Institut Cochin

A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Tran HTN, Ang KS, Chevrier M, Zhang X, Shin Lee YS, Goh M, Chen J. (2020) Genome Biology 21(12)

10/06/2021

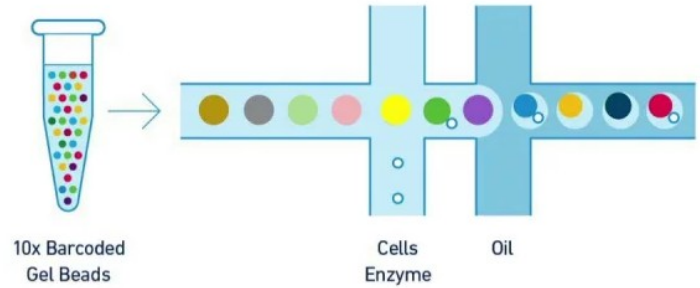
Instituts thématiques Inserm Institut national de la santé et de la recherche médicale

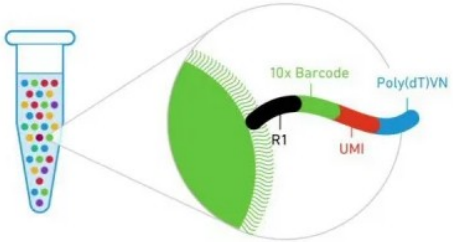
cnrs

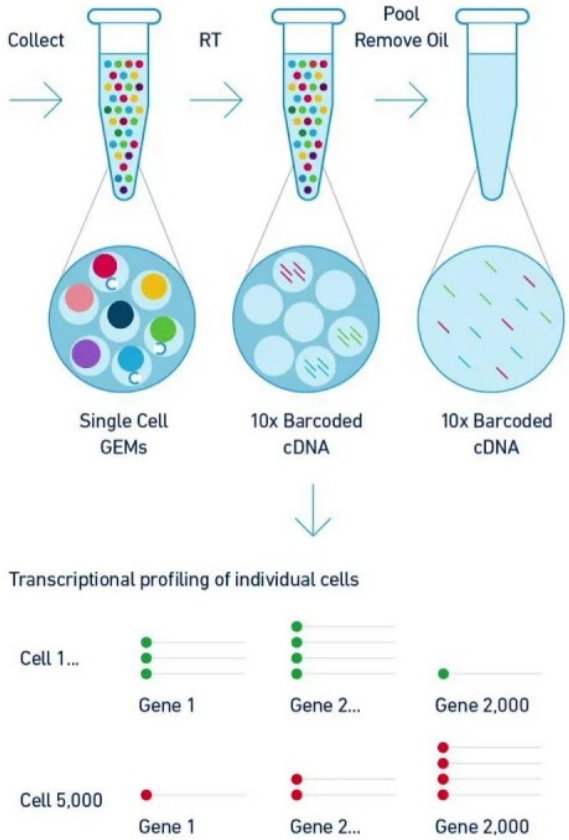
Université de Paris

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

single-cell RNA seq





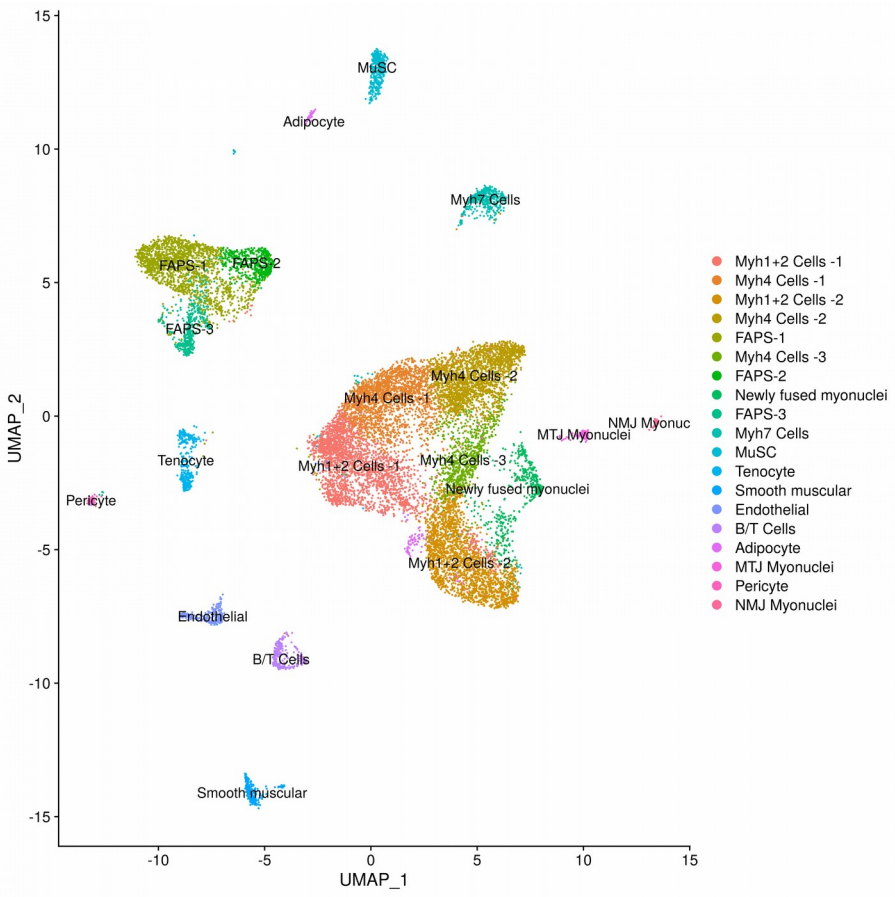


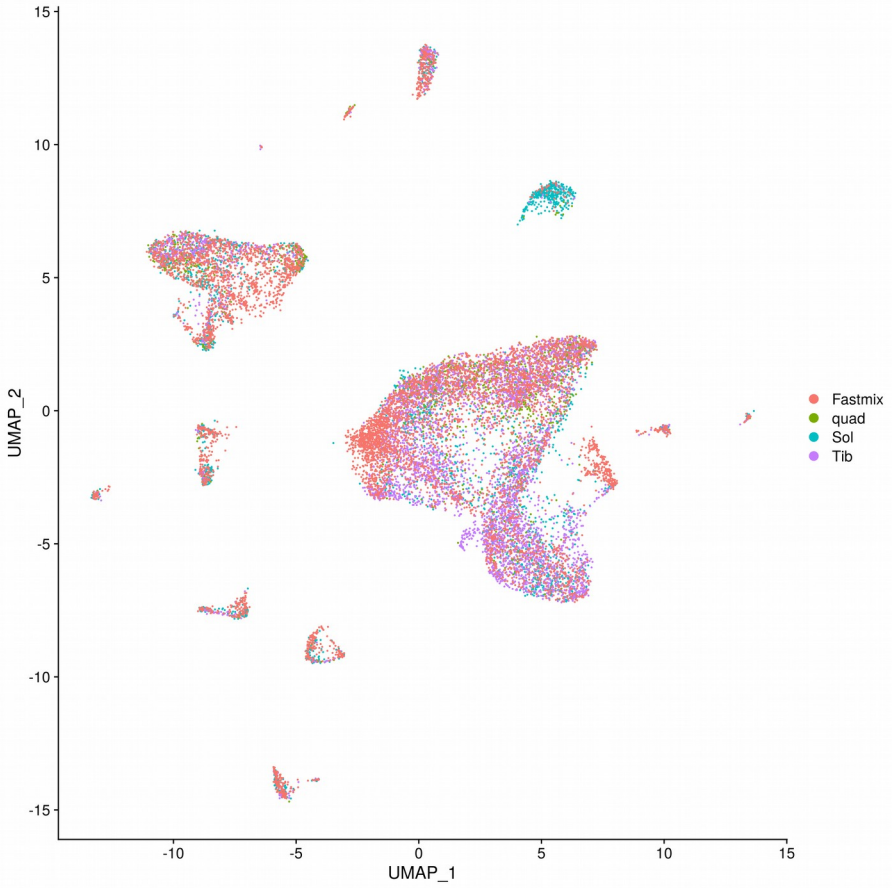
10x GENOMICS

- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

single-cell RNA seq

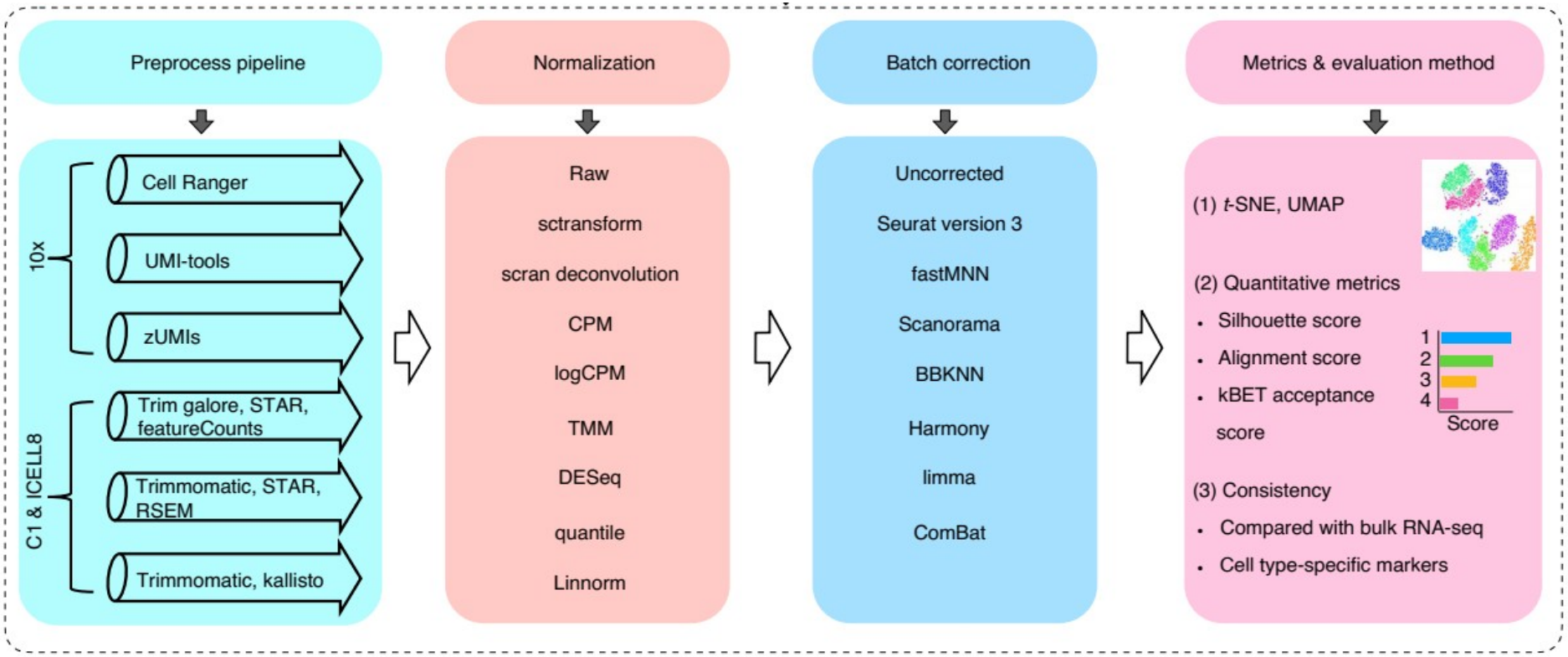




Dos Santos et al., *submitted*

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

single-cell RNA seq



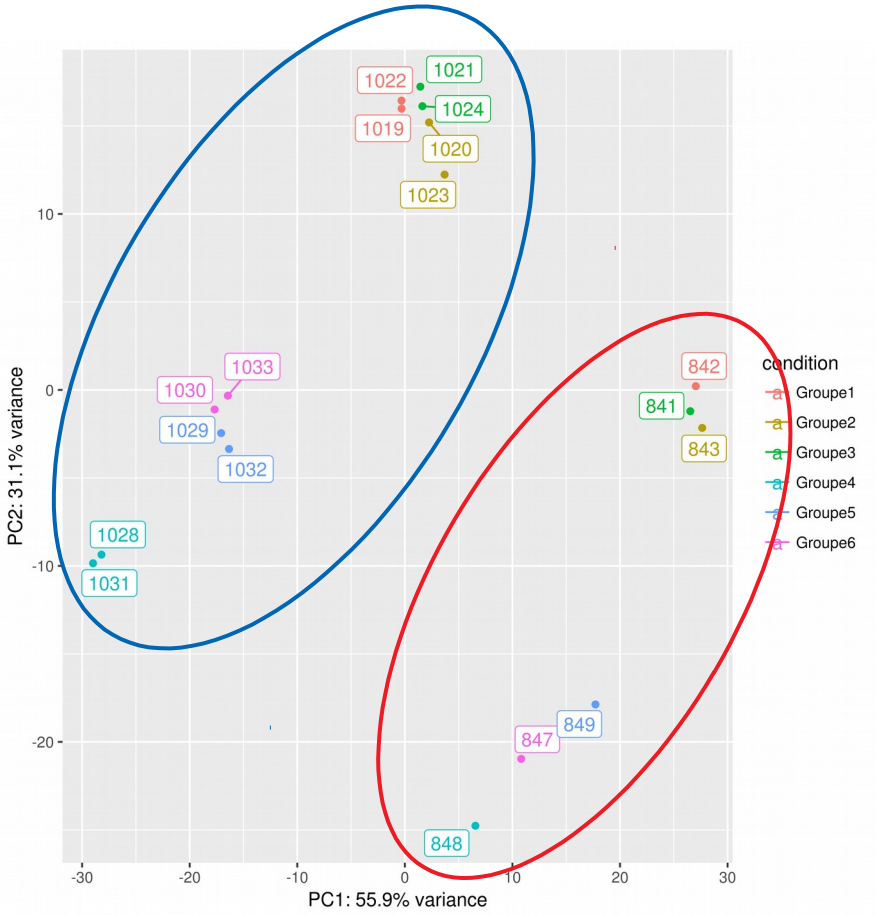
Chen, W., Zhao, Y., Chen, X. et al. (2020)

Batch-effect

Data compiled from multiple experiments

- Capturing times
- Handling personnel
- Reagents lots
- Equipements and technology

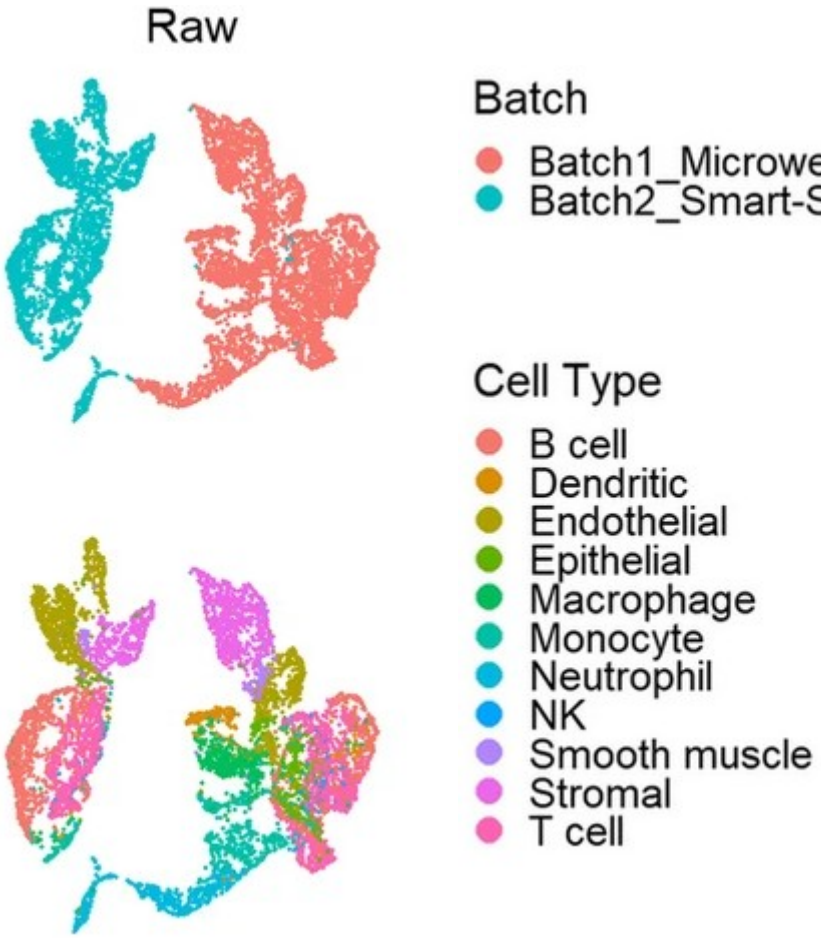
Batch-effect



Plateforme GENOM'IC

- Correctly align different datasets
