Swiss Institute of
Bioinformatics

Dimensionality Reduction

Luciano Cascione, PhD
Bioinformatics Core Unit

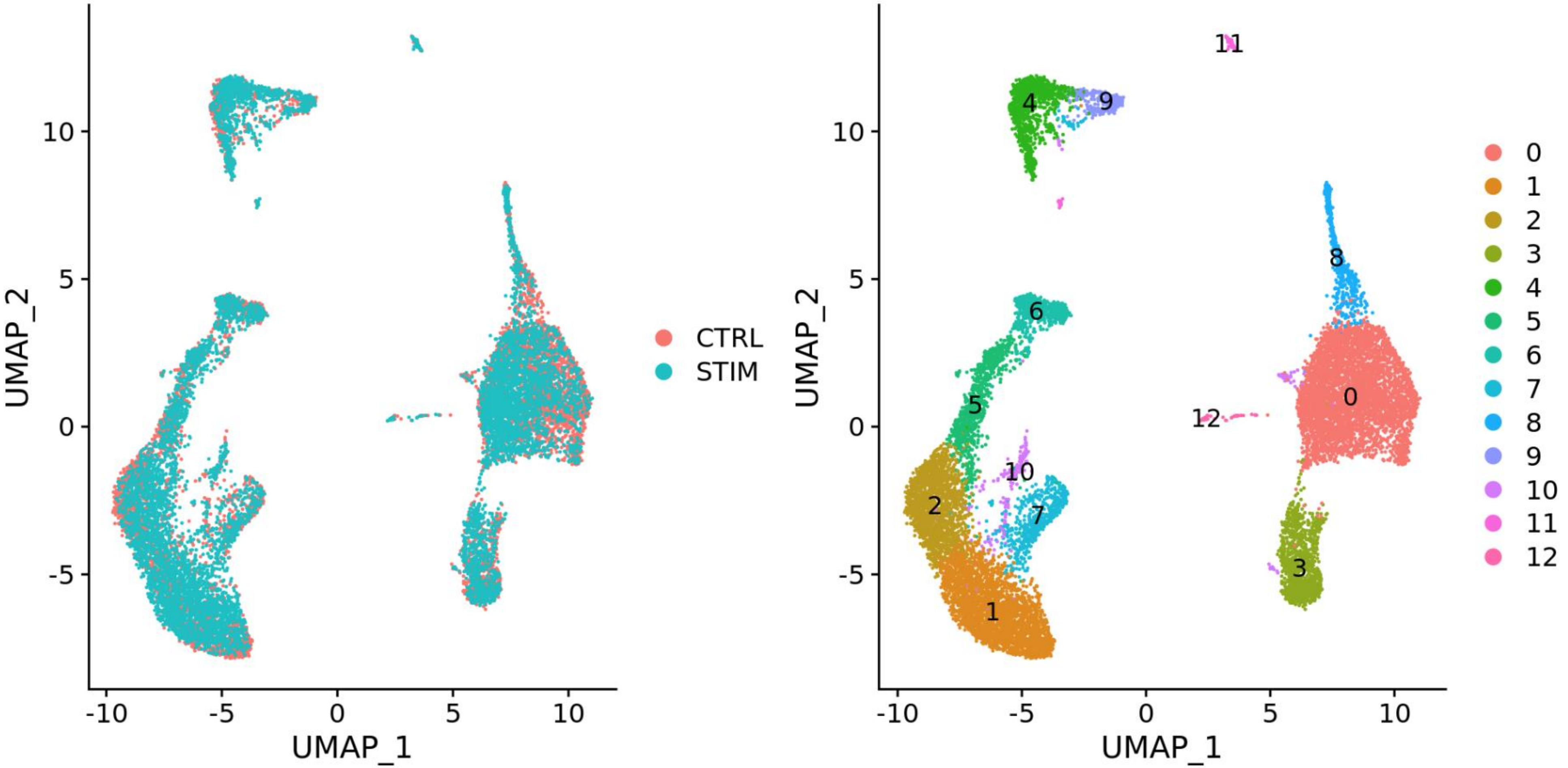
LUCIANO CASCIONE, PHD

BELLINZONA, OCT. 30TH 2024

elixir SWITZERLAND

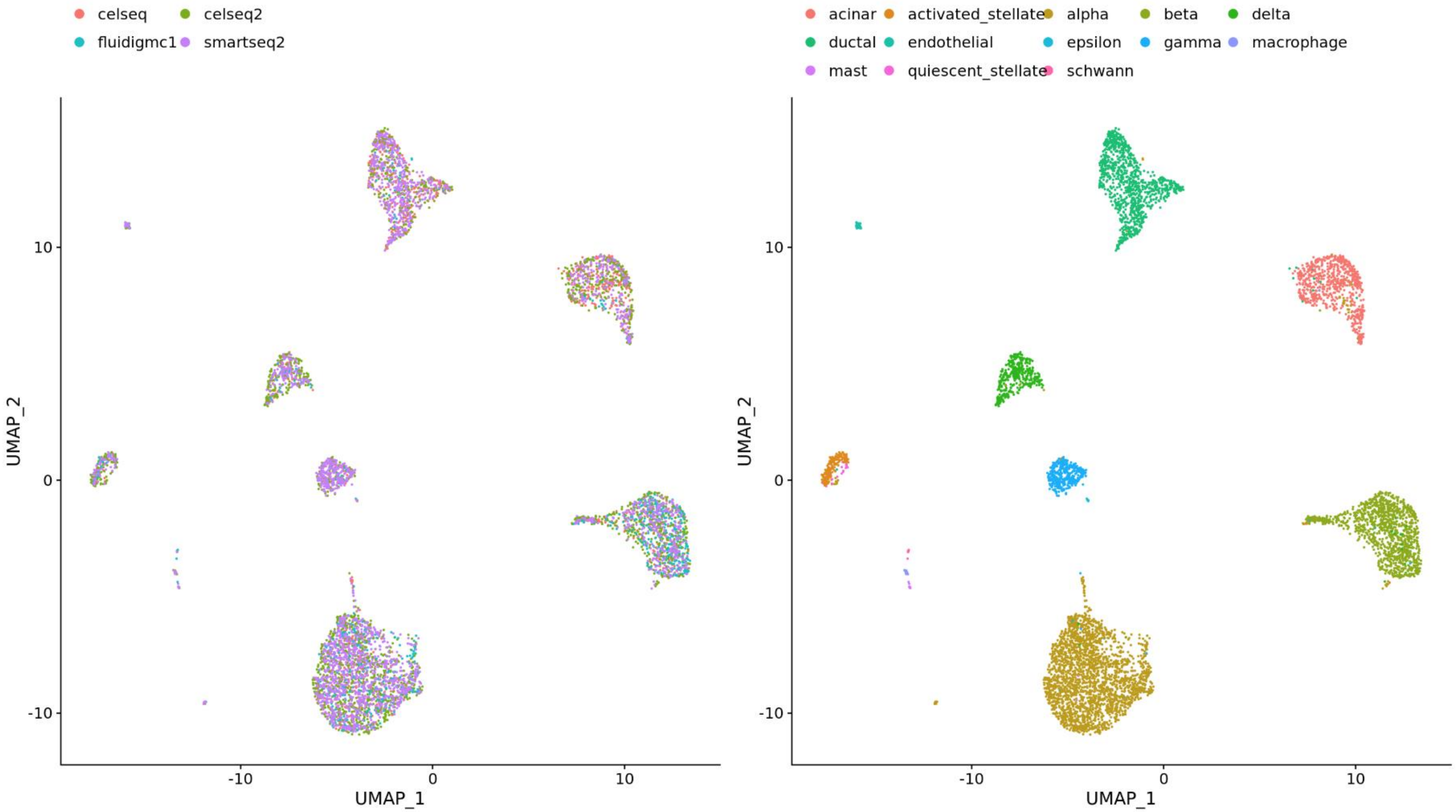
What for ?

Goal: identify shared subpopulations across conditions or datasets



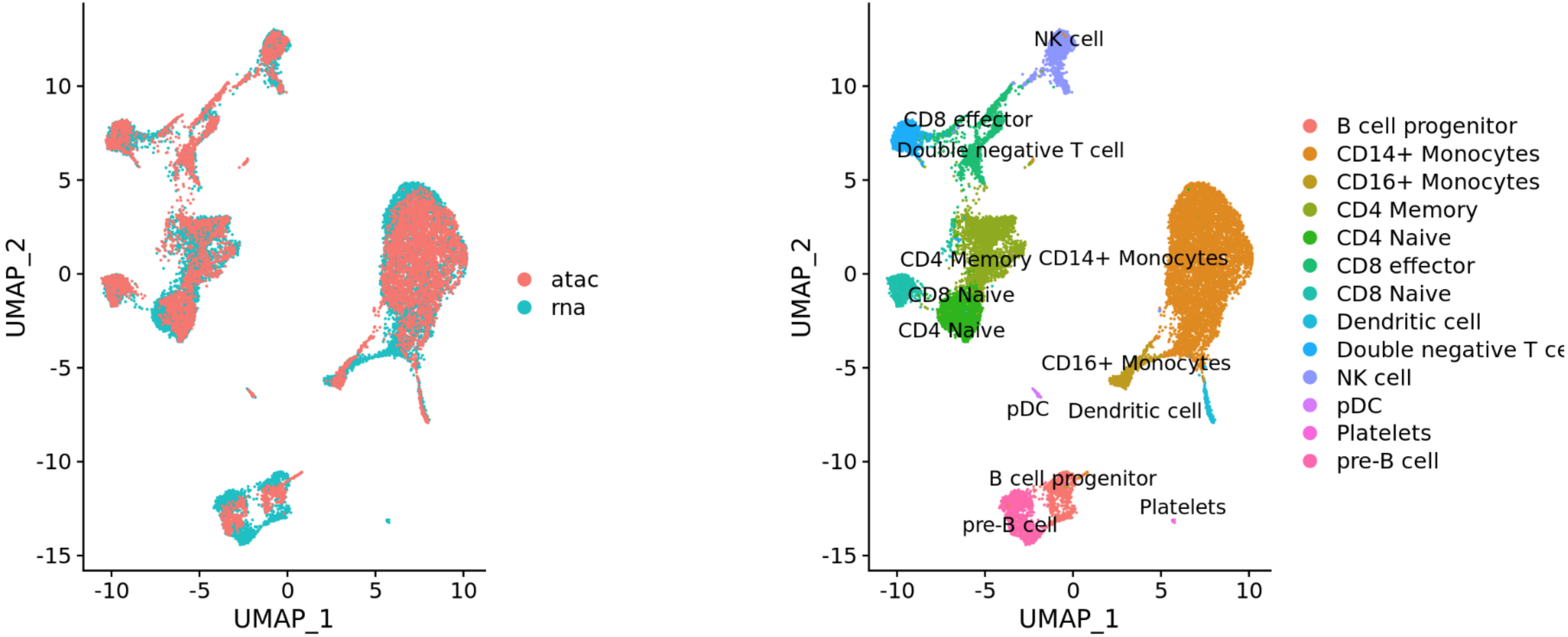
What for ?

