Celltype Prediction

Fariba Roshanzamir
03-Apr-2025

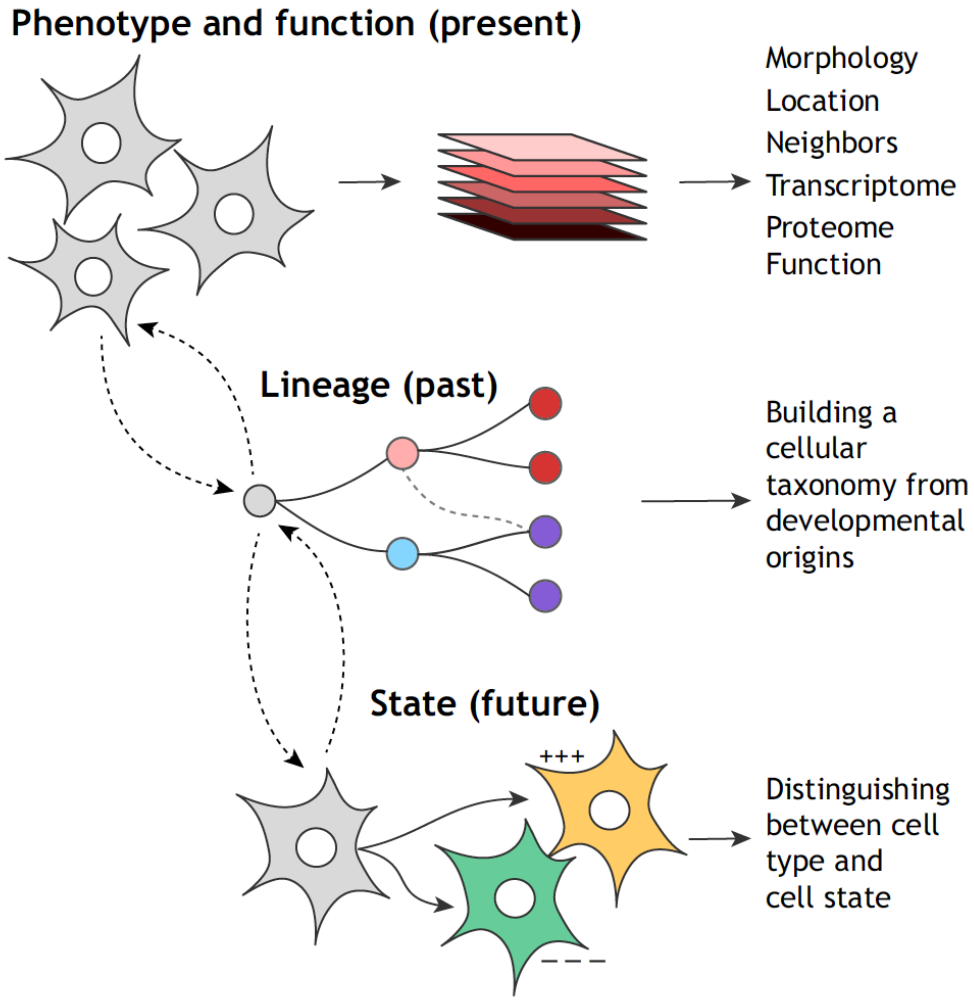
fariba.roshanzamir@scilifelab.se

Åsa Björklund (NBIS, scilifelab)
Ahmed Mahfouz (Leiden University Medical Center, TU Delft)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Cell identity

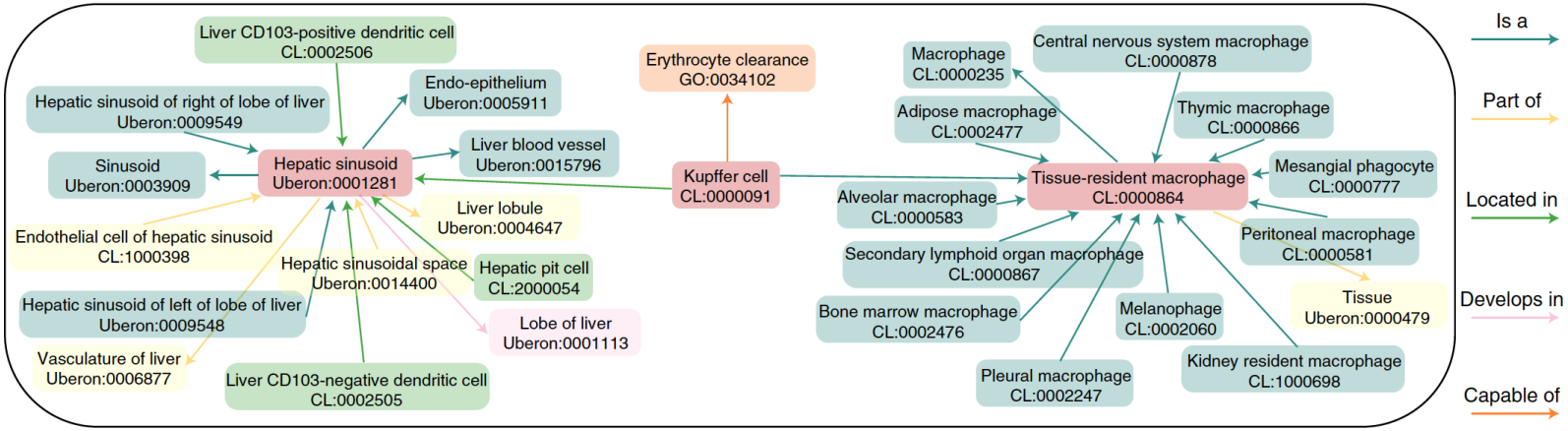


Morris, *Development* 2019

Celltype ontologies

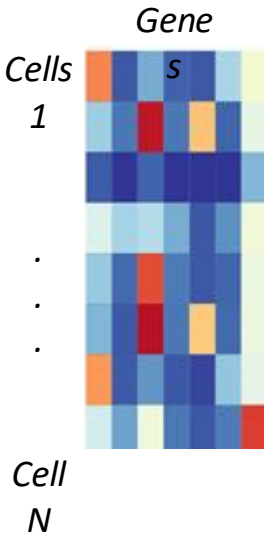
We need a standardized way of classifying celltypes.
Mainly driven by cell atlas projects.

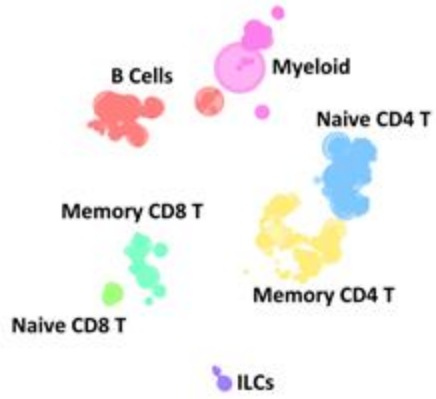
Including HuBMAP, Human Cell Atlas (HCA), cellxgene, Single Cell Expression Atlas, BRAIN Initiative Cell Census Network (BICCN), ArrayExpress, The Cell Image Library, ENCODE, and FANTOM5,



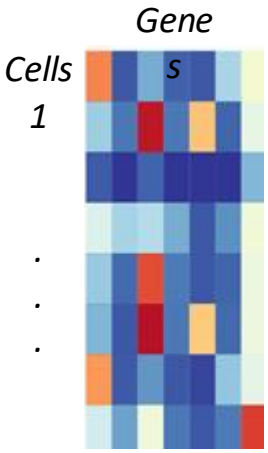
Osumi-Sutherland, *Nature Cell Biol* 2021

How can we identify cell populations?



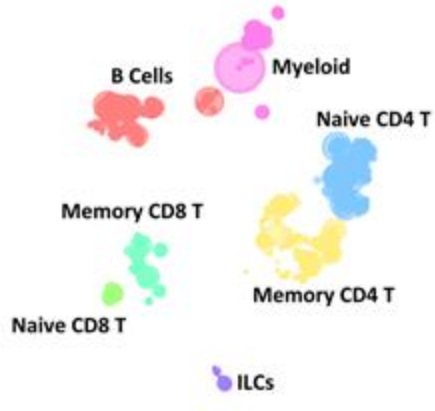


How can we identify cell populations?

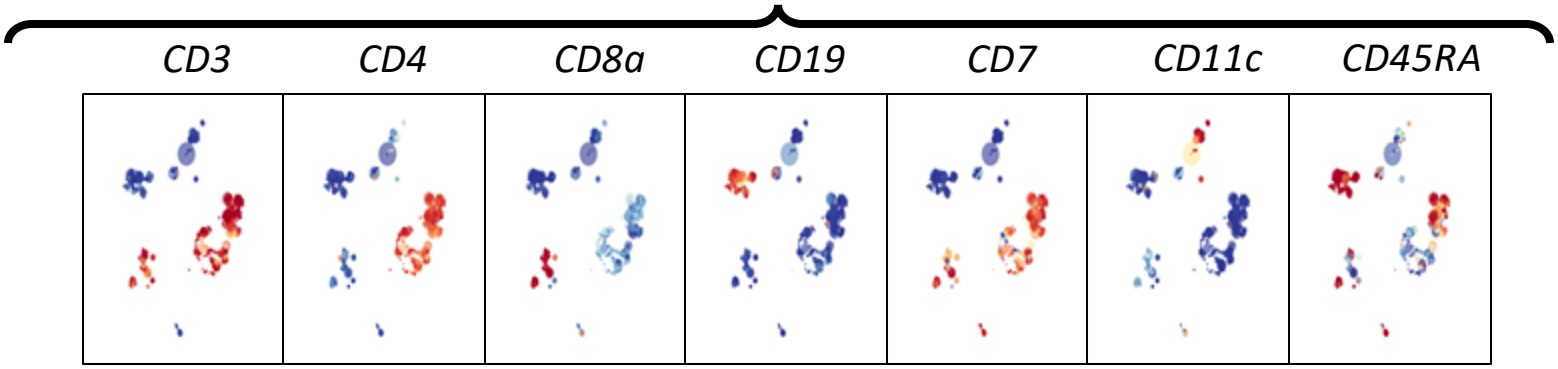


cluster

identify



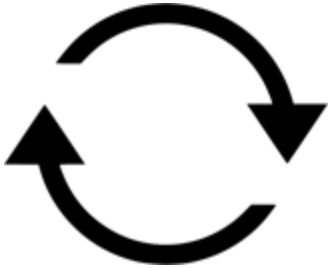
marker genes



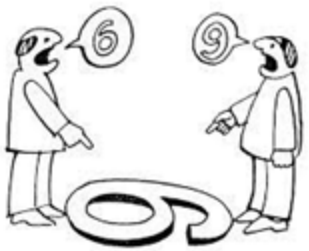
Unsupervised celltype identification is problematic

Time consuming

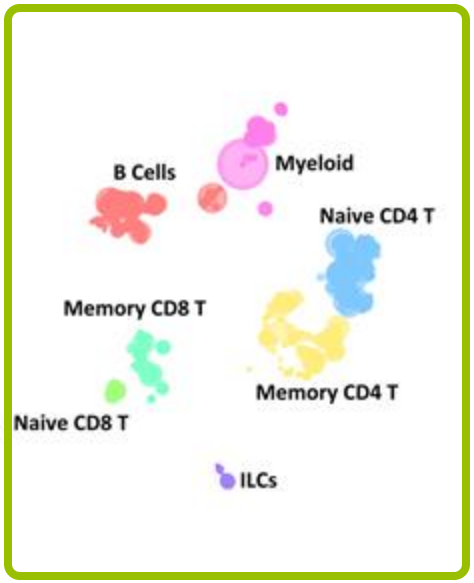
Not reproducible



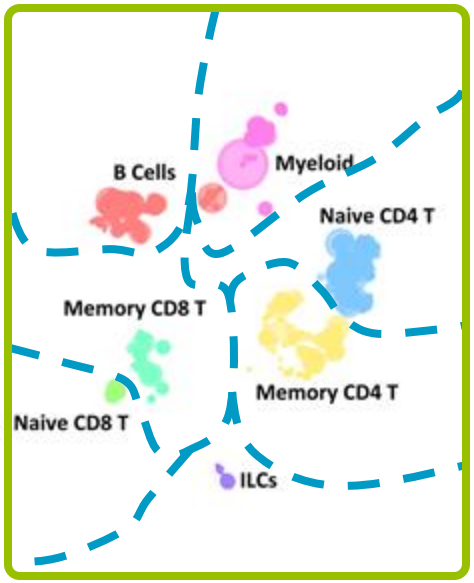
Subjective



Can we automatically identify cell populations?



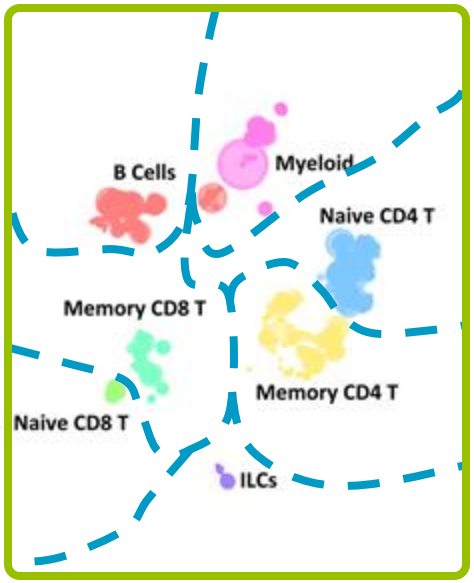
Can we automatically identify cell populations?



Can we automatically identify cell populations?

Clustering

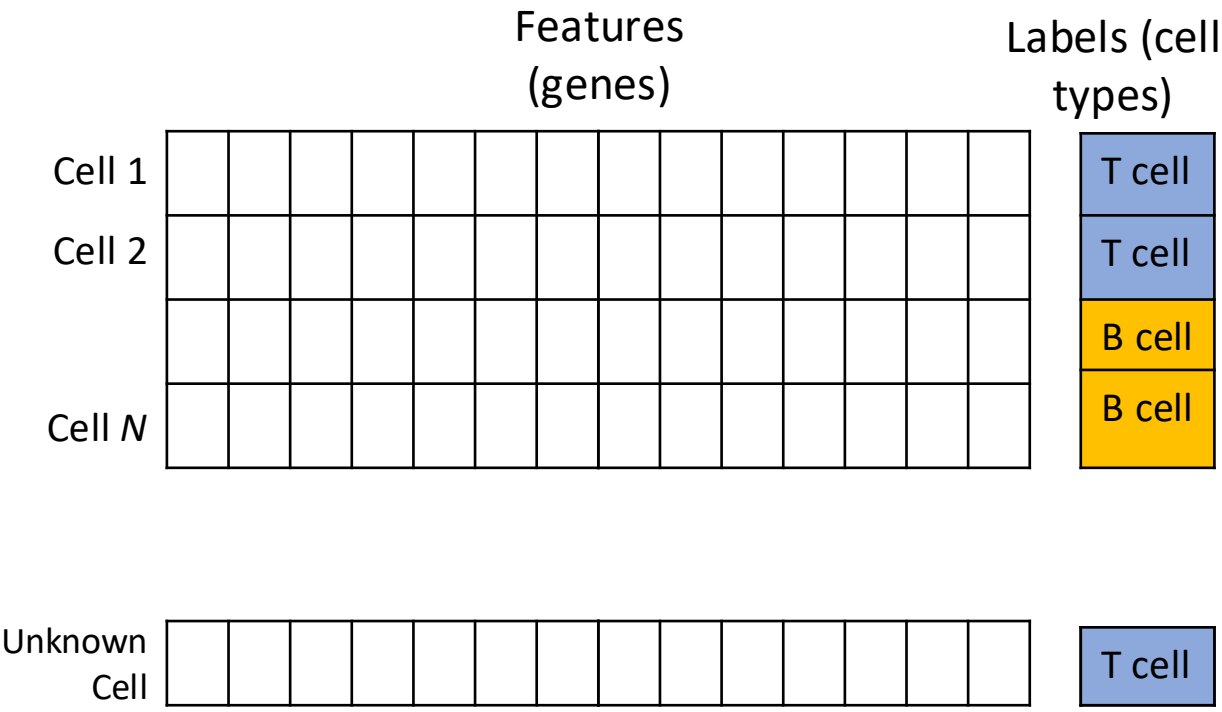
- **Unsupervised** learning
- Discovering structure/relations
- Clusters are defined by a decision boundary

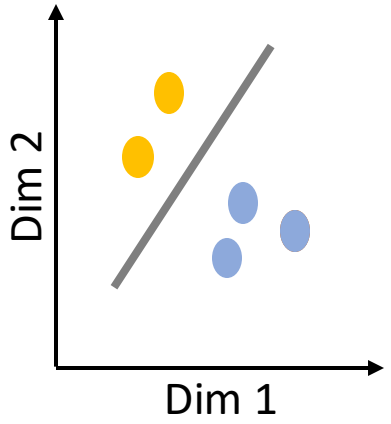


Classification

- **Supervised** learning
- Prior information available about different groups
- Classifiers find descriptions of decision boundaries

