ZMNH Center for Molecular Neurobiology Hamburg

Universitätsklinikum Hamburg-Eppendorf

Uncertainty Estimation for Single-cell Label Transfer

Robin Khatri and Stefan Bonn

COPA, August 25 2022

Single-cell Transcriptomics

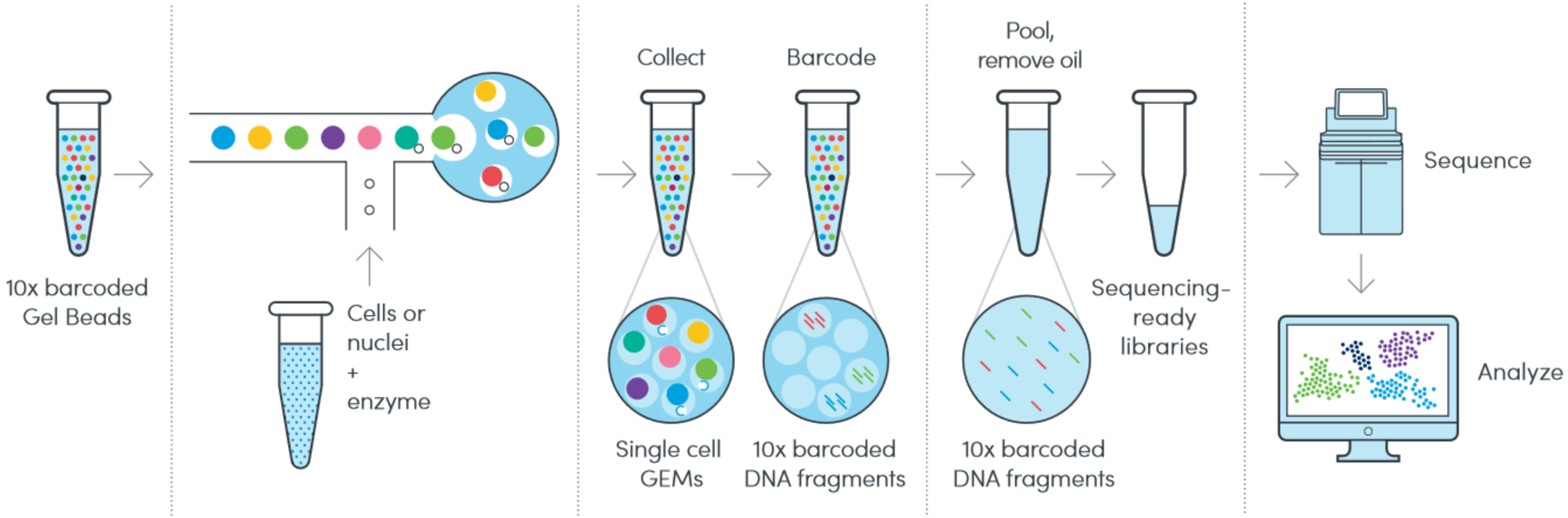
A brief introduction

- Single-cell RNA sequencing (scRNA-seq) measures mRNA expression at a single-cell level

Single-cell Transcriptomics

A brief introduction

- Single-cell RNA sequencing (scRNA-seq) measures mRNA expression at a single-cell level

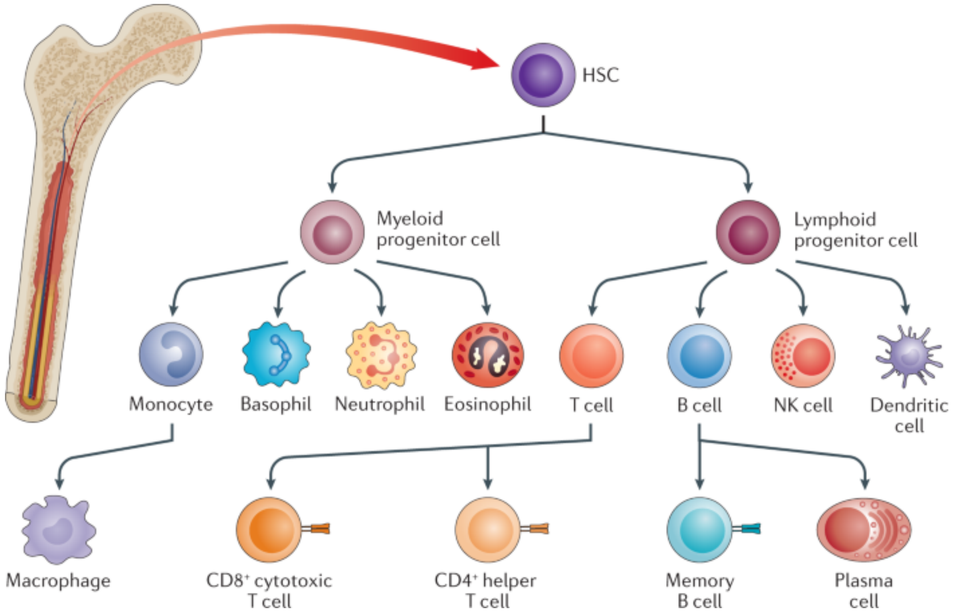


Source: 10X Genomics

Single-cell Transcriptomics

A brief introduction

- Cells can be grouped into groups known as cell-types that share similar features and lineages
- > 2300 cell-types (Osumi-Sutherland et. al., 2021, *Nature Cell Biology*)



