Celltype prediction

Åsa Björklund

asa.bjorklund@scilifelab.se

Ahmed Mahfouz

Leiden University Medical Center / TU Delft

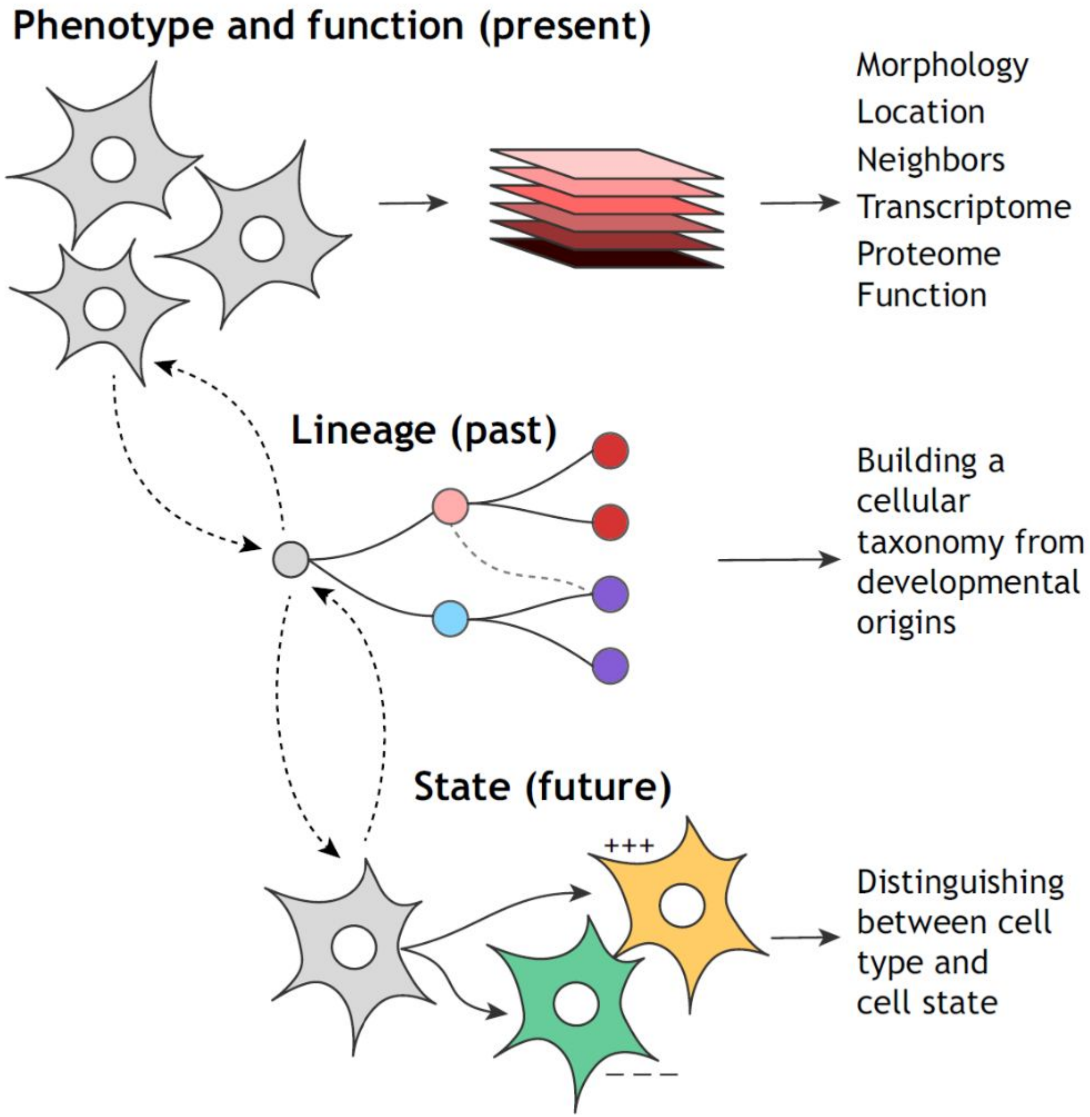
NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Outline

- Introduction
- Normalization
- Removal of confounders
- Gene set selection

Cell identity



Phenotype and function (present)

Morphology
Location
Neighbors
Transcriptome
Proteome
Function

Lineage (past)

Building a cellular taxonomy from developmental origins

State (future)

Distinguishing between cell type and cell state

(Morris, *Development* 2019)

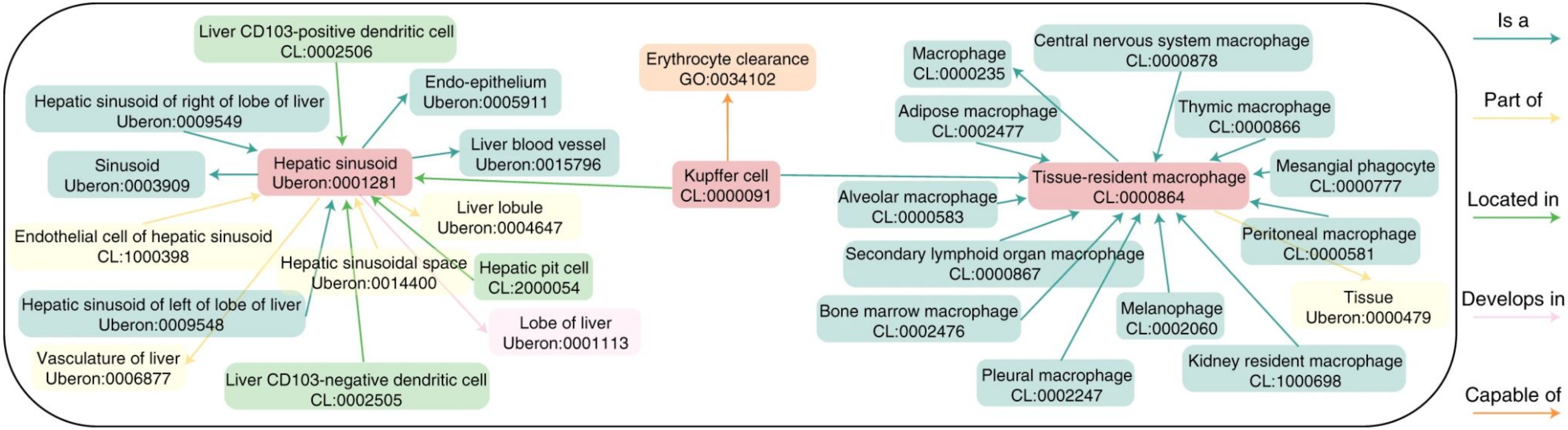
Why do we want to classify celltypes?

- In a novel tissue - what celltypes are there?
- Compare same celltype across conditions.
- Compare abundance of celltypes across conditions.
- Infer communication between celltypes
- …..

Celltype ontologies

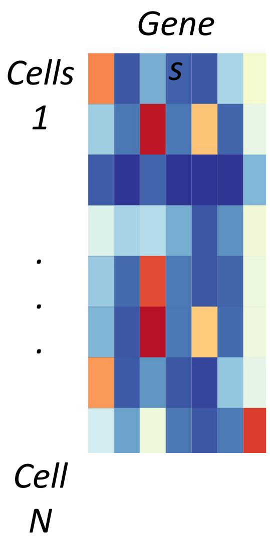
We need a standardized way of classifying celltypes.
Mainly driven by cell atlas projects.

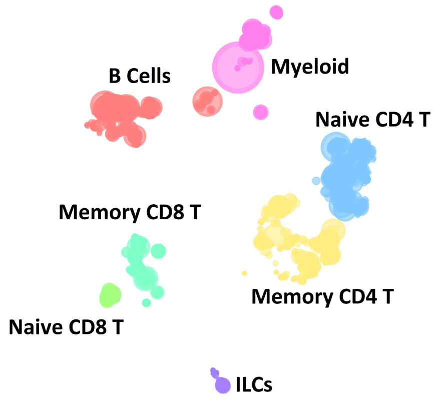
Including HuBMAP, Human Cell Atlas (HCA), cellxgene, Single Cell Expression Atlas, BRAIN Initiative Cell Census Network (BICCN), ArrayExpress, The Cell Image Library, ENCODE, and FANTOM5,


Liver CD103-positive dendritic cell CL:0002506
Hepatic sinusoid of right of lobe of liver Uberon:0009549
Endo-epithelium Uberon:0005911
Erythrocyte clearance GO:0034102
Macrophage CL:0000235
Central nervous system macrophage CL:0000878
Adipose macrophage CL:0002477
Thymic macrophage CL:0000866
Sinusoid Uberon:0003909
Hepatic sinusoid Uberon:0001281
Liver blood vessel Uberon:0015796
Kupffer cell CL:0000091
Tissue-resident macrophage CL:0000864
Mesangial phagocyte CL:0000777
Alveolar macrophage CL:0000583
Liver lobule Uberon:0004647
Endothelial cell of hepatic sinusoid CL:1000398
Hepatic sinusoidal space Uberon:0014400
Hepatic pit cell CL:2000054
Secondary lymphoid organ macrophage CL:0000867
Peritoneal macrophage CL:0000581
Hepatic sinusoid of left of lobe of liver Uberon:0009548
Lobe of liver Uberon:0001113
Bone marrow macrophage CL:0002476
Melanophage CL:0002060
Tissue Uberon:0000479
Vasculature of liver Uberon:0006877
Liver CD103-negative dendritic cell CL:0002505
Pleural macrophage CL:0002247
Kidney resident macrophage CL:1000698
Is a
Part of
Located in
Develops in
Capable of

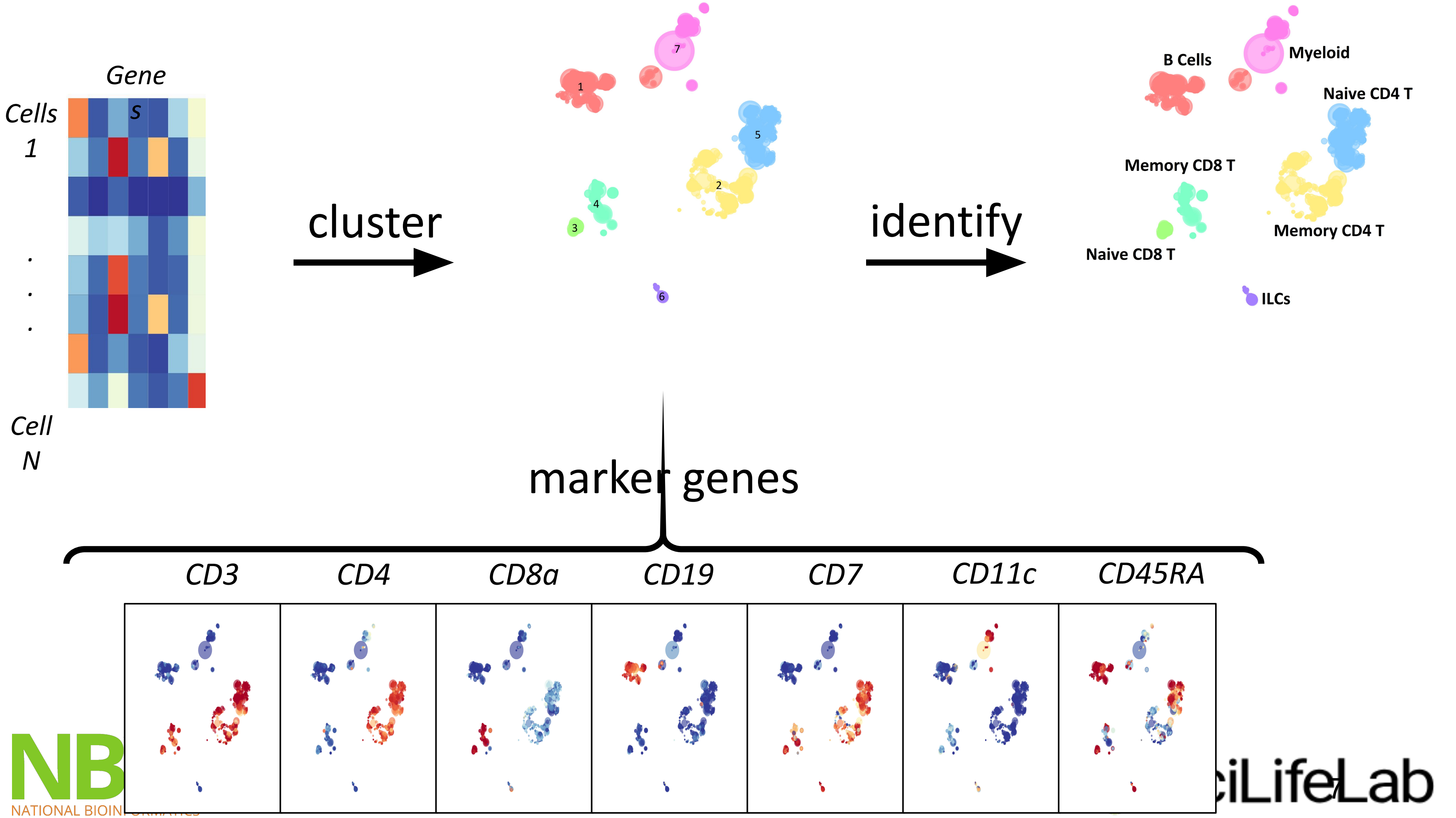
(Osumi-Sutherland, Nature Cell Biol 2021)

How can we identify cell populations?





How can we identify cell populations?



cluster

identify

marker genes

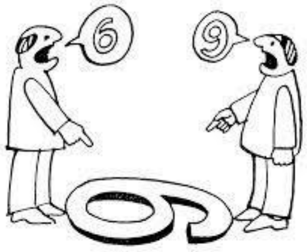
CD3 *CD4* *CD8a* *CD19* *CD7* *CD11c* *CD45RA*

Unsupervised celltype identification is problematic

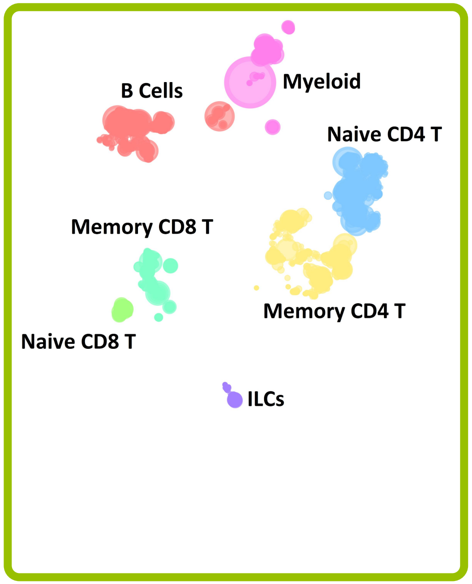
Time consuming

Not reproducible

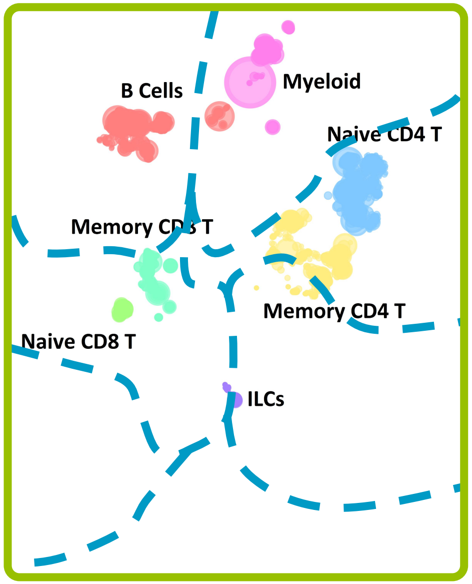
Subjective



Can we automatically identify cell populations?



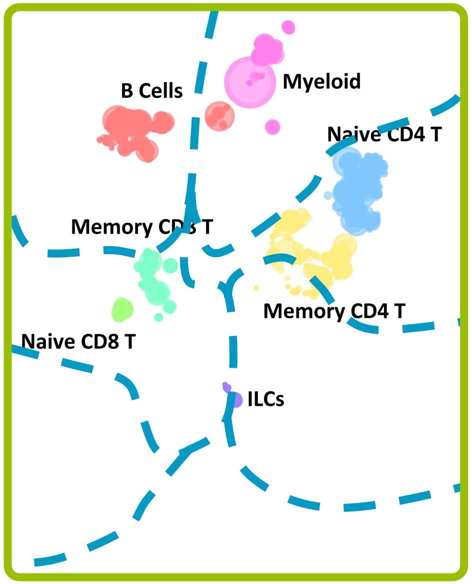
Can we automatically identify cell populations?



Can we automatically identify cell populations?

Clustering

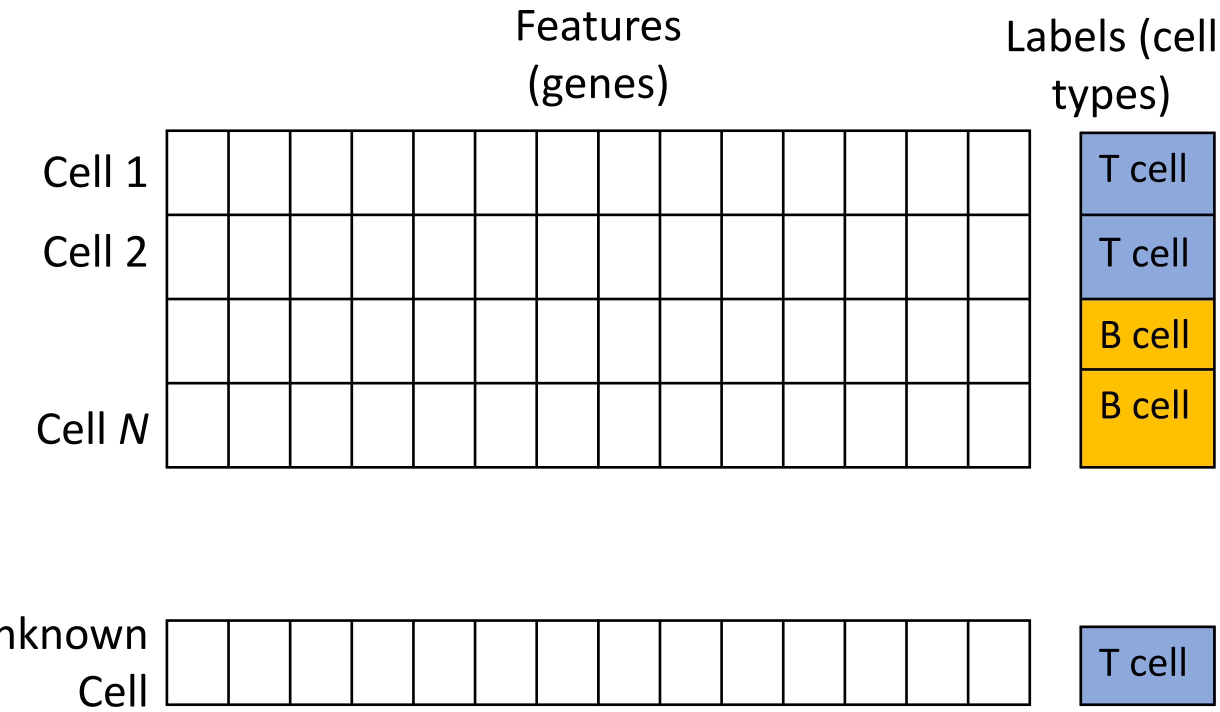
- Unsupervised learning
- Discovering structure/relations
- Clusters are defined by a decision boundary

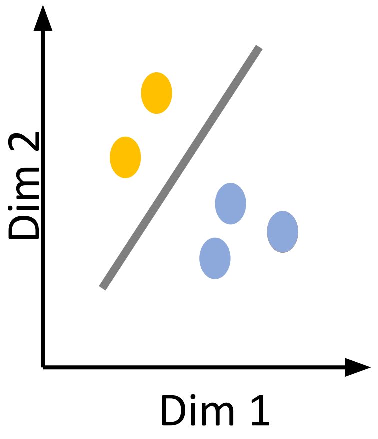


Classification

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

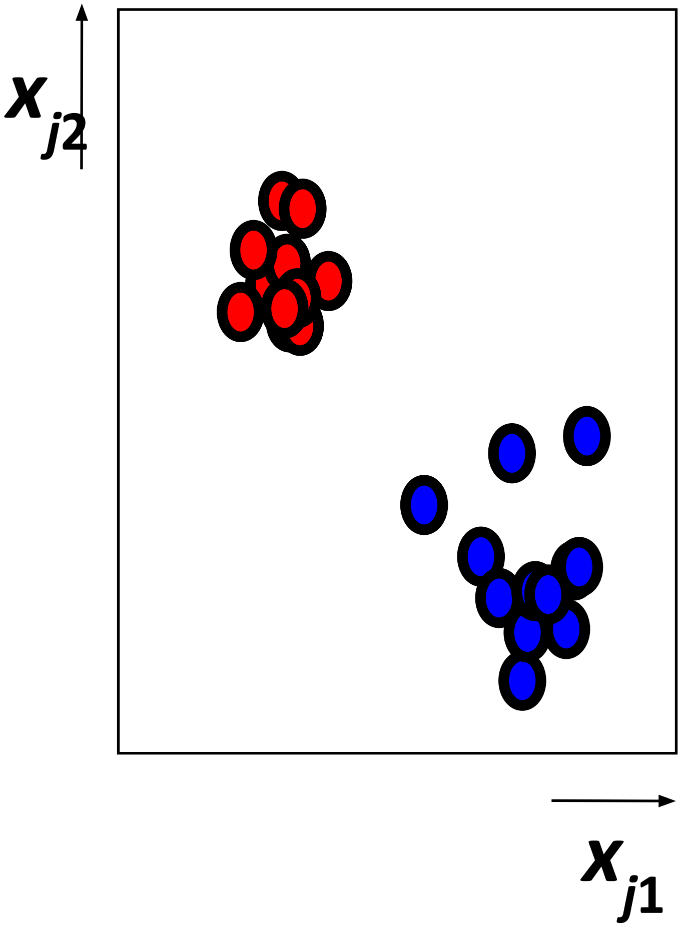
Classification





Classifier training

- Dataset: for *j* th cell:
  - gene expressions $\boldsymbol{x}_j$
  - class label: $y_j \in \{1=\mathrm{T}, -1=\mathrm{B}\}$
- 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

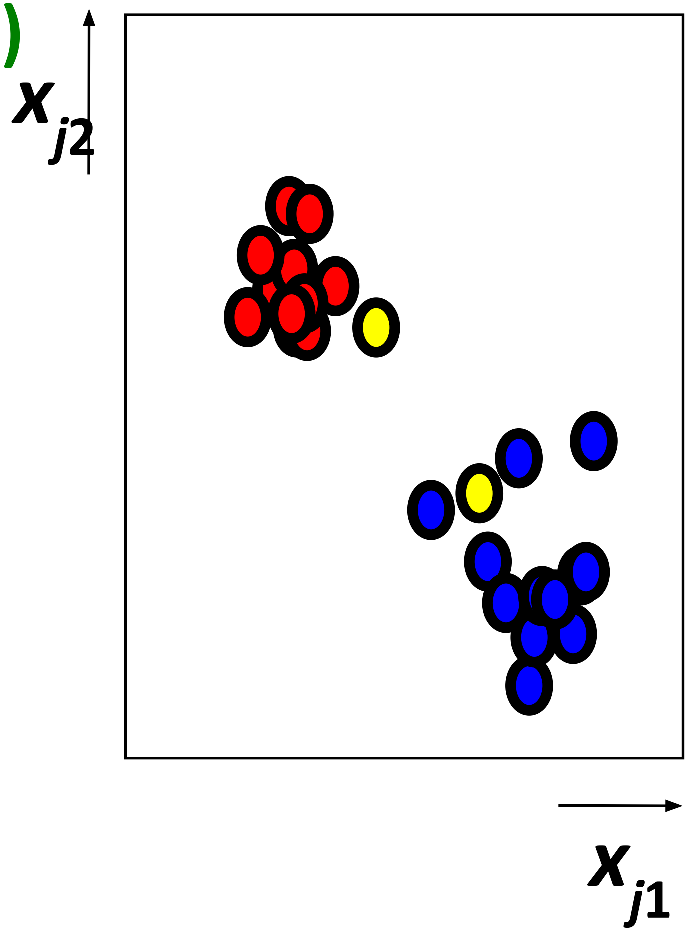


Instance Based Learning (Lazy Classification)

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

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

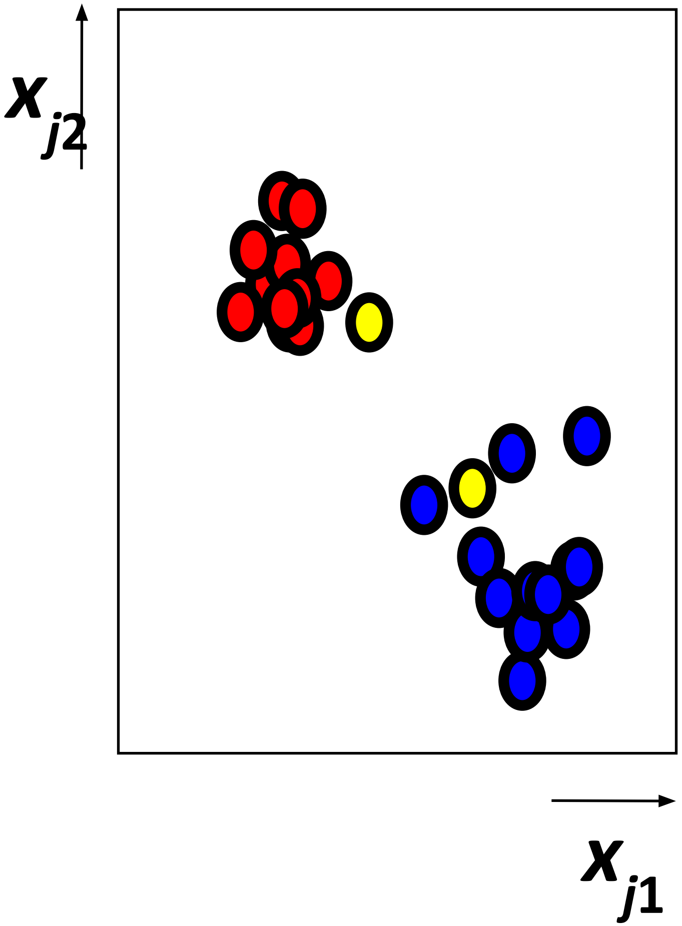
- *No actual training*



Nearest Neighbor (k-NN)

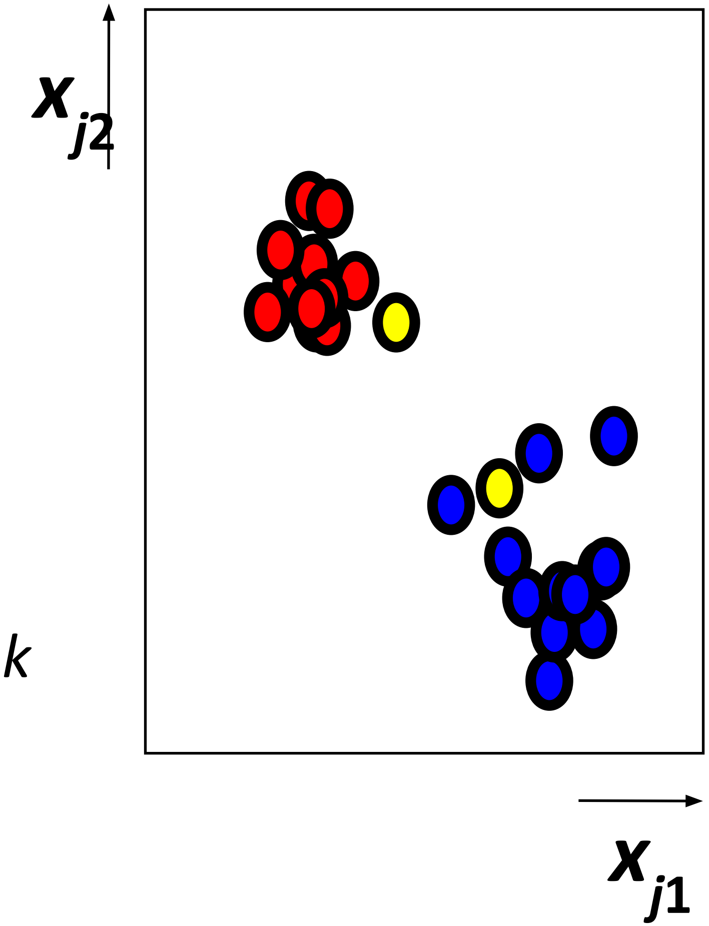
- To make Nearest Neighbor work we need 4 things:

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

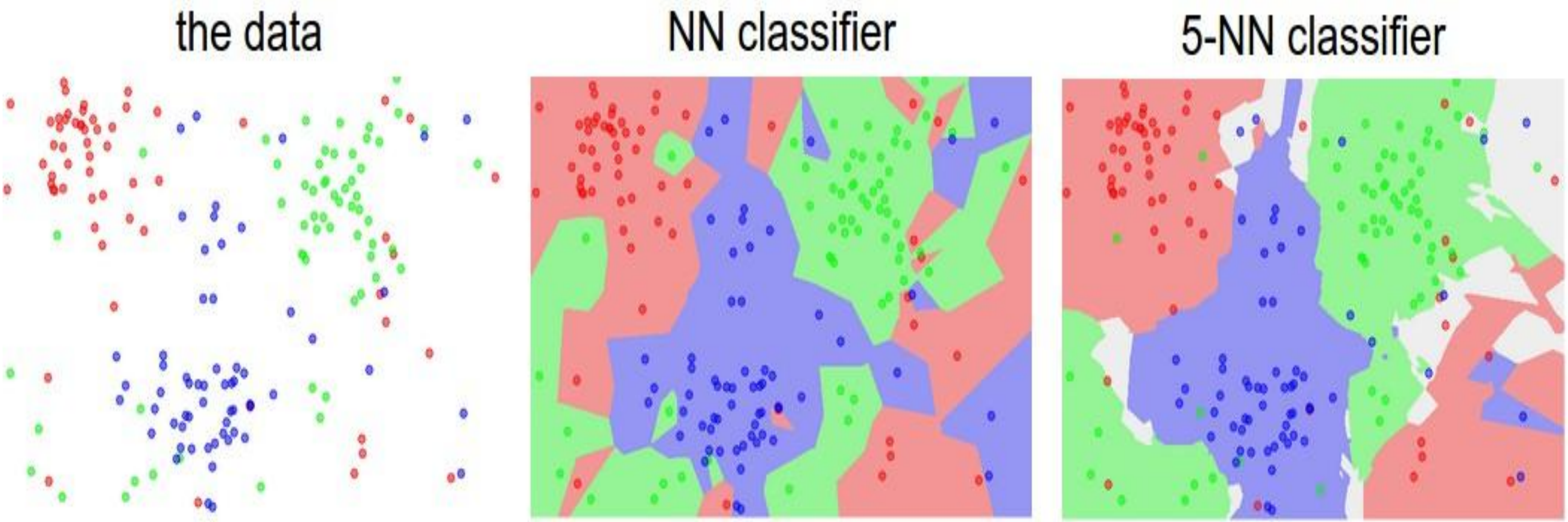


Nearest Neighbor (k-NN)

- 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



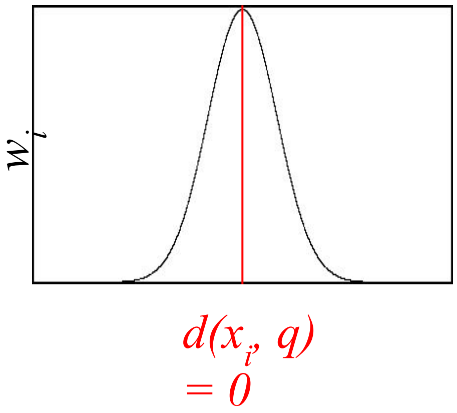
Effect of *k*



Weighted Nearest Neighbor (kernel regression)

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

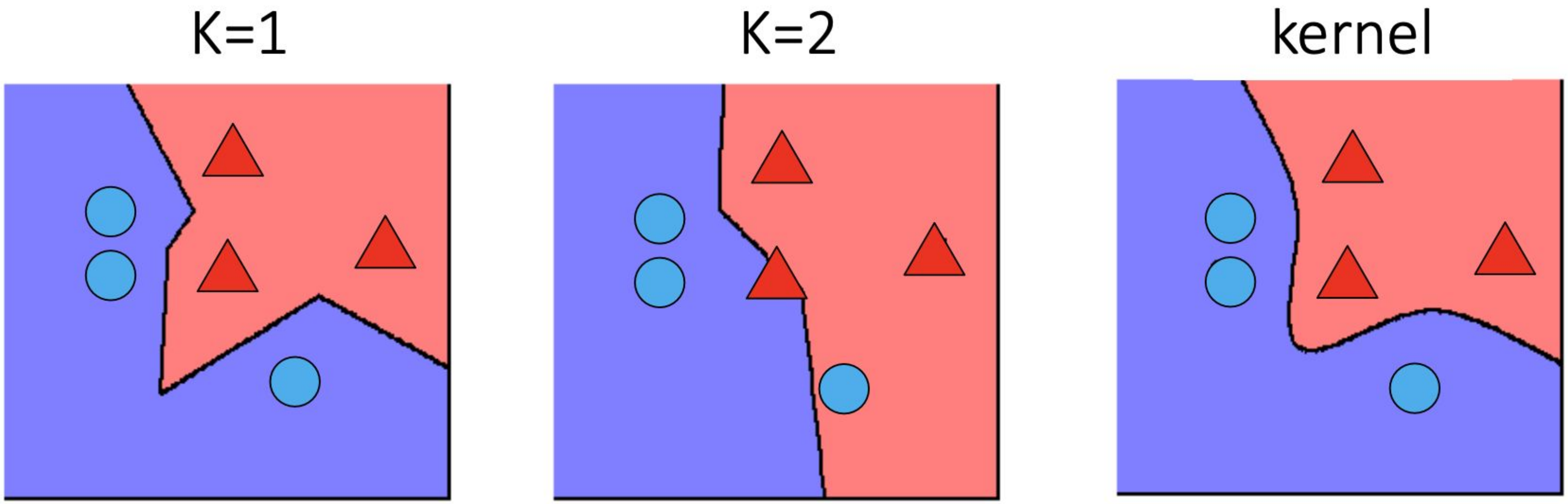
$$w_i = \exp\left(-\frac{d(x_i,q)^2}{K_w}\right)$$



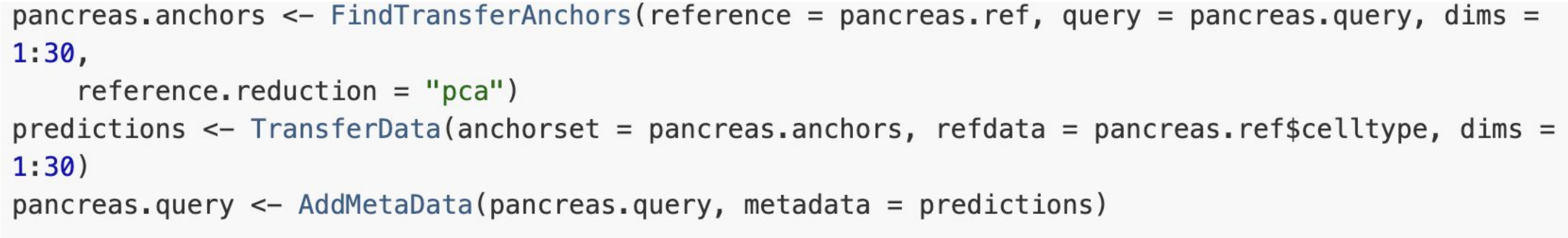
  - Nearby points to a query q are weighted more strongly. K_W: kernel width
- How to fit with the local points?
  - Predict the weighted average

$$\frac{\sum_i w_i y_i}{\sum_i w_i}$$

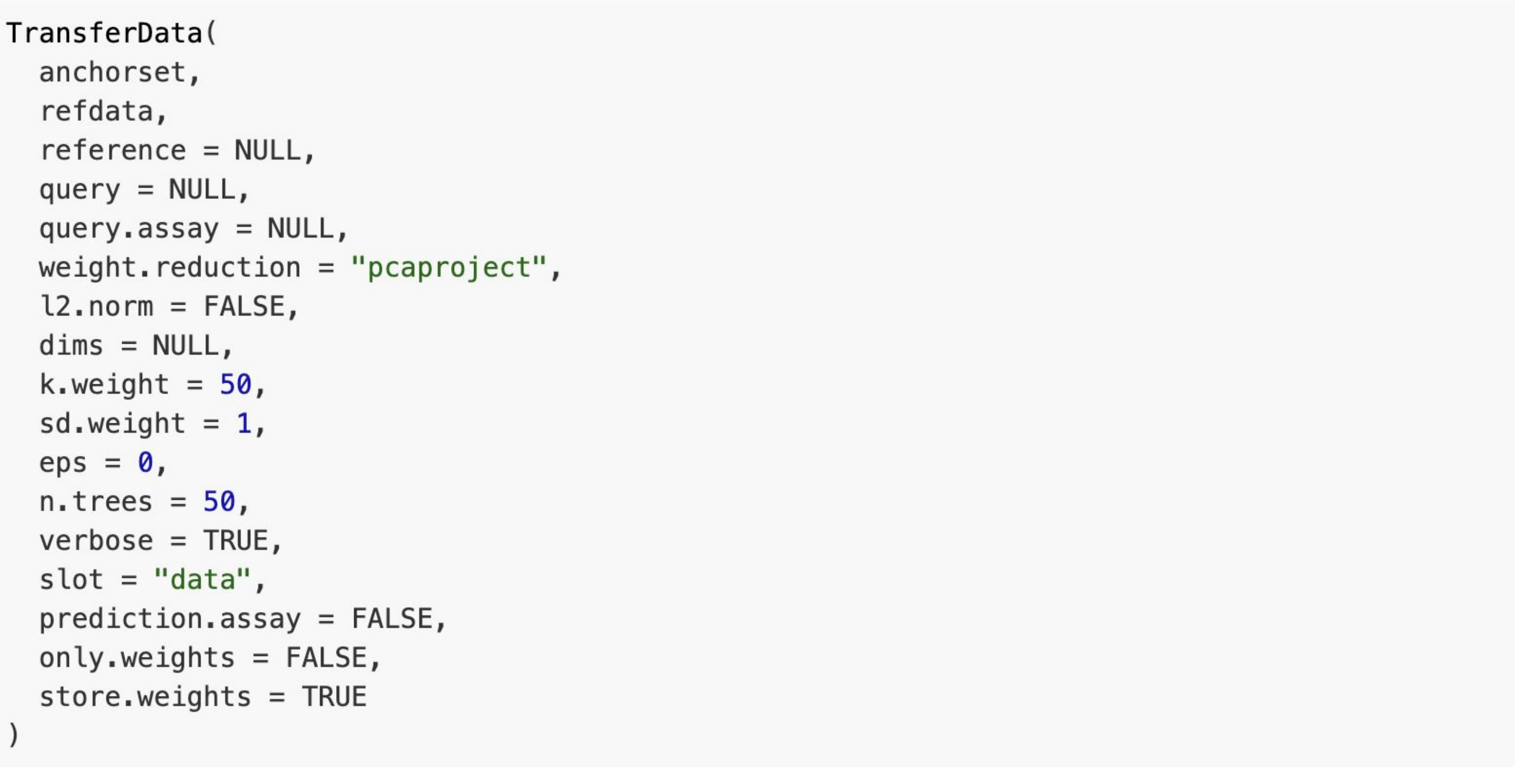
Comparison: K=1, K=2, kernel



Seurat data transfer

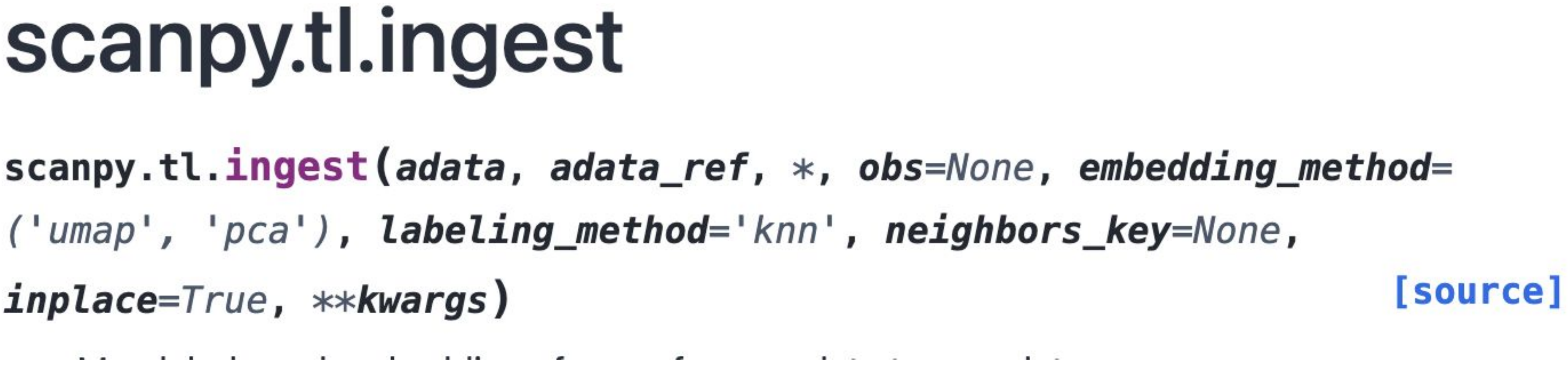


```
pancreas.anchors <- FindTransferAnchors(reference = pancreas.ref, query = pancreas.query, dims =
1:30,
    reference.reduction = "pca")
predictions <- TransferData(anchorset = pancreas.anchors, refdata = pancreas.ref$celltype, dims =
1:30)
pancreas.query <- AddMetaData(pancreas.query, metadata = predictions)
```



```
TransferData(
  anchorset,
  refdata,
  reference = NULL,
  query = NULL,
  query.assay = NULL,
  weight.reduction = "pcaproject",
  l2.norm = FALSE,
  dims = NULL,
  k.weight = 50,
  sd.weight = 1,
  eps = 0,
  n.trees = 50,
  verbose = TRUE,
  slot = "data",
  prediction.assay = FALSE,
  only.weights = FALSE,
  store.weights = TRUE
)
```

Scanpy data transfer

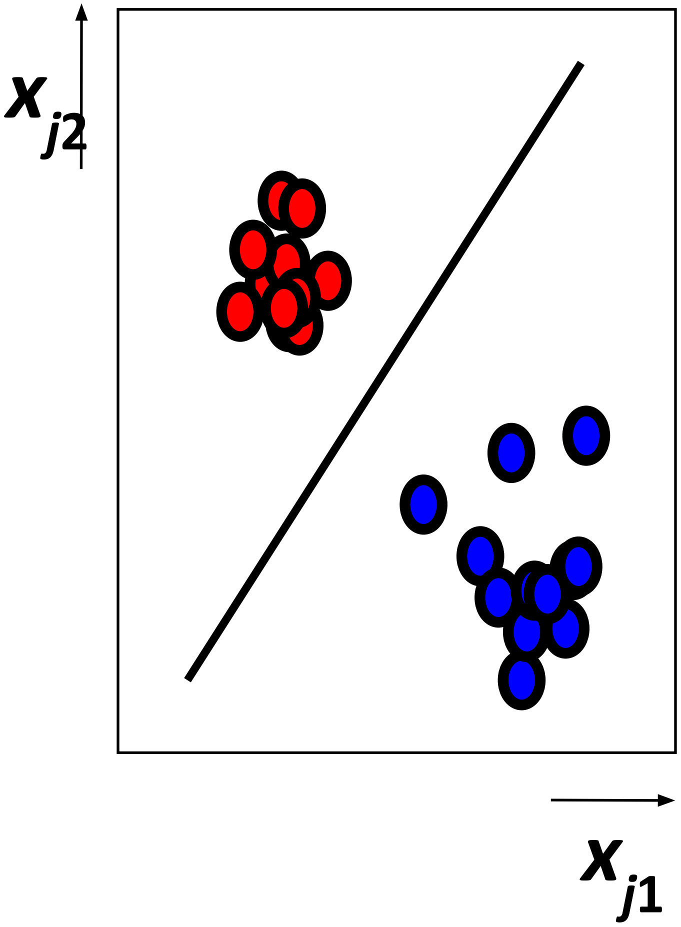


scanpy.tl.ingest

```
scanpy.tl.ingest(adata, adata_ref, *, obs=None, embedding_method=
('umap', 'pca'), labeling_method='knn', neighbors_key=None,
inplace=True, **kwargs)
```

[source]

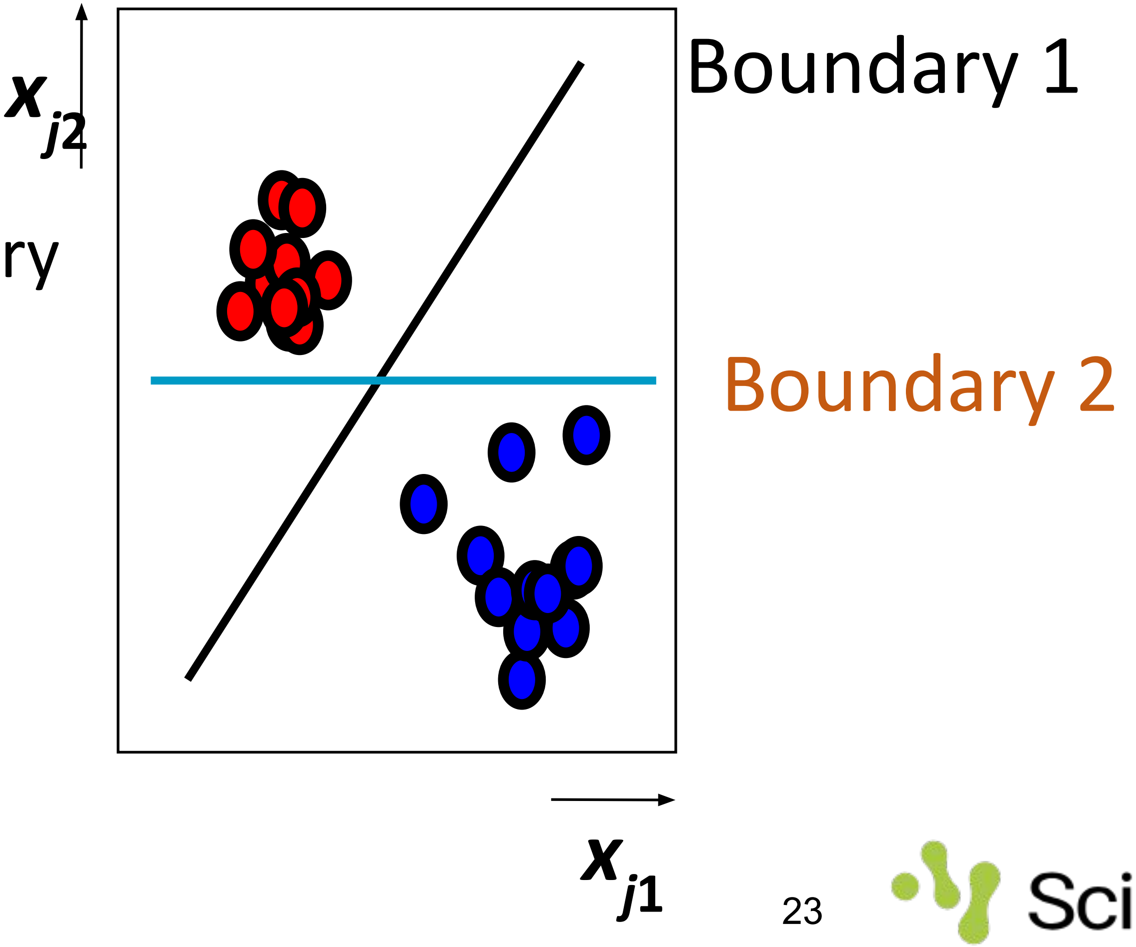
Support Vector Machine (SVM)



Decision Boundary

Support Vector Machine (SVM)

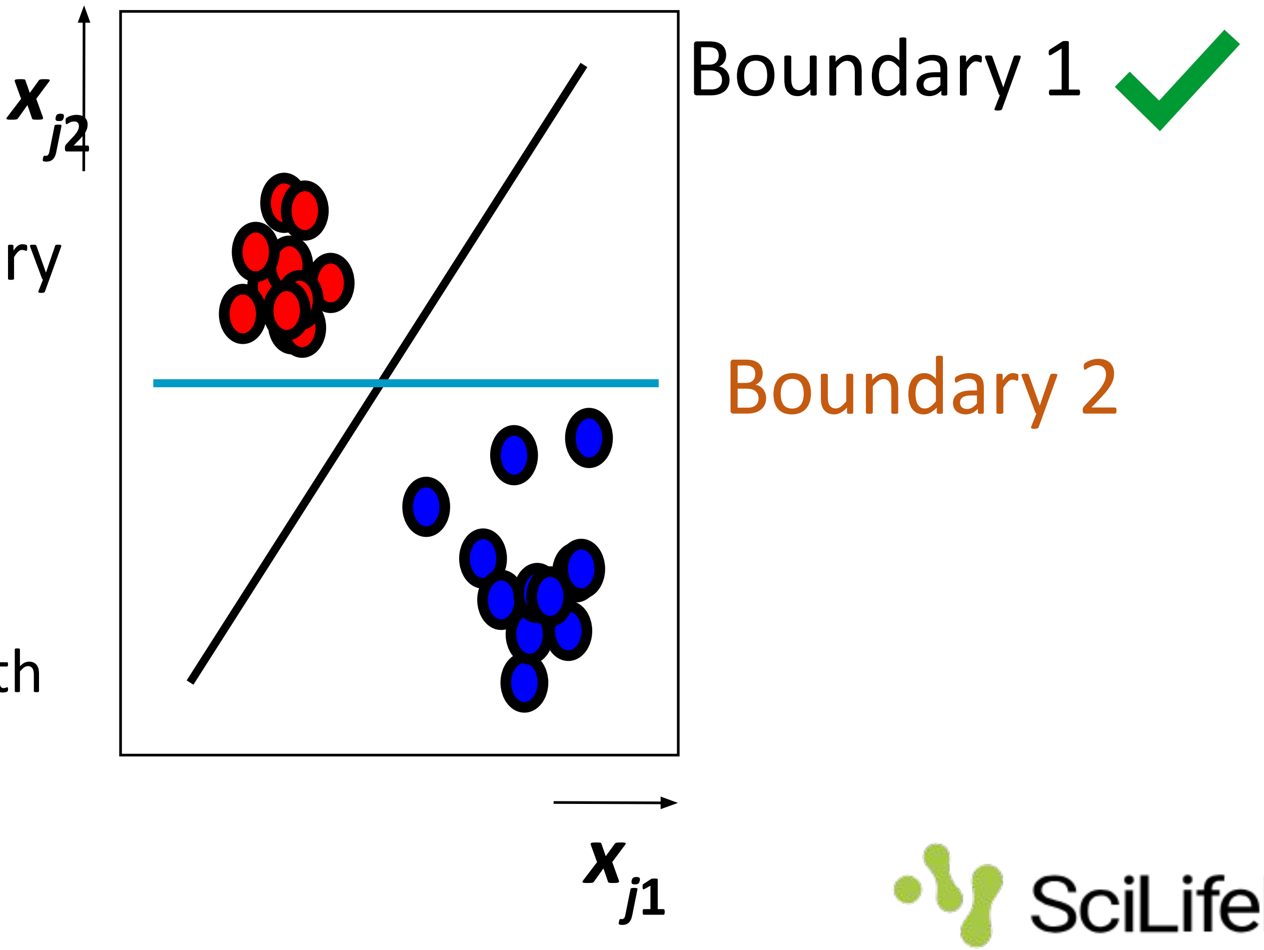
Which boundary is better?



Support Vector Machine (SVM)

Which boundary is better?

The one that maximizes the margins from both labels.



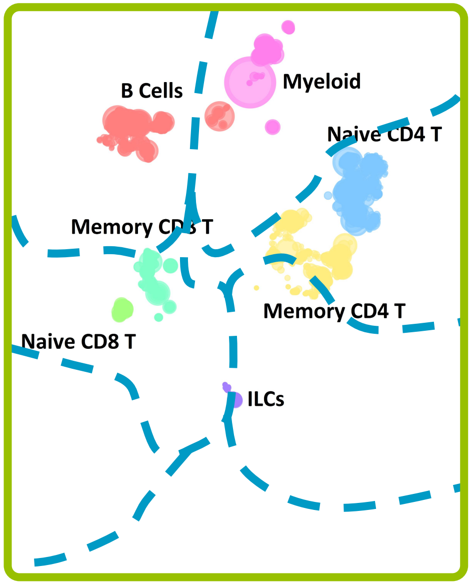
Boundary 1 ✓

Boundary 2

Can we automatically identify cell populations?

Training data

Annotated Cells (e.g. atlas)



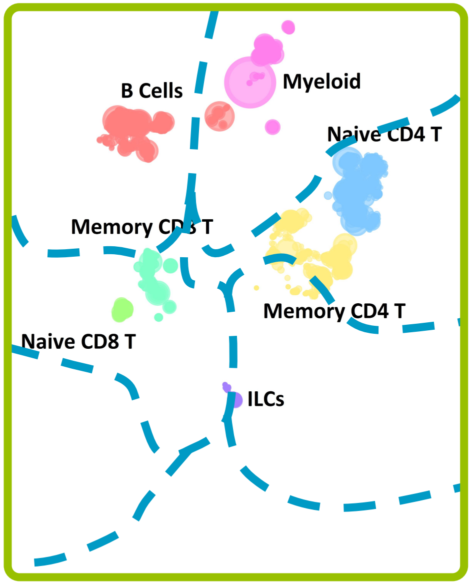
Can we automatically identify cell populations?

Training data

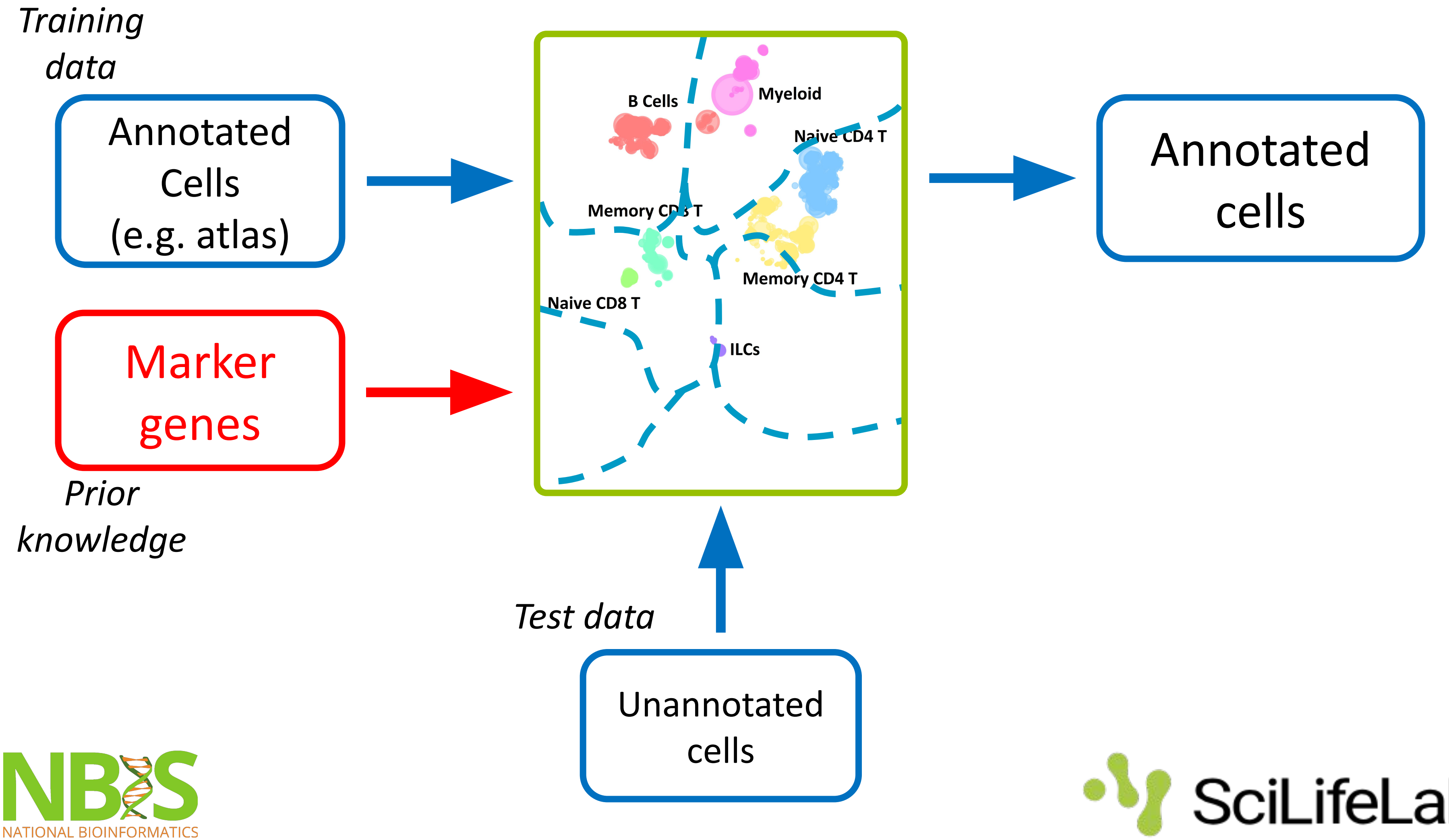
Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge



Can we automatically identify cell populations?



Training data

Annotated Cells (e.g. atlas)

Marker genes

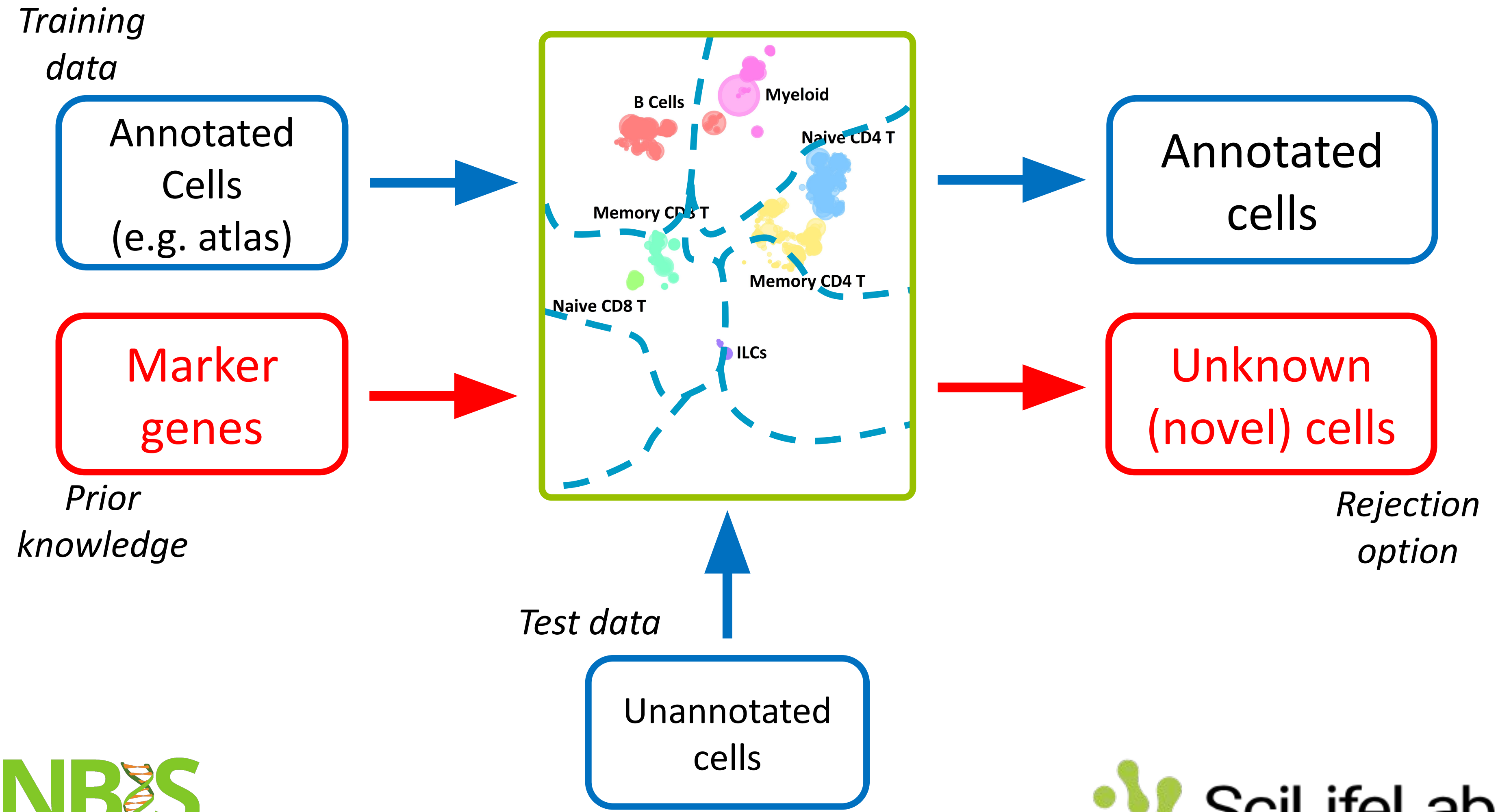
Prior knowledge

Test data

Unannotated cells

Annotated cells

Can we automatically identify cell populations?



Training data

Annotated Cells (e.g. atlas)

Marker genes

Prior knowledge

Test data

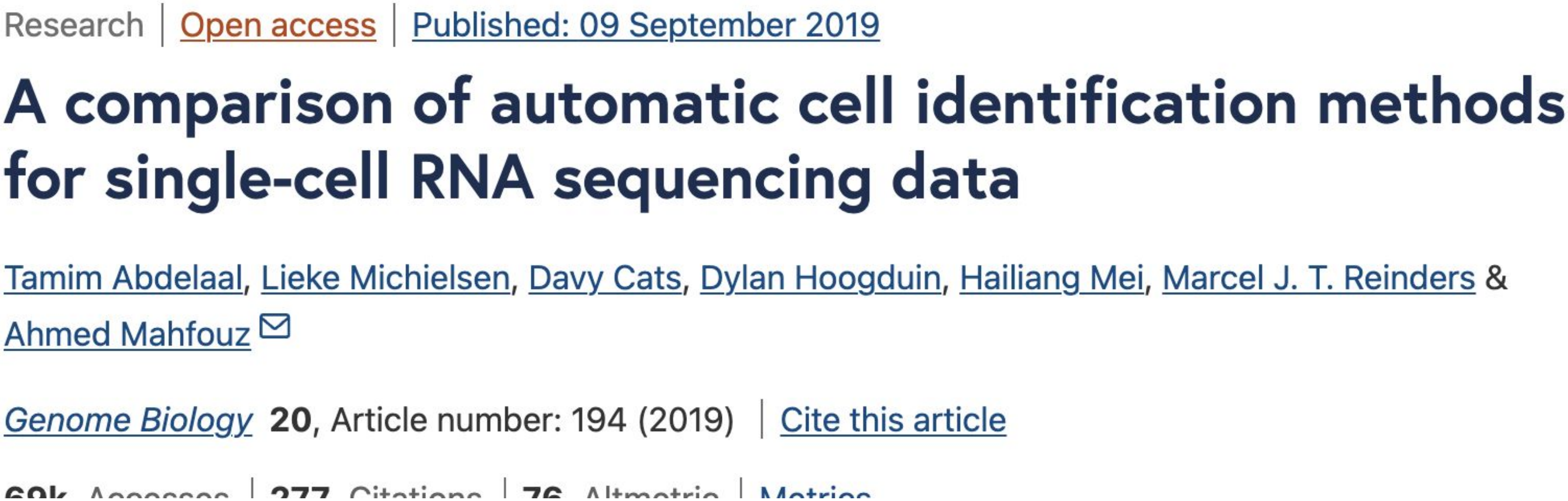
Unannotated cells

Annotated cells

Unknown (novel) cells

Rejection option

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

69k Accesses | 277 Citations | 76 Altmetric | Metrics

(Abdelaal et al. Genome Biology 2019)

16 existing classifiers (April 2019)

CaSTLe

LAmbDA

scPred

scmap$_{\text{cluster}}$

Moana

SingleCell
Net

CHETAH

SingleR

Garnett

scmap$_{\text{cell}}$

SCINA

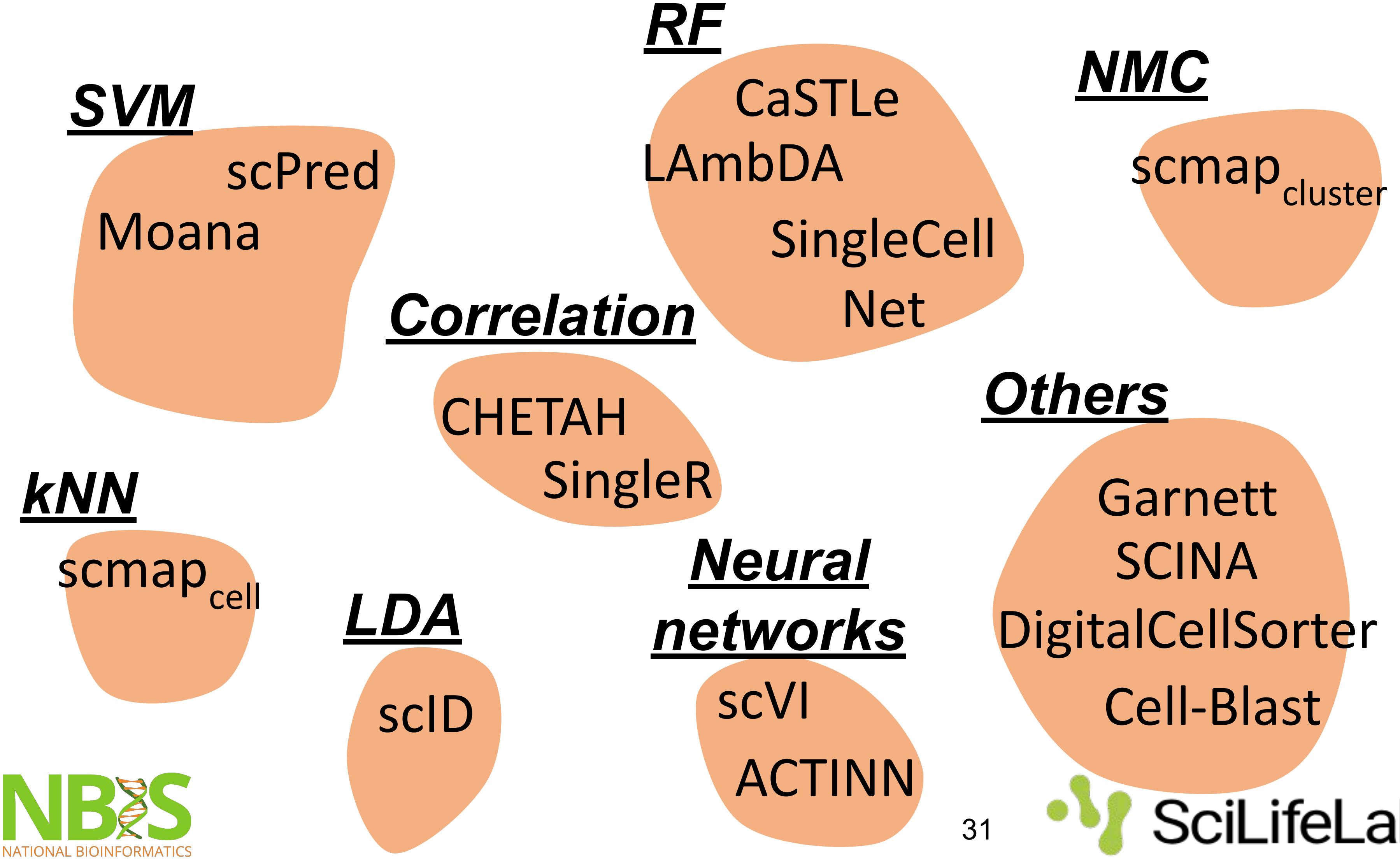
DigitalCellSorter

scVI

scID

Cell-Blast

ACTINN



SVM

scPred

Moana

_RF_

CaSTLe

LAmbDA

SingleCellNet

NMC

$\text{scmap}_{\text{cluster}}$

Correlation

CHETAH

SingleR

kNN

$\text{scmap}_{\text{cell}}$

LDA

scID

Neural networks

scVI

ACTINN

Others

Garnett

SCINA

DigitalCellSorter

Cell-Blast

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

16 existing + 6 off-the-shelf classifiers

SVM

scPred

Moana

SVM

$\text{SVM}_{\text{rejection}}$

RF

CaSTLe

LAmbDA

SingleCell Net

RF

NMC

$\text{scmap}_{\text{cluster}}$

NMC

Correlation

CHETAH

SingleR

kNN

$\text{scmap}_{\text{cell}}$

kNN

LDA

scID

LDA

Neural networks

scVI

ACTINN

Others

Garnett

SCINA

DigitalCellSorter

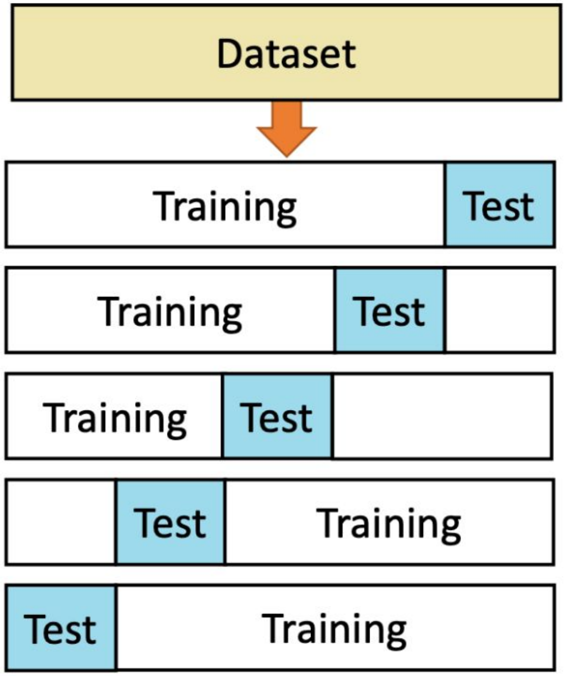
Cell-Blast

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Experiment 1: intra-dataset evaluation

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



Dataset

Training Test

Training Test

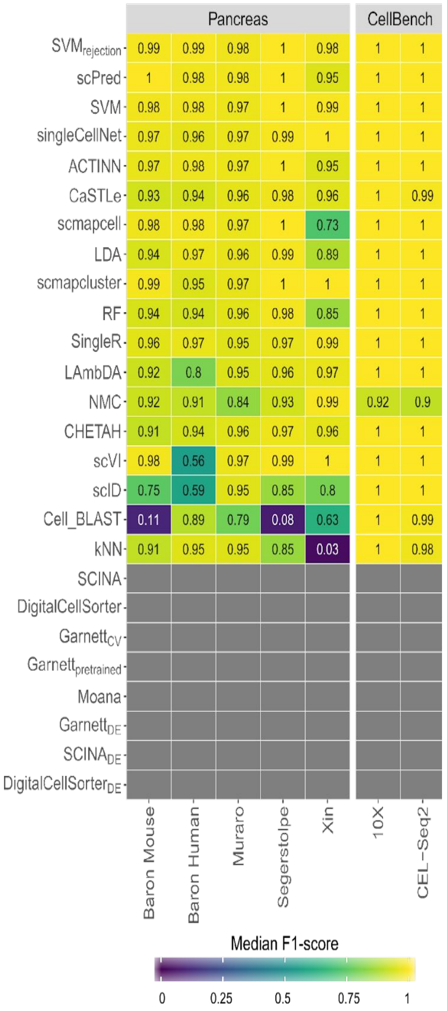
Training Test

Test Training

Test Training

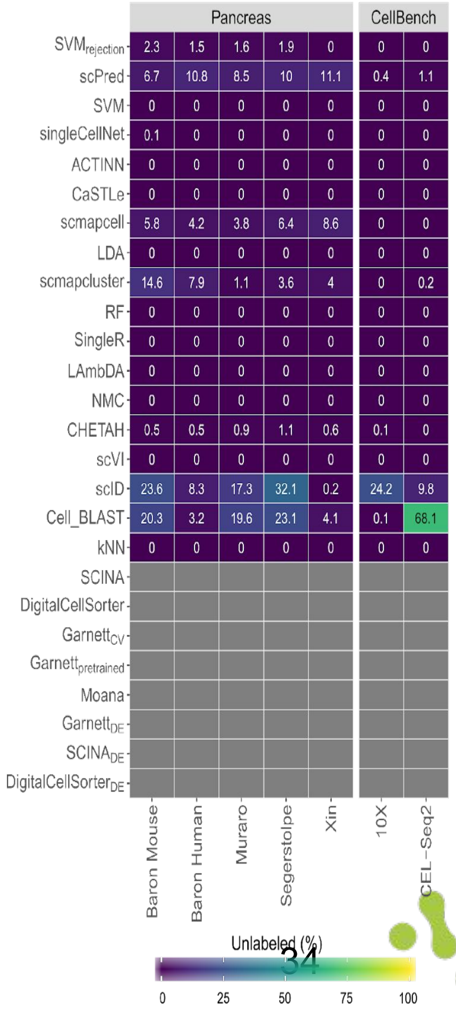
Most classifiers work well

Median F1-score



	Pancreas					CellBench	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2
SVM$_{rejection}$	0.99	0.99	0.98	1	0.98	1	1
scPred	1	0.98	0.98	1	0.95	1	1
SVM	0.98	0.98	0.97	1	0.99	1	1
singleCellNet	0.97	0.96	0.97	0.99	1	1	1
ACTINN	0.97	0.98	0.97	1	0.95	1	1
CaSTLe	0.93	0.94	0.96	0.98	0.96	1	0.99
scmapcell	0.98	0.98	0.97	1	0.73	1	1
LDA	0.94	0.97	0.96	0.99	0.89	1	1
scmapcluster	0.99	0.95	0.97	1	1	1	1
RF	0.94	0.94	0.96	0.98	0.85	1	1
SingleR	0.96	0.97	0.95	0.97	0.99	1	1
LAmbDA	0.92	0.8	0.95	0.96	0.97	1	1
NMC	0.92	0.91	0.84	0.93	0.99	0.92	0.9
CHETAH	0.91	0.94	0.96	0.97	0.96	1	1
scVI	0.98	0.56	0.97	0.99	1	1	1
scID	0.75	0.59	0.95	0.85	0.8	1	1
Cell_BLAST	0.11	0.89	0.79	0.08	0.63	1	0.99
kNN	0.91	0.95	0.95	0.85	0.03	1	0.98
SCINA							
DigitalCellSorter							
Garnett$_{CV}$							
Garnett$_{pretrained}$							
Moana							
Garnett$_{DE}$							
SCINA$_{DE}$							
DigitalCellSorter$_{DE}$							

% Unlabeled

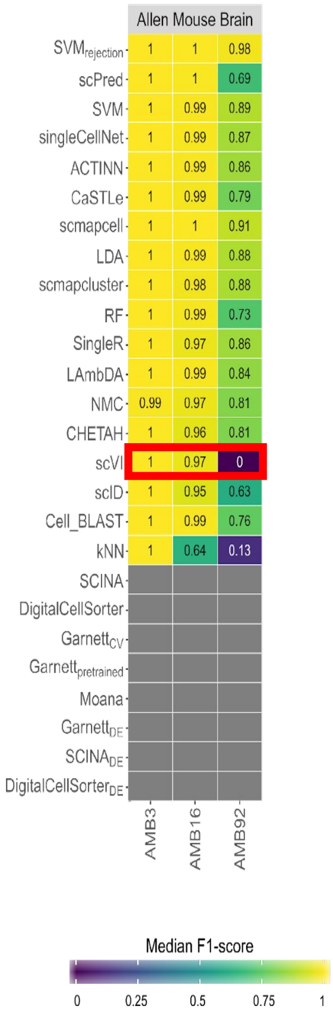


	Pancreas					CellBench	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2
SVM$_{rejection}$	2.3	1.5	1.6	1.9	0	0	0
scPred	6.7	10.8	8.5	10	11.1	0.4	1.1
SVM	0	0	0	0	0	0	0
singleCellNet	0.1	0	0	0	0	0	0
ACTINN	0	0	0	0	0	0	0
CaSTLe	0	0	0	0	0	0	0
scmapcell	5.8	4.2	3.8	6.4	8.6	0	0
LDA	0	0	0	0	0	0	0
scmapcluster	14.6	7.9	1.1	3.6	4	0	0.2
RF	0	0	0	0	0	0	0
SingleR	0	0	0	0	0	0	0
LAmbDA	0	0	0	0	0	0	0
NMC	0	0	0	0	0	0	0
CHETAH	0.5	0.5	0.9	1.1	0.6	0.1	0
scVI	0	0	0	0	0	0	0
scID	23.6	8.3	17.3	32.1	0.2	24.2	9.8
Cell_BLAST	20.3	3.2	19.6	23.1	4.1	0.1	68.1
kNN	0	0	0	0	0	0	0
SCINA							
DigitalCellSorter							
Garnett$_{CV}$							
Garnett$_{pretrained}$							
Moana							
Garnett$_{DE}$							
SCINA$_{DE}$							
DigitalCellSorter$_{DE}$							

NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

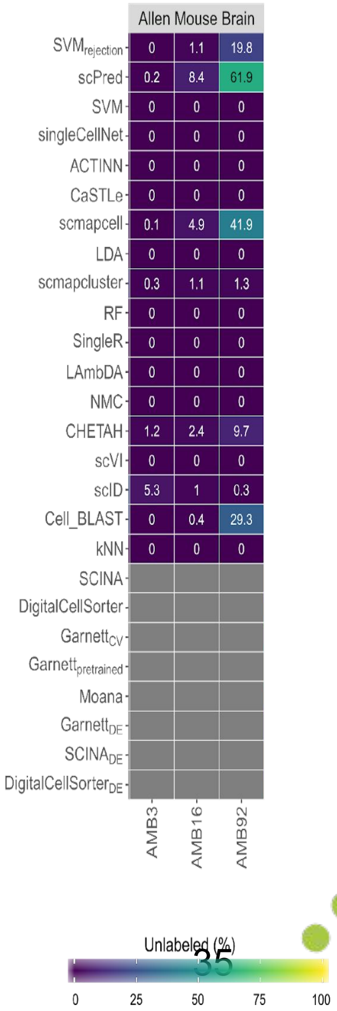
Performance drops with deeper annotation

Median F1-score



	Allen Mouse Brain		
	AMB3	AMB16	AMB92
$SVM_{rejection}$	1	1	0.98
scPred	1	1	0.69
SVM	1	0.99	0.89
singleCellNet	1	0.99	0.87
ACTINN	1	0.99	0.86
CaSTLe	1	0.99	0.79
scmapcell	1	1	0.91
LDA	1	0.99	0.88
scmapcluster	1	0.98	0.88
RF	1	0.99	0.73
SingleR	1	0.97	0.86
LAmbDA	1	0.99	0.84
NMC	0.99	0.97	0.81
CHETAH	1	0.96	0.81
scVI	1	0.97	0
scID	1	0.95	0.63
Cell_BLAST	1	0.99	0.76
kNN	1	0.64	0.13
SCINA			
DigitalCellSorter			
$Garnett_{CV}$			
$Garnett_{pretrained}$			
Moana			
$Garnett_{DE}$			
$SCINA_{DE}$			
$DigitalCellSorter_{DE}$			

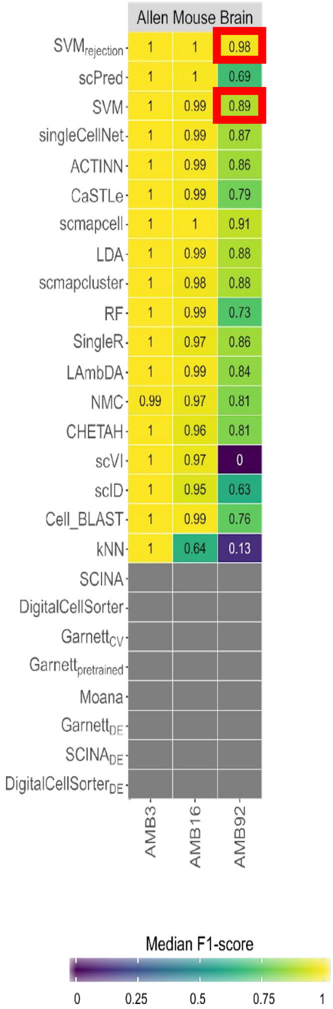
% Unlabeled



	Allen Mouse Brain		
	AMB3	AMB16	AMB92
$SVM_{rejection}$	0	1.1	19.8
scPred	0.2	8.4	61.9
SVM	0	0	0
singleCellNet	0	0	0
ACTINN	0	0	0
CaSTLe	0	0	0
scmapcell	0.1	4.9	41.9
LDA	0	0	0
scmapcluster	0.3	1.1	1.3
RF	0	0	0
SingleR	0	0	0
LAmbDA	0	0	0
NMC	0	0	0
CHETAH	1.2	2.4	9.7
scVI	0	0	0
scID	5.3	1	0.3
Cell_BLAST	0	0.4	29.3
kNN	0	0	0
SCINA			
DigitalCellSorter			
$Garnett_{CV}$			
$Garnett_{pretrained}$			
Moana			
$Garnett_{DE}$			
$SCINA_{DE}$			
$DigitalCellSorter_{DE}$			

Trade-off between high performance and rejecting cells

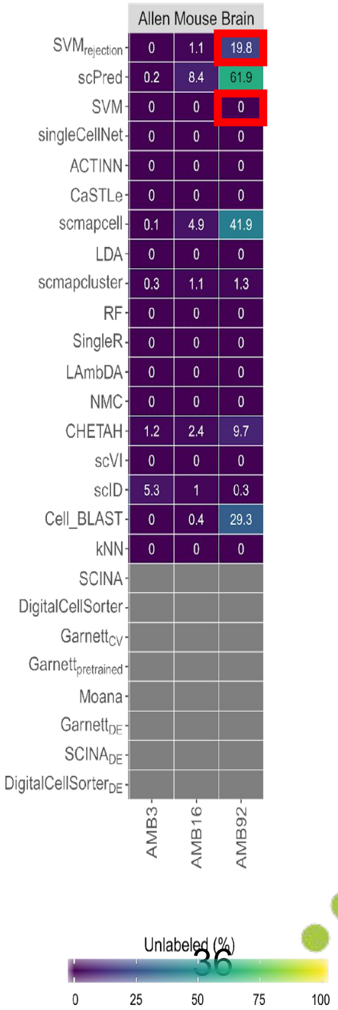
Median F1-score



Allen Mouse Brain

	AMB3	AMB16	AMB92
SVM$_{rejection}$	1	1	0.98
scPred	1	1	0.69
SVM	1	0.99	0.89
singleCellNet	1	0.99	0.87
ACTINN	1	0.99	0.86
CaSTLe	1	0.99	0.79
scmapcell	1	1	0.91
LDA	1	0.99	0.88
scmapcluster	1	0.98	0.88
RF	1	0.99	0.73
SingleR	1	0.97	0.86
LAmbDA	1	0.99	0.84
NMC	0.99	0.97	0.81
CHETAH	1	0.96	0.81
scVI	1	0.97	0
scID	1	0.95	0.63
Cell_BLAST	1	0.99	0.76
kNN	1	0.64	0.13
SCINA			
DigitalCellSorter			
Garnett$_{CV}$			
Garnett$_{pretrained}$			
Moana			
Garnett$_{DE}$			
SCINA$_{DE}$			
DigitalCellSorter$_{DE}$			

% Unlabeled



Allen Mouse Brain

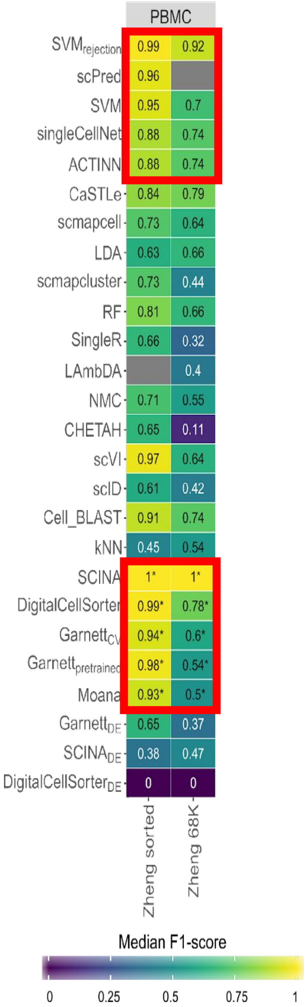
	AMB3	AMB16	AMB92
SVM$_{rejection}$	0	1.1	19.8
scPred	0.2	8.4	61.9
SVM	0	0	0
singleCellNet	0	0	0
ACTINN	0	0	0
CaSTLe	0	0	0
scmapcell	0.1	4.9	41.9
LDA	0	0	0
scmapcluster	0.3	1.1	1.3
RF	0	0	0
SingleR	0	0	0
LAmbDA	0	0	0
NMC	0	0	0
CHETAH	1.2	2.4	9.7
scVI	0	0	0
scID	5.3	1	0.3
Cell_BLAST	0	0.4	29.3
kNN	0	0	0
SCINA			
DigitalCellSorter			
Garnett$_{CV}$			
Garnett$_{pretrained}$			
Moana			
Garnett$_{DE}$			
SCINA$_{DE}$			
DigitalCellSorter$_{DE}$			

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

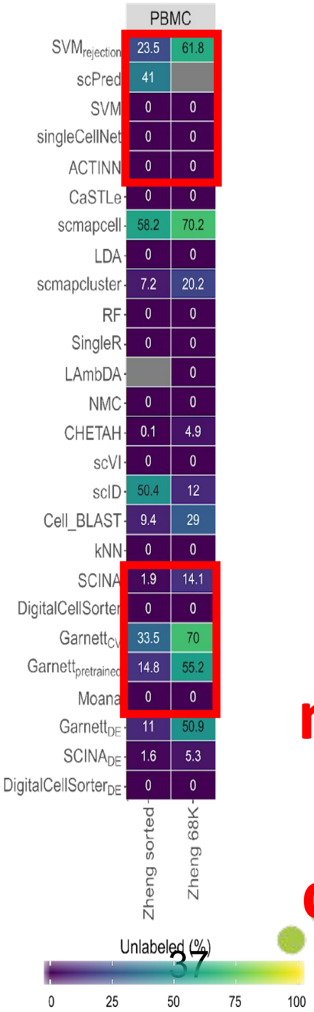
SciLifeLab

Prior knowledge is not always beneficial

Median F1-score



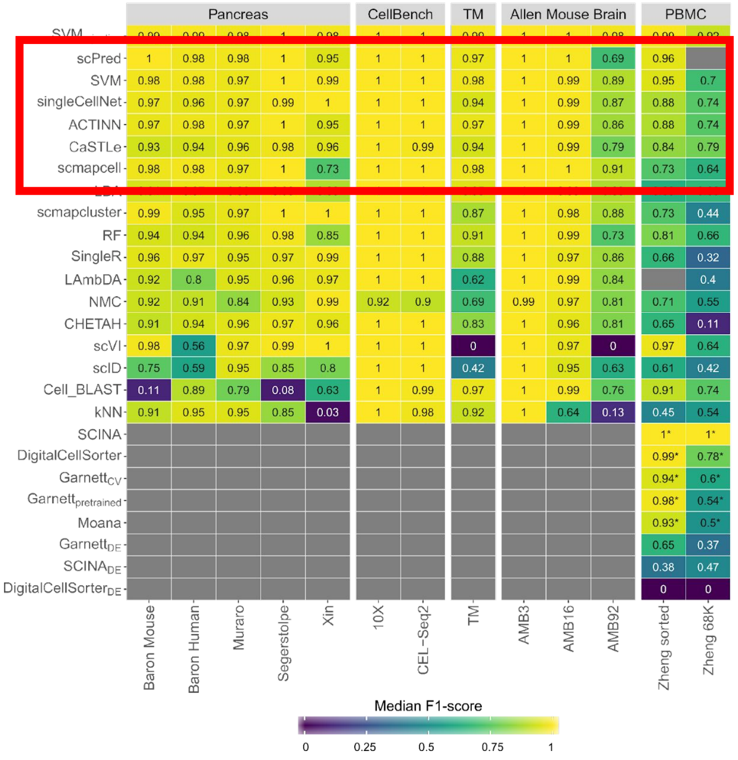
% Unlabeled



Lower number of classes!

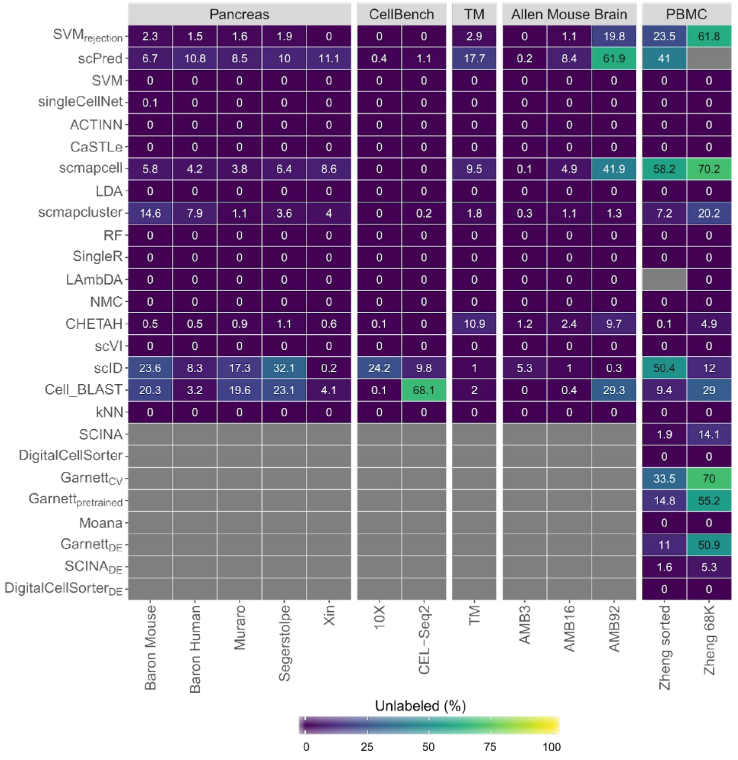
Off-the-shelf SVM outperforms dedicated single cell classifiers

Median F1-score



	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K
$SVM_{rejection}$	0.99	0.99	0.98	1	0.98	1	1	0.99	1	1	0.98	0.99	0.92
scPred	1	0.98	0.98	1	0.95	1	1	0.97	1	1	0.69	0.96	
SVM	0.98	0.98	0.97	1	0.99	1	1	0.98	1	0.99	0.89	0.95	0.7
singleCellNet	0.97	0.96	0.97	0.99	1	1	1	0.94	1	0.99	0.87	0.88	0.74
ACTINN	0.97	0.98	0.97	1	0.95	1	1	0.97	1	0.99	0.86	0.88	0.74
CaSTLe	0.93	0.94	0.96	0.98	0.96	1	0.99	0.94	1	0.99	0.79	0.84	0.79
scmapcell	0.98	0.98	0.97	1	0.73	1	1	0.98	1	1	0.91	0.73	0.64
LDA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
scmapcluster	0.99	0.95	0.97	1	1	1	1	0.87	1	0.98	0.88	0.73	0.44
RF	0.94	0.94	0.96	0.98	0.85	1	1	0.91	1	0.99	0.73	0.81	0.66
SingleR	0.96	0.97	0.95	0.97	0.99	1	1	0.88	1	0.97	0.86	0.66	0.32
LAmbDA	0.92	0.8	0.95	0.96	0.97	1	1	0.62	1	0.99	0.84		0.4
NMC	0.92	0.91	0.84	0.93	0.99	0.92	0.9	0.69	0.99	0.97	0.81	0.71	0.55
CHETAH	0.91	0.94	0.96	0.97	0.96	1	1	0.83	1	0.96	0.81	0.65	0.11
scVI	0.98	0.56	0.97	0.99	1	1	1	0	1	0.97	0	0.97	0.64
scID	0.75	0.59	0.95	0.85	0.8	1	1	0.42	1	0.95	0.63	0.61	0.42
Cell_BLAST	0.11	0.89	0.79	0.08	0.63	1	0.99	0.97	1	0.99	0.76	0.91	0.74
kNN	0.91	0.95	0.95	0.85	0.03	1	0.98	0.92	1	0.64	0.13	0.45	0.54
SCINA												1*	1*
DigitalCellSorter												0.99*	0.78*
$Garnett_{CV}$												0.94*	0.6*
$Garnett_{pretrained}$												0.98*	0.54*
Moana												0.93*	0.5*
$Garnett_{DE}$												0.65	0.37
$SCINA_{DE}$												0.38	0.47
$DigitalCellSorter_{DE}$												0	0

% Unlabeled



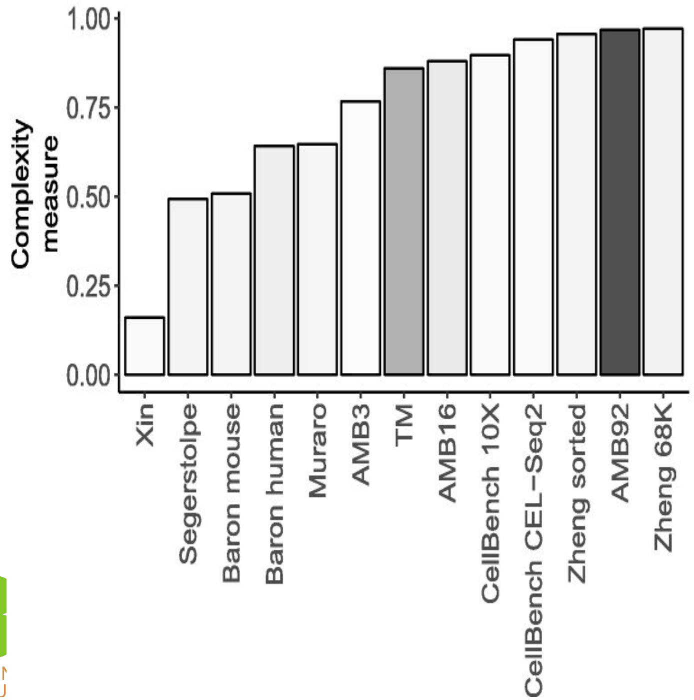
	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K
$SVM_{rejection}$	2.3	1.5	1.6	1.9	0	0	0	2.9	0	1.1	19.8	23.5	61.8
scPred	6.7	10.8	8.5	10	11.1	0.4	1.1	17.7	0.2	8.4	61.9	41	
SVM	0	0	0	0	0	0	0	0	0	0	0	0	0
singleCellNet	0.1	0	0	0	0	0	0	0	0	0	0	0	0
ACTINN	0	0	0	0	0	0	0	0	0	0	0	0	0
CaSTLe	0	0	0	0	0	0	0	0	0	0	0	0	0
scmapcell	5.8	4.2	3.8	6.4	8.6	0	0	9.5	0.1	4.9	41.9	58.2	70.2
LDA	0	0	0	0	0	0	0	0	0	0	0	0	0
scmapcluster	14.6	7.9	1.1	3.6	4	0	0.2	1.8	0.3	1.1	1.3	7.2	20.2
RF	0	0	0	0	0	0	0	0	0	0	0	0	0
SingleR	0	0	0	0	0	0	0	0	0	0	0	0	0
LAmbDA	0	0	0	0	0	0	0	0	0	0	0		0
NMC	0	0	0	0	0	0	0	0	0	0	0	0	0
CHETAH	0.5	0.5	0.9	1.1	0.6	0.1	0	10.9	1.2	2.4	9.7	0.1	4.9
scVI	0	0	0	0	0	0	0	0	0	0	0	0	0
scID	23.6	8.3	17.3	32.1	0.2	24.2	9.8	1	5.3	1	0.3	50.4	12
Cell_BLAST	20.3	3.2	19.6	23.1	4.1	0.1	68.1	2	0	0.4	29.3	9.4	29
kNN	0	0	0	0	0	0	0	0	0	0	0	0	0
SCINA												1.9	14.1
DigitalCellSorter												0	0
$Garnett_{CV}$												33.5	70
$Garnett_{pretrained}$												14.8	55.2
Moana												0	0
$Garnett_{DE}$												11	50.9
$SCINA_{DE}$												1.6	5.3
$DigitalCellSorter_{DE}$												0	0

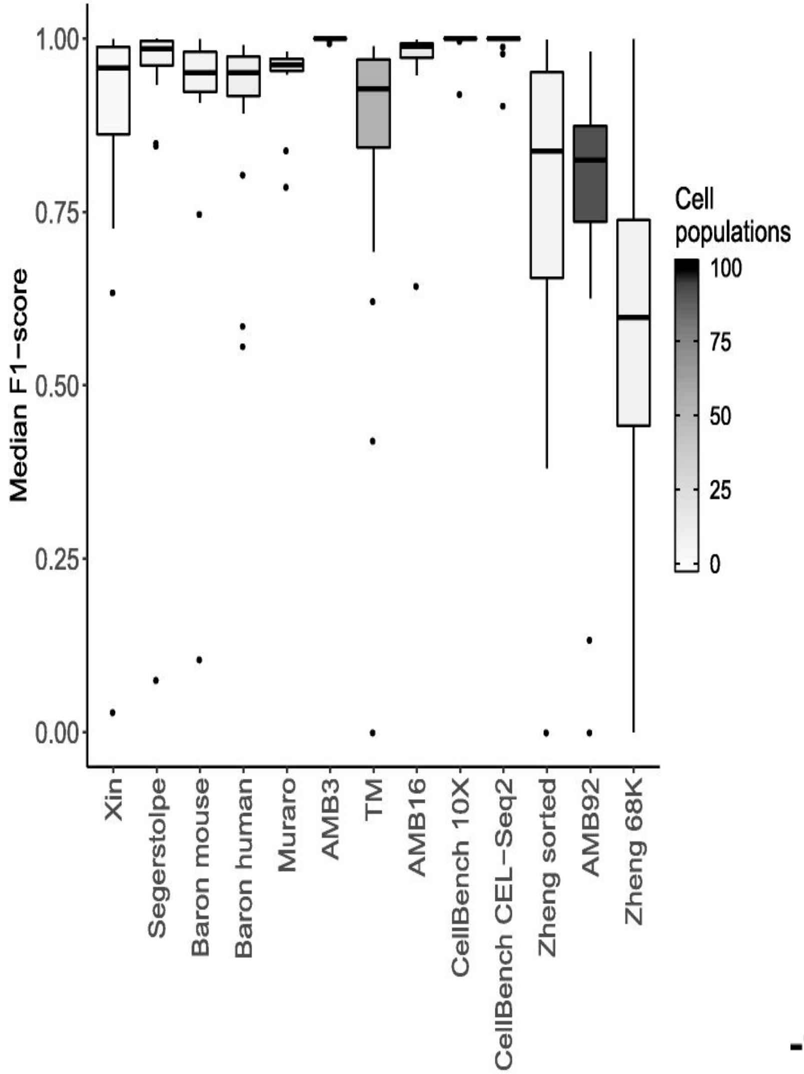
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Performance depends on dataset complexity

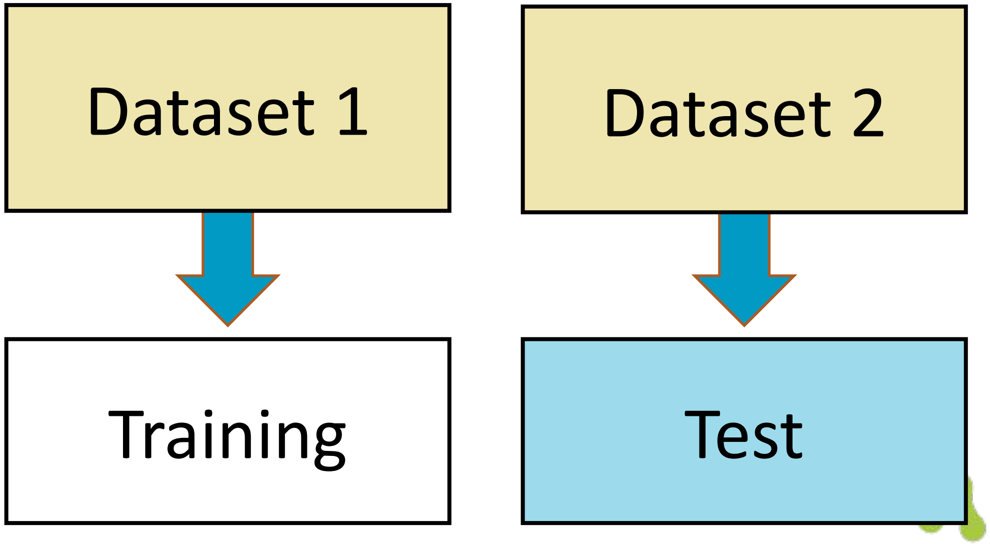
$$\text{Complexity} = \text{mean}\left(\max_{\forall i, i \neq j} \underset{\forall i,j}{\text{corr}}\left(\text{avg}_{C_i}, \text{avg}_{C_j}\right)\right)$$



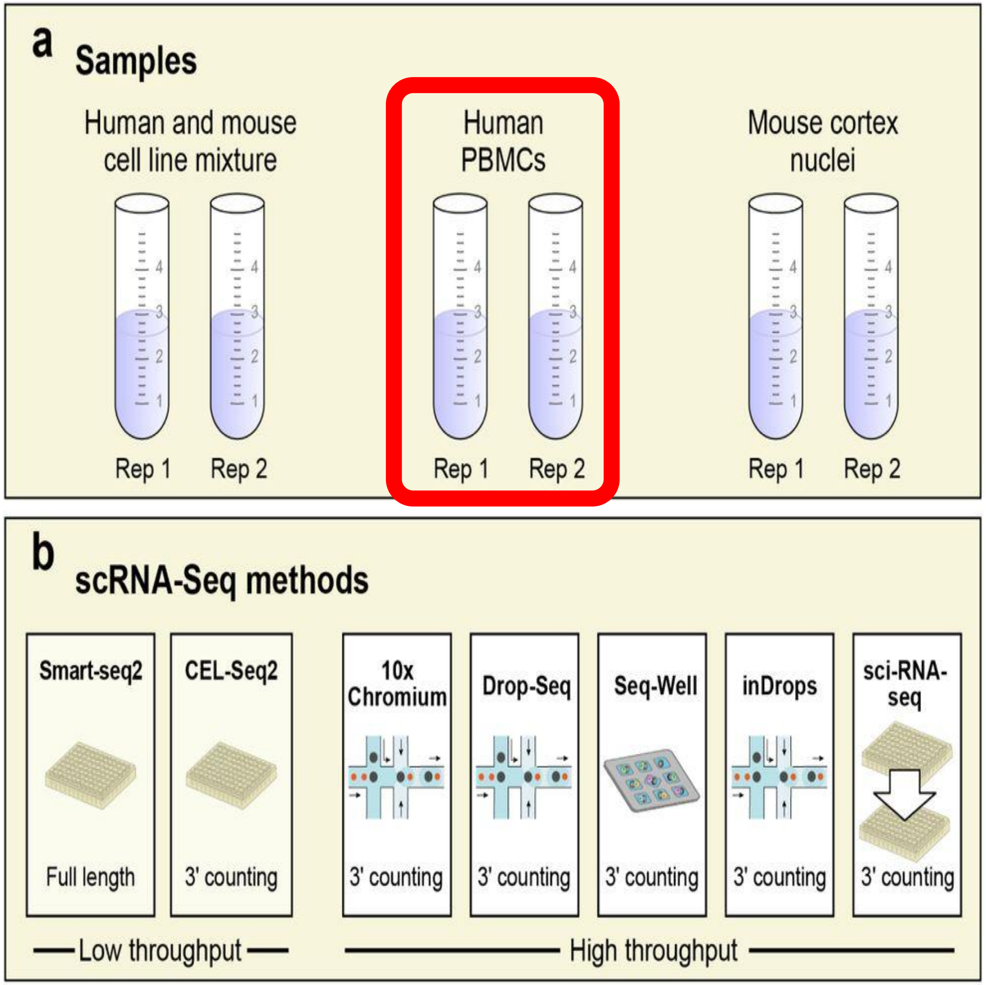


Experiment 2: inter-dataset evaluation

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

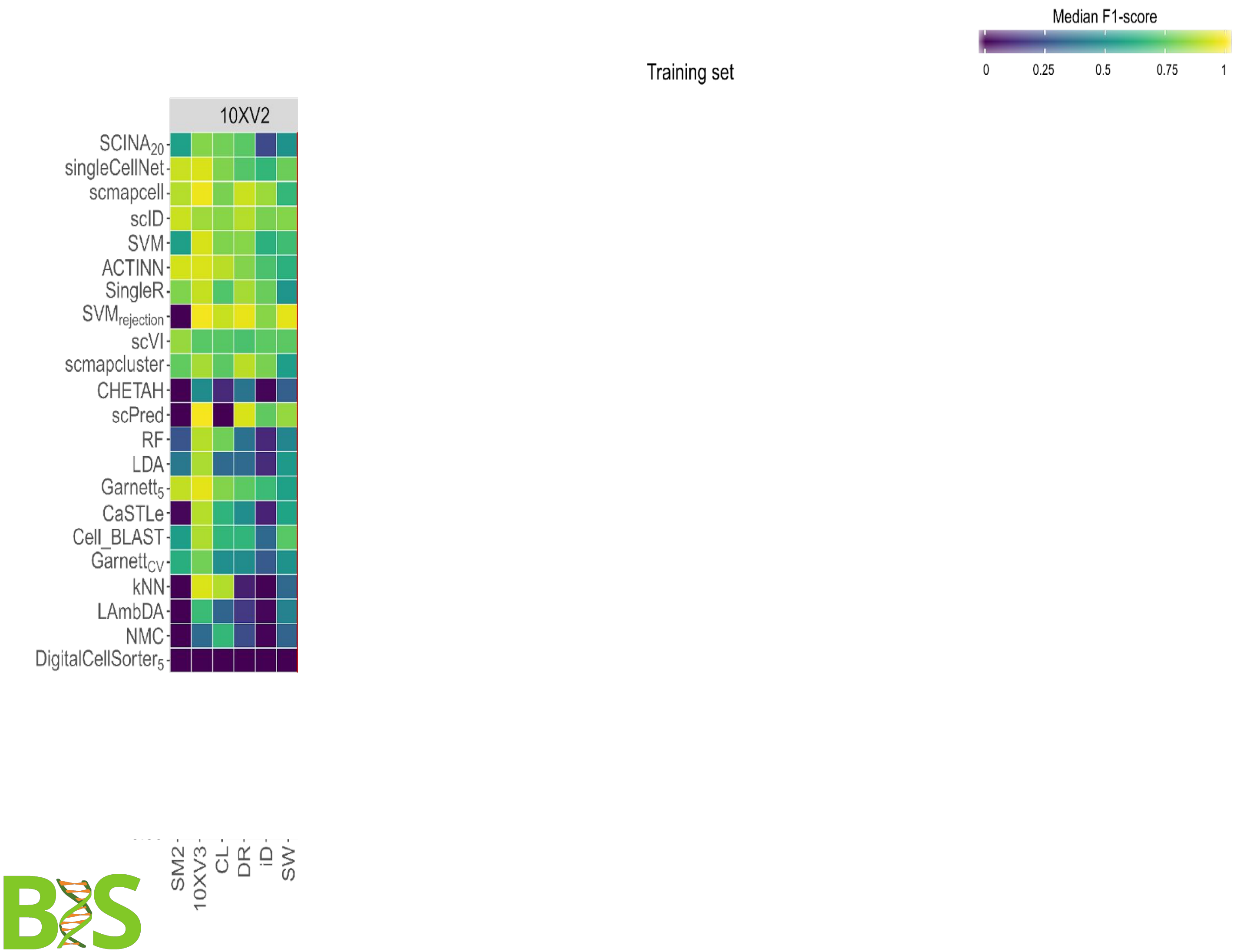


Experiment 2: inter-dataset evaluation

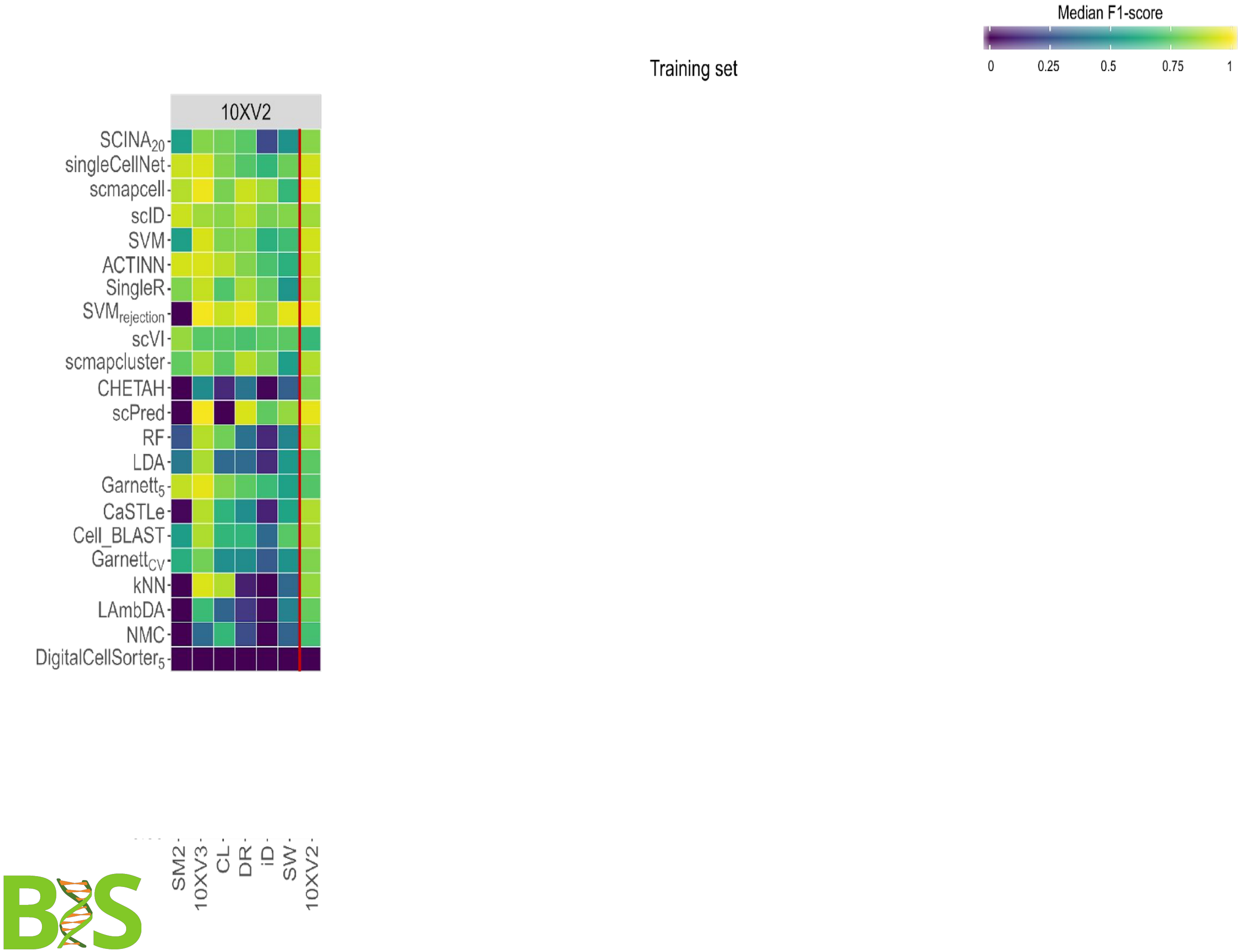


Jiarui Ding et al. *Nature Biotechnology* 2020

Prediction across protocols



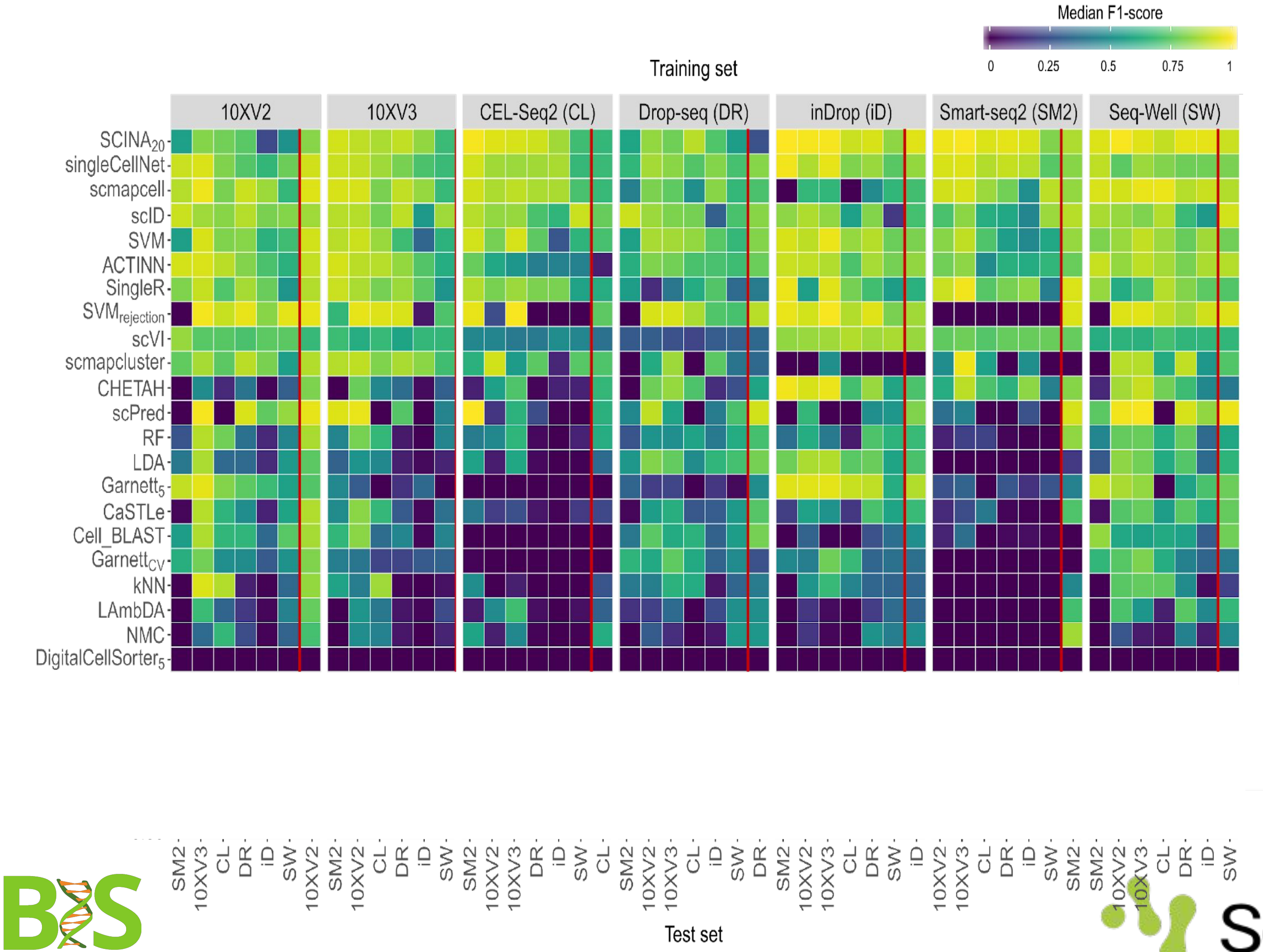
Prediction across protocols



NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

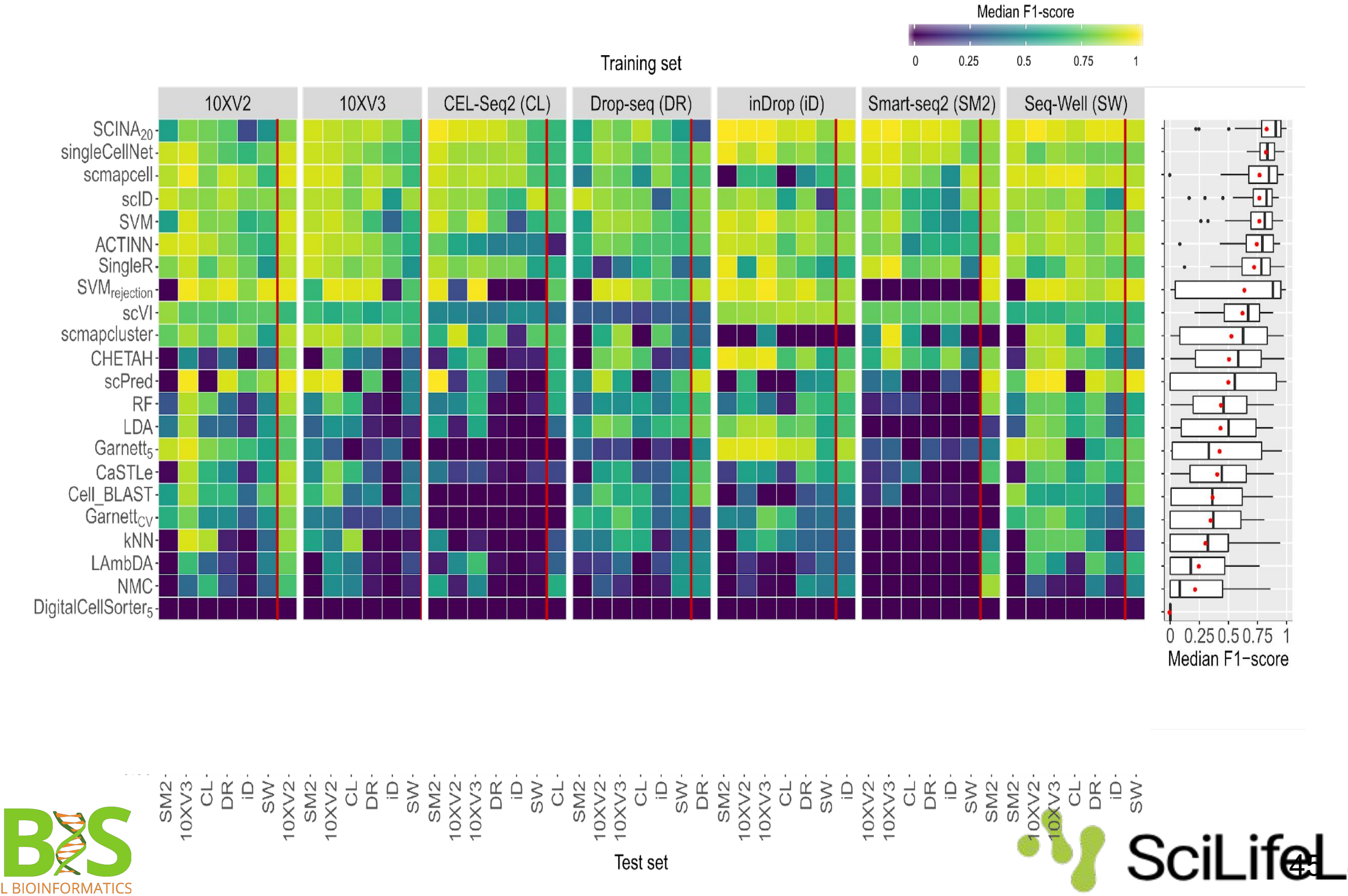
.ab

Prediction across protocols

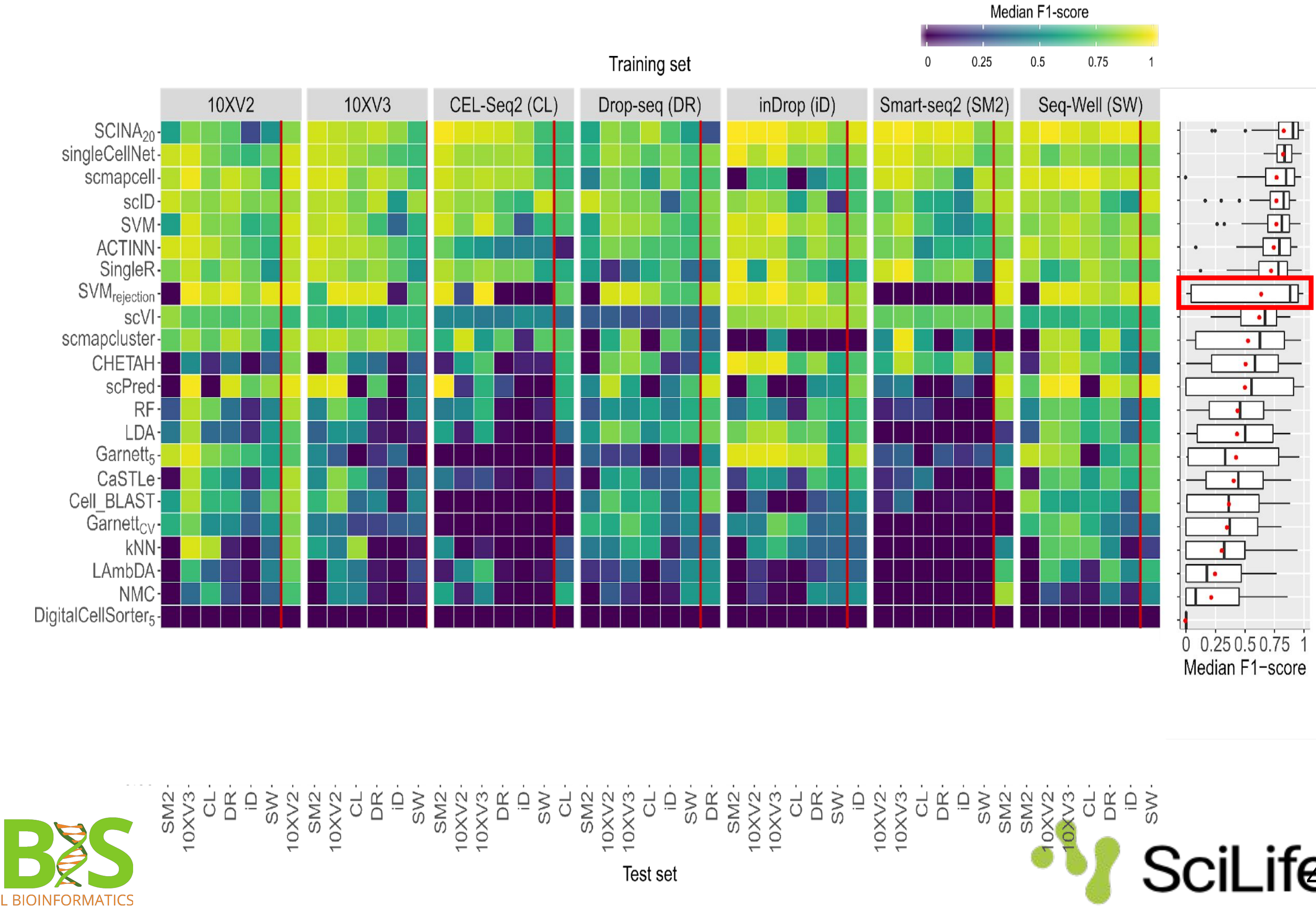


NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

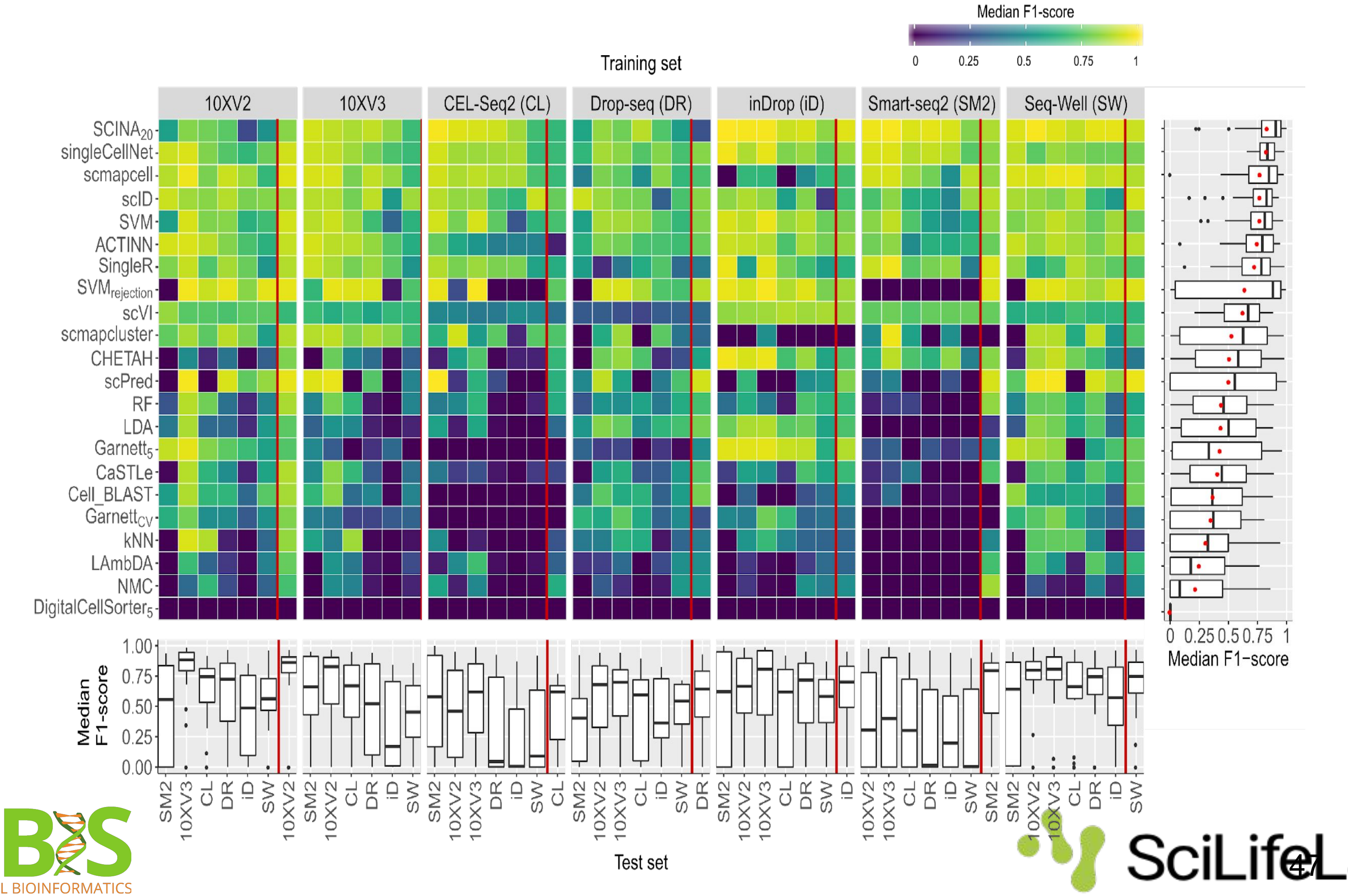
Prediction across protocols



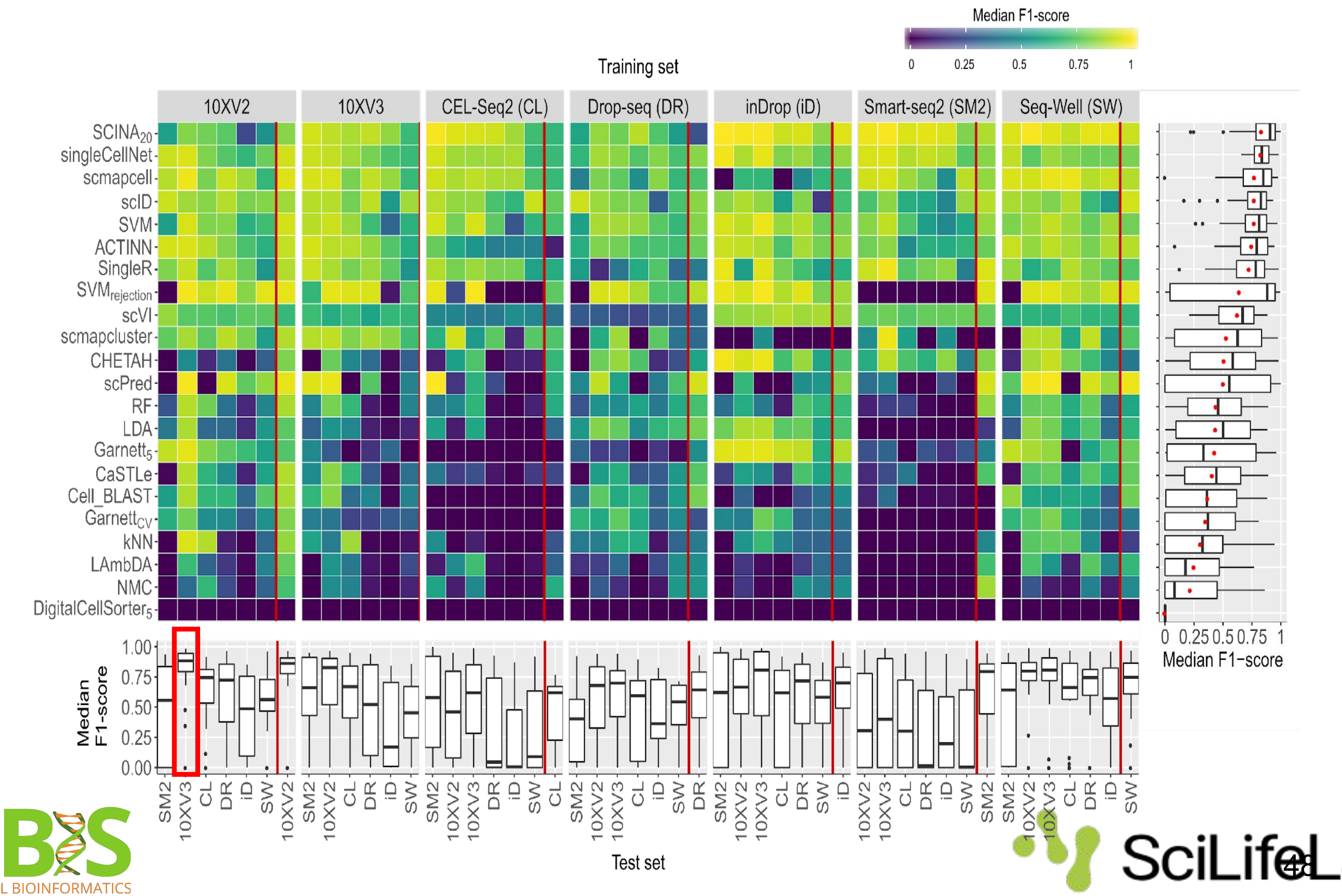
Prediction across protocols



Prediction across protocols



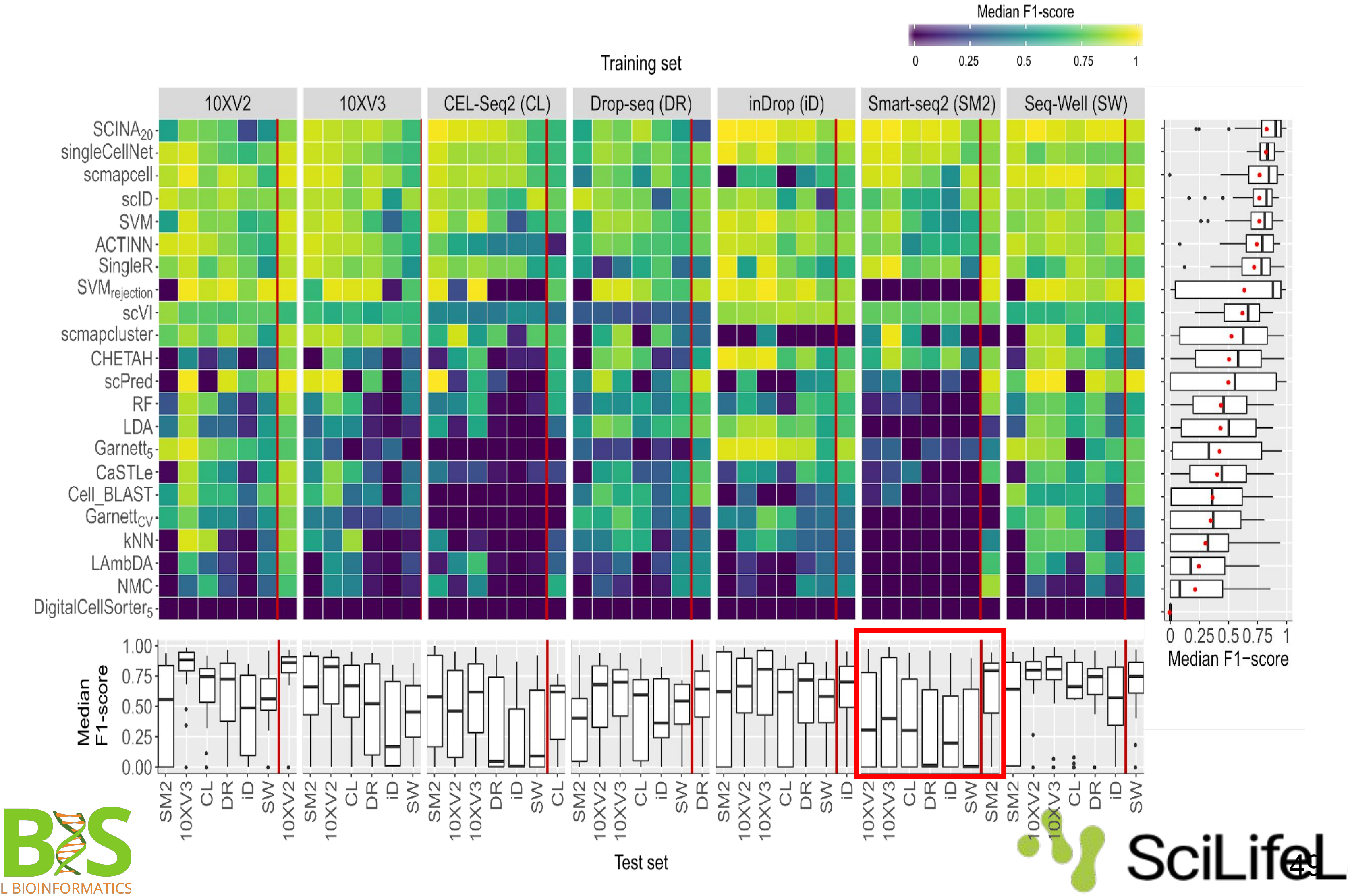
Prediction across protocols



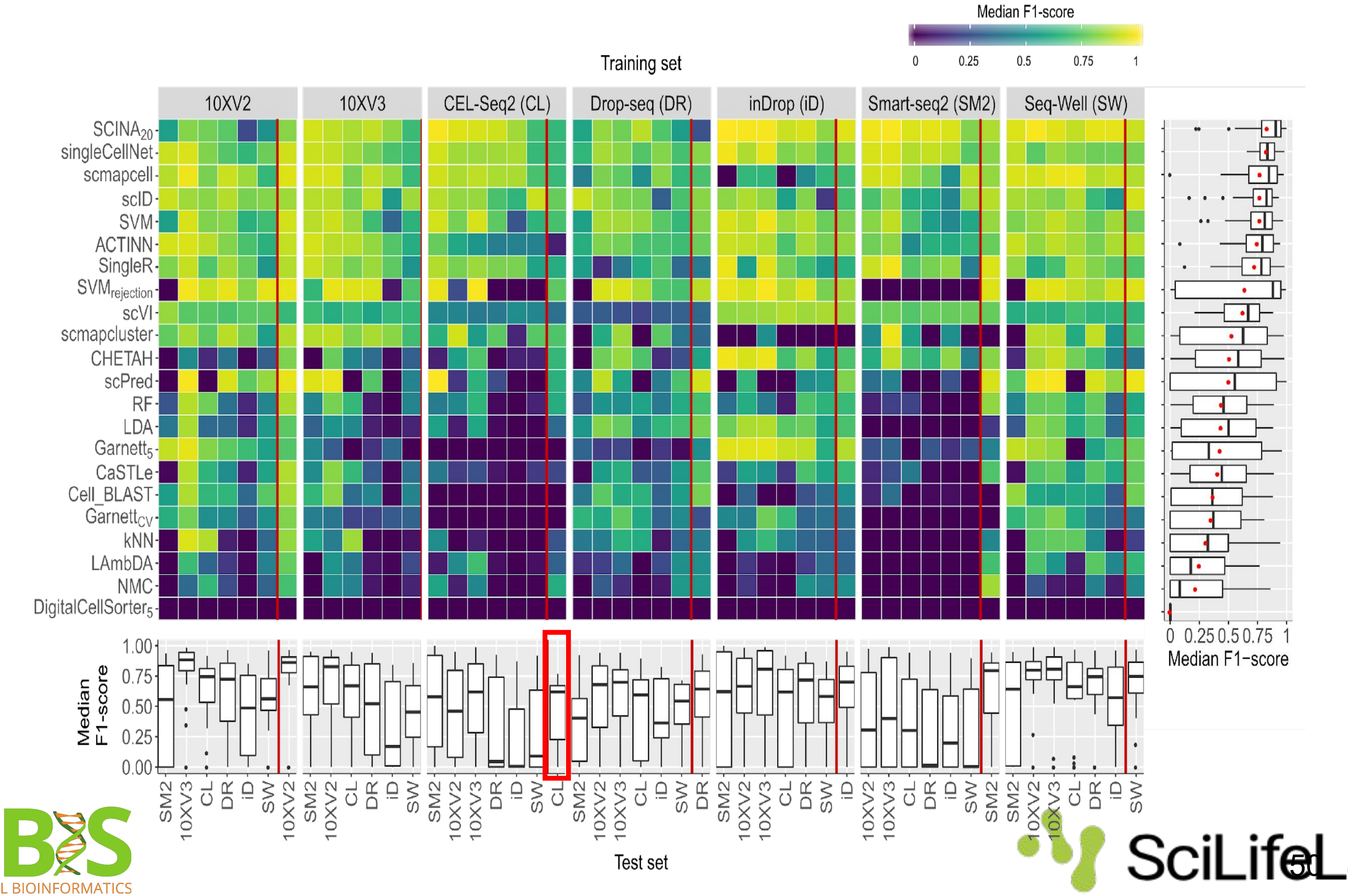
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

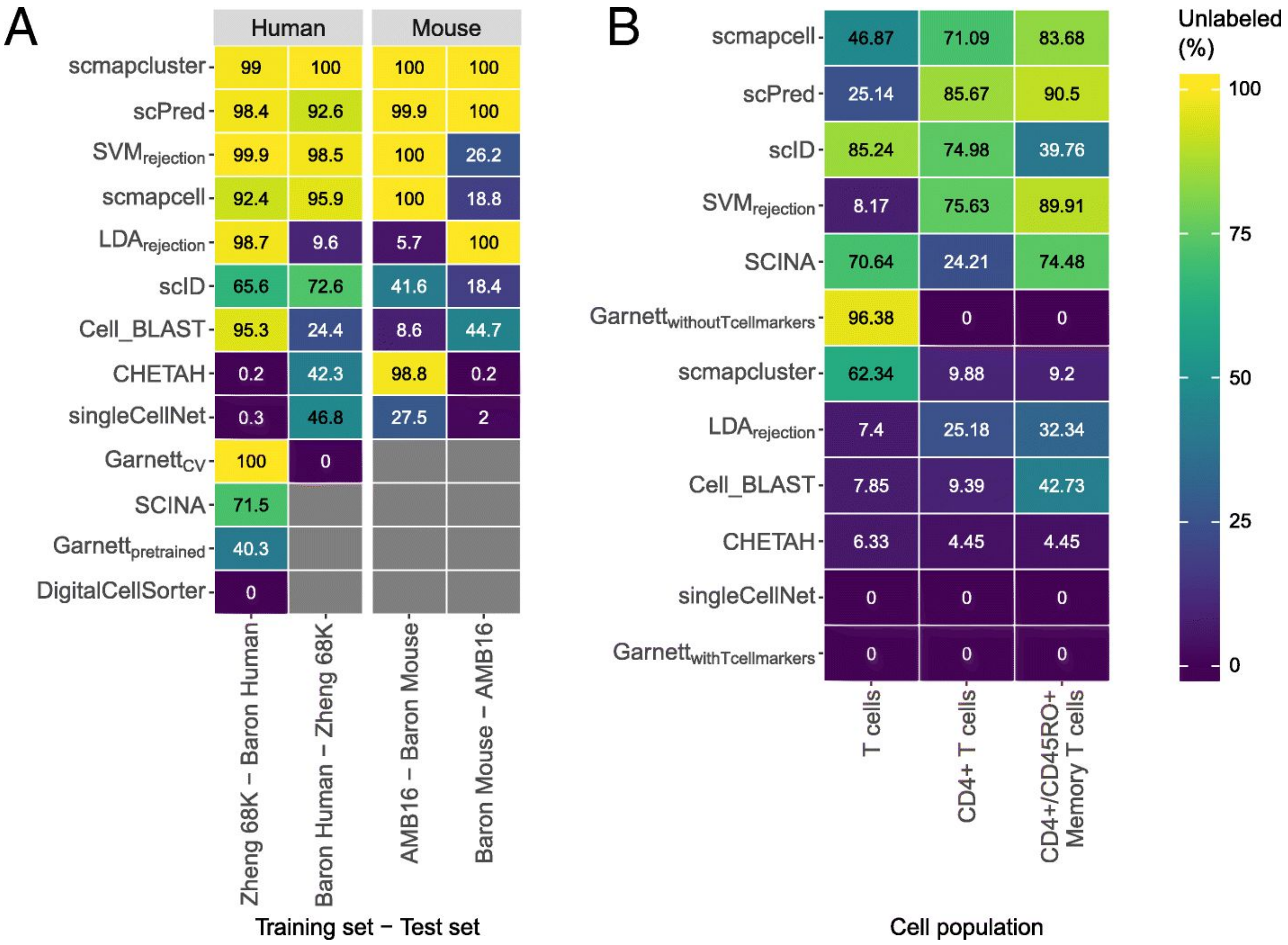
Prediction across protocols



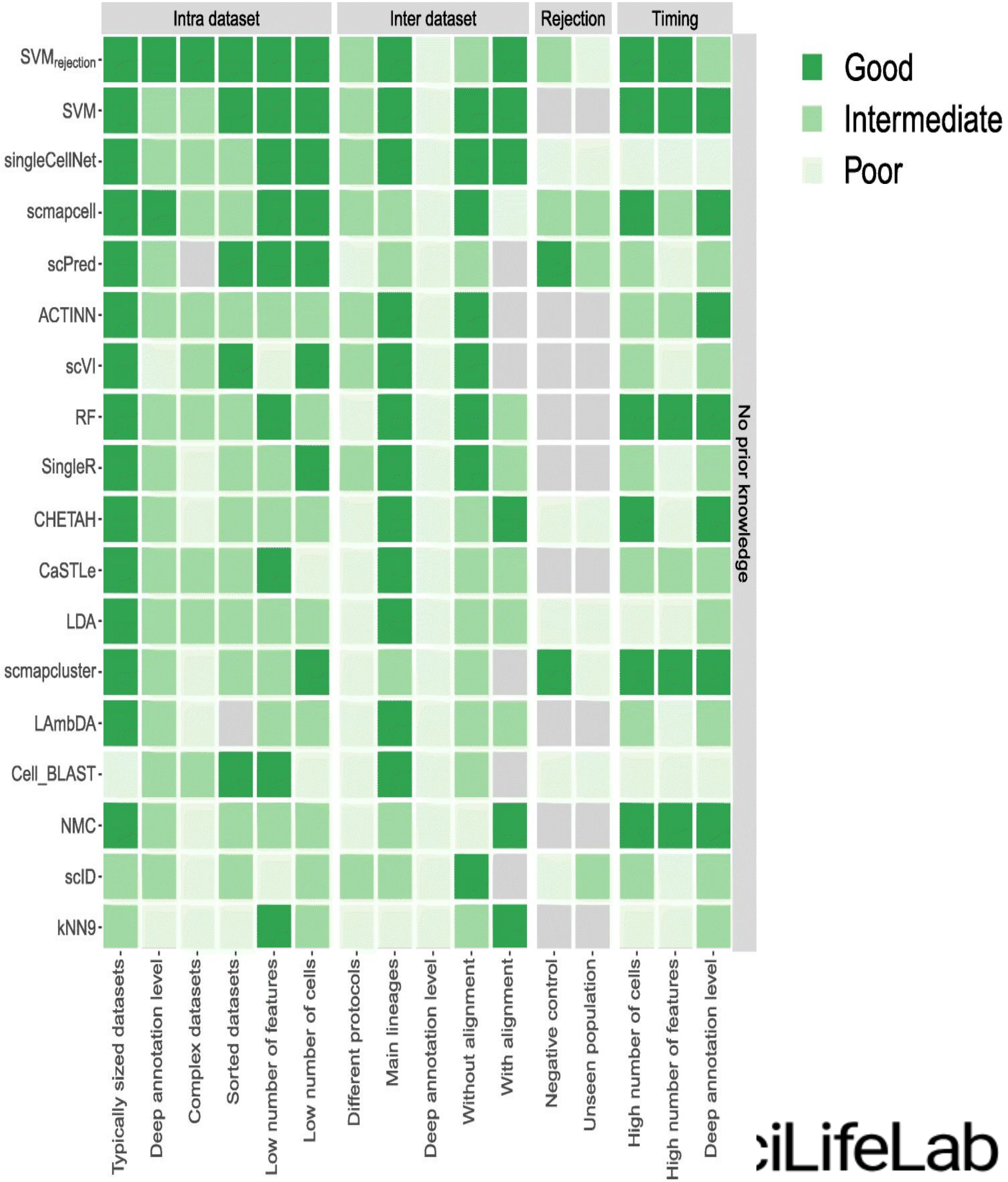
Prediction across protocols



Experiment 3: rejection evaluation



Performance Summary

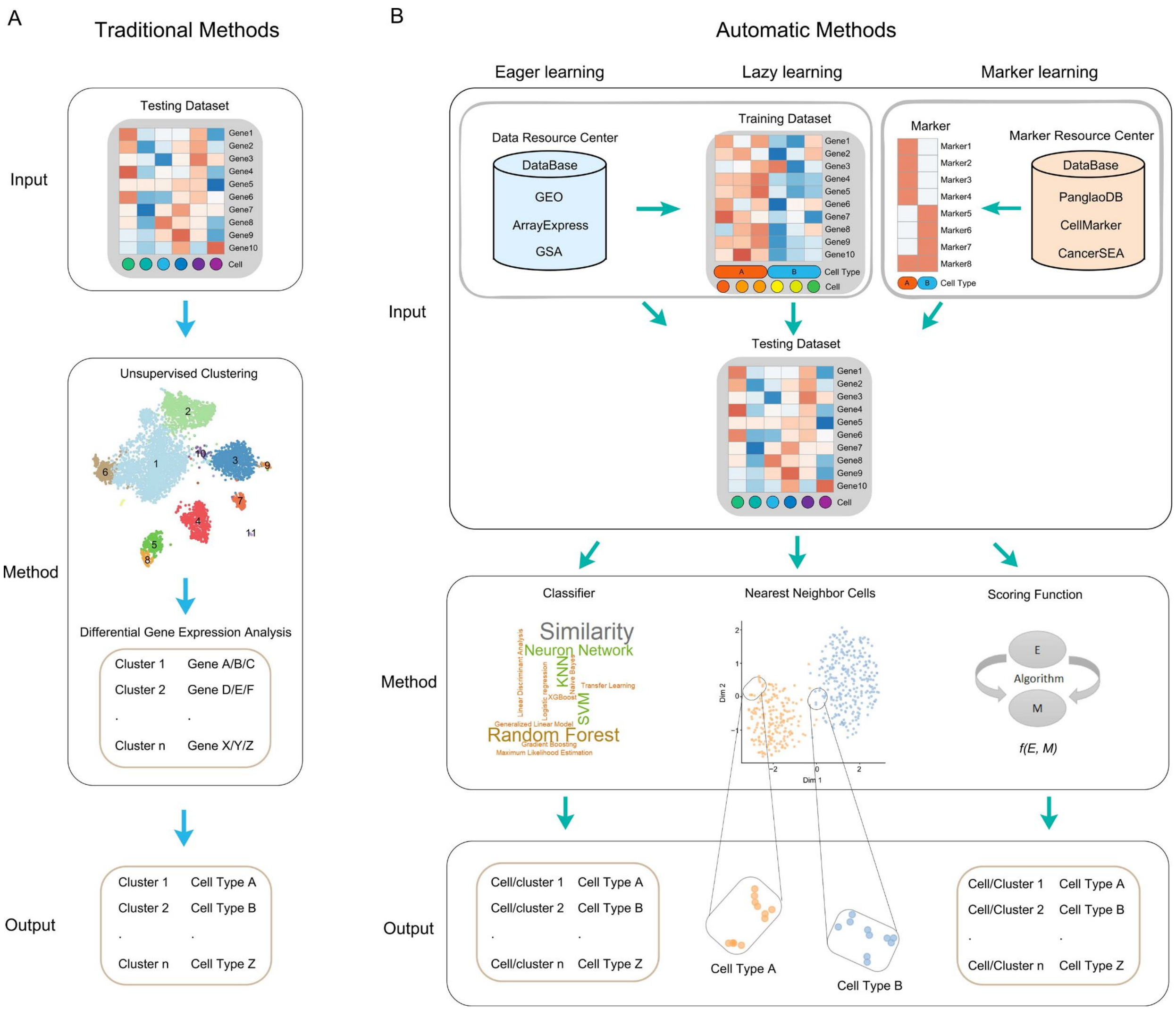


Conclusions so far

- 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*

Benchmark paper 2021



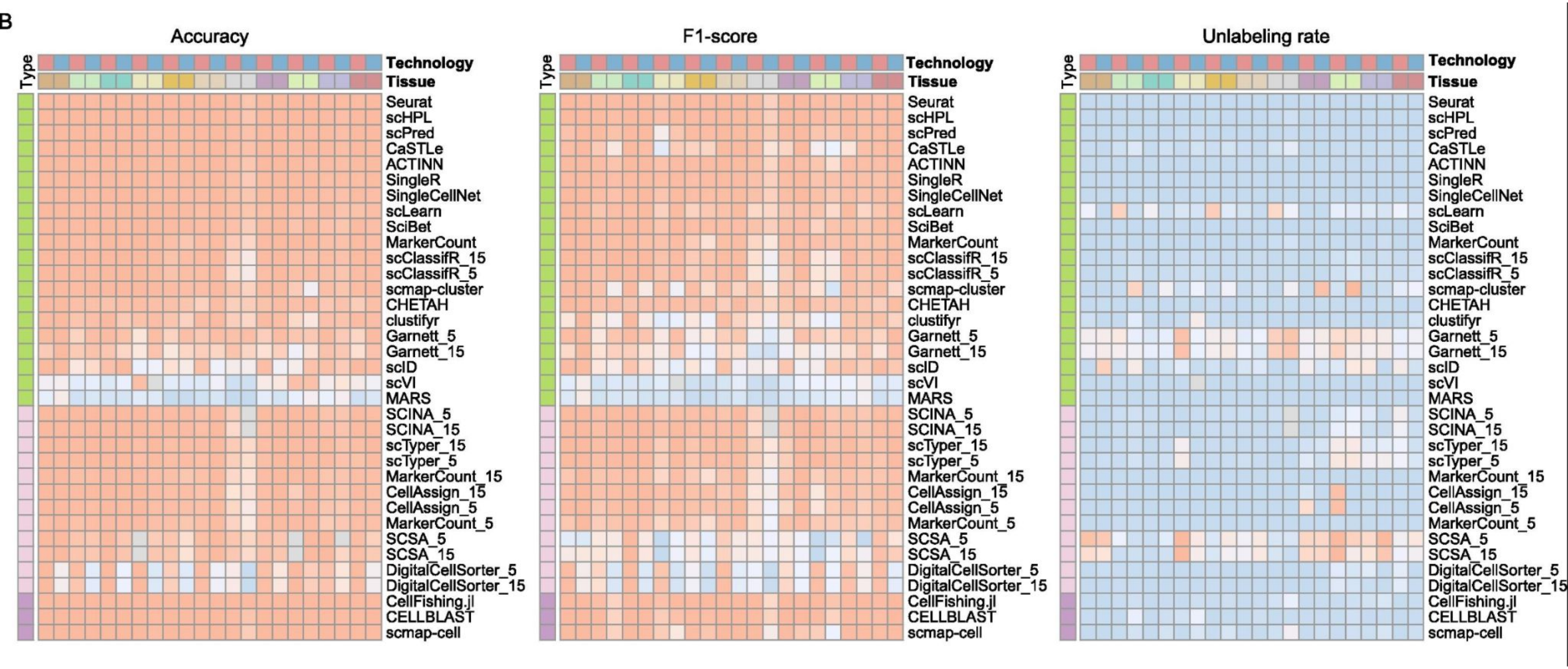
(Xie et al. Comp. Struct. Biotech J. 2021)

Table with all the methods

https://www.csbj.org/action/showFullTableHTML?isHtml=true&tableId=t0005&pii=S2001-0370%2821%2900449-9

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

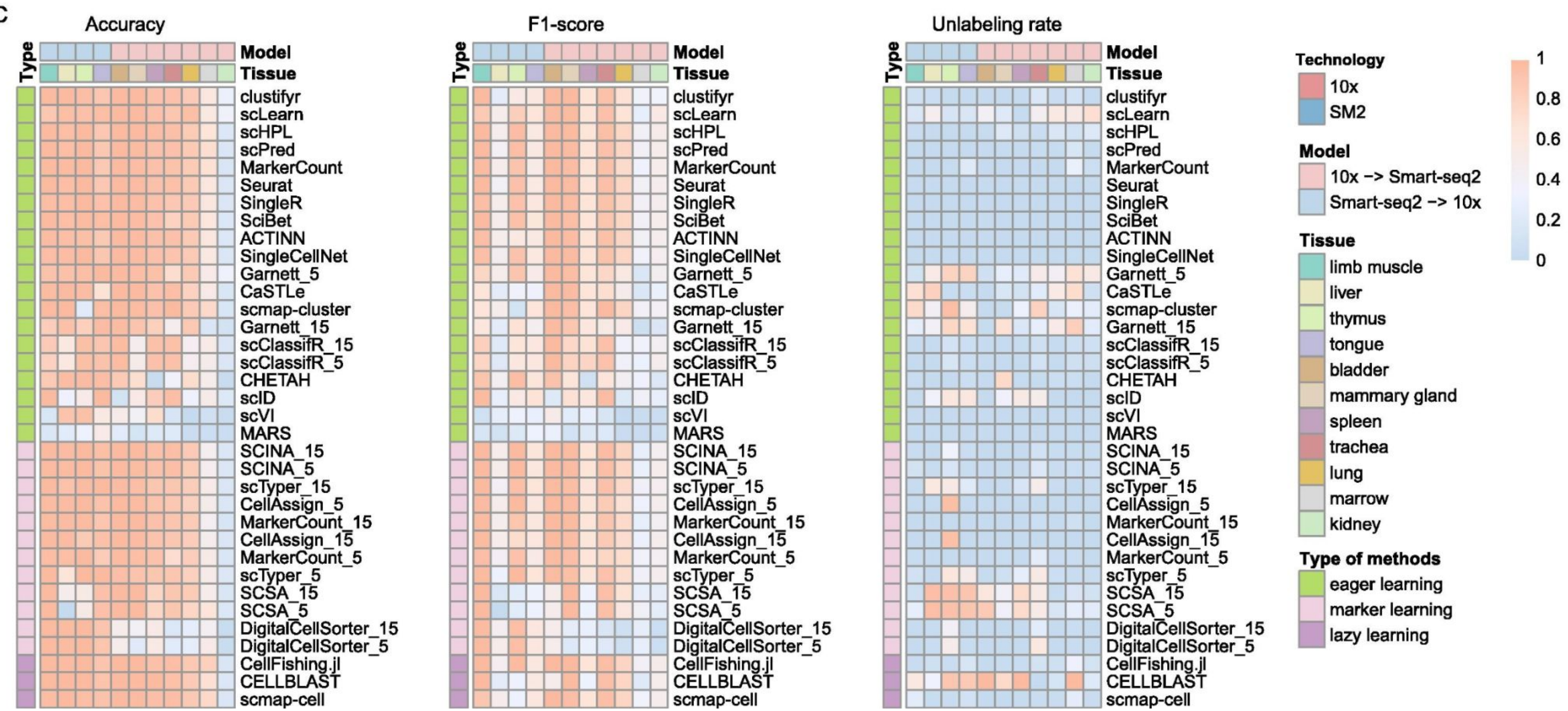
Within dataset training/testing with cross-validation



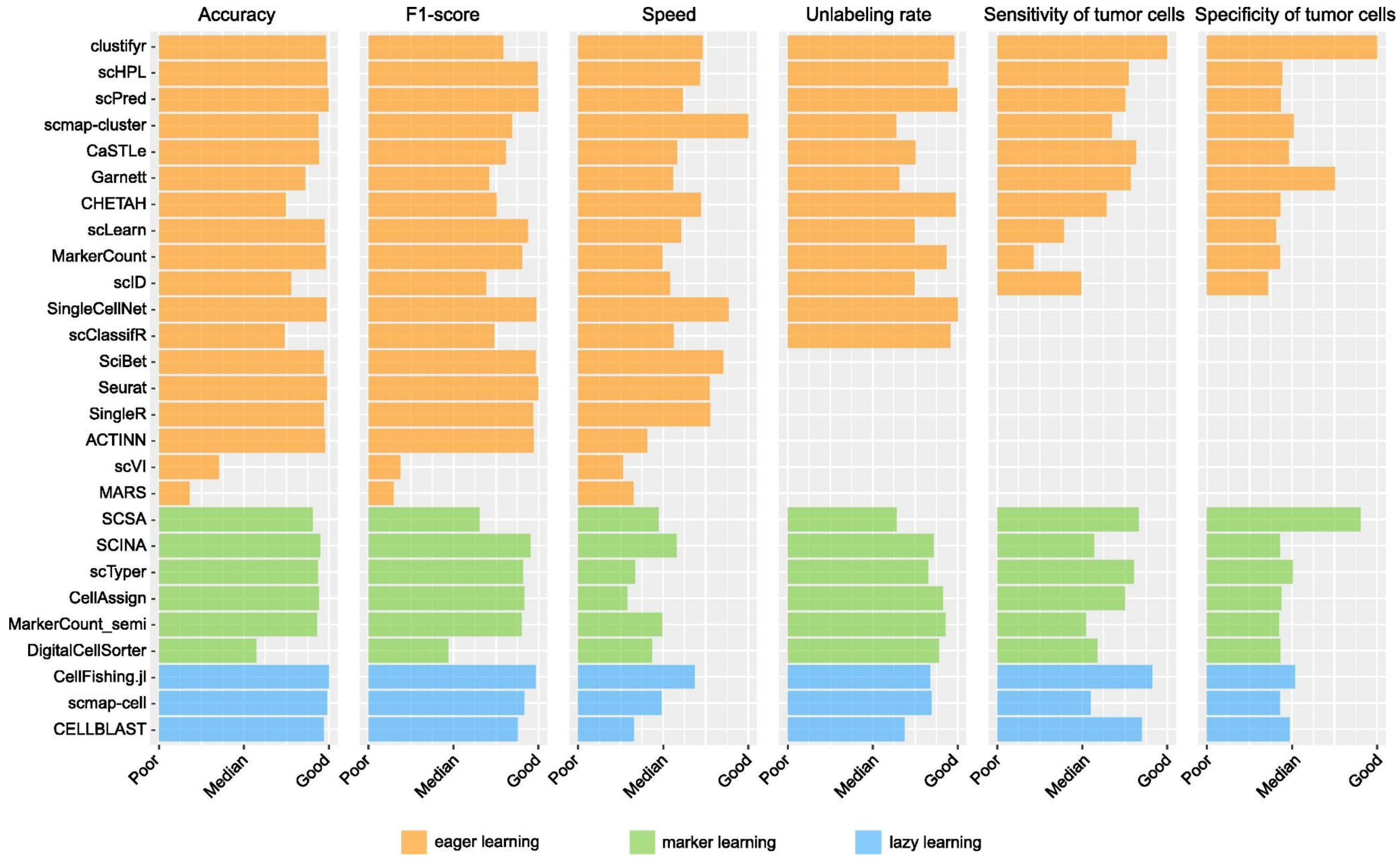
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Across technologies



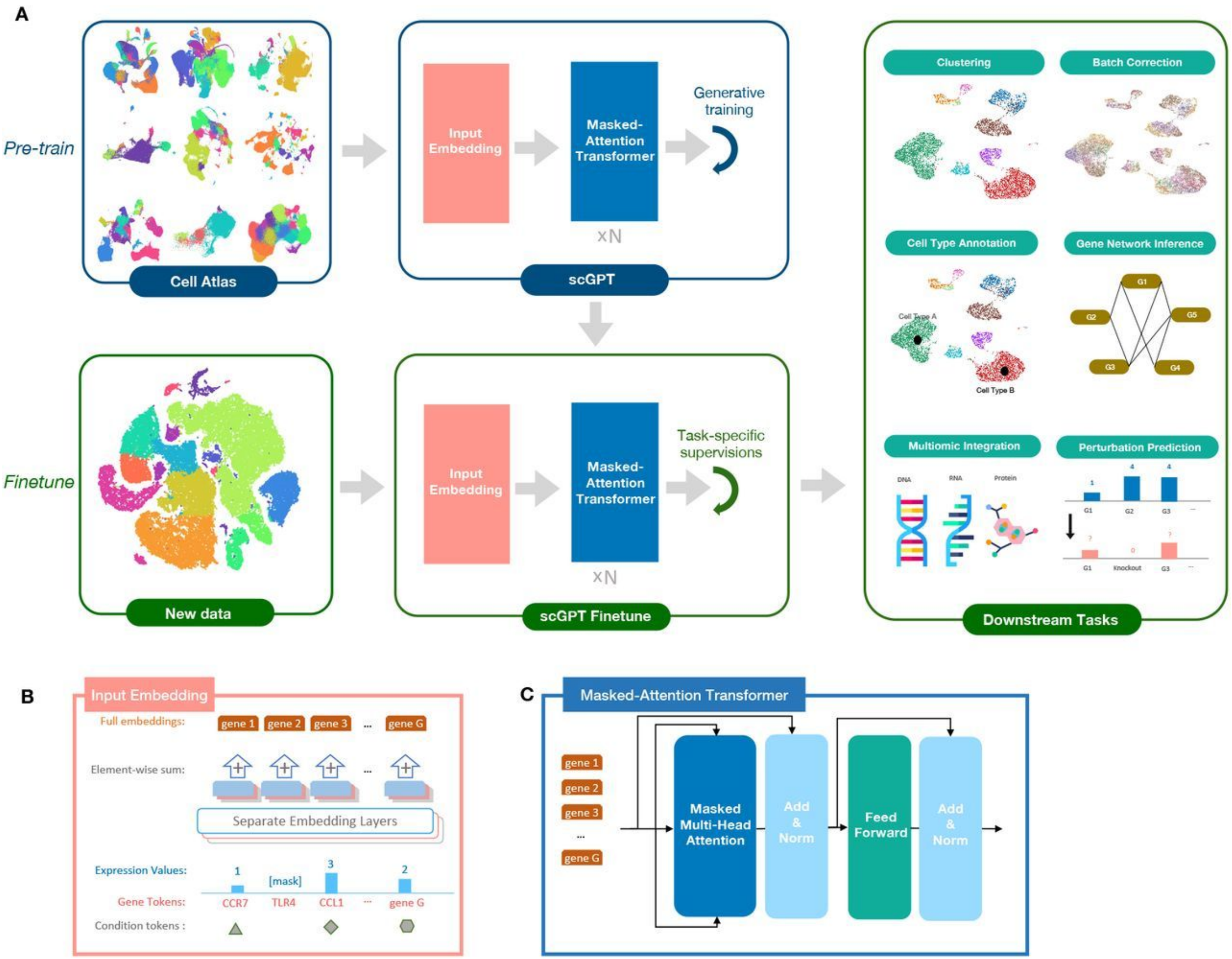
Summary



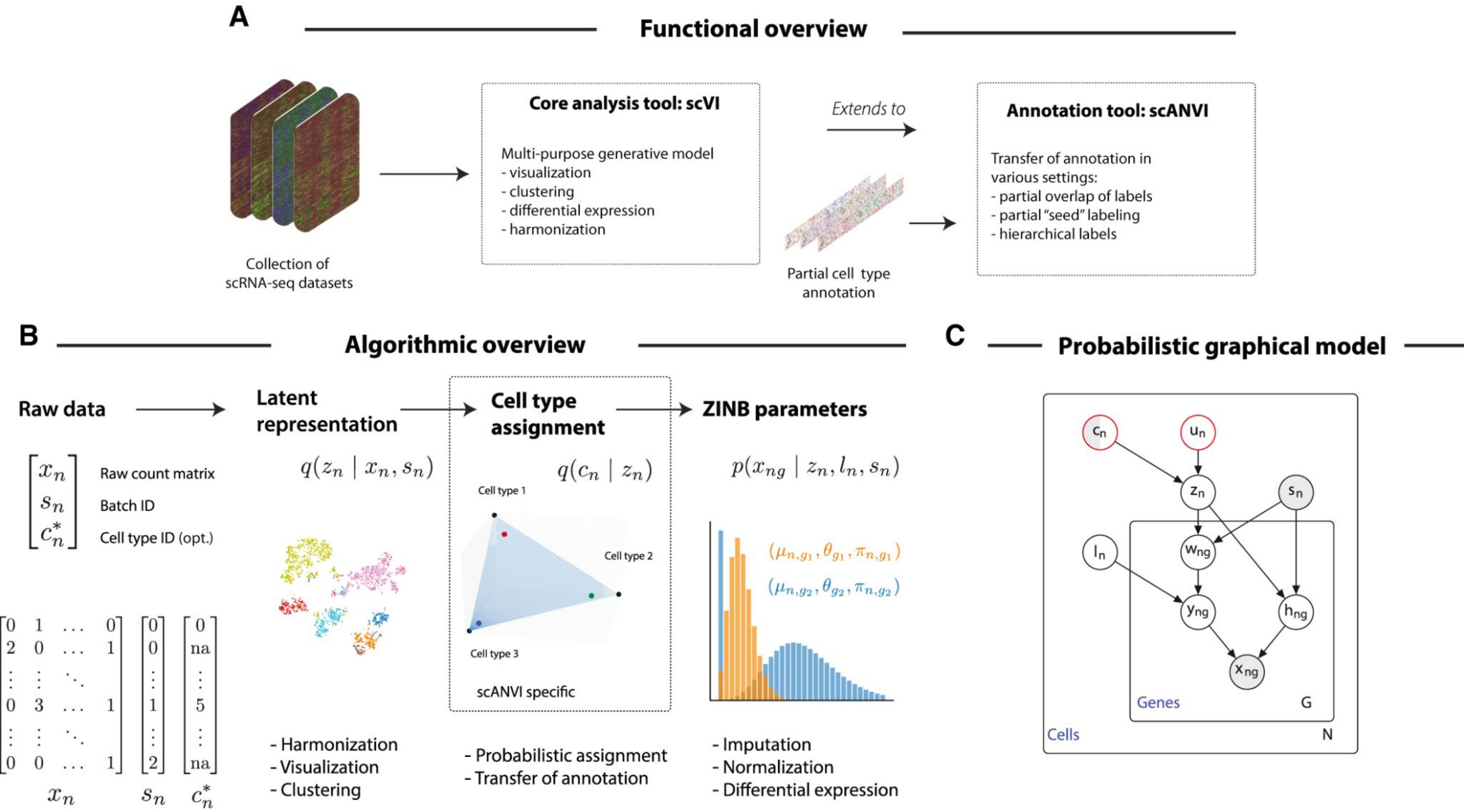
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Generative learning is the next big thing? scGPT



Generative learning is the next big thing? scANVI



A

Functional overview

Collection of scRNA-seq datasets

Core analysis tool: scVI

Multi-purpose generative model
- visualization
- clustering
- differential expression
- harmonization

Extends to

Partial cell type annotation

Annotation tool: scANVI

Transfer of annotation in various settings:
- partial overlap of labels
- partial "seed" labeling
- hierarchical labels

B

Algorithmic overview

Raw data

x_n Raw count matrix

s_n Batch ID

c_n^* Cell type ID (opt.)

$$x_n = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 2 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \\ 0 & 3 & \dots & 1 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad s_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 2 \end{bmatrix} \quad c_n^* = \begin{bmatrix} 0 \\ \text{na} \\ \vdots \\ 5 \\ \vdots \\ \text{na} \end{bmatrix}$$

Latent representation

$q(z_n \mid x_n, s_n)$

- Harmonization
- Visualization
- Clustering

Cell type assignment

$q(c_n \mid z_n)$

Cell type 1

Cell type 2

Cell type 3

scANVI specific

- Probabilistic assignment
- Transfer of annotation

ZINB parameters

$p(x_{ng} \mid z_n, l_n, s_n)$

$(\mu_{n,g_1}, \theta_{g_1}, \pi_{n,g_1})$

$(\mu_{n,g_2}, \theta_{g_2}, \pi_{n,g_2})$

- Imputation
- Normalization
- Differential expression

C

Probabilistic graphical model

Some useful resources

- Azimuth - Seurat label transfer to reference sets
  - https://azimuth.hubmapconsortium.org/
  - online or R package
- DISCO - CellMapper to several tissues
  - https://www.immunesinglecell.org/
- Celltypist - Regularised linear models with Stochastic Gradient Descent
  - https://www.celltypist.org/
  - online or python package

Summary

- Cell identification is moving from unsupervised (clustering/visualization) to supervised (classification) learning
- Check what reference you are using!
  - The more similar 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 makes sense