Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

BTEP Presentation: scRNA – Cell Type Annotation

Keyur Talsania

CCR-SF Bioinformatics Group
Advanced Biomedical and Computational Sciences

Biomedical Informatics and Data Science (BIDS) Directorate

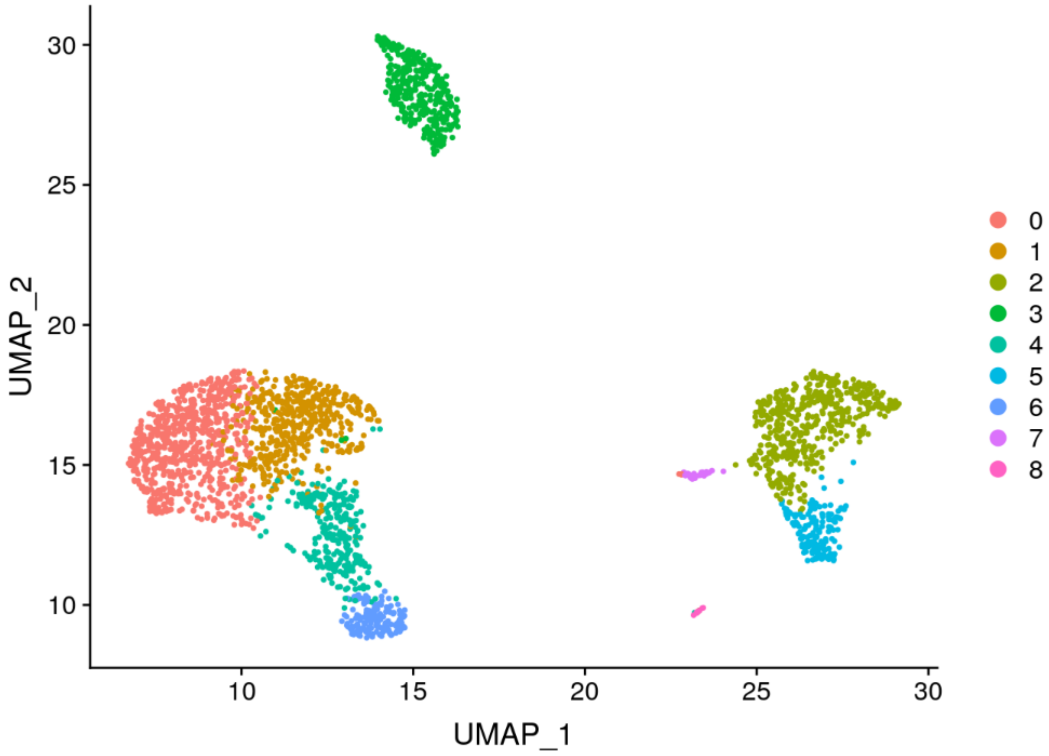
Frederick National Laboratory for Cancer Research

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute
The Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

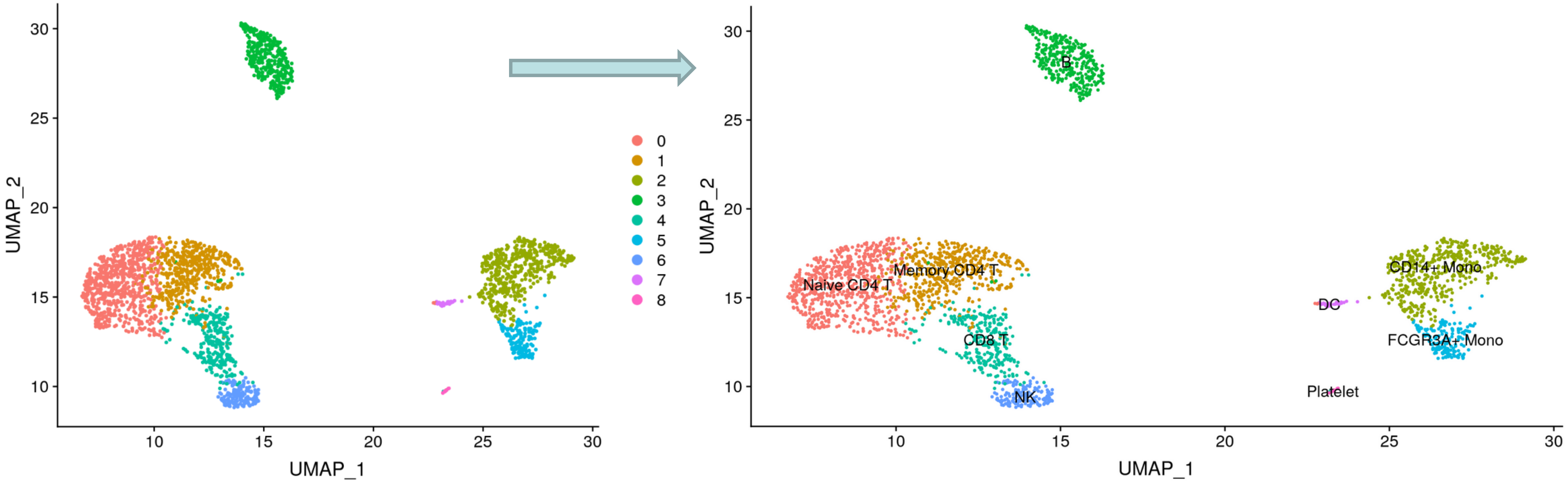
Outline

- **How it works**
  - Tools available
  - Publications
- **Tools we use**
  - SingleR, DigitalCellSorter
    - Label Transfer - SingleR
- **Concept of Automated Panels**
  - SF Pipeline’s Automated Panels

Why we need cell type annotation?



Why we need cell type annotation?



Tools Available - https://www.scrna-tools.org

- ACTINN
- AltAnalyze
- bigSCale2
- cardelino
- CaSTLe
- celaref
- Cell-BLAST
- cellassign
- CellFishing
- CellO
- CHETAH
- ClusterMap
- DigitalCellSorter
- DistMap
- DropLasso
- FateID
- Garnett
- hscScore
- LAmbDA
- matchSCore2
- MetaNeighbor
- MIMOSCA
- Moana
- ParaDPMM
- scdney
- scID
- SCINA
- scmap
- scMatch
- scMCA
- scPred
- scVI
- SingleCellNet
- SingleR

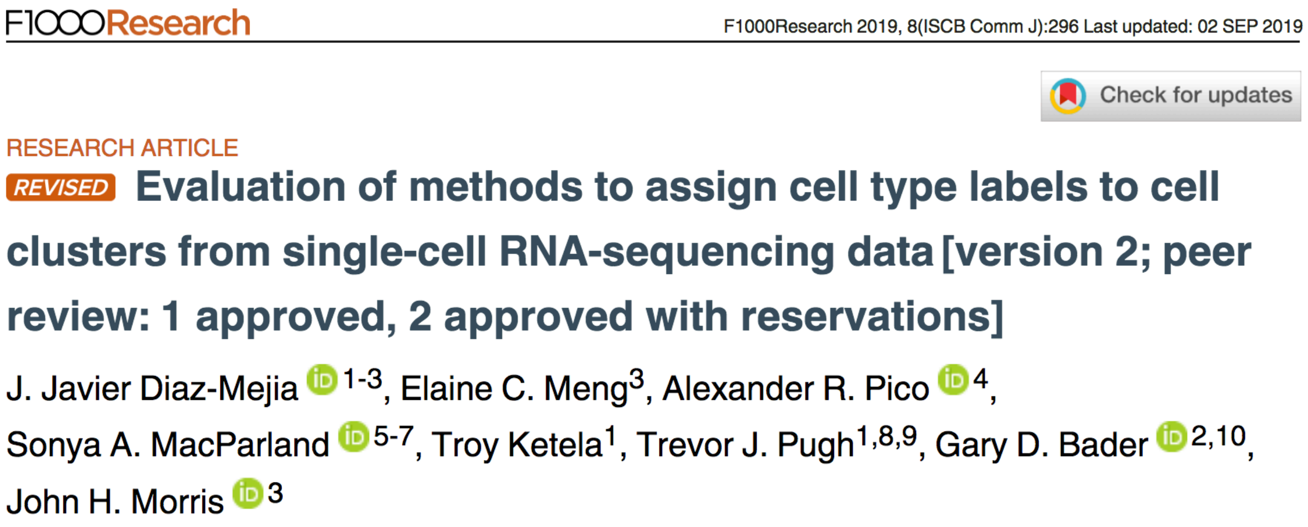
Types of Cell Annotation Tools

- **Supervised methods :**
  - which require a training dataset labeled with the corresponding cell populations in order to train the classifier
  - SingleR, SVM, ACTINN, scPred, CaSTle

- **Prior-knowledge based methods:**
  - for which either a marker gene file is required as an input or a pretrained classifier for specific cell populations is provided.
  - Garnett, DigitalCellSorter, Moana

Articles of comparison of Celltype Annotation tools

Frederick National Laboratory for Cancer Research



F1000Research

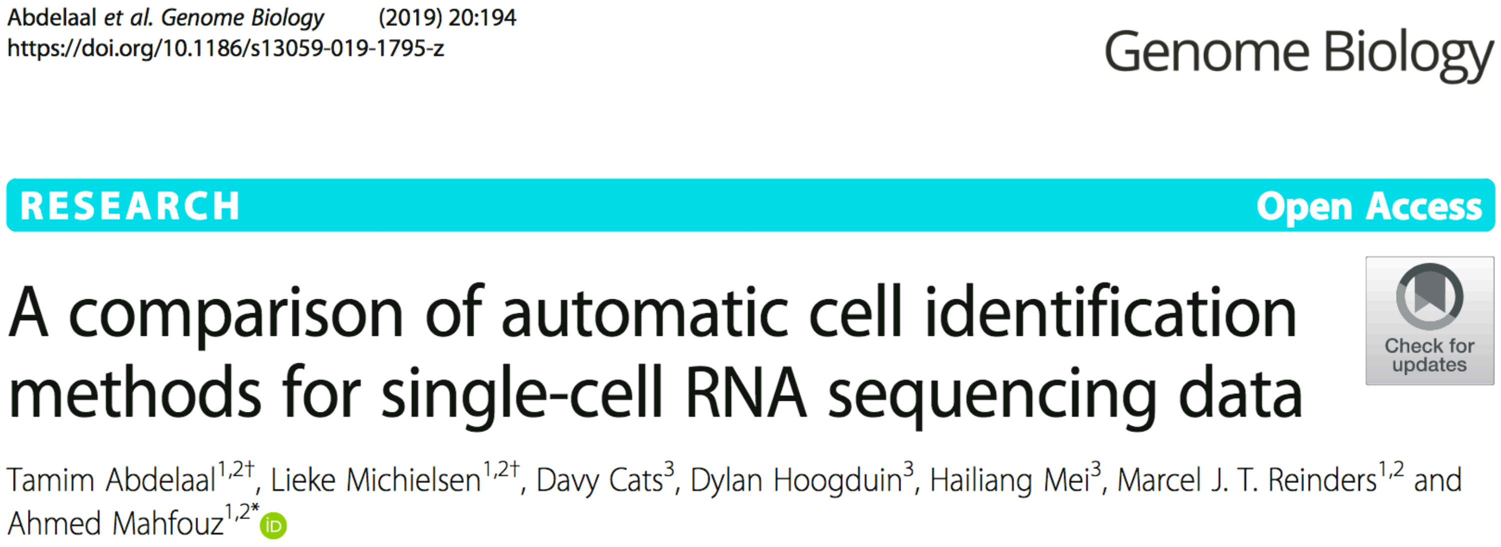
F1000Research 2019, 8(ISCB Comm J):296 Last updated: 02 SEP 2019

Check for updates

RESEARCH ARTICLE

REVISED Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data [version 2; peer review: 1 approved, 2 approved with reservations]

J. Javier Diaz-Mejia[1-3], Elaine C. Meng[3], Alexander R. Pico[4], Sonya A. MacParland[5-7], Troy Ketela[1], Trevor J. Pugh[1,8,9], Gary D. Bader[2,10], John H. Morris[3]



Abdelaal *et al. Genome Biology* (2019) 20:194
https://doi.org/10.1186/s13059-019-1795-z

Genome Biology

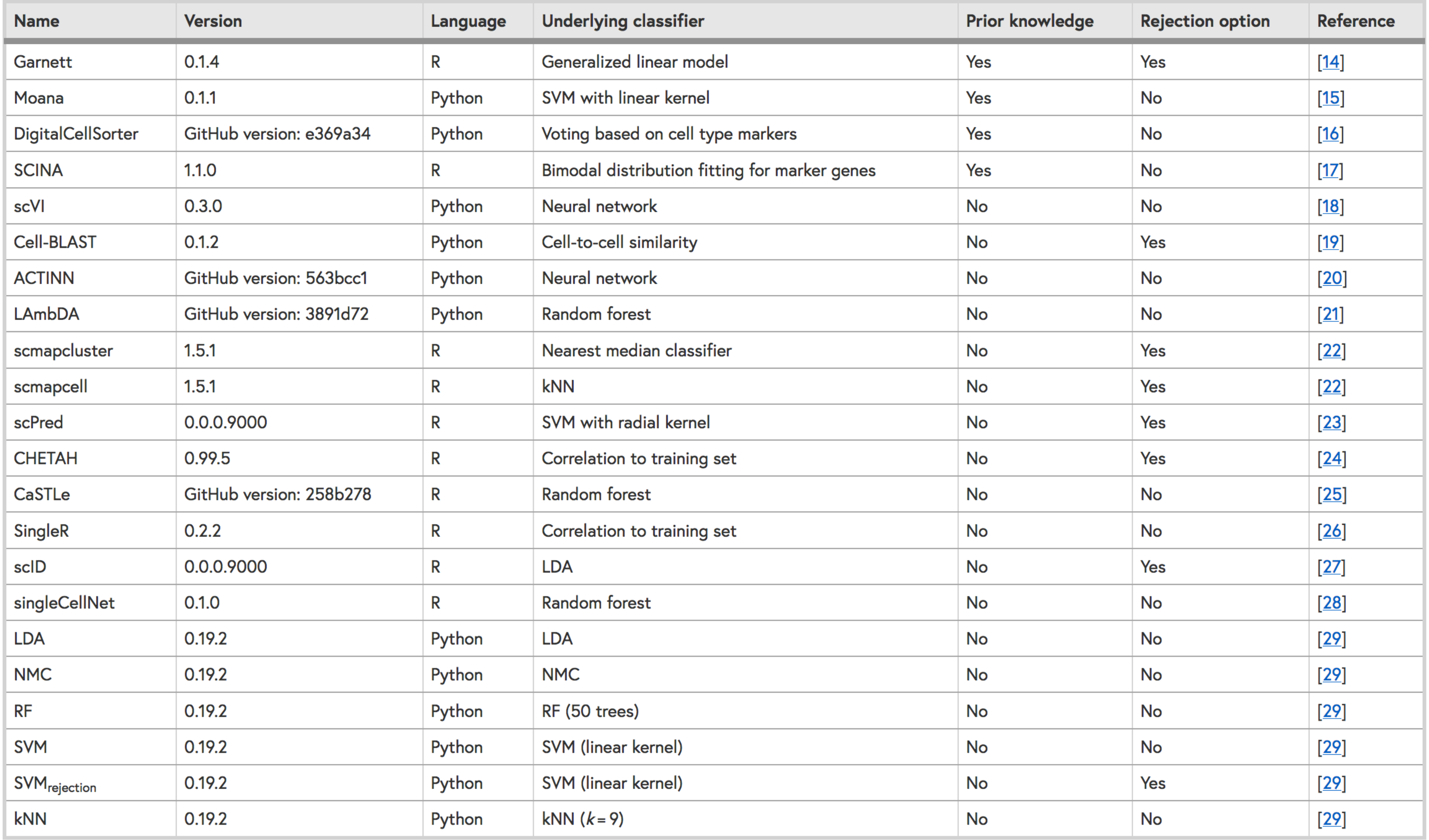
RESEARCH Open Access

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

Check for updates

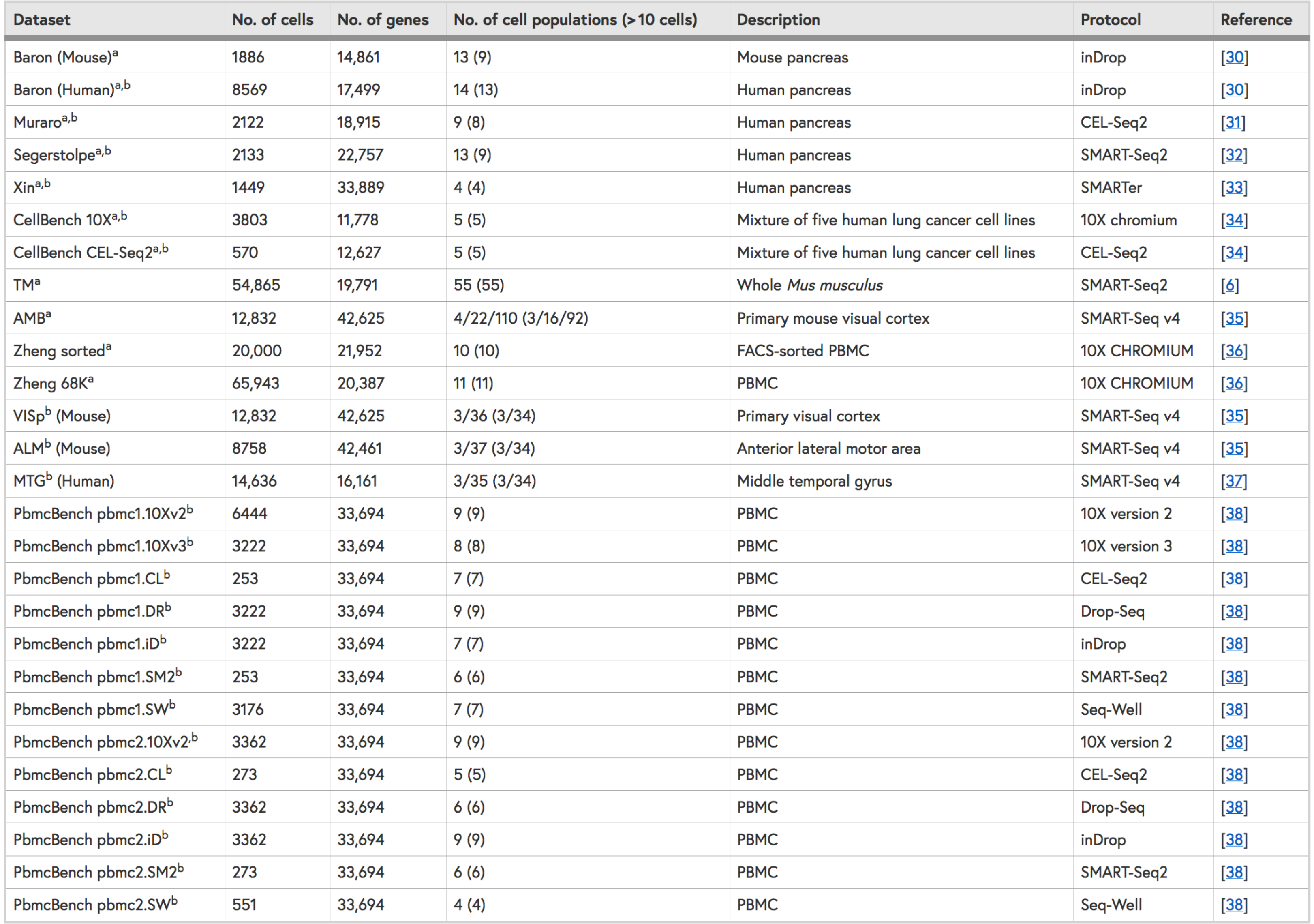
Tamim Abdelaal[1,2†], Lieke Michielsen[1,2†], Davy Cats[3], Dylan Hoogduin[3], Hailiang Mei[3], Marcel J. T. Reinders[1,2] and Ahmed Mahfouz[1,2*]

Abdelaal et al.



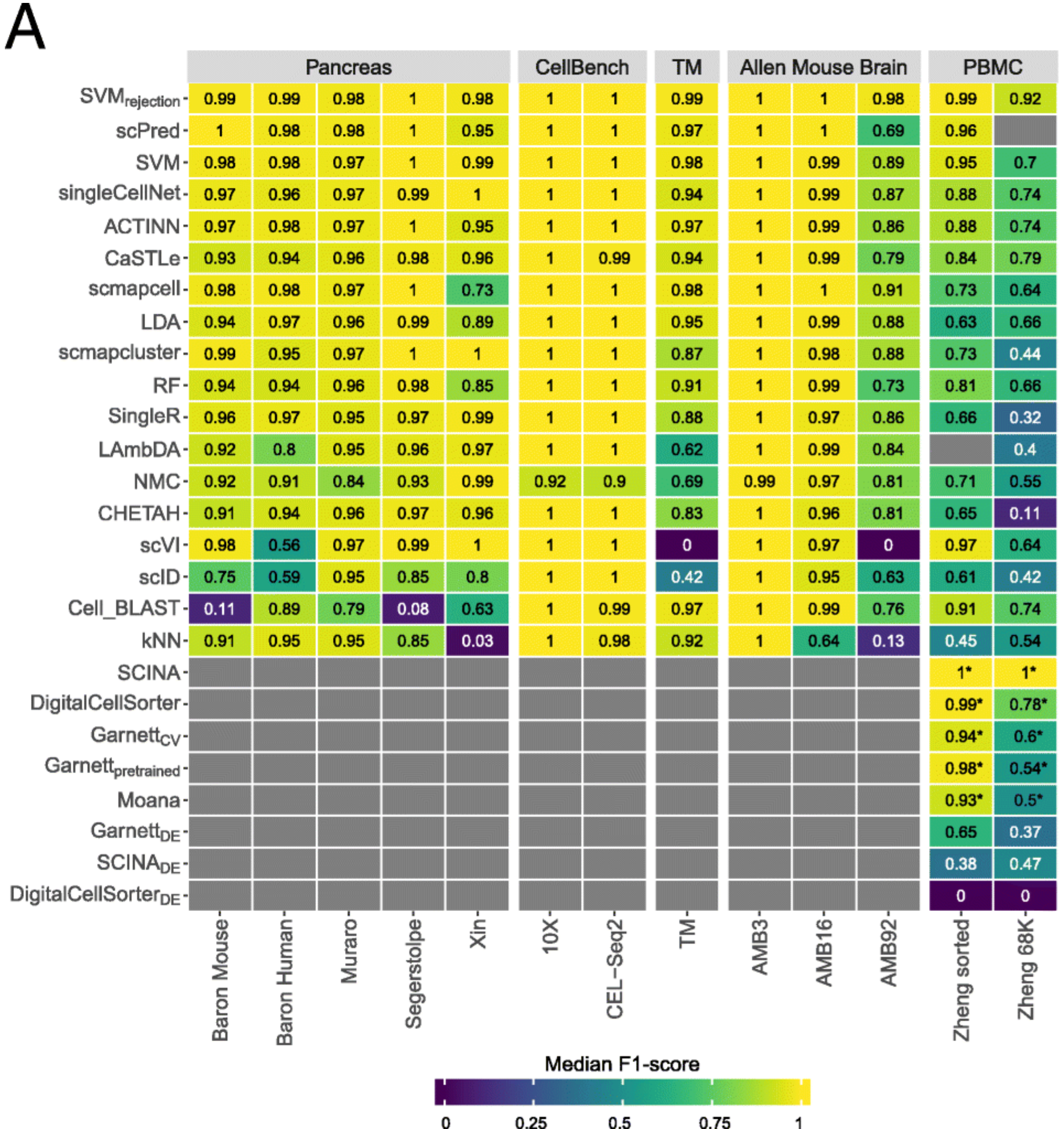
Name	Version	Language	Underlying classifier	Prior knowledge	Rejection option	Reference
Garnett	0.1.4	R	Generalized linear model	Yes	Yes	[14]
Moana	0.1.1	Python	SVM with linear kernel	Yes	No	[15]
DigitalCellSorter	GitHub version: e369a34	Python	Voting based on cell type markers	Yes	No	[16]
SCINA	1.1.0	R	Bimodal distribution fitting for marker genes	Yes	No	[17]
scVI	0.3.0	Python	Neural network	No	No	[18]
Cell-BLAST	0.1.2	Python	Cell-to-cell similarity	No	Yes	[19]
ACTINN	GitHub version: 563bcc1	Python	Neural network	No	No	[20]
LAmbDA	GitHub version: 3891d72	Python	Random forest	No	No	[21]
scmapcluster	1.5.1	R	Nearest median classifier	No	Yes	[22]
scmapcell	1.5.1	R	kNN	No	Yes	[22]
scPred	0.0.0.9000	R	SVM with radial kernel	No	Yes	[23]
CHETAH	0.99.5	R	Correlation to training set	No	Yes	[24]
CaSTLe	GitHub version: 258b278	R	Random forest	No	No	[25]
SingleR	0.2.2	R	Correlation to training set	No	No	[26]
scID	0.0.0.9000	R	LDA	No	Yes	[27]
singleCellNet	0.1.0	R	Random forest	No	No	[28]
LDA	0.19.2	Python	LDA	No	No	[29]
NMC	0.19.2	Python	NMC	No	No	[29]
RF	0.19.2	Python	RF (50 trees)	No	No	[29]
SVM	0.19.2	Python	SVM (linear kernel)	No	No	[29]
$\text{SVM}_{\text{rejection}}$	0.19.2	Python	SVM (linear kernel)	No	Yes	[29]
kNN	0.19.2	Python	kNN ($k = 9$)	No	No	[29]

Abdelaal et al.

