Swiss Institute of Bioinformatics

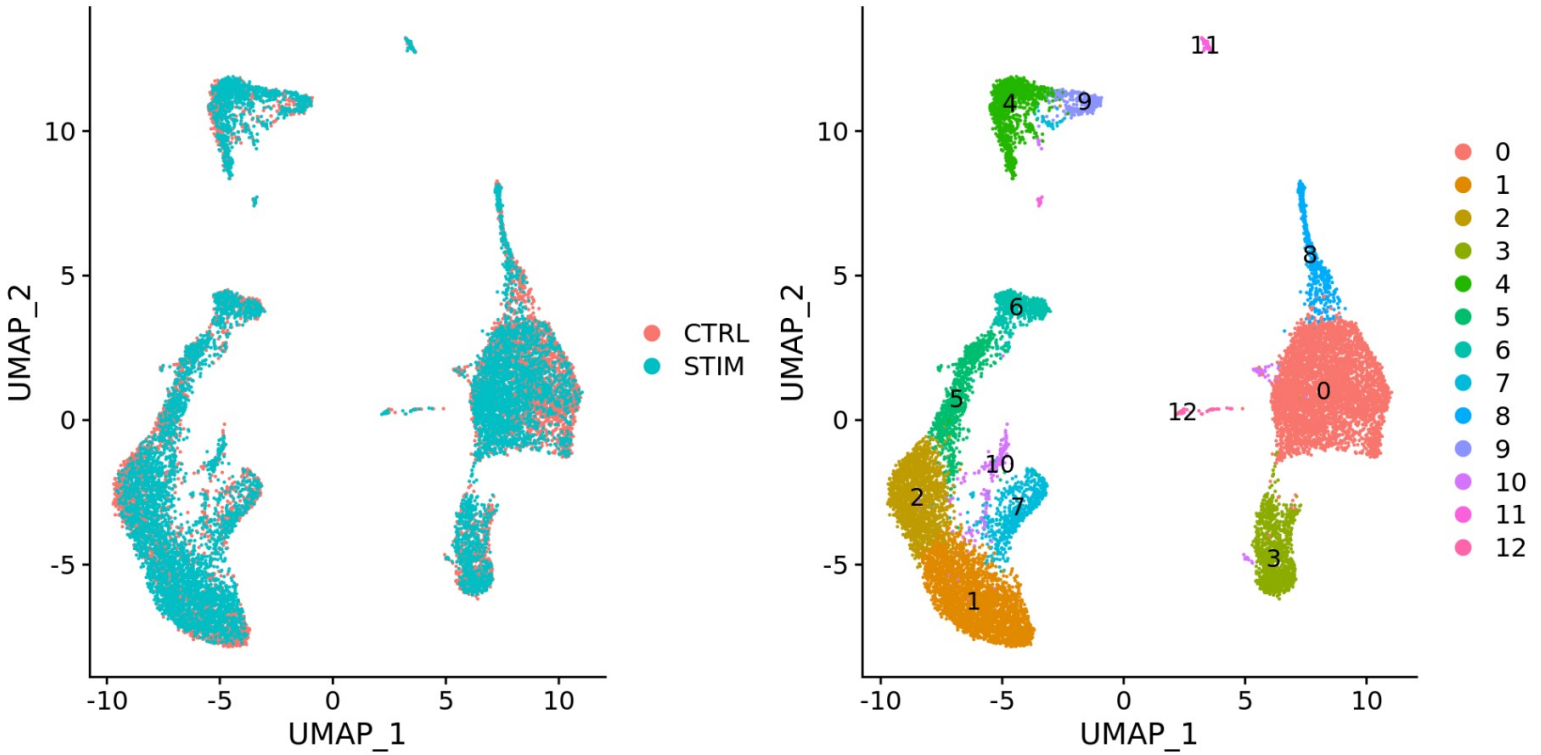
Integration

Luciano Cascione, PhD
Bioinformatics Core Unit

LUCIANO CASCIONE, PHD
BELLINZONA, OCT. 30TH 2024

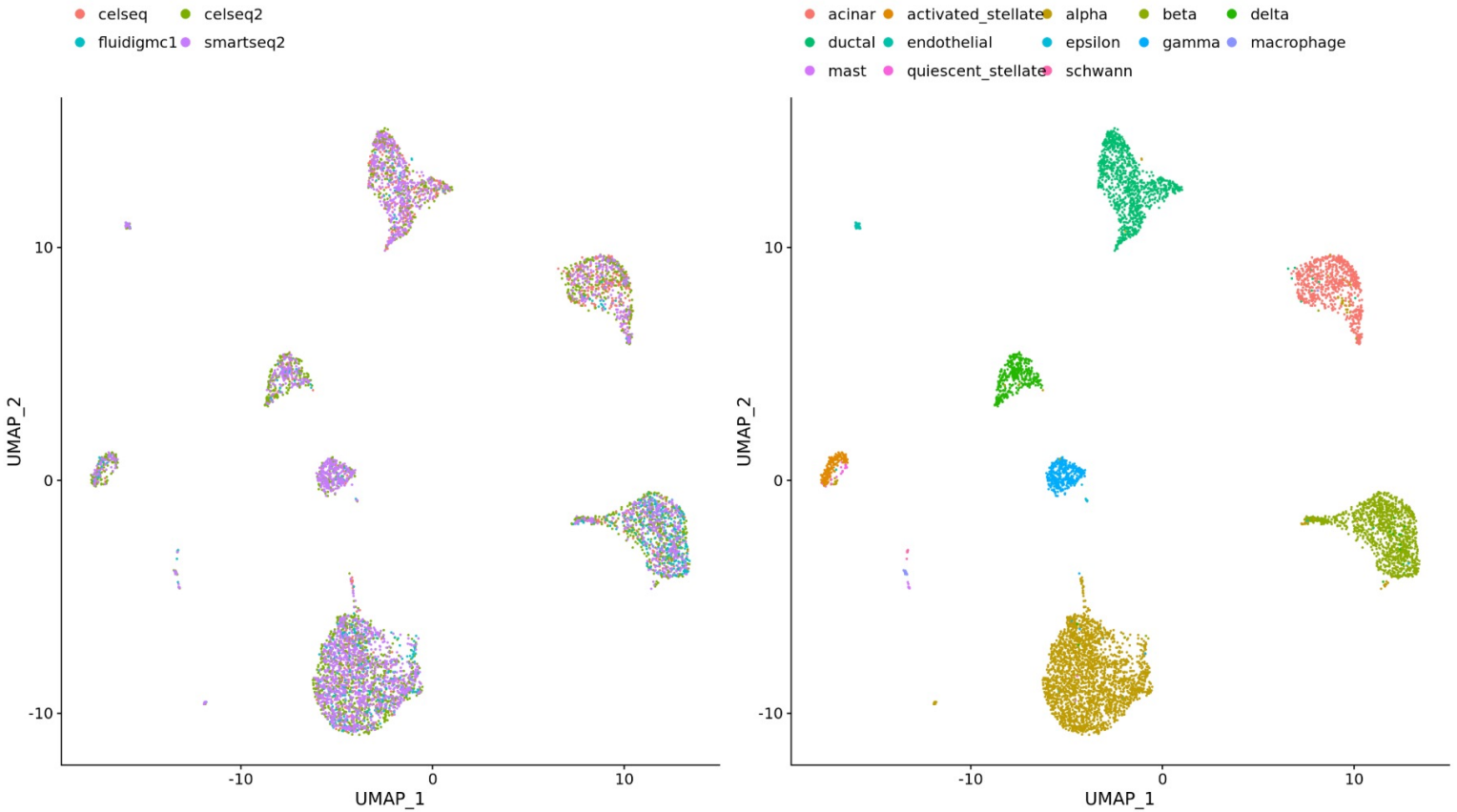
What for ?

Goal: identify shared subpopulations across conditions or datasets



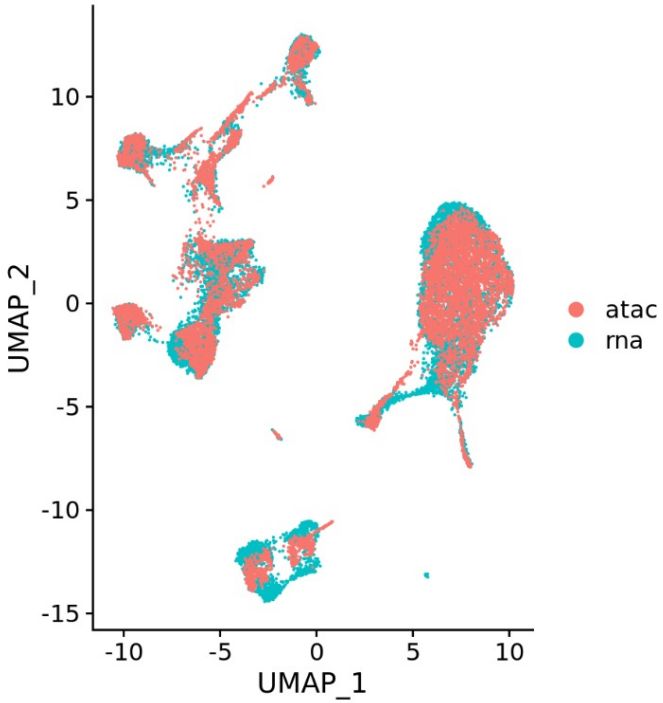
What for ?

