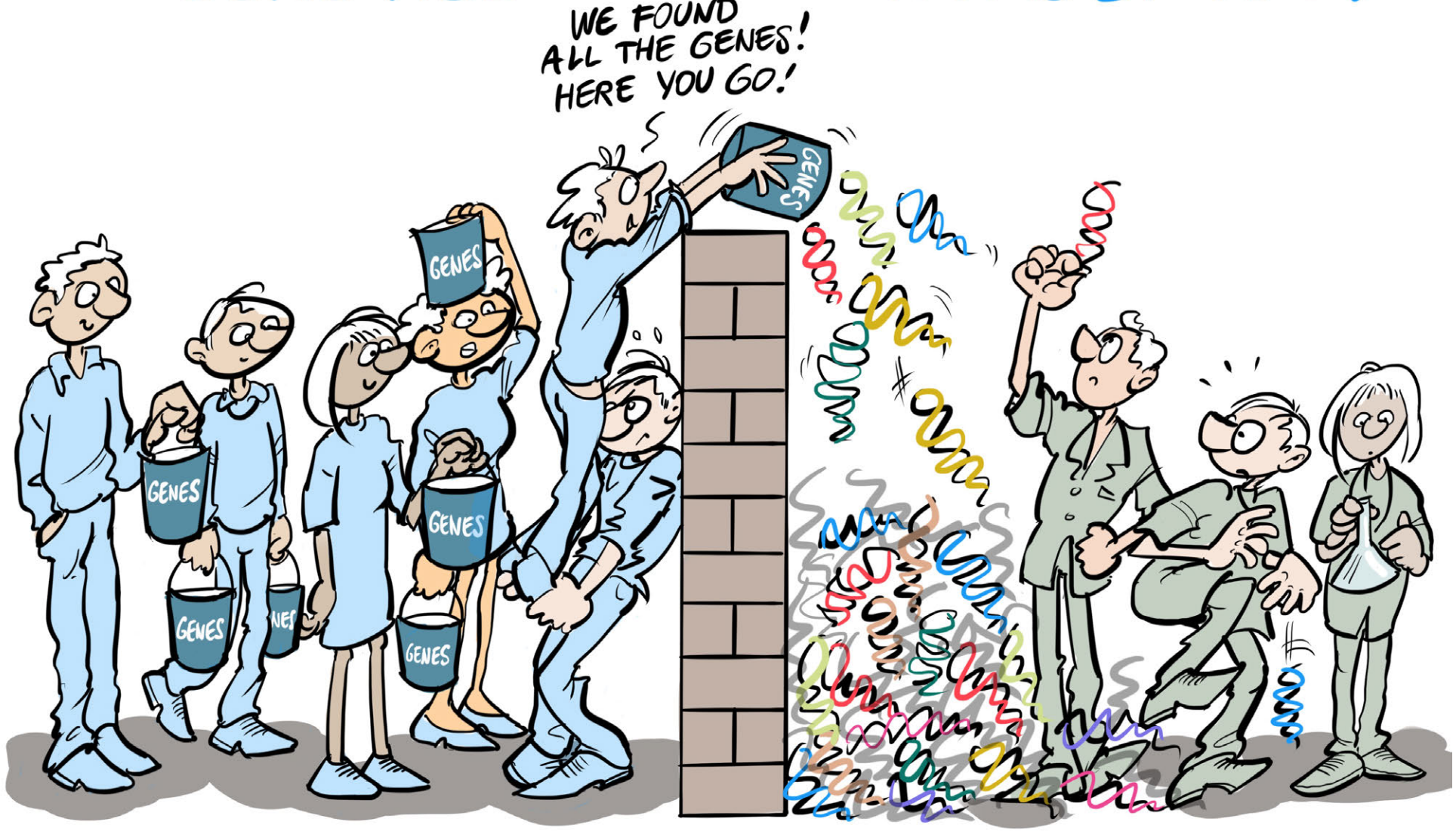
Cell type prediction

Ahmed Mahfouz
Department of Human Genetics, Leiden University Medical Center
Pattern Recognition and Bioinformatics, TU Delft

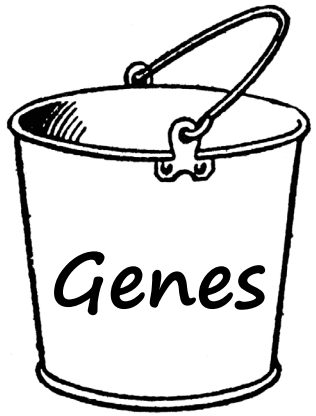
@ahmedElkoussy
mahfouzlab.org

GENETICS NEUROBIOLOGY

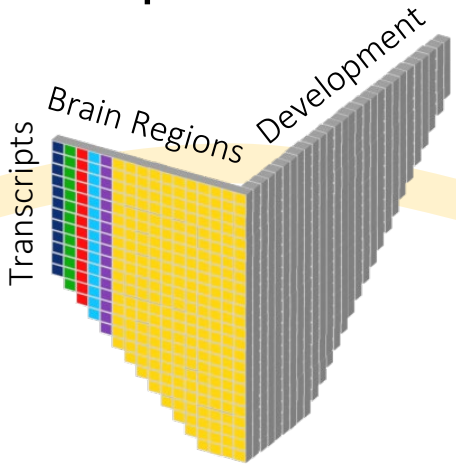


BRAINSCAPES

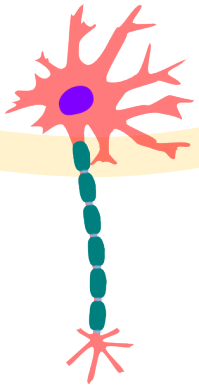
A functional genomics approach

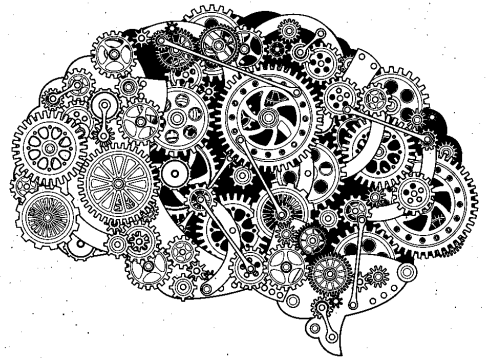


Transcriptome atlases

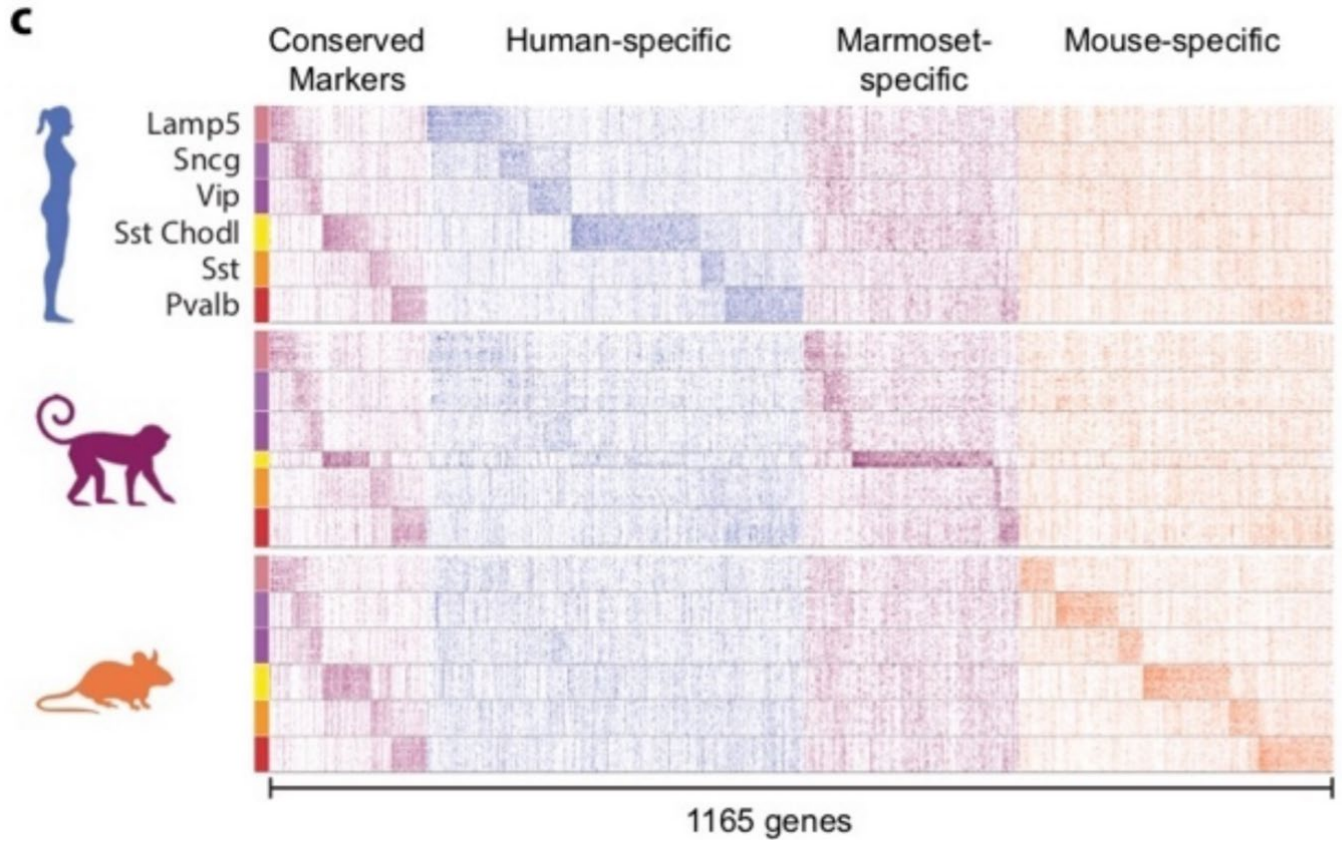


Single cell data



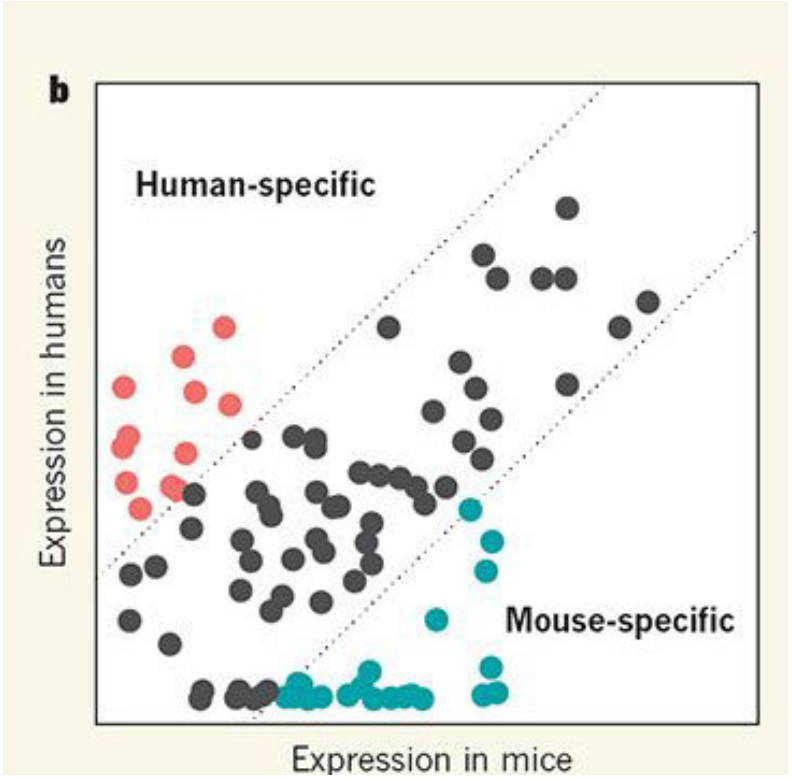


Comparing cells across studies/species/conditions…



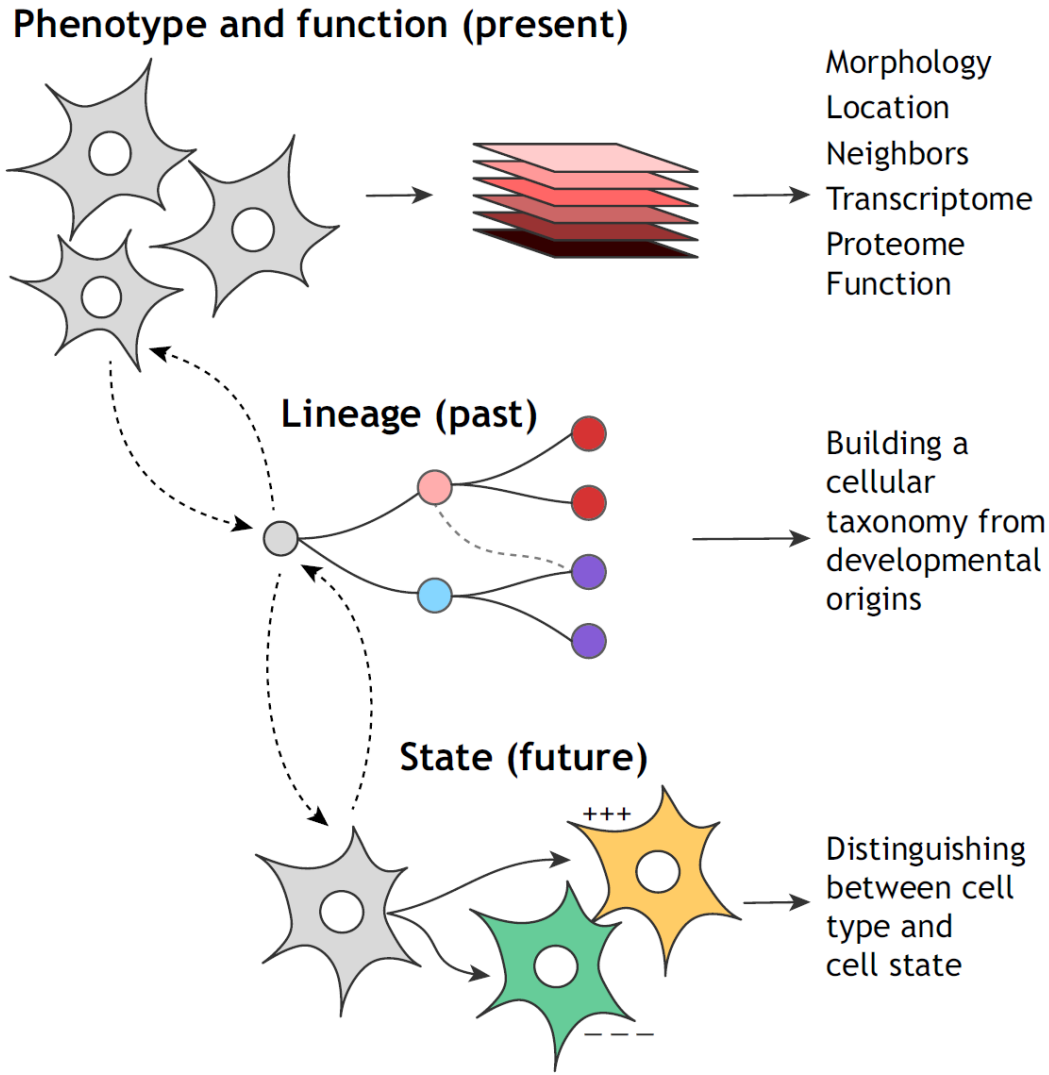
Bakken et al. bioRxiv 2020

Several marker genes are species-specific



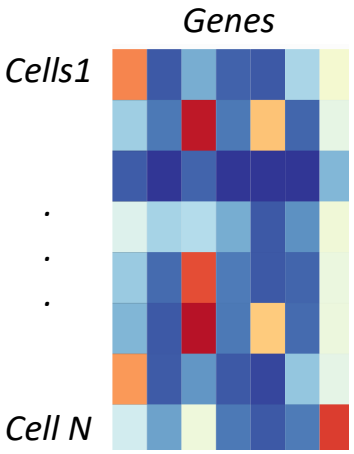
Hodge, Bakken et al. Nature 2019
Keefe & Nowakowski, Nature 2019

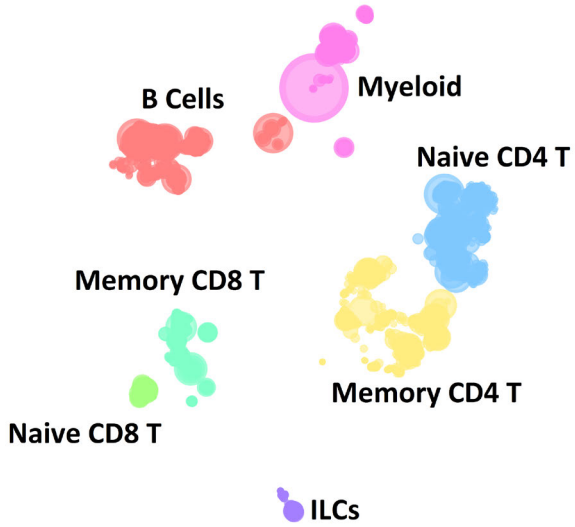
Cell identity



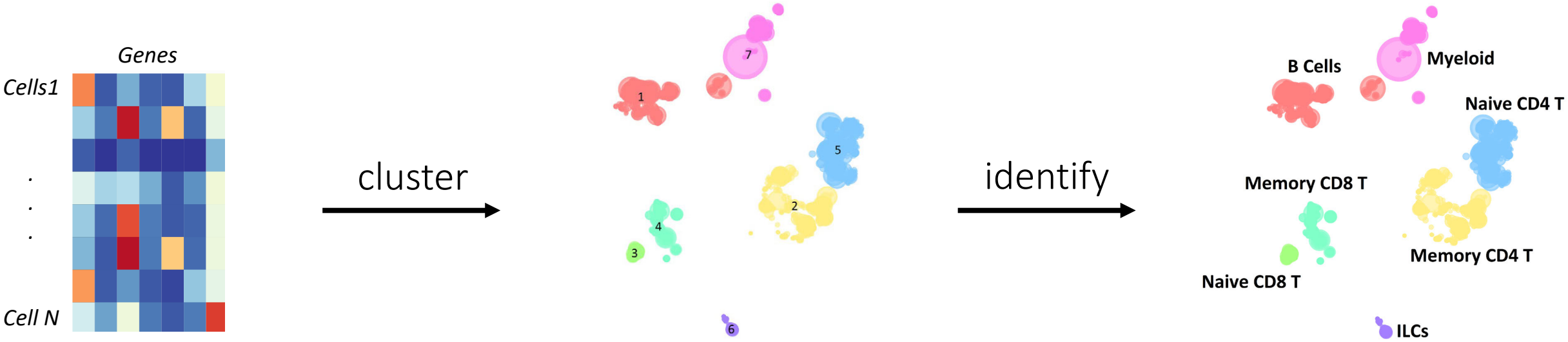
Morris *Development* 2019

How can we identify cell populations?

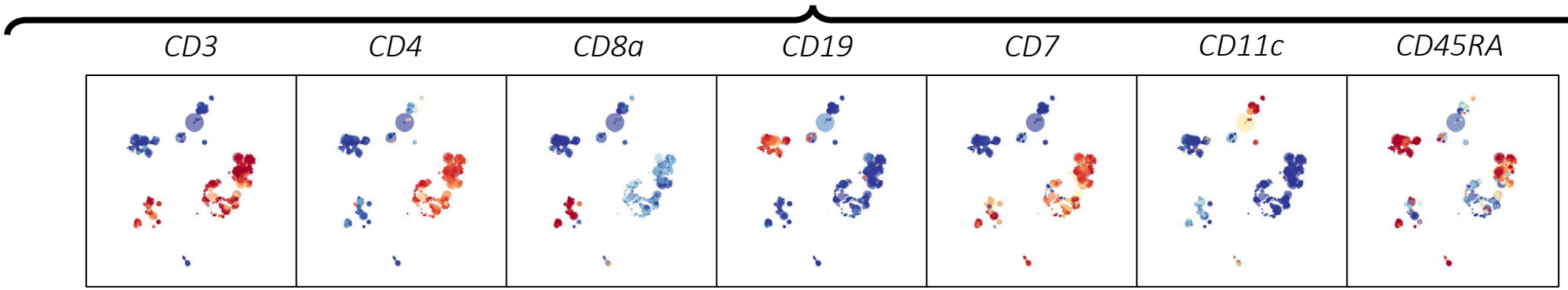




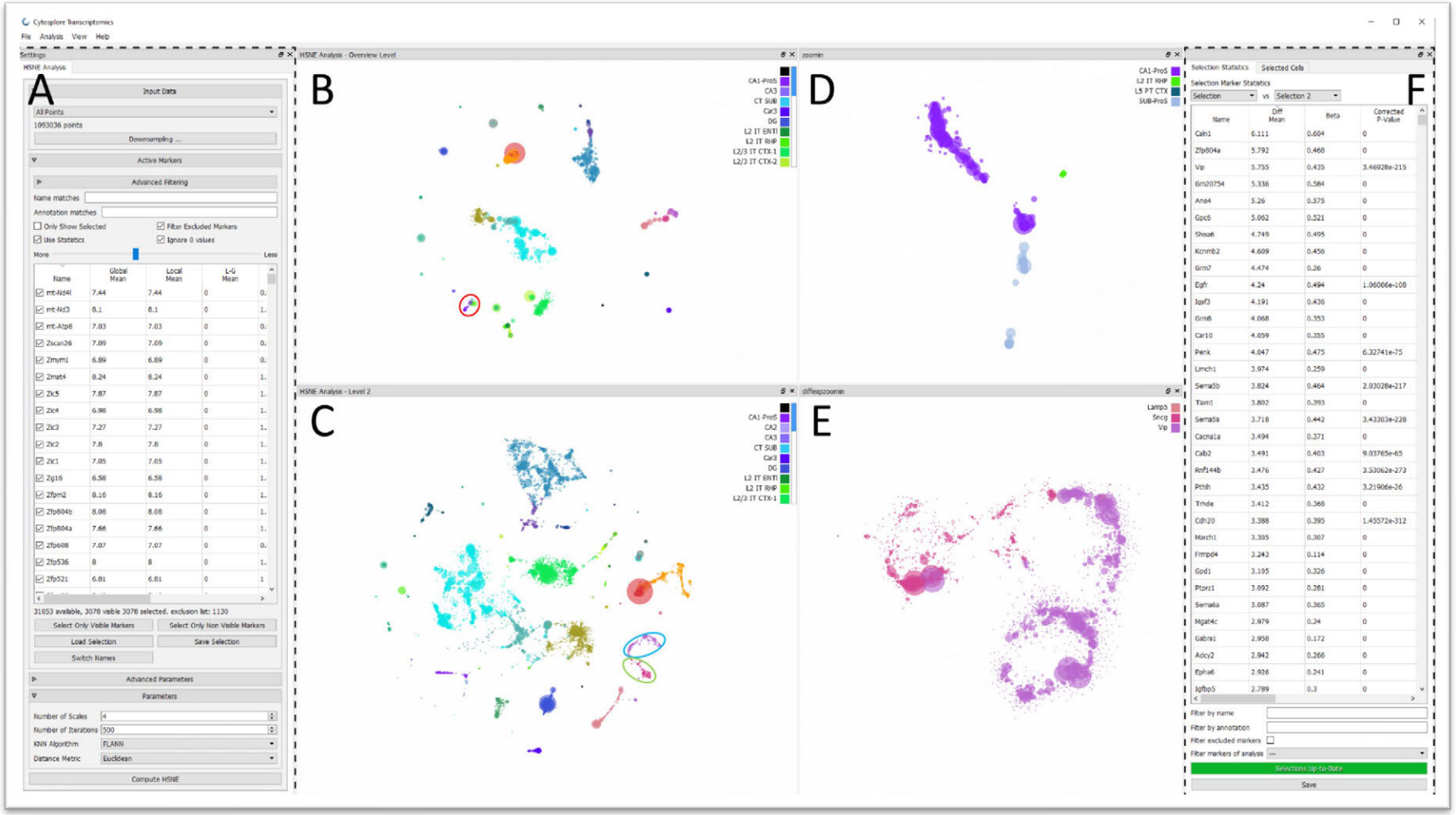
How can we identify cell populations?