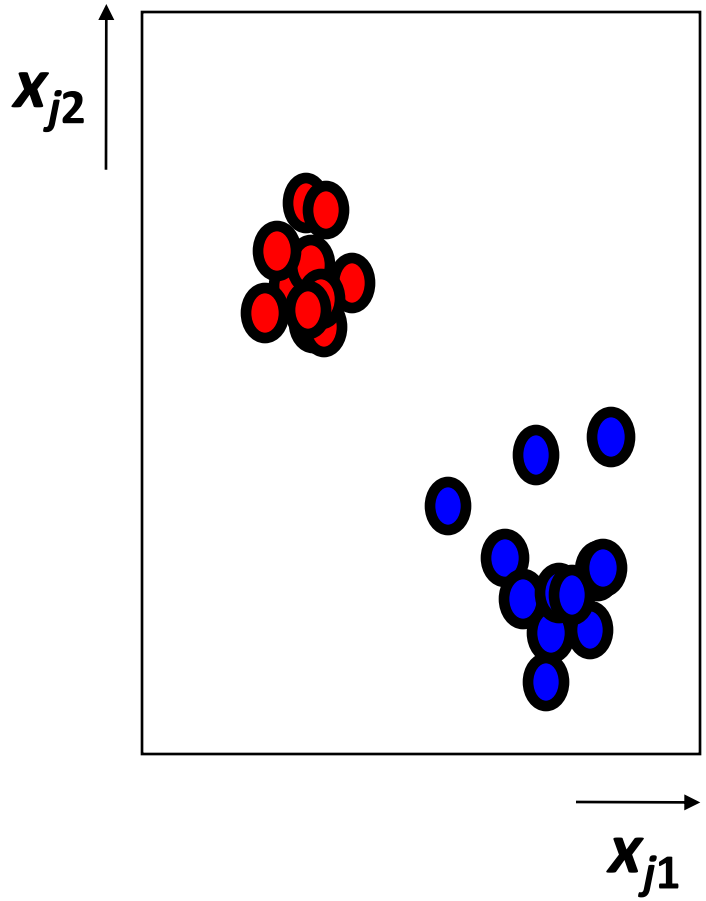
Classification





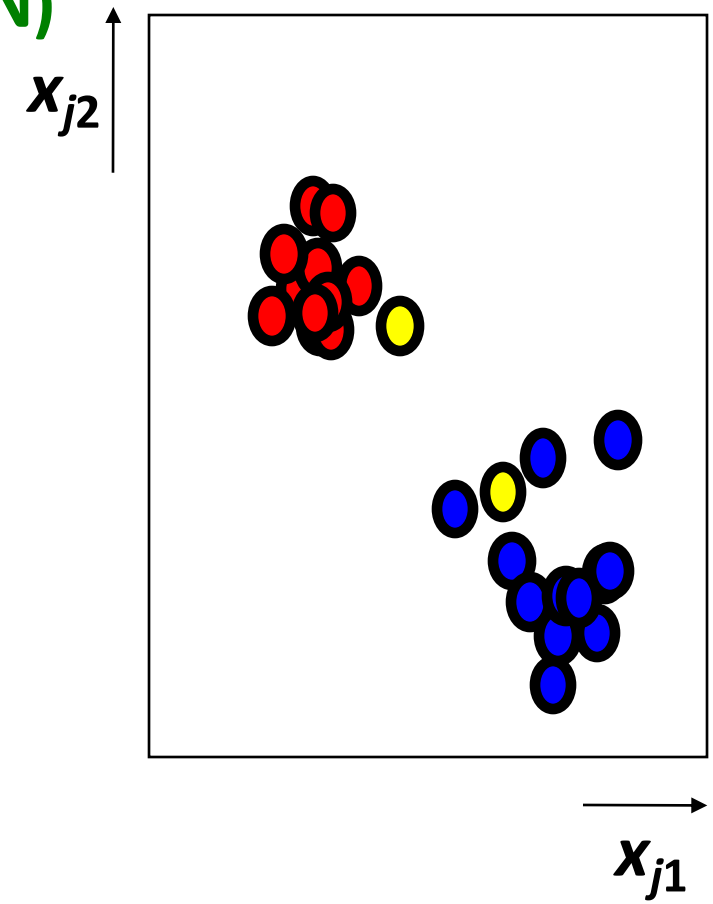
Classifier training

- Dataset: for *j* th cell:
  - gene expressions $\boldsymbol{x}_j$
  - class label: $y_j \in \{1=T,-1=B\}$
- Classifier: $\hat{y}_j = W(x_j)$
- Errors: $E = \mathrm{sum}(E_j) \quad E_j = \begin{cases} 1 & \text{if} \quad \hat{y}_j \neq y_j \\ 0 & \text{if} \quad \hat{y}_j = y_j \end{cases}$
- Place decision boundary (i.e. change *W*) s.t. *E* is minimal



Instance Based Learning (Lazy Classification)

- **Example: Nearest neighbor (k-NN)**



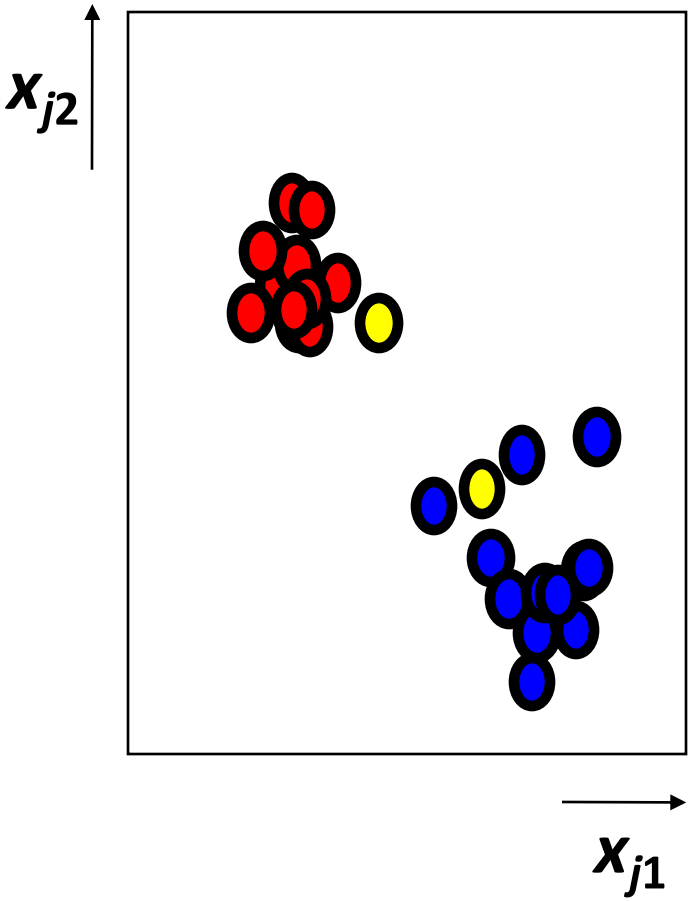
- Keep the whole training dataset
- A query example (vector) comes
- Find closest example(s)
- Predict

- *No actual training*

Nearest Neighbor (k-NN)

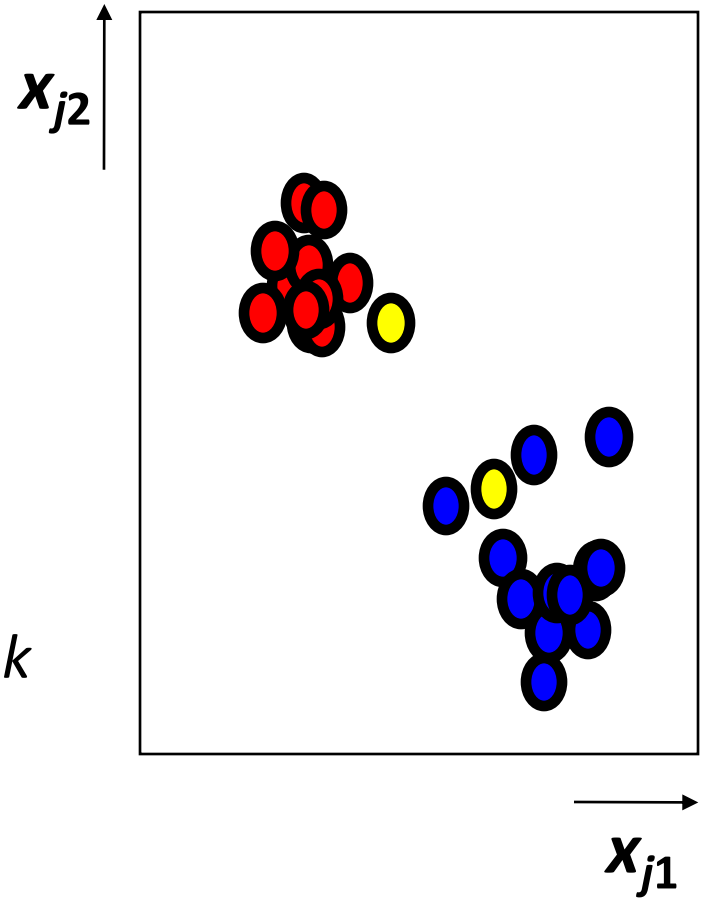
- To make Nearest Neighbor work we need 4 things:

1) Distance metric:
2) How many neighbors to look at?
3) Weighting function (optional)
4) How to fit with the local points?

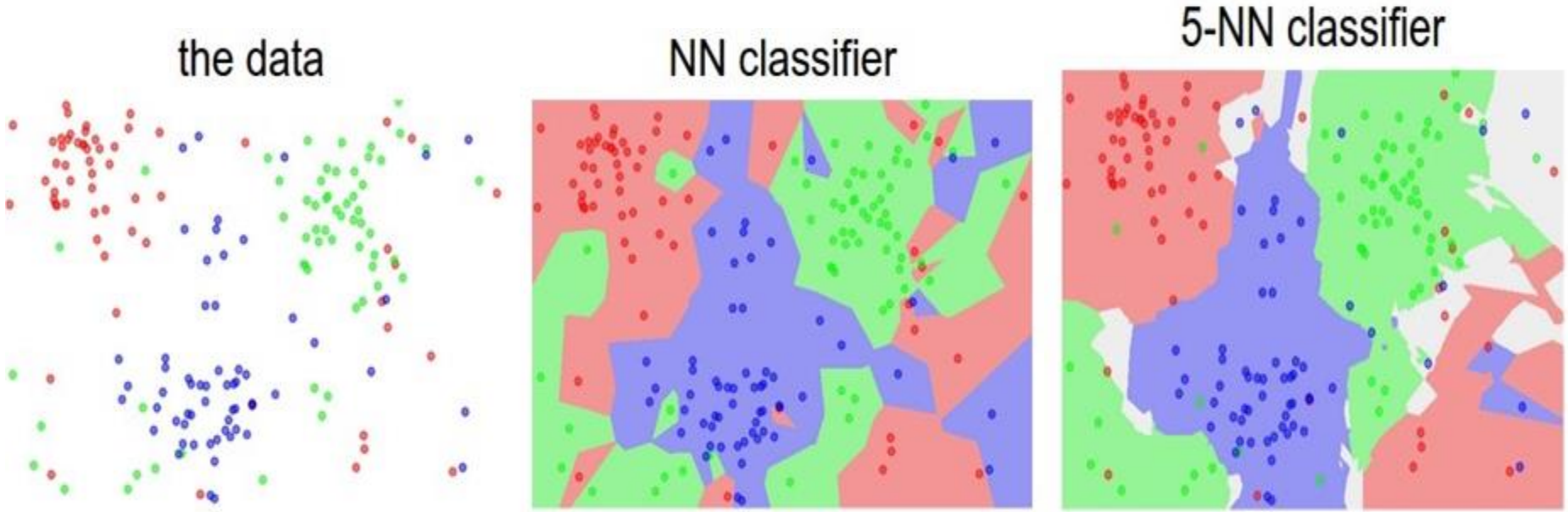


Nearest Neighbor (k-NN)

- Distance metric:
  - Euclidean
- How many neighbors to look at?
  - *k*
- Weighting function (optional):
  - Unused
- How to fit with the local points?
  - Predict the average output among *k* nearest neighbors



Effect of *k*

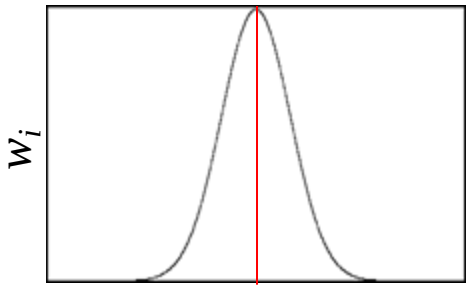


Weighted Nearest Neighbor (kernel regression)

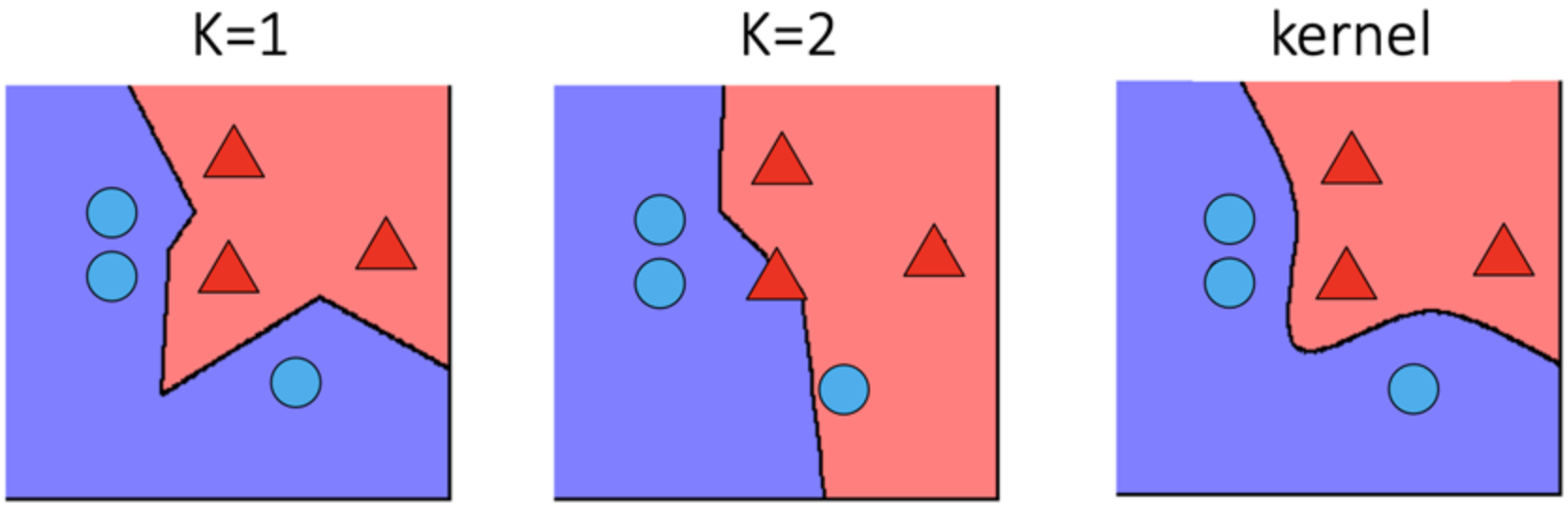
- Distance metric:
  - Euclidean
- How many neighbors to look at?
  - All of them!
- Weighting function:

$$w_i = \exp(-\frac{d(x_i,q)^2}{K_w})$$

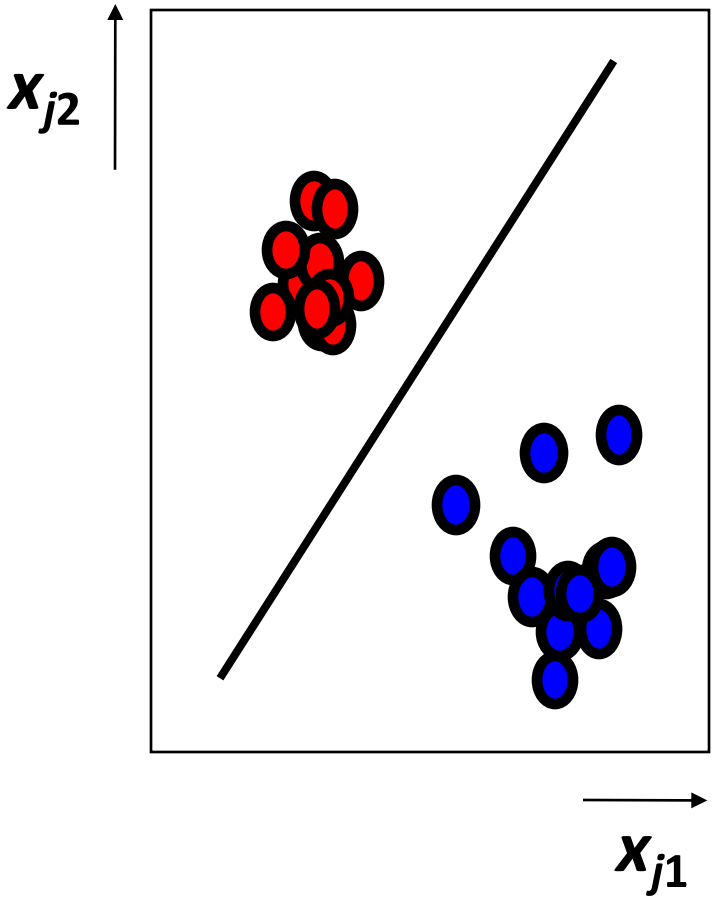
  - Nearby points to a query q are weighted more strongly. K_W: kernel width
- How to fit with the local points?
  - Predict the weighted average $\frac{\sum_i w_i y_i}{\sum_i w_i}$



Comparison: K=1, K=2, kernel



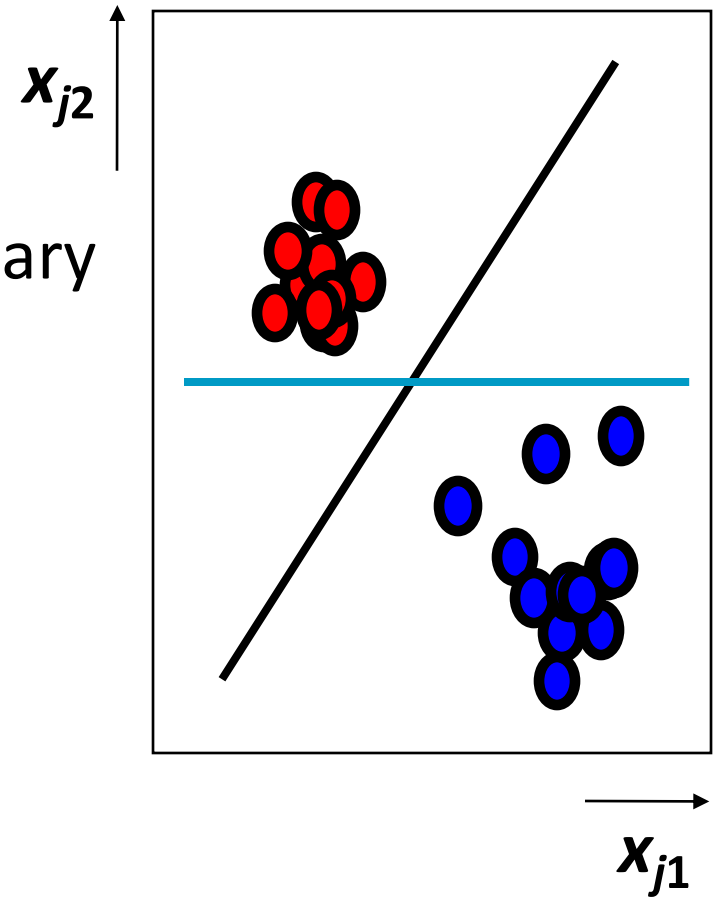
Support Vector Machine (SVM)



Decision Boundary

Support Vector Machine (SVM)

Which boundary is better?