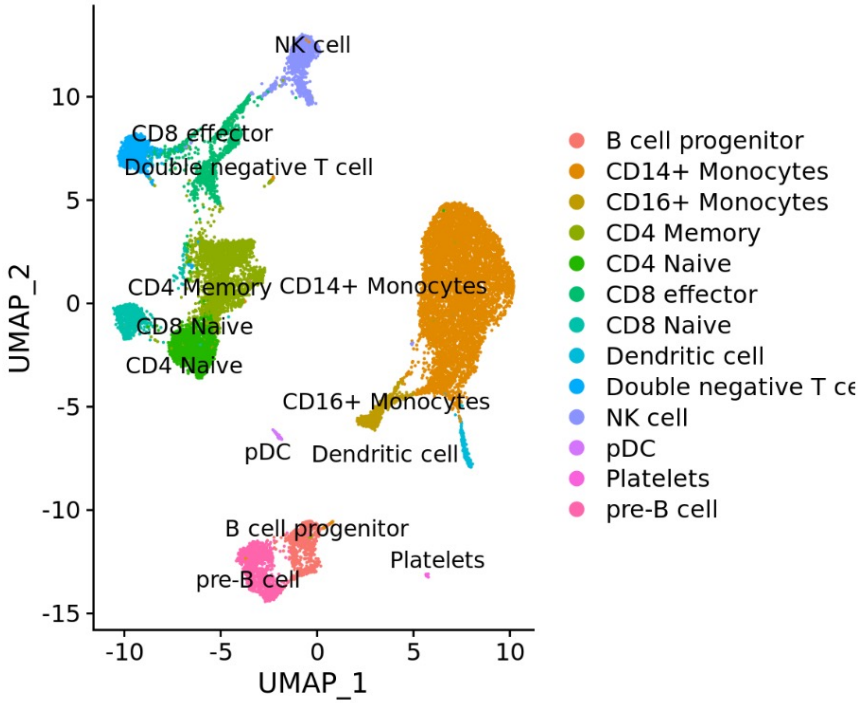
Goal: identify shared subpopulations across conditions or datasets

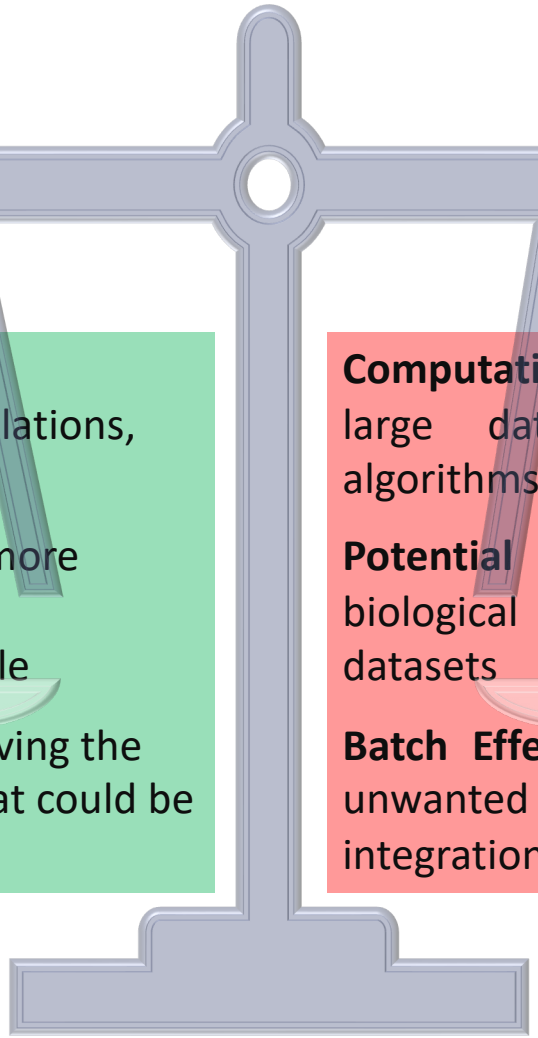


What for ?

Goal: identify shared subpopulations across conditions or datasets enabling comprehensive analysis







Pro

Enhanced Resolution: a more comprehensive view of cell populations, e.g. identify rare cell types

Improved Robustness: findings more robust across different biological conditions and more generalizable

Greater Statistical Power: improving the ability to detect subtle trends that could be missed in smaller datasets.

Cons

Computational Complexity: Integrating large datasets requires sophisticated algorithms

Potential Loss of Information: Masking biological signals specific to individual datasets

Batch Effects: batches effects introduce unwanted variability that complicates integration and analysis.

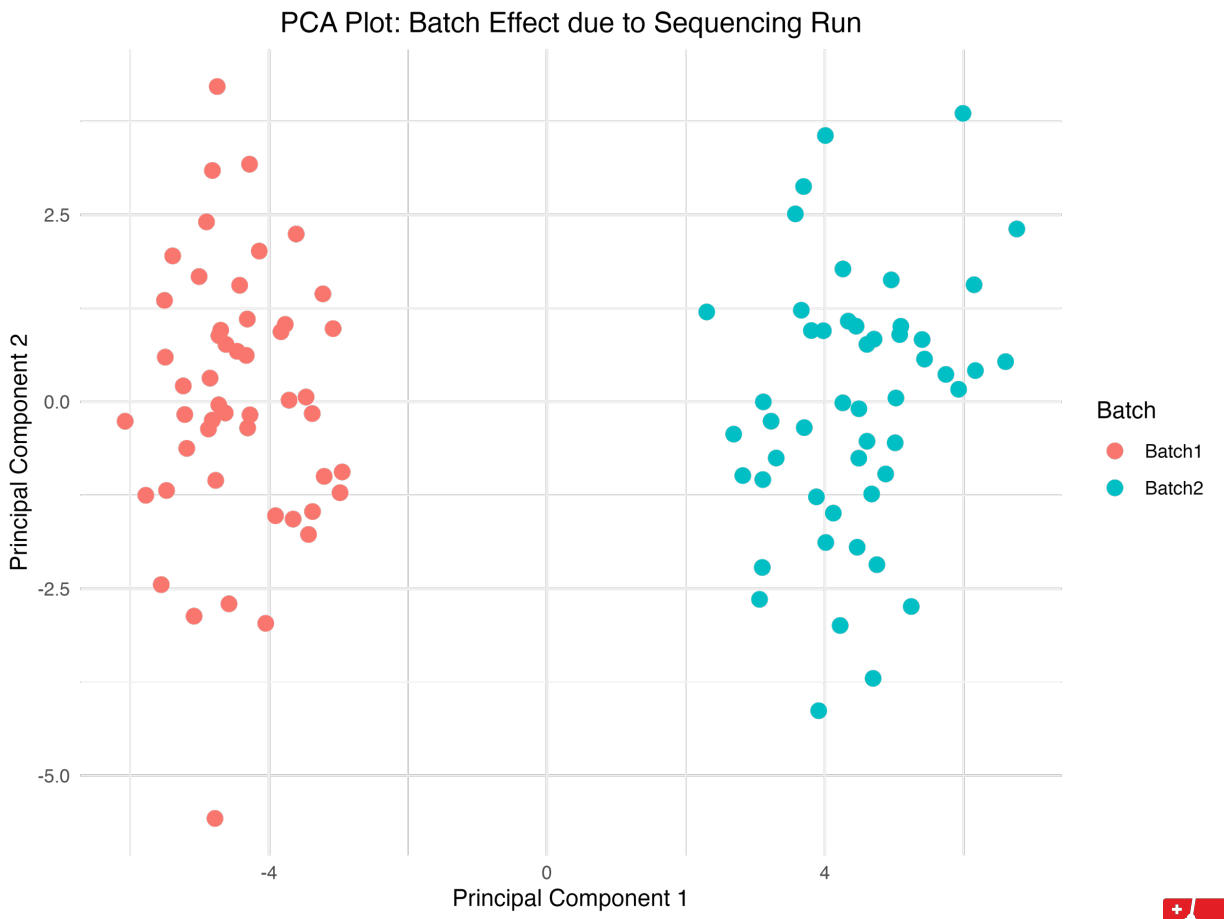
Unwanted Sources of Variation

Batch Effect is systematic techincal variations due to differences in:

a) cell isolation and handling protocols,

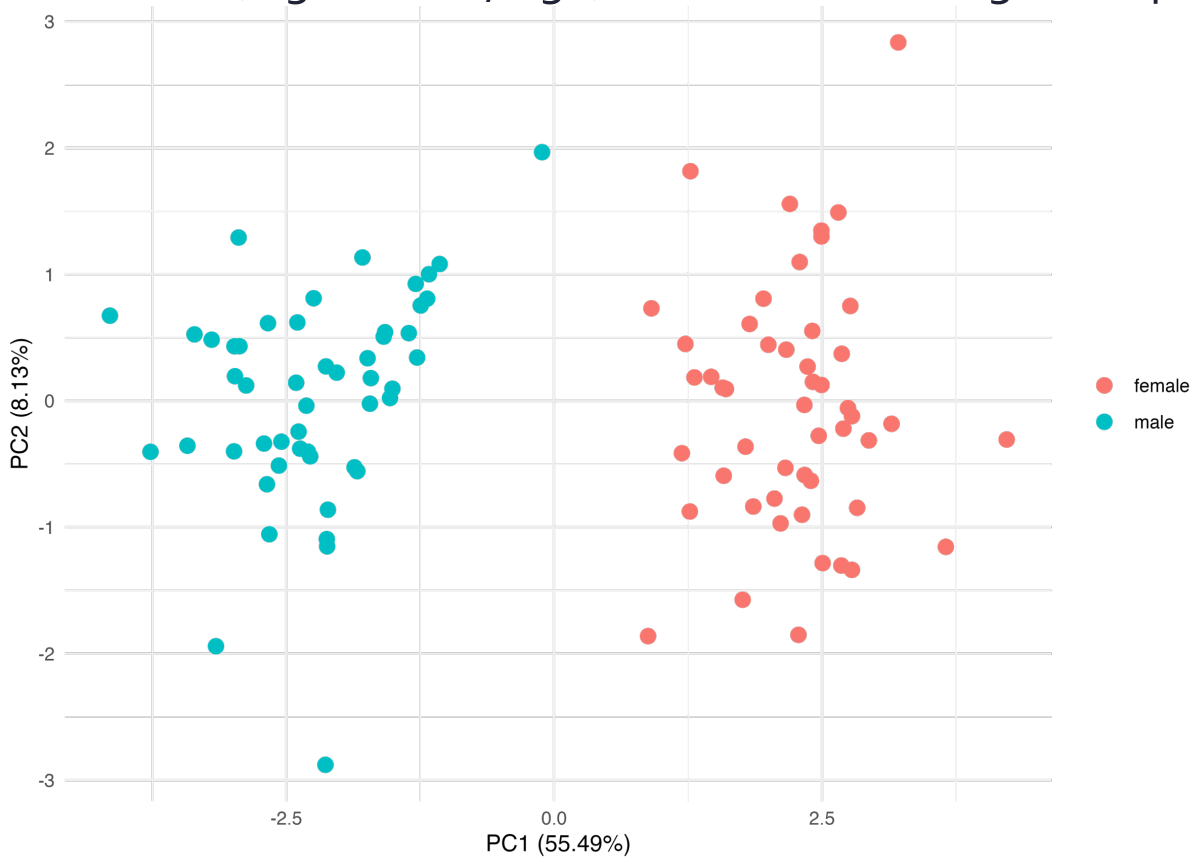
b) library preparation technology, and sequencing platforms

