scRNAseq2021

Data integration and batch correction

Paulo Czarnewski, **ELIXIR-Sweden (NBIS)**

Ahmed Mahfouz, **ELIXIR-Netherlands**

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

Why integrate?

Sources of variance in SC-RNAseq data

Biological:

- Cell Type Heterogeneity
- Genetics
- Cell State/Microenvir.
- GExpr Stochasticity
- Cell Cycle Dynamics
- Transcriptional Bursts
- Oscillations

...

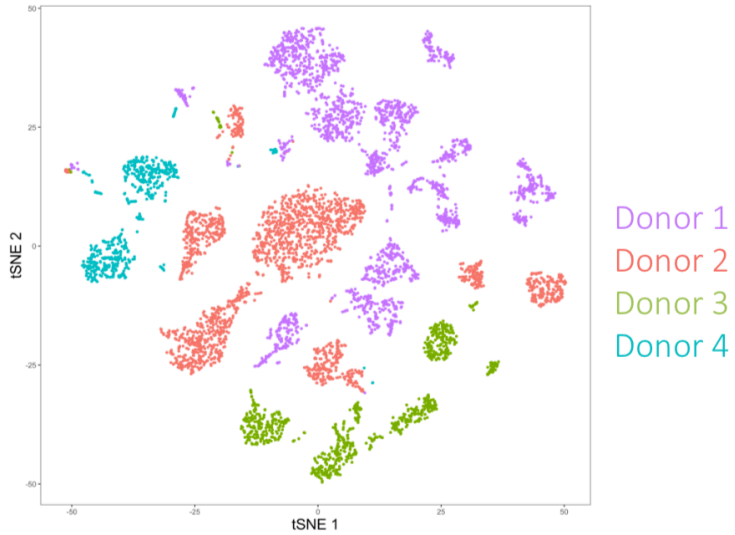
Technical:

- Capture Efficiency
- Amplification Bias
- PCR artifacts
- Contamination
- Cell Doublets
- Cell Damage
- Sampling (Jackpot Effects)

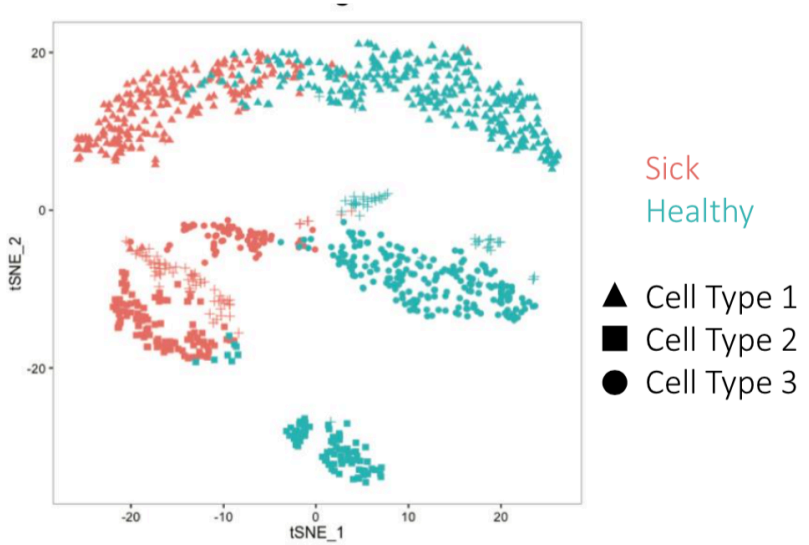
...

Why integrate?

Sources of variance in SC-RNAseq data



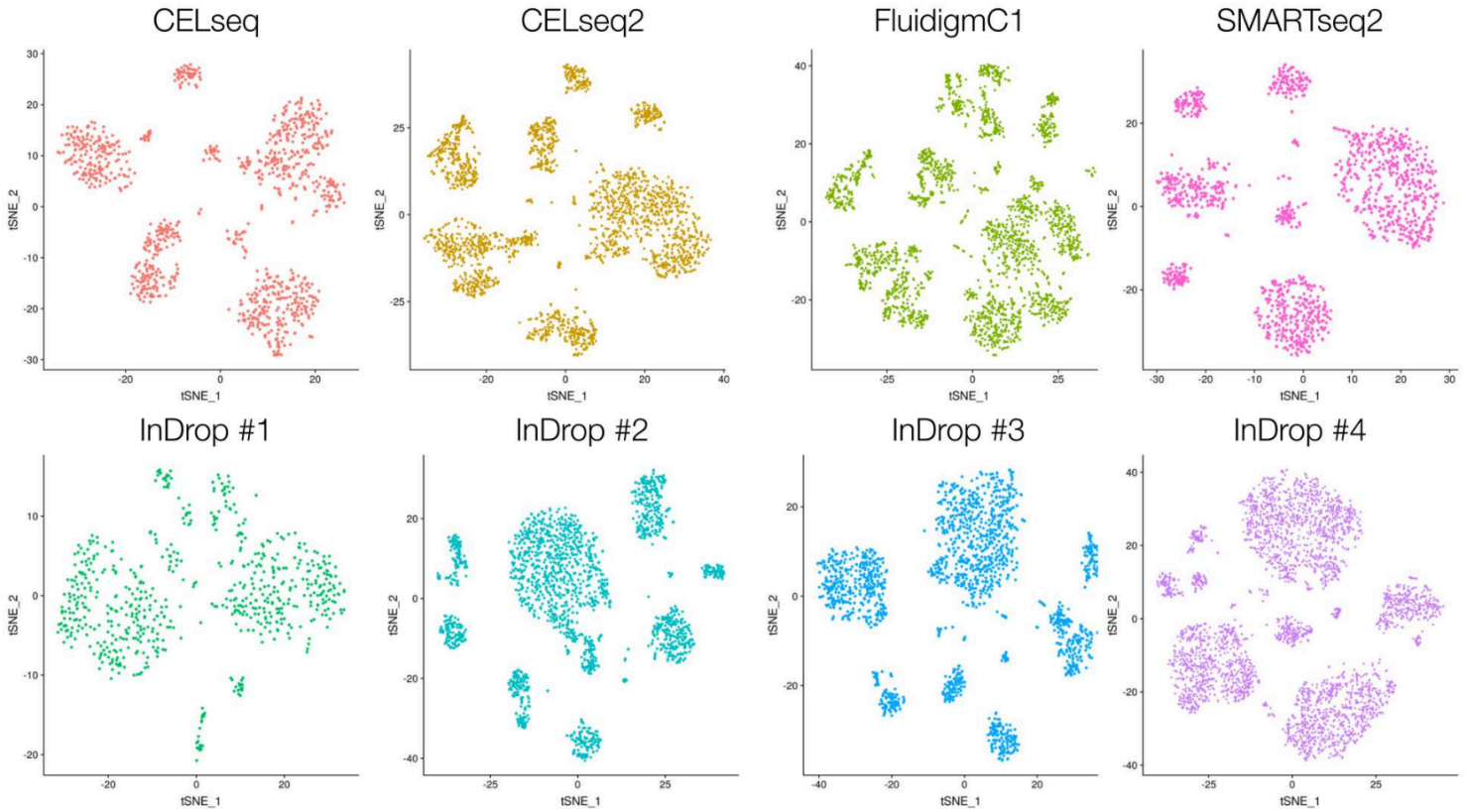
Same tissue from different donors



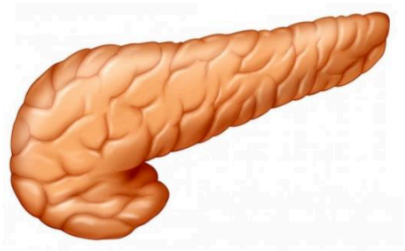
Cross condition comparisons

Why integrate?

Building a cell atlas: 8 maps of the human pancreas



Human pancreas



Baron et al. 2016, *Cell Syst.*

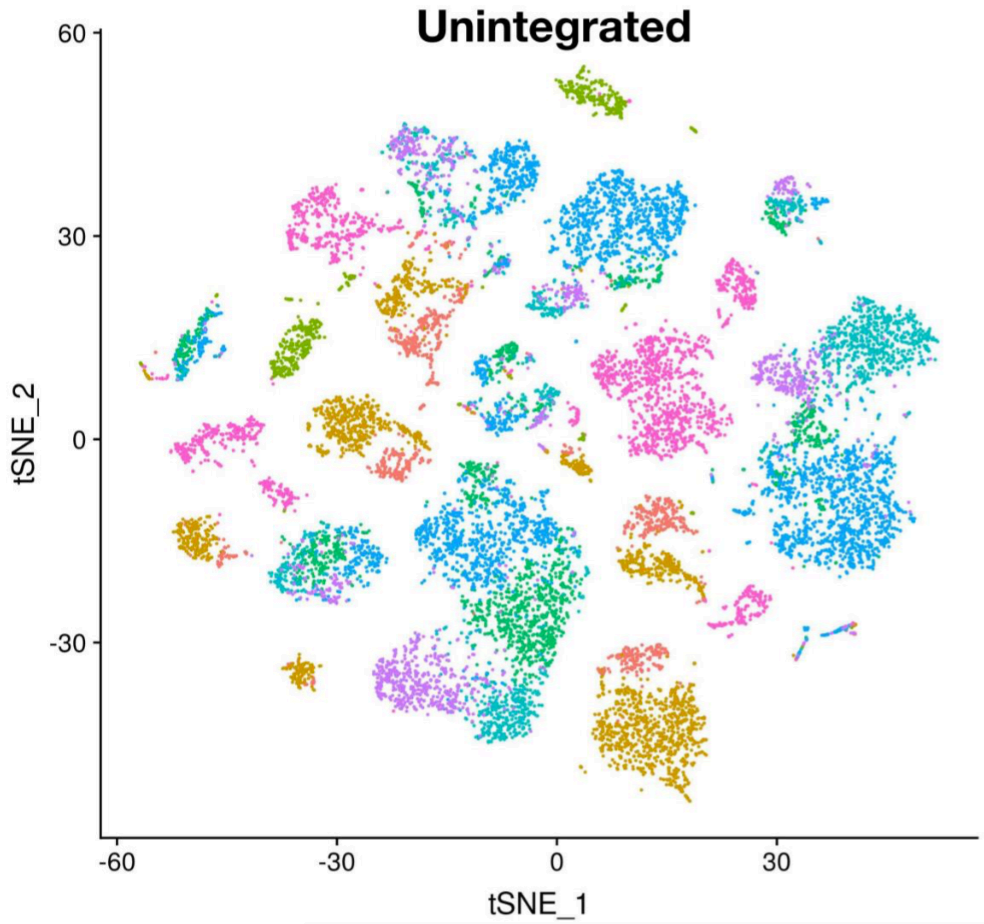
Lawlor et al. 2017, *Genome Res.*

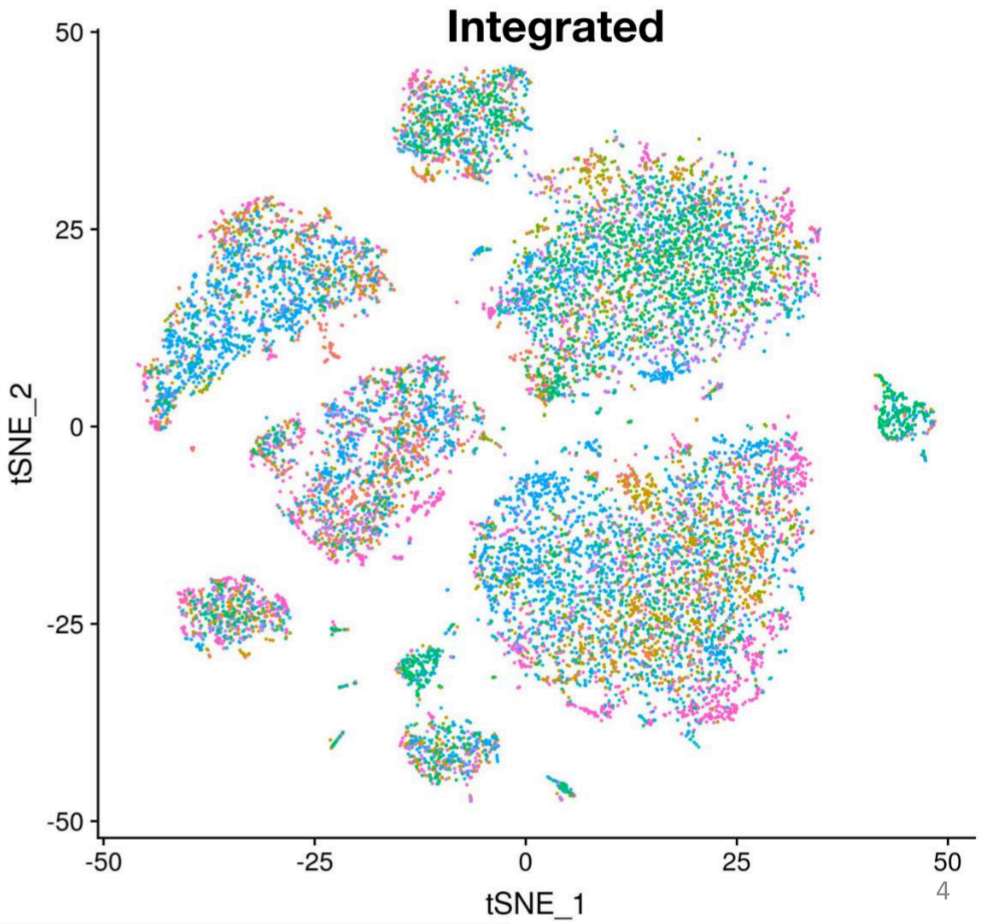
Grun et al. 2016, *Cell Stem Cell*

Muraro et al. 2016, *Cell Syst.*

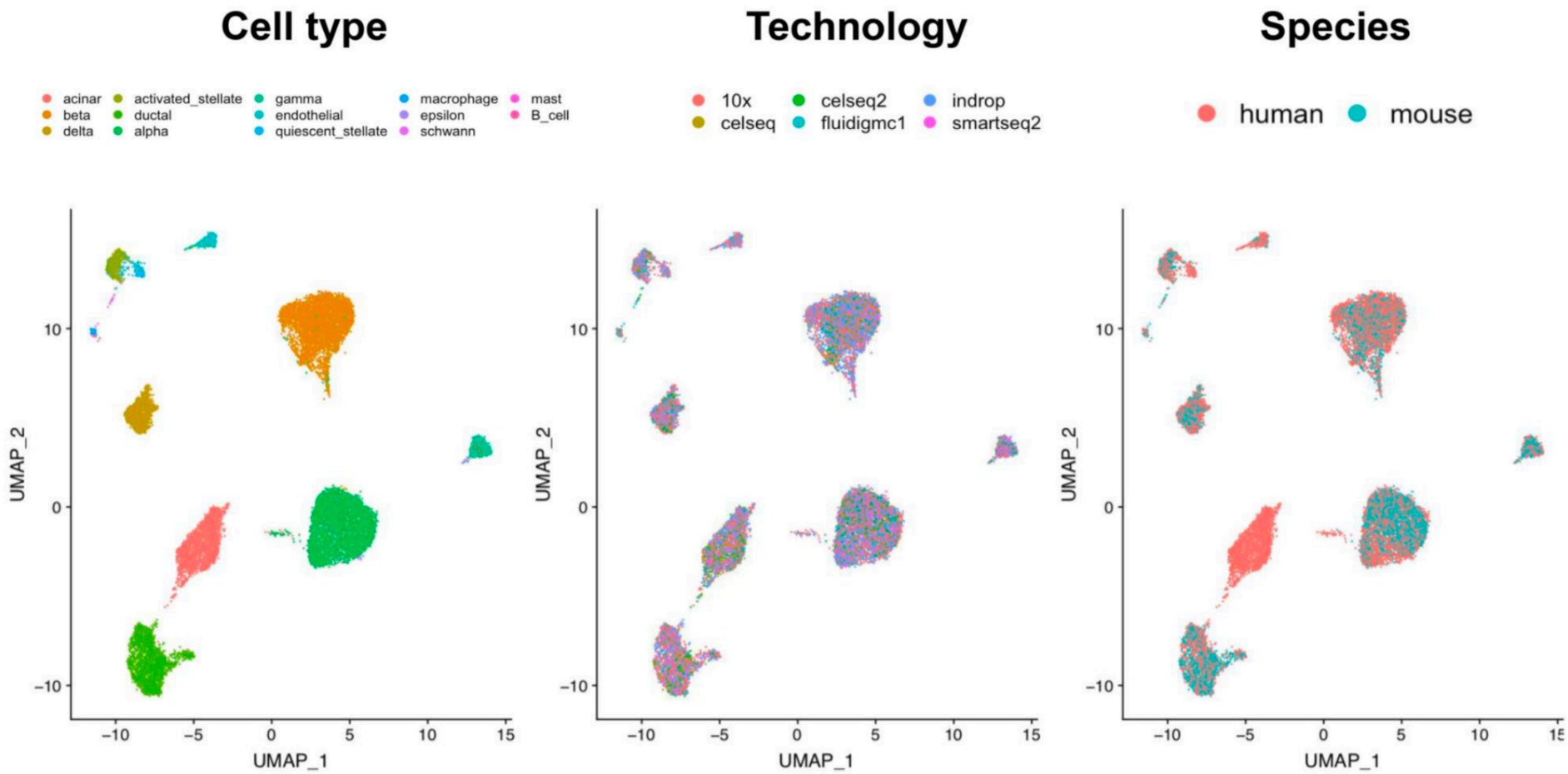
Why integrate?

