Single Cell RNA-seq Data Integration

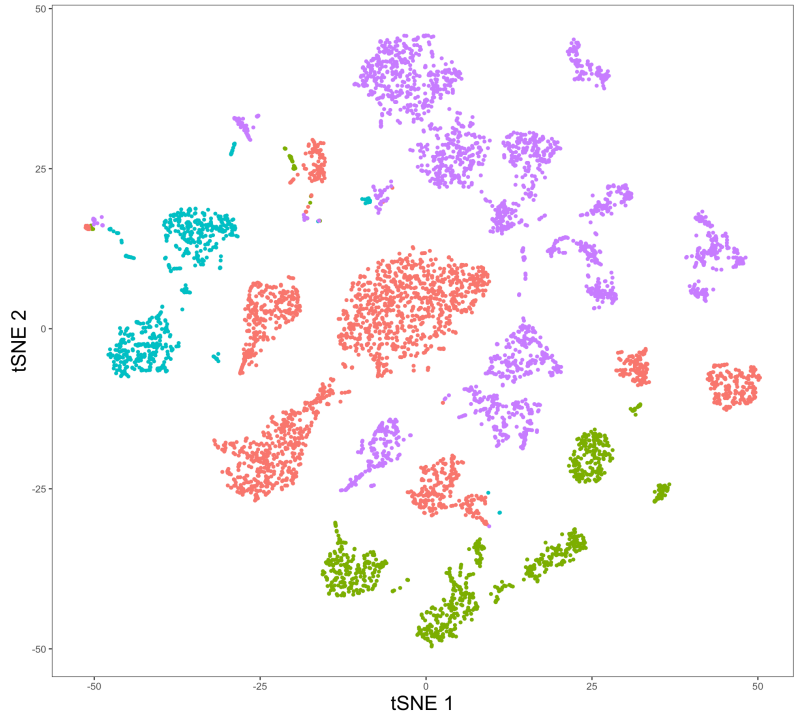
Tamim Abdelaal
Leiden Computational Biology Center, LUMC
Delft Bioinformtaics Lab, TU Delft

Leiden
Computational Biology Center

LUMC

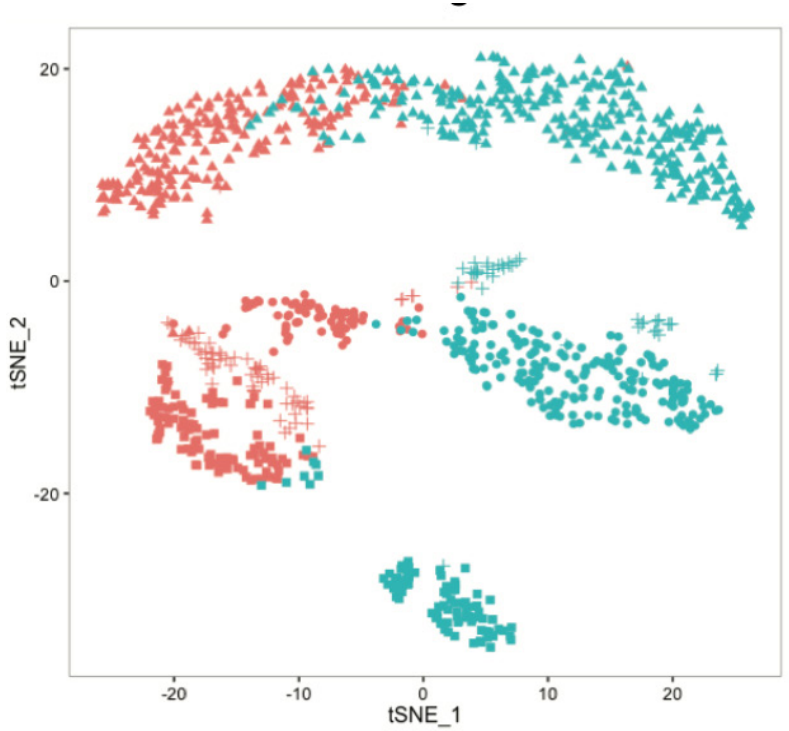
TU Delft

Why integrate?



Donor 1
Donor 2
Donor 3
Donor 4

Same tissue from different donors



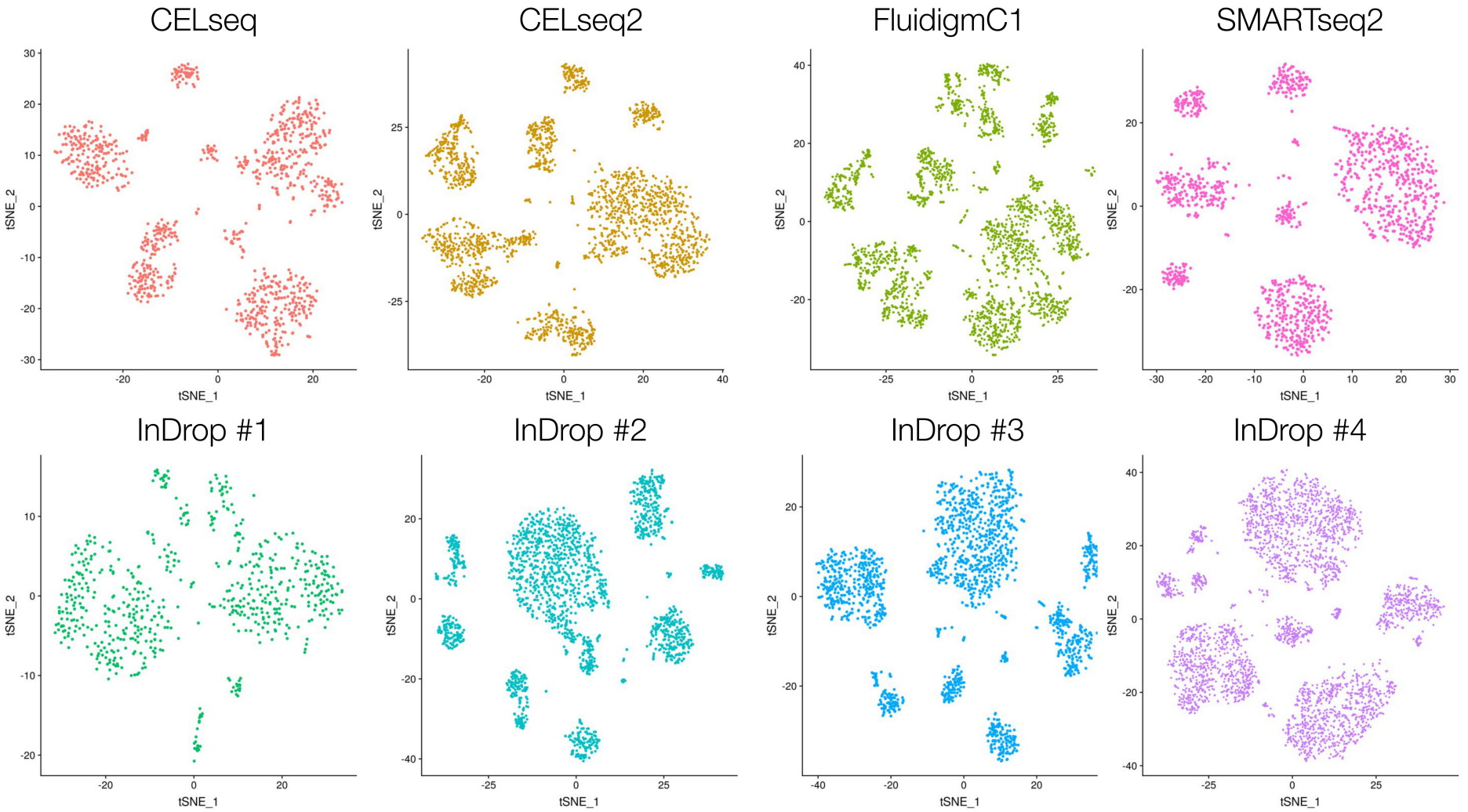
Sick
Healthy

▲ Cell Type 1
■ Cell Type 2
● Cell Type 3

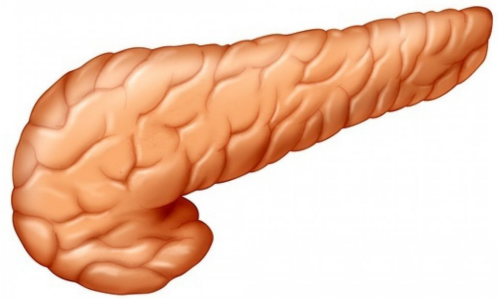
Cross condition comparisons

Building a cell atlas

8 maps of the human pancreas



Human pancreas



Baron et al. 2016, *Cell Syst.*

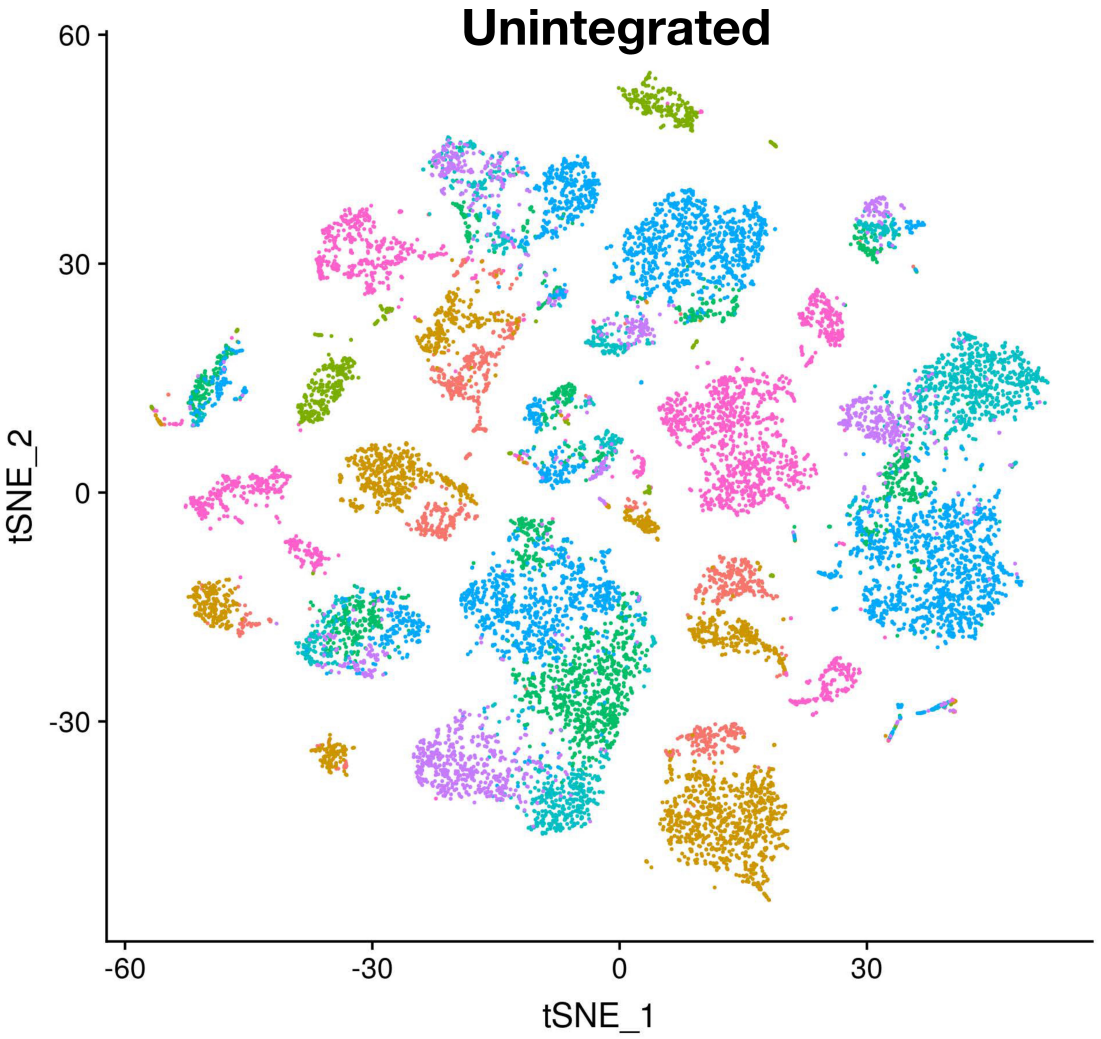
Lawlor et al. 2017, *Genome Res.*

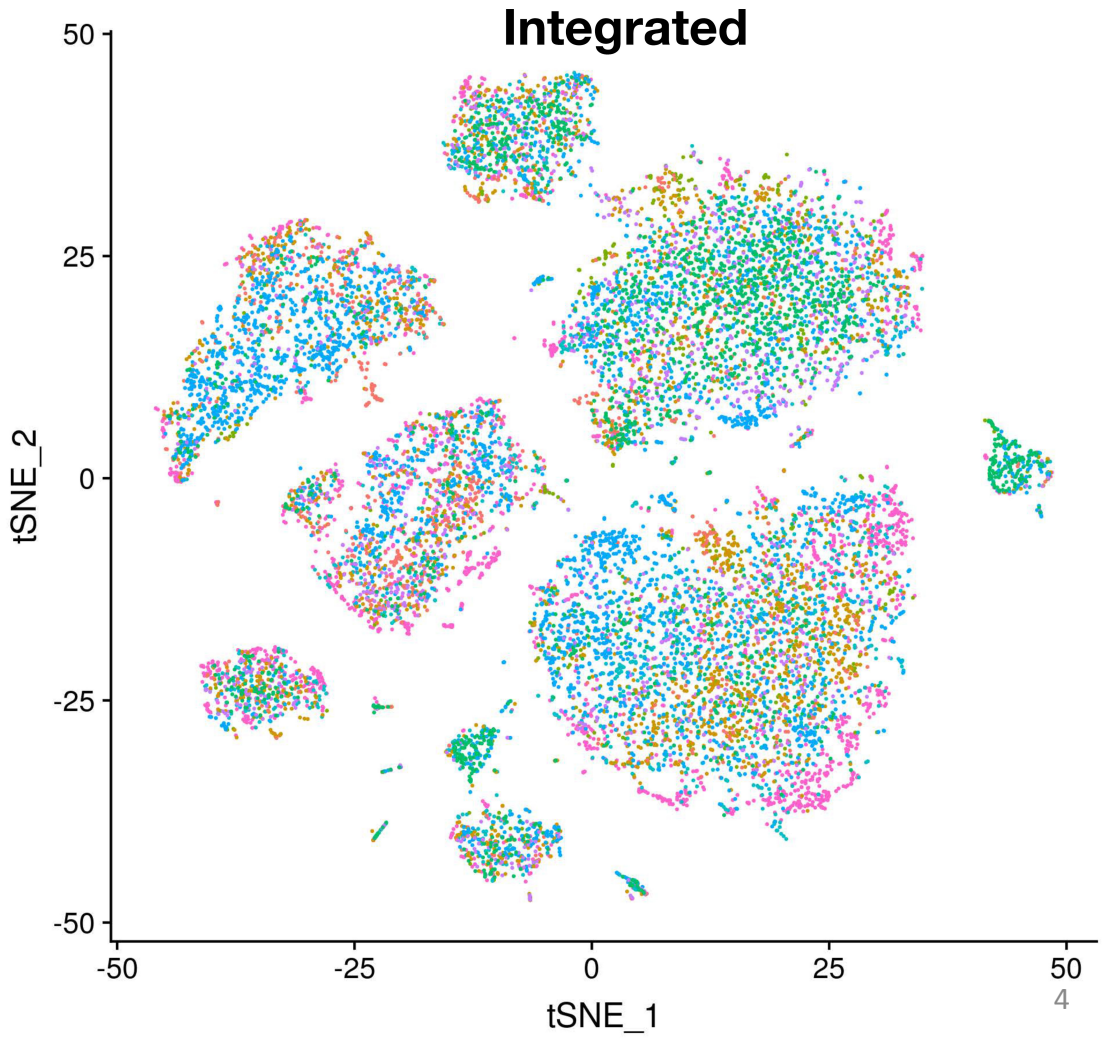
Grun et al. 2016, *Cell Stem Cell*

Muraro et al. 2016, *Cell Syst.*

Building a cell atlas

8 maps of the human pancreas

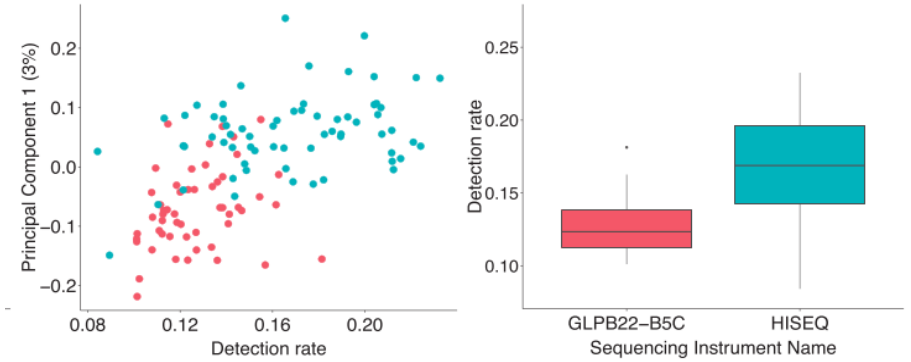




Confounders and batch effects

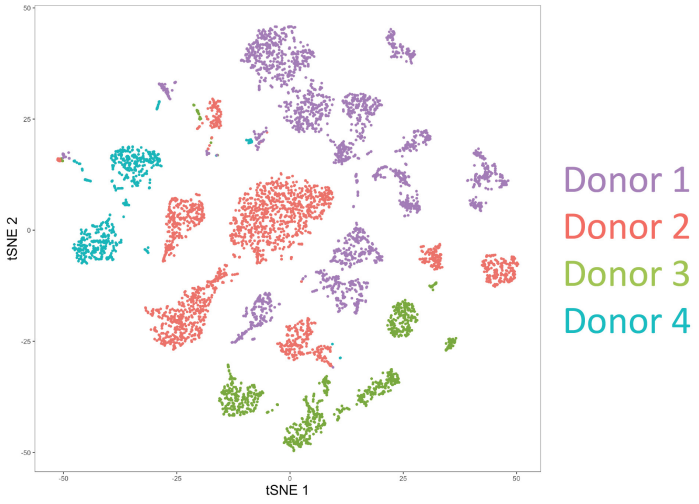
1. Technical variability
   - Changes in sample quality/processing
   - Library prep or sequencing technology
   - ‘Experimental reality’

Technical ‘batch effects’ confound downstream analysis



2. Biological variability
   - Patient differences
   - Environmental/genetic perturbation
   - Evolution! (cross-species analysis)

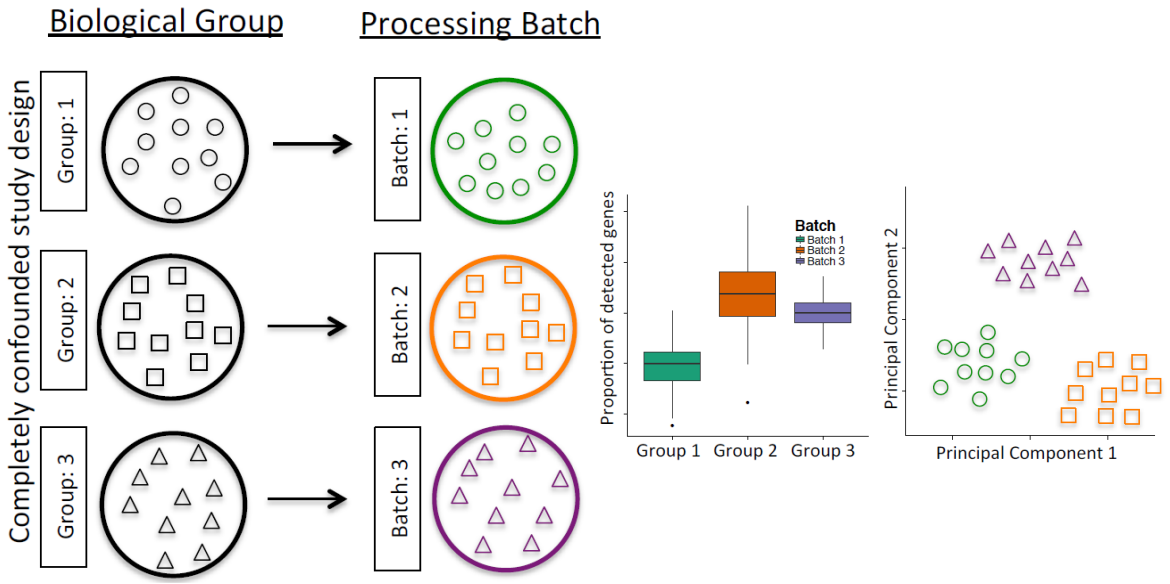
Biological ‘batch effects’ confound comparisons of scRNA-seq data



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

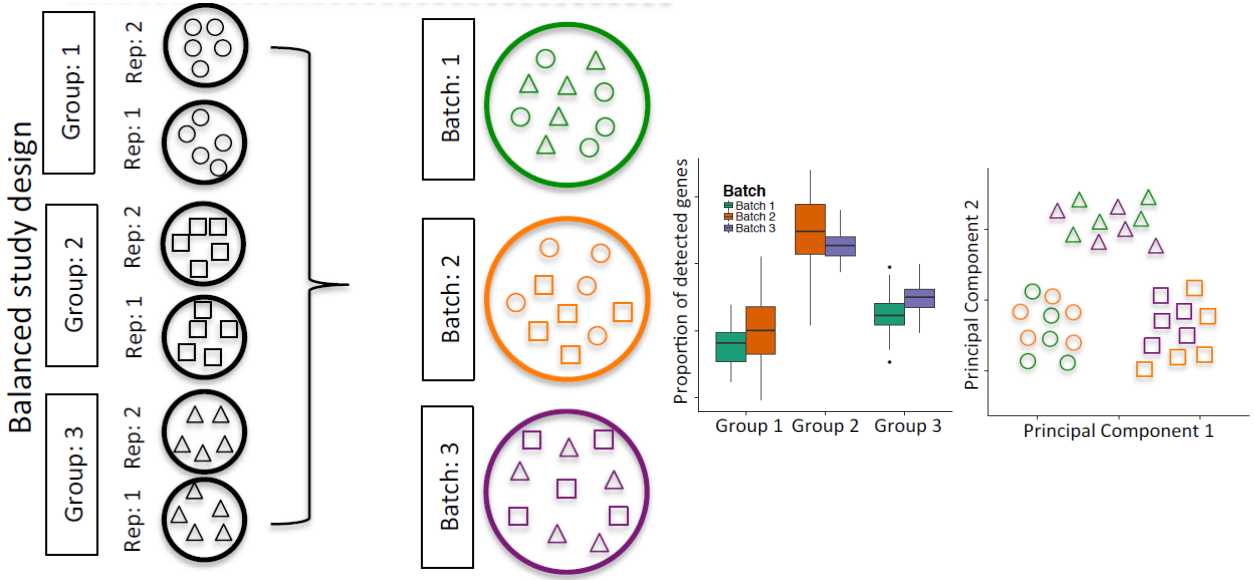
Confounders and batch effects

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.

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

Our agenda

- Single cell batch correction methods
- Performance assessment
- SpaGE – integration of scRNA-seq and spatial transcriptomics

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

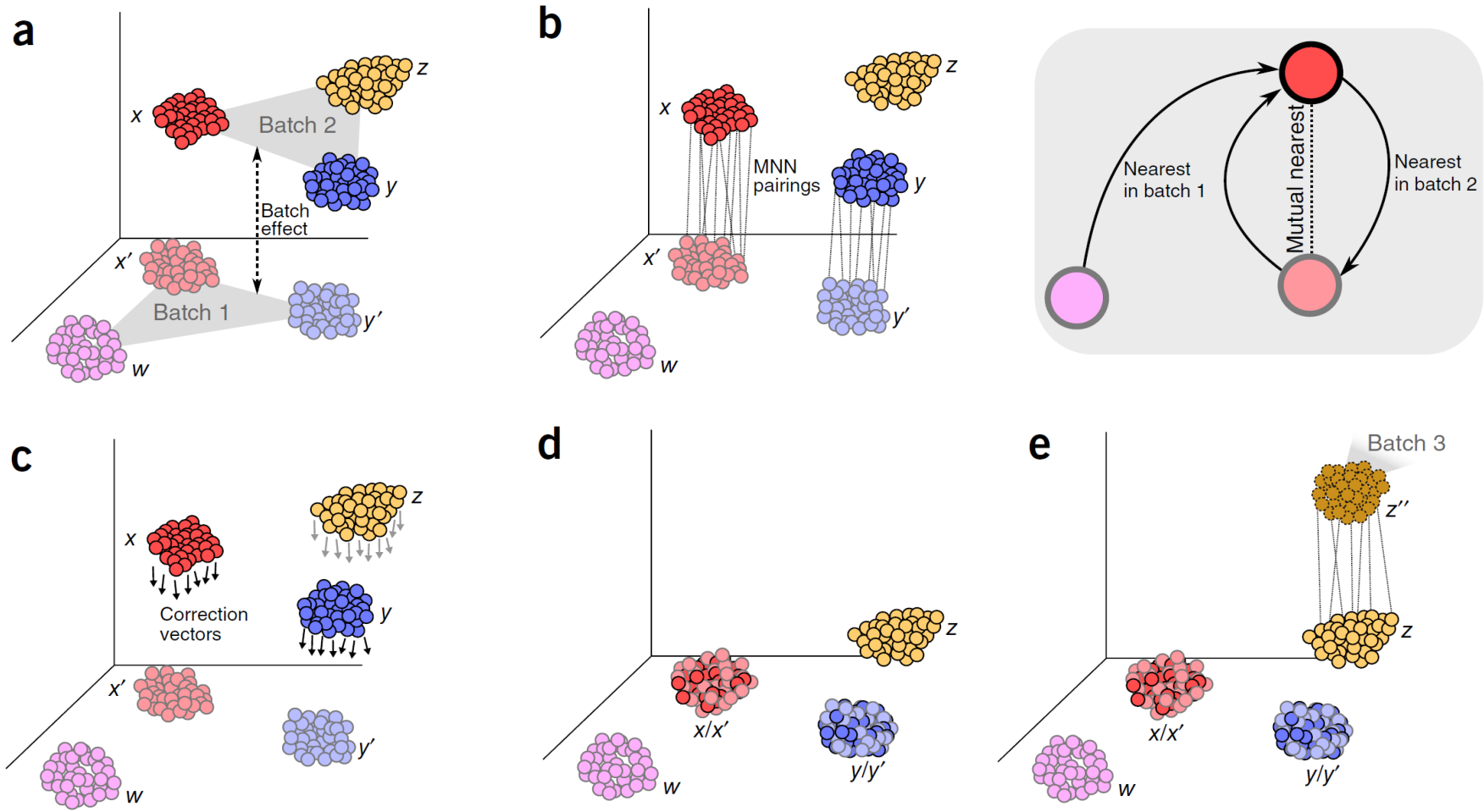
Batch correction methods

- **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:

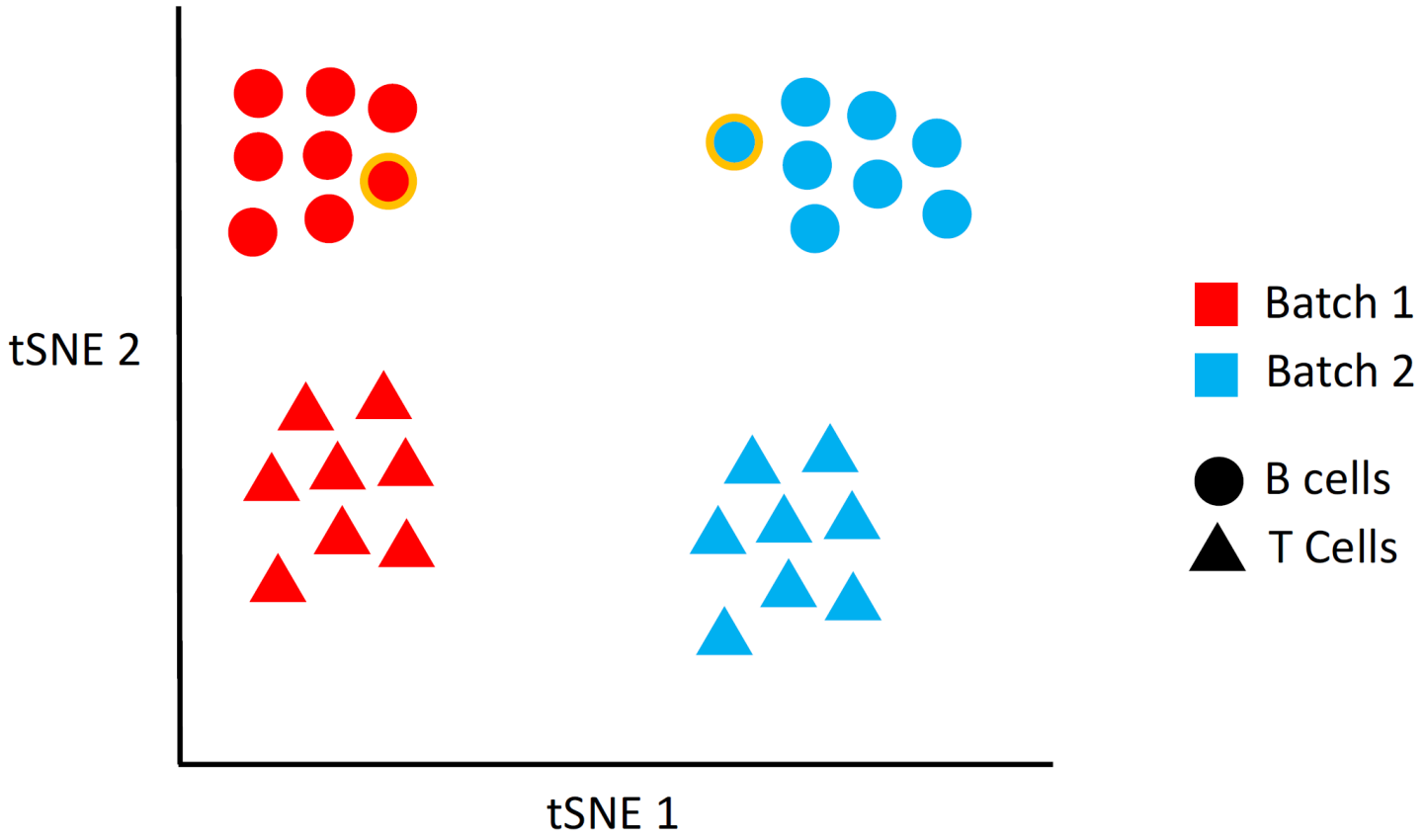
- Joint dimension reduction
- Graph-based approaches

Mutual Nearest Neighbors (MNN)

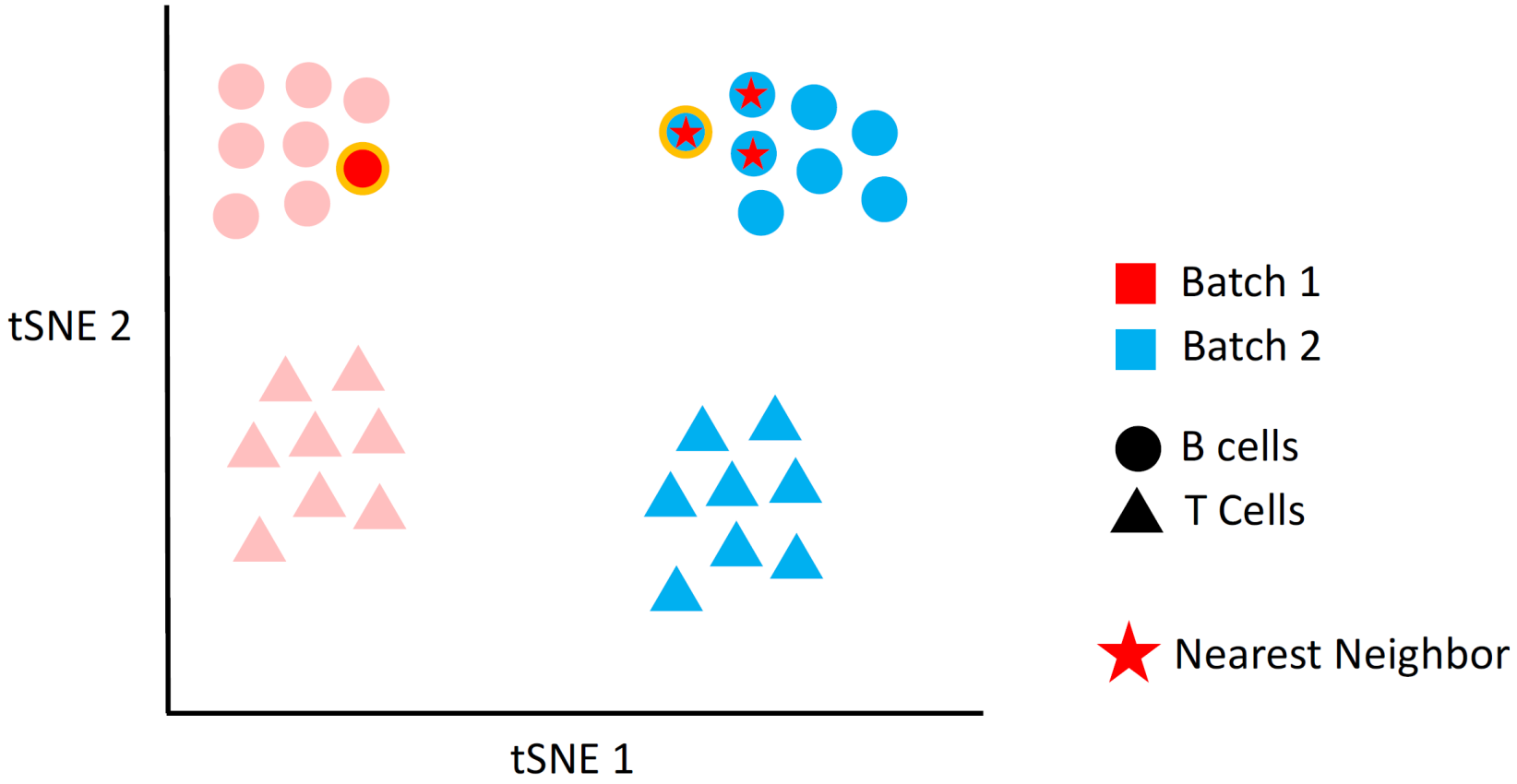


Haghverdi et al. (https://doi.org/10.1038/nbt.4091)

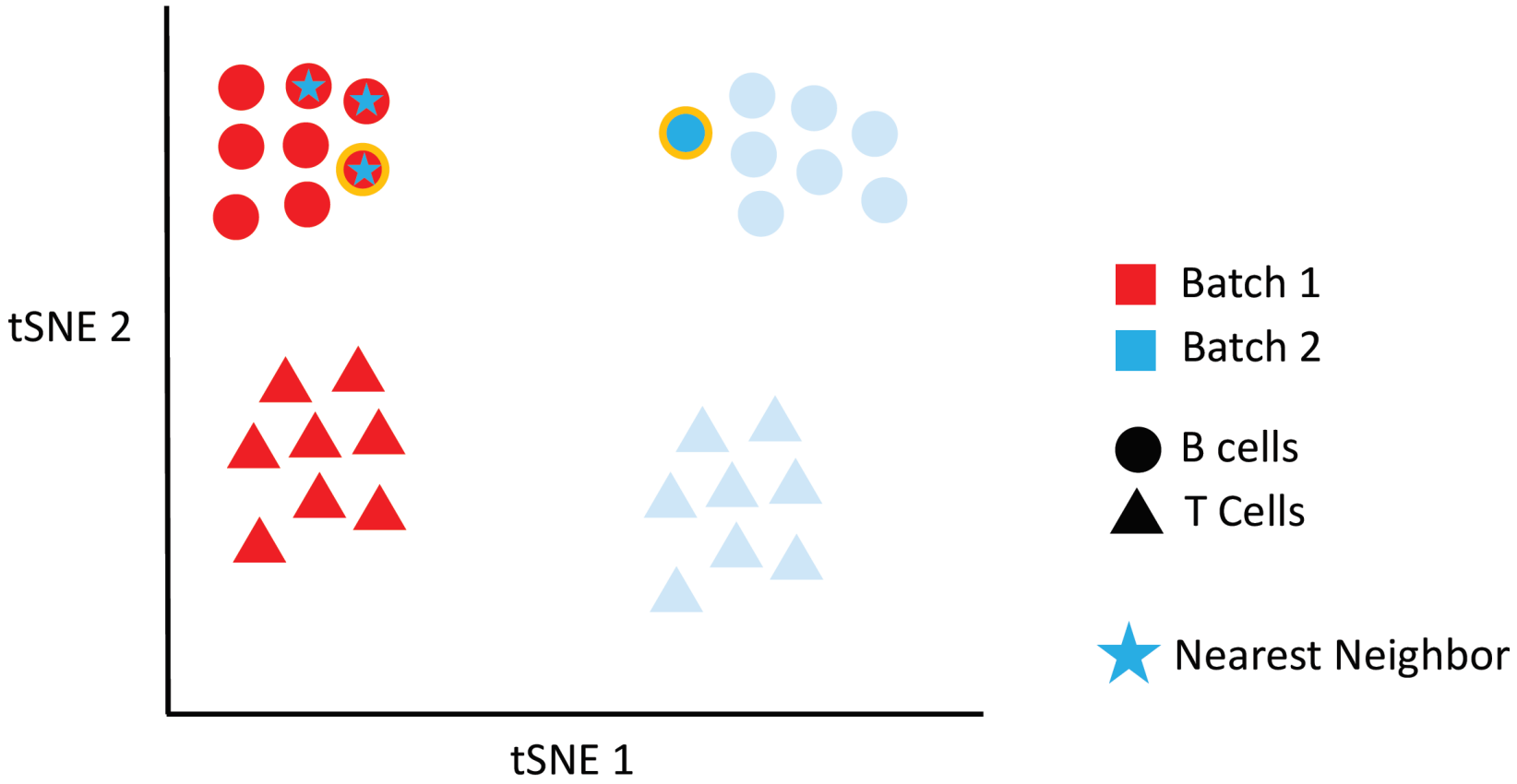
Mutual Nearest Neighbors (MNN)



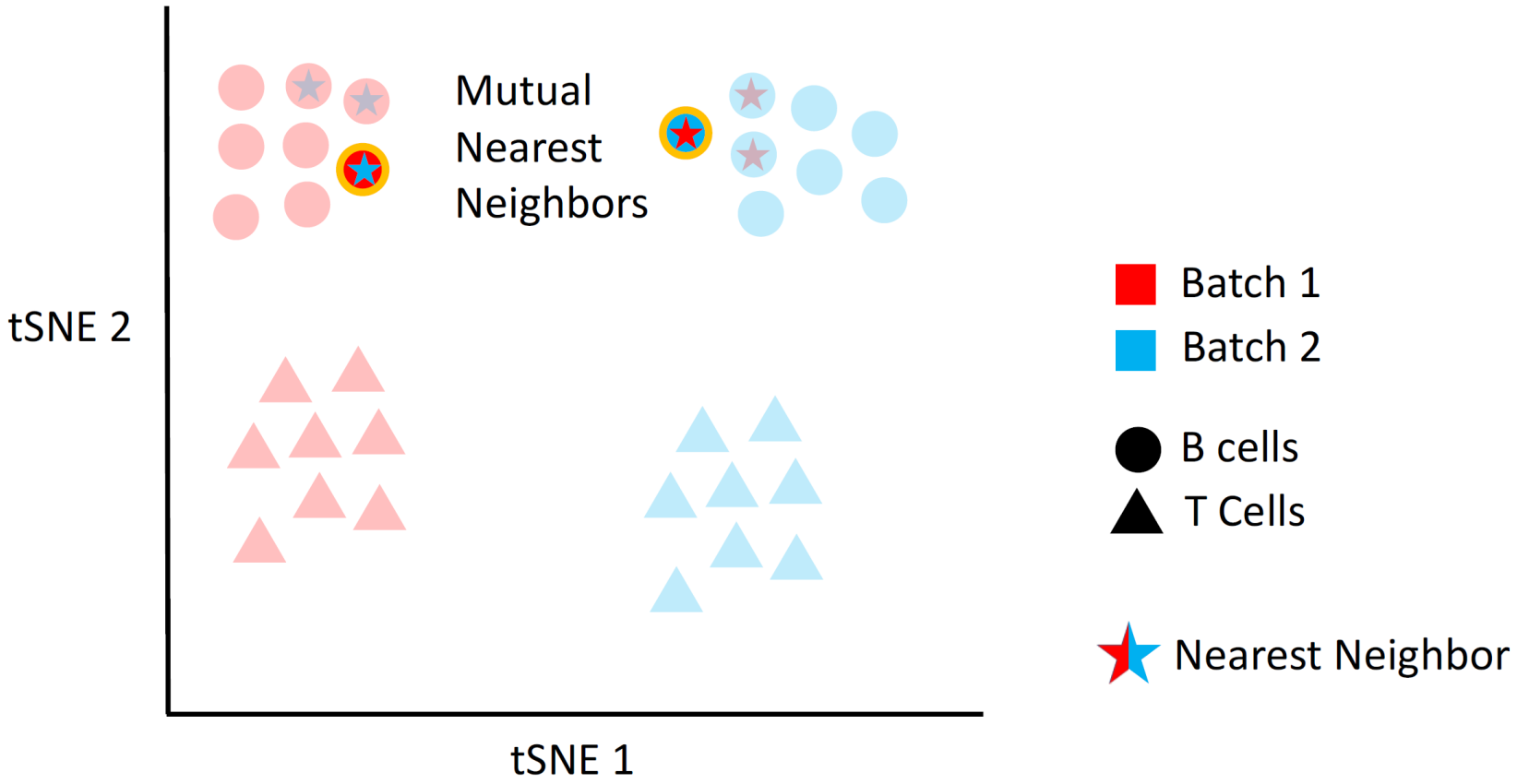
Mutual Nearest Neighbors (MNN)



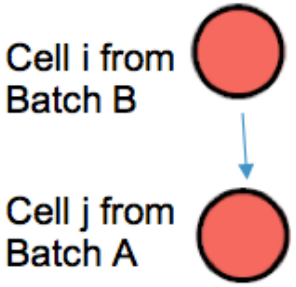
Mutual Nearest Neighbors (MNN)



Mutual Nearest Neighbors (MNN)



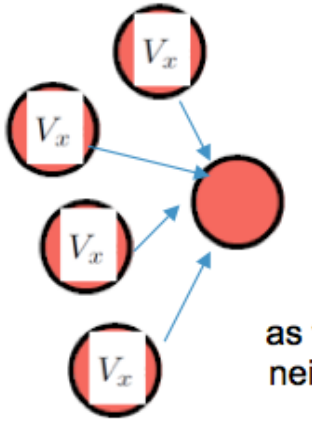
Mutual Nearest Neighbors (MNN)



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.

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

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.

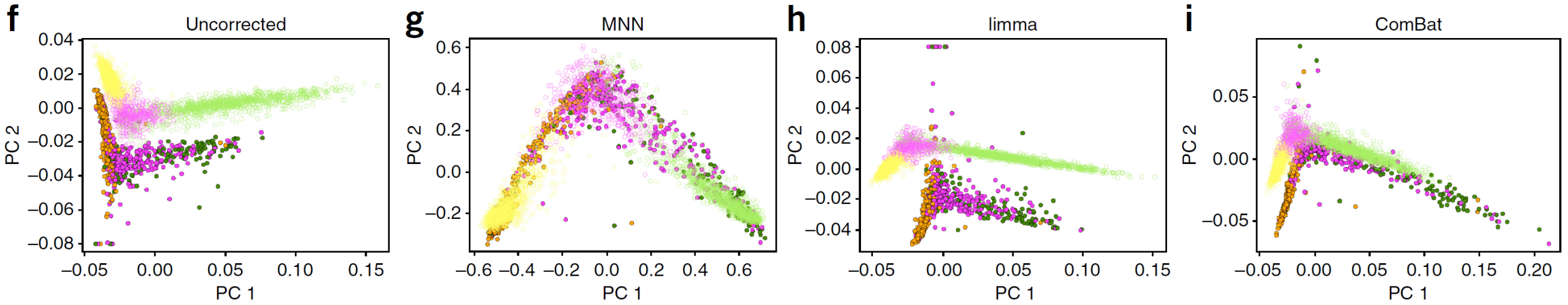


Gaussian Kernel Smoothing

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

Batch Correction vector for each cell

Mutual Nearest Neighbors (MNN)



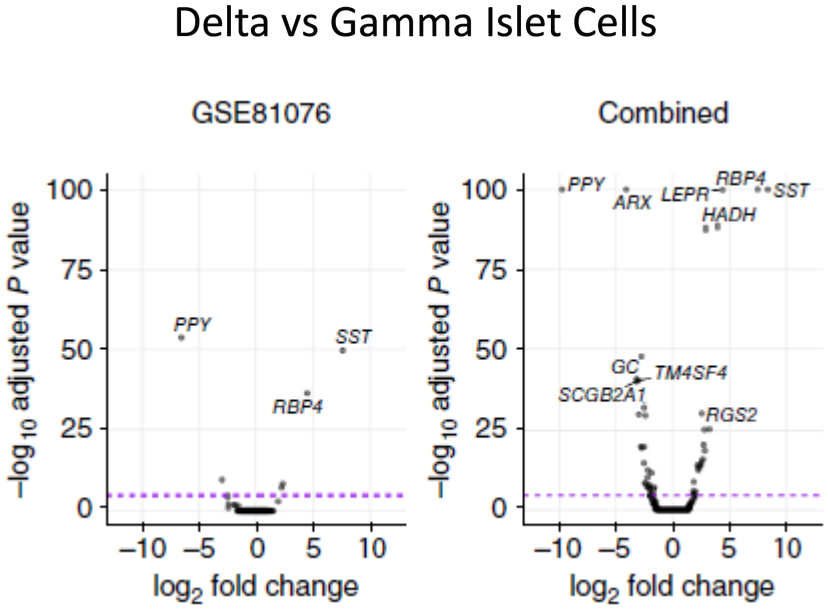
SMART-seq2
- MEP
- GMP
- CMP

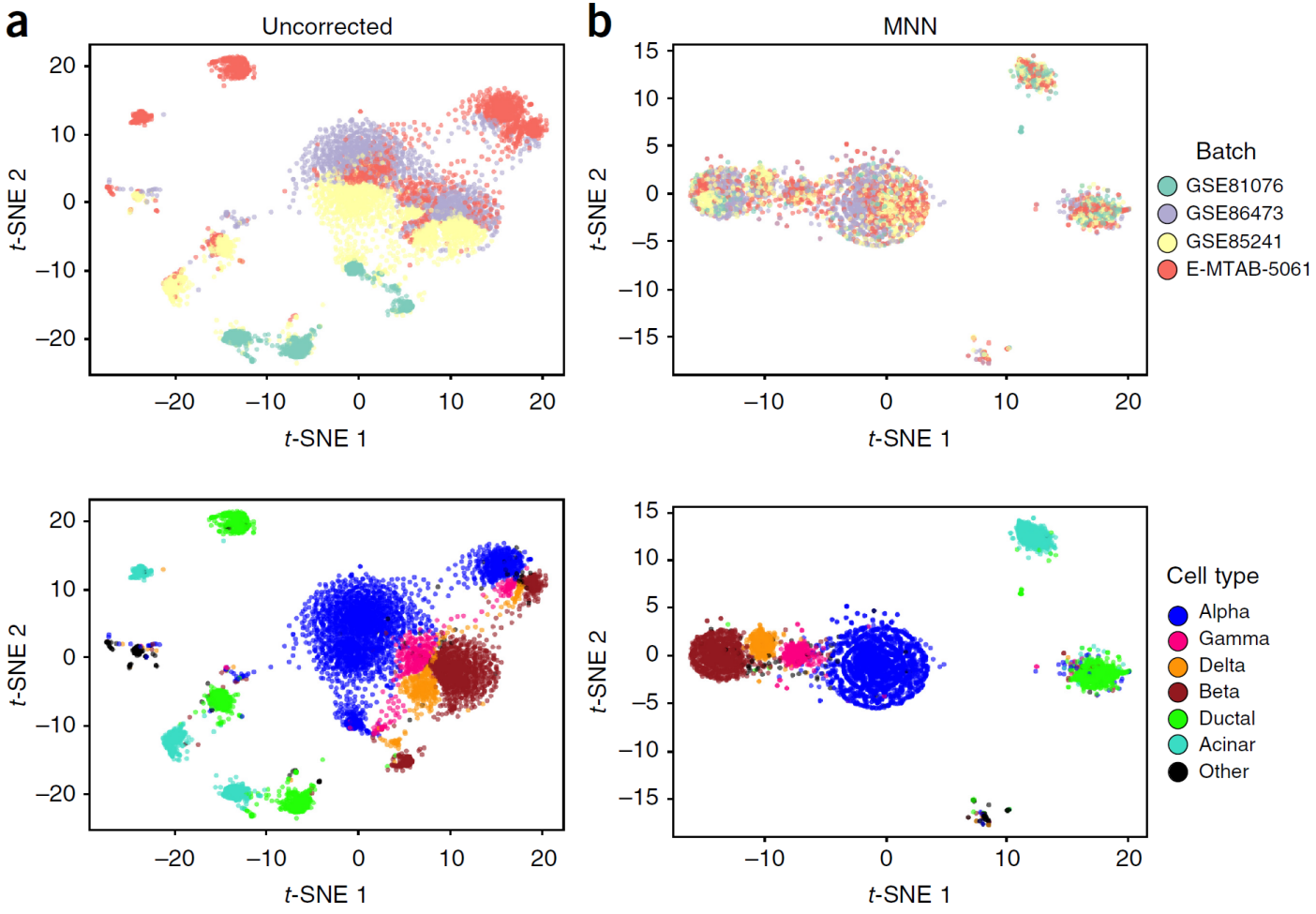
MARS-seq
- MEP
- GMP
- CMP

MEPs: megakaryocyte–erythrocyte progenitors
GMPs: granulocyte–monocyte progenitors
CMPs: common myeloid progenitors

Mutual Nearest Neighbors (MNN)

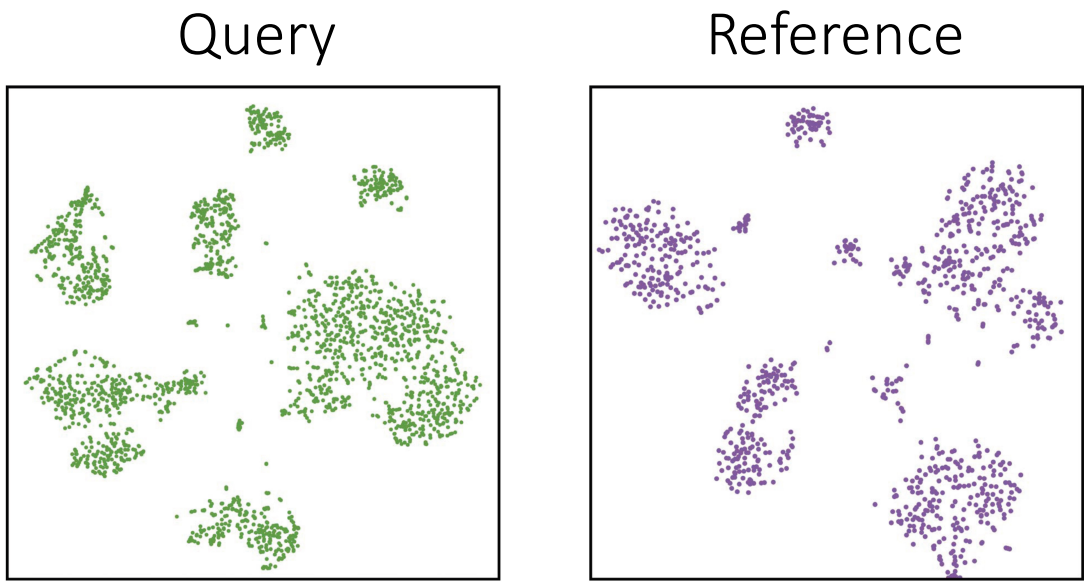
- Pooling experiments -> increased statistical power





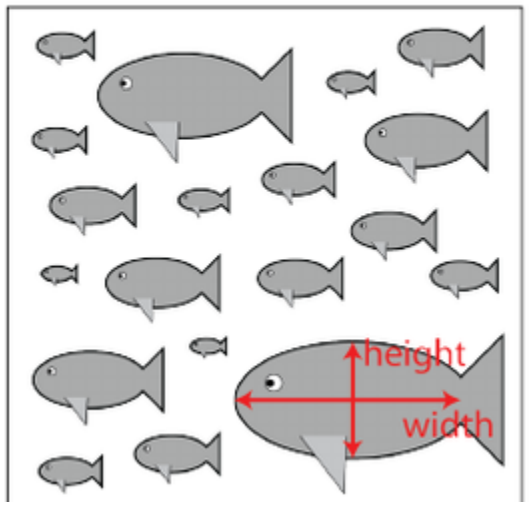
CCA + anchors (Seurat v3)

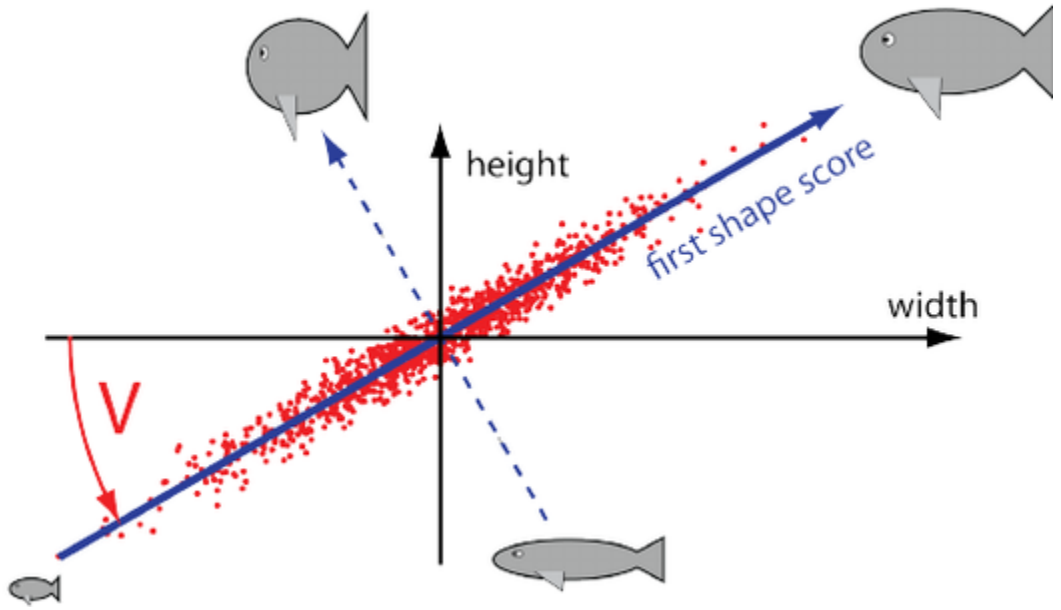
1. Find corresponding cells across datasets
2. Compute a data adjustment based on correspondences between cells
3. Apply the adjustment



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

Principle component analysis

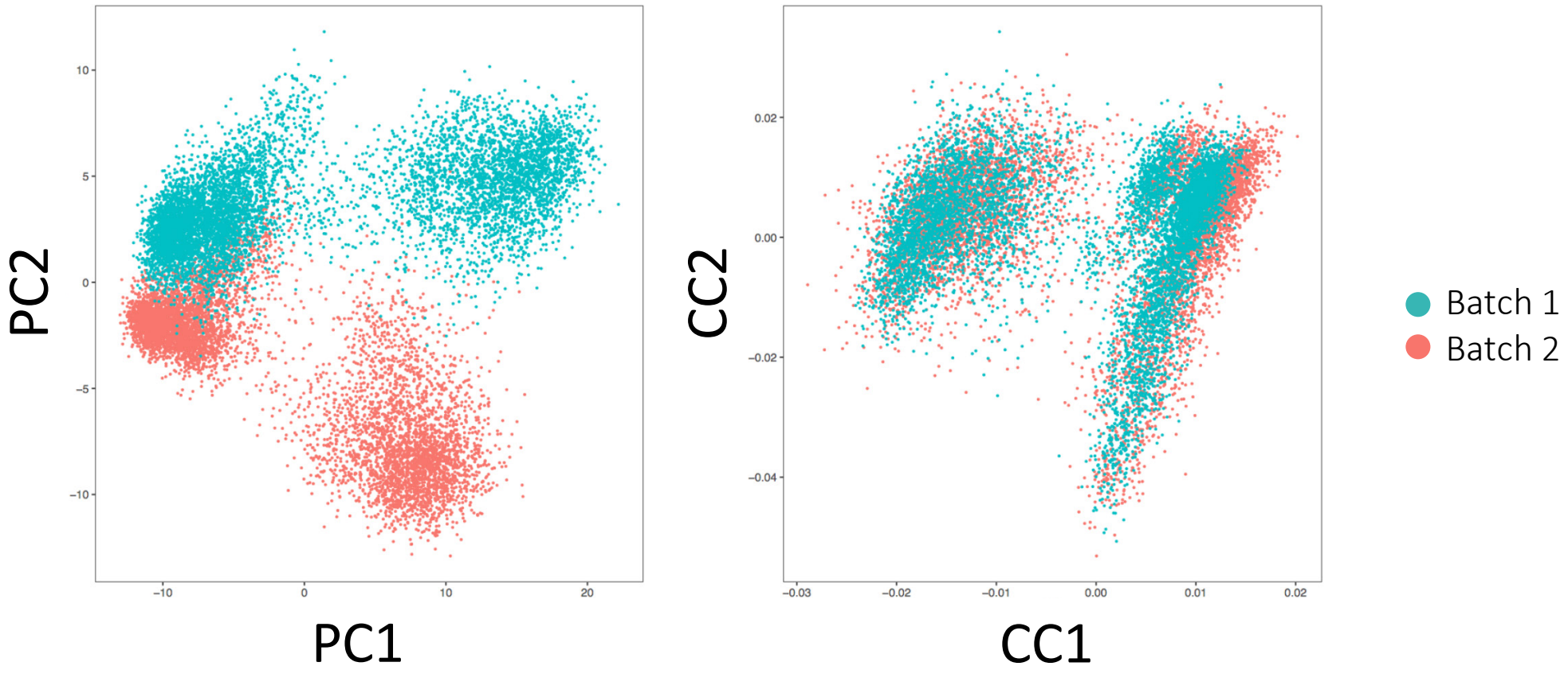




Werner et al. (https://doi.org/10.1371/journal.pone.0113083)

Finding corresponding cells

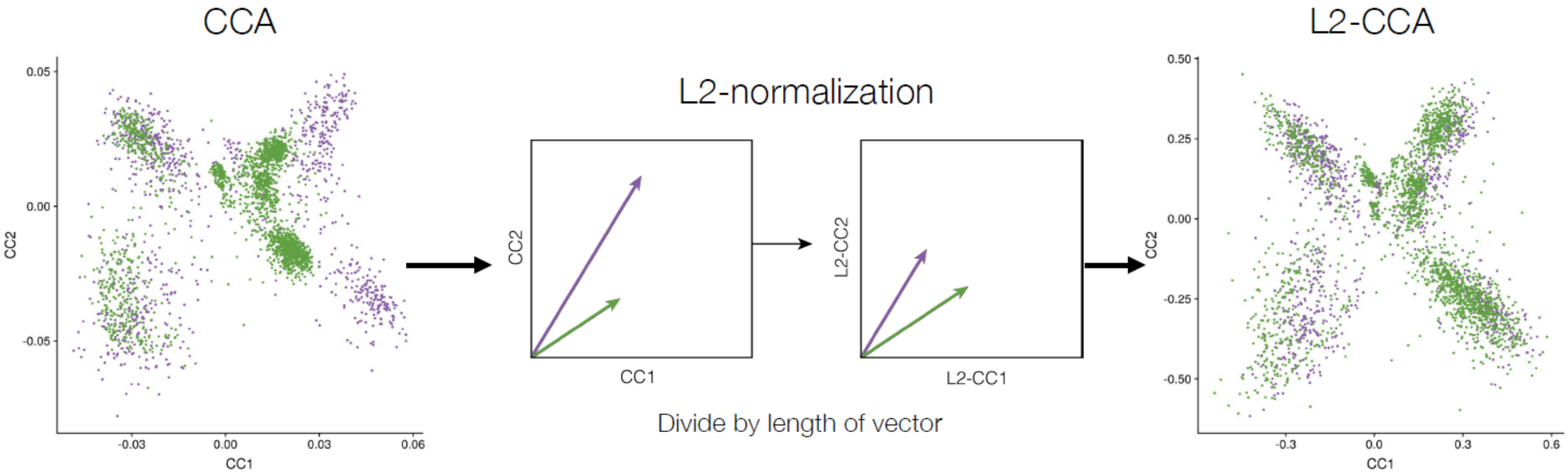
Canonical correlation analysis and normalization



CCA captures correlated sources of variation between two datasets

Finding corresponding cells

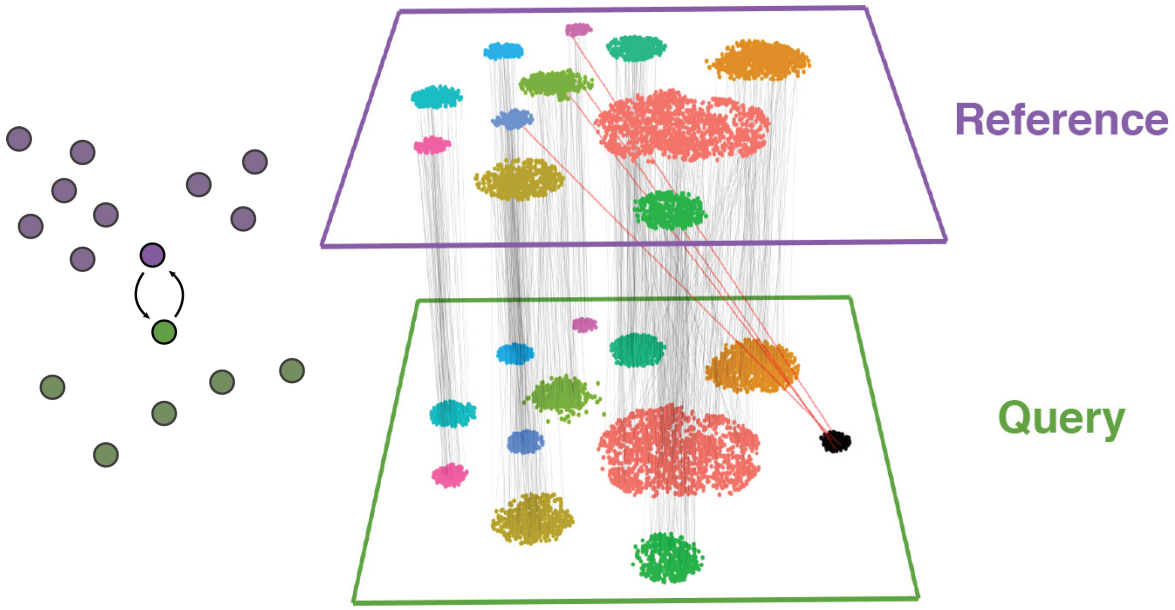
Canonical correlation analysis and normalization

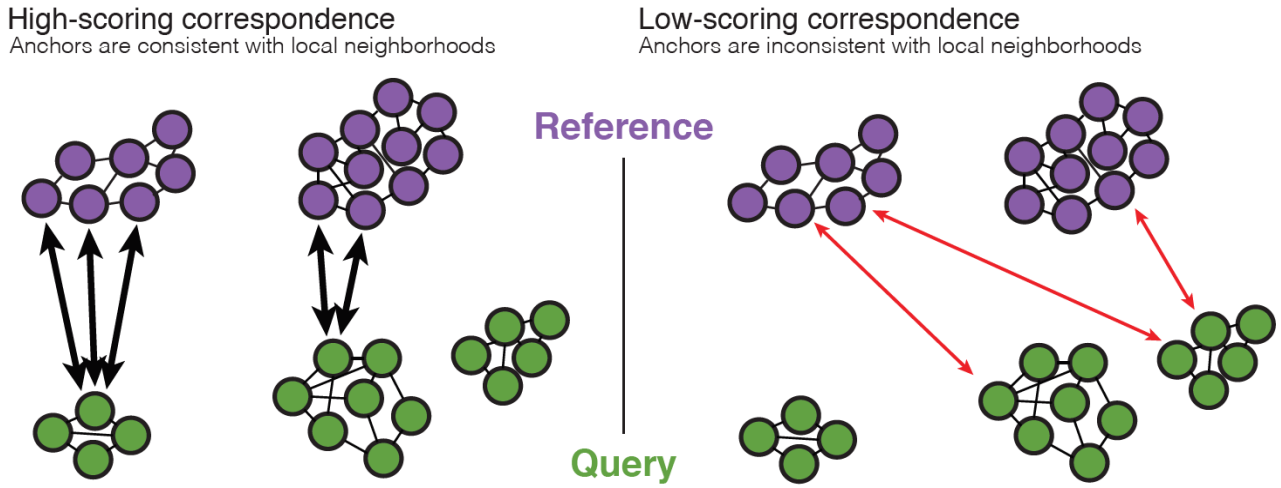


L2-normalization corrects for differences in scale

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

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

2. Construct a weight matrix W that defines the strength of association between each query cell c, and each anchor i

$$W_{c,i} = \frac{\tilde{D}_{c,i}}{\sum_{1}^{j=k.weight} \tilde{D}_{c,j}}$$

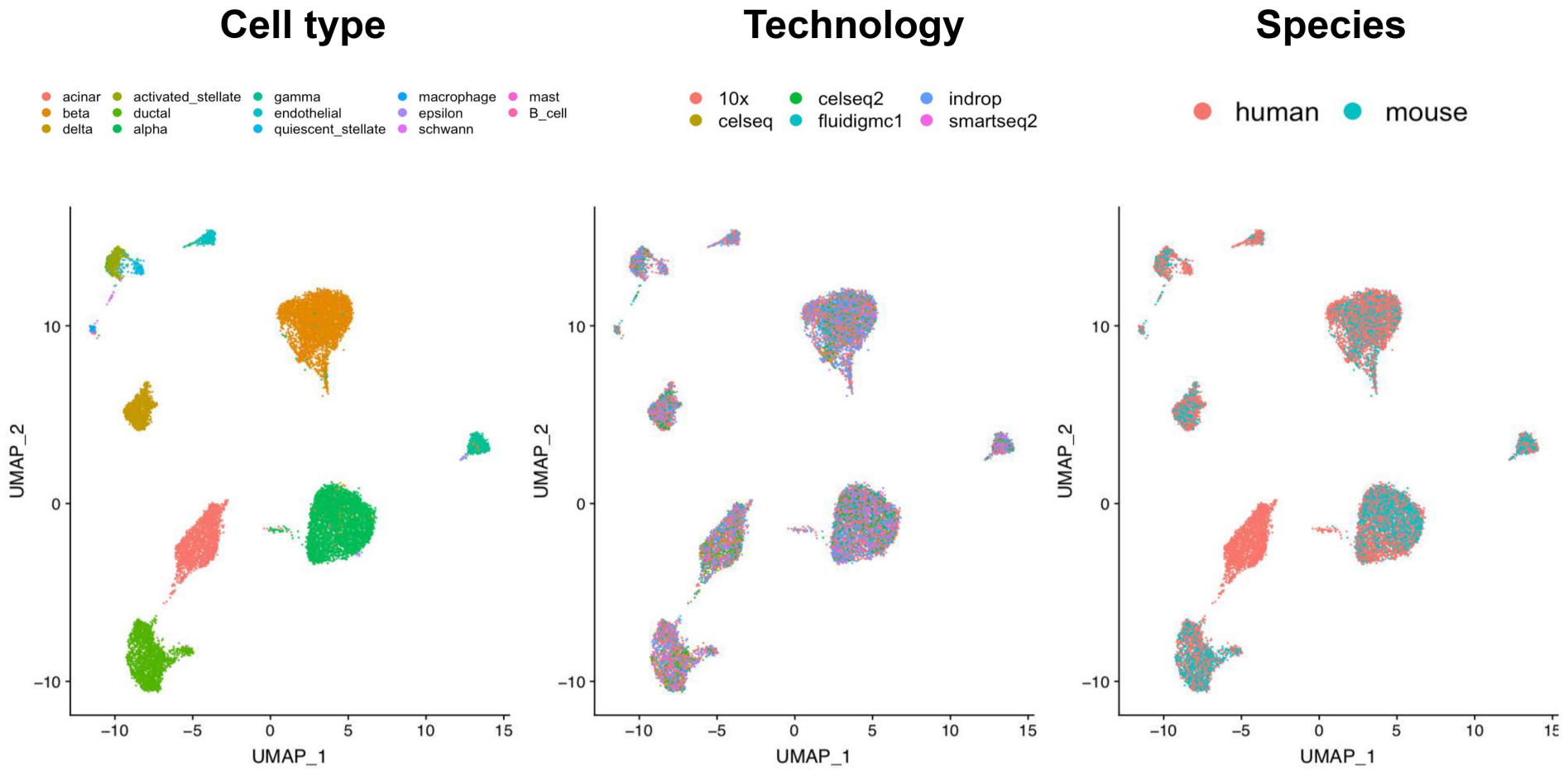
3. Calculate a transformation matrix C using the previously computed weights matrix and the integration matrix as

$$C = BW^T$$

4. Subtract the transformation matrix C from the original expression matrix Y to produce the integrated expression matrix $\hat{Y}$

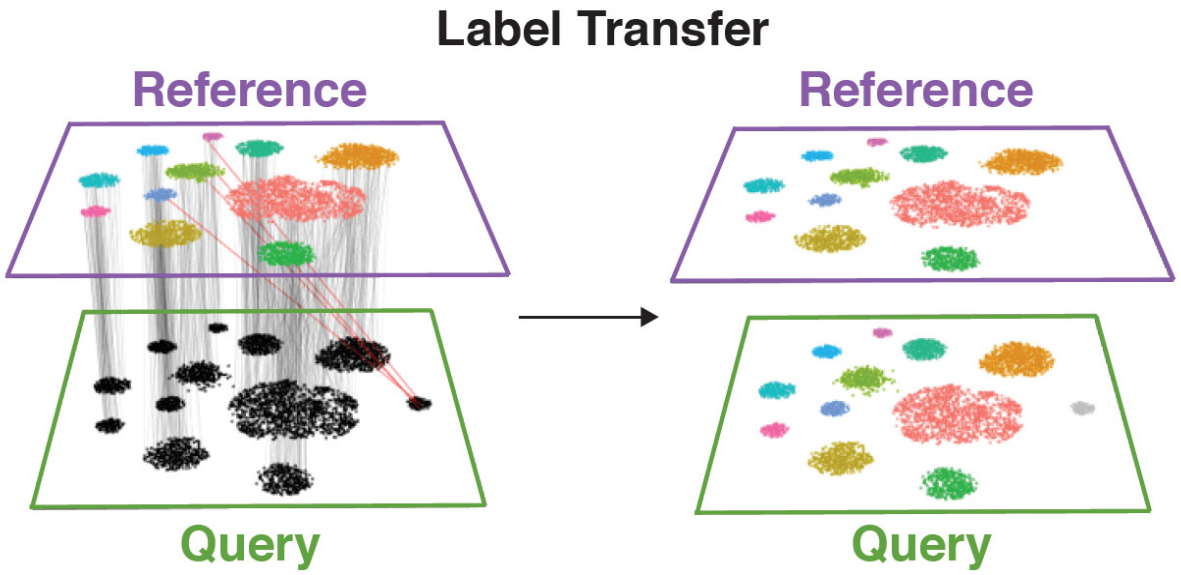
$$\hat{Y} = Y - C$$

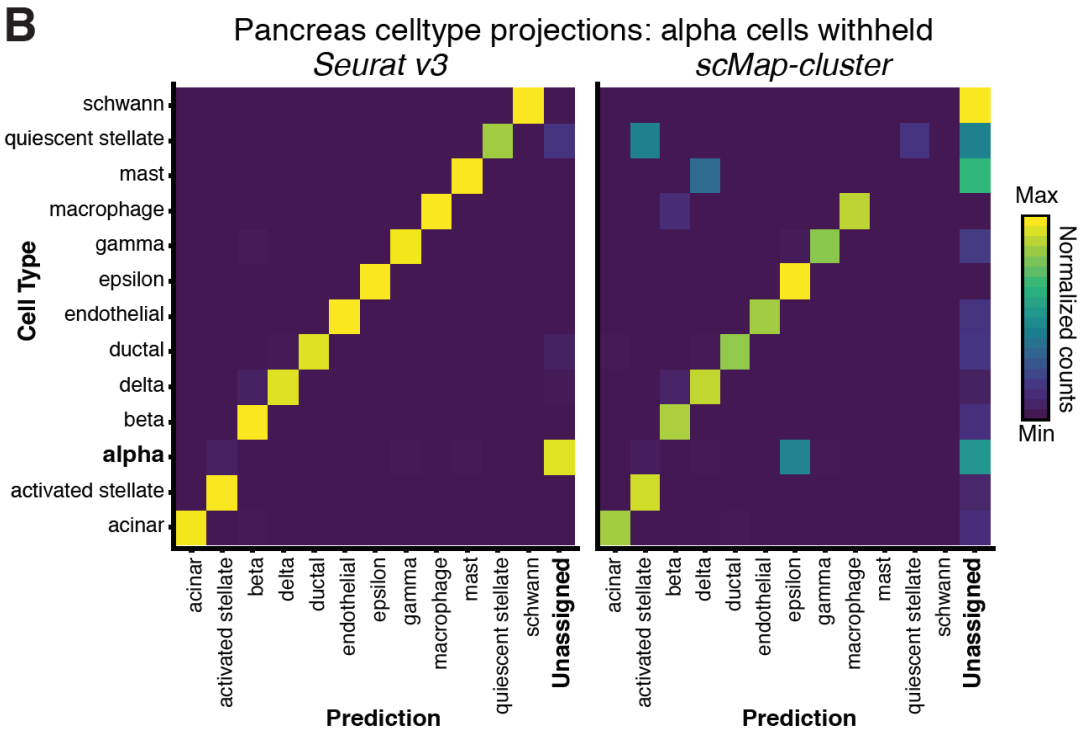
CCA + anchors (Seurat v3)



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

Label transfer (classification)

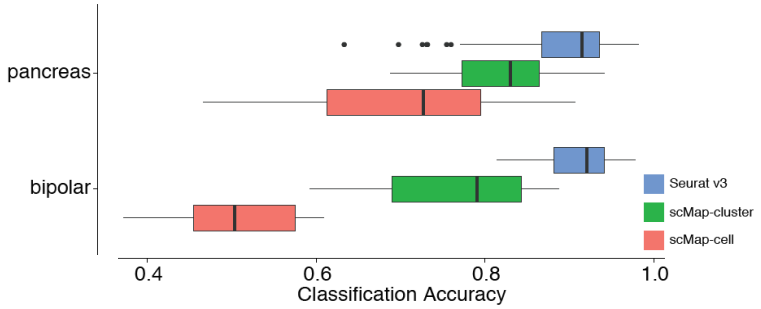




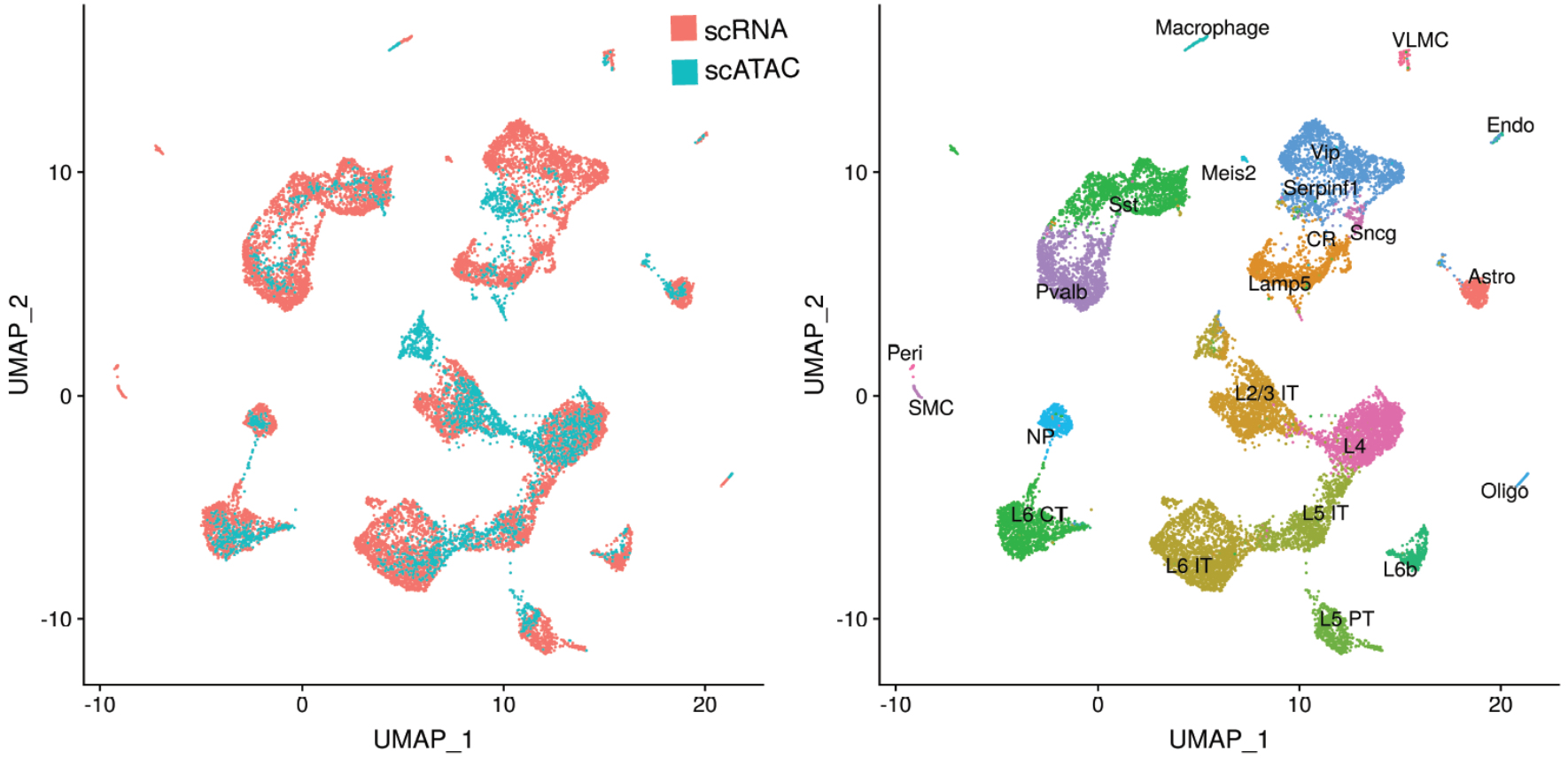
Weighted vote classifier

What is the classification of each cells nearest anchors?

$$P_l = LW^T$$



Integration across modalities



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*

Tran et al. (https://doi.org/10.1186/s13059-019-1850-9)

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.

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

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

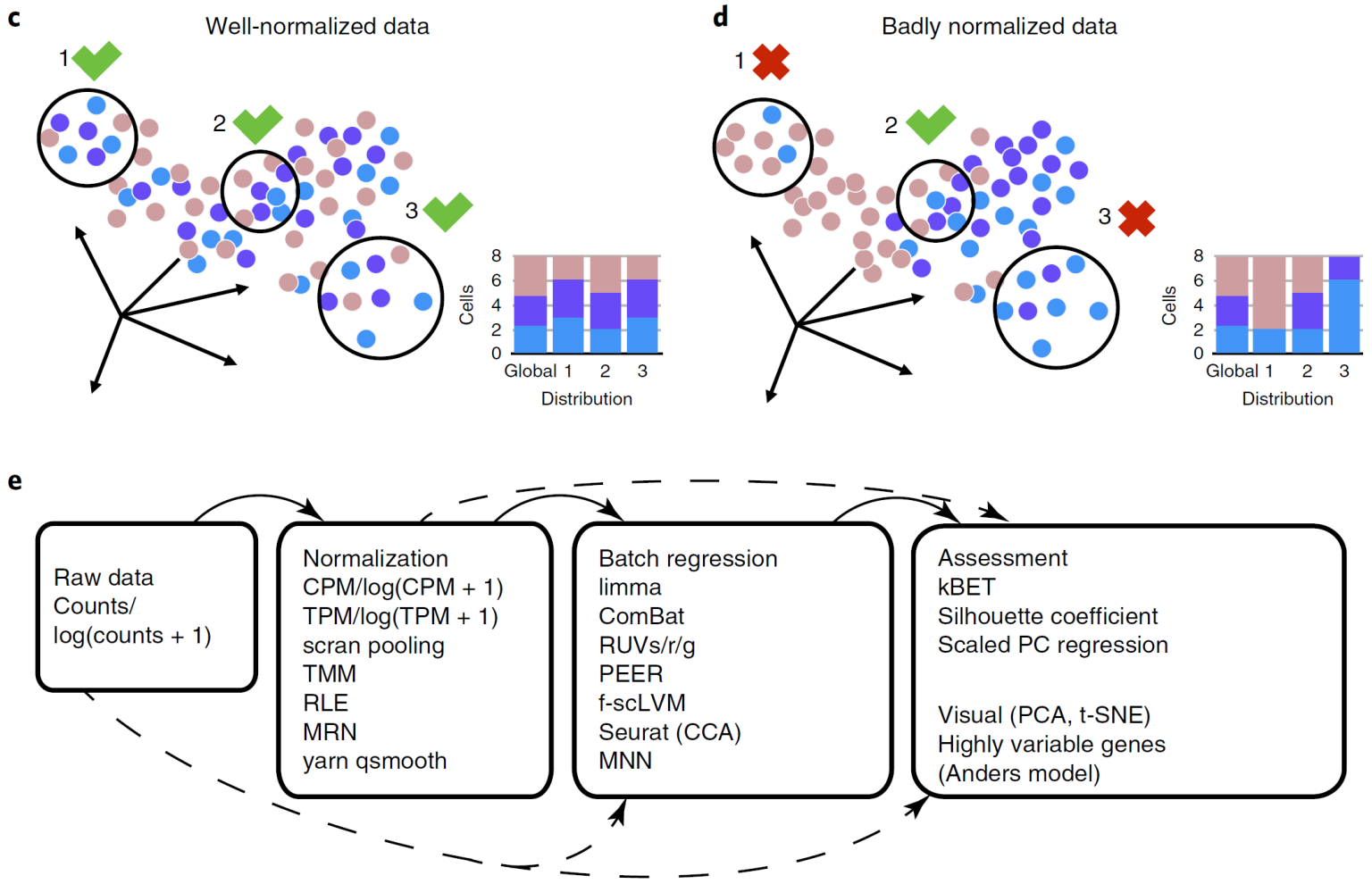
$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j)$$

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

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

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

kBET: *k*-nearest-neighbor batch-effect test

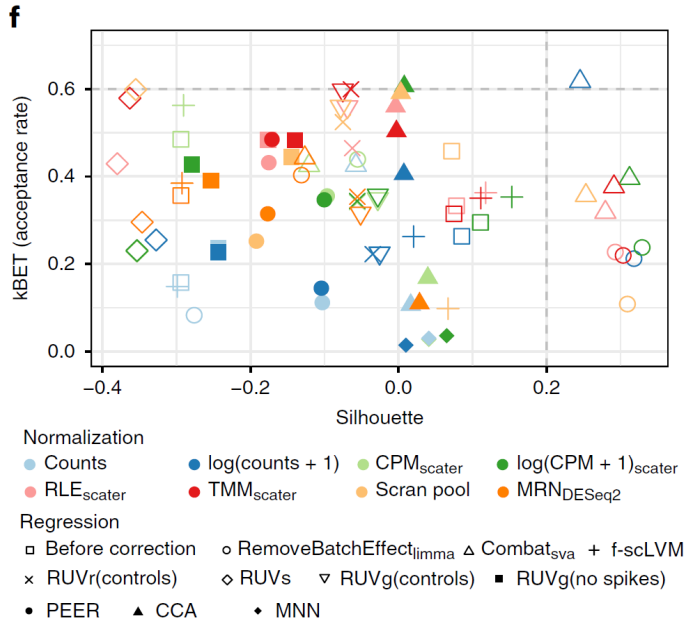


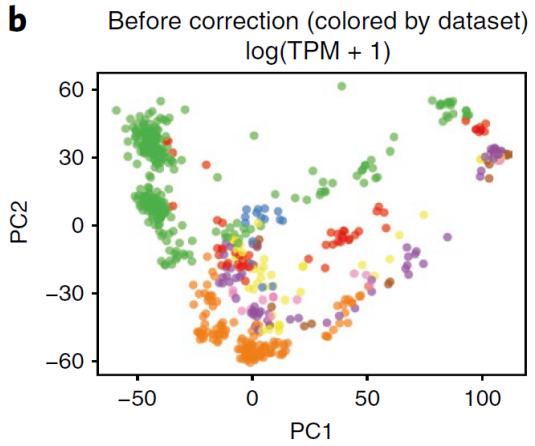
Büttner et al. (https://doi.org/10.1038/s41592-018-0254-1)

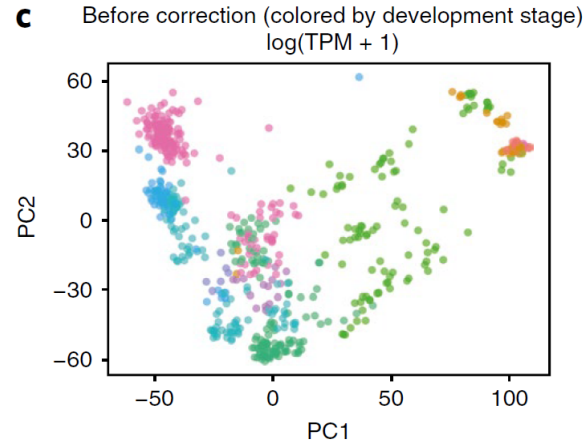
kBET assesses data-integration quality

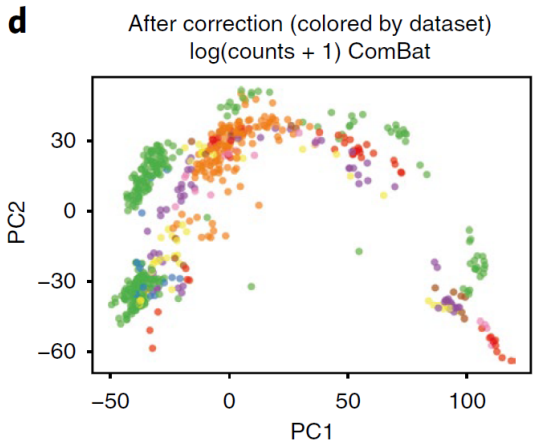
Development stage

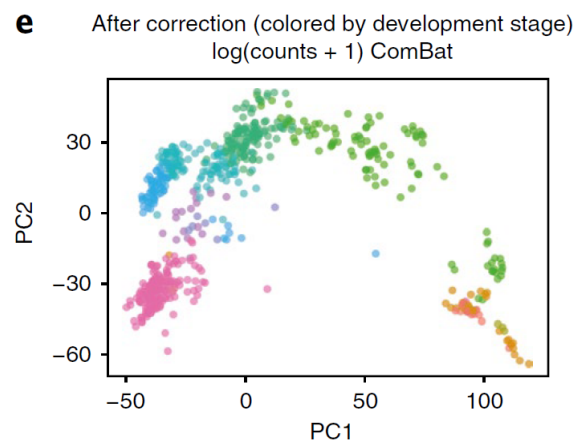
Dataset	Oocyte	Zygote	Pronucleus	2-cell	4-cell	8-cell	16-cell	32-cell	Morula	Blastocyst
Biase et al. 2014[16]	9			20	20					7
Boroviak et al. 2015[20]						3				12
Deng et al. 2014[21]	4			38	14	47	58			133
Fan et al. 2015[19]	9	9		13	7	7			7	13
Goolam et al. 2016[18]				16	64	32	6	6		
Liu et al. 2016[17]	6	5		5	6	6			6	15
Wu et al. 2016[23]	2	2		4	2	2				3
Xue et al. 2013[22]	2		3	3	3	3			3	











Summary (so far)

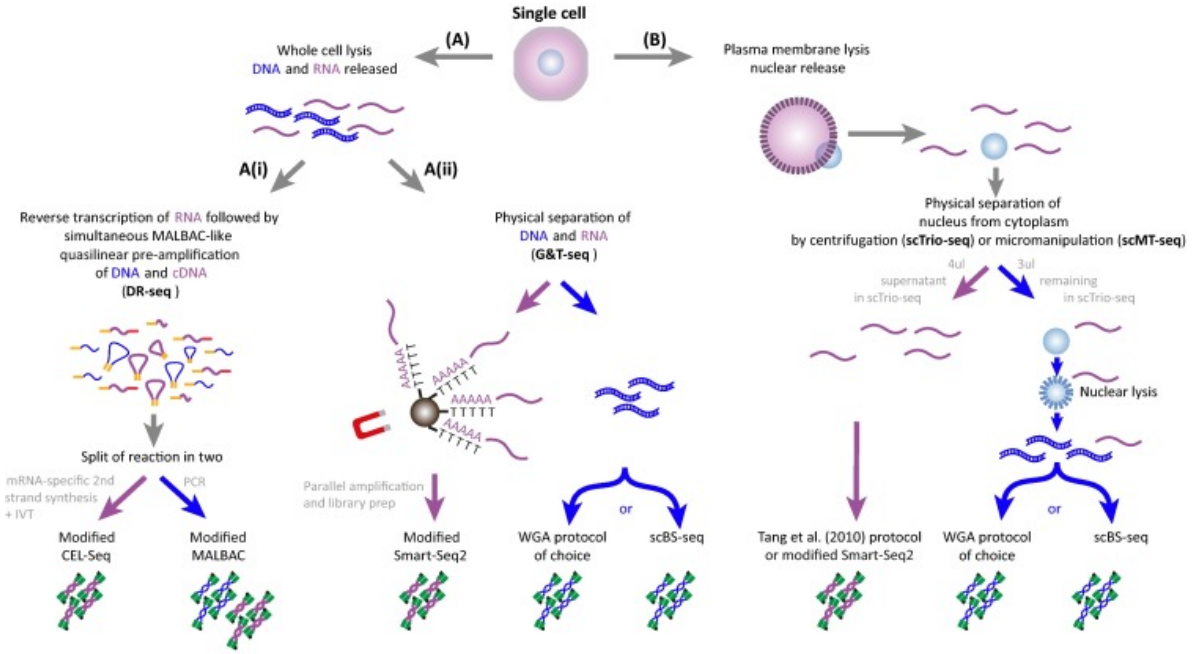
- Integration can allow us to **improve the interpretation** of single-cell data, and build a **multi-modal view** of the tissue

- 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.

Single cell Multi-omics

Same cell



(C)	Loss of nucleic acids	Nature of RNA-seq	Nature of gDNA-seq	Shown amenable to bisulphite-sequencing
DR-seq	Minimal risk of loss	3'end tag transcript seq	MALBAC-like amplified gDNA, contaminated with co-amplified cDNA	no
G&T-seq (like)	Potential loss of mRNA and DNA molecules	Full-length transcript seq	In line with chosen WGA	yes
scTrio-seq	Loss of nearly half of cytoplasmic and all nuclear mRNA-molecules	Full-length transcript seq	reduced representation bisulphite-seq	yes
scMT-seq	Loss of some cytoplasmic and all nuclear mRNA-molecules during micromanipulation	Full-length transcript seq	reduced representation bisulphite-seq	yes

Trends in Genetics

Macaulay et al. (https://doi.org/10.1016/j.tig.2016.12.003)

nature biotechnology
LETTERS
https://doi.org/10.1038/s41587-019-0290-0

High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell

Song Chen, Blue B. Lake and Kun Zhang*

Simultaneous epitope and transcriptome measurement in single cells

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija & Peter Smibert

Nature Methods **14**, 865–868(2017) | Cite this article

nature methods
ARTICLES
https://doi.org/10.1038/s41592-019-0547-z

Simultaneous profiling of 3D genome structure and DNA methylation in single human cells

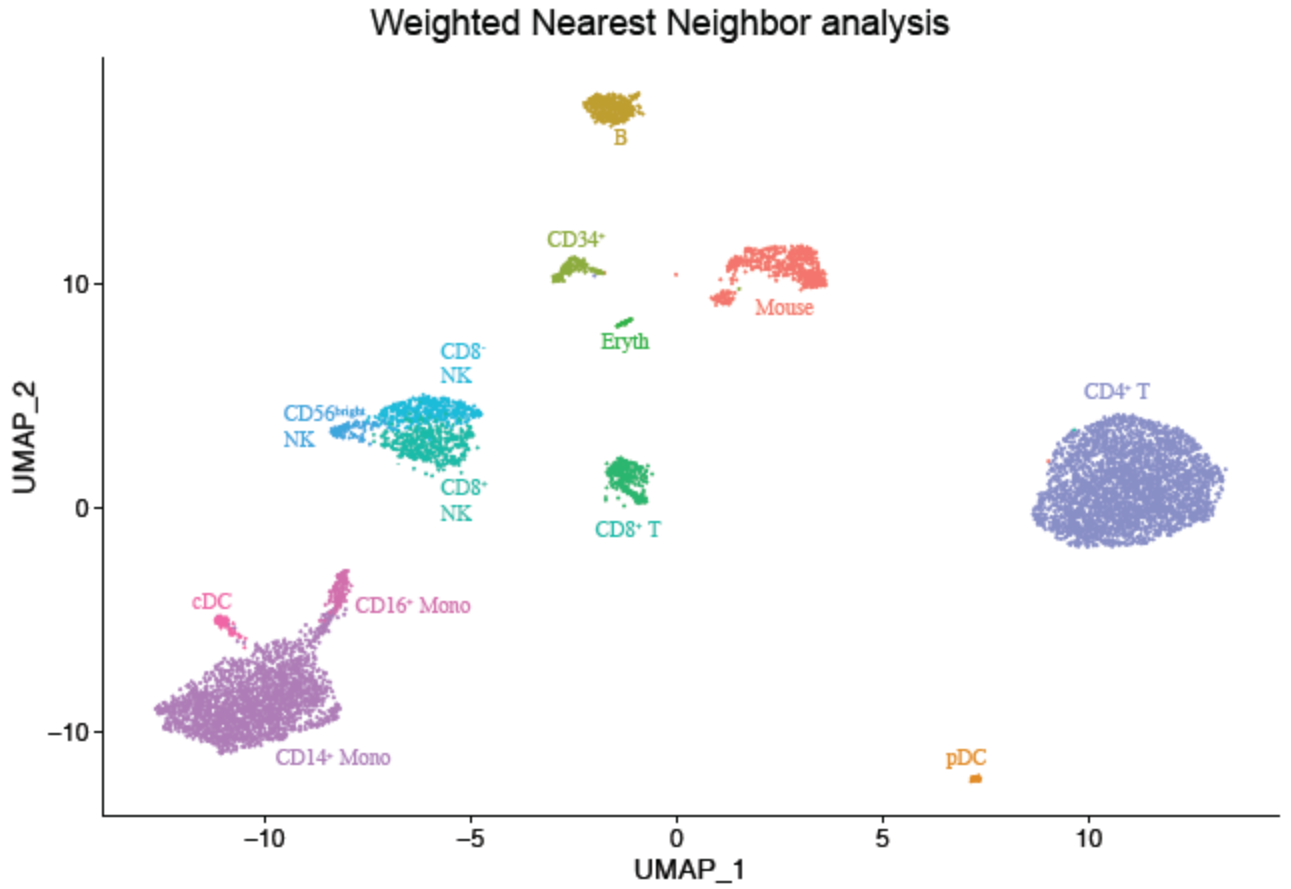
Dong-Sung Lee[1,5], Chongyuan Luo[2,3,5], Jingtian Zhou[2,5], Sahaana Chandran[1], Angeline Rivkin[2], Anna Bartlett[2], Joseph R. Nery[2], Conor Fitzpatrick[4], Carolyn O'Connor[4], Jesse R. Dixon[1*] and Joseph R. Ecker[2,3*]

scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells

Stephen J. Clark, Ricard Argelaguet, ChantriolInt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle & Wolf Reik

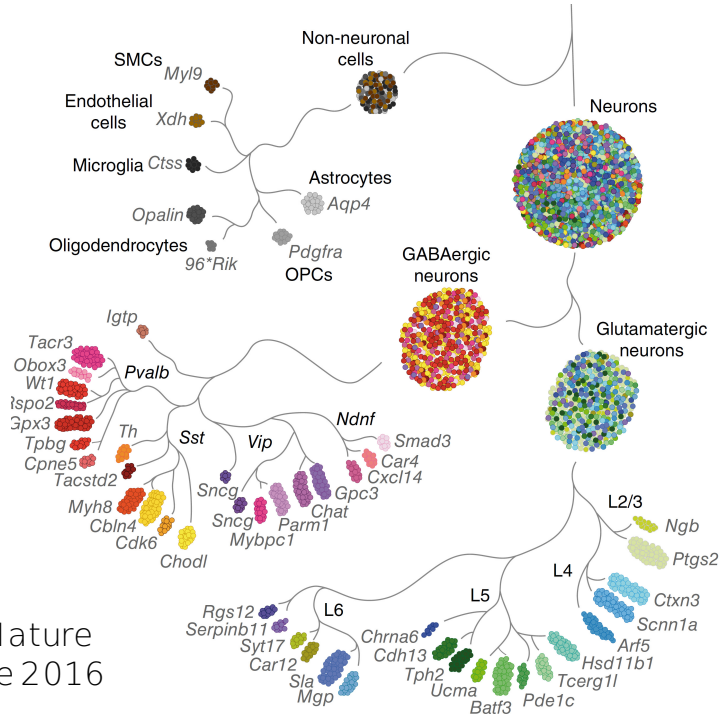
Nature Communications **9**, Article number: 781 (2018) | Cite this article

Seurat v4



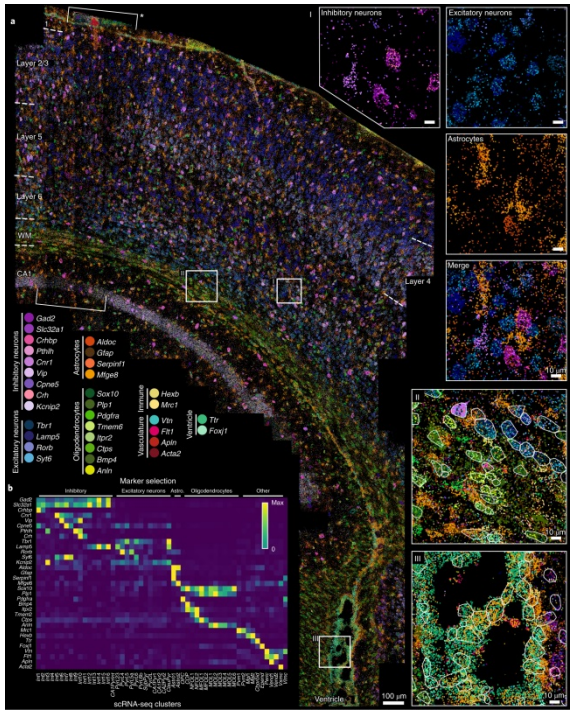
Yuhan Hao et al. (https://www.biorxiv.org/content/10.1101/2020.10.12.335331v1)

SpaGE: Spatial Gene Enhancement using scRNA-seq



Tasic et al, Nature Neuroscience 2016

Codeluppi et al, Nature methods 2018



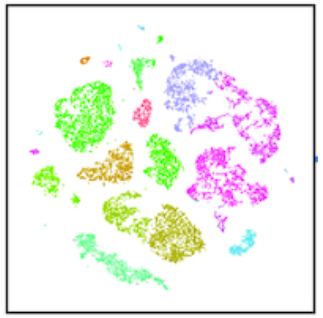
scRNA-seq
Whole transcriptome
Loss of spatial information

Spatial transcriptomics
Retain spatial information
Limited in the number of genes

Abdelaal et al., *NAR* (https://doi.org/10.1093/nar/gkaa740)

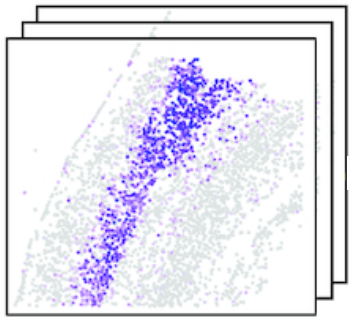
SpaGE overview

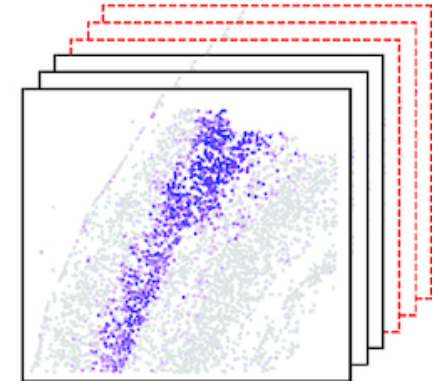
scRNA-seq (reference)
(~25,000 genes)



+

Spatial transcriptomics (query)
(~ 10-1,000 genes)





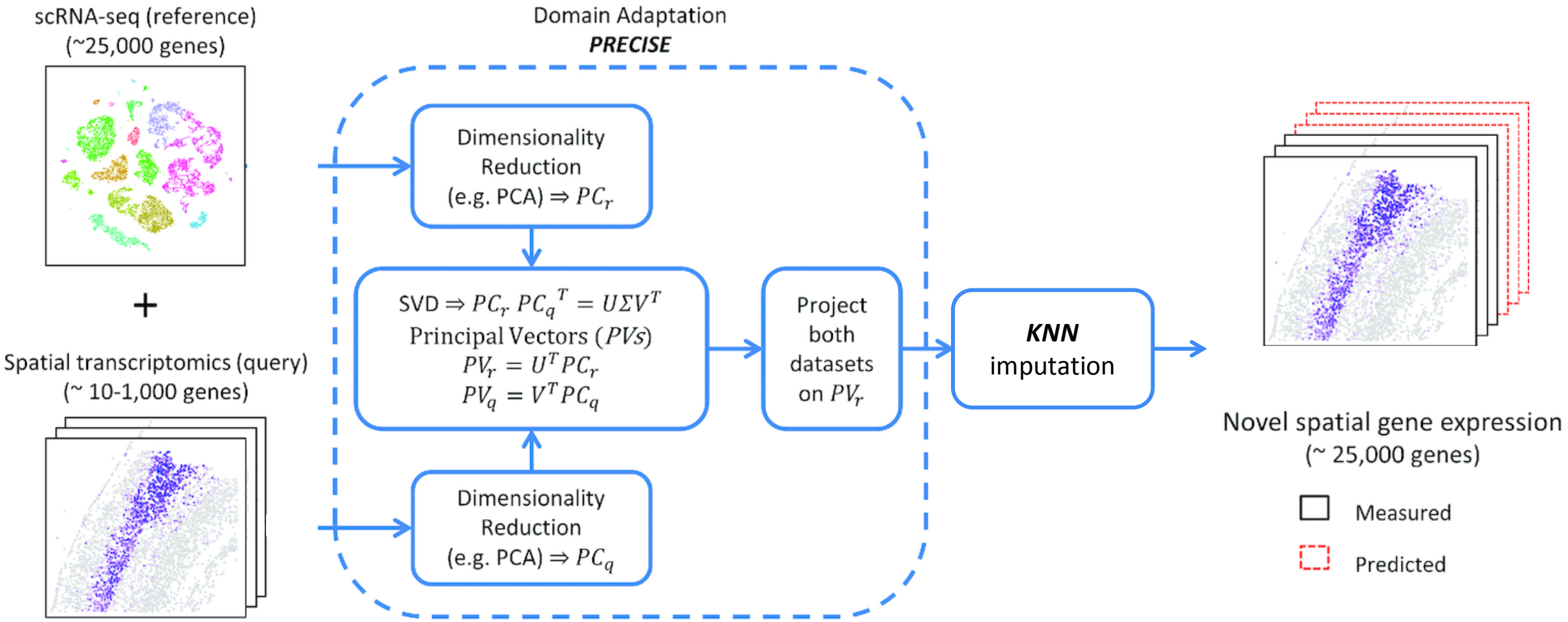
Novel spatial gene expression
(~ 25,000 genes)

Measured

Predicted

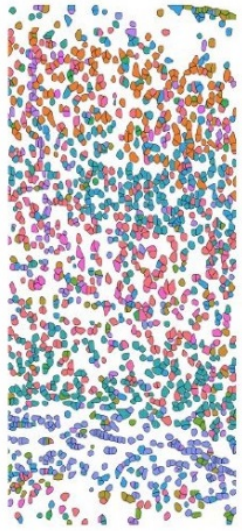
PRECISE: Mourragui et al, Bioinformatics. 2019

SpaGE overview

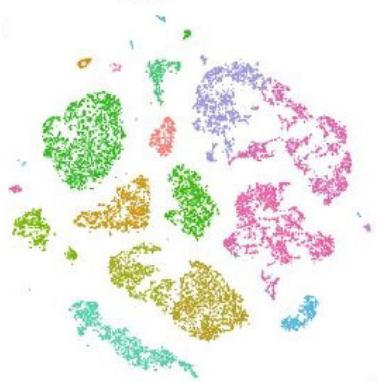


PRECISE: Mourragui et al, Bioinformatics. 2019

SpaGE on primary visual cortex (VISp)

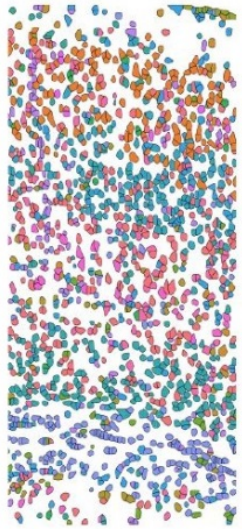


STARmap
1,549 cells
1,020 genes
Wang et al. Science 2018

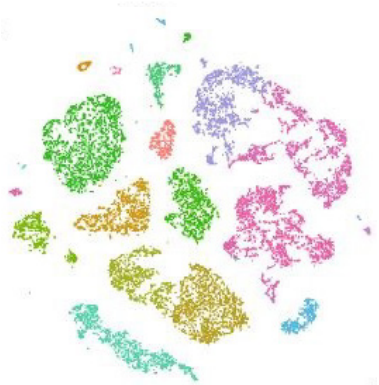


scRNA-seq
14,249 cells
34,617 transcripts
Tasic et al. Nature 2018

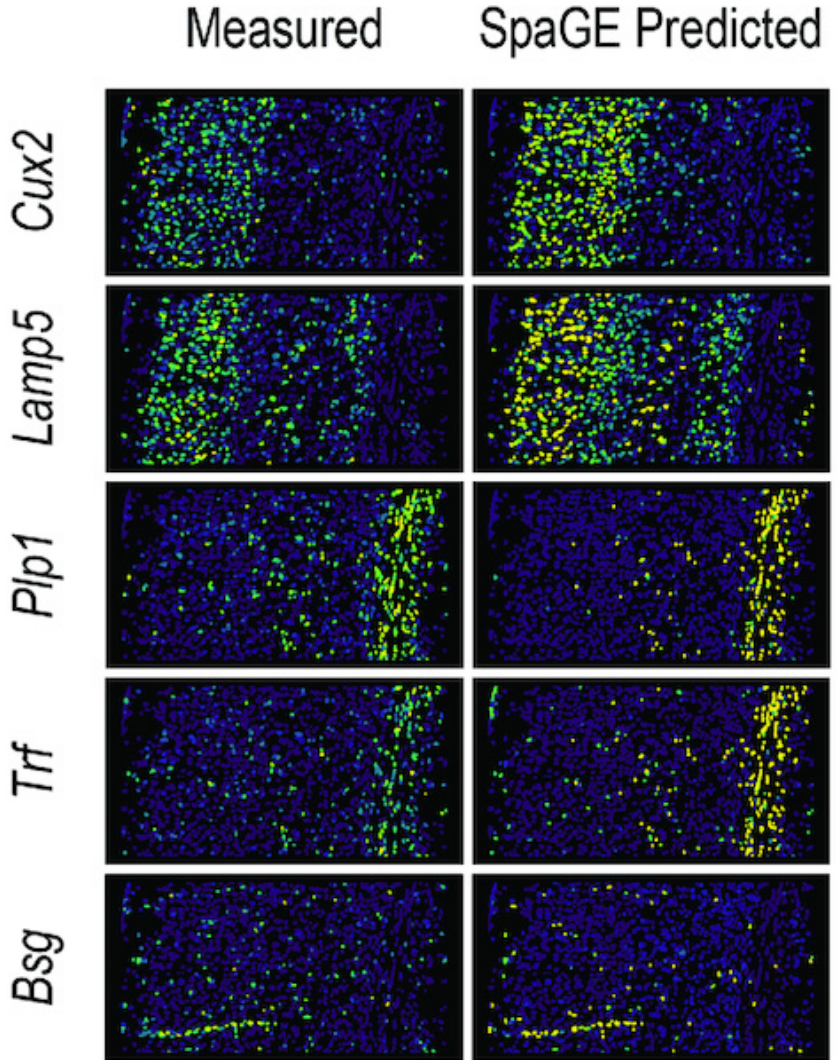
SpaGE on primary visual cortex (VISp)



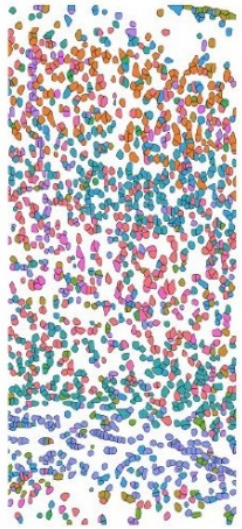
STARmap
1,549 cells
1,020 genes
Wang et al. Science 2018



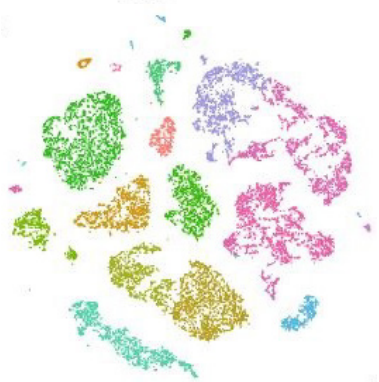
scRNA-seq
14,249 cells
34,617 transcripts
Tasic et al. Nature 2018



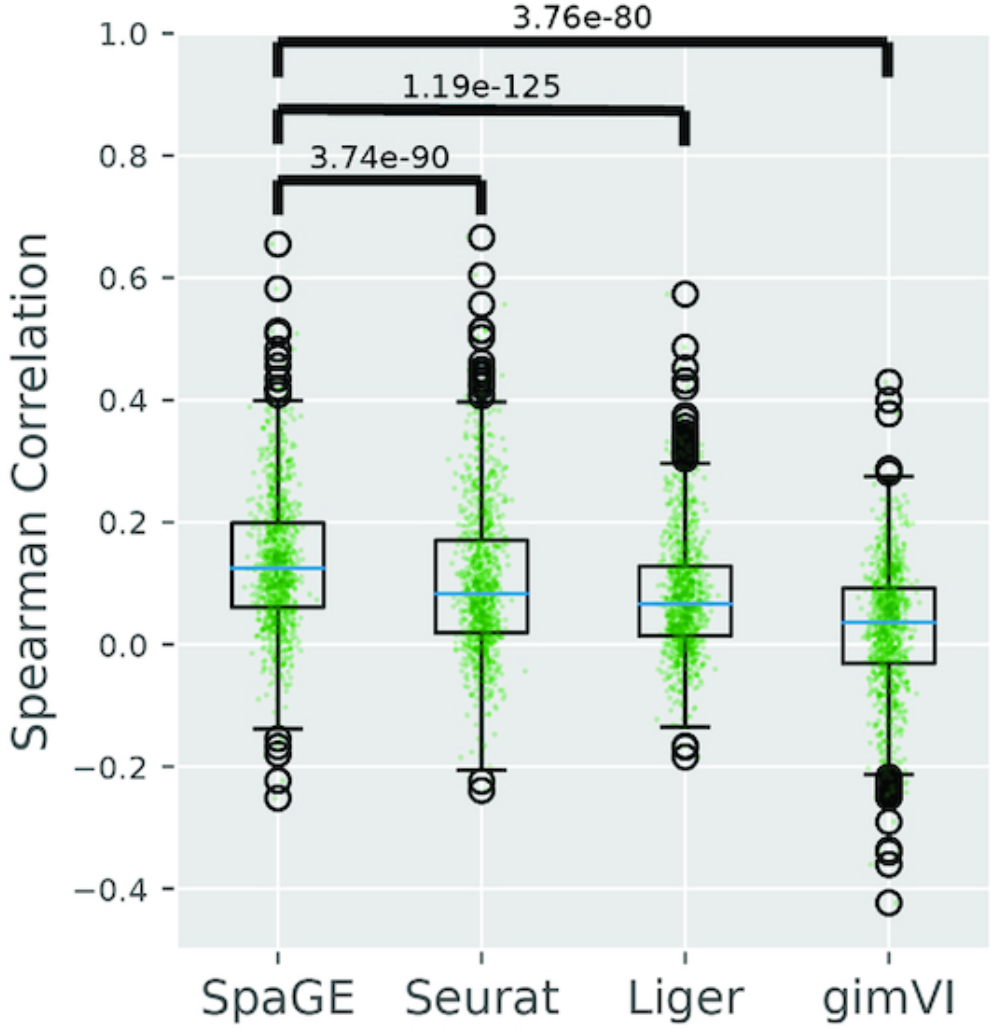
SpaGE outperforms Seurat, Liger and gimVI



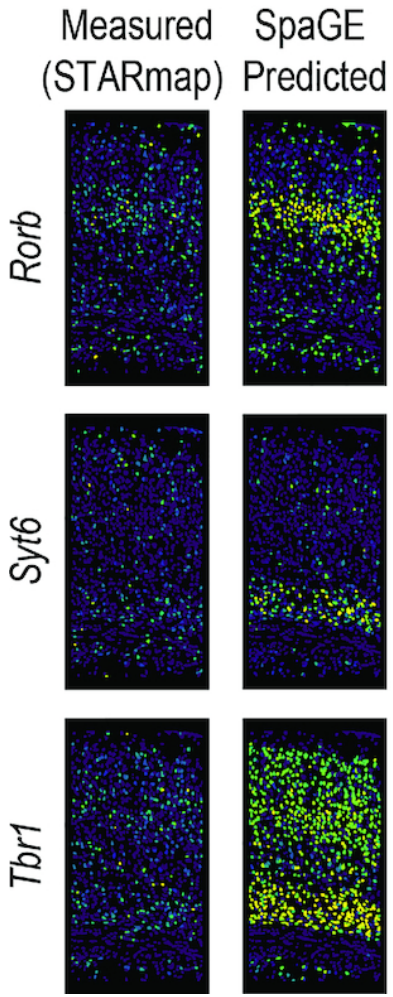
STARmap
1,549 cells
1,020 genes
Wang et al. Science 2018



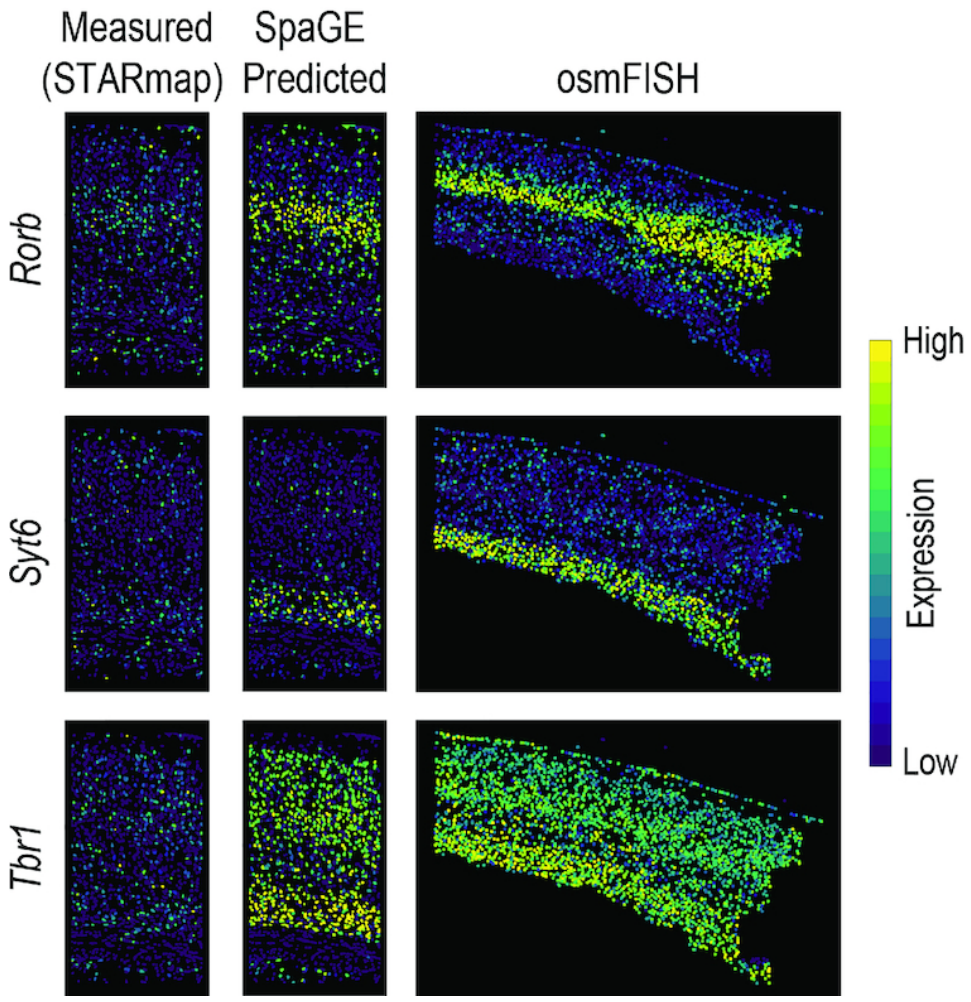
scRNA-seq
14,249 cells
34,617 transcripts
Tasic et al. Nature 2018



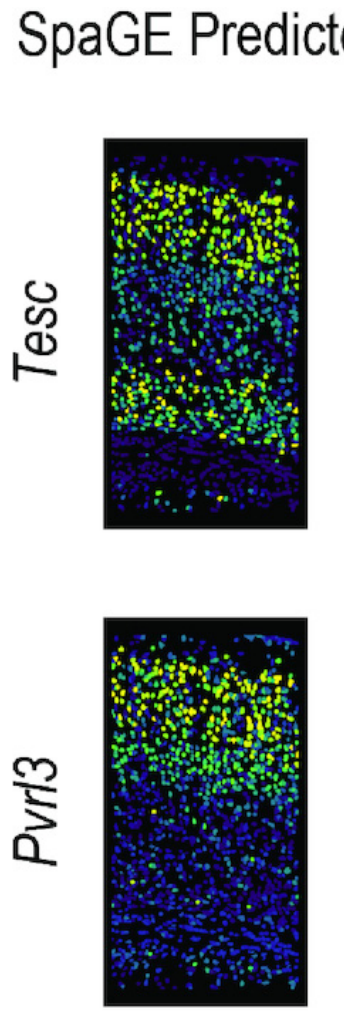
Wrongly measured, correctly predicted



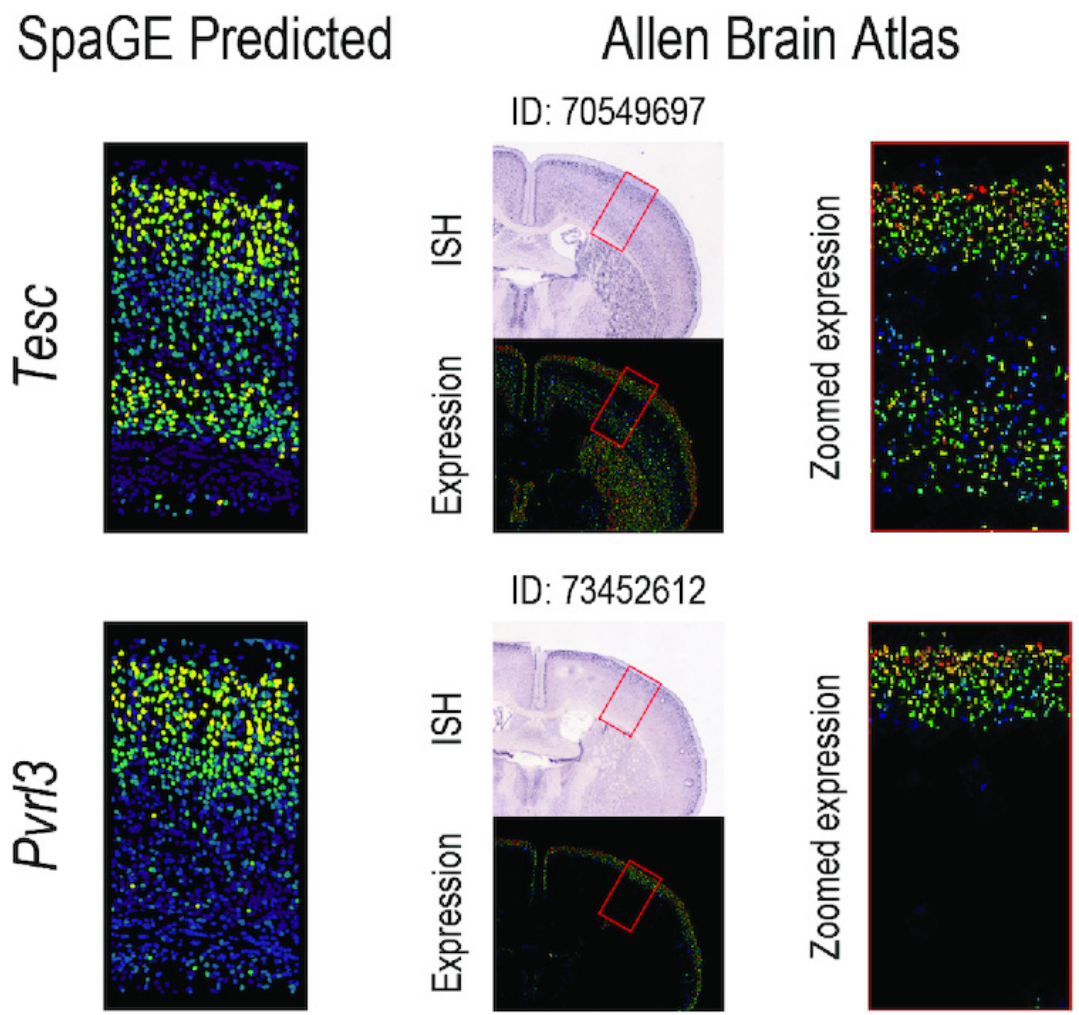
Wrongly measured, correctly predicted



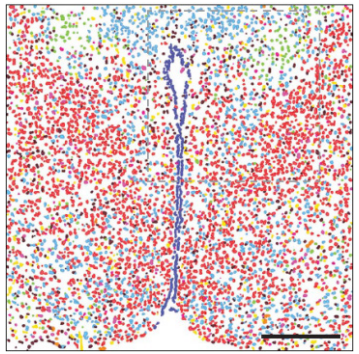
Predicting unmeasured genes



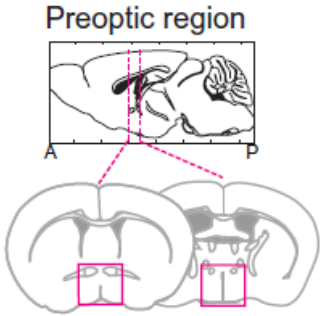
Predicting unmeasured genes



SpaGE scales to large spatial datasets

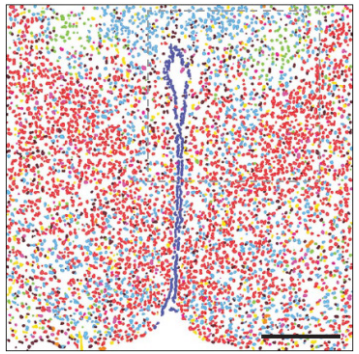


MERFISH
64,373 cells
155 genes
Moffit et al. Science 2018

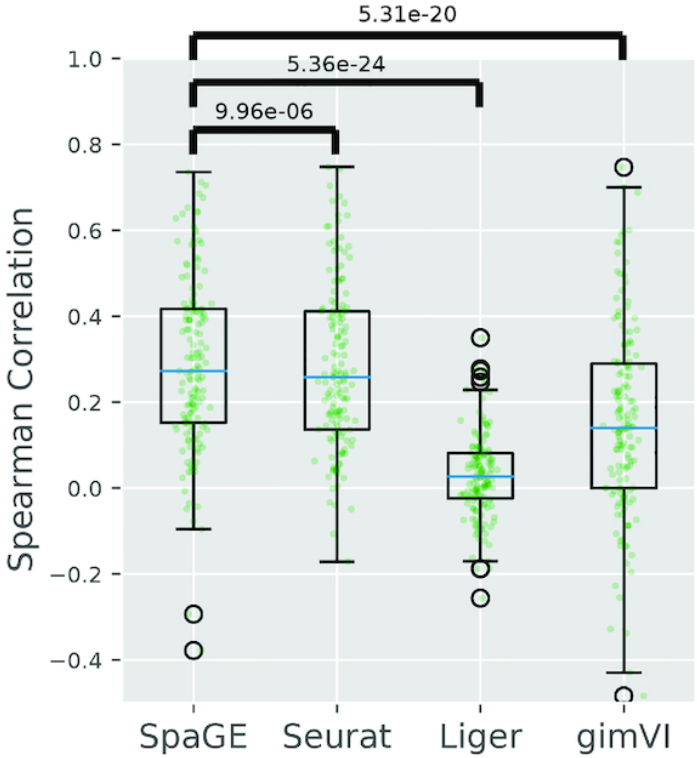


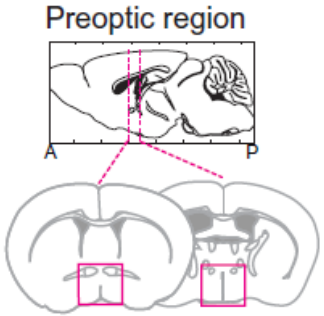
scRNA-seq
31,299 cells
18,646 transcripts
Moffit et al. Nature 2018

SpaGE scales to large spatial datasets



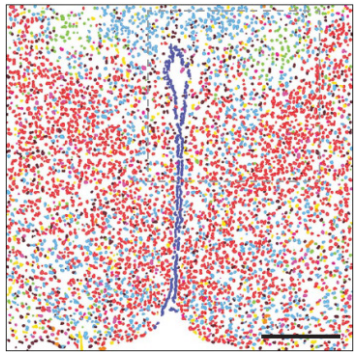
MERFISH
64,373 cells
155 genes
Moffit et al. Science 2018



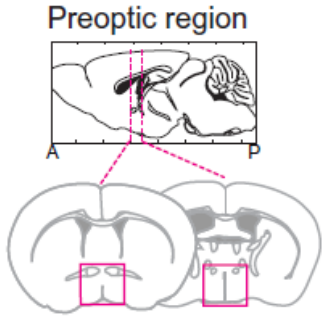


scRNA-seq
31,299 cells
18,646 transcripts
Moffit et al. Nature 2018

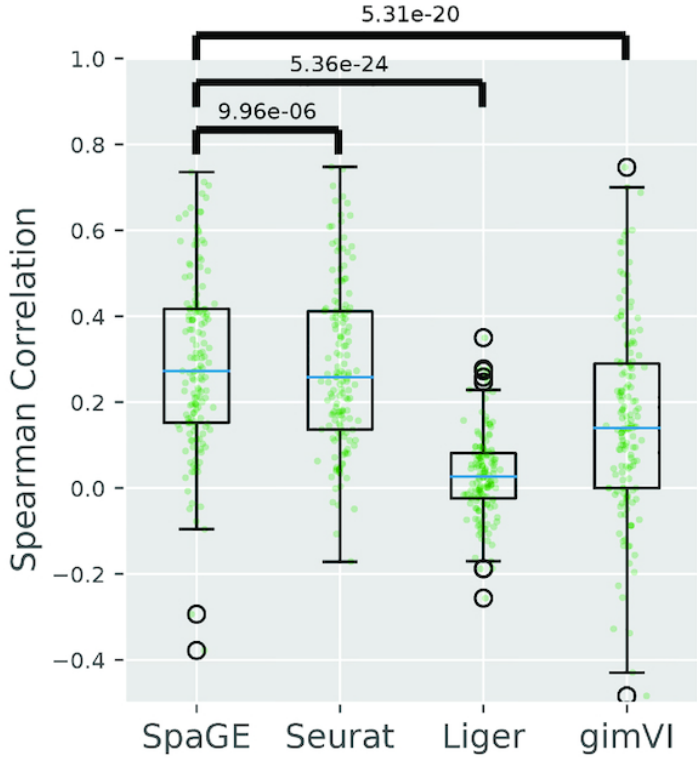
SpaGE scales to large spatial datasets

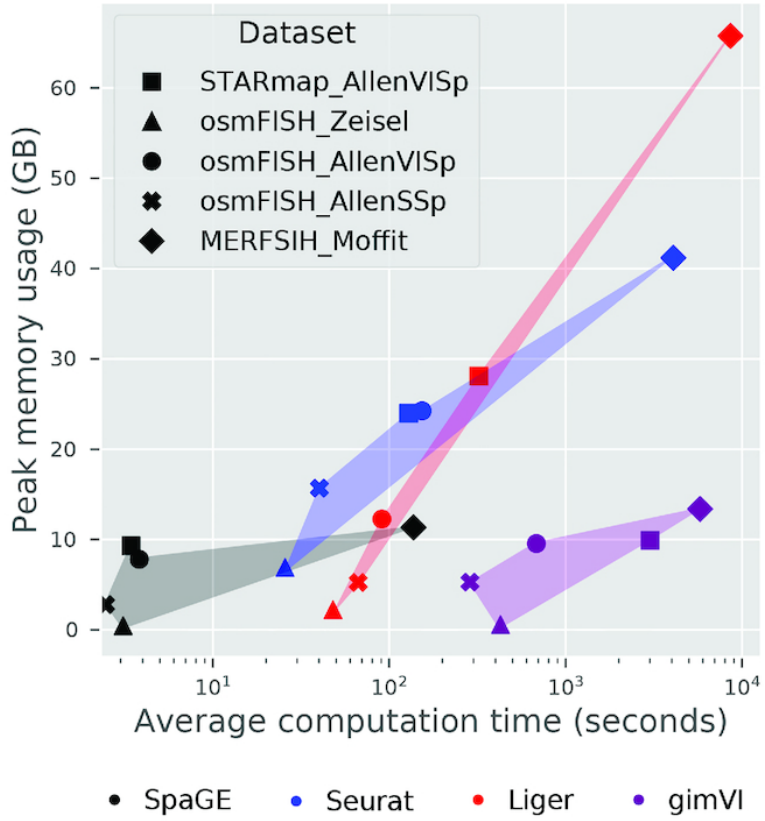


MERFISH
64,373 cells
155 genes
Moffit et al. Science 2018



scRNA-seq
31,299 cells
18,646 transcripts
Moffit et al. Nature 2018





Summary

- SpaGE predicts spatial expression of unmeasured genes at the single cell level

- SpaGE outperforms state-of-the-art methods using is simpler, scalable and more flexible approach.

Data integration practical

- MNN correction
- Seurat v3
- Four pancreatic datasets

Resources

- Stuart et al. “Comprehensive integration of single cell data”

https://www.biorxiv.org/content/10.1101/460147v1

- Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”

https://doi.org/10.1038/nbt.4091

- 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/

- Hemberg’s group course: Analysis of single cell RNA-seq data

https://scrnaseq-course.cog.sanger.ac.uk/website/index.html

- Seurat Integration and Label Transfer tutorial

https://satijalab.org/seurat/v3.0/pancreas_integration_label_transfer.html

Thank You!

t.r.m.abdelaal-1@tudelft.nl