Batch effects can obscure true biological signals, making it difficult to compare datasets



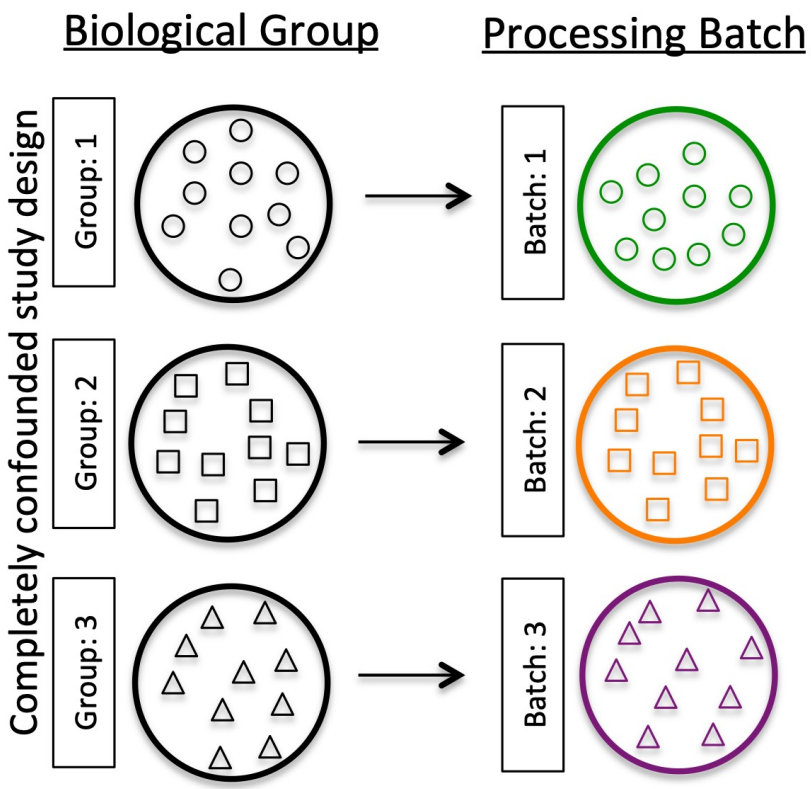
Unwanted Sources of Variation

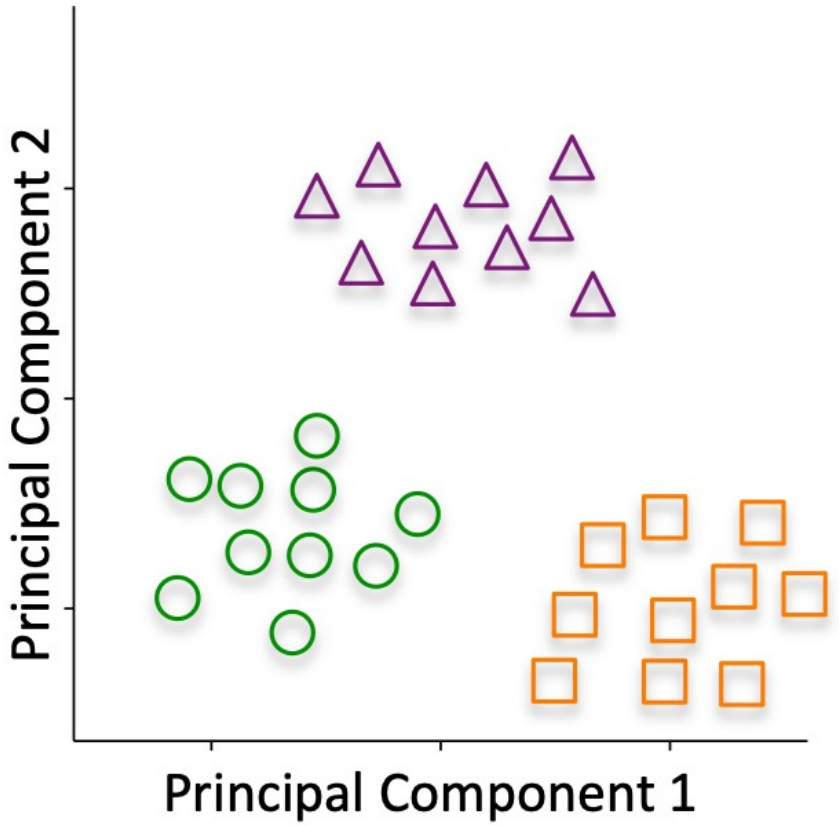
Confounders are variables (e.g. Gender, Age) that could influce gene expression



If they are not properly accounted for in the analysis they could potentially lead to misleading associations.

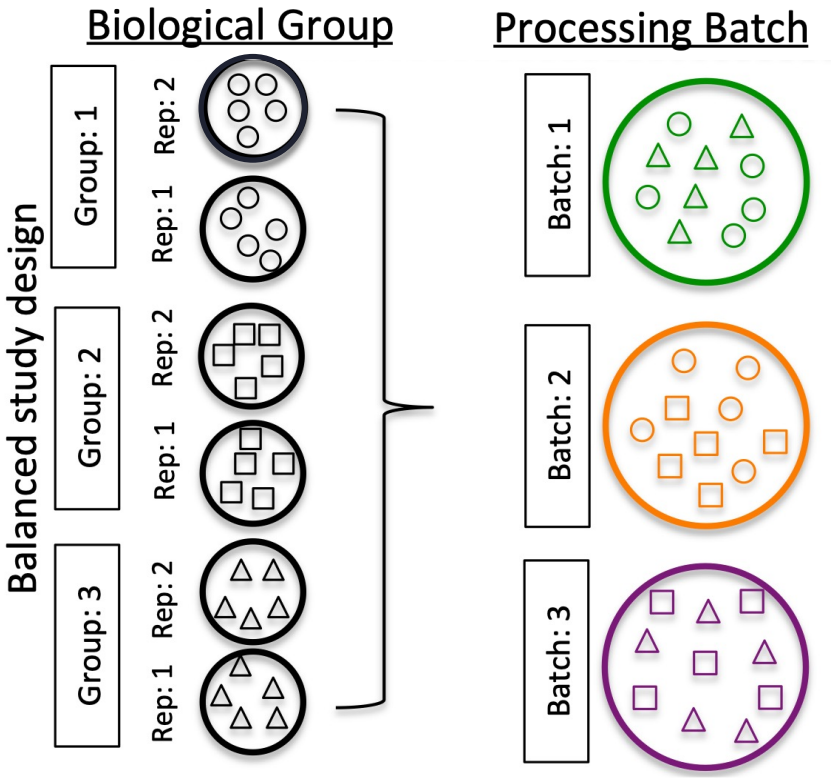
Experimental Design metters

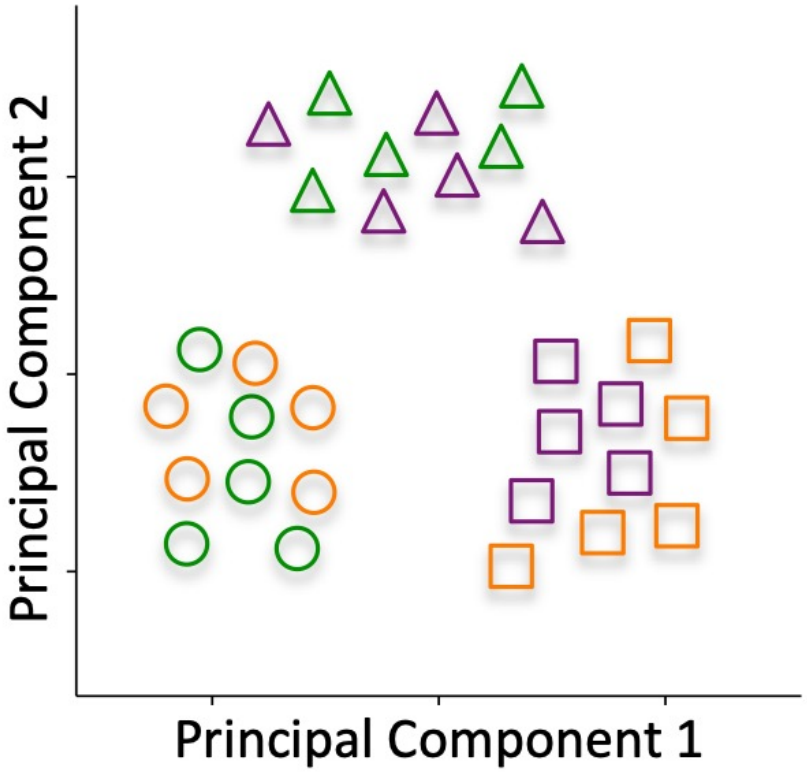




Adapted from: Hicks et al.

