



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](http://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

...

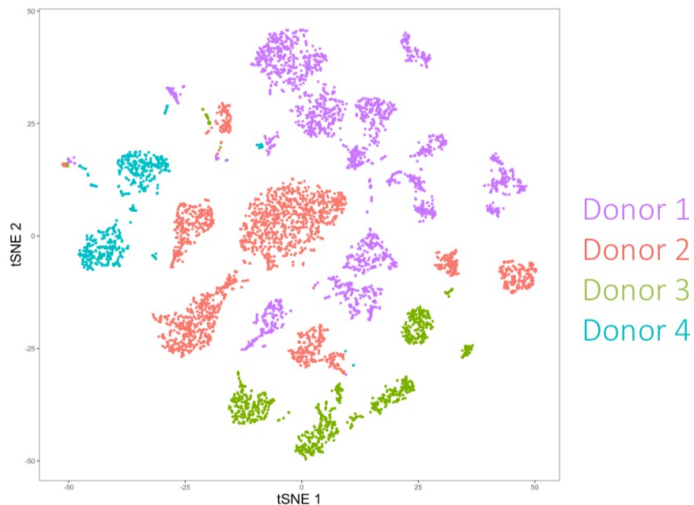
### Technical:

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

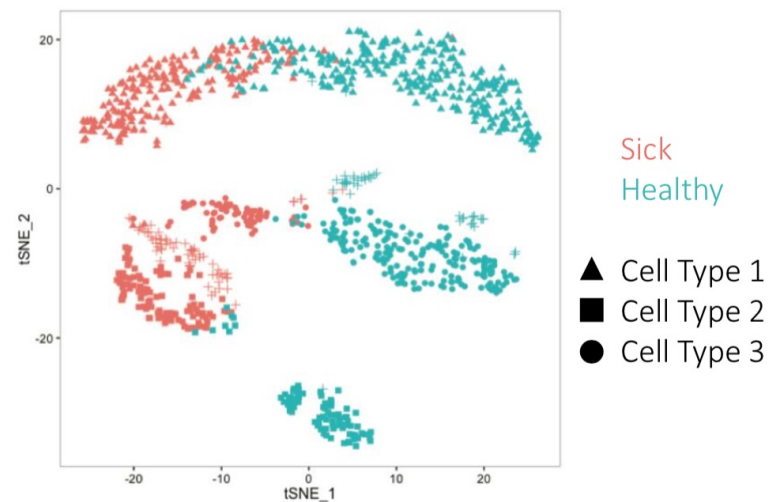
...

# Why integrate?

## Sources of variance in SC-RNAseq data



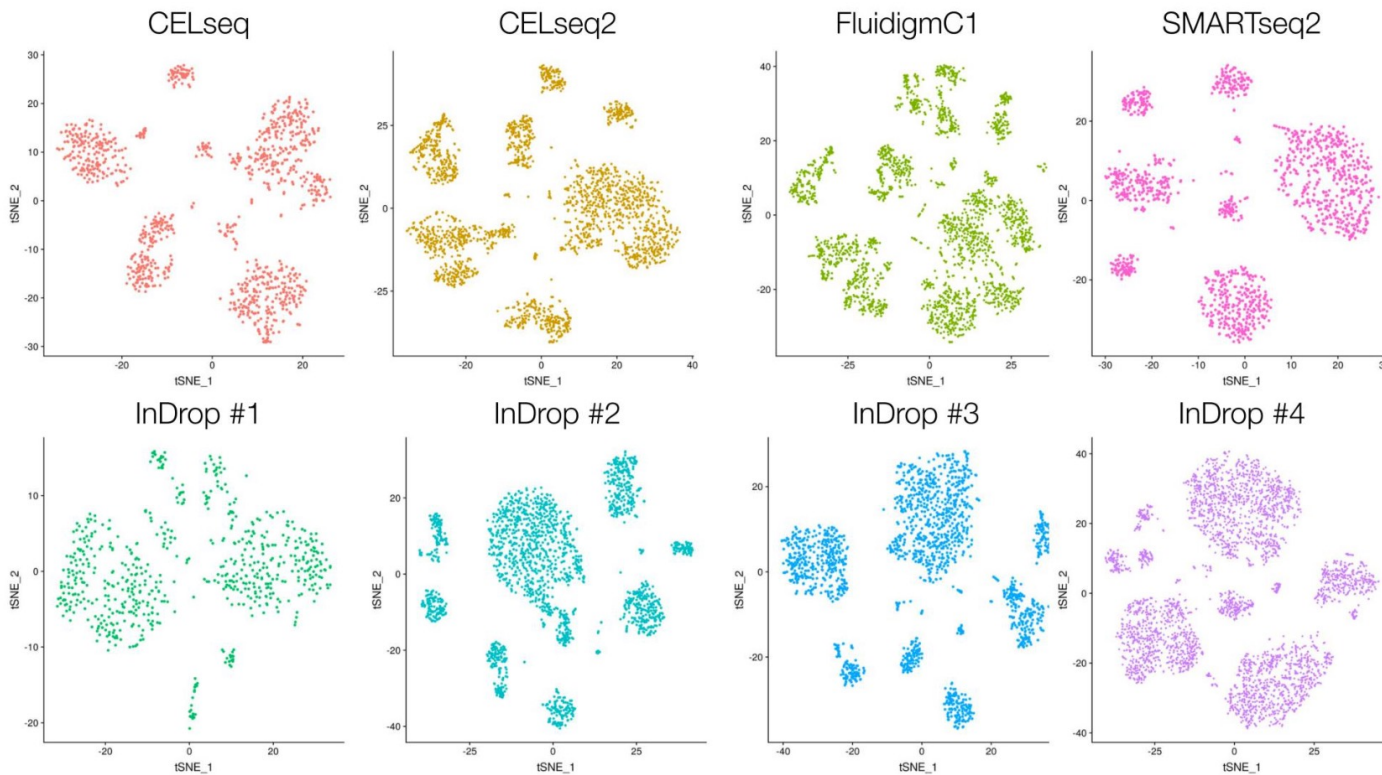
Same tissue from different donors



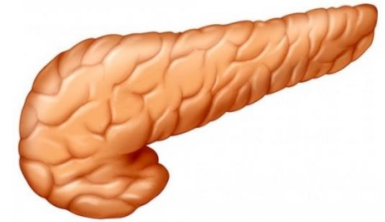
Cross condition comparisons

# Why integrate?

Building a cell atlas: 8 maps of the human pancreas



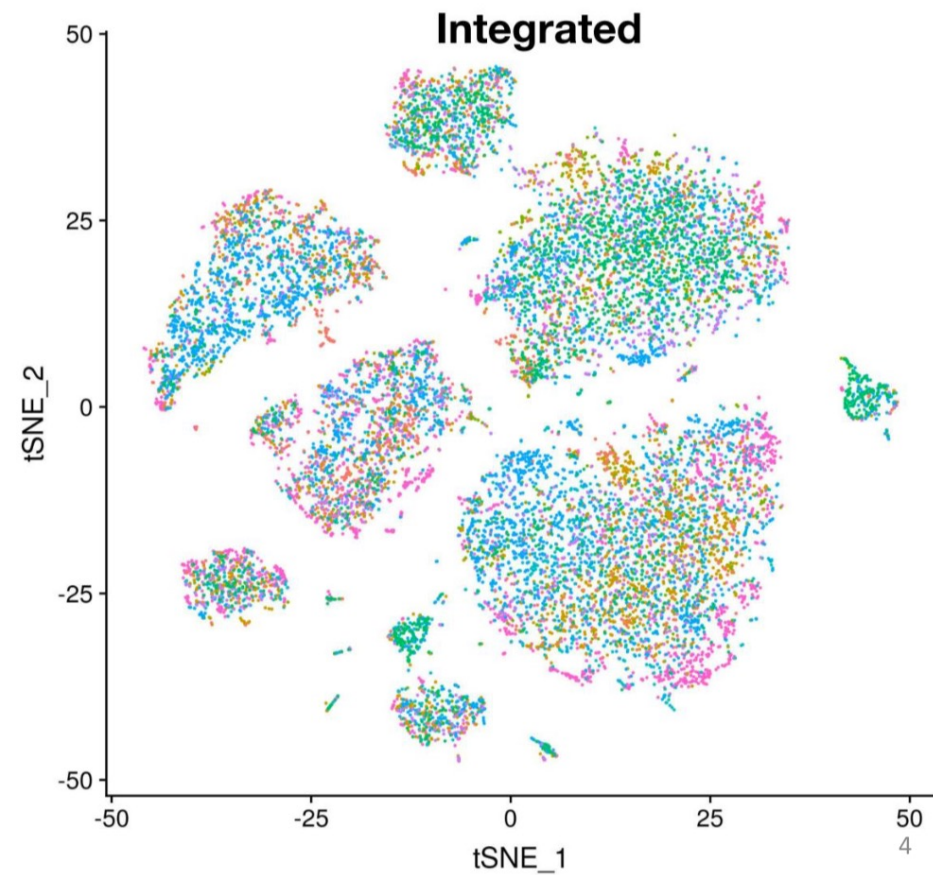
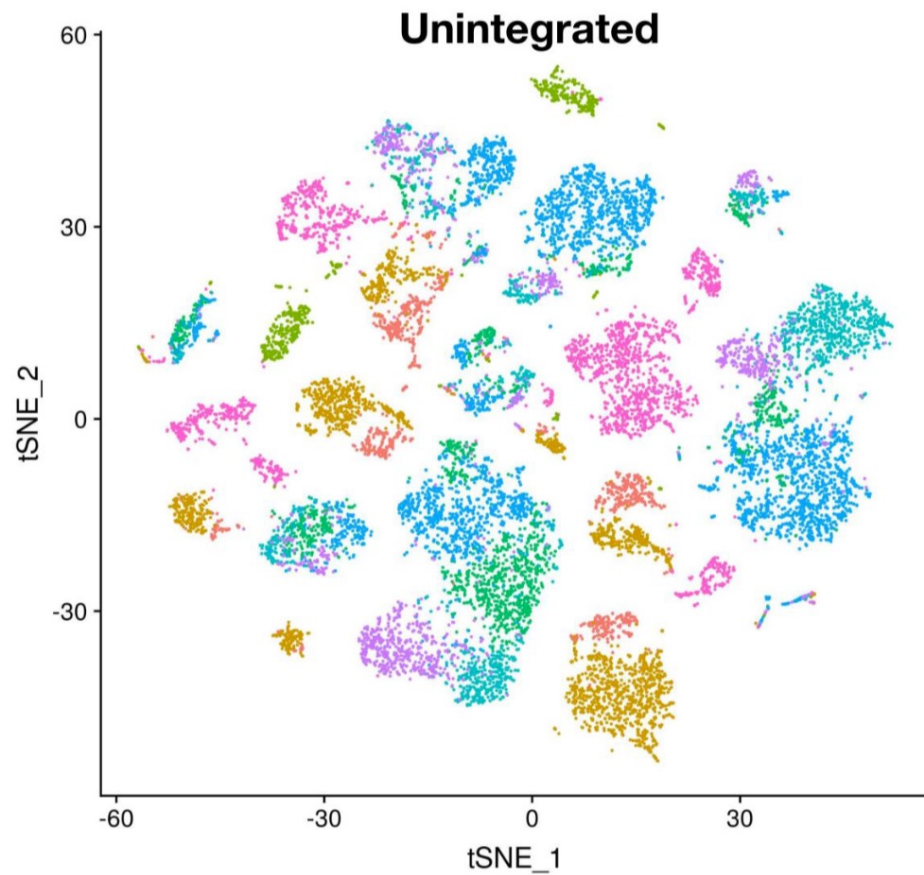
**Human pancreas**