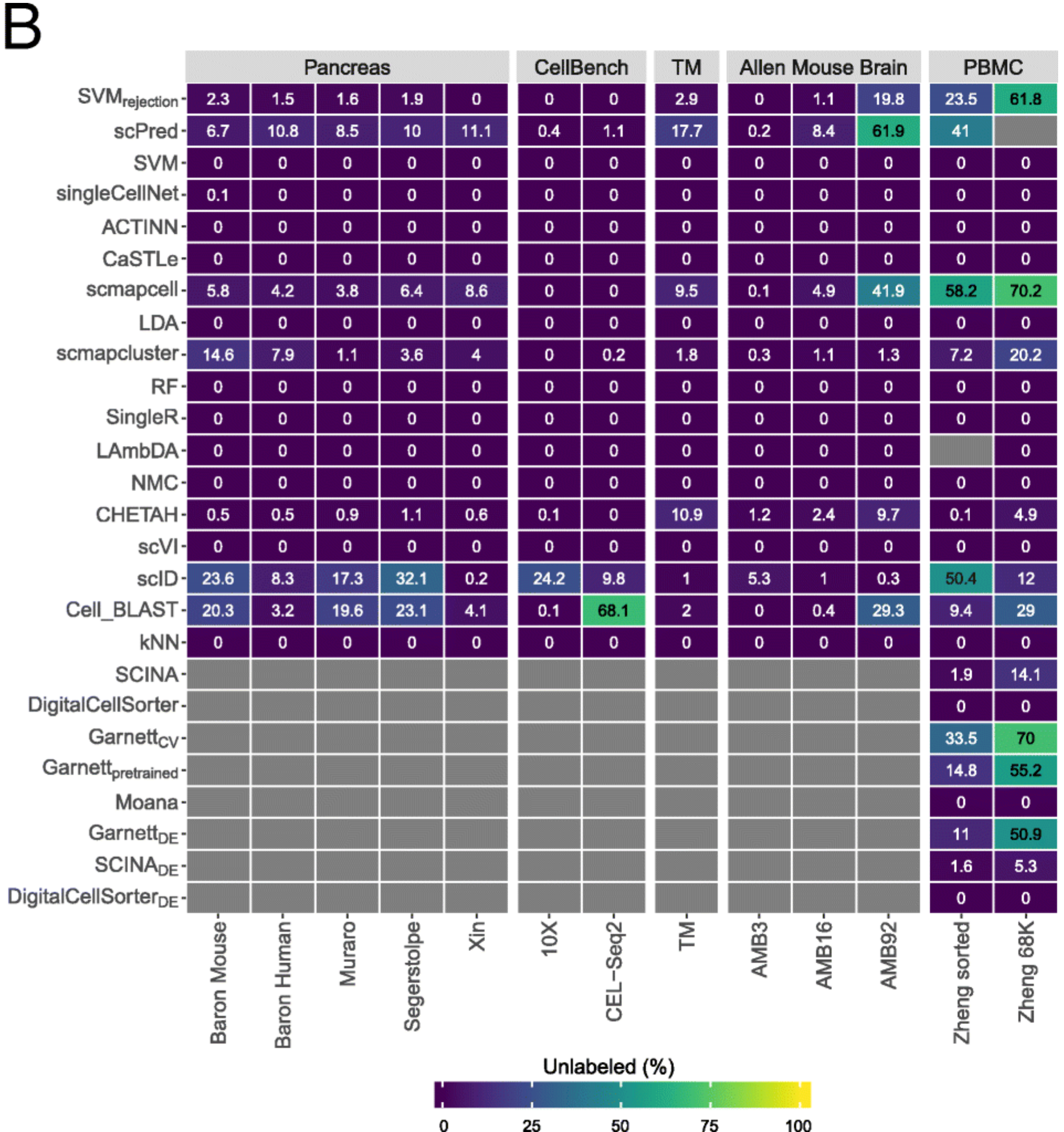


Dataset	No. of cells	No. of genes	No. of cell populations (> 10 cells)	Description	Protocol	Reference
Baron (Mouse)[a]	1886	14,861	13 (9)	Mouse pancreas	inDrop	[30]
Baron (Human)[a,b]	8569	17,499	14 (13)	Human pancreas	inDrop	[30]
Muraro[a,b]	2122	18,915	9 (8)	Human pancreas	CEL-Seq2	[31]
Segerstolpe[a,b]	2133	22,757	13 (9)	Human pancreas	SMART-Seq2	[32]
Xin[a,b]	1449	33,889	4 (4)	Human pancreas	SMARTer	[33]
CellBench 10X[a,b]	3803	11,778	5 (5)	Mixture of five human lung cancer cell lines	10X chromium	[34]
CellBench CEL-Seq2[a,b]	570	12,627	5 (5)	Mixture of five human lung cancer cell lines	CEL-Seq2	[34]
TM[a]	54,865	19,791	55 (55)	Whole *Mus musculus*	SMART-Seq2	[6]
AMB[a]	12,832	42,625	4/22/110 (3/16/92)	Primary mouse visual cortex	SMART-Seq v4	[35]
Zheng sorted[a]	20,000	21,952	10 (10)	FACS-sorted PBMC	10X CHROMIUM	[36]
Zheng 68K[a]	65,943	20,387	11 (11)	PBMC	10X CHROMIUM	[36]
VISp[b] (Mouse)	12,832	42,625	3/36 (3/34)	Primary visual cortex	SMART-Seq v4	[35]
ALM[b] (Mouse)	8758	42,461	3/37 (3/34)	Anterior lateral motor area	SMART-Seq v4	[35]
MTG[b] (Human)	14,636	16,161	3/35 (3/34)	Middle temporal gyrus	SMART-Seq v4	[37]
PbmcBench pbmc1.10Xv2[b]	6444	33,694	9 (9)	PBMC	10X version 2	[38]
PbmcBench pbmc1.10Xv3[b]	3222	33,694	8 (8)	PBMC	10X version 3	[38]
PbmcBench pbmc1.CL[b]	253	33,694	7 (7)	PBMC	CEL-Seq2	[38]
PbmcBench pbmc1.DR[b]	3222	33,694	9 (9)	PBMC	Drop-Seq	[38]
PbmcBench pbmc1.iD[b]	3222	33,694	7 (7)	PBMC	inDrop	[38]
PbmcBench pbmc1.SM2[b]	253	33,694	6 (6)	PBMC	SMART-Seq2	[38]
PbmcBench pbmc1.SW[b]	3176	33,694	7 (7)	PBMC	Seq-Well	[38]
PbmcBench pbmc2.10Xv2,[b]	3362	33,694	9 (9)	PBMC	10X version 2	[38]
PbmcBench pbmc2.CL[b]	273	33,694	5 (5)	PBMC	CEL-Seq2	[38]
PbmcBench pbmc2.DR[b]	3362	33,694	6 (6)	PBMC	Drop-Seq	[38]
PbmcBench pbmc2.iD[b]	3362	33,694	9 (9)	PBMC	inDrop	[38]
PbmcBench pbmc2.SM2[b]	273	33,694	6 (6)	PBMC	SMART-Seq2	[38]
PbmcBench pbmc2.SW[b]	551	33,694	4 (4)	PBMC	Seq-Well	[38]

Abdelaal et al. – All data



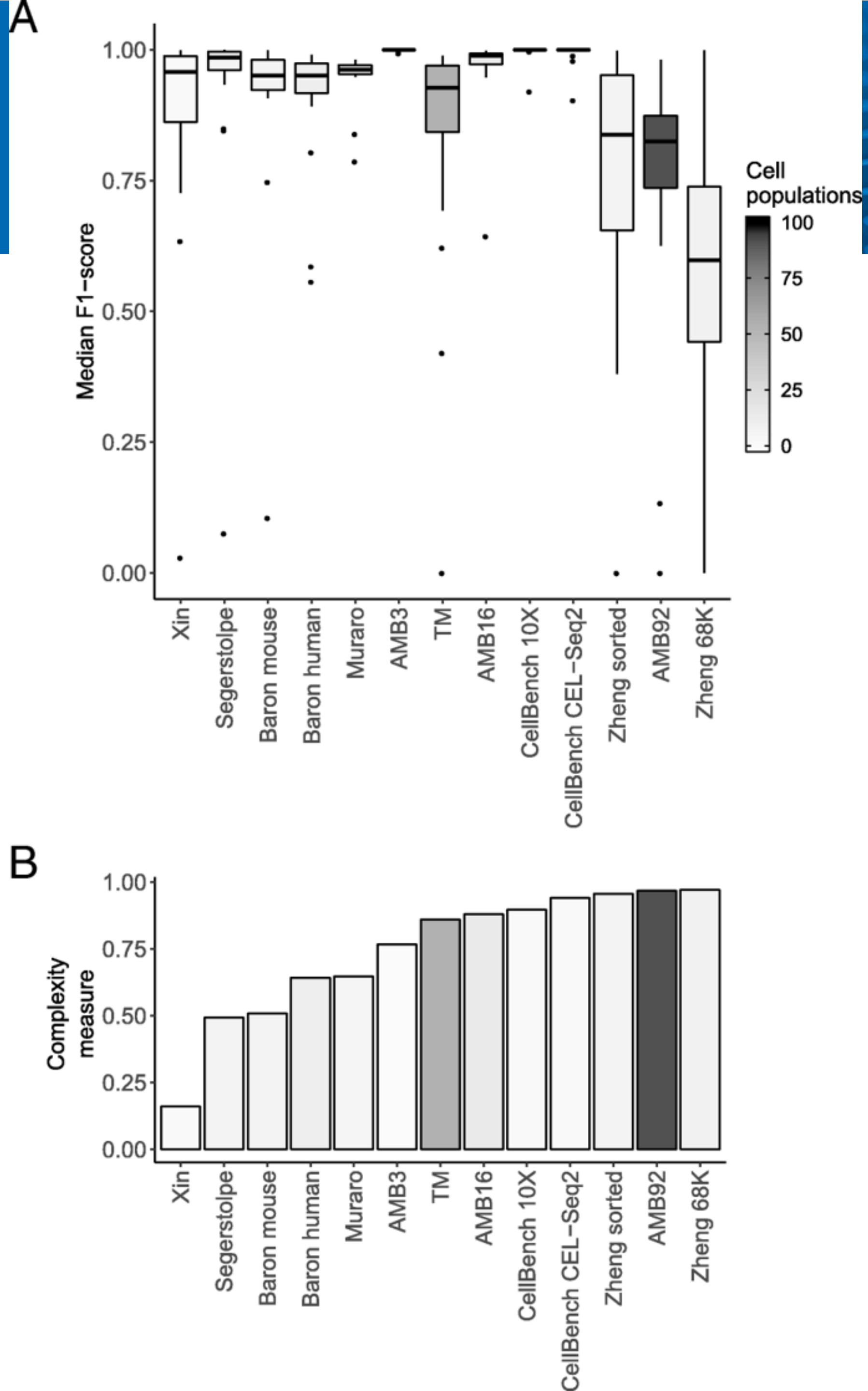
	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K
SVM$_{rejection}$	0.99	0.99	0.98	1	0.98	1	1	0.99	1	1	0.98	0.99	0.92
scPred	1	0.98	0.98	1	0.95	1	1	0.97	1	1	0.69	0.96	
SVM	0.98	0.98	0.97	1	0.99	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	0.94	0.97	0.96	0.99	0.89	1	1	0.95	1	0.99	0.88	0.63	0.66
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



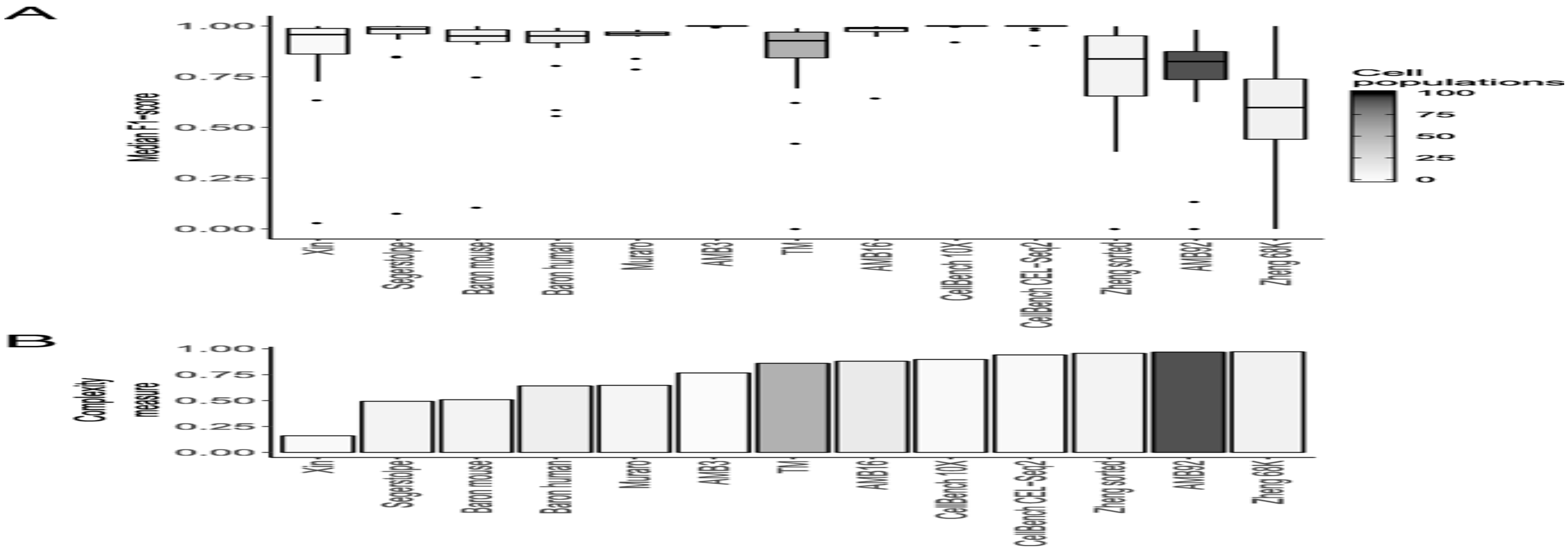
	Pancreas					CellBench		TM	Allen Mouse Brain			PBMC	
	Baron Mouse	Baron Human	Muraro	Segerstolpe	Xin	10X	CEL−Seq2	TM	AMB3	AMB16	AMB92	Zheng sorted	Zheng 68K
SVM$_{rejection}$	2.3	1.5	1.6	1.9	0	0	0	2.9	0	1.1	19.8	23.5	61.8
scPred	6.7	10.8	8.5	10	11.1	0.4	1.1	17.7	0.2	8.4	61.9	41	
SVM	0	0	0	0	0	0	0	0	0	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

Abdelaal et al. – All data

Frederick National Laboratory for Cancer Research

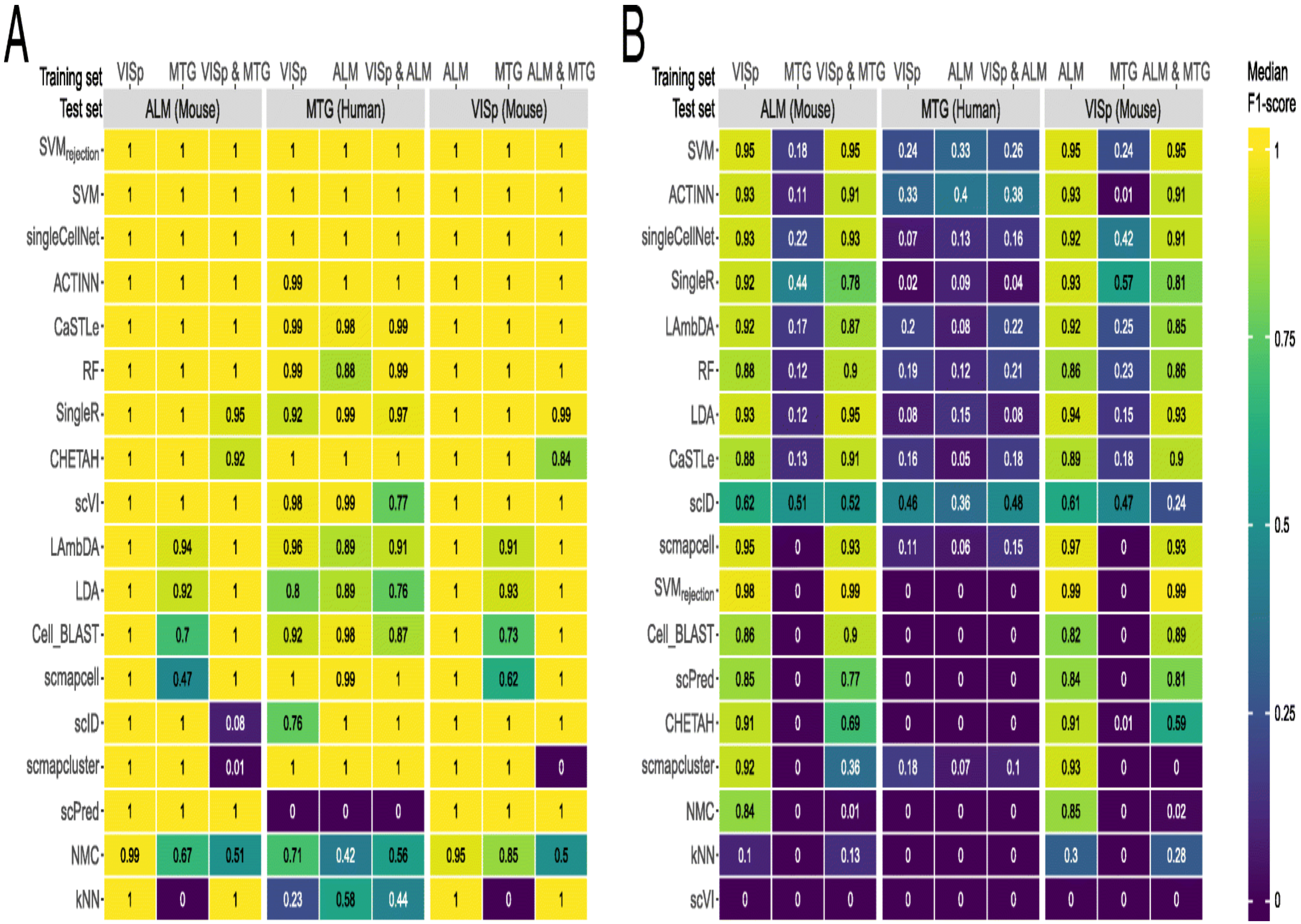


Abdelaal et al. - Brain



Abdelaal et al. - Pancreas

Frederick National Laboratory for Cancer Research



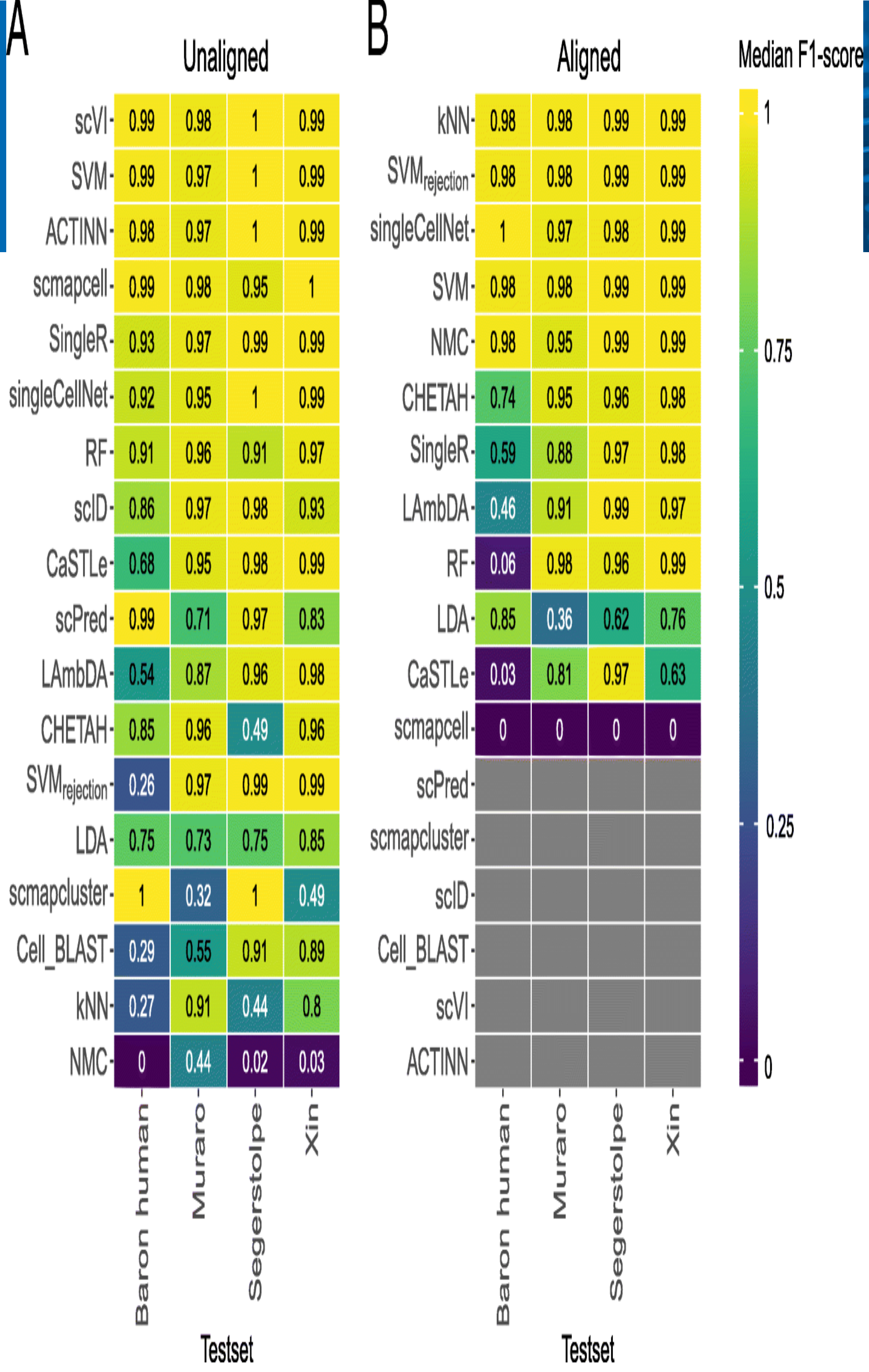
A

Test set	ALM (Mouse)			MTG (Human)			VISp (Mouse)		
Training set	VISp	MTG	VISp & MTG	VISp	ALM	VISp & ALM	ALM	MTG	ALM & MTG
$SVM_{rejection}$	1	1	1	1	1	1	1	1	1
SVM	1	1	1	1	1	1	1	1	1
singleCellNet	1	1	1	1	1	1	1	1	1
ACTINN	1	1	1	0.99	1	1	1	1	1
CaSTLe	1	1	1	0.99	0.98	0.99	1	1	1
RF	1	1	1	0.99	0.88	0.99	1	1	1
SingleR	1	1	0.95	0.92	0.99	0.97	1	1	0.99
CHETAH	1	1	0.92	1	1	1	1	1	0.84
scVI	1	1	1	0.98	0.99	0.77	1	1	1
LAmbDA	1	0.94	1	0.96	0.89	0.91	1	0.91	1
LDA	1	0.92	1	0.8	0.89	0.76	1	0.93	1
Cell_BLAST	1	0.7	1	0.92	0.98	0.87	1	0.73	1
scmapcell	1	0.47	1	1	0.99	1	1	0.62	1
scID	1	1	0.08	0.76	1	1	1	1	1
scmapcluster	1	1	0.01	1	1	1	1	1	0
scPred	1	1	1	0	0	0	1	1	1
NMC	0.99	0.67	0.51	0.71	0.42	0.56	0.95	0.85	0.5
kNN	1	0	1	0.23	0.58	0.44	1	0	1

B

Test set	ALM (Mouse)			MTG (Human)			VISp (Mouse)		
Training set	VISp	MTG	VISp & MTG	VISp	ALM	VISp & ALM	ALM	MTG	ALM & MTG
SVM	0.95	0.18	0.95	0.24	0.33	0.26	0.95	0.24	0.95
ACTINN	0.93	0.11	0.91	0.33	0.4	0.38	0.93	0.01	0.91
singleCellNet	0.93	0.22	0.93	0.07	0.13	0.16	0.92	0.42	0.91
SingleR	0.92	0.44	0.78	0.02	0.09	0.04	0.93	0.57	0.81
LAmbDA	0.92	0.17	0.87	0.2	0.08	0.22	0.92	0.25	0.85
RF	0.88	0.12	0.9	0.19	0.12	0.21	0.86	0.23	0.86
LDA	0.93	0.12	0.95	0.08	0.15	0.08	0.94	0.15	0.93
CaSTLe	0.88	0.13	0.91	0.16	0.05	0.18	0.89	0.18	0.9
scID	0.62	0.51	0.52	0.46	0.36	0.48	0.61	0.47	0.24
scmapcell	0.95	0	0.93	0.11	0.06	0.15	0.97	0	0.93
$SVM_{rejection}$	0.98	0	0.99	0	0	0	0.99	0	0.99
Cell_BLAST	0.86	0	0.9	0	0	0	0.82	0	0.89
scPred	0.85	0	0.77	0	0	0	0.84	0	0.81
CHETAH	0.91	0	0.69	0	0	0	0.91	0.01	0.59
scmapcluster	0.92	0	0.36	0.18	0.07	0.1	0.93	0	0
NMC	0.84	0	0.01	0	0	0	0.85	0	0.02
kNN	0.1	0	0.13	0	0	0	0.3	0	0.28
scVI	0	0	0	0	0	0	0	0	0

Abdelaal et al.