marker genes



Cytosplore Transcriptomics

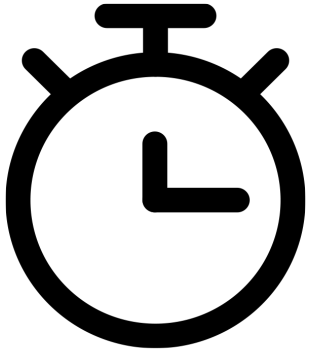


Abdelaal et al., bioRxiv 2020

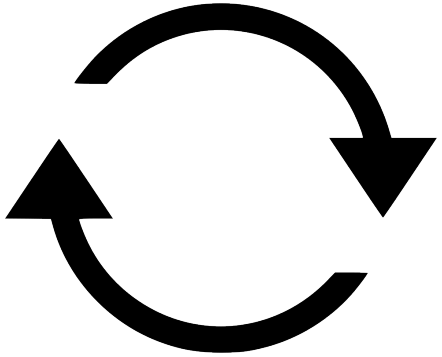
https://transcriptomics.cytosplore.org/

Unsupervised cell identification is problematic

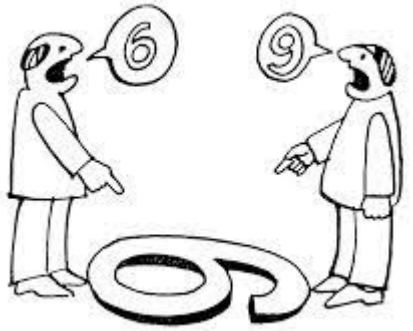
Time consuming



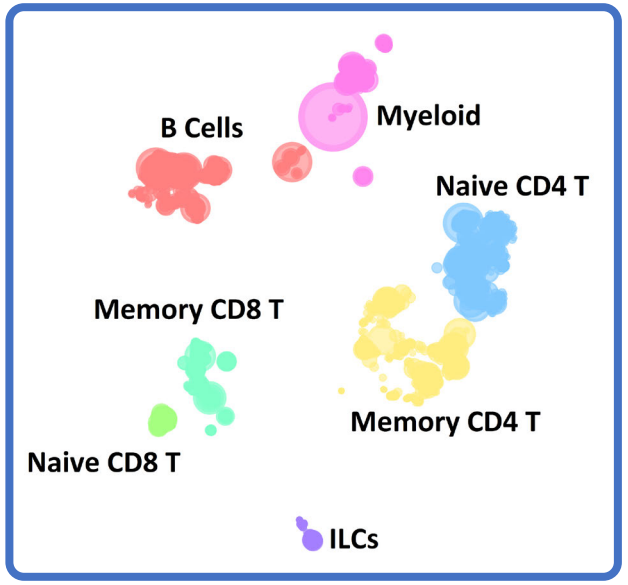
Not reproducible



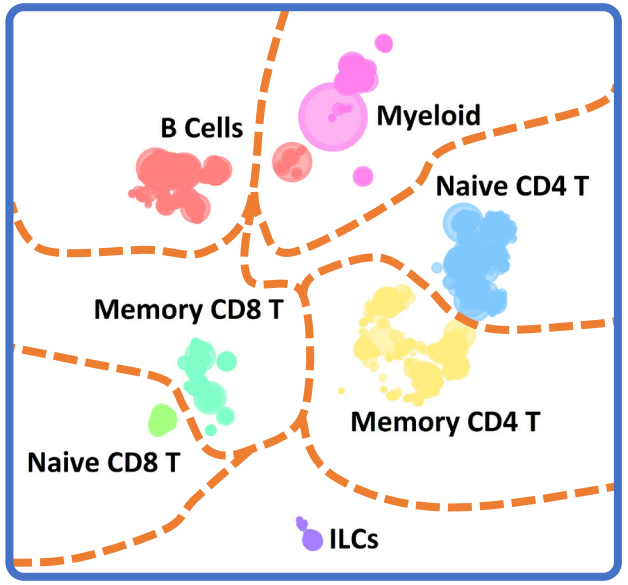
Subjective



Can we automatically identify cell populations?



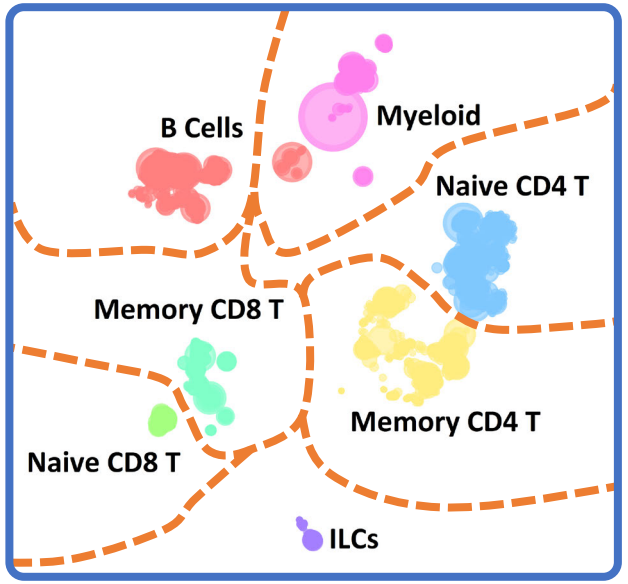
Can we automatically identify cell populations?



Can we automatically identify cell populations?

Clustering

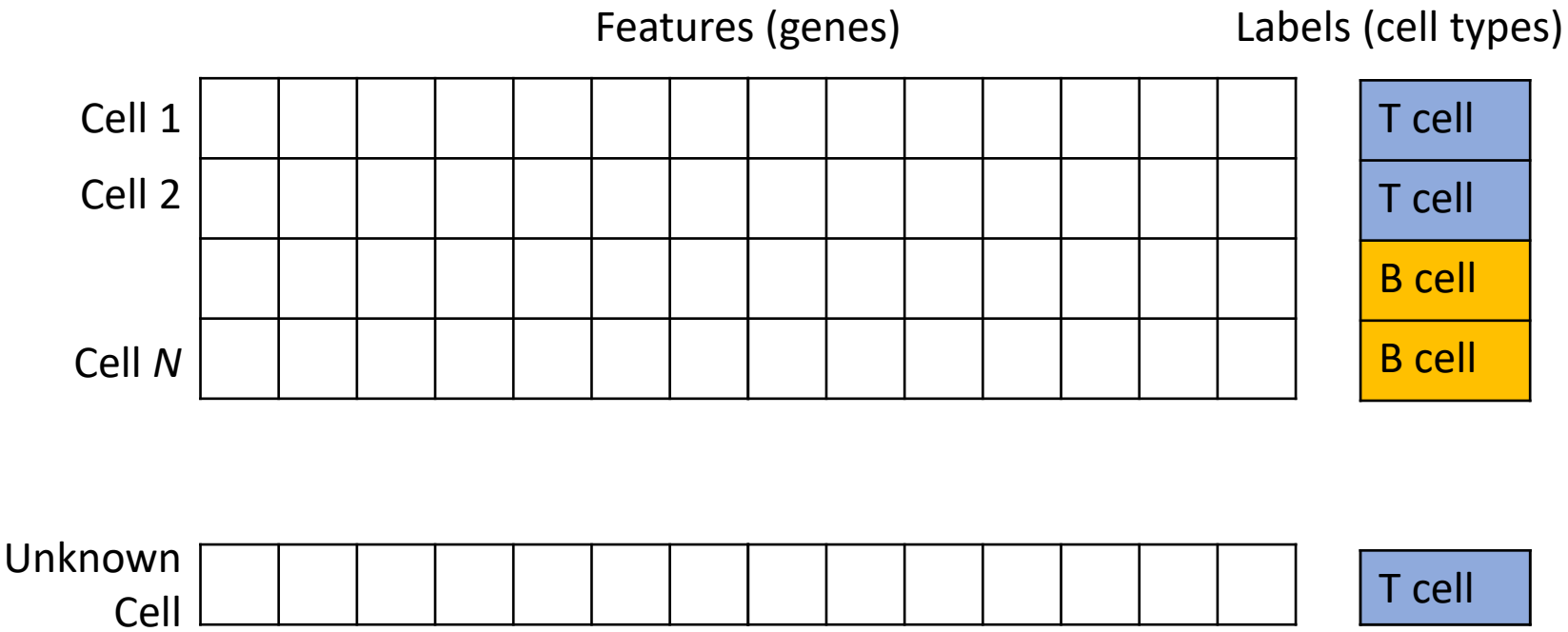
- Unsupervised learning
- Discovering structure/relations
- Clusters are defined by a decision boundary

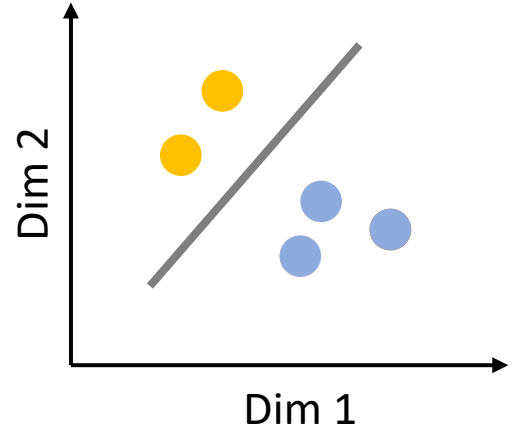


Classification

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

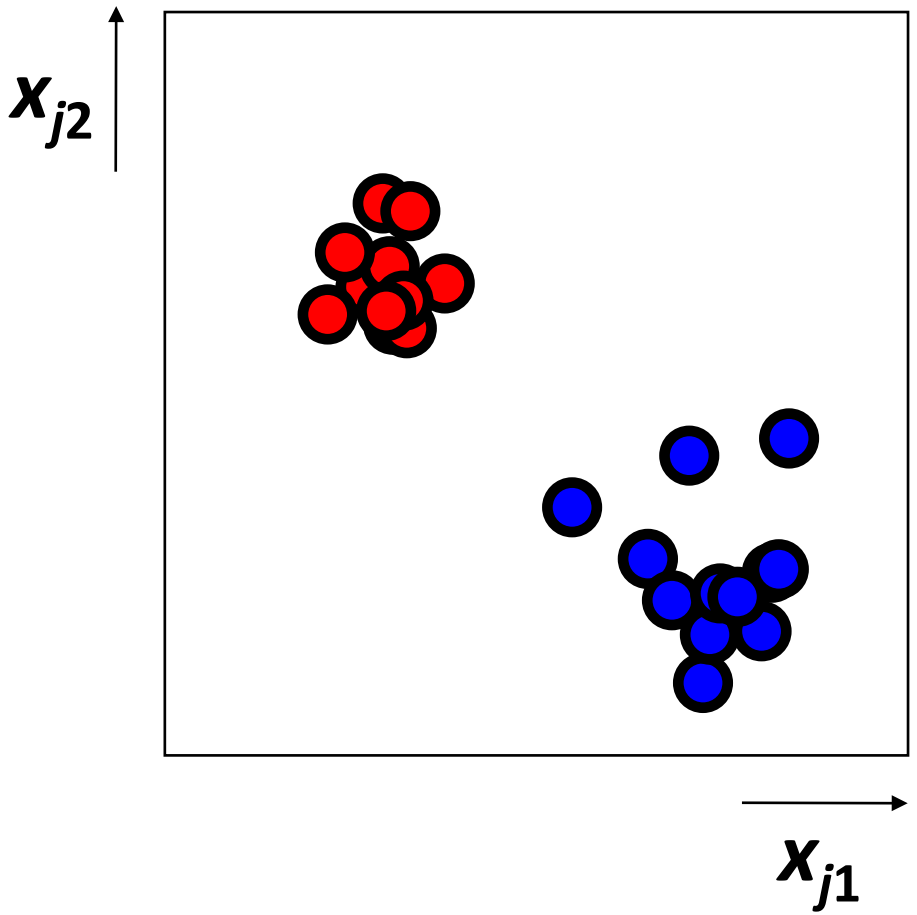
Classification





Classifier training

- Dataset: for j^{th} cell:
  - gene expressions 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

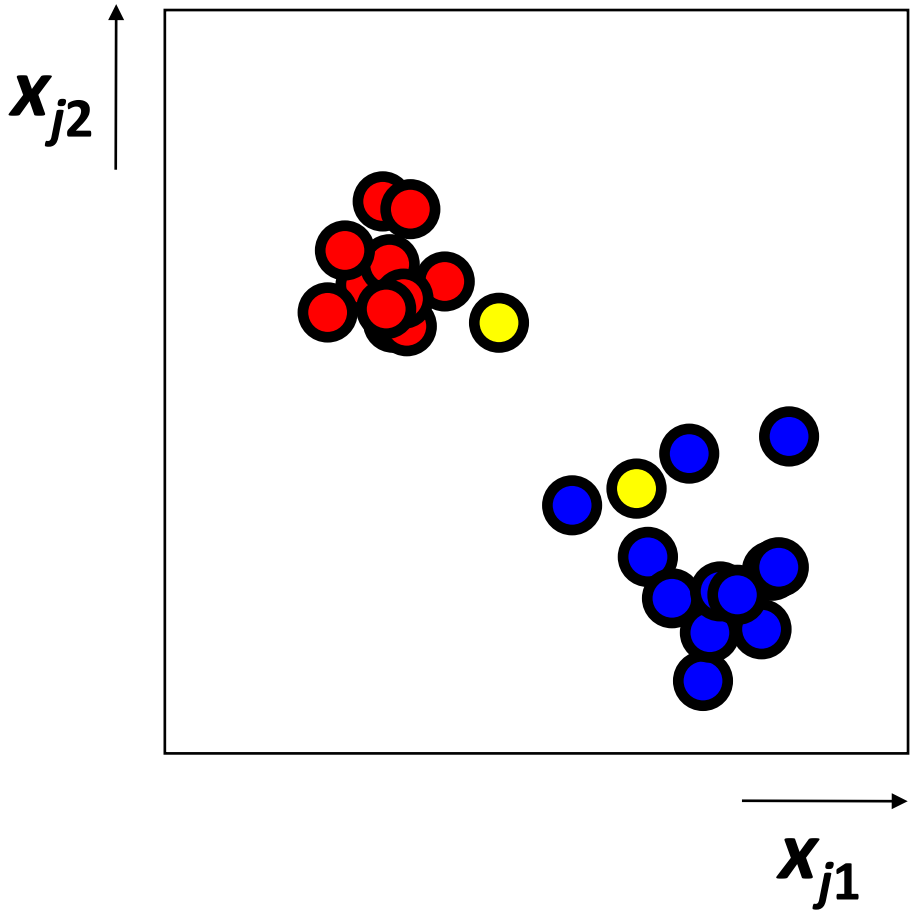


Instance Based Learning (Lazy Classification)

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

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

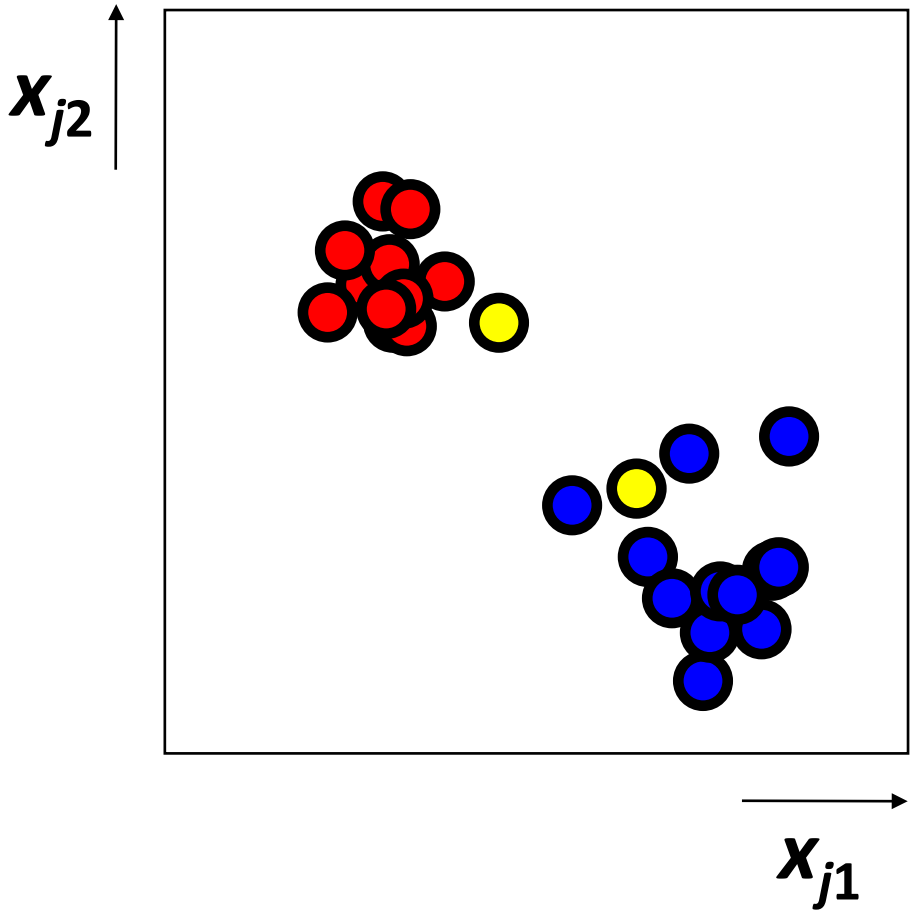
- *No actual training*



Nearest Neighbor (k-NN)

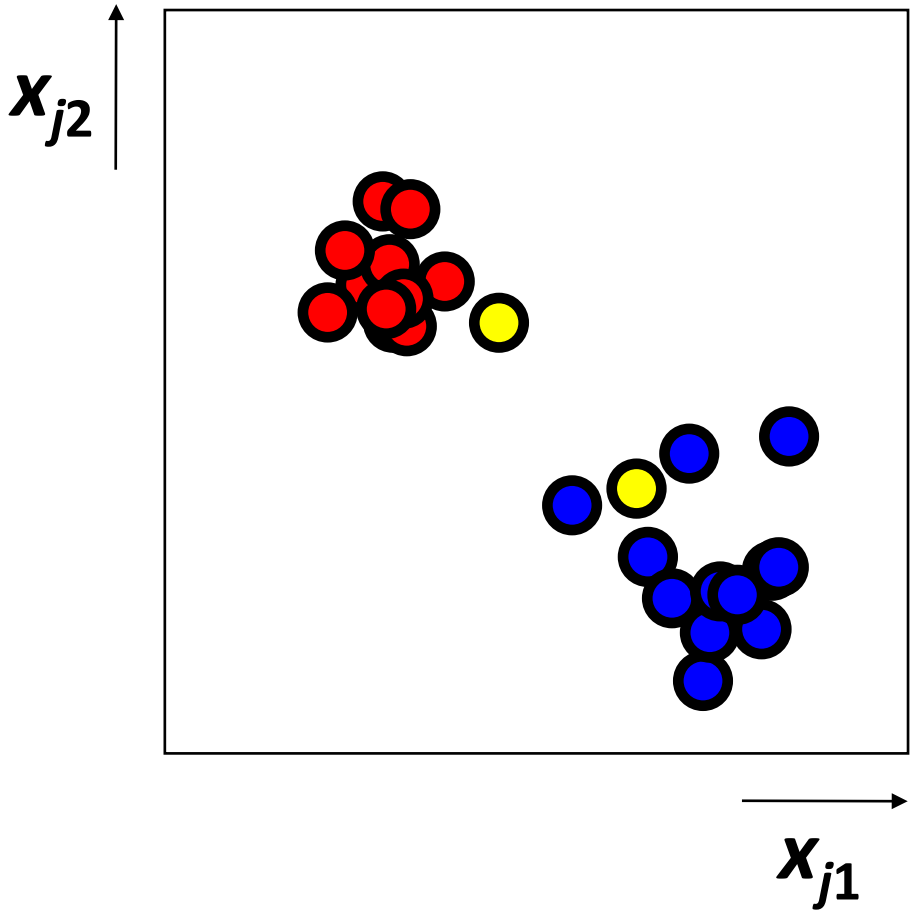
- To make Nearest Neighbor work we need 4 things:

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

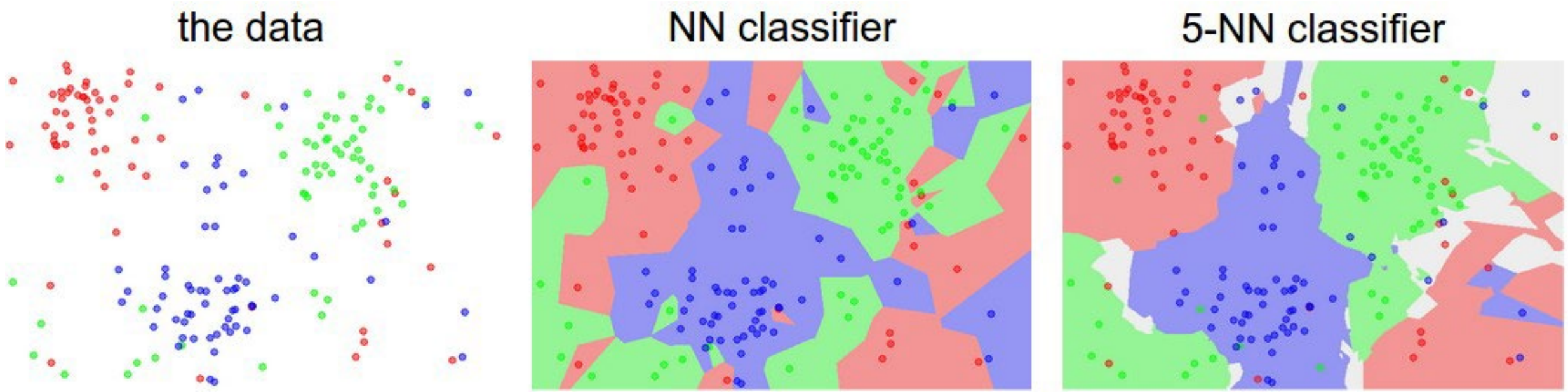


Nearest Neighbor (k-NN)