- Conserving key biological variations
- Handle large data with drop-out effect

Batch-effect



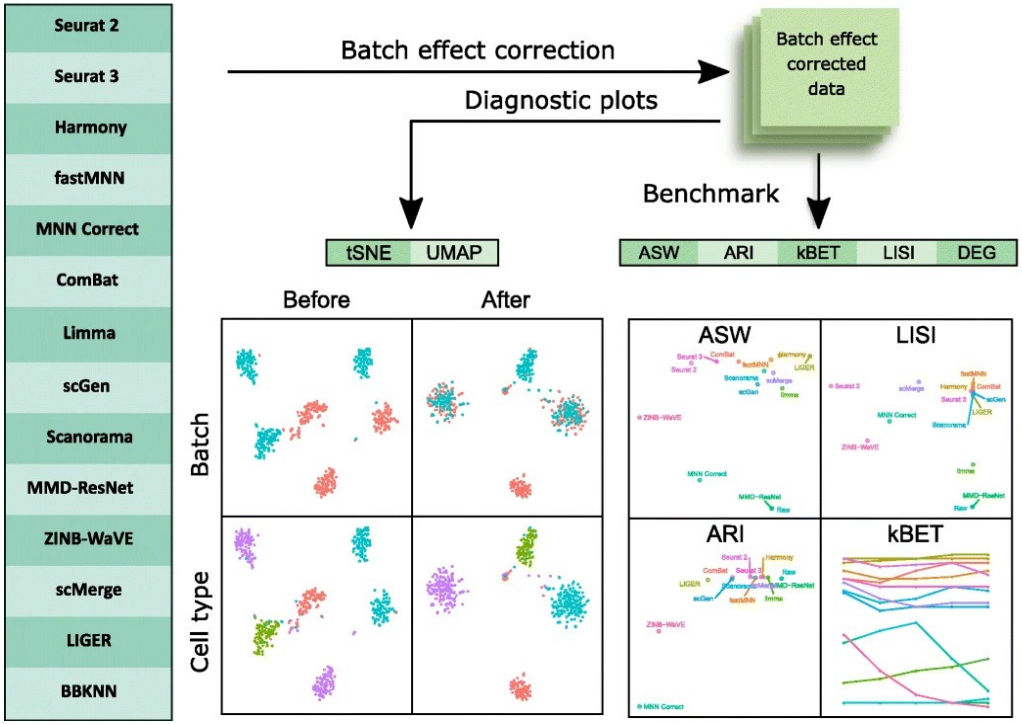
- Correctly align different datasets
- Conserving key biological variations
- Handle large data with drop-out effect

Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Benchmarking 14 methods on ten datasets using five evaluation metrics

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

A



B

Dataset	Description	Number of batches	Total cell number	Technologies
1	Human Dendritic Cells	2	576	Smart-Seq2
2	Mouse Cell Atlas	2	6,954	Microwell-Seq
				Smart-Seq2
3	Simulation	Refer to Simulation table		
4	Human Pancreas	5	14,767	inDrop
				CEL-Seq2
				Smart-Seq2
				SMARTer
				SMARTer
5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3'
				10x 5'
6	Cell line	3	9,530	10x
7	Mouse Retina	2	71,638	Drop-seq
8	Mouse Brain	2	833,206	Drop-seq
				SPLiT-seq
9	Human Cell Atlas	2	621,466	10x
10	Mouse Haematopoietic Stem and Progenitor Cells	2	4,649	MARS-seq
				Smart-Seq2

Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Benchmarking 14 methods on ten datasets using five evaluation metrics

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

Aim : answer to 5 scenarios

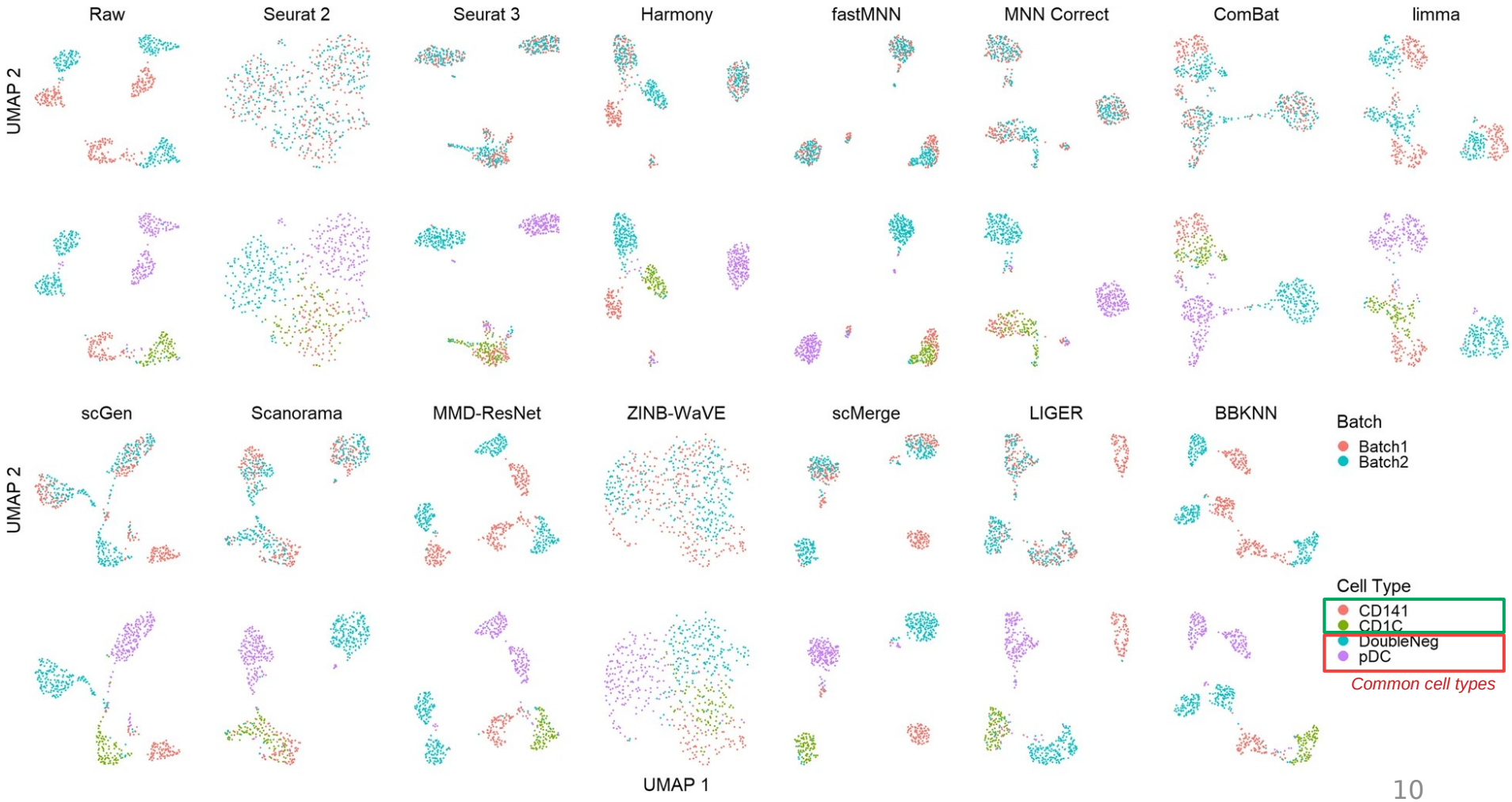
- Identical cell types, different technologies
- Simulation
- non-identical cell types
- Big data
- Multiples batches

Tools : 5 evaluation metrics

- Adjusted Rand Index (ARI)
- k-nearest-neighbor batch-effect test (kBET)
- local inverse Simpson's index (LISI)
- Differentially expressed genes (DEG)
- average silhouette width (ASW)

Non-identical cell types

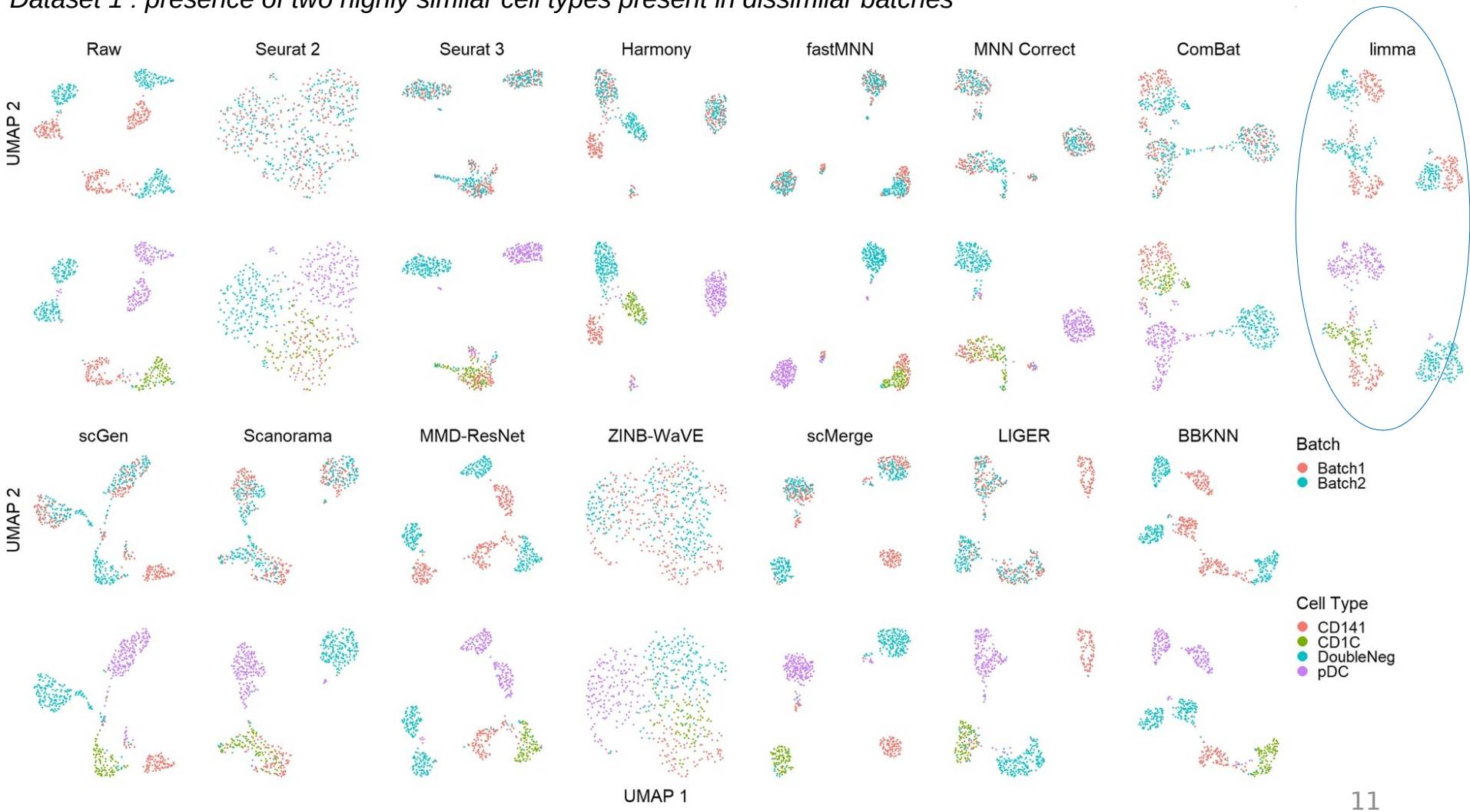
Dataset 1 : presence of two highly similar cell types present in dissimilar batches



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Non-identical cell types

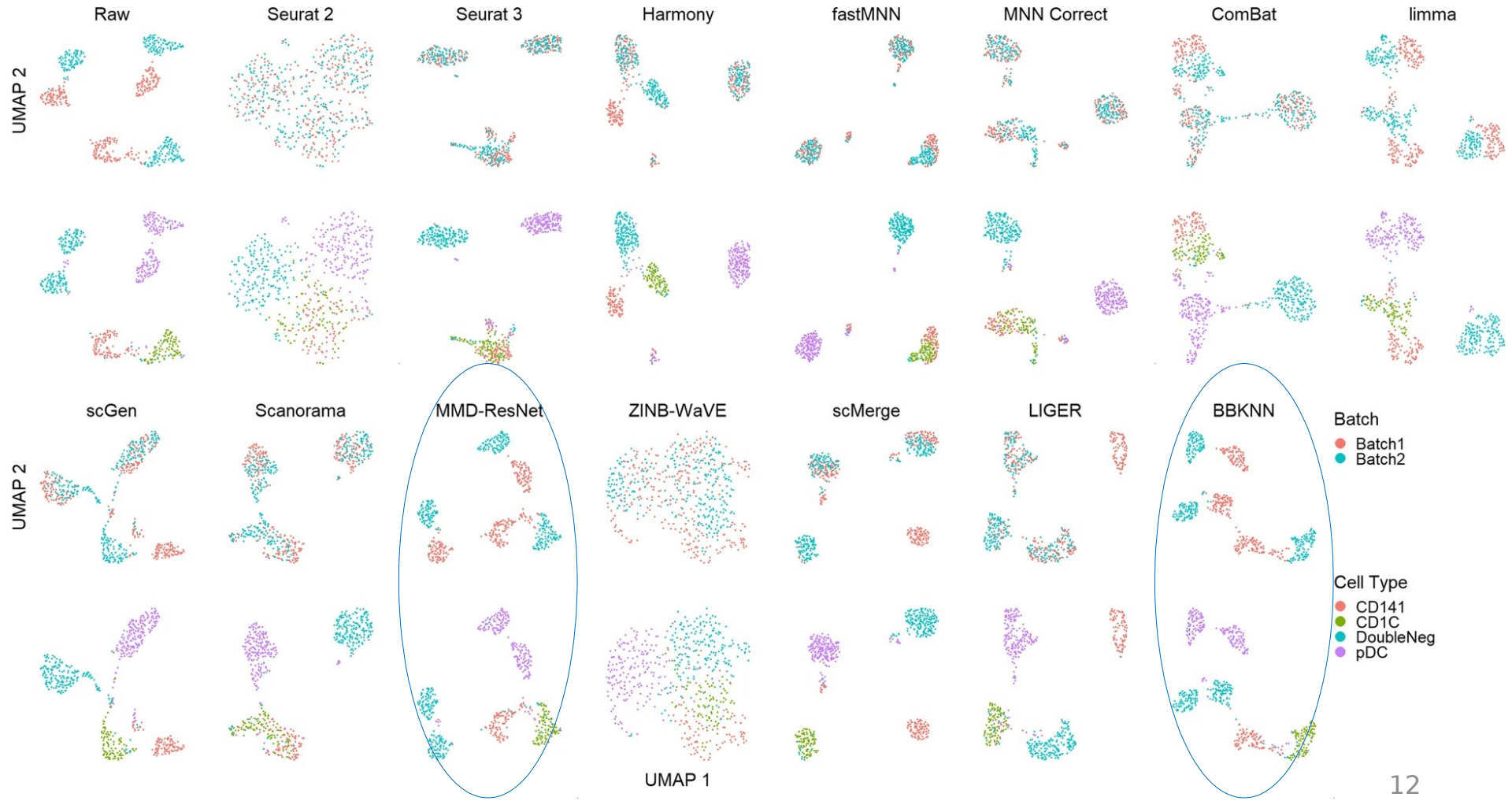
Dataset 1 : presence of two highly similar cell types present in dissimilar batches



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Non-identical cell types

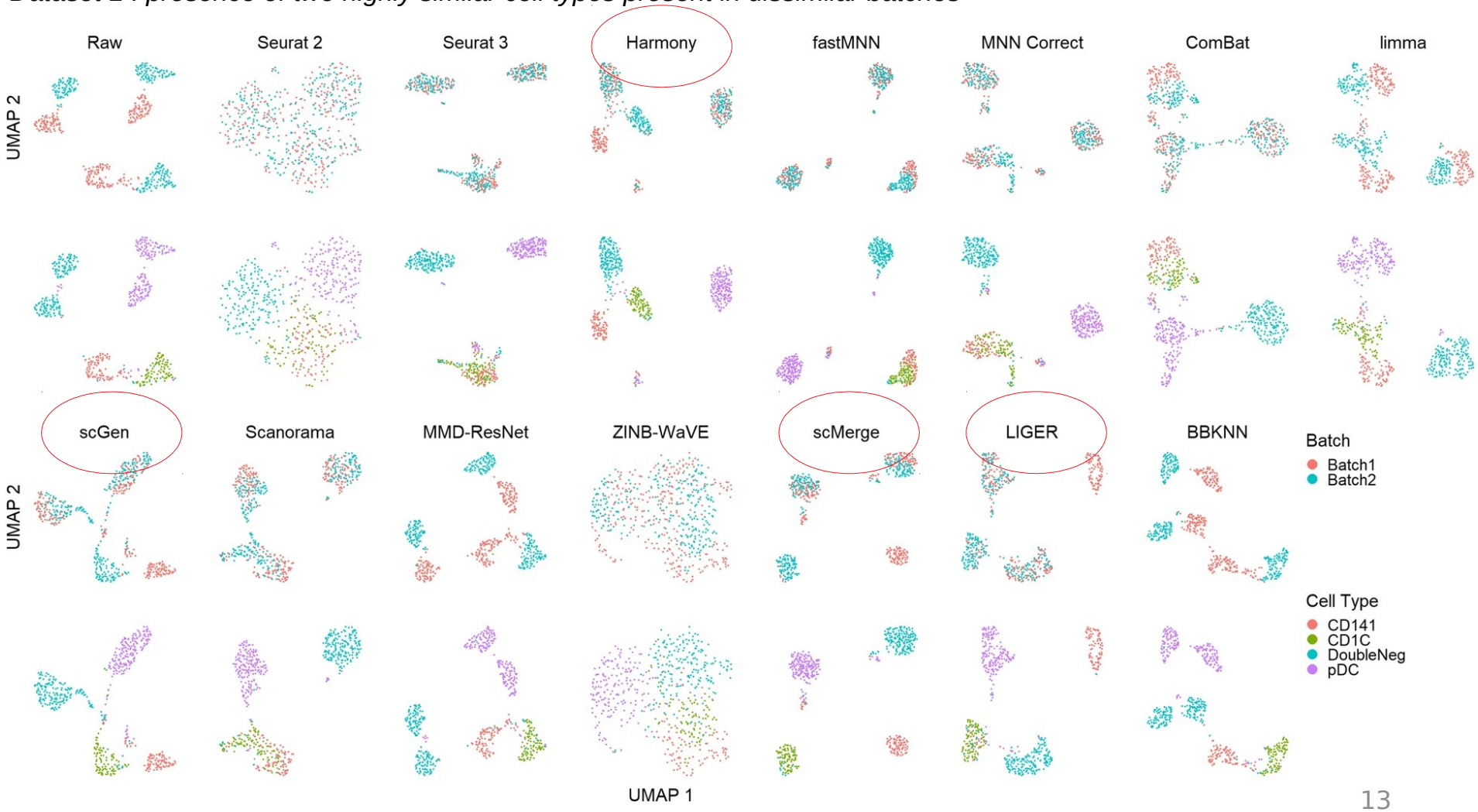
Dataset 1 : presence of two highly similar cell types present in dissimilar batches



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Non-identical cell types

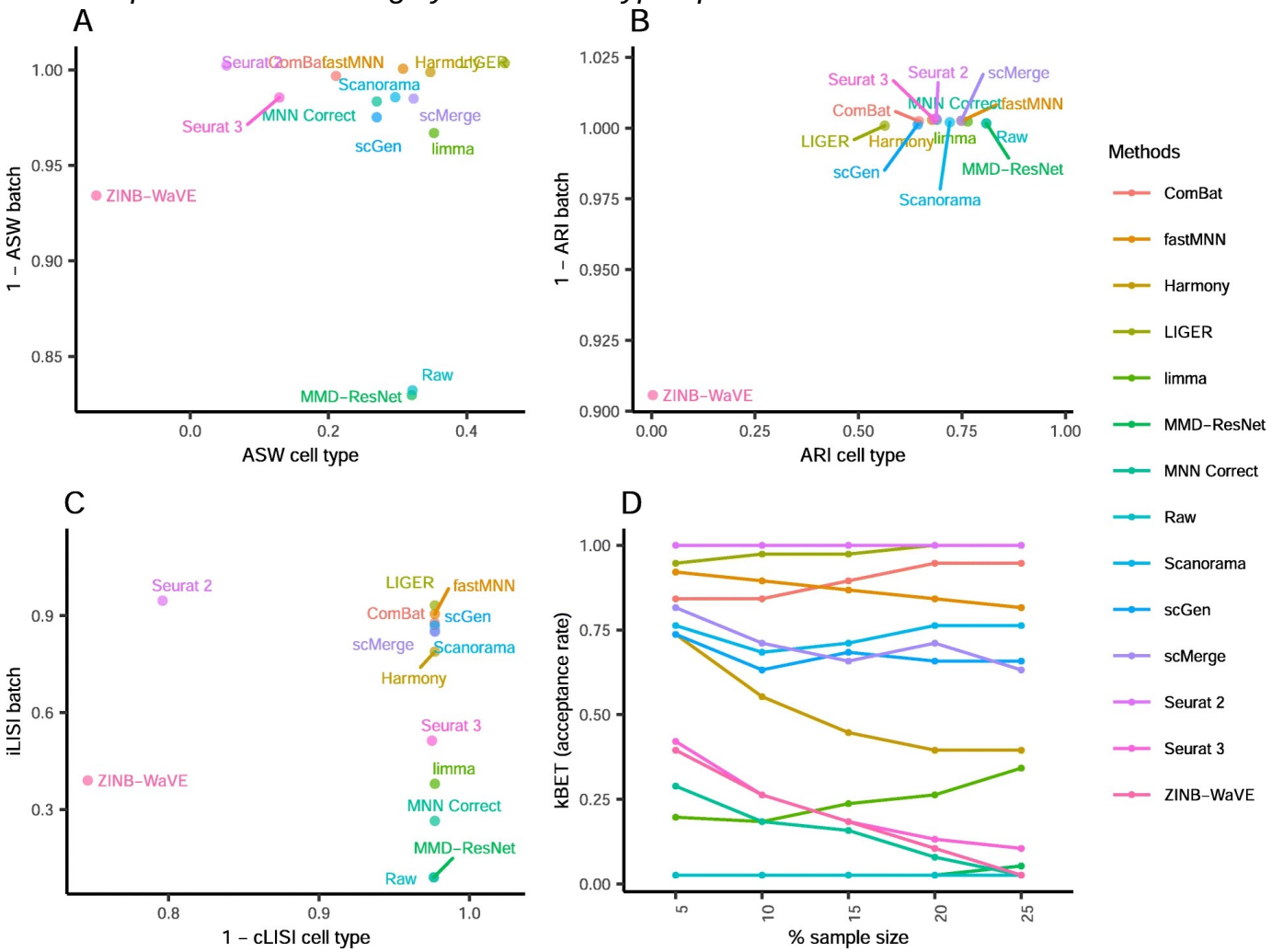
Dataset 1 : presence of two highly similar cell types present in dissimilar batches



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Non-identical cell types

Dataset 1 : presence of two highly similar cell types present in dissimilar batches



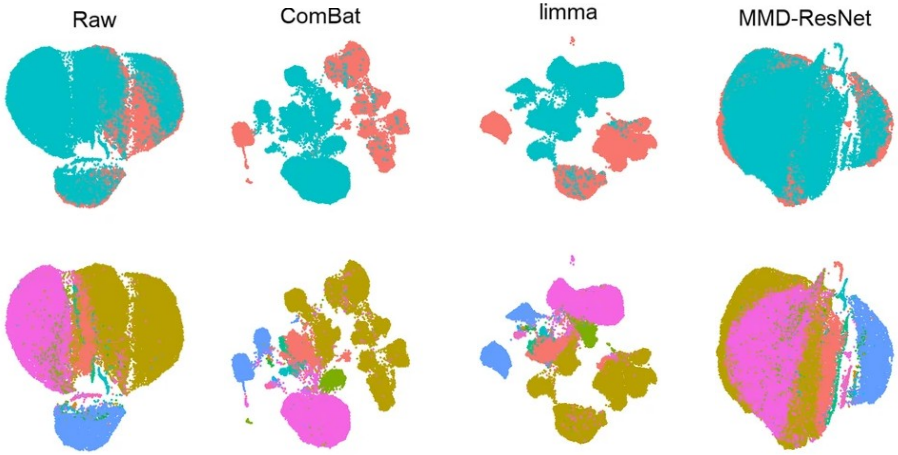
Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Non-identical cell types

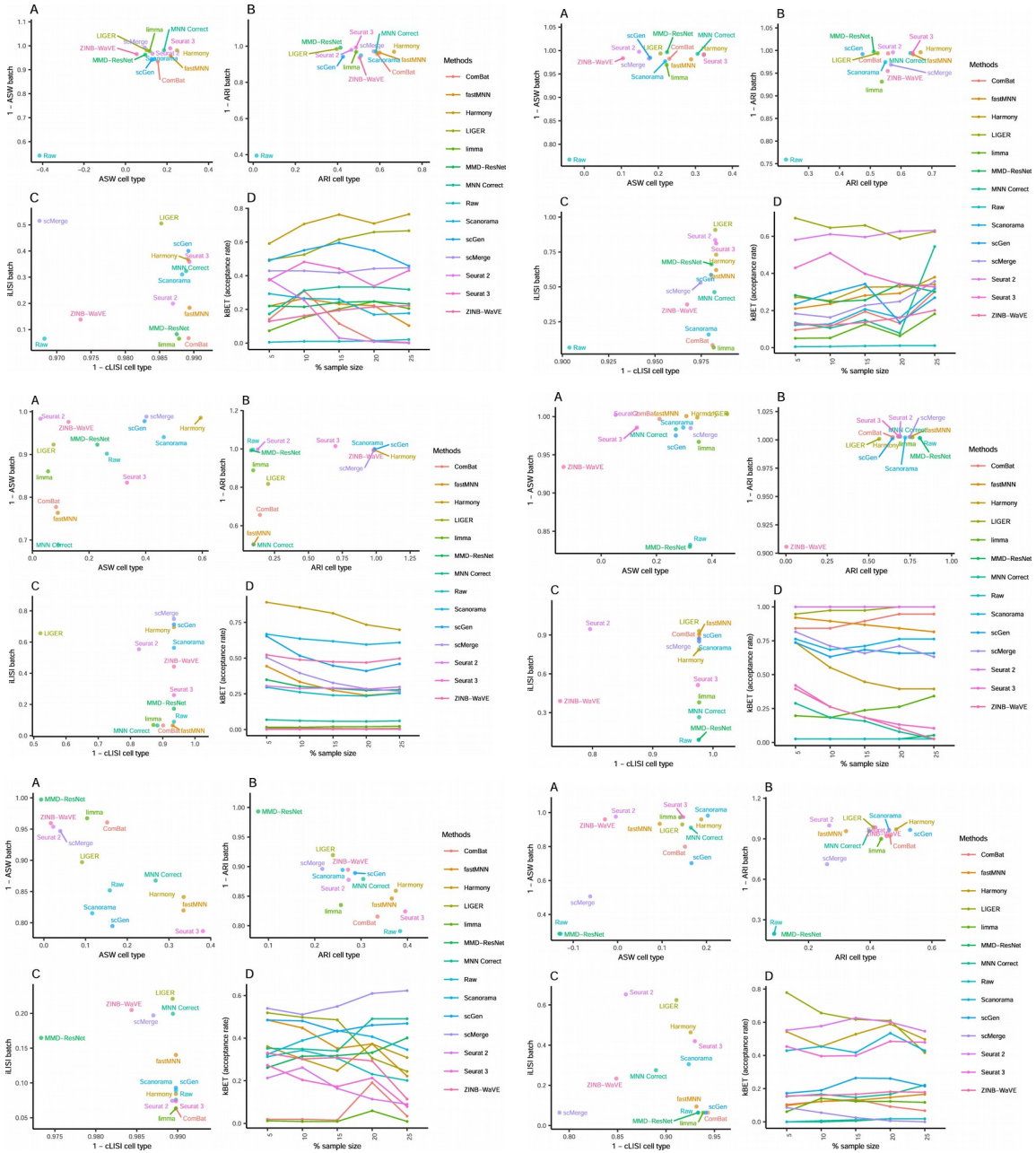
Dataset 1 : presence of two highly similar cell types present in dissimilar batches

Dataset 6 : two cell types, three batches (the third is the mixed one)

Dataset 7 : cell counts across cell types highly uneven within batches

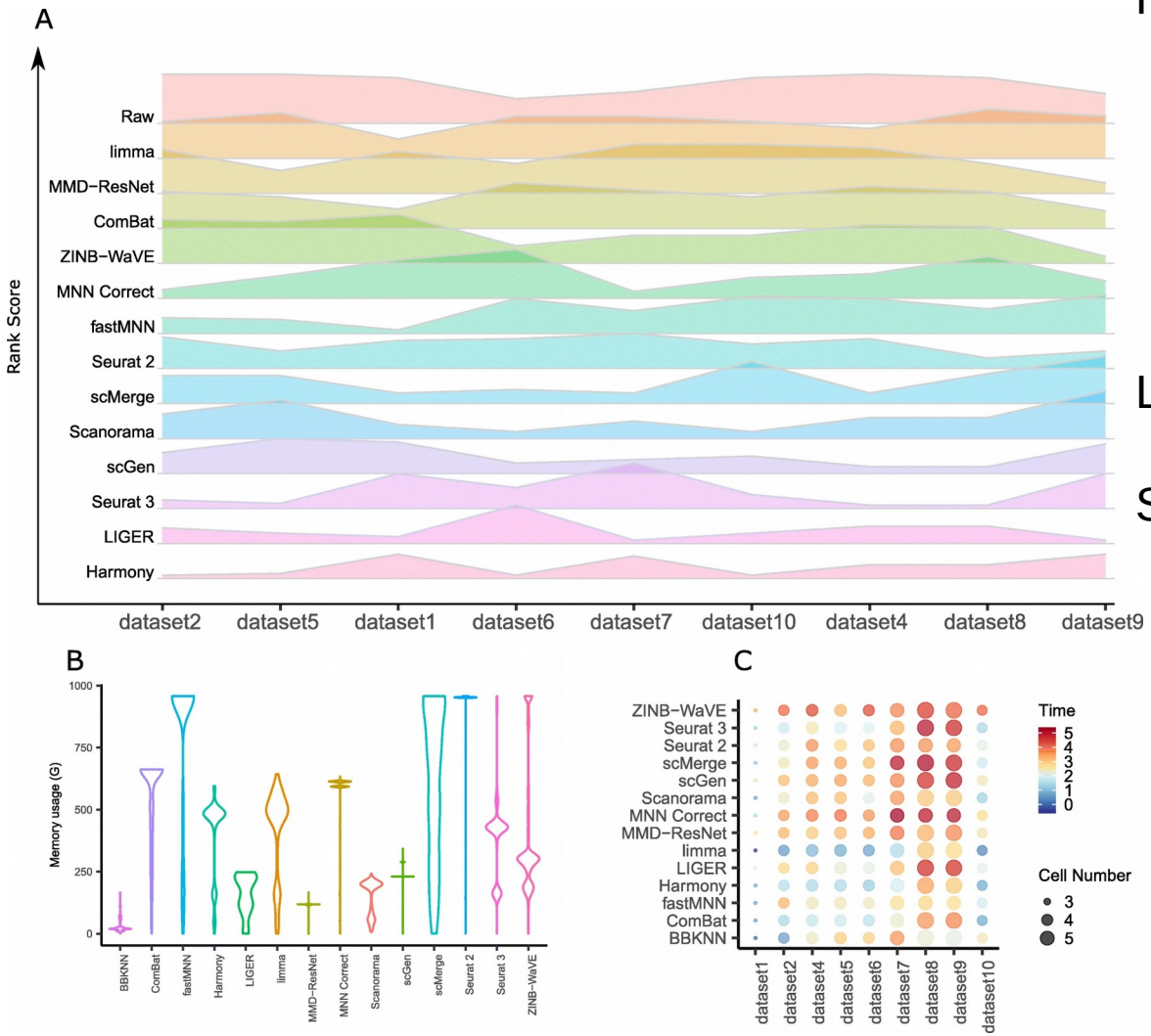


Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Discussion on rank & performance



Harmony

- Recomended for non-identical cell types or different technologies
- Not that good for large datasets

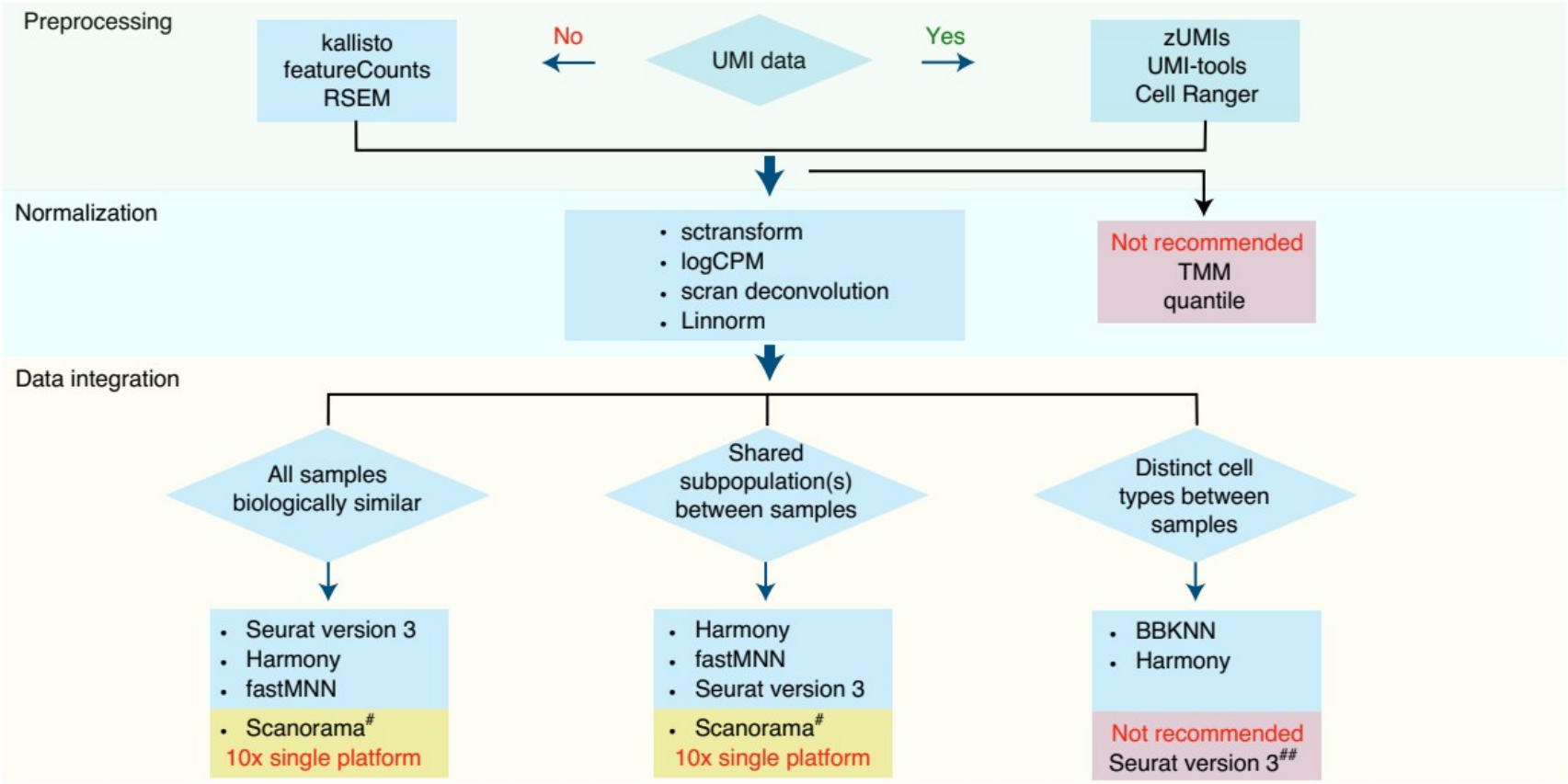
LIGER

Seurat 3

ComBat, MMD-ResNet, and limma were the worst performing methods.

Tran, H.T.N., Ang, K.S., Chevrier, M. et al. (2020)

Discussion



Chen, W., Zhao, Y., Chen, X. et al. (2020)

Discussion

Box 1 | Best practice recommendations

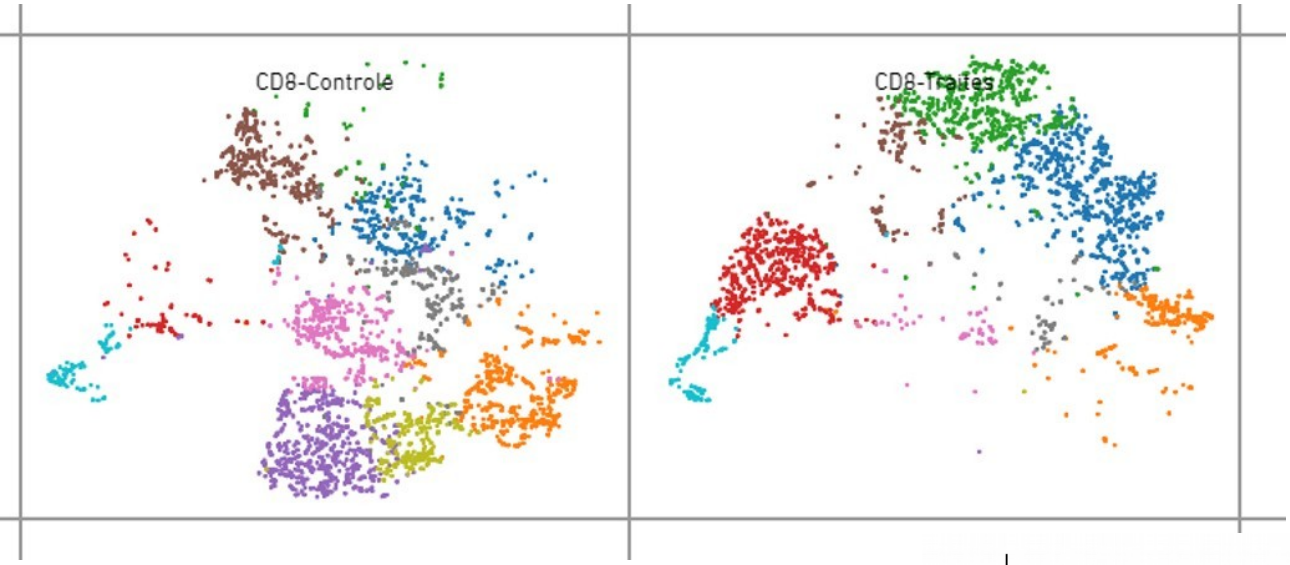
We summarize below 11 best practice recommendations for the community based on our analysis.

1. There were large variations across different scRNA-seq platforms and centers.
2. While most of the genes and cells detected were consistent between the different methods, we observed variations for low-expression genes and cells with low mRNA content across different methods. However, these differences did not affect our analyses of cell classification or mixability.
3. Normalization algorithms alone could not remove batch effects.
4. Different normalization strategies performed differently across datasets and platforms; sctransform, scran, logCPM and Linnorm performed well for either 3′- or full-length-transcript scRNA-seq platforms, but TMM and quantile performed poorly and are not recommended.
5. Seurat version 3, Harmony, BBKNN, fastMNN and Scanorama all could correct and remove batch variations in specific sample and dataset scenarios; we recommend users apply appropriate batch-effect correction methods depending on the characteristics of their datasets (for example, cellular and sample heterogeneity and composition, platforms used; Fig. 6e).
6. BBKNN, fastMNN and Harmony ranked best for clusterability/cell type classification, whereas Seurat version 3, Harmony and fastMNN performed best for mixability.
7. fastMNN, BBKNN and Harmony removed batch variations effectively across different platforms, including both mixed and unmixed distinct samples, but the order of importing the datasets into the pipeline and the requirement for a mixed sample was critical for MNN and fastMNN, whereas BBKNN and Harmony performed well regardless of the inclusion of mixed heterogeneous biologically distinct samples across platforms and batches; thus, for MNN and fastMMN, we recommend including a mixed sample and importing the mixed data into the pipeline first.
8. CCA/Seurat version 3 had superior mixability for biologically similar samples but overcorrected batch effects and misclassified cells (that is, poor clusterability/cell type classification) when large proportions of distinct cell types were present. However, Seurat version 3 performed well both for clusterability and mixability for datasets when only a small fraction of dissimilar cells (for example, 5–10%) was present. Thus, we do not recommend using CCA Seurat version 3 for scenarios containing large fractions of biologically distinct cell types.
9. BBKNN performed best in clusterability and cell type classification, but it ranked low in mixability, particularly in heterogeneous cell samples.
10. The current version of Scanorama performed well only for the 10x Genomics data and did not work for non-10x platforms; thus, we do not recommend it for data from non-10x platforms.
11. We observed good consistency between Cell Ranger 3.1 and 2.0 preprocessed data; however, Cell Ranger 3.1 can detect some extra cells with very few transcripts; this may affect batch-effect corrections in certain scenarios.

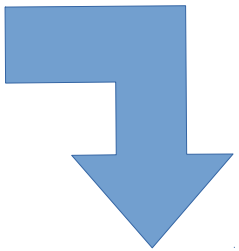
Chen, W., Zhao, Y., Chen, X. et al. (2020)

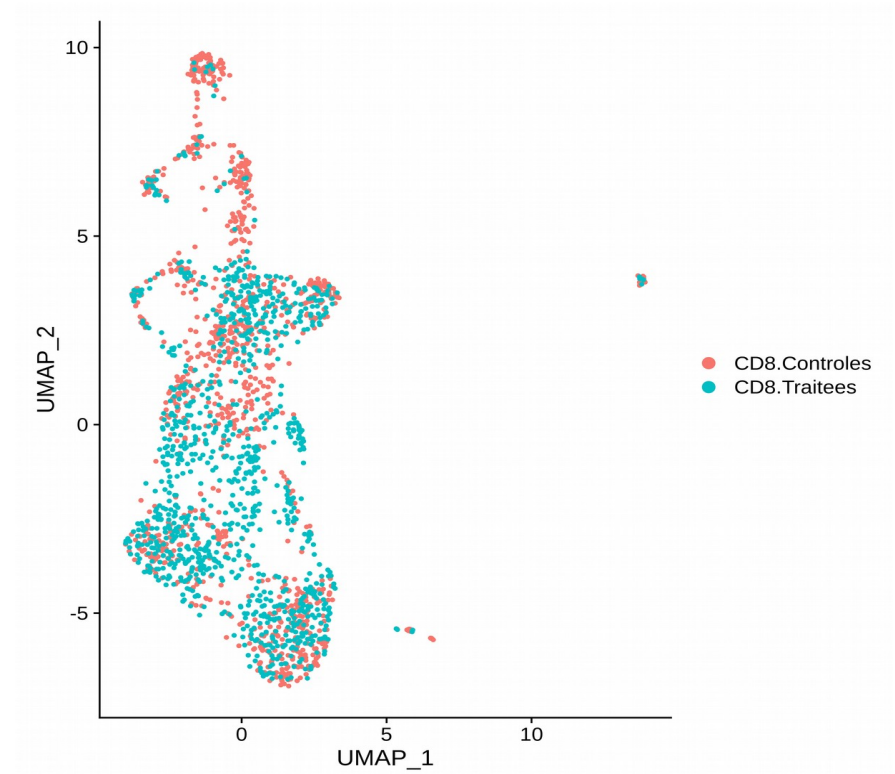
GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

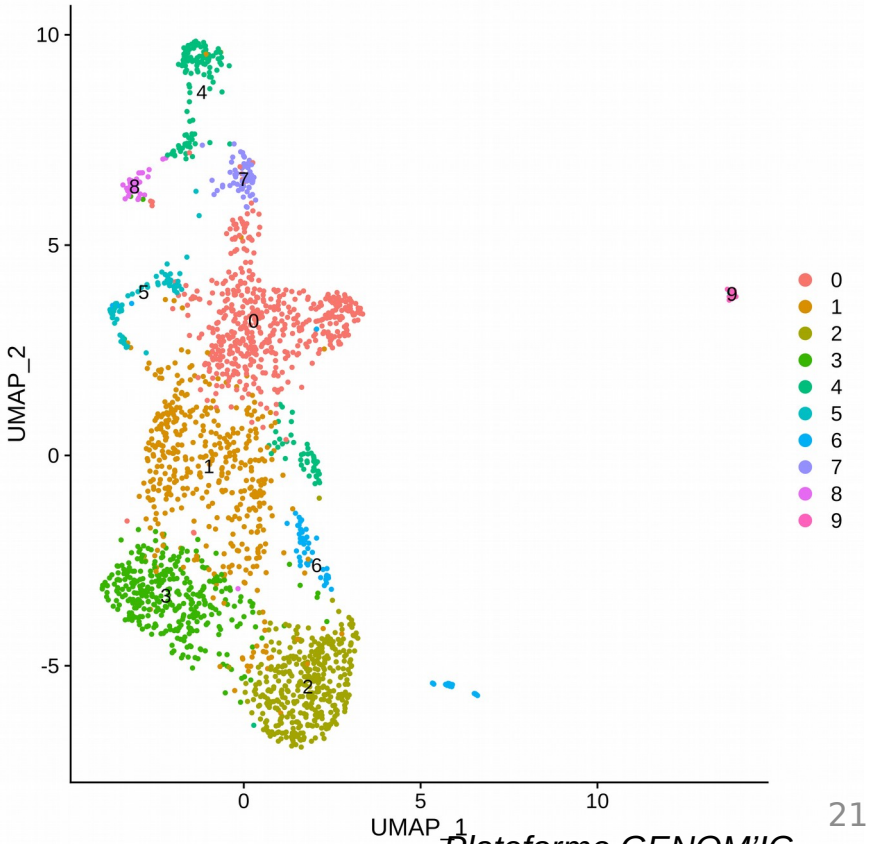
On our plateform : cellranger vs Seurat3



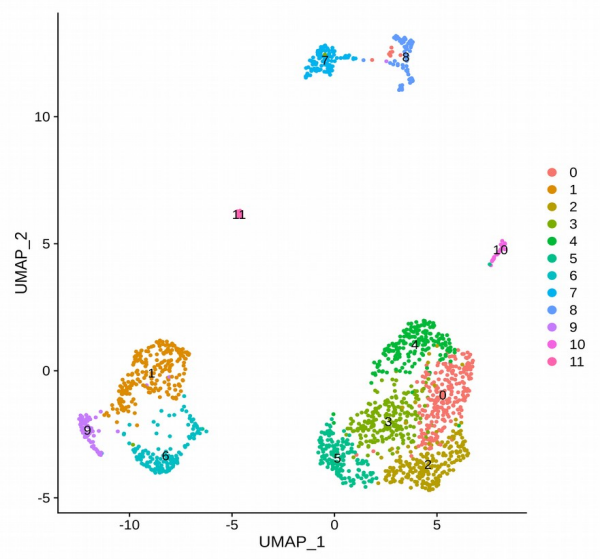
GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

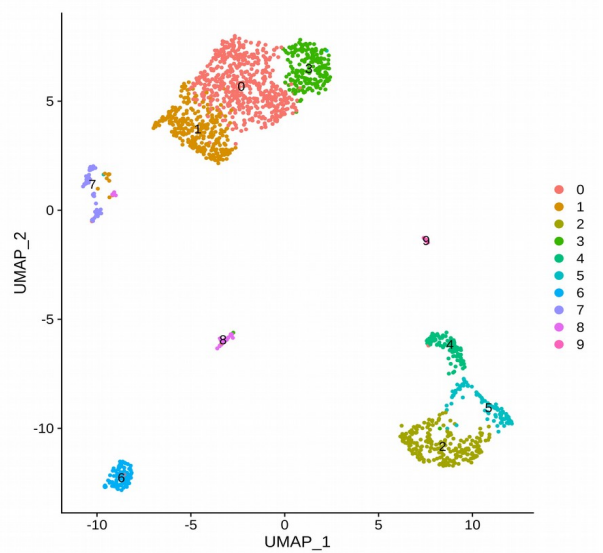


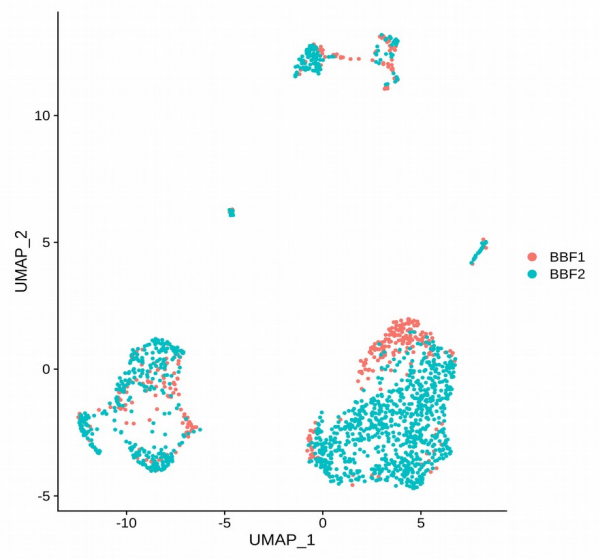


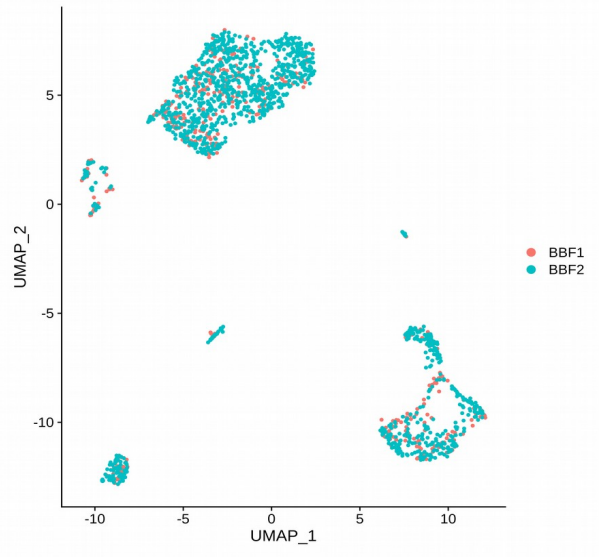


NGS20-090 : Before and After integration (Seurat3)

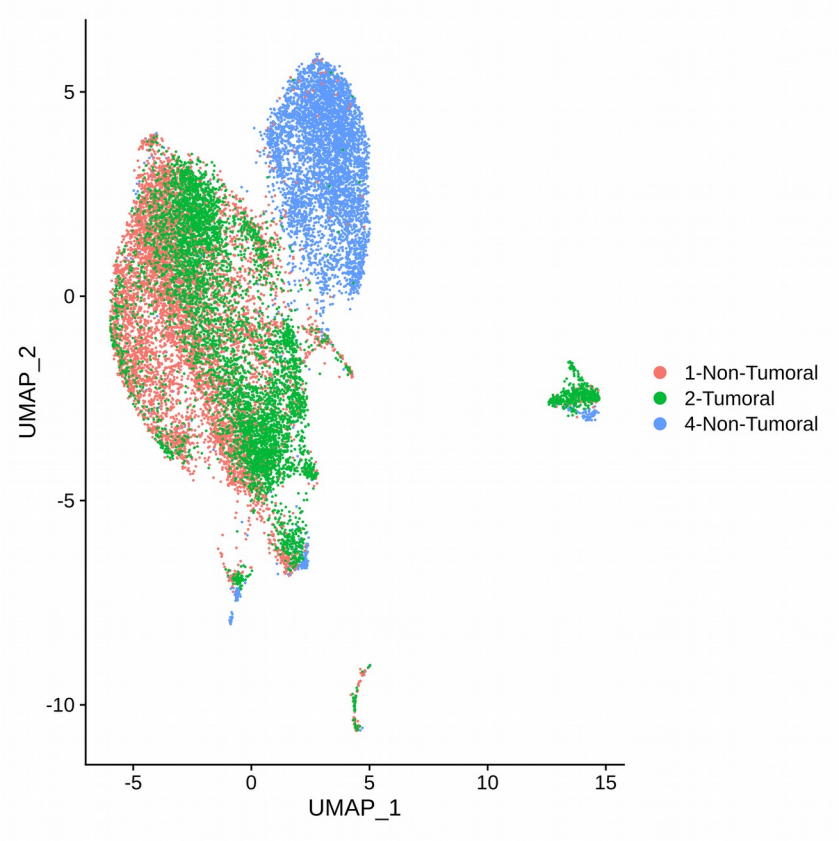


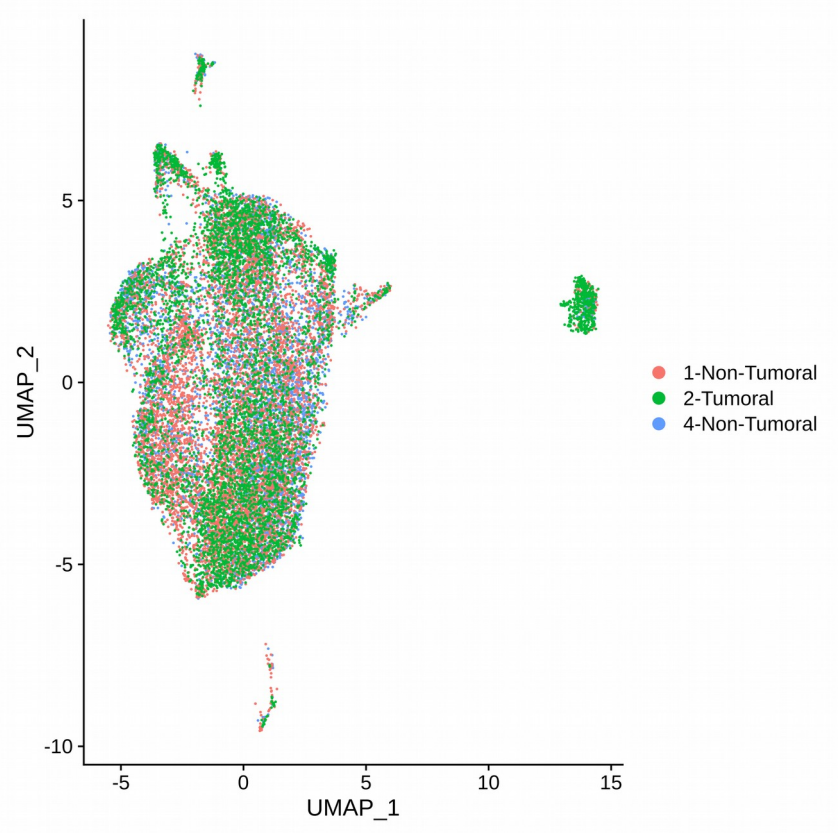






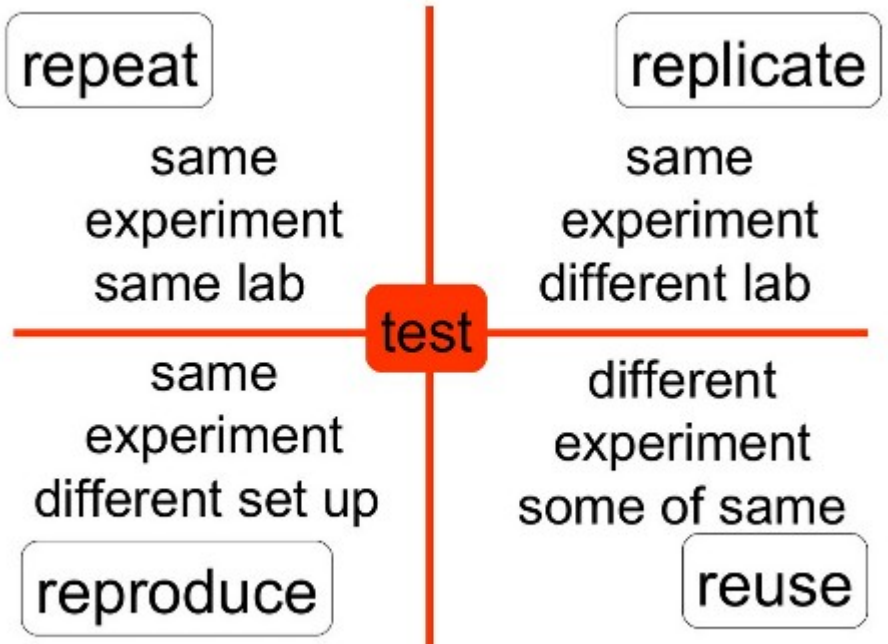
NGS20-007 : Before and After integration (Seurat3)





Plateforme GENOM'IC

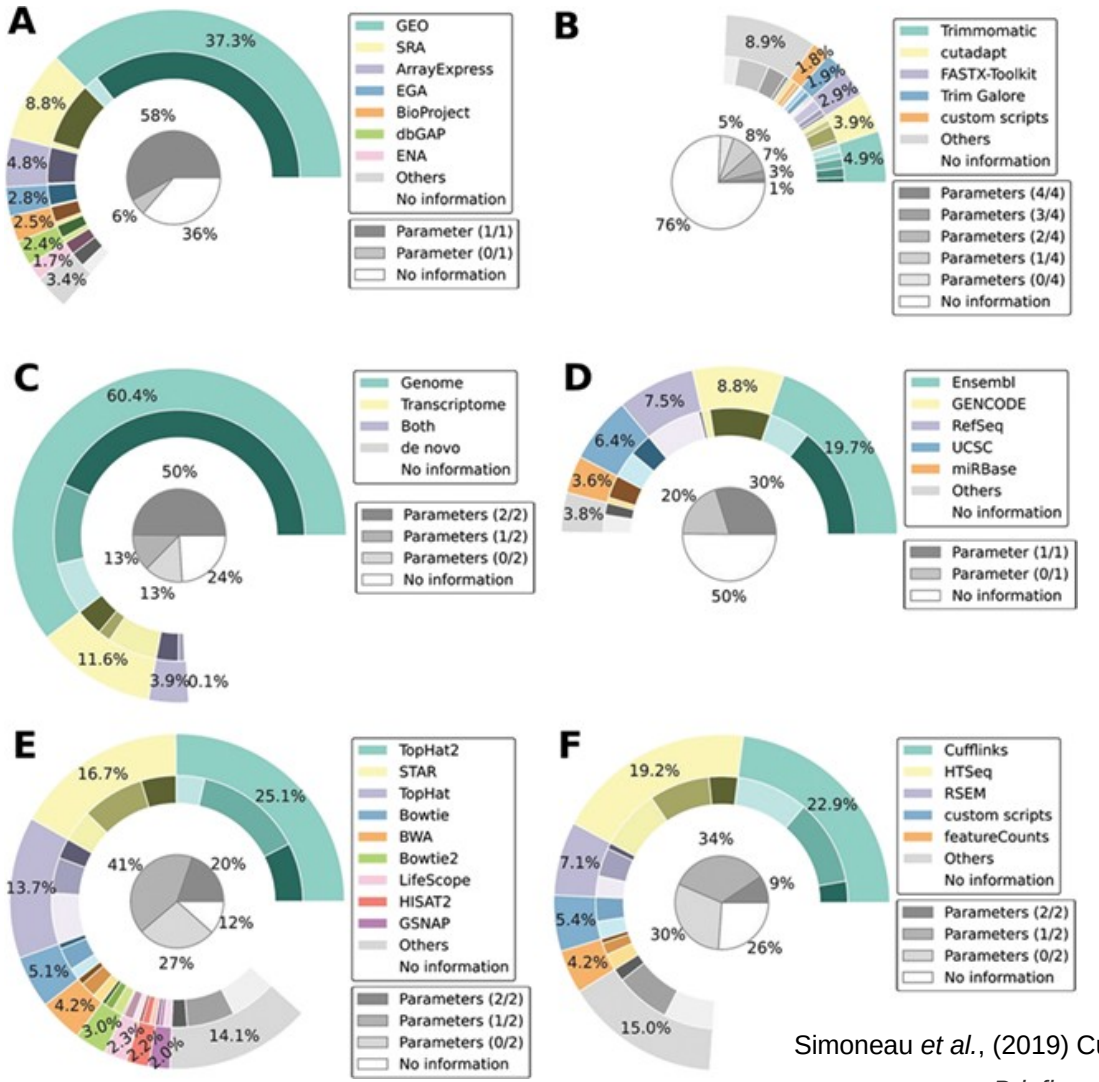
A word on reproducibility



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

- Script & Command-line
  - Github, for example
  - Readme & Comments
- Data in raw format
  - Fastq, or at least count files (for RNA-seq experiment)
- Versions
  - Docker, conda
  - Targz archives

A word on reproducibility



465 articles, information about the RNA-seq pipeline, from FASTQ files to gene or isoform count matrices

which tools & parameters ?

Simoneau *et al.*, (2019) Current RNA-seq methodology reporting limits reproducibility

Briefings in Bioinformatics, bbz124, https://doi.org/10.1093/bib/bbz124

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

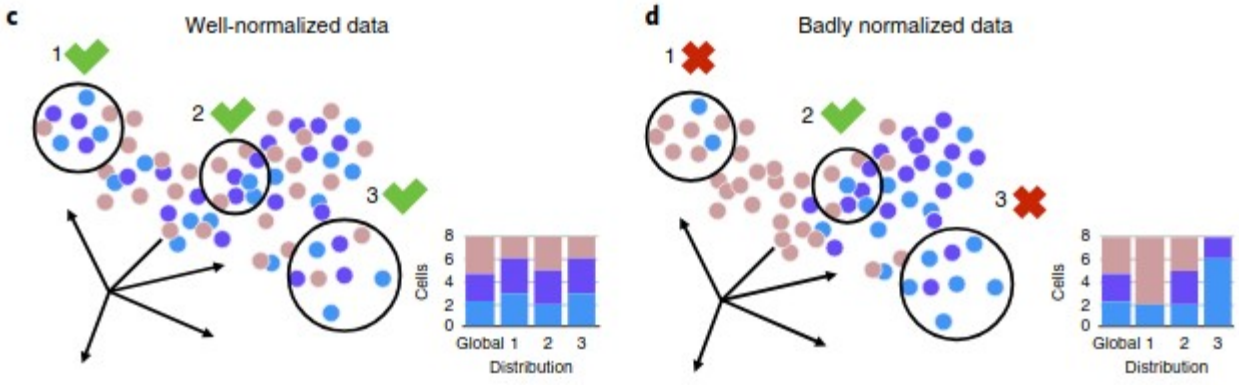
Thanks for your attention

GENOM'IC
PLATEFORME GÉNOMIQUE INSTITUT COCHIN

Additional files

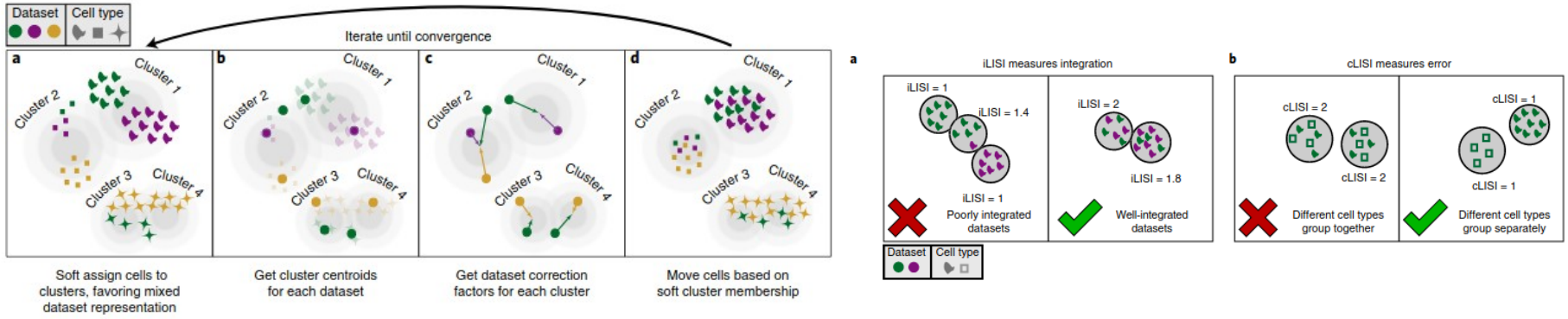
5 evaluation metrics

k-nearest-neighbor batch-effect test (kBET)



*Büttner, M., Miao, Z., Wolf, F.A. et al. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods **16,** 43–49 (2019).* https://doi.org/10.1038/s41592-018-0254-1

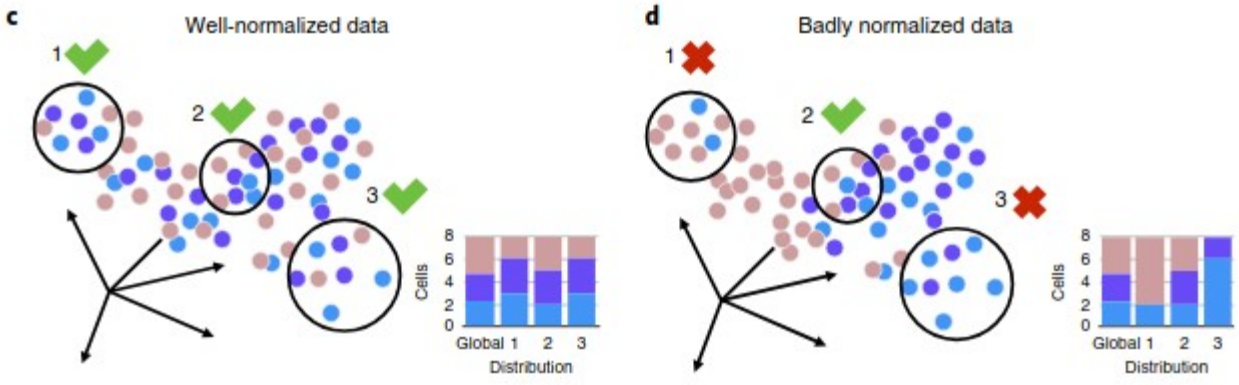
local inverse Simpson's index (LISI)



*Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods **16,** 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0*

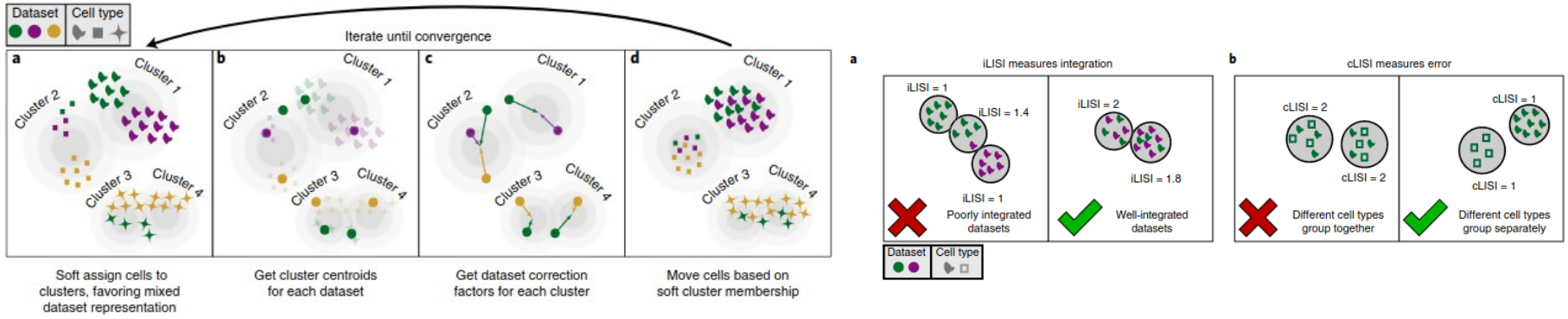
5 evaluation metrics

k-nearest-neighbor batch-effect test (kBET)



*Büttner, M., Miao, Z., Wolf, F.A. et al. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods **16,** 43–49 (2019).* https://doi.org/10.1038/s41592-018-0254-1

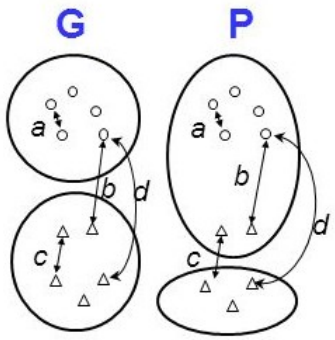
local inverse Simpson's index (LISI)



*Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods **16,** 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0*

5 evaluation metrics

Adjusted Rand Index (ARI)



Agreement: a, d
Disagreement: b, c

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

average silhouette width (ASW)

Differentially expressed genes (DEG)