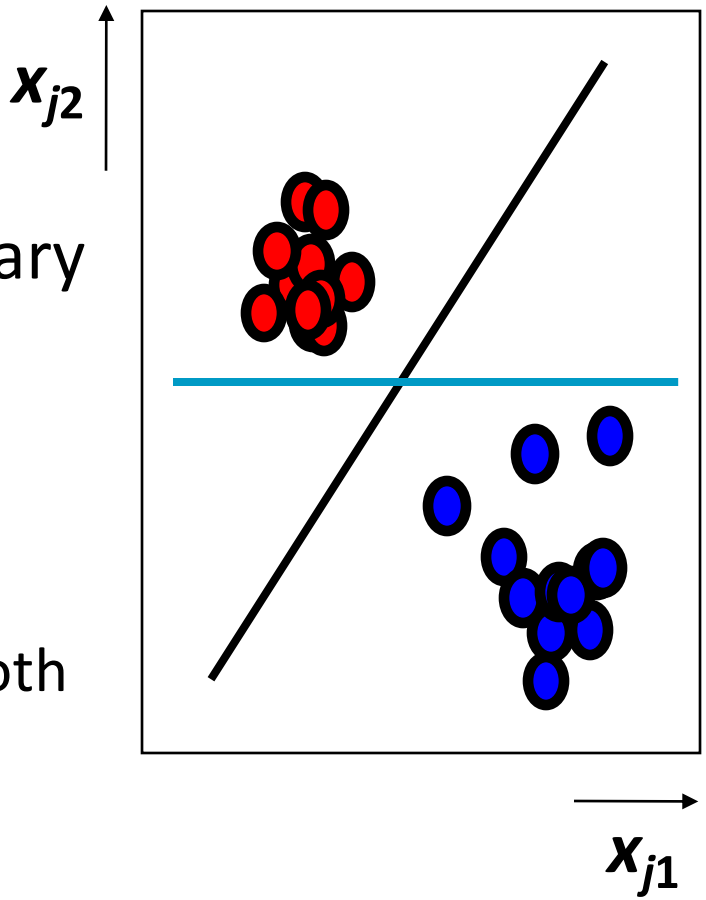
Boundary 1

Boundary 2

Support Vector Machine (SVM)

Which boundary is better?

The one that maximizes the margins from both labels.



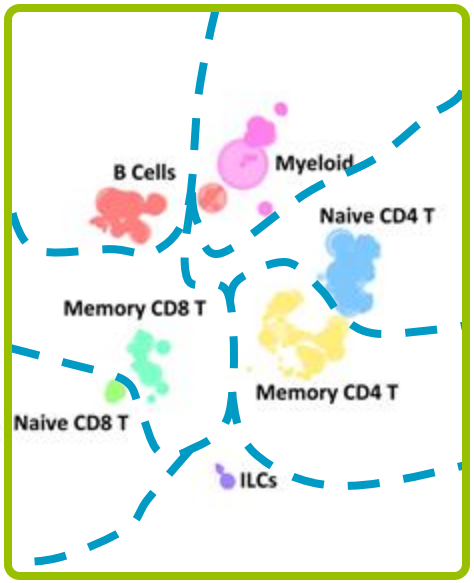
Boundary 1

Boundary 2

Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)



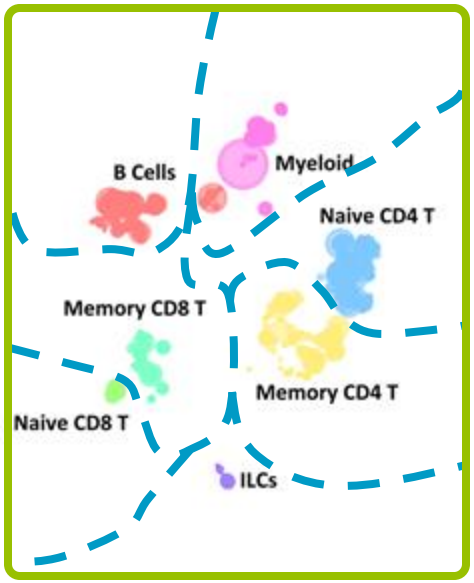
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



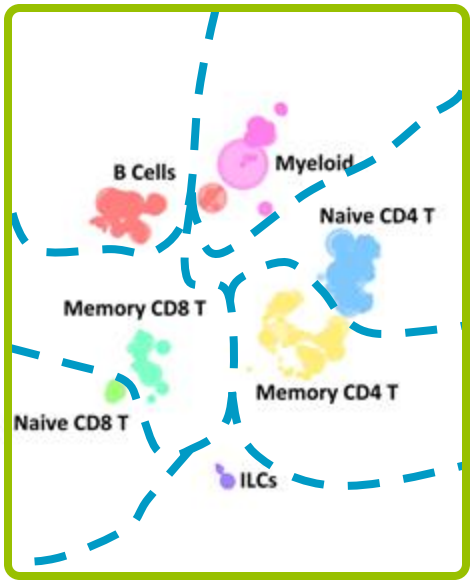
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



Annotated cells

Test data

Unannotated cells

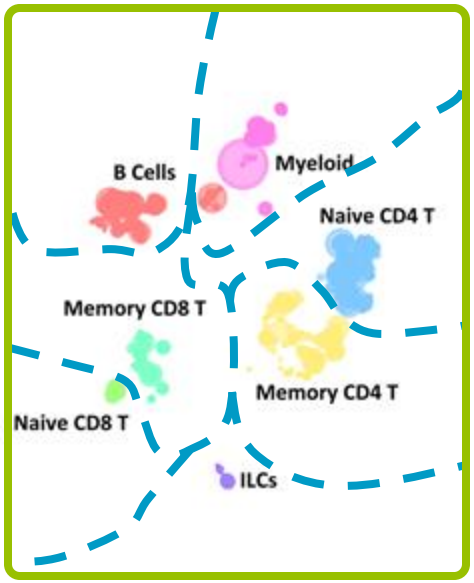
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



Annotated cells

Unknown (novel) cells

Rejection option

Test data

Unannotated cells

Benchmark paper 2019

Research | Open access | Published: 09 September 2019

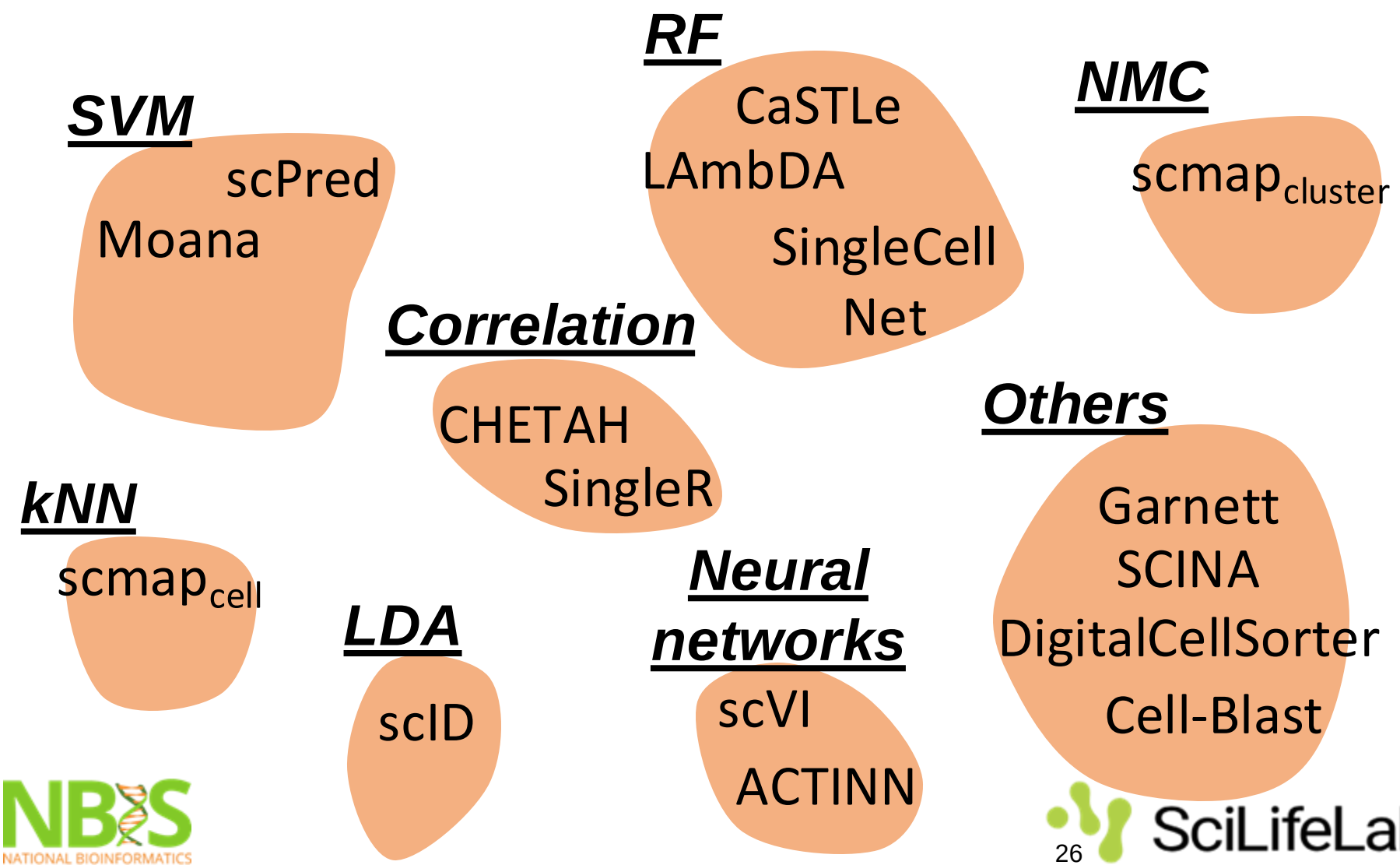
A comparison of automatic cell identification methods for single-cell RNA sequencing data

Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders & Ahmed Mahfouz

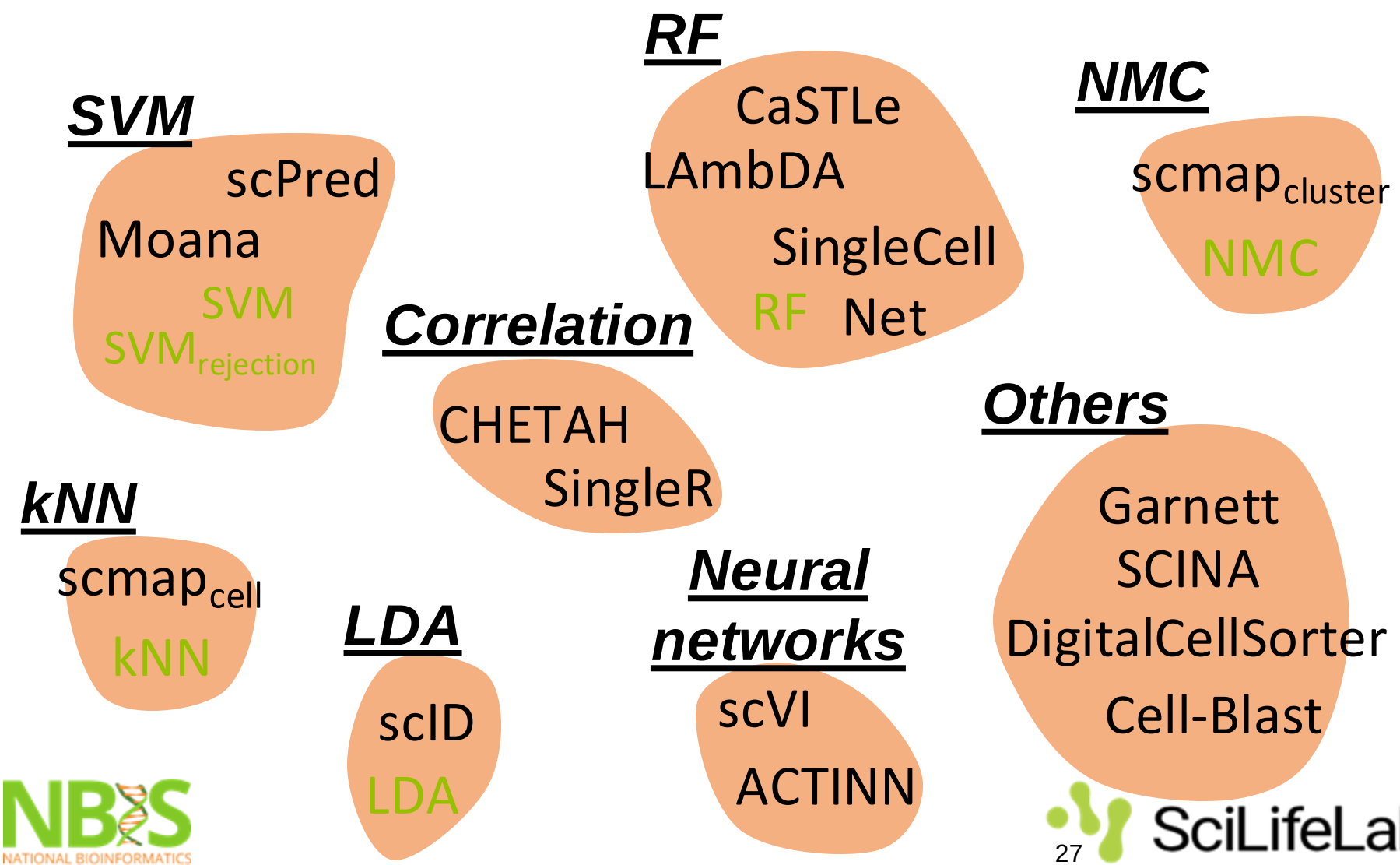
Genome Biology **20**, Article number: 194 (2019) | Cite this article

(Abdelaal et al. Genome Biology 2019)

16 existing classifiers (April 2019)



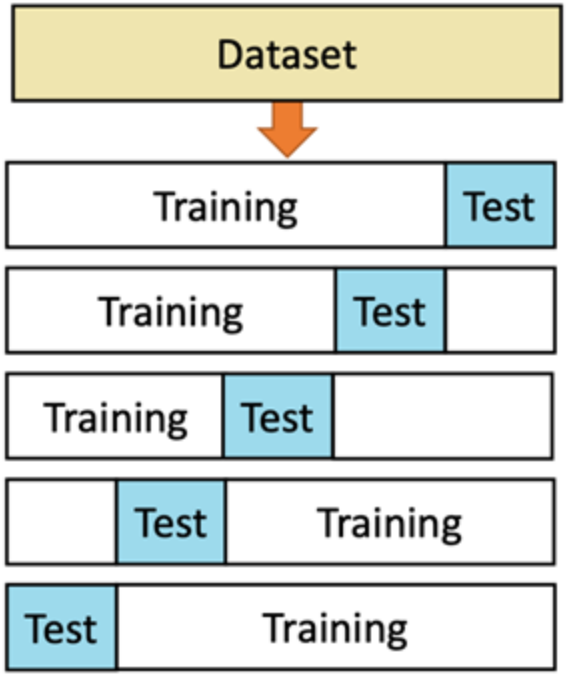
16 existing + 6 off-the-shelf classifiers



Experiment 1: intra-dataset evaluation

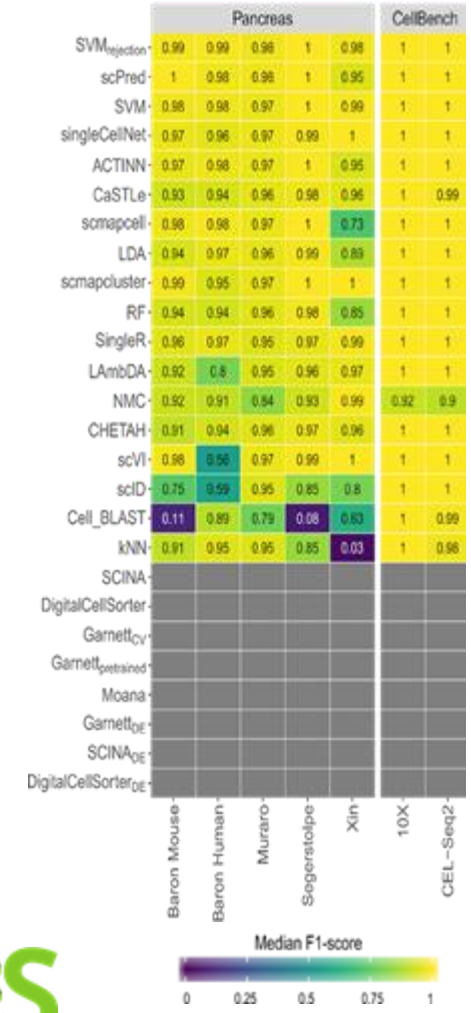
- Stratified 5-fold cross validation
- Performance evaluation
  - Median F1-score: $F1 = 2\frac{precision.recall}{precision+recall}$
  - % unlabelled cells