- 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

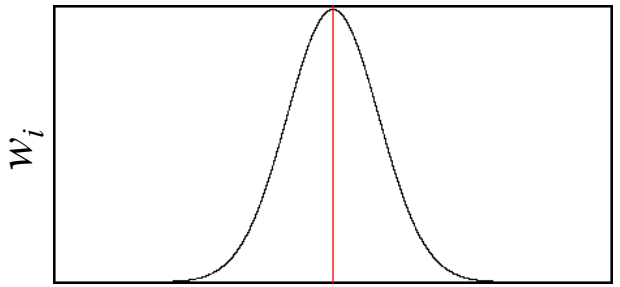


Effect of *k*

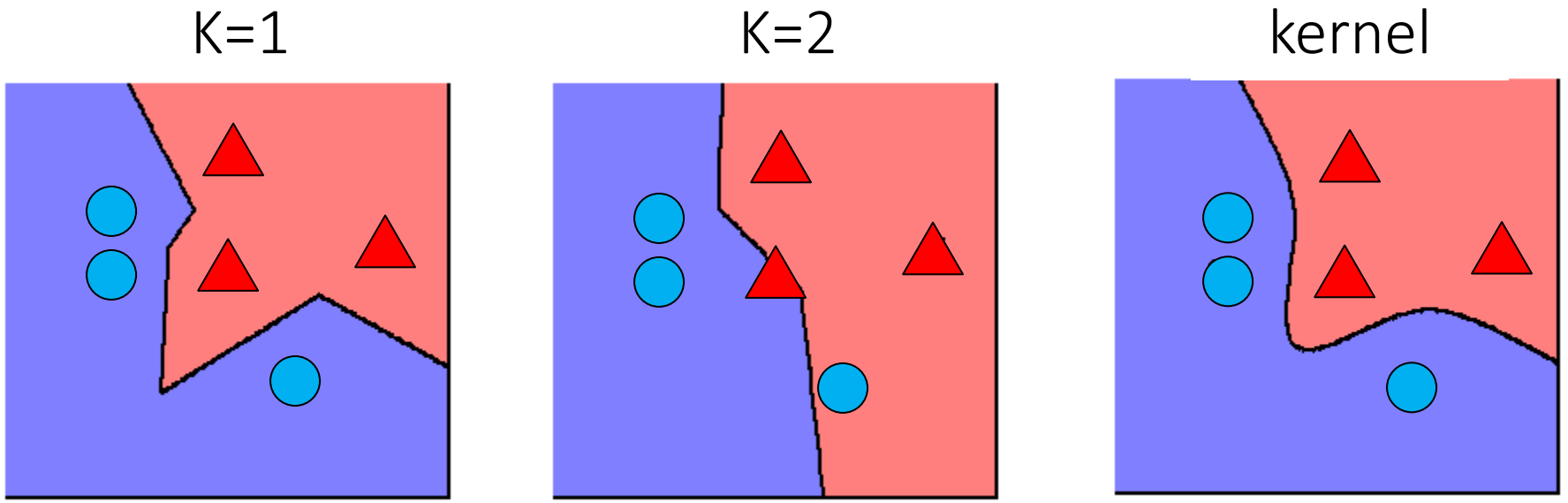


Weighted Nearest Neighbor
(kernel regression)

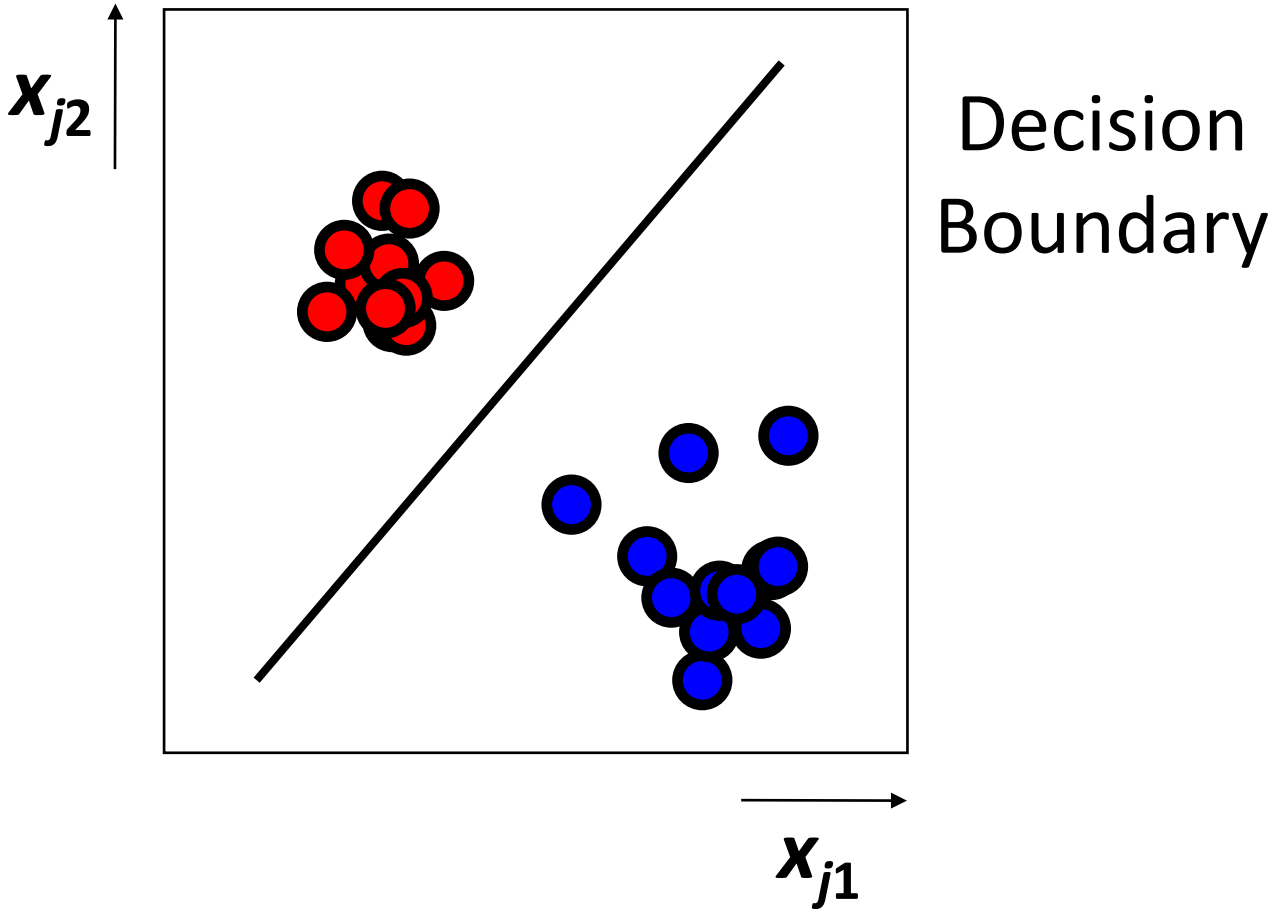
- Distance metric:
  - Euclidean
- How many neighbors to look at?
  - All of them (!)
- Weighting function:
  - $w_i = \exp(-\frac{d(x_i,q)^2}{K_w})$
  - Nearby points to query q are weighted more strongly. K_w:kernel width.
- How to fit with the local points?
  - Predict weighted average: $\frac{\sum_i w_i y_i}{\sum_i w_i}$



Comparison: K=1, K=2, kernel

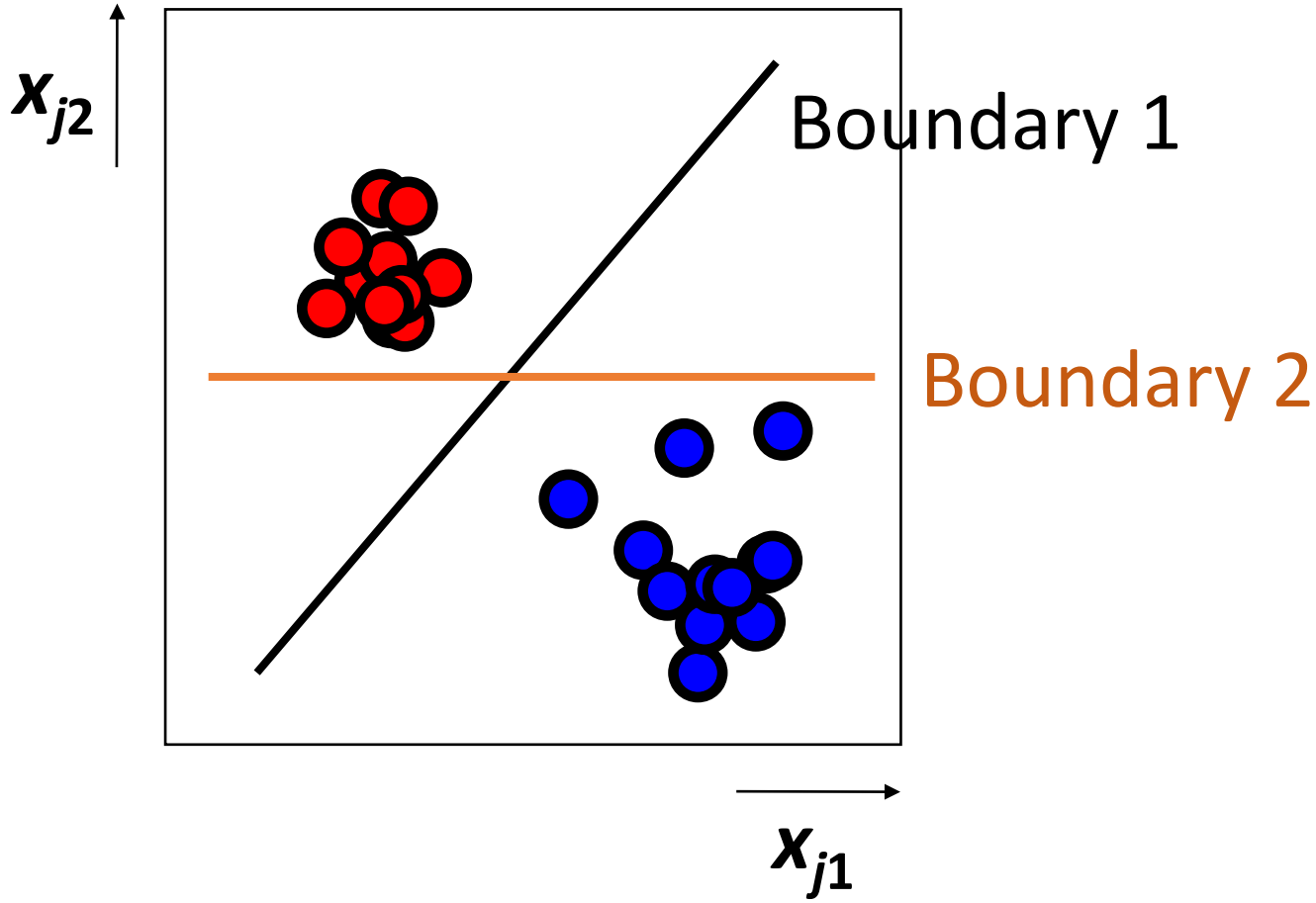


Support Vector Machine (SVM)



Support Vector Machine (SVM)

Which boundary is better?

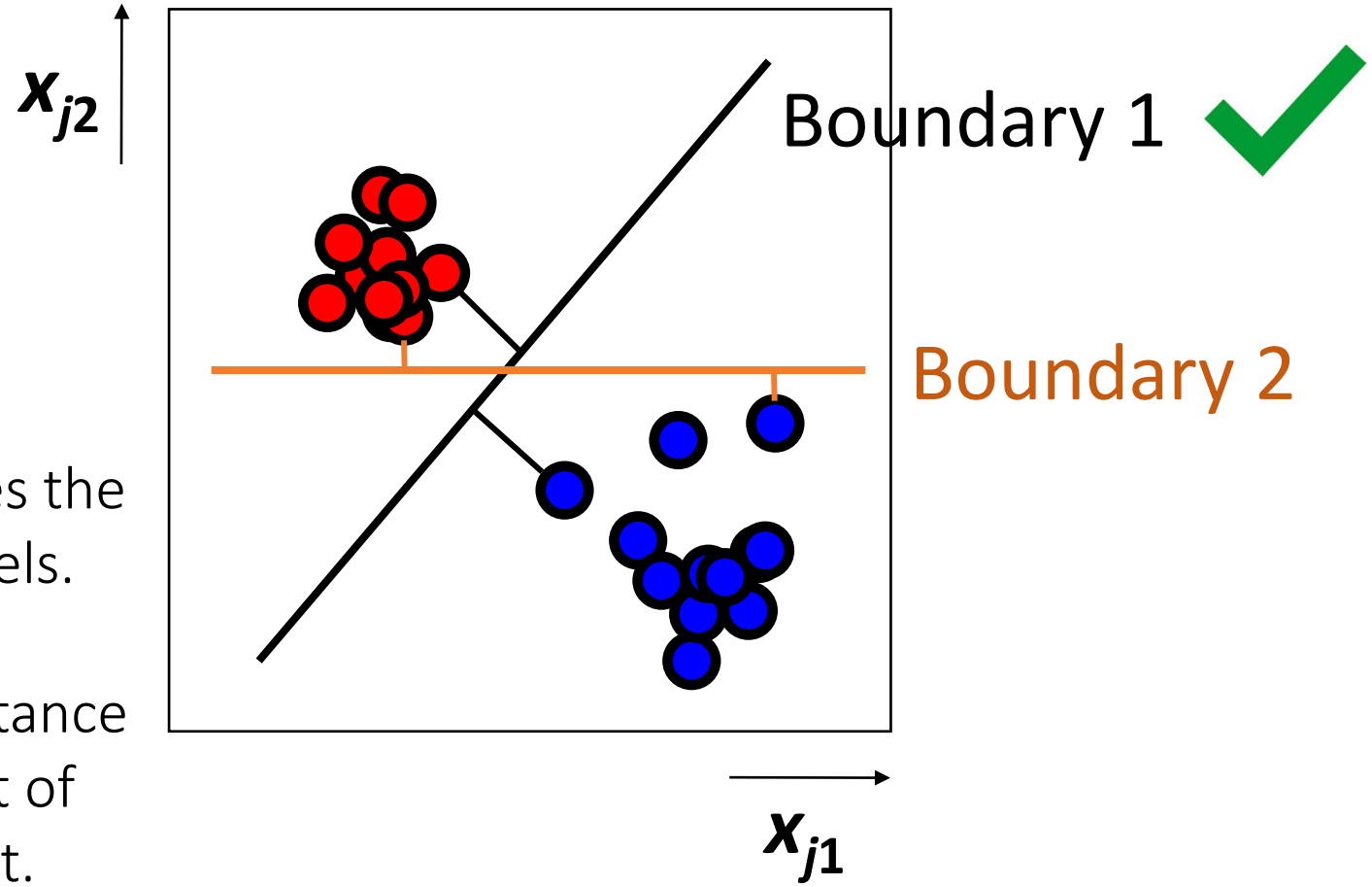


Support Vector Machine (SVM)

Which boundary is better?

The one that maximizes the margins from both labels.

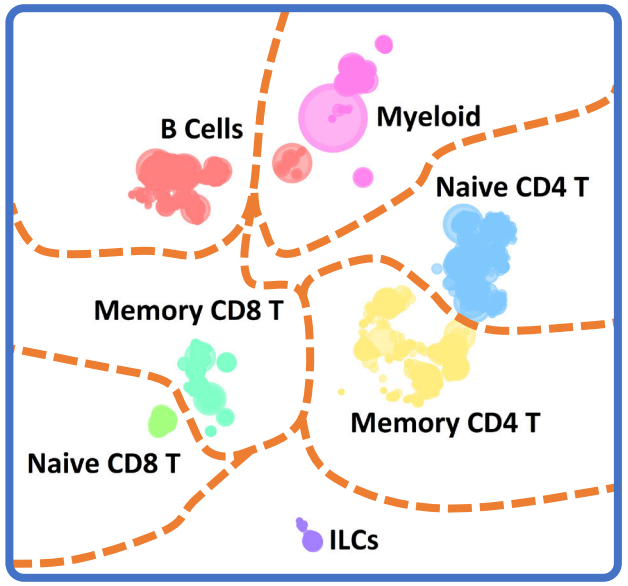
i.e. The one whose distance to the nearest element of each label is the largest.



Can we automatically identify cell populations?

Training data

Annotated Cells
(e.g. atlas)



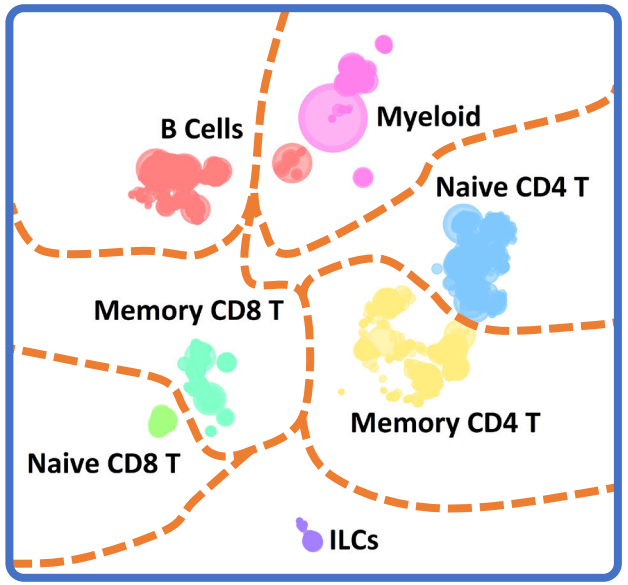
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



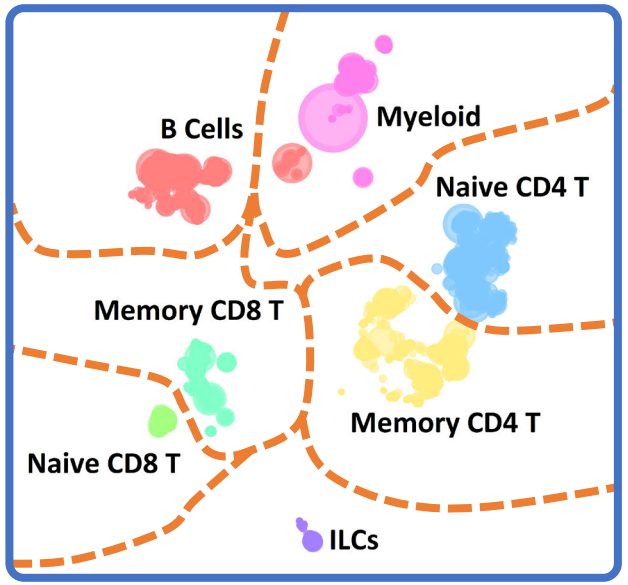
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



Annotated cells

Test data

Unannotated cells

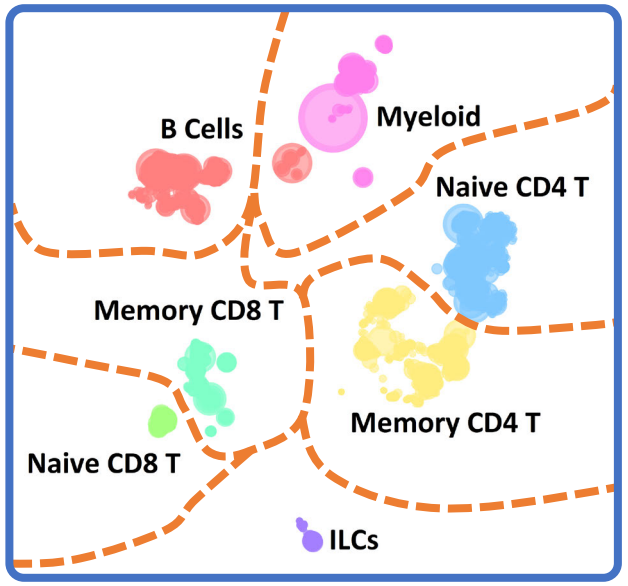
Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



Test data

Unannotated cells

Annotated cells

Unknown (novel) cells

Rejection option

16 existing classifiers (April 2019)