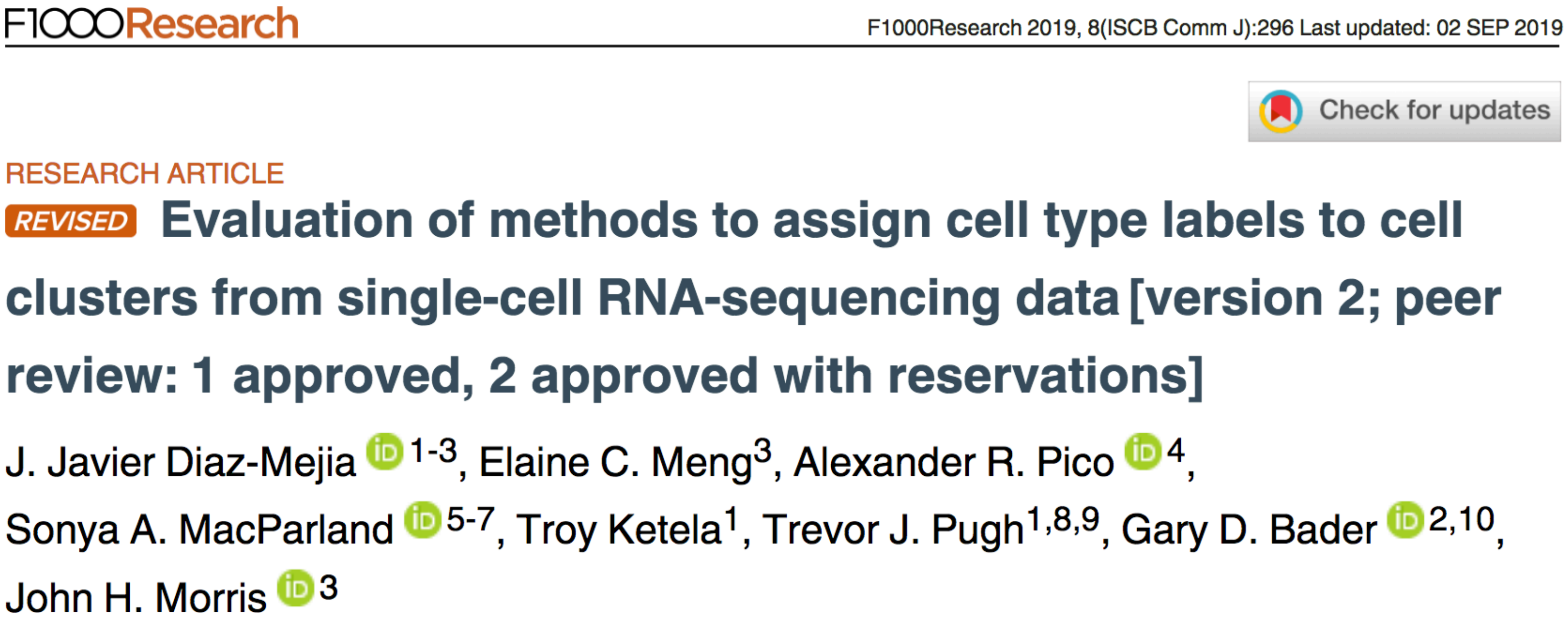


A — Unaligned

	Baron human	Muraro	Segerstolpe	Xin
scVI	0.99	0.98	1	0.99
SVM	0.99	0.97	1	0.99
ACTINN	0.98	0.97	1	0.99
scmapcell	0.99	0.98	0.95	1
SingleR	0.93	0.97	0.99	0.99
singleCellNet	0.92	0.95	1	0.99
RF	0.91	0.96	0.91	0.97
scID	0.86	0.97	0.98	0.93
CaSTLe	0.68	0.95	0.98	0.99
scPred	0.99	0.71	0.97	0.83
LAmbDA	0.54	0.87	0.96	0.98
CHETAH	0.85	0.96	0.49	0.96
$SVM_{rejection}$	0.26	0.97	0.99	0.99
LDA	0.75	0.73	0.75	0.85
scmapcluster	1	0.32	1	0.49
Cell_BLAST	0.29	0.55	0.91	0.89
kNN	0.27	0.91	0.44	0.8
NMC	0	0.44	0.02	0.03

B — Aligned

	Baron human	Muraro	Segerstolpe	Xin
kNN	0.98	0.98	0.99	0.99
$SVM_{rejection}$	0.98	0.98	0.99	0.99
singleCellNet	1	0.97	0.98	0.99
SVM	0.98	0.98	0.99	0.99
NMC	0.98	0.95	0.99	0.99
CHETAH	0.74	0.95	0.96	0.98
SingleR	0.59	0.88	0.97	0.98
LAmbDA	0.46	0.91	0.99	0.97
RF	0.06	0.98	0.96	0.99
LDA	0.85	0.36	0.62	0.76
CaSTLe	0.03	0.81	0.97	0.63
scmapcell	0	0	0	0
scPred				
scmapcluster				
scID				
Cell_BLAST				
scVI				
ACTINN				



F1000Research

F1000Research 2019, 8(ISCB Comm J):296 Last updated: 02 SEP 2019

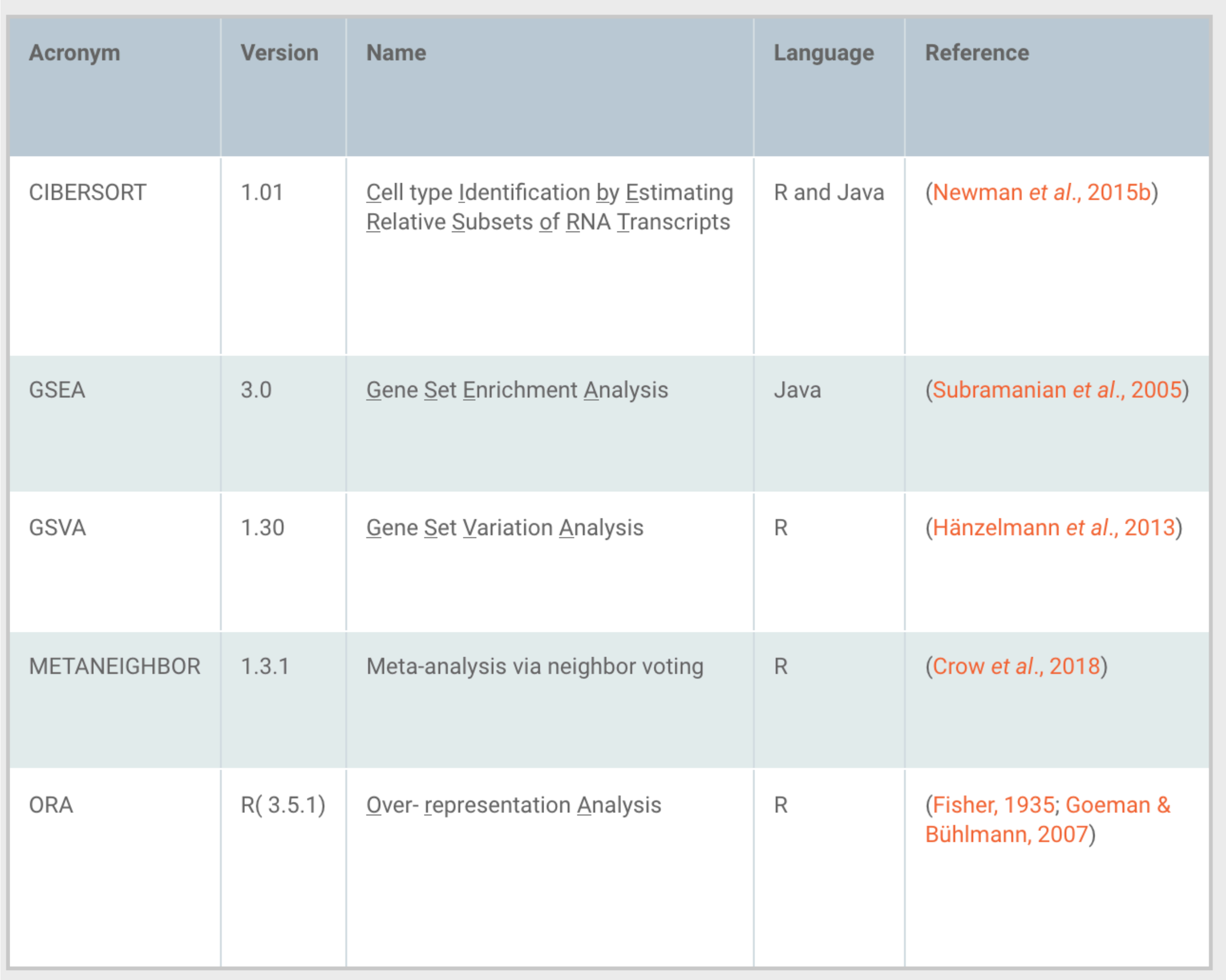
Check for updates

RESEARCH ARTICLE

REVISED Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data [version 2; peer review: 1 approved, 2 approved with reservations]

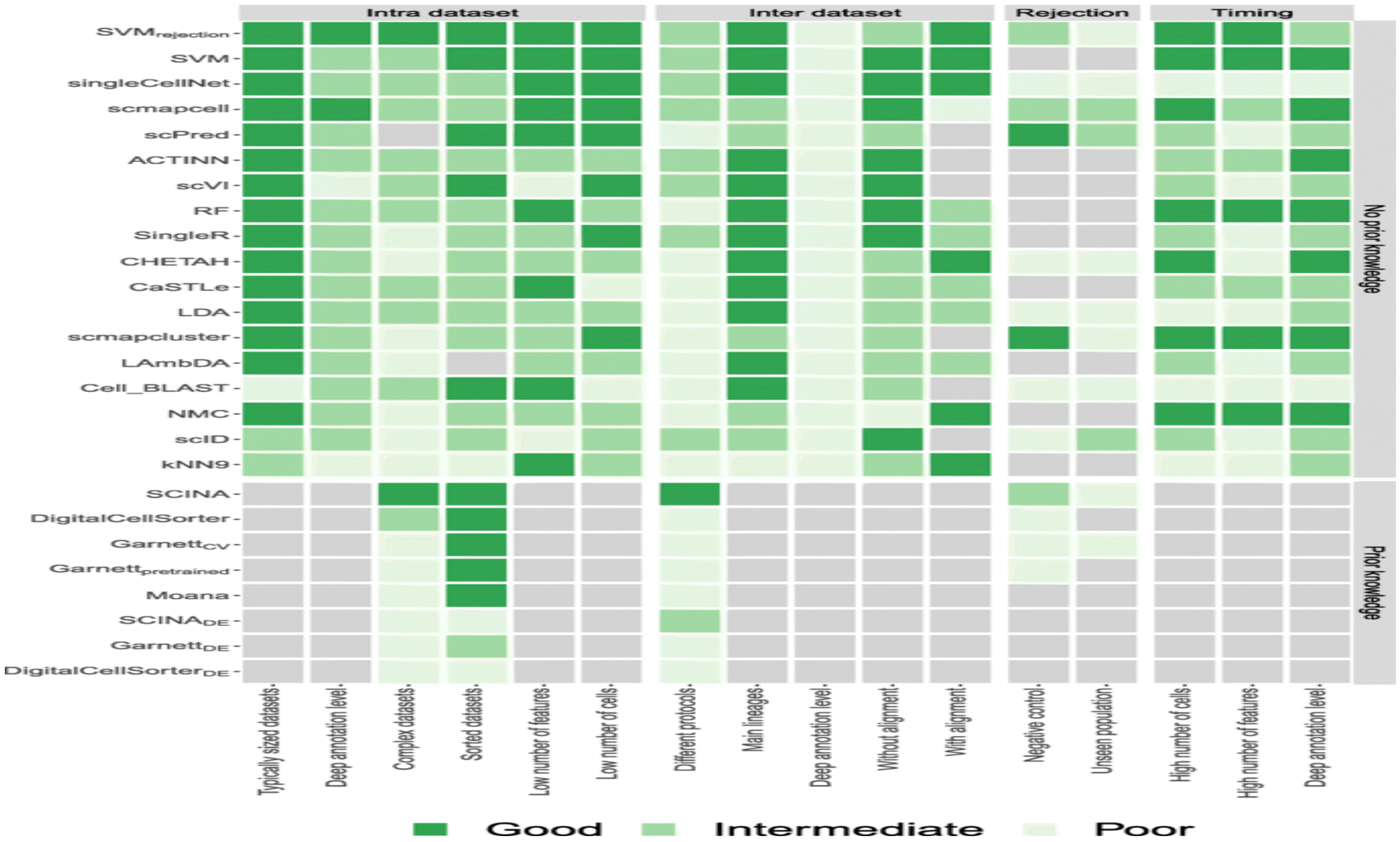
J. Javier Diaz-Mejia[1-3], Elaine C. Meng[3], Alexander R. Pico[4], Sonya A. MacParland[5-7], Troy Ketela[1], Trevor J. Pugh[1,8,9], Gary D. Bader[2,10], John H. Morris[3]

Diaz-Mejia et al.

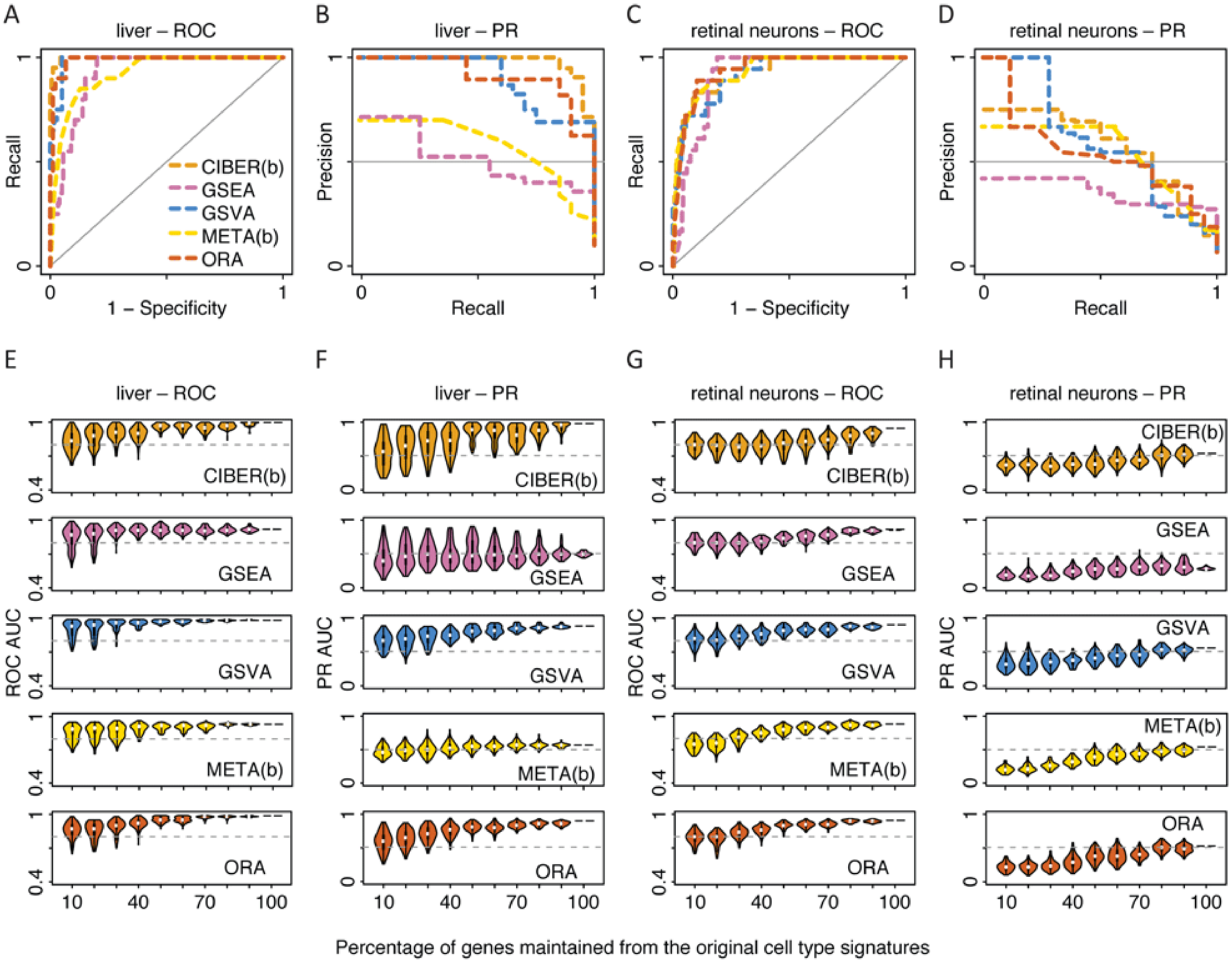


Acronym	Version	Name	Language	Reference
CIBERSORT	1.01	Cell type Identification by Estimating Relative Subsets of RNA Transcripts	R and Java	(Newman *et al.*, 2015b)
GSEA	3.0	Gene Set Enrichment Analysis	Java	(Subramanian *et al.*, 2005)
GSVA	1.30	Gene Set Variation Analysis	R	(Hänzelmann *et al.*, 2013)
METANEIGHBOR	1.3.1	Meta-analysis via neighbor voting	R	(Crow *et al.*, 2018)
ORA	R(3.5.1)	Over- representation Analysis	R	(Fisher, 1935; Goeman & Bühlmann, 2007)

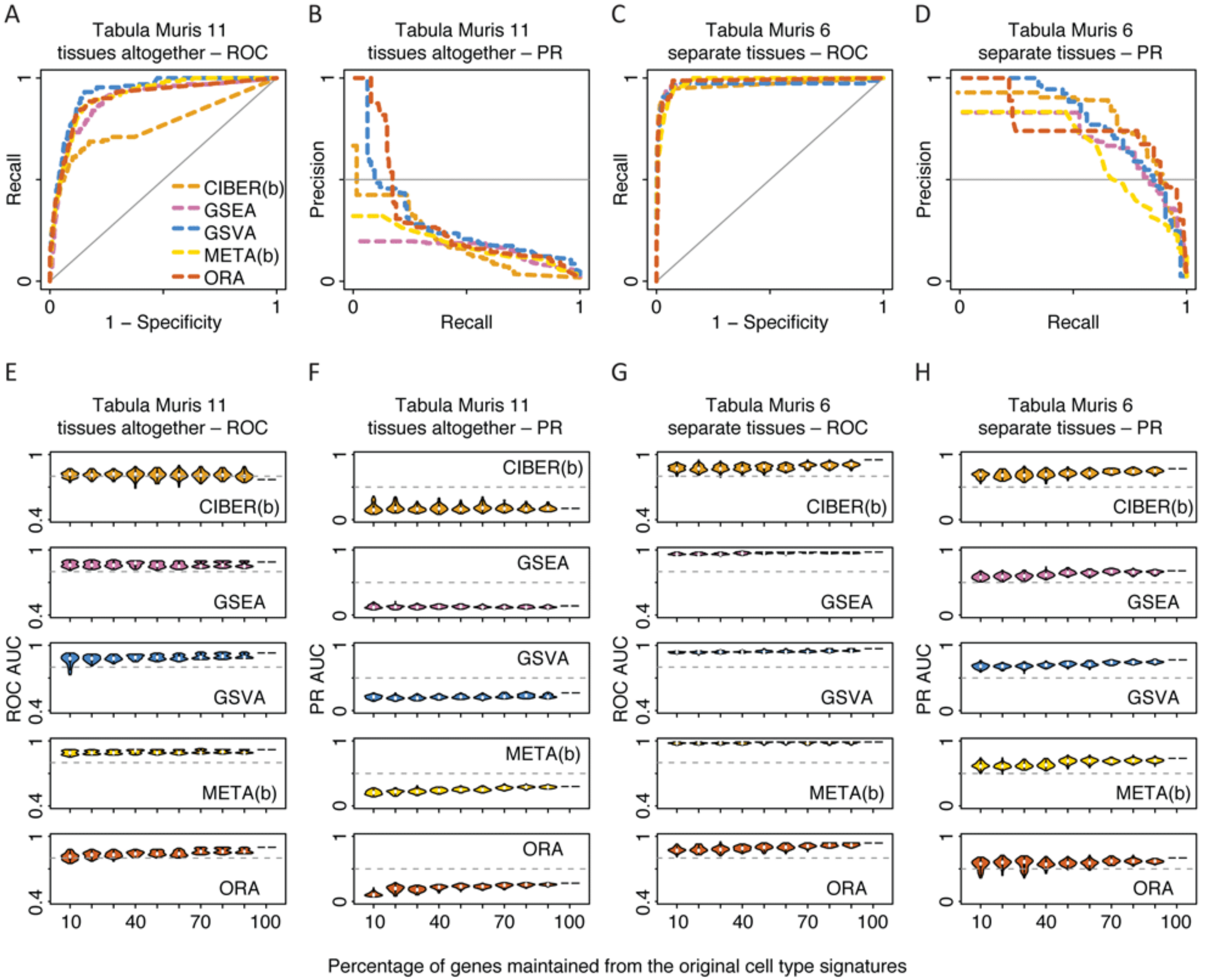
Diaz-Mejia et al.



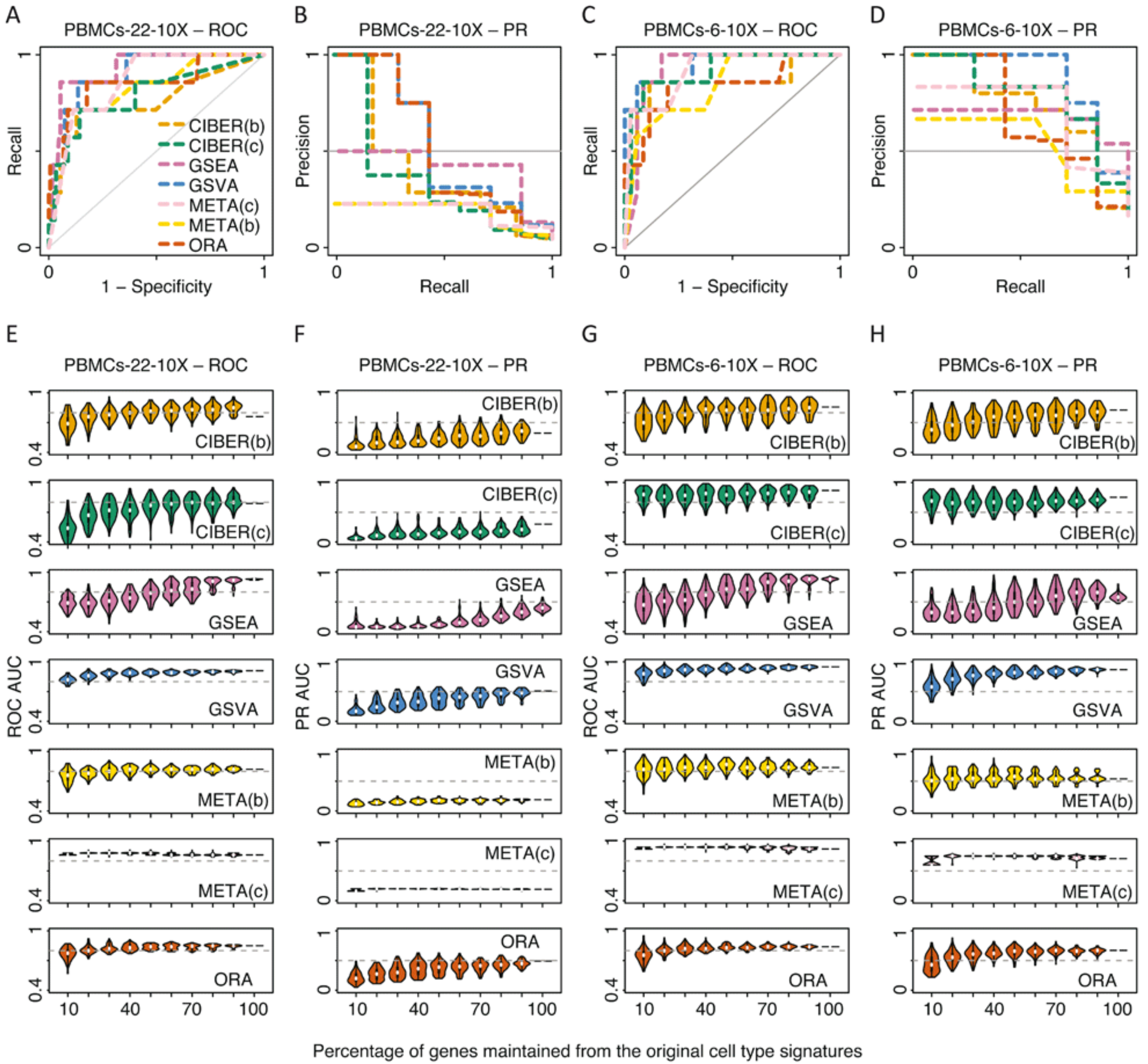
Diaz-Mejia et al.



