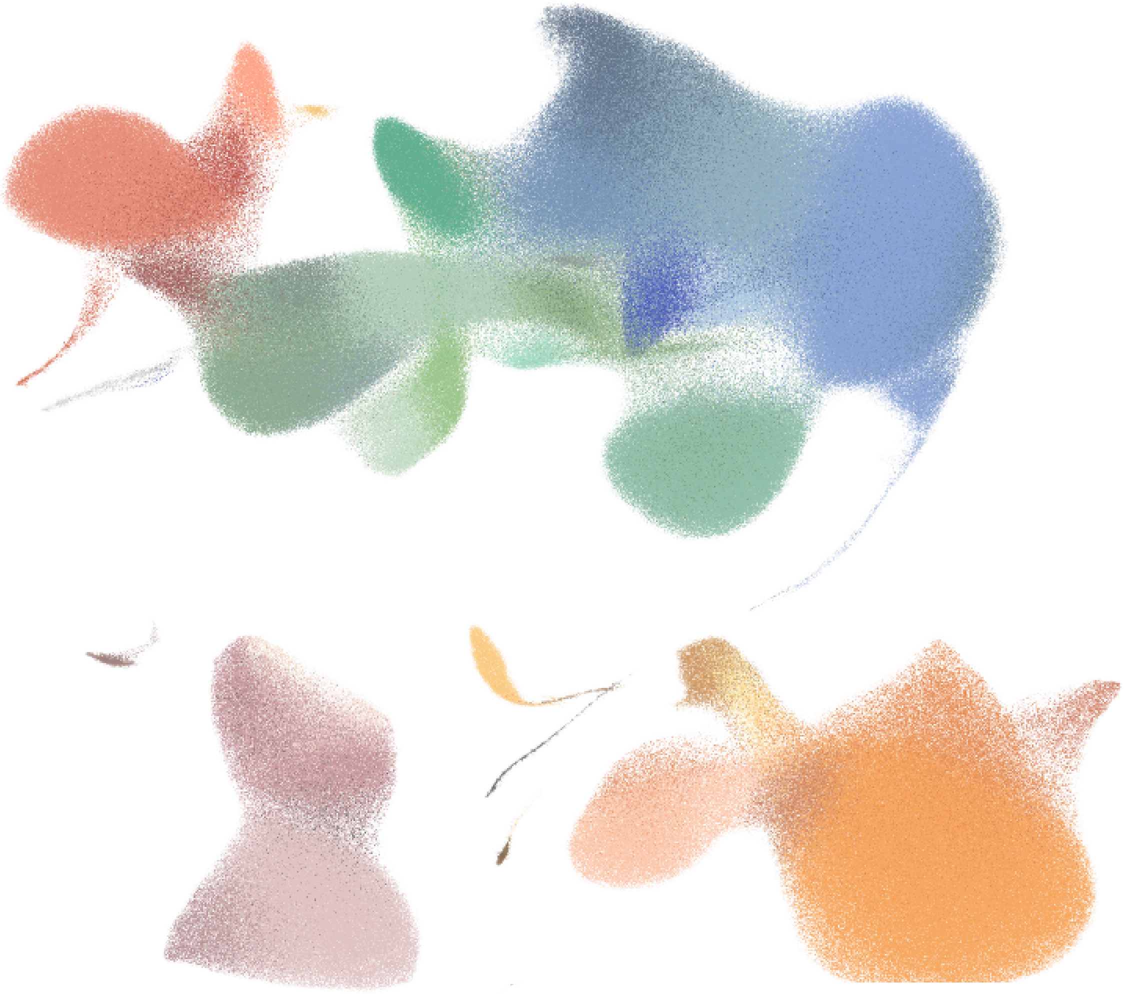
Data Integration



Yusuf Caglar Odabasi

December 1.-3. 2025

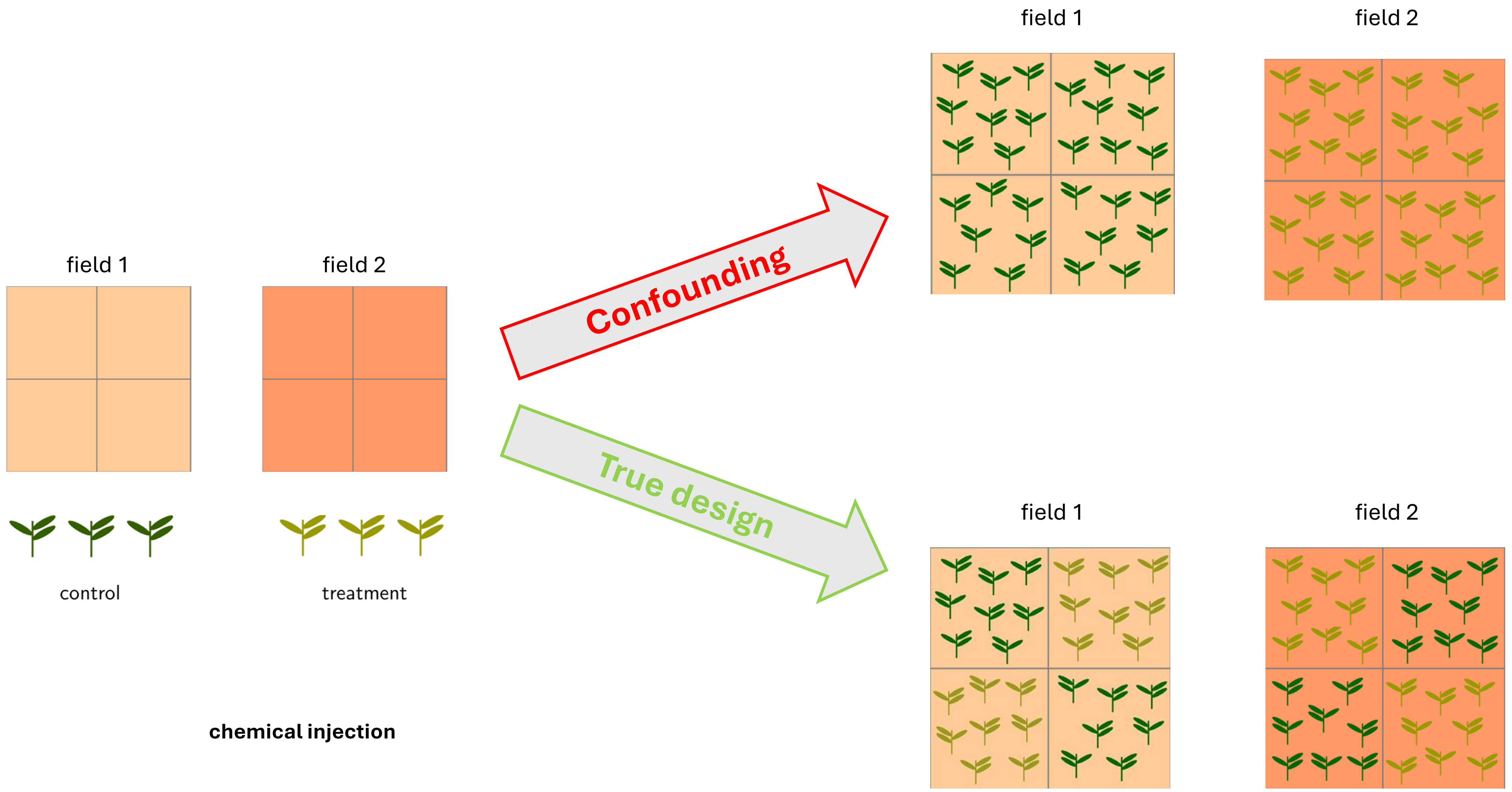
Course on scRNA-seq Data Analysis

elixir CZECH REPUBLIC

ibt

IMG

Experiment Design: Confounders and Batch effects



field 1

field 2

control

treatment

chemical injection

Confounding

True design

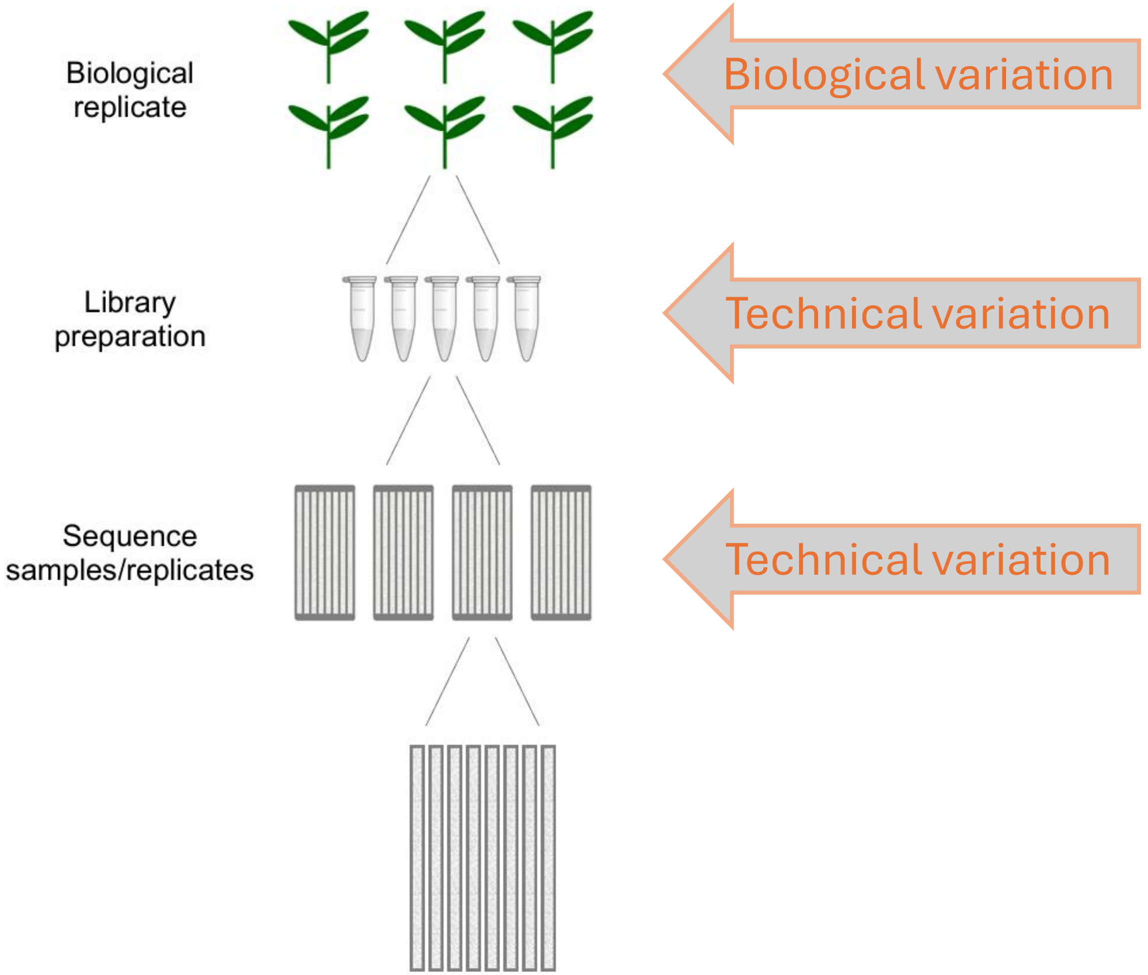
field 1

field 2

field 1

field 2

Experiment Design: Confounders and Batch effects