It ranges from 0 to 1, with 1 indicating perfect precision and recall

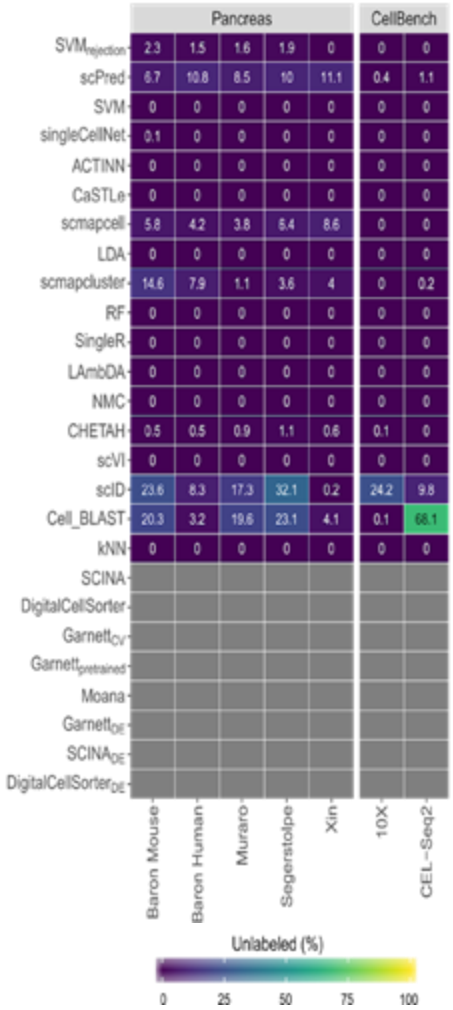


Most classifiers work well

Median F1-score



% Unlabeled

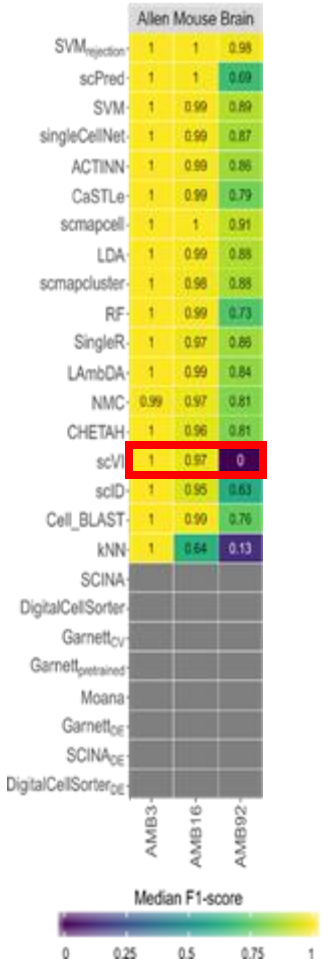


NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

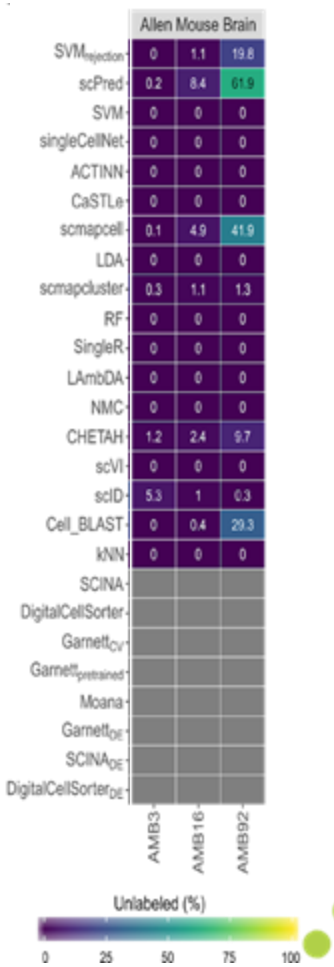
SciLifeLab

Performance drops with deeper annotation

Median F1-score

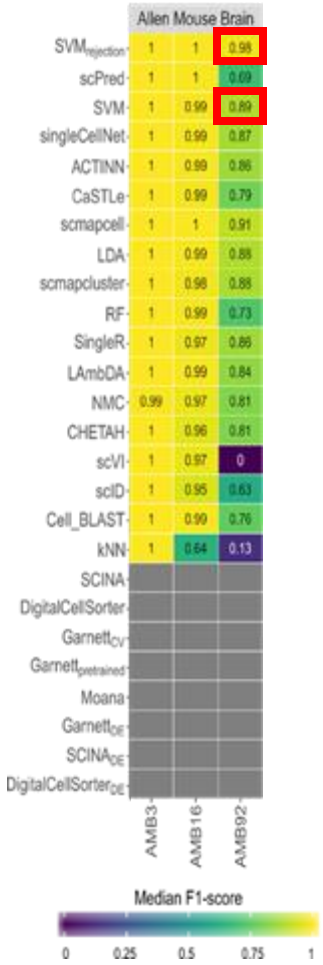


% Unlabeled

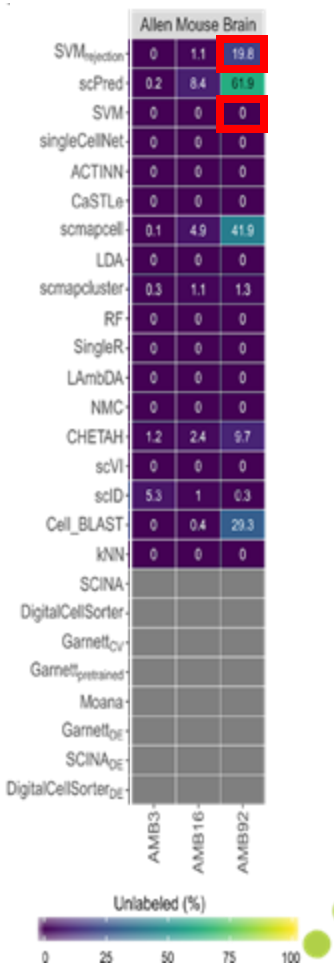


Trade-off between high performance and rejecting cells

Median F1-score

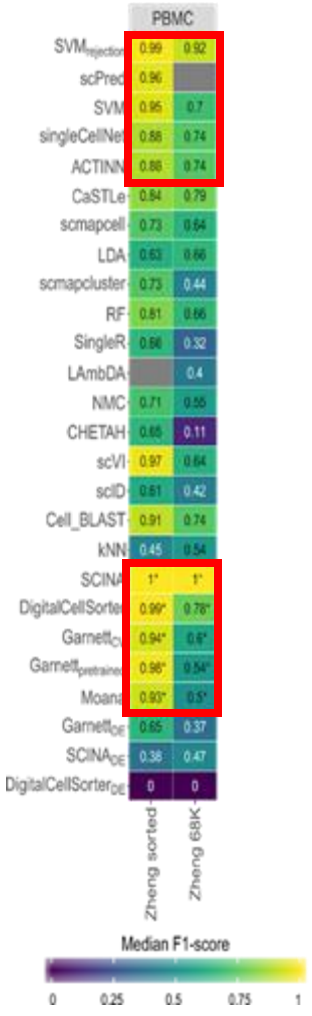


% Unlabeled

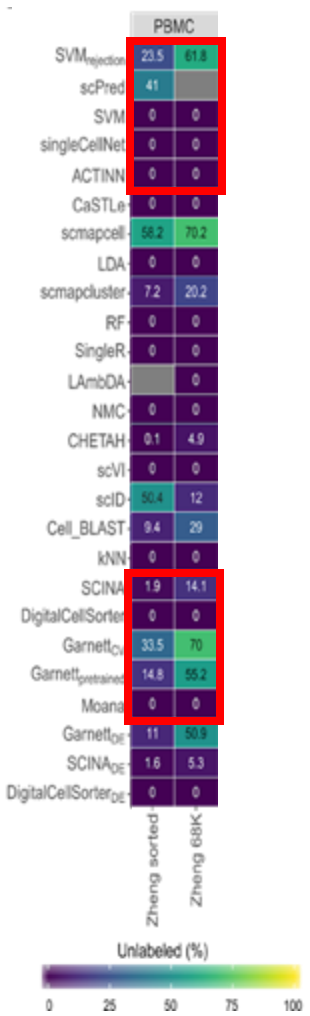


Prior knowledge is not always beneficial

Median F1-score



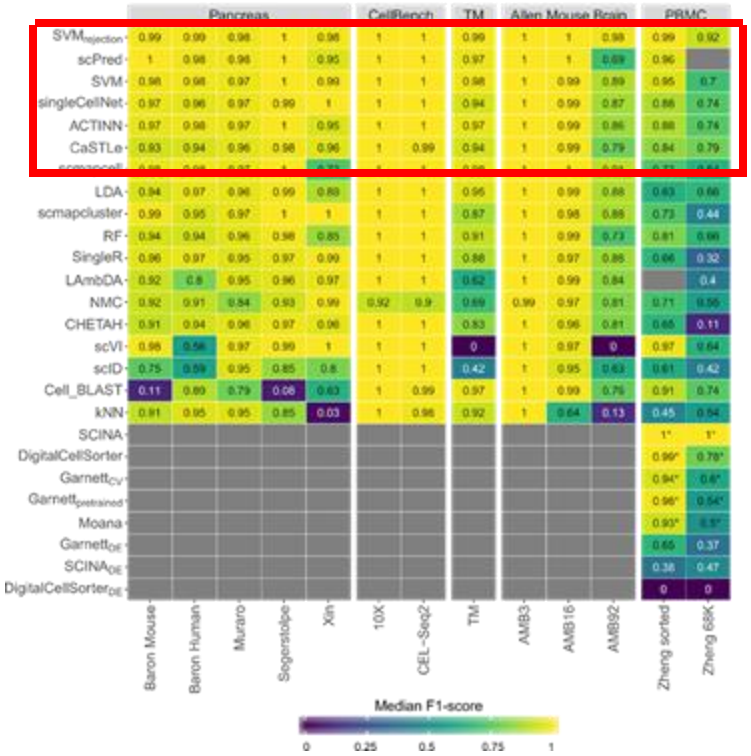
% Unlabeled



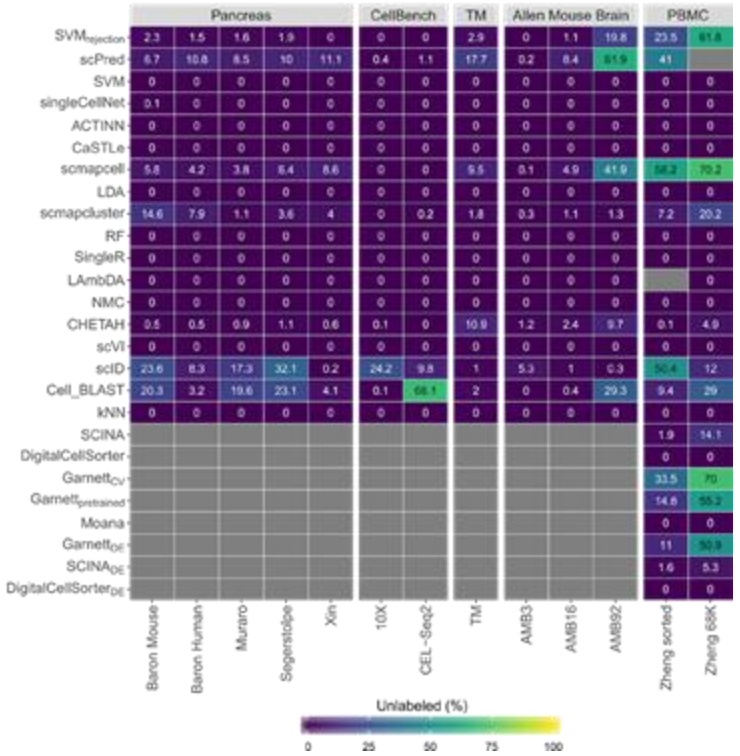
Lower number of classes!

Off-the-shelf SVM outperforms dedicated single cell classifiers

Median F1-score

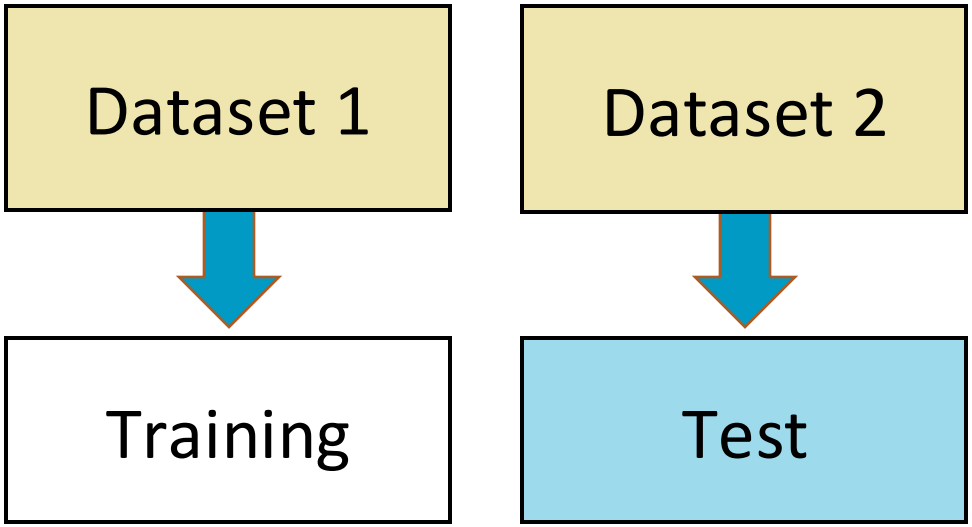


% Unlabeled



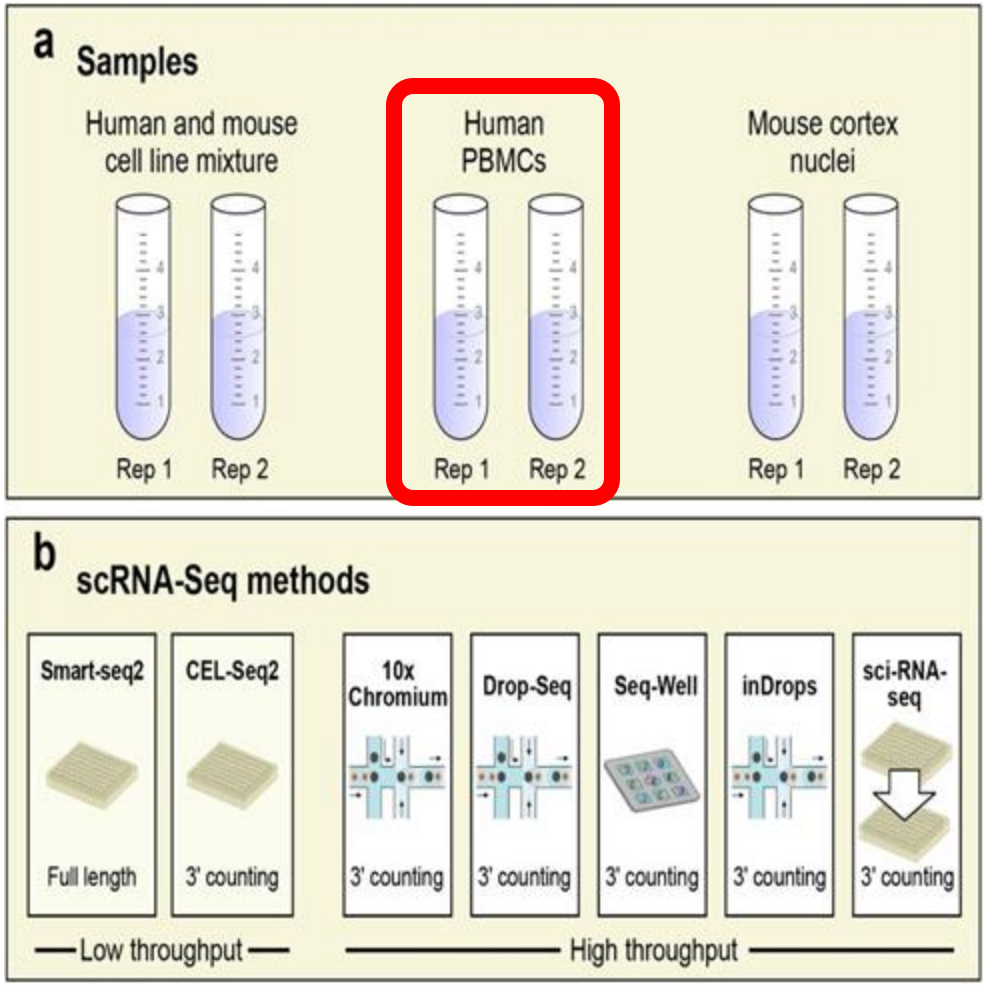
Experiment 2: inter-dataset evaluation

- Train on one dataset, evaluate on another
- More realistic scenario
- More challenging, data is not aligned