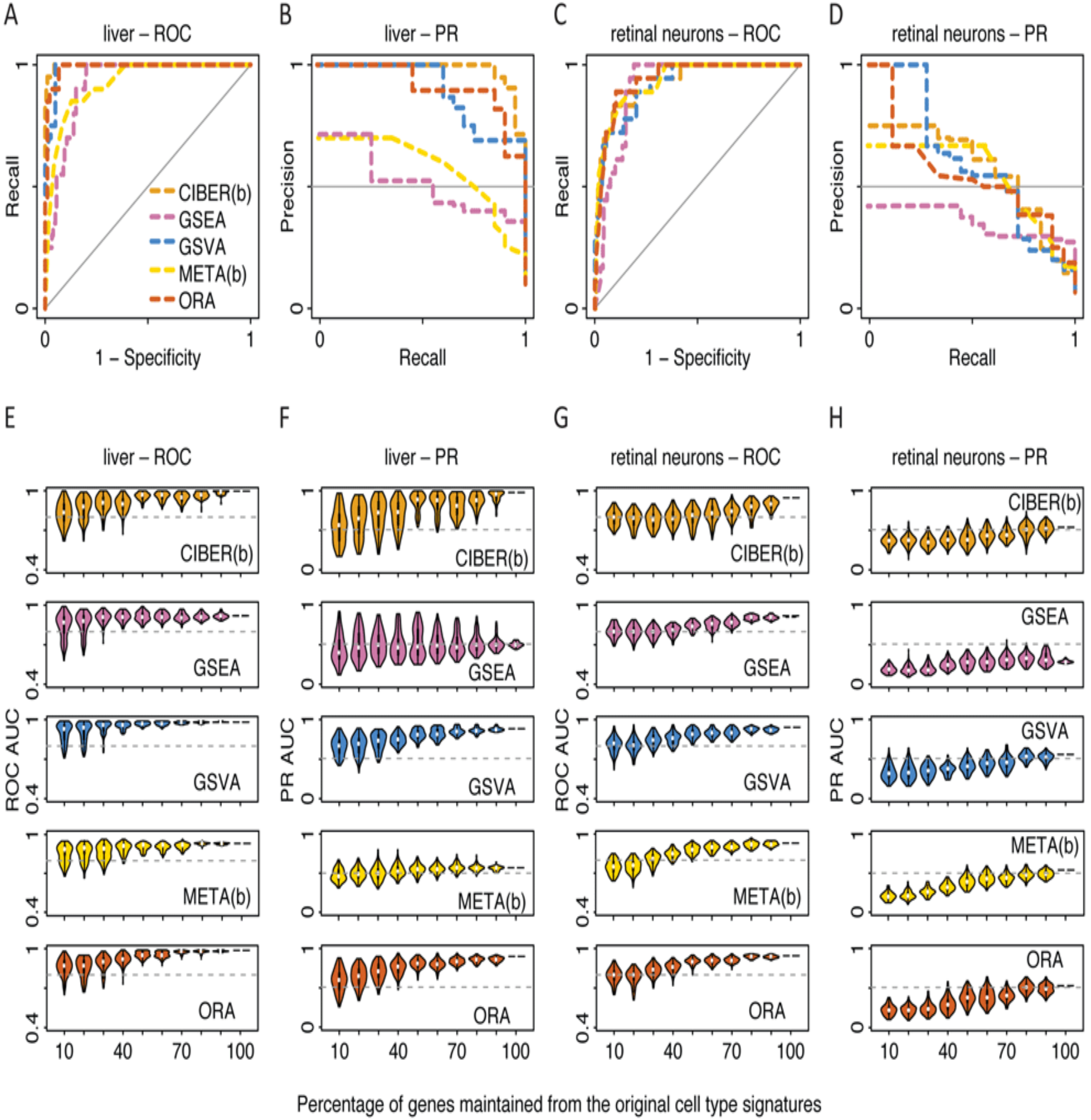
Diaz-Mejia et al.



Diaz-Mejia et al.

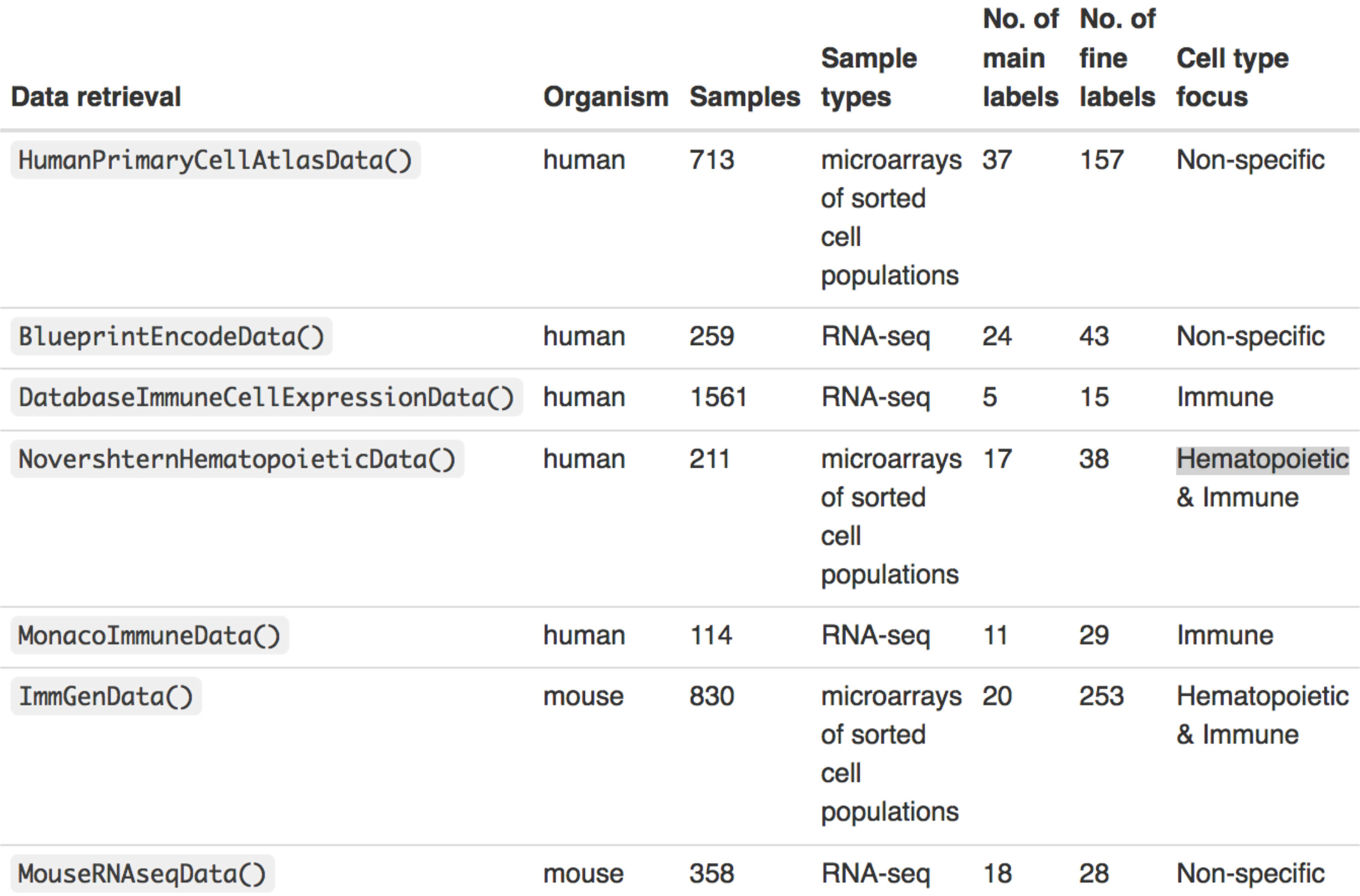


Diaz-Mejia et al.



Tools we use….

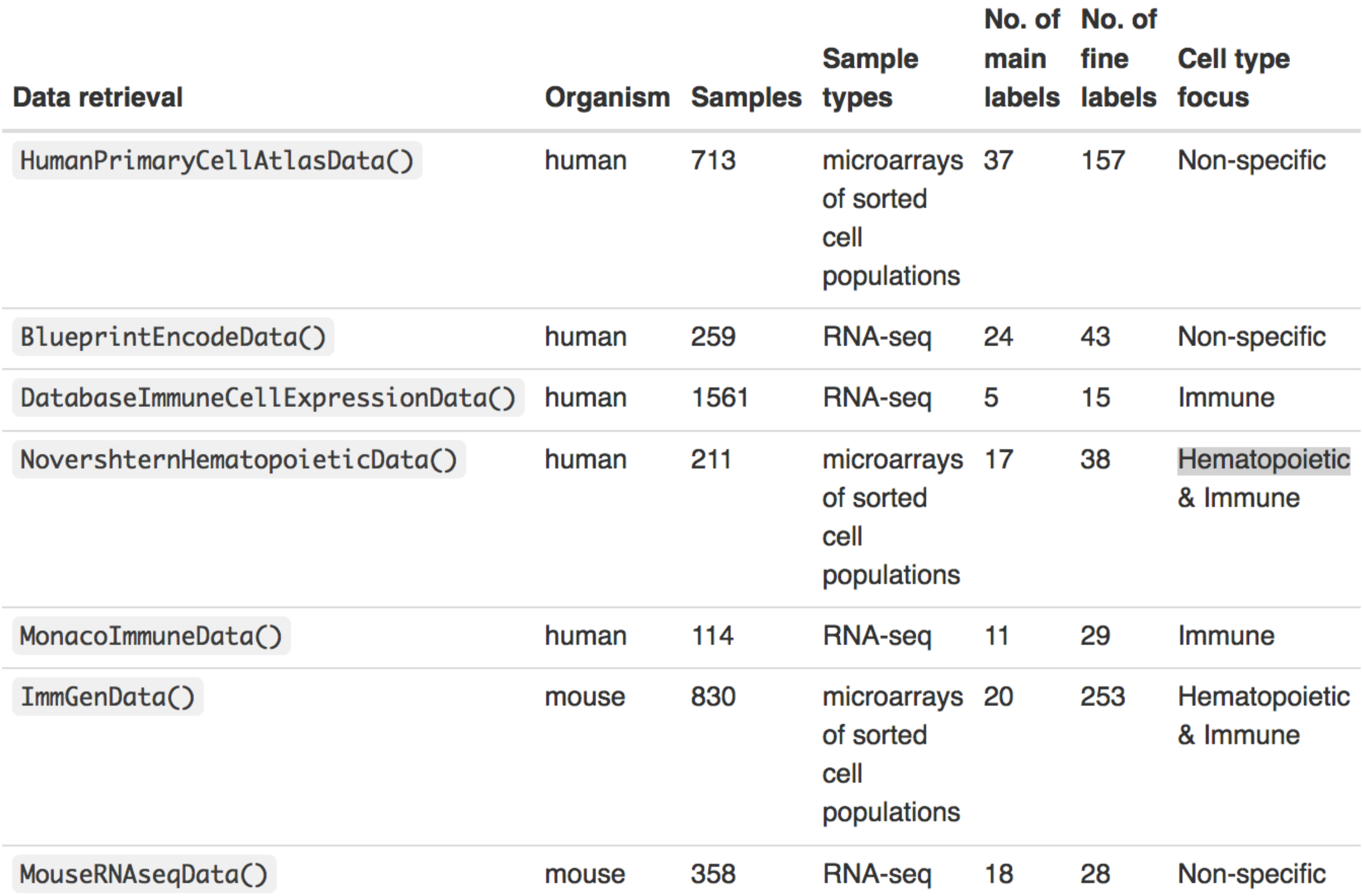
SingleR



Data retrieval	Organism	Samples	Sample types	No. of main labels	No. of fine labels	Cell type focus
`HumanPrimaryCellAtlasData()`	human	713	microarrays of sorted cell populations	37	157	Non-specific
`BlueprintEncodeData()`	human	259	RNA-seq	24	43	Non-specific
`DatabaseImmuneCellExpressionData()`	human	1561	RNA-seq	5	15	Immune
`NovershternHematopoieticData()`	human	211	microarrays of sorted cell populations	17	38	Hematopoietic & Immune
`MonacoImmuneData()`	human	114	RNA-seq	11	29	Immune
`ImmGenData()`	mouse	830	microarrays of sorted cell populations	20	253	Hematopoietic & Immune
`MouseRNAseqData()`	mouse	358	RNA-seq	18	28	Non-specific

Aran, Looney, Liu et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology (2019)

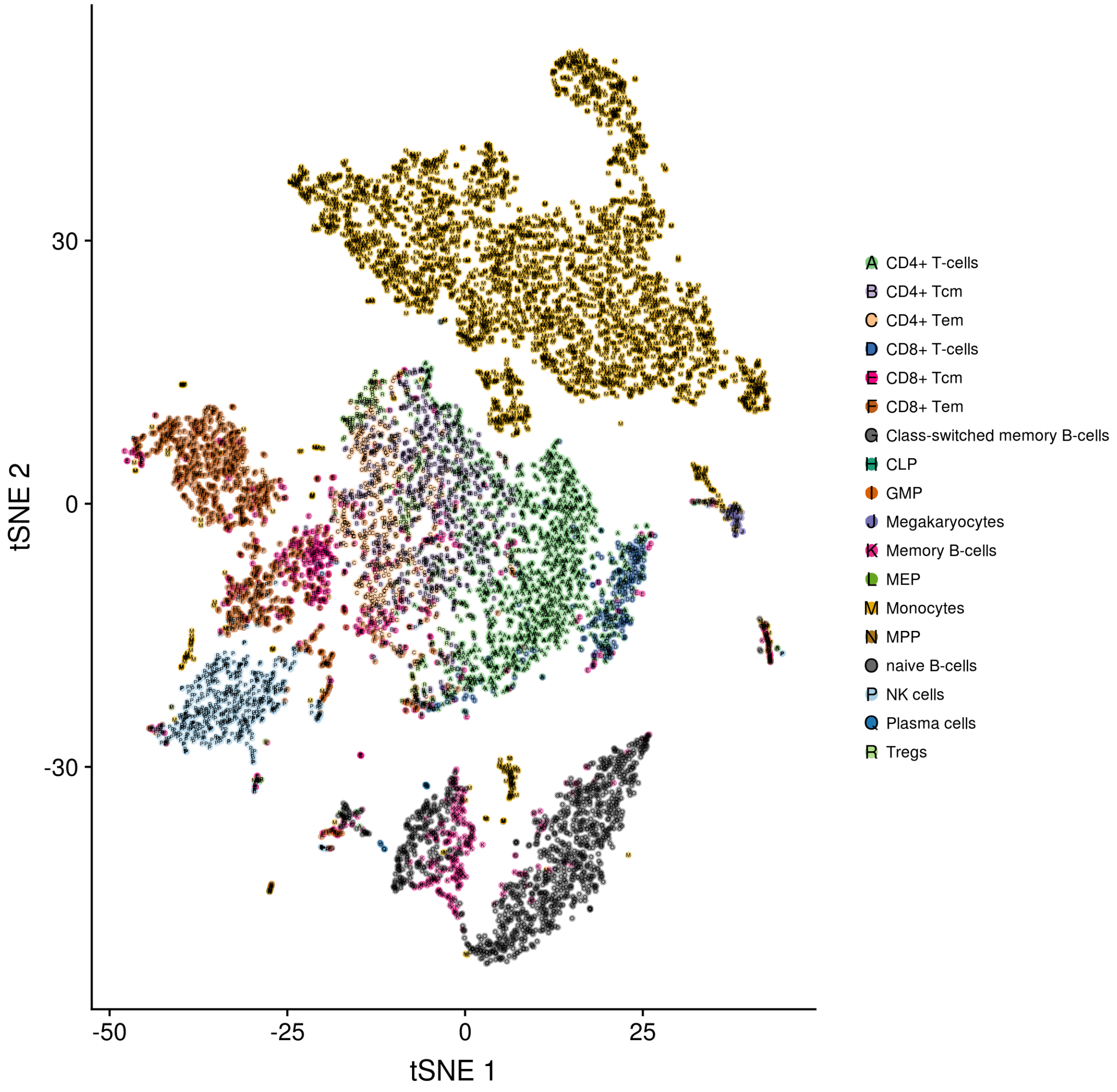
SingleR

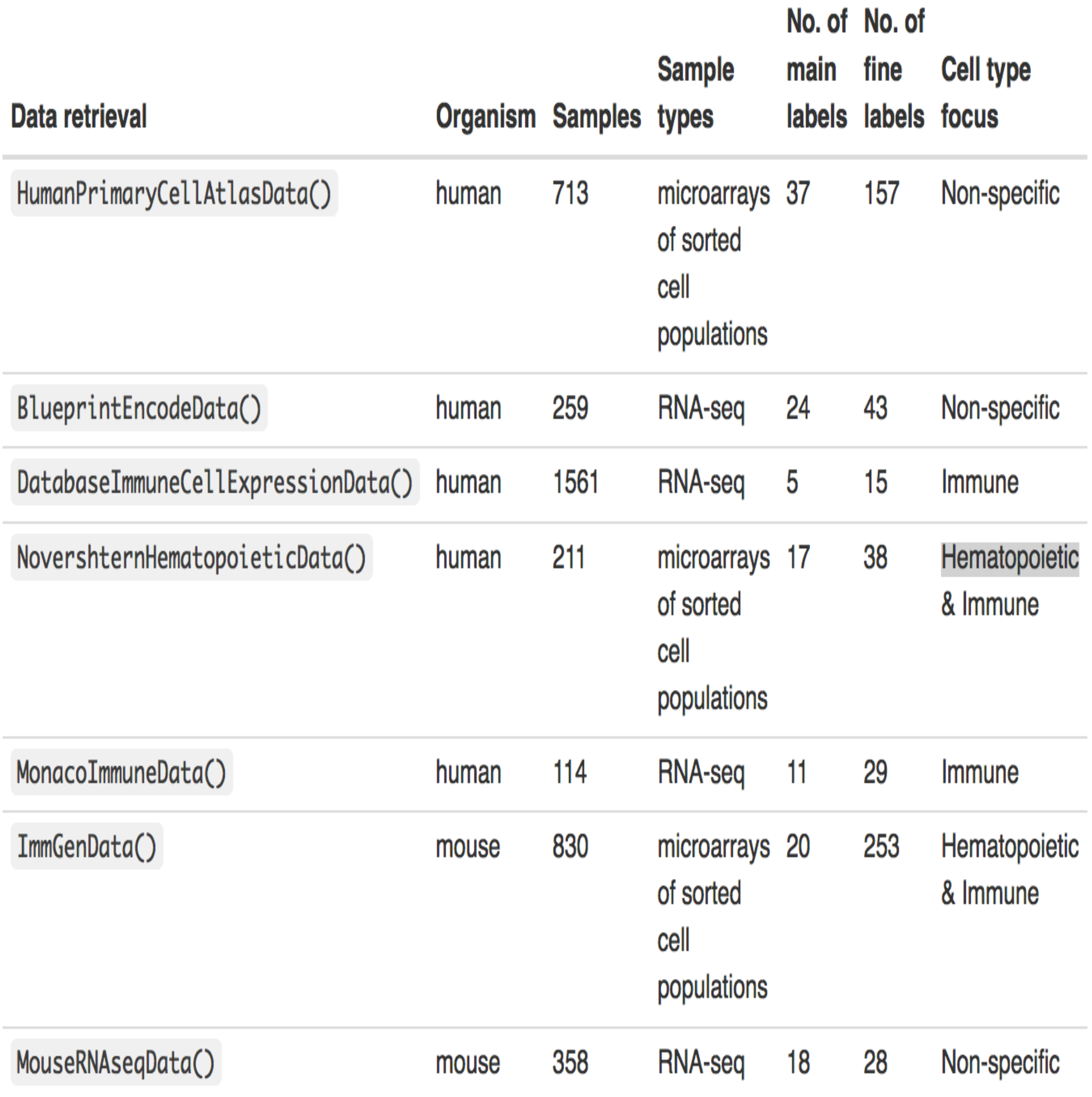


Data retrieval	Organism	Samples	Sample types	No. of main labels	No. of fine labels	Cell type focus
`HumanPrimaryCellAtlasData()`	human	713	microarrays of sorted cell populations	37	157	Non-specific
`BlueprintEncodeData()`	human	259	RNA-seq	24	43	Non-specific
`DatabaseImmuneCellExpressionData()`	human	1561	RNA-seq	5	15	Immune
`NovershternHematopoieticData()`	human	211	microarrays of sorted cell populations	17	38	Hematopoietic & Immune
`MonacoImmuneData()`	human	114	RNA-seq	11	29	Immune
`ImmGenData()`	mouse	830	microarrays of sorted cell populations	20	253	Hematopoietic & Immune
`MouseRNAseqData()`	mouse	358	RNA-seq	18	28	Non-specific

+ Any dataset which is annotated!