Biological variation

Technical variation

Technical variation

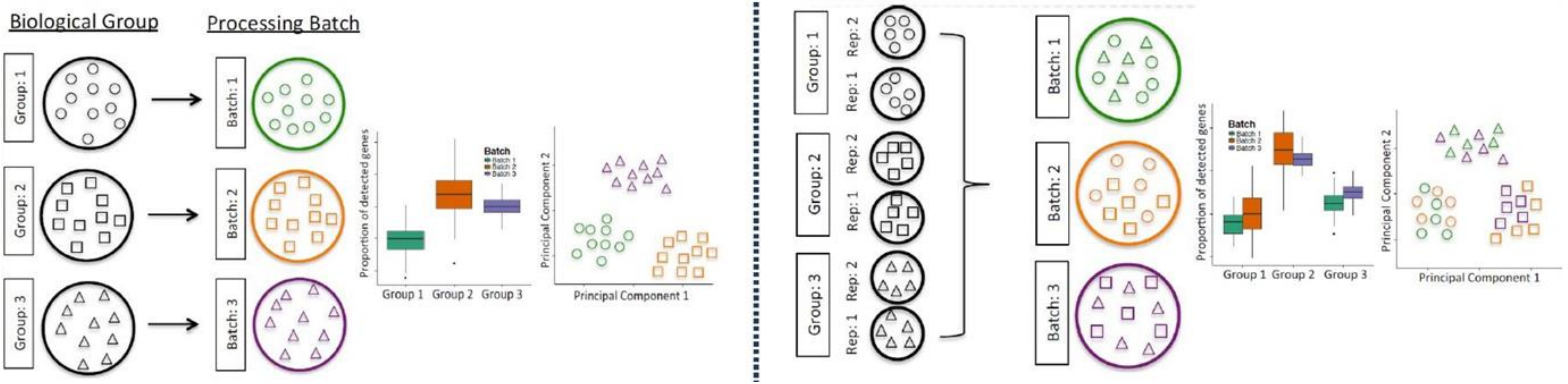
1. Technical variability
   - Changes in sample quality/processing
   - Library prep or sequencing technology

Technical ‘batch effects’ confound downstream analysis

2. Biological variability
   - Patient differences
   - Evolution! (cross-species analysis)

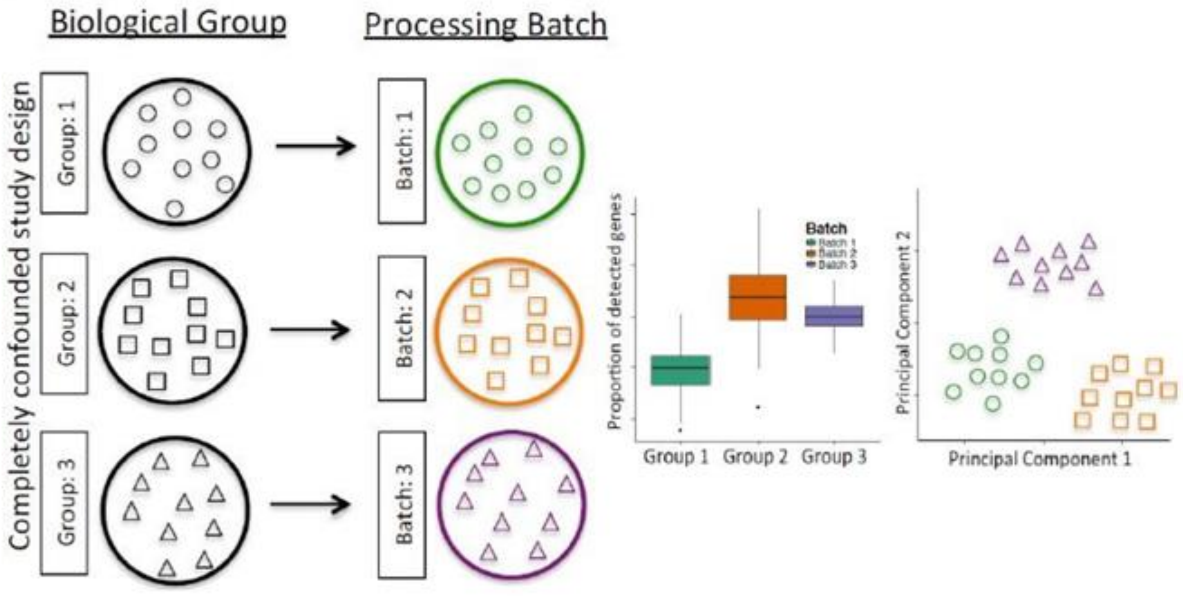
Biological ‘batch effects’ confound comparisons of scRNA-seq data