Goal: identify shared subpopulations across conditions or datasets

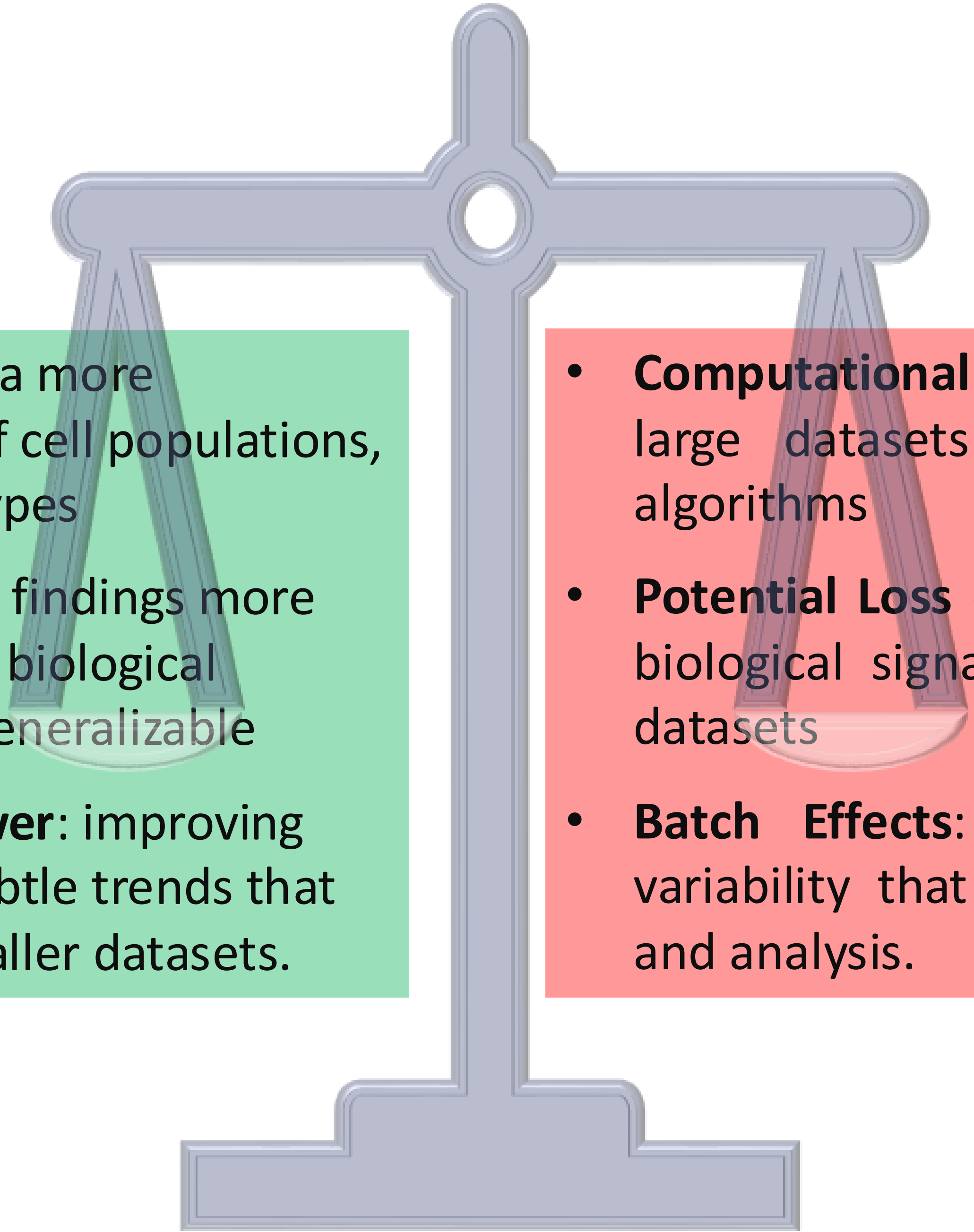


What for ?

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



Pro and Cons



- **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.

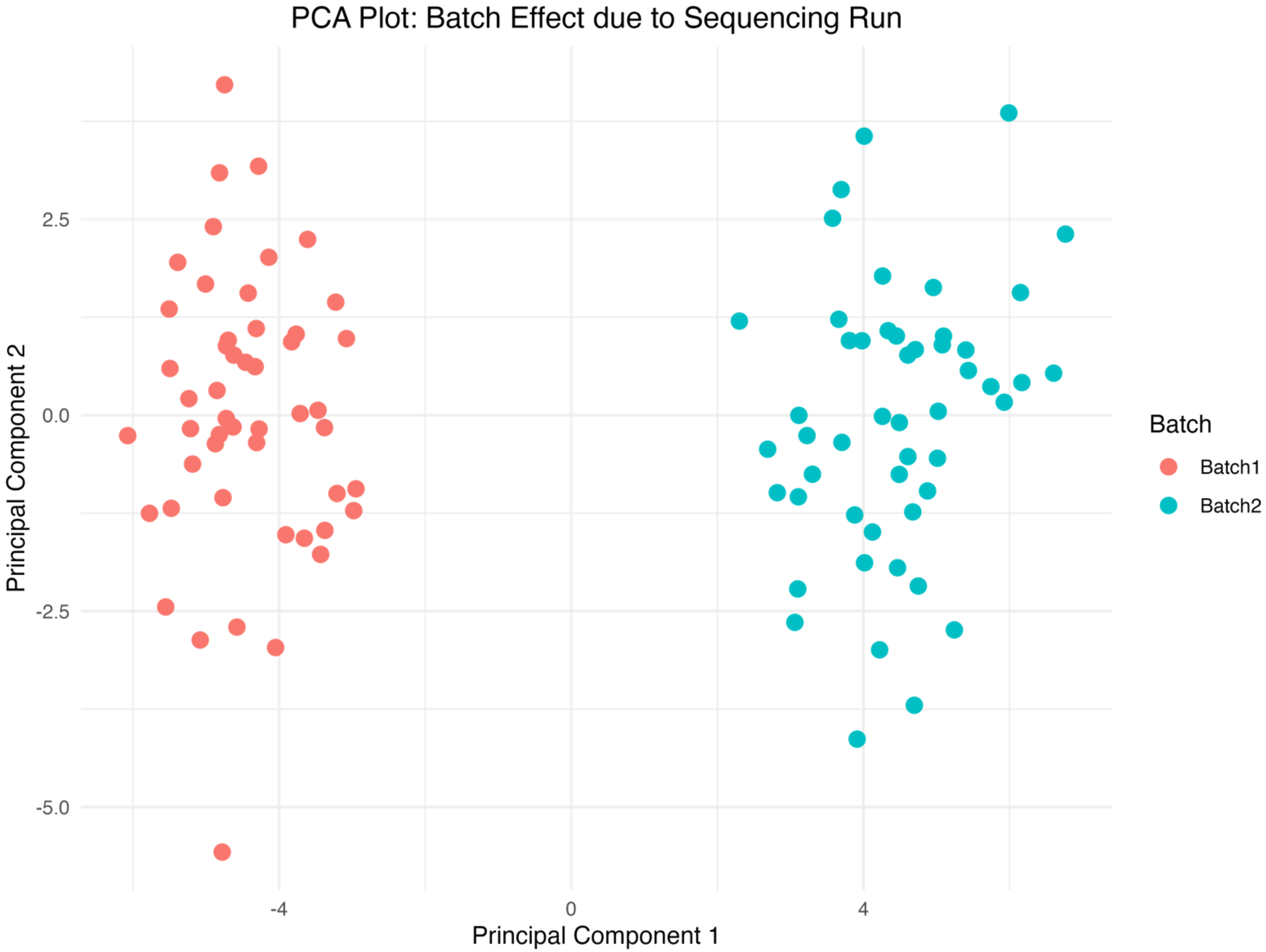
- **Computational Complexity**: Integrating large datasets requires sophisticated algorithms
- **Potential Loss of Information**: Masking biological signals specific to individual datasets
- **Batch Effects**: Introducing unwanted variability that complicates integration and analysis.

Unwanted Sources of Variation

Batch Effects: systematic techincal variations due to differences in:

a) cell isolation and handling protocols,

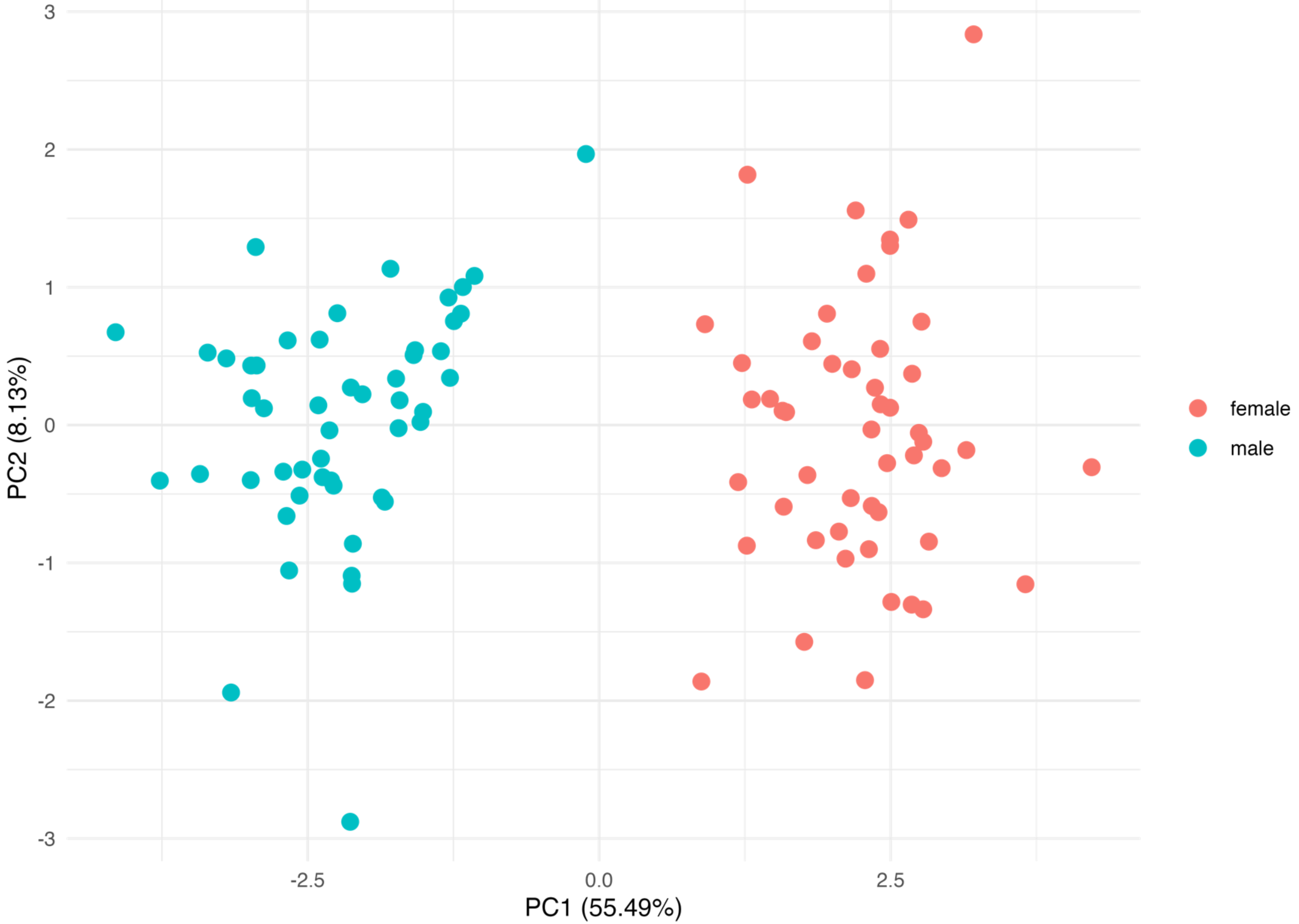
b) library preparation technology, and sequencing platforms



