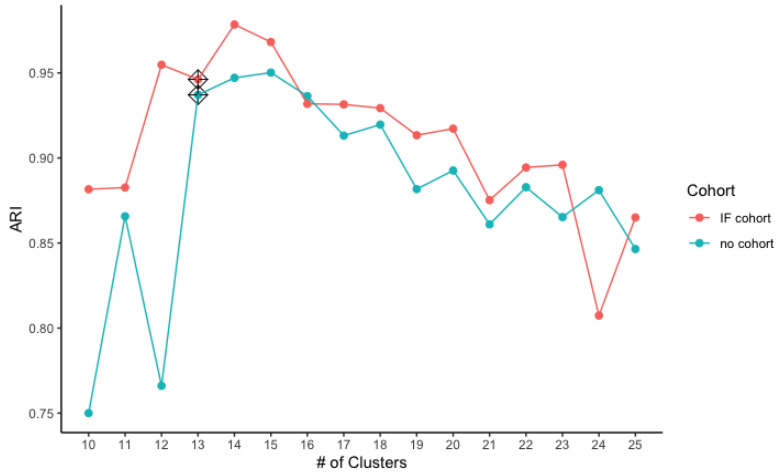


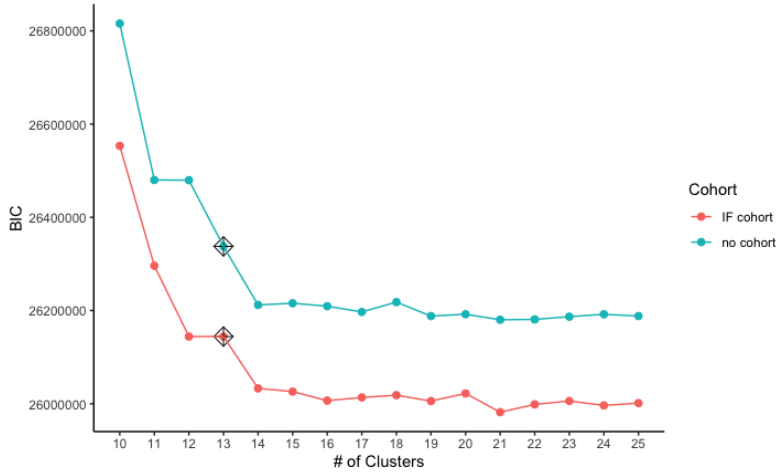


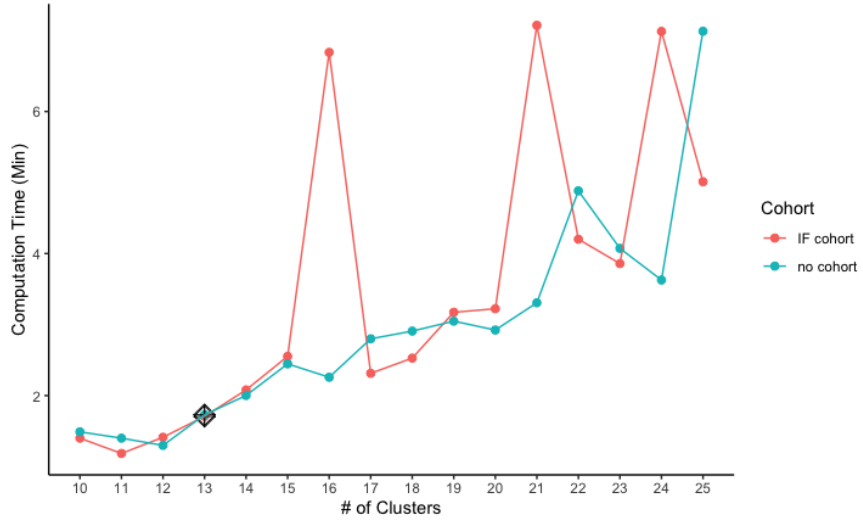
Results –

unsupervised cell typing performance (number of clusters)

Half gene data set







Results –

unsupervised cell typing performance (number of clusters)

10% Highly Variable Gene