Experiment Design: Confounders and Batch effects

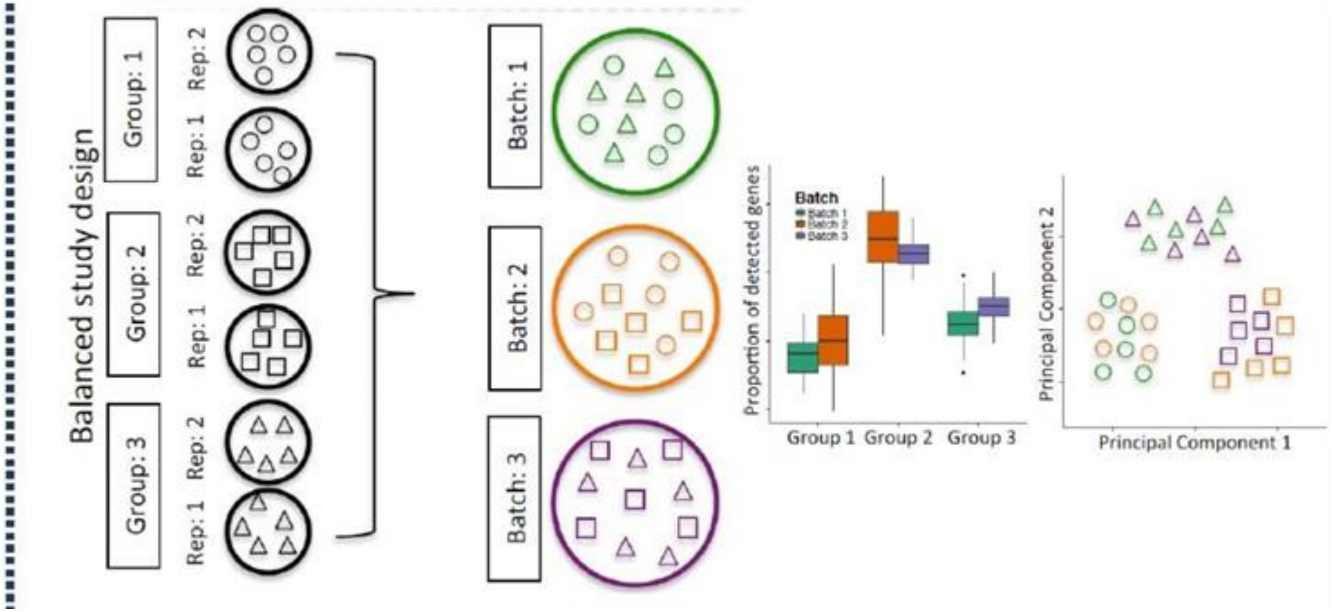


Experiment Design: Confounders and Batch effects

Confounded design

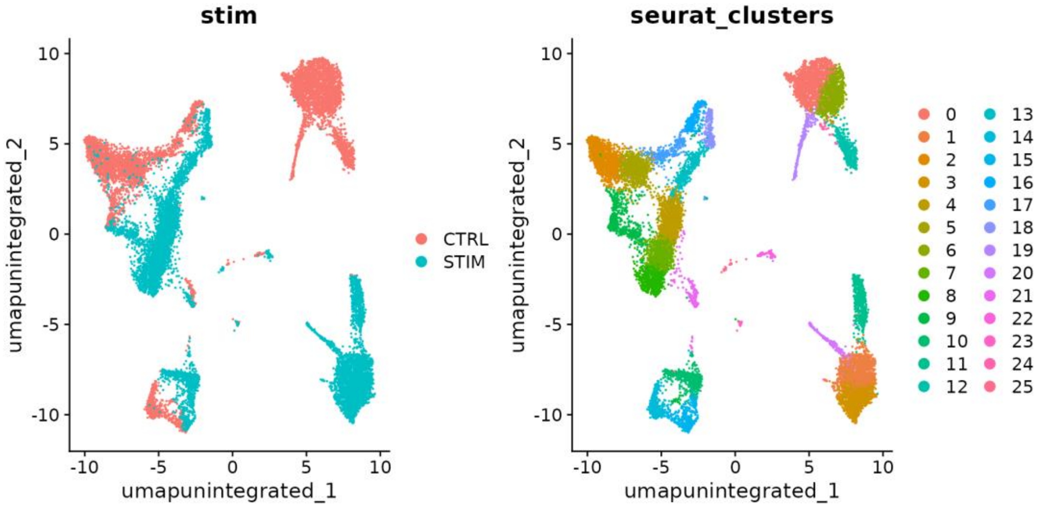


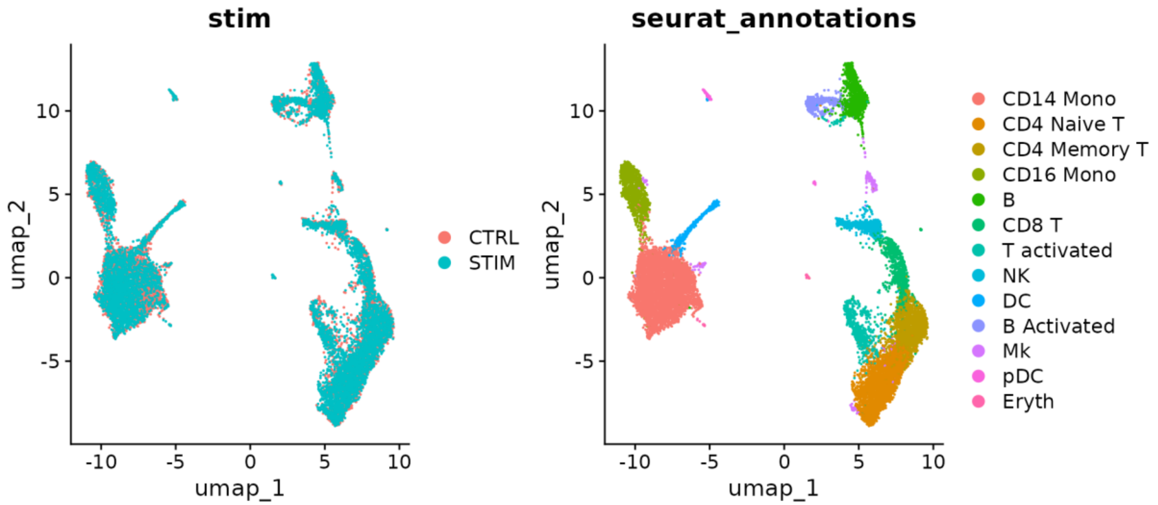
Not confounded design

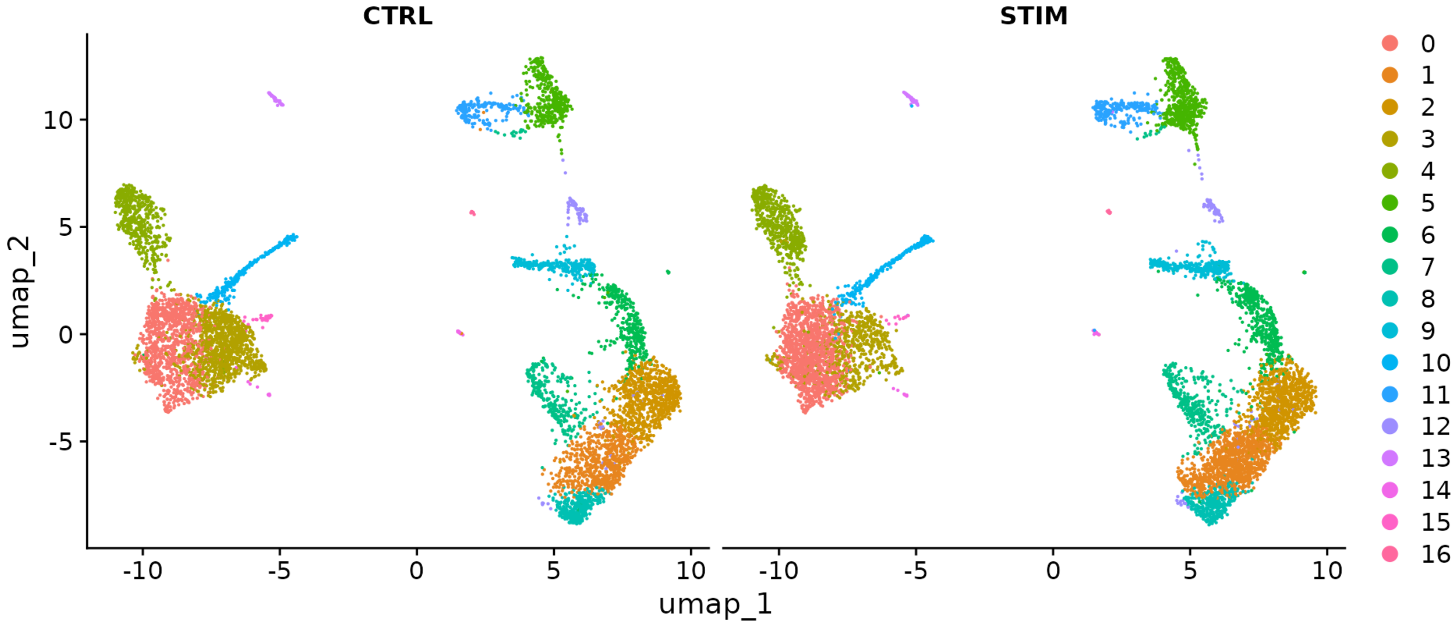


Good experimental design *does not remove batch effects*, it prevents them from biasing your results.

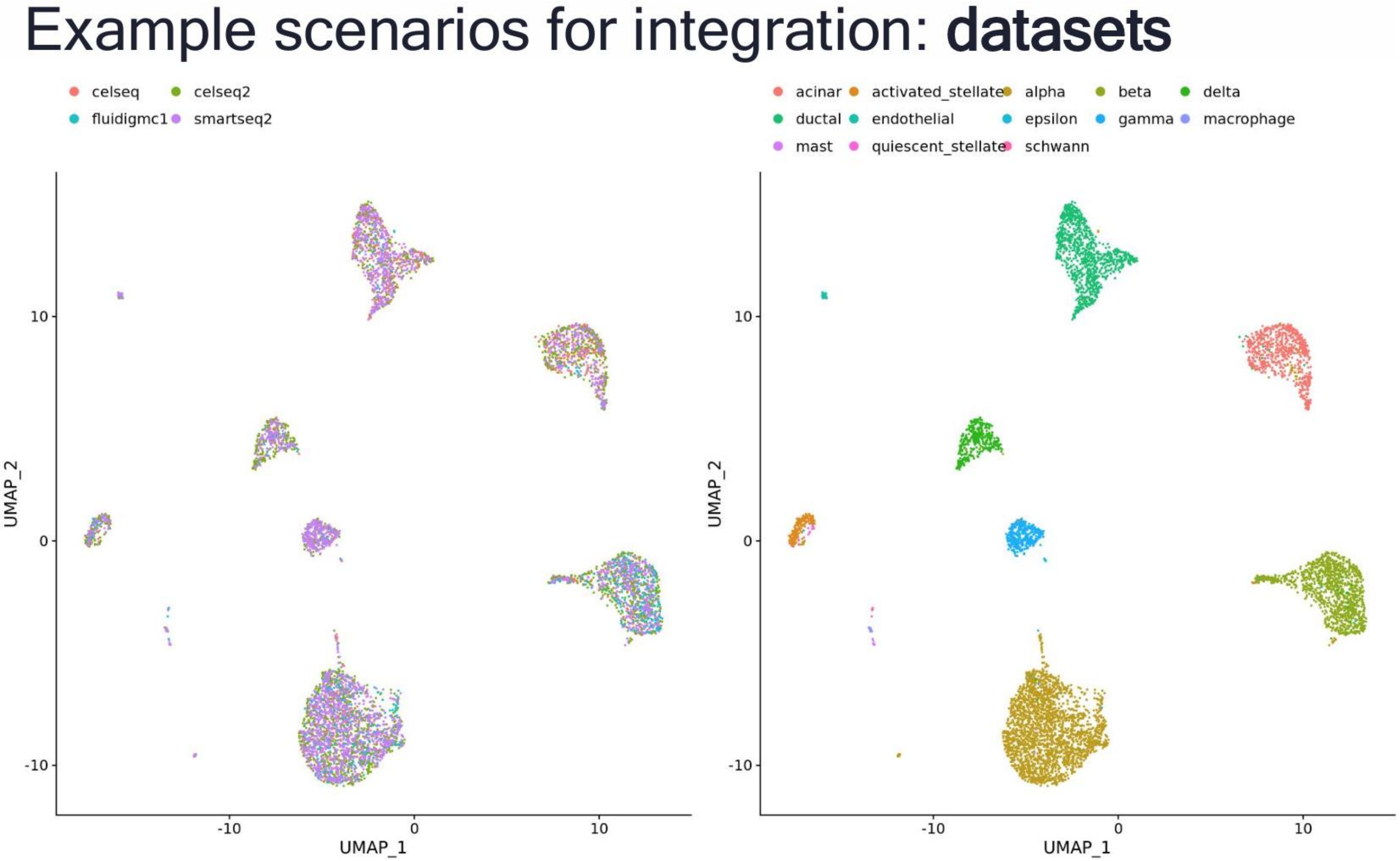
Why do we need integration?







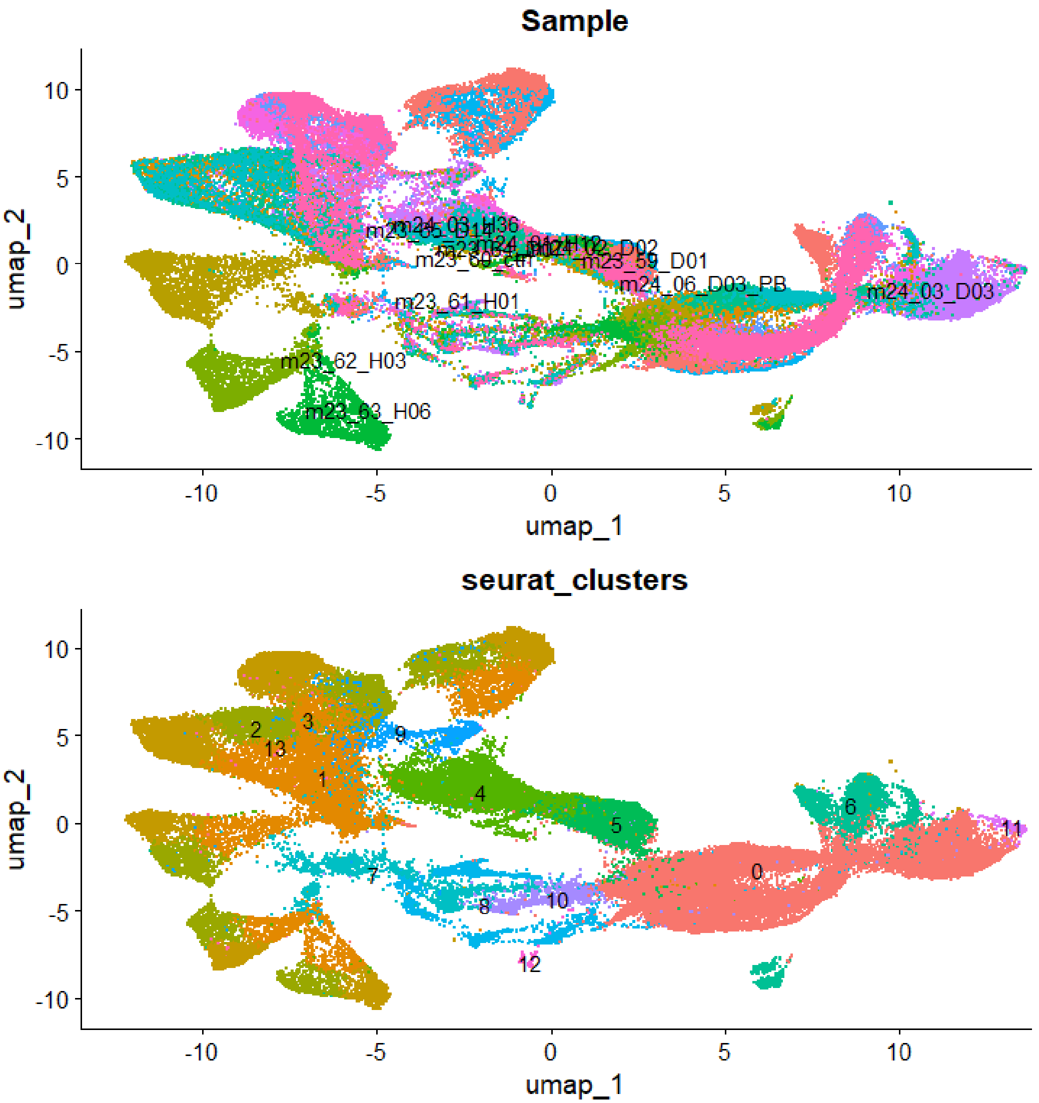
Why do we need integration?



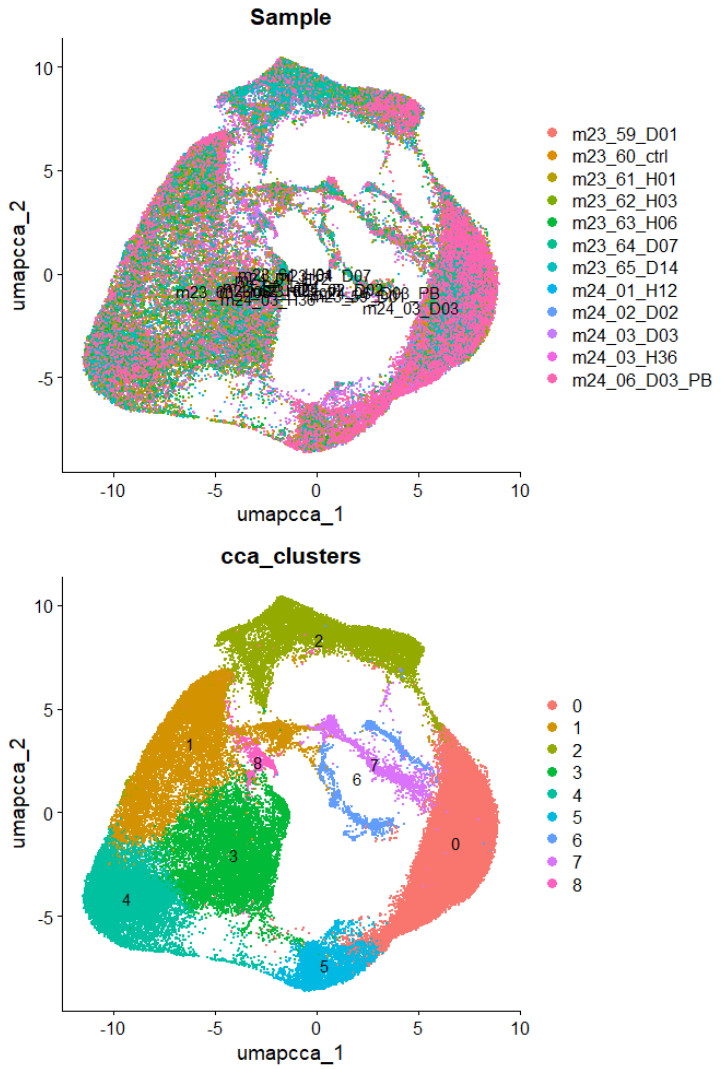
Example scenarios for integration: **datasets**

Why do we need integration?

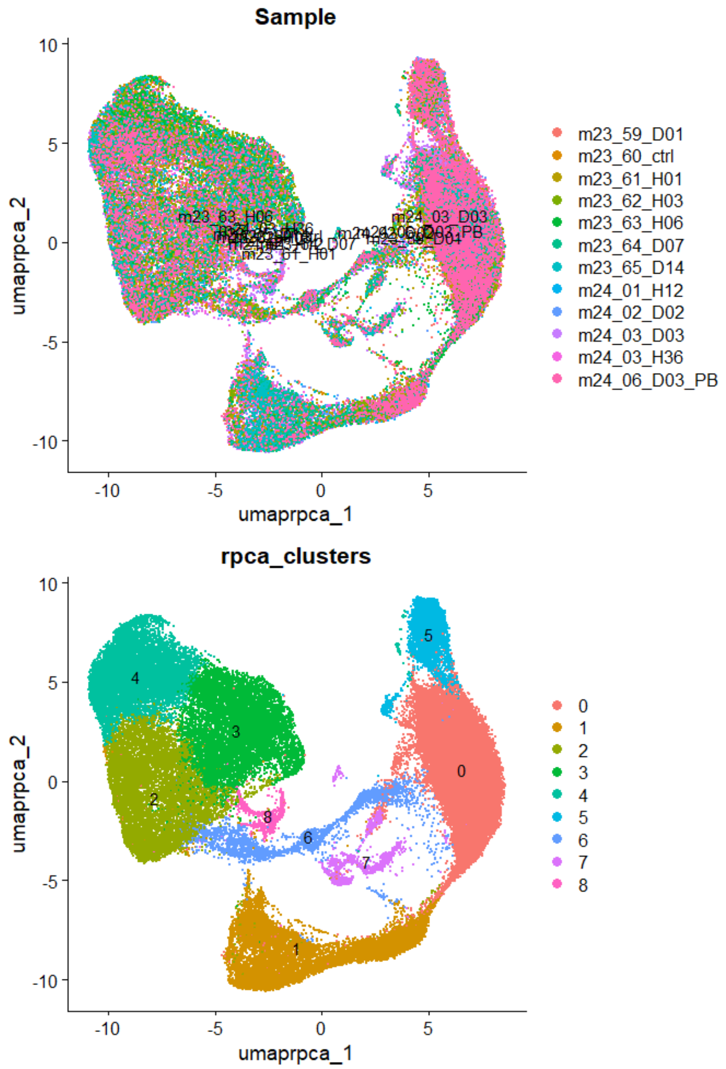
Before integration



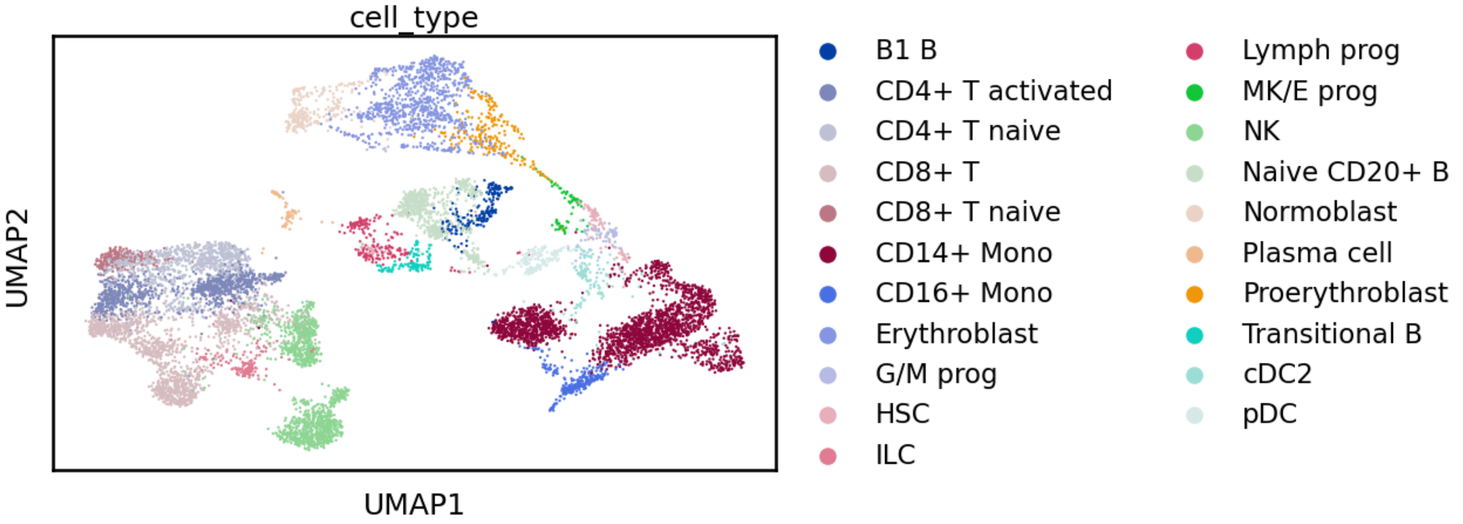
CCA integration

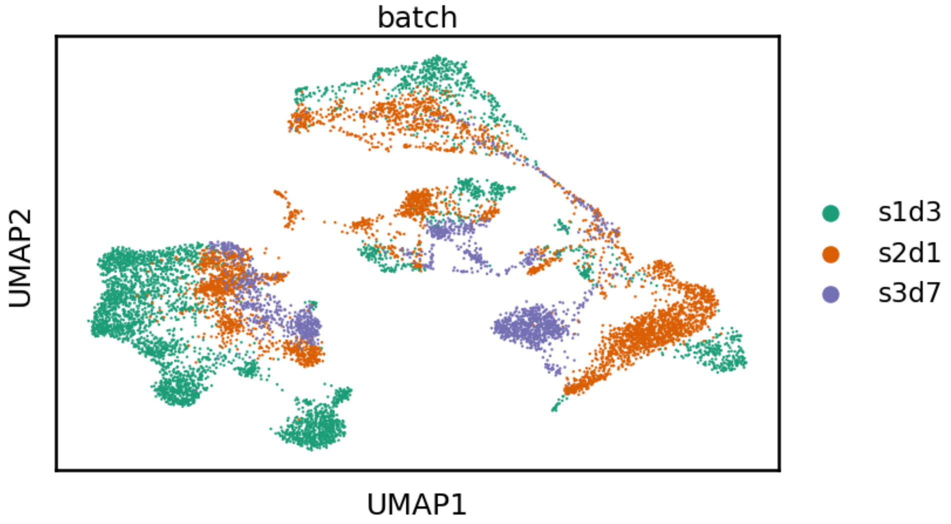


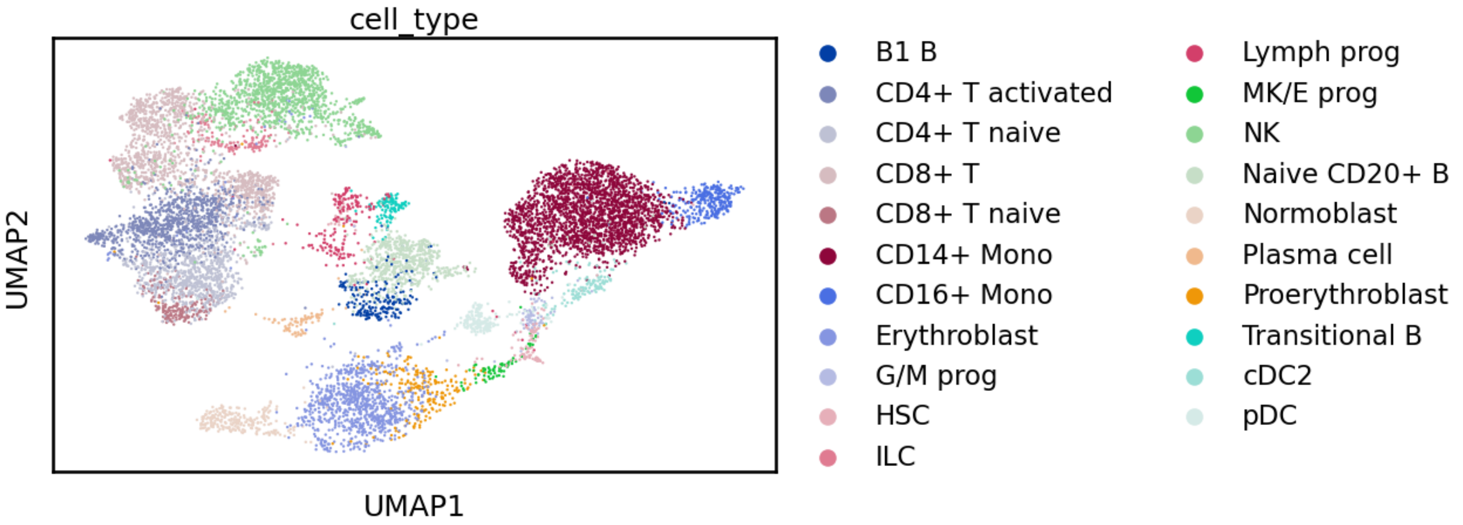
RPCA integration

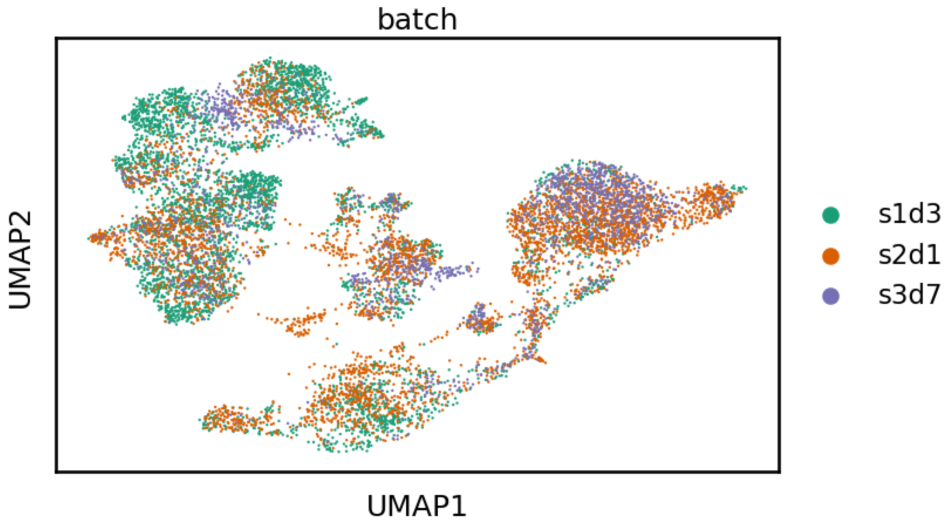


Why do we need integration?









Types of integration models

Global models

- Fit regression model with batch effect covariate

Residuals (often using linear regression):

$$\hat{n}_{gc} = f_D(B_c, \dots)$$

$$r_{gc} = n_{gc} - \hat{n}_{gc} = n_{gc} - (\beta_0 + \beta_1 B_c)$$

in linear model case

Example:

sc.tl.regress_out()

Correct for fitted batch effect:

$$n_{gcb} = \alpha_g + X\beta_g + \gamma_{gb} + \delta_{gb}\epsilon_{gcb}$$

bio design matrix

additive batch effect

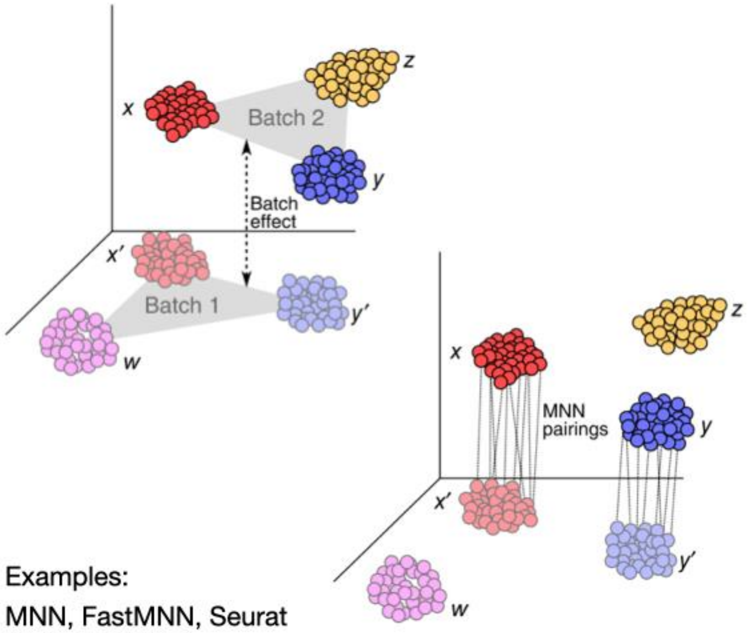
multiplicative batch effect

Example:

ComBat - *scanpy.pp.combat()*

Linear embedding models

- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector

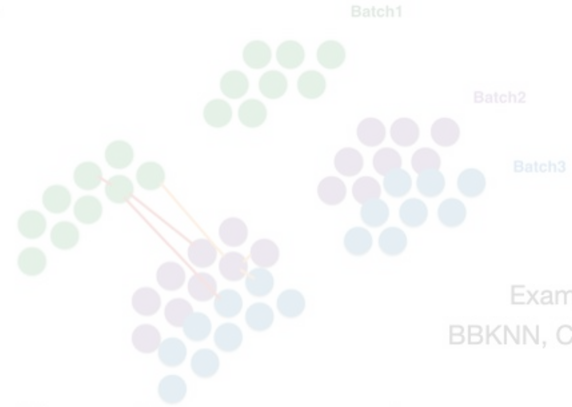


Examples:

MNN, FastMNN, Seurat v3, Scanorama

Graph-based methods & Deep learning

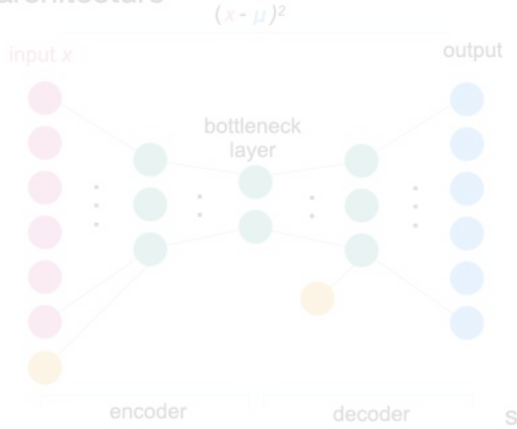
Enforce graph connections between different batches



Examples:

BBKNN, Conos

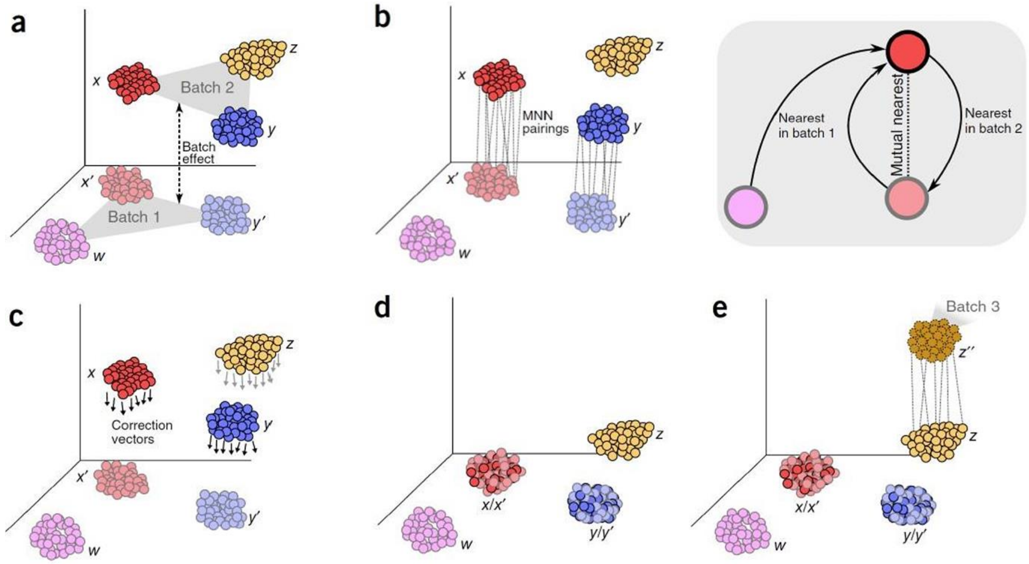
Add condition node into auto-encoder architecture

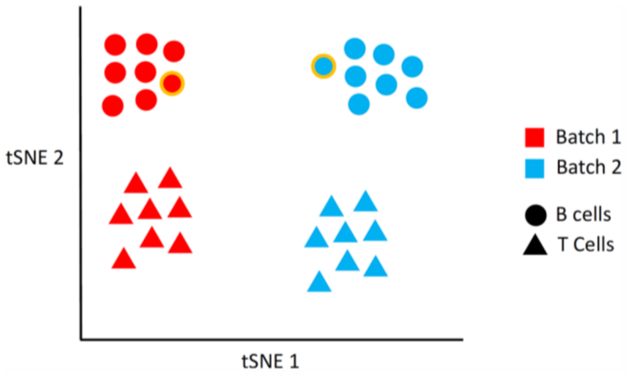


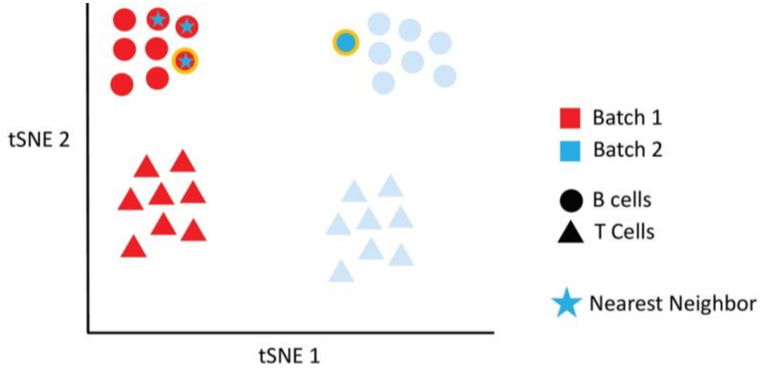
Examples:

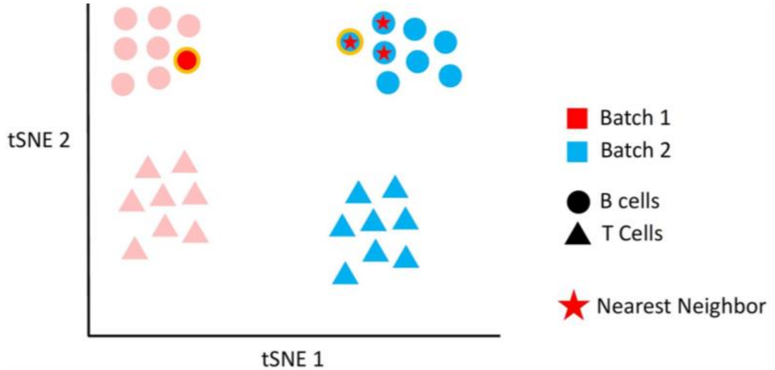
scVI, trVAE, SAUCIE

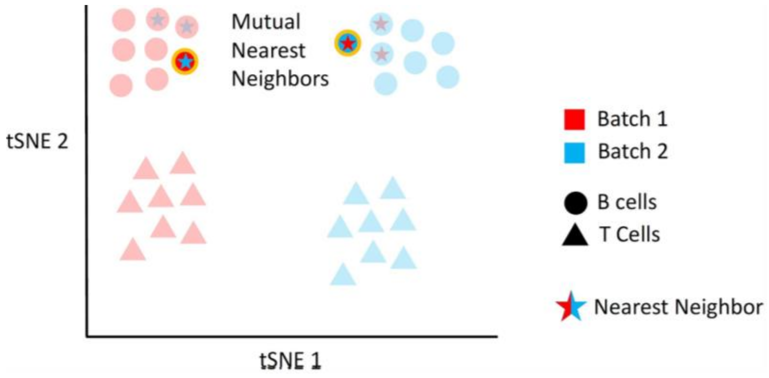
Integration using Mutual Nearest Neighbors (MNN)





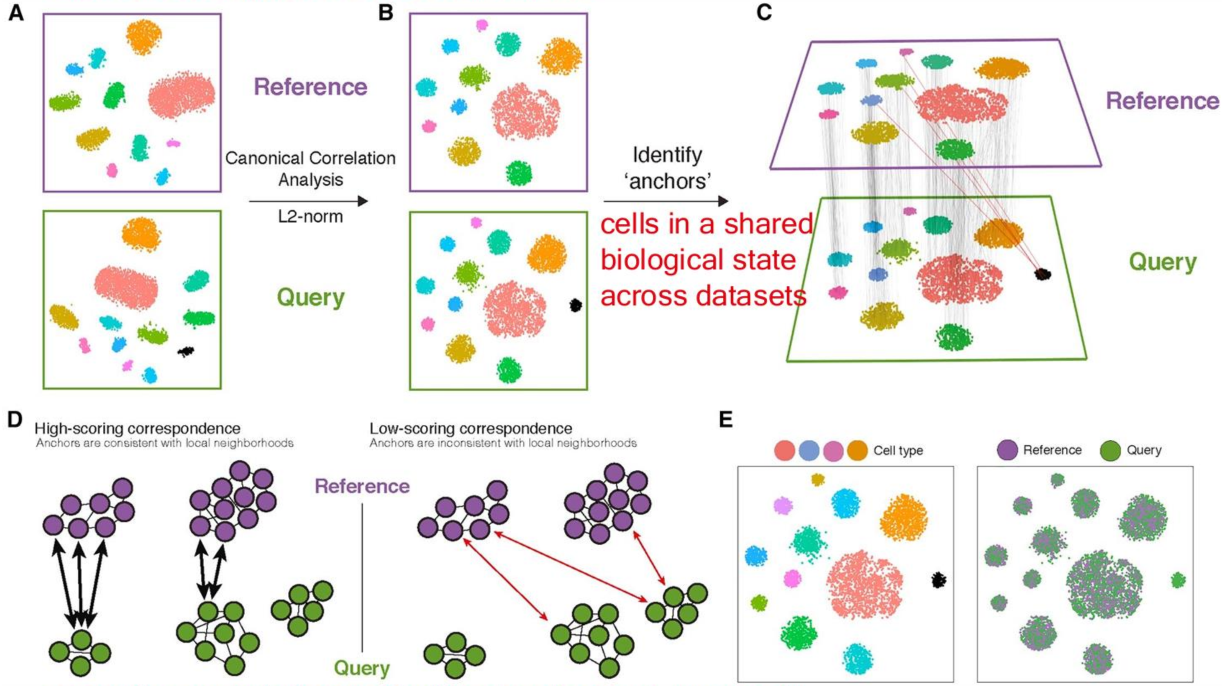






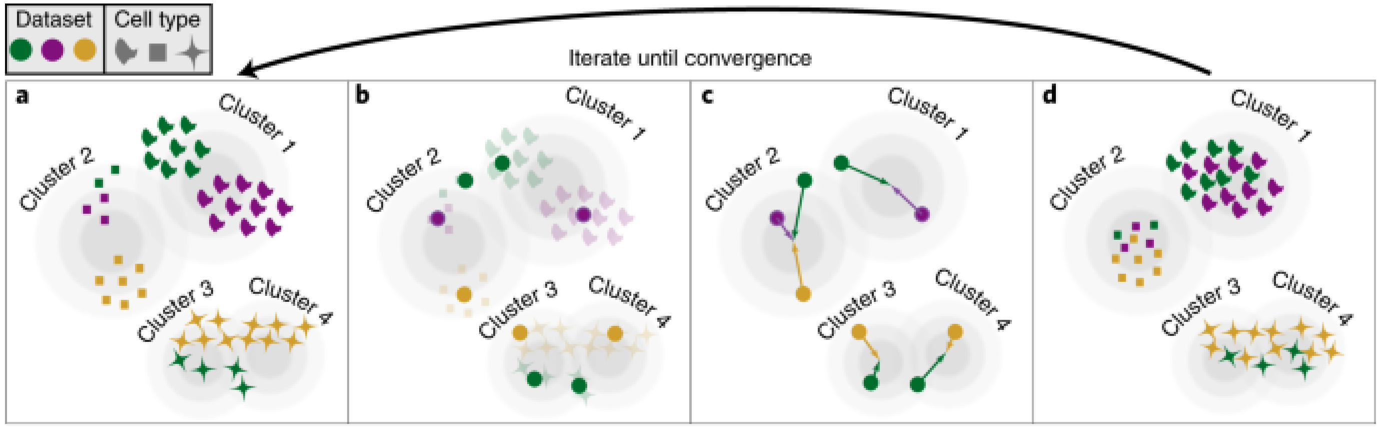
Concept of integration

Datasets coming from 2 different platforms


A
B
C
Reference
Query
Canonical Correlation Analysis
L2-norm
Identify 'anchors'
Reference
Query
D
High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query
E
Cell type
Reference
Query

cells in a shared biological state across datasets

scores to compute “correction” vectors for each query cell, transforming its expression so it can be jointly analyzed



Soft assign cells to clusters, favoring mixed dataset representation

Get cluster centroids for each dataset

Get dataset correction factors for each cluster

Move cells based on soft cluster membership

- https://satijalab.org/seurat/articles/integration_introduction
- https://www.sc-best-practices.org/cellular_structure/integration.html
- https://www.singlecellcourse.org/biological-analysis.html#clustering-introduction
- https://bioconductor.org/books/3.12/OSCA/clustering.html#k-means-clustering
- https://github.com/quadbio/scRNAseq_analysis_vignette/blob/master/Tutorial.md#step-2-3-data-integration-using-liger