SingleR - PBMC

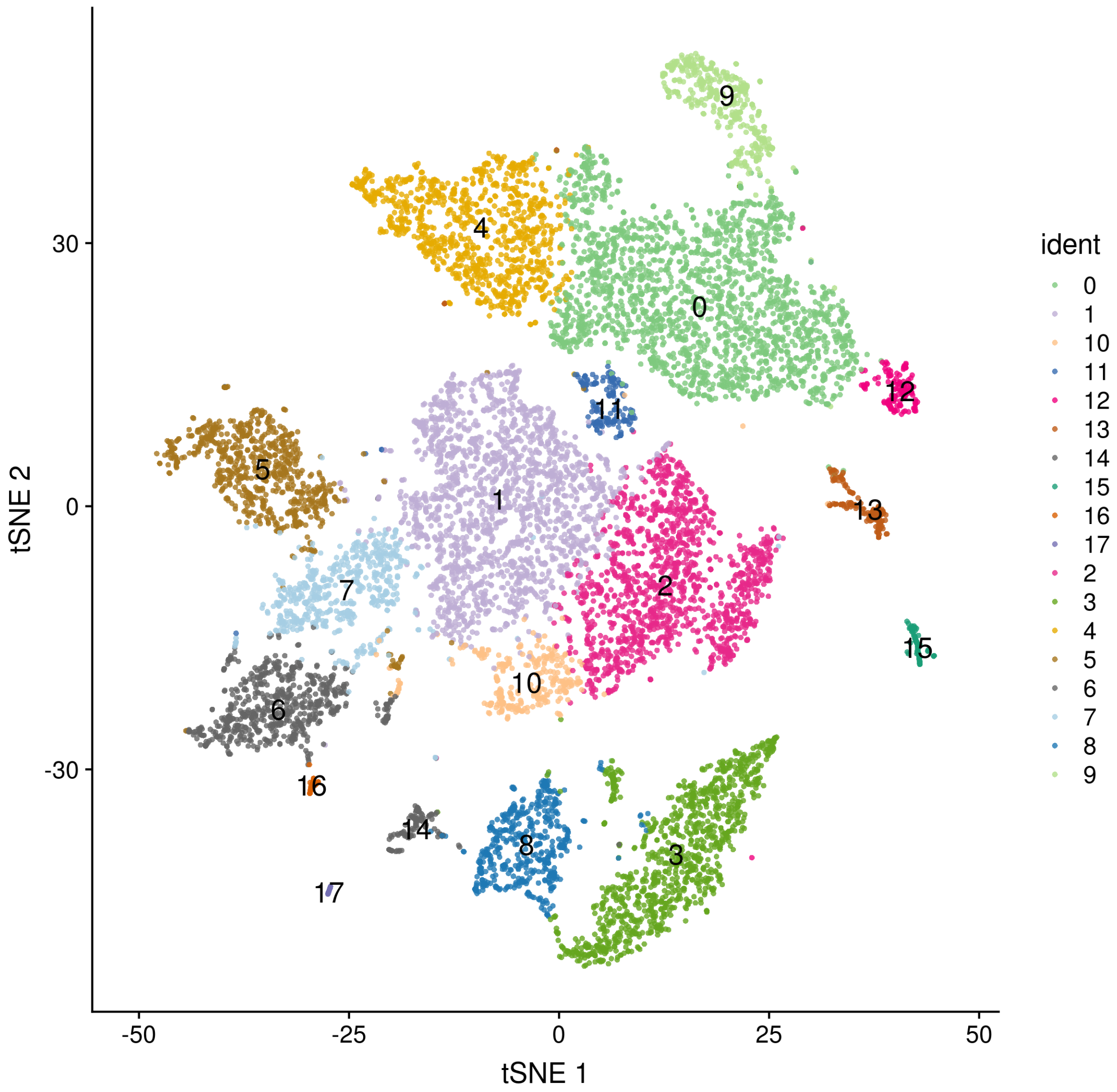


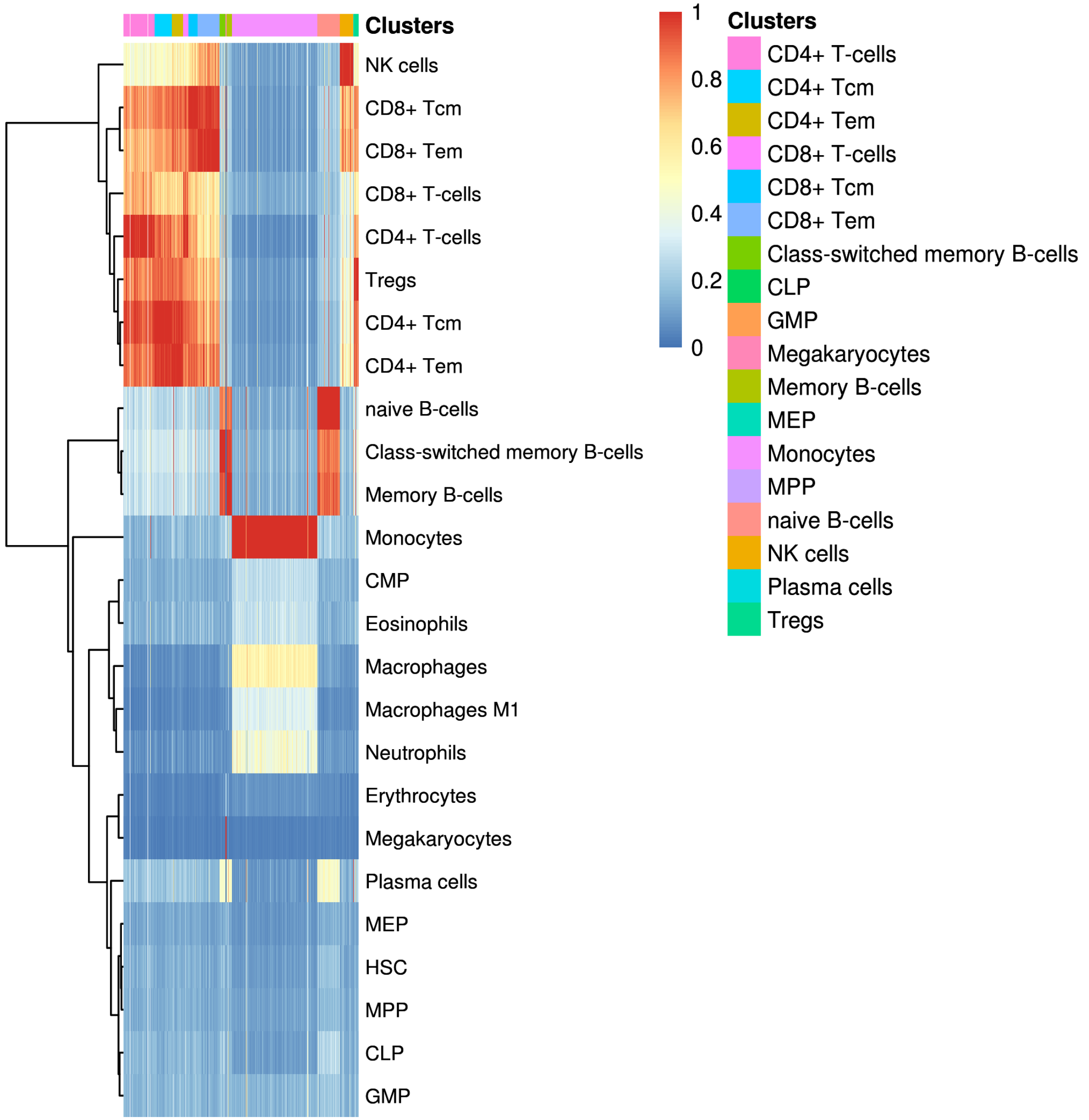


Data retrieval	Organism	Samples	Sample types	No. of main labels	No. of fine labels	Cell type focus
`HumanPrimaryCellAtlasData()`	human	713	microarrays of sorted cell populations	37	157	Non-specific
`BlueprintEncodeData()`	human	259	RNA-seq	24	43	Non-specific
`DatabaseImmuneCellExpressionData()`	human	1561	RNA-seq	5	15	Immune
`NovershternHematopoieticData()`	human	211	microarrays of sorted cell populations	17	38	Hematopoietic & Immune
`MonacoImmuneData()`	human	114	RNA-seq	11	29	Immune
`ImmGenData()`	mouse	830	microarrays of sorted cell populations	20	253	Hematopoietic & Immune
`MouseRNAseqData()`	mouse	358	RNA-seq	18	28	Non-specific

SingleR - PBMC

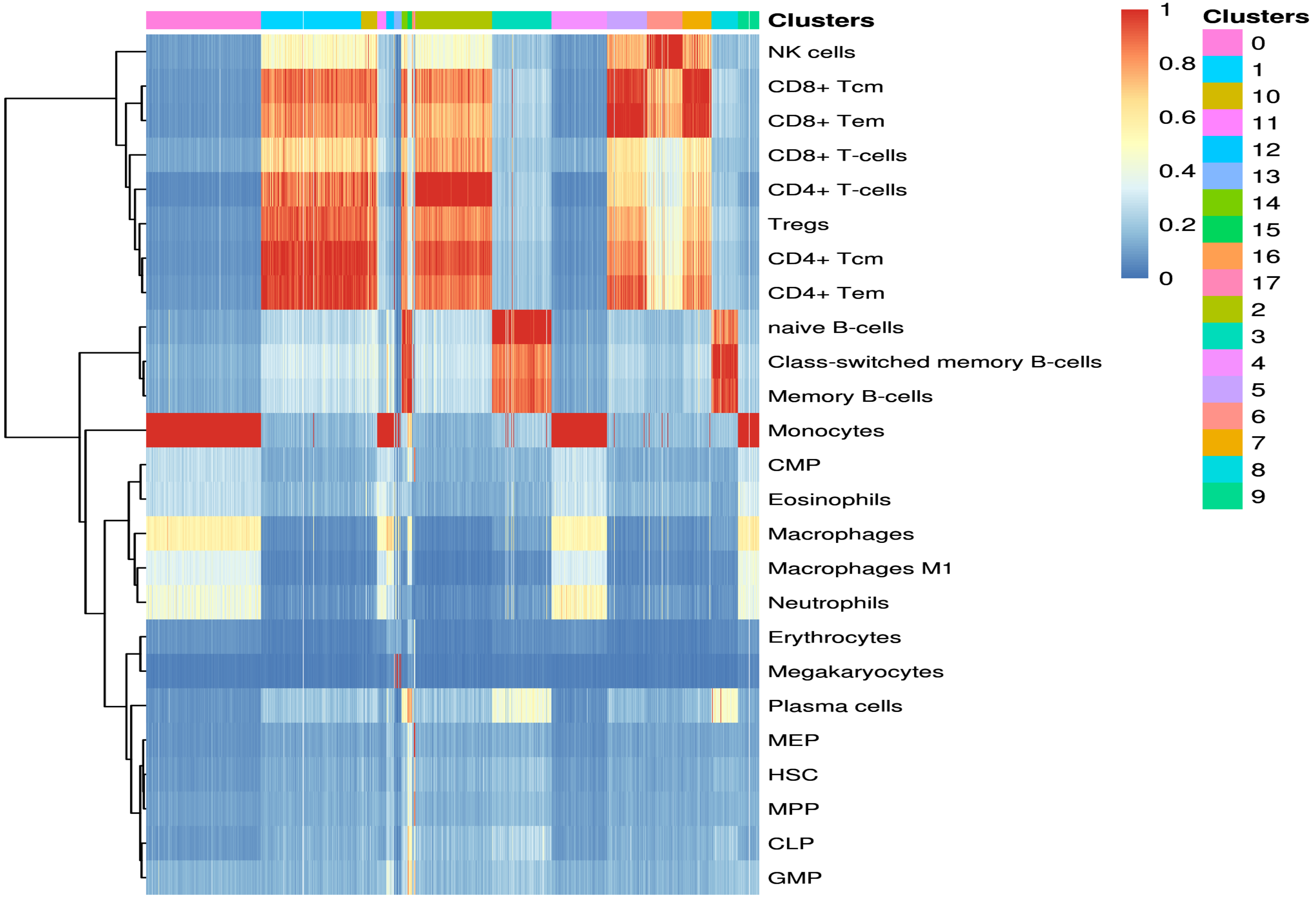
Frederick National Laboratory for Cancer Research



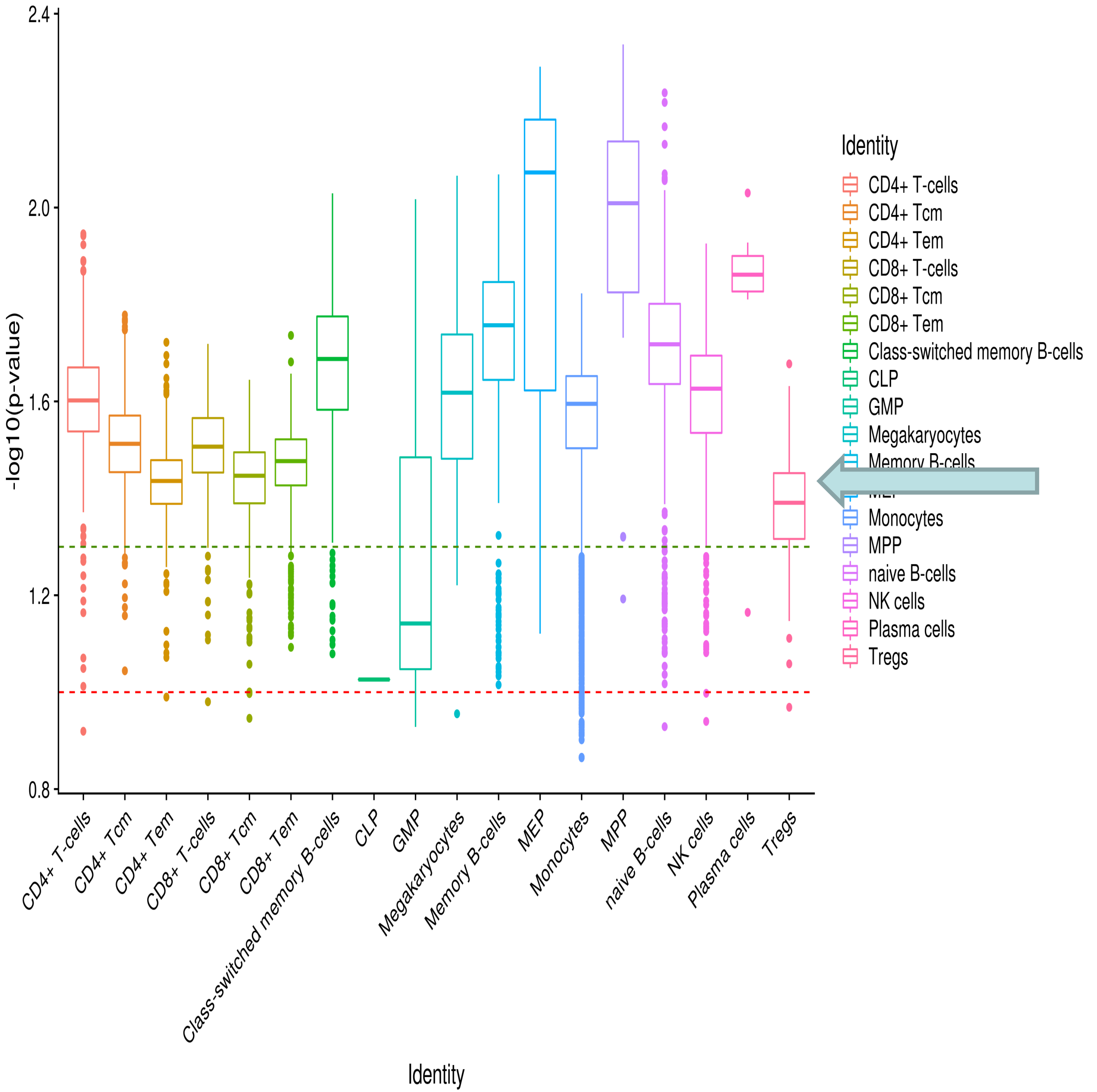


SingleR - PBMC

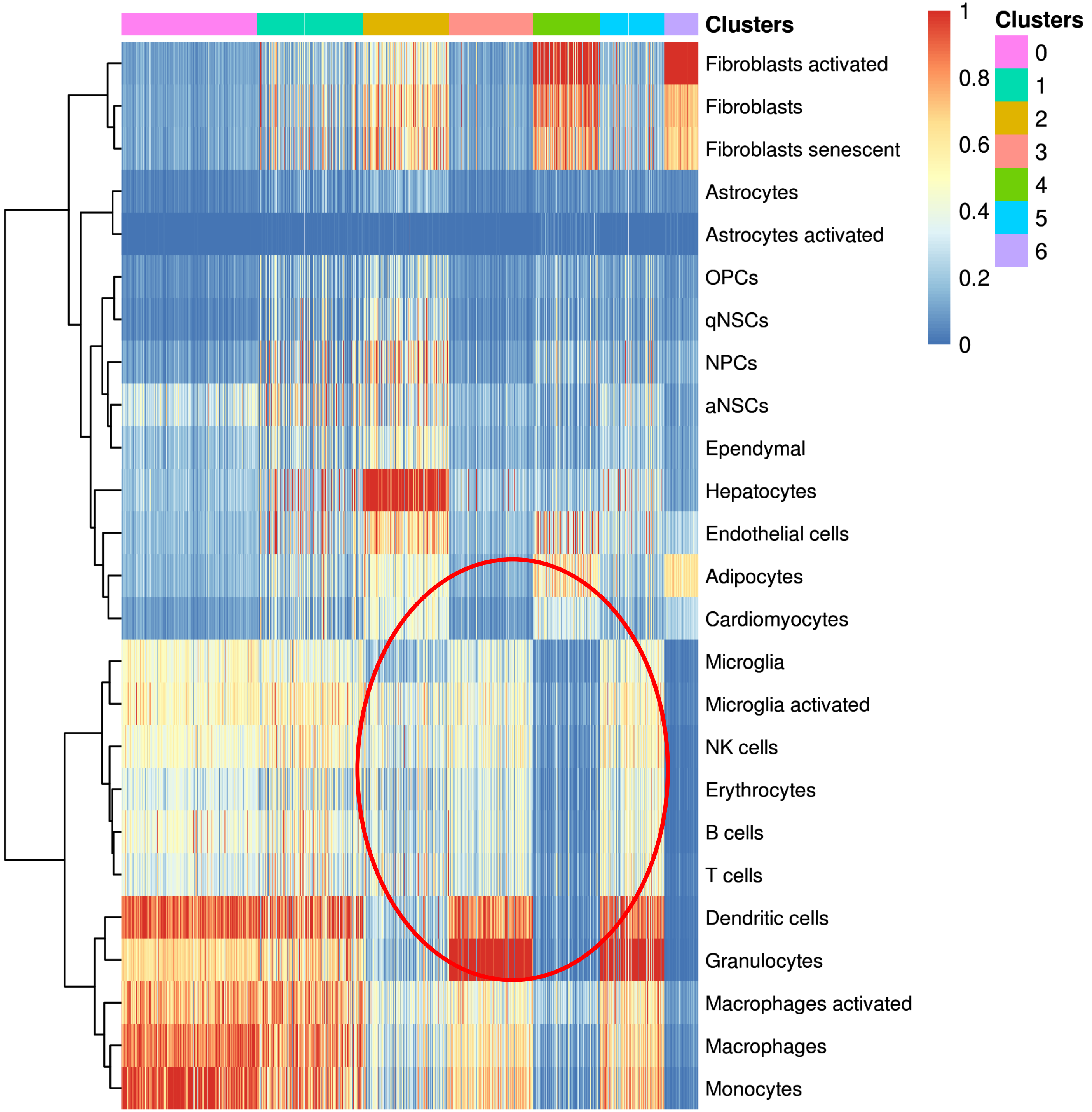
Frederick National Laboratory for Cancer Research



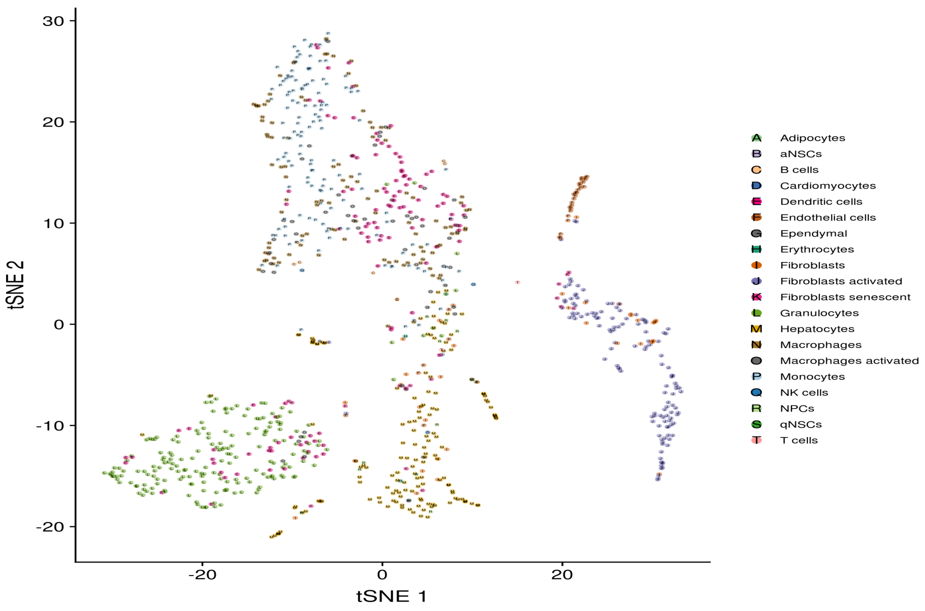
SingleR - Pancreas

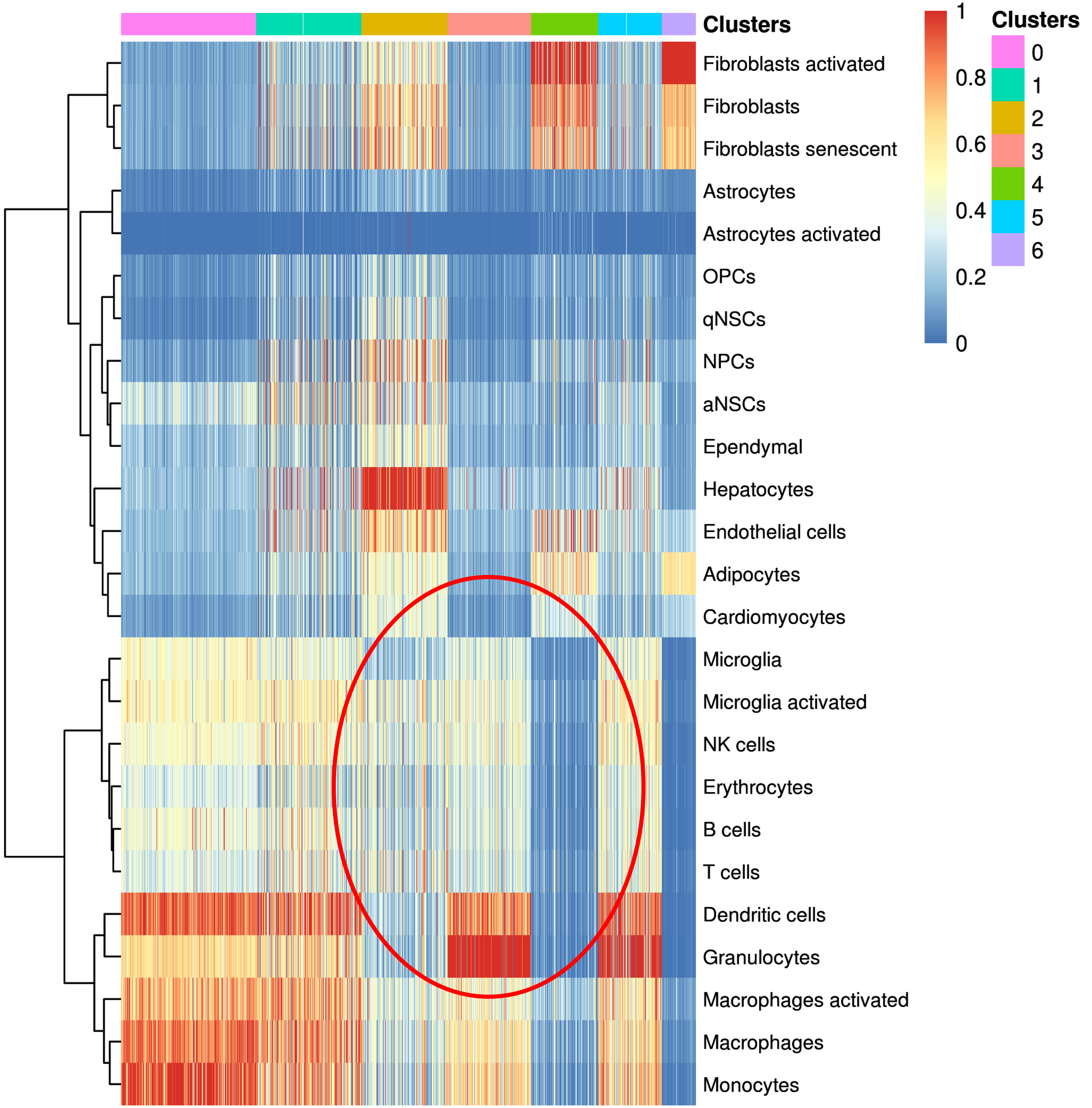


SingleR - Pancreas



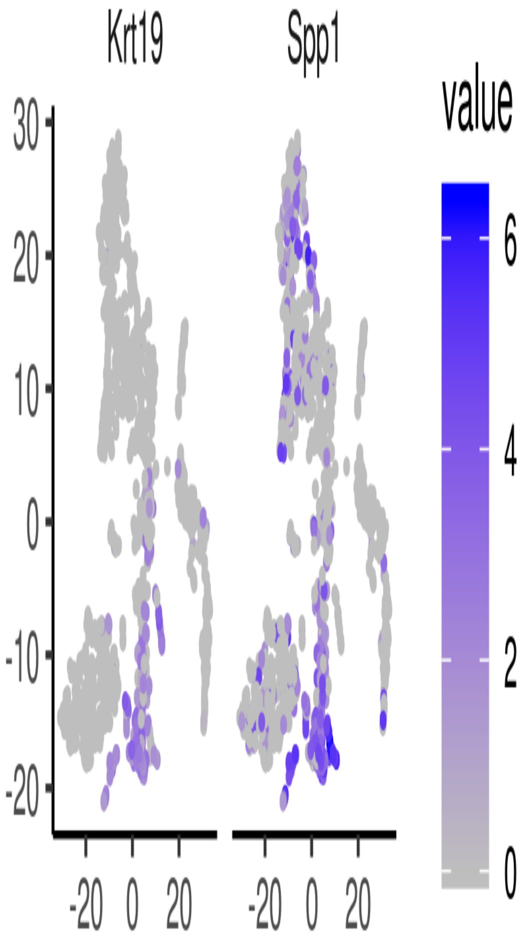
SingleR - Pancreas





DigitalCellSorter

Sergii Domanskyi, Anthony Szedlak, Nathaniel T Hawkins, Jiayin Wang, Giovanni Paternostro & Carlo Piermarocchi, *BMC Bioinformatics* volume 20, Article number: 369 (**2019**)

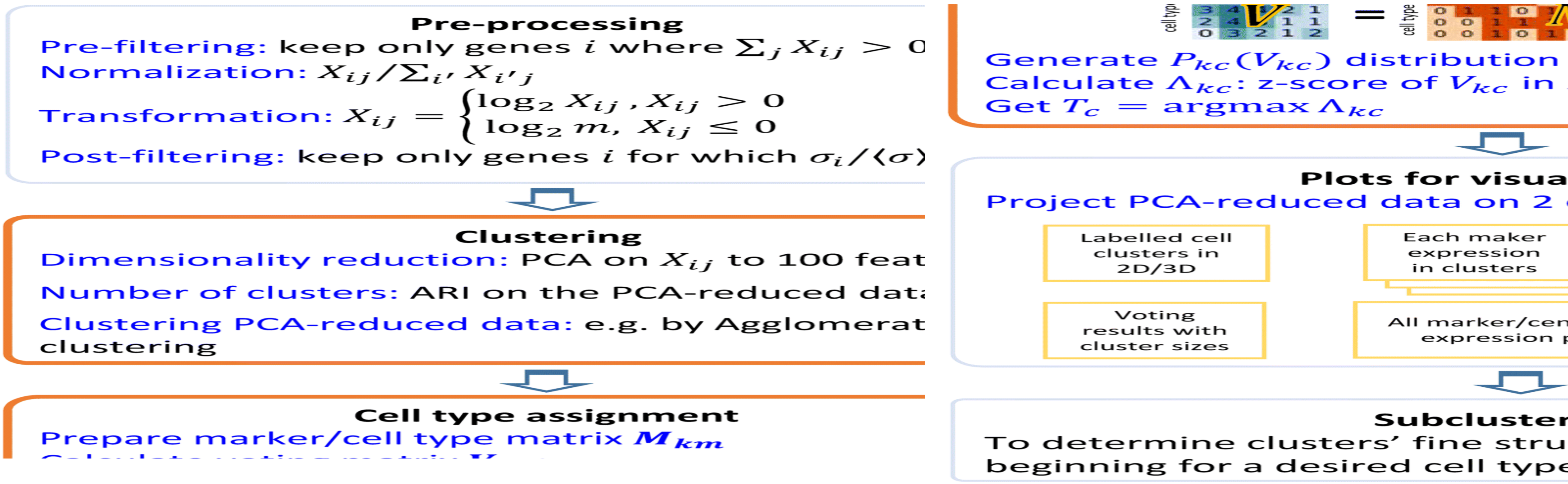


DigitalCellSorter

1. **Pre-preprocessing** of single cell mRNA sequencing data (gene expression data)
    1. Cleaning: filling in missing values, zemoving all-zero genes and cells, converting gene index to a desired convention, etc.
    2. Normalizing: rescaling all cells expression, log-transforming, etc.
2. **Quality control**
3. **Batch effects correction**
4. **Cells anomaly score evaluation**
5. **Dimensionality reduction**
6. **Clustering** (Hierarchical, K-Means, knn-graph-based, etc.)
7. **Annotating cell types**
8. **Vizualization**
    1. t-SNE layout plot
    2. Quality Control histogram plot
    3. Marker expression t-SNE subplot
    4. Marker-centroids expression plot
    5. Voting results matrix plot
    6. Cell types stacked barplot
    7. Anomaly scores plot
    8. Histogram null distribution plot
    9. New markers plot
    10. Sankey diagram (a.k.a. river plot)
9. **Post-processing** functions, e.g. extract cells of interest, find significantly expressed genes, plot marker expression of the cells of interest, etc.