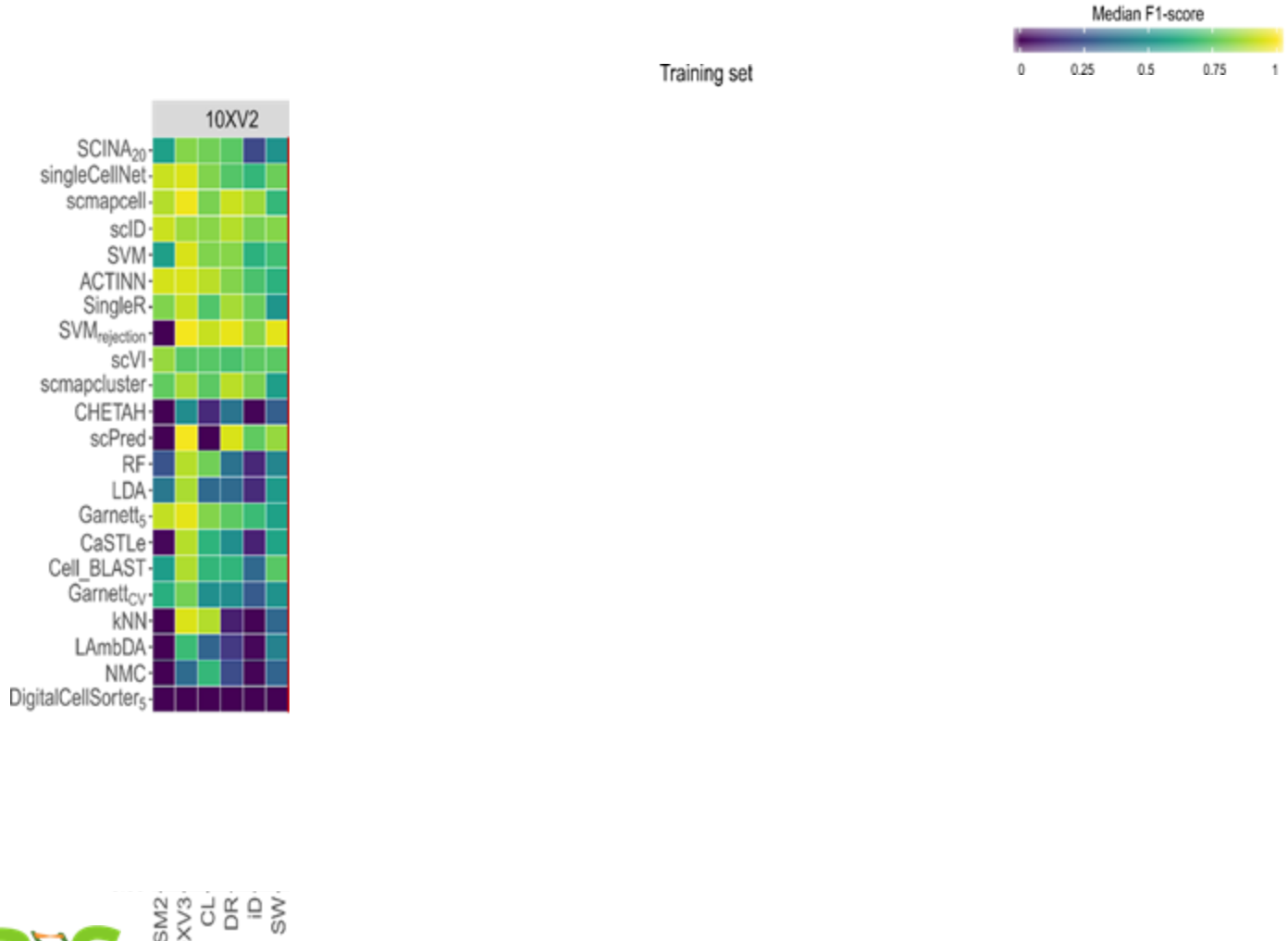


34

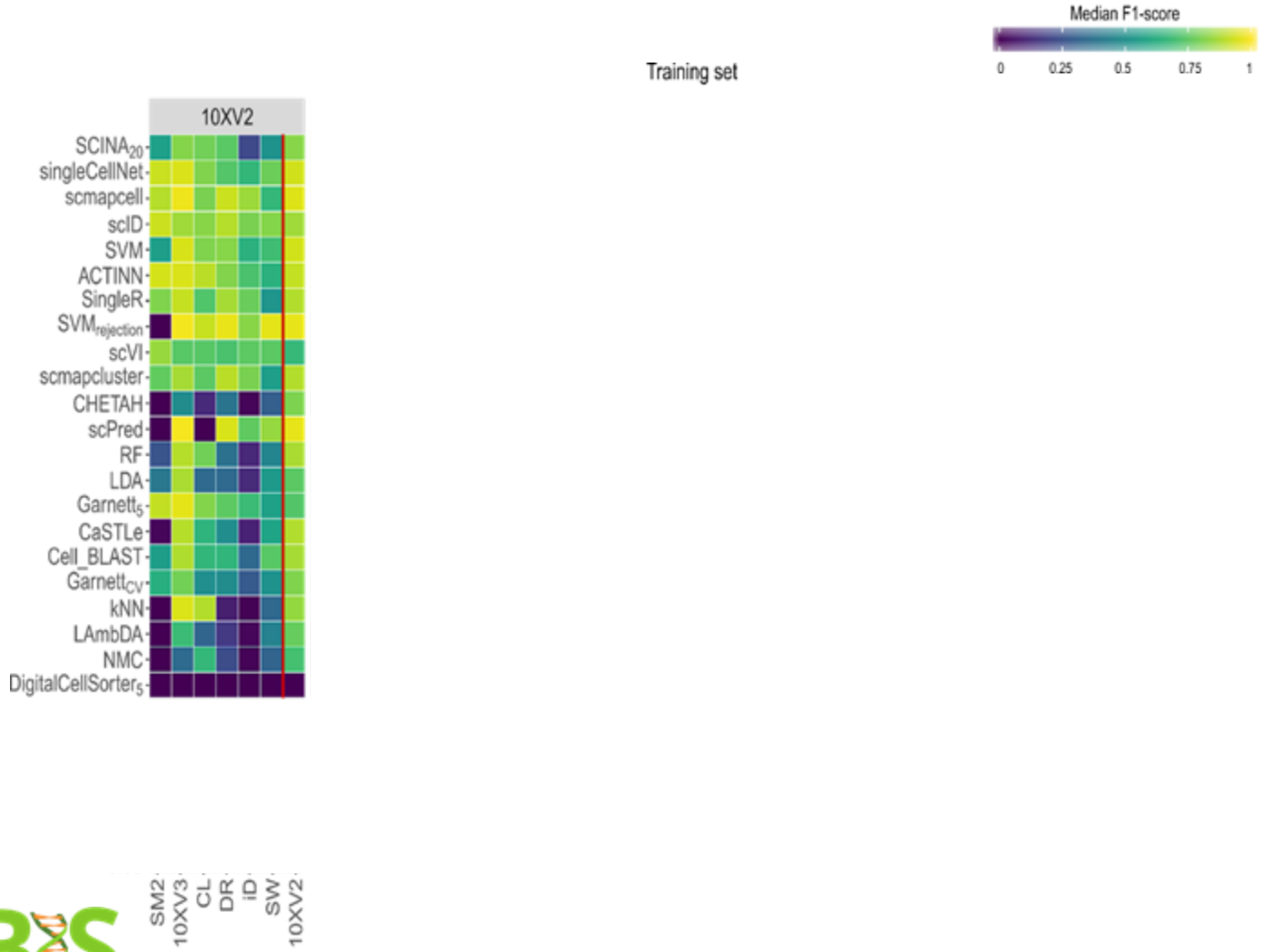
Experiment 2: inter-dataset evaluation



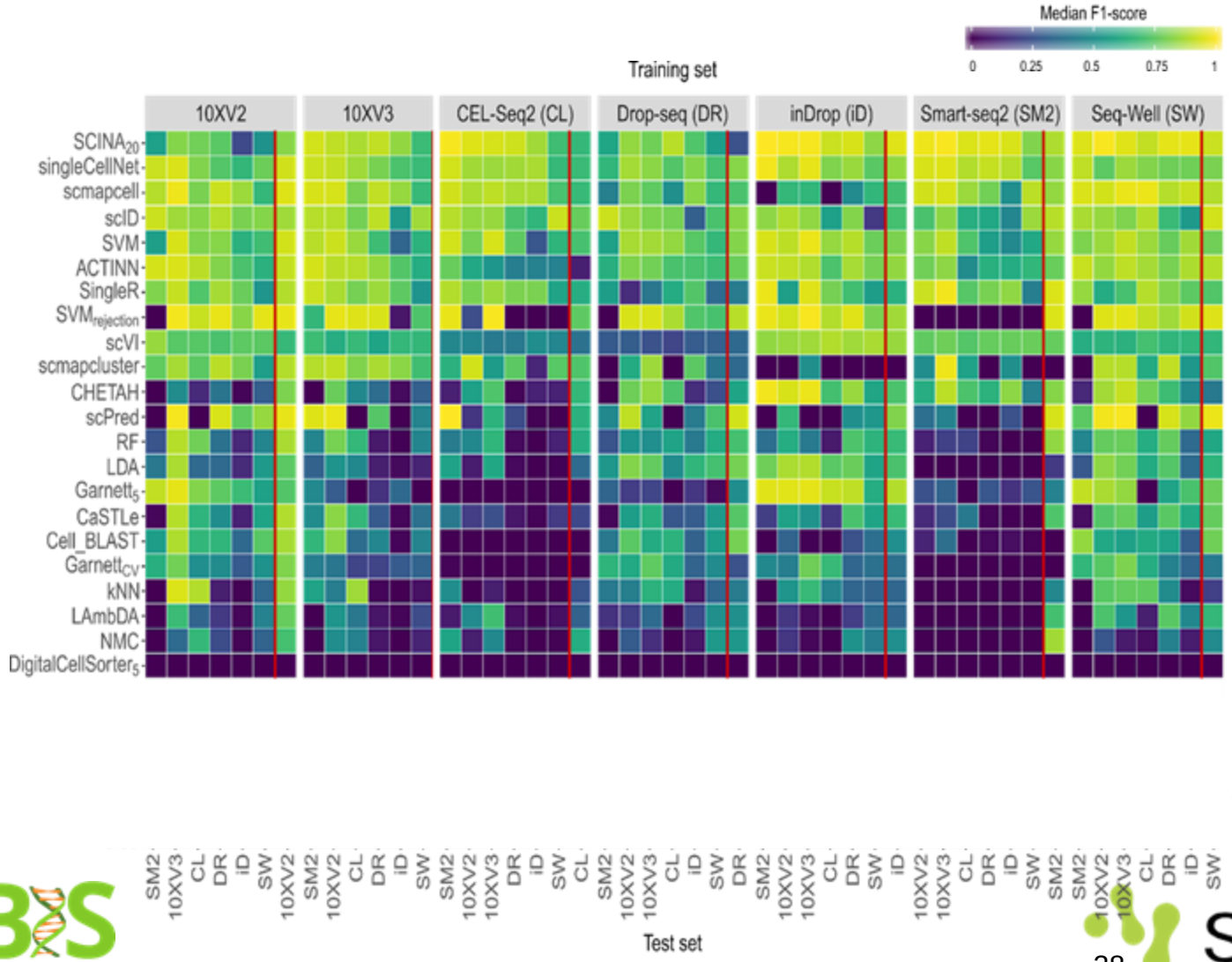
Prediction across protocols



Prediction across protocols



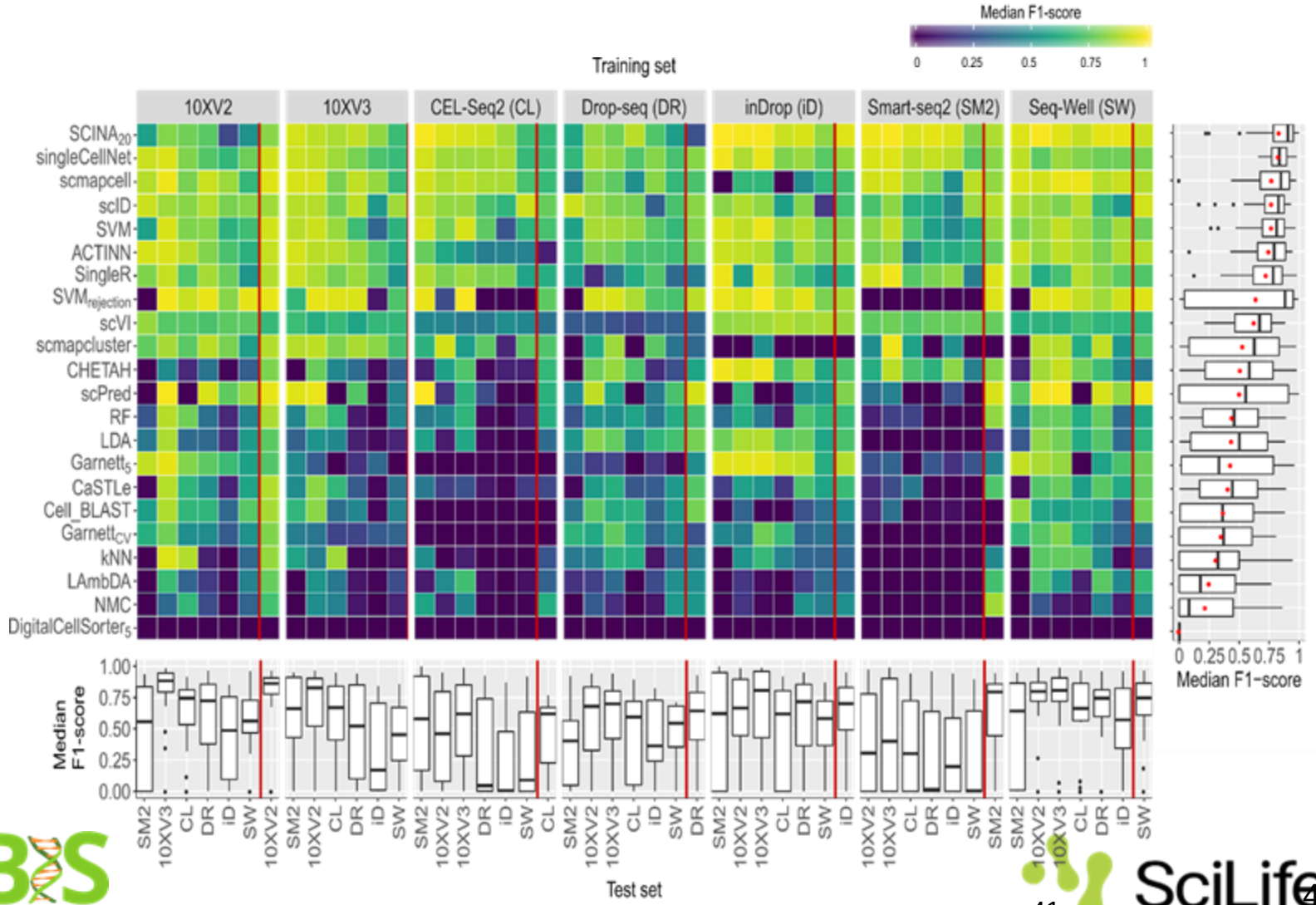
Prediction across protocols



Prediction across protocols

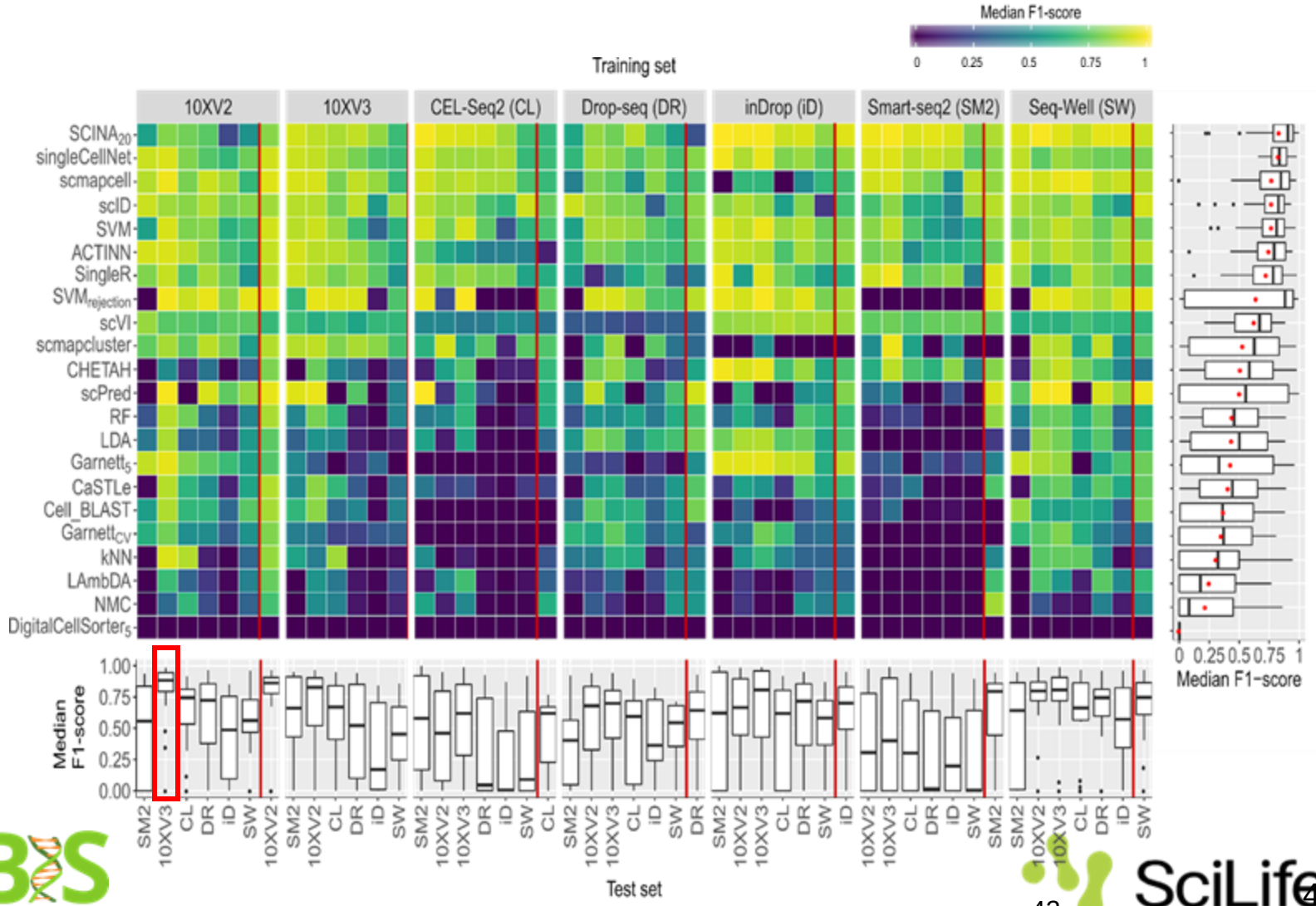
Prediction across protocols

Prediction across protocols



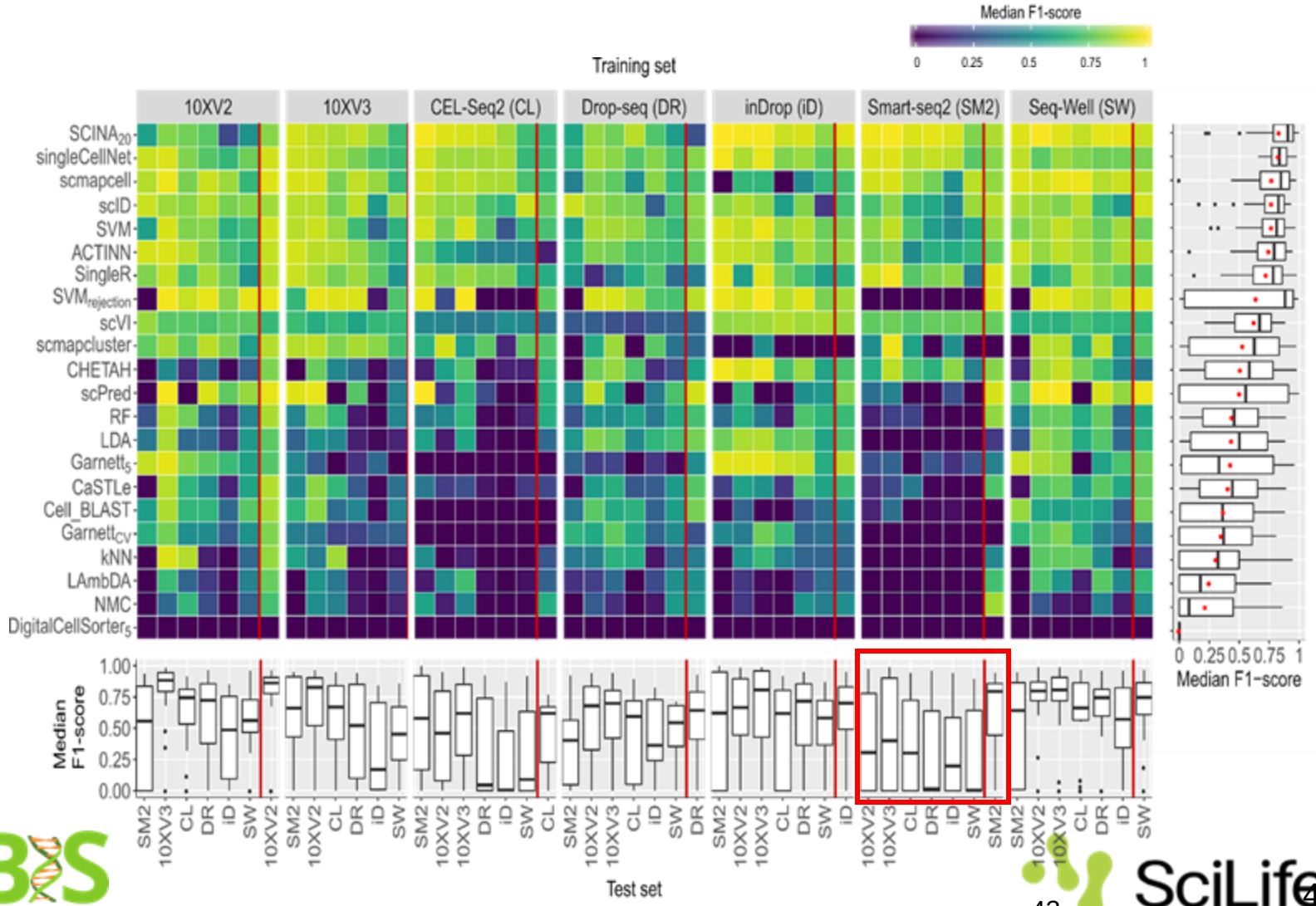
41

Prediction across protocols



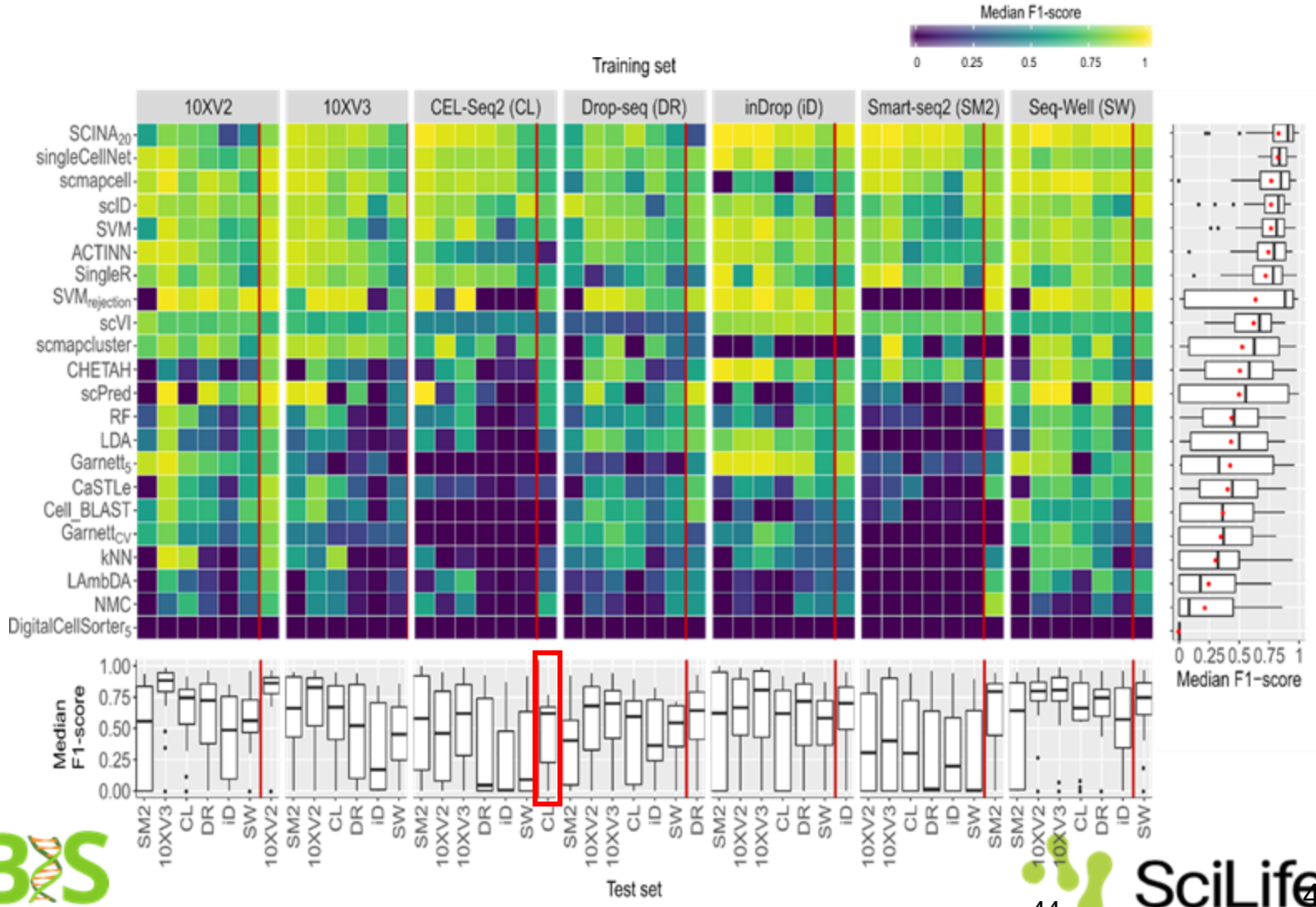
42

Prediction across protocols



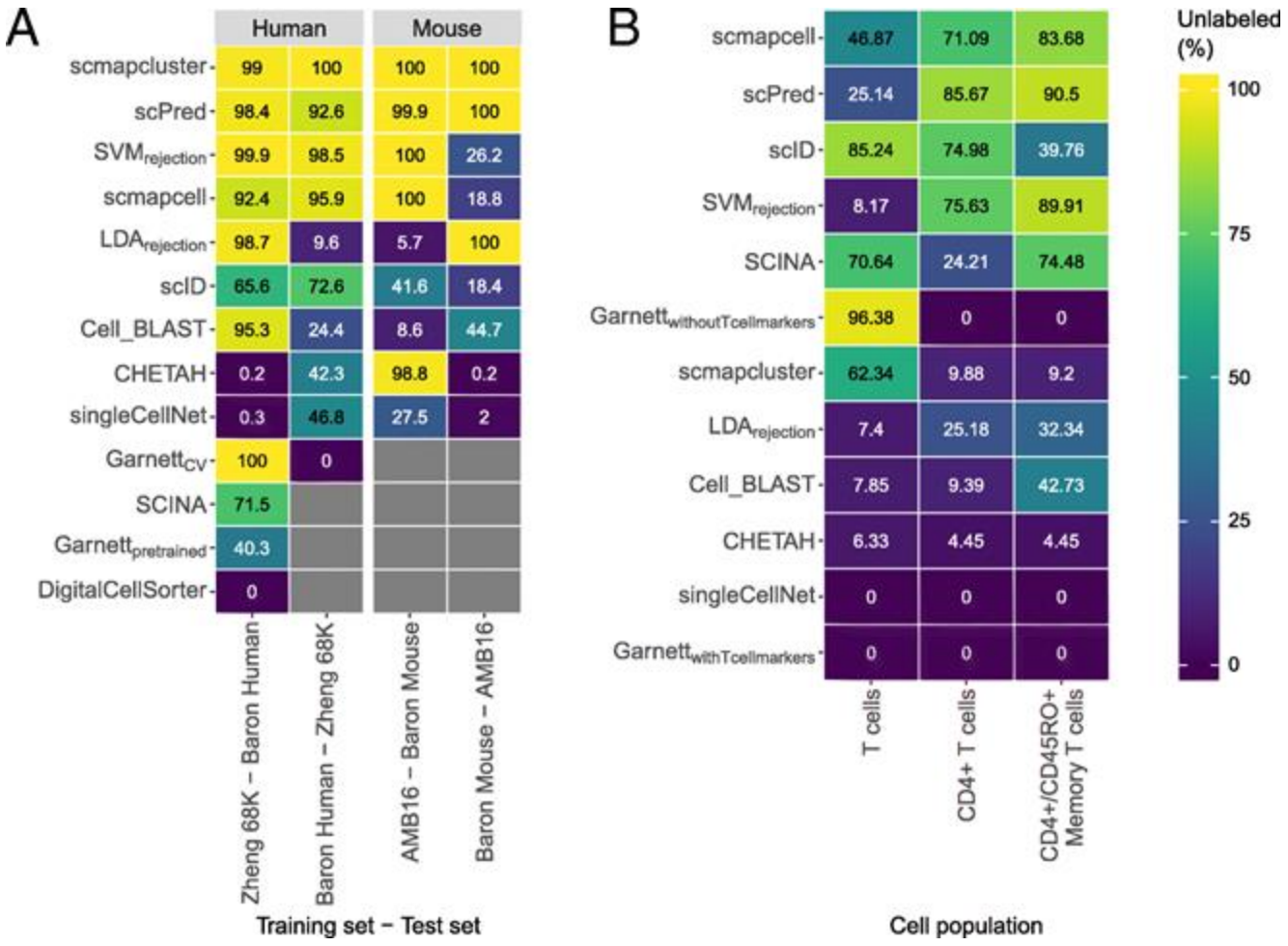