CaSTLe

scPred

LAmbDA

$scmap_{cluster}$

Moana

SingleCellNet

CHETAH

SingleR

Garnett

$scmap_{cell}$

SCINA

DigitalCellSorter

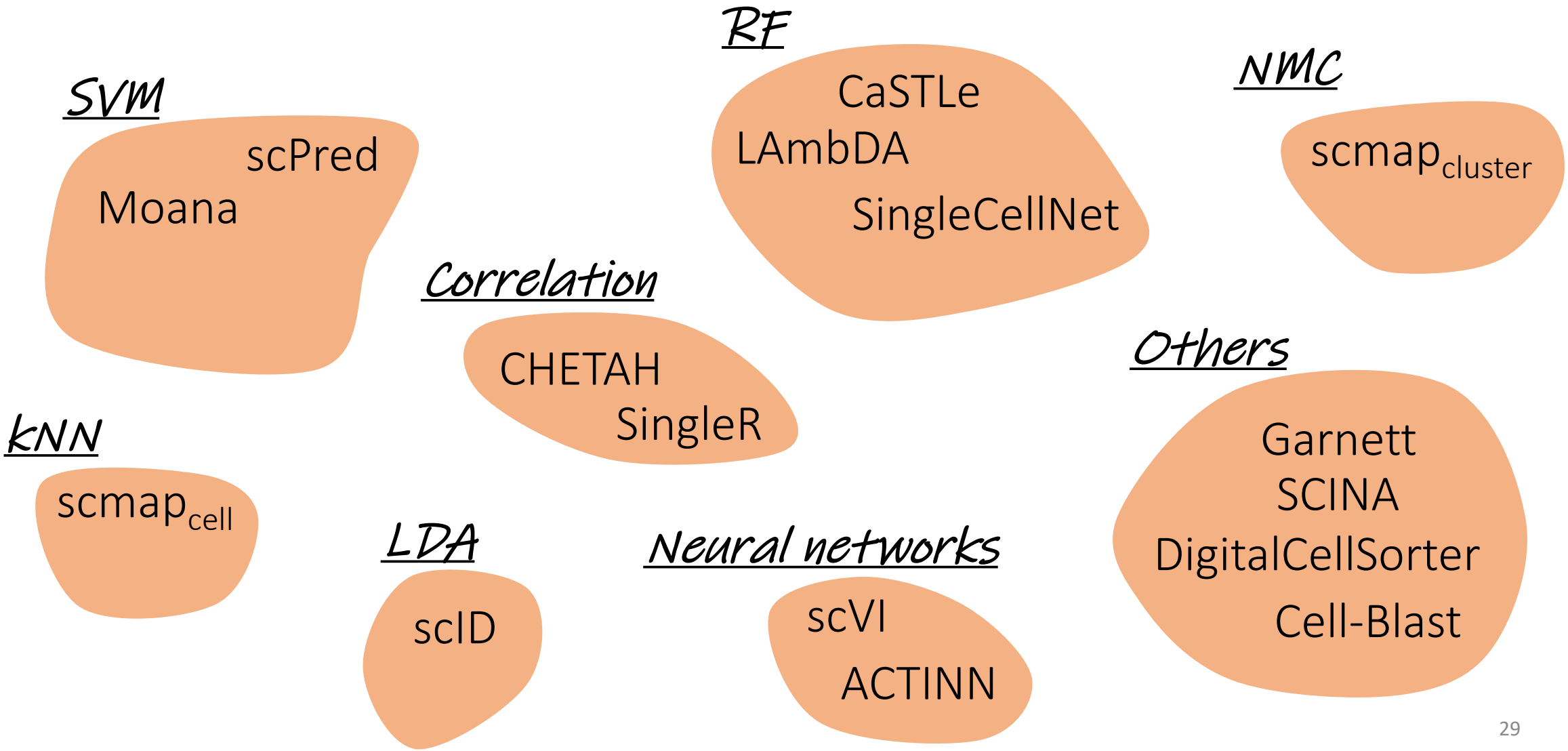
scID

scVI

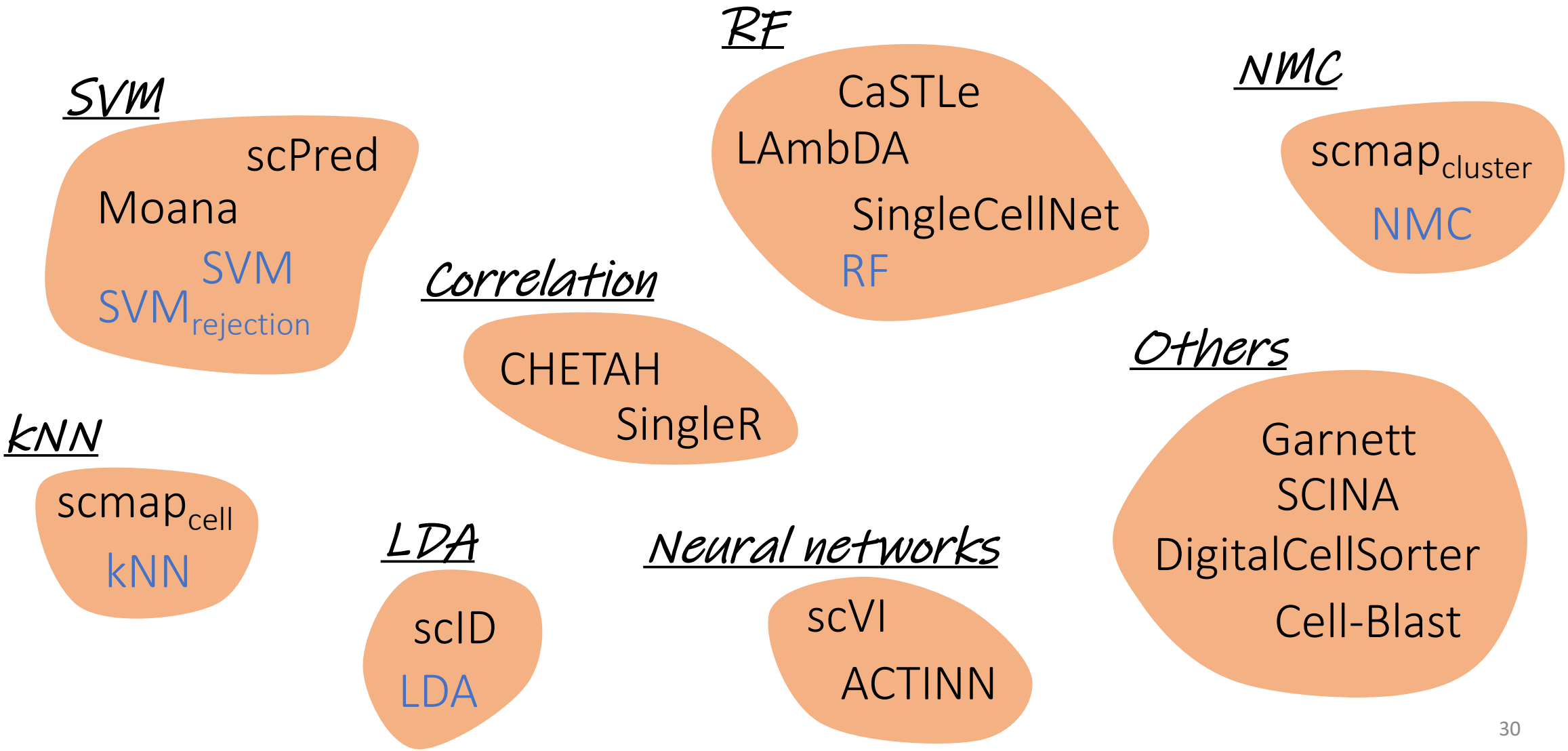
Cell-Blast

ACTINN

16 existing classifiers (April 2019)



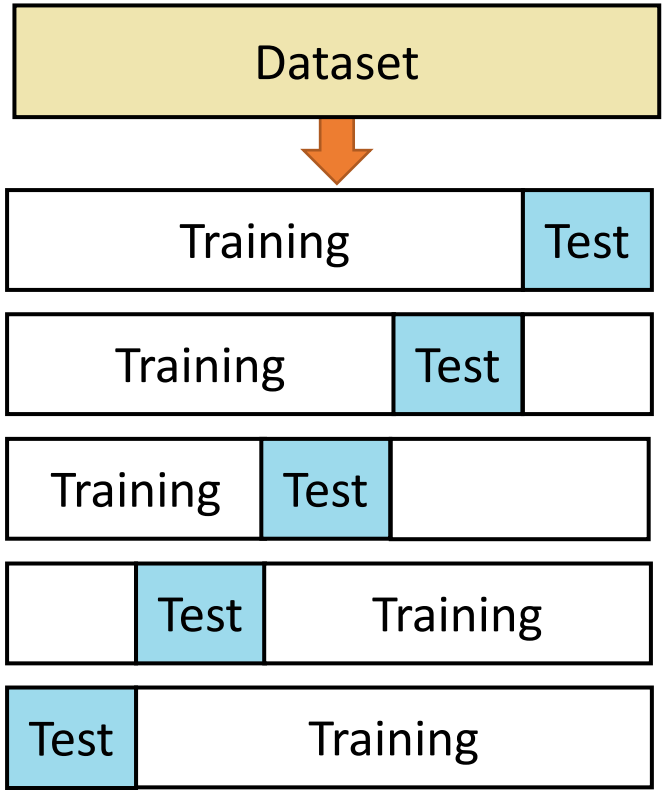
16 existing + 6 off-the-shelf classifiers