DigitalCellSorter

Frederick National Laboratory for Cancer Research



Pre-processing

Pre-filtering: keep only genes i where $\sum_j X_{ij} > 0$

Normalization: $X_{ij}/\sum_{i'} X_{i'j}$

Transformation: $X_{ij} = \begin{cases} \log_2 X_{ij}, X_{ij} > 0 \\ \log_2 m, \ X_{ij} \leq 0 \end{cases}$

Post-filtering: keep only genes i for which $\sigma_i/\langle\sigma\rangle$

Clustering

Dimensionality reduction: PCA on X_{ij} to 100 feat

Number of clusters: ARI on the PCA-reduced data

Clustering PCA-reduced data: e.g. by Agglomerat clustering

Cell type assignment

Prepare marker/cell type matrix M_{km}

Calculate voting matrix V

Generate $P_{kc}(V_{kc})$ distribution

Calculate Λ_{kc}: z-score of V_{kc} in

Get $T_c = \text{argmax}\,\Lambda_{kc}$

Plots for visua

Project PCA-reduced data on 2

Labelled cell clusters in 2D/3D

Each maker expression in clusters

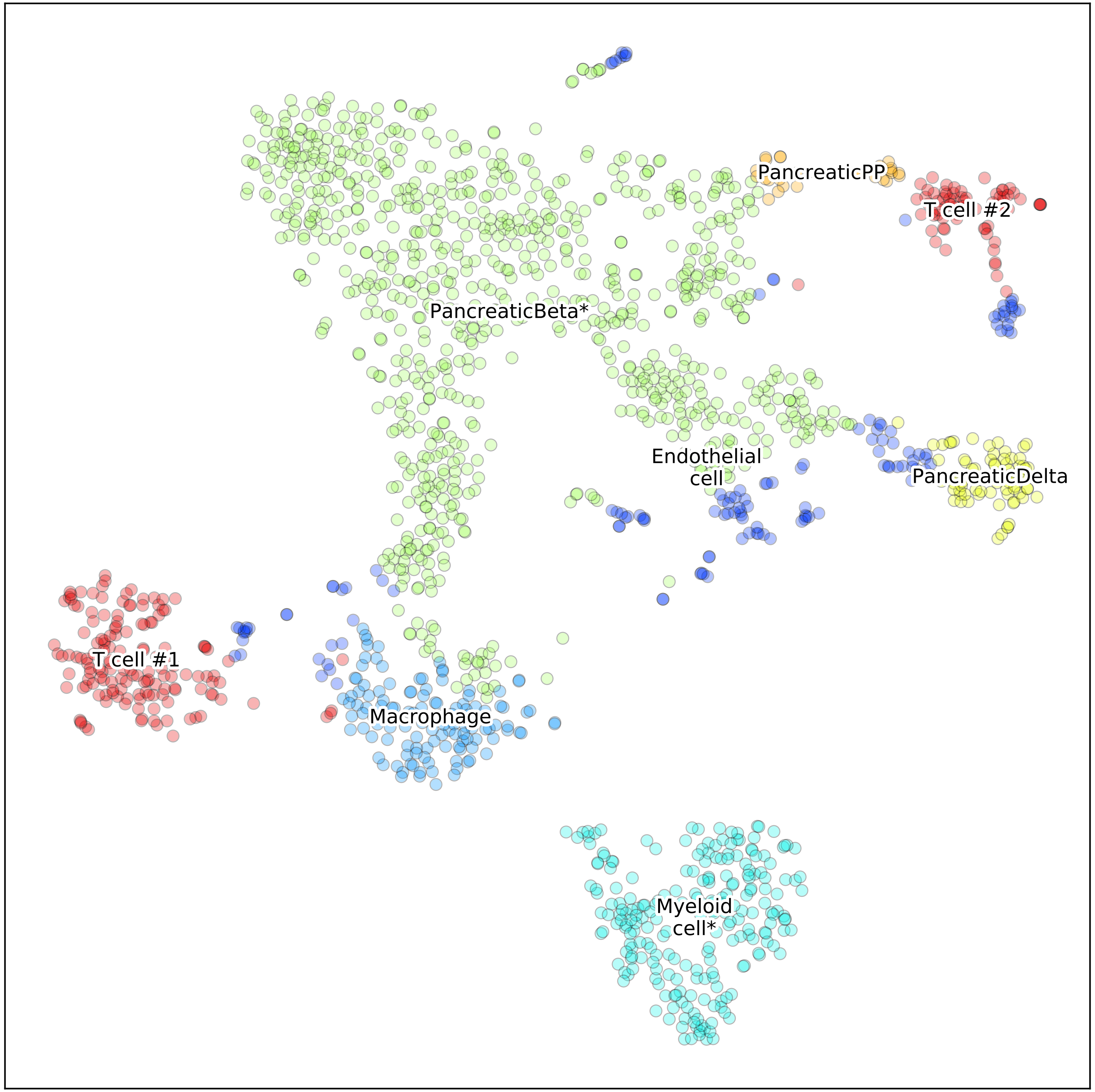
Voting results with cluster sizes

All marker/cen expression

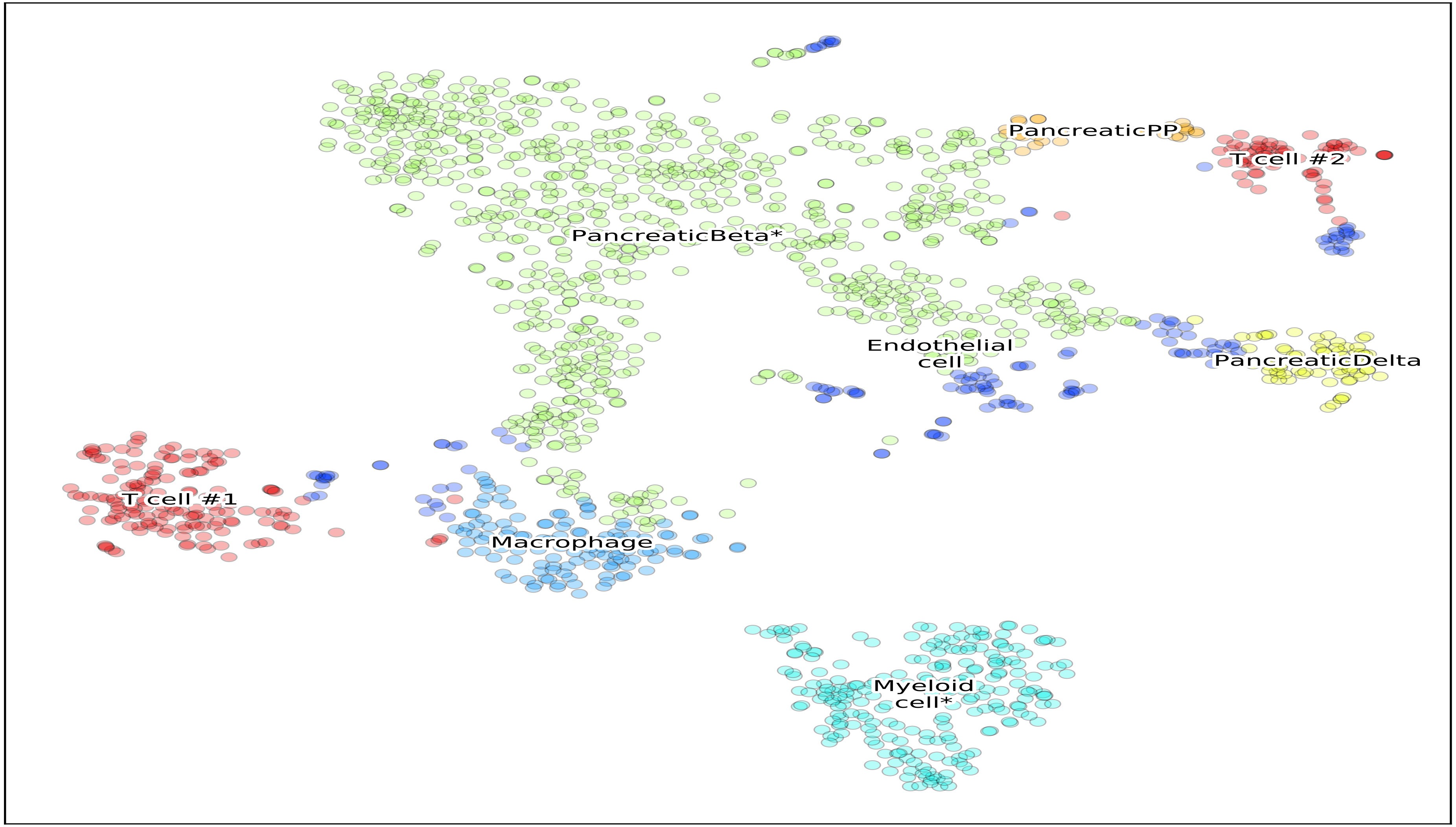
Subcluster

To determine clusters' fine stru beginning for a desired cell type

DigitalCellSorter – Pancreas



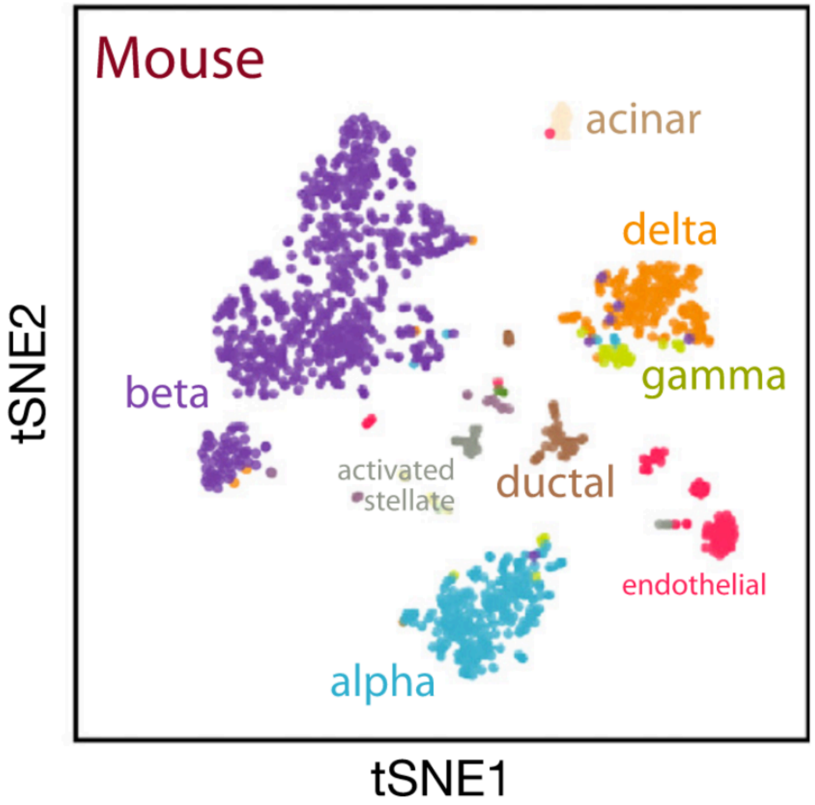
DigitalCellSorter



scRNASeq – r package

Study	***Type***	***Cells***
Lun et al. (2017)	416B cells	192
La Manno et al. (2016)	Human embryonic midbrain	1,977
La Manno et al. (2016)	Human embryonic stem cells	1,715
Messmer et al. (2019)	Human embryonic stem cells	1,344
La Manno et al. (2016)	Human embryonic midbrain	1,907
La Manno et al. (2016)	Human induced pluripotent stem cells	337
Baron et al. (2016)	Human pancreas	8,569
Grun et al. (2016)	Human pancreas	1,728
Lawlor et al. (2017)	Human pancreas	638
Muraro et al. (2016)	Human pancreas	3,072
Segerstolpe et al. (2016)	Human pancreas	3,514
Xin et al. (2016)	Human pancreas	1,600
La Manno et al. (2016)	Mouse adult dopaminergic neurons	243
Campbell et al. (2017)	Mouse brain	21,086
Chen et al. (2017)	Mouse brain	14,437
Marques et al. (2016)	Mouse brain	5,069
Romanov et al. (2017)	Mouse brain	2,881
Usoskin et al. (2015)	Mouse brain	864
Tasic et al. (2016)	Mouse brain	1,809
Zeisel et al. (2015)	Mouse brain	3,005
Richard et al. (2018)	Mouse CD8+ T cells	572
Grun et al. (2016)	Mouse haematopoietic stem cells	1,915
Nestorowa et al. (2016)	Mouse haematopoietic stem cells	1,920
Bach et al. (2017)	Mouse mammary gland	25,806
Kolodziejczyk et al. (2015)	Mouse mebryonic stem cells	704
Baron et al. (2016)	Mouse pancreas	1,886
Macosko et al. (2015)	Mouse retina	49,300
Shekhar et al. (2016)	Mouse retina	44,994
Lun et al. (2017)	Mouse trophoblasts	192
Aztekin et al. (2019)	Xenopus tail	13,199

scRNASeq – r package



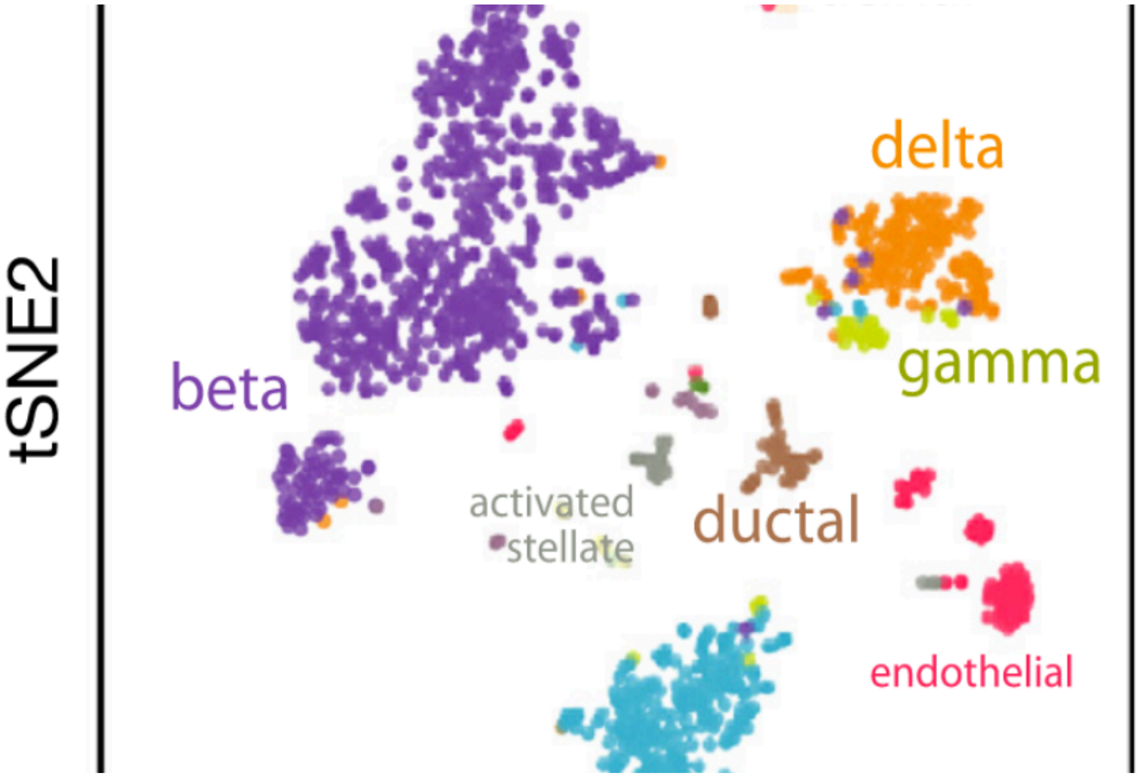
Baron et al. (2016)

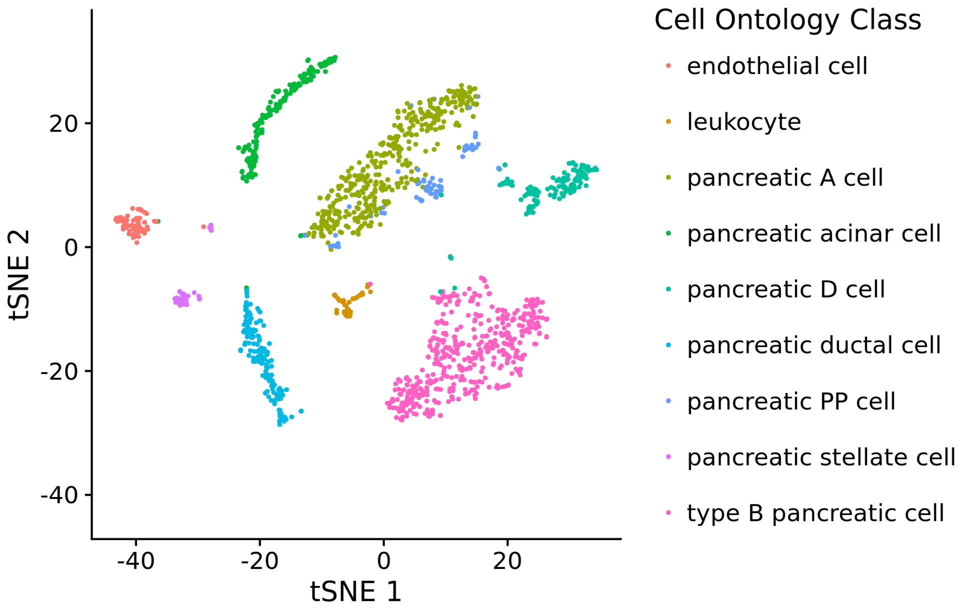
Study	*Type*	*Cells*
Lun et al. (2017)	416B cells	192
La Manno et al. (2016)	Human embryonic midbrain	1,977
La Manno et al. (2016)	Human embryonic stem cells	1,715
Messmer et al. (2019)	Human embryonic stem cells	1,344
La Manno et al. (2016)	Human embryonic midbrain	1,907
La Manno et al. (2016)	Human induced pluripotent stem cells	337
Baron et al. (2016) - BM	Human pancreas	8,569
Grun et al. (2016)	Human pancreas	1,728
Lawlor et al. (2017)	Human pancreas	638
Muraro et al. (2016)	Human pancreas	3,072
Segerstolpe et al. (2016)	Human pancreas	3,514
Xin et al. (2016)	Human pancreas	1,600
La Manno et al. (2016)	Mouse adult dopaminergic neurons	243
Campbell et al. (2017)	Mouse brain	21,086
Chen et al. (2017)	Mouse brain	14,437
Marques et al. (2016)	Mouse brain	5,069
Romanov et al. (2017)	Mouse brain	2,881
Usoskin et al. (2015)	Mouse brain	864
Tasic et al. (2016)	Mouse brain	1,809
Zeisel et al. (2015)	Mouse brain	3,005
Richard et al. (2018)	Mouse CD8+ T cells	572
Grun et al. (2016)	Mouse haematopoietic stem cells	1,915
Nestorowa et al. (2016)	Mouse haematopoietic stem cells	1,920
Bach et al. (2017)	Mouse mammary gland	25,806
Kolodziejczyk et al. (2015)	Mouse mebryonic stem cells	704
Baron et al. (2016)	Mouse pancreas	1,886
Macosko et al. (2015)	Mouse retina	49,300
Shekhar et al. (2016)	Mouse retina	44,994
Lun et al. (2017)	Mouse trophoblasts	192
Aztekin et al. (2019)	Xenopus tail	13,199

Tabula Muris Consortium (TM)

The ***Chan Zuckerberg Biohub*** recently released ***Tabula Muris***, a compendium of single cell transcriptome data from the mouse containing nearly ***100,000 cells from 20 organs and tissues***. The data allow for direct and controlled comparison of gene expression in cell types shared between tissues, such as immune cells from distinct anatomical locations. They also allow for a comparison of two distinct technical approaches:

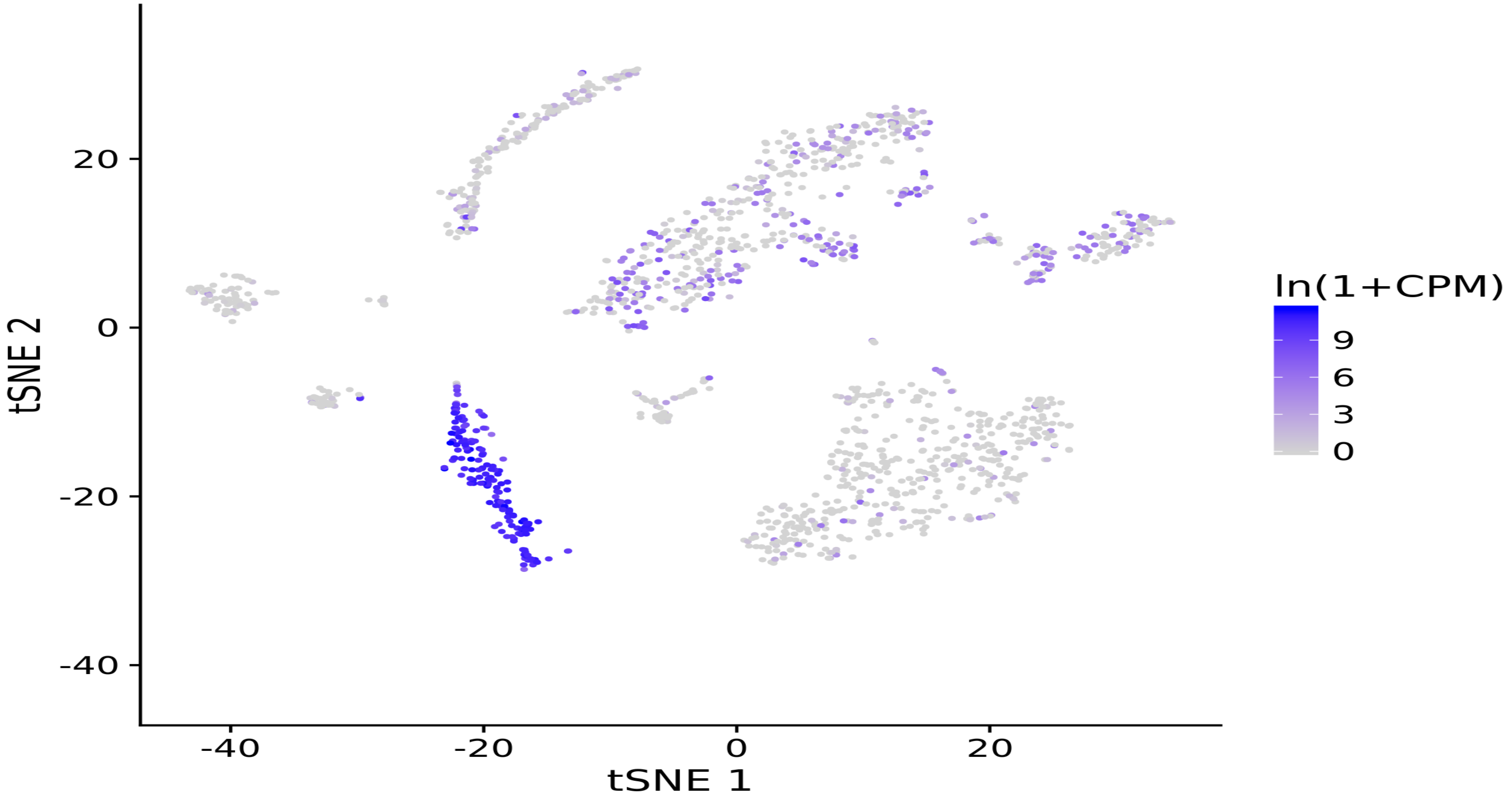
•microfluidic droplet-based 3’-end counting, which provides a survey of thousands of cells per organ at relatively low coverage.
•FACS-based full length transcript analysis, which provides higher sensitivity and coverage.