43

Prediction across protocols

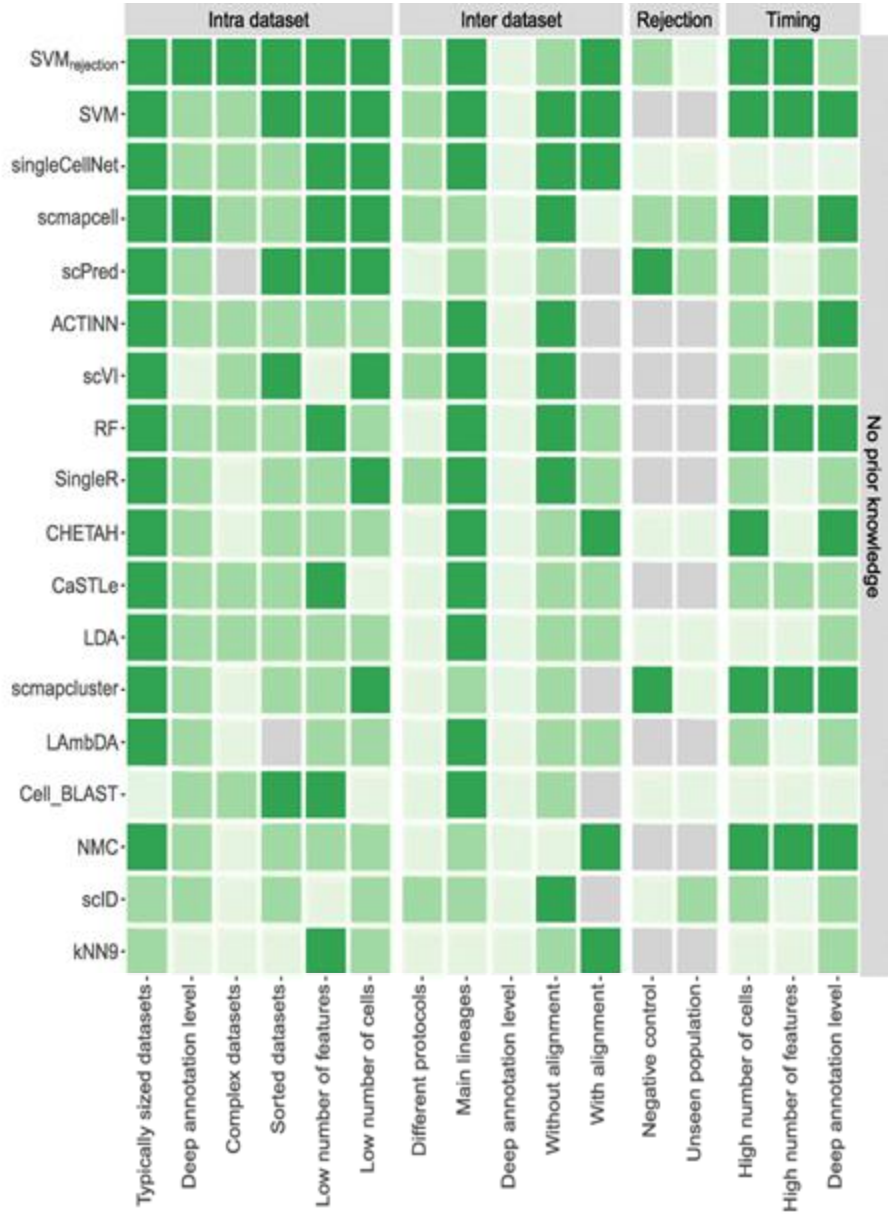


44

Experiment 3: rejection evaluation



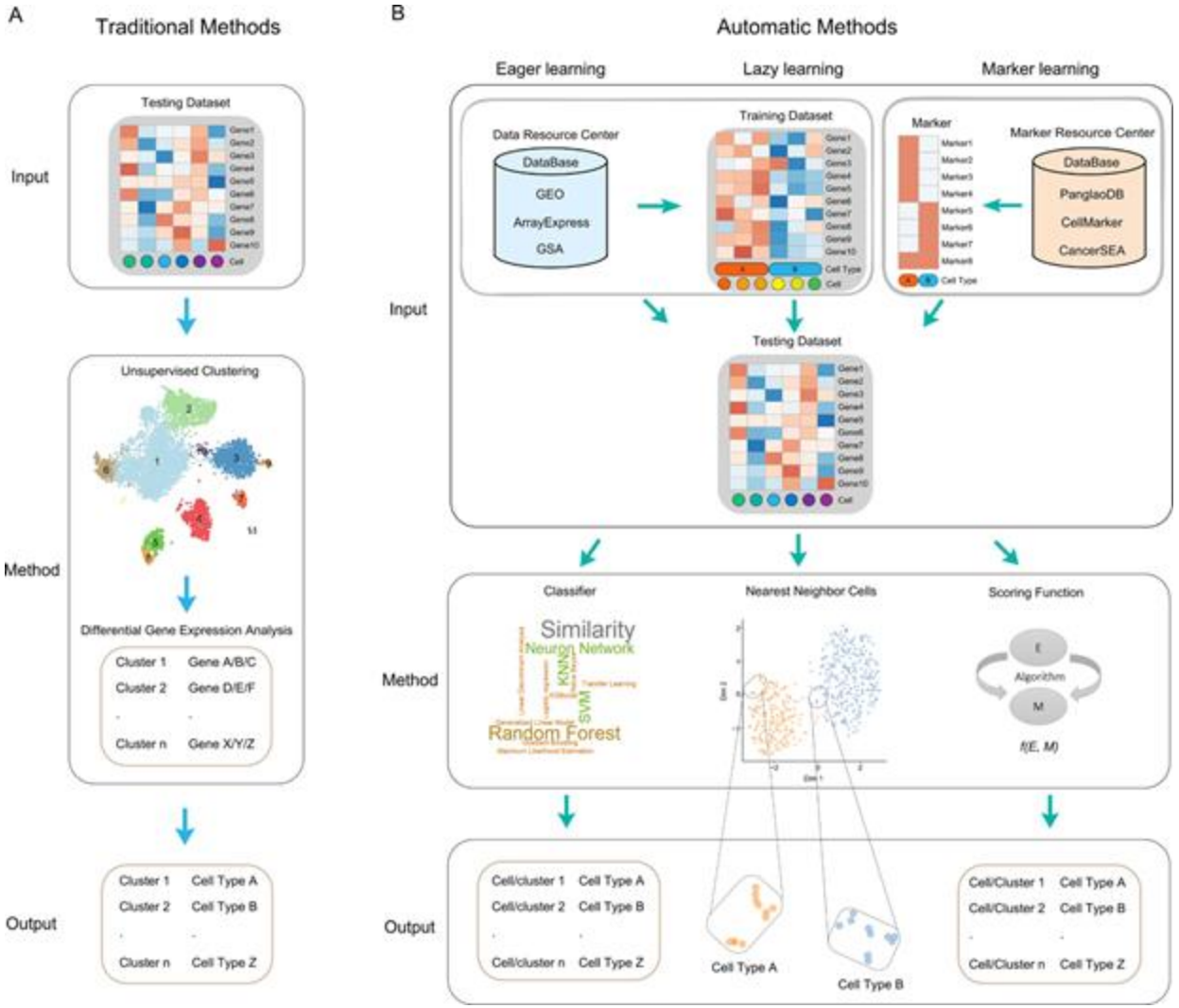
Performance Summary



Conclusions so far

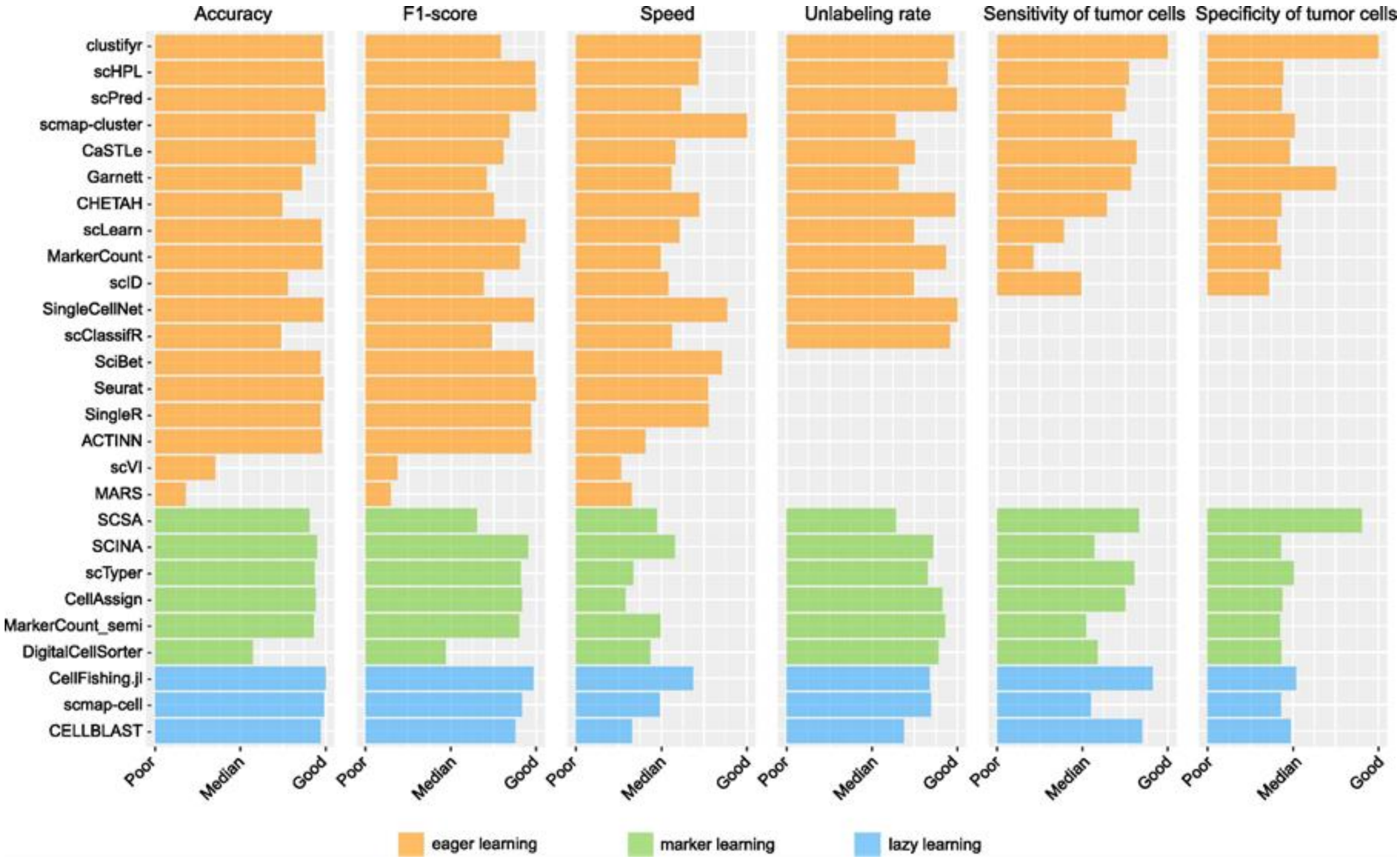
- Simple, off-the-shelf classifiers outperform dedicated single cell methods (see also Köhler et al. bioRxiv 2019)
- Prior-knowledge does not improve performance (highly dependent on selected markers)
- Rejection is difficult
- SnakeMake pipeline: *https://github.com/tabdelaal/scRNAseq_Benchmark/*

Benchmark paper 2021



Xie et al. *Comp. Struct. Biotech J.* 2021

Summary

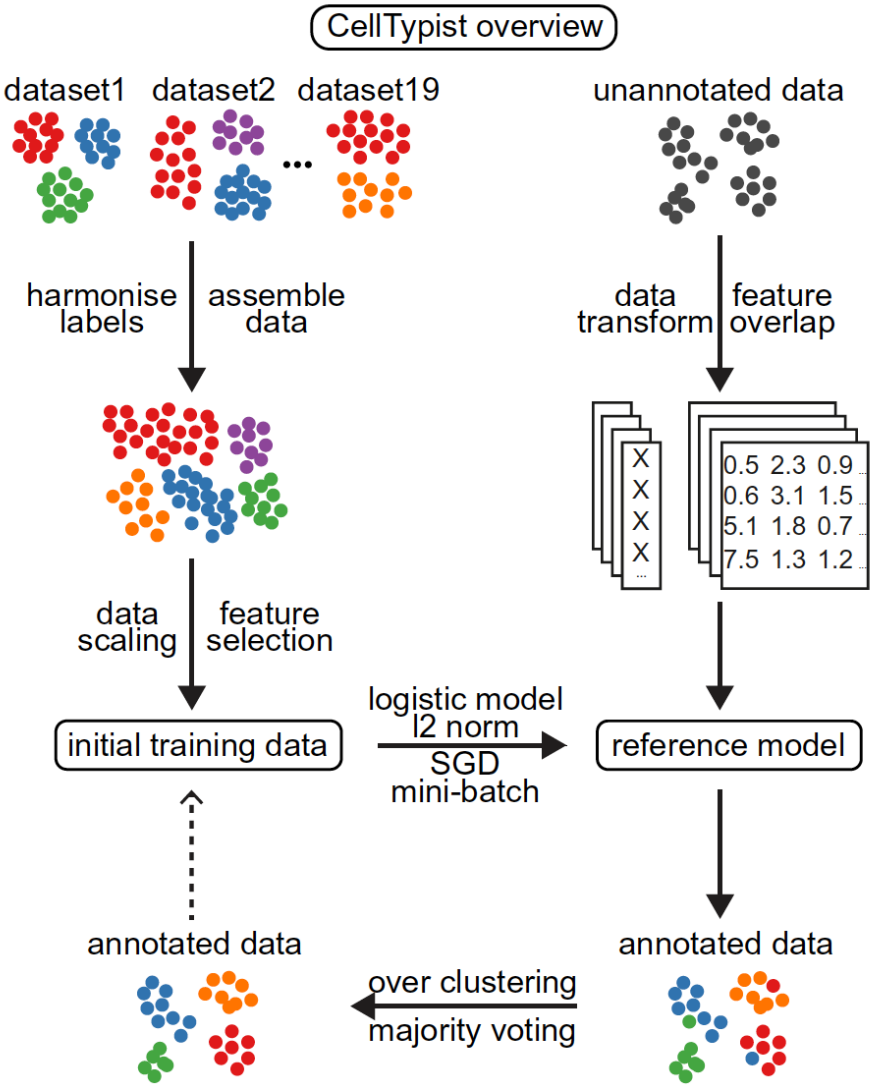


Xie et al. *Comp. Struct. Biotech J.* 2021

More recent methods/approaches?

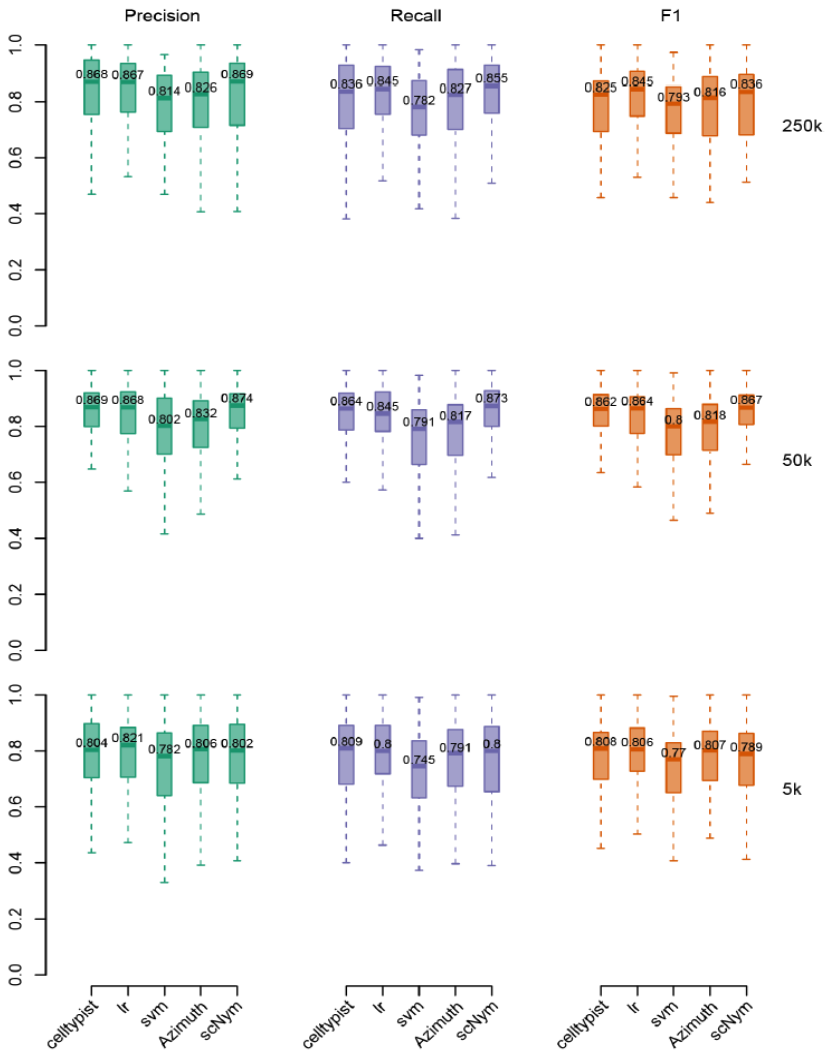
CellTypist

- 20 different tissues from 19 datasets
- Immune cells across different organs
- Logistic regression with stochastic gradient descent learning



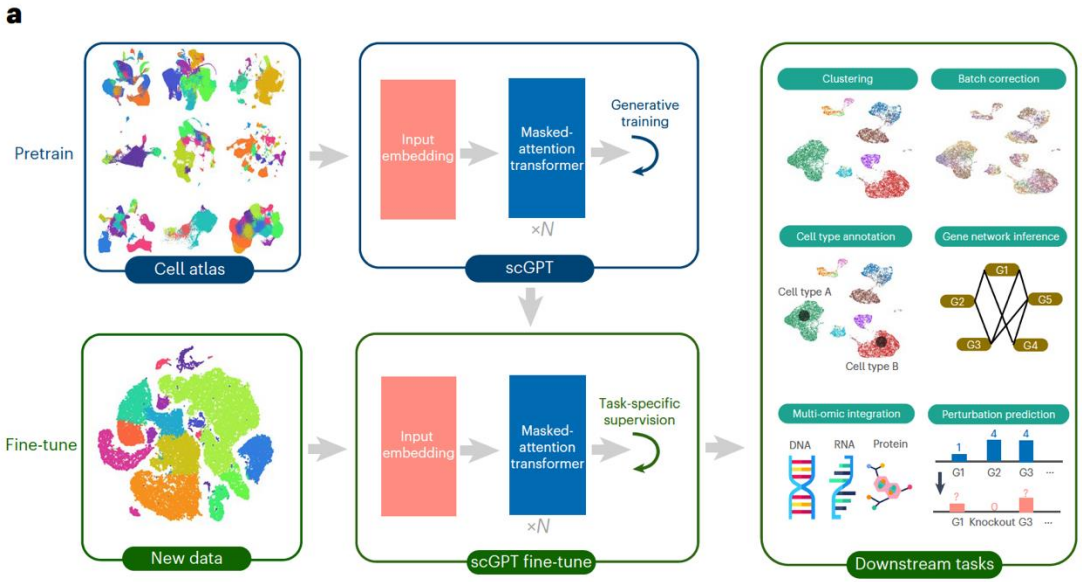
CellTypist

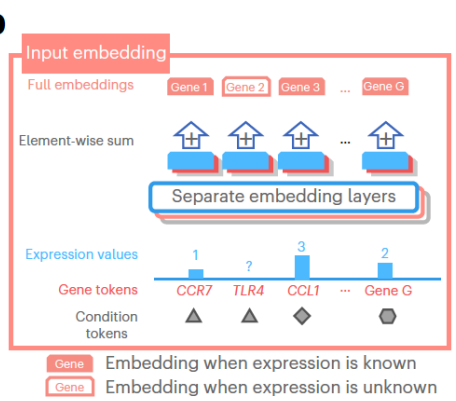
Benchmarking with other label-transferring methods

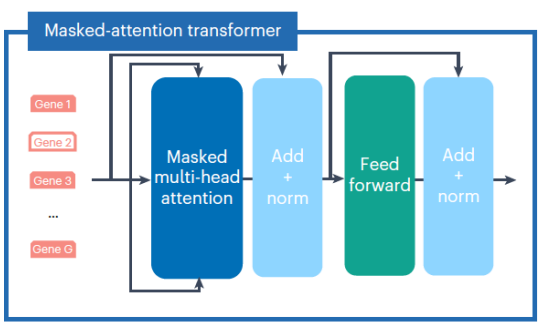


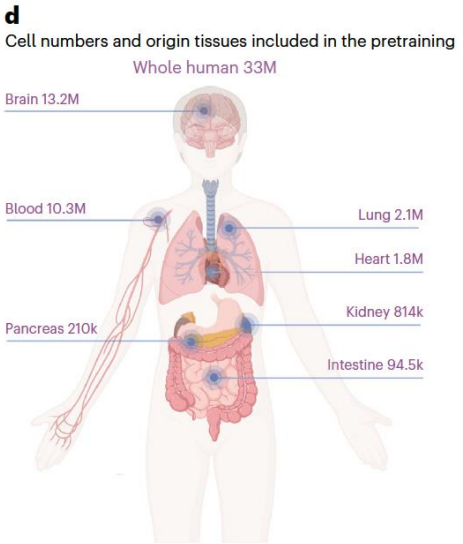
Domínguez Conde et al., *Science 376, eabl5197* 2022