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

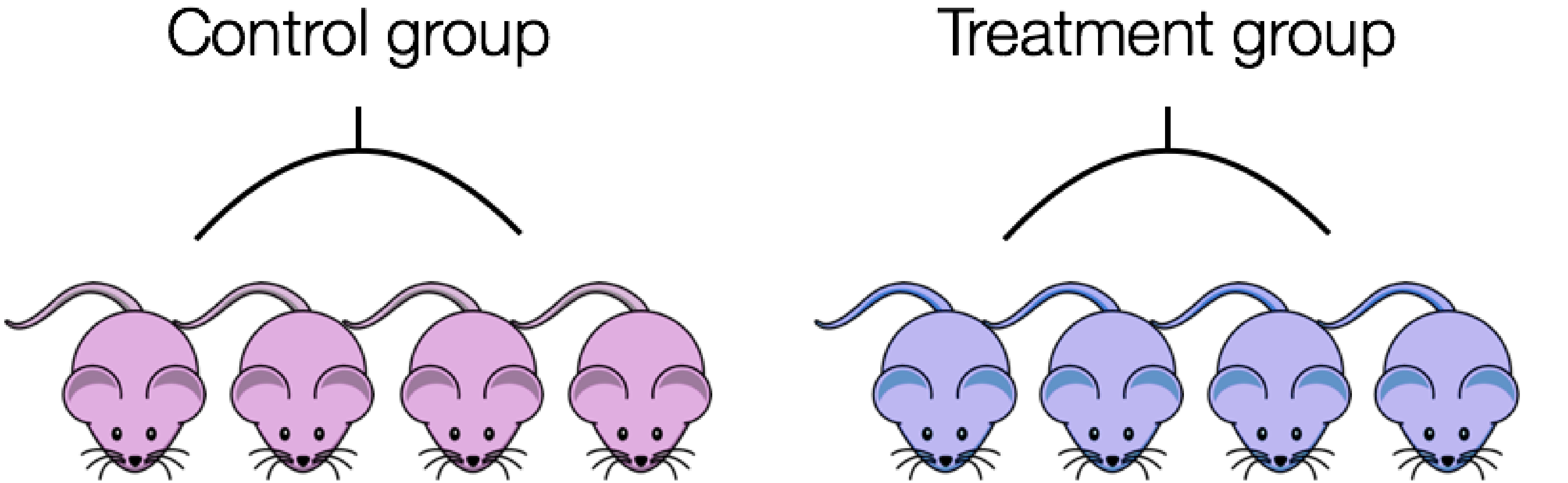
Unwanted Sources of Variation

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



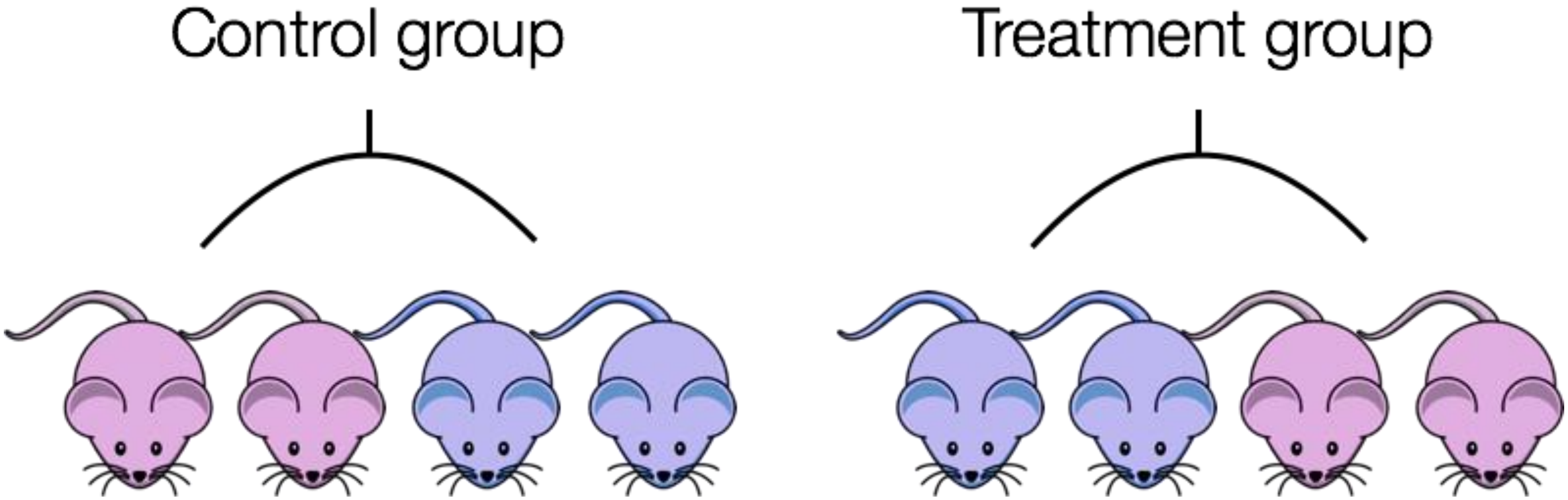
If they are not properly accounted for in the analysis they could potentially lead to misleading associations.

Experimental Design metters



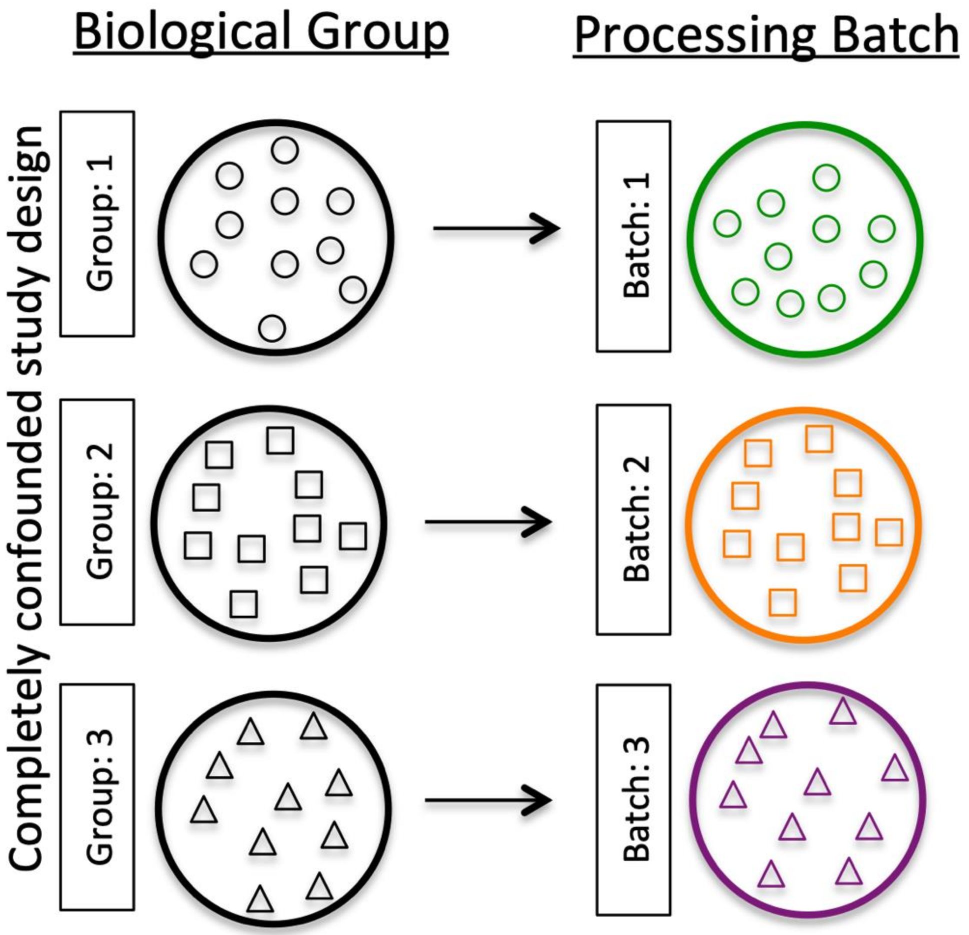
We could not differentiate the effect of treatment from the effect of sex

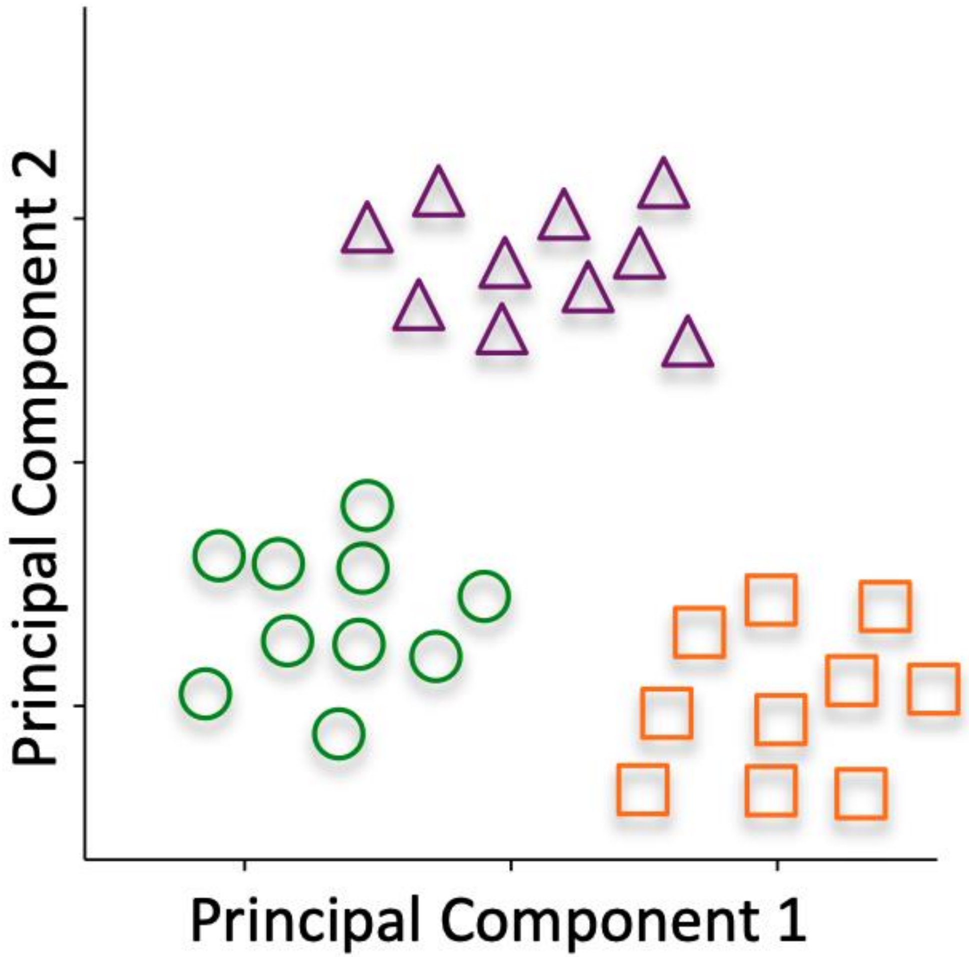
Experimental Design metters



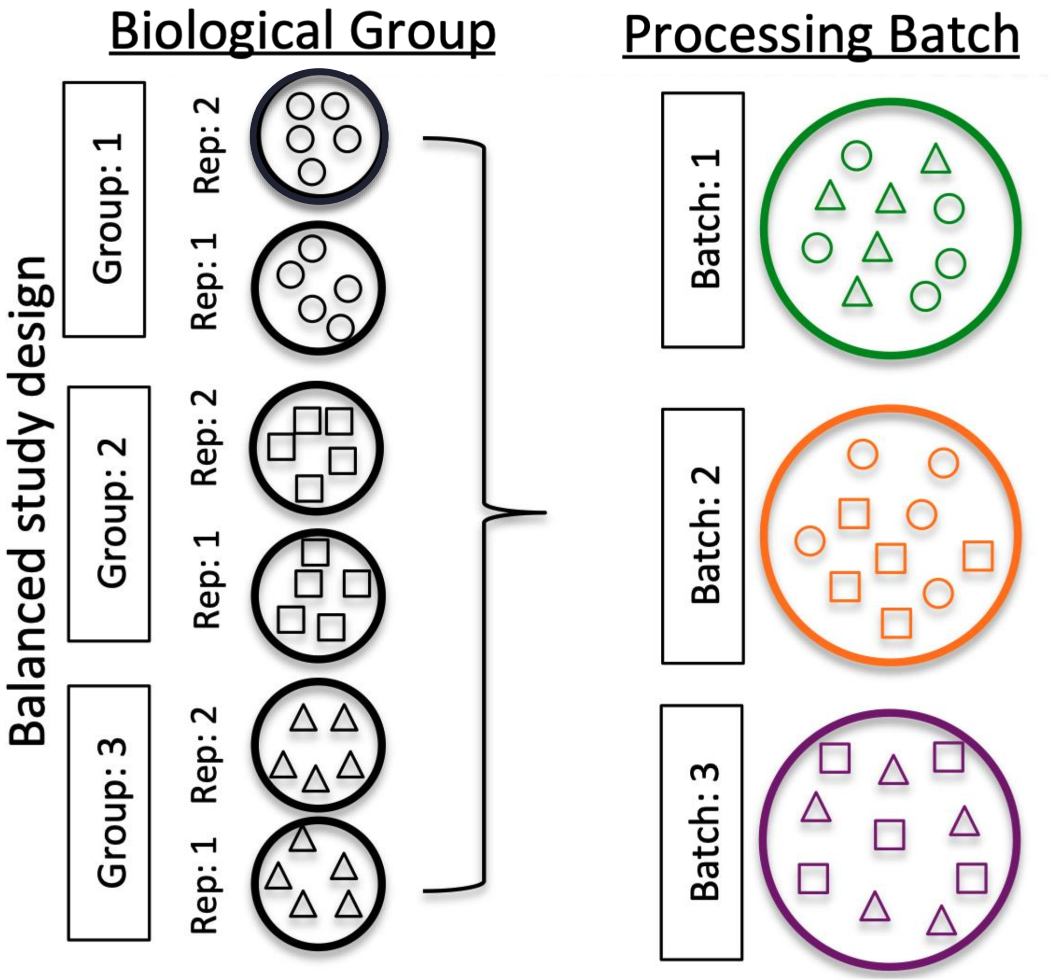
split the animals/samples equally between conditions