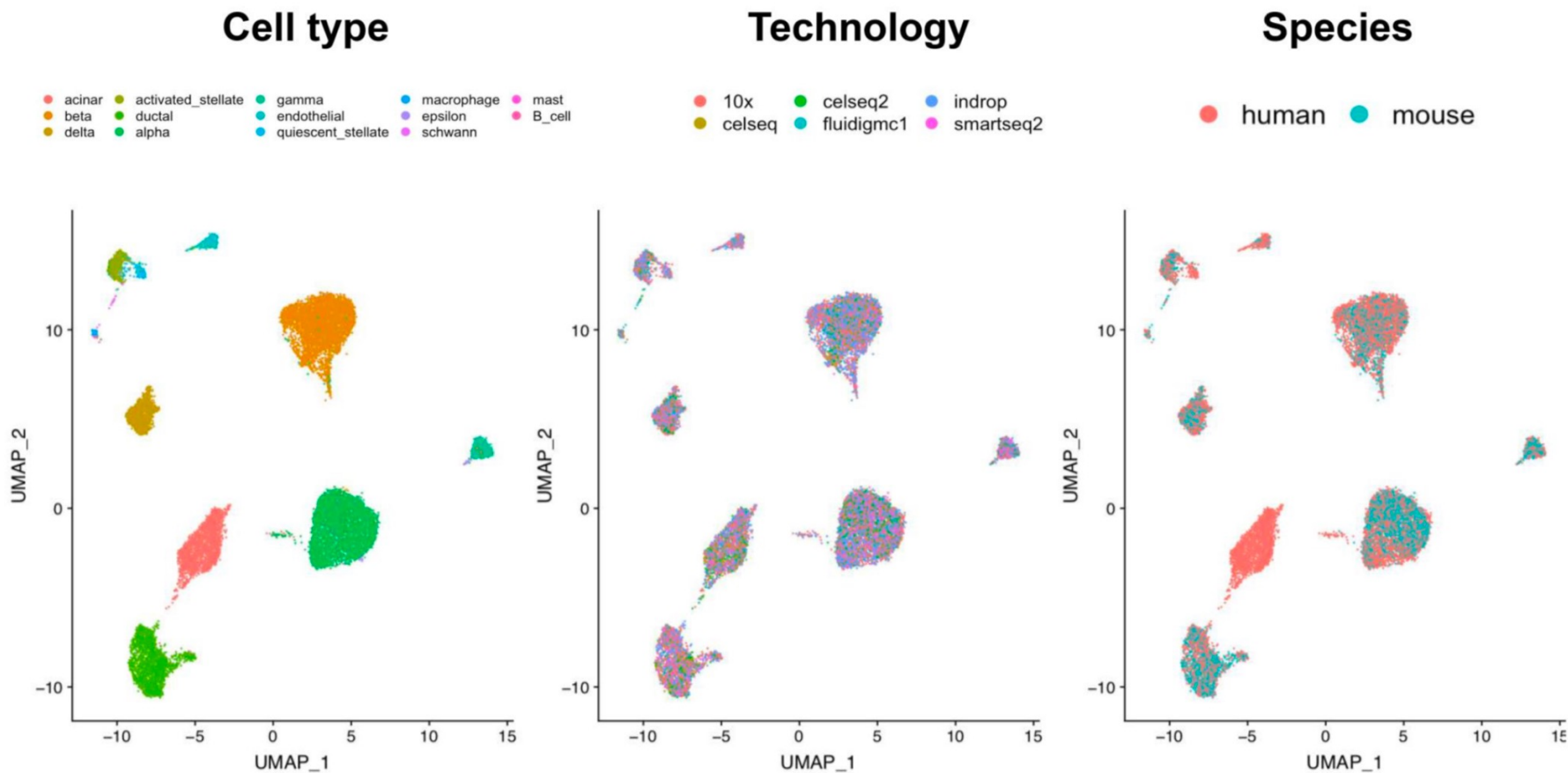
Baron et al. 2016, *Cell Syst.*  
Lawlor et al. 2017, *Genome Res.*  
Grun et al. 2016, *Cell Stem Cell*  
Muraro et al. 2016, *Cell Syst.*