Generative learning is the next big thing? scGPT



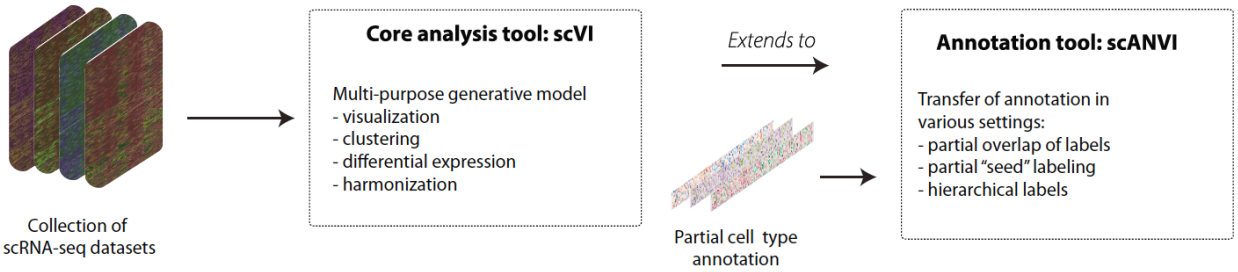


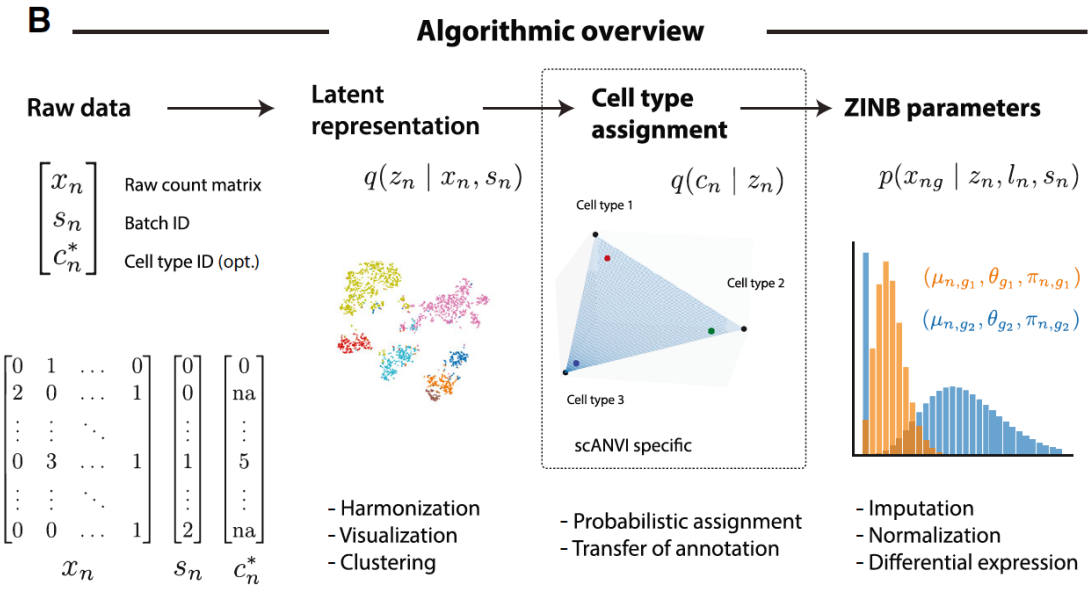


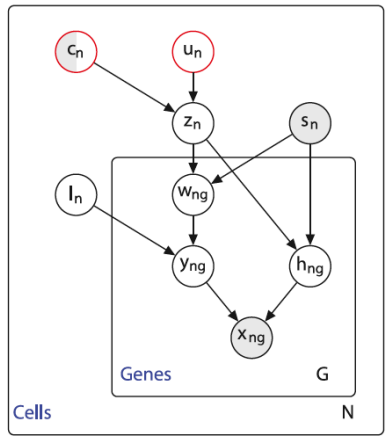


Cue H., et al., *Nature Methods*, 2024

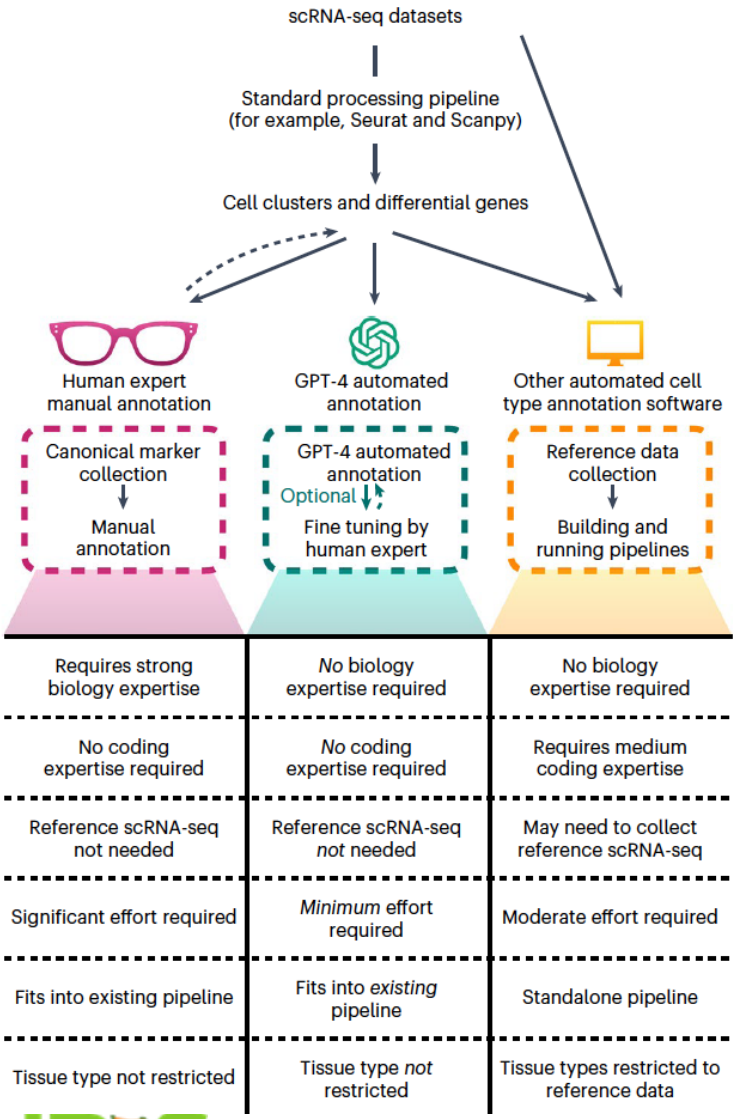
Generative learning is the next big thing? scANVI







GPTCelltype



nature methods

Brief Communication

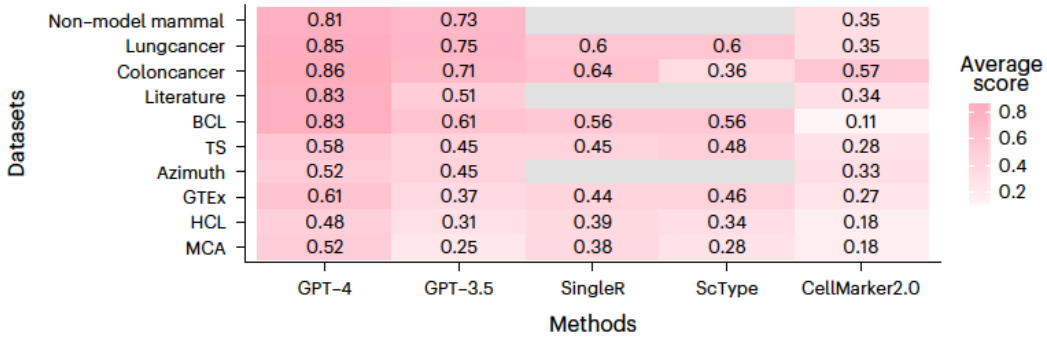
https://doi.org/10.1038/s41592-024-02235-4

Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis

Received: 16 April 2023

Accepted: 5 March 2024

Wenpin Hou [1] & Zhicheng Ji [2]



Challenges of automated cell type annotation?

- There is no ground truth for cell type annotation within a specific dataset
- Biology is complex, and cell states vary continuously
- Delineations between cell types are imprecise
- Cancer cell annotation is even more challenging due to their heterogeneity

➢ It is crucial that annotation methods highlight areas of uncertainty that may require manual scrutiny

➢ Any solution?

popular Vote (popV)

nature genetics

Article

https://doi.org/10.1038/s41588-024-01993-3

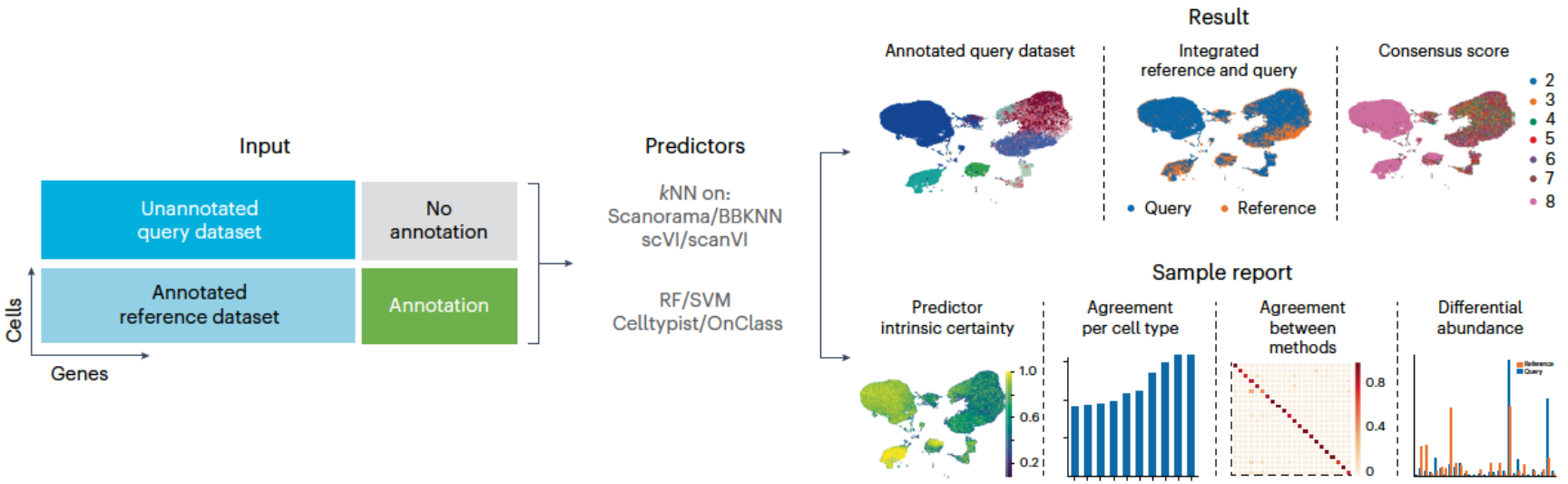
Consensus prediction of cell type labels in single-cell data with popV

Received: 2 September 2023

Accepted: 18 October 2024

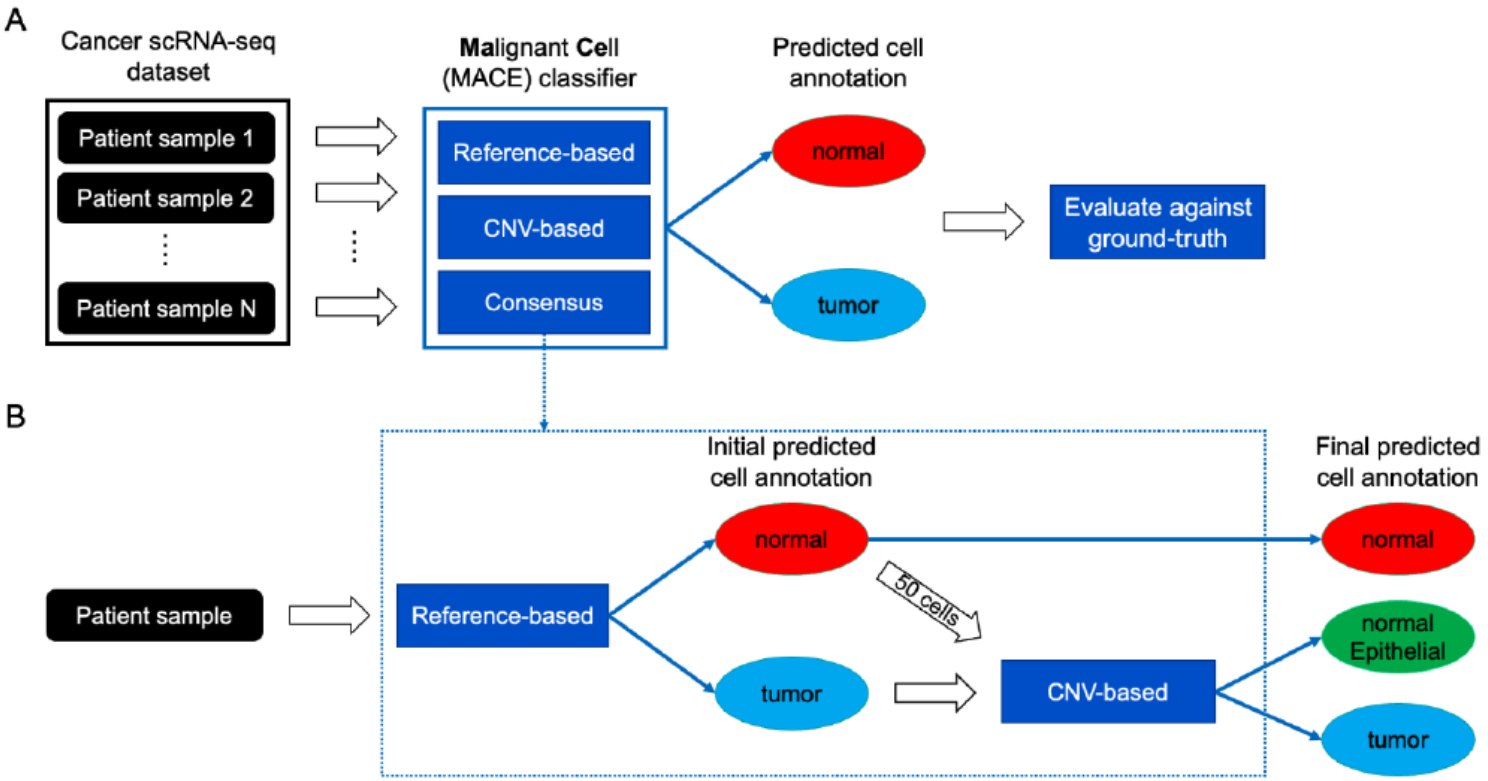
Published online: 20 November 2024

Can Ergen[1,2], Galen Xing[1,3,7], Chenling Xu[1], Martin Kim[2], Michael Jayasuriya[2], Erin McGeever[3], Angela Oliveira Pisco[3,8], Aaron Streets[1,3,4] & Nir Yosef[1,2,5,6]



Malignant Cell (MACE) annotation

- CNV-based methods: *SCEVAN*
- Reference-based method: *scATOMIC*

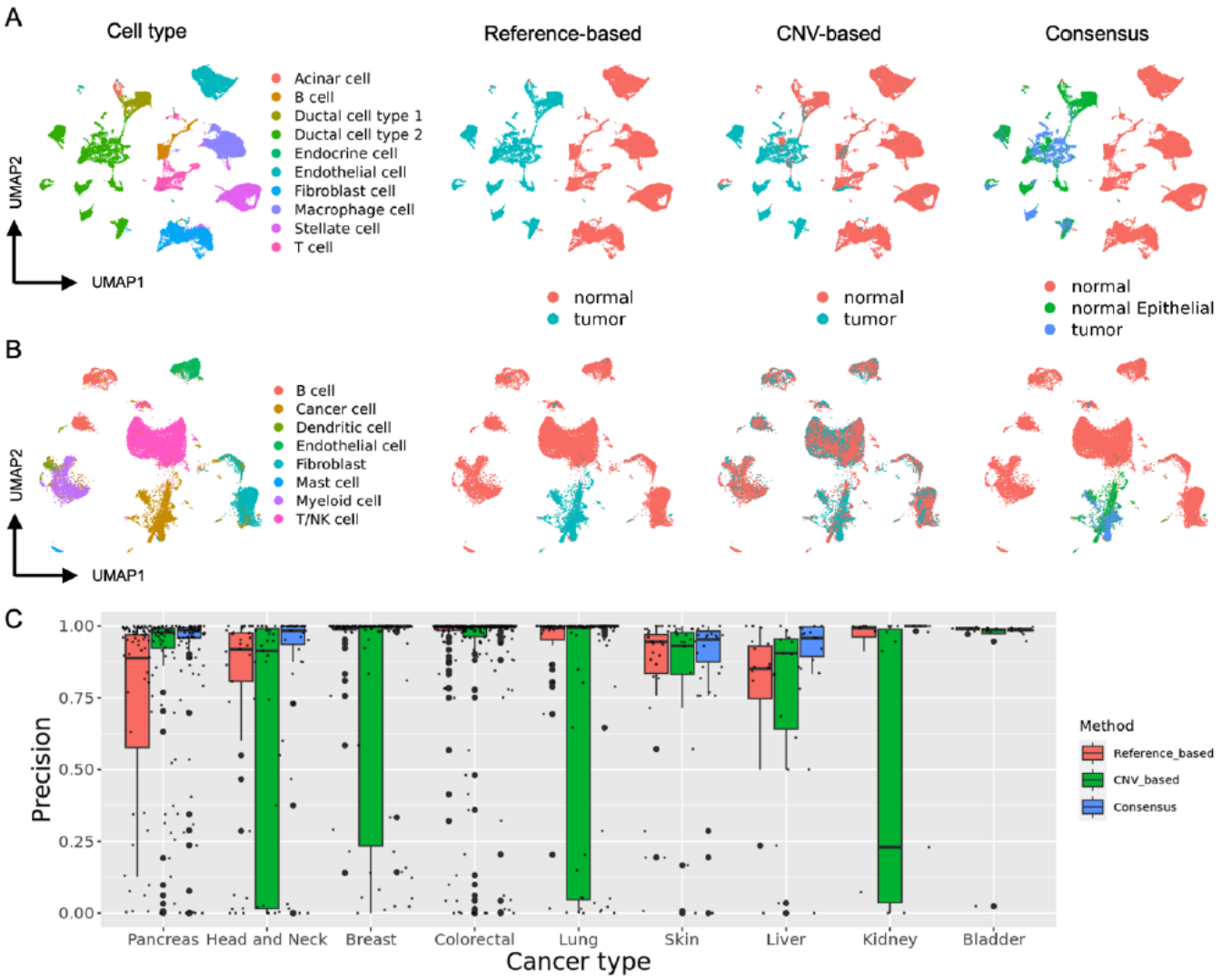


Abdelaal T., et al. *BioRxiv,* 2025

Malignant Cell (MACE) annotation

Peng_et_al_PDAC

Qian_et_al_Breast



Abdelaal T., et al. *BioRxiv,* 2025

Summary

- Cell identification is moving from unsupervised (clustering/visualization) to supervised (classification) learning
- Check what reference you are using!
  - The more similar the reference is to your data - the better the prediction.
  - Same technology matters
  - Do you trust their celltype annotations?
- Atlases do not contain all tissues/celltype and especially not all disease states of cells.
- Also, look at DGE and known markers and check that predictions make sense
- Consensus annotation methods combine the power of other automated methods and can be beneficial when addressing new, unknown cell types or disease samples.

Some useful resources

- Azimuth - Seurat label transfer to reference sets
  - https://azimuth.hubmapconsortium.org/
  - online or R package
- DISCO - CellMapper to several tissues (correlation-based)
  - https://www.immunesinglecell.org/
- Celltypist - Regularised linear models with Stochastic Gradient Descent
  - https://www.celltypist.org/
  - online or python package
- scATOMIC (Random Forest) Pan-cancer TME cell type classifier
  - https://github.com/abelson-lab/scATOMIC

Acknowledgment

Most of the slides were adapted from previous presentations by
Åsa Björklund (NBIS, scilifelab)
Ahmed Mahfouz (Leiden University Medical Center, Björklund TU Delft)

62