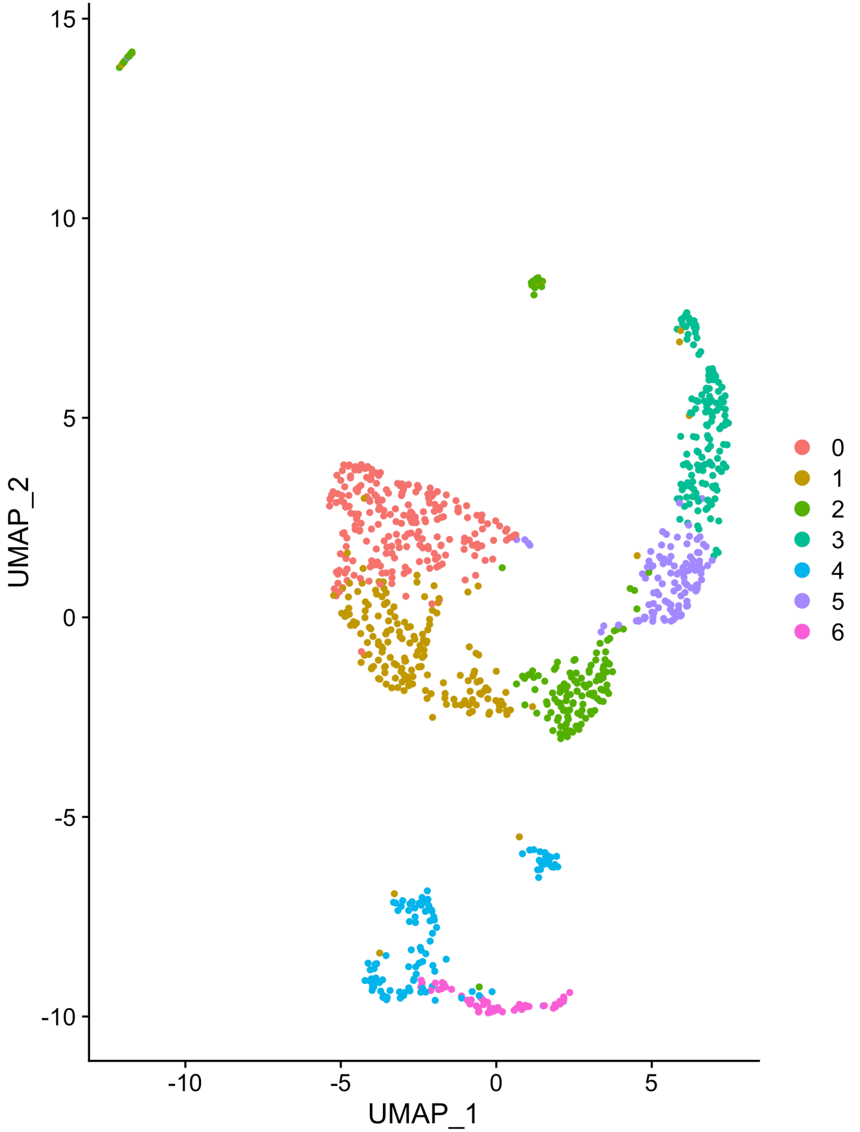


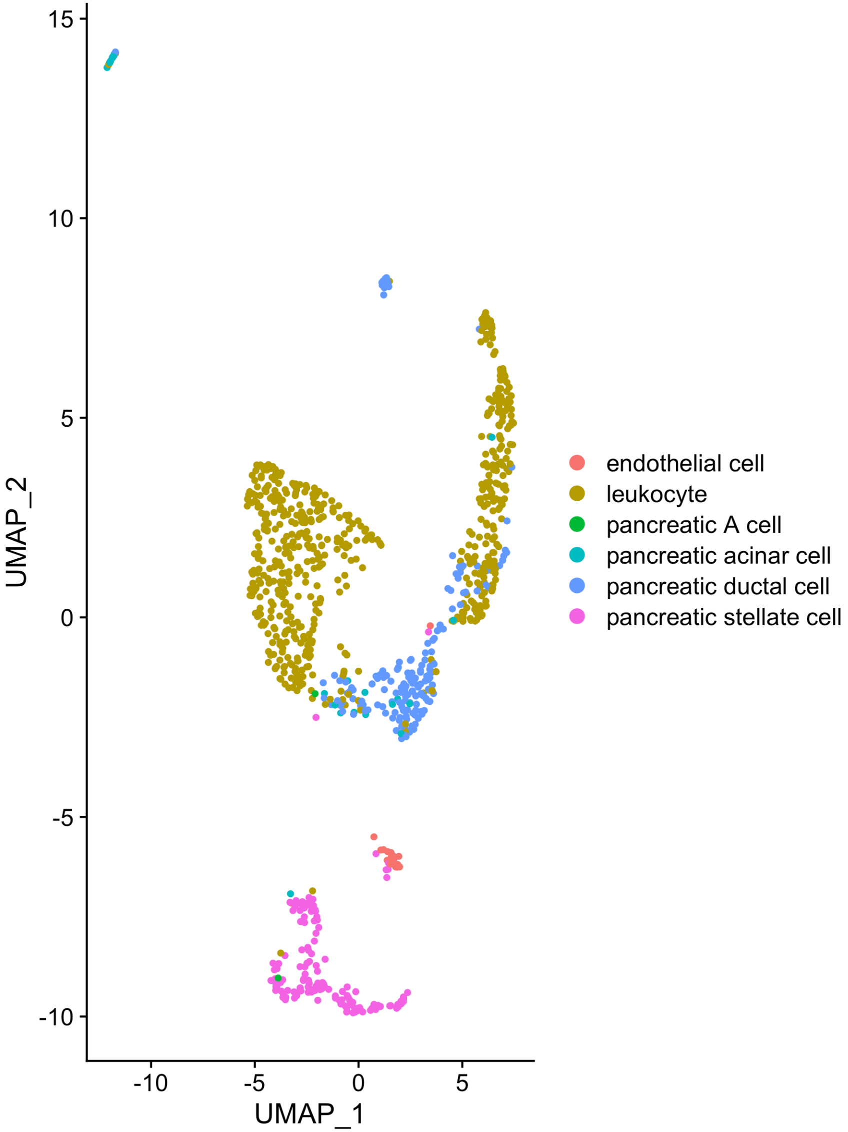
SingleR – Label Transfer - Tabula Muris (TM)

Frederick National Laboratory for Cancer Research

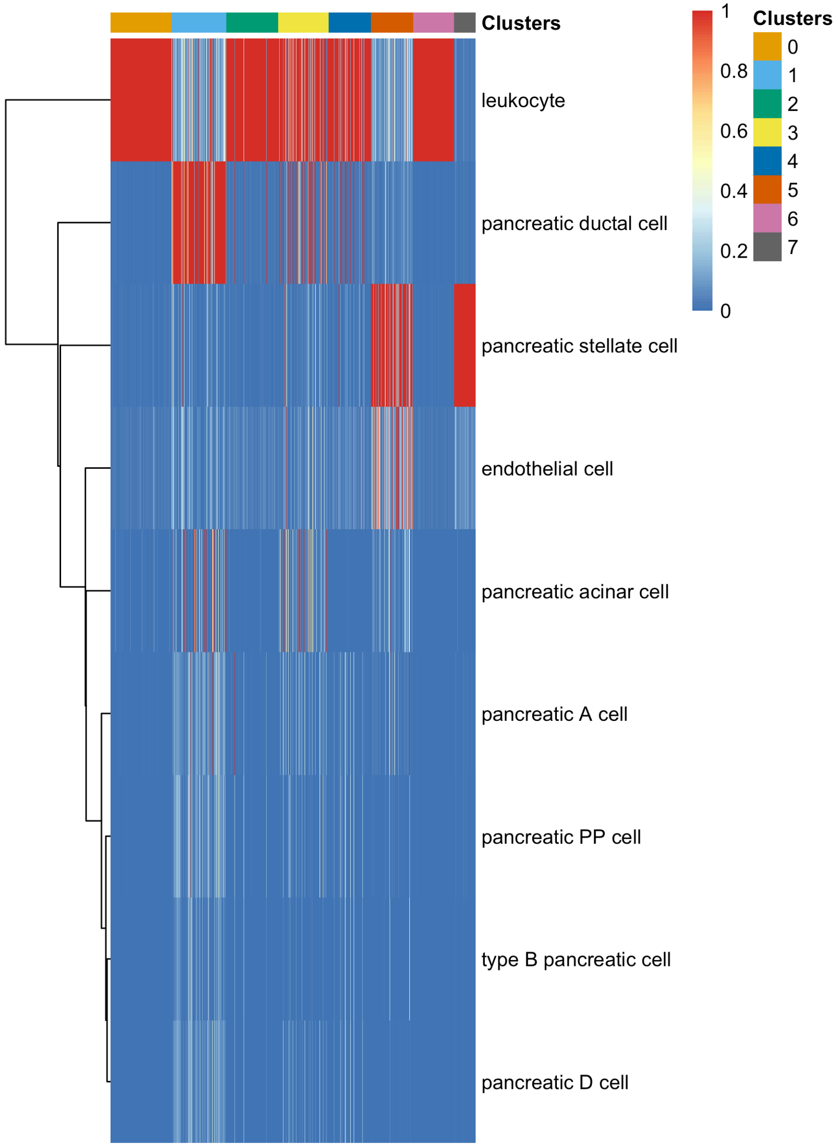


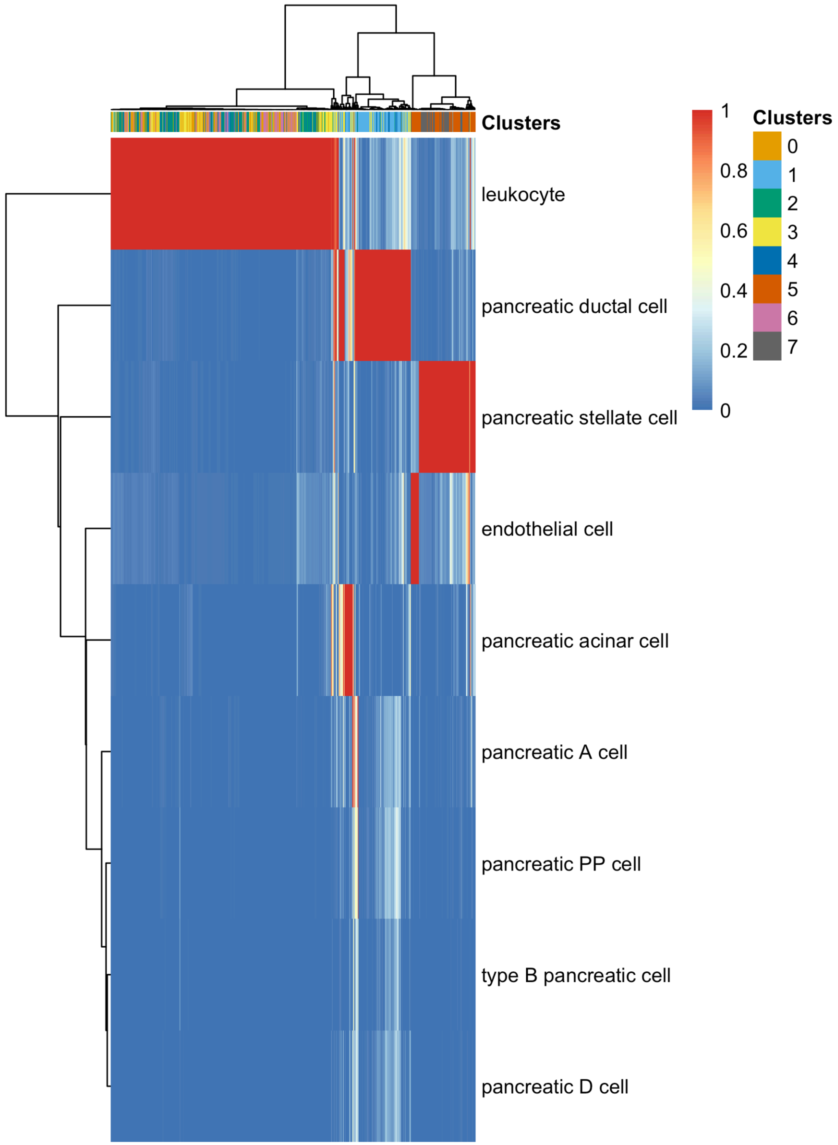
SingleR – Label Transfer - Tabula Muris (TM)



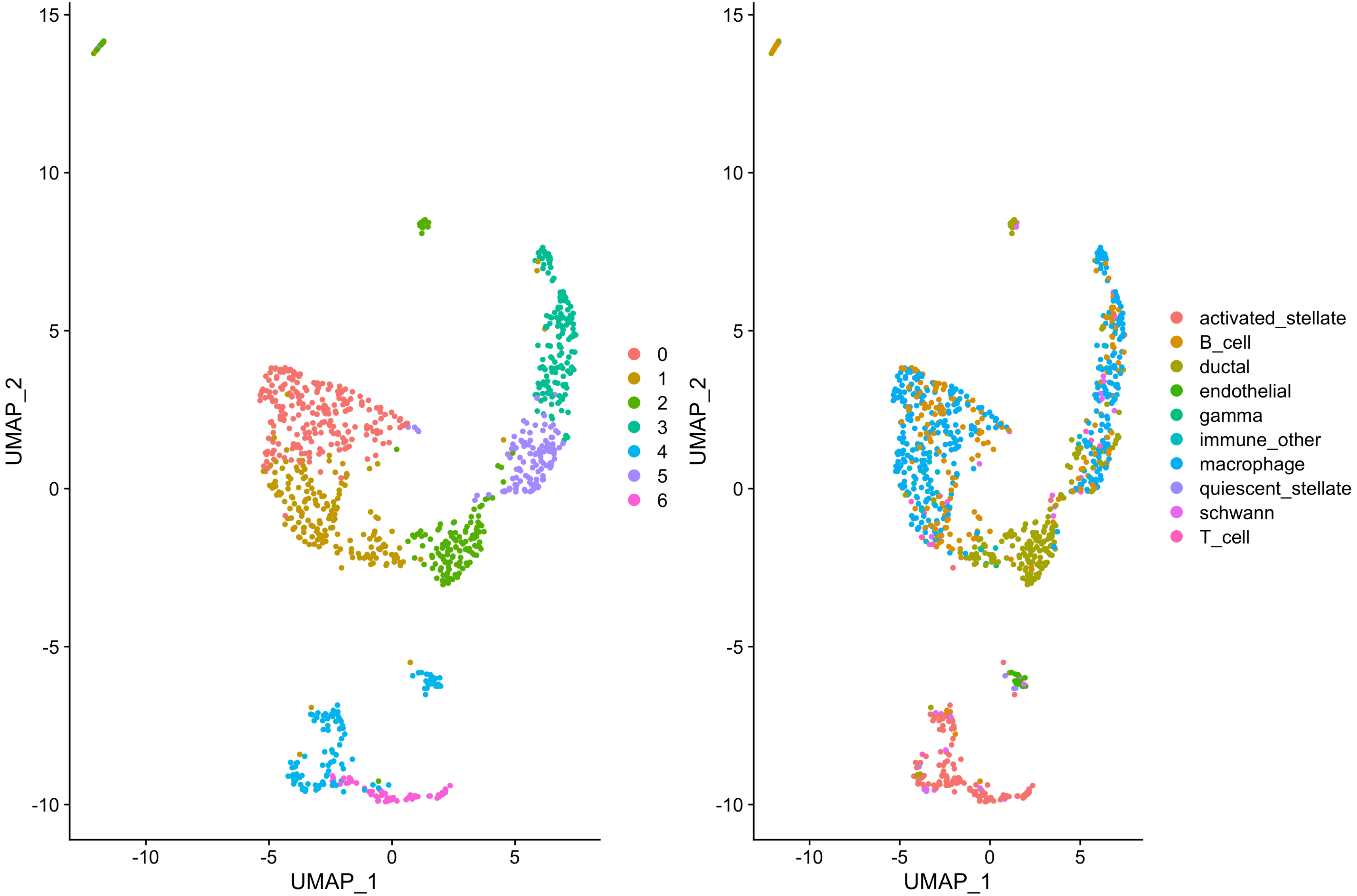


SingleR – Label Transfer – Barron et al (BM)

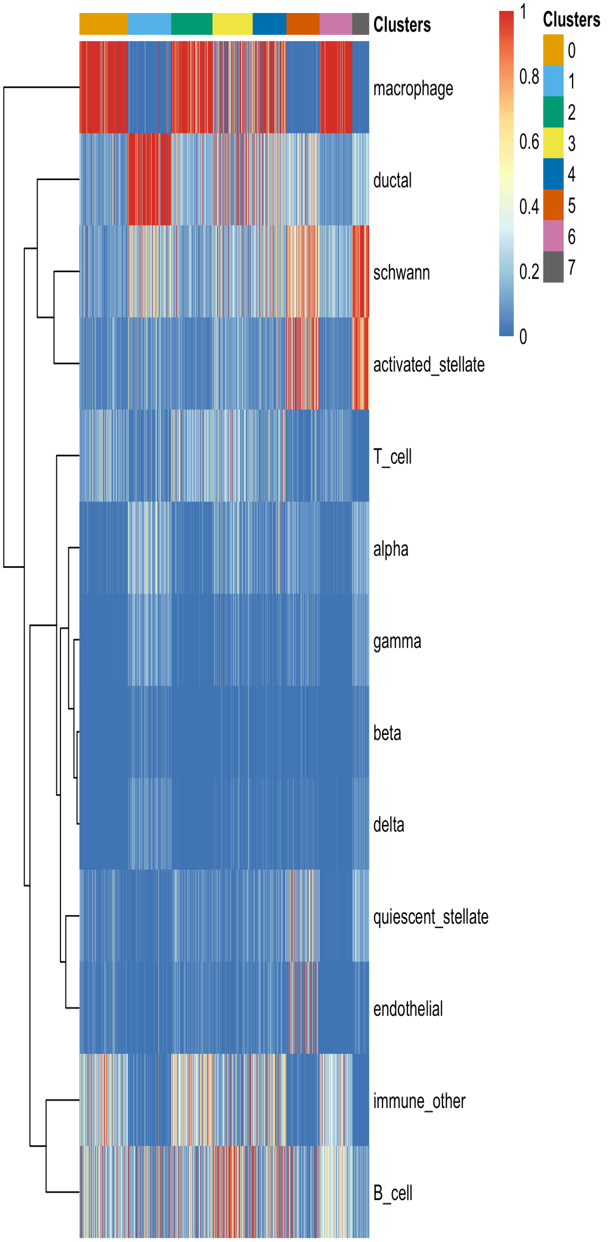


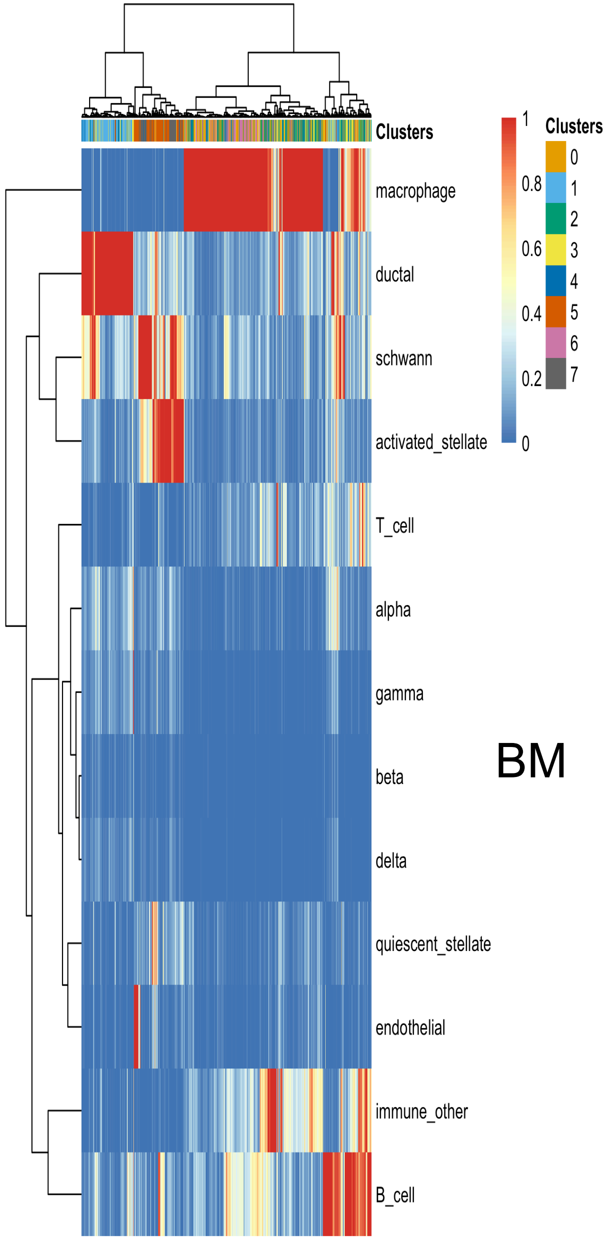


SingleR – Label Transfer – Barron et al (BM)