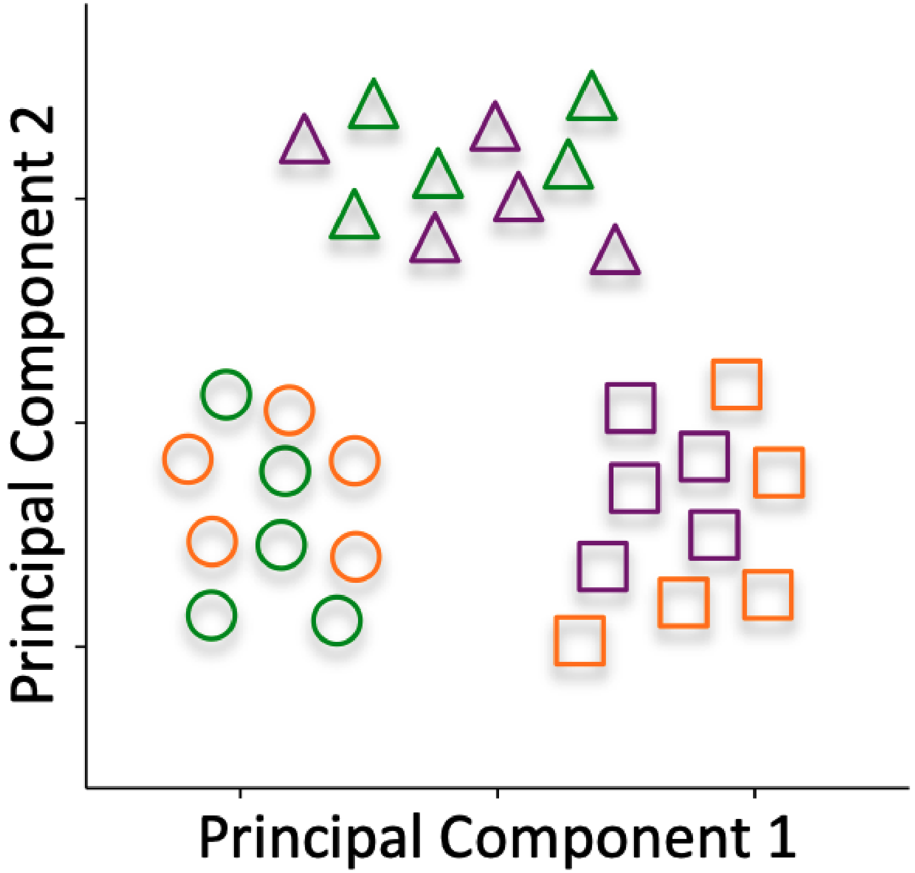
Experimental Design metters





Experimental Design metters





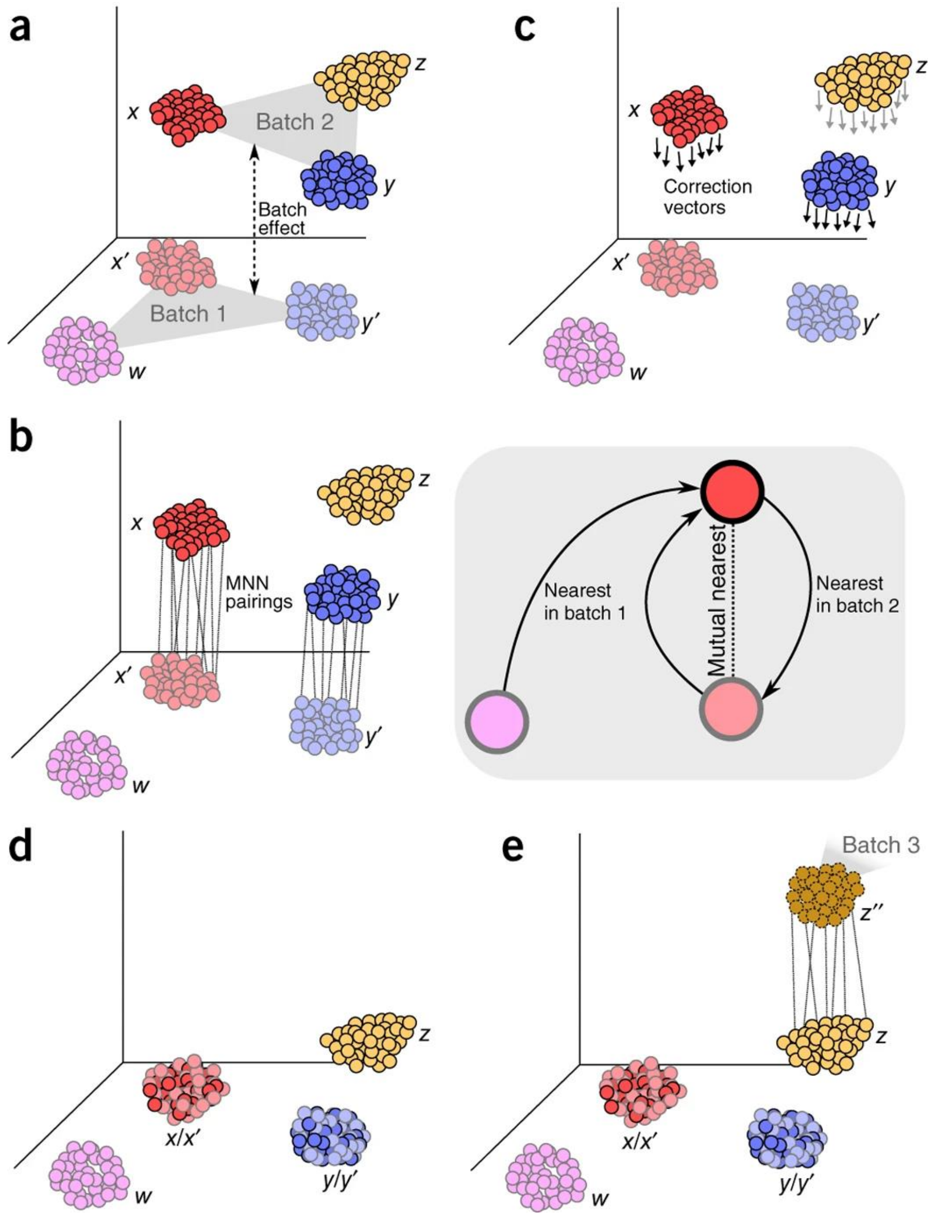
How to integrate

1. Find corresponding cells across datasets (by computing a **distance between cells** in a certain space)

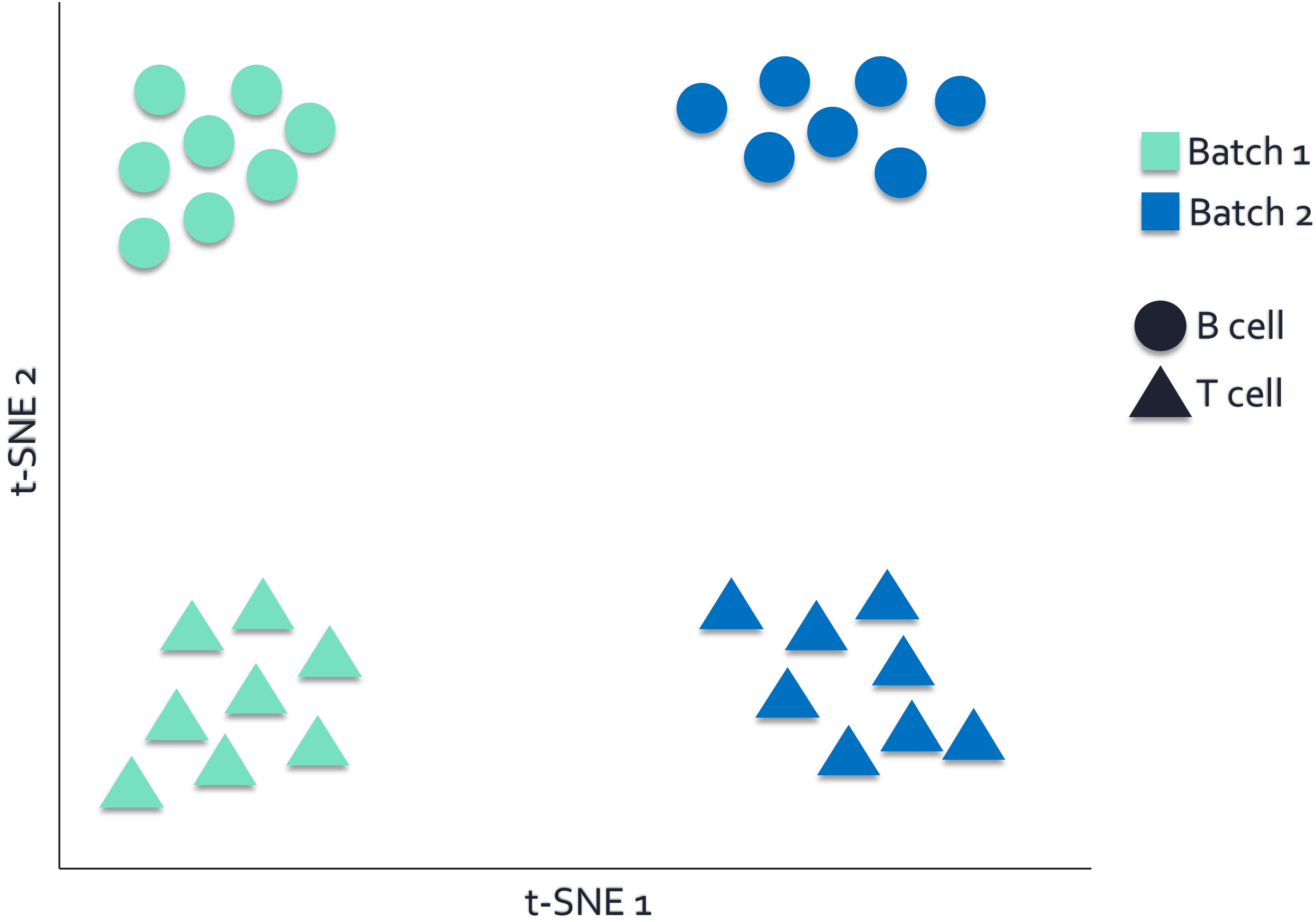
2. Compute a data adjustment based on correspondences between cells

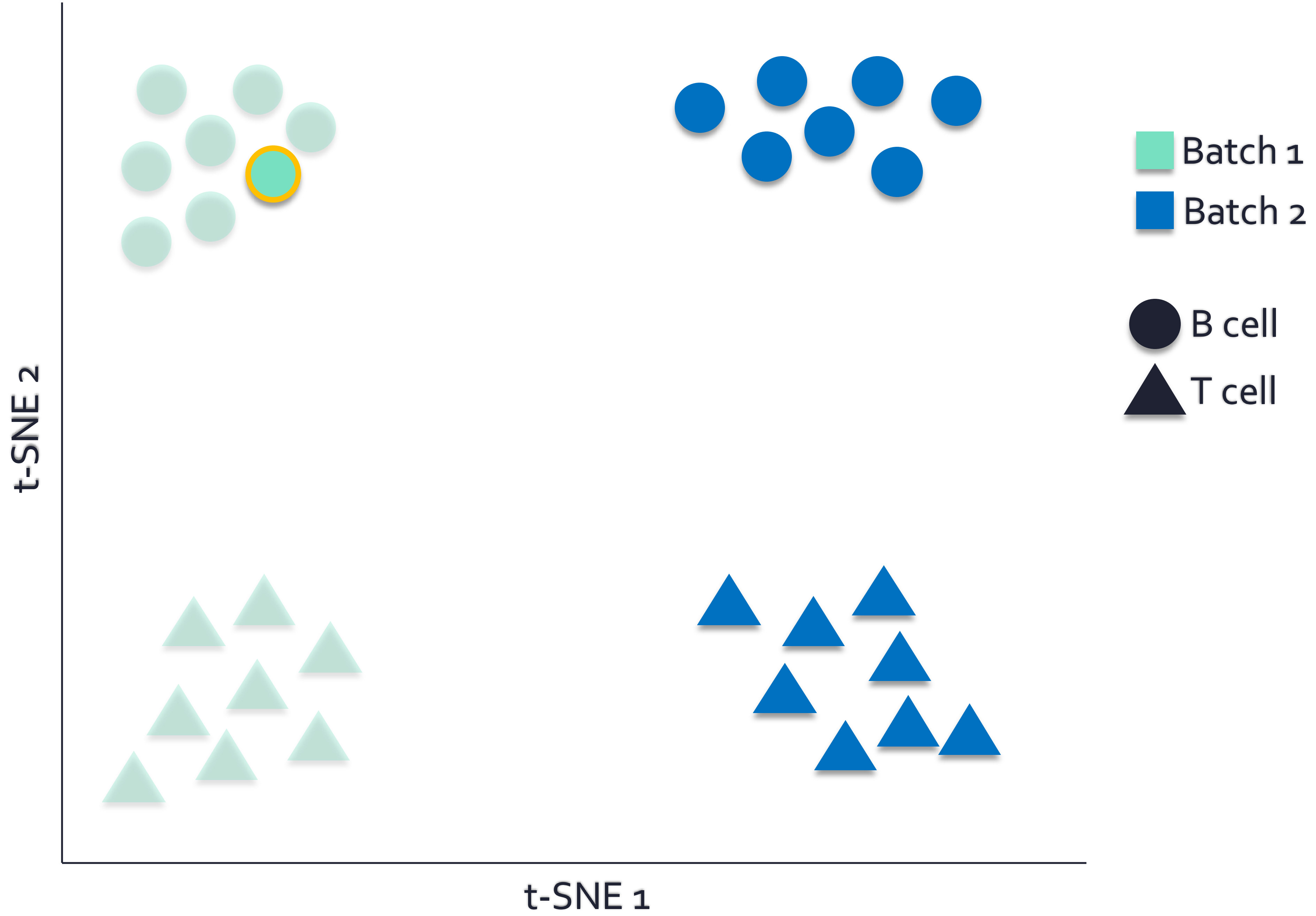
3. Apply the adjustment

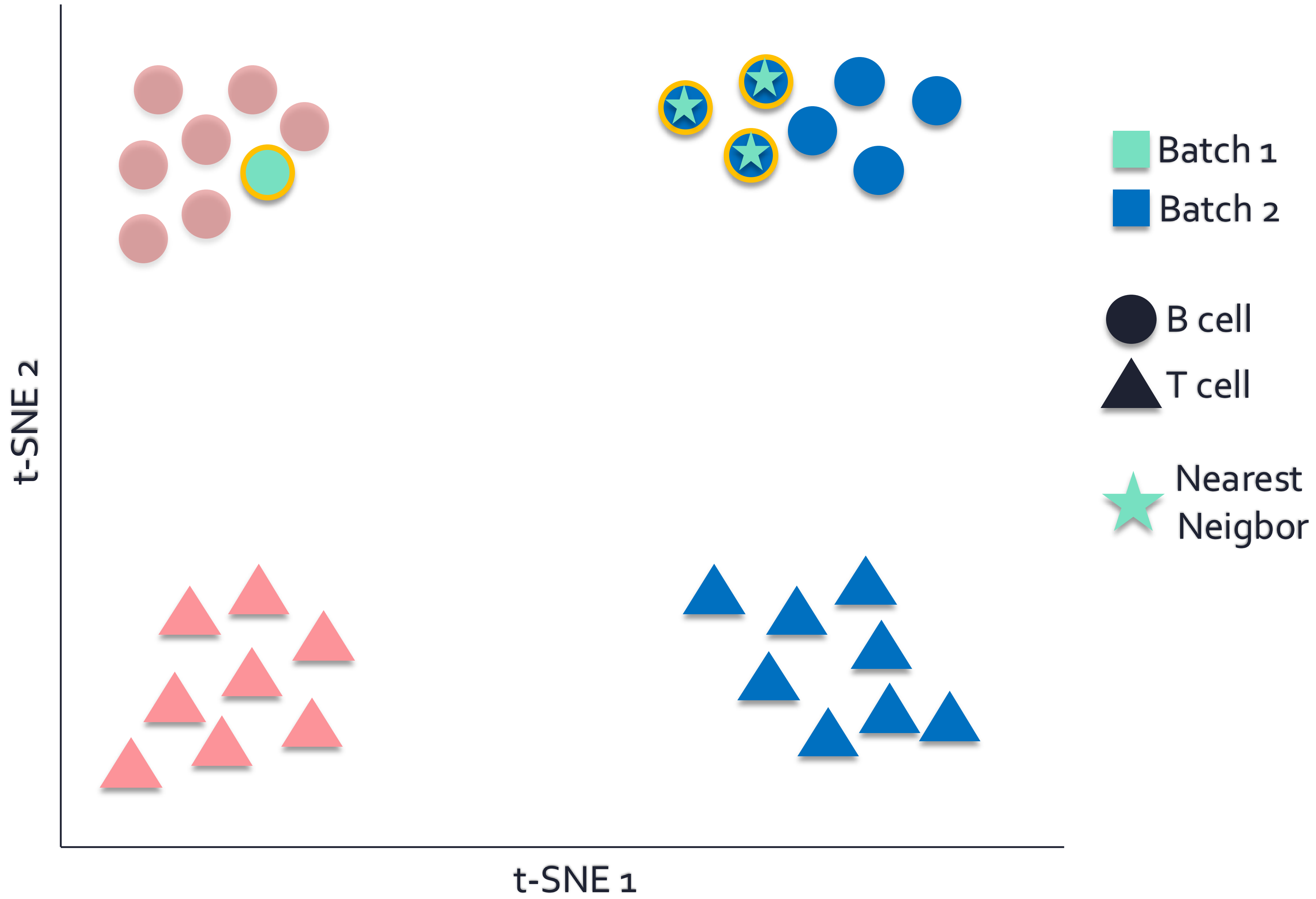
Step 1

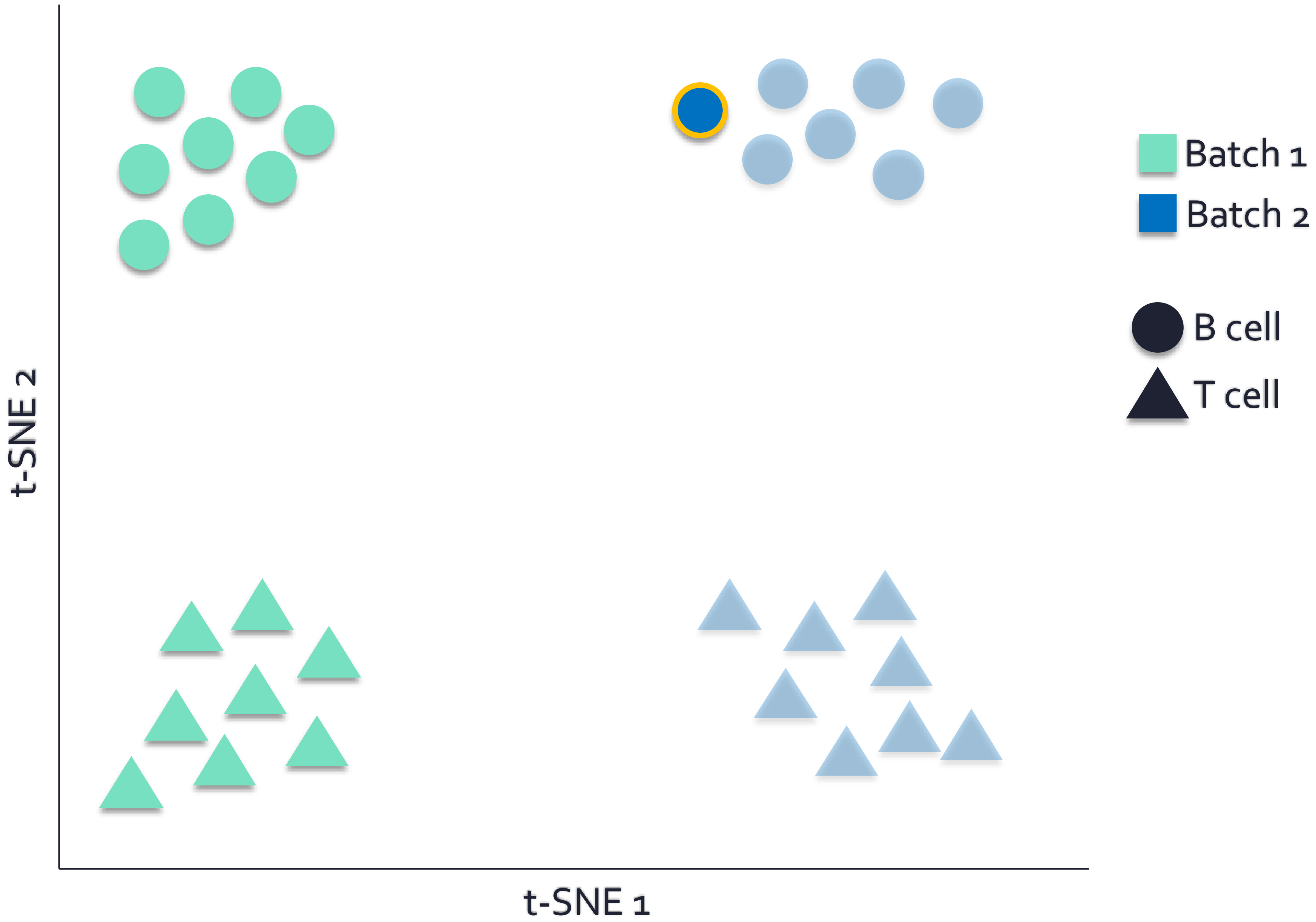

a
x
z
Batch 2
y
Batch effect
x′
Batch 1
y′
w
c
x
z
Correction vectors
y
x′
y′
w
b
x
z
MNN pairings
y
x′
y′
w
Nearest in batch 1
Mutual nearest
Nearest in batch 2
d
z
x/x′
w
y/y′
e
Batch 3
z″
z
x/x′
w
y/y′

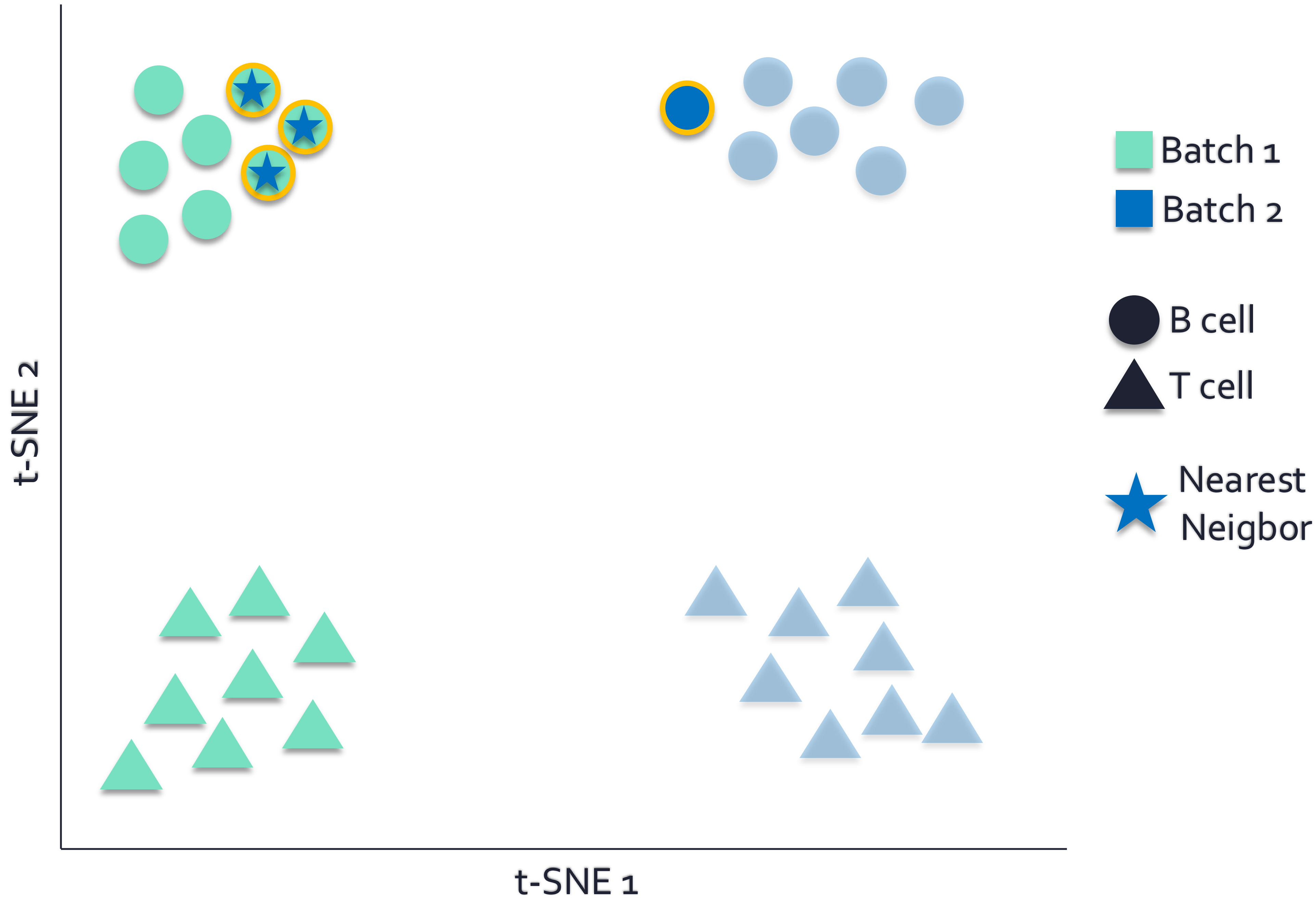
Example

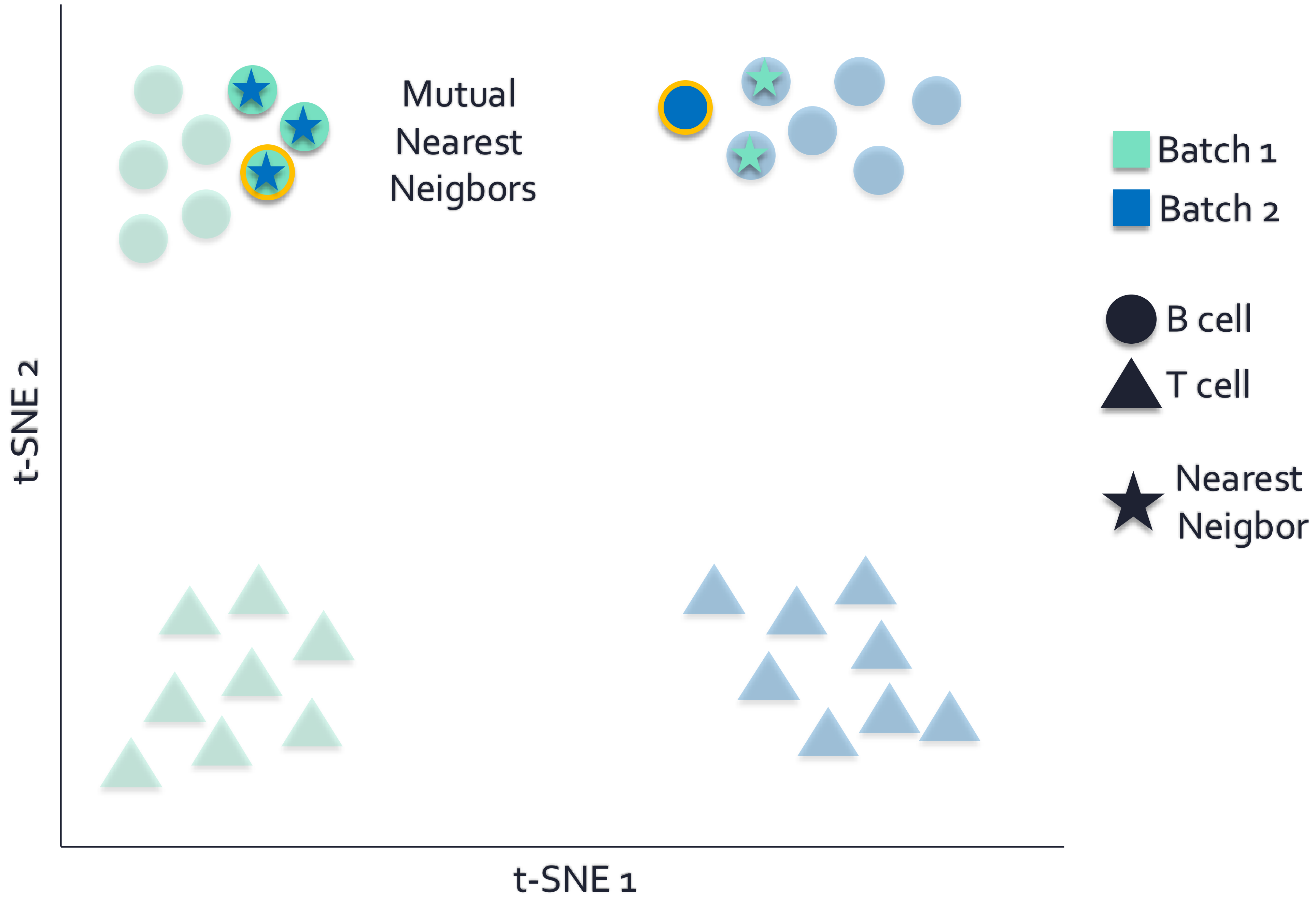



t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell

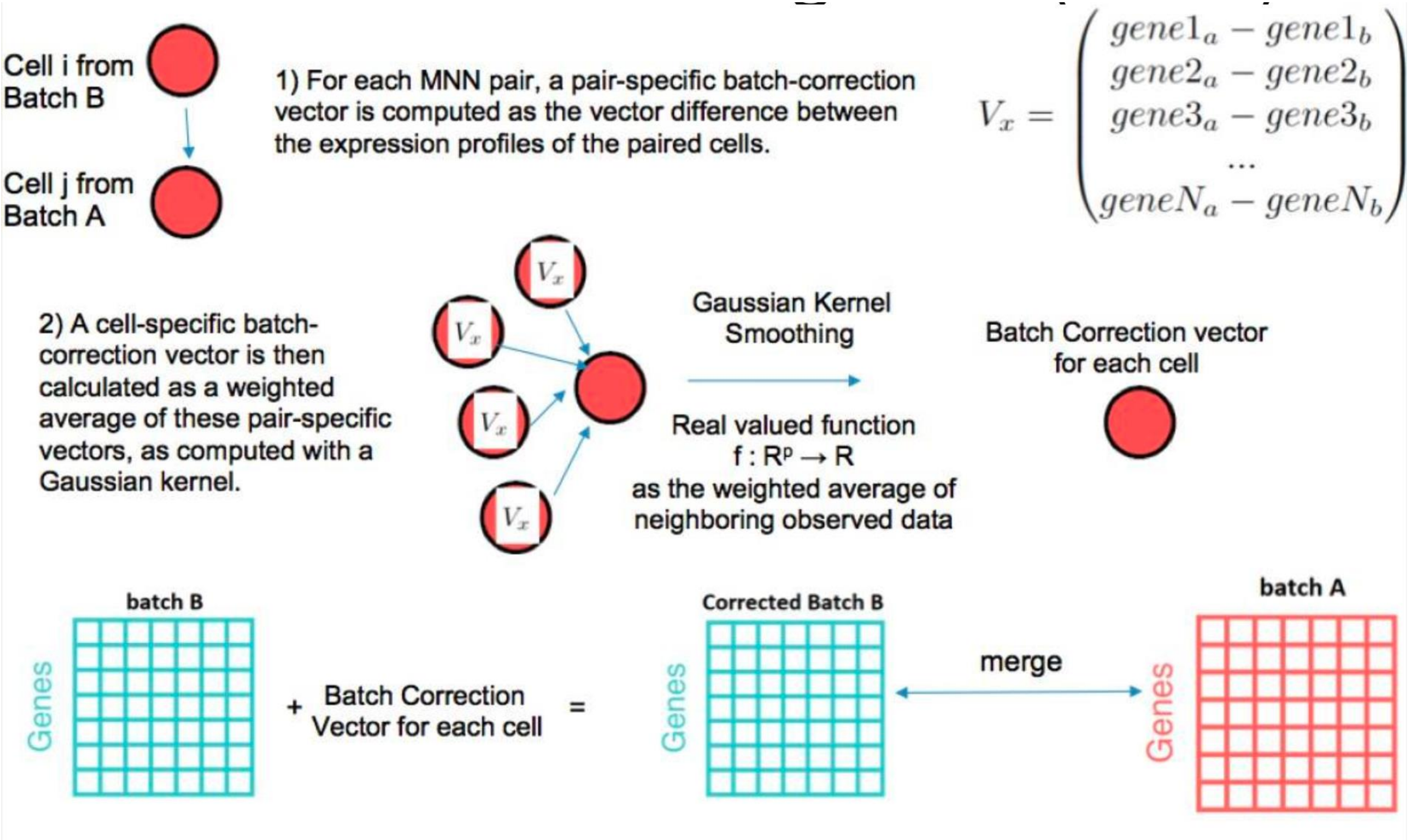

t-SNE 2
t-SNE 1
Batch 1
Batch 2
B cell
T cell
Nearest
Neigbor


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



1) For each MNN pair, a pair-specific batch-correction vector is computed as the vector difference between the expression profiles of the paired cells.

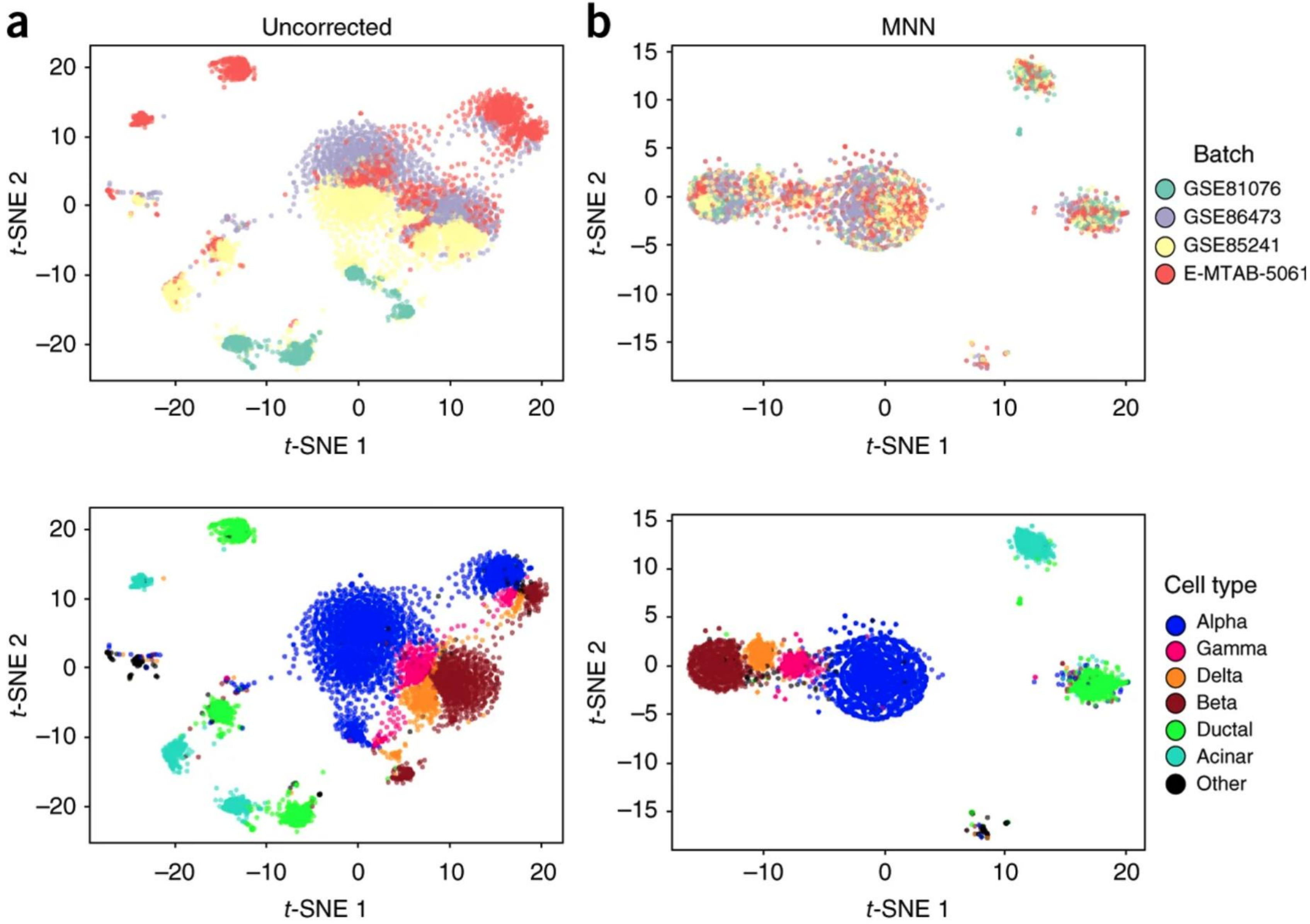
$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \cdots \\ geneN_a - geneN_b \end{pmatrix}$$

2) A cell-specific batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with a Gaussian kernel.

Real valued function $f : R^p \rightarrow R$ as the weighted average of neighboring observed data

batch B + Batch Correction Vector for each cell = Corrected Batch B

Final Integration

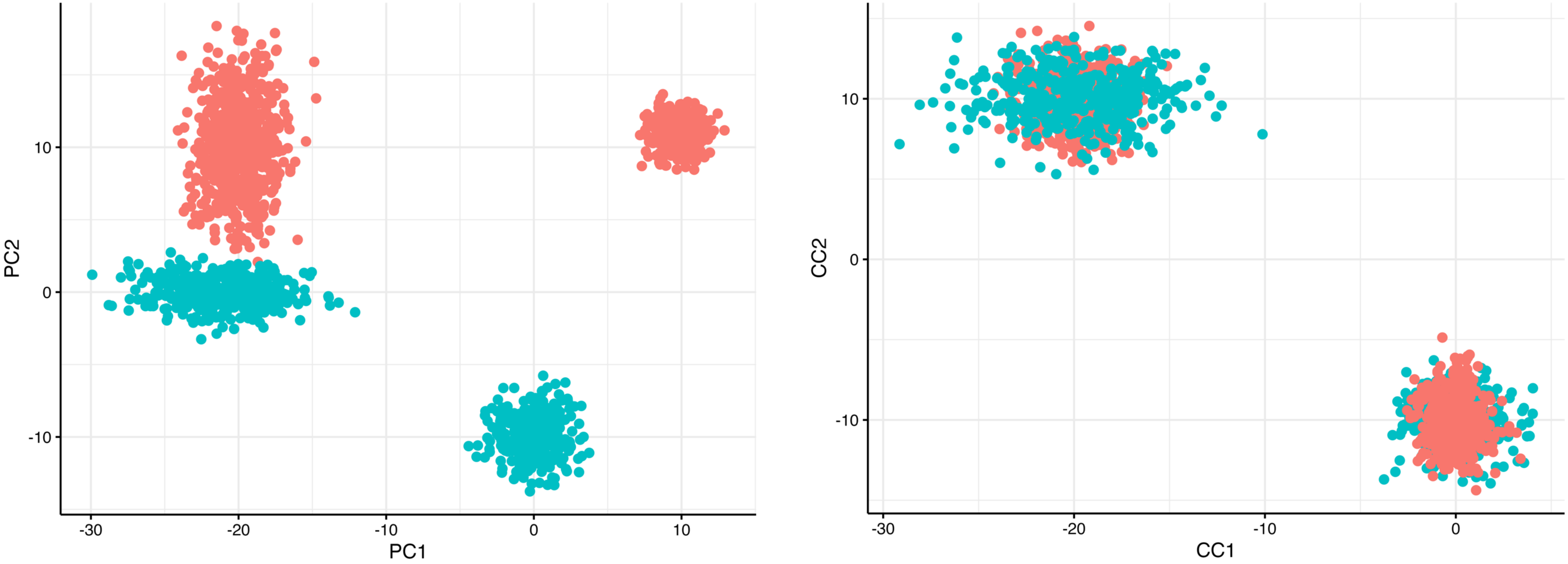


Canonical Correlation Analysis + anchors

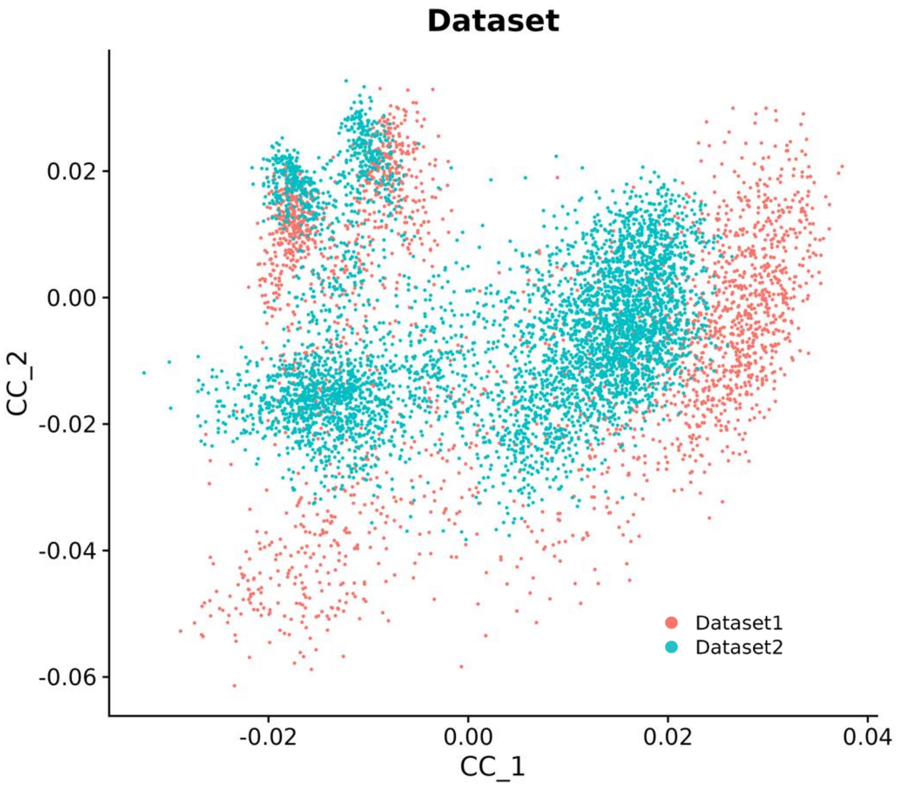
1. Find corresponding cells across datasets (anchors) in (canonical-correlation analysis) L2-normalized CCA space and euclidean distance
2. Compute a data adjustment based on correspondences between cells
3. Apply the adjustment

Step1

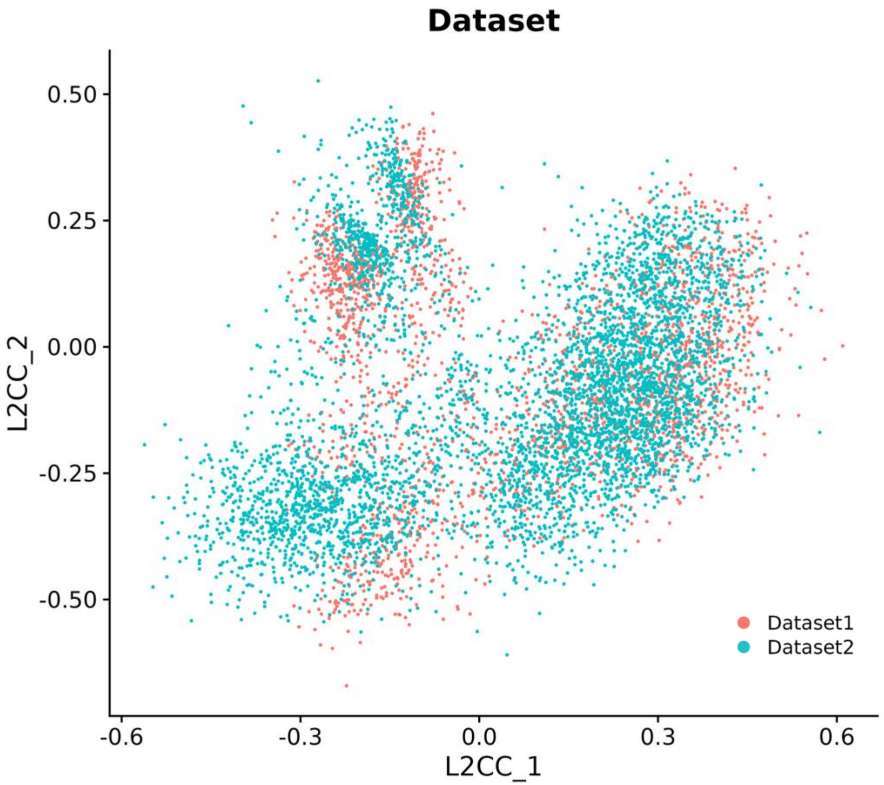
Find corresponding cells across datasets



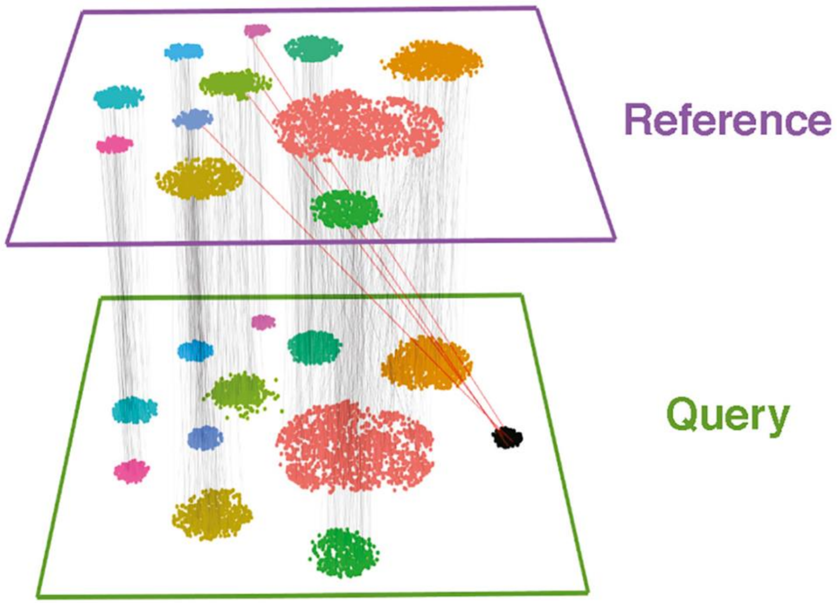
L2-normalization of CCs

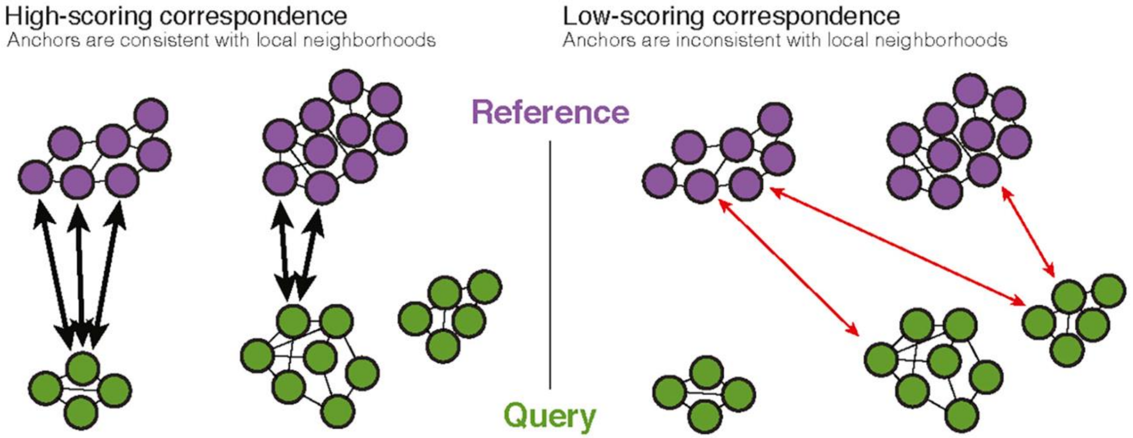


L2-Normalization

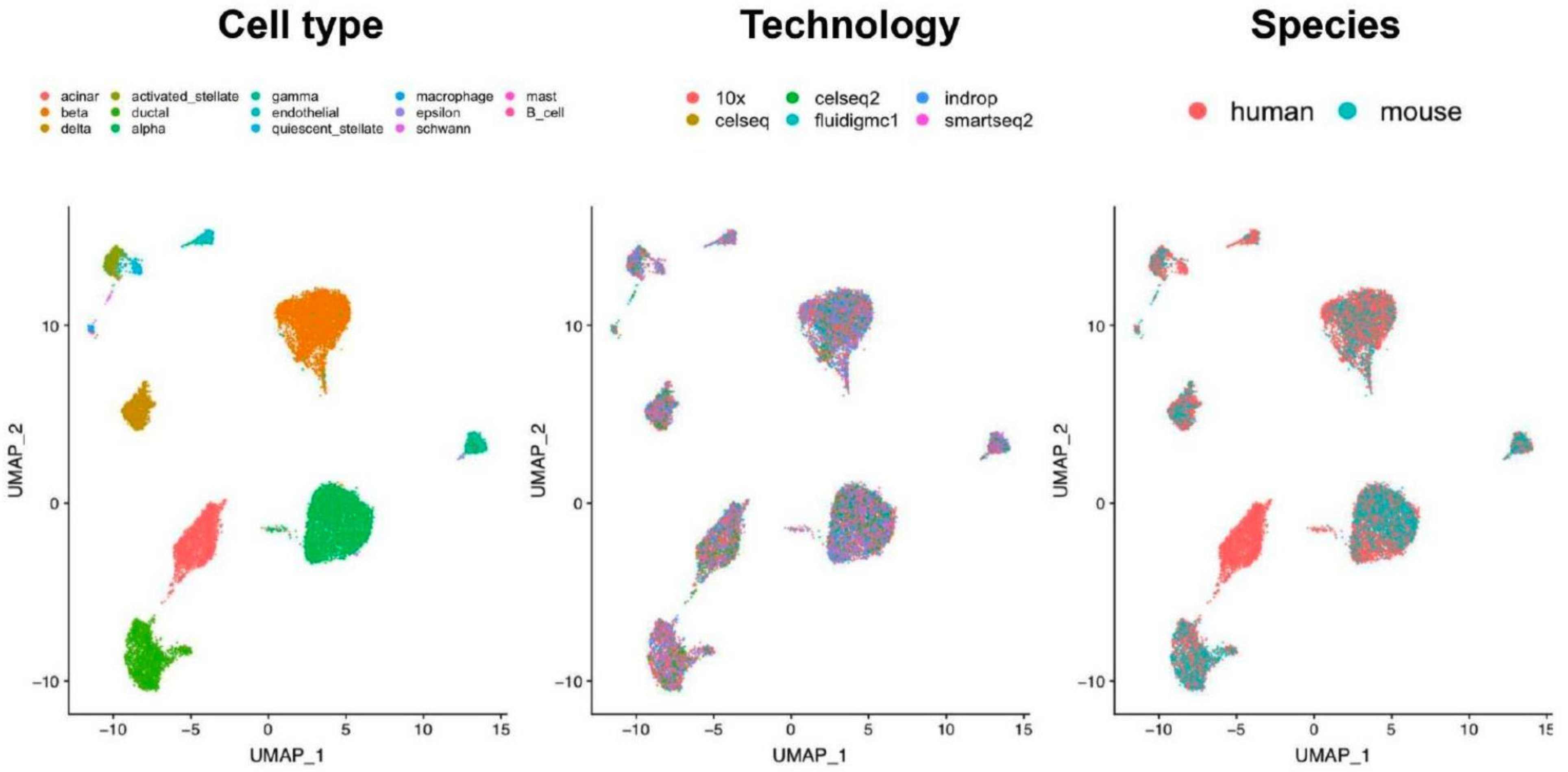


Anchors Identification



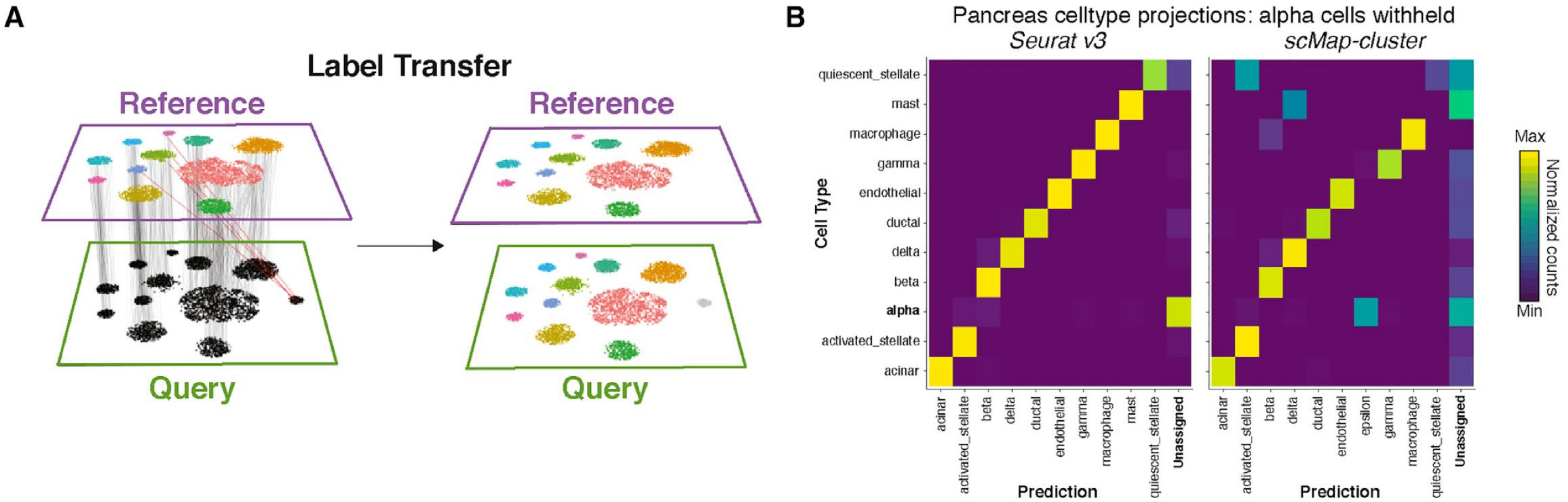

High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query

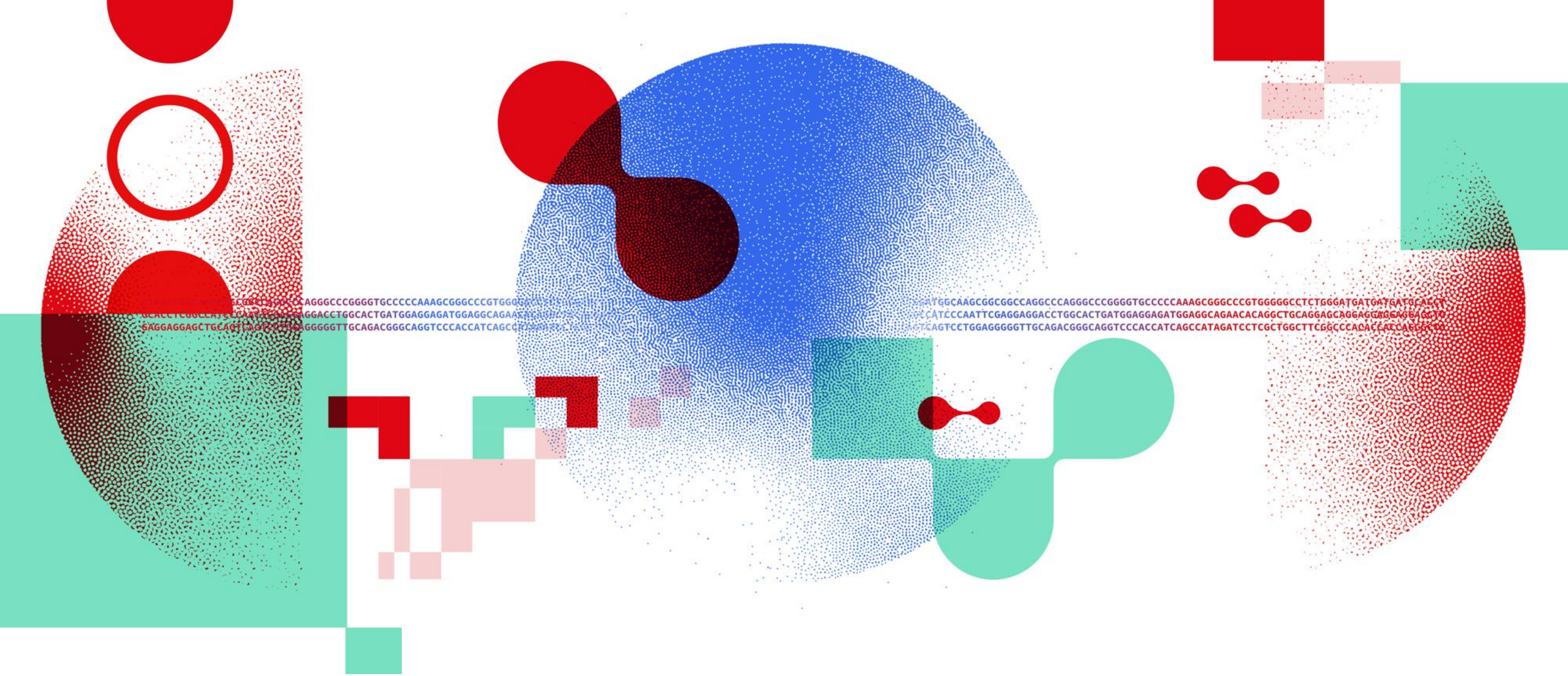
Good performarce



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

Label transfer: CCA + anchor




GGATGGCAAGCGGCGGCCAGGCCCAGGGCCCGGGGTGCCCCCAAAGCGGGCCCGTGGGGGCCTCTGGGATGATGATGATGCACCT
CCATCCCAATTCGAGGAGGACCTGGCACTGATGGAGGAGATGGAGGCAGAACACAGGCTGCAGGAGCAGGAGGAGGAGGAGGTG
AGTCAGTCCTGGAGGGGGTTGCAGACGGGCAGGTCCCACCATCAGCCATAGATCCTCGCTGGCTTCGGCCCACACCACCAGG

Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss