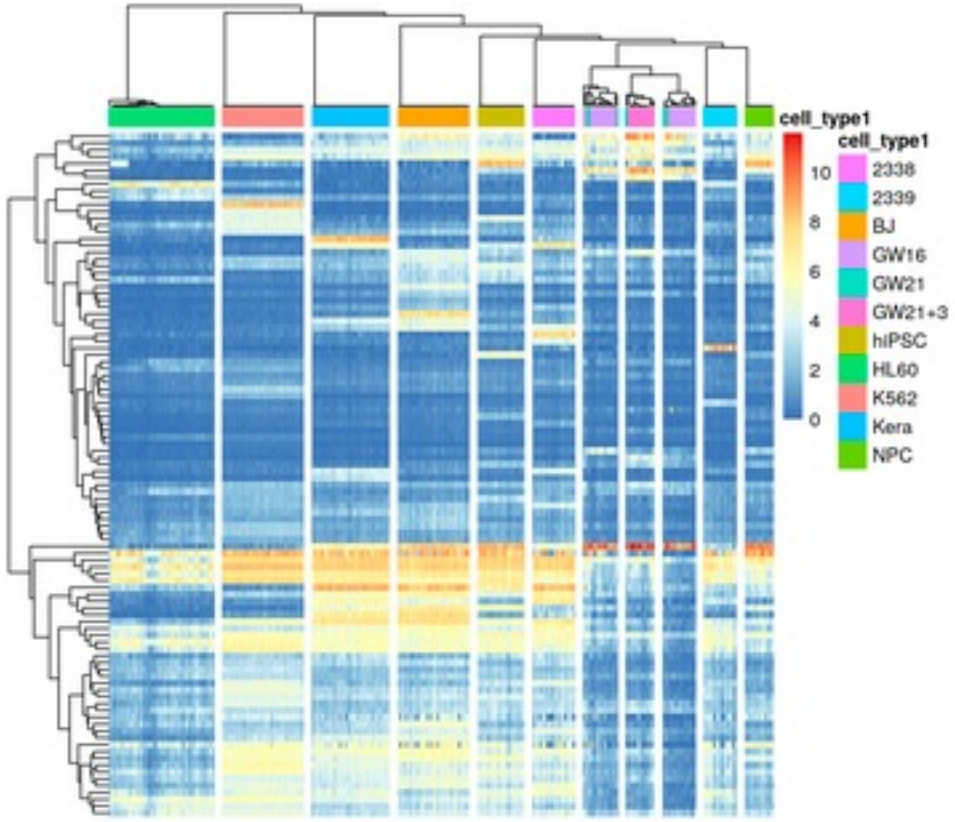
Analysis of Single Cell RNA-Seq Data: Data integration and batch effect correction

Orr Ashenberg, Jacques Serizay, Fabricio Almeida-Silva

November 2026

Physalia

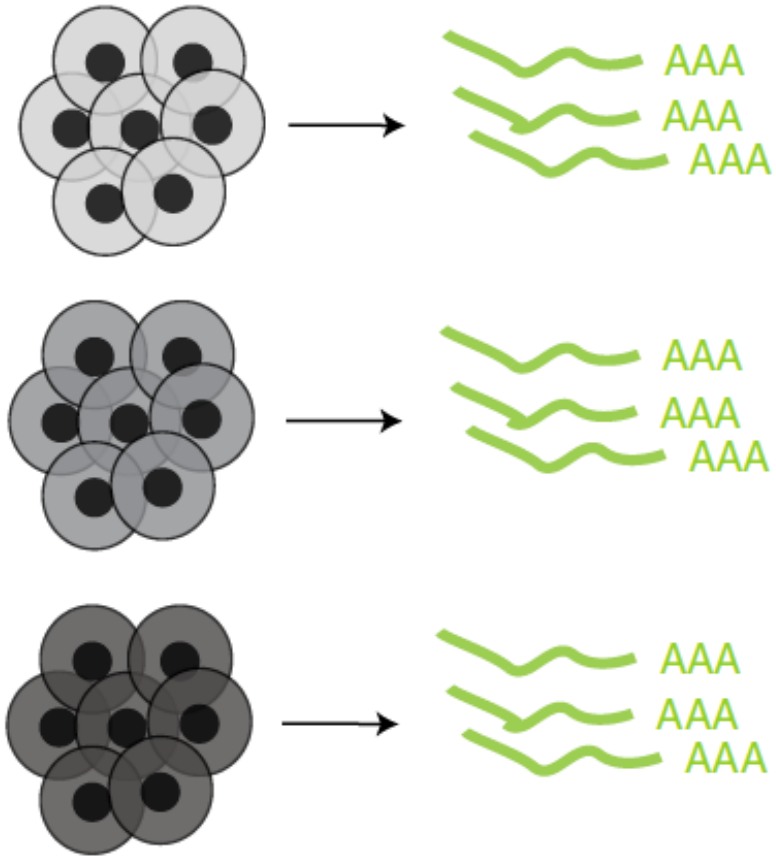
Courses

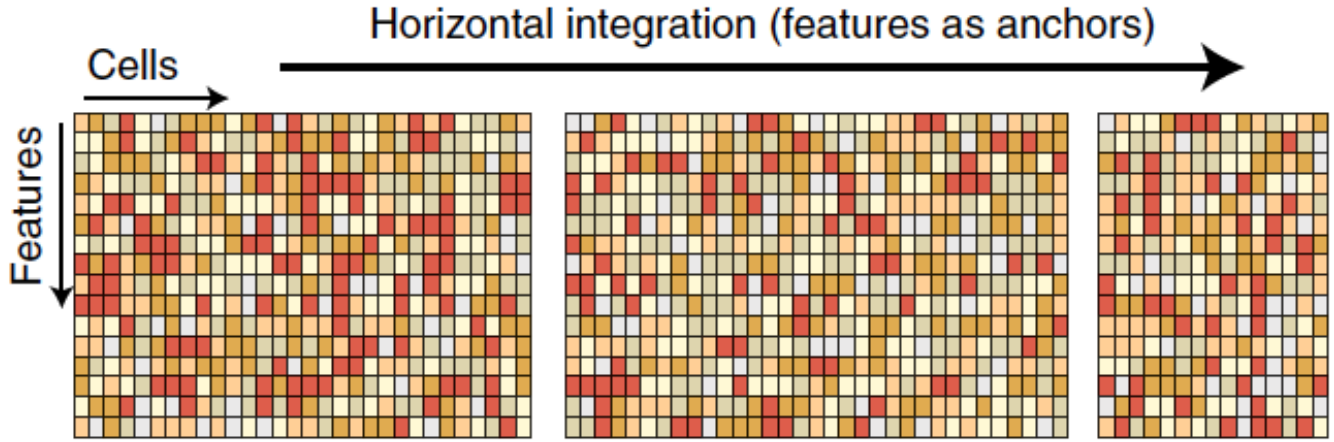


Outline: Batch effects and data integration

- Types of data integration.
- Where batch effects come from.
- Computational approaches to correct for batch effects and integrate data.
- Making biological comparisons after data integration.

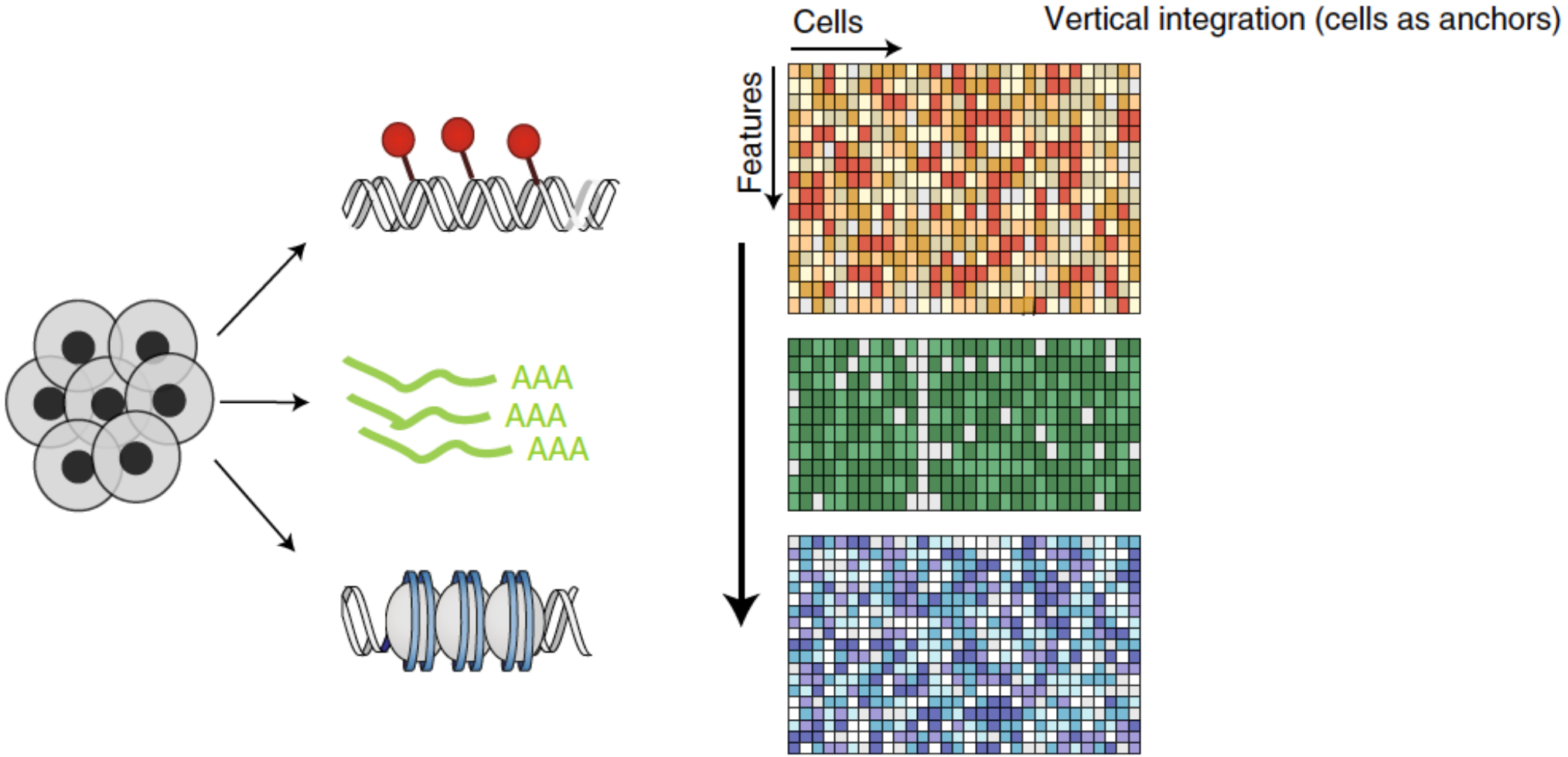
Data integration
Shared features across different batches of cells





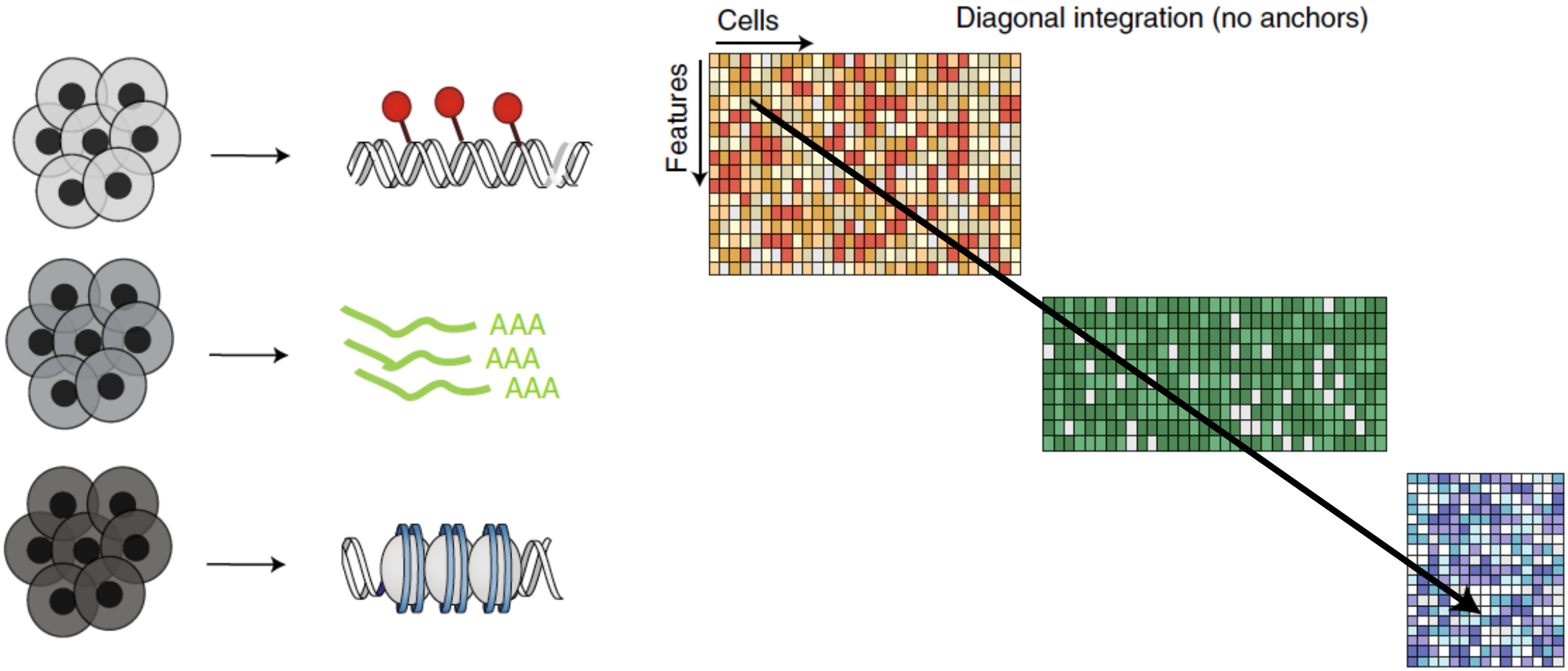
Argelaguet R. et al. (2022) *Nature Methods*.

Data integration
Different features measured from same cells



Argelaguet R. et al. (2022) *Nature Methods*.

Data integration
Different cell batches and features in each experiment

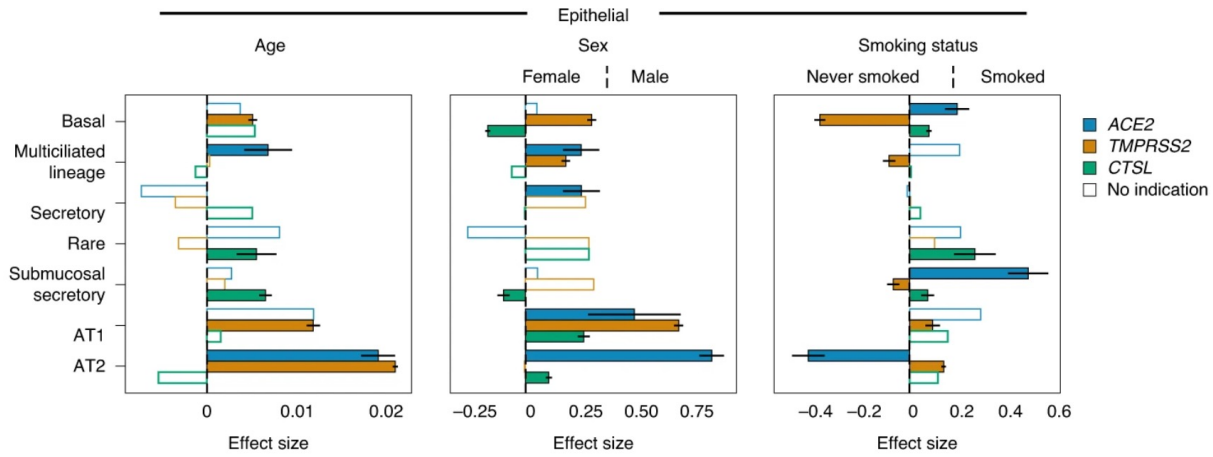


Efremova M. et al. (2020) *Nature Methods.* 17:11-20.

Why perform data integration and batch effect correction?

We often require large cohorts of individuals ($n > 10$) to make biological comparisons (case vs control, mutant vs wildtype etc…).

- As an example, *how do smoking status and age contribute to SARS-CoV-2 entry-related genes in lung cells*?



1,320,896 cells from 228 individuals

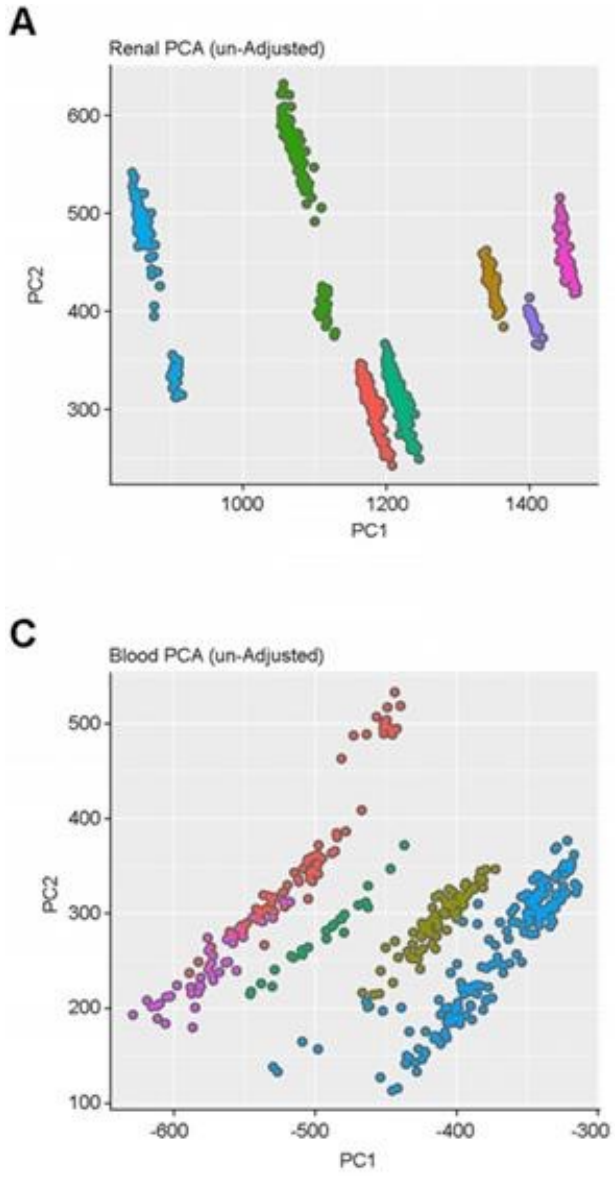
Challenges

- Technical batch effects confound biological variation of interest (e.g. site and time of collection, experimental protocol used…)
- Different sources of biological variation confound biological variation of interest (e.g. cell cycle, treatment status…)

Muus C. et al. (2021) *Nature Medicine.*

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

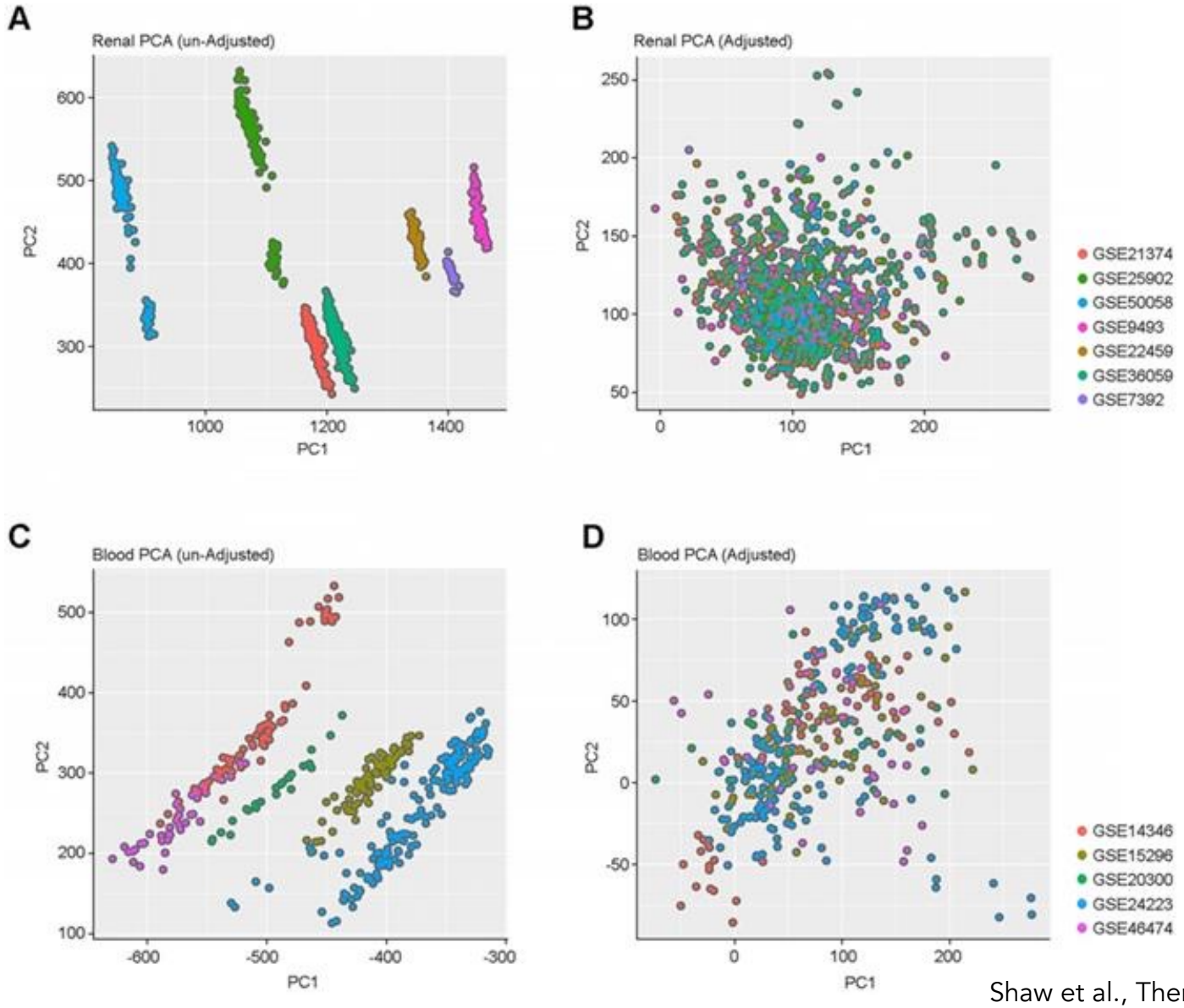
Correcting for batch effects allows us to **combine datasets and boost biological signal**, while reducing technical confounders



Shaw et al., Theranostics 2020

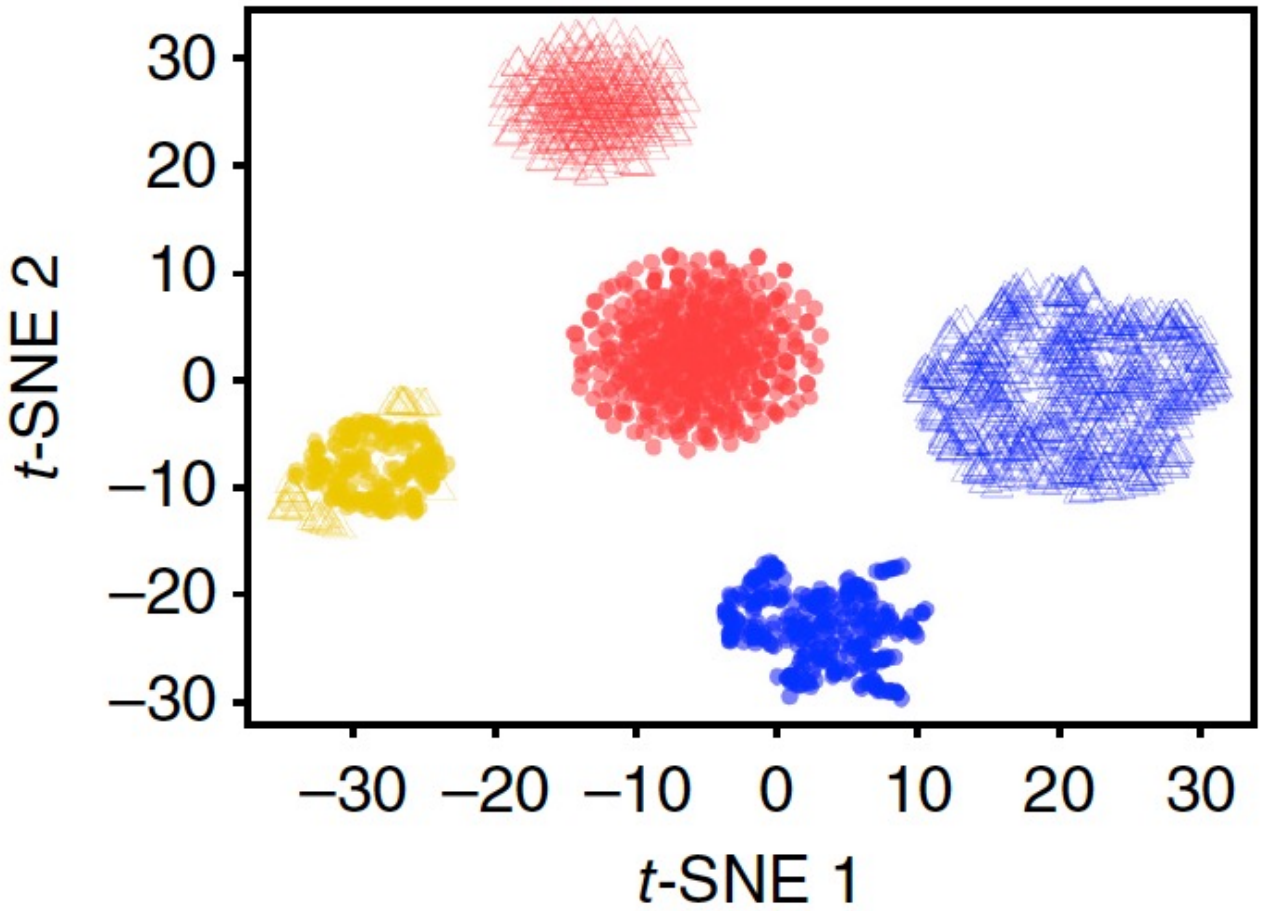
Distinguishing biological effects from technical, batch effects is a difficult problem

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



Shaw et al., Theranostics 2020

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

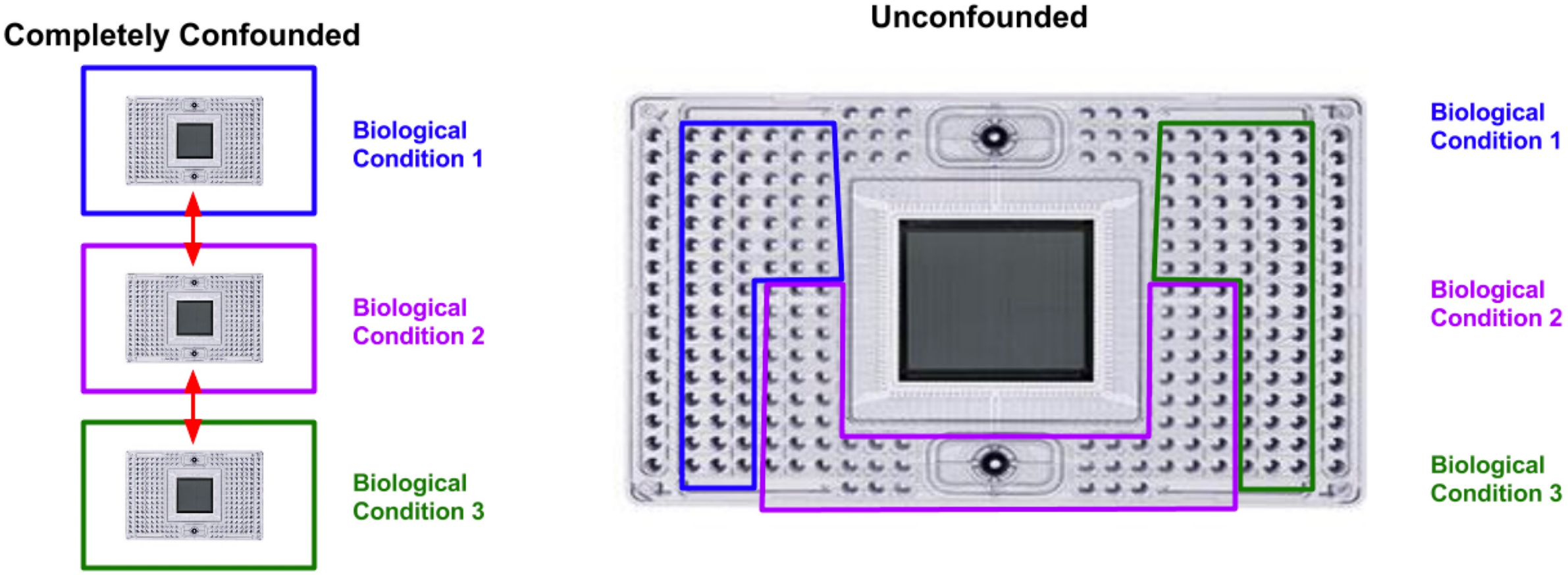


Cells are colored by cell type.

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

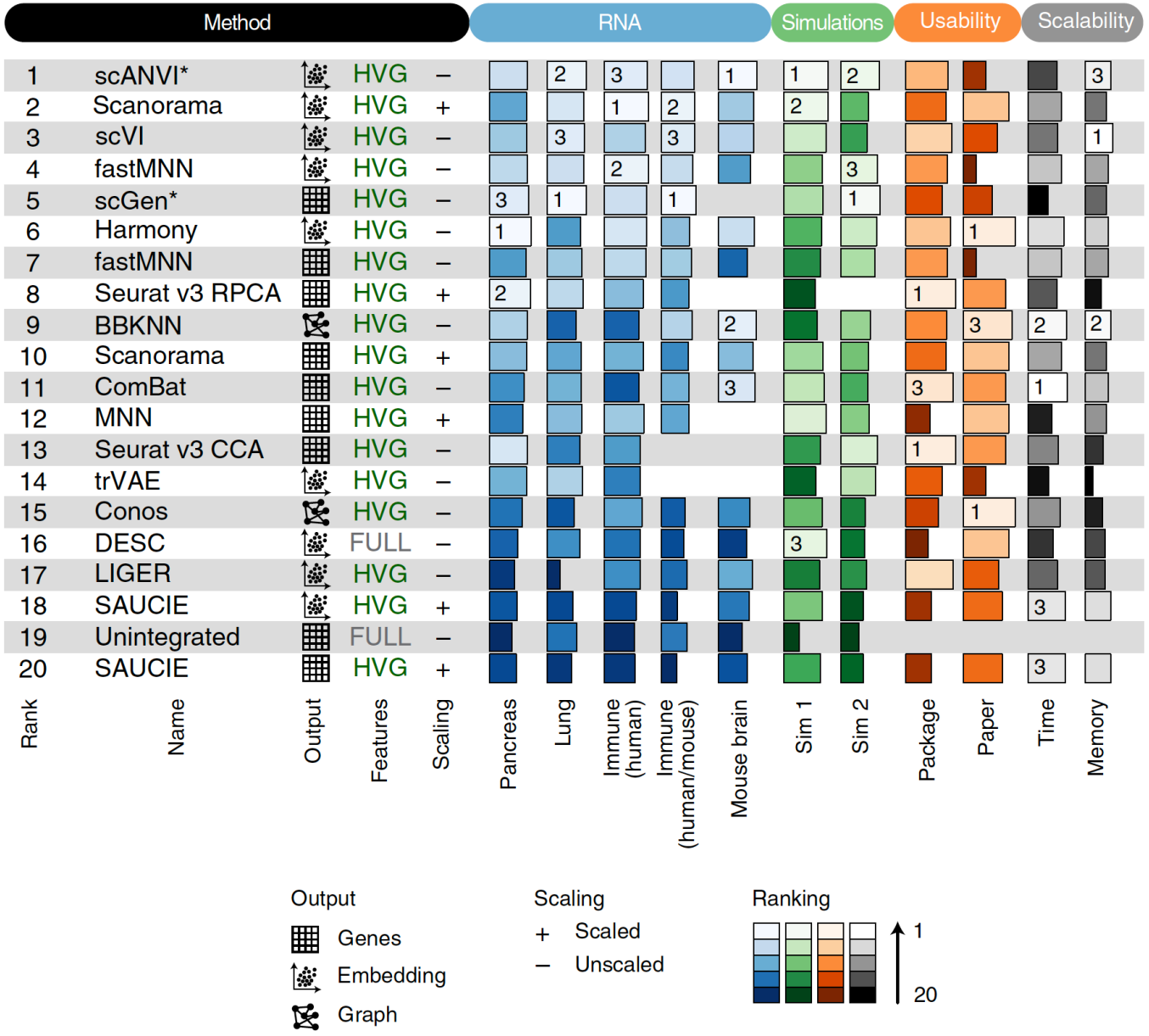
HiSeq 4000

4 billion reads

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)

Benchmarking of data integration methods



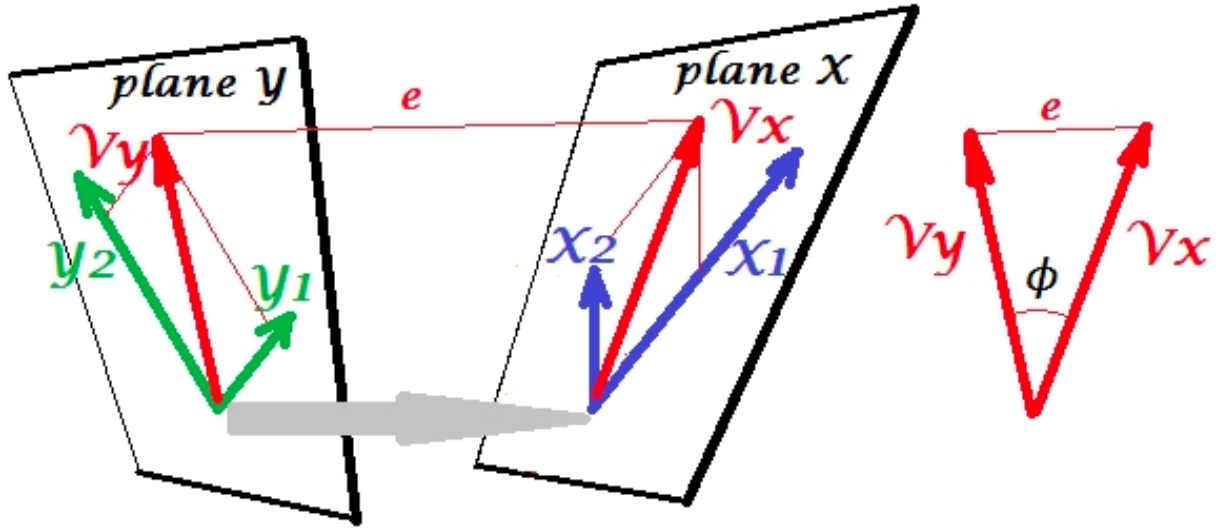
Luecken M. D. et al. (2022) *Nature Methods.*

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.

Canonical Correlation Analysis (CCA)

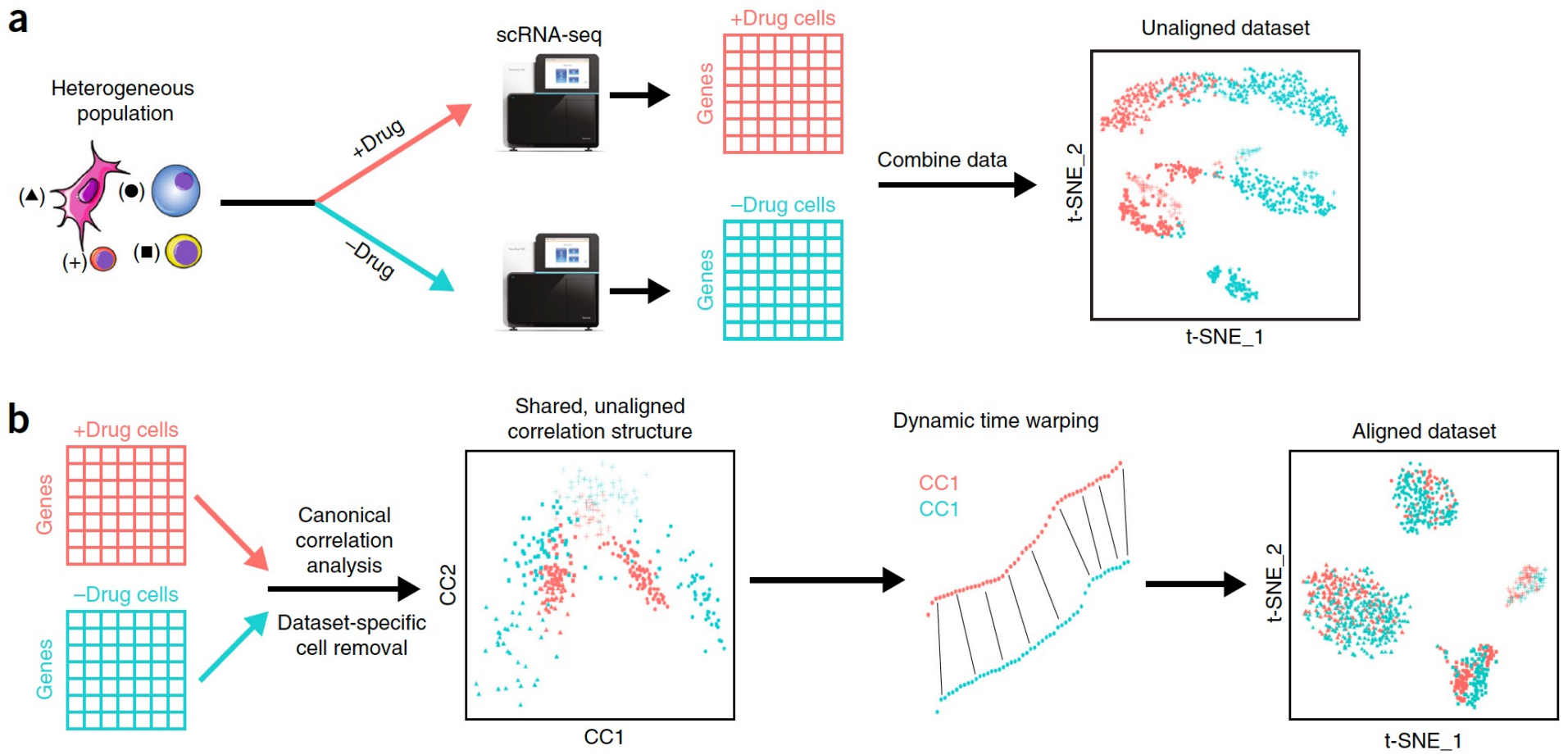
- CCA finds the linear combinations of variables across two datasets that are maximally correlated with one another.
  - The first pair of canonical variables maximizes the correlation across datasets.
  - The second pair of canonical variables maximizes the correlation subject to the constraint of not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

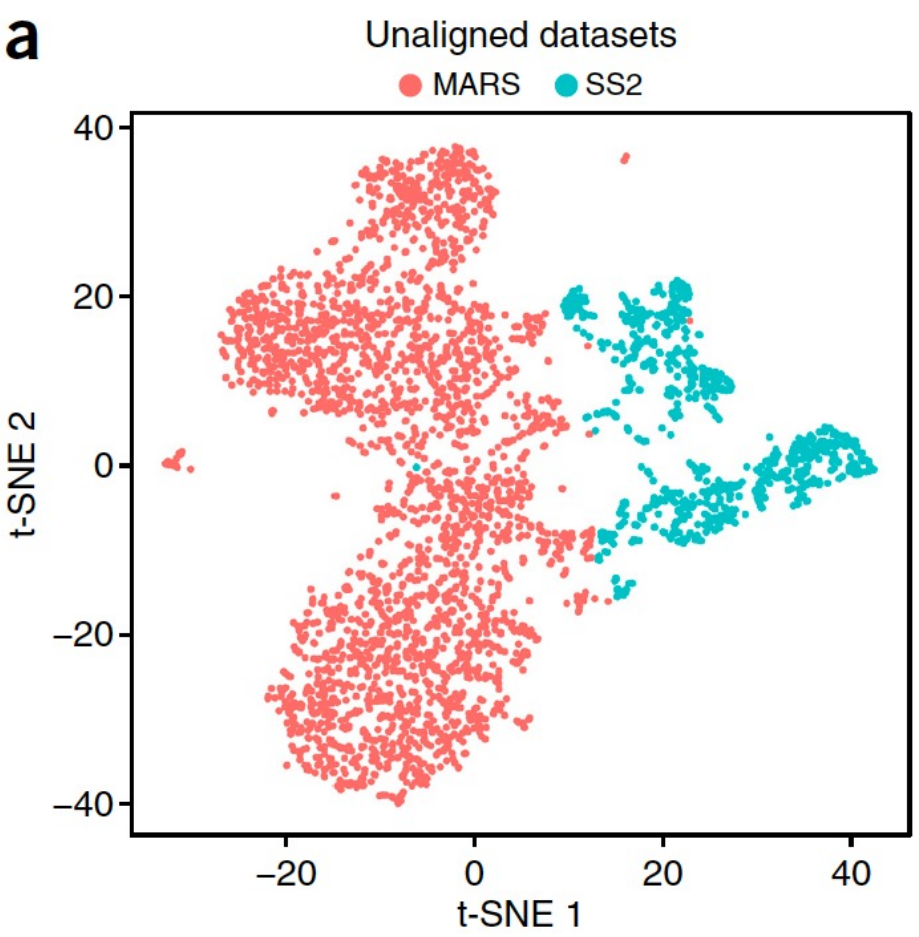
Canonical Correlation Analysis (CCA)

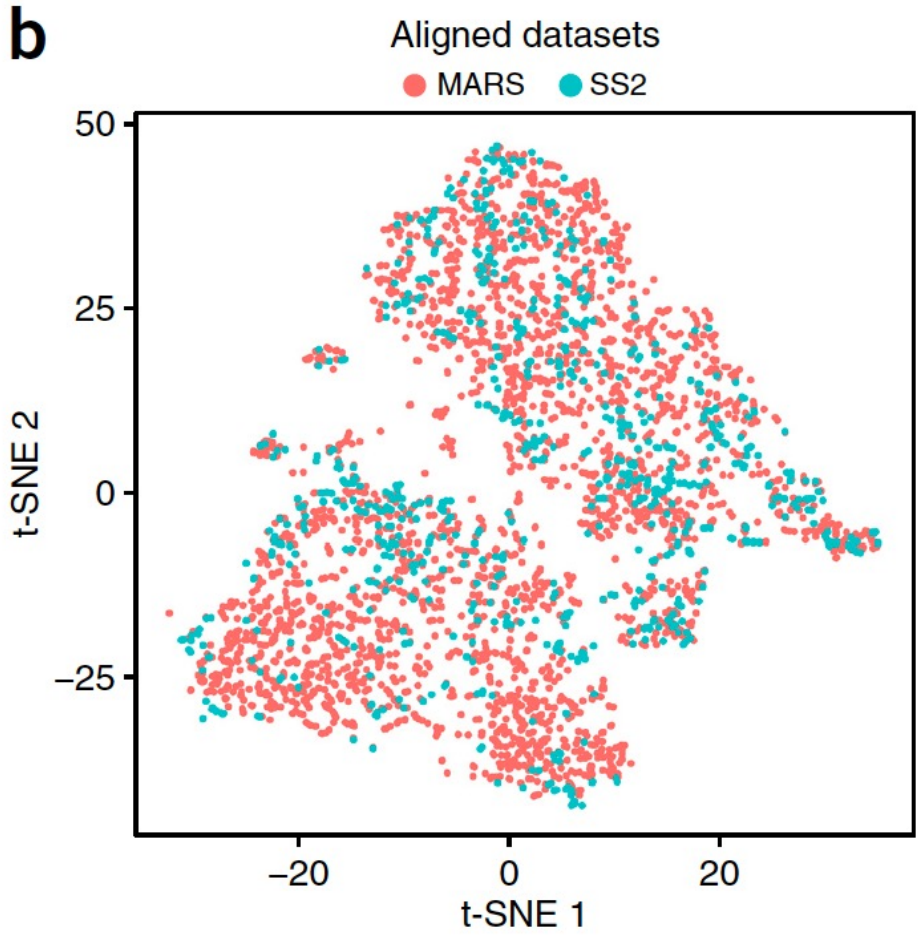


"Effectively, we treat the data sets as multiple measurements of a gene–gene covariance structure, and search for patterns that are common to the data sets."

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

Canonical Correlation Analysis (CCA)





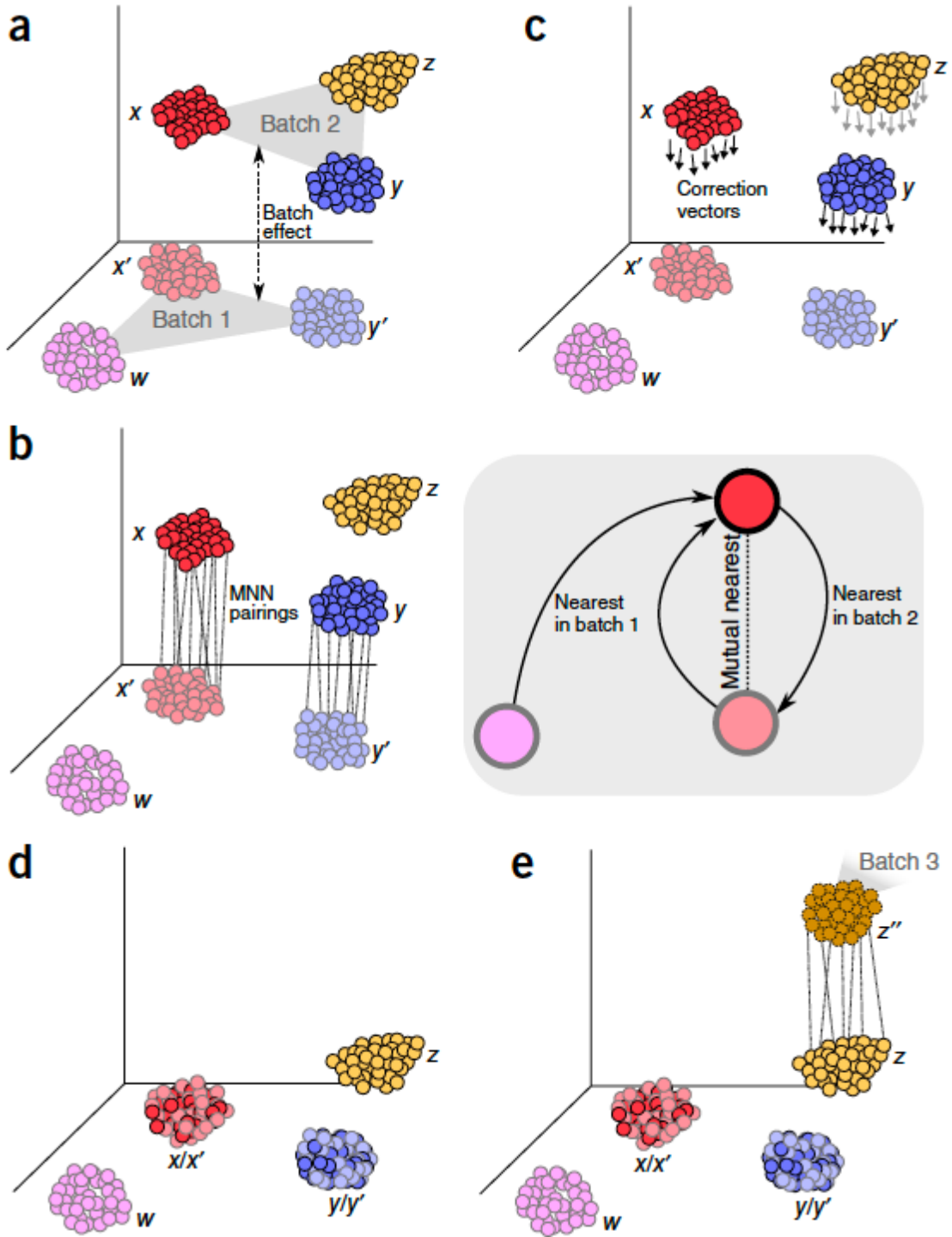
Assessing the performance of batch correction

"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

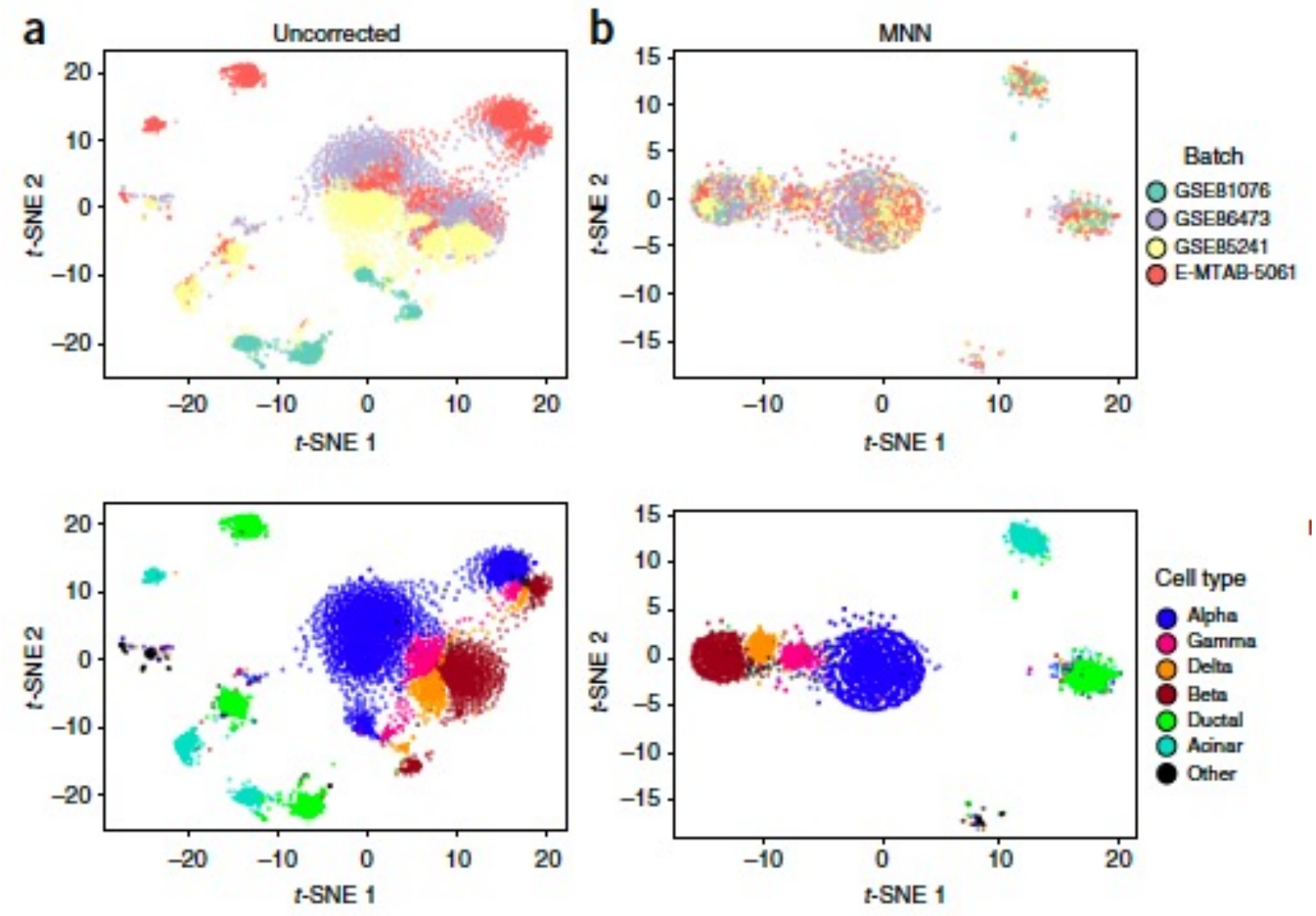
"If a pair of cells from each batch is contained in each other's set of nearest neighbors, those cells are considered to be mutual nearest neighbors. We interpret these pairs as containing cells that belong to the same cell type or state despite being generated in different batches. Thus, any systematic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



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

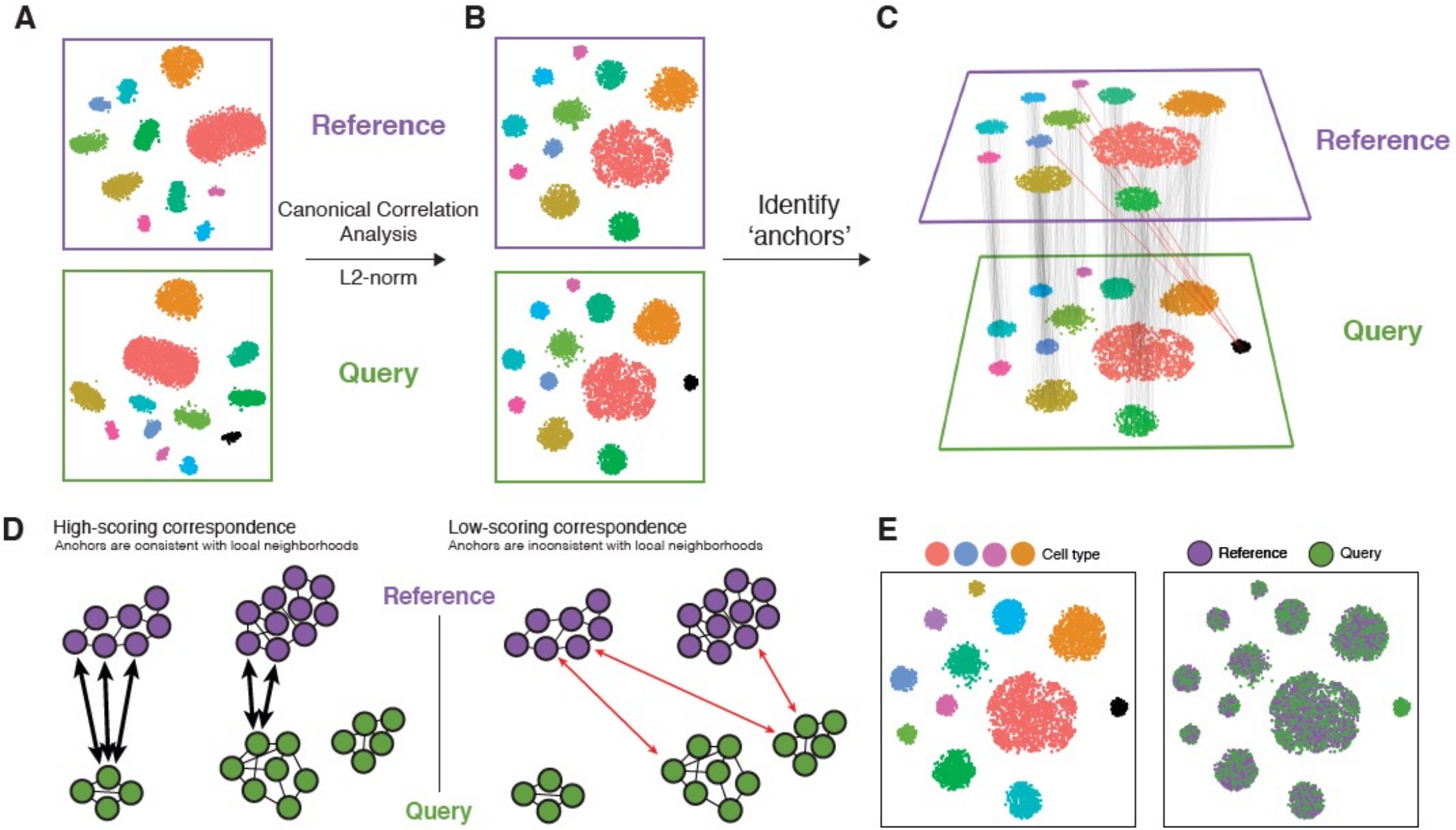
Mutual Nearest Neighbors

4 pancreas datasets



Haghverdi, L, et al. *Nature Biotechnology* (2018).

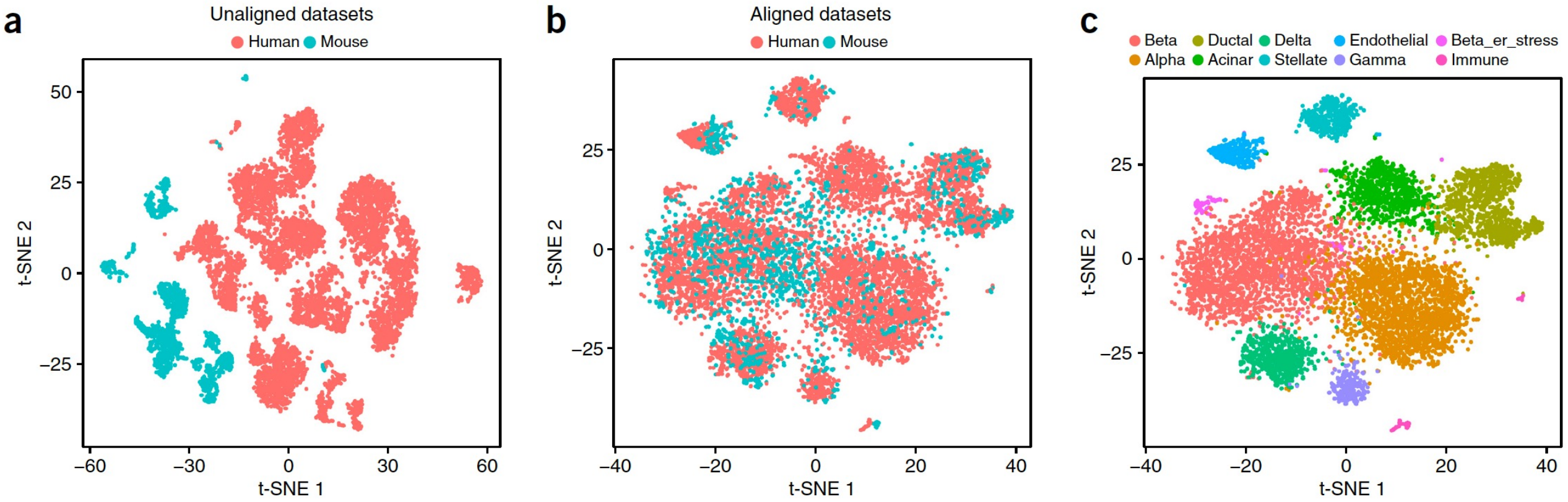
Combining CCA and Mutual Nearest Neighbors (Seurat v3)



Stuart et al. *Cell* (2019).

Combining CCA and Mutual Nearest Neighbors (Seurat v3)

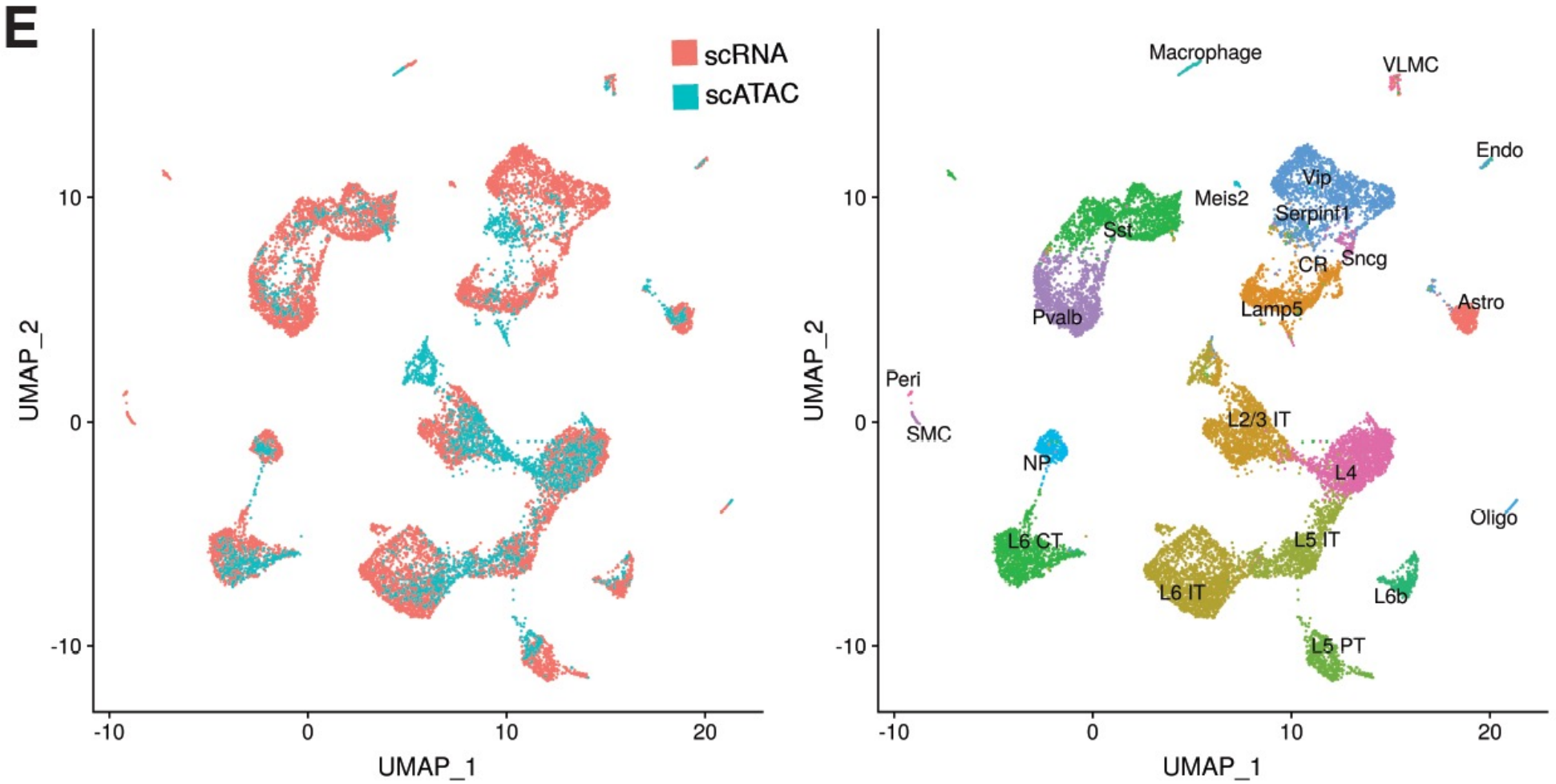
Human and mouse pancreas datasets



Stuart et al. *Cell* (2019).

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

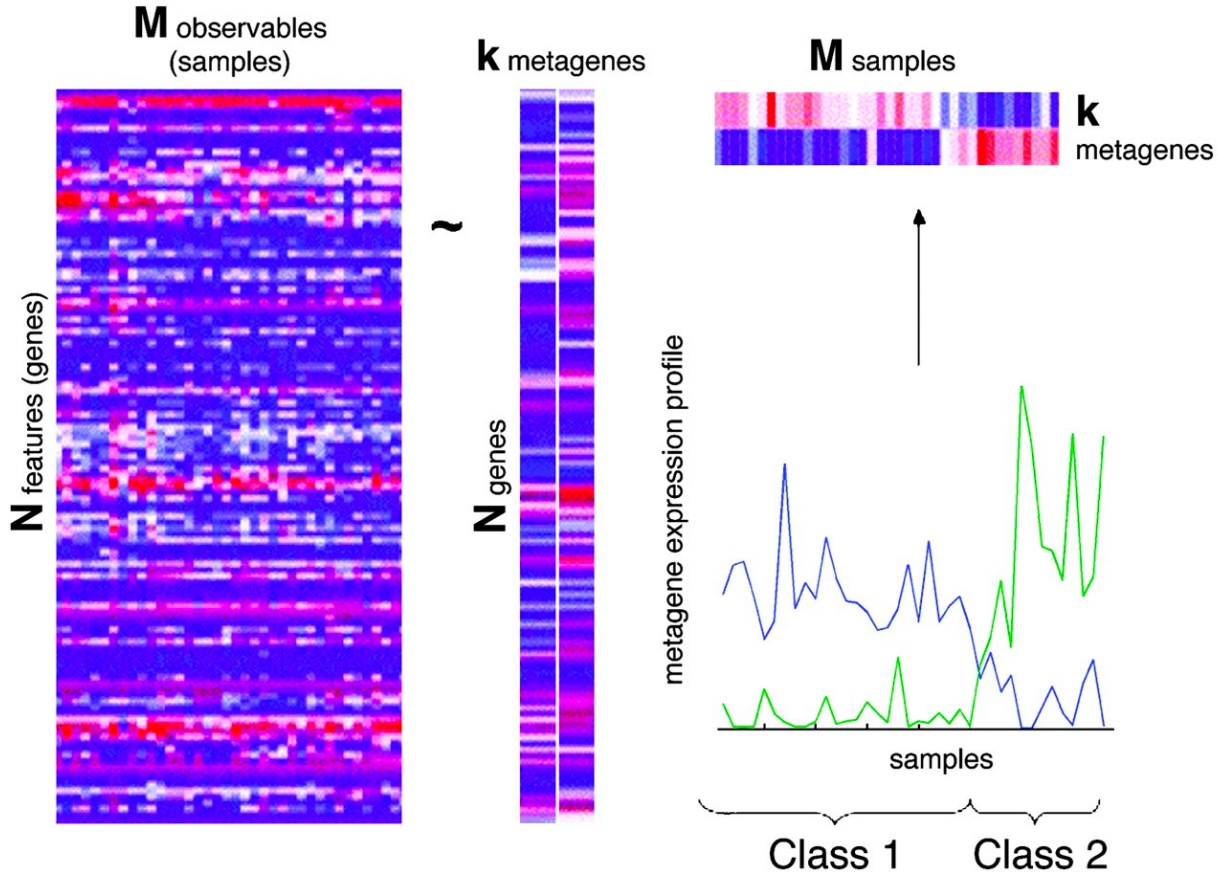


Stuart et al. *Cell* (2019).

Batch correction and data modality integration

Nonnegative Matrix Factorization (NMF)

$A_{\text{(rank M)}} \sim W \quad H_{\text{(rank k=2)}}$

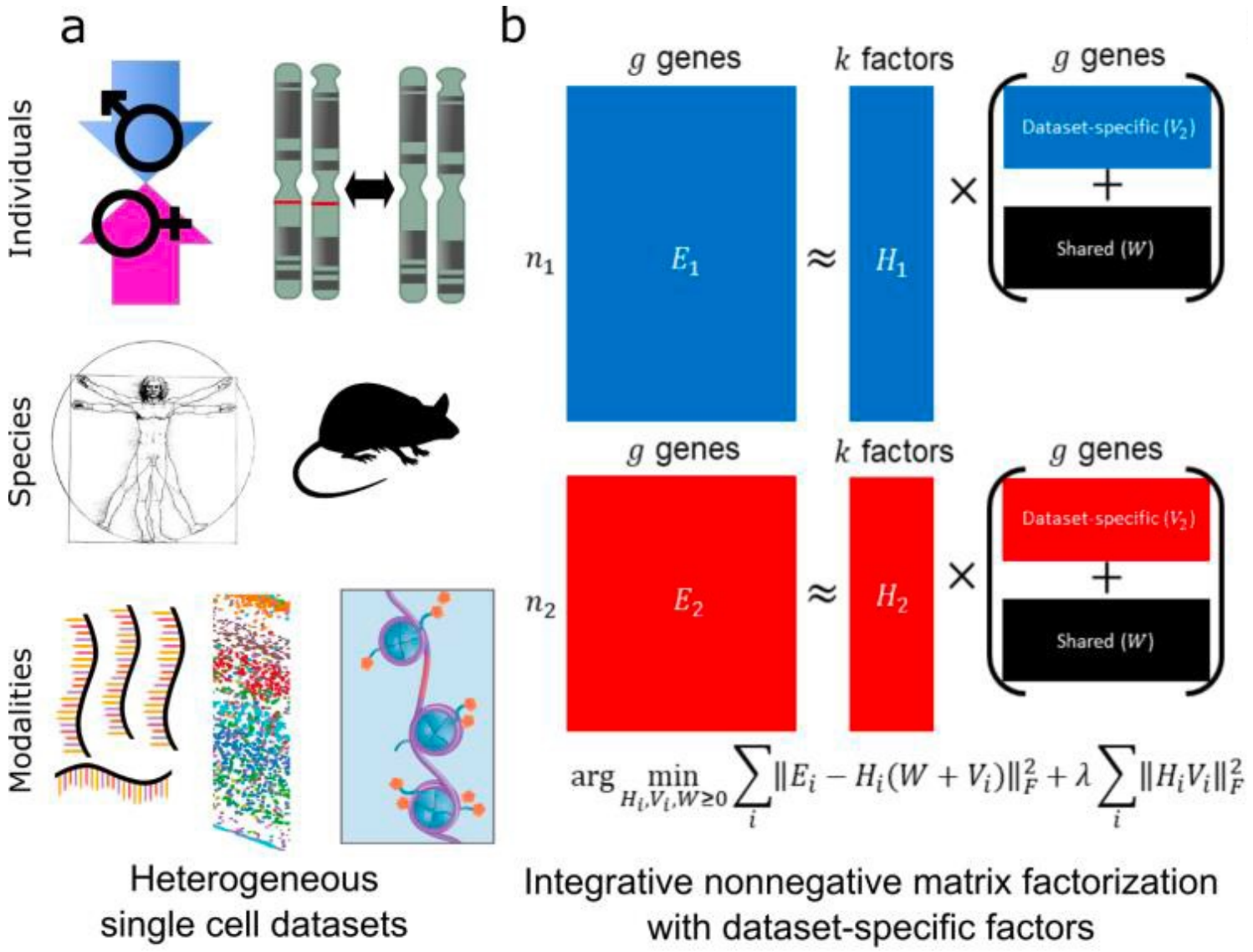


"Our goal is to find a small number of metagenes, each defined as a positive linear combination of the *N* genes. We can then approximate the gene expression pattern of samples as positive linear combinations of these metagenes. Mathematically, this corresponds to factoring matrix *A* into two matrices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

Brunet J. et *al*. (2004) *PNAS*..

Batch correction and data modality integration using LIGER

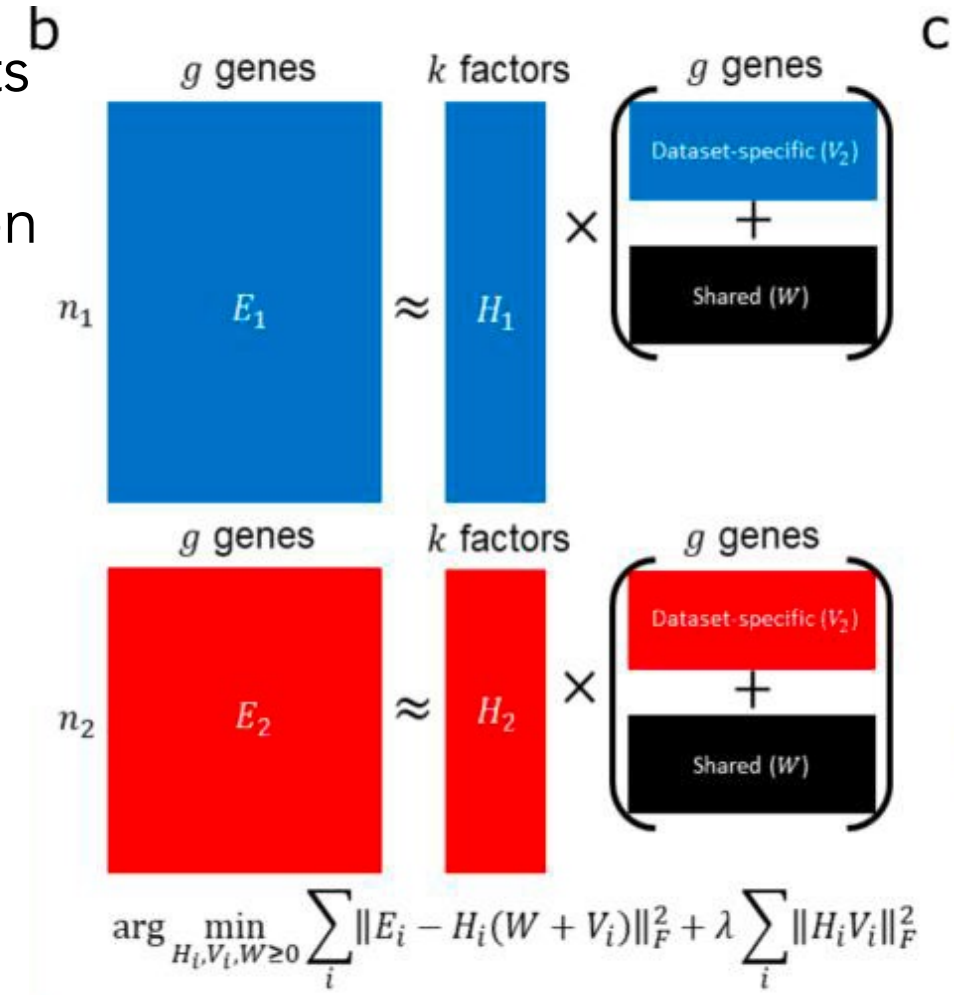
LIGER implements non-negative matrix factorization



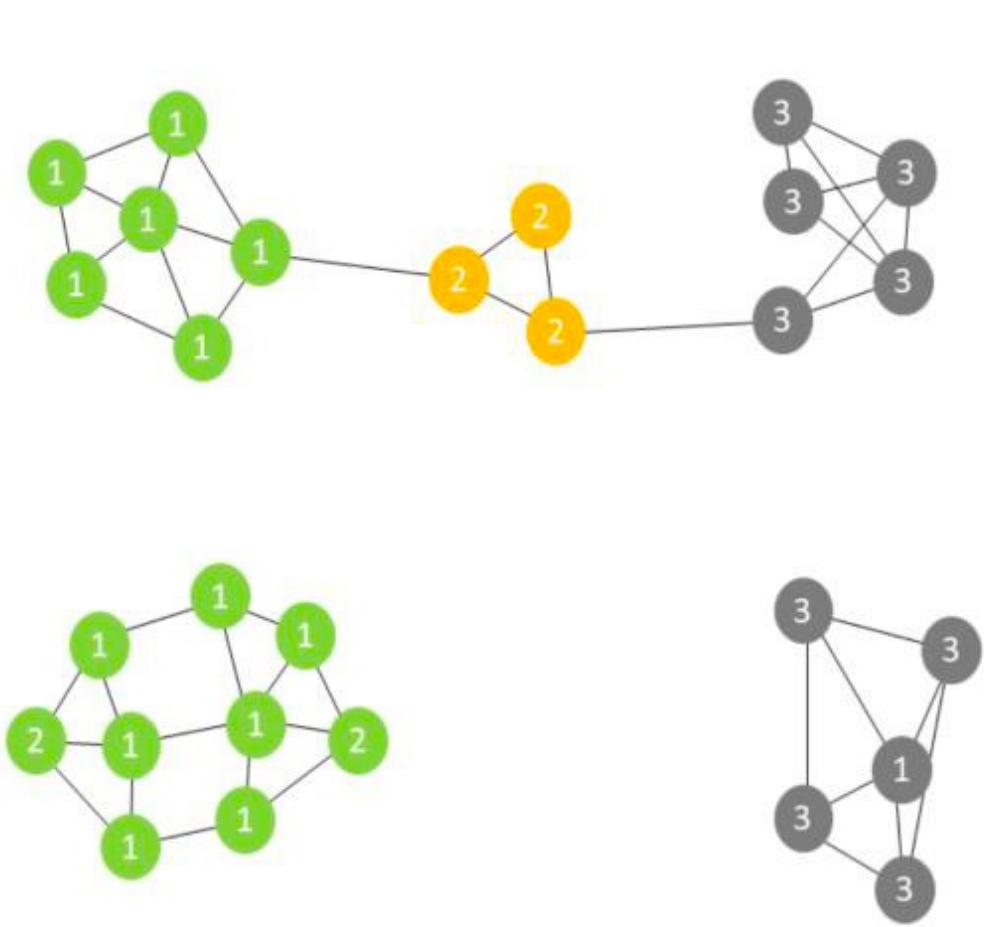
Welch, et al. *Cell* (2019).

Batch correction and data modality integration using LIGER

LIGER implements non-negative matrix factorization



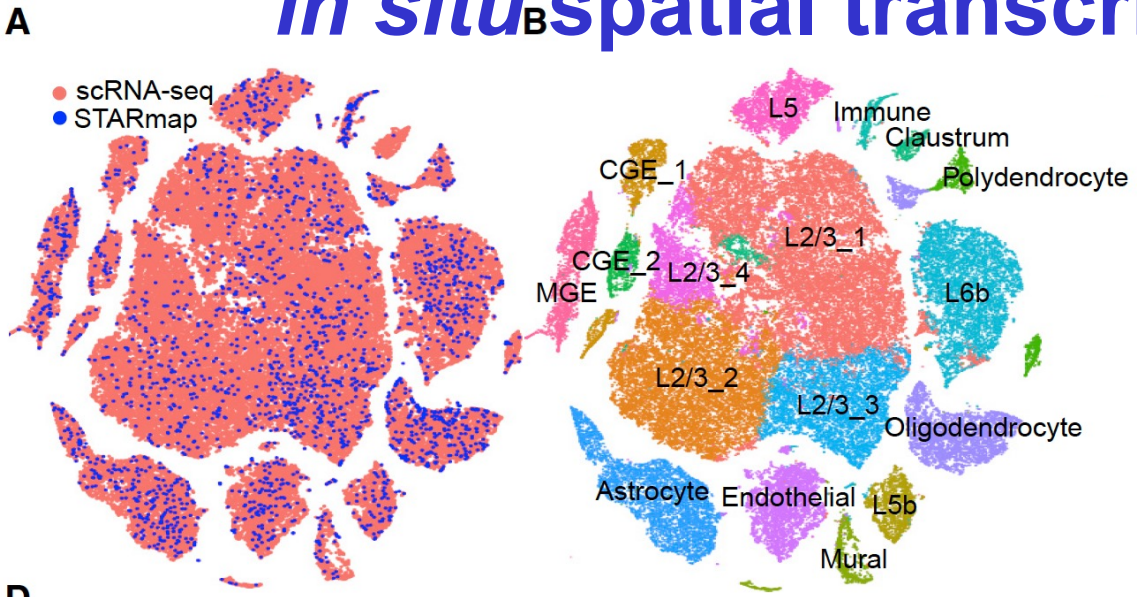
Integrative nonnegative matrix factorization with dataset-specific factors



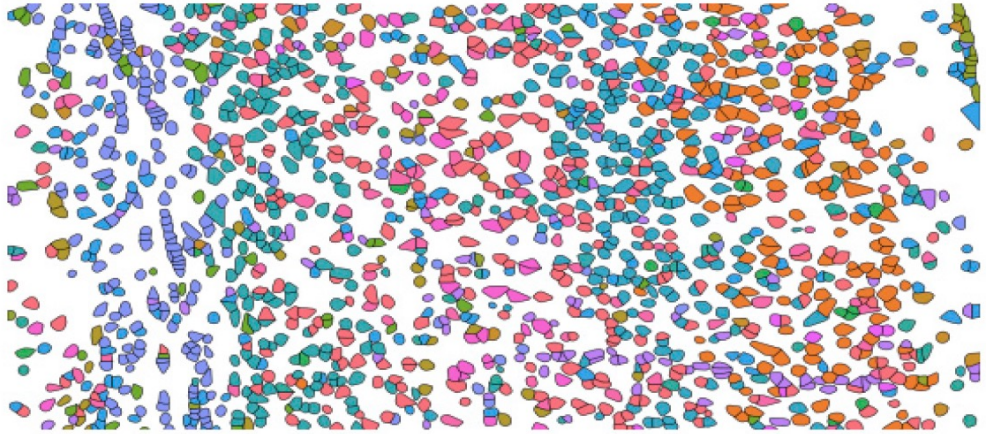
Joint clustering using shared factor neighborhood graph

Welch, et al. *Cell* (2019).

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



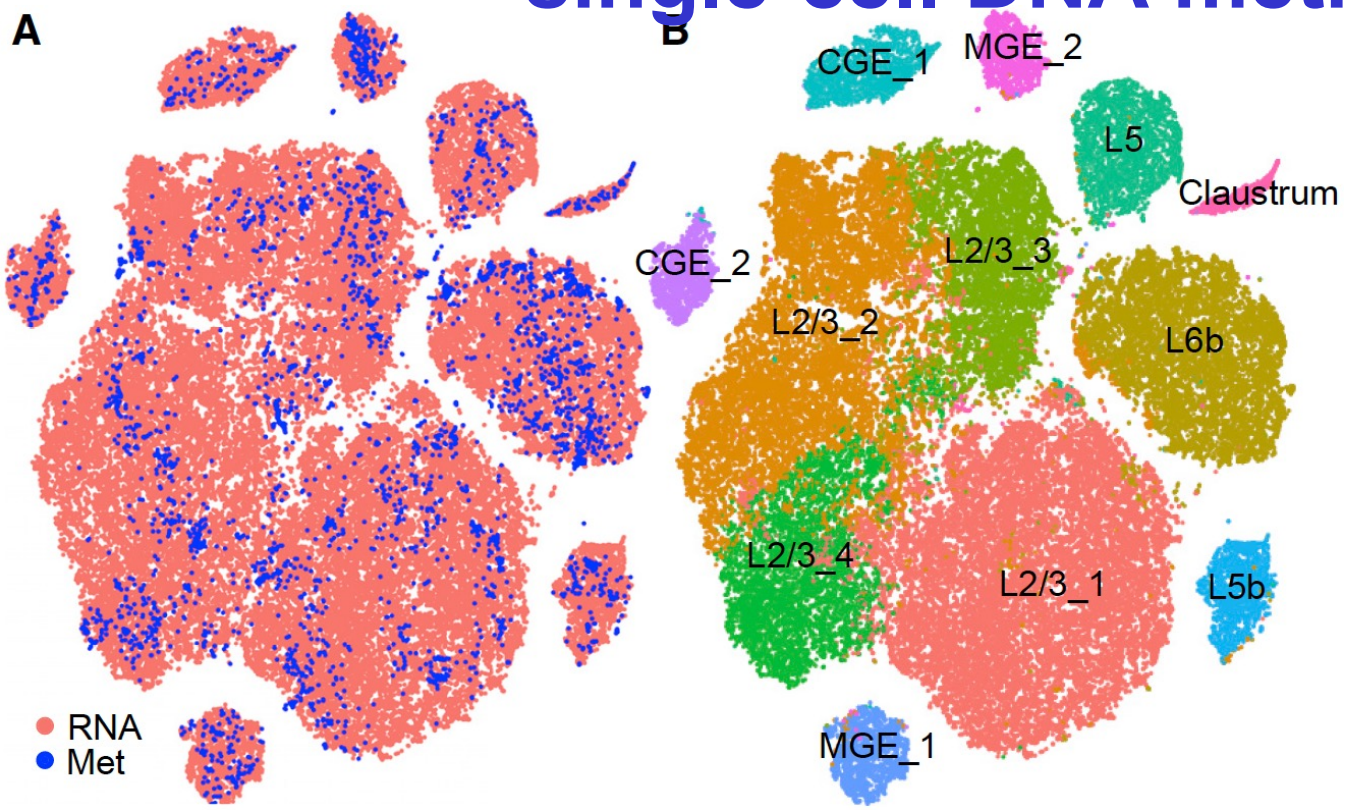
71,000 cells from scRNA-seq
2,500 cells from STARmap



What are advantages of integrated analysis of these 2 datasets?

Welch, et al. *Cell* (2019).

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

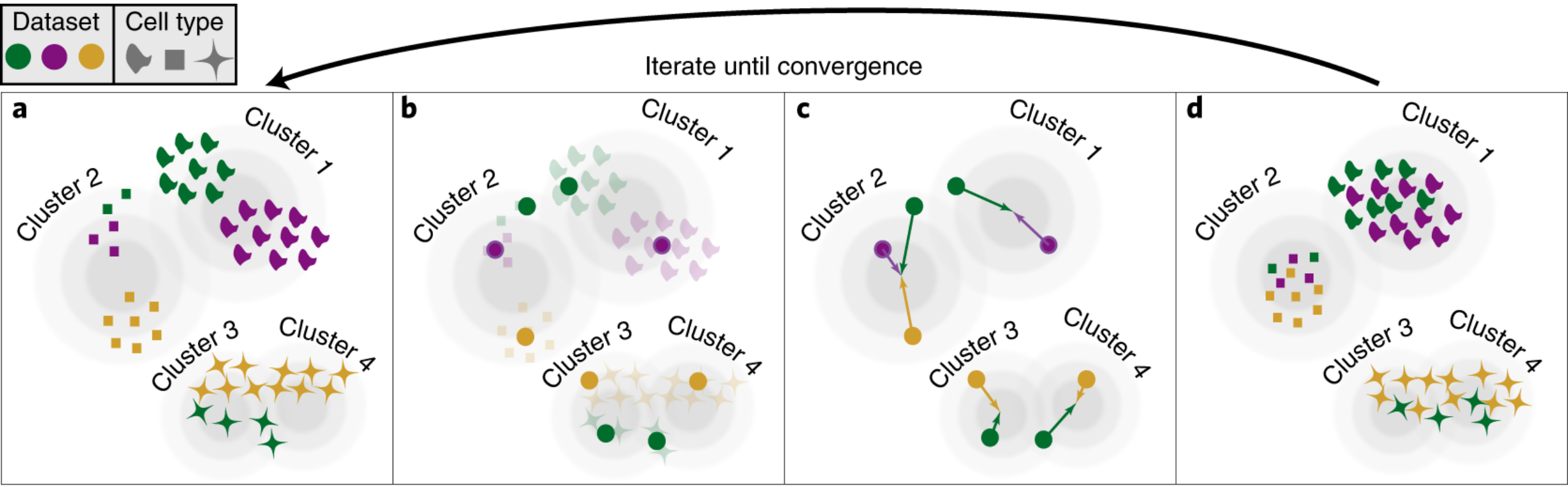


56,000 cells from scRNA-seq
3,000 cells from DNA methylation

"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).

Batch correction and data modality integration using Harmony



Soft assign cells to clusters, favoring mixed dataset representation

Get cluster centroids for each dataset

Get dataset correction factors for each cluster

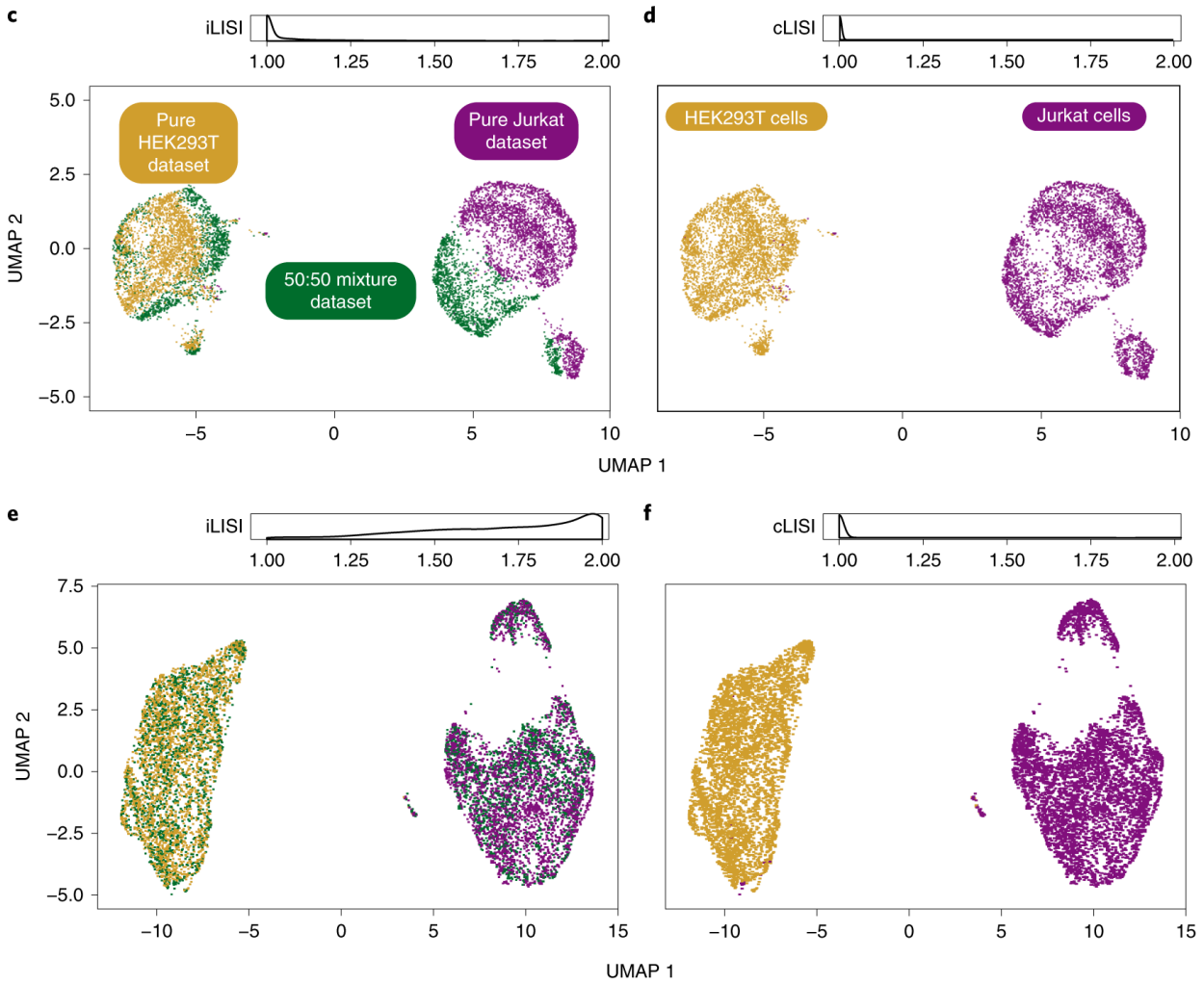
Move cells based on soft cluster membership

Korsunksy I. et al. *Nature Methods* (2019).

Harmony integrates different datasets while maintaining cell type differences

Before integration

After integration



Korsunksy I. et al. *Nature Methods* (2019).

Batch correction and biological differences

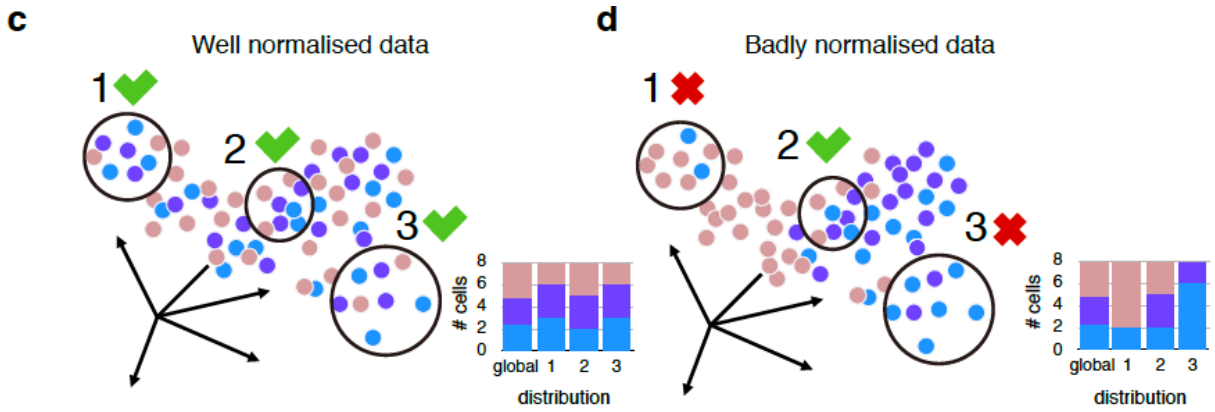
Batch correction in single-cell transcriptomics is designed to **remove technical variations that arise due to differences in the experimental procedures such as sample preparation, sequencing, and data analysis**. The goal of batch correction is to reduce technical noise in the data and allow biological differences between cells to be better distinguished.

However, batch correction can **potentially remove some biological differences in the data, particularly if the biological differences are correlated with the batch effect.**

For example, if cells from different batches are isolated from different tissues or at different stages of development, then the batch correction algorithm may inadvertently adjust the expression levels of genes that are differentially expressed between these groups.

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

Luecken M. D. et al. (2022) *Nature Methods.*

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

Comparison between discussed approaches

Aspect	CCA	MNN	NMF	LIGER	Harmony	scVI
Type	Linear	Distance-based	Linear	Linear+NMF	Linear	Non-linear (neural)
Computational Cost	Low-Medium	Low	Medium	Medium	Low-Medium	High (GPU recommended)
Scalability	Small	Medium	Medium	Good	Excellent	Excellent
# Batches	2 main	2+	2+	2+	**3+ (optimal)**	2+
Batch Assumptions	Linear	Uniform	Additive	Shared factors	Density-based	Complex/non-linear
Interpretability	Low	Anchors clear	**High (factors)**	**High (factors)**	Medium	Low (black box)
Preserves Structure	Modest	**Excellent**	Moderate	**Good**	**Excellent**	Good
Handles Dropout	No	No	Partially	Partially	No	**Yes**

Once data is integrated, how do we compare across biological conditions?

When detecting gene differentially expressed across different biological conditions (e.g. case vs control, treated vs untreated, etc…), what is the replicate? A single cell? A sample?

Strongly consider using pseudobulk models or mixed effect models when comparing samples across conditions.

Remember: go back to normalized (non batch-corrected) gene expression values for DE analyses! Add covariates for your batches in the DE statistical model.