Building a cell atlas: 8 maps of the human pancreas





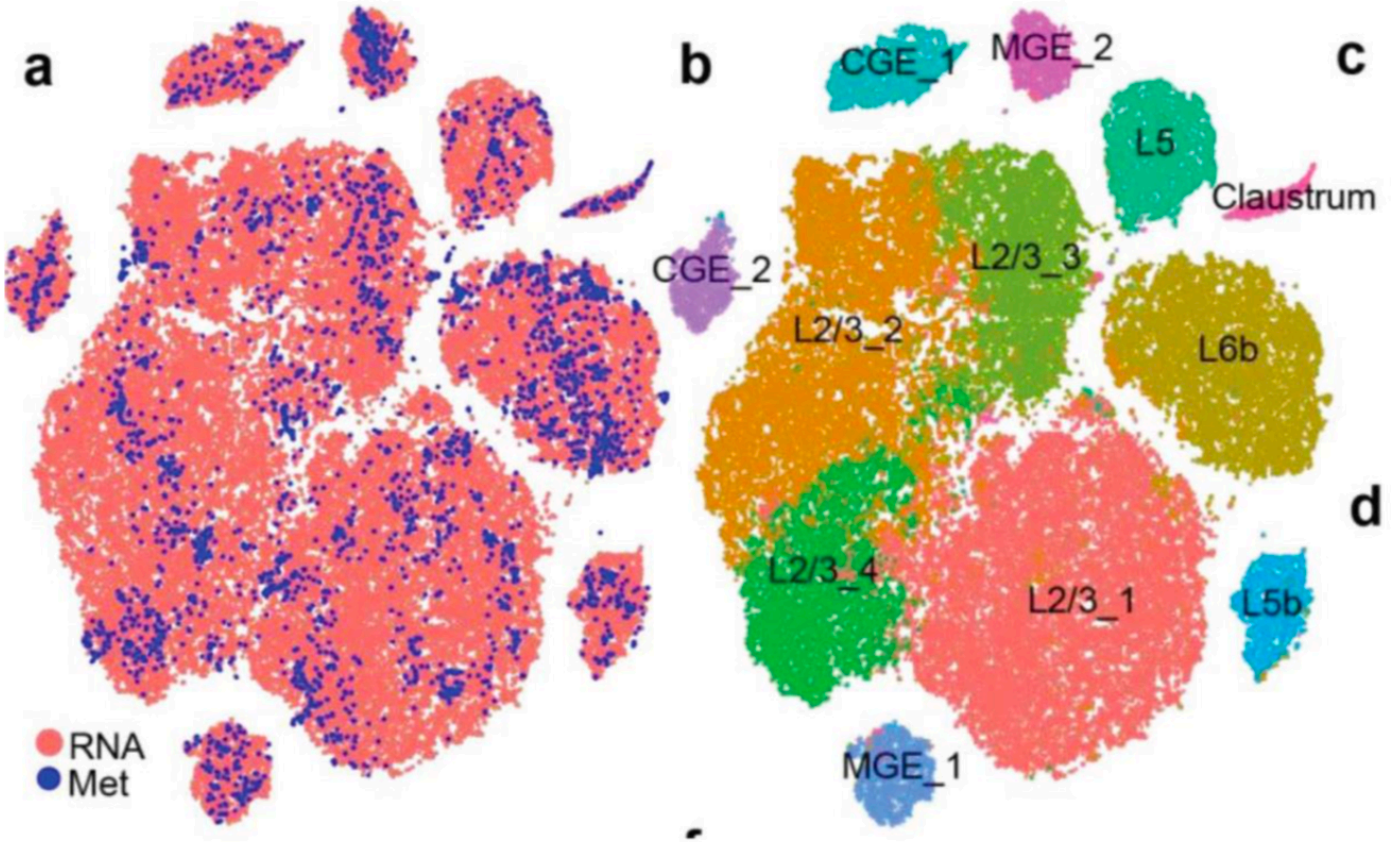
Integration across modalities



Retinal bipolar datasets: 51K cells, 6 technologies, 2 Species

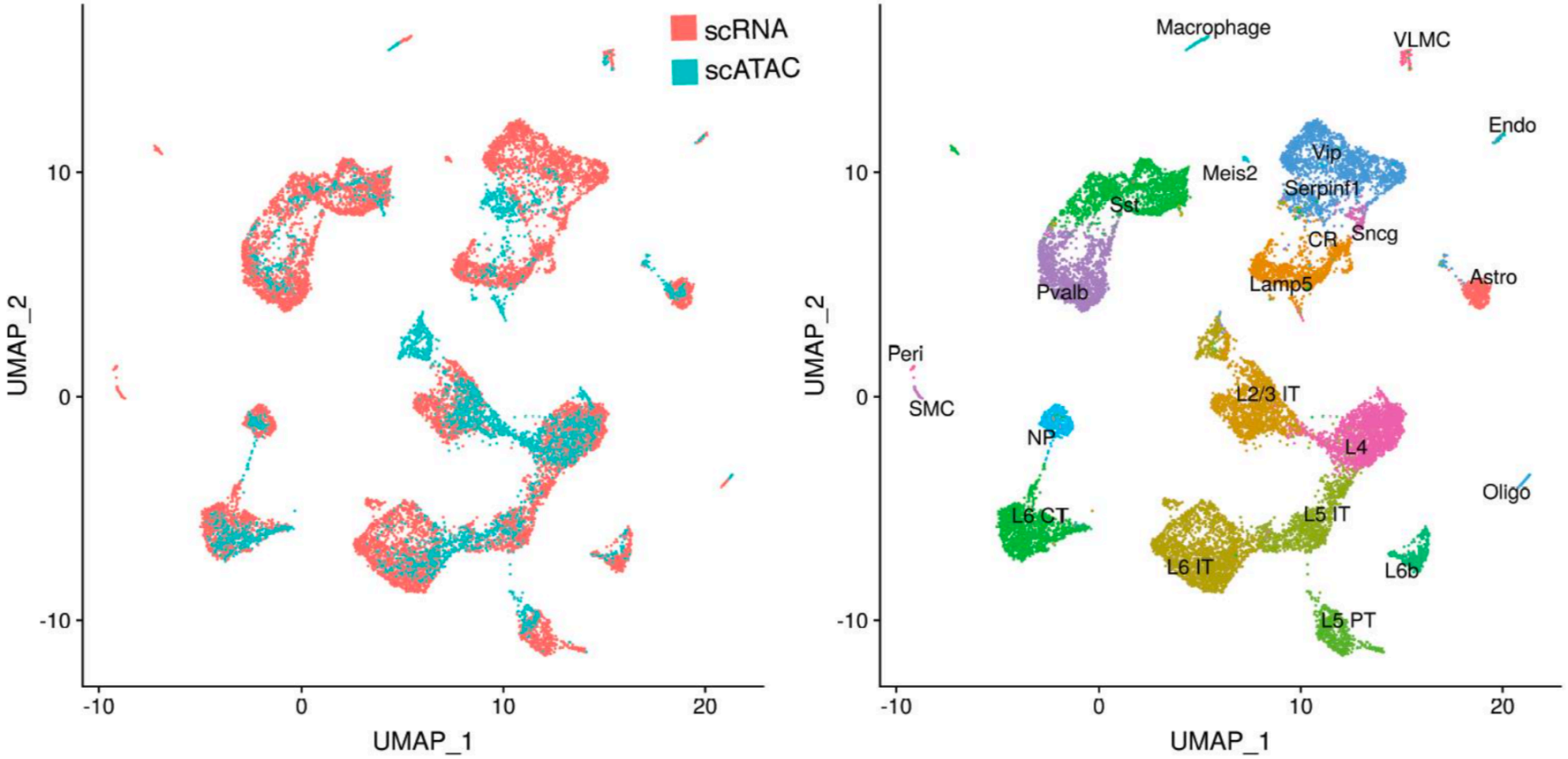
Integration across modalities

RNA-seq and methylation



Integration across modalities

RNA-seq and ATAC-seq

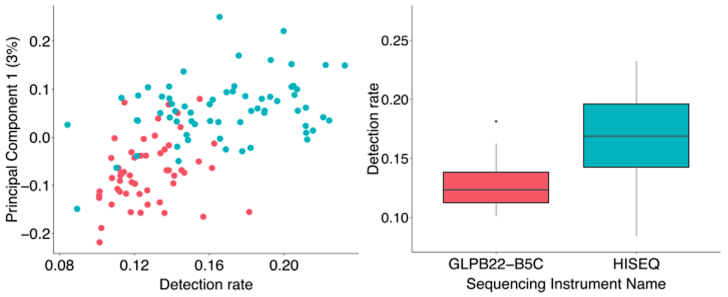


Confounders and batch effects

1. Technical variability

- Changes in sample quality/processing
- Library prep or sequencing technology
- 'Experimental reality'

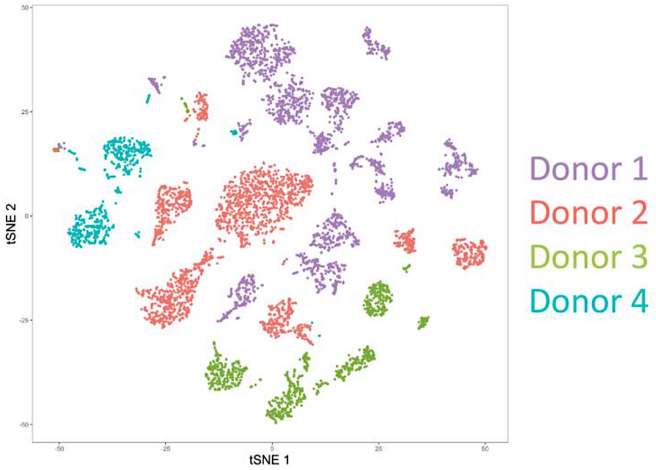
Technical '<u>batch effects</u>' confound downstream analysis



2. Biological variability

- Patient differences
- Environmental/genetic perturbation
- Evolution! (cross-species analysis)

Biological '<u>covariates</u>' confound comparisons of scRNA-seq data

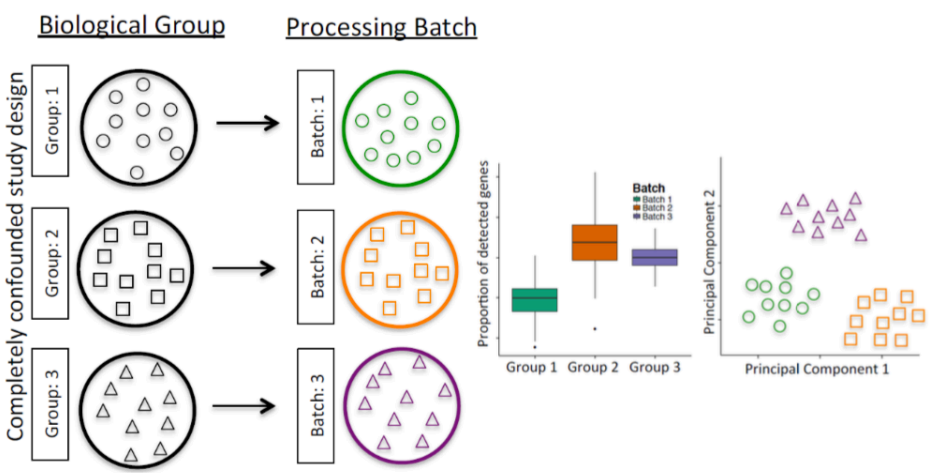


Shaham et al. (https://doi.org/10.1093/bioinformatics/btx196)

Confounders and batch effects

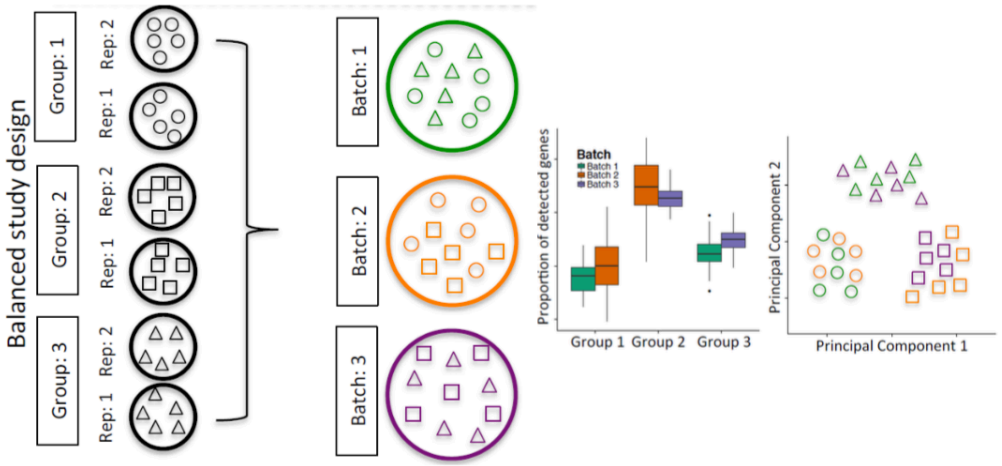
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN SciLifeLab

Confounded design



Don’t design your experiment like this!!!

Not confounded design



Good experimental design *does not remove batch effects*, it prevents them from biasing your results.

Hicks et al. (https://doi.org/10.1093/biostatistics/kxx053)

Integration methods overview

Batch correction methods

Regression–based correction:

- Regression via GLM
- ComBat (doi.org/10.1093/biostatistics/kxj037)
- RUVseq (10.1038/nbt.2931)

Joint dimensionality reduction:

- common PCA / CPCA (doi.org/10.1006/jmva.2000.1908)
- contrastive PCA / cPCA (https://doi.org/10.1038/s41467-018-04608-8)
- LIGER (https://doi.org/10.1101/459891)
- zinb-wave (10.1038/s41467-017-02554-5)
- scMerge (https://doi.org/10.1073/pnas.1820006116)
- Harmony (https://doi.org/10.1101/461954)

Graph-based joint clustering:

- MNNcorrect (https://doi.org/10.1038/nbt.4091)
- Conos (https://doi.org/10.1101/460246)

Joint dimensionality reduction + Graph-based joint clustering

- CCA + anchors (Seurat v3) (https://doi.org/10.1101/460147)
- CCA + dynamic time warping (Seurat v2) (https://doi.org/10.1038/nbt.4096)
- Scanorama (https://doi.org/10.1101/371179)
- fastMMN (https://doi.org/10.1038/nbt.4091)

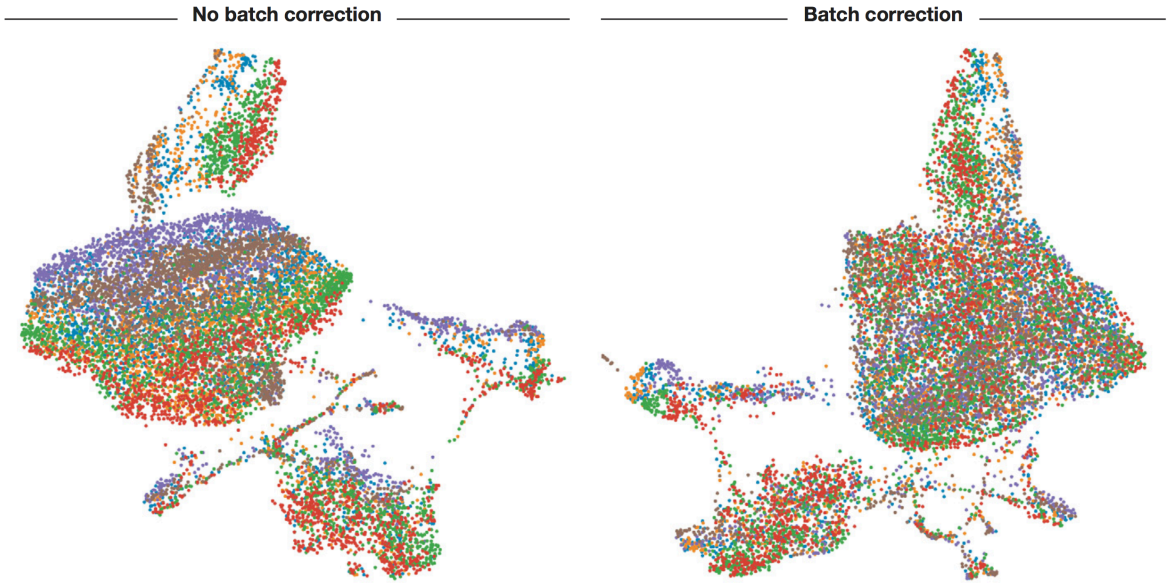
And many many others ….

ComBat

ComBat

- Uses empirical Bayes regression on shared gene factors
- Works well on simpler small-medium datasets
- All datasets need to be similar in cell type composition

- Will fail in large datasets with complex mixture of cell type



Luecken, Theis (2019) Molecular Systems biology https://doi.org/10.15252/msb.20188746

ComBat

Major issues of regression-based batch correction methods:

- limma::**removeBatchEffect**()
- seurat::**ScaleData**() #using the regression parameter
- sva::**combat**()
- batchelor::**rescaleBatches**()

1. Do not account for differences in population composition

2. Assume batch effect is additive

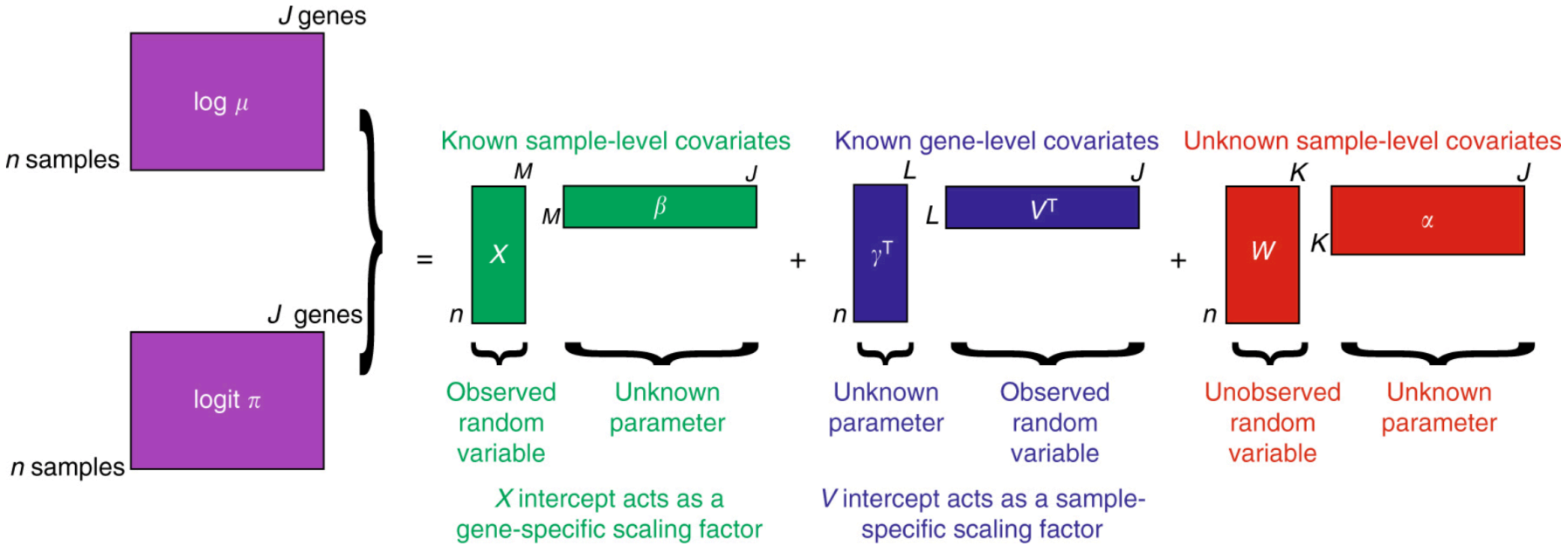
3. Prone to overcorrection (in cases of partial confounding)

zinb-wave

Zimwave

- Applies a matrix factorization model to accommodate both gene and cell covariates
- Similar to ZIFA (zero-inflated factor analysis)
- **Works well on simpler small-medium datasets**

- **It will be slow on large datasets**



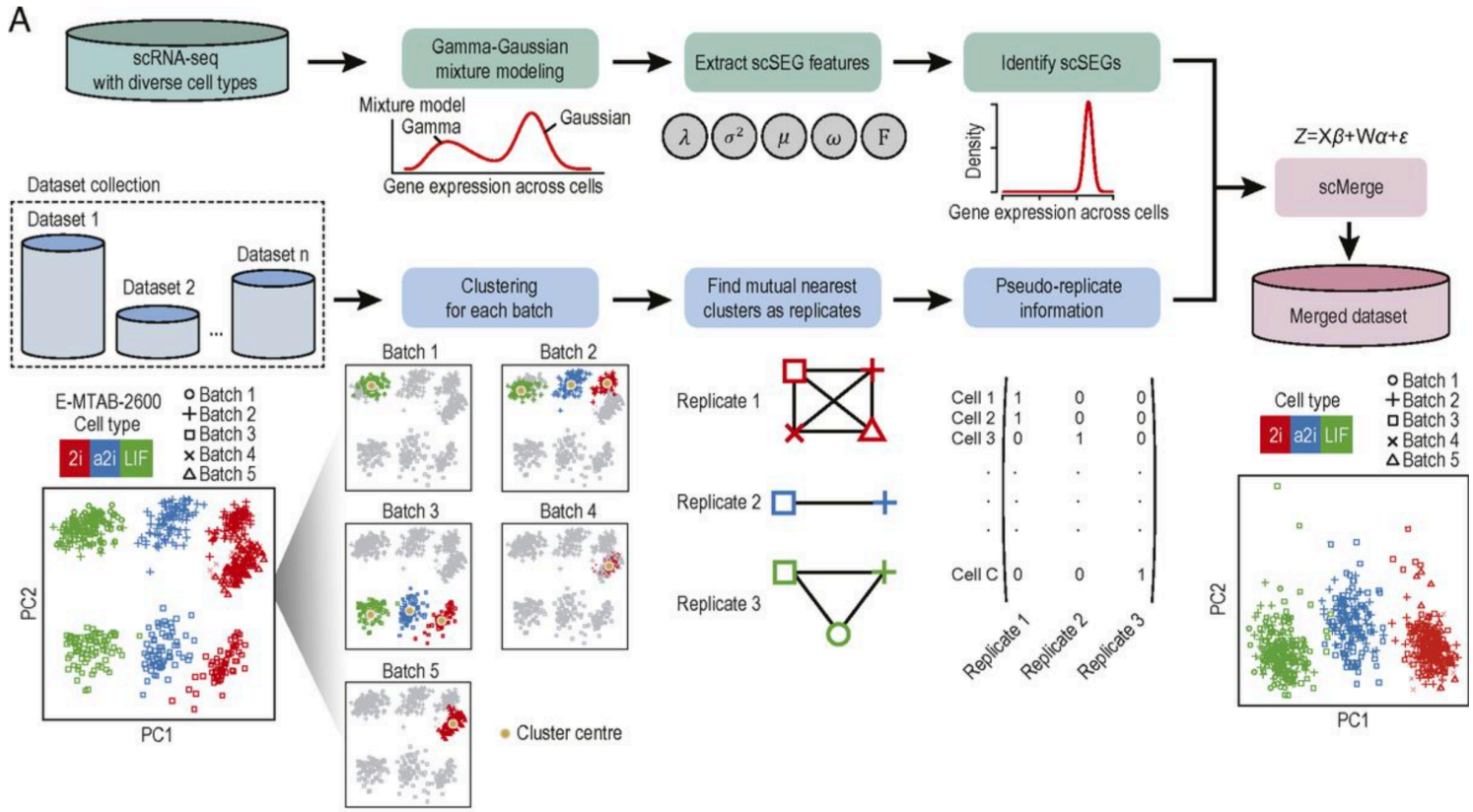
Risso et al (2018) Nat Communicaiton

scMerge

scMerge

- Identifies single-cell stably expressed genes (scSEGs)
- Uses a fast implementation of RUV-seq to scale other genes based on the scSEGs
- **Works well on simpler small-medium datasets**

- **It will be slow on large datasets**



Lin et al (2019) PNAS https://doi.org/10.1073/pnas.1820006116

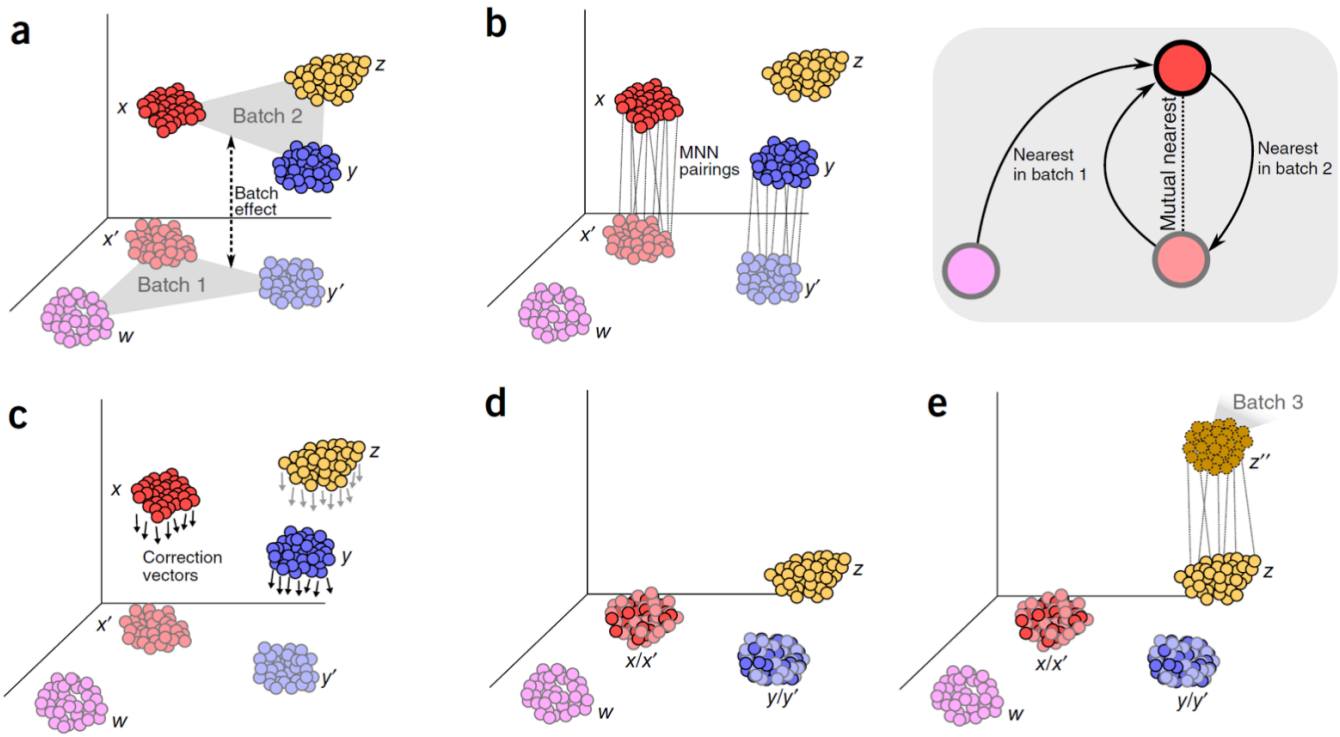
Mutual Nearest Neighbors (MNN)

THE “turning point” method

Mutual Nearest Neighbors (MNN)

- Dimensionality reduction via multibatch PCA with all datasets
- Find K-NN across datasets
- Compute merging vectors

- **It scales well on large datasets**



Haghverdi, L., Lun, A., Morgan, M. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36,** 421–427 (2018). https://doi.org/10.1038/nbt.4091

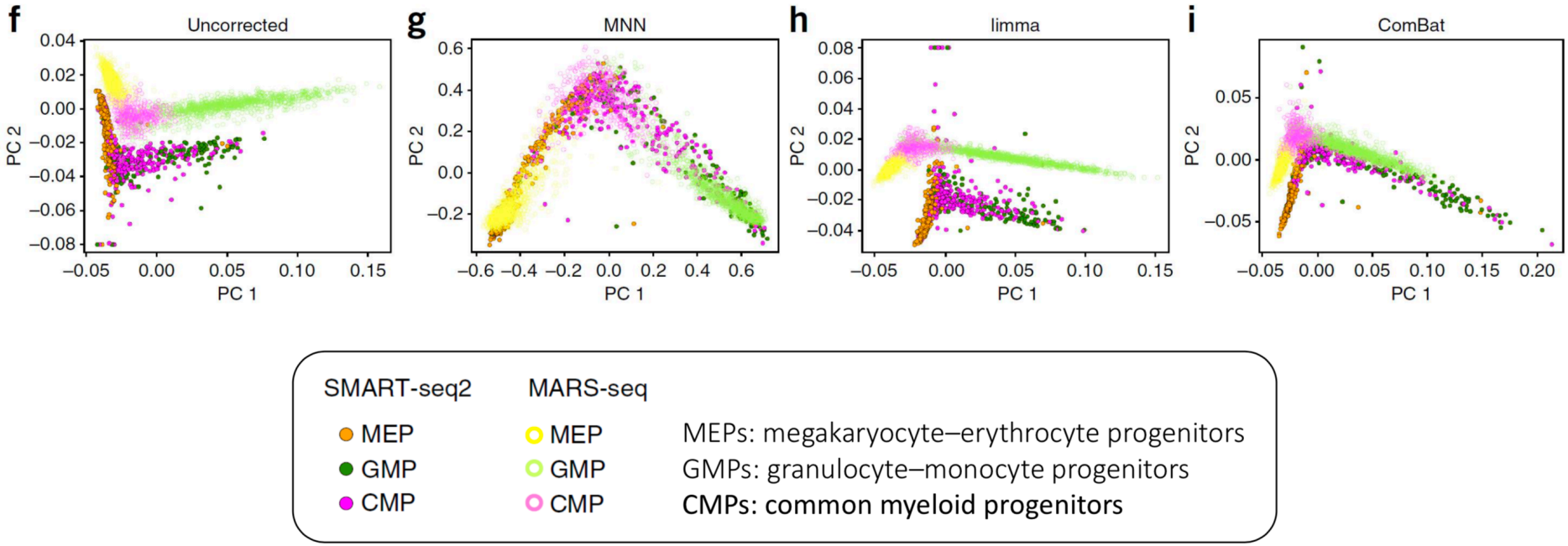
Mutual Nearest Neighbors (MNN)

Model assumptions

1. There is at least one cell population that is present in both batches,

2. The batch effect is almost orthogonal to the biological subspace, and

3. Batch effect variation is much smaller than the biological effect variation between different cell types



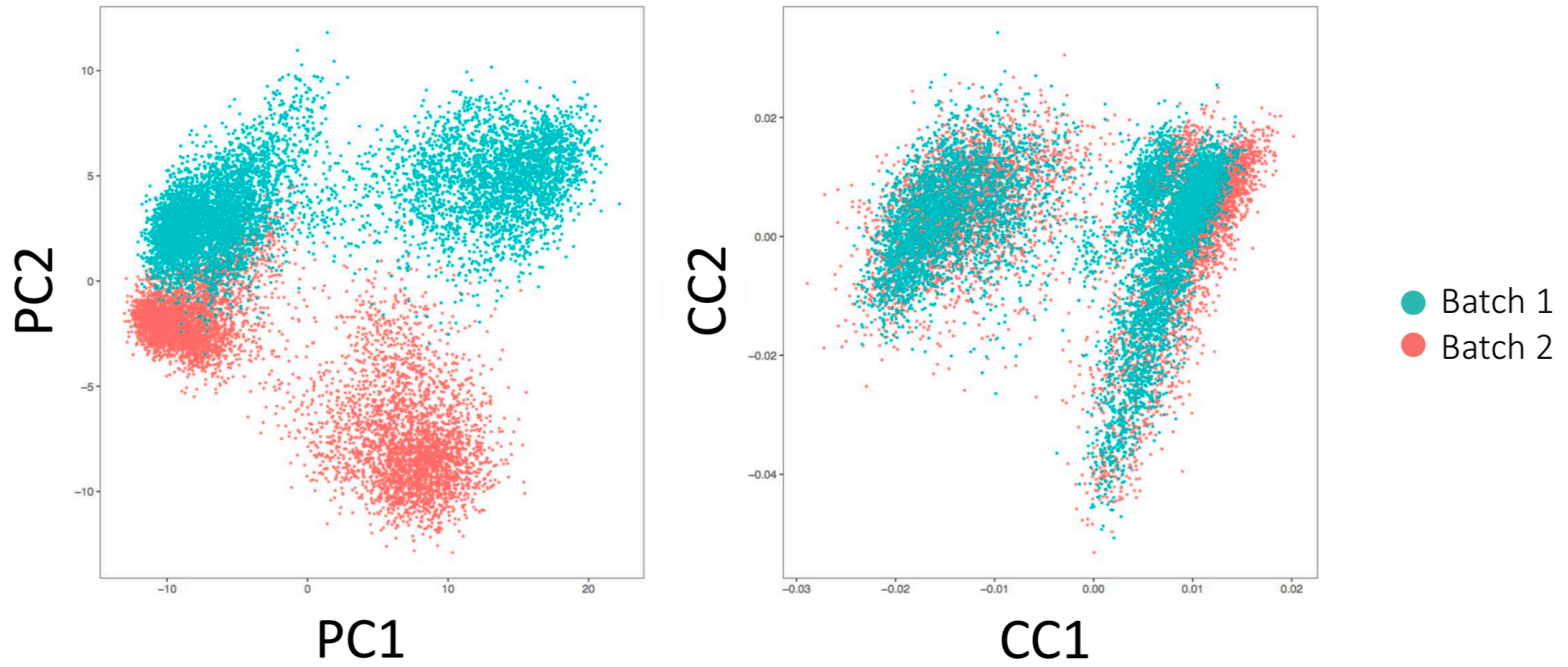
Haghverdi, L., Lun, A., Morgan, M. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36,** 421–427 (2018). https://doi.org/10.1038/nbt.4091

CCA + anchors (Seurat v3)

How CCA works?

Canonical correlation analysis

CCA captures correlated sources of variation between two datasets



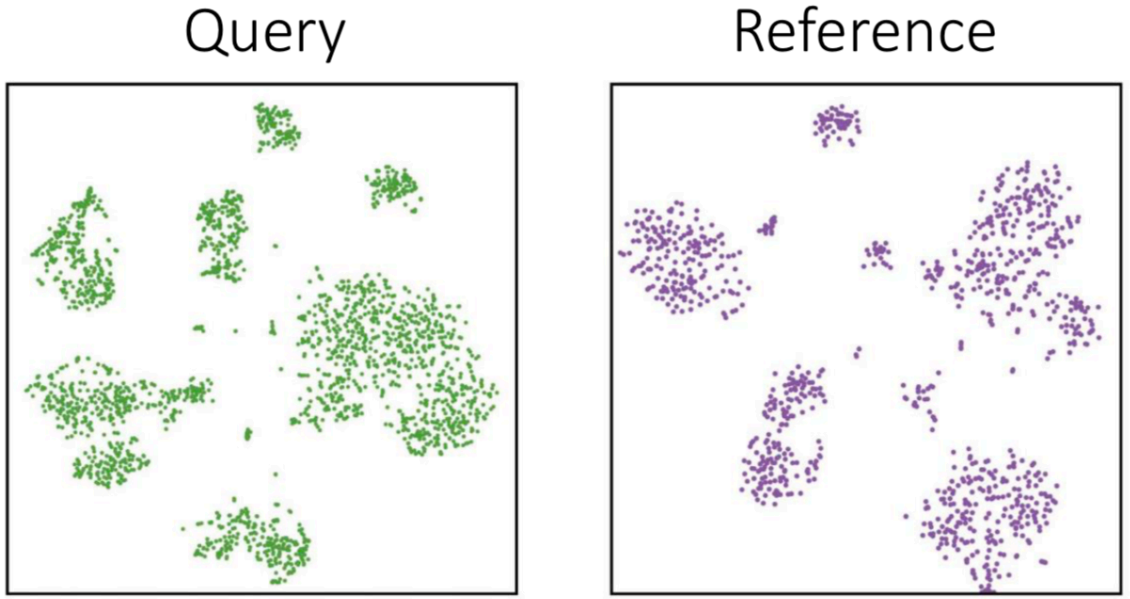
Werner et al. (https://doi.org/10.1371/journal.pone.0113083)

CCA + anchors (Seurat v3)

1. Find corresponding cells across datasets

2. Compute a data adjustment based on correspondences between cells

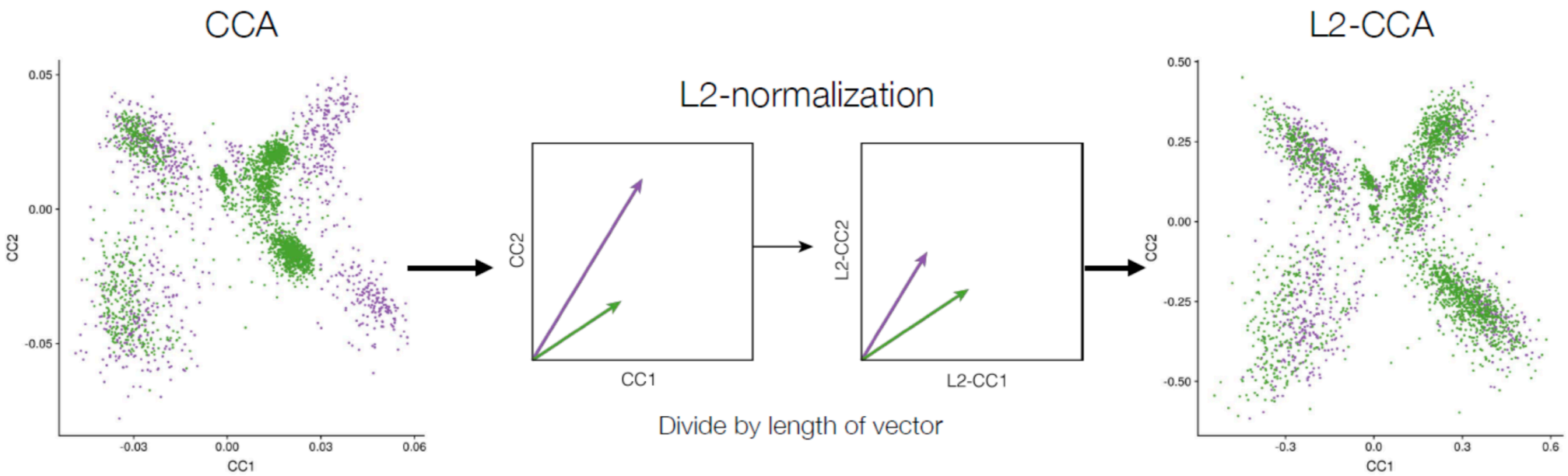
3. Apply the adjustment



Stuart et al (2019) Cell doi.org/10.1016/j.cell.2019.05.031

Mutual Nearest Neighbors (MNN)

L2-normalization corrects for differences in scale

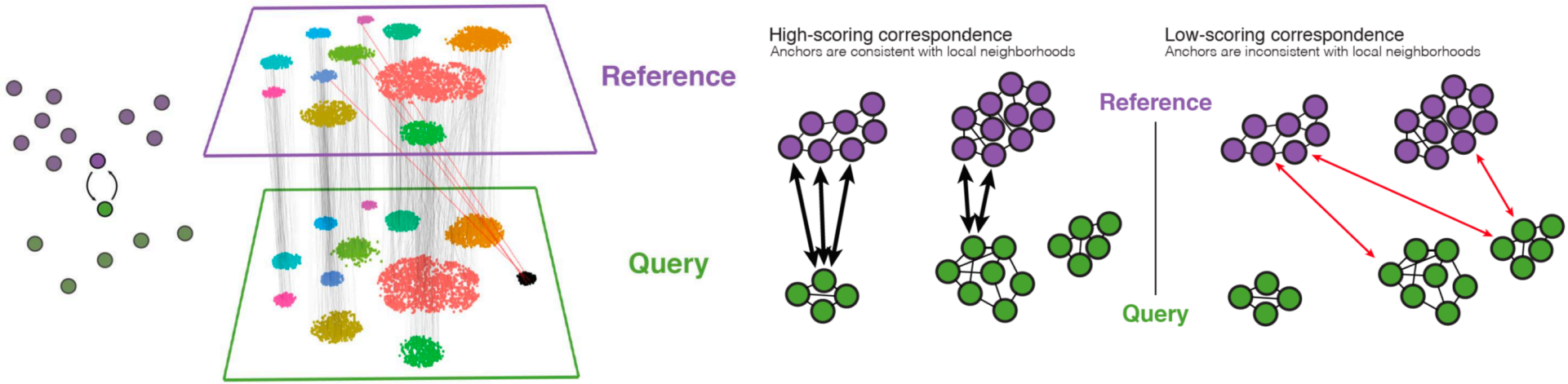


Stuart et al (2019) Cell doi.org/10.1016/j.cell.2019.05.031

Finding corresponding cells

Anchors: mutual nearest neighbours (MNN)

- **It scales well on large datasets**



Stuart et al (2019) Cell doi.org/10.1016/j.cell.2019.05.031

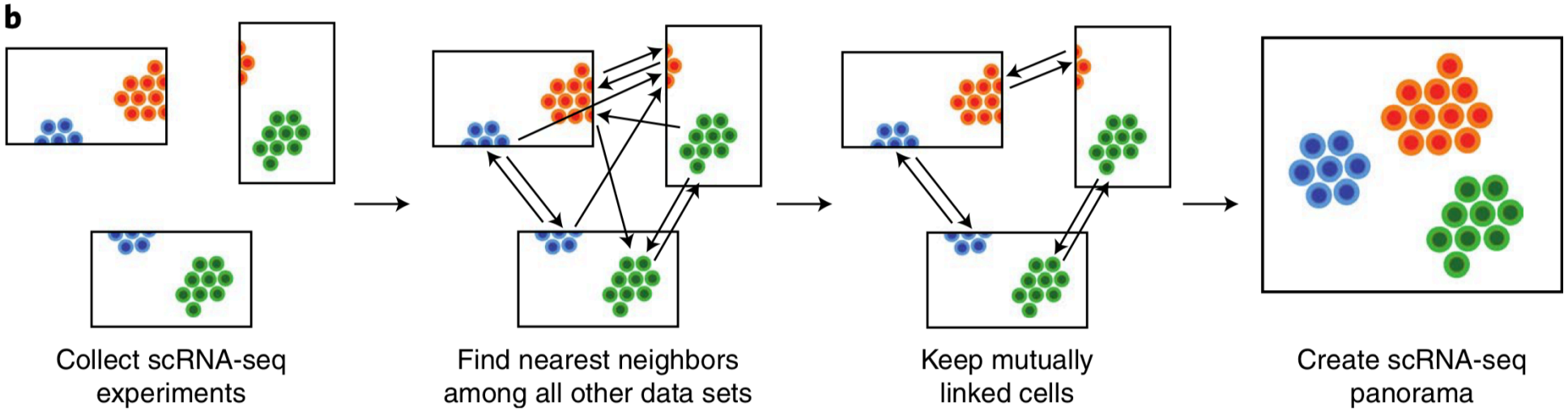
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN
SciLifeLab

Scanorama

Scanorama

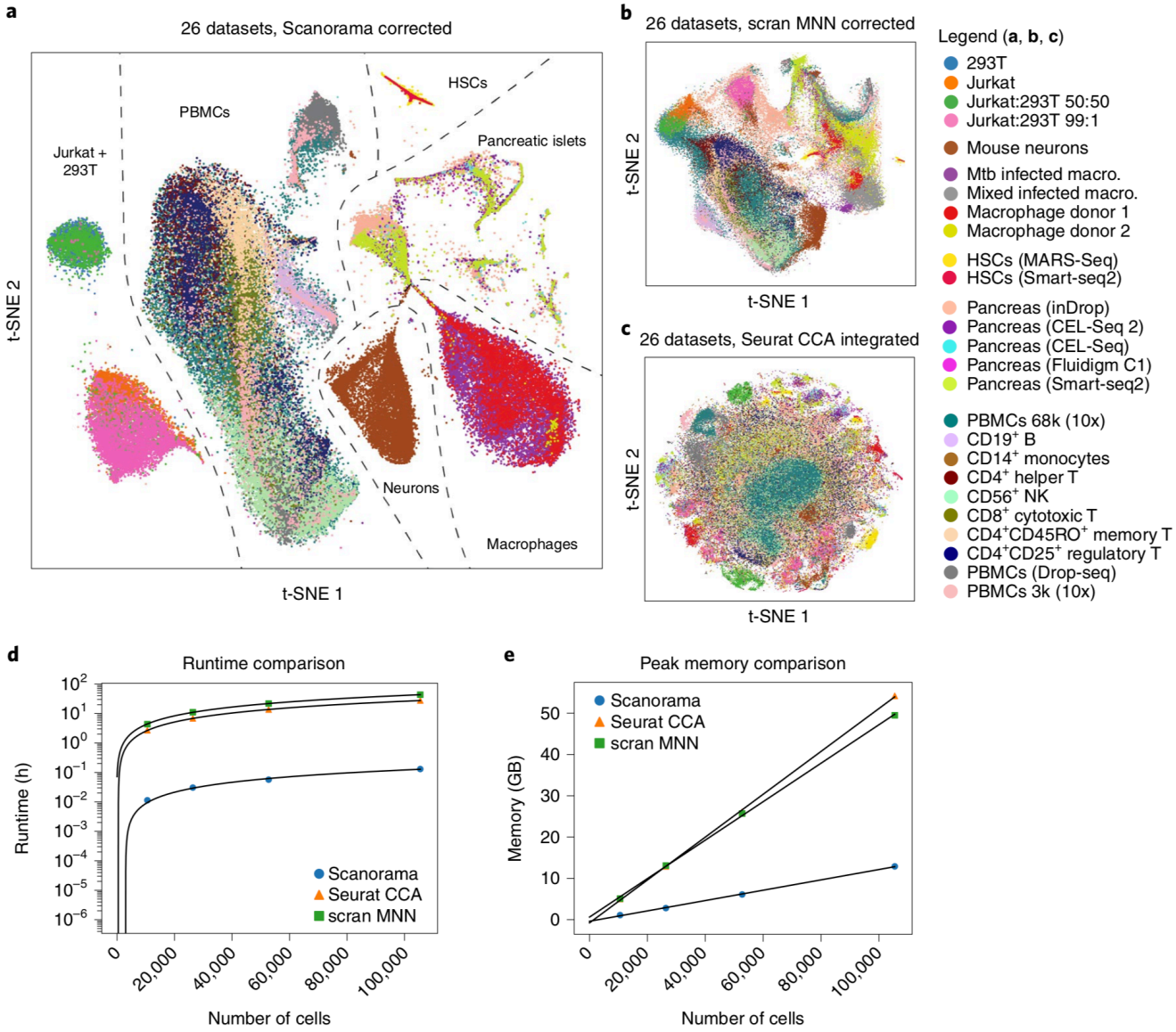
- Python implementation of fastMNN ?
  - Dimensionality reduction via SVD with all datasets
  - Find K-NN across datasets
  - Compute merging vectors

- **It scales well on large datasets**



Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37,** 685–691 (2019). https://doi.org/10.1038/s41587-019-0113-3

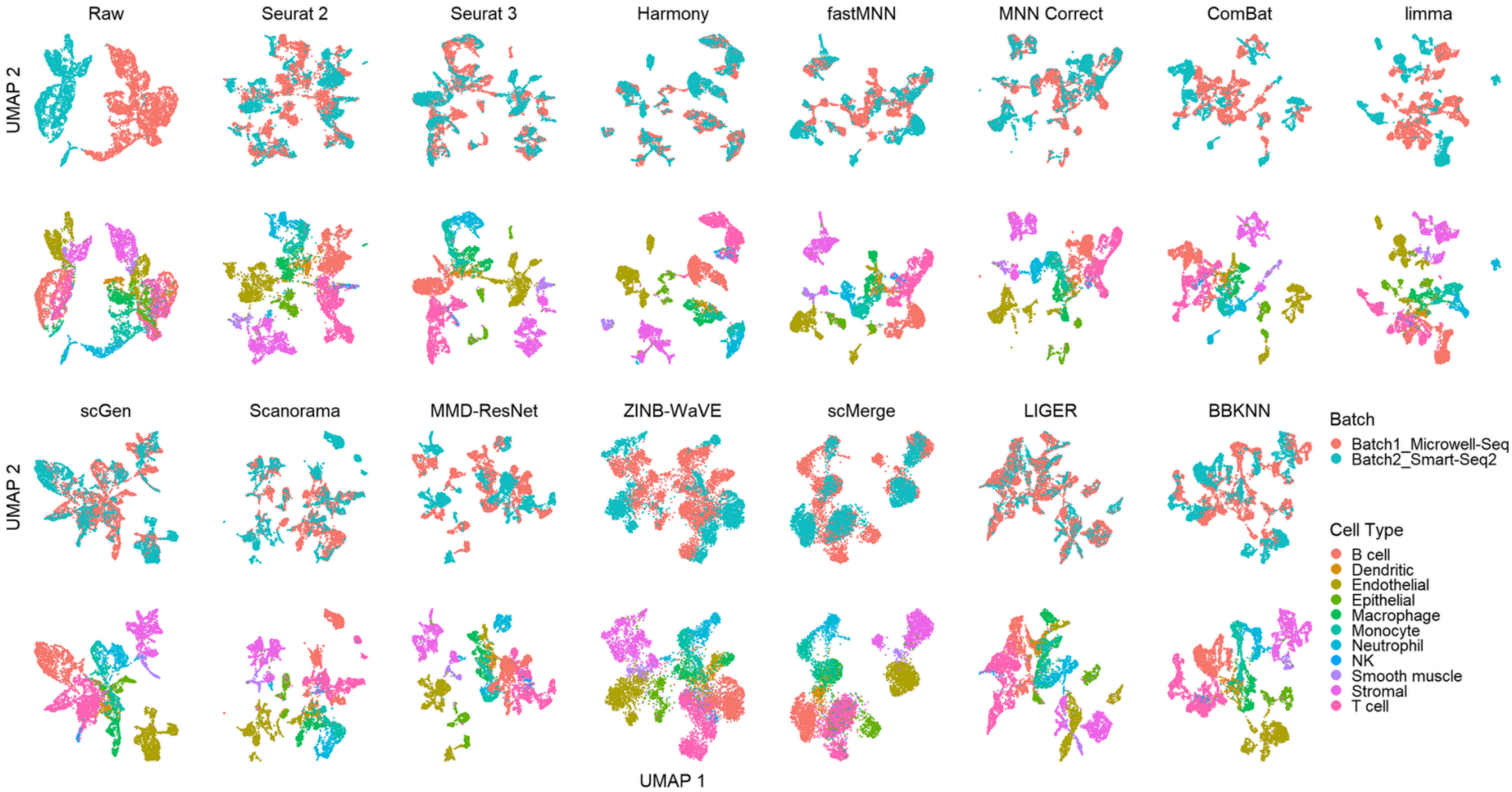
Scanorama



Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37,** 685–691 (2019). https://doi.org/10.1038/s41587-019-0113-3

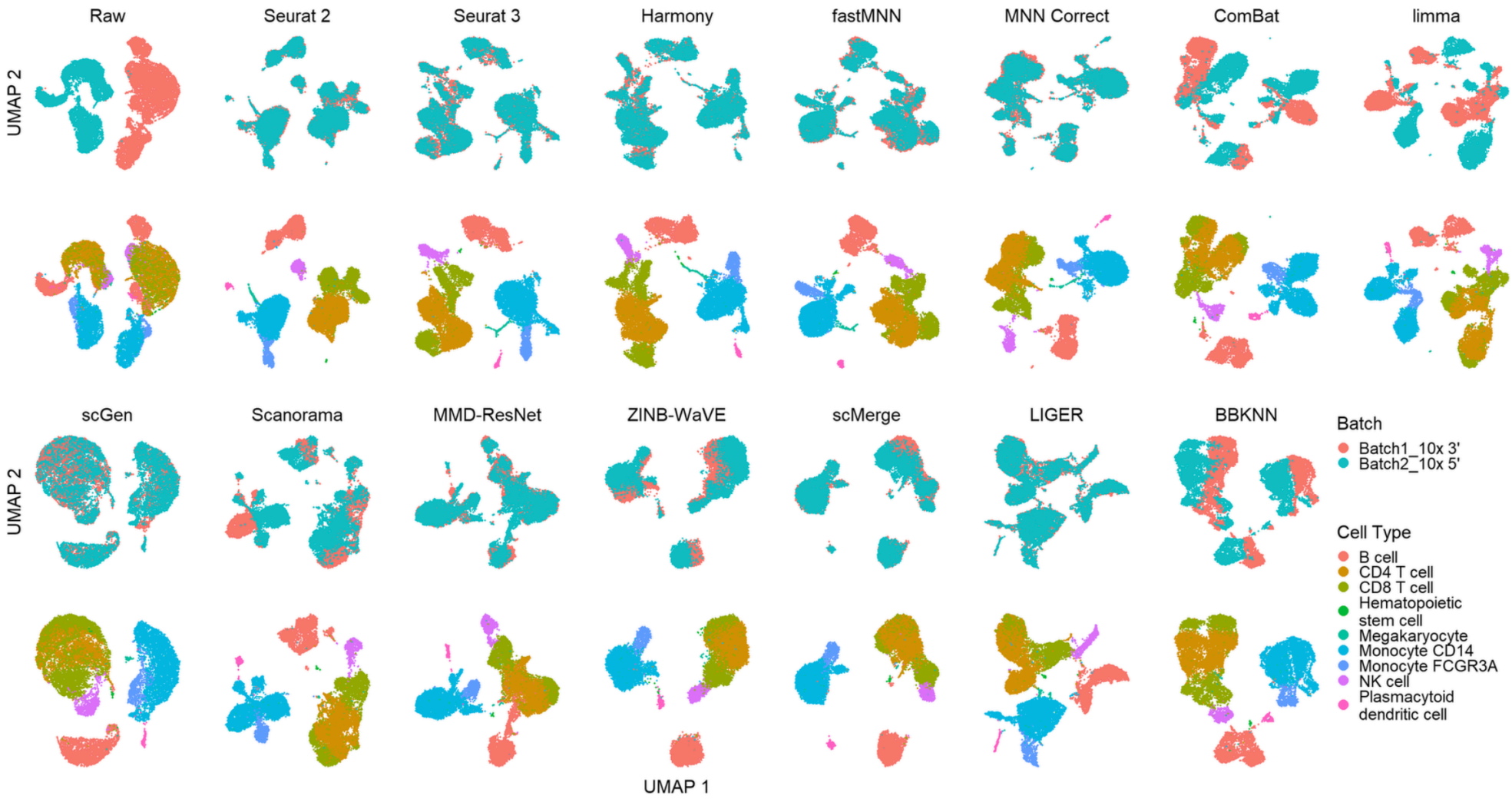
Evaluation of batch correction efficiency

Batch-correction performance assessment



Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21,** 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

Batch-correction performance assessment



Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21,** 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

Batch-correction performance assessment

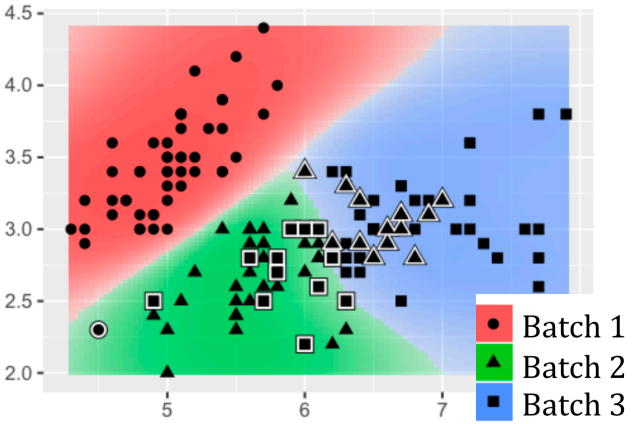
1. Evaluate mixing efficiency (Goal A)

- How well mixed are the obtained clusters post-batch correction?
- How well does a classifier (i.e. SVM) perform pre/post-correction?

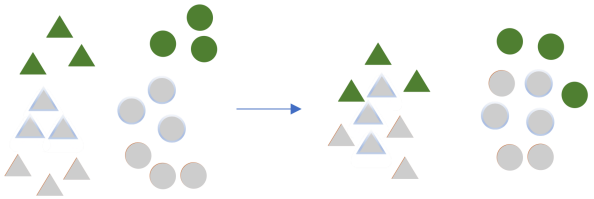
2. Evaluate preservation of remaining variance (Goals B, C)

- Evaluate proportion of removed variance, overlap of HVGs
- Evaluate preservation of within-batch cell topologies

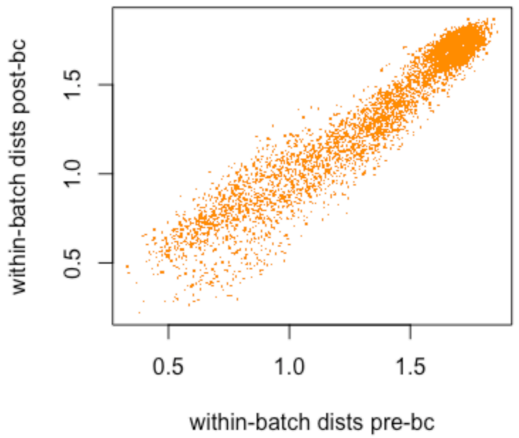
Classifiers must fail to find batches



Local structure preservation



Global structure preservation

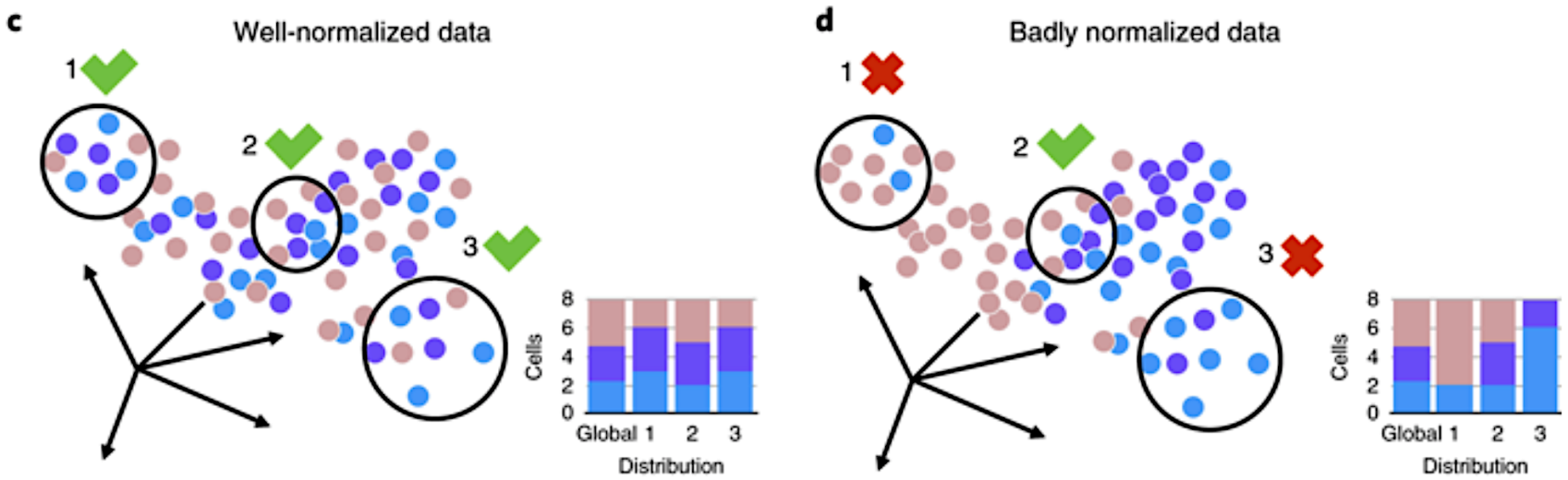


CellMixS: https://bioconductor.org/packages/release/bioc/manuals/CellMixS/man/CellMixS.pdf kBet: *Nature Methods* **volume 16**, pages43–49 (2019), https://github.com/theislab/kBET

Batch-correction performance assessment

We wish to obtain corrected data where the following goals are met:

Goal:	What it practically means:
1.The batch-originating variance is erased	Similar cell types are intermixed across batches
2.Meaningful heterogeneity is preserved within batches)	We are not mixing distinct cell types (across or
3.No artefactual variance is introduced	We do not separate similar cells within batches



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

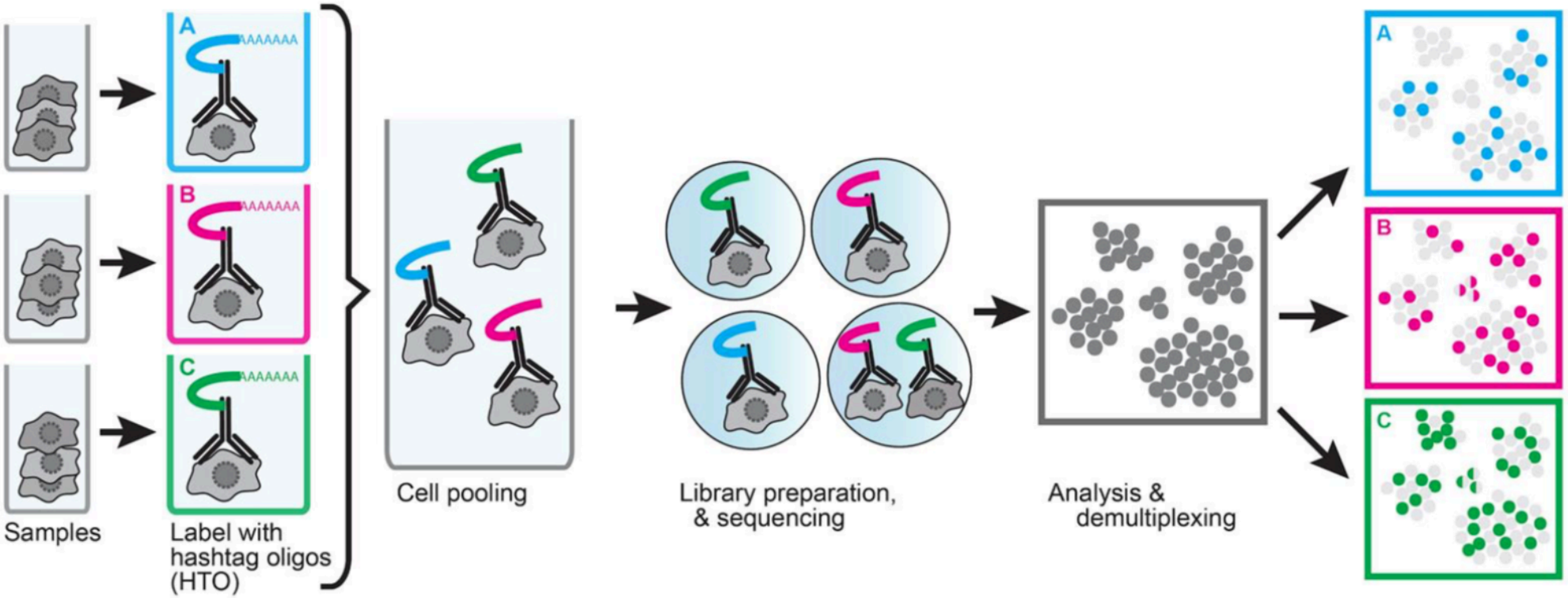
Summary

We wish to obtain corrected data where the following goals are met:

Goal:	**What it practically means:**
1.The batch-originating variance is erased	Similar cell types are intermixed across batches
2.Meaningful heterogeneity is preserved	We are not mixing distinct cell types (across or within batches)
3.No artefactual variance is introduced	We do not separate similar cells within batches

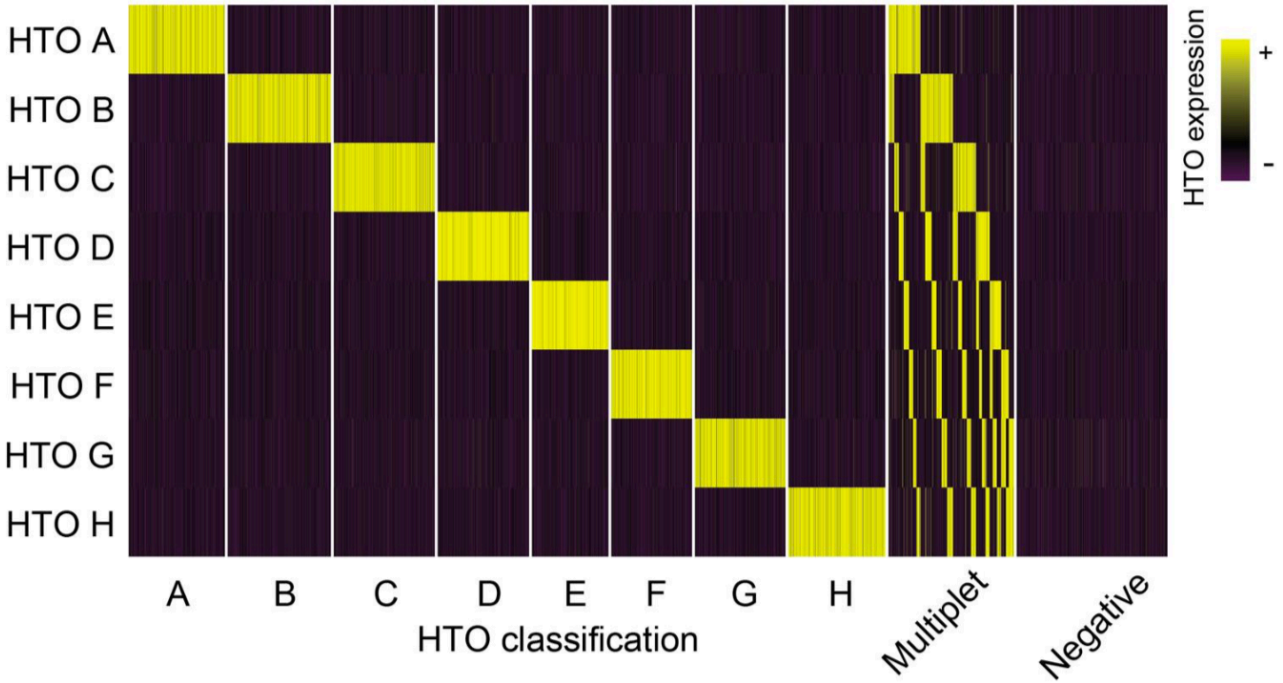
Avoiding batches

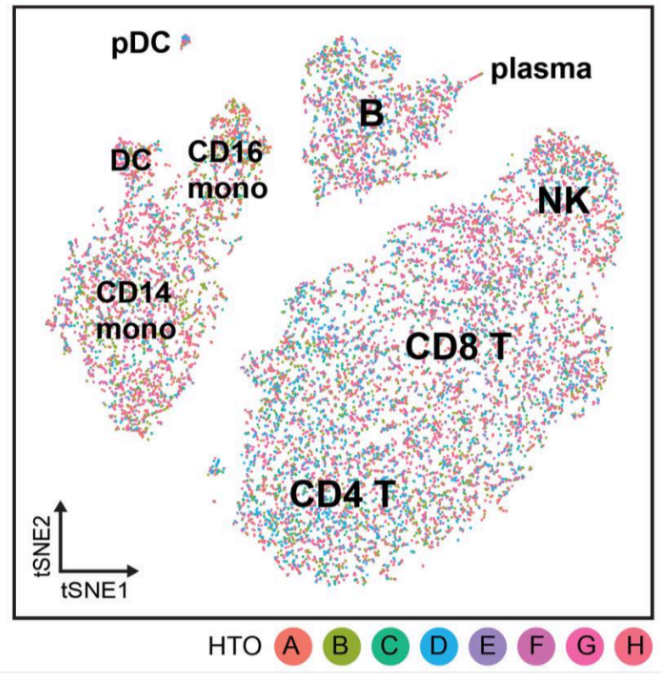
Cell hashing



Stoeckius et al. (https://doi.org/10.1186/s13059-018-1603-1)

Cell hashing





Summary

Summary

- Batch effects sometimes are not avoidable
- Many batch correction/integration methods available, mainly using joint dimension reduction, or joint clustering, or a combination of both
- Joint dimension reduction can yield interpretable factors and aid in the identification of equivalent states, but is computationally expensive
- Graph-based methods alone can be extremely fast, but may struggle when technical differences are on a similar scale to biological differences
- Performance assessment is challenging
- Sample multiplexing can help alleviate batch effects
- Simultaneous mRNA and protein profiling: REAP-seq and CITE-seq
- Several single cell multi-omics technologies