Experiment 1: intra-dataset evaluation

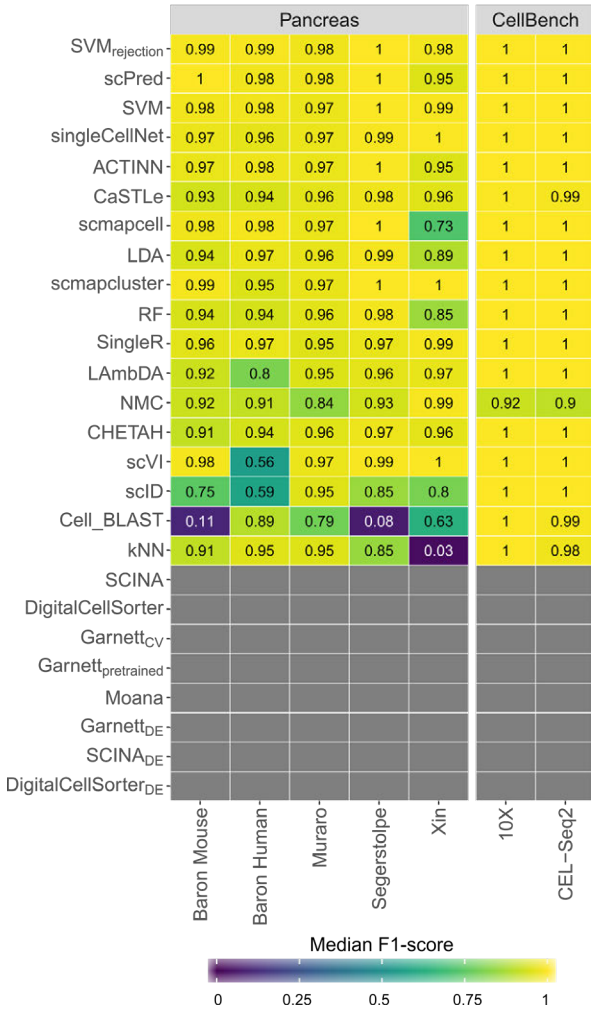
- Stratified 5-fold cross validation

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



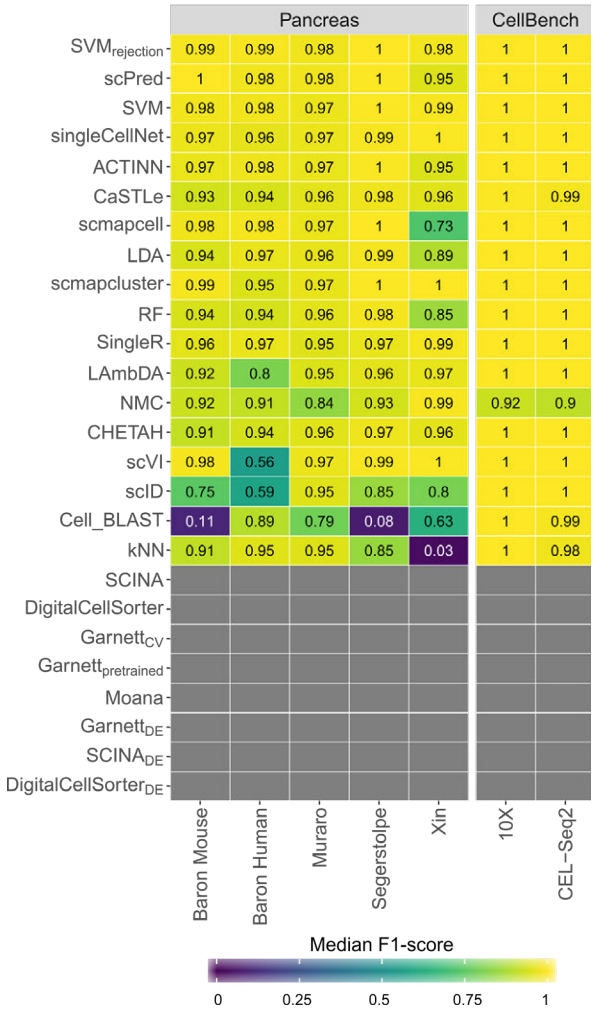
Most classifiers work well

Median F1-score

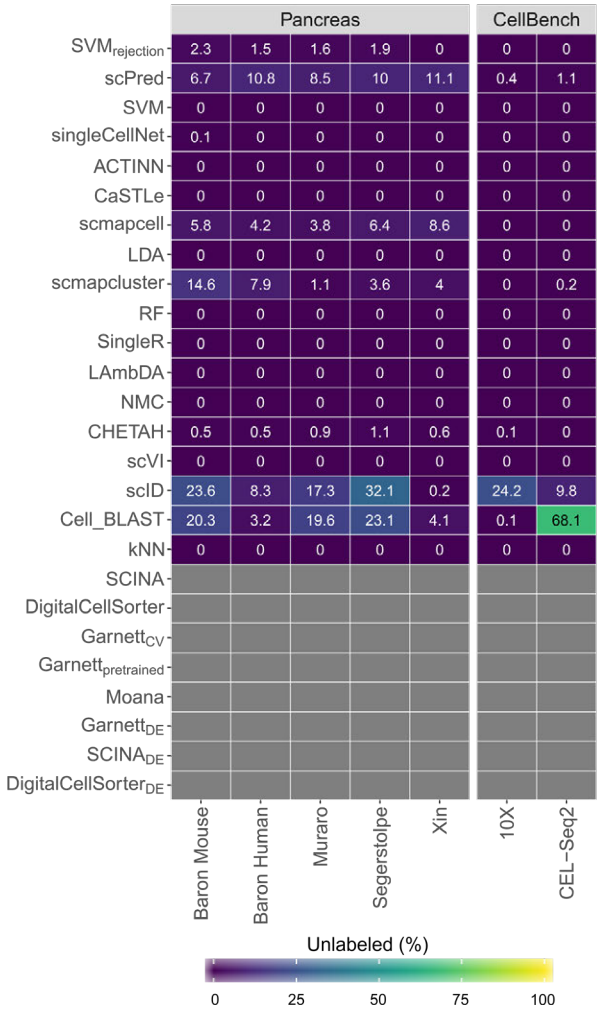


Most classifiers work well

Median F1-score

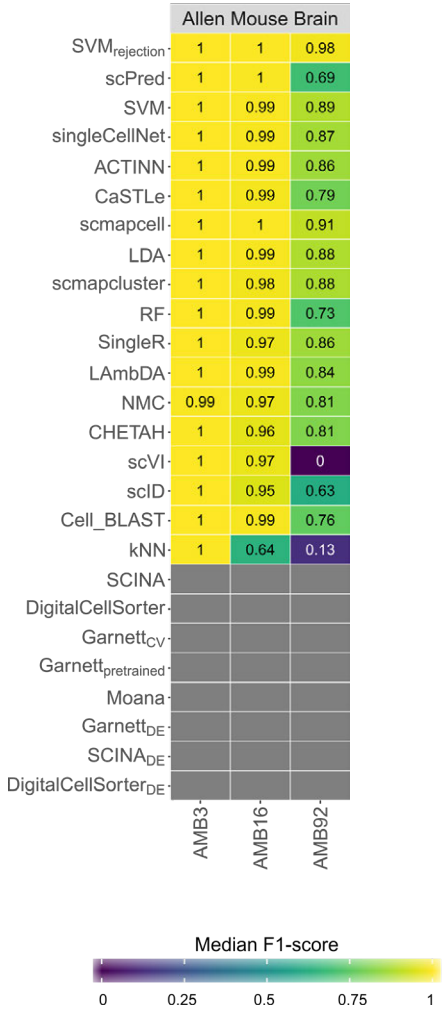


% Unlabeled

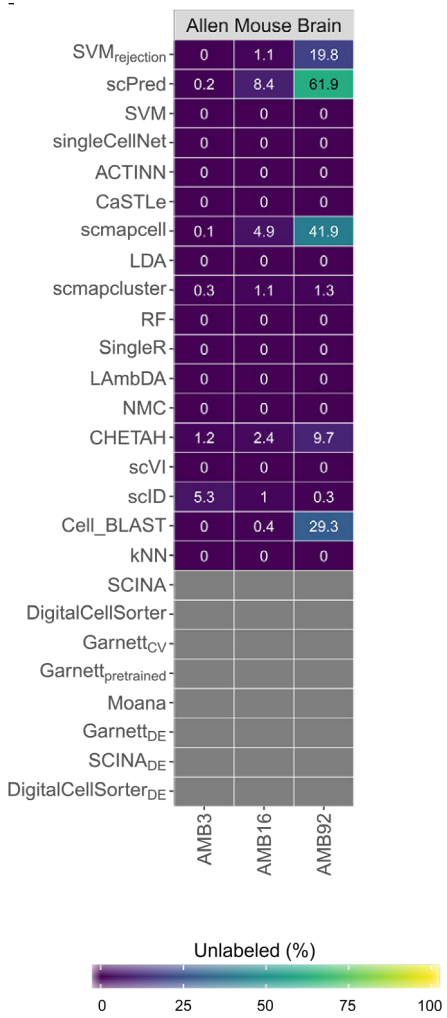


Performance drops with deeper annotation

Median F1-score

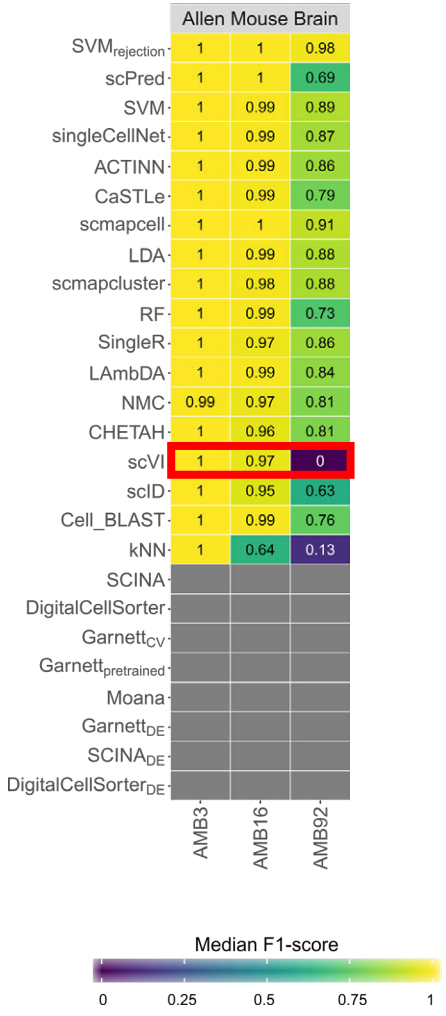


% Unlabeled

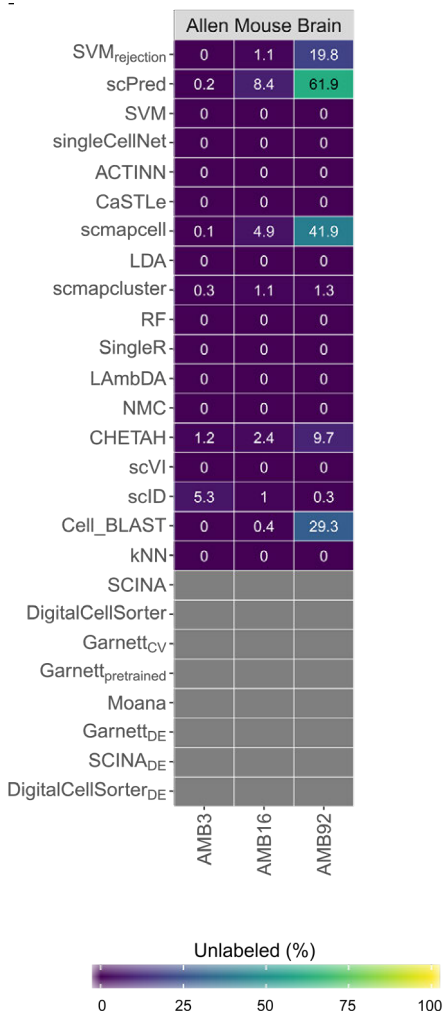


Performance drops with deeper annotation

Median F1-score

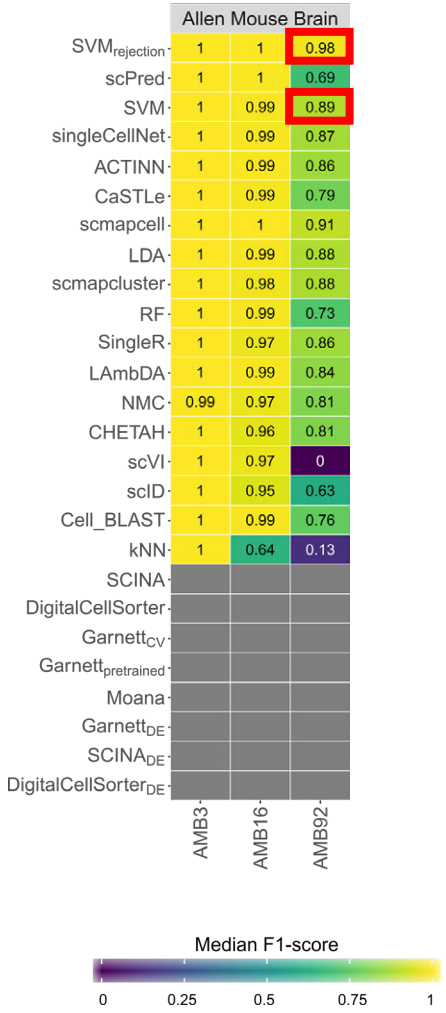


% Unlabeled

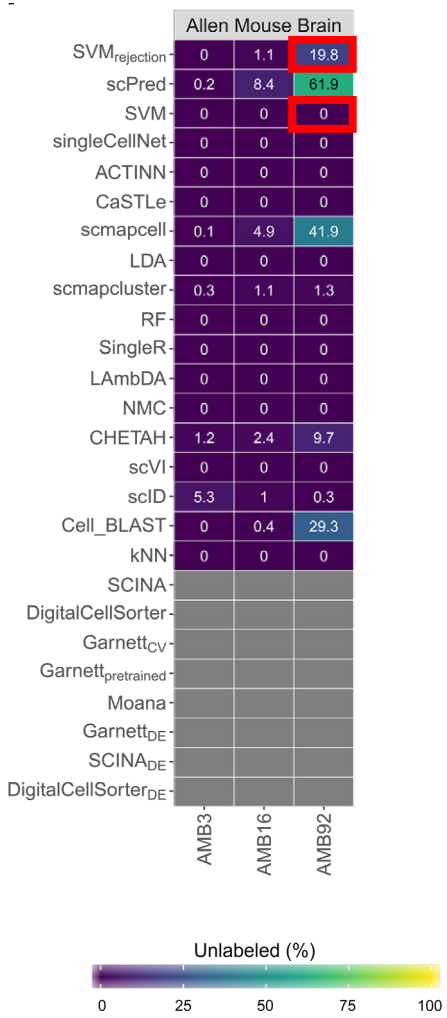


Trade-off between high performance and rejecting cells

Median F1-score

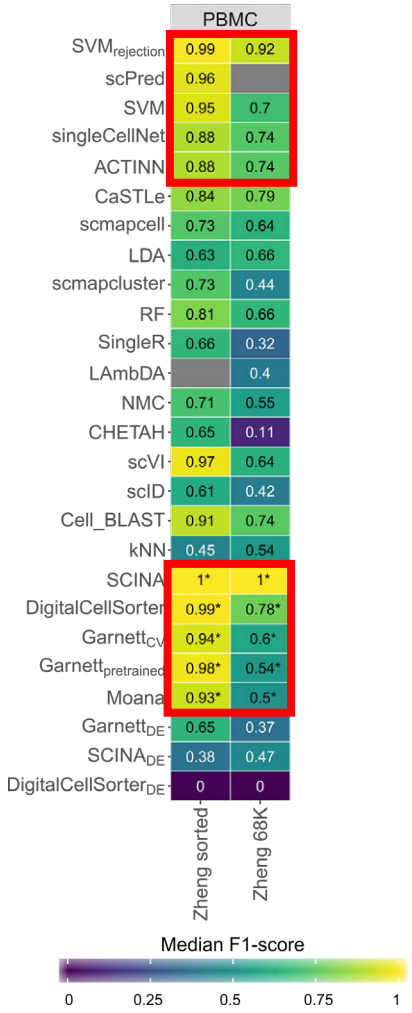


% Unlabeled

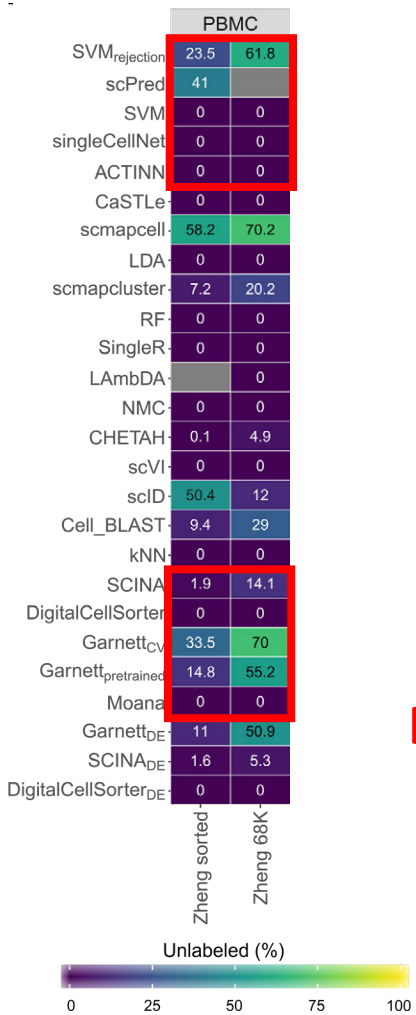


Prior knowledge is not beneficial

Median F1-score



% Unlabeled

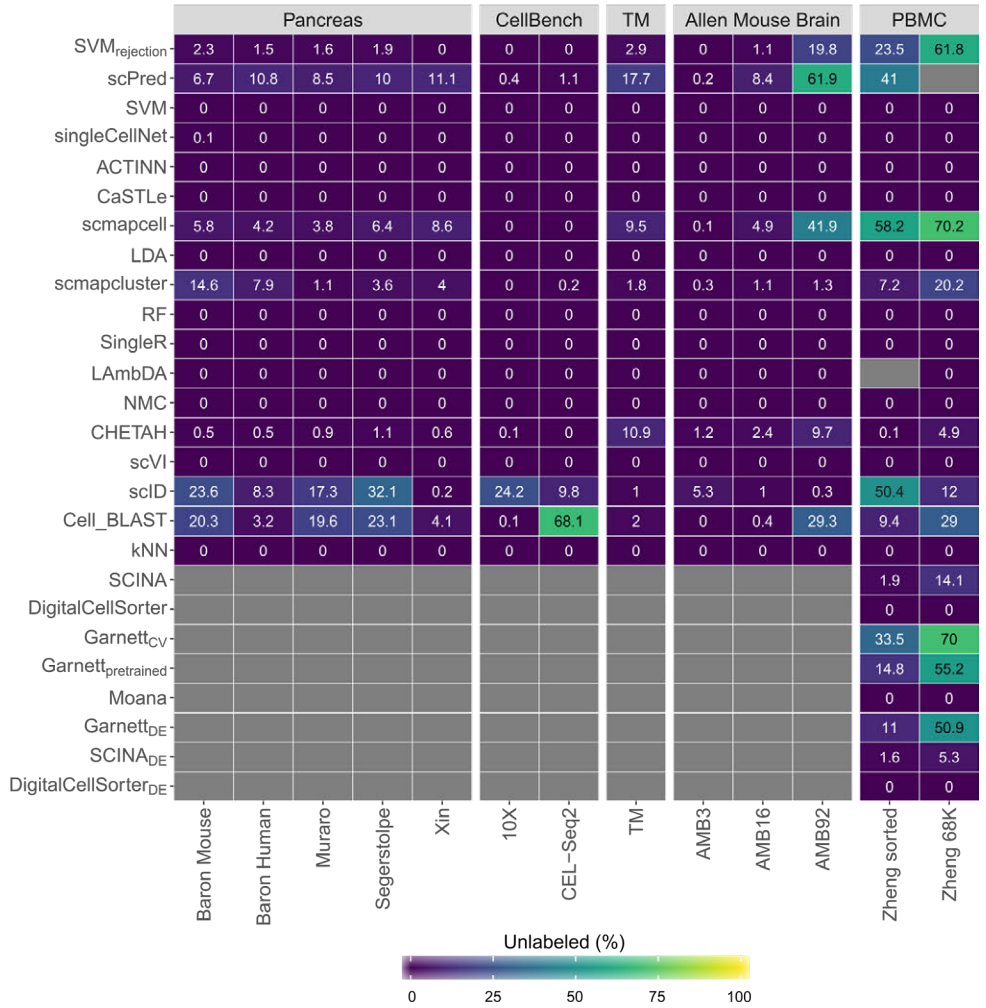


Lower number of classes!

Off-the-shelf SVM outperforms dedicated single cell classifiers

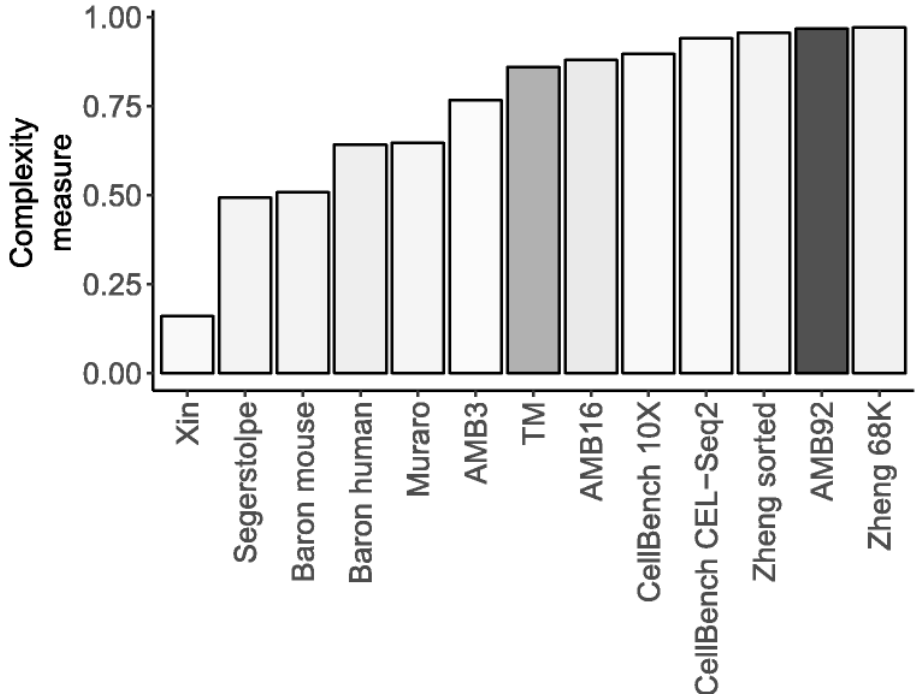
Median F1-score

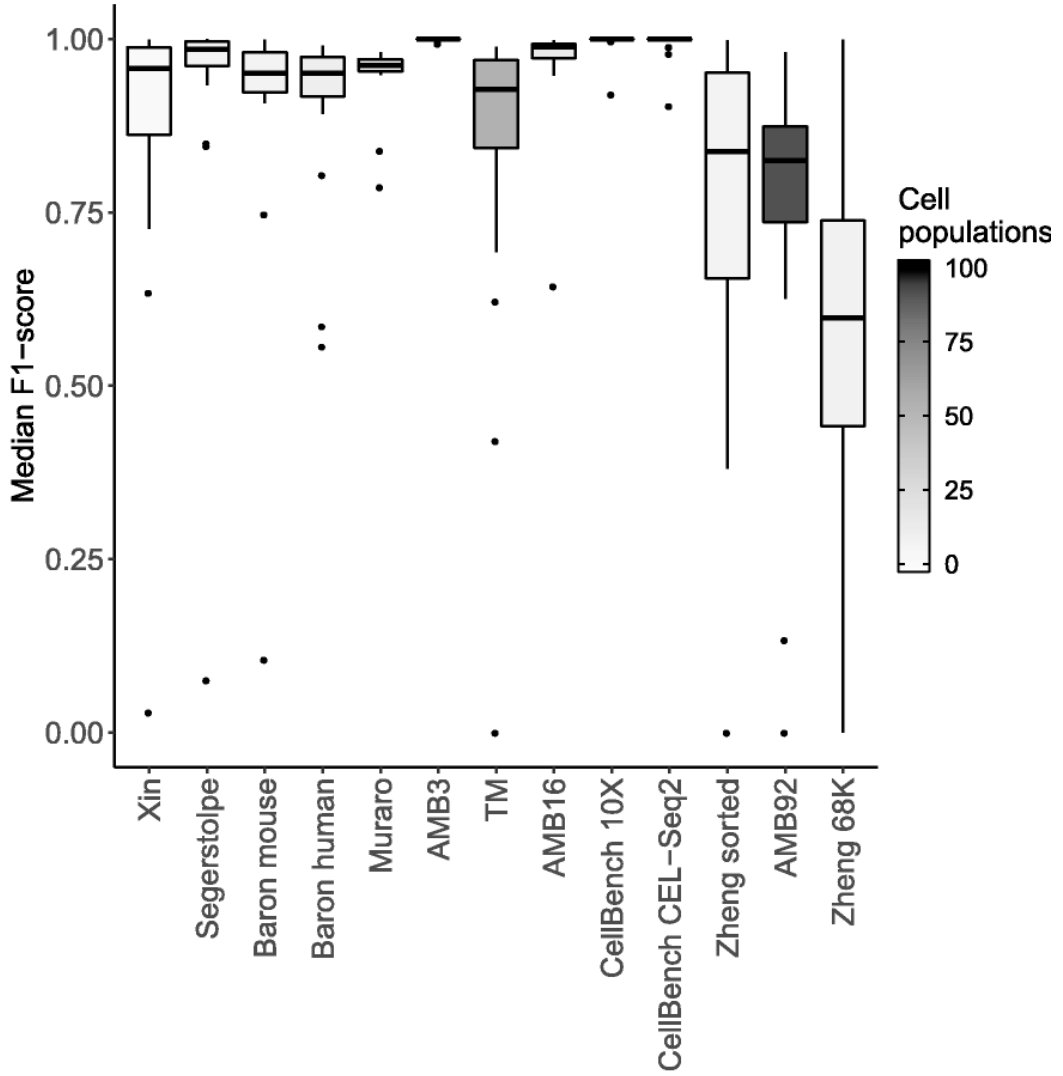
% Unlabeled



Performance depends on dataset complexity

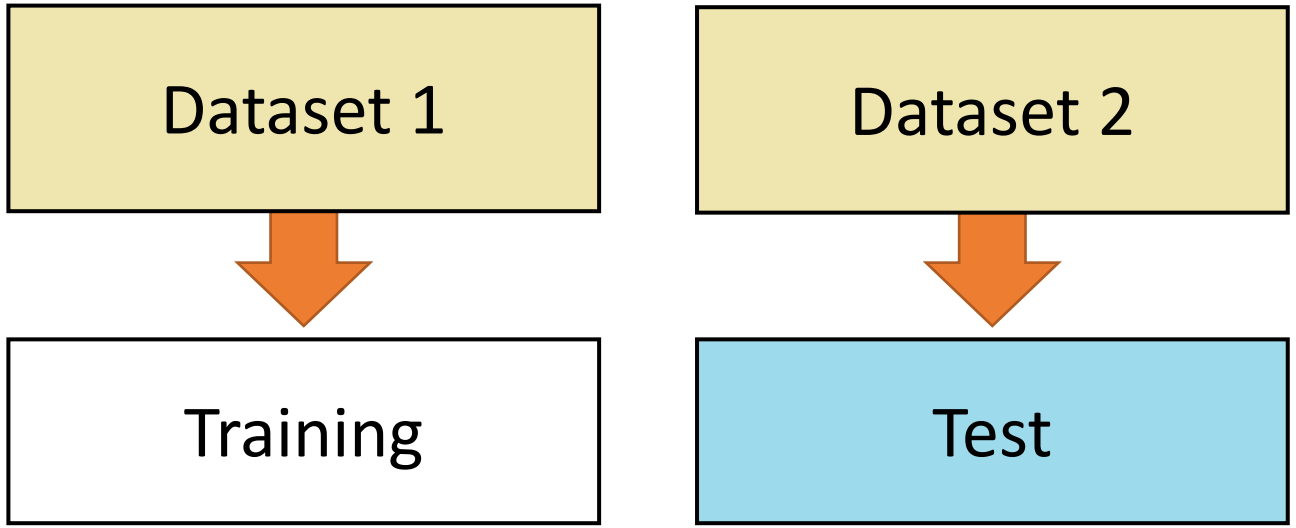
$$\text{Complexity} = \text{mean}\left(\max_{\forall i, i \neq j} \text{corr}_{\forall i,j}\left(\text{avg}_{C_i}, \text{avg}_{C_j}\right)\right)$$