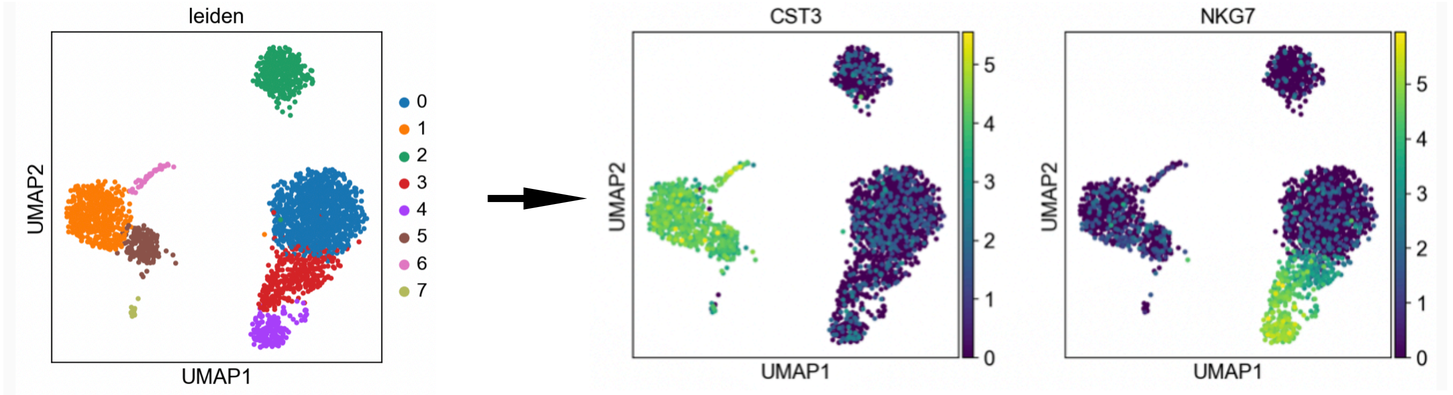
Haematopoiesis in Bone Marrow
Source: Kim et. al., 2019, *Nature Reviews Genetics*

Cell type identification

Workflow

Classical approach

Clustering (PCA, UMAP, tSNE) → Expression of marker genes → Cell type labels

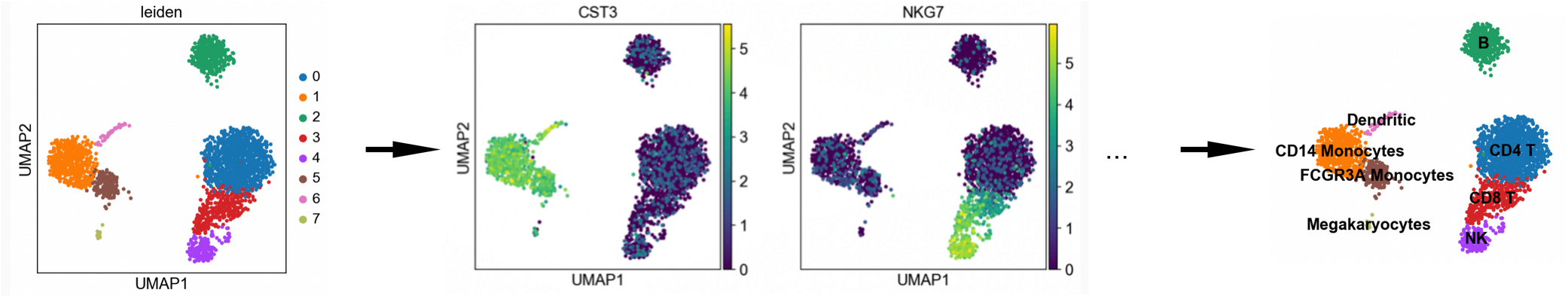


Cell type identification

Workflow

Classical approach

Clustering (PCA, UMAP, tSNE) → Expression of marker genes → Cell type labels



- Dependence on clustering algorithm, Requires very good (and few) marker genes per cell type, not suitable for identifying rare cell types.

+ Detailed control, expert-based

Cell type identification

Label transfer

Idea: Use already available and annotated datasets

+ Larger and larger reference datasets are being generated *e.g.* Human Cell Atlas (https://www.humancellatlas.org), Allen Brain Atlas (https://portal.brain-map.org) etc.

+ Machine Learning algorithms are getting better

- No (human) expert

- Many sources of error - Batch effects, data integration (batch correction), data quality issues, partial references, dynamic nature of cell states

Cell type identification

Label transfer

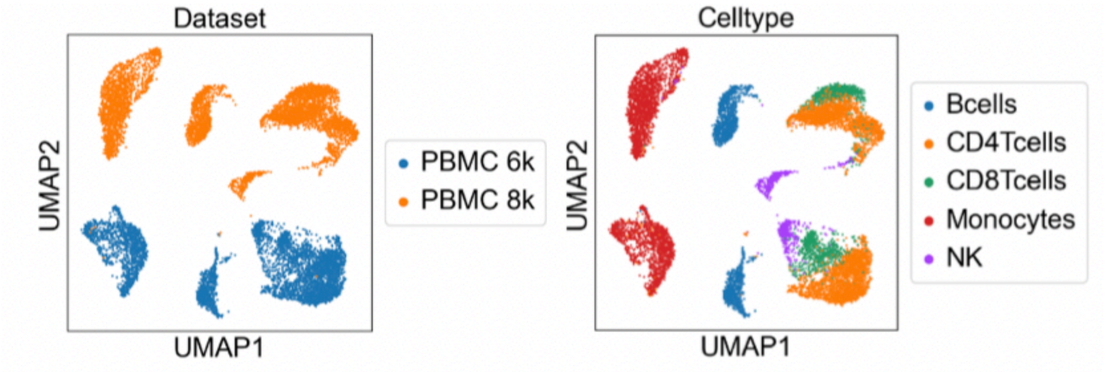
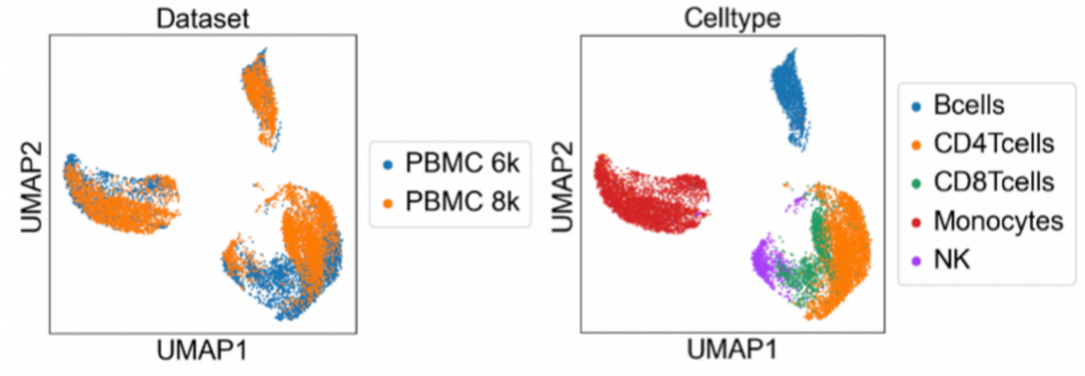


Illustration of Batch Effect on two peripheral blood mononuclear cell (PBMC) datasets.



Post-batch correction using Harmony (Karunsky et. al., 2019, *Nature Methods*)

Conformal prediction for label-transfer

Integration of conformal prediction in the workflows

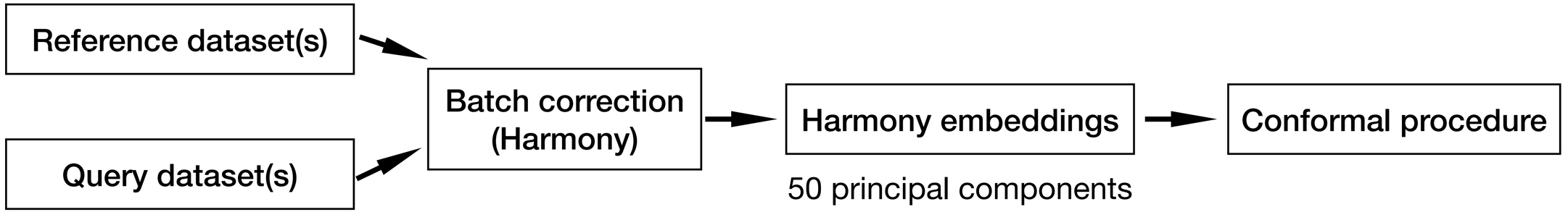
Why use conformal prediction?

- Model agnostic
- Interpretable
- Simple and general

Goals: Explore conformal prediction, investigate how it could help in single-cell label transfer and how it could identify previously unseen cell types

Conformal prediction for label-transfer

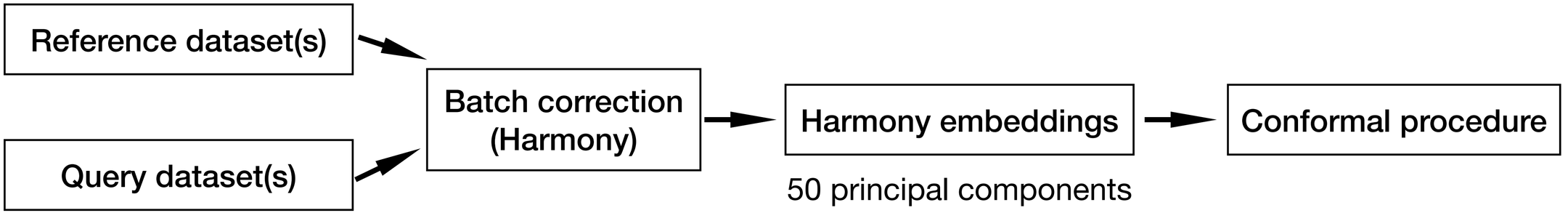
Integration of conformal prediction in the workflows



Conformal prediction for label-transfer

Integration of conformal prediction in the workflows

Inductive conformal classifiers (ICPs)
Base classifier: SVM

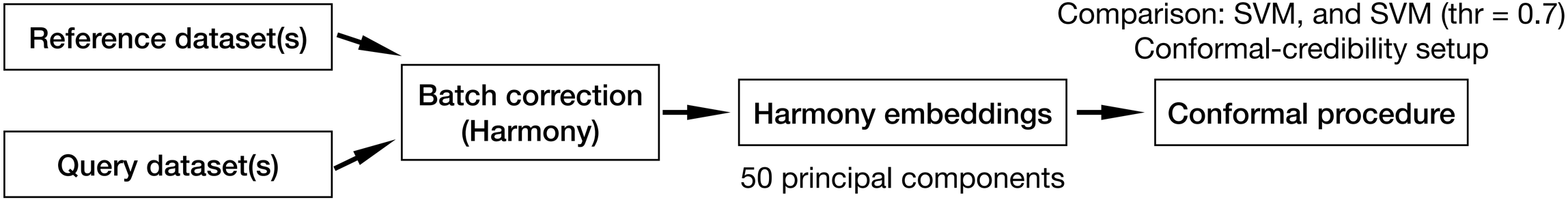


Conformal prediction for label-transfer

Integration of conformal prediction in the workflows

Inductive conformal classifiers (ICPs)
Base classifier: SVM

Comparison: SVM, and SVM (thr = 0.7)
Conformal-credibility setup

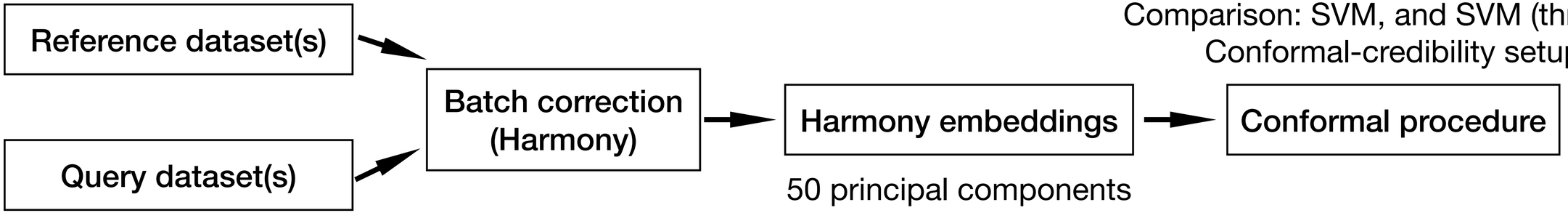


Conformal prediction for label-transfer

Integration of conformal prediction in the workflows

Inductive conformal classifiers (ICPs)
Base classifier: SVM

Comparison: SVM, and SVM (thr = 0.7)
Conformal-credibility setup



PBMCs

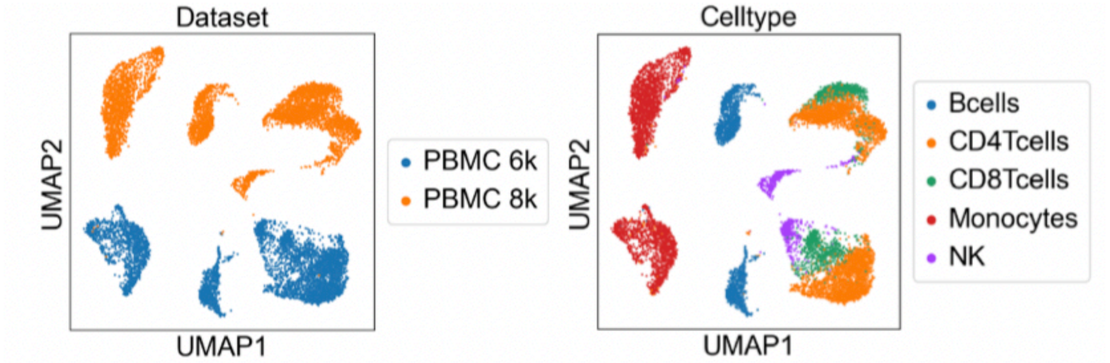
Dataset	B	CD4T	CD8T	Monocytes	NK	Total
PBMC 6k	704	2,240	714	1,397	2,240	5,356
PBMC 8k	992	1,975	890	1,870	320	6,047

Pancreas

Dataset	Alpha	Beta	Gamma	Delta	Acinar	Ductal	Endo.	Stellate	Total
inDrop	2,249	3,048	260	613	272	898	689	362	8,391
SS2	1,109	796	219	142	103	462	67	63	2,961
CEL-Seq 2	885	600	125	199	170	315	31	101	2,426
Fluidigm C1	241	300	12	21	6	27	12	13	632
CEL-Seq	220	341	21	66	162	402	37	22	1,271

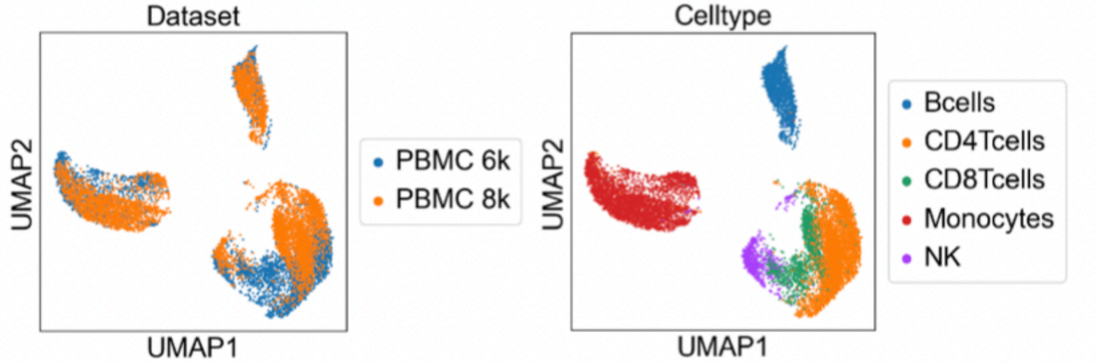
Results

PBMCs



	Bcells	CD4Tcells	CD8Tcells	Monocytes	NK
Bcells	0.88	0.11	0	0.01	0
CD4Tcells	0	0.99	0	0.01	0
CD8Tcells	0	0.99	0	0.01	0
Monocytes	0	0.04	0	0.96	0
NK	0	0.52	0	0.21	0.27

Before batch correction



	Bcells	CD4Tcells	CD8Tcells	Monocytes	NK
Bcells	0.99	0.01	0	0	0
CD4Tcells	0	0.97	0.02	0.01	0
CD8Tcells	0	0.88	0.12	0	0
Monocytes	0	0.02	0	0.98	0
NK	0	0.04	0.15	0.02	0.79

After batch correction

SVM performance before and after batch correction with Harmony
Training set: PBMC 6k, Test set: PBMC 8k

Results

PBMCs

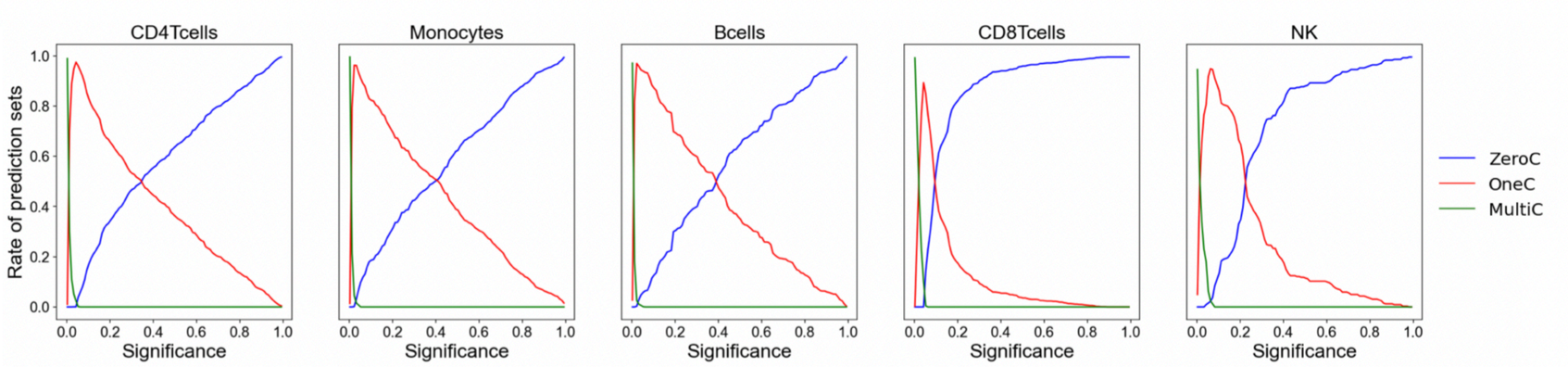
Method	Test set	Average accuracy	Overall accuracy
SVM	PBMC 8k	0.771	0.844
SVM	PBMC 6k	0.616	0.834
SVM (thr 0.7)	PBMC 8k	0.842	0.868
SVM (thr 0.7)	PBMC 6k	0.742	0.863
ICP (With SVM)	PBMC 8k	0.935	0.854
ICP (With SVM)	PBMC 6k	0.790	0.868

Peformance on PBMCs
ICP at significance: 0.025

Results

PBMCs

PBMC 8k

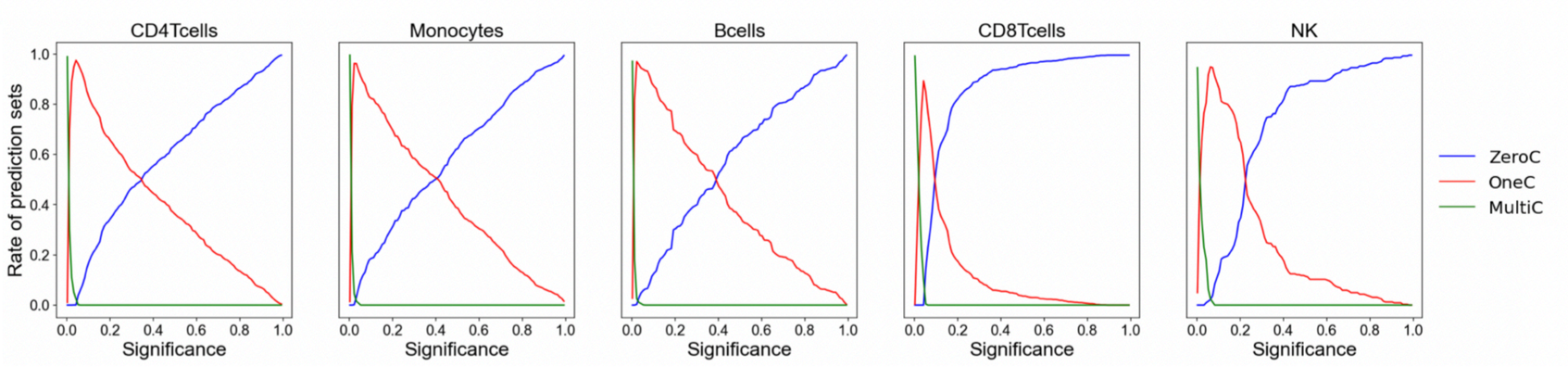


Rate of prediction sets on PBMC 8k

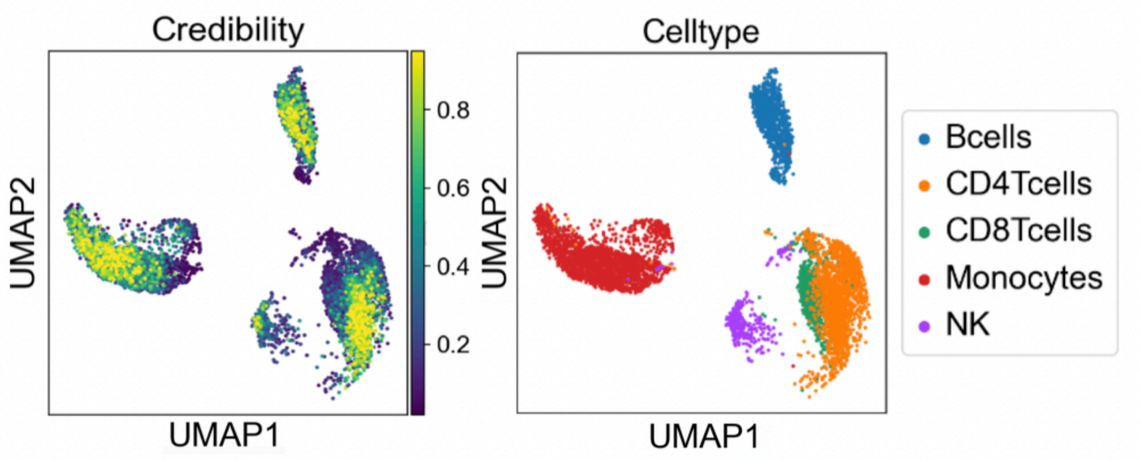
Results

PBMCs

PBMC 8k



Rate of prediction sets on PBMC 8k

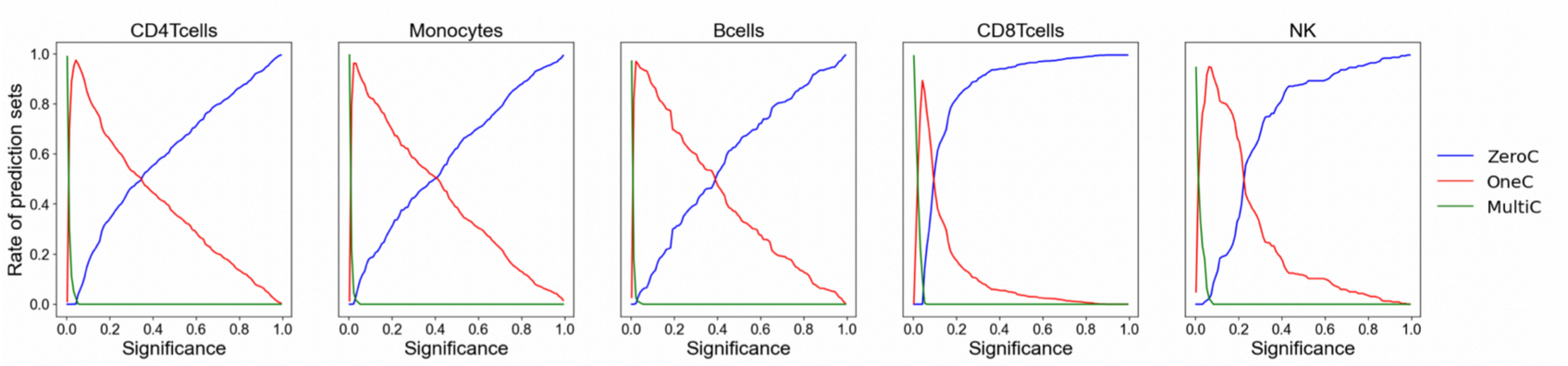


Average credibility of each cell

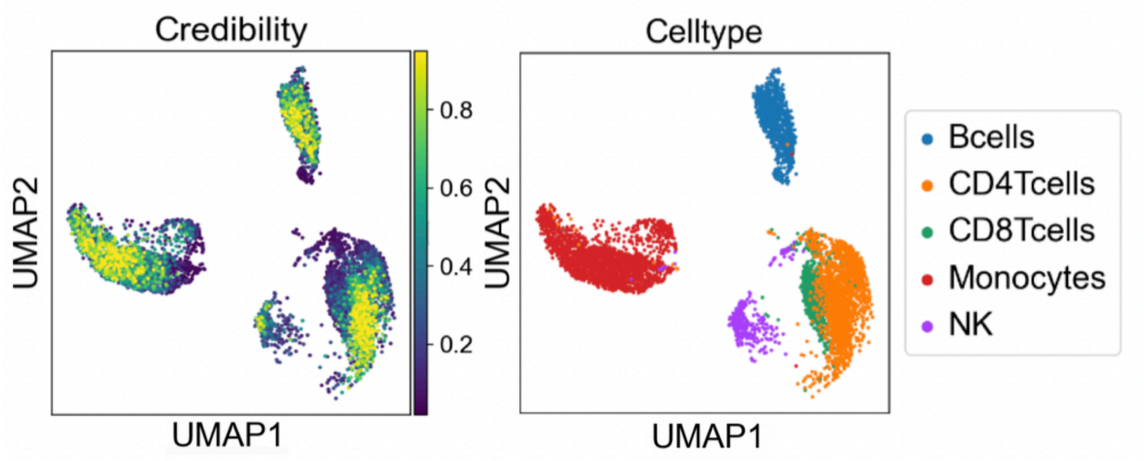
Results

PBMCs

PBMC 8k



Rate of prediction sets on PBMC 8k



Average credibility of each cell

Test set	cell type	Average credibility
PBMC 8k	Bcells	0.417
PBMC 8k	CD4Tcells	0.415
PBMC 8k	CD8Tcells	0.143
PBMC 8k	NK	0.280
PBMC 8k	Monocytes	0.419

Average credibility of each cell type

Results

PBMCs

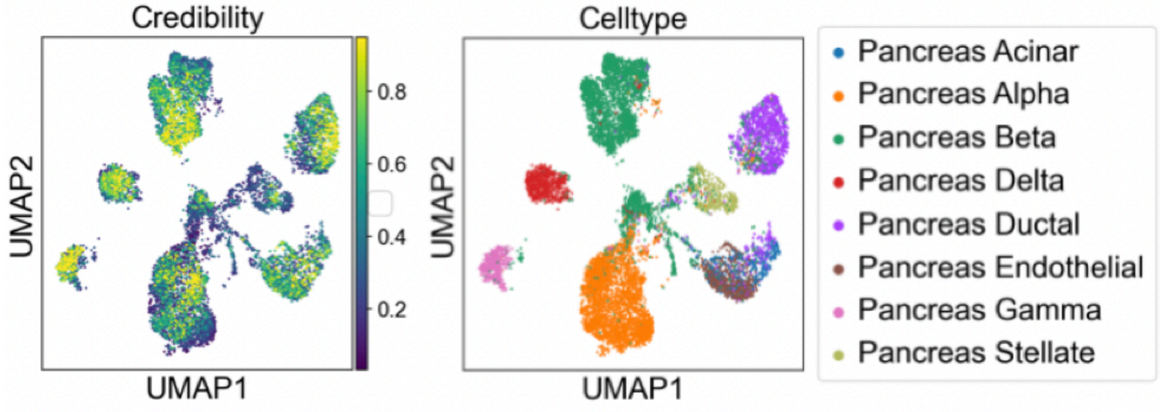
PBMC 8k

Method	Cell type	Classification rate	Accuracy
SVM (thr 0.7)	Bcells	0.982	1
SVM (thr 0.7)	CD4Tcells	0.956	0.988
SVM (thr 0.7)	CD8Tcells	0.808	0.068
SVM (thr 0.7)	NK	0.94	0.987
SVM (thr 0.7)	Monocytes	0.968	0.975
CC-ICP	Bcells	0.968	0.99
CC-ICP	CD4Tcells	0.564	0.98
CC-ICP	CD8Tcells	0.11	0.946
CC-ICP	NK	0.638	0.931
CC-ICP	Monocytes	0.649	0.99

Desired error rate = 0.025 % per cell type

Results

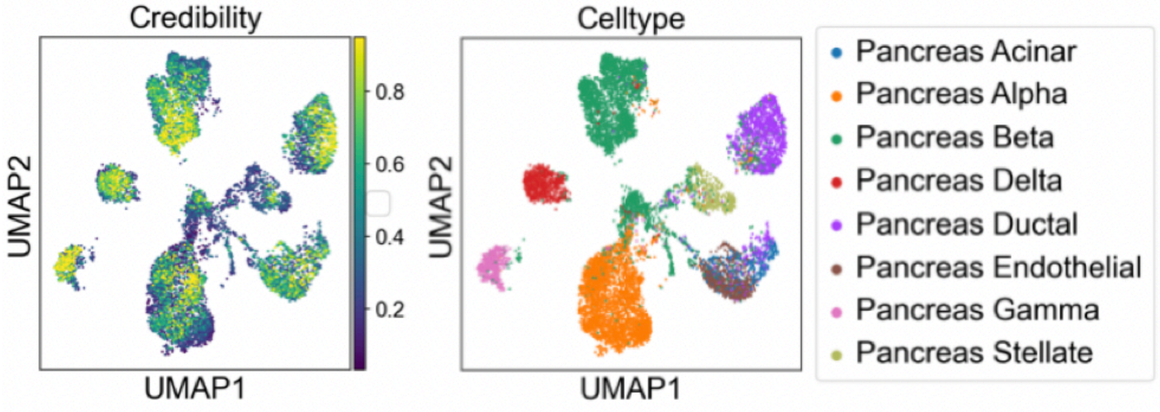
PBMCs



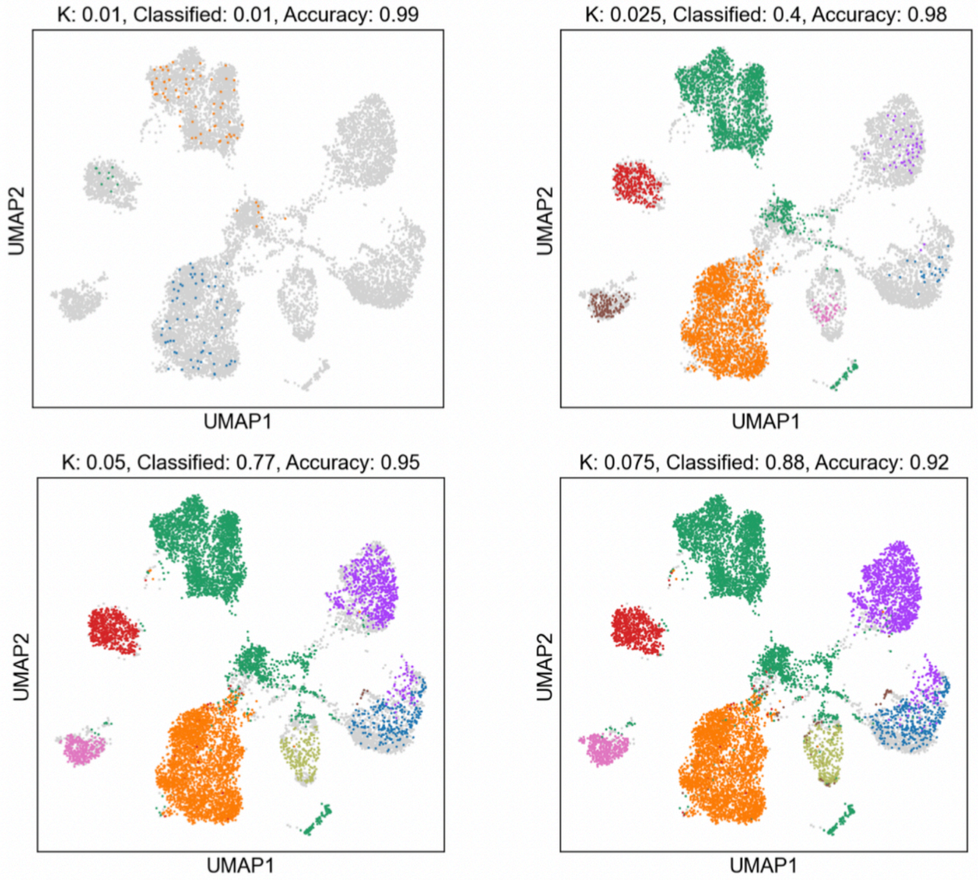
Pancreas datasets after batch correction

Results

PBMCs



Pancreas datasets after batch correction



Desired error rates vs classification and accuracy

Results

PBMCs

Unknow cell type	Assignment on Unknown	Average assignments on known
Alpha	0.08	0.42
Delta	0.59	0.64
Beta	0.68	0.82
Ductal	0.514	0.762
Acinar	0.501	0.761
Gamma	0.455	0.636
Endothelial	0.445	0.748
Stellate	0.379	0.705

Assignments of held-out (unknown) cell types

Results

Impact of batch correction

Batch transfer may itself cause error rather than the model itself

Evaluated two other algorithm on PBMCs:

- Scanorama (Hie et. al., 2019, *Nature Biotechnology*)
- Scgen (Lotfollahi et. al., 2019, *Nature Methods*)

Results

Impact of batch correction

Batch transfer may itself cause error rather than the model itself

Evaluated two other algorithm on PBMCs:

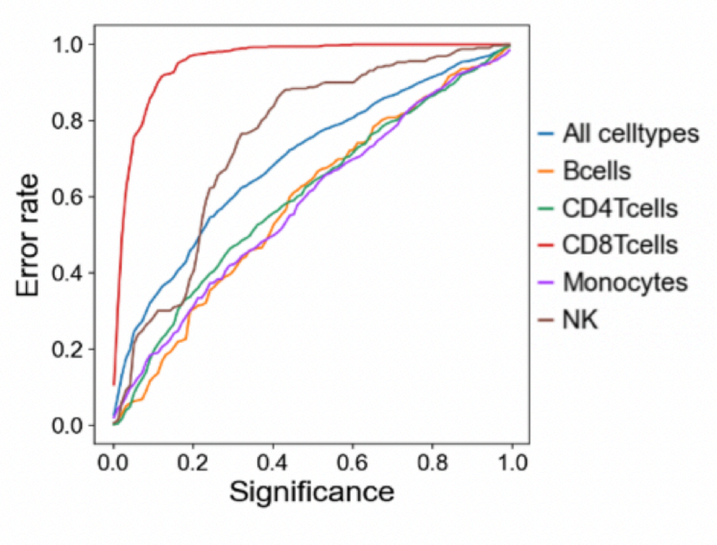
- Scanorama (Hie et. al., 2019, *Nature Biotechnology*)
- Scgen (Lotfollahi et. al., 2019, *Nature Methods*)

Method	**Homogeneity score**
Harmony	0.613
Scanorama	0.812
scGen	0.795

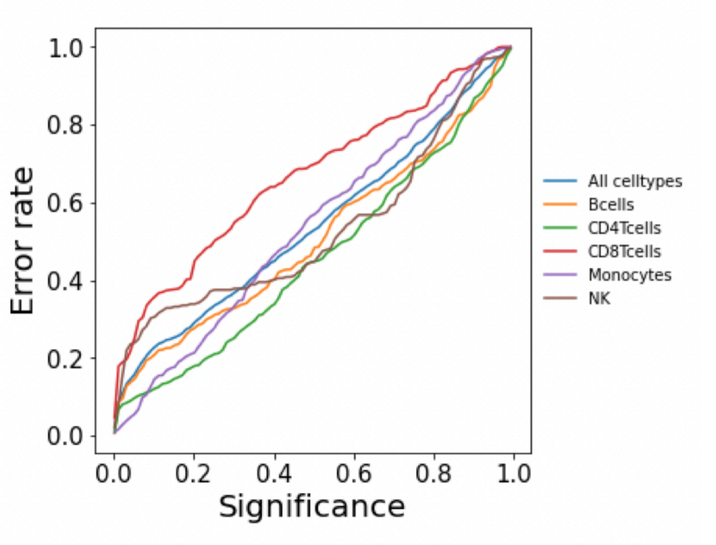
Homogeneity scores for integrating PBMC datasets

Results

Impact of batch correction



Test set error rates
Batch correction: Harmony



Test set error rates
Batch correction: Scanorama

Should we use Scanorama always? Perhaps not. Quality of batch correction methods is data-dependent.

Lessons learned and outook

- Conformal-credibility setup provides an approach for accurate label transfer
- Uncertainties drastically affected by the batch correction method in use

Lessons learned and outook

- Conformal-credibility setup provides an approach for accurate label transfer
- Uncertainties drastically affected by the batch correction method in use
- Progress to be made on three ends:
  - Technical:
    - Evaluating of a set of NCMs and further development of algorithms
    - Extensive benchmarking with multiple batch correction methods *(Ongoing)*
    - Conformal anomaly detection for discovering unknown cell types

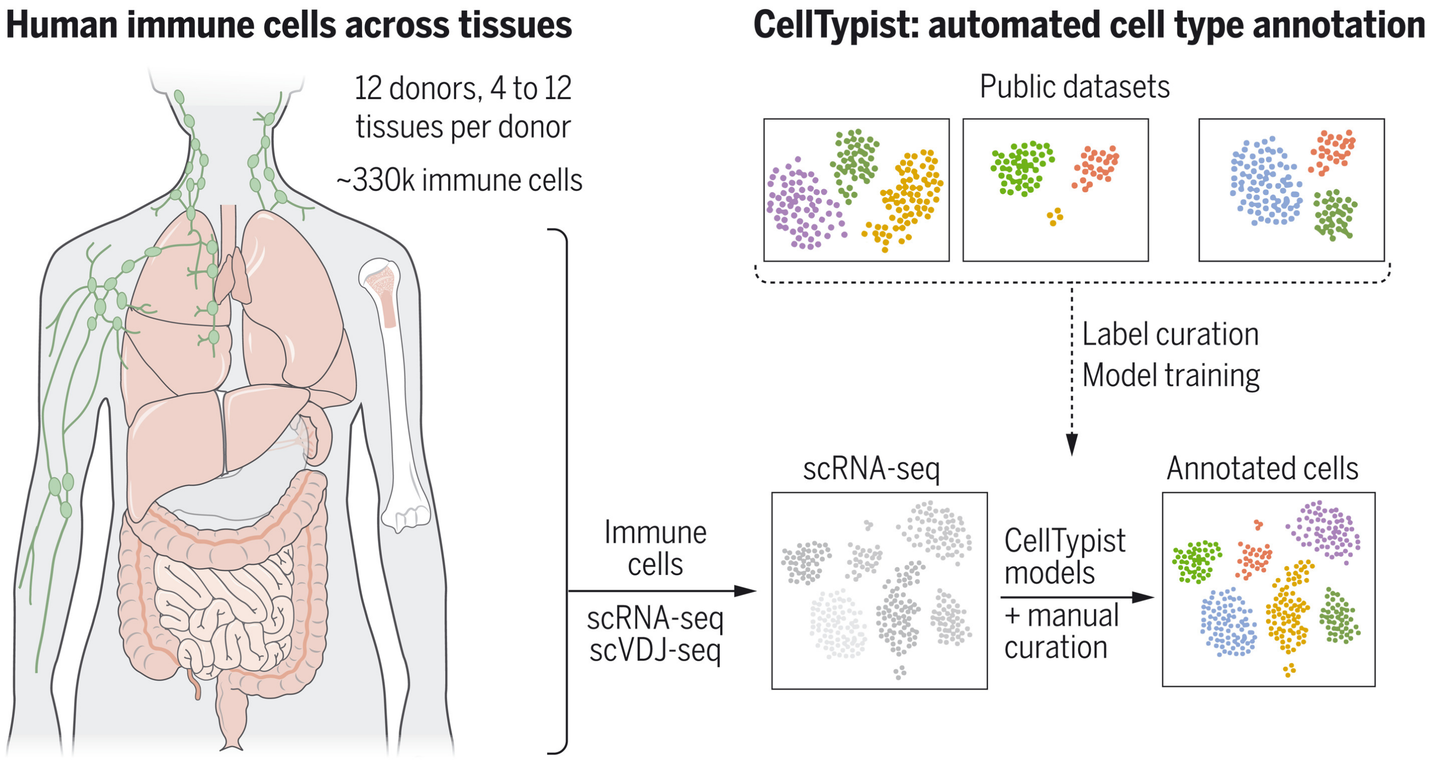
Lessons learned and outook

- Conformal-credibility setup provides an approach for accurate label transfer
- Uncertainties drastically affected by the batch correction method in use
- Progress to be made on three ends:
  - Technical:
    - Evaluating of a set of NCMs and further development of algorithms
    - Extensive benchmarking with multiple batch correction methods *(Ongoing)*
    - Conformal anomaly detection for discovering unknown cell types
  - Biological:
    - Going deeper into cellular subsets (*Ongoing)*
    - Discovering cell states

Lessons learned and outook

- Conformal-credibility setup provides an approach for accurate label transfer
- Uncertainties drastically affected by the batch correction method in use
- Progress to be made on three ends:
  - Technical:
    - Evaluating of a set of NCMs and further development of algorithms
    - Extensive benchmarking with multiple batch correction methods *(Ongoing)*
    - Conformal anomaly detection for discovering unknown cell types
  - Biological:
    - Going deeper into cellular subsets (*Ongoing)*
    - Discovering cell states
  - Datasets:
    - Use atlas-level datasets (Some exciting recent developments and ongoing global efforts)

Recent developments



C. DOMÍNGUEZ CONDE et. al., 2022, *Science*

Classification without the use of batch correction (https://www.celltypist.org), thanks to curation of vast amounts of cell types and careful harmonization of multiple public datasets