# Why integrate?

Building a cell atlas: 8 maps of the human pancreas



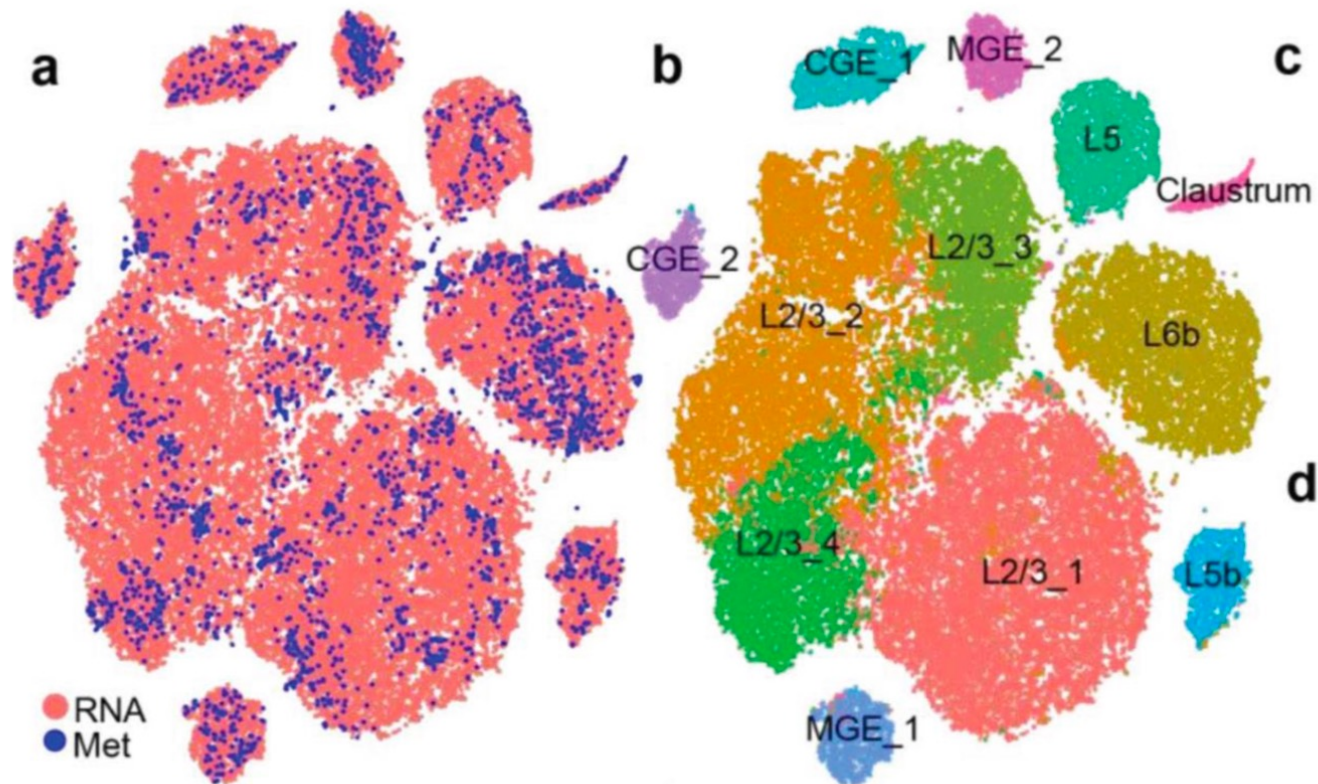
# Integration across modalities



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

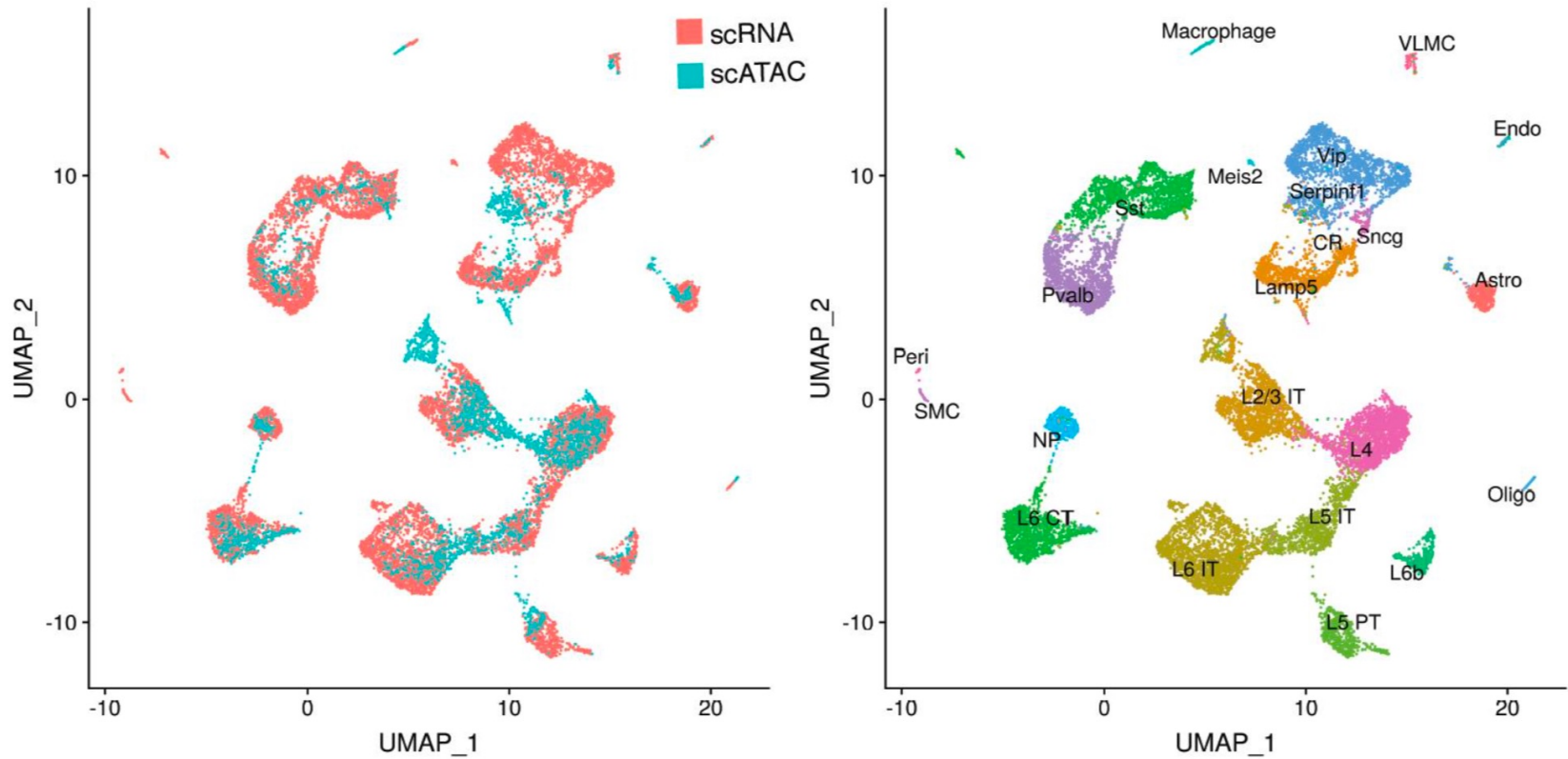
# Integration across modalities

## RNA-seq and methylation



# Integration across modalities

## RNA-seq and ATAC-seq



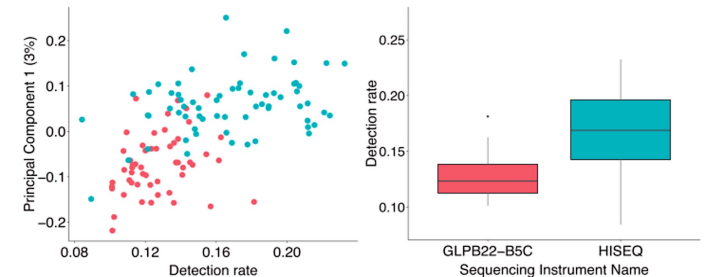


# Confounders and batch effects

## 1. Technical variability

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

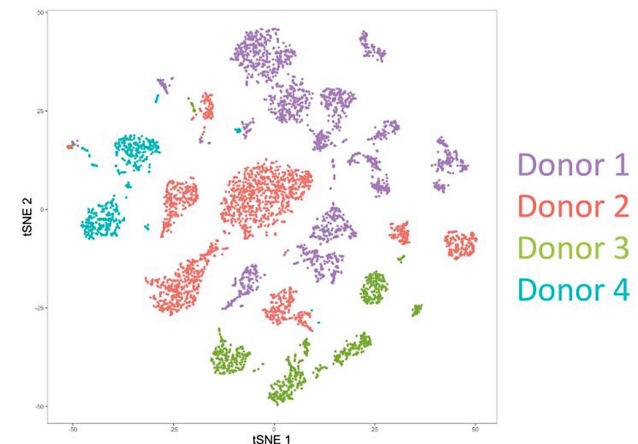
Technical ‘batch effects’ confound downstream analysis



## 2. Biological variability

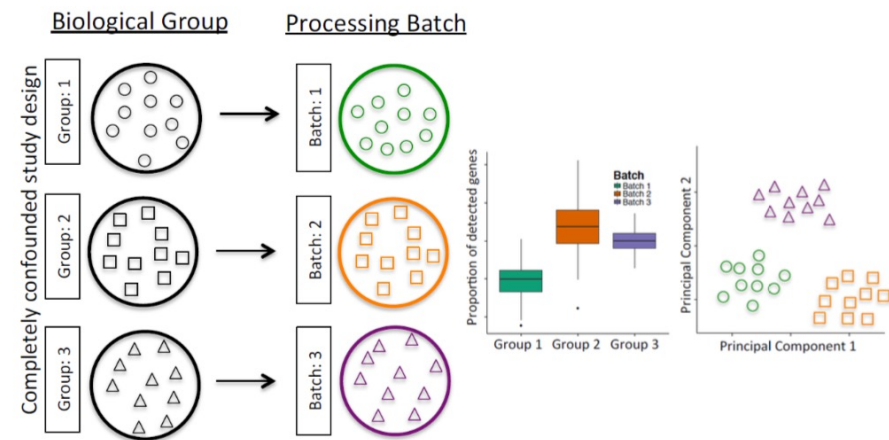
- Patient differences
- Environmental/genetic perturbation
- Evolution! (cross-species analysis)

Biological ‘covariates’ confound comparisons of scRNA-seq data



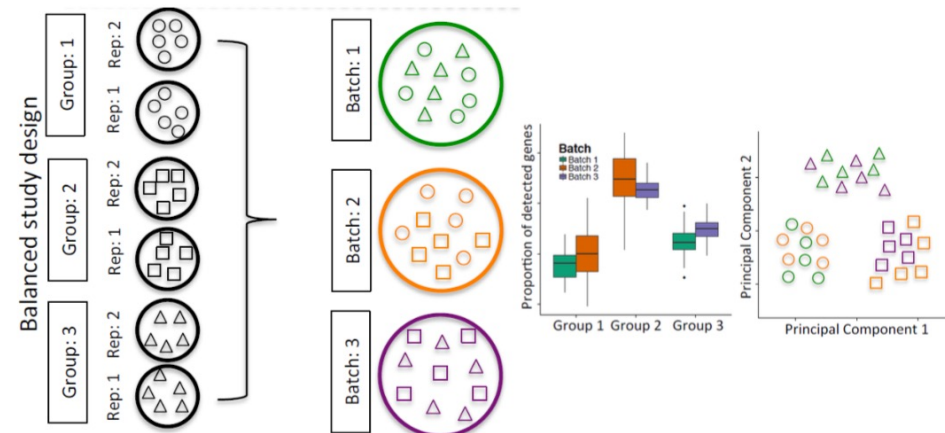
# Confounders and batch effects

## 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.

# Integration methods overview

# Batch correction methods

## Regression-based correction:

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

## Joint dimensionality reduction:

- common PCA / CPCA ([doi.org/10.1006/jmva.2000.1908](https://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](https://doi.org/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>)