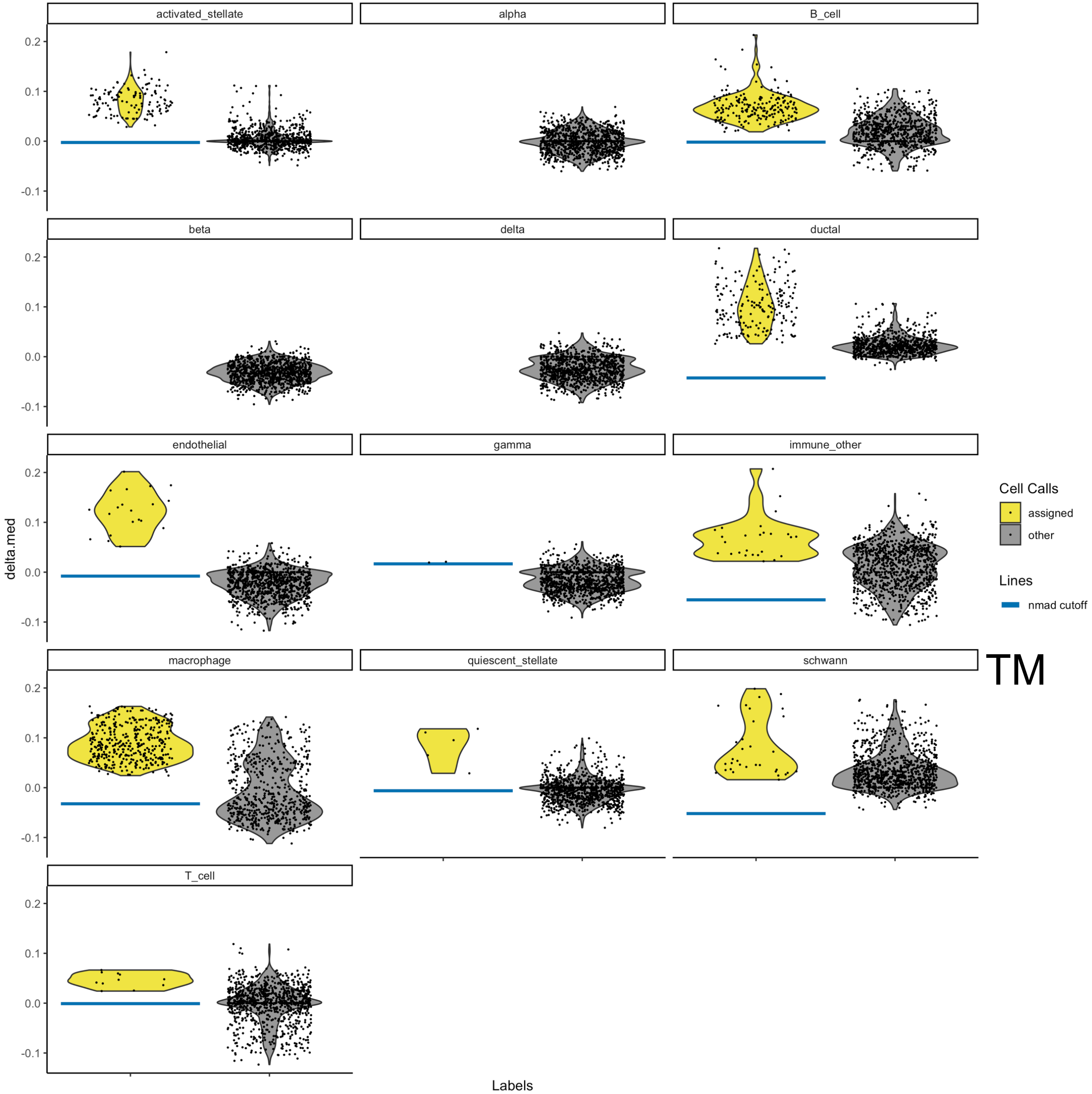


SingleR – Label Transfer – BM vs TM



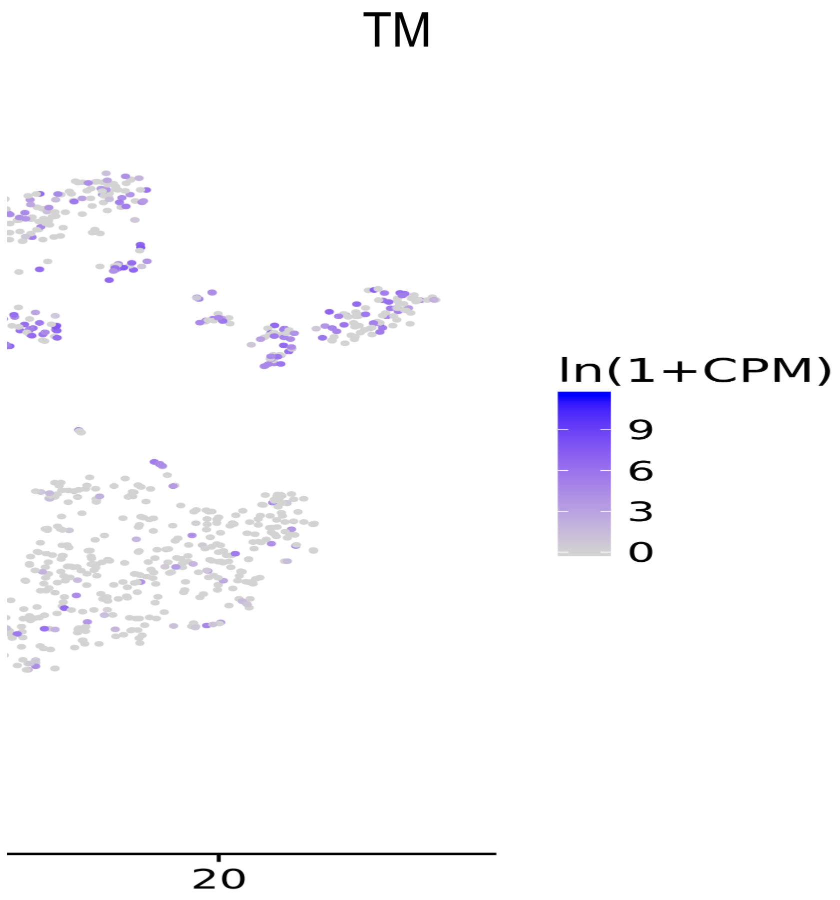


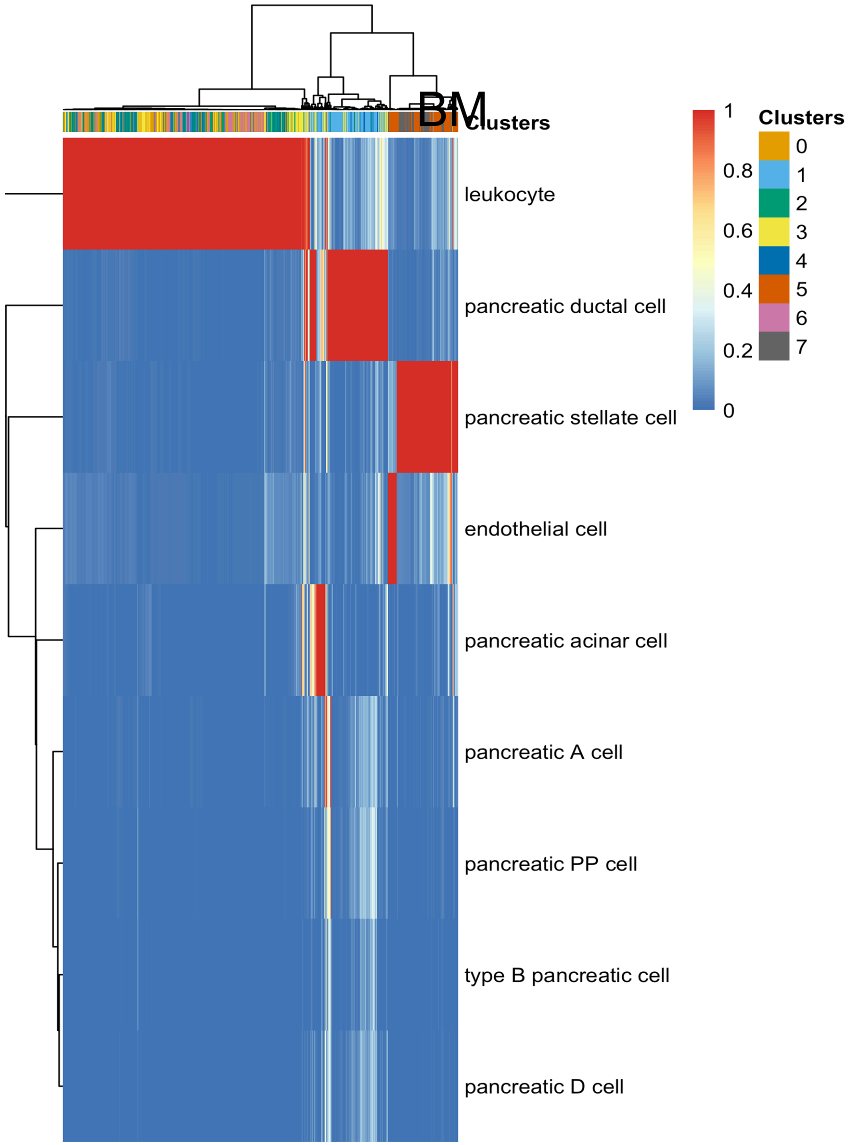
BM



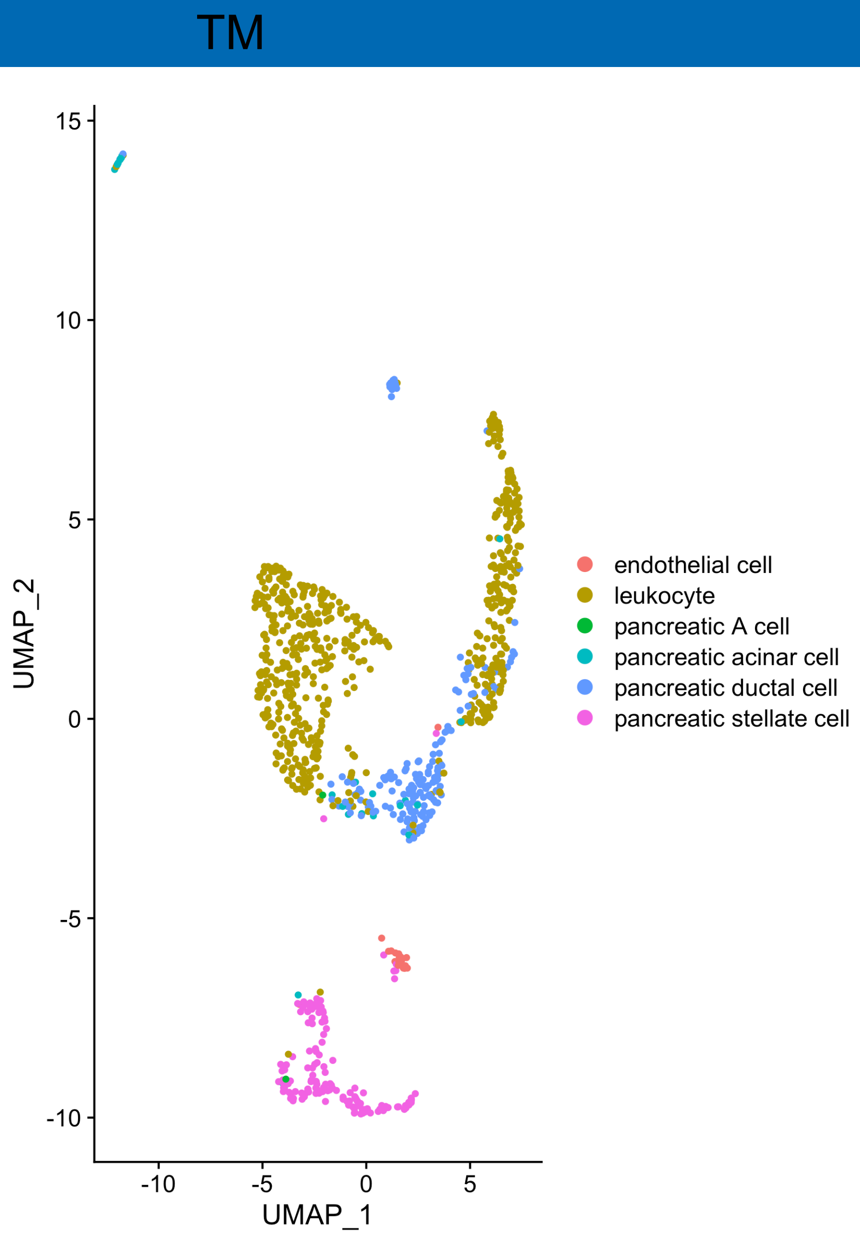
TM

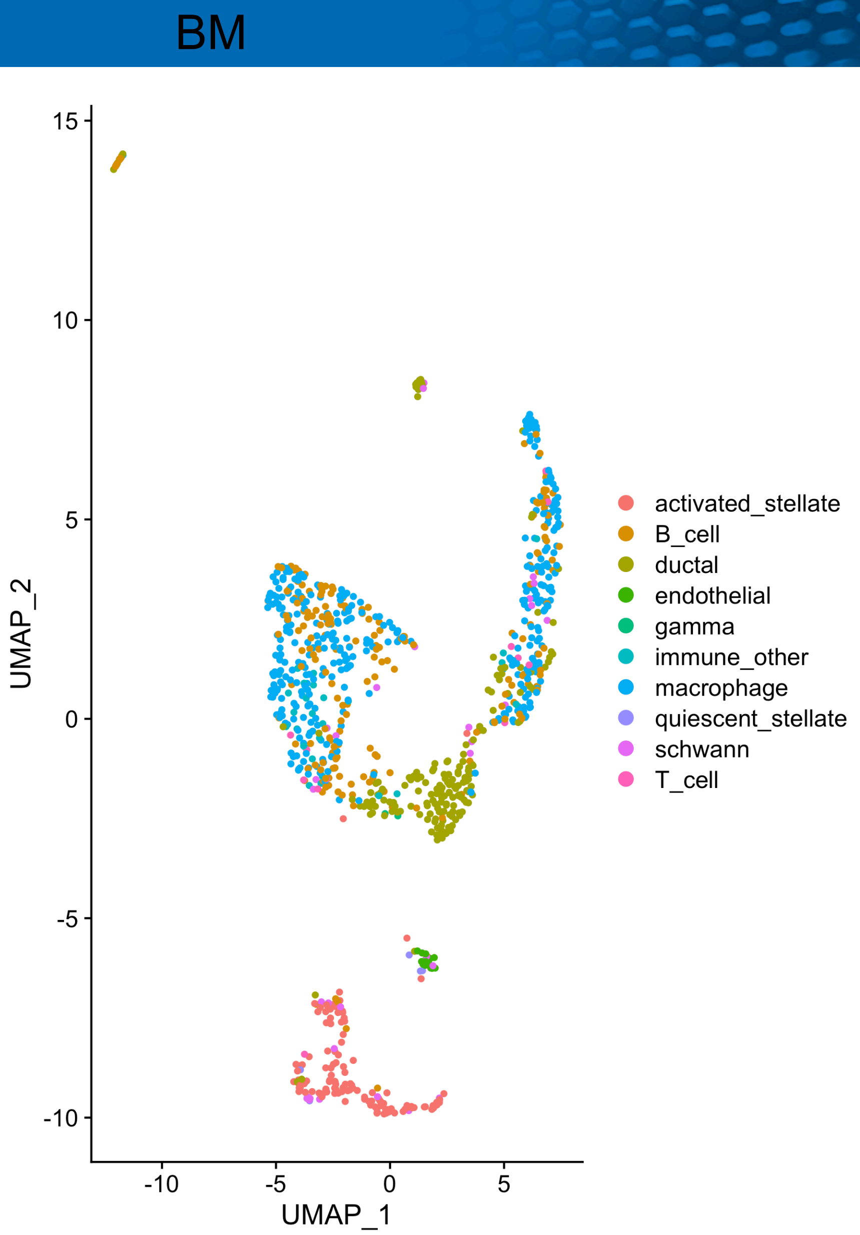
SingleR – Label Transfer – BM vs TM





SingleR – Label Transfer – BM vs TM





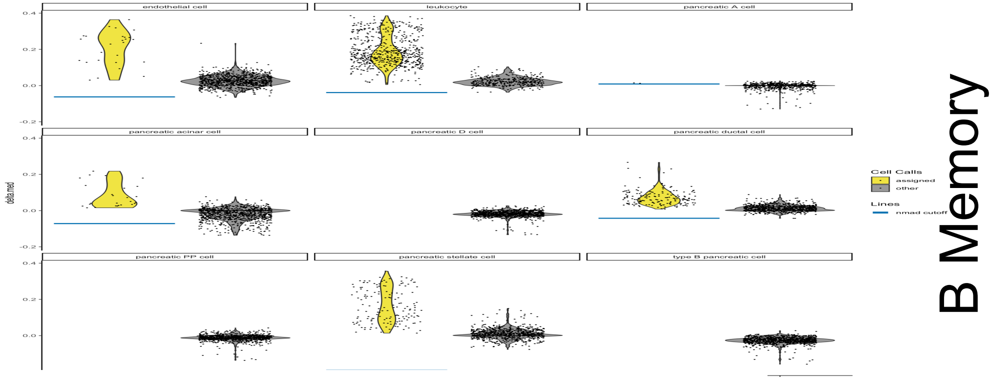
SF scRNA Pipeline Automated Panels

Frederick National Laboratory for Cancer Research

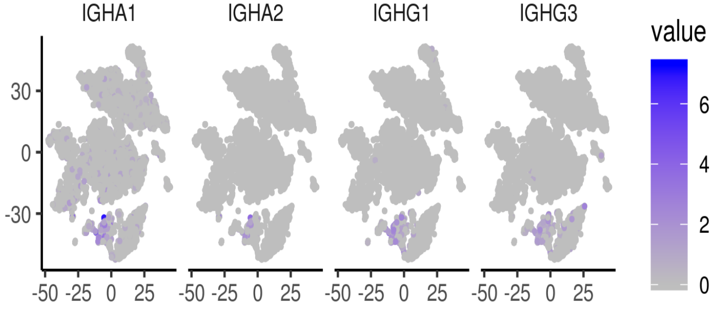
- B Pan
- B Memory
- B Naive
- Macrophage
- NK
- T Central Memory
- T Cytotoxic
- T Effector Memory
- NKT
- T Pan
- T Regulatory
- T Tfh
- T Th1
- T Th1
- T Th17
- T Th2
- T gd
- Monocyte Ly-6C hi
- Monocyte Ly-6C lo
- Cell Cycle M
- Cell Cycle S
- Pancreas
- Pancreatic Alpha
- Pancreatic Beta
- Pancreatic Delta
- Pancreatic PP
- Pancreatic Duct

SF scRNA Pipeline Automated Panels

Frederick National Laboratory for Cancer Research

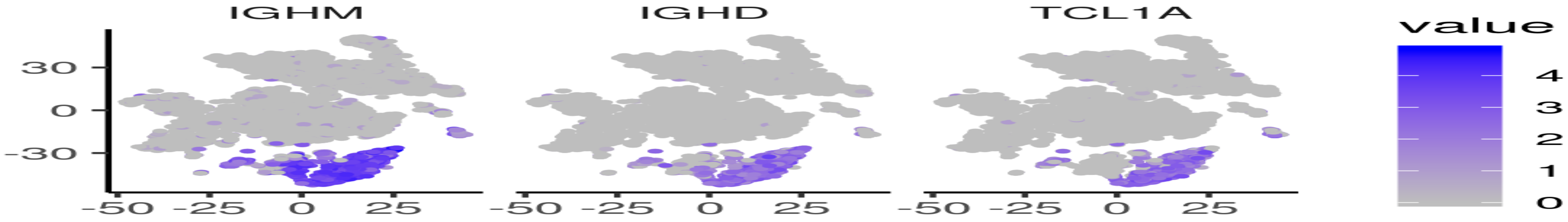


B Memory

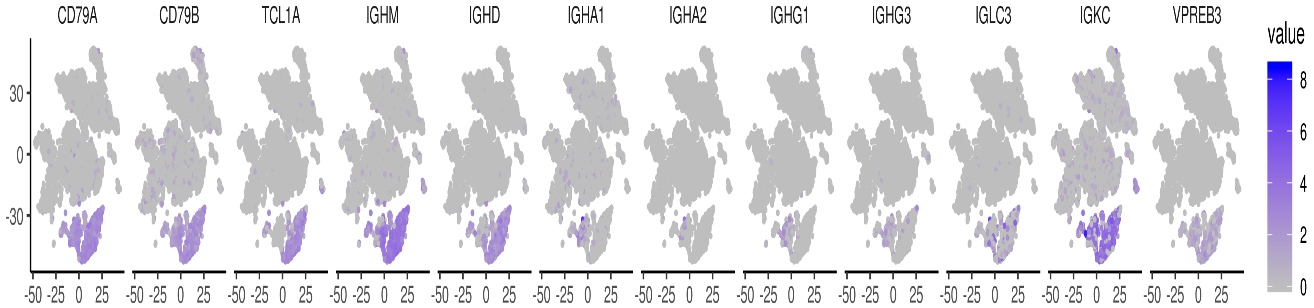


B Naive

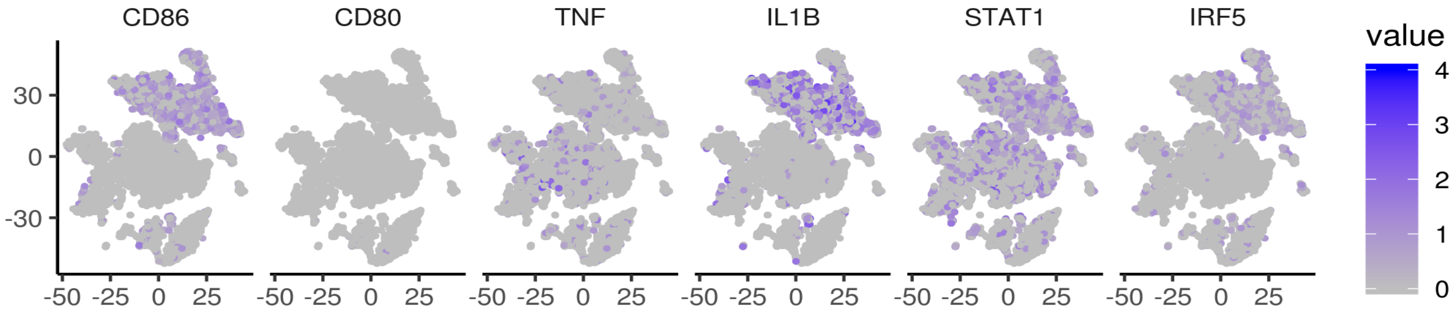
B Pan



SF scRNA Pipeline Automated Panels



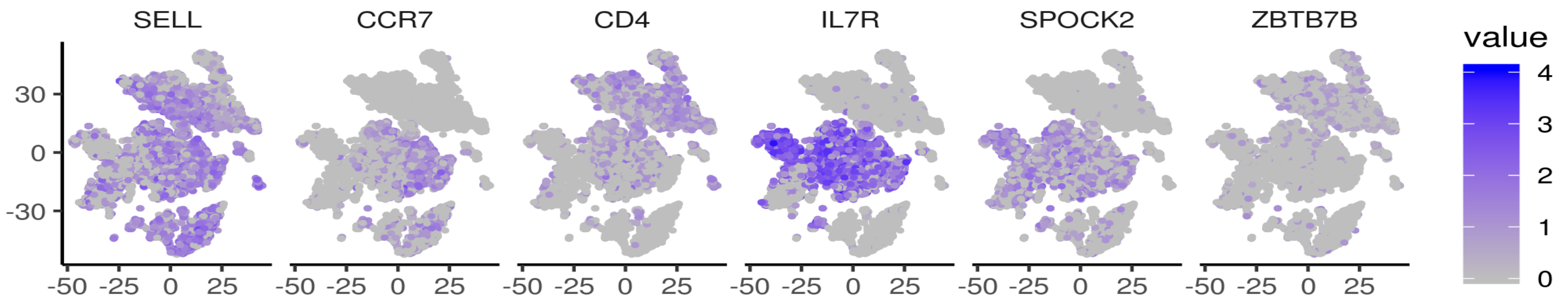
Macrophage – M1



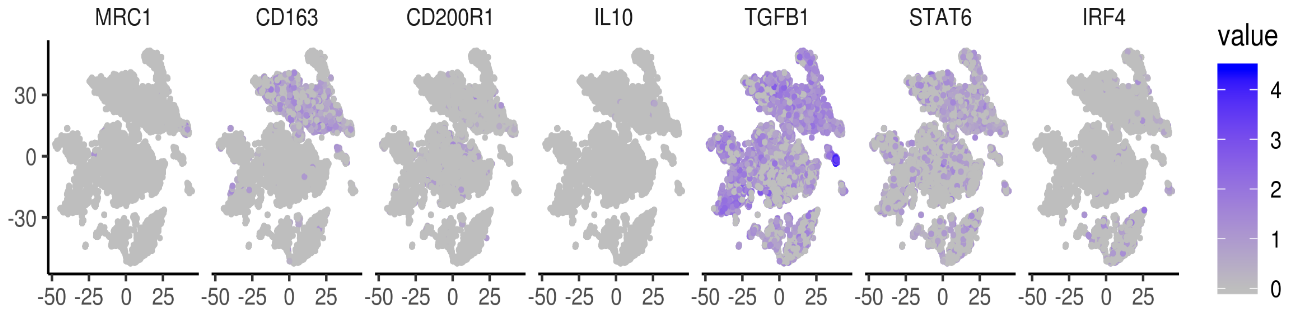
Macrophage – M2

SF scRNA Pipeline Automated Panels

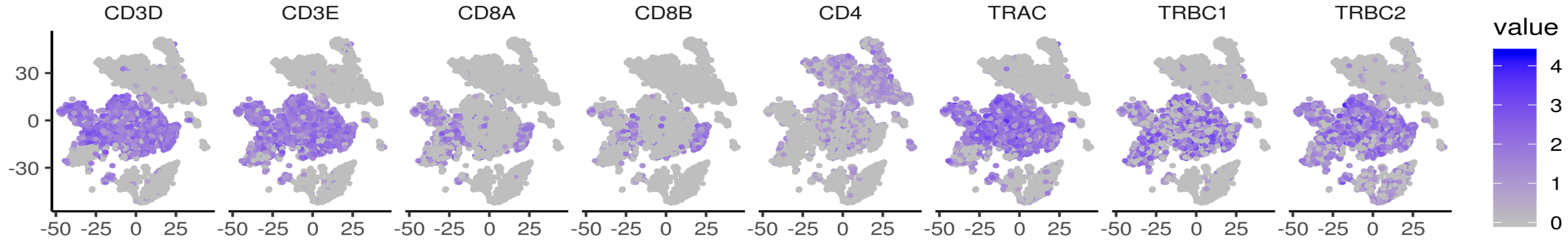
Frederick National Laboratory for Cancer Research



T - Pan



T - Naive



NKT

Question?