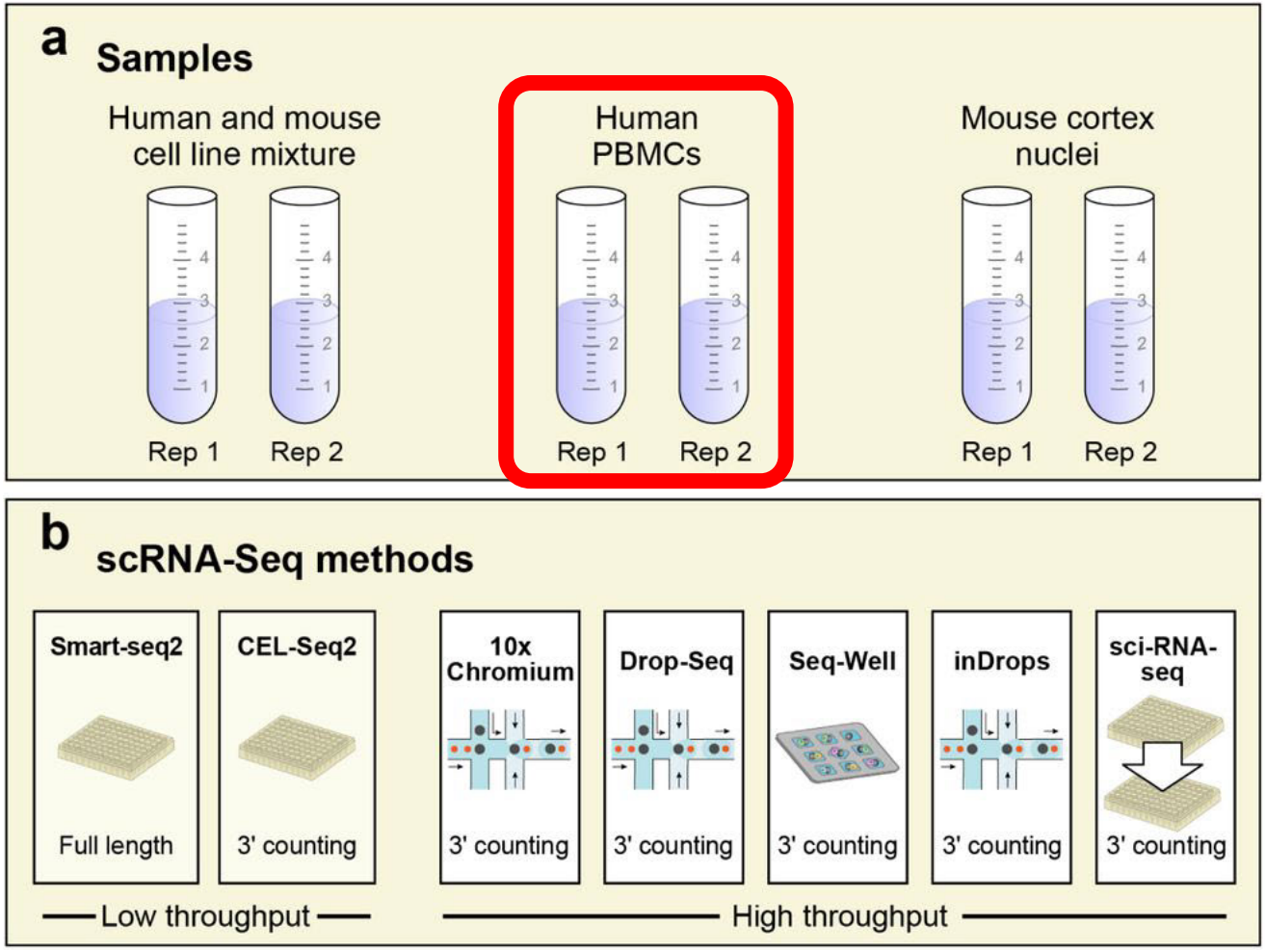


Experiment 2: inter-dataset evaluation

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

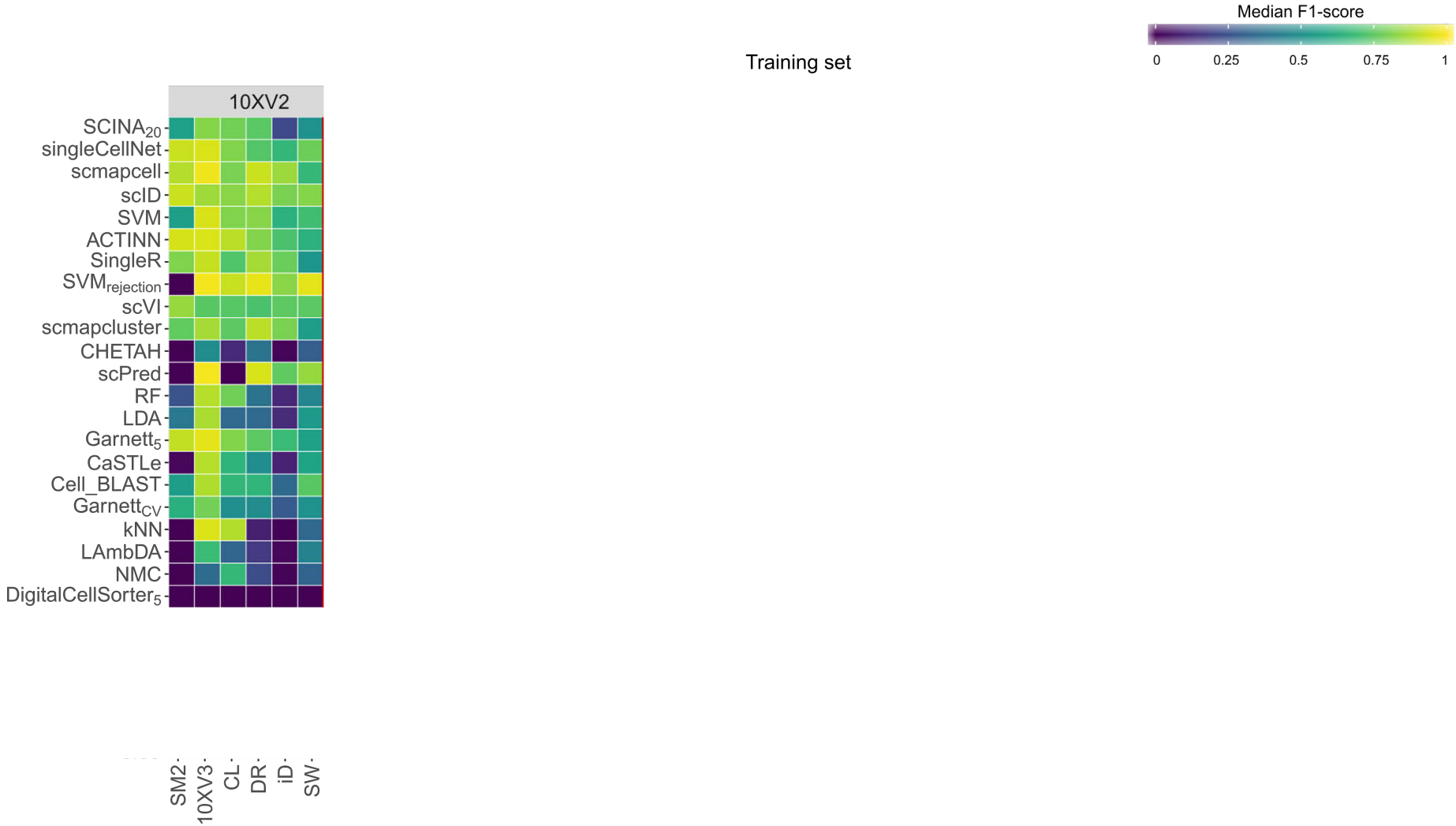


Experiment 2: inter-dataset evaluation

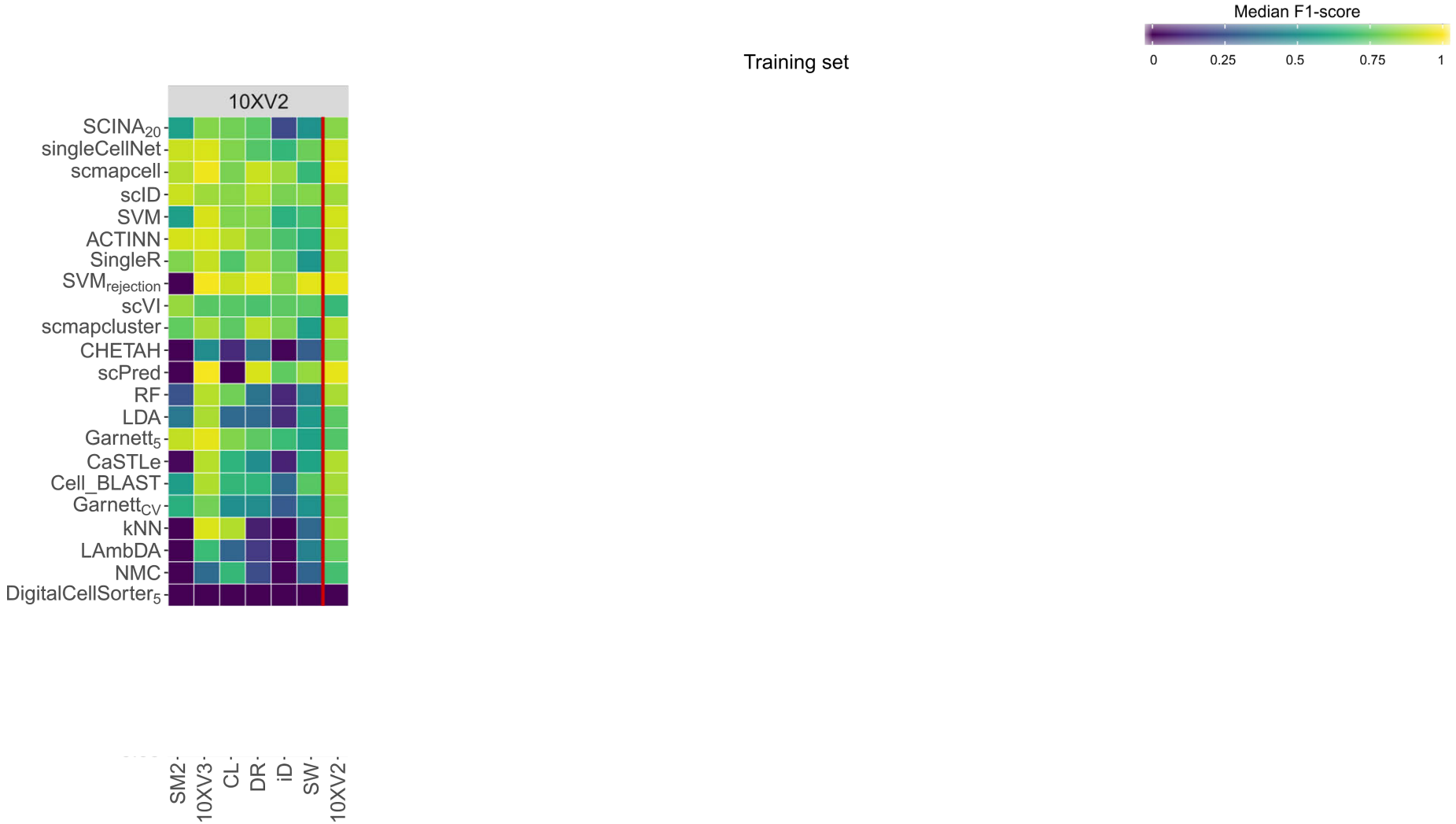


Jiarui Ding et al. *Nature Biotechnology* 2020

Prediction across protocols

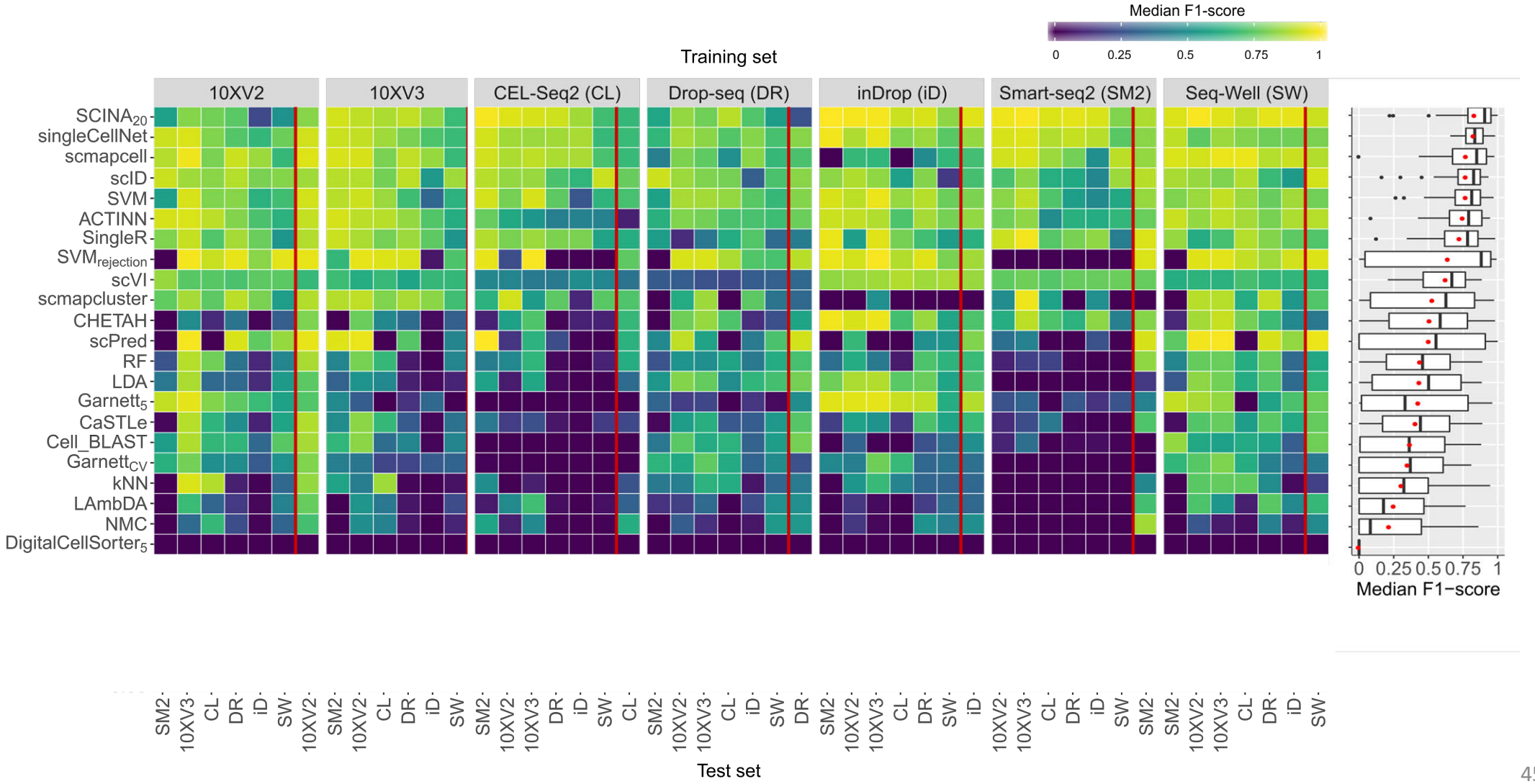


Prediction across protocols

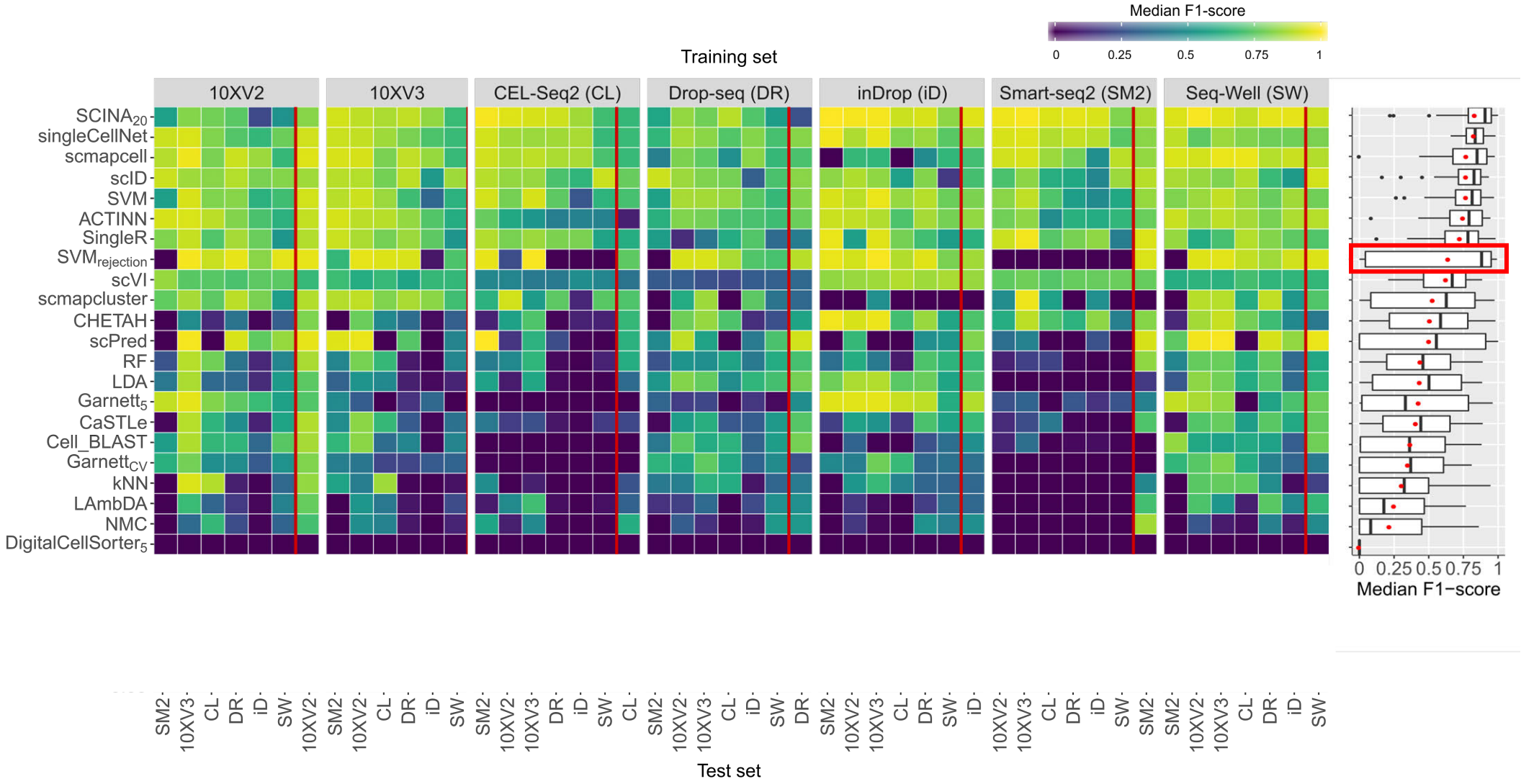


Prediction across protocols

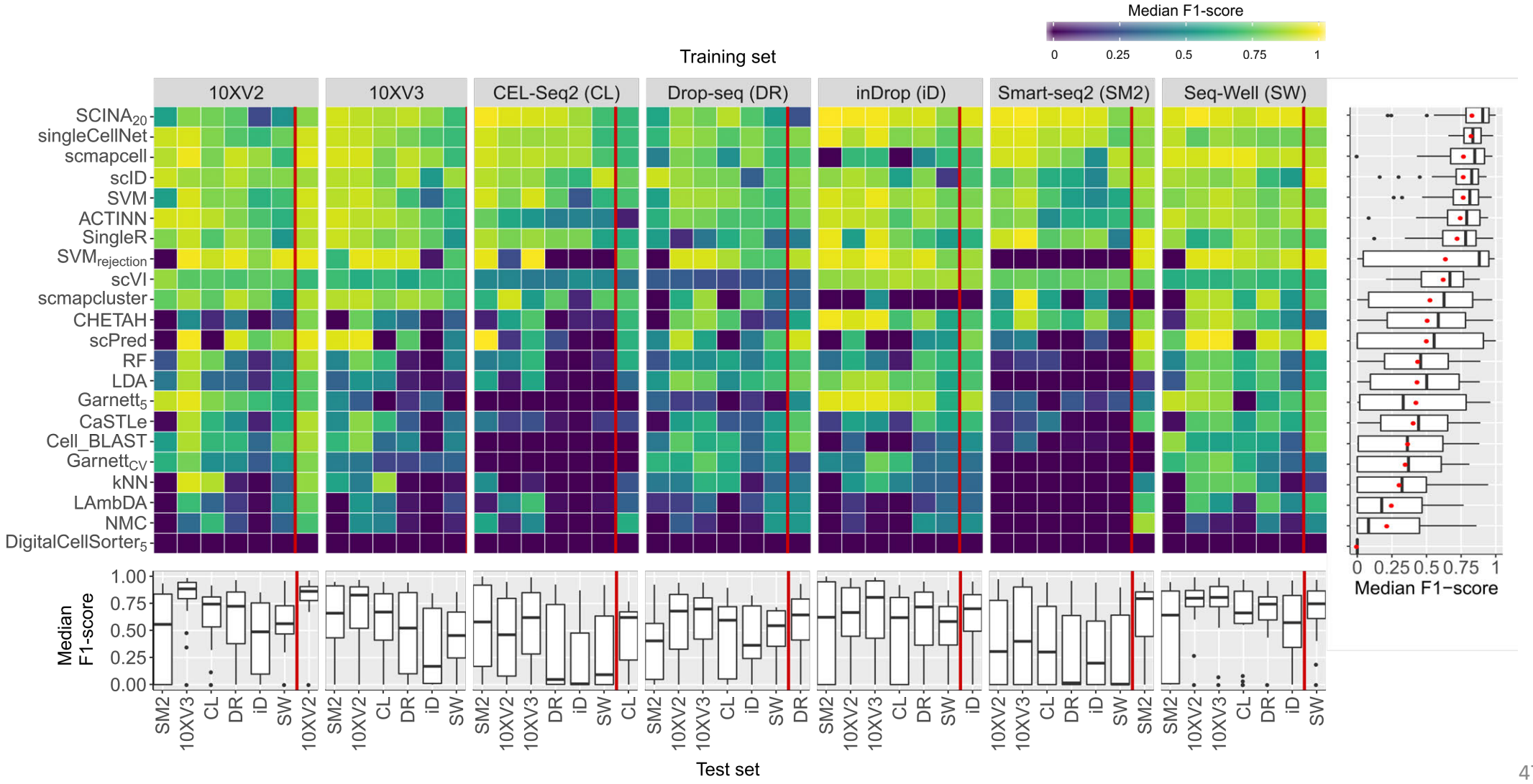
Prediction across protocols



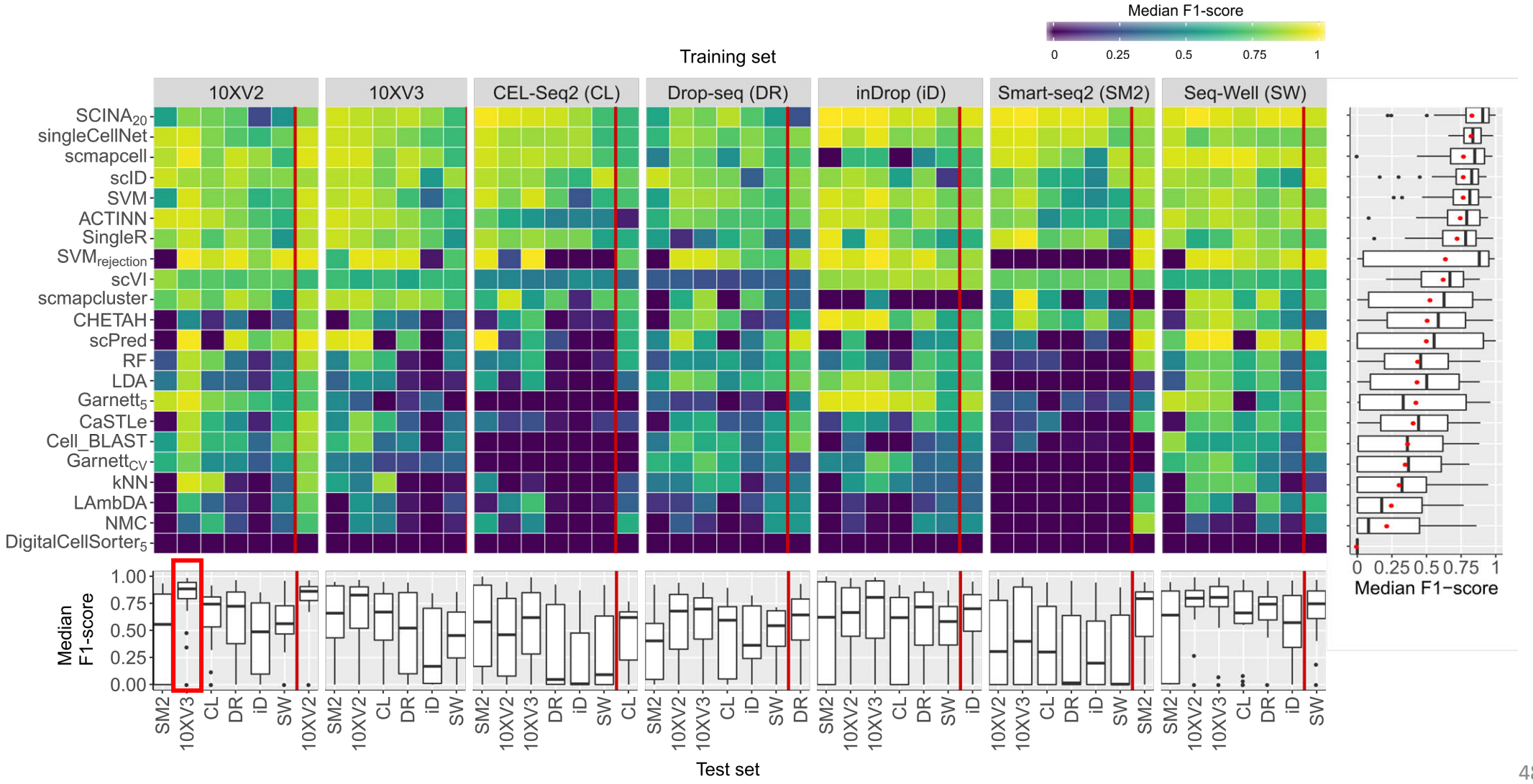
Prediction across protocols



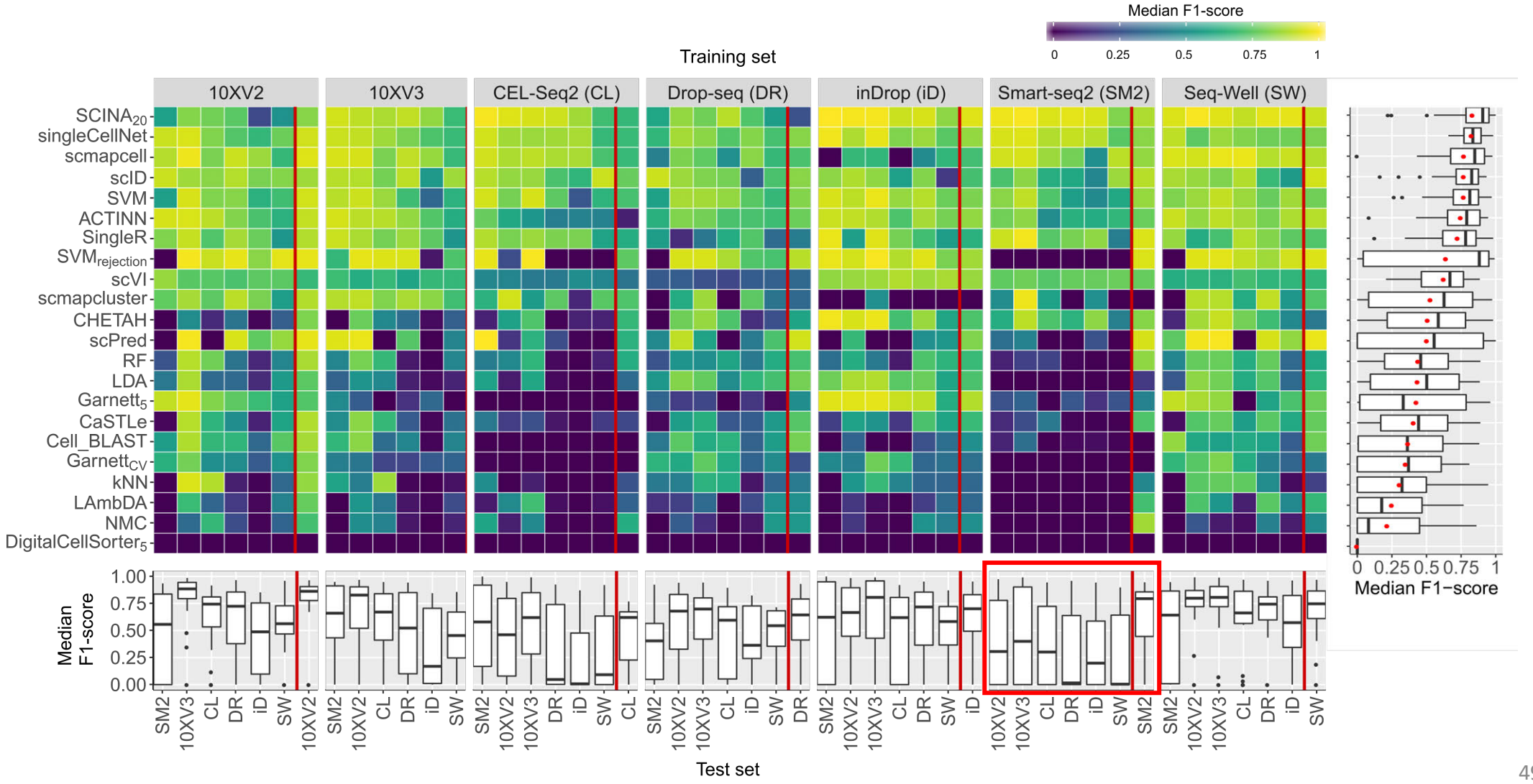
Prediction across protocols



