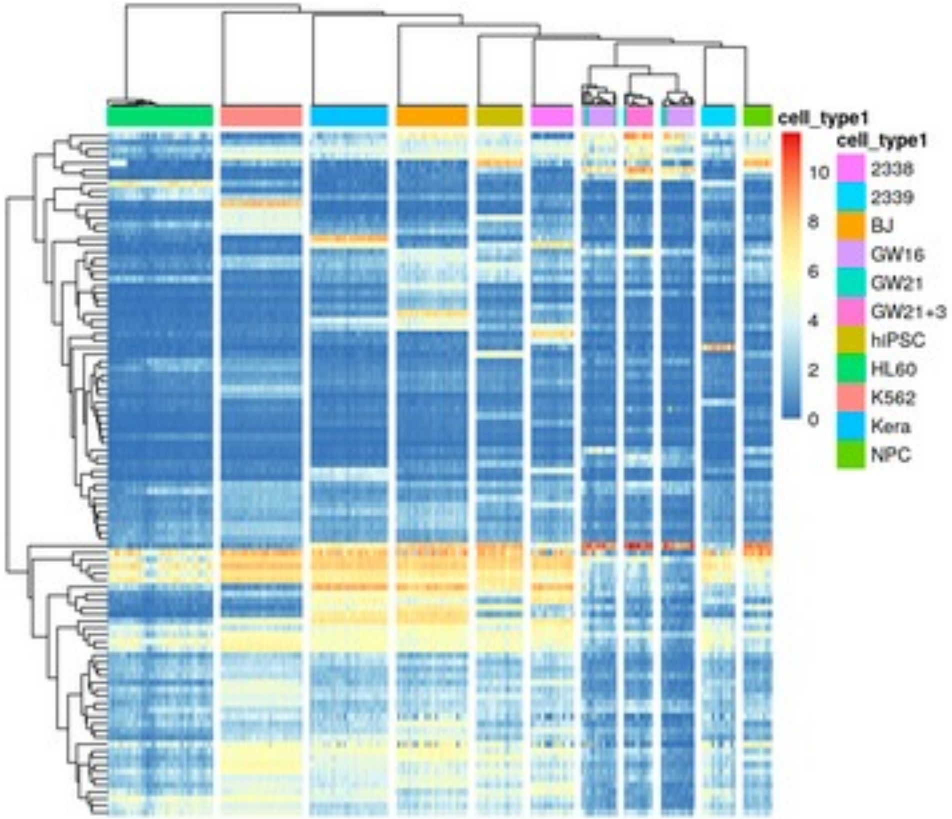
Analysis of Single Cell RNA-Seq Data: Data integration and batch effect correction

Orr Ashenberg, Jacques Serizay, Arnav Mehta

June 8, 2021

Physalia Courses



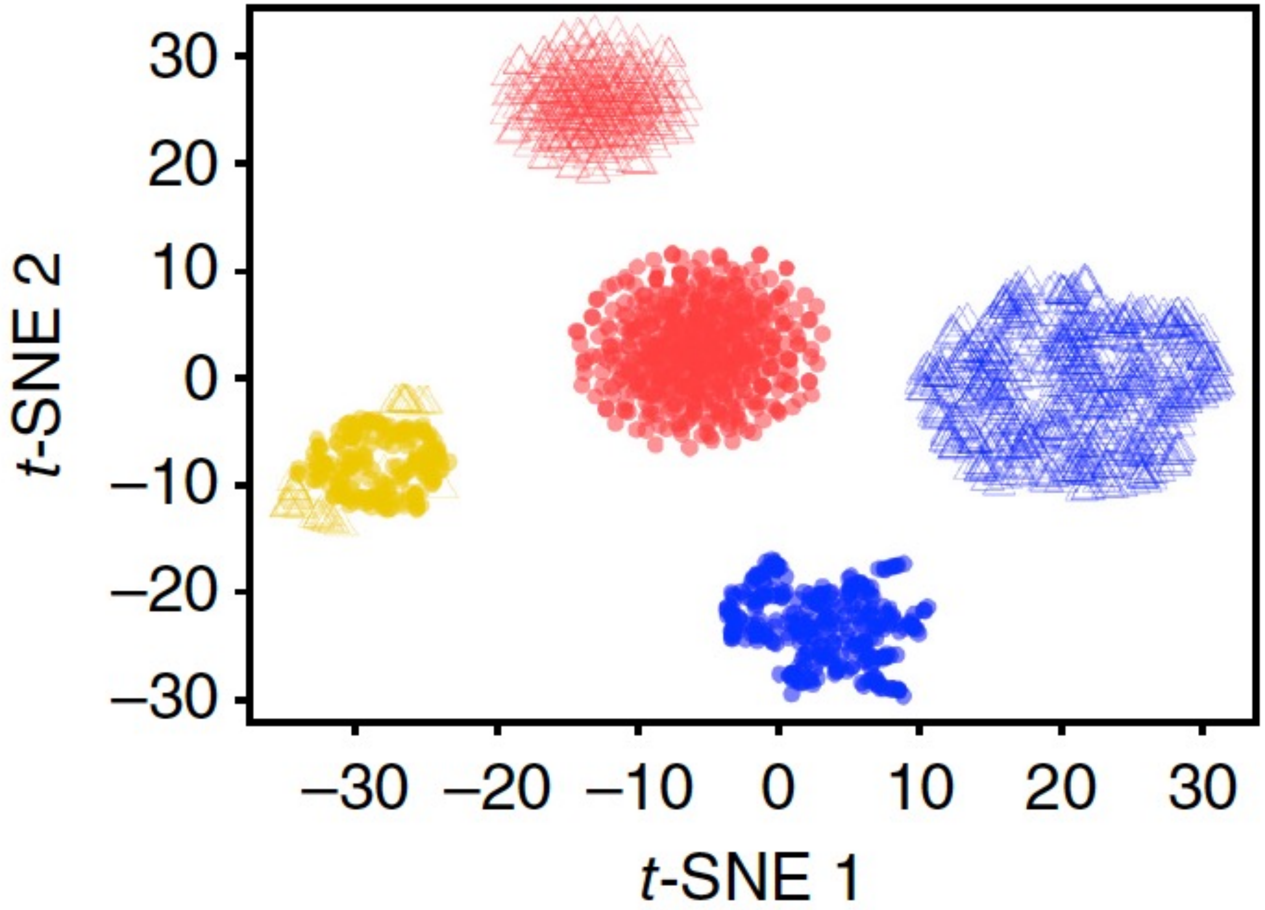
Batch Effects Outline

- Sources of batch effects

- Computational approaches to correct for batch effects and integrate data

- Assessing batch effect corrections and the assumptions of these methods

Distinguishing biological effects from technical, batch effects is a difficult problem

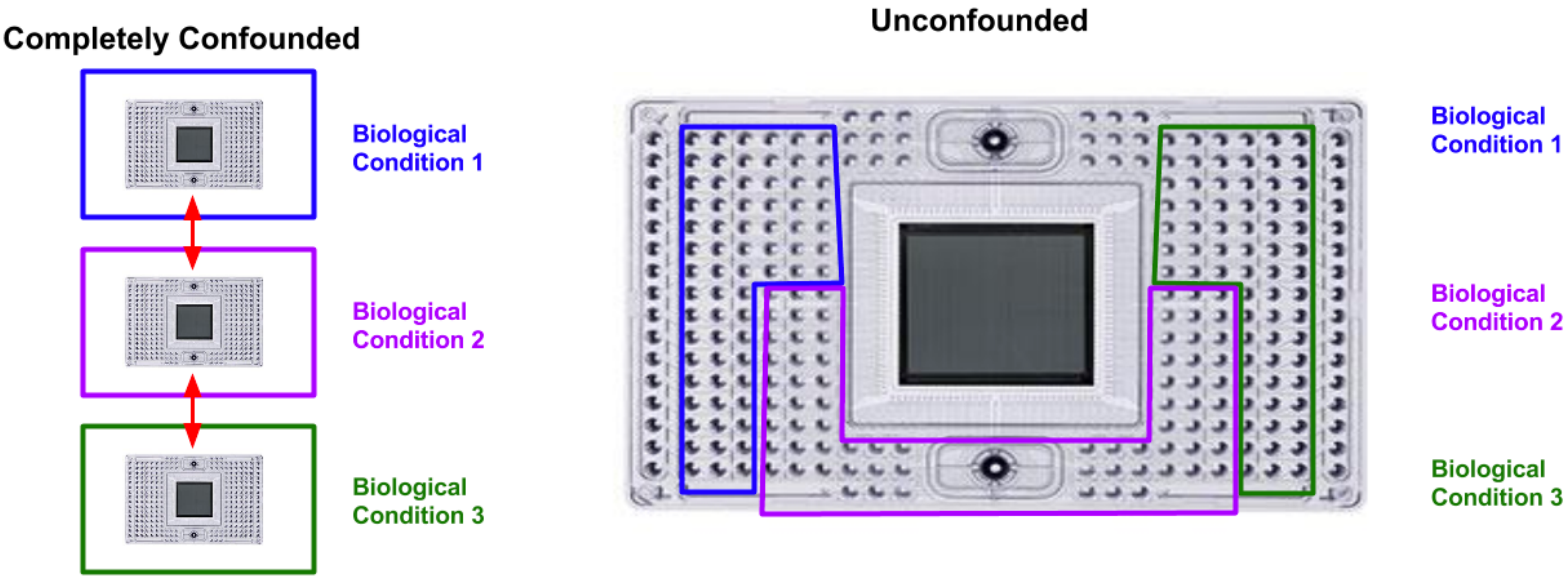


Cells are colored by cell type.

Symbols represent different batches.

Correcting for batch effects allows us to combine datasets and boost biological signal, while reducing technical confounders

The most powerful way to control batch effects is with careful experimental design



Sound experimental design : Replication, Randomization and Blocking

- R. A. Fisher, 1935

https://scrnaseq-course.cog.sanger.ac.uk/website/ideal-scrnaseq-pipeline-as-of-oct-2017.html

Batch effects: technical sources

- Differences in how samples are sequenced
  - sequencing depth and saturation
  - sequencing instrument

Miseq

~ 20M reads total

Nextseq

~ 500M reads total

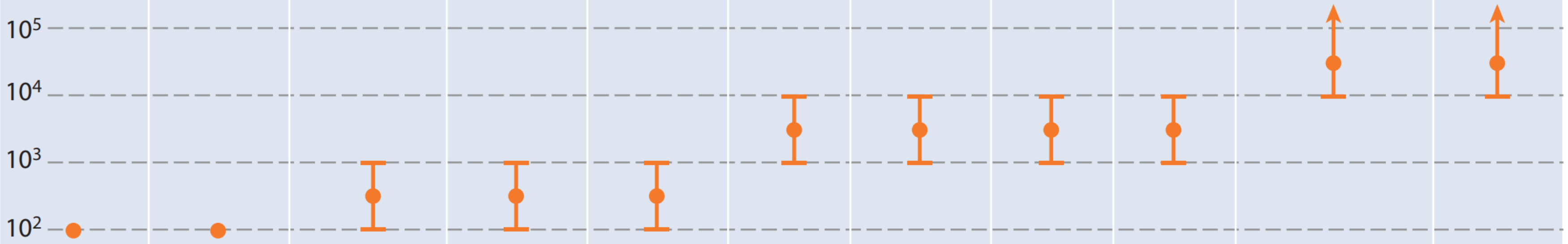
HiSeq 4000

4 billion reads

Batch effects: technical sources

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′

Number of cells per assay



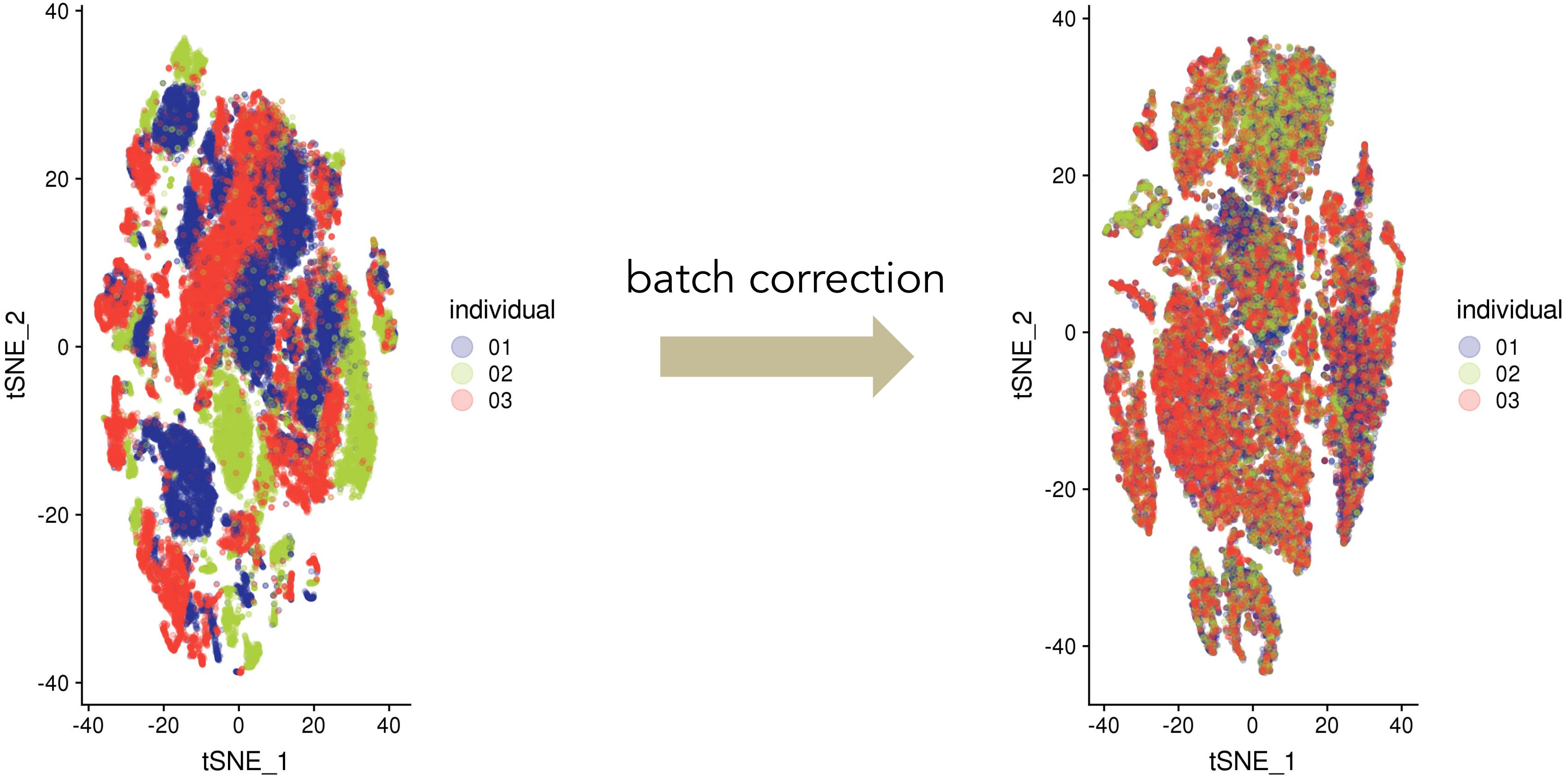
Chen X. et *al.* (2018) *Annual Review of Biomedical Data Science* Vol. 1:29-51

A few basic approaches to batch correction

- Down sampling of sequencing reads
- Normalization
- Using variable genes common to multiple samples
- Removing genes correlated with batch
- Regression of residuals with technical covariates
  - batch id
  - number of UMI per cell
  - number of genes per cell
  - % mitochondrial reads
- ComBat (developed for microarray experiments)

Batch correction and data modality integration

Batch effects often arise when patient samples are analyzed together



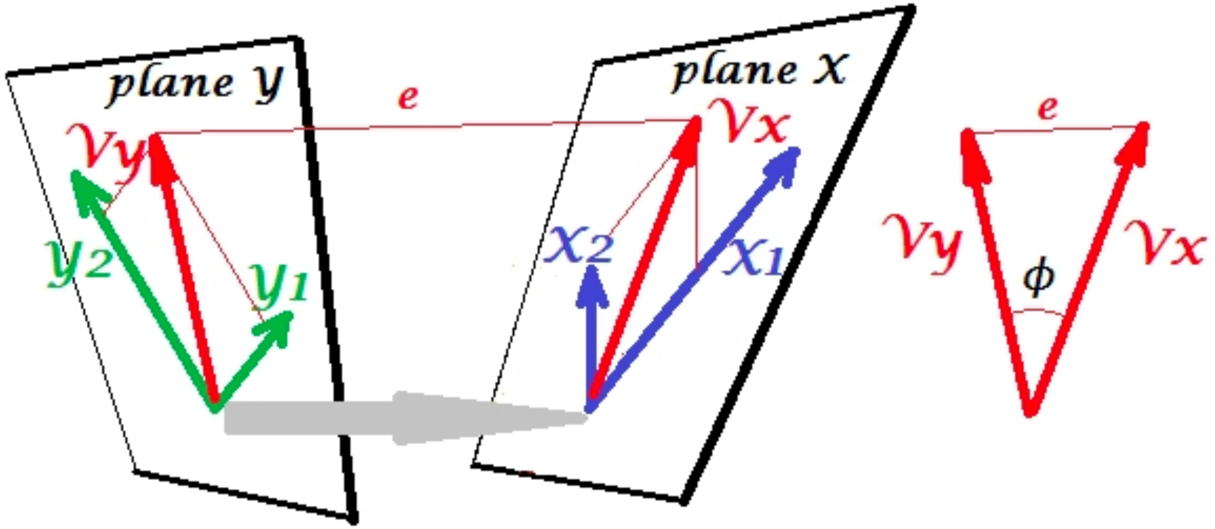
Batch correction and data modality integration

- Seurat v3
- LIGER (Linked Inference of Genomic Experimental Relationships)
- Conos (Clustering on Network of Samples)

IMPORTANT: We typically need to use the raw gene expression data for downstream analyses (e.g. differential expression) and not the corrected gene expression matrices after the batch correction. Why?

Canonical Correlation Analysis (CCA)

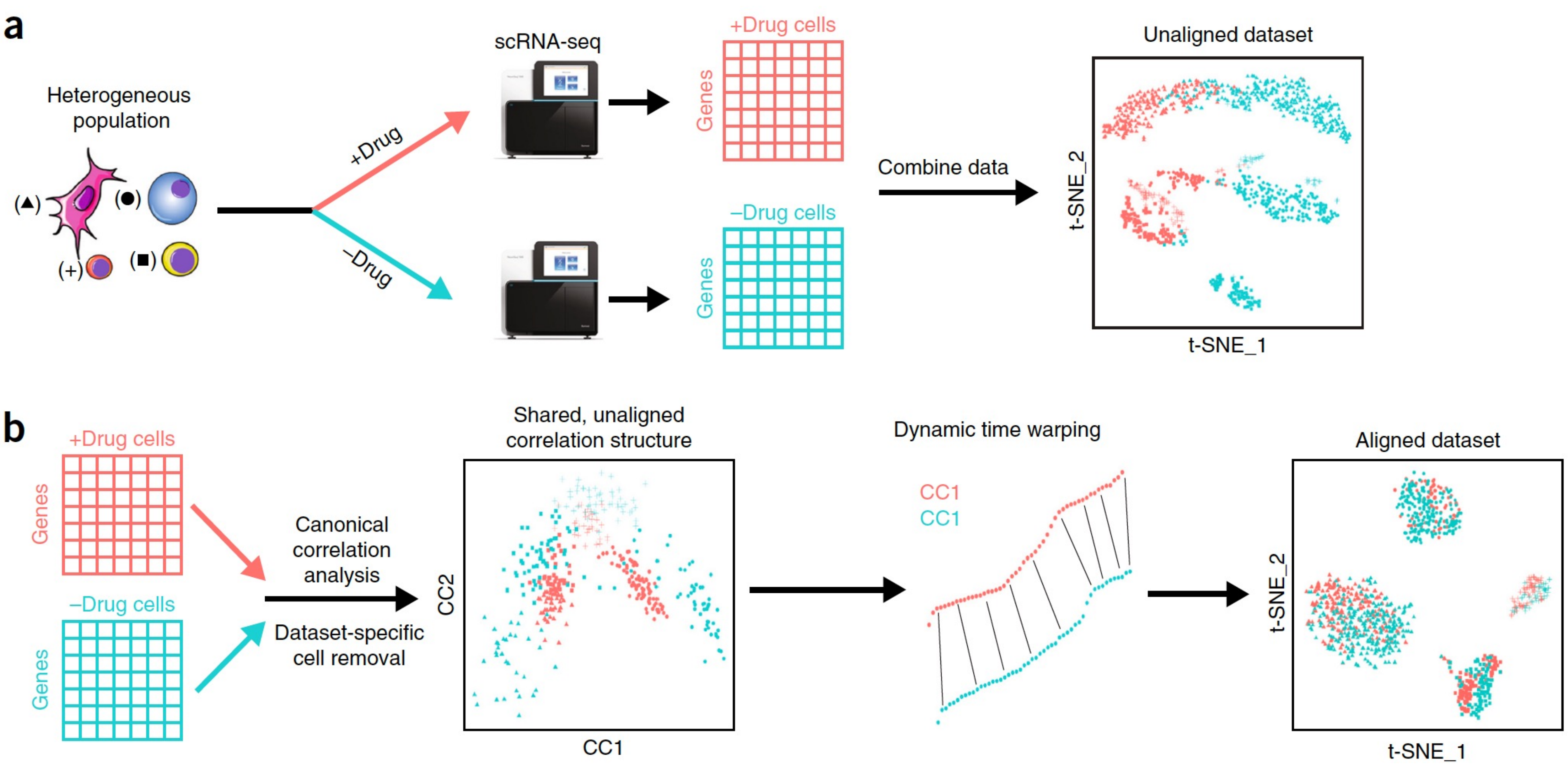
- CCA finds the linear combinations of variables across two datasets that are maximally correlated with one another.
  - The first pair of canonical variables maximizes the correlation across datasets.
  - The second pair of canonical variables maximizes the correlation subject to the constraint of not being correlated with the first pair, and so on.



- Goals of CCA
  - Similar to Principal Components Analysis (PCA)
  - Dimensional reduction: explain covariation between datasets with a small number of linear combinations of variables

https://www.mathworks.com
https://github.com/mhaghighat/ccaFuse

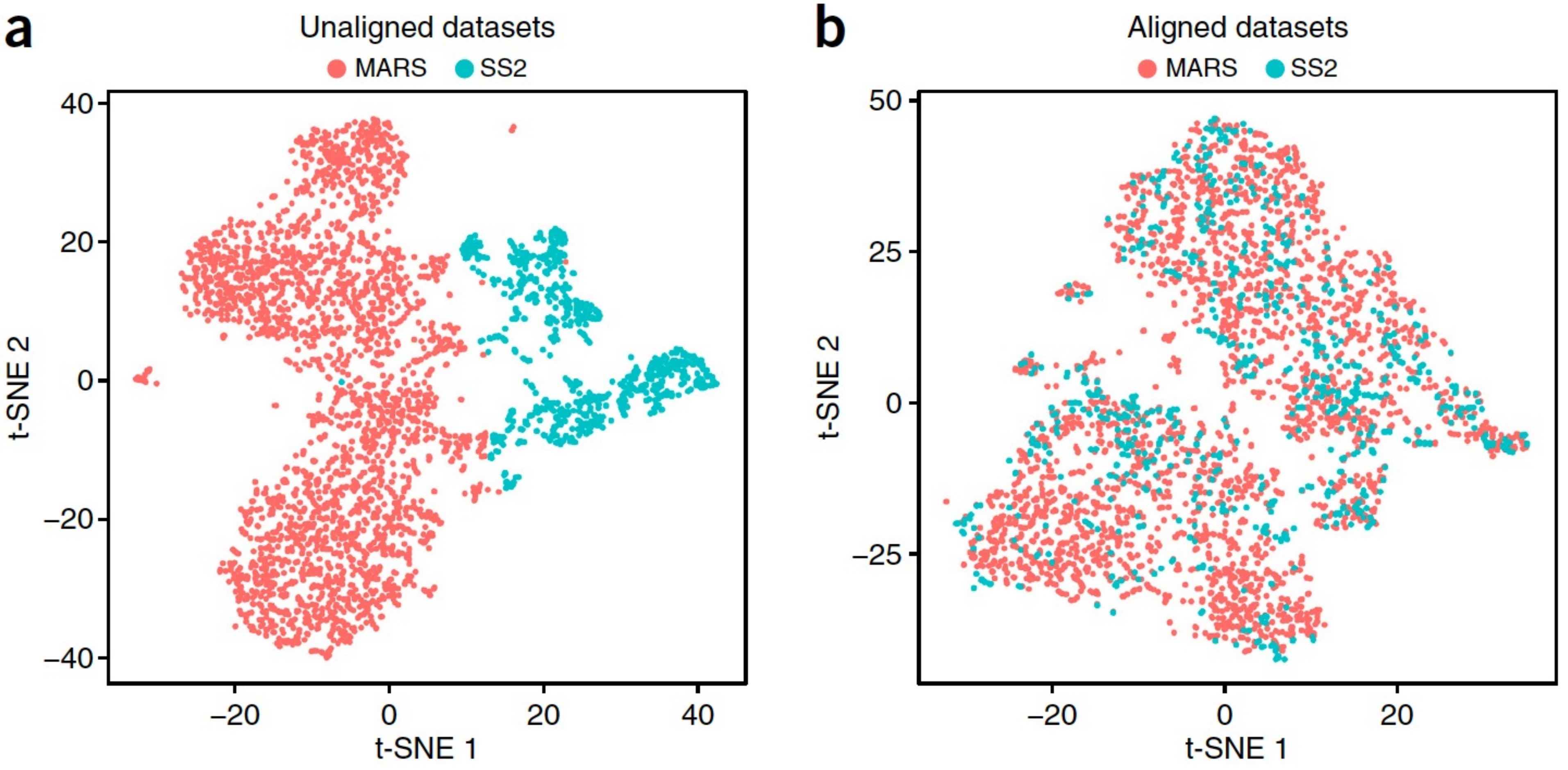
Canonical Correlation Analysis (CCA)



"Effectively, we treat the data sets as multiple measurements of a gene–gene covariance structure, and search for patterns that are common to the data sets."

Butler, A, et al. *Nature Biotechnology* 36.5 (2018): 411.

Canonical Correlation Analysis (CCA)



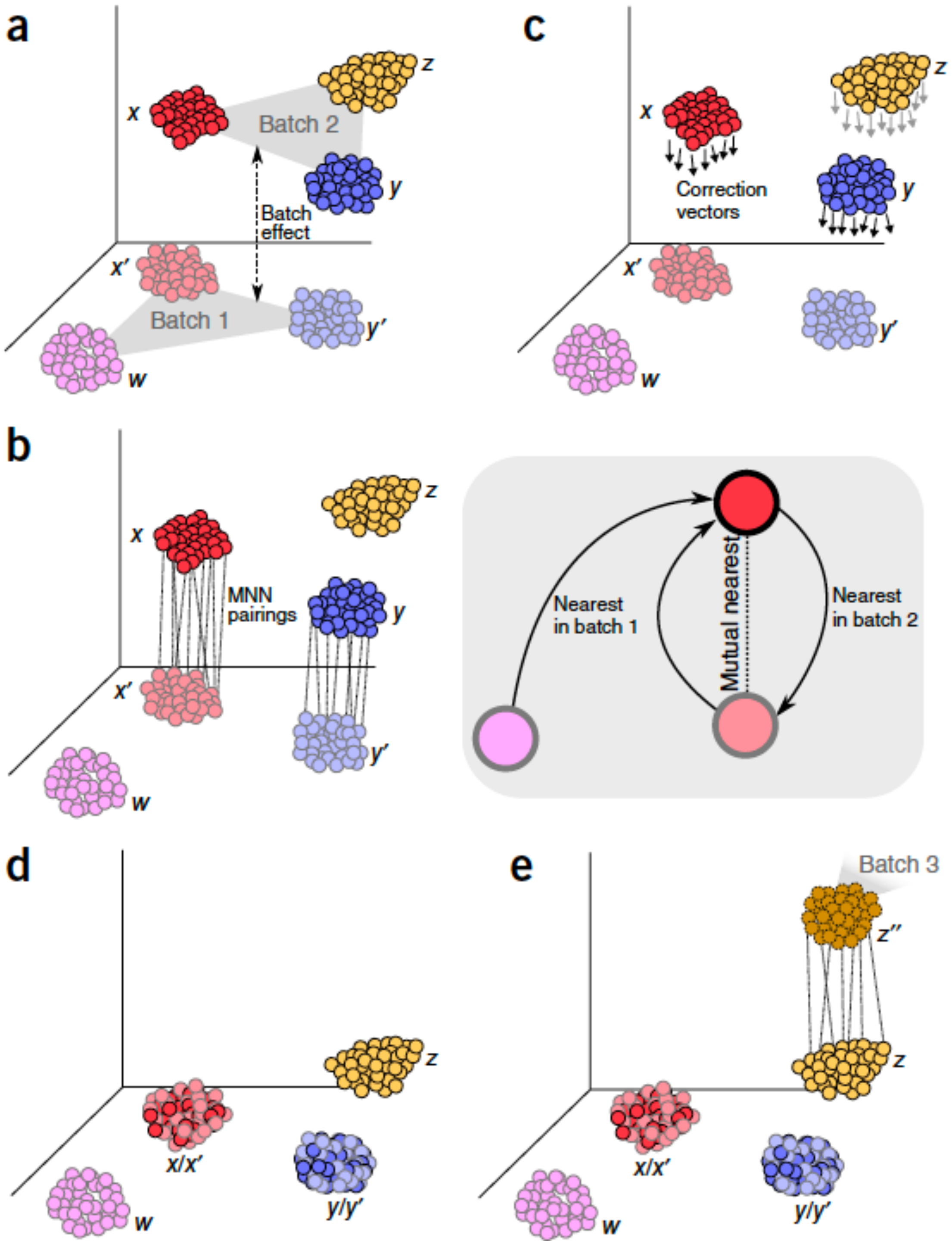
Assessing the performance of batch correction

"For every cell, we calculate how many of its k nearest-neighbors belong to the same data set and average this over all cells. If the data sets are well-aligned, we would expect that each cells' nearest neighbors would be evenly shared across all data sets."

Butler, A, et al. *Nature Biotechnology* 36.5 (2018): 411.

Mutual Nearest Neighbors

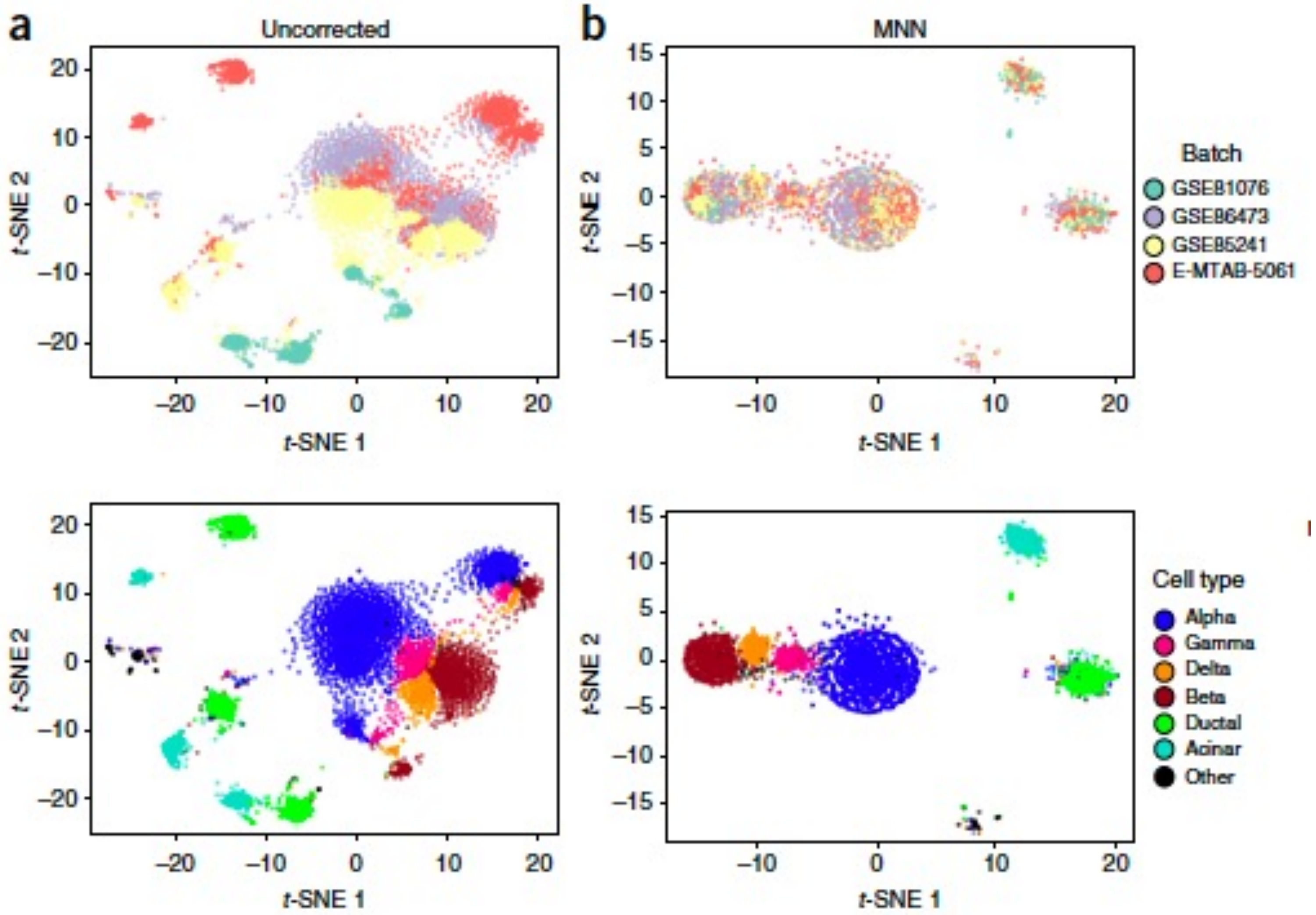
“If a pair of cells from each batch is contained in each other’s set of nearest neighbors, those cells are considered to be mutual nearest neighbors. We interpret these pairs as containing cells that belong to the same cell type or state despite being generated in different batches. Thus, any systematic differences in expression level between cells in MNN pairs should represent the batch effect.”



Haghverdi, L, et al. *Nature biotechnology* 36.5 (2018): 421.

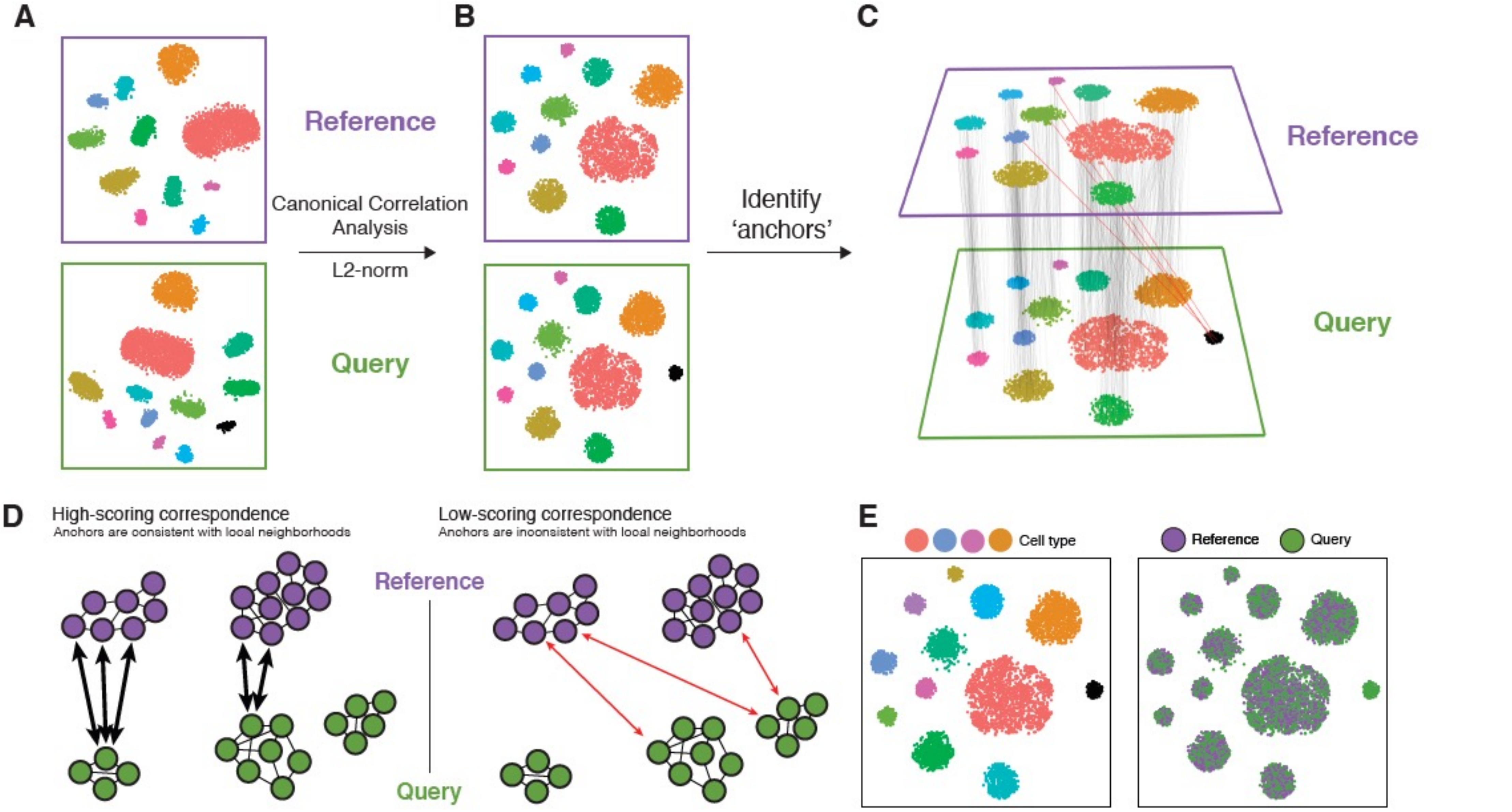
Mutual Nearest Neighbors

4 pancreas datasets



Haghverdi, L, et al. *Nature biotechnology* 36.5 (2018): 421.

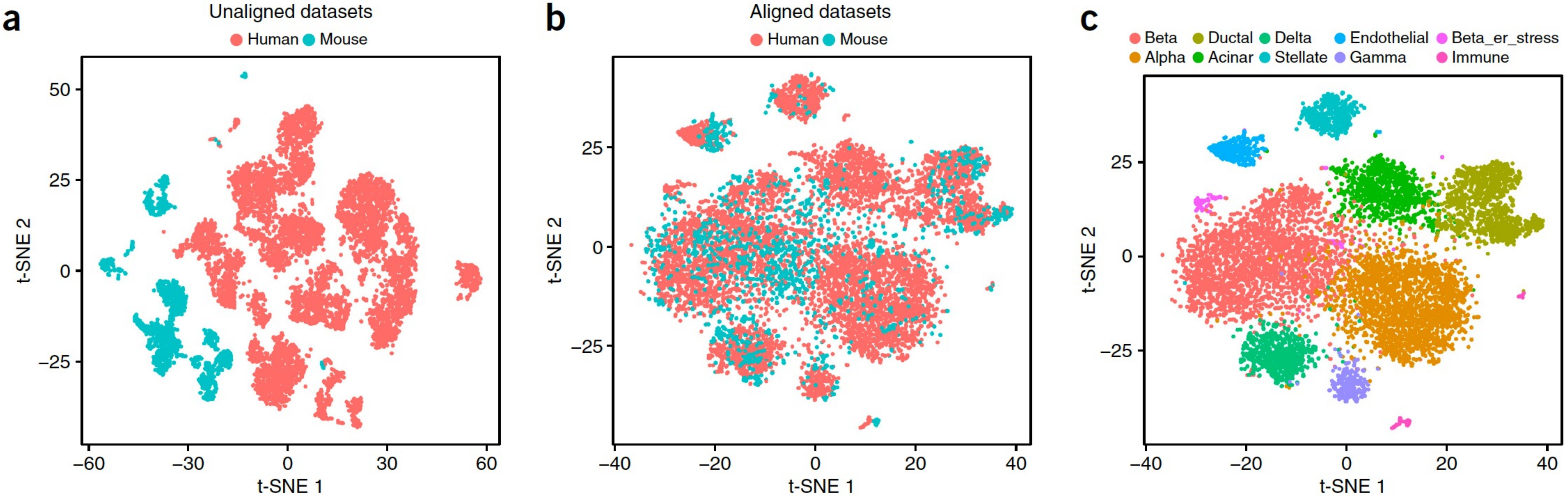
Combining CCA and Mutual Nearest Neighbors (Seurat v3)


A
B
C
Reference
Query
Canonical Correlation Analysis
L2-norm
Identify 'anchors'
Reference
Query
D
High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query
E
Cell type
Reference
Query

Stuart et al. Cell (2019) 177(7):1888-1902.

Combining CCA and Mutual Nearest Neighbors (Seurat v3)

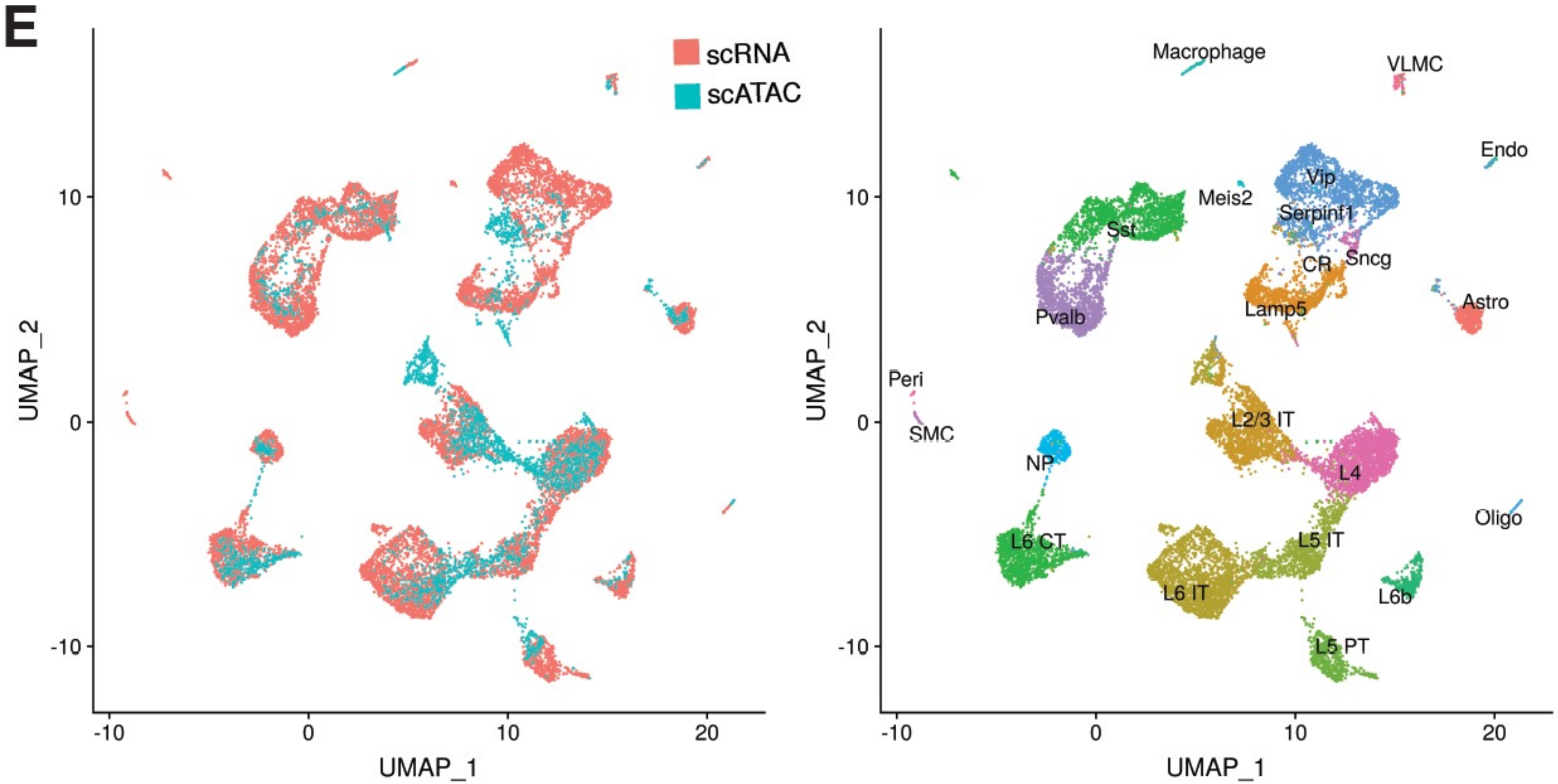
Human and mouse pancreas datasets



Stuart et al. Cell (2019) 177(7):1888-1902.

Classifying nuclei from a single cell ATAC-seq experiment using single cell RNA-seq data as a reference

Integrating data modalities
14,249 cells from scRNA-seq and 2,548 cells from scATAC-seq



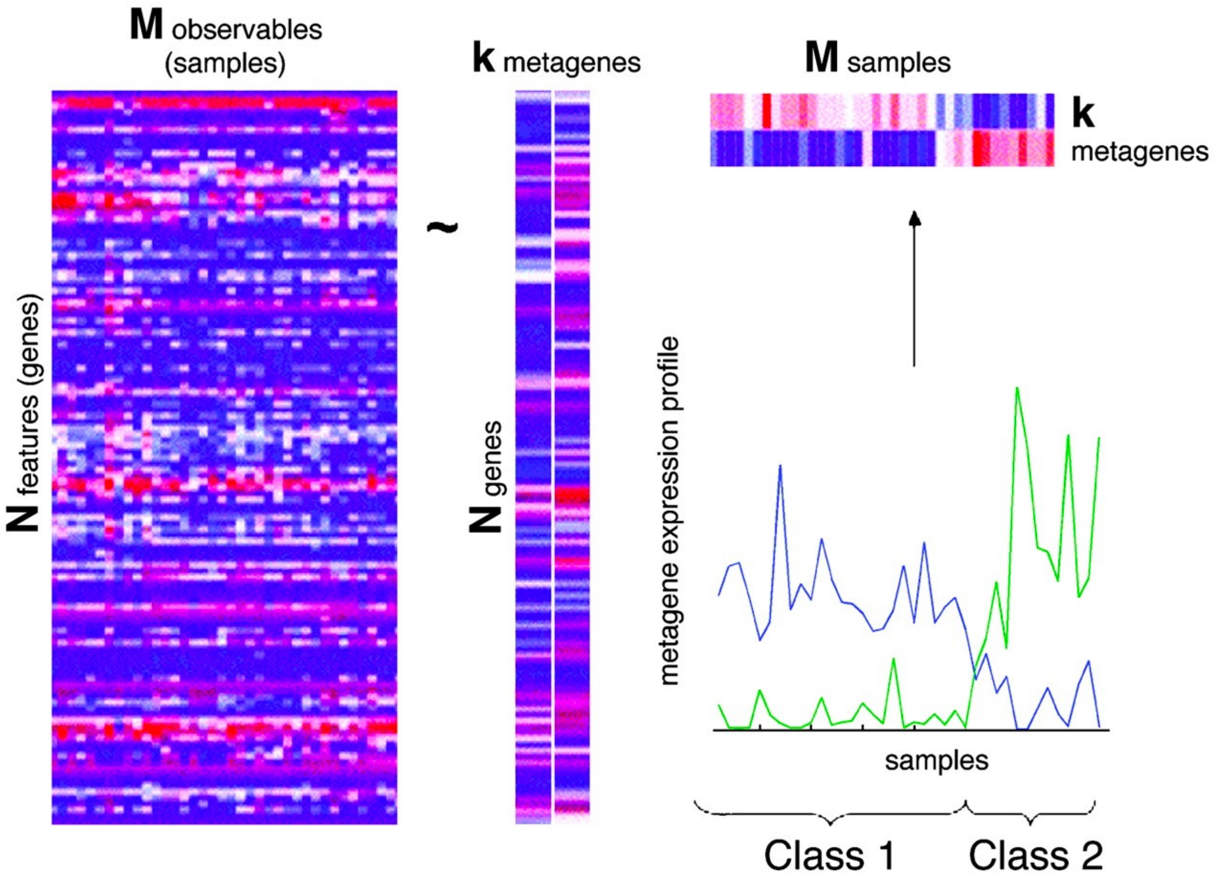
Stuart et al. Cell (2019) 177(7):1888-1902.

Batch correction and data modality integration

Nonnegative Matrix Factorization (NMF)

A (rank M) ~ W H (rank k=2)

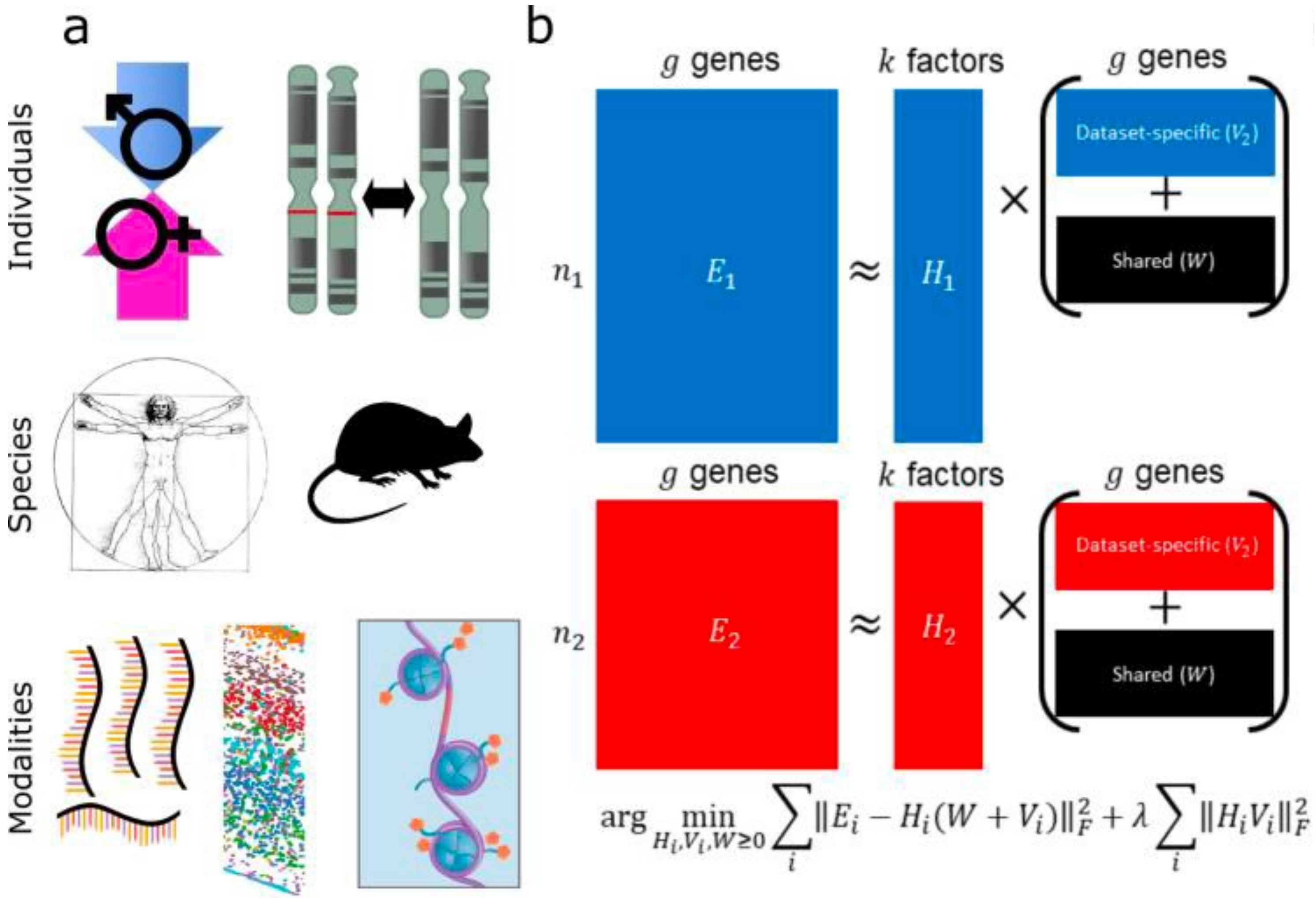
“Our goal is to find a small number of metagenes, each defined as a positive linear combination of the *N* genes. We can then approximate the gene expression pattern of samples as positive linear combinations of these metagenes. Mathematically, this corresponds to factoring matrix *A* into two matrices with positive entries, *A* ~ *WH*.”



Brunet J. et *al.* (2004) *PNAS*. 101:12, 4164-4169.

Batch correction and data modality integration using LIGER

LIGER implements non-negative matrix factorization



$$\arg\min_{H_i,V_i,W\geq 0}\sum_i \|E_i - H_i(W+V_i)\|_F^2 + \lambda\sum_i \|H_iV_i\|_F^2$$

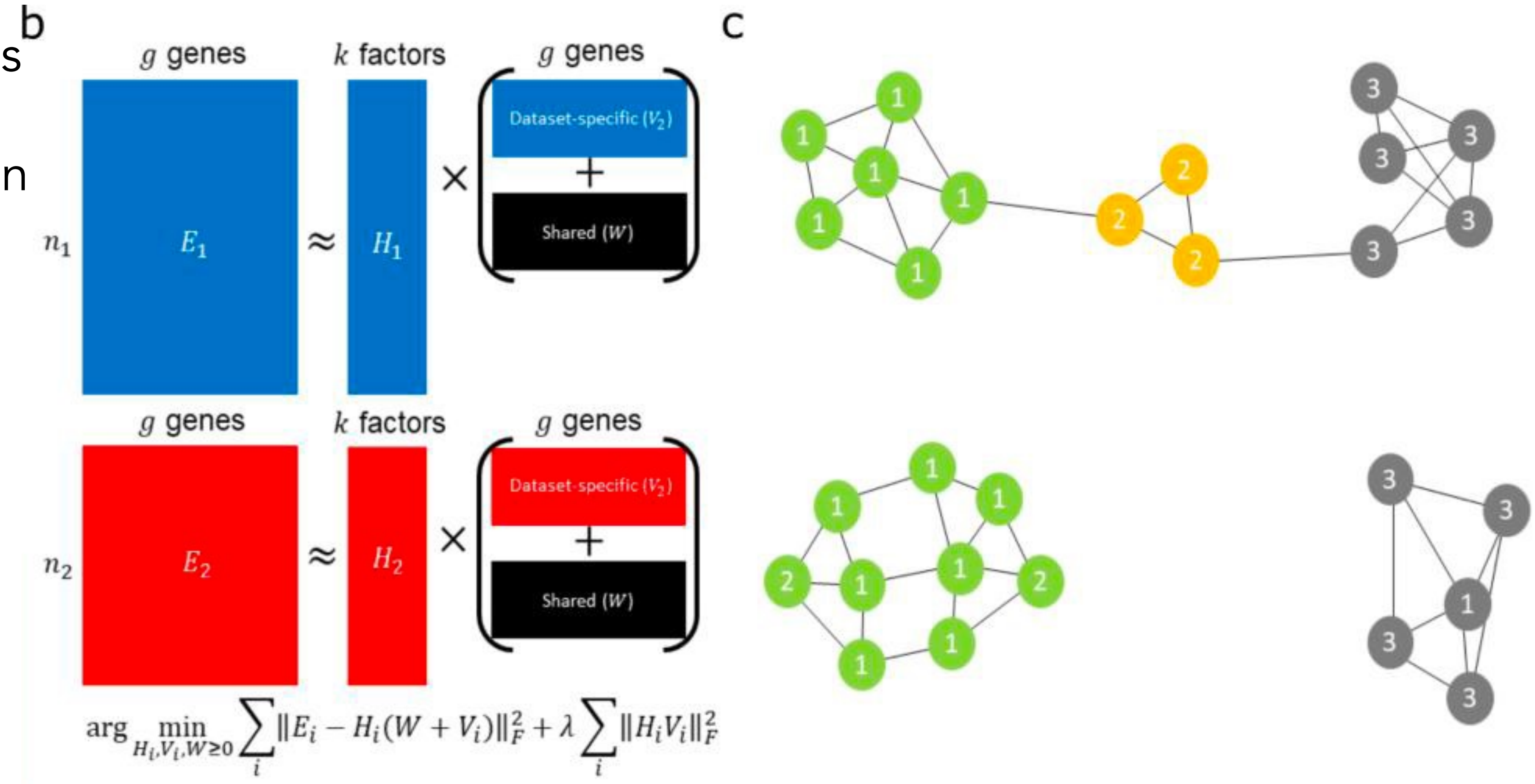
Heterogeneous single cell datasets

Integrative nonnegative matrix factorization with dataset-specific factors

Welch, et al. Cell (2019) 177(7): 1873-1887.

Batch correction and data modality integration using LIGER

LIGER implements non-negative matrix factorization



$$\arg\min_{H_i, V_i, W \geq 0} \sum_i \|E_i - H_i(W + V_i)\|_F^2 + \lambda \sum_i \|H_i V_i\|_F^2$$

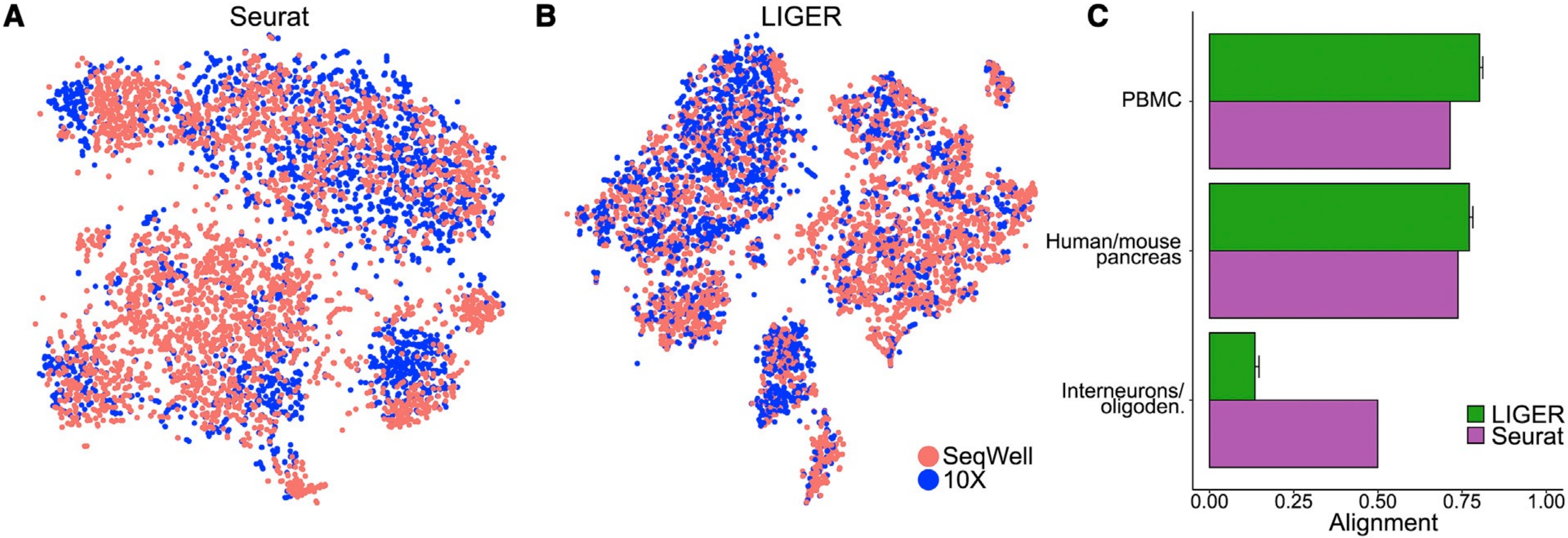
Integrative nonnegative matrix factorization with dataset-specific factors

Joint clustering using shared factor neighborhood graph

Welch, et al. Cell (2019) 177(7): 1873-1887.

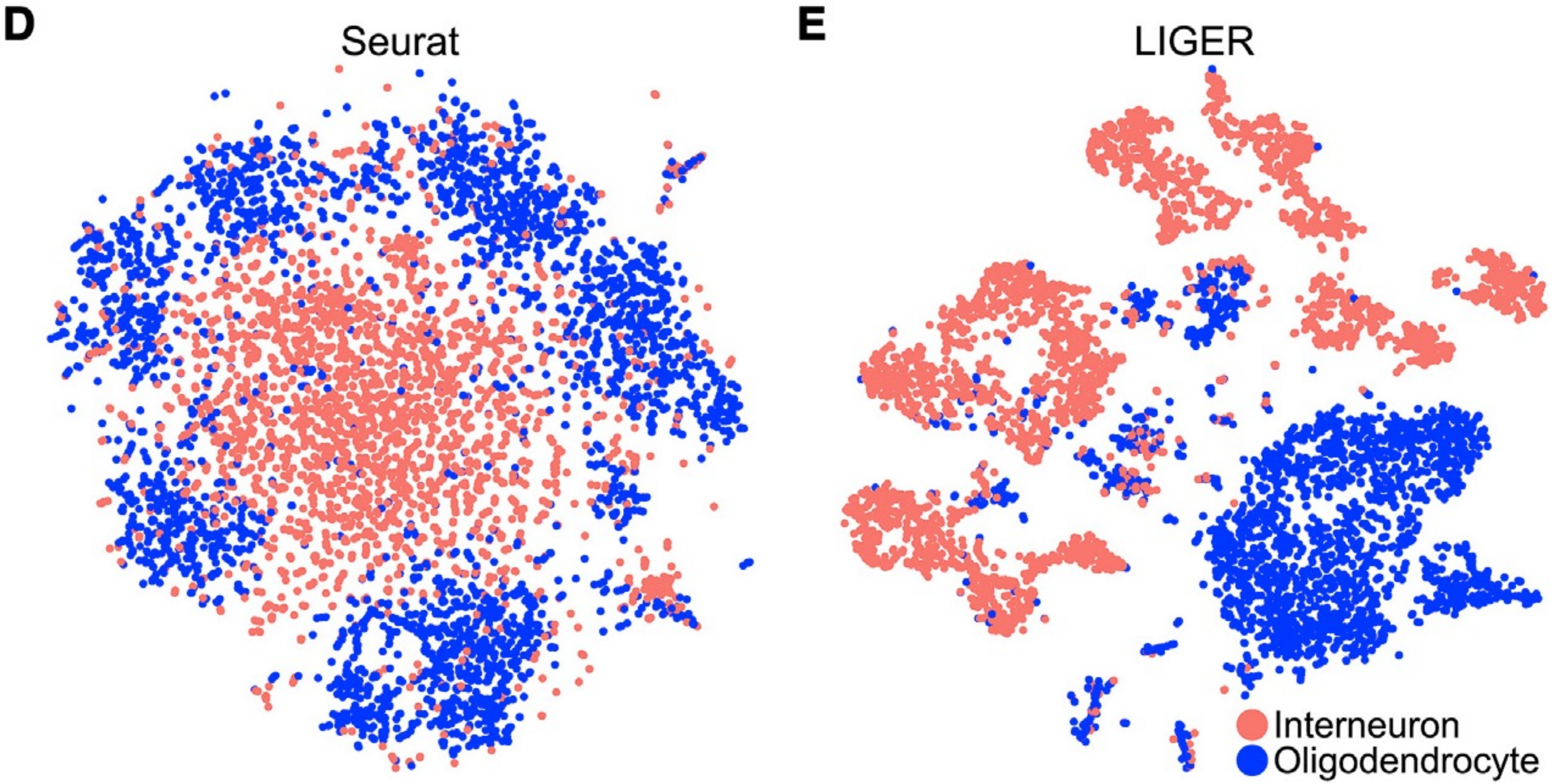
Integrating blood cell datasets using LIGER

Higher alignment score = better data integration



Welch, et al. Cell (2019) 177(7): 1873-1887.

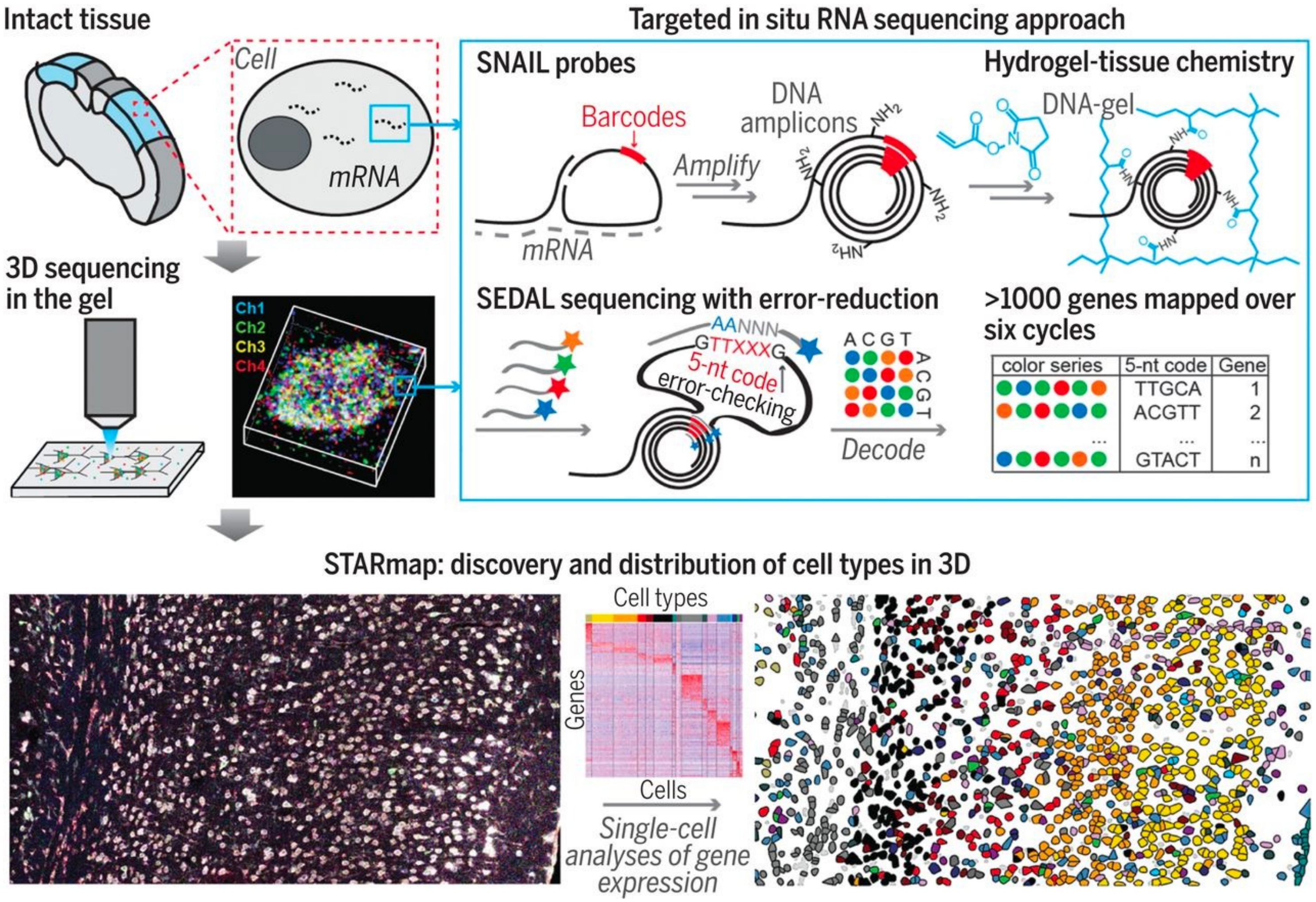
Integrating blood cell datasets using LIGER



Ideally, divergent cell types should not cluster together after batch correction.

Welch, et al. Cell (2019) 177(7): 1873-1887.

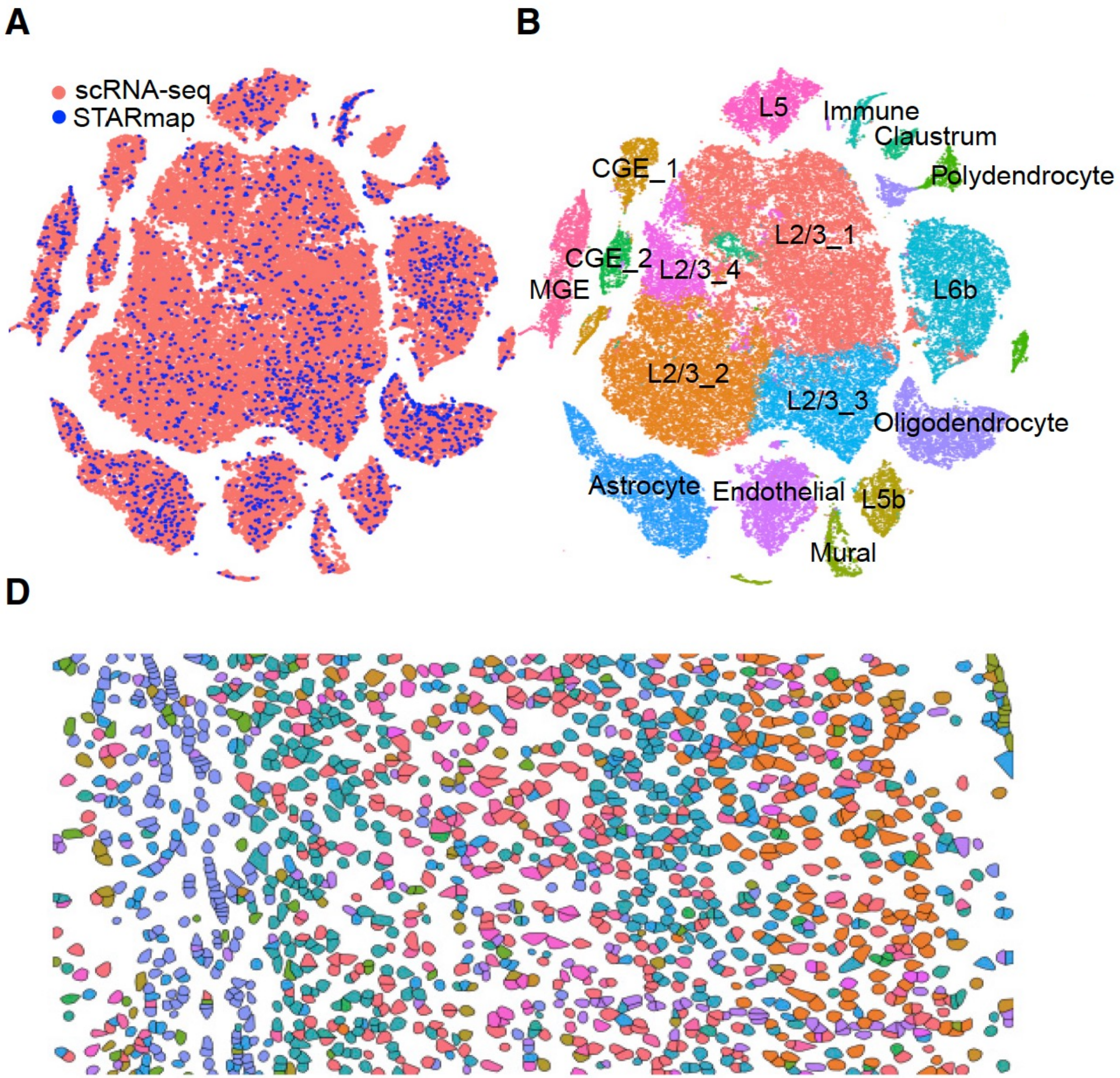
In situ spatial transcriptomic data in mouse frontal cortex
STARmap


Intact tissue
Cell
mRNA
Targeted in situ RNA sequencing approach
SNAIL probes
Barcodes
mRNA
Amplify
DNA amplicons
NH_2
Hydrogel-tissue chemistry
DNA-gel
3D sequencing in the gel
Ch1
Ch2
Ch3
Ch4
SEDAL sequencing with error-reduction
AANNN
GTTXXXG
5-nt code
error-checking
A C G T
A C G T
Decode
>1000 genes mapped over six cycles
STARmap: discovery and distribution of cell types in 3D
Cell types
Genes
Cells
Single-cell analyses of gene expression

color series	5-nt code	Gene
🟢🔵🟢🔴🟢🟠	TTGCA	1
🟠🟢🔴🟢🔵🟢	ACGTT	2
...	...	...
🔵🟢🔴🟢🟠🟢	GTACT	n

Wang, Xiao, et al. *Science* (2018): eaat5691.

Using LIGER to integrate single-cell transcriptomic and *in situ* spatial transcriptomic data

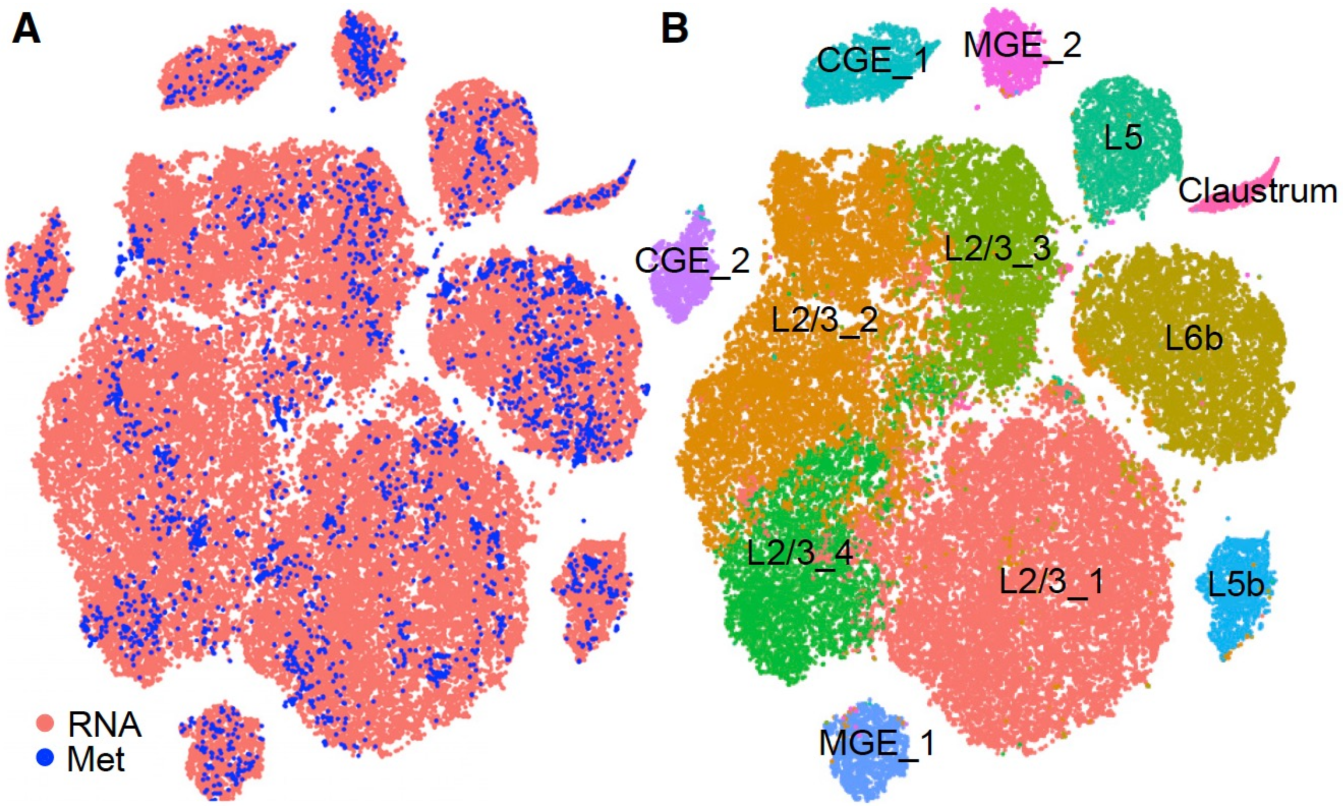


71,000 cells from scRNA-seq
2,500 cells from STARmap

What are advantages of integrated analysis of these 2 datasets?

Welch, et al. Cell (2019) 177(7): 1873-1887.

Using LIGER to integrate single-cell transcriptomic and single-cell DNA methylation data



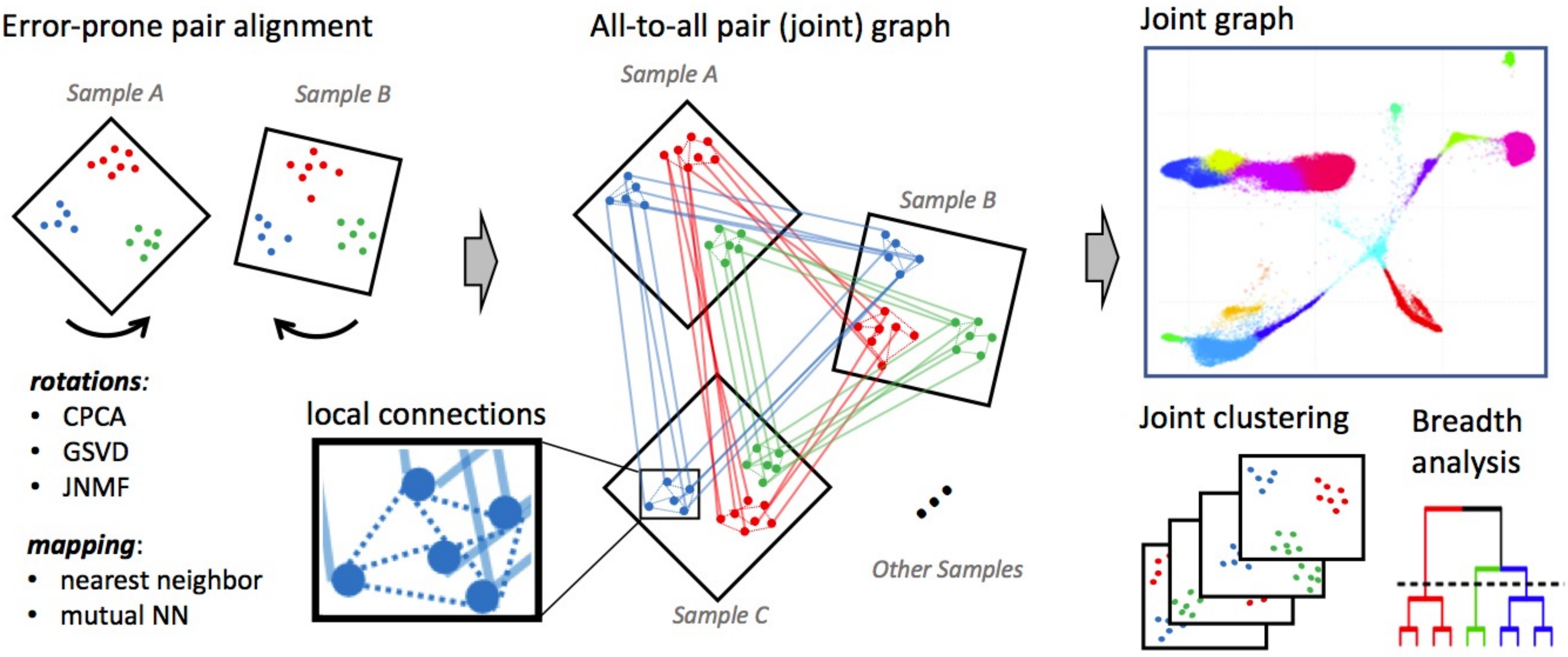
56,000 cells from scRNA-seq
3,000 cells from DNA methylation

"We reasoned that, because gene body methylation is generally anticorrelated with gene expression, reversing the direction of the methylation signal would allow joint analysis."

Welch, et al. Cell (2019) 177(7): 1873-1887.

Batch correction and data modality integration using Conos

Conos (Clustering on Network of Samples)



Error-prone pair alignment

rotations:

- CPCA
- GSVD
- JNMF

mapping:

- nearest neighbor
- mutual NN

local connections

All-to-all pair (joint) graph

Joint graph

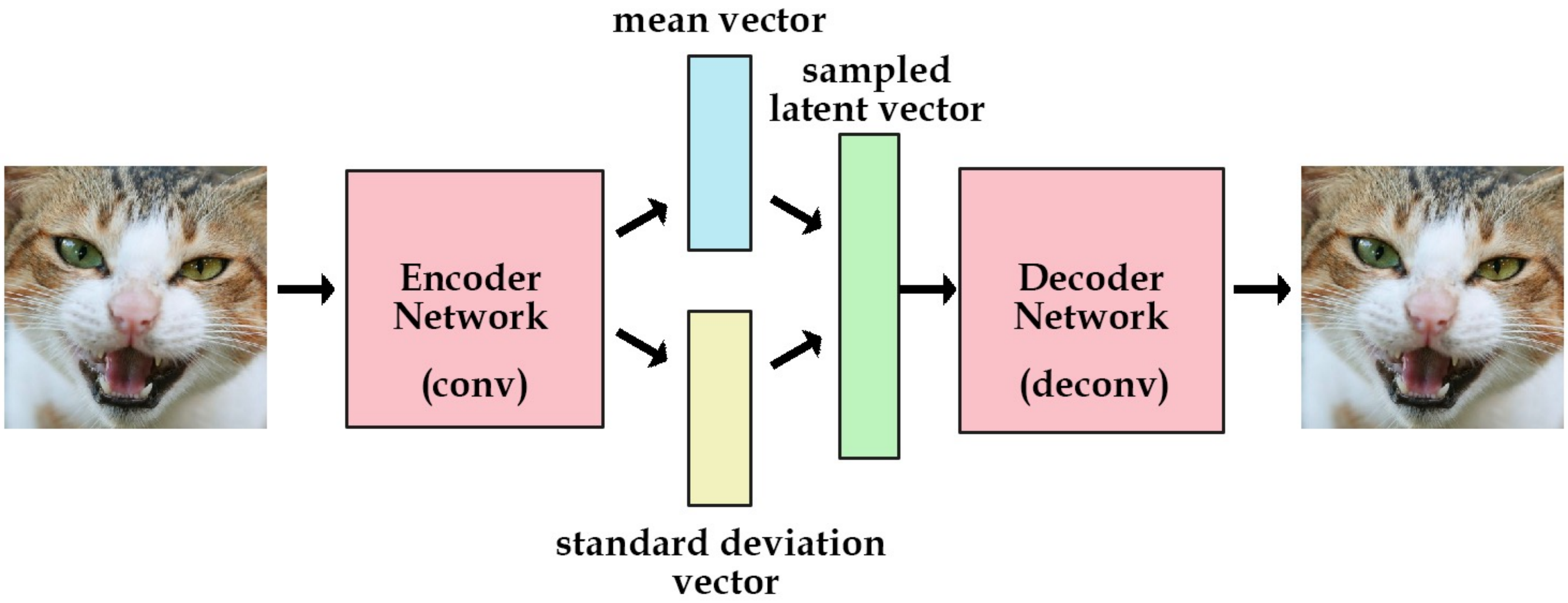
Joint clustering

Breadth analysis

Barkas, et al. Nature Methods (2019) 16, 695–698.

Batch correction with deep learning approaches

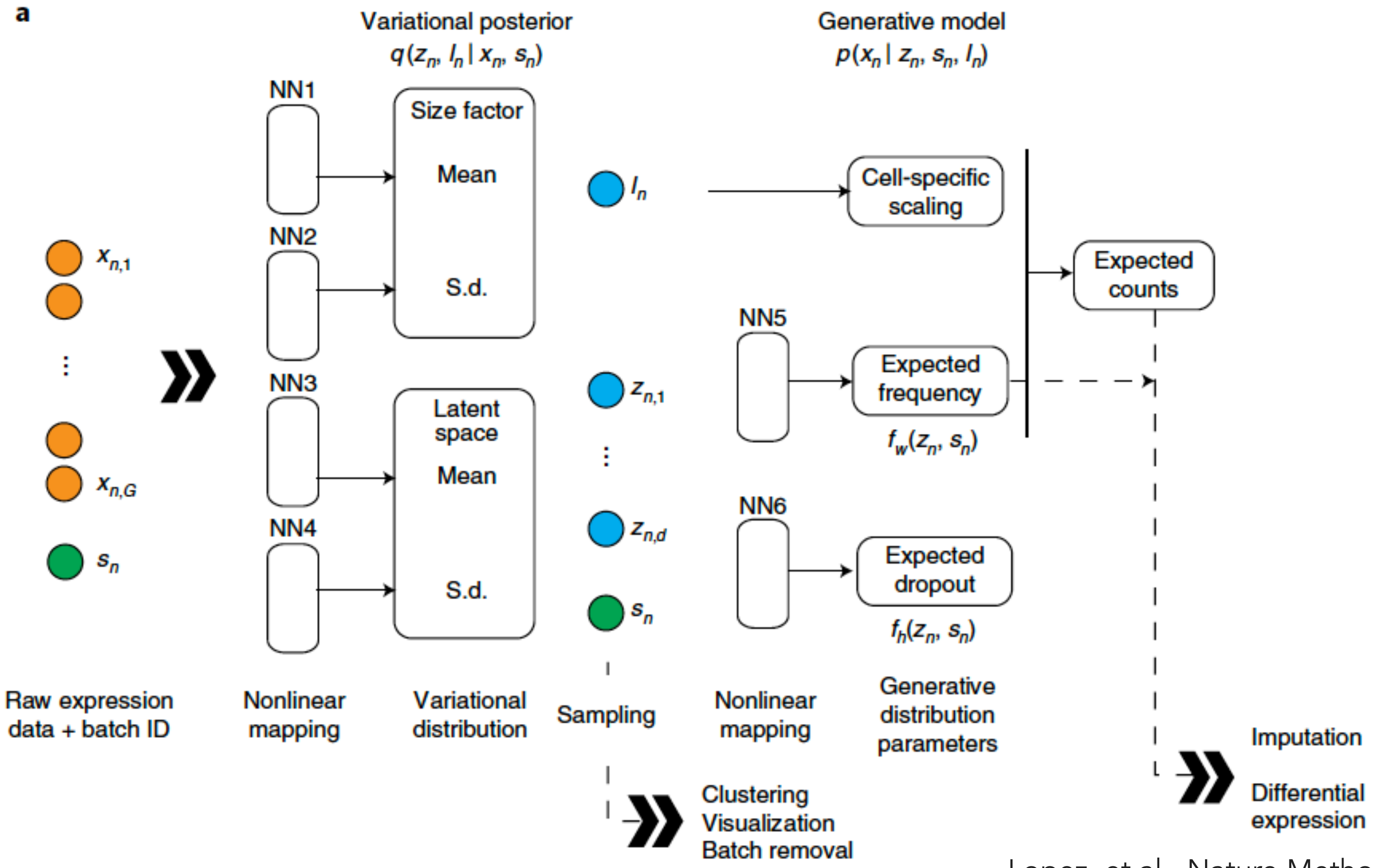
Autoencoders and variational autoencoders are popular frameworks for embedding single cell genomics datasets.



http://kvfrans.com/variational-autoencoders-explained/

Batch correction with deep learning approaches

single-cell variation inference (scVI)

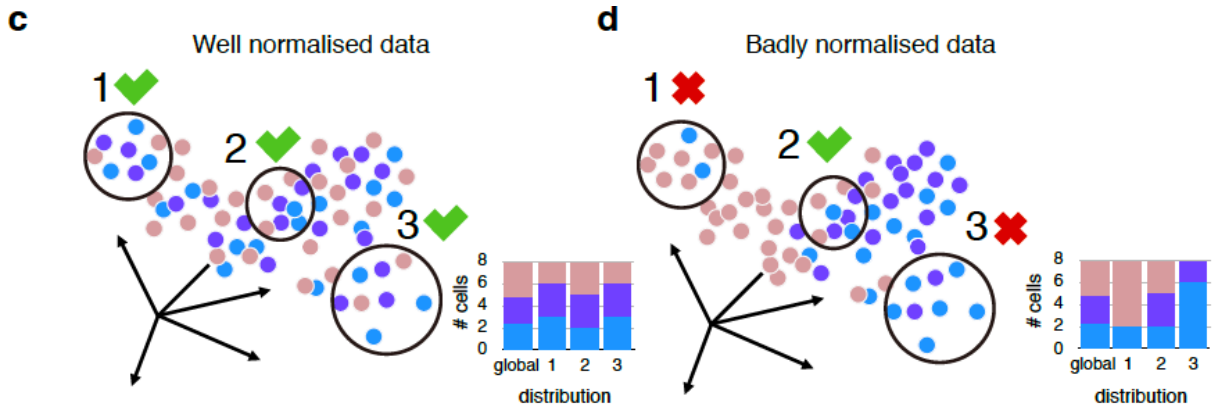


Lopez, et al. Nature Methods (2018) 15, 1053-1058.

Methods to assess performance of batch correction

- Entropy of batch mixing

- kBET - k-nearest neighbor batch effect test



- Silhouette coefficient
  - cells of the same cell type are close together and far from other cells of a different type

- Adjusted rand index
  - do batch labels and cluster labels agree with one another?

- Biological significance

Common assumptions during batch correction

- At least one cell population is found shared in both datasets

- The batch effect and the biological differences do not overlap with one another (ie orthogonal)

- The magnitude and variation of the biological effect you care about is greater than that of the batch effect

Recent helpful articles on data integration methods

Argelaguet, Ricard, et al. "Computational principles and challenges in single-cell data integration." *Nature biotechnology* (2021).

- Review

Luecken, Malte D., et al. "Benchmarking atlas-level data integration in single-cell genomics." *Nature methods* (2022).

- Methods comparison

Tran, Hoa Thi Nhu, et al. "A benchmark of batch-effect correction methods for single-cell RNA sequencing data." *Genome biology* (2020).

- Methods comparison