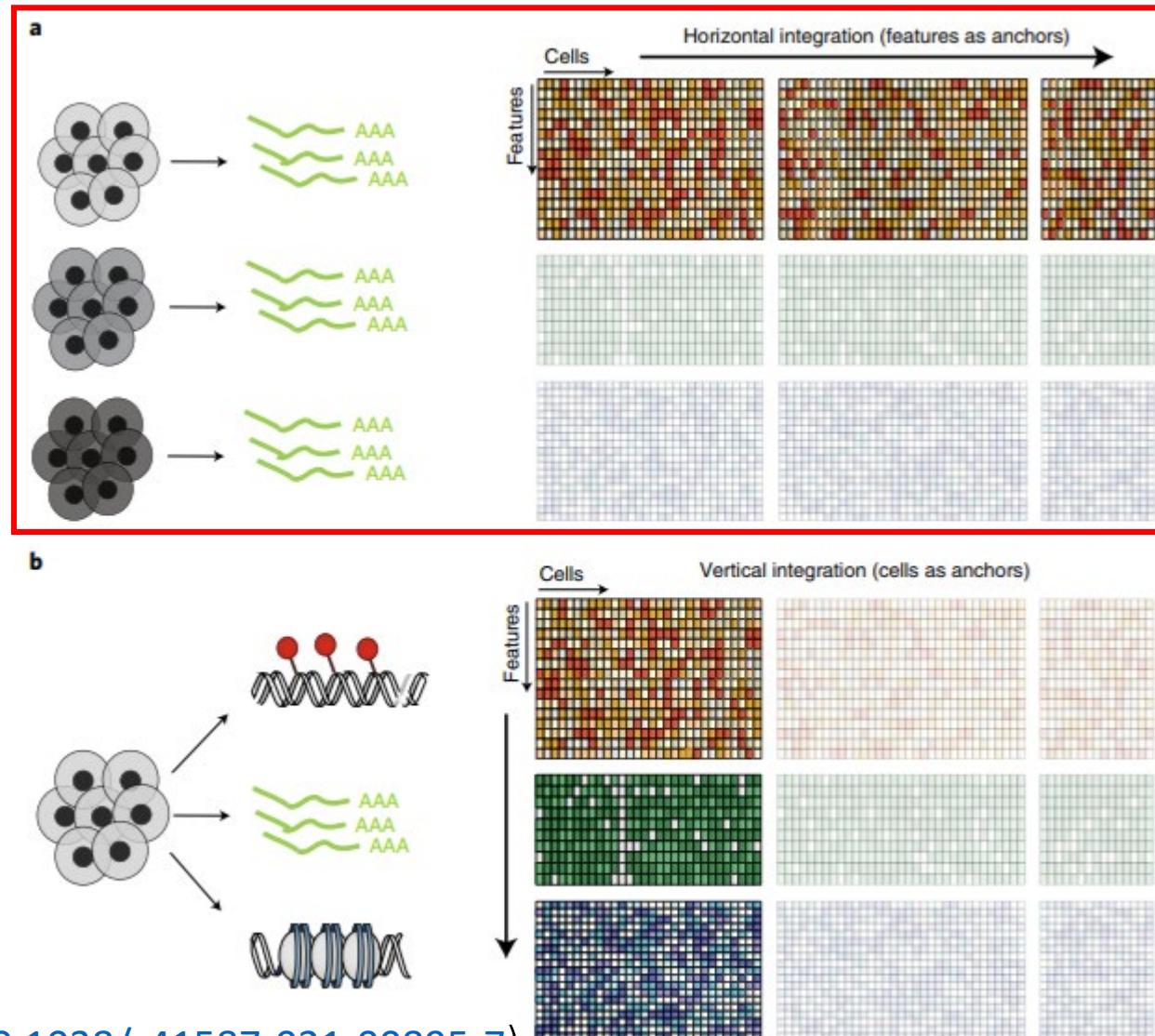


Single Cell RNA-seq Data Integration

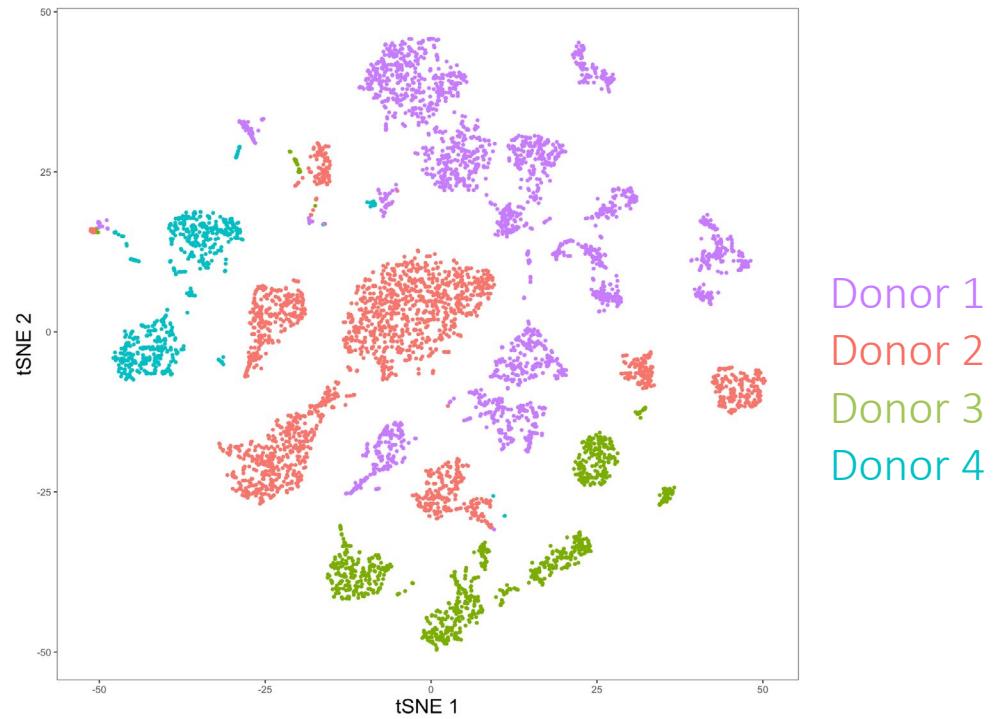
Tamim Abdelaal, PhD

Department of Radiology, LUMC
Delft Bioinformatics Lab, TU Delft

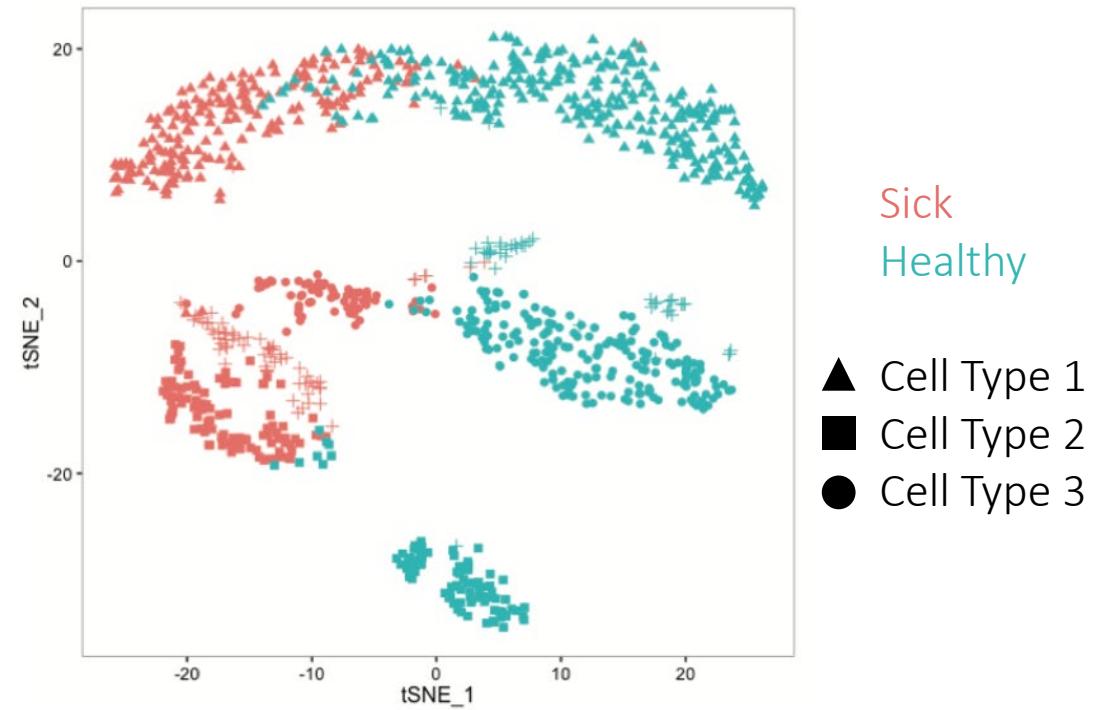
Single cell data integration



Why integrate?



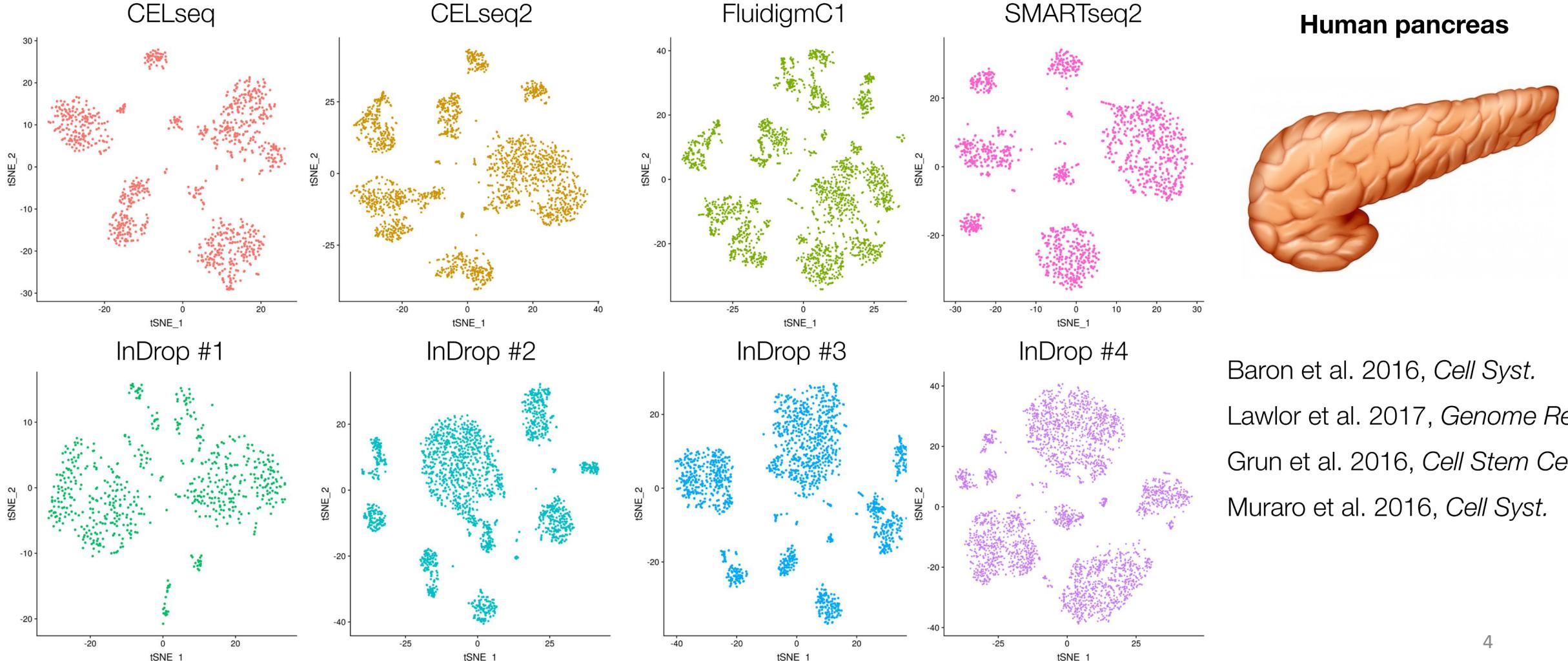
Same tissue from different donors



Cross condition comparisons

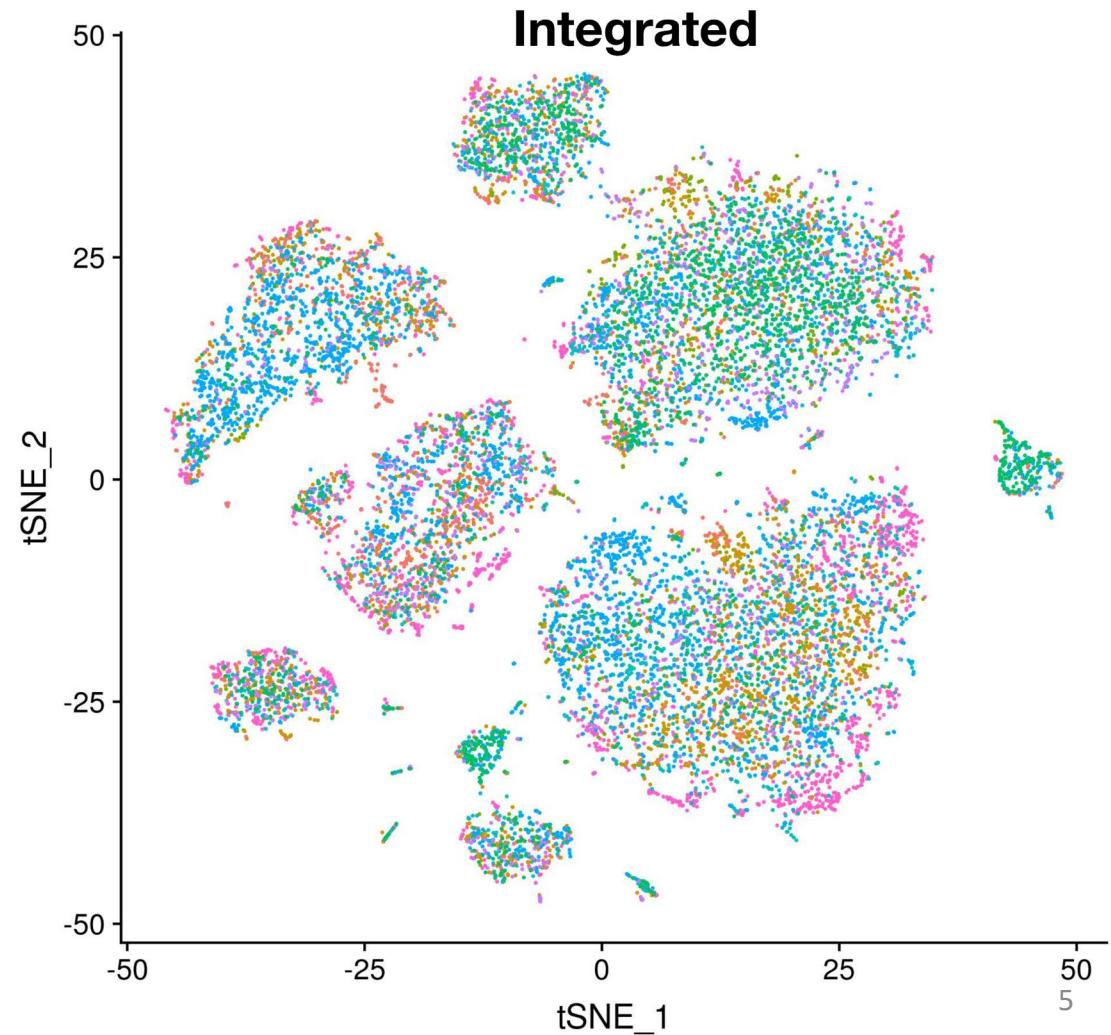
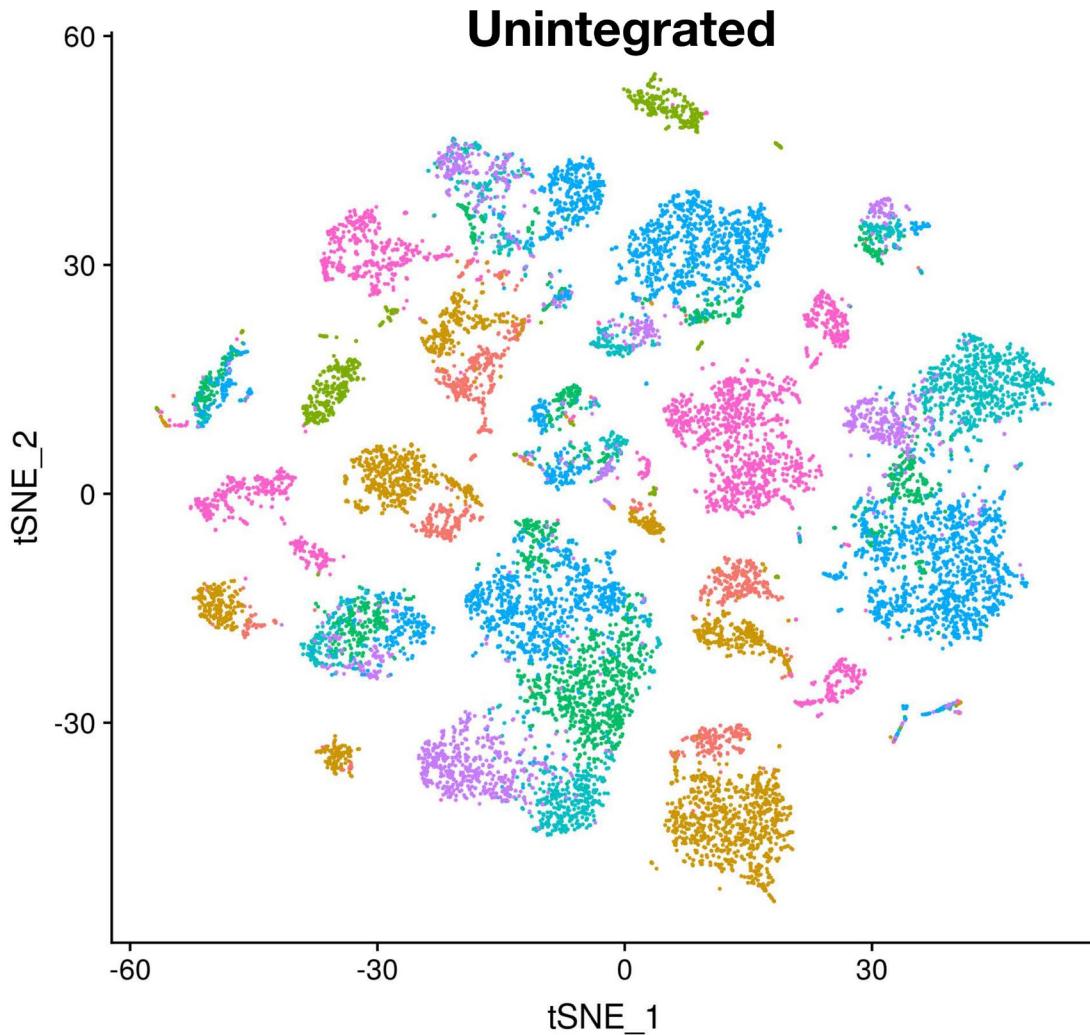
Building a cell atlas

8 maps of the human pancreas



Building a cell atlas

8 maps of the human pancreas

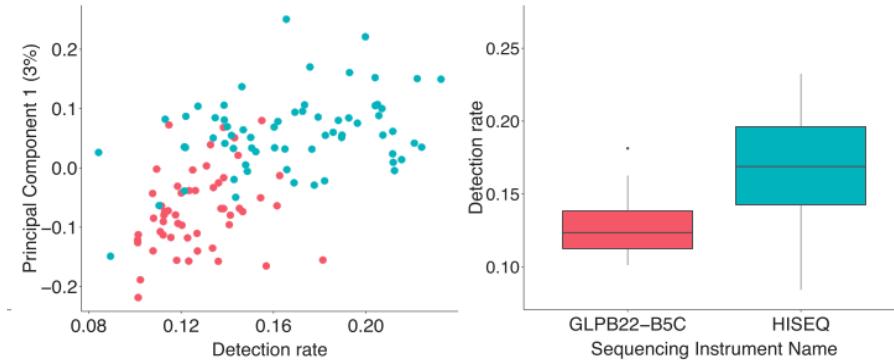


Confounders and batch effects

1. Technical variability

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

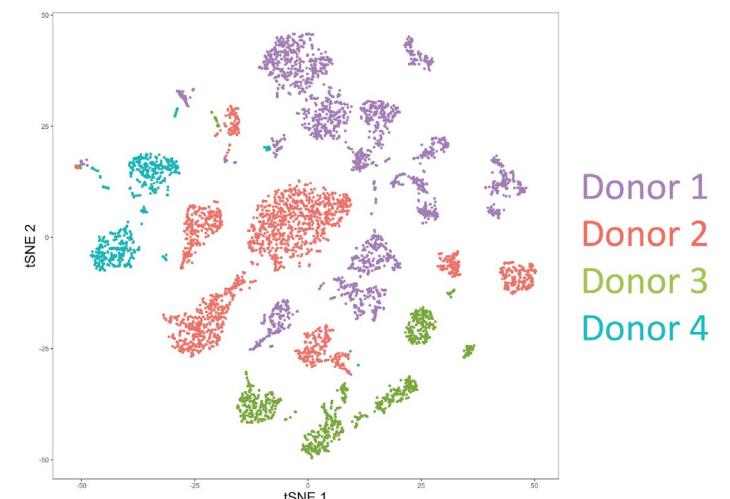
Technical ‘batch effects’ confound downstream analysis



2. Biological variability

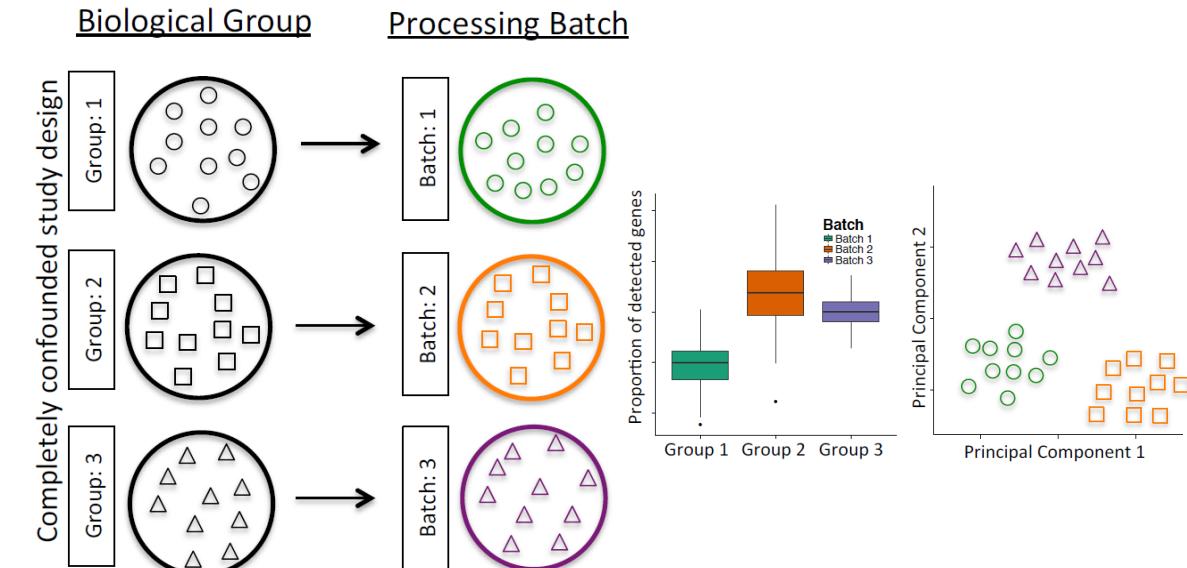
- Patient differences
- Evolution! (cross-species analysis)

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



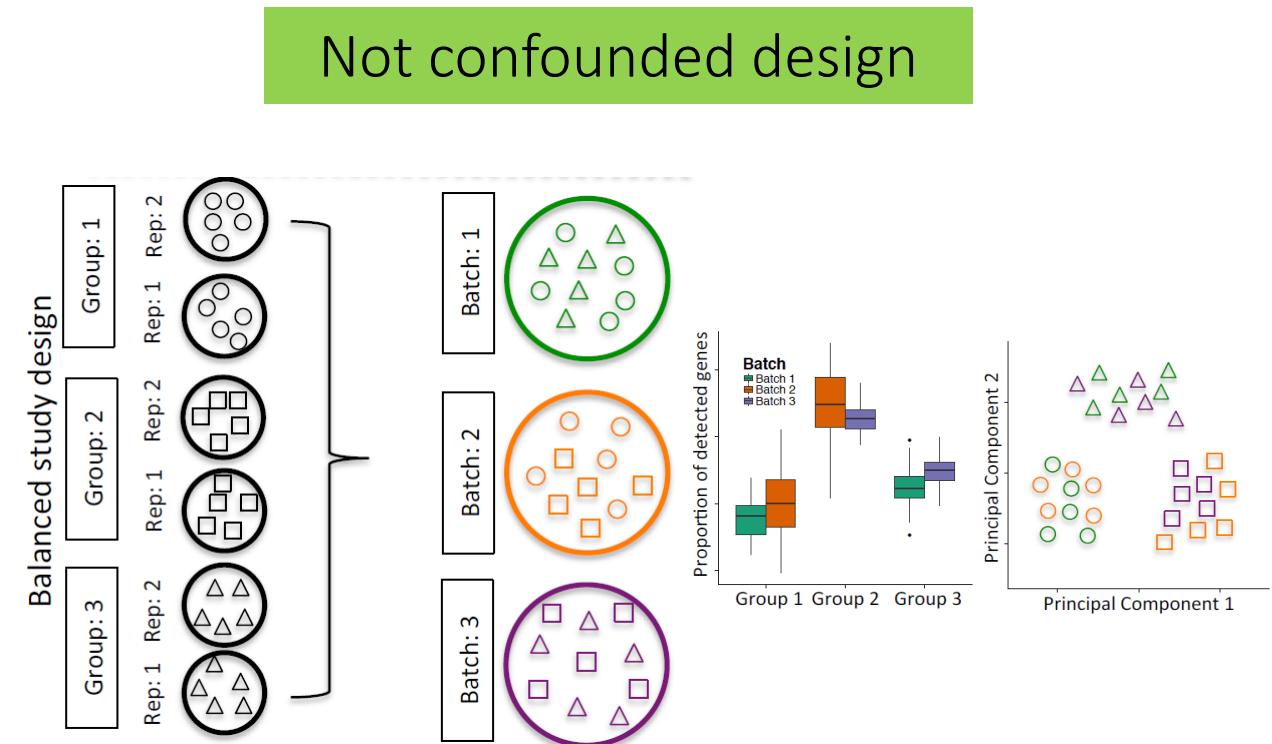
Confounders and batch effects

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.

Our agenda

- Single cell batch correction methods
- Performance assessment

Batch correction methods

- Many good options have been developed for bulk RNA-seq data:
 - RUVseq() or svaseq()
 - Linear models with e.g. removeBatchEffect() in limma or scater
 - ComBat() in sva
 - ...
- But bulk RNA-seq methods make modelling assumptions that are likely to be violated in scRNAseq data
 - The composition of cell populations are either known or the same across batches
 - Batch effect is additive: batch-induced fold-change in expression is the same across different cell subpopulations for any given gene

Batch correction methods

- MNNcorrect (<https://doi.org/10.1038/nbt.4091>)
- CCA + anchors (Seurat v3) (<https://doi.org/10.1101/460147>)
- CCA + dynamic time warping (Seurat v2) (<https://doi.org/10.1038/nbt.4096>)
- LIGER (<https://doi.org/10.1101/459891>)
- Harmony (<https://doi.org/10.1101/461954>)
- Scanorama (<https://doi.org/10.1101/371179>)
- scMerge (<https://doi.org/10.1073/pnas.1820006116>)
- BBKNN (<https://doi.org/10.1093/bioinformatics/btz625>)
- scGen (<https://doi.org/10.1038/s41592-019-0494-8>)
- ...

Two broad strategies:

- Joint dimension reduction
- Graph-based approaches

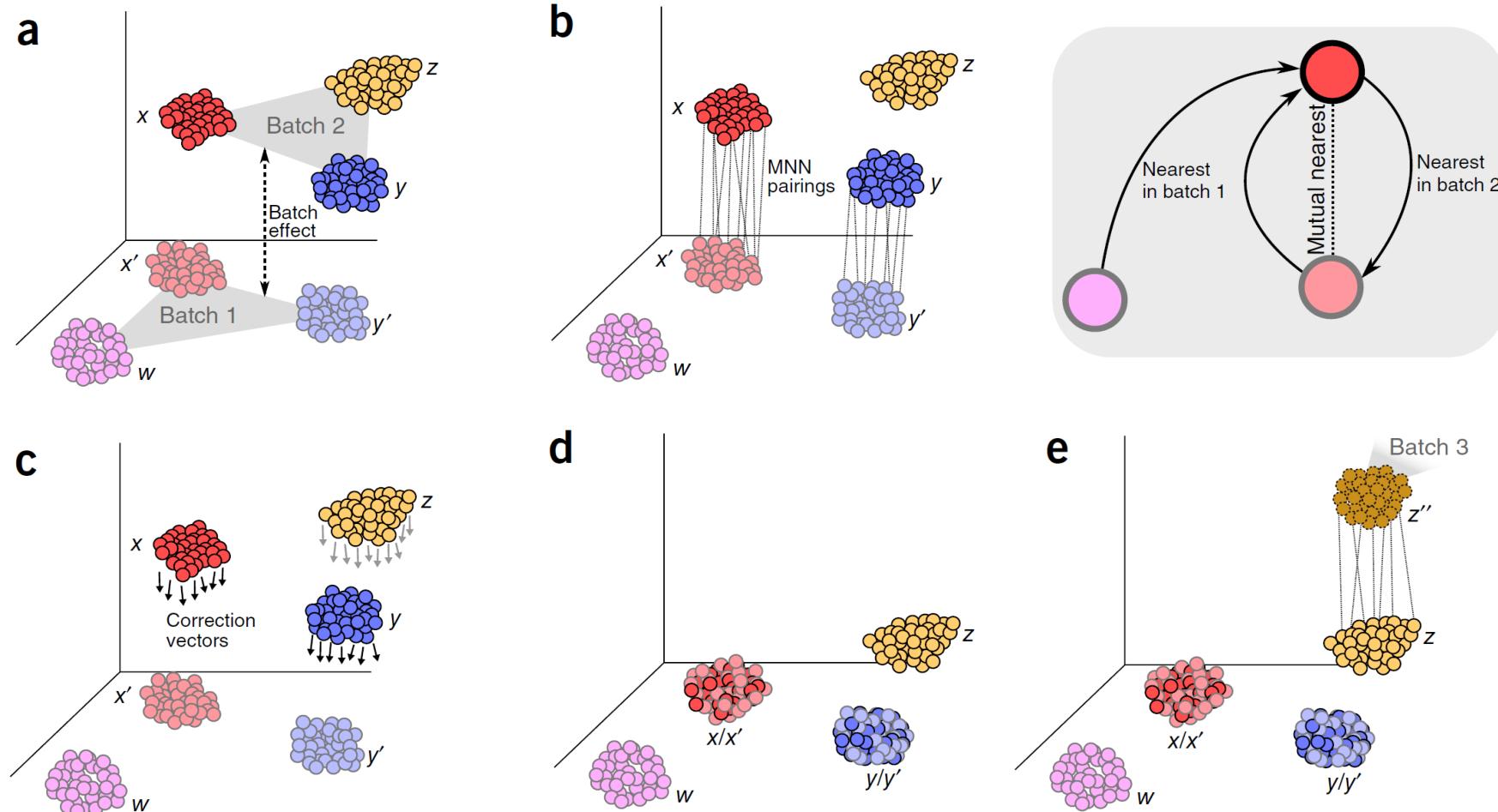
Batch correction methods

- **MNNcorrect** (<https://doi.org/10.1038/nbt.4091>)
- **CCA + anchors (Seurat v3)** (<https://doi.org/10.1101/460147>)
- CCA + dynamic time warping (Seurat v2) (<https://doi.org/10.1038/nbt.4096>)
- LIGER (<https://doi.org/10.1101/459891>)
- **Harmony** (<https://doi.org/10.1101/461954>)
- Scanorama (<https://doi.org/10.1101/371179>)
- scMerge (<https://doi.org/10.1073/pnas.1820006116>)
- BBKNN (<https://doi.org/10.1093/bioinformatics/btz625>)
- scGen (<https://doi.org/10.1038/s41592-019-0494-8>)
- ...

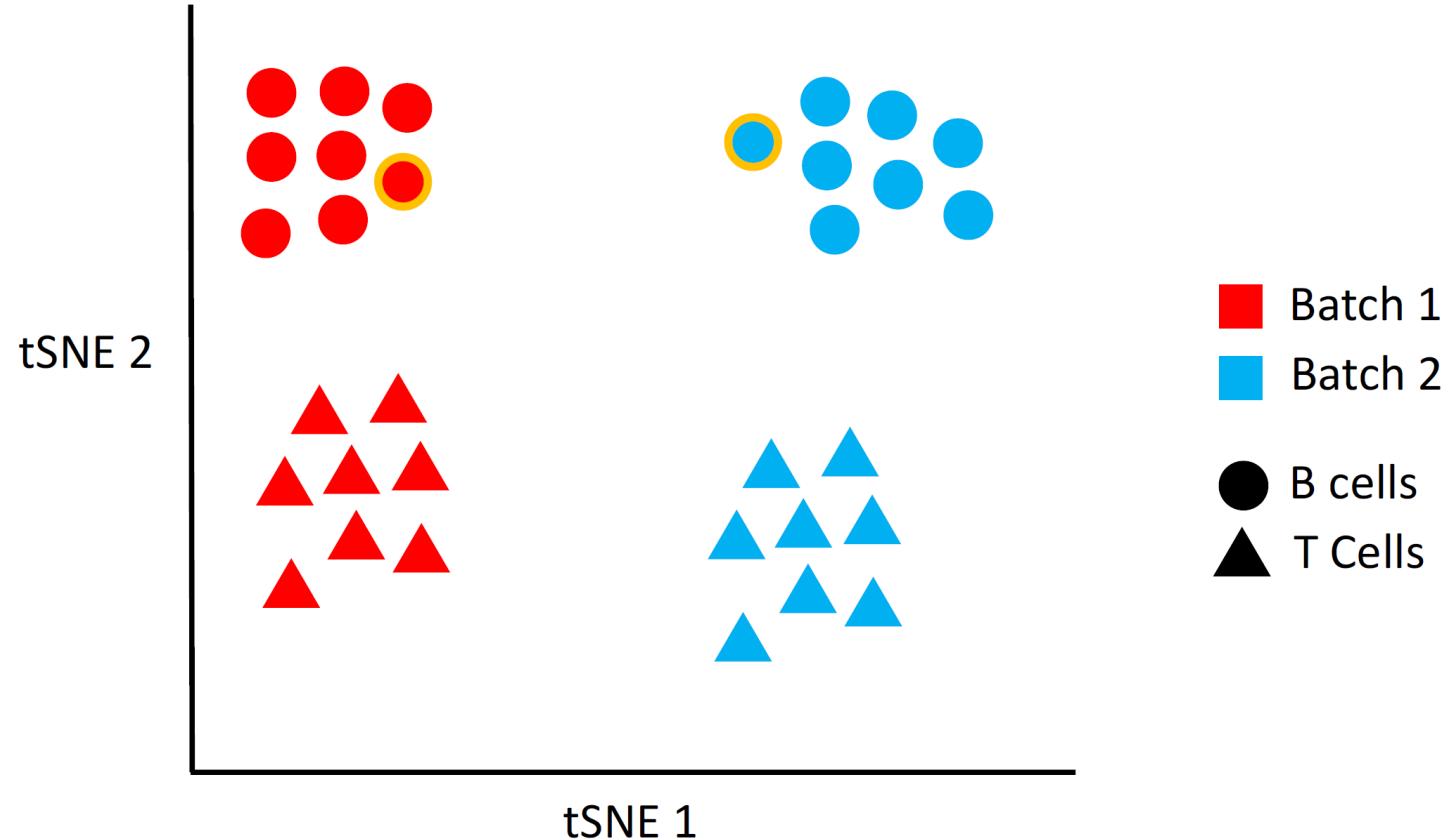
Two broad strategies:

- Joint dimension reduction
- Graph-based approaches

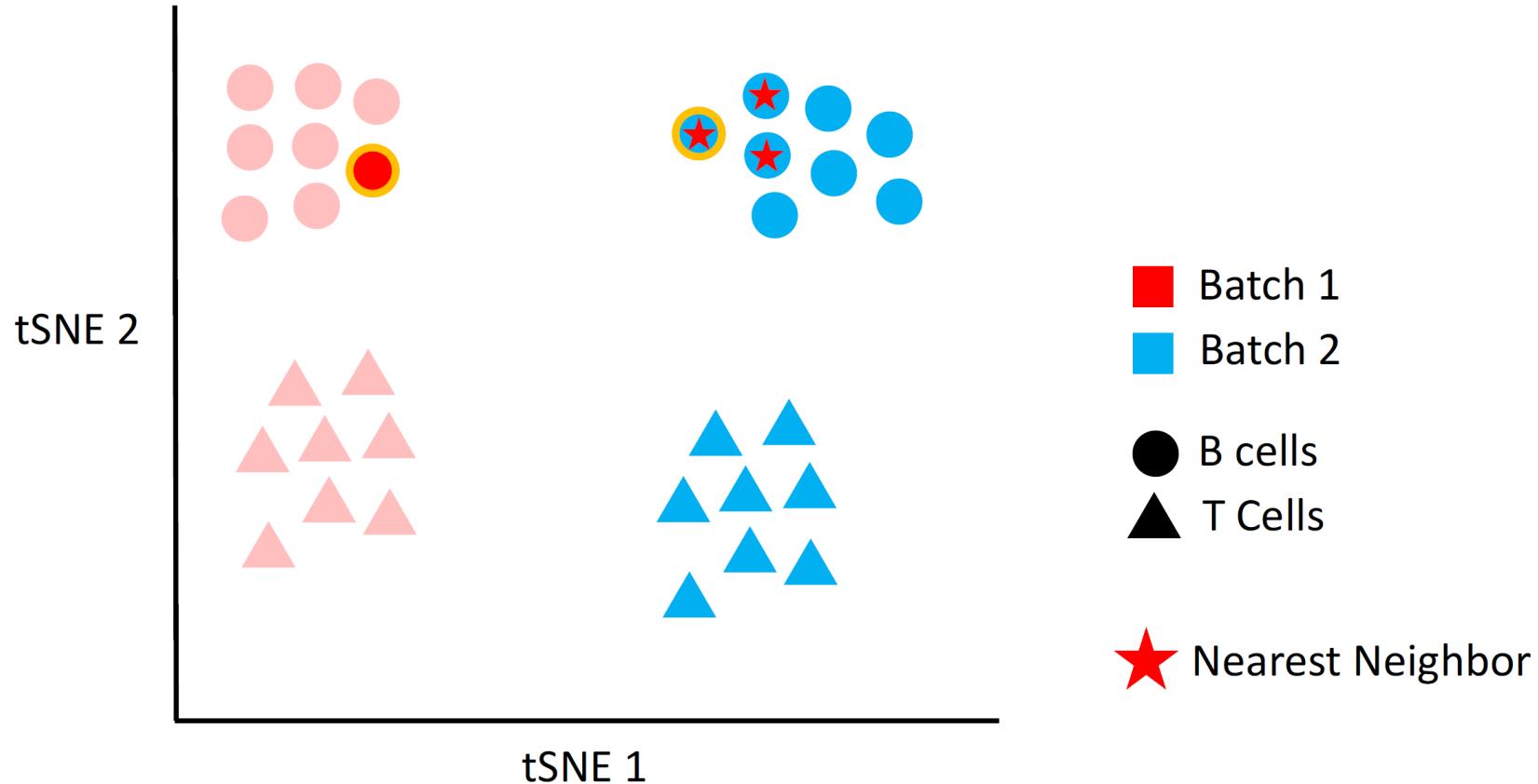
Mutual Nearest Neighbors (MNN)



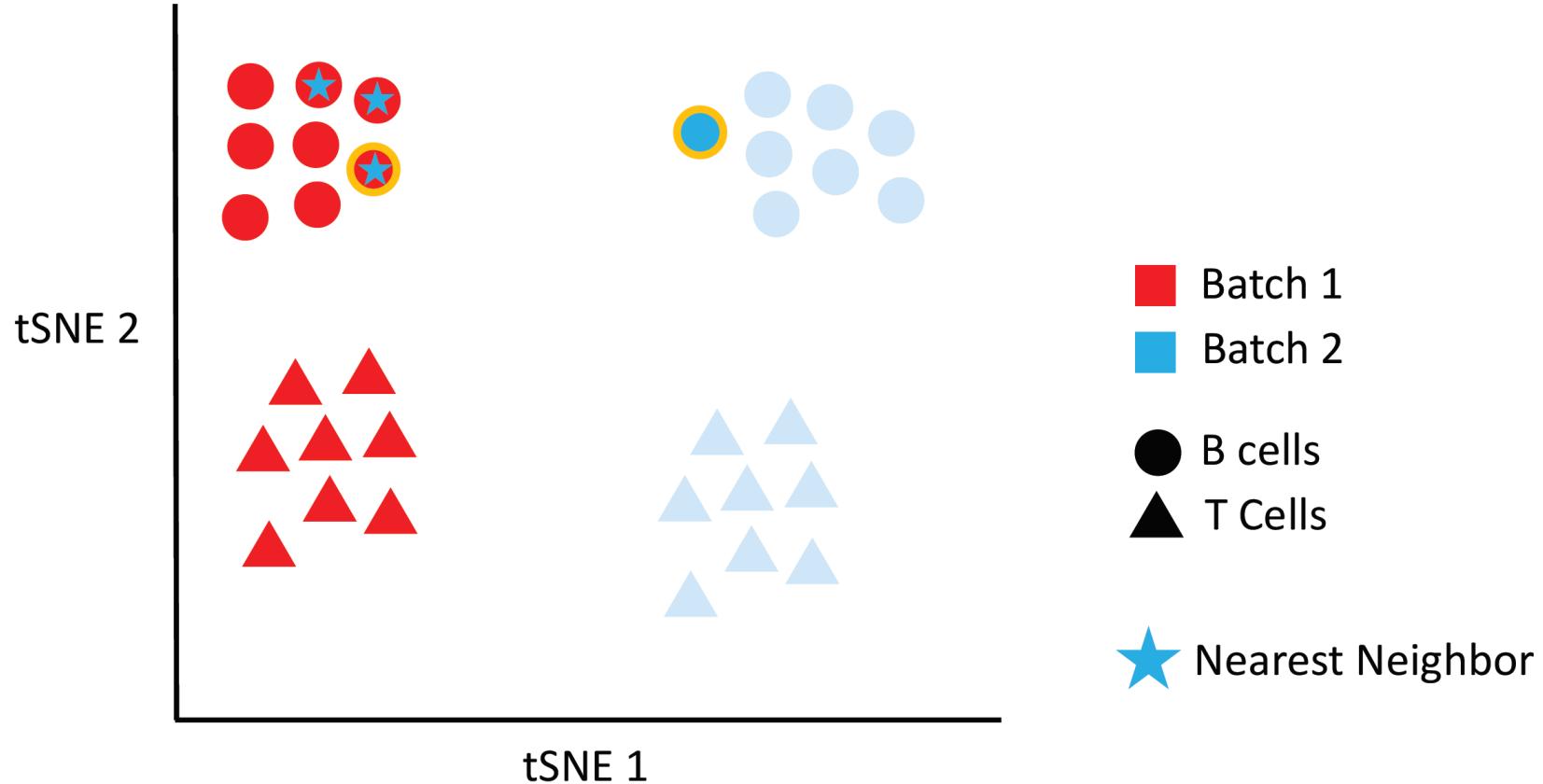
Mutual Nearest Neighbors (MNN)



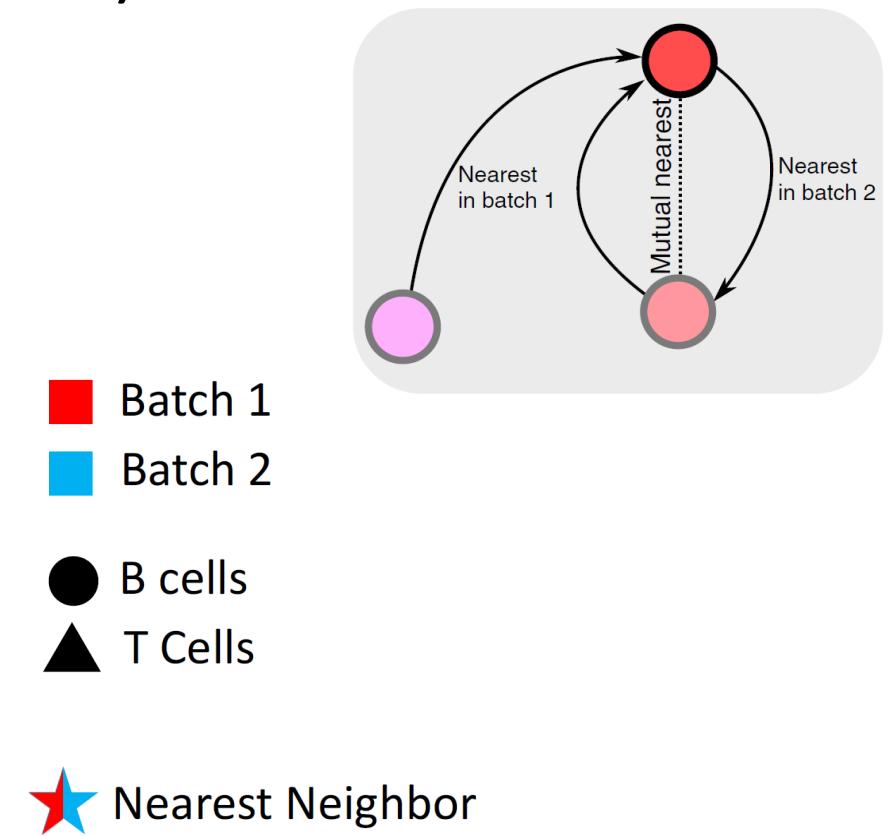
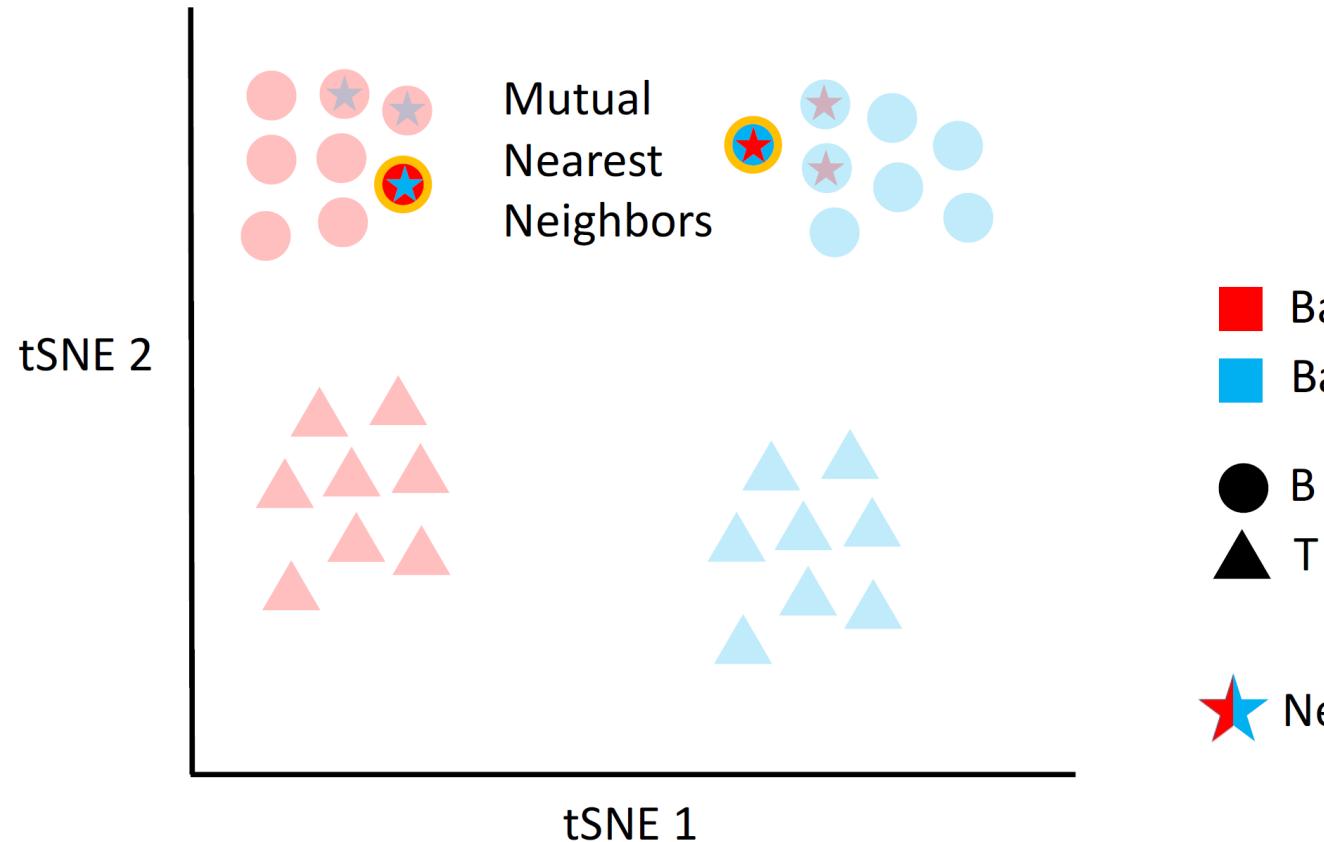
Mutual Nearest Neighbors (MNN)



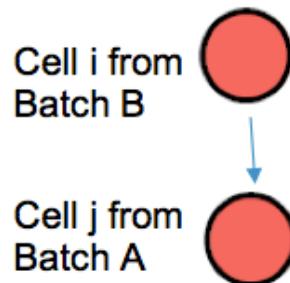
Mutual Nearest Neighbors (MNN)



Mutual Nearest Neighbors (MNN)

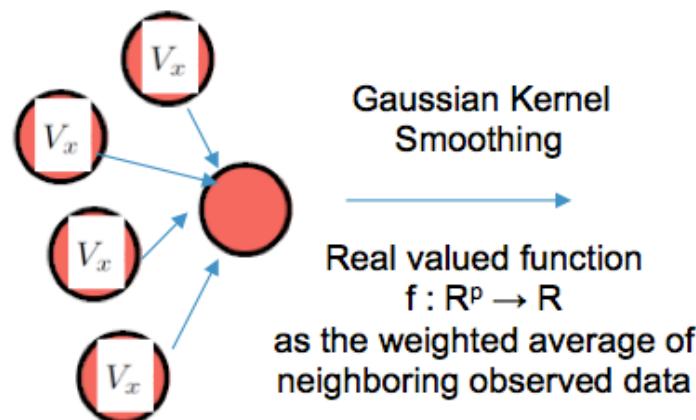


Mutual Nearest Neighbors (MNN)



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.

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.



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

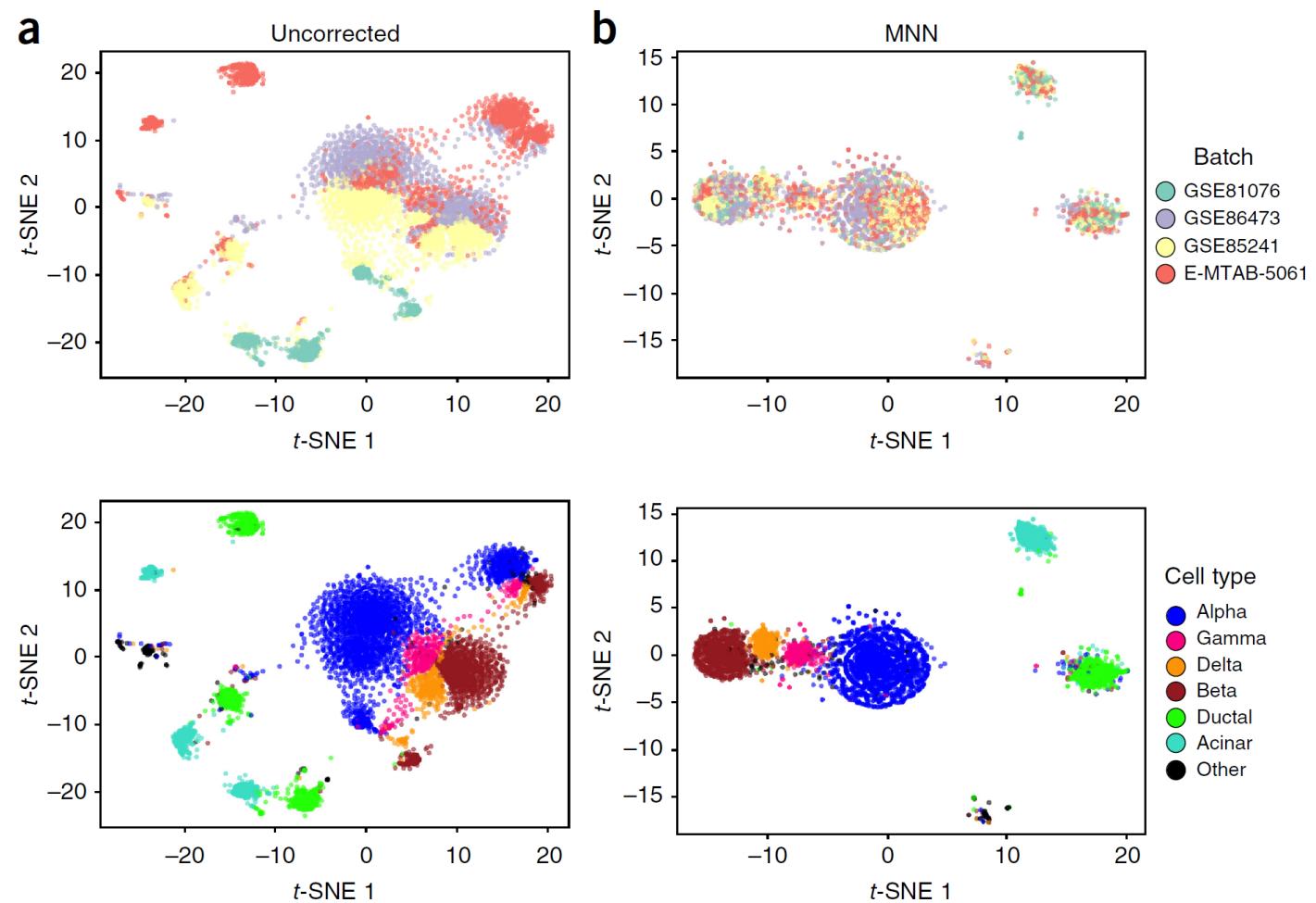
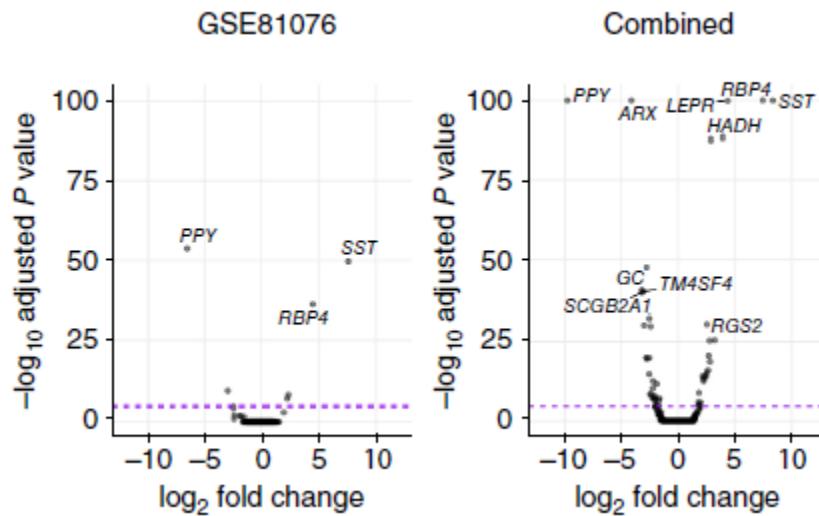
Batch Correction vector for each cell



Mutual Nearest Neighbors (MNN)

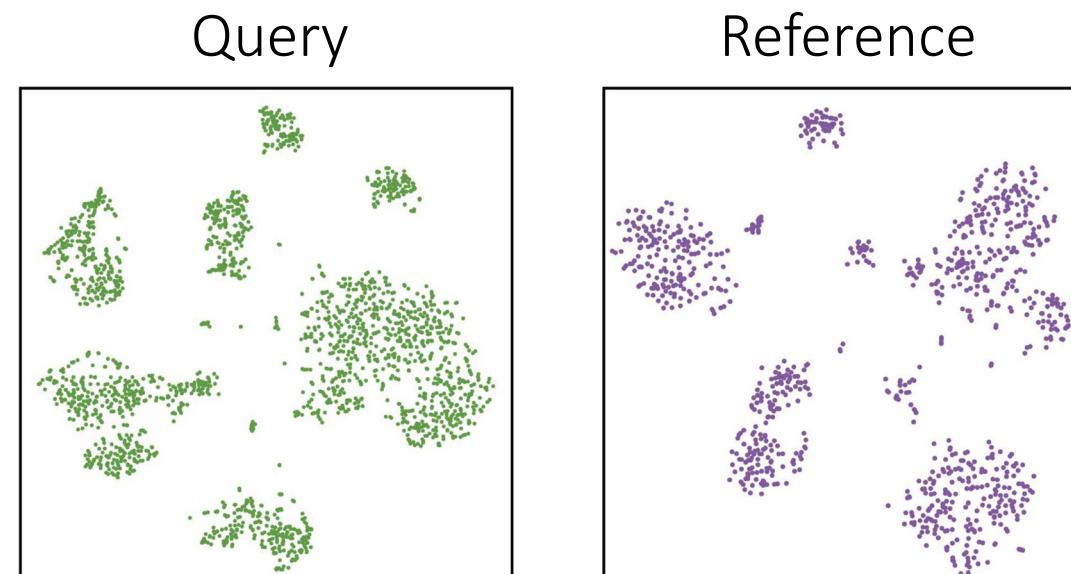
- Pooling experiments -> increased statistical power

Delta vs Gamma Islet Cells

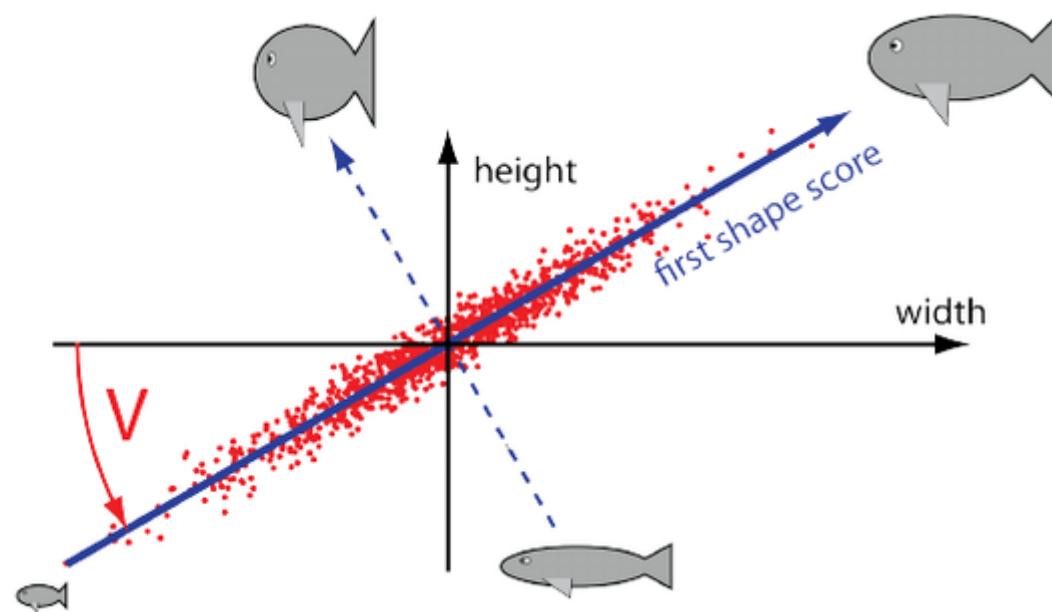
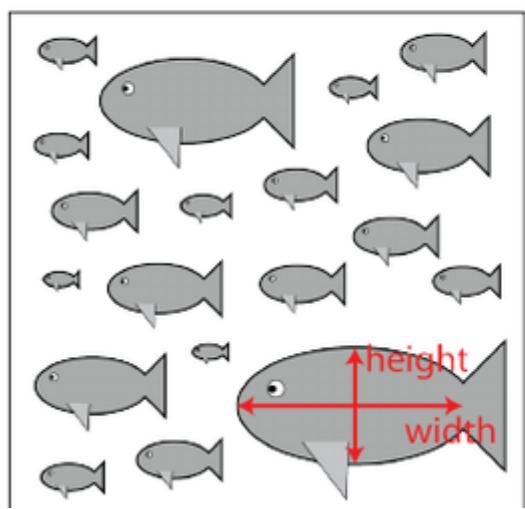


CCA + anchors (Seurat v3)

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

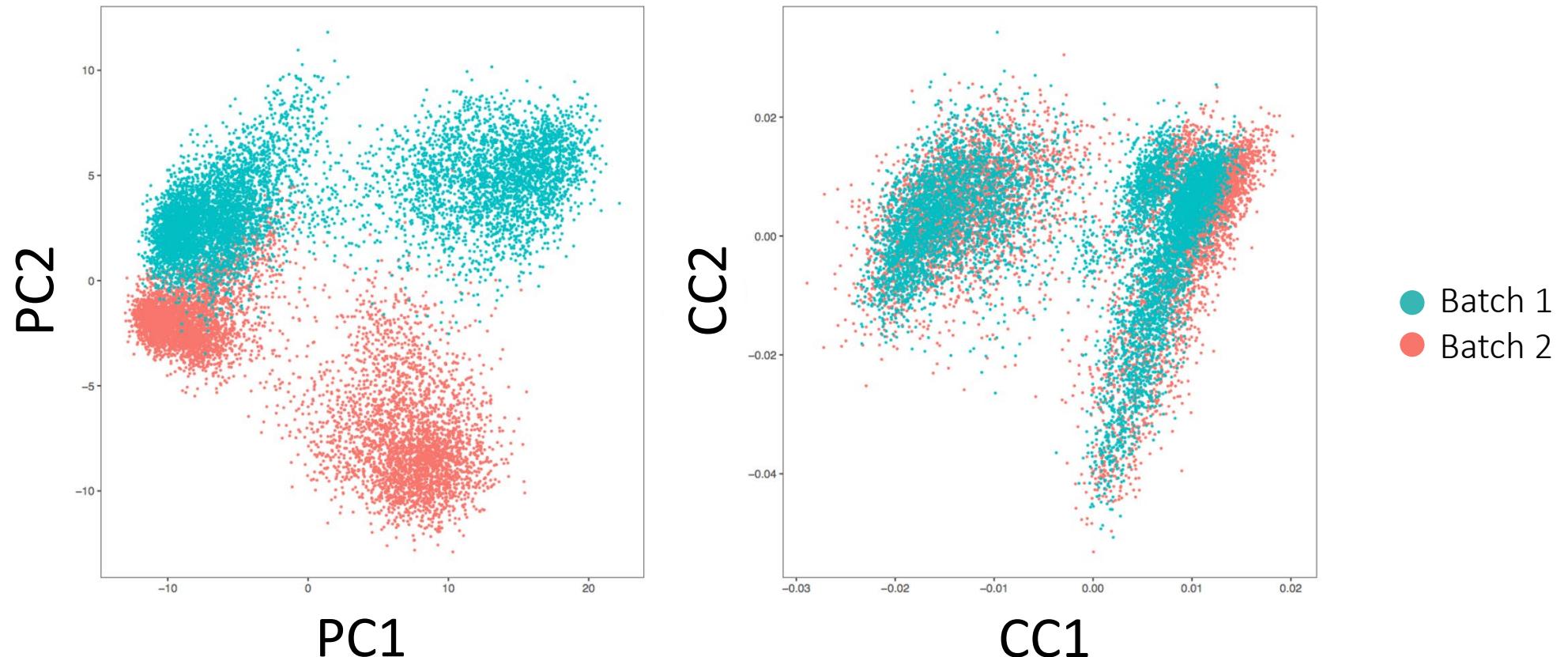


Principle component analysis



Finding corresponding cells

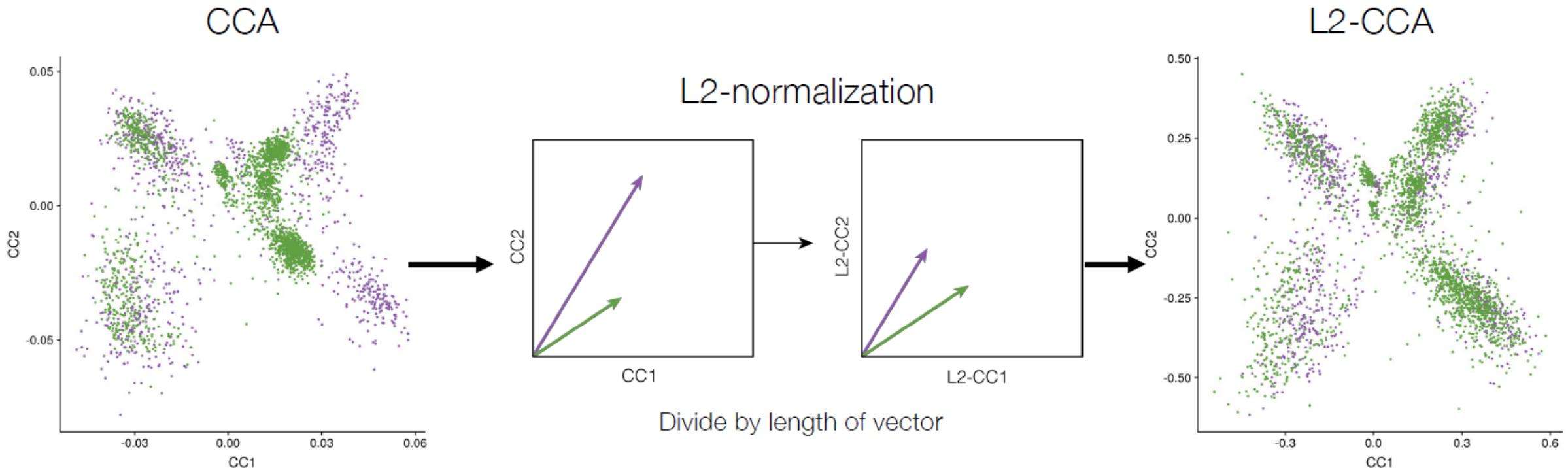
Canonical correlation analysis and normalization



CCA captures correlated sources of variation between two datasets

Finding corresponding cells

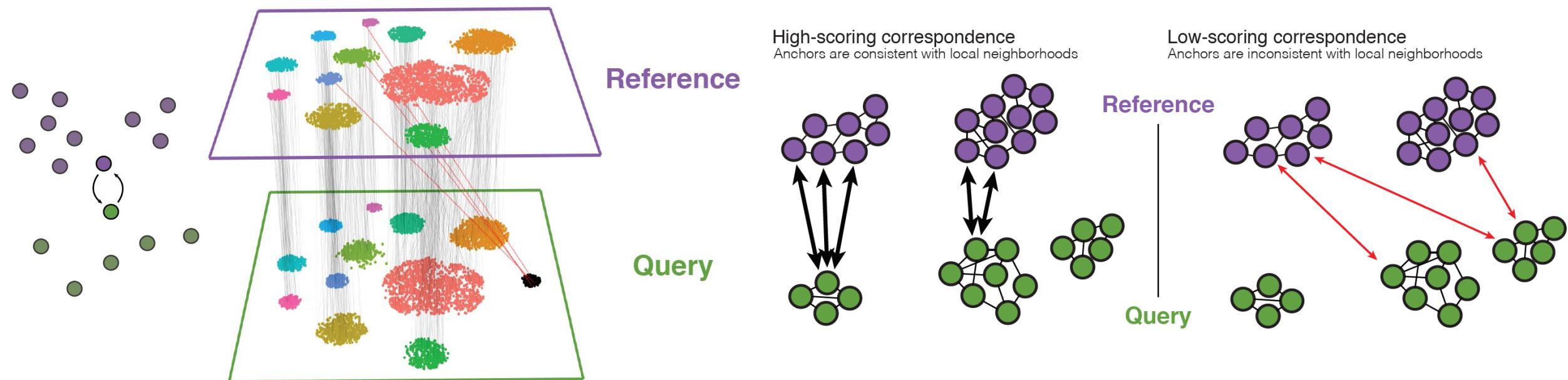
Canonical correlation analysis and normalization



L2-normalization corrects for differences in scale

Finding corresponding cells

Anchors: mutual nearest neighbors



Finding corresponding cells

Data integration

1. Calculate the matrix B , where each column represents the difference between the two expression vectors for every pair of anchor cells a
2. Construct a weight matrix W that defines the strength of association between each query cell c , and each anchor i
3. Calculate a transformation matrix C using the previously computed weights matrix and the integration matrix as
4. Subtract the transformation matrix C from the original expression matrix Y to produce the integrated expression matrix \hat{Y}

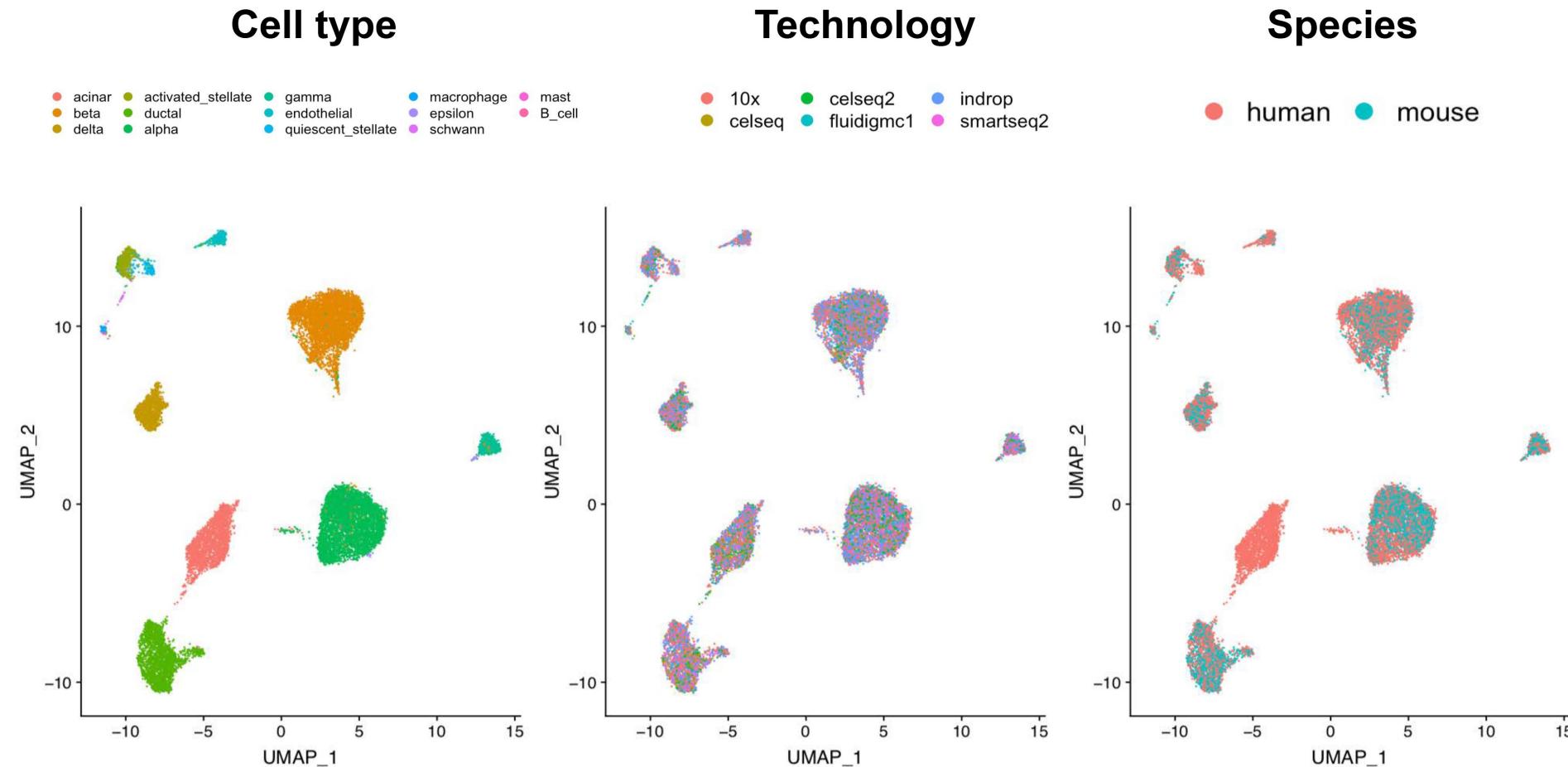
$$B = X[, a] - Y[, a]$$

$$W_{c,i} = \frac{\tilde{D}_{c,i}}{\sum_1^{j=k.weight} \tilde{D}_{c,j}}$$

$$C = BW^T$$

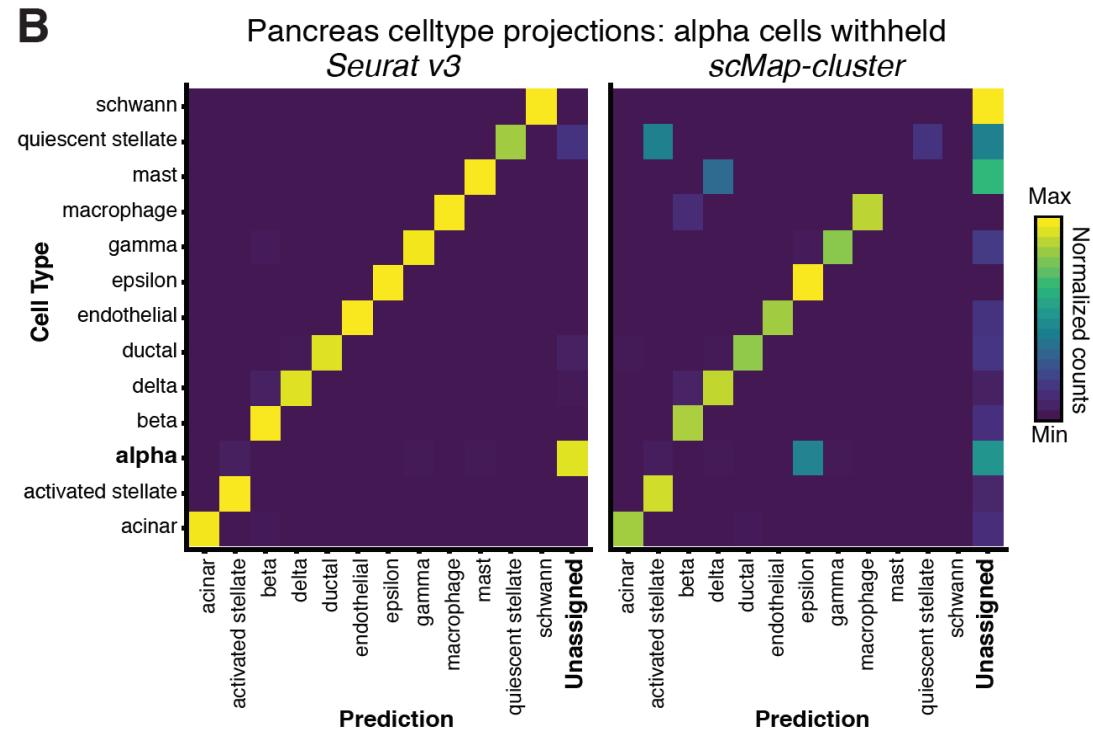
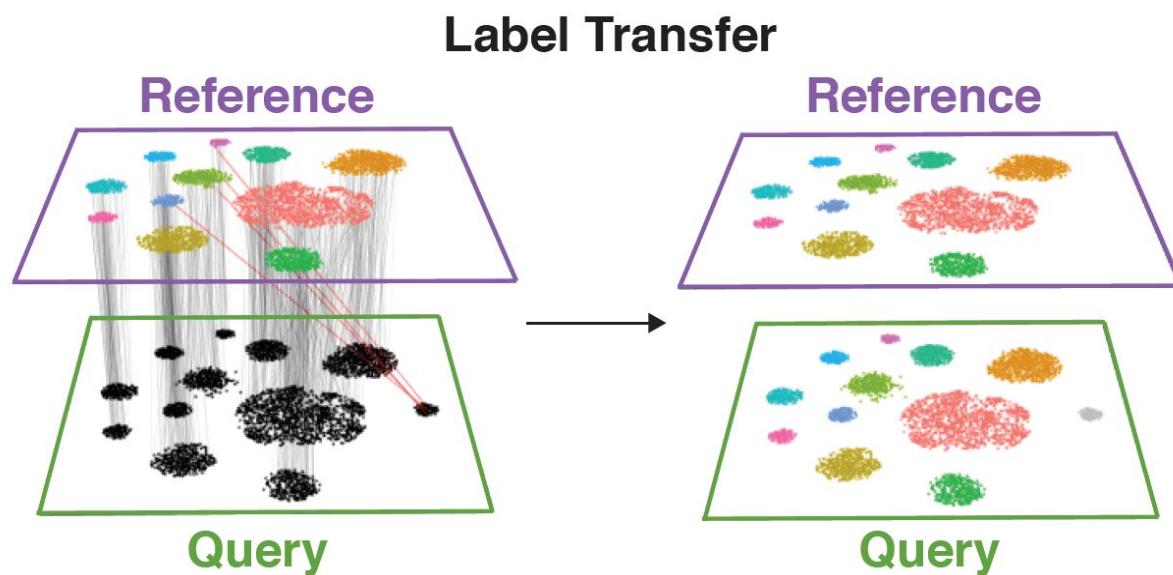
$$\hat{Y} = Y - C$$

CCA + anchors (Seurat v3)



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

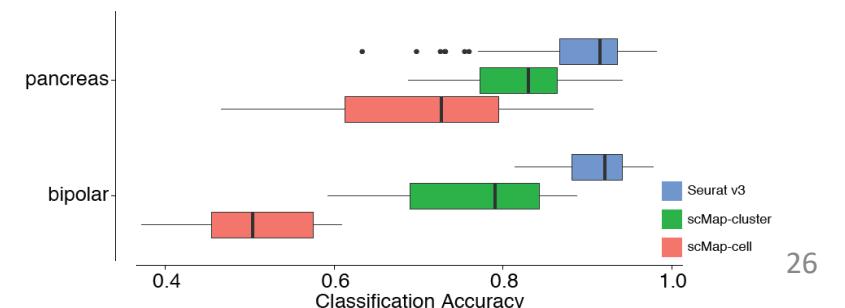
Label transfer (classification)



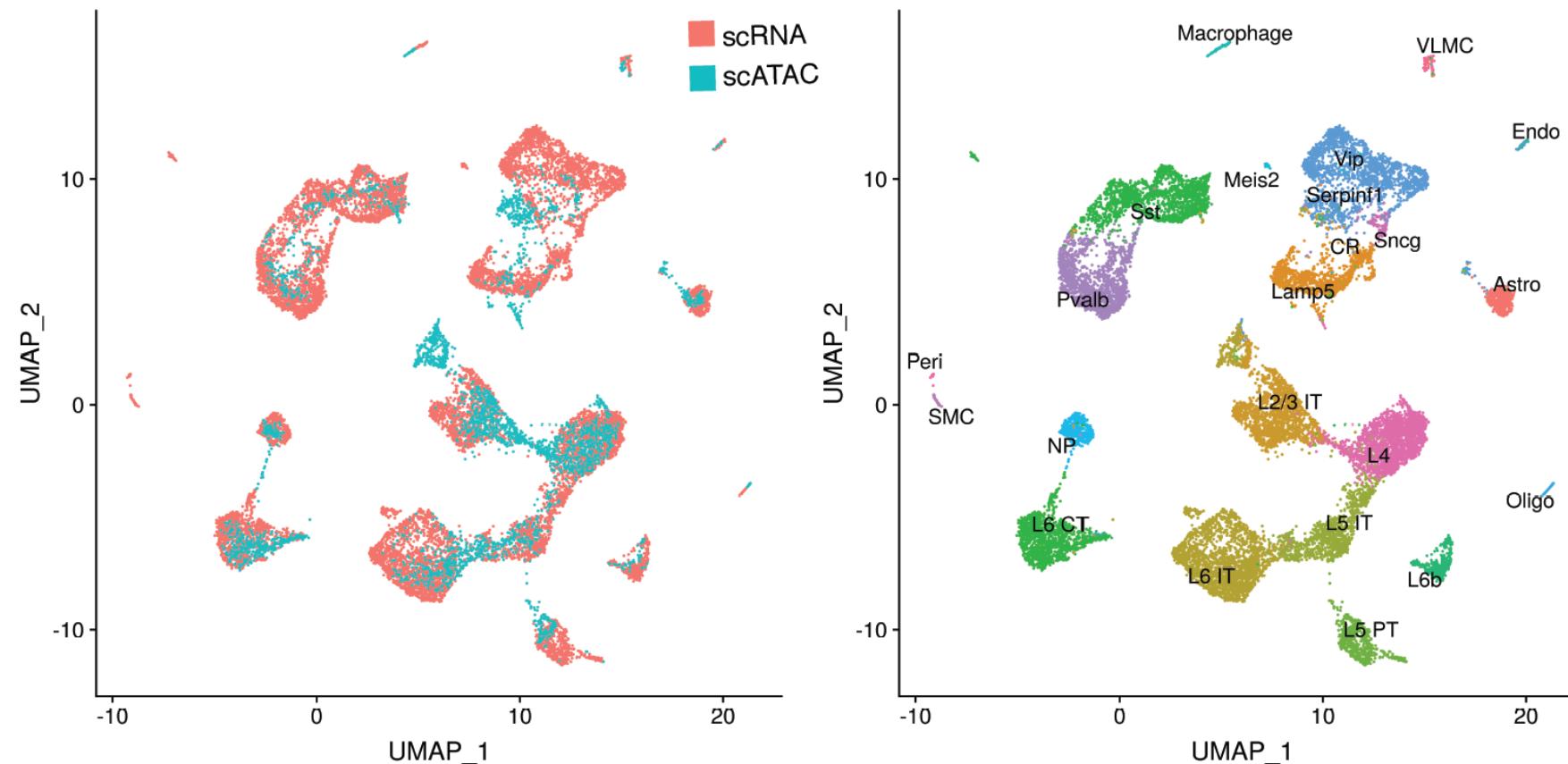
Weighted vote classifier

What is the classification of each cells nearest anchors?

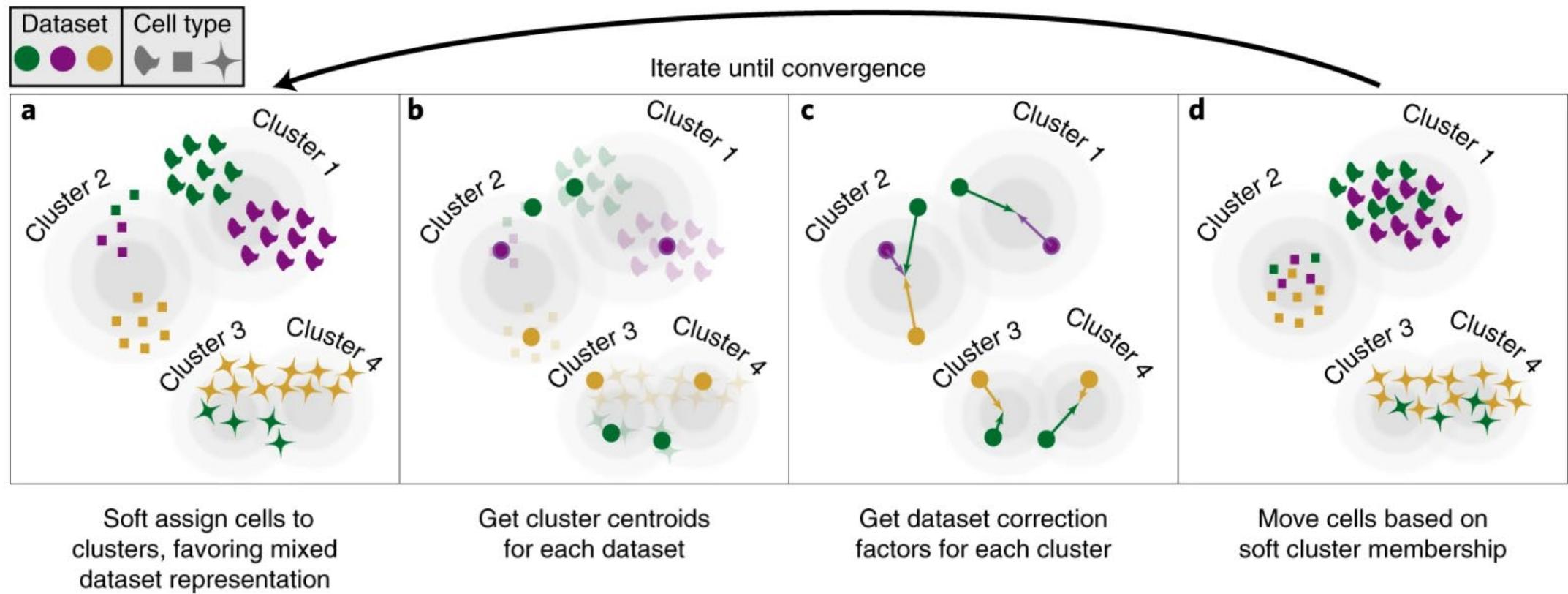
$$P_i = LW^T$$



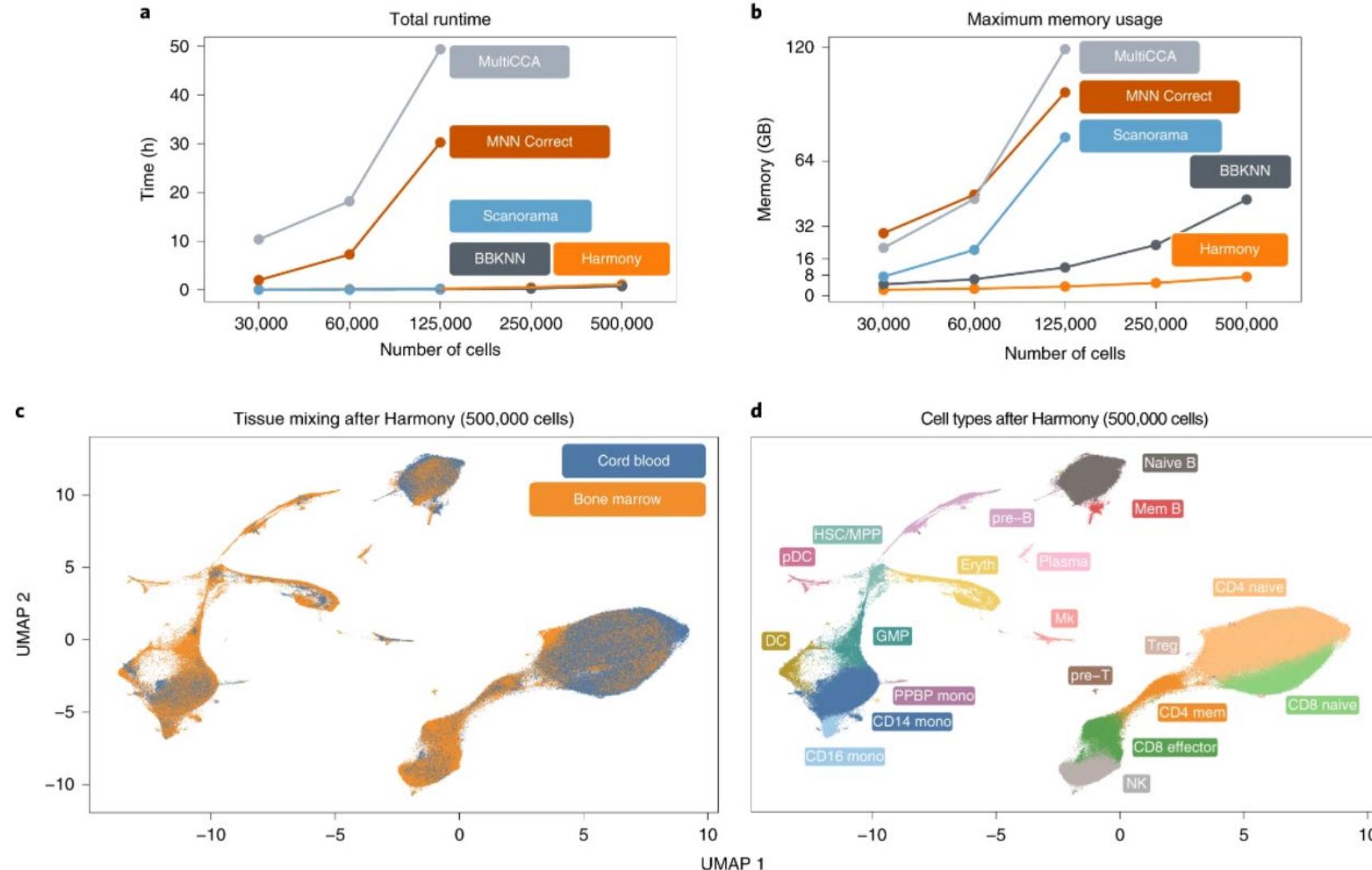
Integration across modalities



Harmony



Harmony



Using the corrected values

- Batch correction facilitates cell-based analysis of population heterogeneity in a consistent manner across batches.
 - No need to identify mappings between separate clusterings
 - Increased number of cells allows for greater resolution of population structure
- BUT...
- It is not recommended to use the corrected expression values for gene-based analyses (e.g. differential expression)
- Arbitrary correction algorithms are not obliged to preserve the magnitude (or even direction) of differences in per-gene expression when attempting to align multiple batches

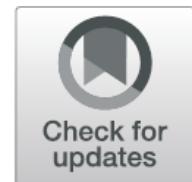
Using the corrected values

- Batch correction facilitates cell-based analysis of population heterogeneity in a consistent manner across batches.
 - No need to identify mappings between separate clusterings
 - Increased number of cells allows for greater resolution of population structure
- BUT... (except **scMerge**, MNNcorrect, Seurat v3, Scanorama and scGen)
- It is not recommended to use the corrected expression values for gene-based analyses (e.g. differential expression)
- Arbitrary correction algorithms are not obliged to preserve the magnitude (or even direction) of differences in per-gene expression when attempting to align multiple batches

RESEARCH

Open Access

A benchmark of batch-effect correction methods for single-cell RNA sequencing data



Hoa Thi Nhu Tran[†], Kok Siong Ang[†], Marion Chevrier[†], Xiaomeng Zhang[†], Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen^{*} 

Performance assessment

- Qualitative (visualization)
- Quantitative:
 - Silhouette score
 - kBET: k-nearest-neighbor batch-effect test
 - ...

Silhouette score

A score for each cell that assesses the separation of cell types, with a high score suggesting that cells of the same cell type are close together and far from other cells of a different type.

$a(i)$ is the average distance of cell i to all other cells within i 's cluster.

$b(i)$ is the average distance of i to all cells in the nearest cluster to which i does not belong.

Silhouette score:

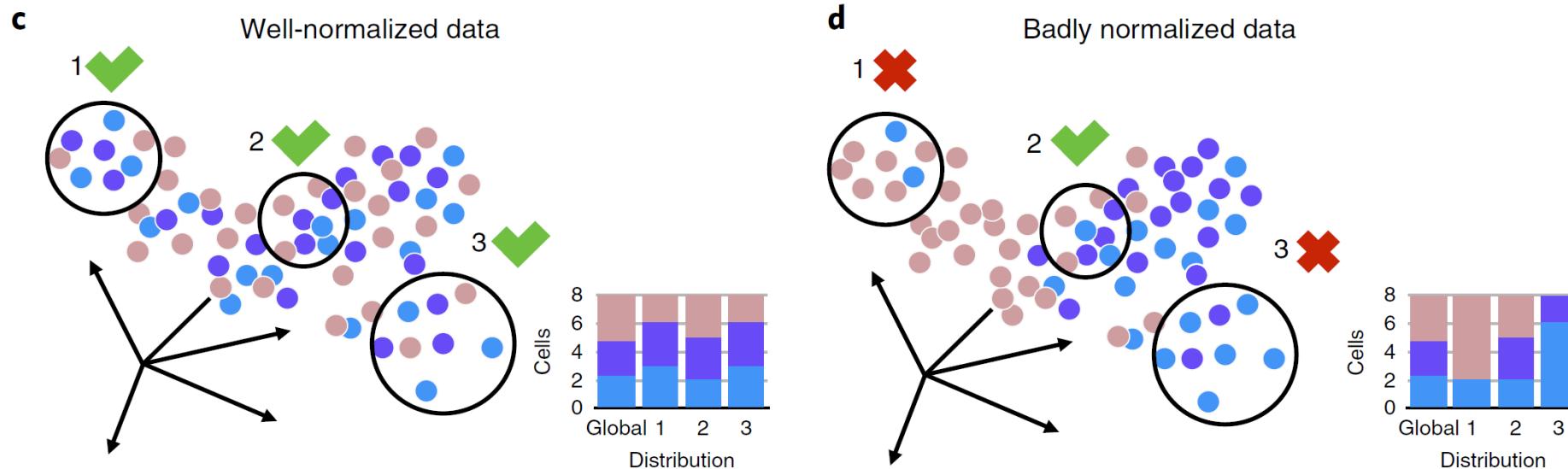
$$S = \frac{1}{N} \sum s(i)$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

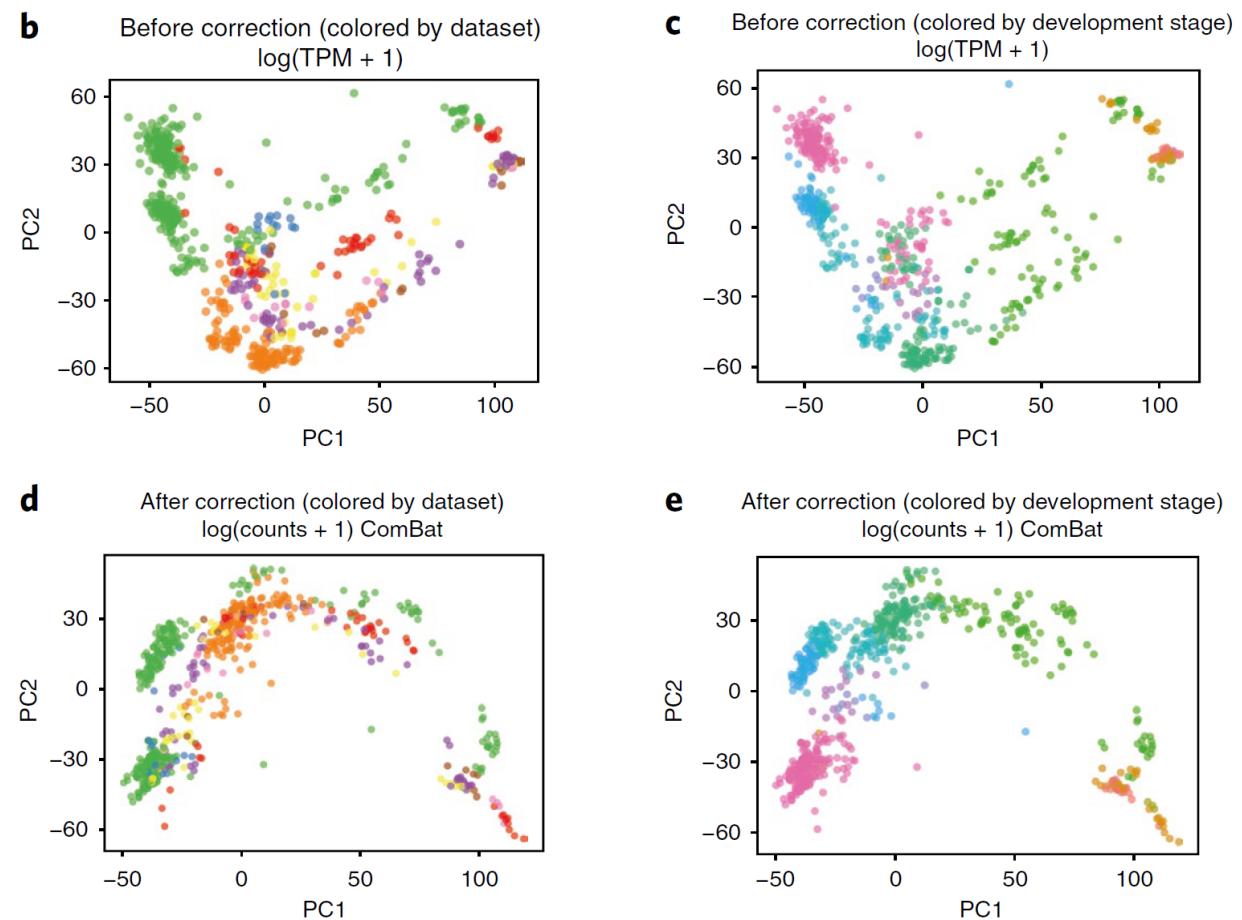
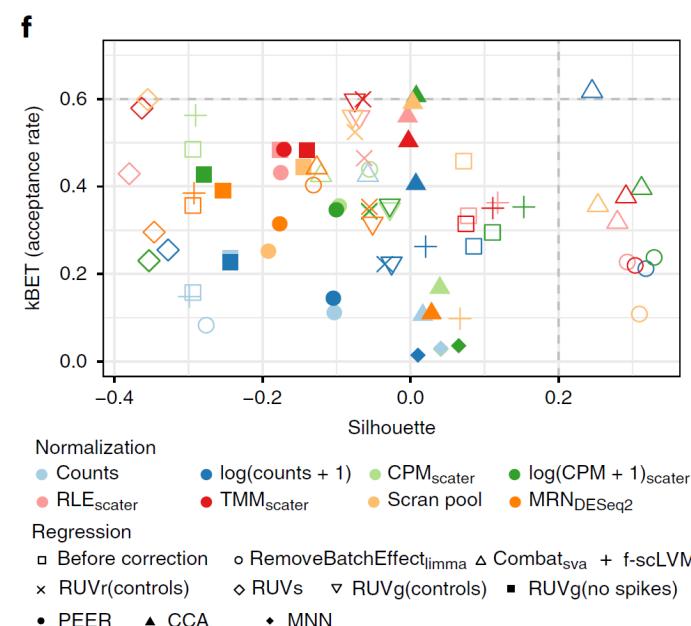
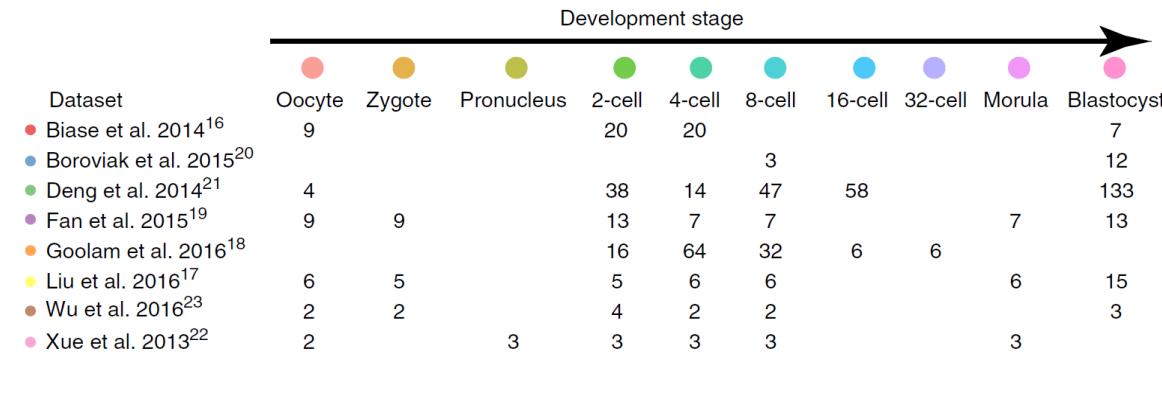
$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j)$$

$$b(i) = \min_{\forall j, j \notin C_i} d(x_i, x_j)$$

kBET: k -nearest-neighbor batch-effect test



kBET assesses data-integration quality



Summary (so far)

- Integration can allow us to **improve the interpretation** of single-cell data, and build a **multi-modal view** of the tissue
- Numerous methods now available for integration, mainly using **joint dimension reduction**, or **joint clustering**, or a combination of both
- Methods yielding **corrected expression matrix** can be used for further downstream analysis.

Data integration practical

- Seurat
- Harmony
- PBMC datasets

Resources

- Stuart et al. “Comprehensive integration of single-cell data”
<https://doi.org/10.1016/j.cell.2019.05.031>
- Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”
<https://doi.org/10.1038/s41592-019-0619-0>
- Tim Stuart “Integration and harmonization of single-cell data” (Satija Lab single cell genomics day 2019)
<https://satijalab.org/scgd/>
- Andrew Butler “Batch Correction and Data Integration for Single Cell Transcriptomics” (Satija Lab single cell genomics day 2018)
<https://satijalab.org/scgd18/>
- Orchestrating Single-Cell Analysis with Bioconductor
<https://osca.bioconductor.org/>
- Seurat Integration and Label Transfer tutorial
https://satijalab.org/seurat/v3.0/pancreas_integration_label_transfer.html
- Harmony portal <https://portals.broadinstitute.org/harmony/>
- Azimuth portal <https://azimuth.hubmapconsortium.org/>

Thank You!

 t.r.m.abdelaal@lumc.nl