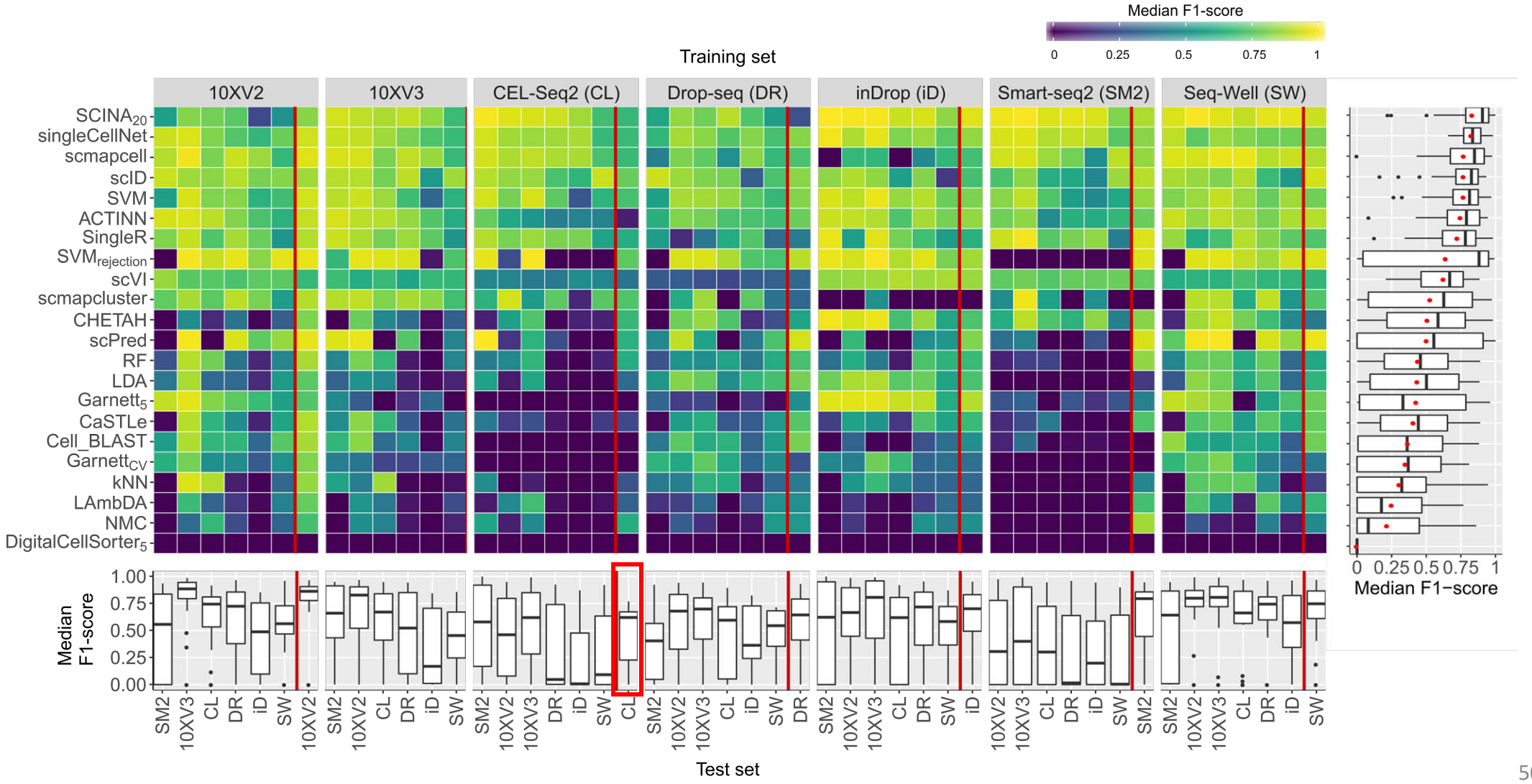
Prediction across protocols



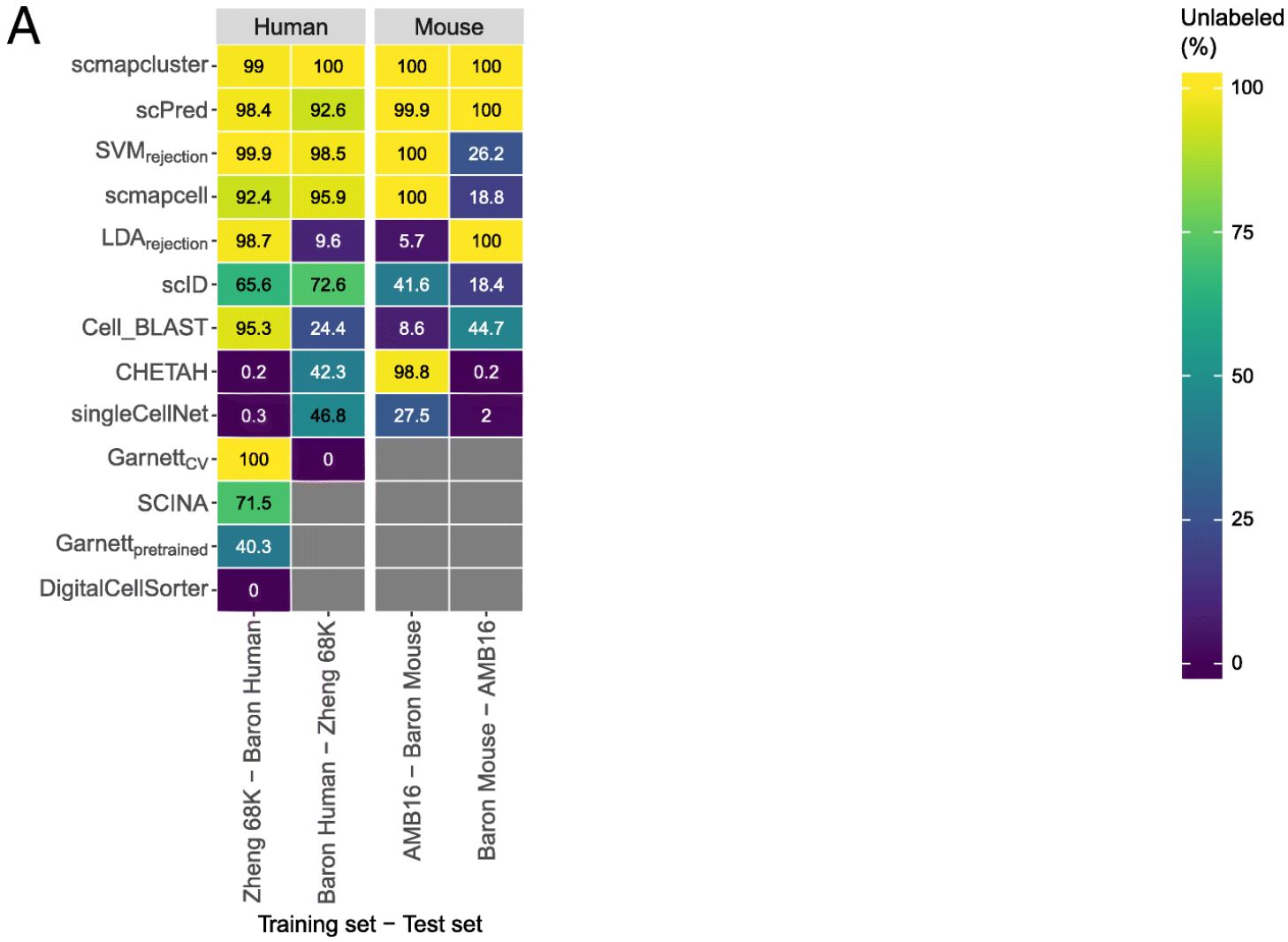
Prediction across protocols



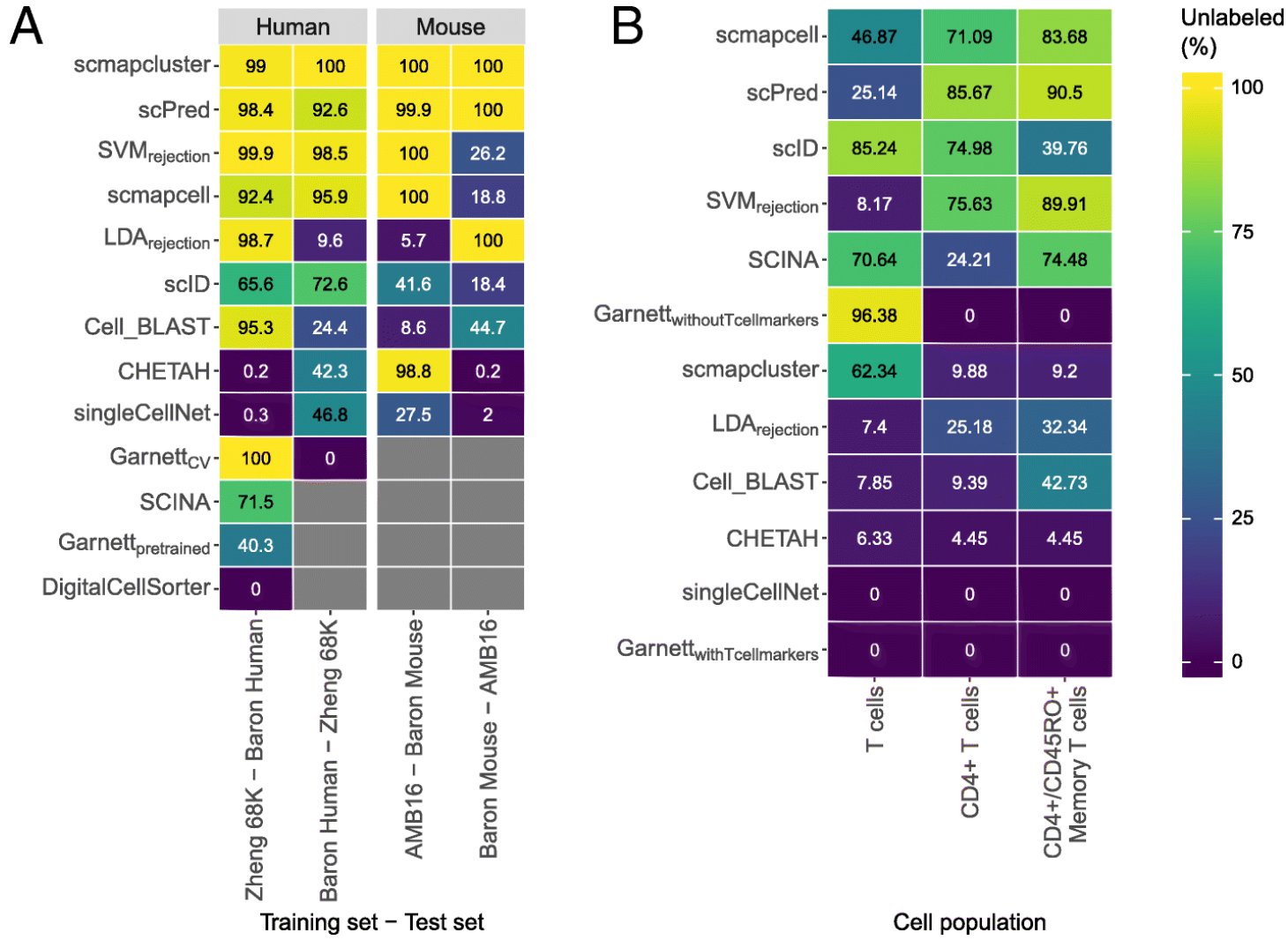
Prediction across protocols



Experiment 3: rejection evaluation



Experiment 3: rejection evaluation



Performance Summary

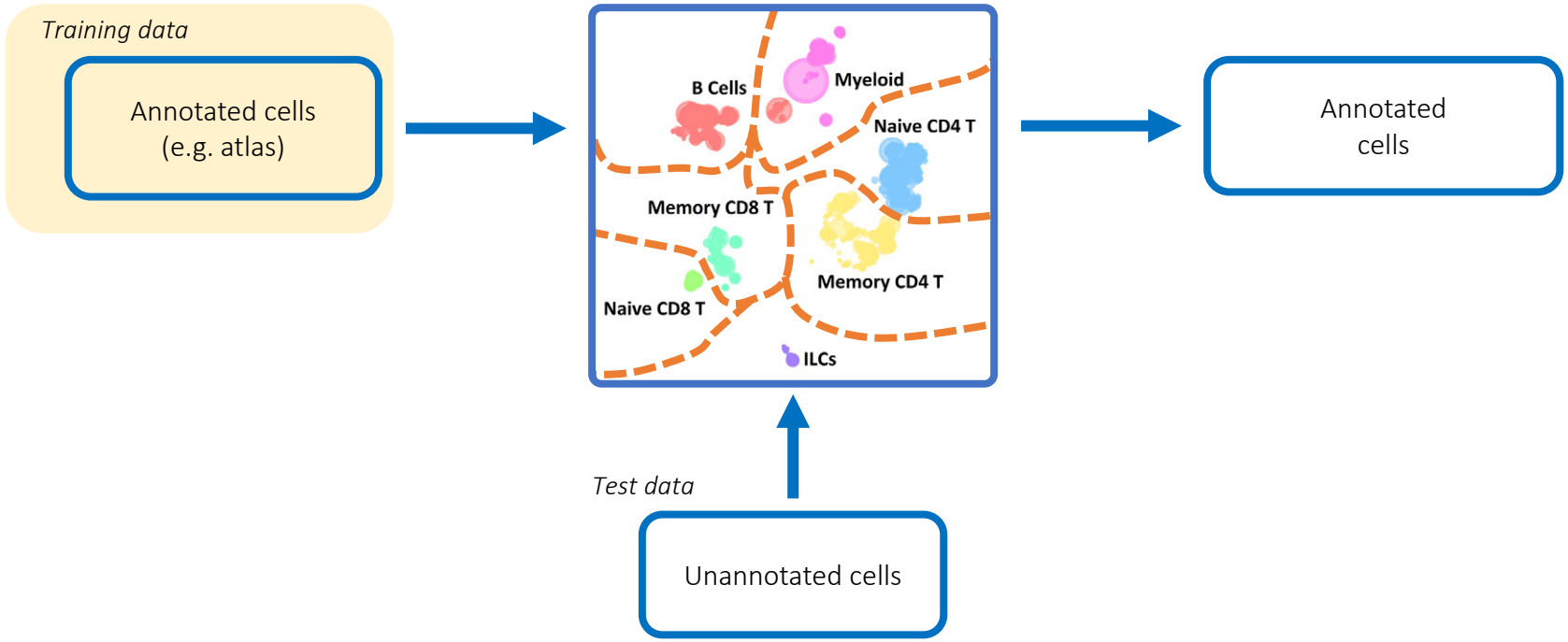
Conclusions so far

- 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/

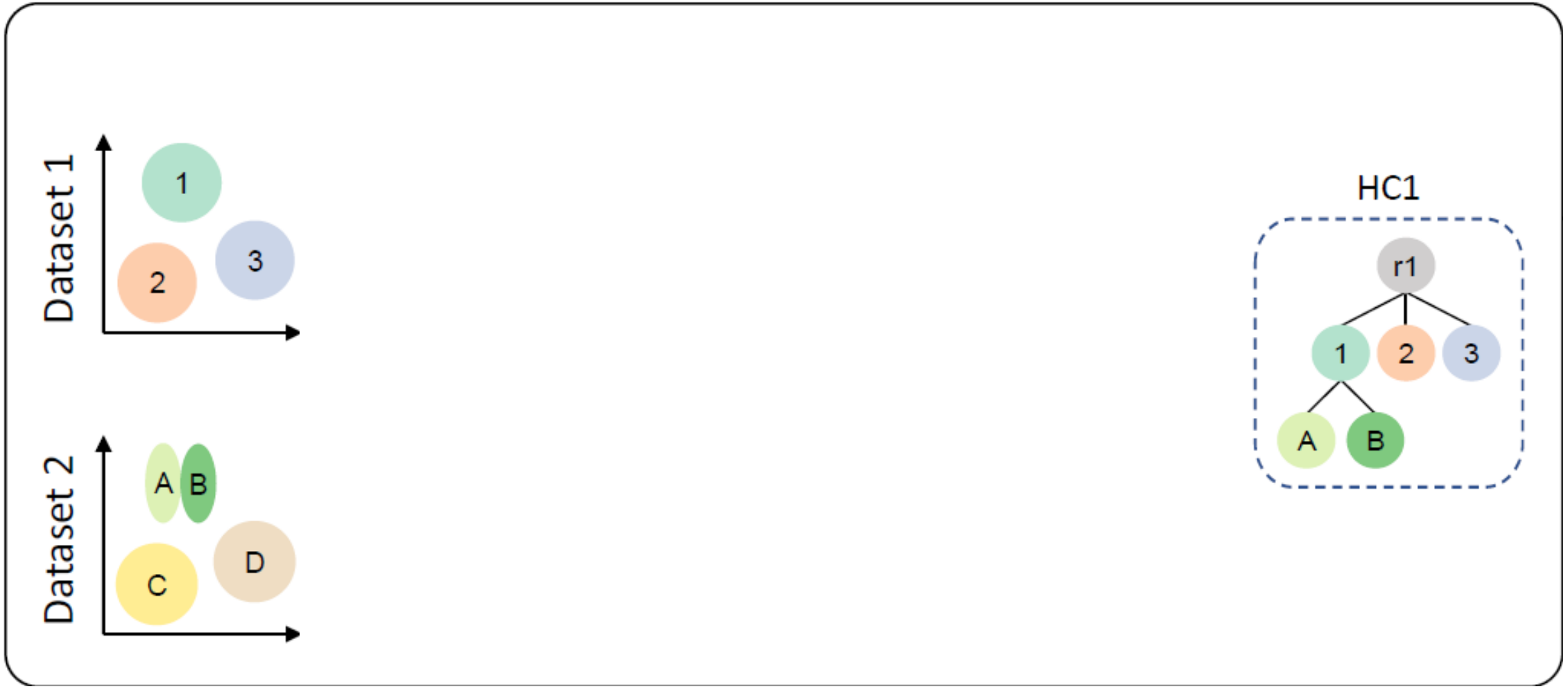
Abdelaal*, Michielsen* et al. *Genome Biology* 2019

Still, challenges remain

- Incomplete/missing reference atlas
- Inconsistent labels across datasets
- Sharing data is an issue (scArches, Lotfollahi et a;. bioRxiv 2020)

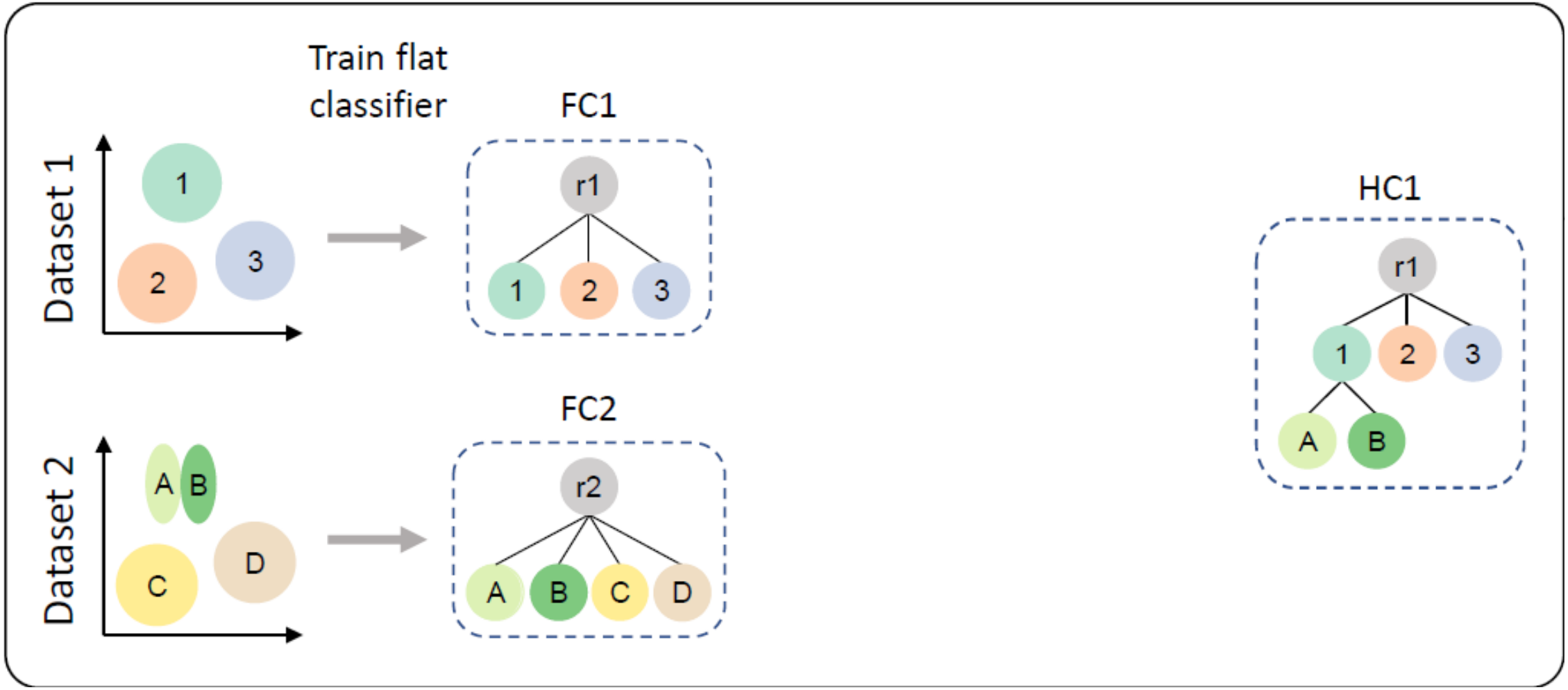


Hierarchical Progressive Learning



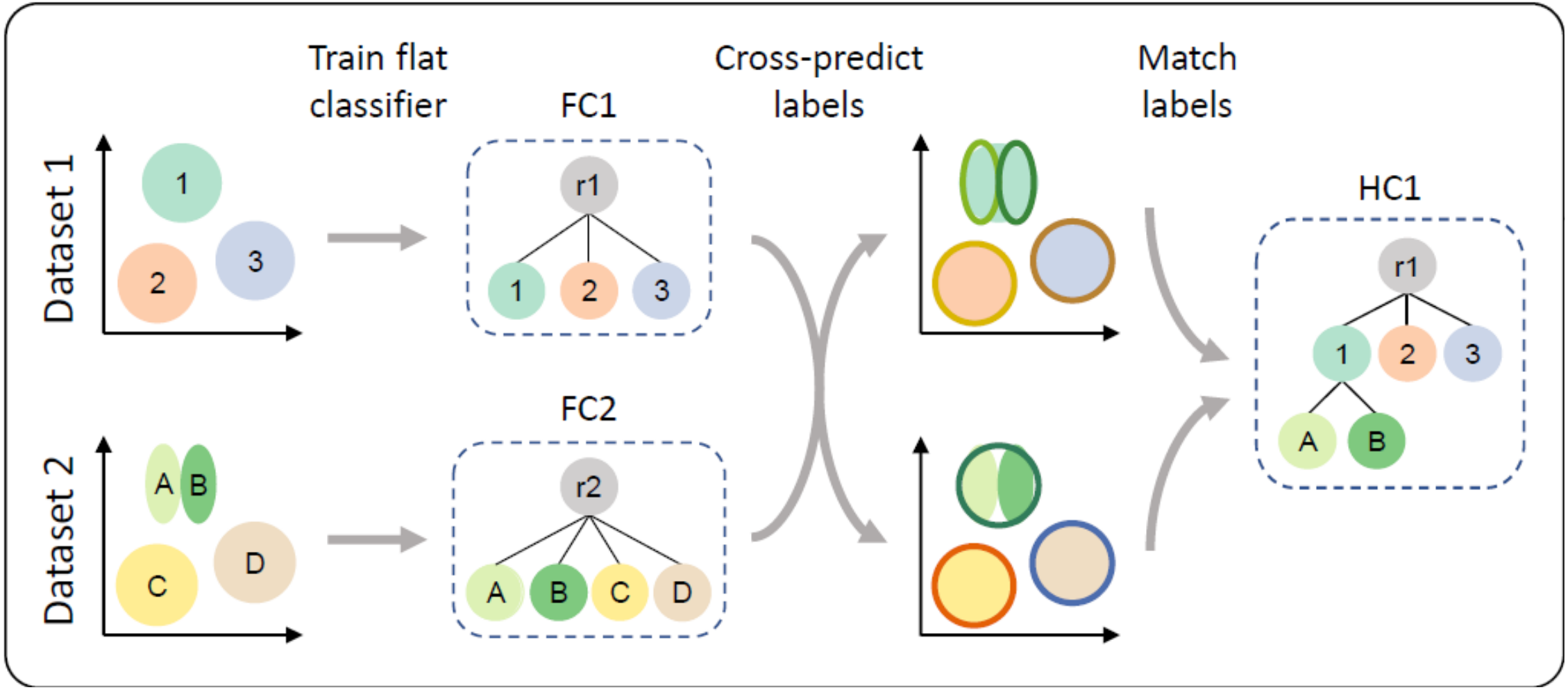
Michielsen et al. *bioRxiv* 2020

Hierarchical Progressive Learning



Michielsen et al. *bioRxiv* 2020

Hierarchical Progressive Learning

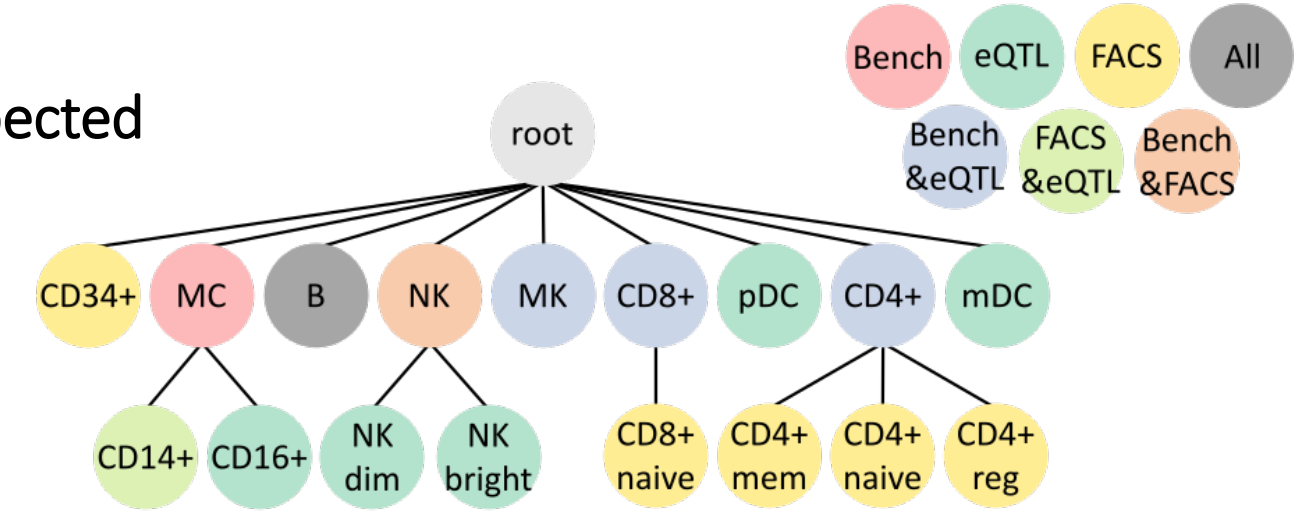


Michielsen et al. *bioRxiv* 2020

Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000

Expected

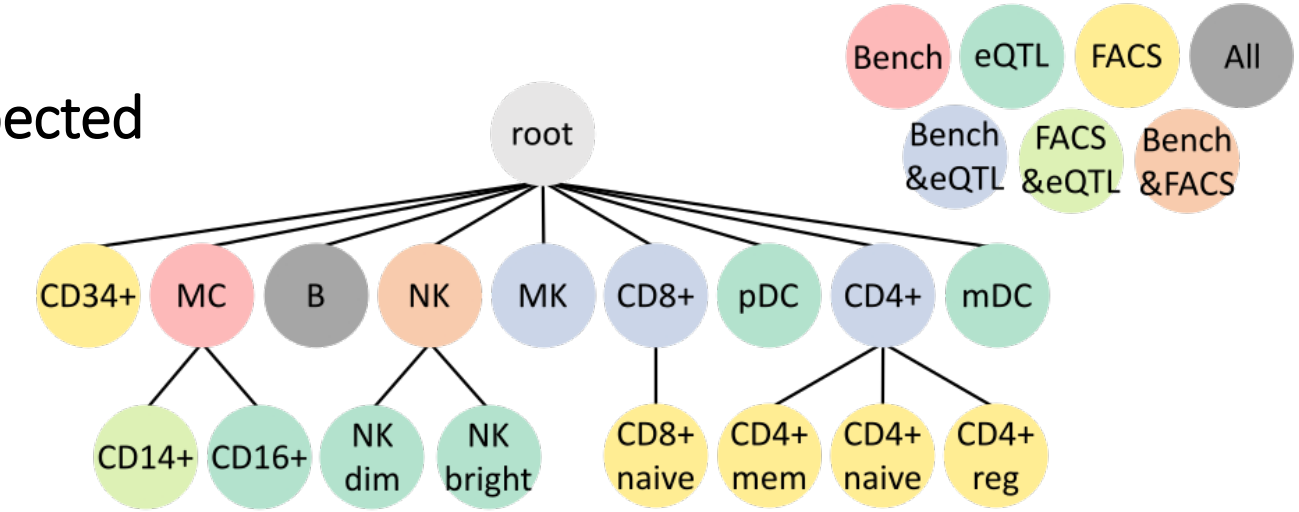


Michielsen et al. *bioRxiv* 2020

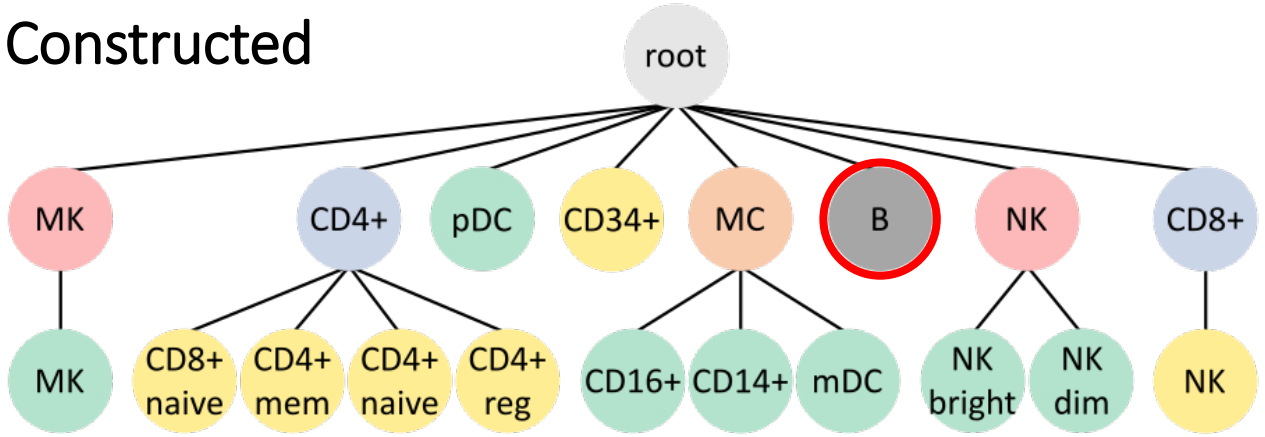
Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000

Expected



Constructed

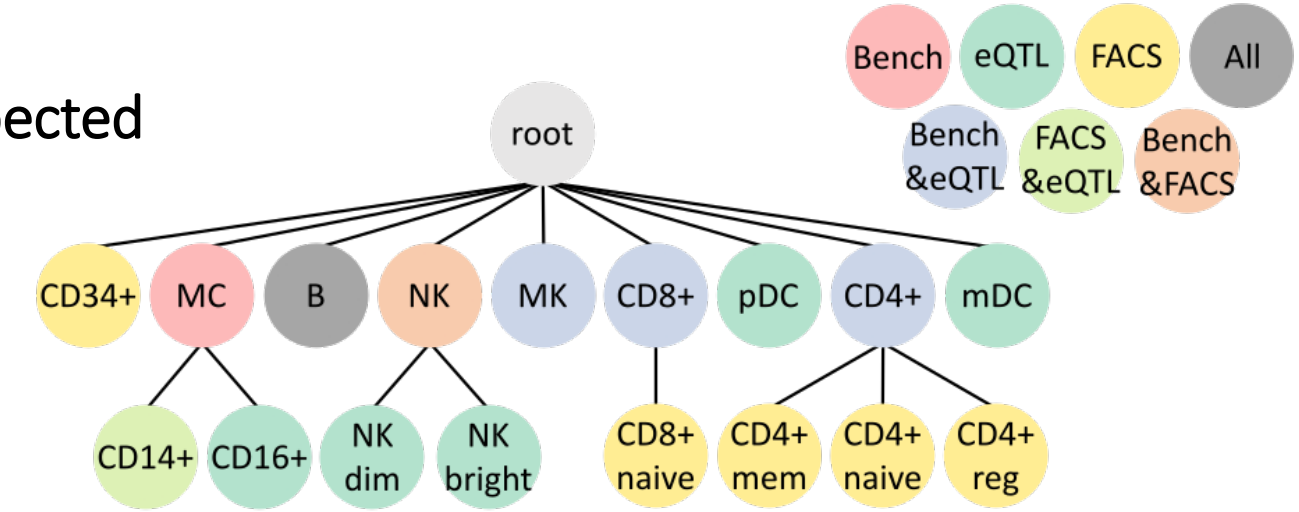


Michielsen et al. *bioRxiv* 2020

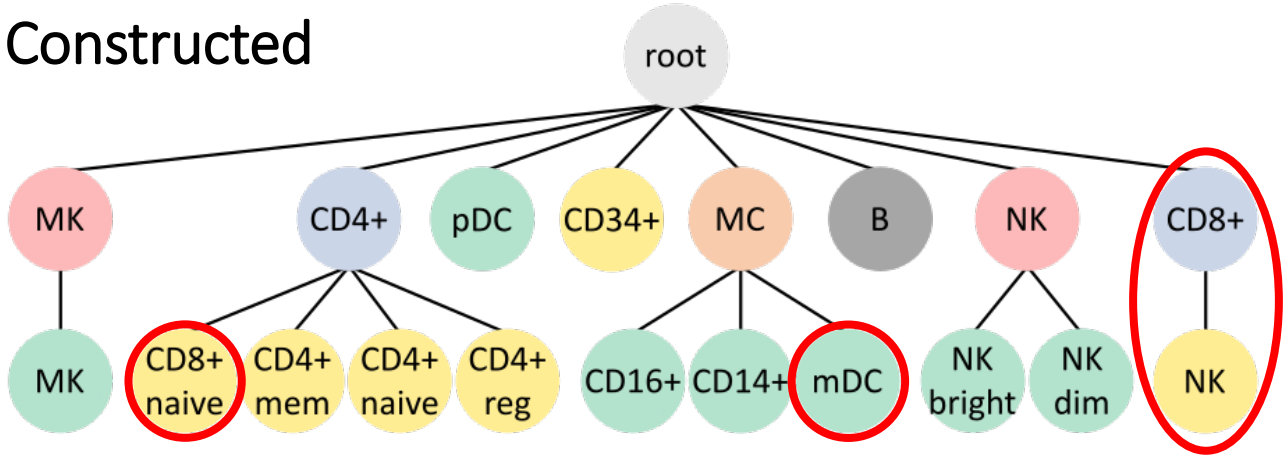
Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000

Expected

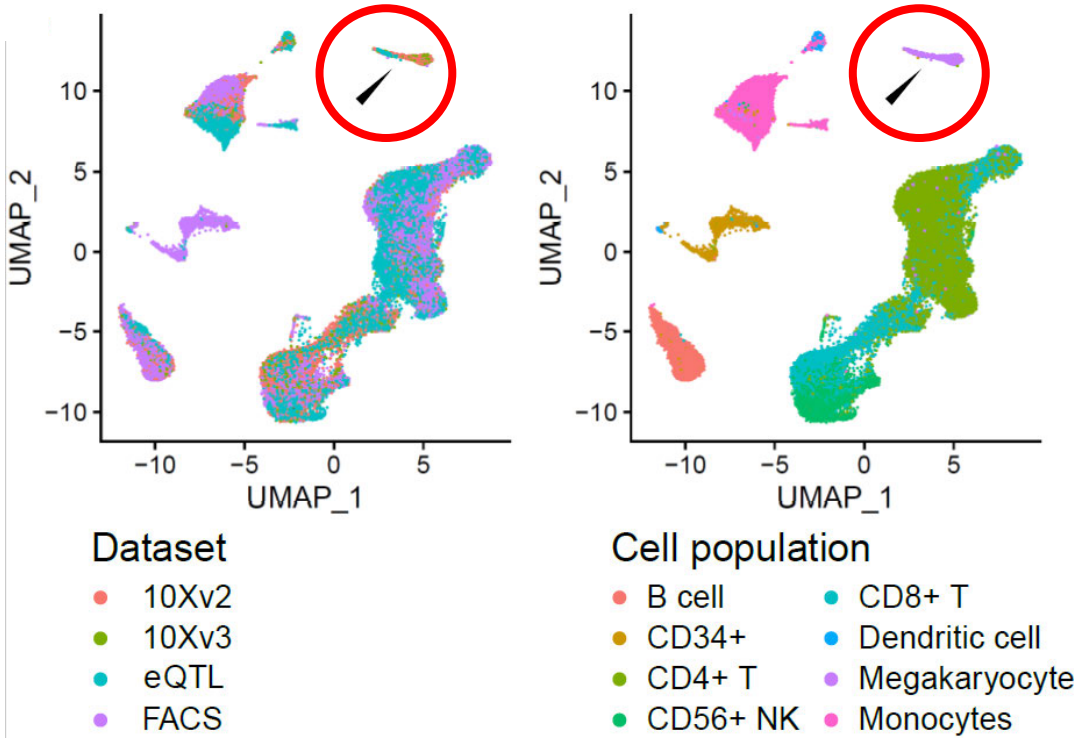


Constructed

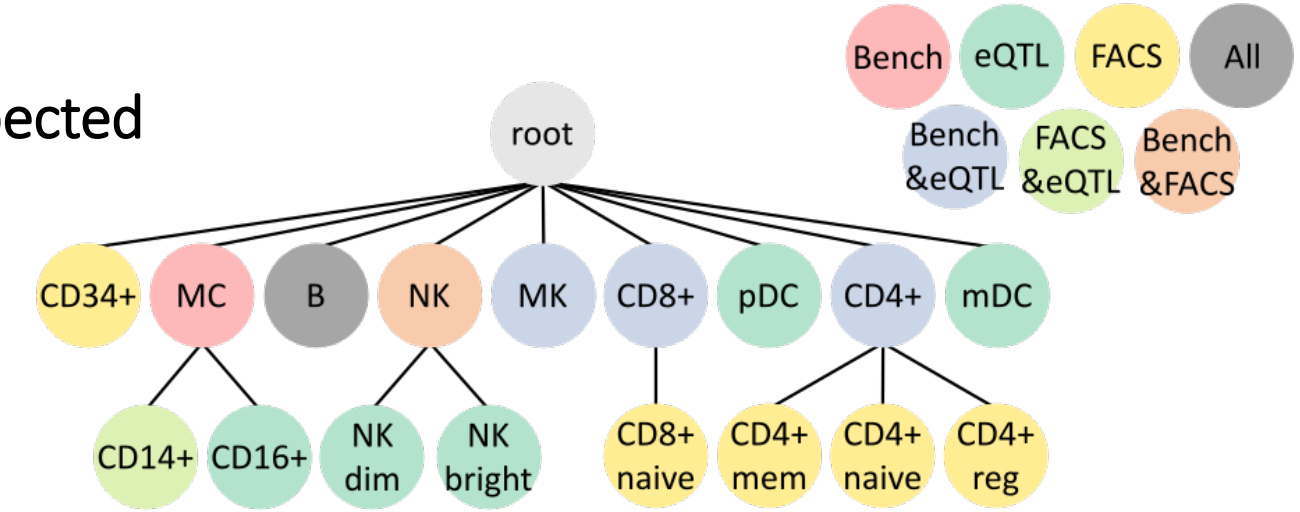


Michielsen et al. *bioRxiv* 2020

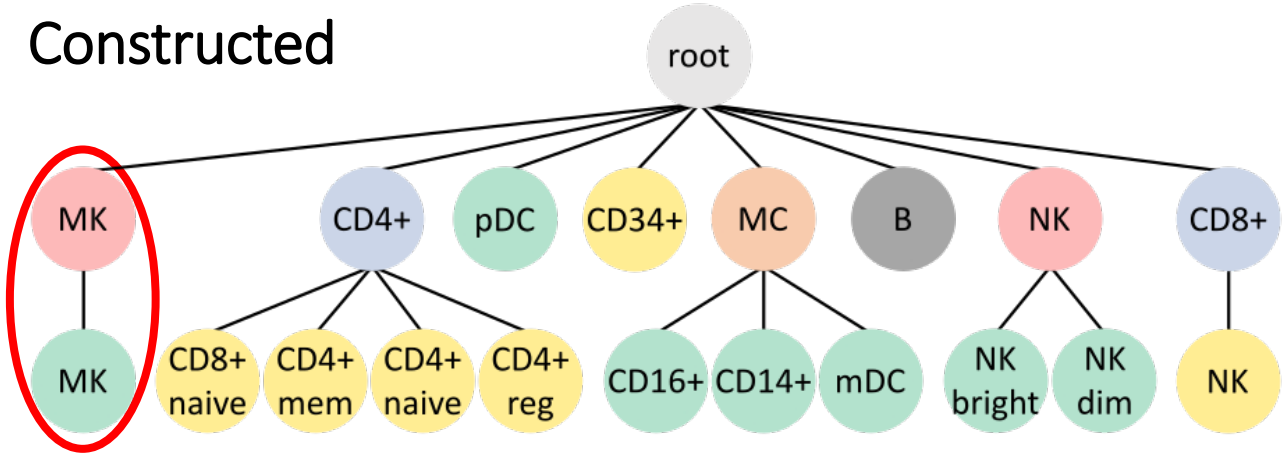
Tree construction



Expected



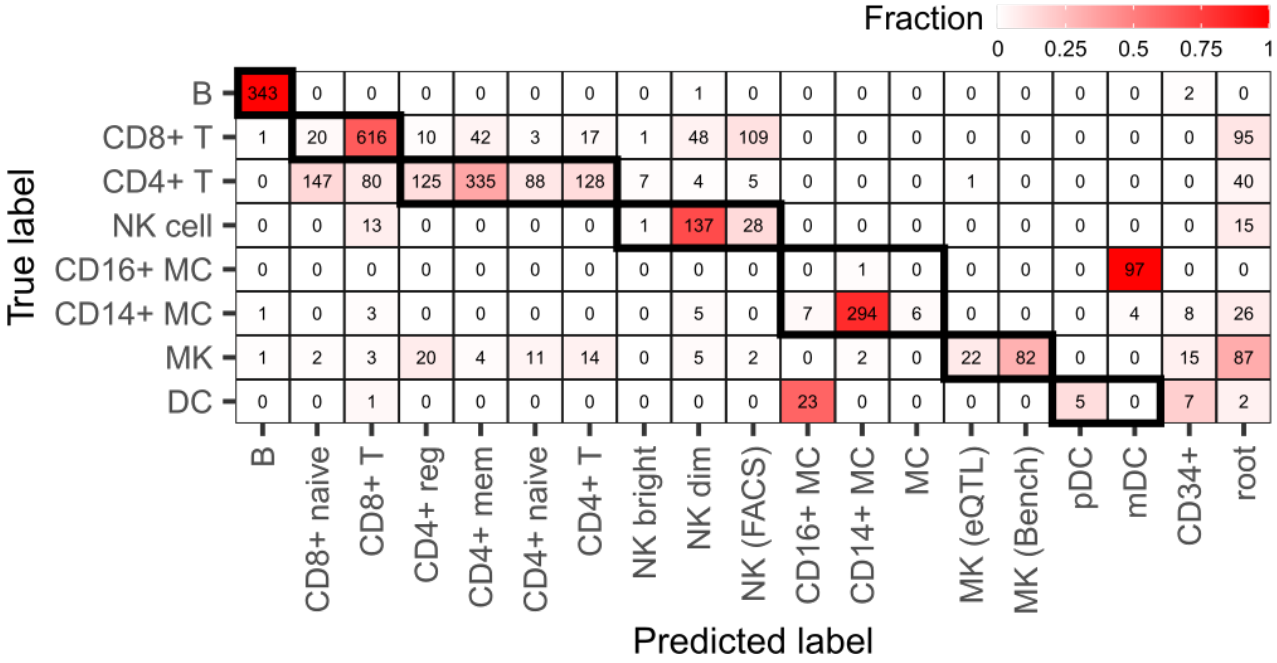
Constructed



Michielsen et al. *bioRxiv* 2020

Classification performance

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS	Testset Bench 10Xv3
CD19+ B	812	676	2,000	346
Monocytes (MC)		1,194		
CD14+	2,081		2,000	354
CD16+	274			98
CD4+ T	13,523	1,458		960
Reg.			2,000	
Naive			2,000	
Memory			2,000	
CD8+ T	4,195	2,128		962
Naive			2,000	
Megakaryocyte (MK)	142	433		270
NK cell		429	2,000	194
CD56+ bright	355			
CD56+ dim	2,415			
Dendritic				38
Plasmacytoid (pDC)	101			
Myeloid (mDC)	455			
CD34+			2,000	



Michielsen et al. *bioRxiv* 2020

Summary

- Cell identification is moving from unsupervised (clustering/visualization) to supervised (classification) learning

- Comprehensive benchmark of classifiers for single-cell RNA-seq data helps both users and developers

- Continuous learning from a growing reference atlas by combining multiple annotated datasets into a hierarchical classifier (scHPL)

Thank You!

a.mahfouz@lumc.nl
mahfouzlab.org
@ahmedElkoussy

NWO BRAINSCAPES CHAN ZUCKERBERG INITIATIVE IMID ZonMw Single-cell eQTL Consortium