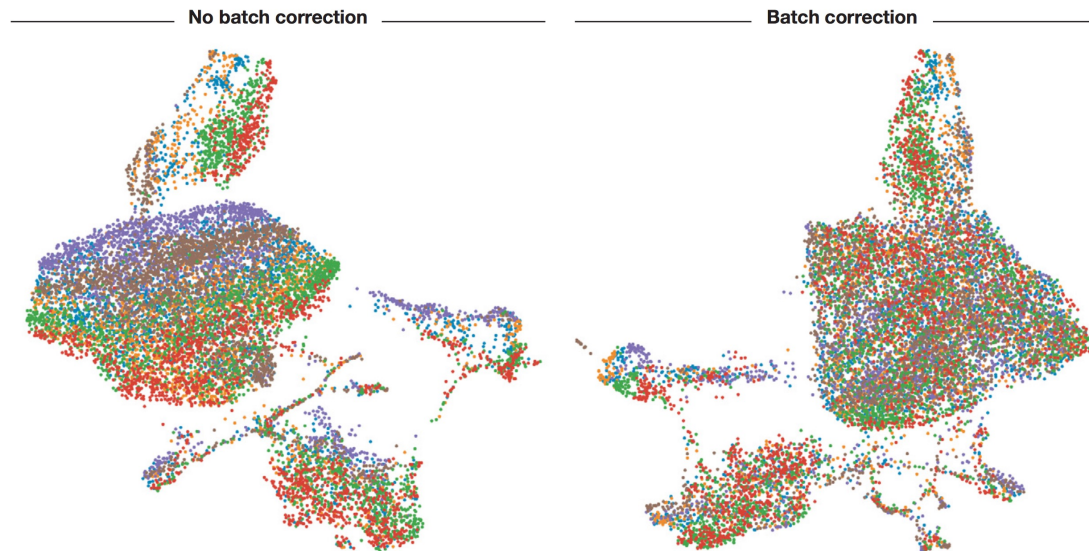
## Joint dimensionality reduction + Graph-based joint clustering

- 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

- 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



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**

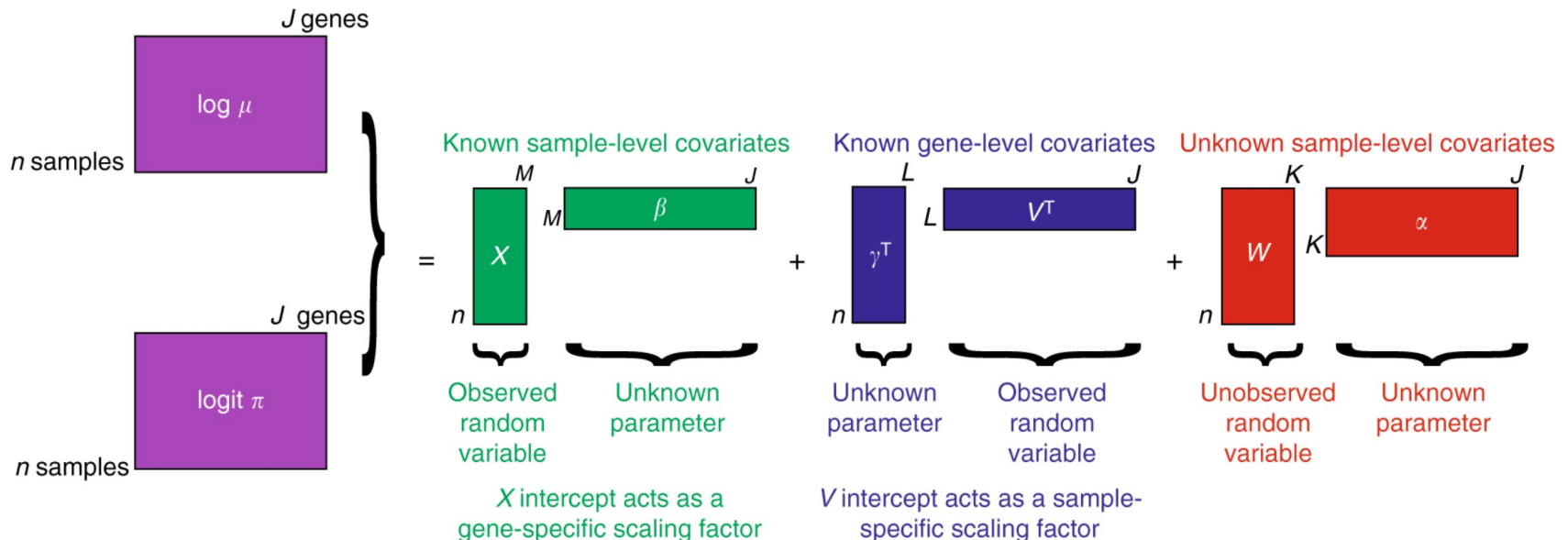
**2. Assume batch effect is additive**

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

**zinb-wave**

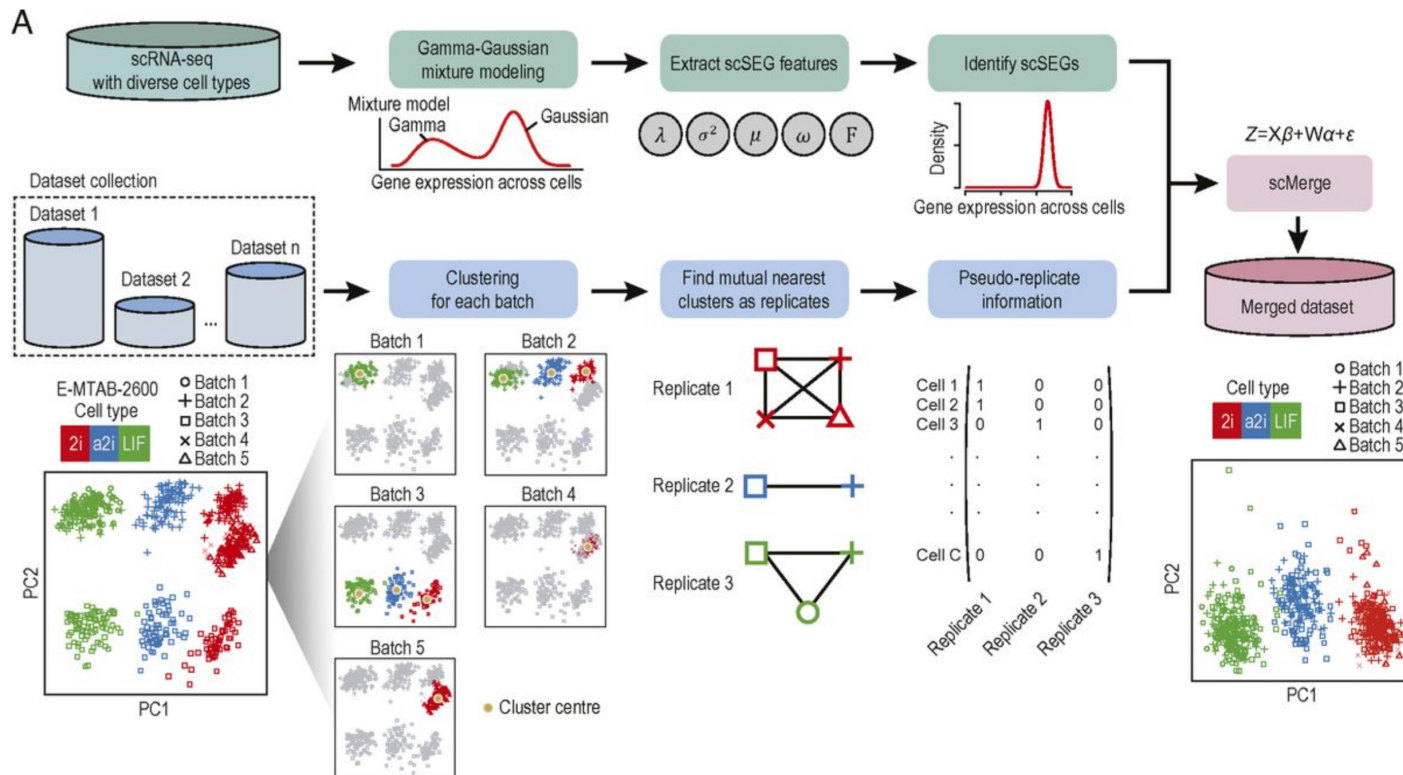


- Applies a matrix factorization model to accommodate both gene and cell covariates
- Similar to ZIFA (zero-inflated factor analysis)
- **Works well on simpler small-medium datasets**
- **It will be slow on large datasets**



# 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**

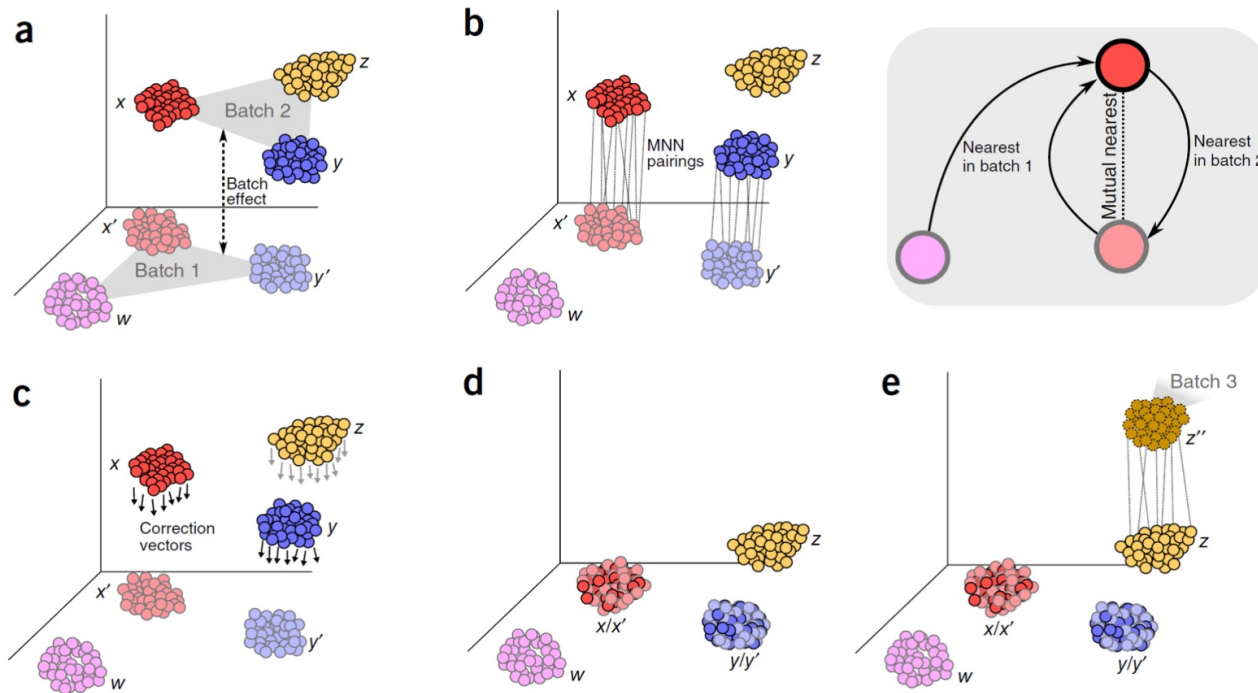


# Mutual Nearest Neighbors (MNN)

THE “turning point” method

# Mutual Nearest Neighbors (MNN)

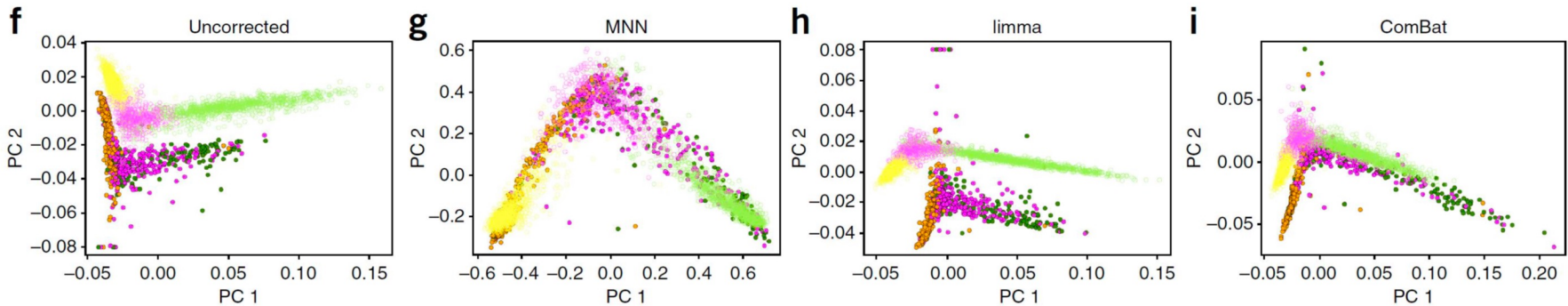
- Dimensionality reduction via multibatch PCA with all datasets
- Find K-NN across datasets
- Compute merging vectors
- **It scales well on large datasets**



# 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



SMART-seq2

● MEP

● GMP

● CMP

MARS-seq

● MEP

● GMP

● CMP

MEPs: megakaryocyte–erythrocyte progenitors

GMPs: granulocyte–monocyte progenitors

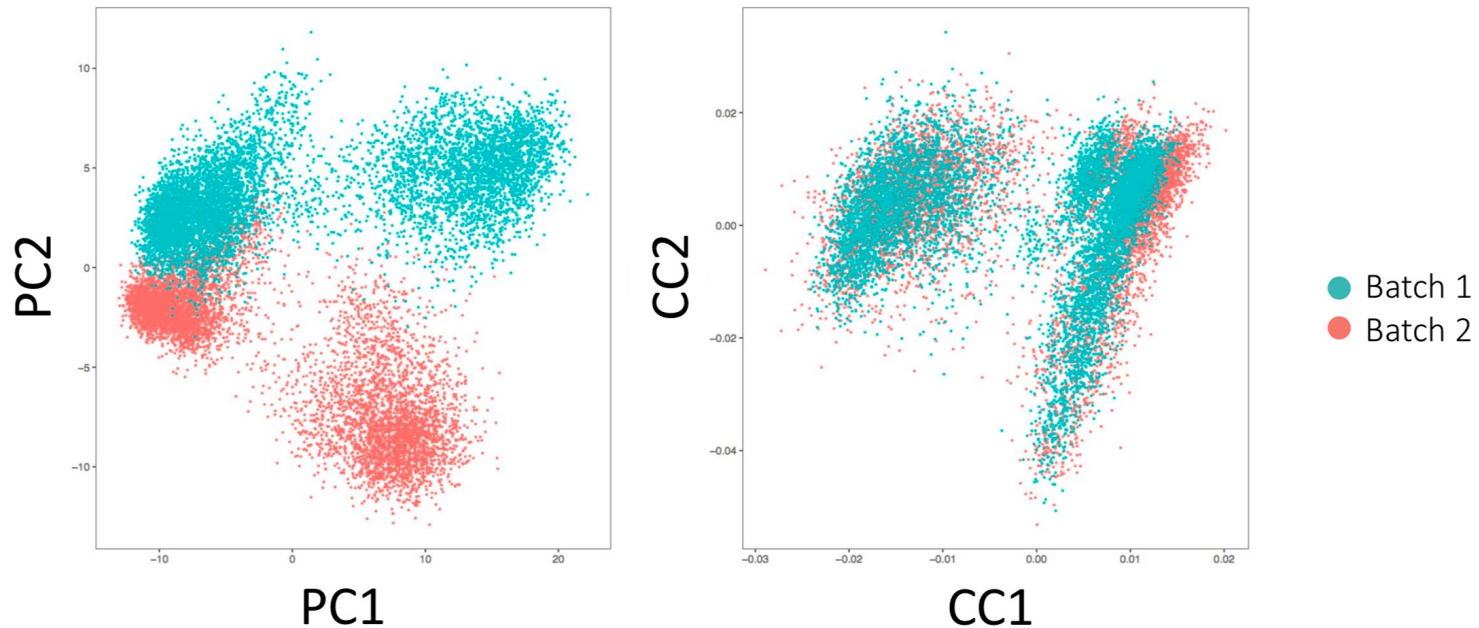
CMPs: common myeloid progenitors

# CCA + anchors (Seurat v3)

# How CCA works?

Canonical correlation analysis

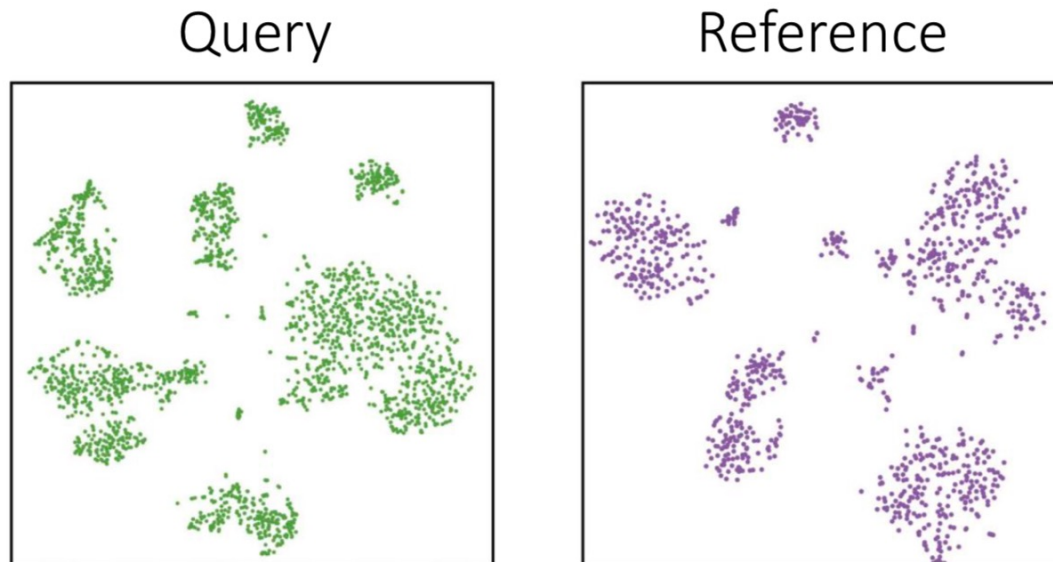
CCA captures correlated sources of variation between two datasets





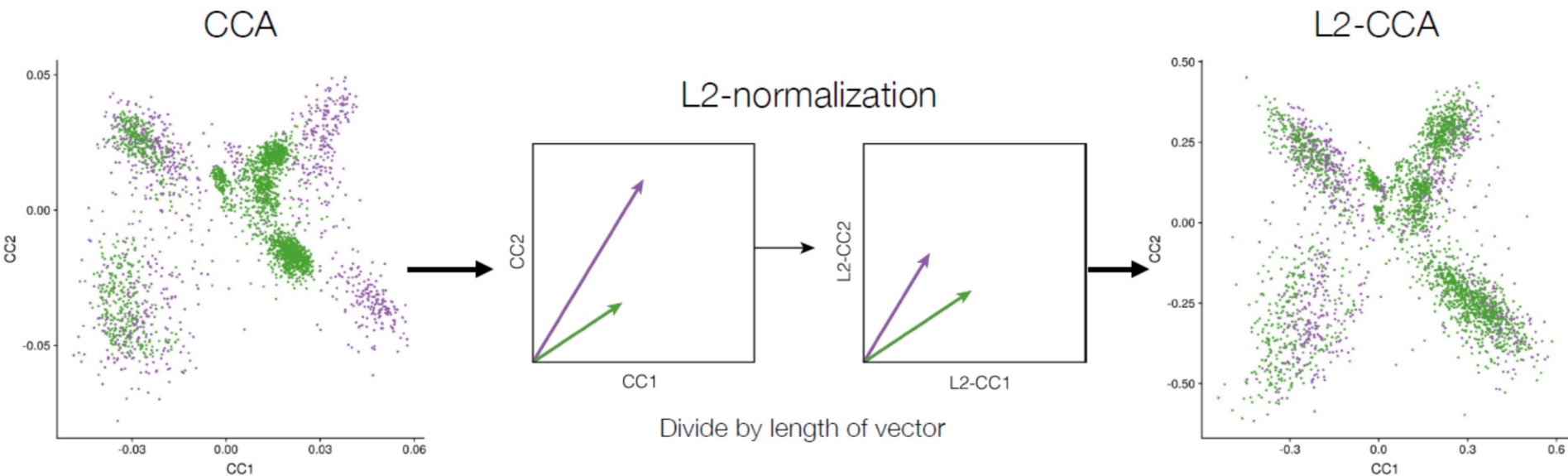
# CCA + anchors (Seurat v3)

1. Find corresponding cells across datasets
2. Compute a data adjustment based on correspondences between cells
3. Apply the adjustment



# Mutual Nearest Neighbors (MNN)

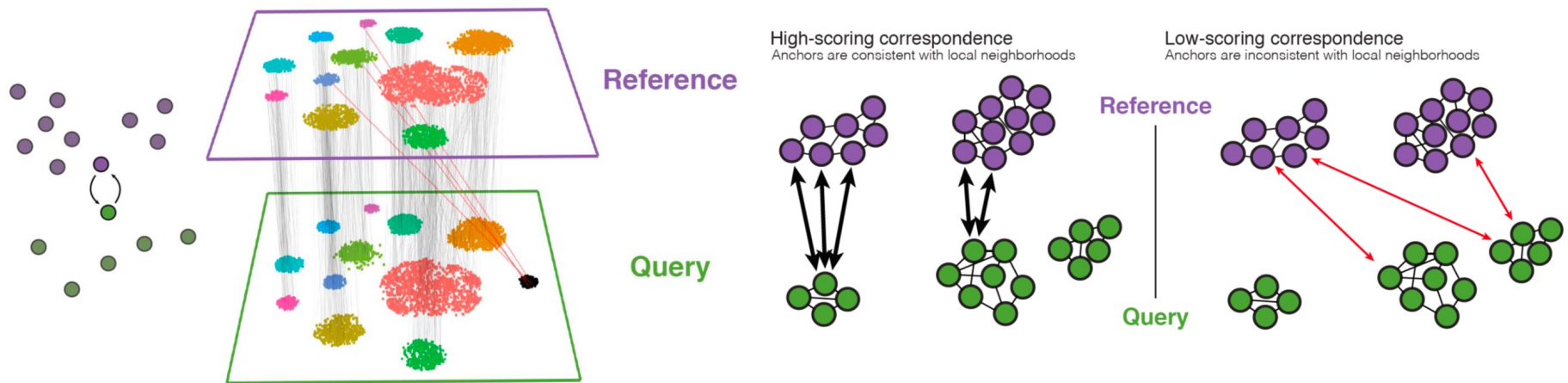
L2-normalization corrects for differences in scale



# Finding corresponding cells

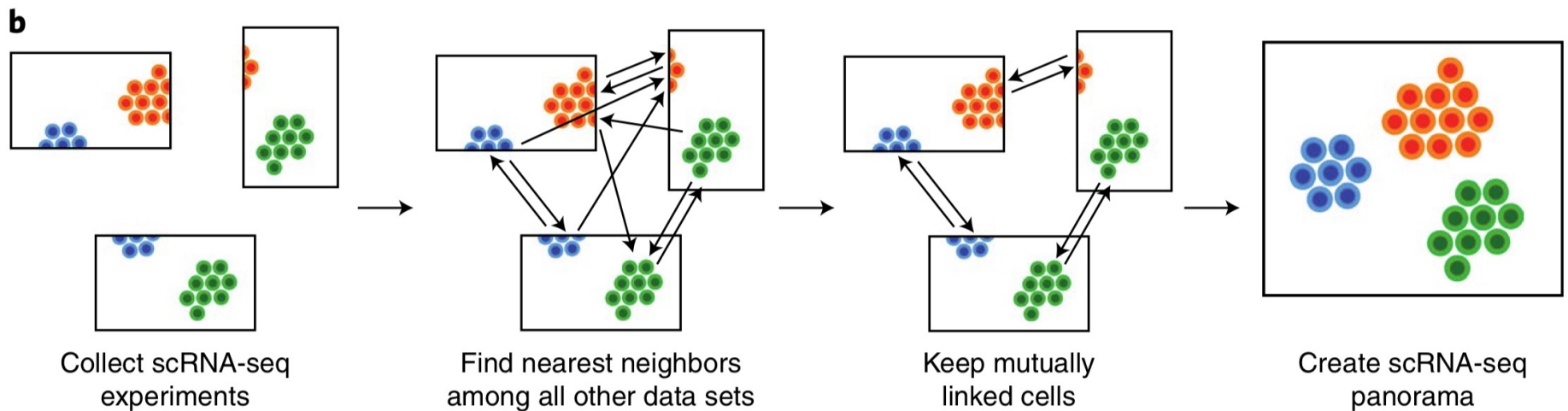
Anchors: mutual nearest neighbours (MNN)