Experimental Design metters



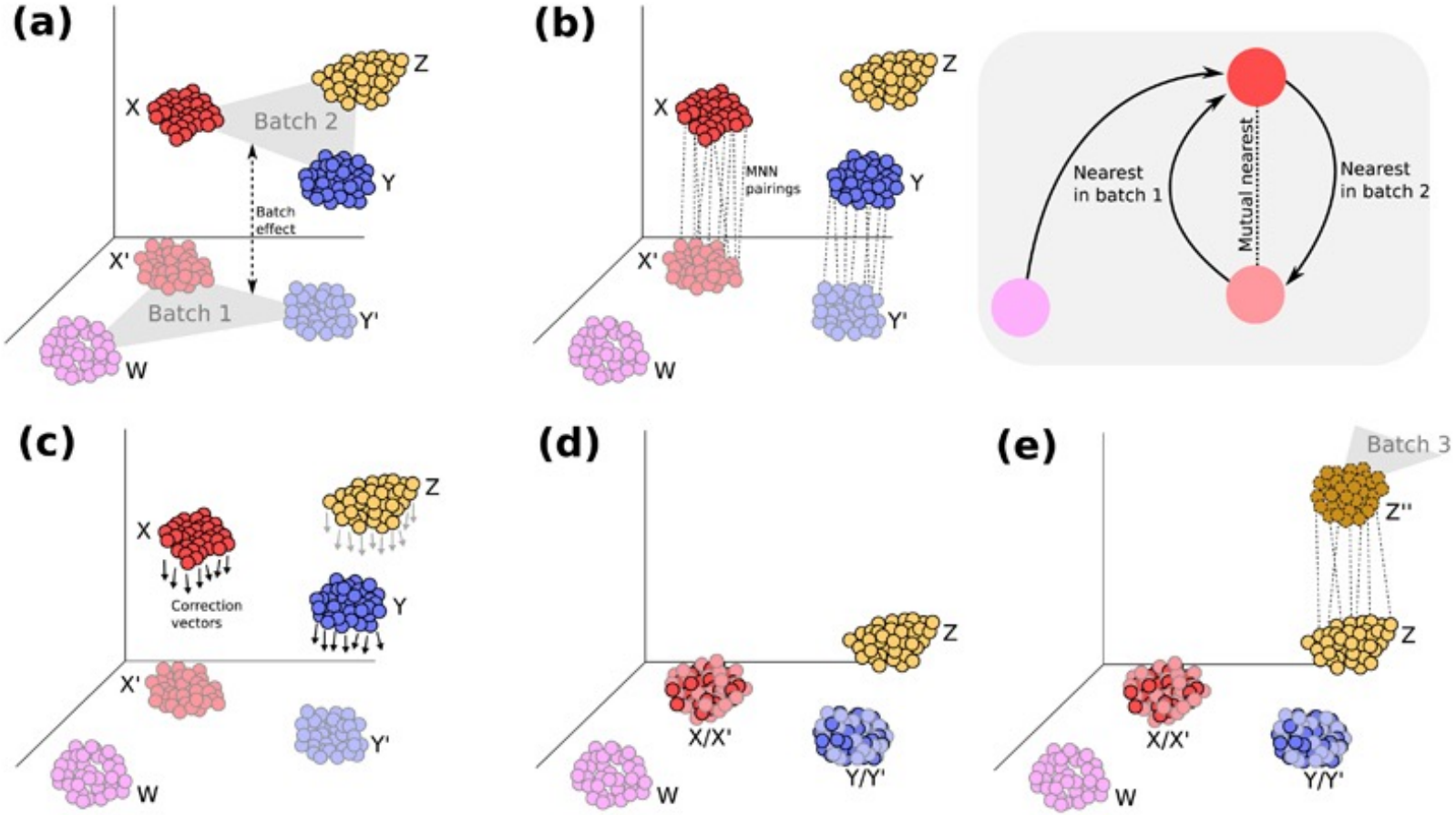


Hicks et al.

How to integrate

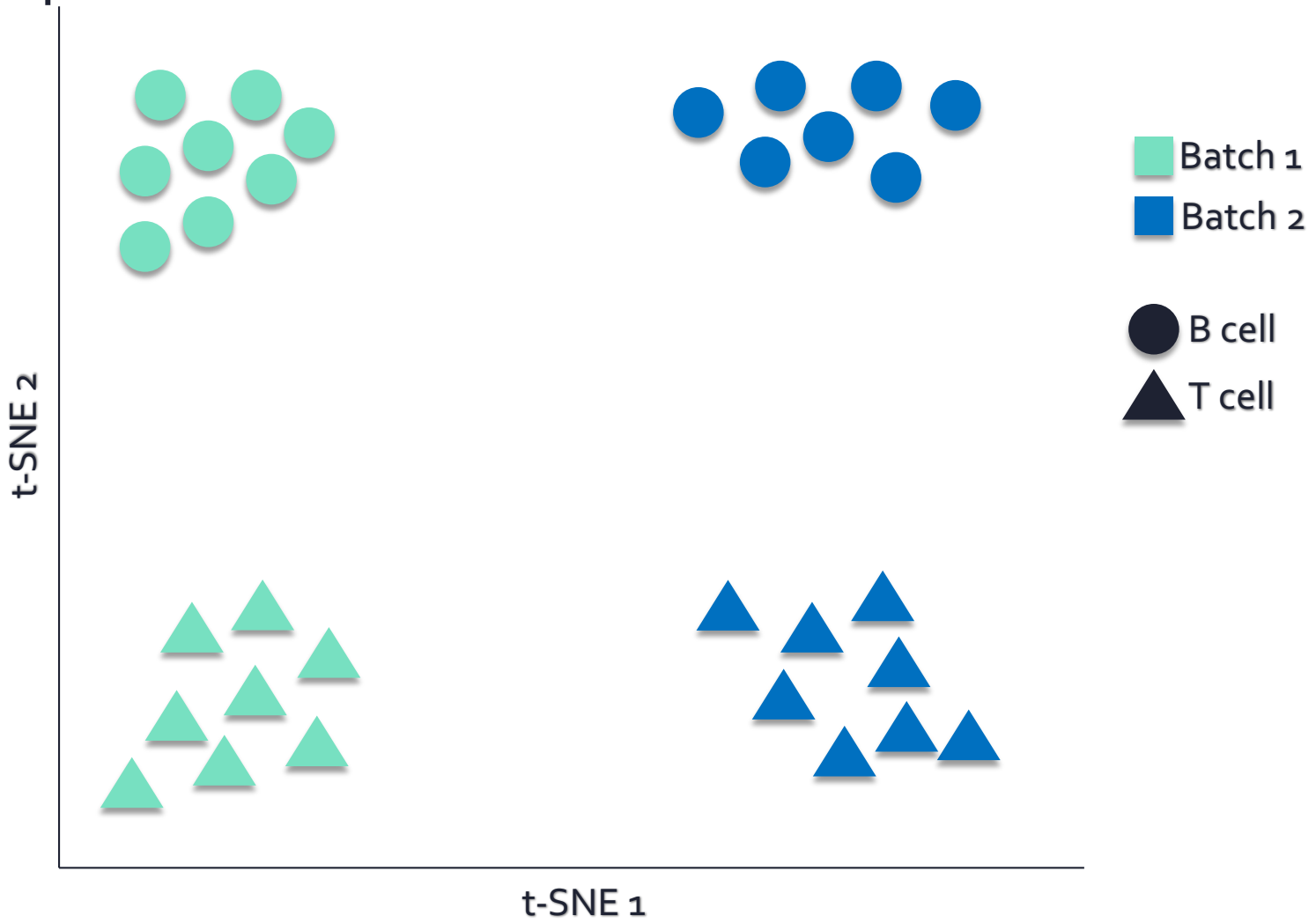
1. Find corresponding cells across datasets (by computing a **distance between cells** in a certain space)
2. Compute a data adjustment based on correspondences between cells
3. Apply the adjustment

