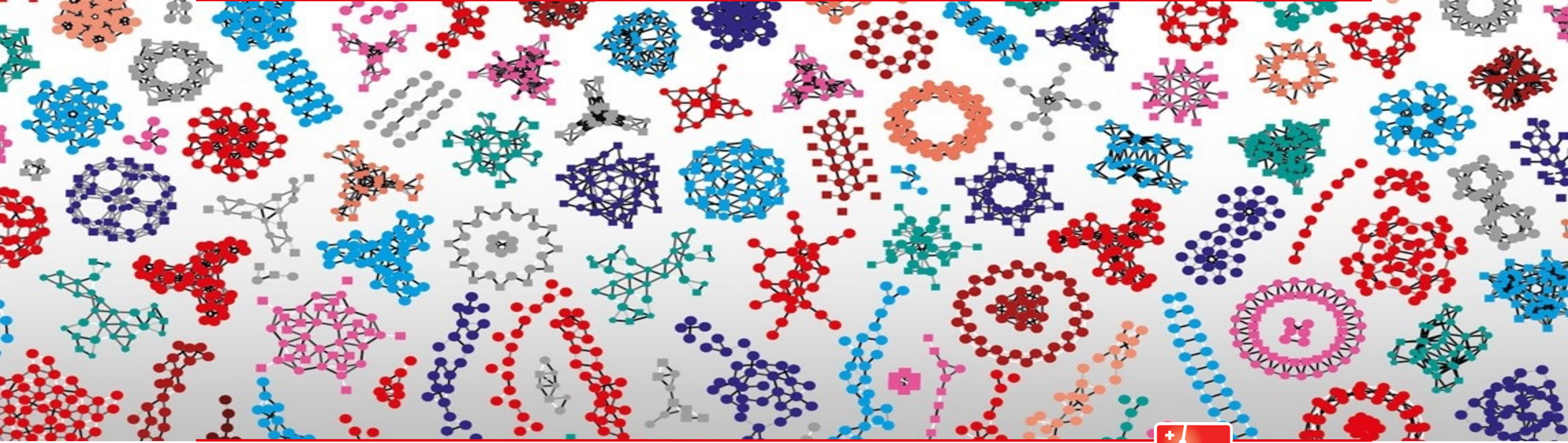


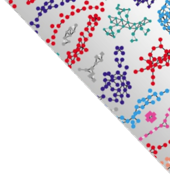
Swiss Institute of Bioinformatics

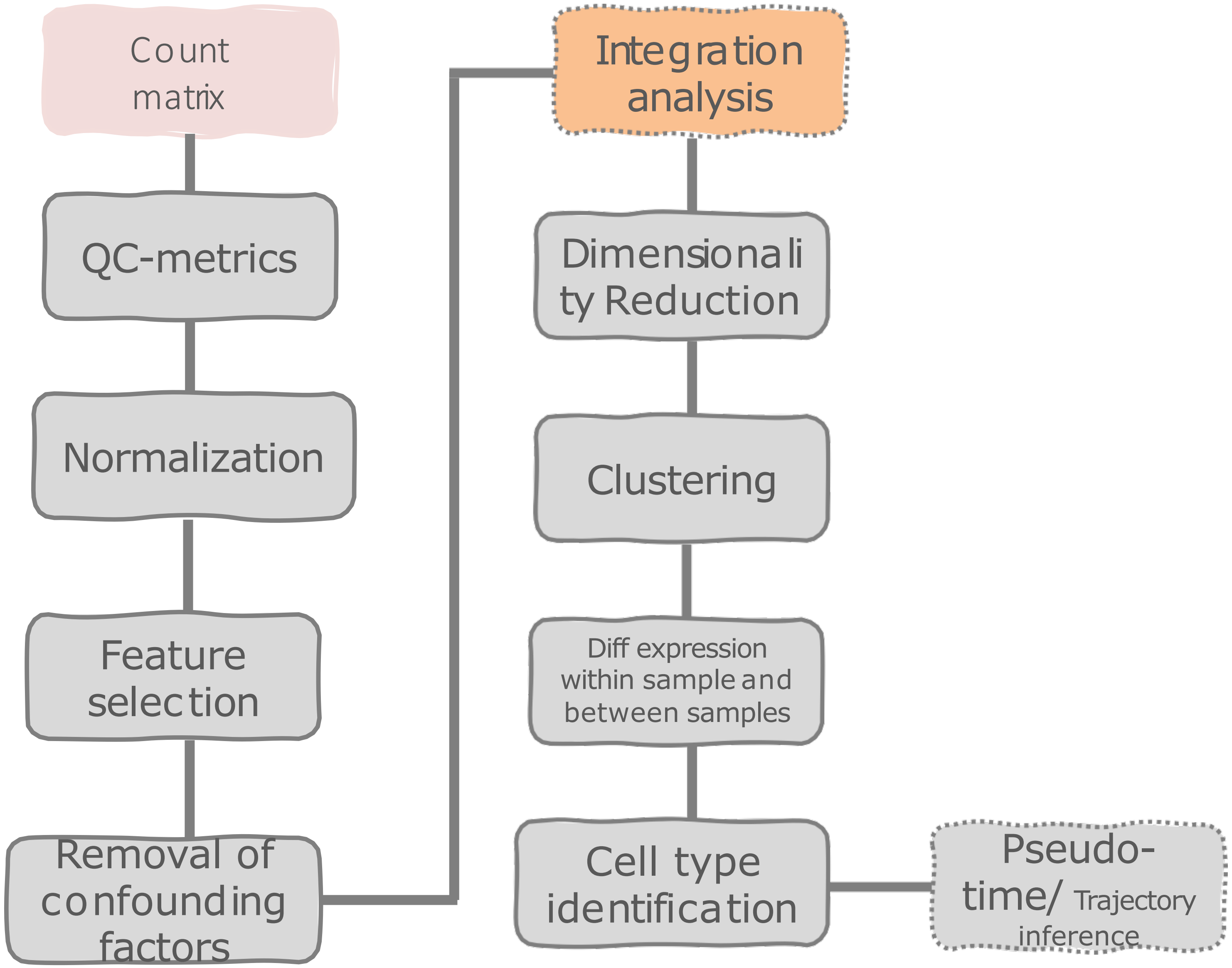
Day 2 : Single cell RNA sequencing: The bioinformatic downstream analysis

Geert van Geest, Rachel Marcone, Tania Wyss

www.sib.swiss

Integration





Count matrix

QC-metrics

Normalization

Feature selection

Removal of confounding factors

Integration analysis

Dimensionality Reduction

Clustering

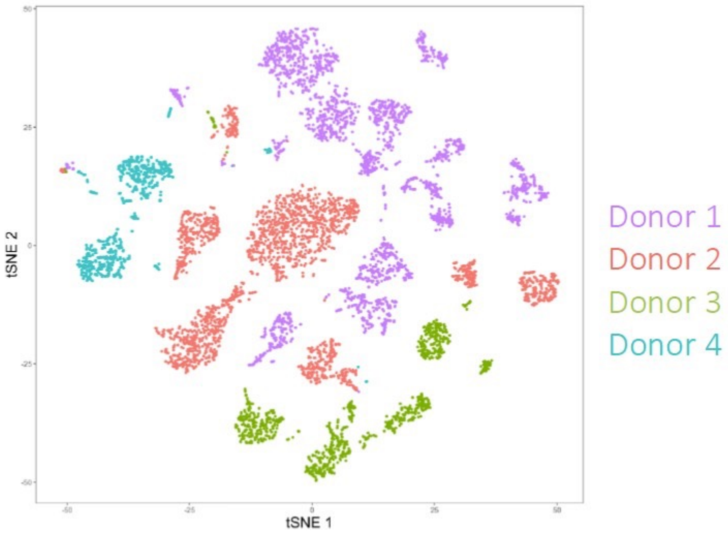
Diff expression within sample and between samples

Cell type identification

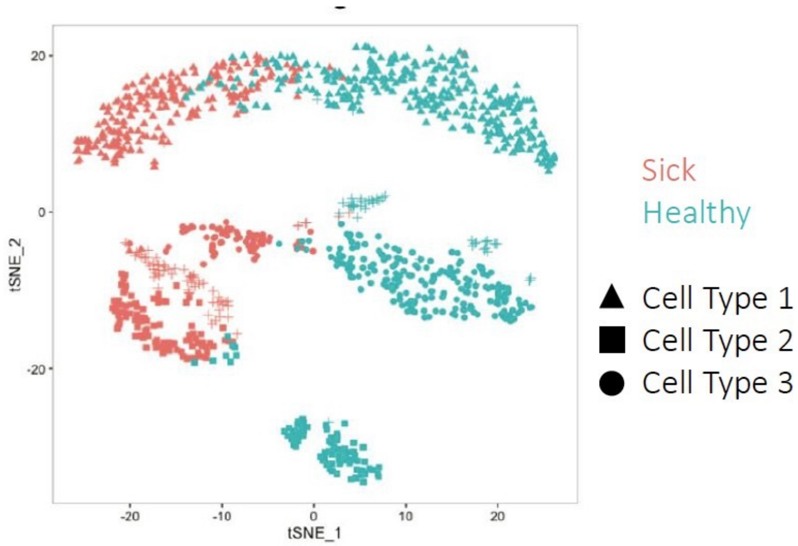
Pseudo-time/ Trajectory inference

Integration analysis

- Why do we integrate?



Same tissue from different donors

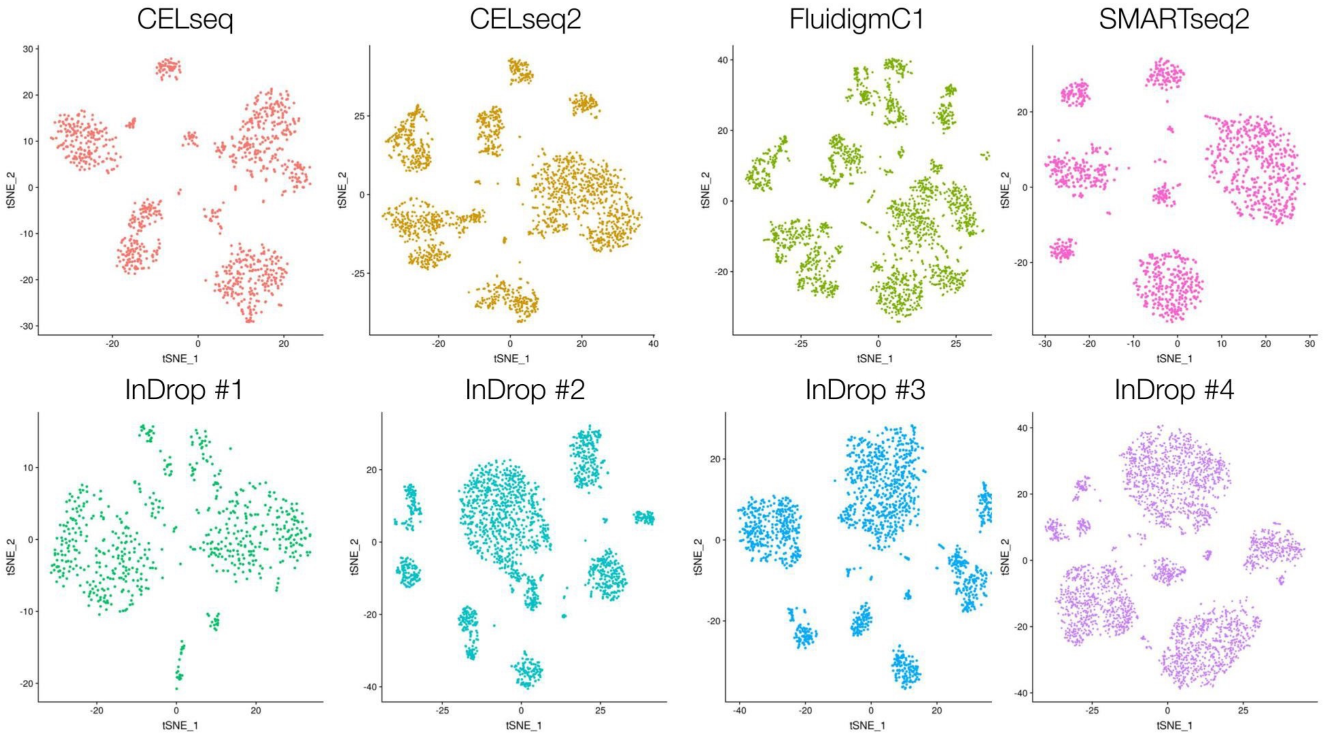


Cross condition comparisons

Courtesy form Ahmed Mahfouz

Integration analysis

- 8 maps from the human pancreas (Seurat tutorial)



Baron et al. 2016, *Cell Syst.*

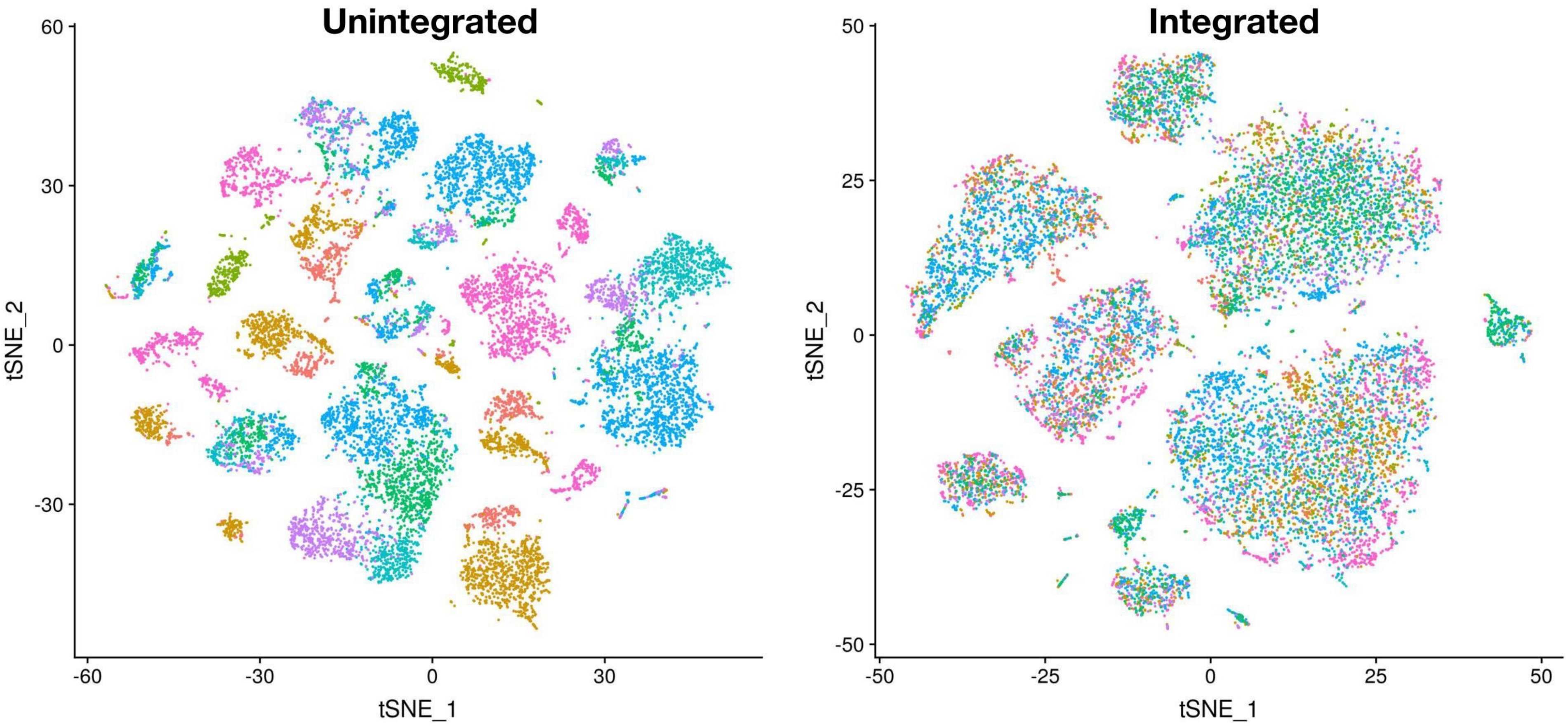
Lawlor et al. 2017, *Genome Res.*

Grun et al. 2016, *Cell Stem Cell*

Muraro et al. 2016, *Cell Syst.*

Integration analysis

- 8 maps from the human pancreas (Seurat tutorial)



Integration analysis: Confounders and batch effect

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

Technical 'batch effects' confound downstream analysis

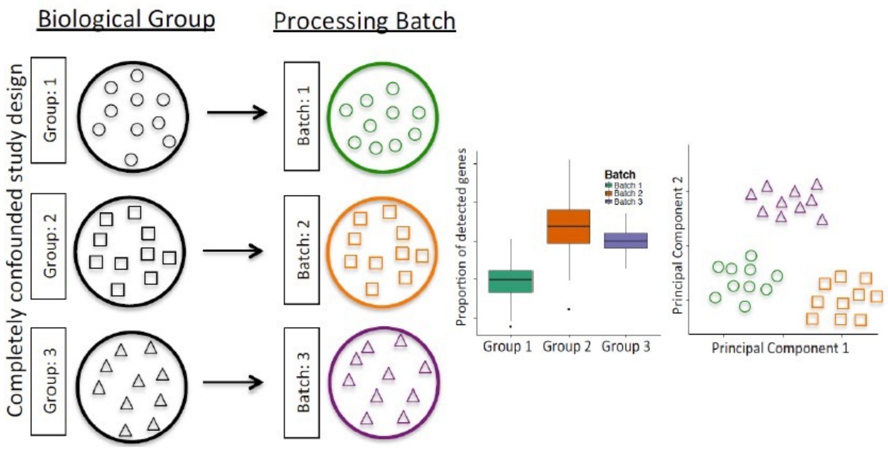
2. Biological variability
   - Patient differences
   - Evolution! (cross-species analysis)

Biological 'batch effects' confound comparisons of scRNA-seq data

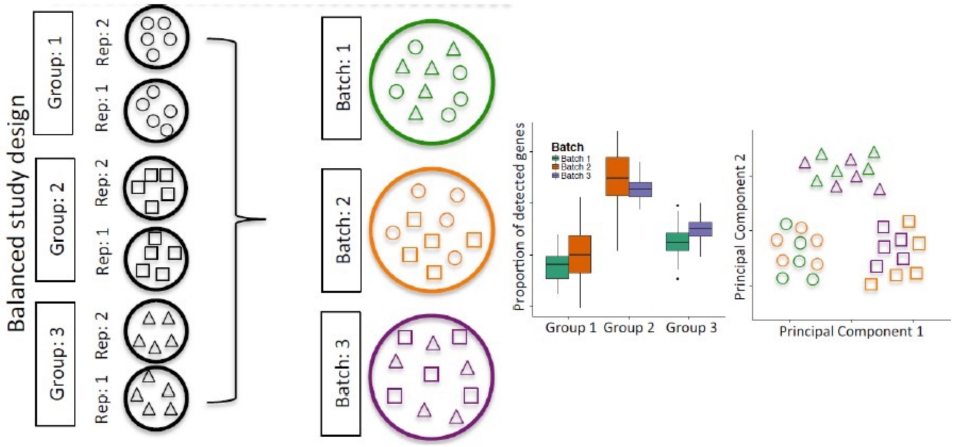
Shaham et al. (https://doi.org/10.1093/bioinformatics/btx196)

Integration analysis:
Confounders and batch effect

Confounded design



Not confounded design



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

Hicks et al. (https://doi.org/10.1093/biostatistics/kxx053)

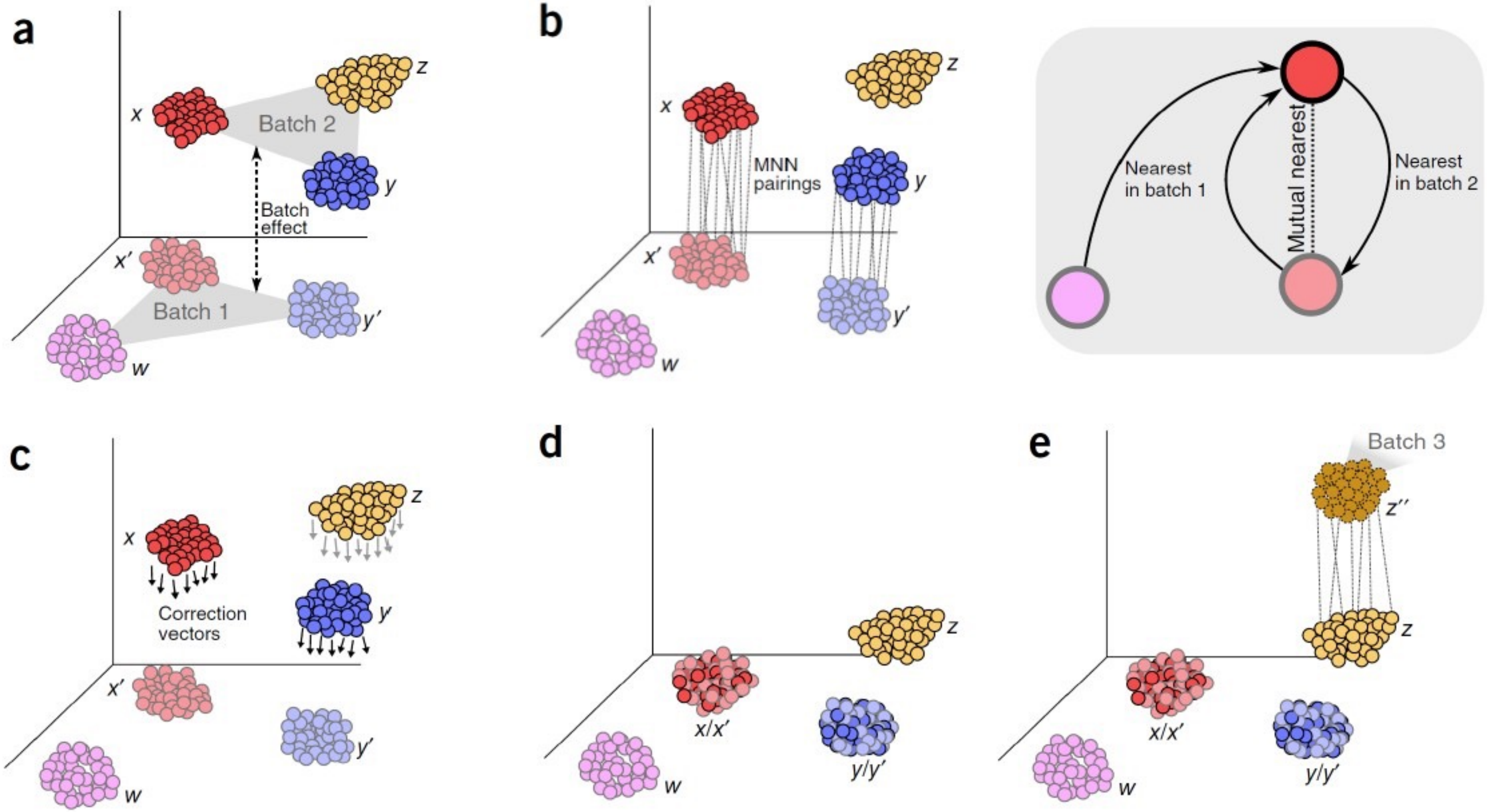
Integration analysis: Batch correction method

- 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)
- Conos (https://doi.org/10.1101/460246)
- Scanorama (https://doi.org/10.1101/371179)
- scMerge (https://doi.org/10.1073/pnas.1820006116)

Integration analysis: Batch correction method

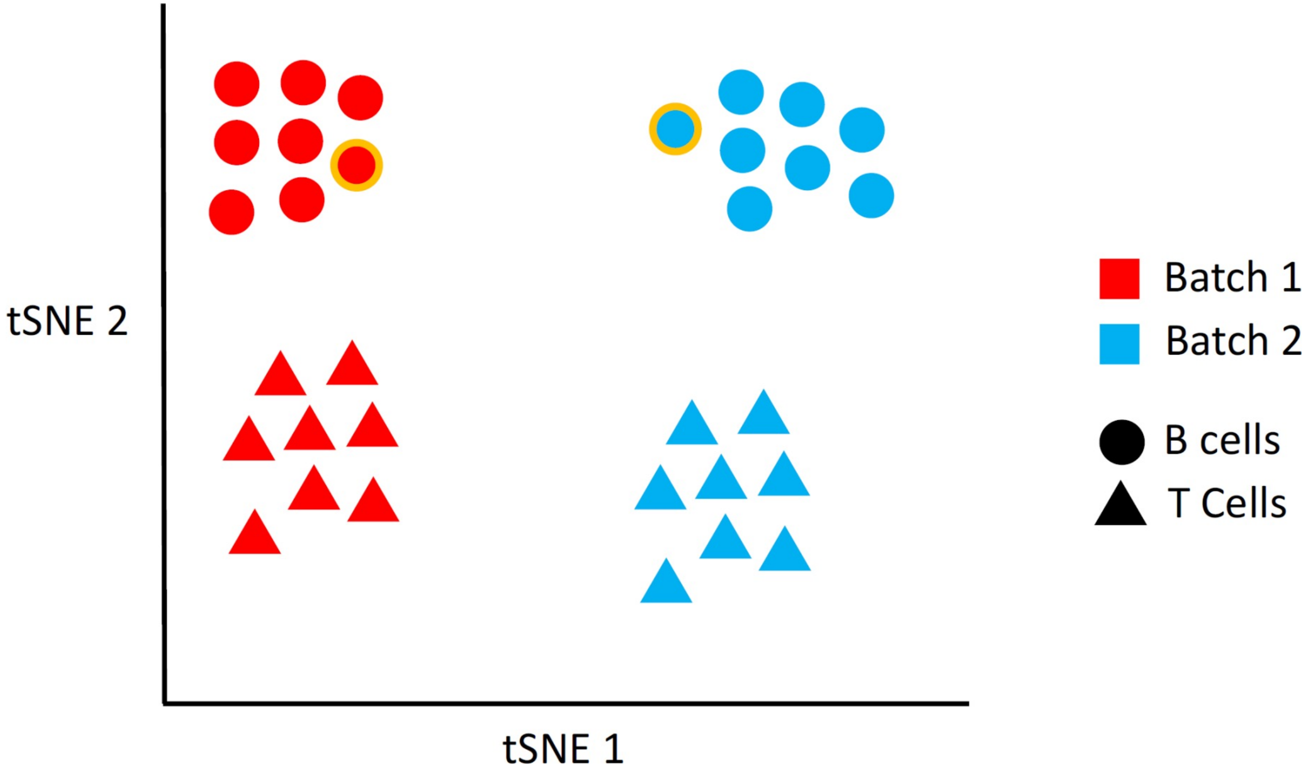
- MNNcorrect (https://doi.org/10.1038/nbt.4091)
- CCA +anchors (Seurat v3) (https://doi.org/10.1101/460147)
- CCA +dynamic time warping (Seurat v2) (https://doi.org/10.1038/nbt.4096)
- LIGER (https://doi.org/10.1101/459891)
- Harmony (https://doi.org/10.1101/461954)
- Conos (https://doi.org/10.1101/460246)
- Scanorama (https://doi.org/10.1101/371179)
- scMerge (https://doi.org/10.1073/pnas.1820006116)

Integration analysis:
Mutual Nearest Neighbors (MNN)

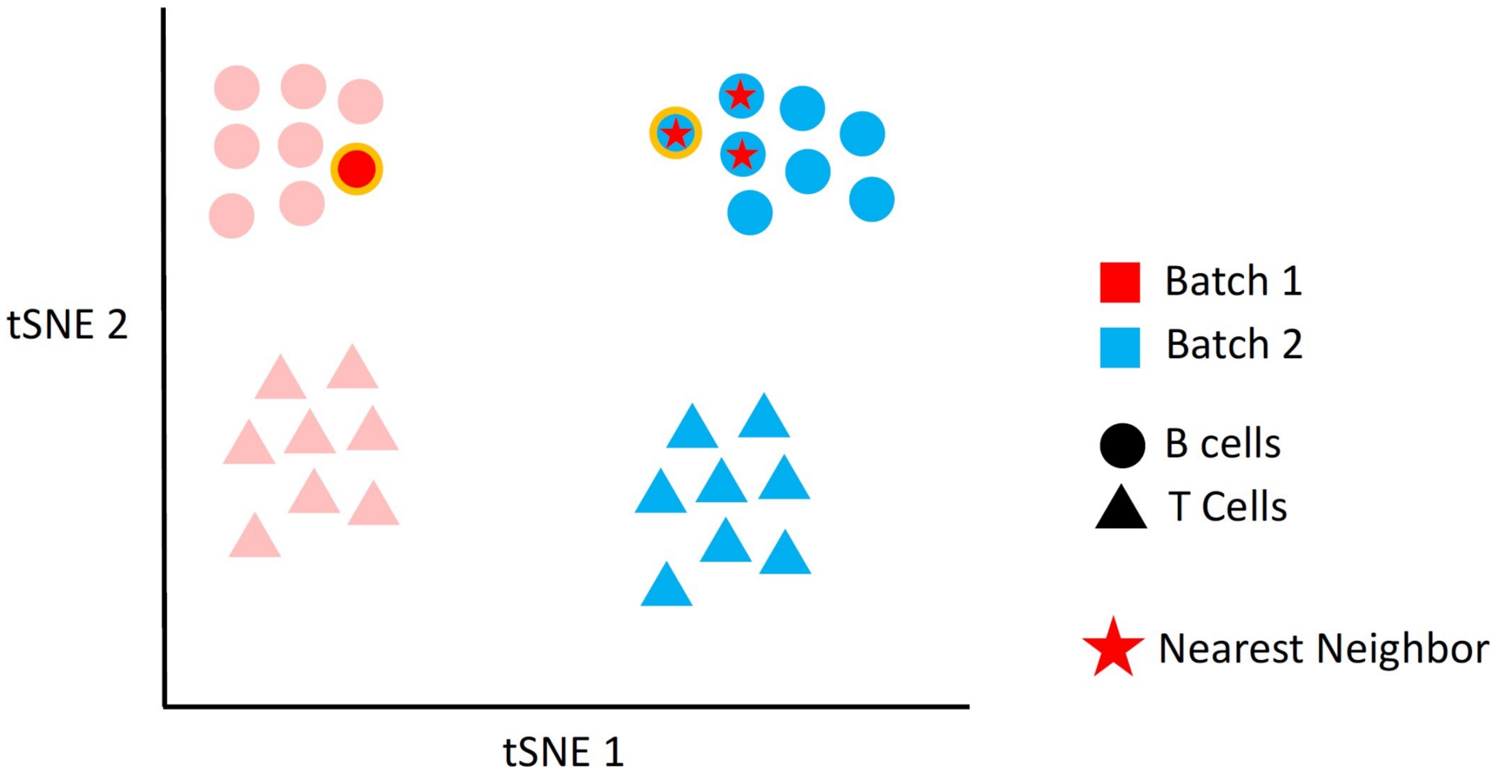


Haghverdi (https://doi.org/10.1038/nbt.4091)

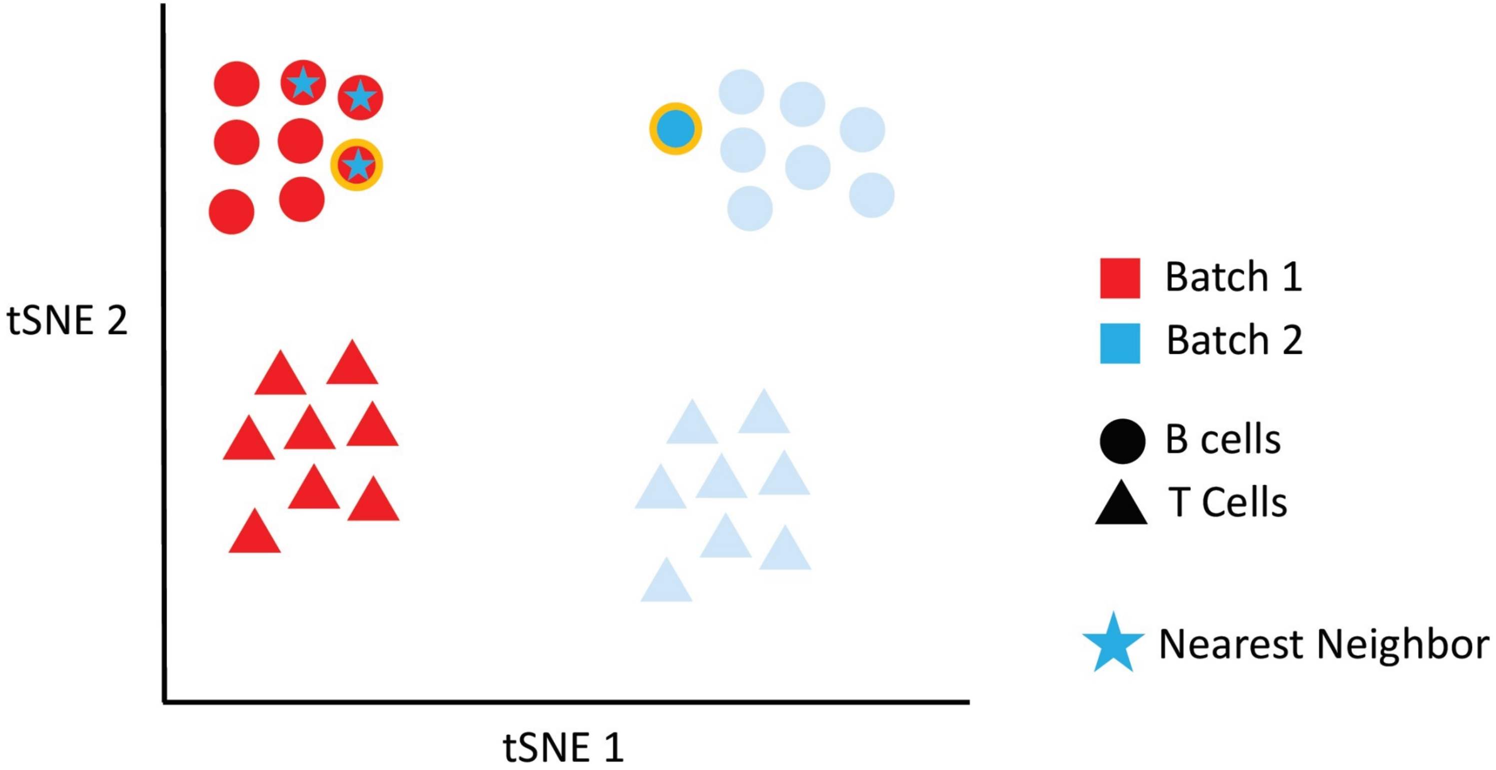
Integration analysis:
Mutual Nearest Neighbors (MNN)



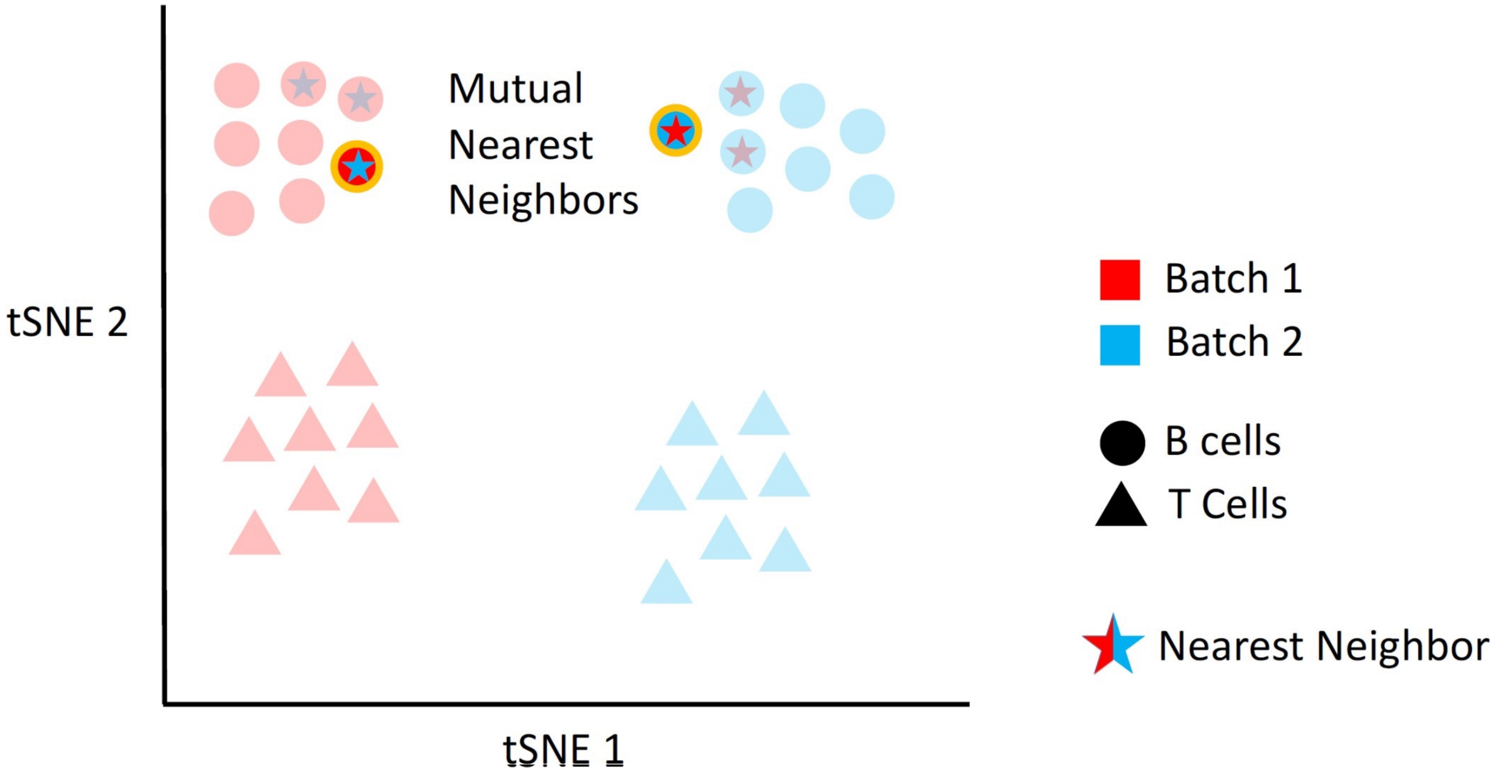
Integration analysis:
Mutual Nearest Neighbors (MNN)



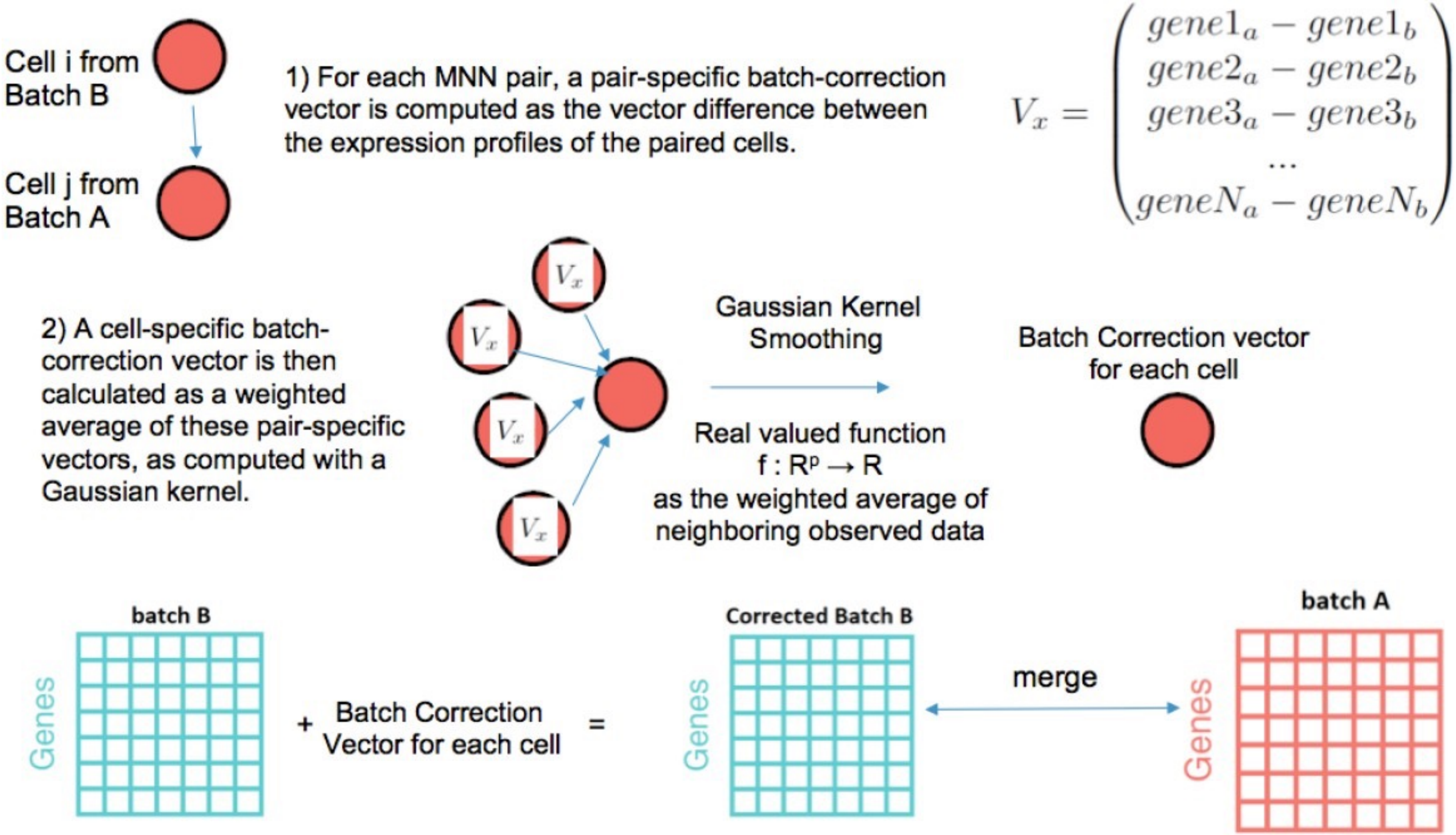
Integration analysis:
Mutual Nearest Neighbors (MNN)



Integration analysis:
Mutual Nearest Neighbors (MNN)



Integration analysis:
Mutual Nearest Neighbors (MNN)



Cell i from Batch B

Cell j from Batch A

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.

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

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.

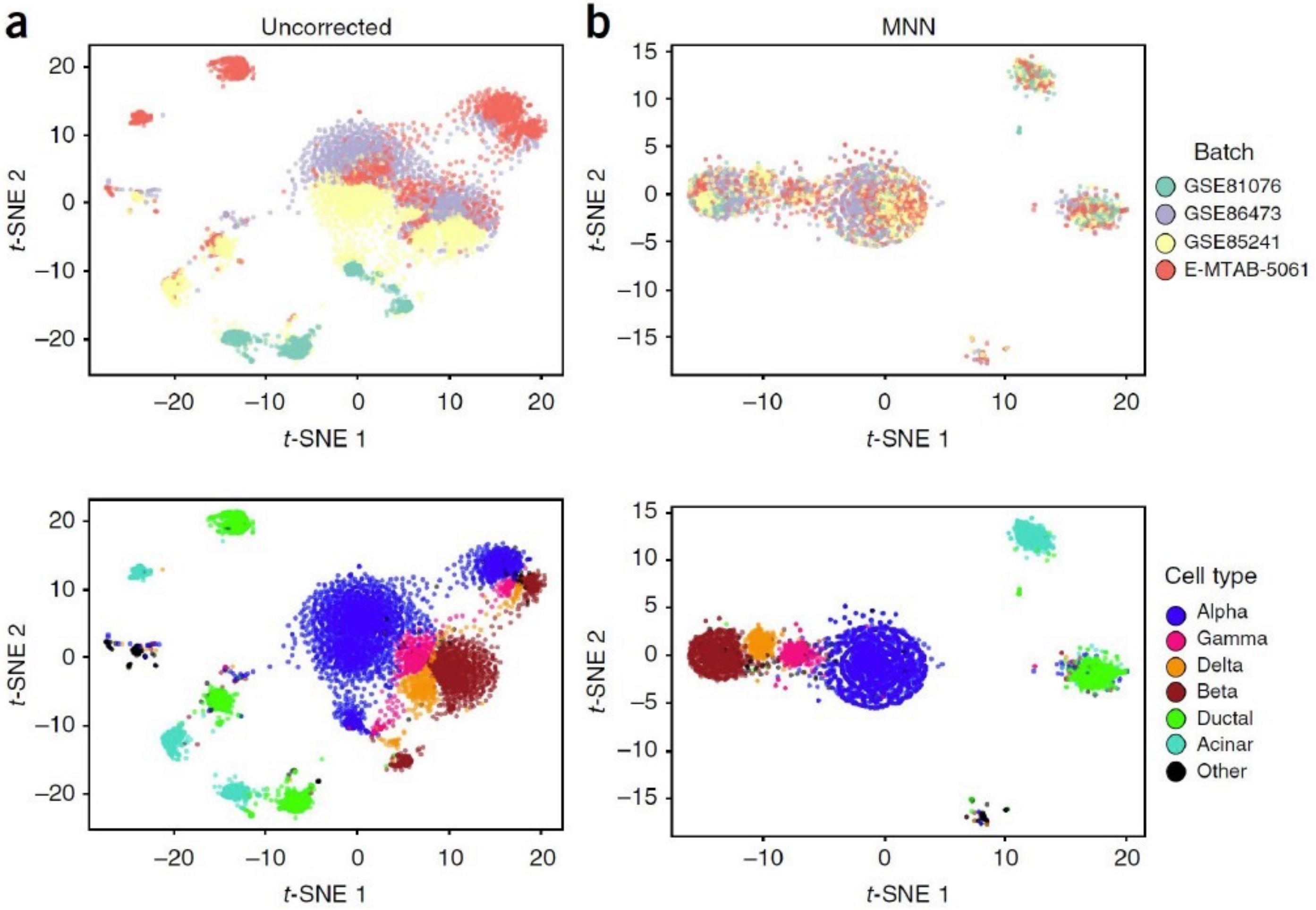
Gaussian Kernel Smoothing

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

Batch Correction vector for each cell

\+ Batch Correction Vector for each cell =

Integration analysis:
Mutual Nearest Neighbors (MNN)



Haghverdi (https://doi.org/10.1038/nbt.4091)

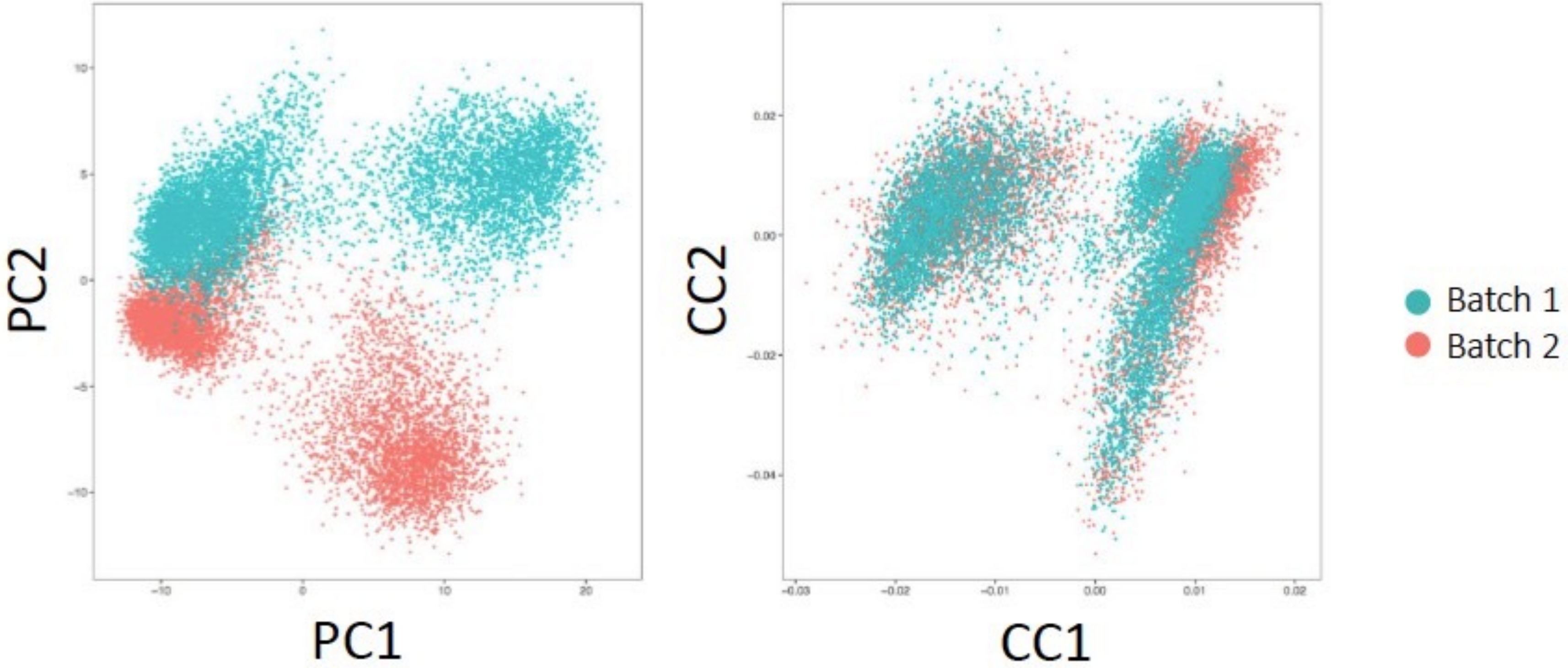
Integration analysis: CCA +anchors (Seurat v3)

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

Stuart et al. (https://doi.org/10.1101/460147)

Integration analysis: CCA +anchors (Seurat v3)

1. Find corresponding cells across datasets

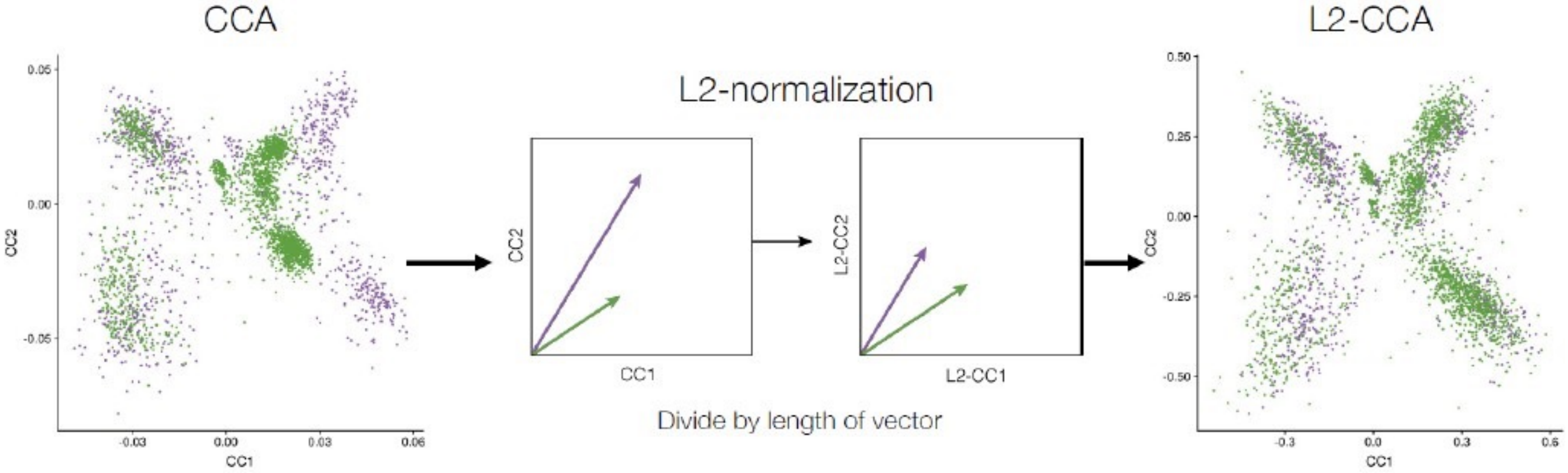


CCA captures correlated sources of variation between two datasets

Stuart et al. (https://doi.org/10.1101/460147)

Integration analysis: CCA +anchors (Seurat v3)

1. Find corresponding cells across datasets

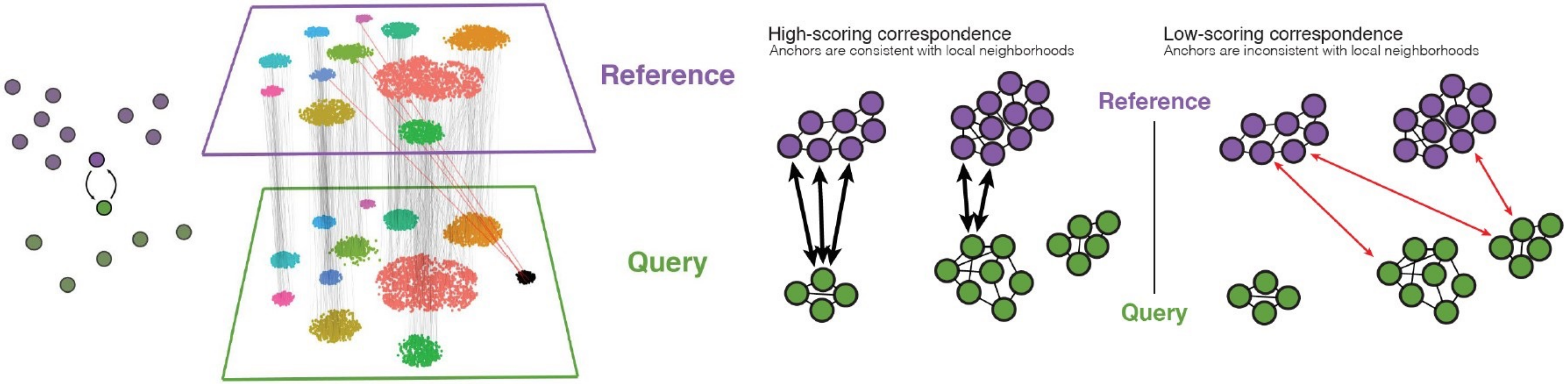


L2-normalization corrects for differences in scale

Integration analysis: CCA +anchors (Seurat v3)

1. Find corresponding cells across datasets

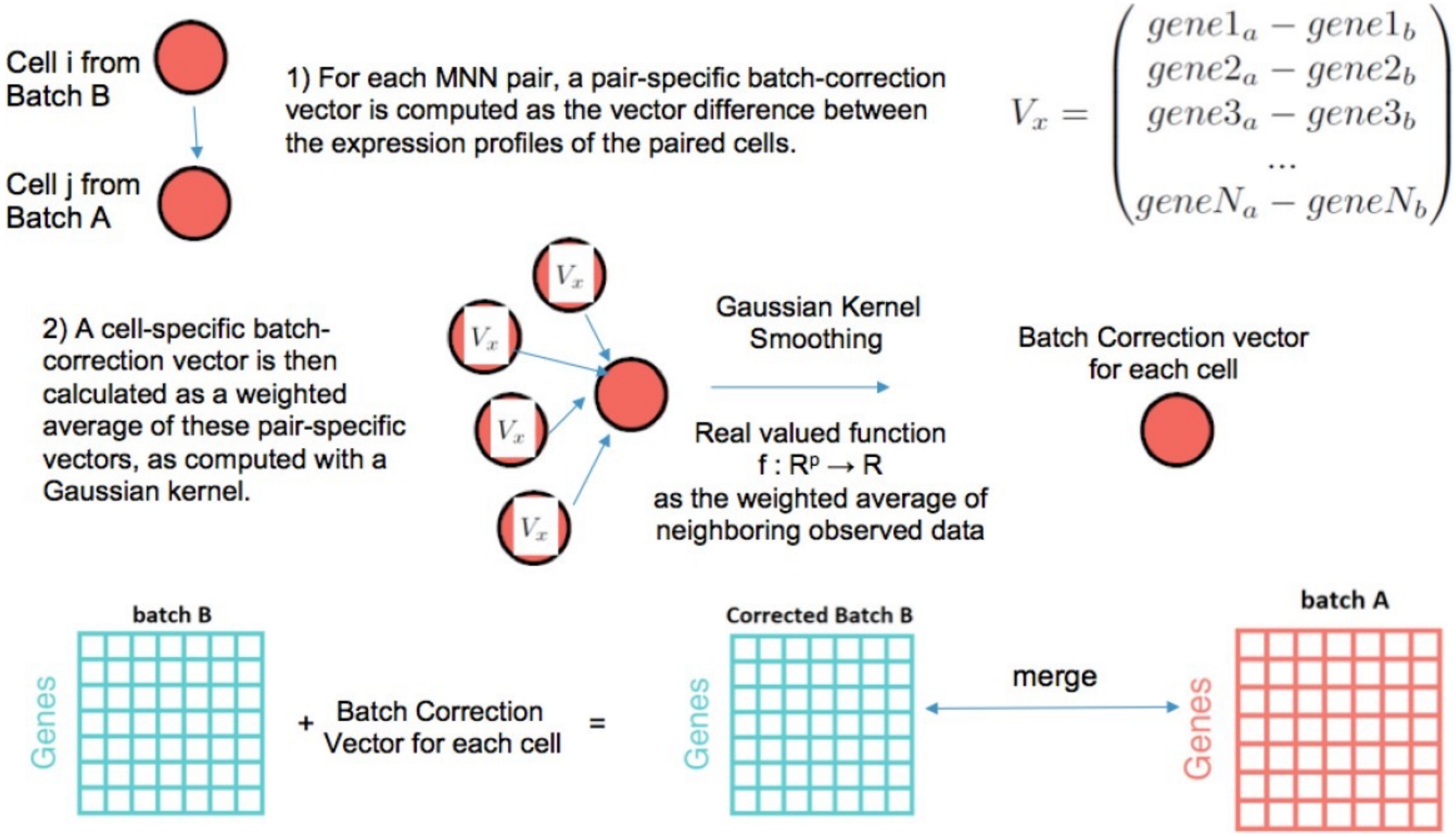
Anchors: Mutual nearest neighbors



`Seurat(v3)::FindIntegrationAnchors()`

Integration analysis: CCA +anchors (Seurat v3)

2. Data integration



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.

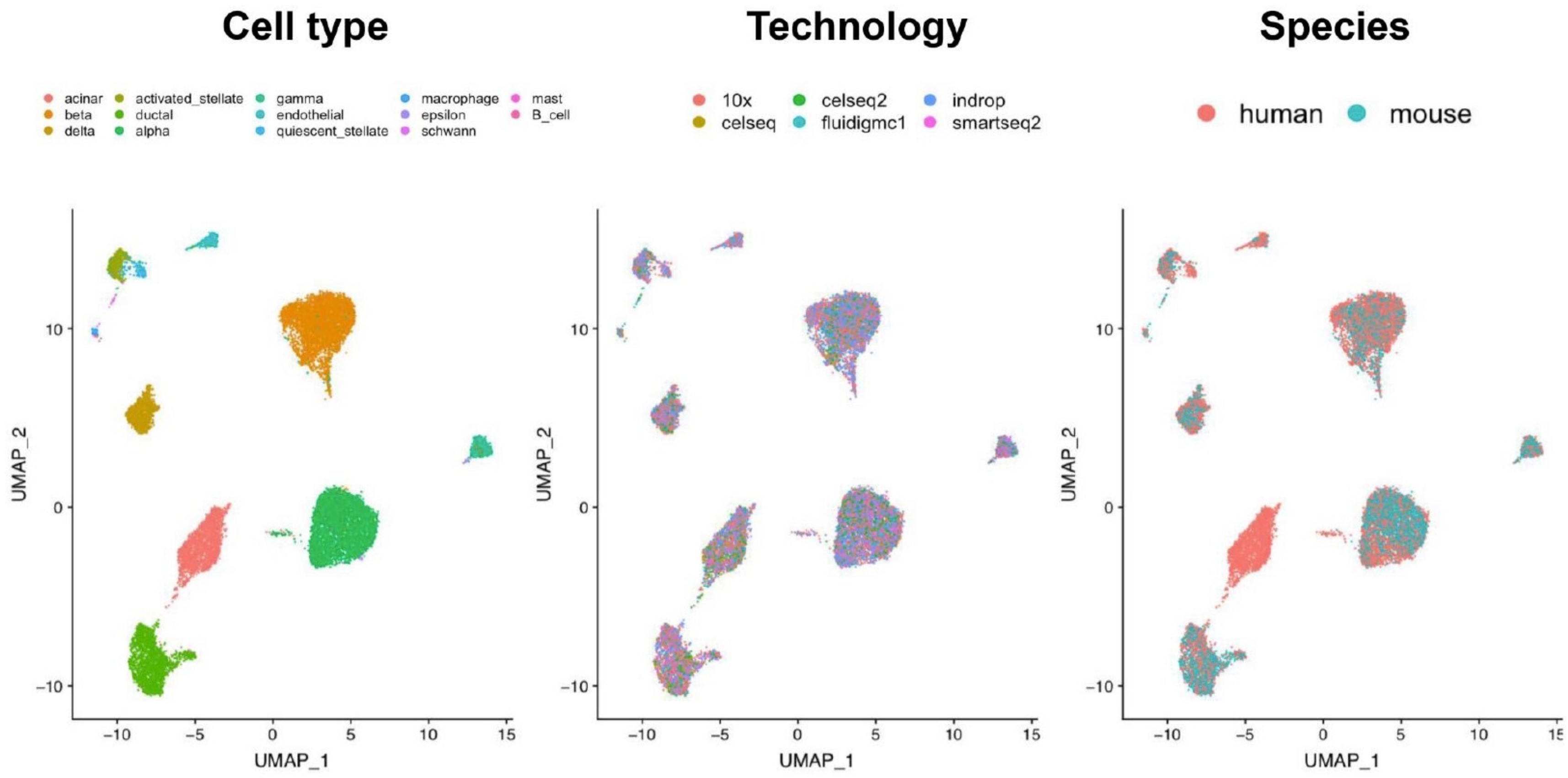
$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \cdots \\ geneN_a - geneN_b \end{pmatrix}$$

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

`Seurat(v3)::IntegrateData()`

Integration analysis:
CCA +anchors (Seurat v3)



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