- It scales well on large datasets

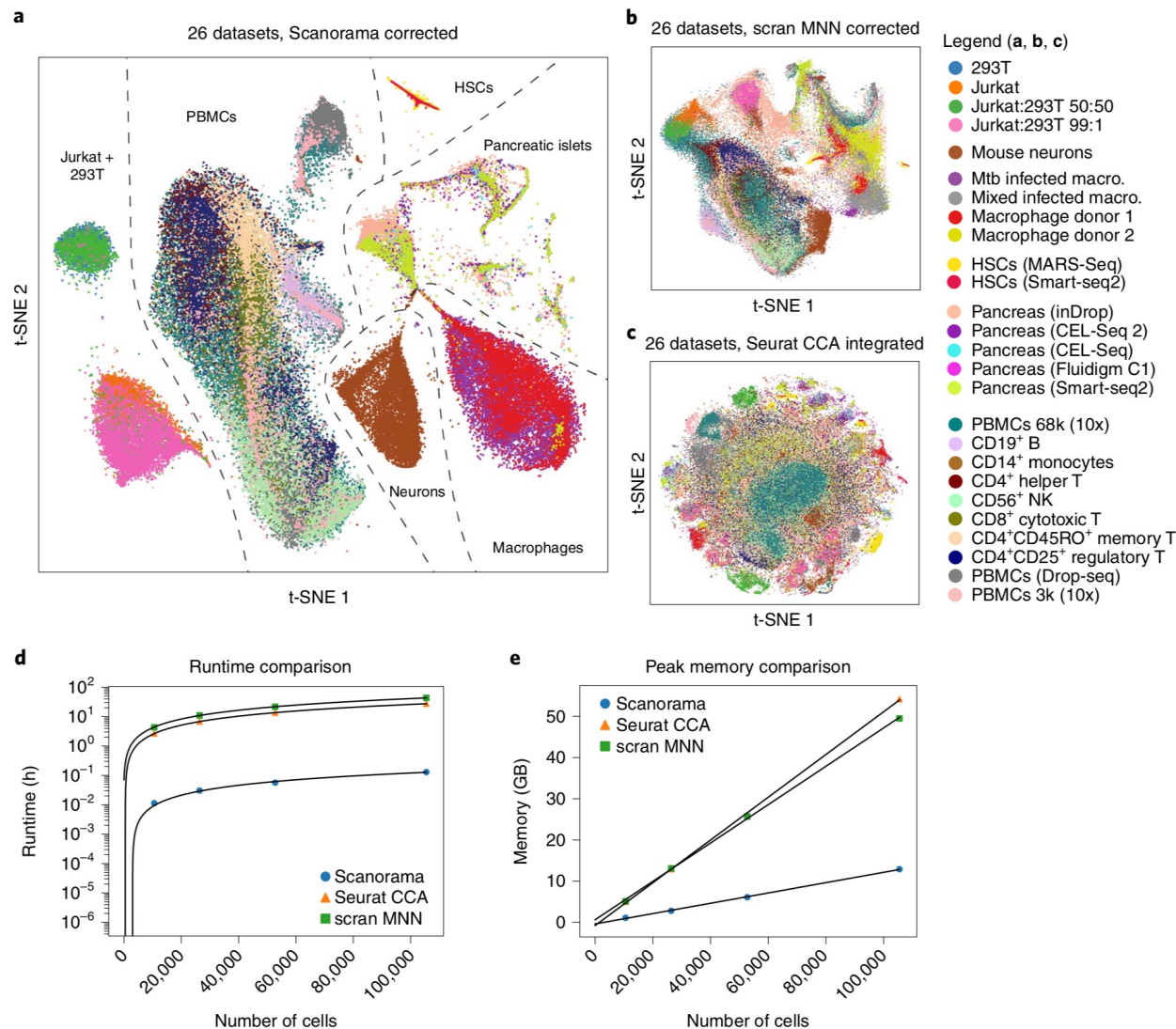


# Scanorama

- 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**

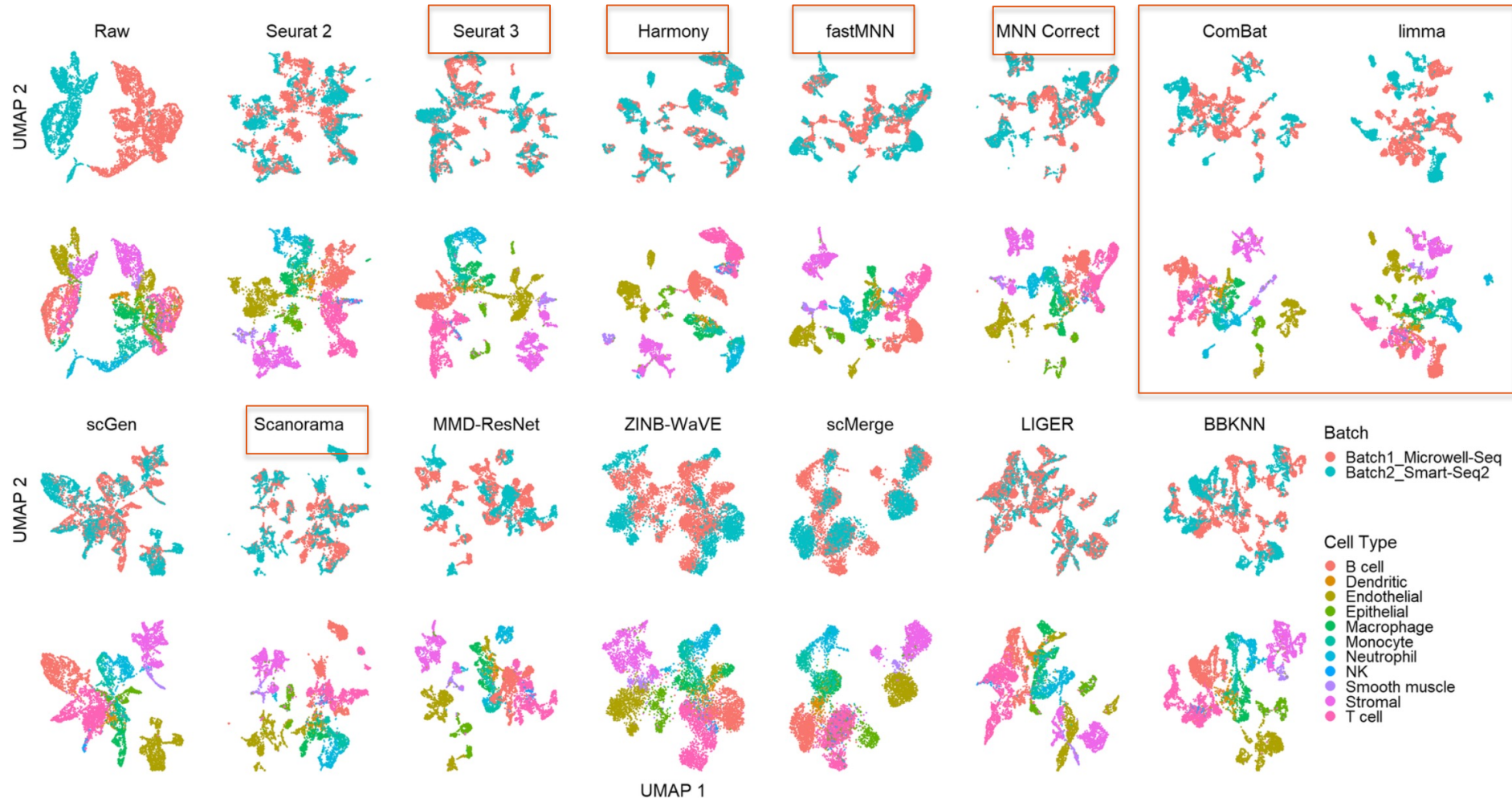


# Scanorama



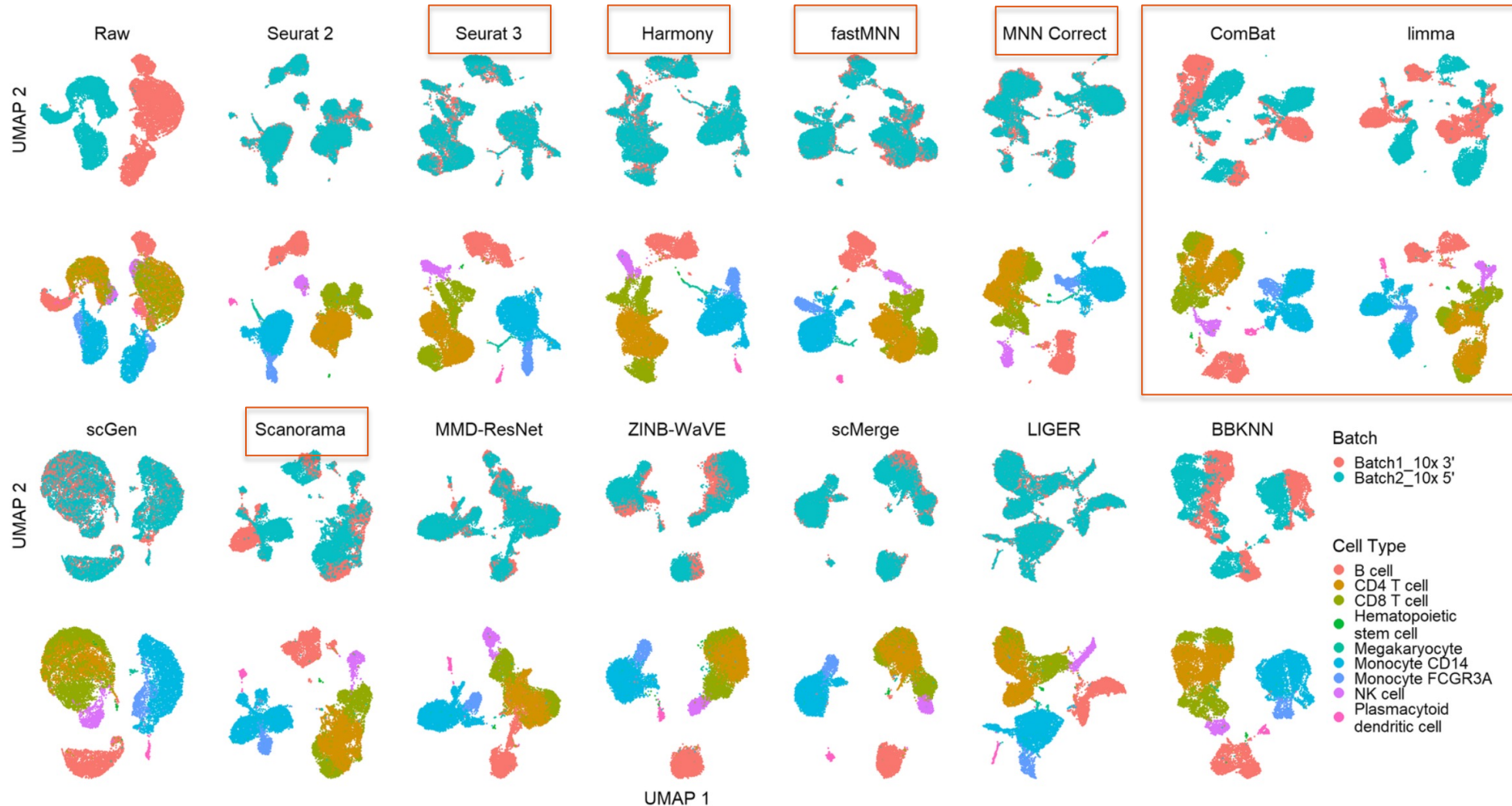
# Evaluation of batch correction efficiency

# Batch-correction performance assessment





# Batch-correction performance assessment



# Batch-correction performance assessment

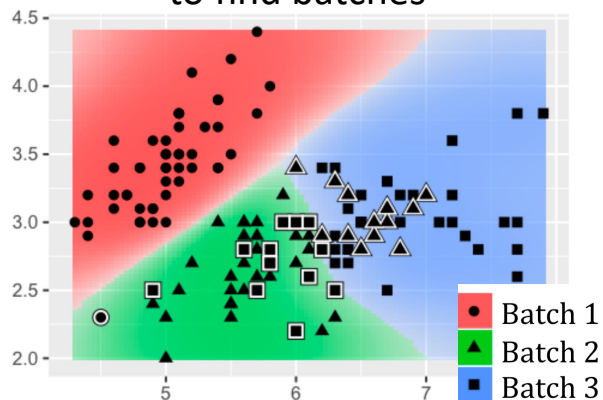
## 1. Evaluate mixing efficiency (Goal A)

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

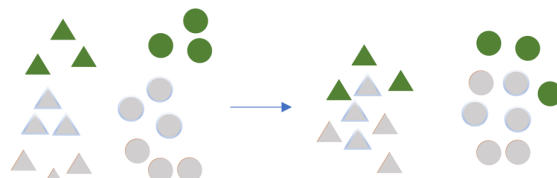
## 2. Evaluate preservation of remaining variance (Goals B, C)

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

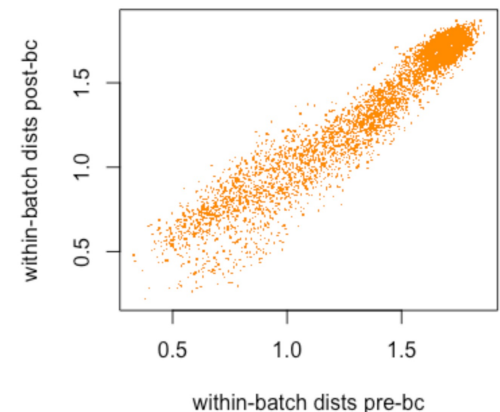
Classifiers must fail  
to find batches



Local structure  
preservation



Global structure  
preservation



# Batch-correction performance assessment

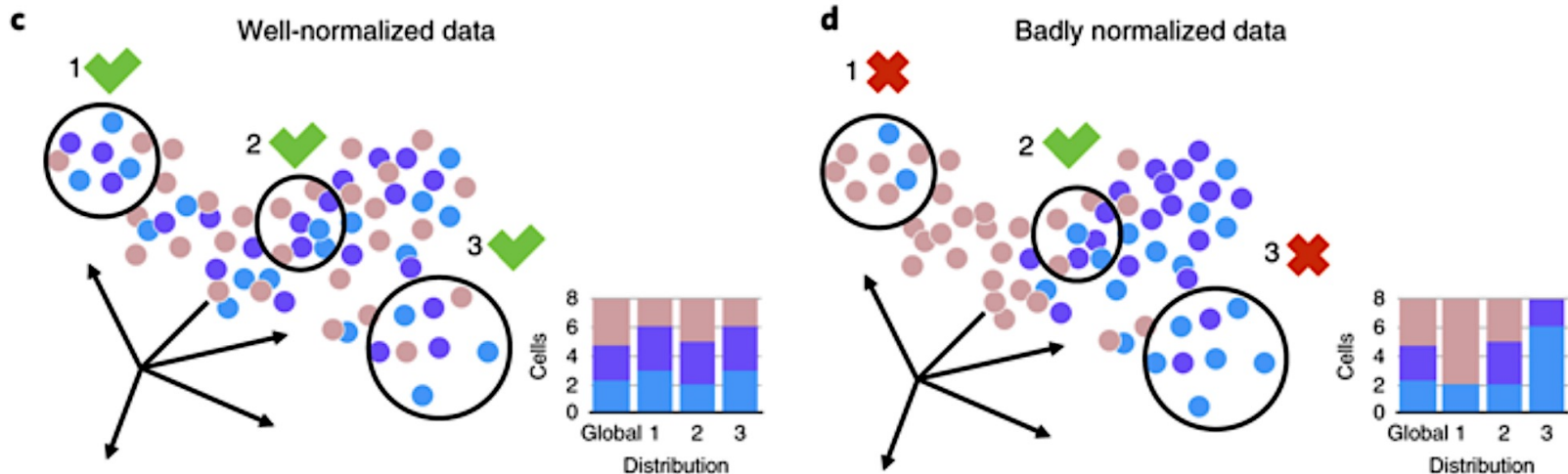
We wish to obtain corrected data where the following goals are met:

## Goal:

1. The batch-originating variance is erased
2. Meaningful heterogeneity is preserved (within batches)
3. No artefactual variance is introduced

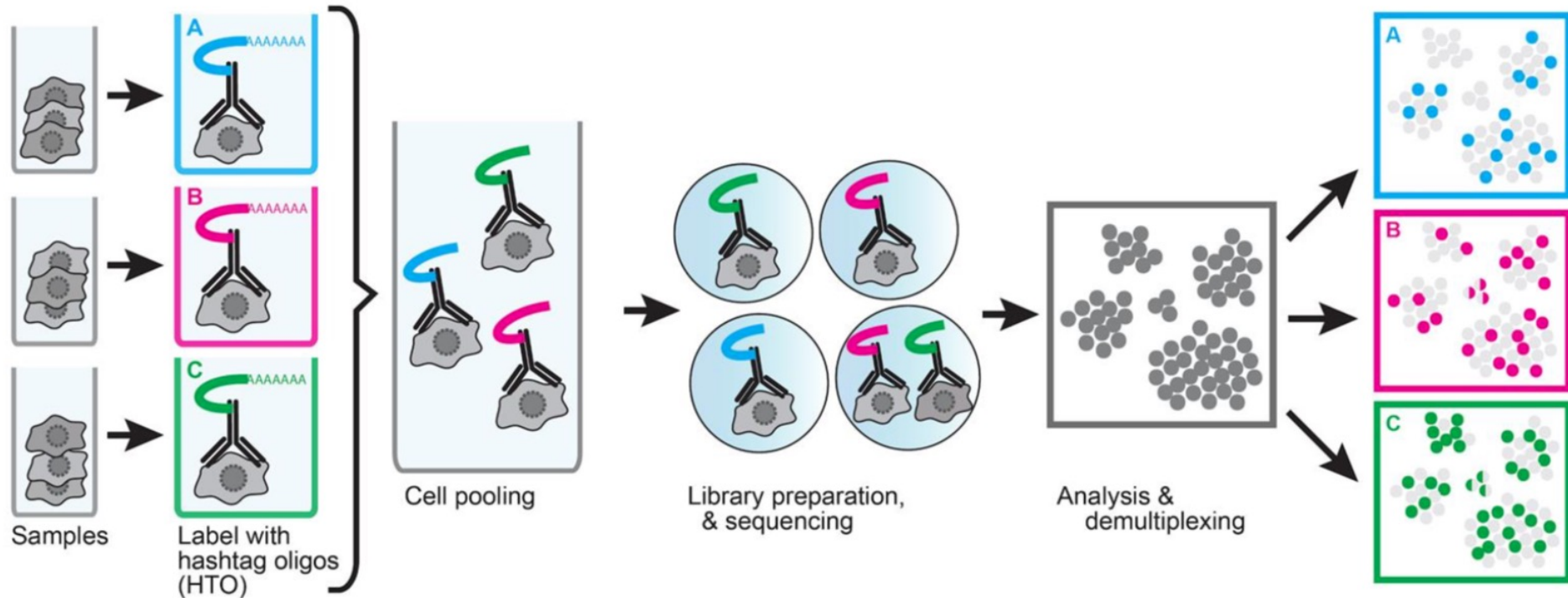
## What it practically means:

- Similar cell types are intermixed across batches
- We are not mixing distinct cell types (across or within batches)
- We do not separate similar cells within batches

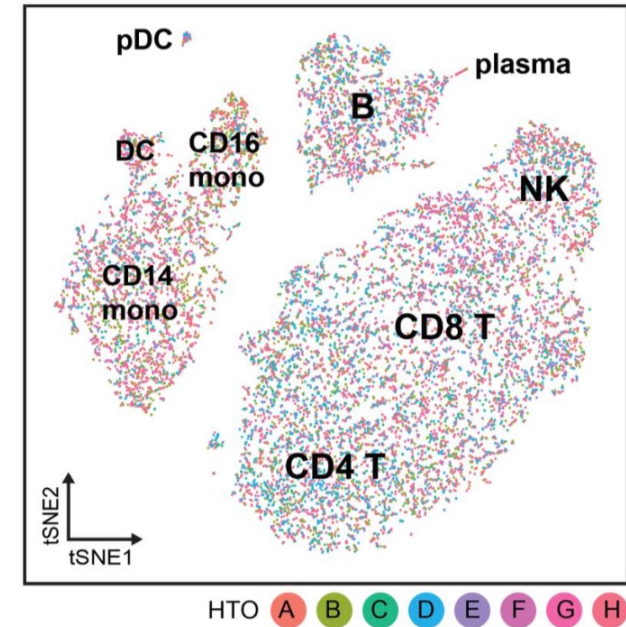
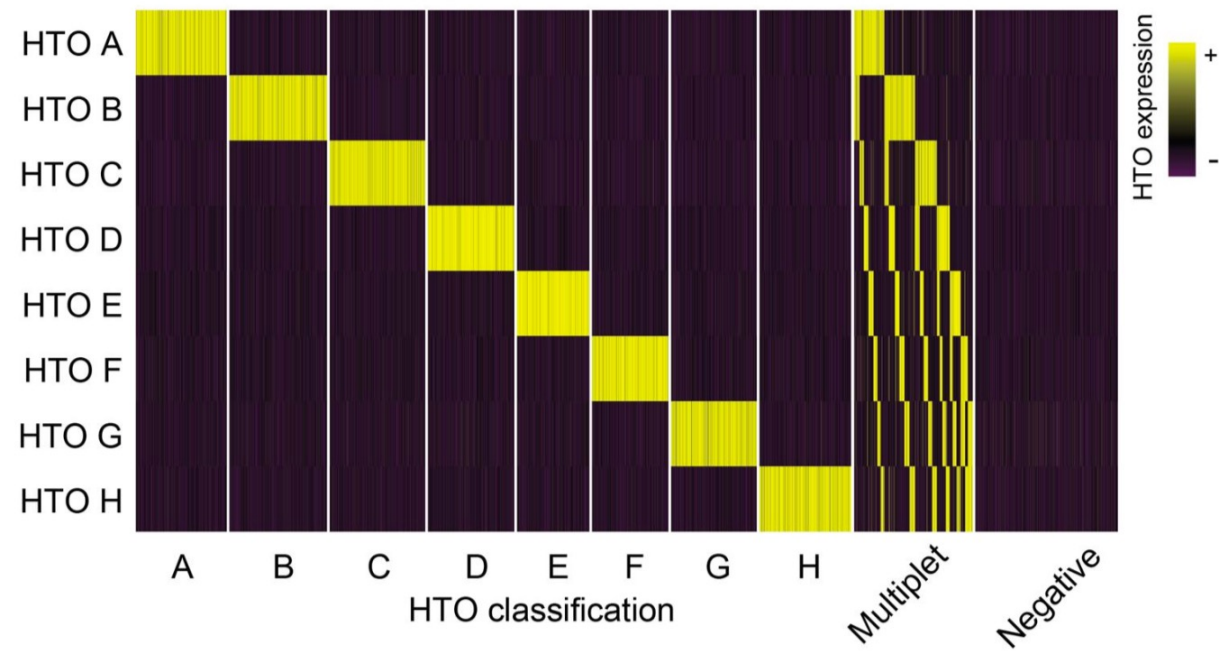


# Avoiding batches

# Cell hashing



# 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 can be integrated