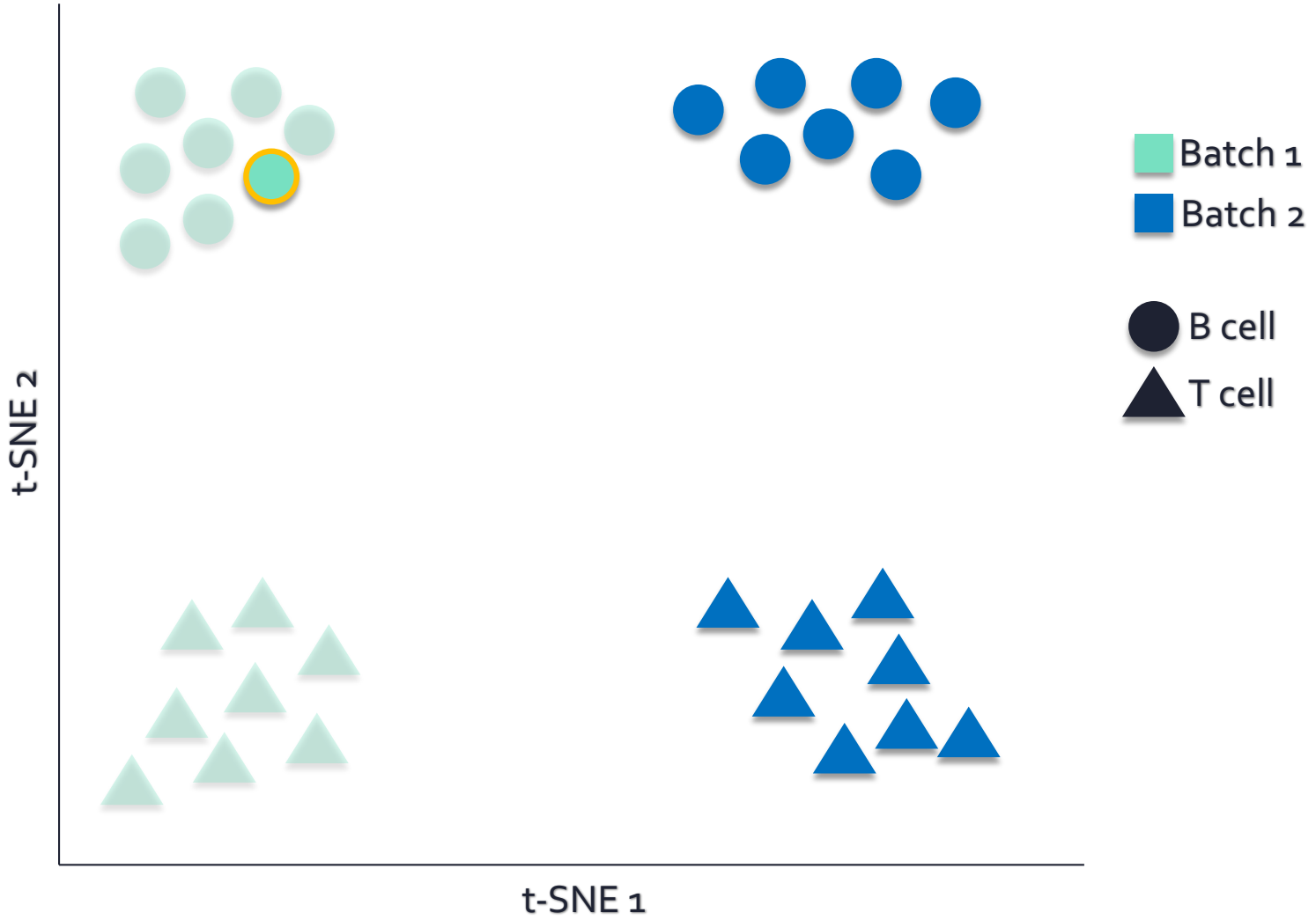
Mutual Nearest Neighbours

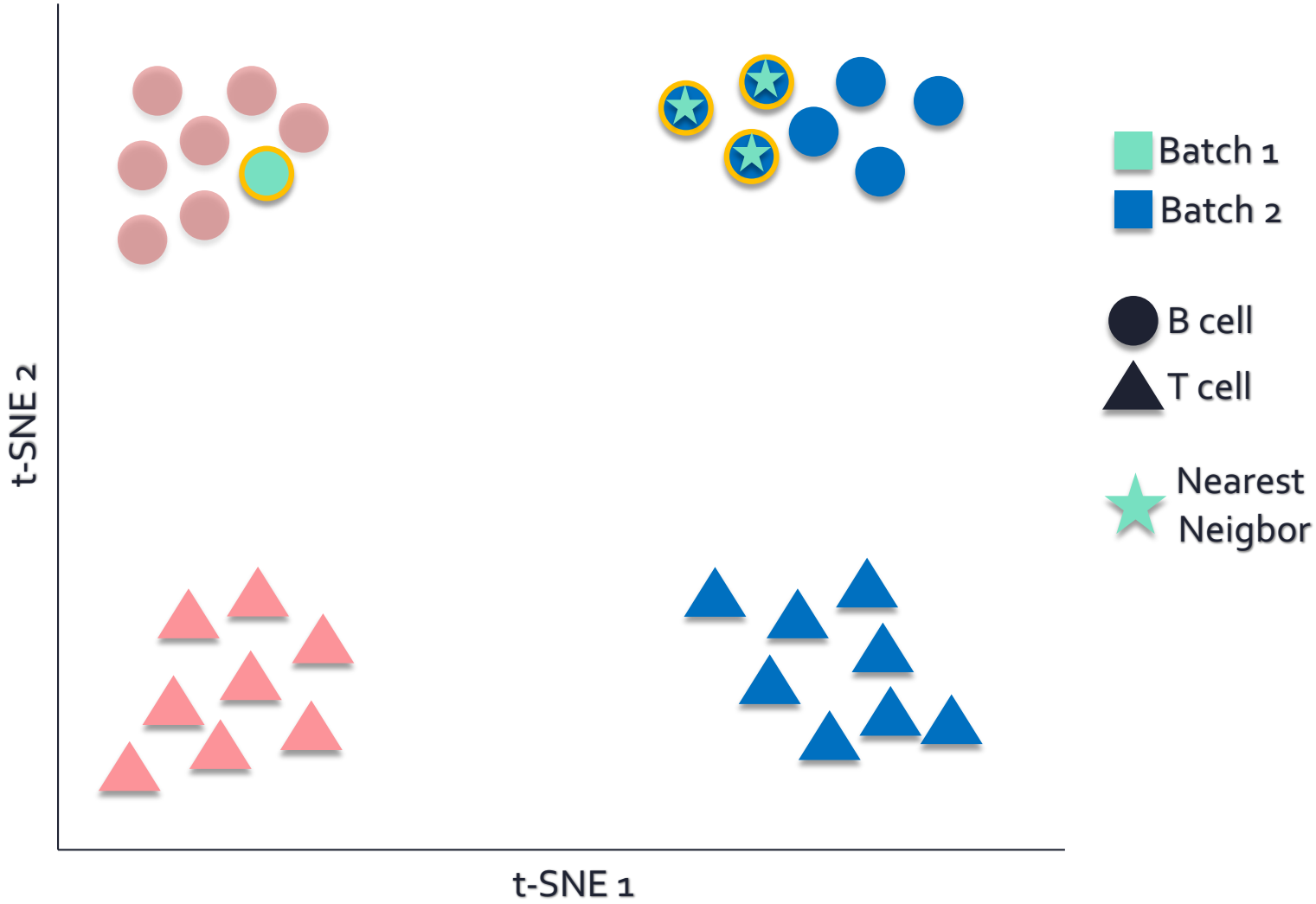


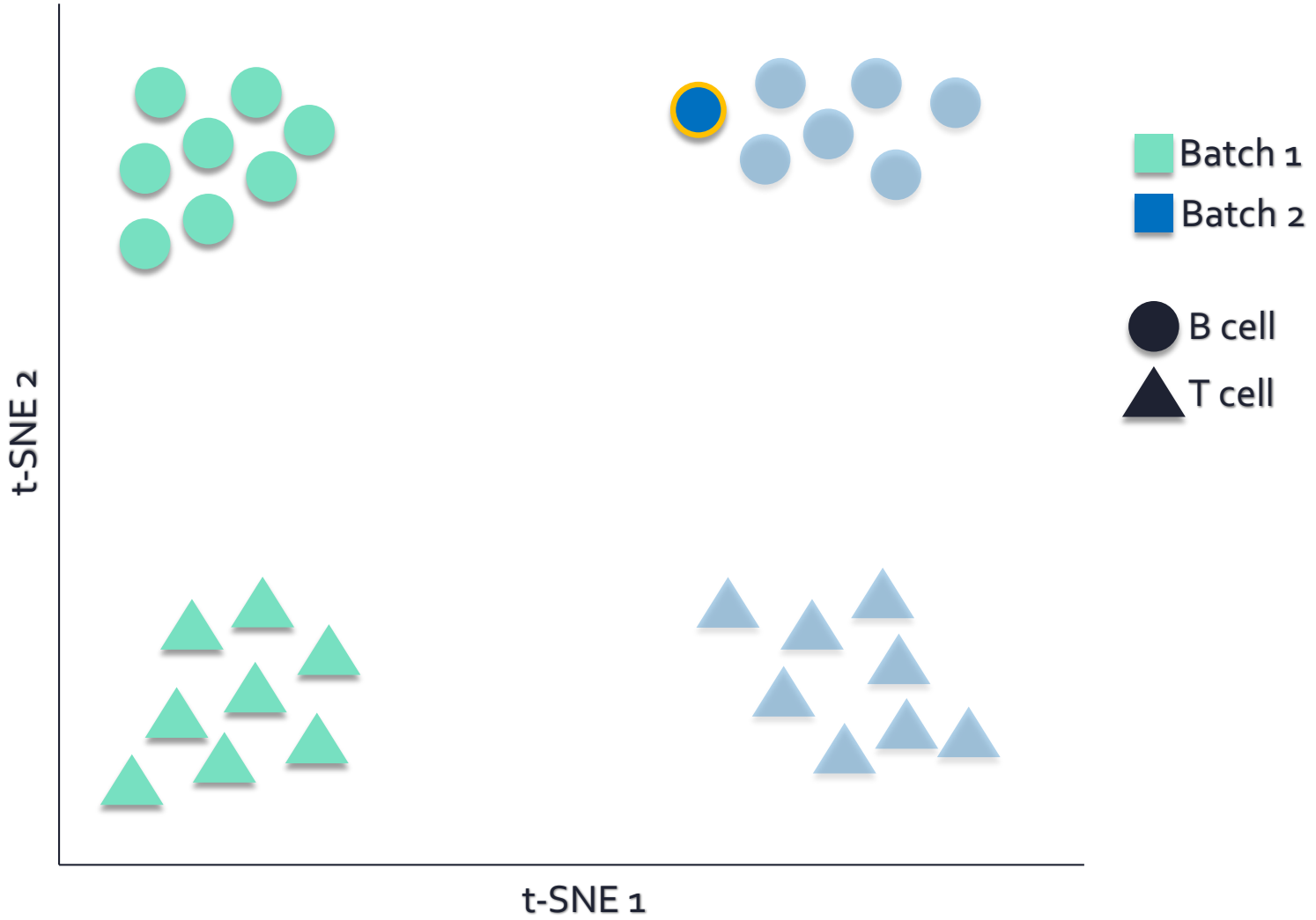
Adapted from: Haghverdi et al. Nat Biotechnol. 2018

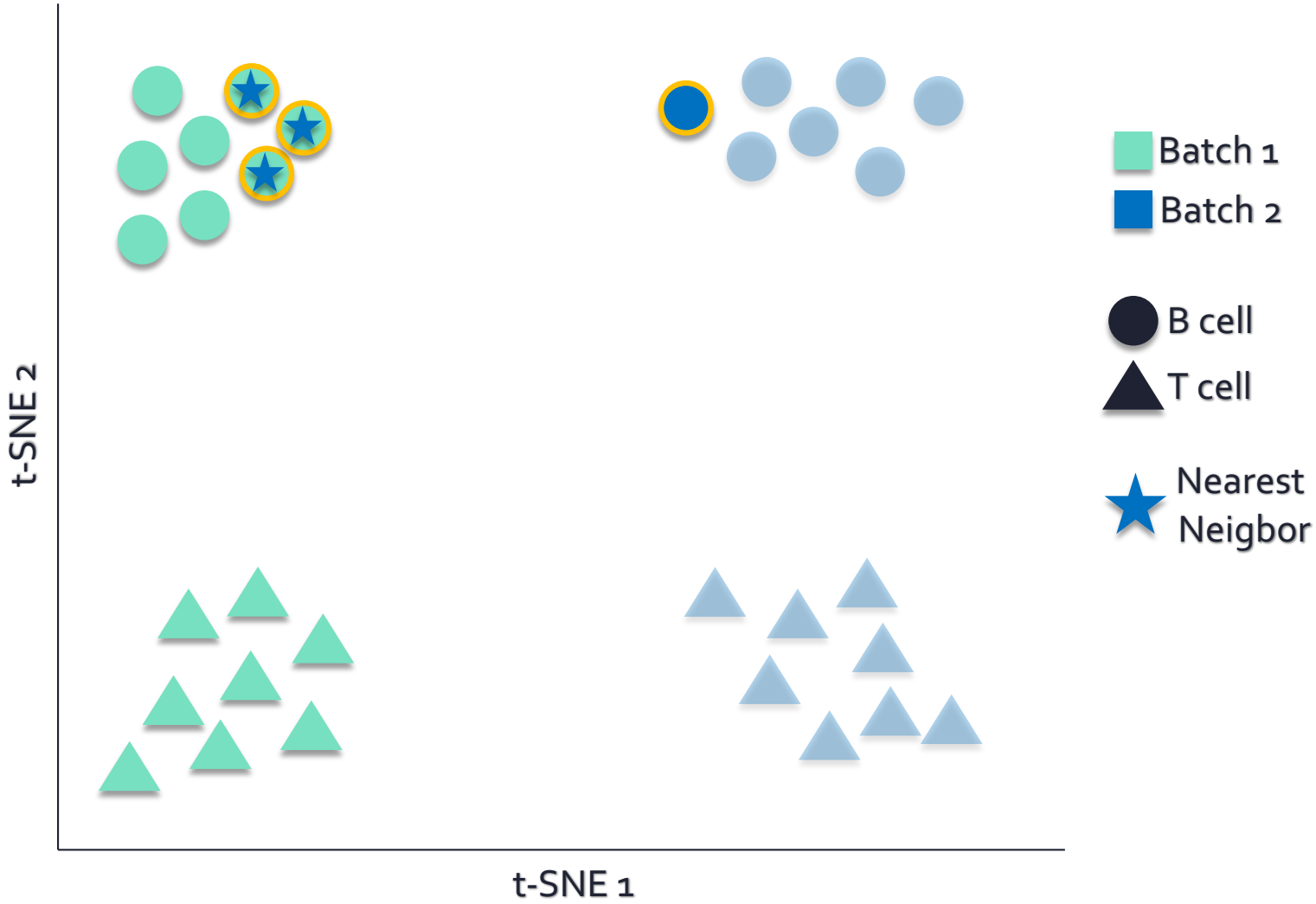
Example

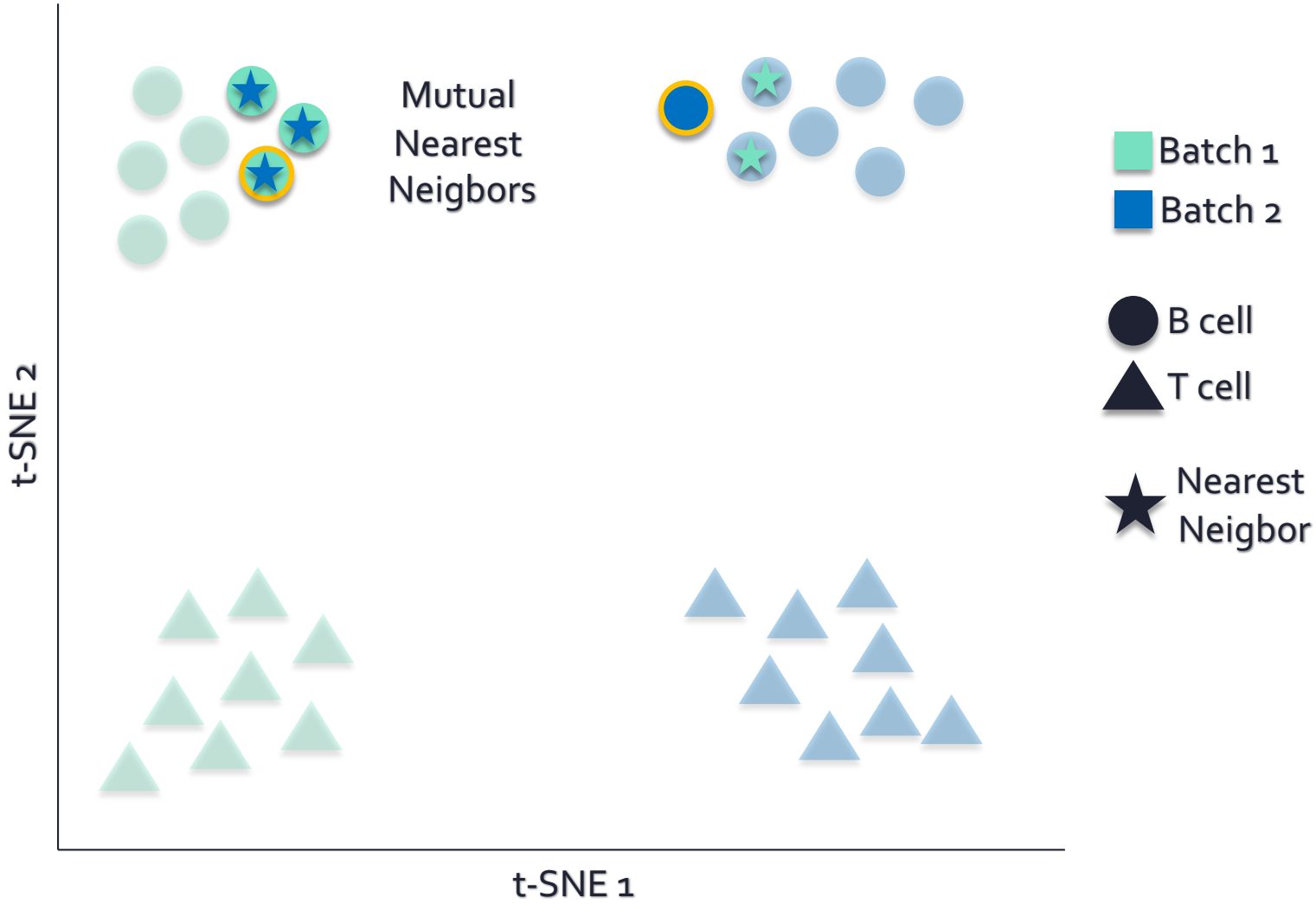



t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell

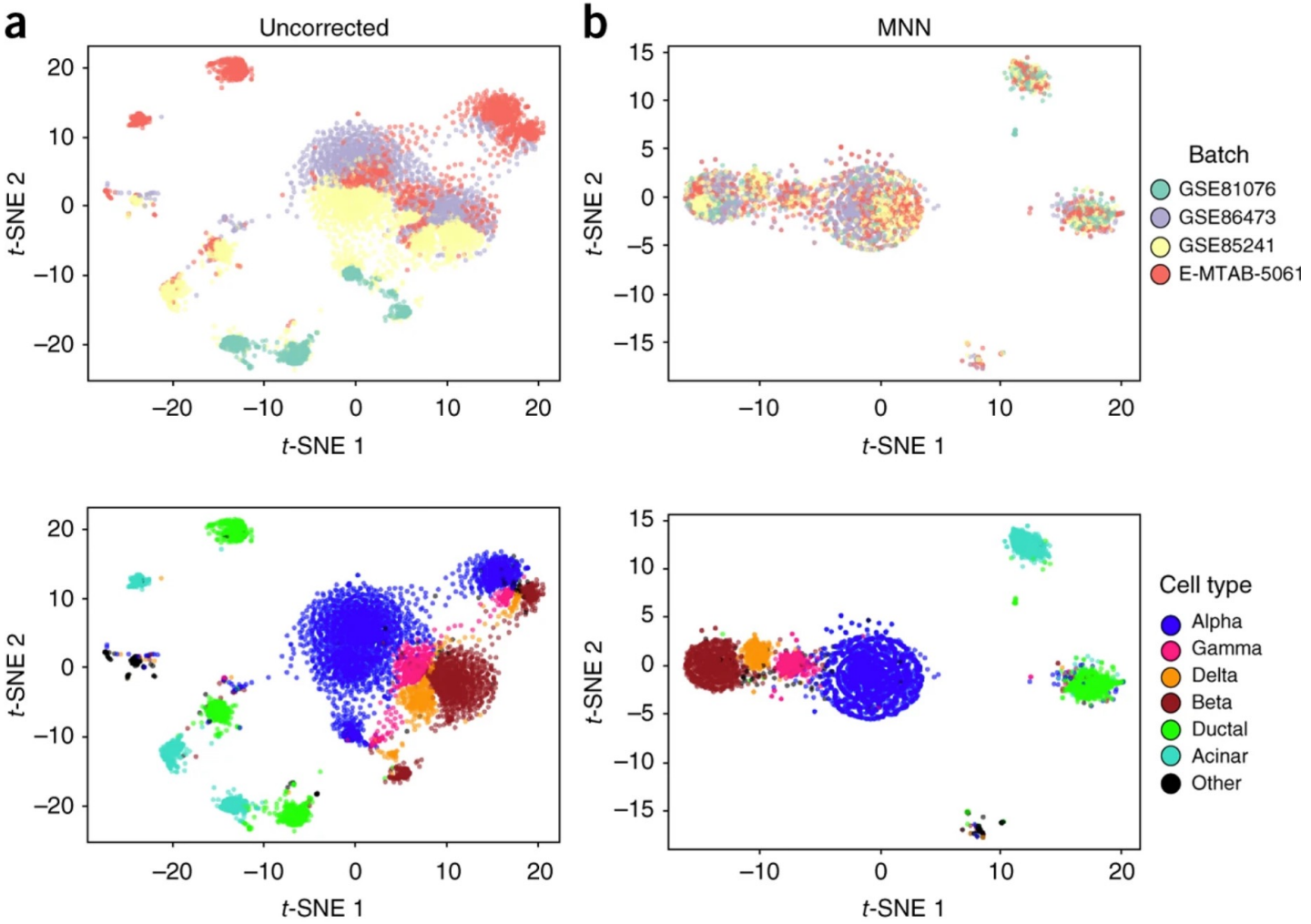

t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell
Nearest
Neigbor


t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell


t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell
Nearest
Neigbor


Mutual
Nearest
Neigbors
Batch 1
Batch 2
B cell
T cell
Nearest
Neigbor
t-SNE 2
t-SNE 1

Final Integration

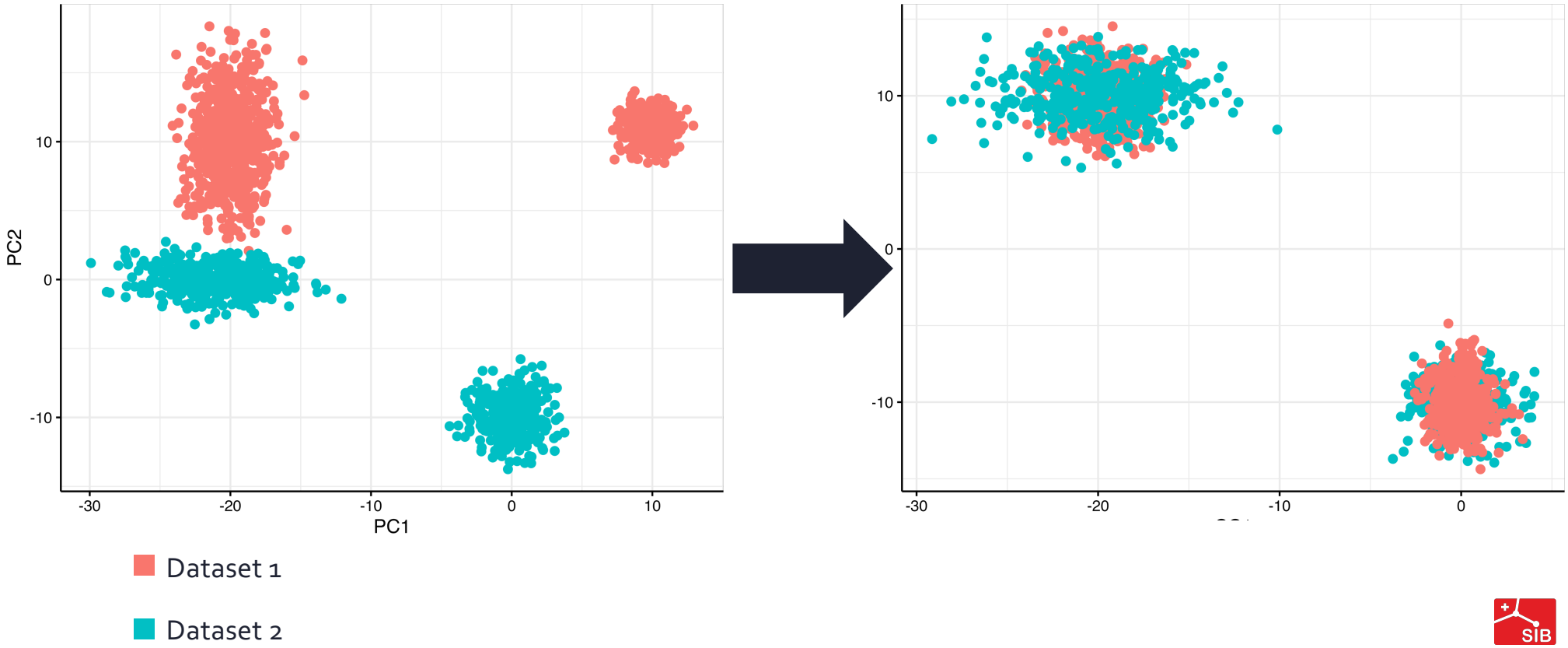


Canonical Correlation Analysis (CCA) + anchors

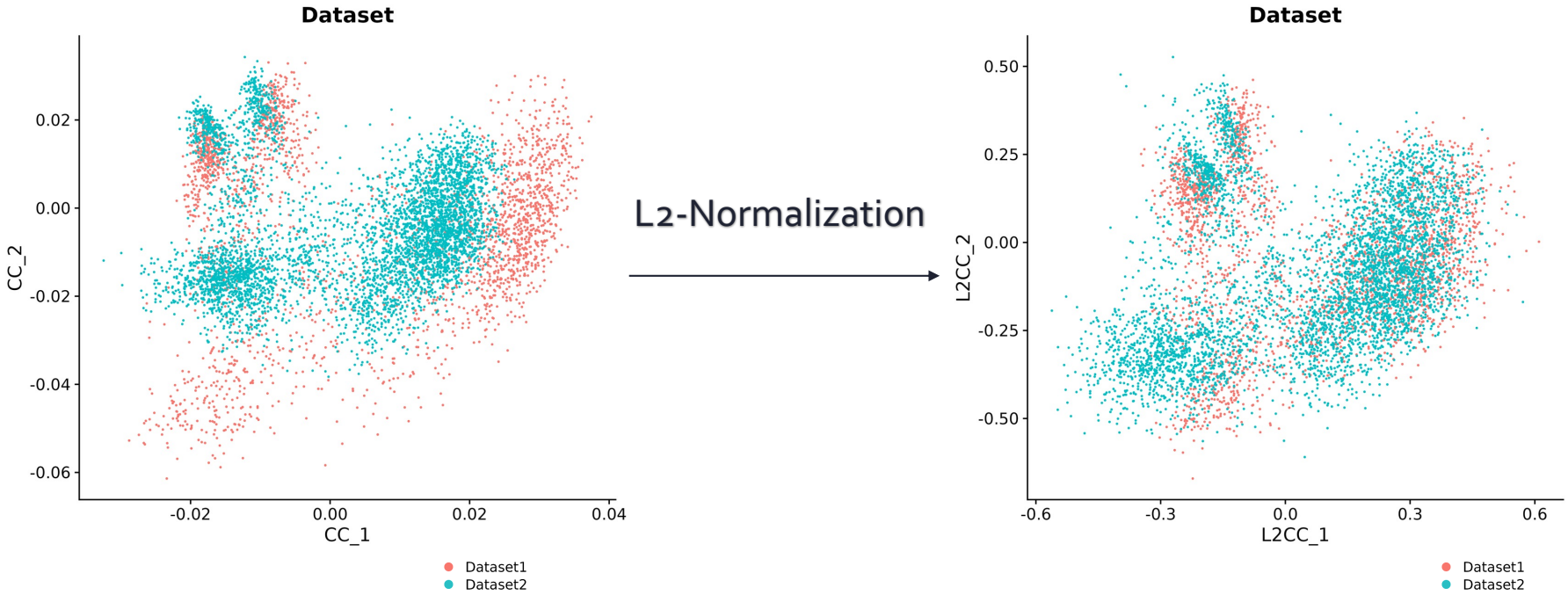
Find	Compute	Apply
Find corresponding cells across datasets (anchors) in L2-normalized CCA	Compute a data adjustment based on correspondences between cells	Apply the adjustment

Step1

Find corresponding cells across datasets



L2-normalization of CCs

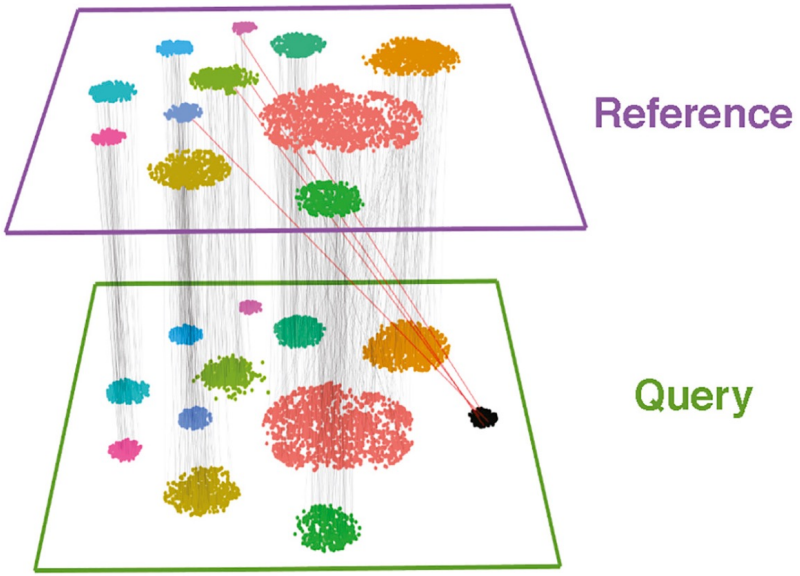


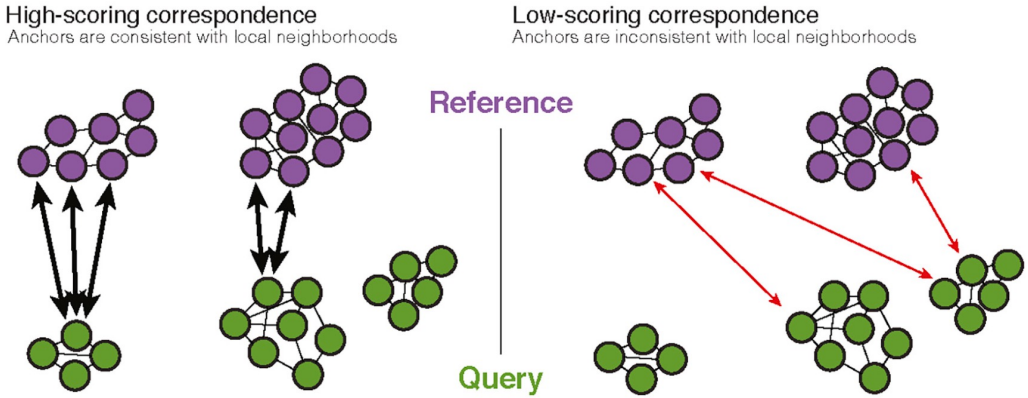
Imagine you have two datasets, A and B, each with a set of genes. CCA tries to find **linear combinations of genes** in A that correlate with corresponding combinations in B.

CCA looks for **pairs of "canonical" variables** (one from each dataset) that are maximally correlated with each other, capturing the shared structure

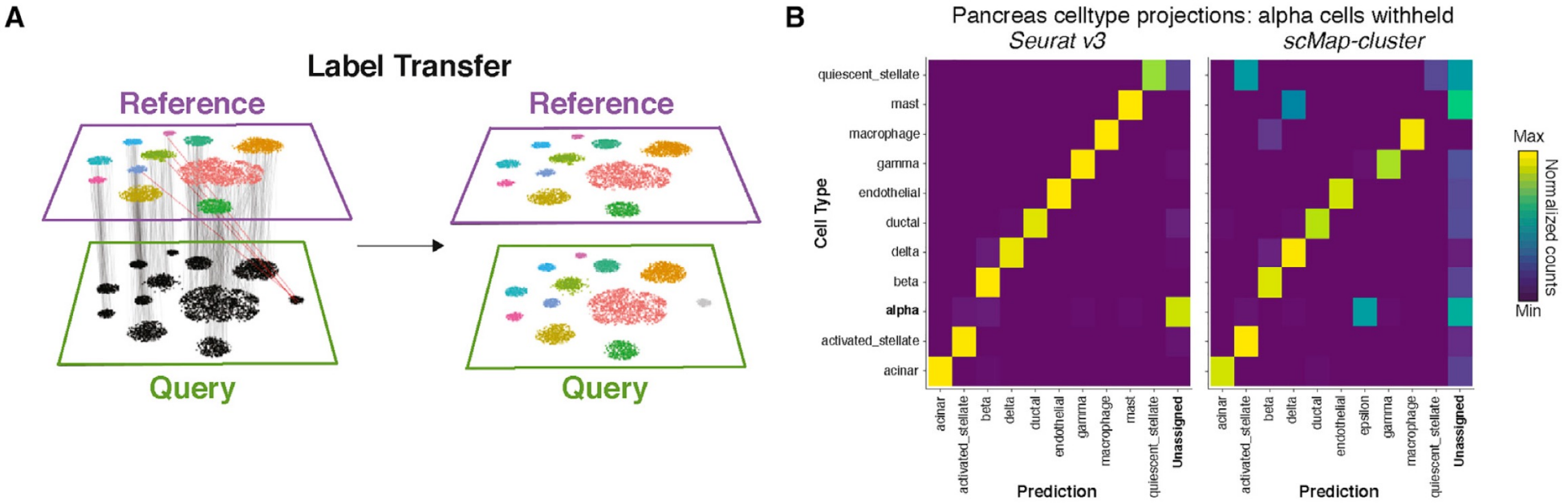
Anchors Identification

Seurat first performs **mutual nearest neighbours (MNN) matching** to identify cell pairs that are closest in gene expression space across datasets.

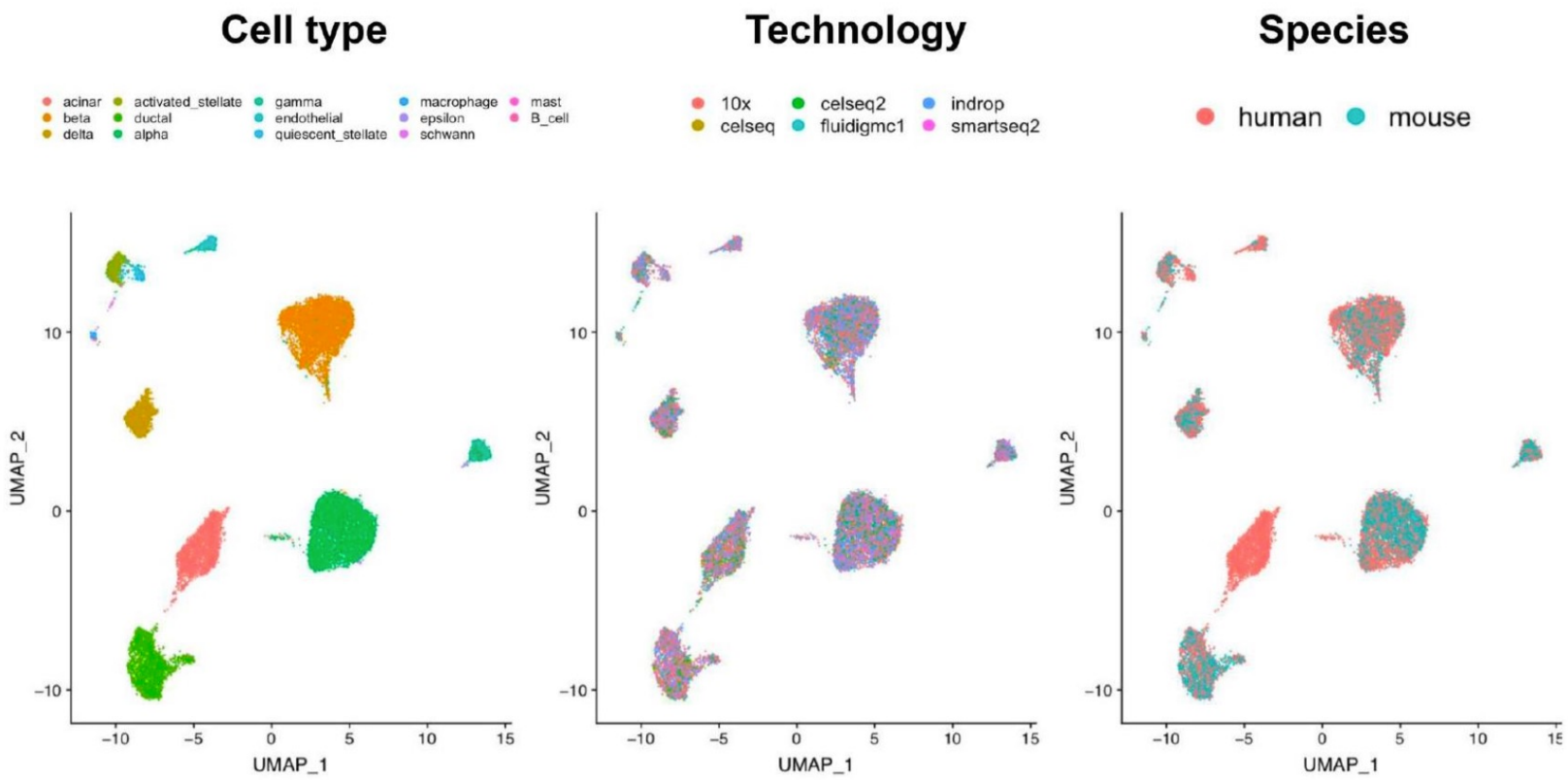




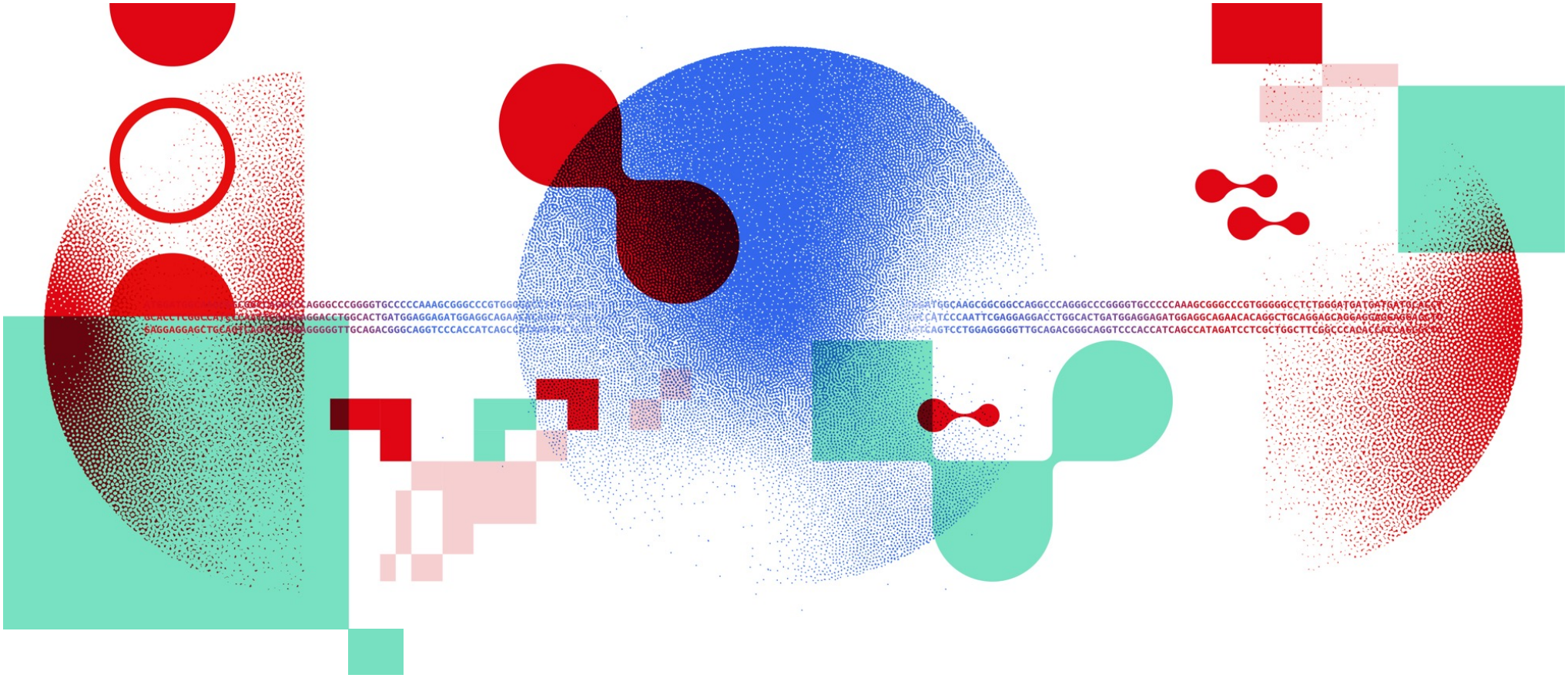
Label transfer: CCA + anchor



Good performarce



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



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss