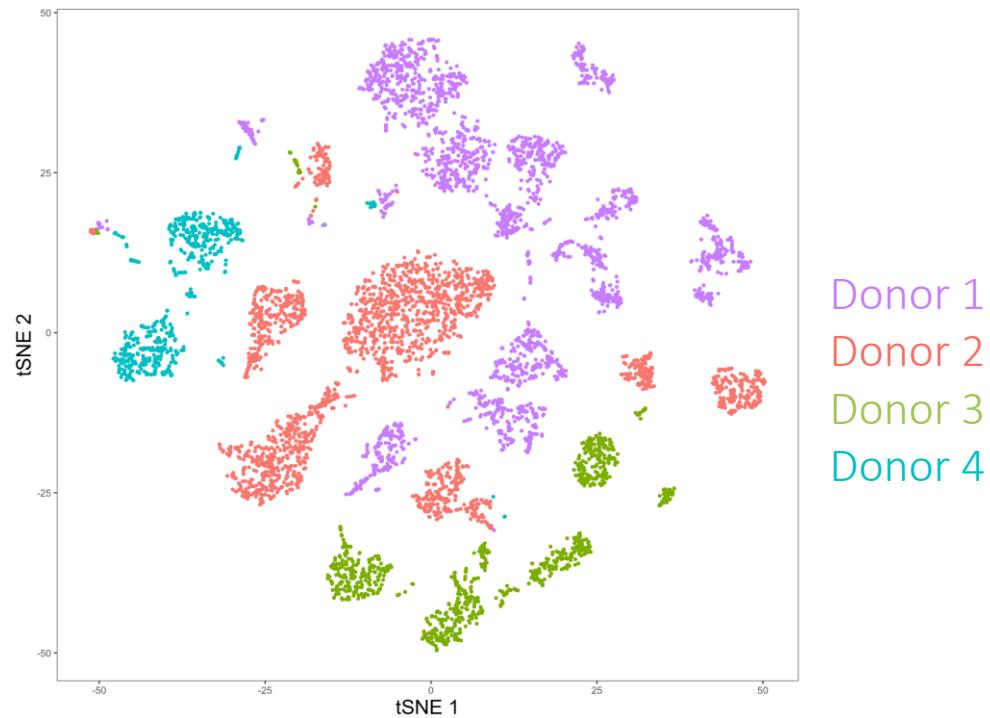


Single Cell RNA-seq Data Integration

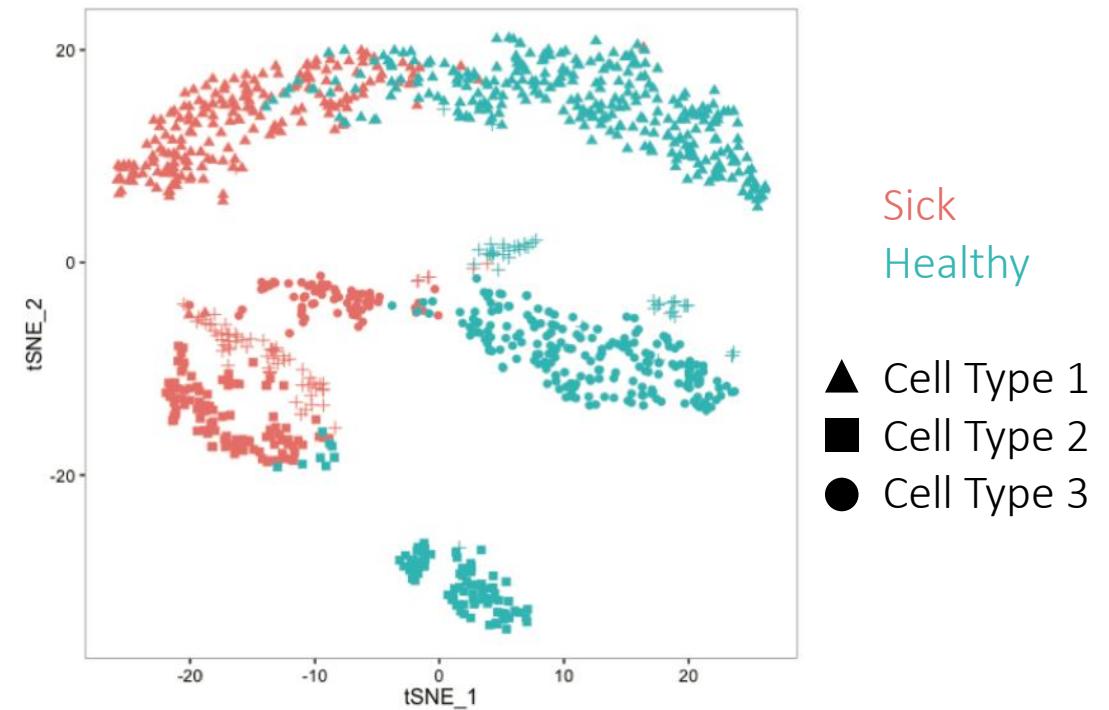
Ahmed Mahfouz

Leiden Computational Biology Center, LUMC
Delft Bioinformatics Lab, TU Delft

Why integrate?



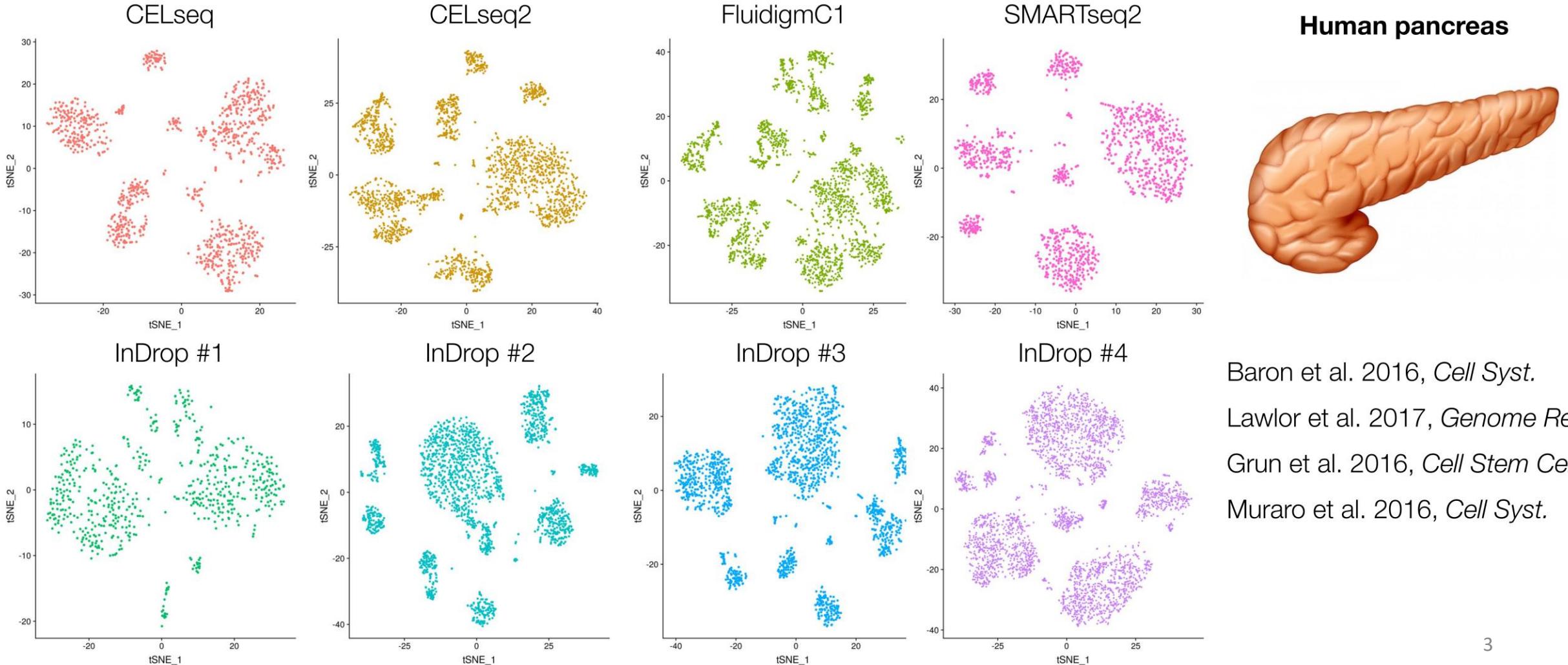
Same tissue from different donors



Cross condition comparisons

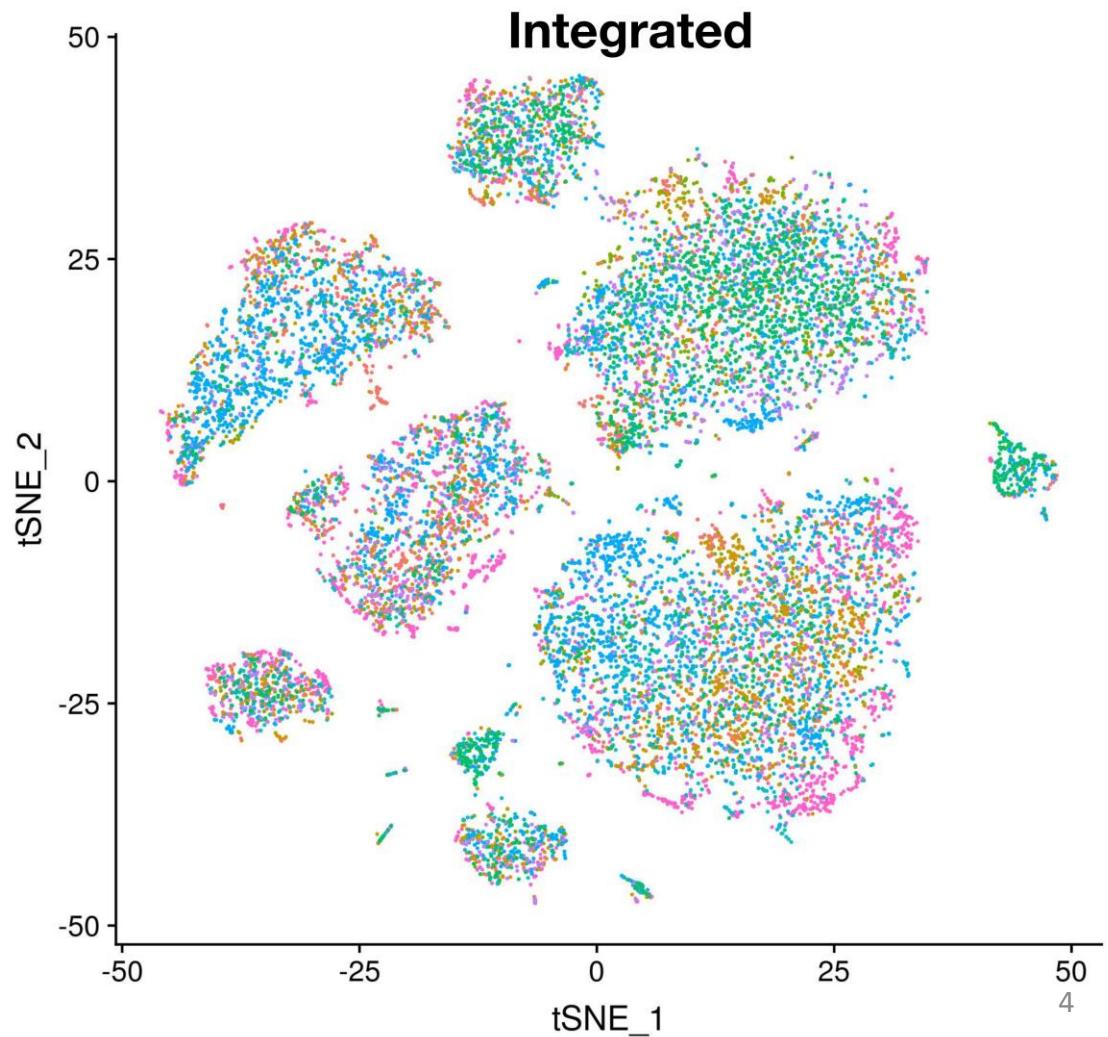
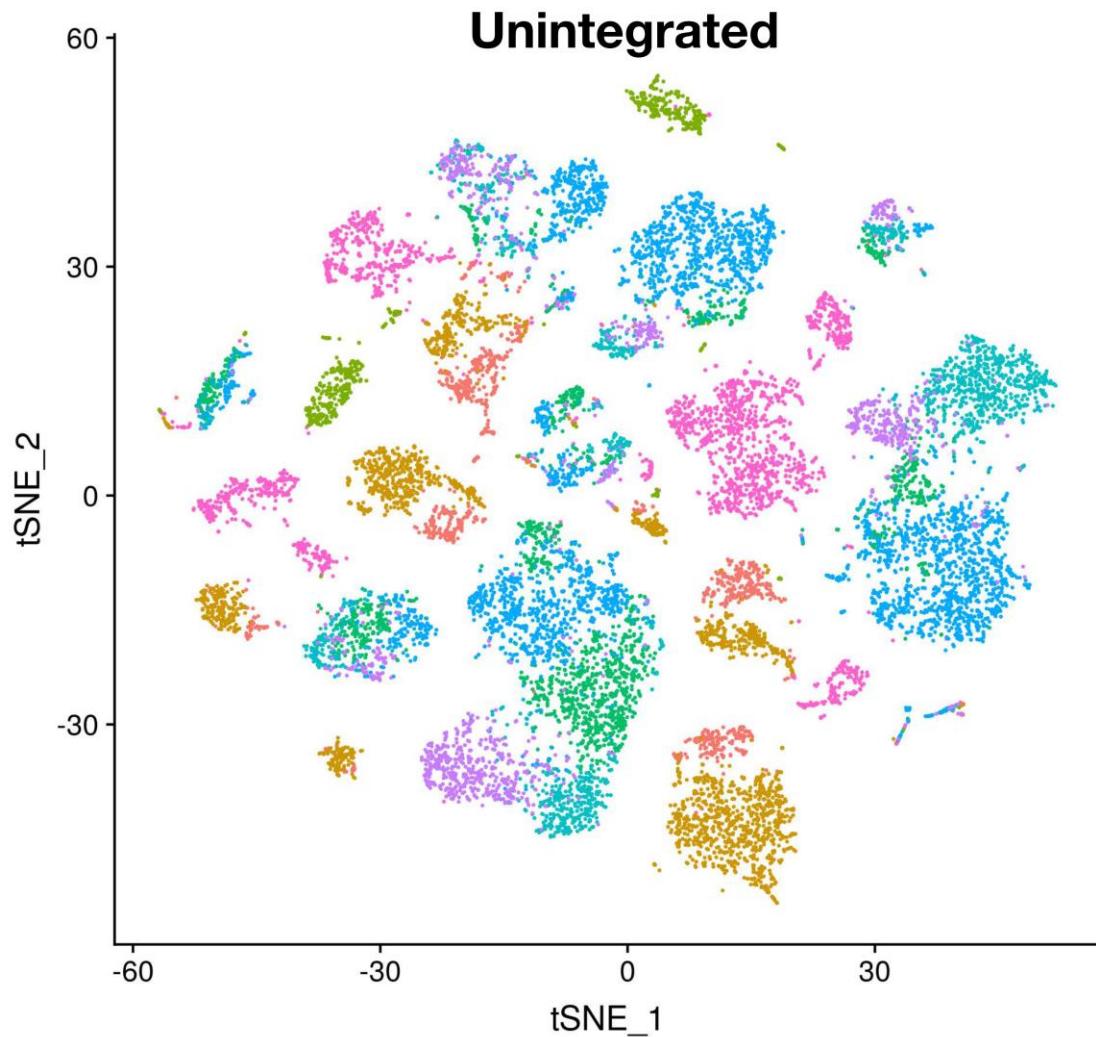
Building a cell atlas

8 maps of the human pancreas



Building a cell atlas

8 maps of the human pancreas

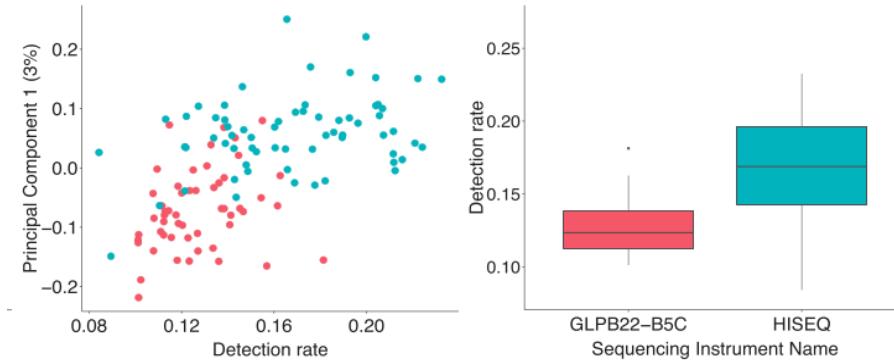


Confounders and batch effects

1. Technical variability

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

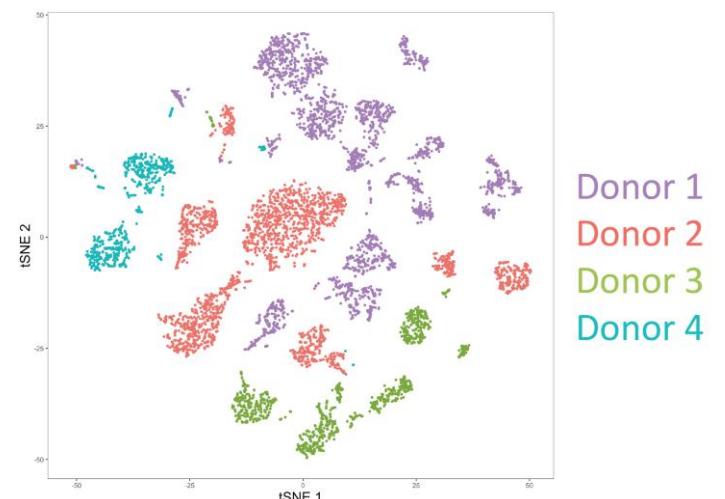
Technical ‘batch effects’ confound downstream analysis



2. Biological variability

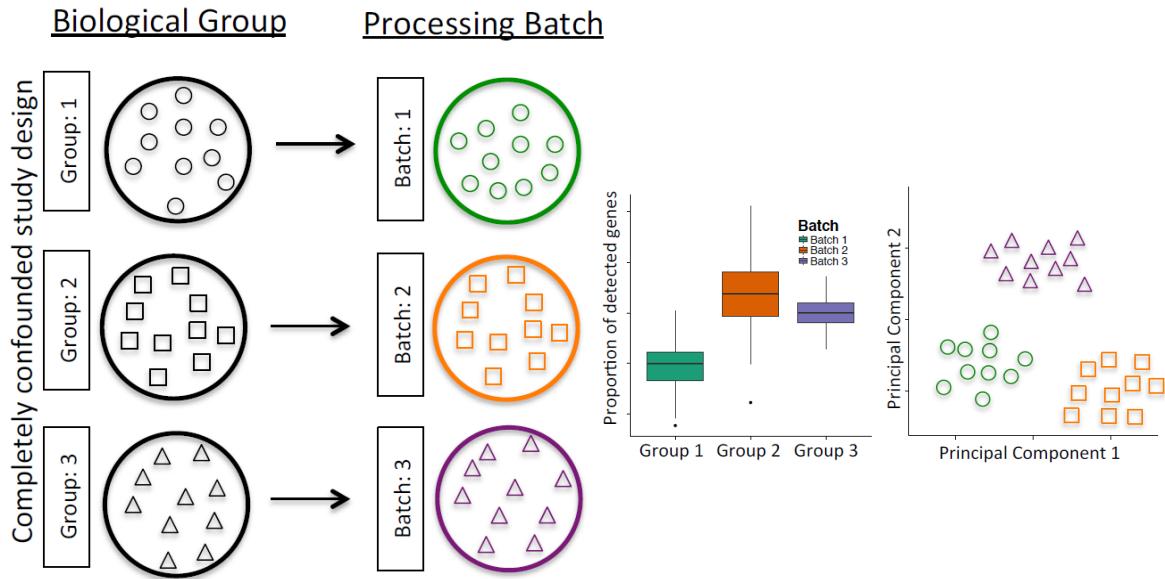
- Patient differences
- Environmental/genetic perturbation
- Evolution! (cross-species analysis)

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



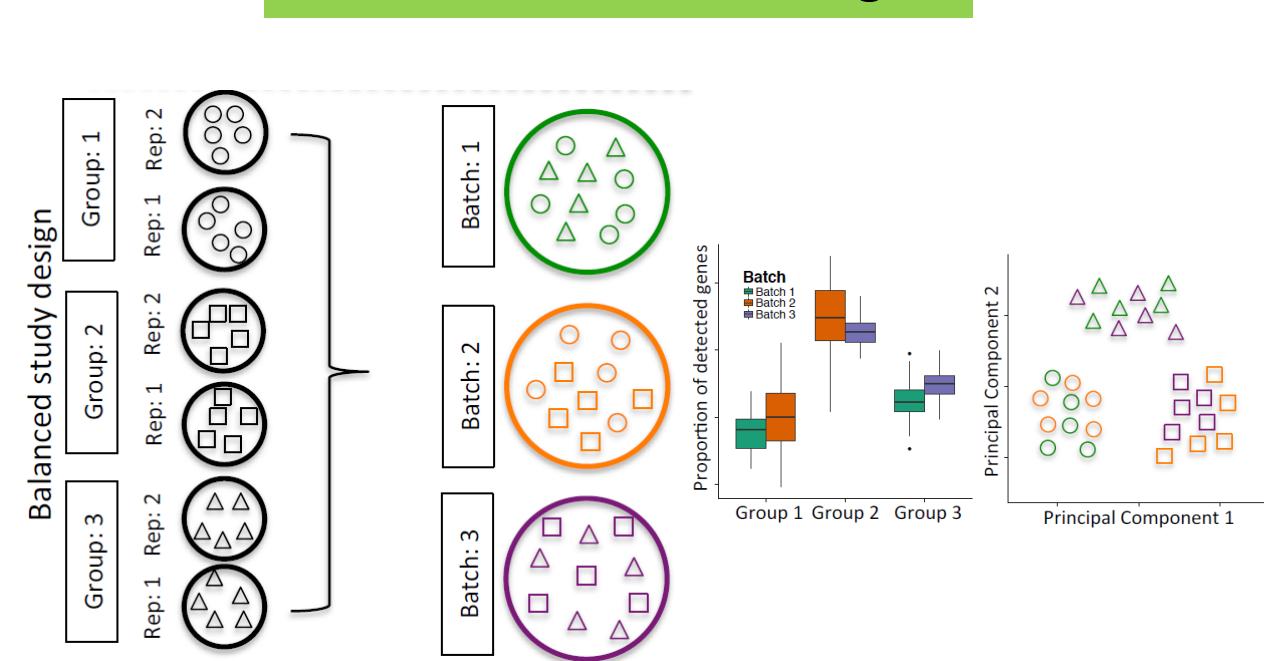
Confounders and batch effects

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.

Outline

- Single cell batch Correction methods:
- Performance assessment
- Sample multiplexing
- Simultaneous mRNA and protein profiling: REAP-seq and CITE-seq

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 (do they?)

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

Two broad strategies:

- Joint dimension reduction
- Graph-based joint clustering

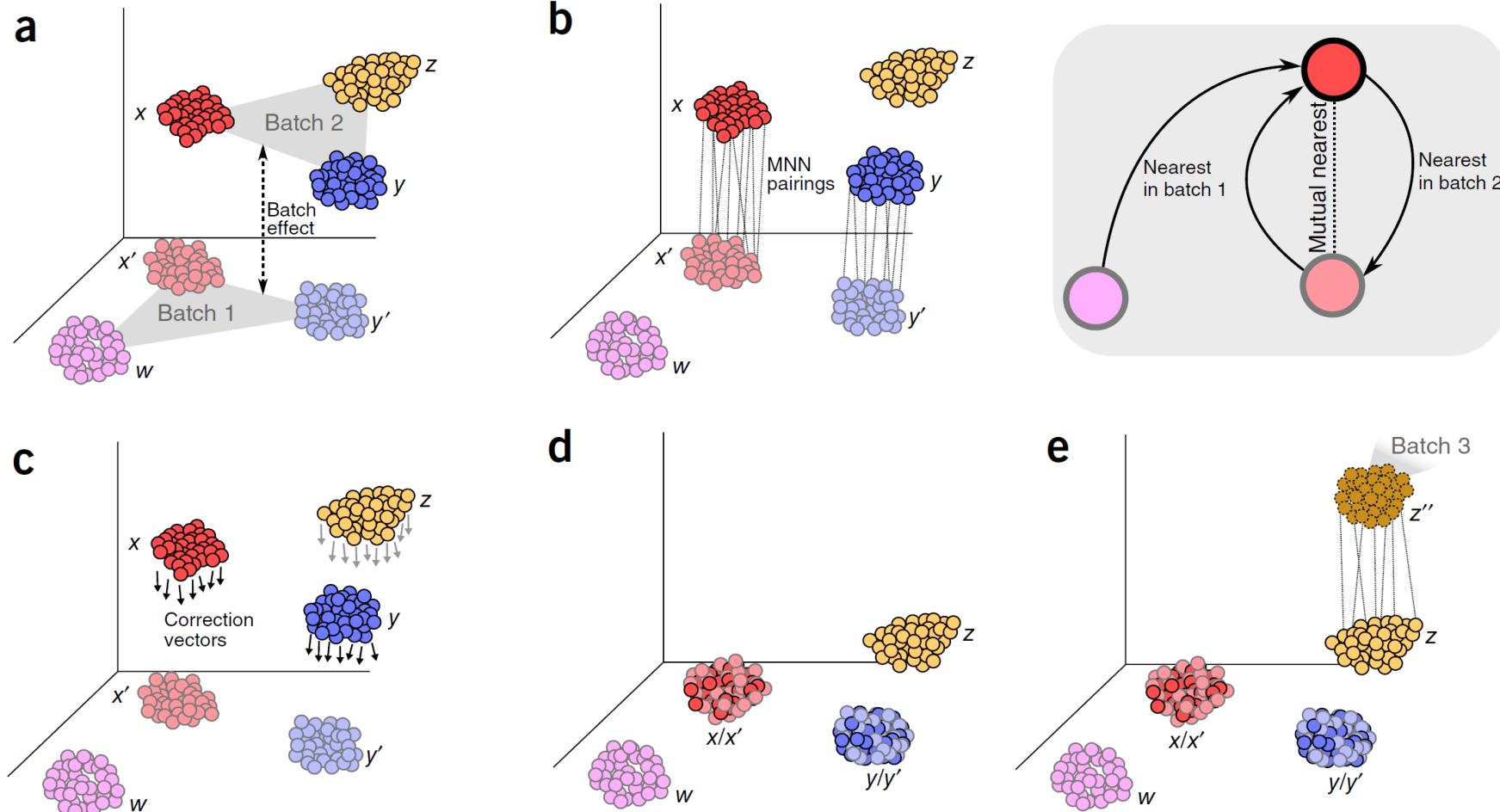
Batch correction methods

- MNcorrect (<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>)
- ...

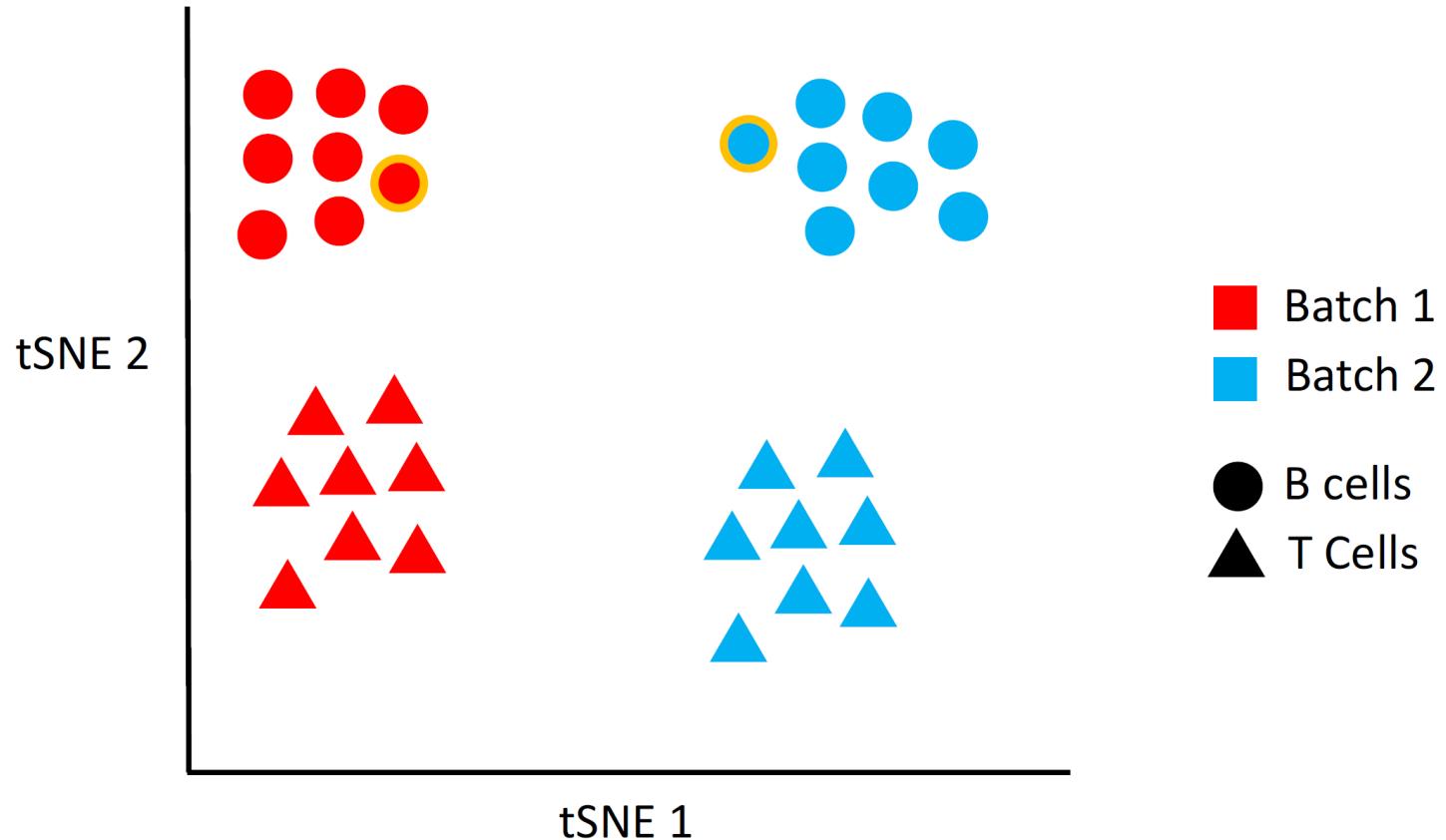
Two broad strategies:

- Joint dimension reduction
- Graph-based joint clustering

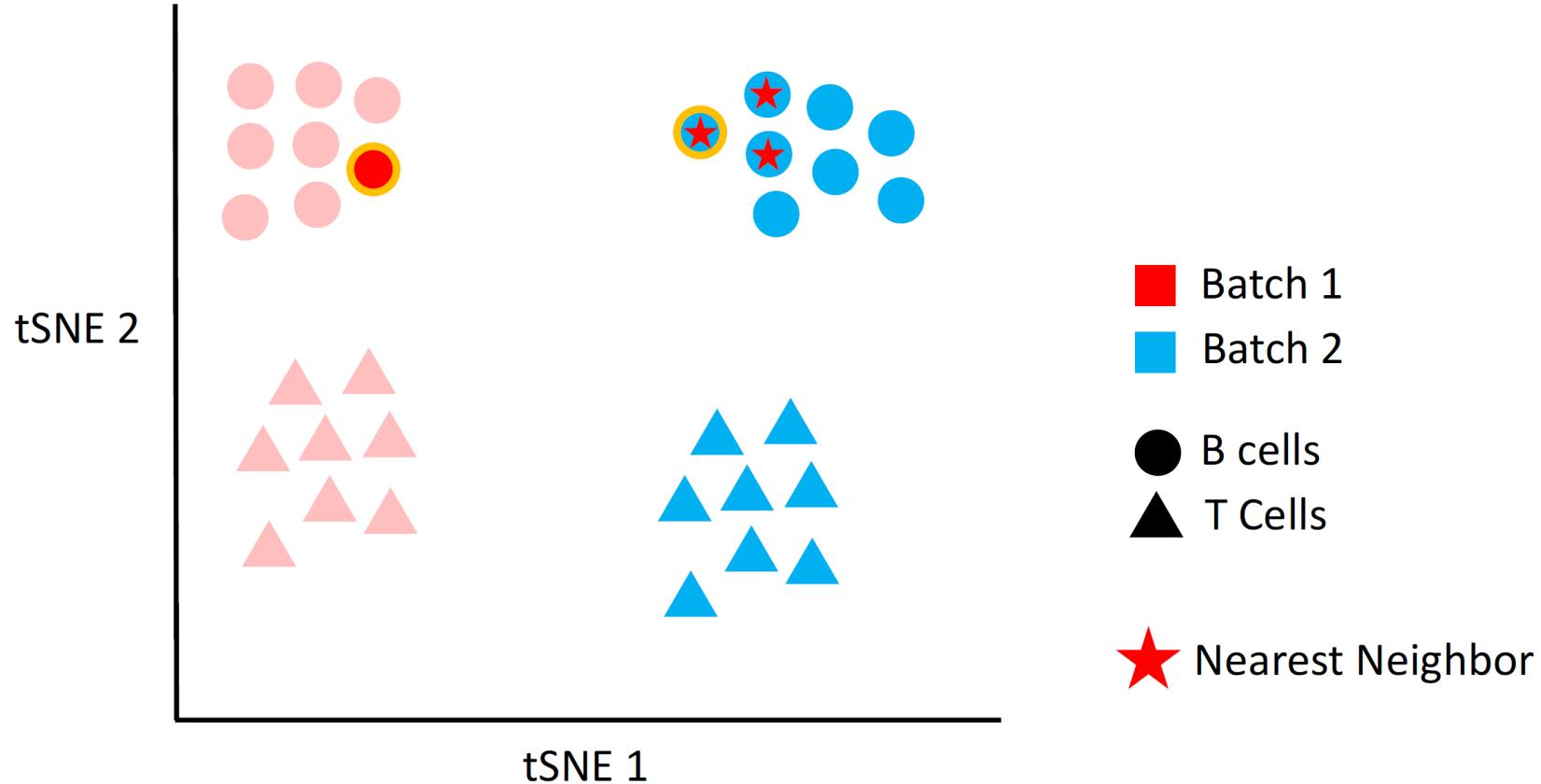
Mutual Nearest Neighbors (MNN)



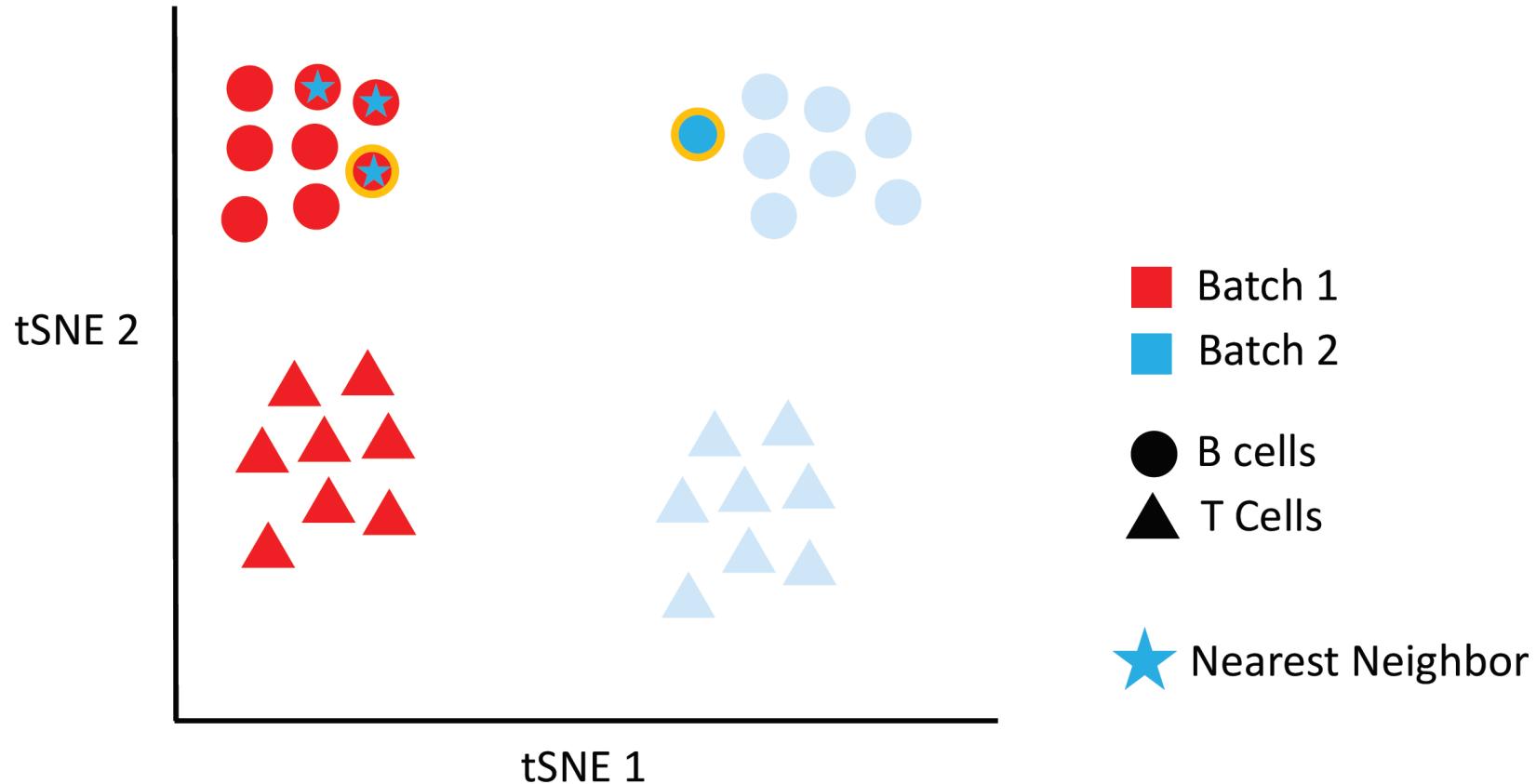
Mutual Nearest Neighbors (MNN)



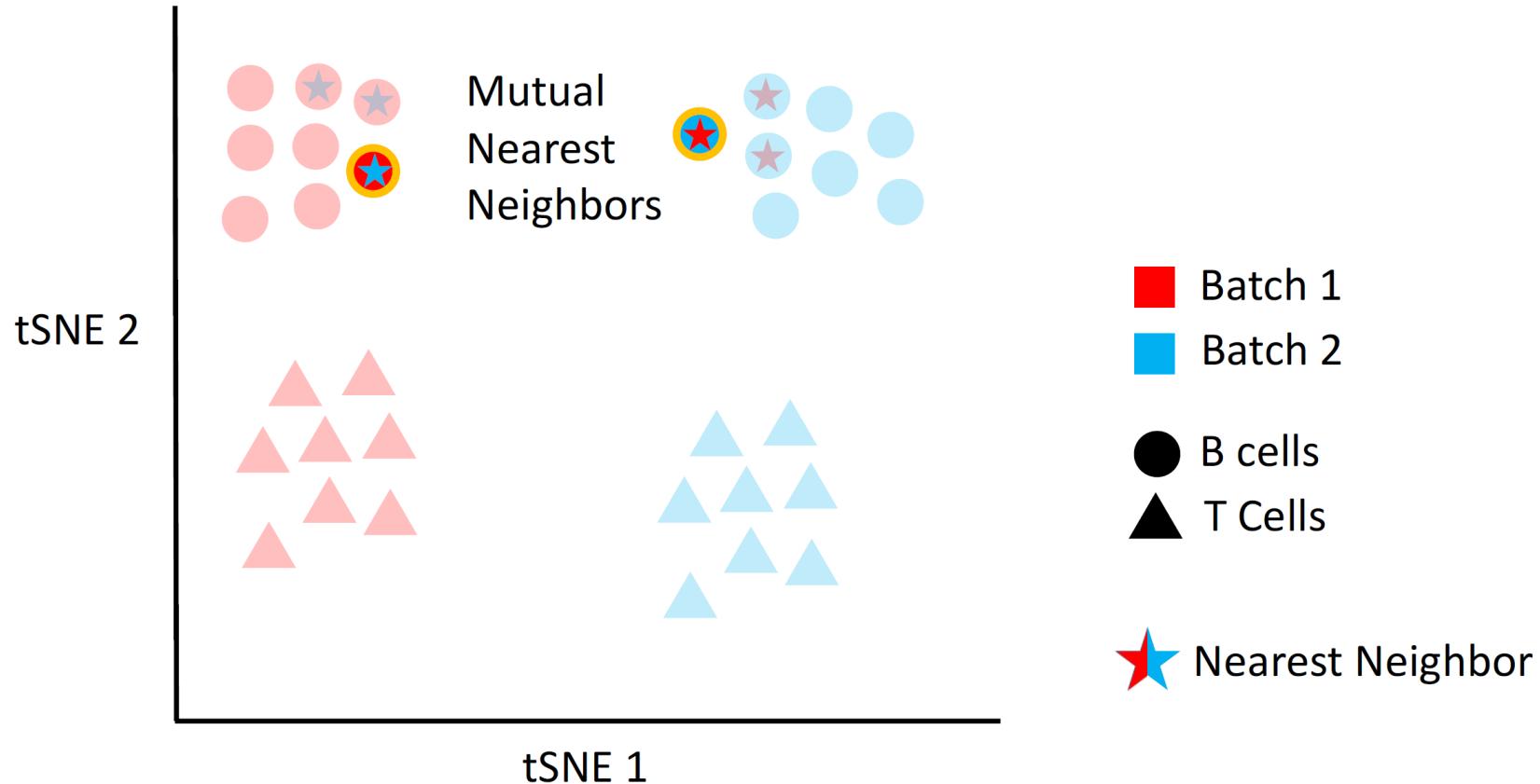
Mutual Nearest Neighbors (MNN)



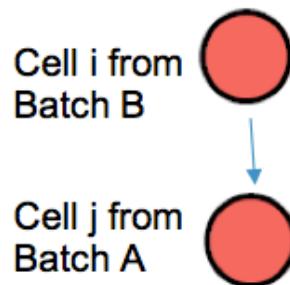
Mutual Nearest Neighbors (MNN)



Mutual Nearest Neighbors (MNN)

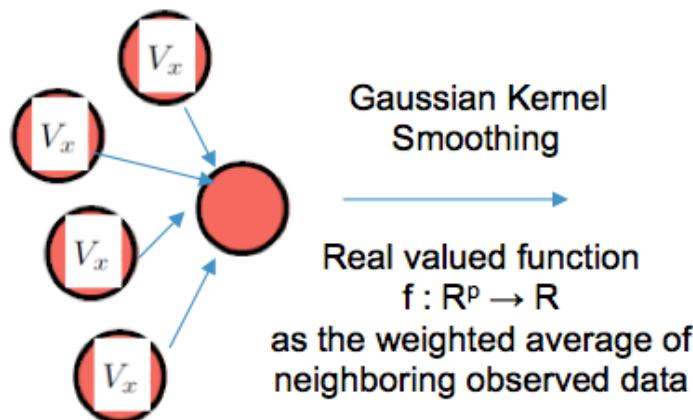


Mutual Nearest Neighbors (MNN)



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.

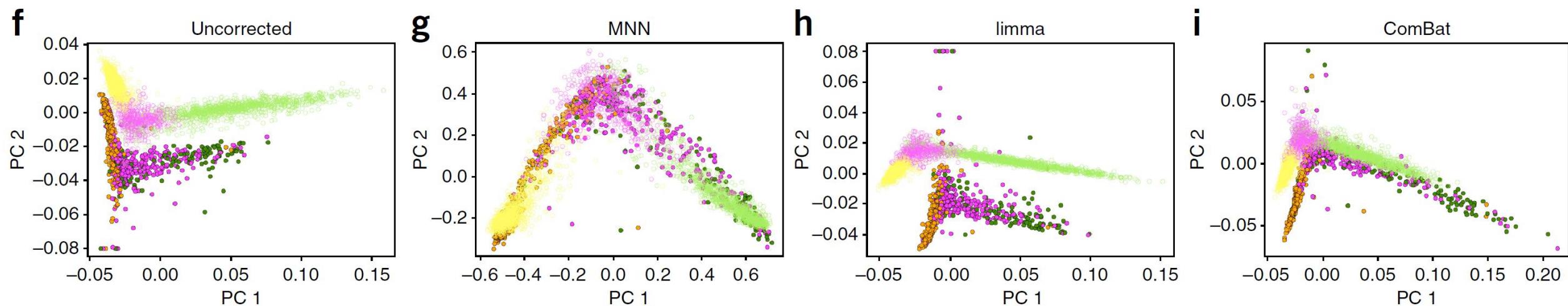


$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \dots \\ geneN_a - geneN_b \end{pmatrix}$$

Batch Correction vector for each cell



Mutual Nearest Neighbors (MNN)



SMART-seq2

- MEP
- GMP
- CMP

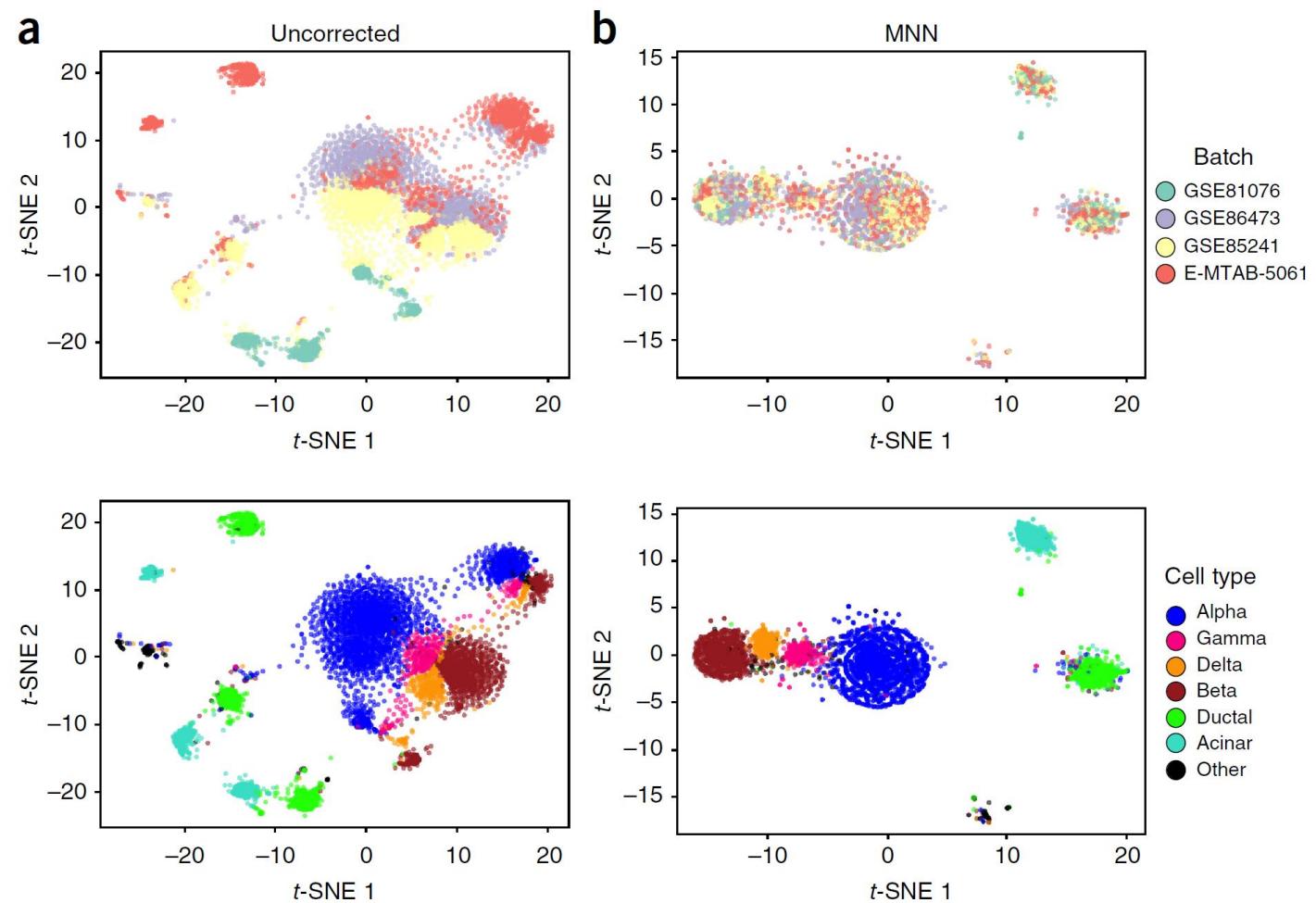
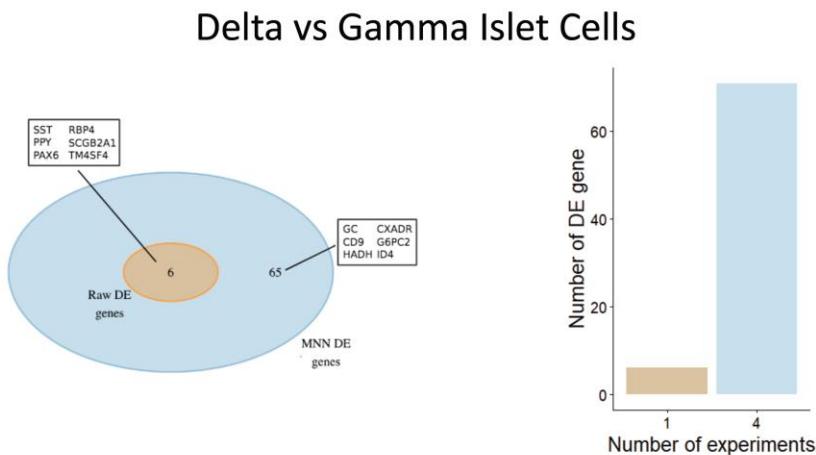
MARS-seq

- MEP
- GMP
- CMP

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

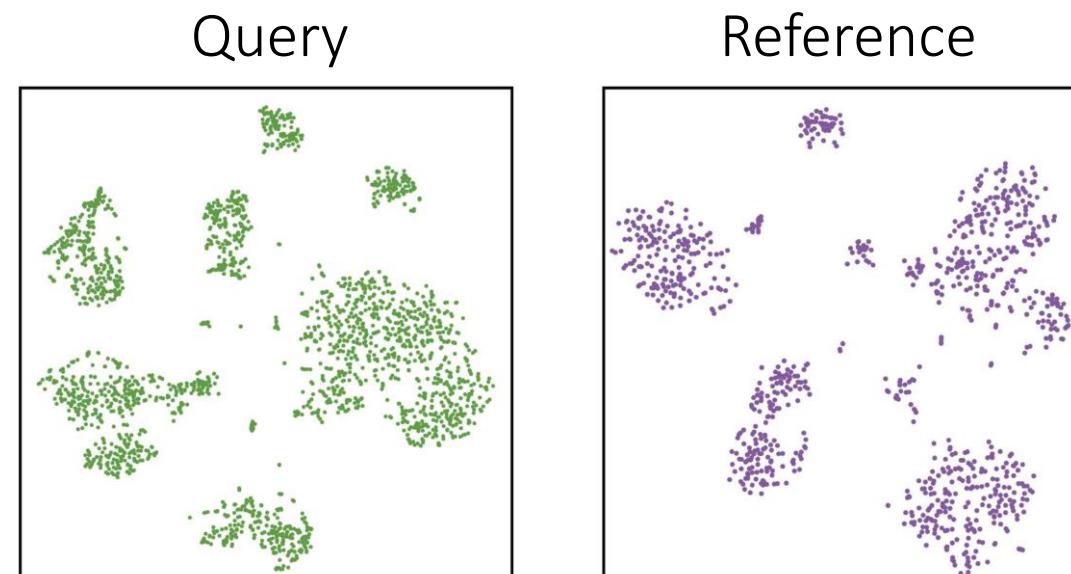
Mutual Nearest Neighbors (MNN)

- Pooling experiments -> increased statistical power

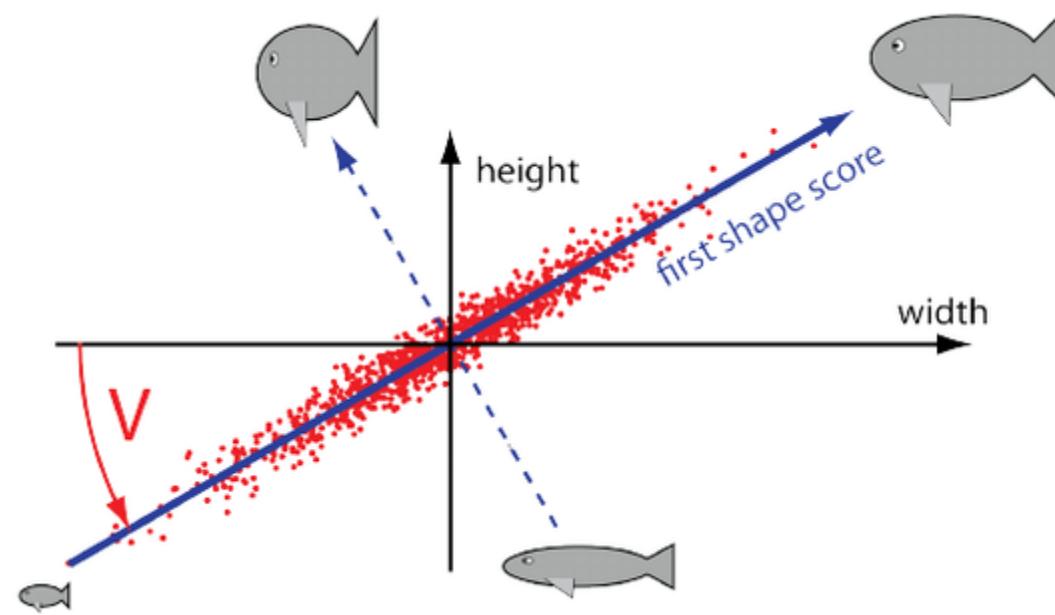
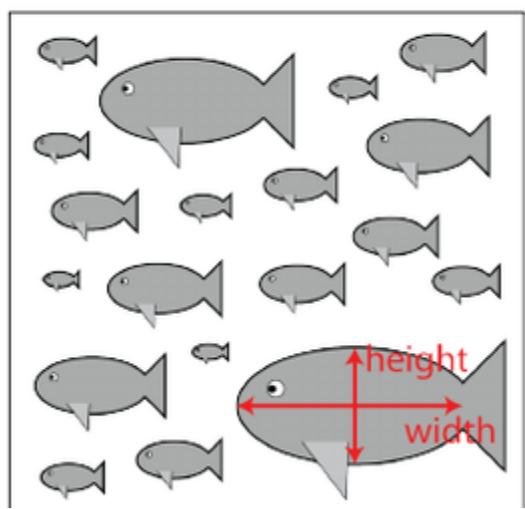


CCA + anchors (Seurat v3)

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

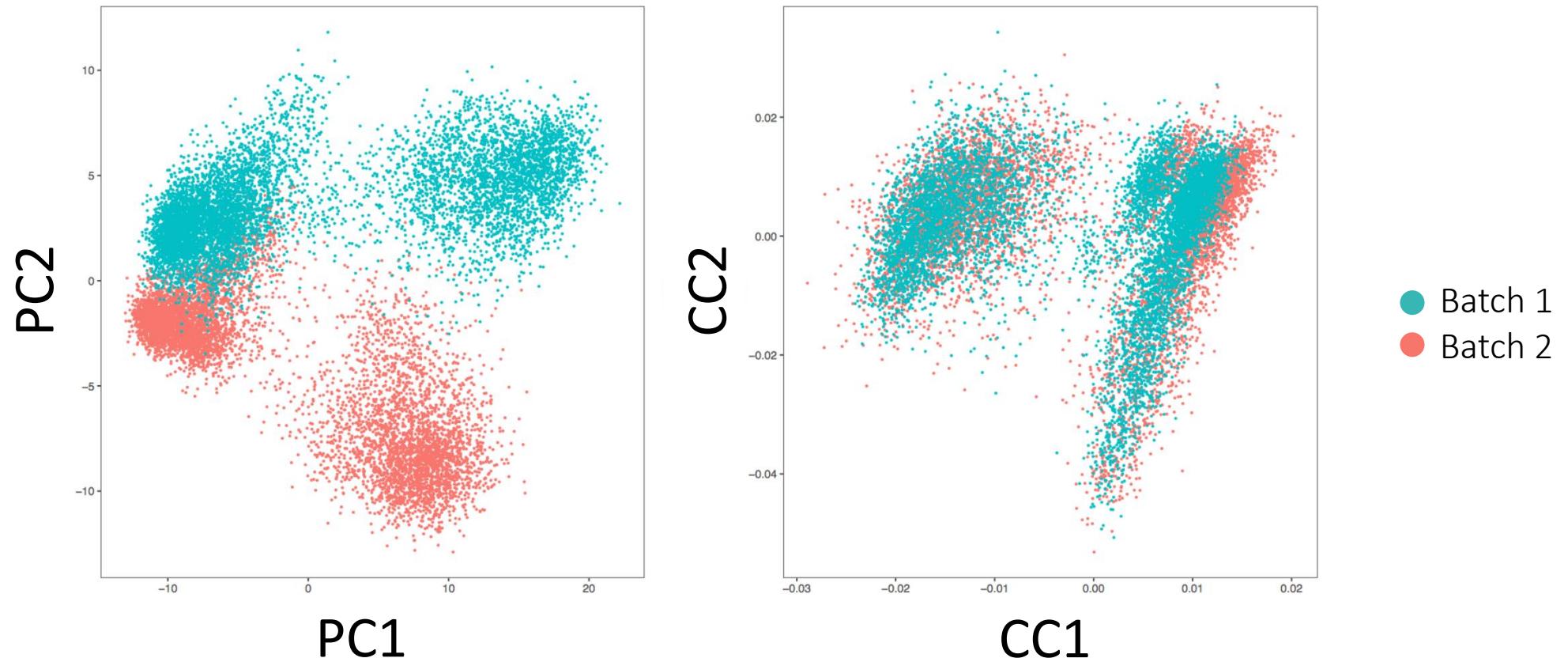


Principle component analysis



Finding corresponding cells

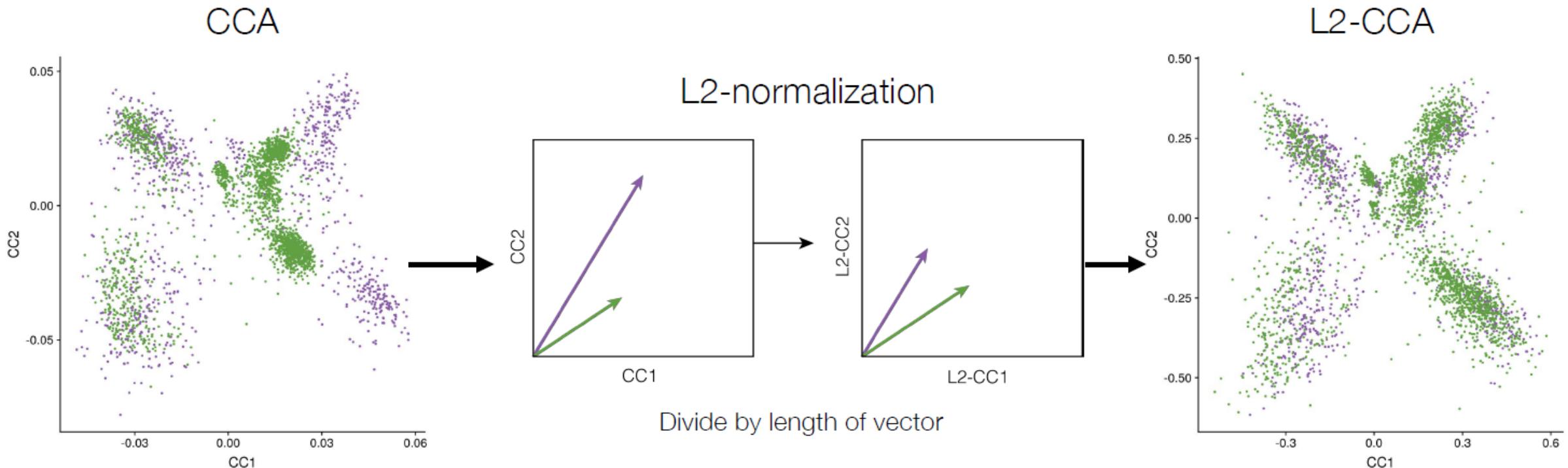
Canonical correlation analysis and normalization



CCA captures correlated sources of variation between two datasets

Finding corresponding cells

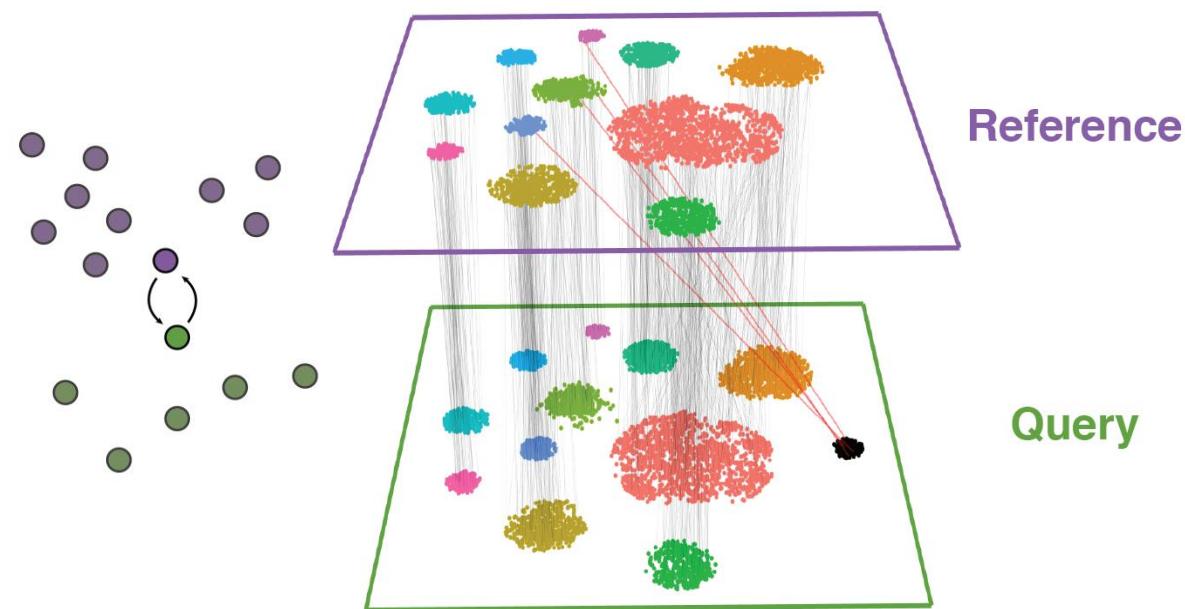
Canonical correlation analysis and normalization



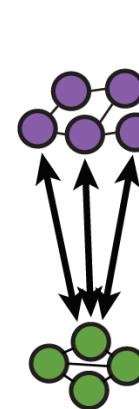
L2-normalization corrects for differences in scale

Finding corresponding cells

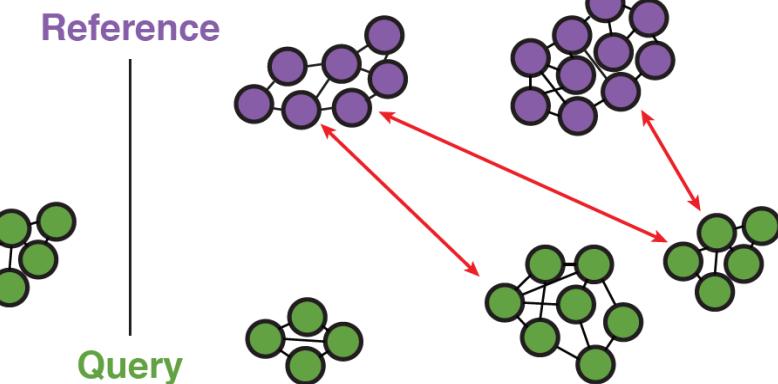
Anchors: mutual nearest neighbors



High-scoring correspondence
Anchors are consistent with local neighborhoods



Low-scoring correspondence
Anchors are inconsistent with local neighborhoods



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}

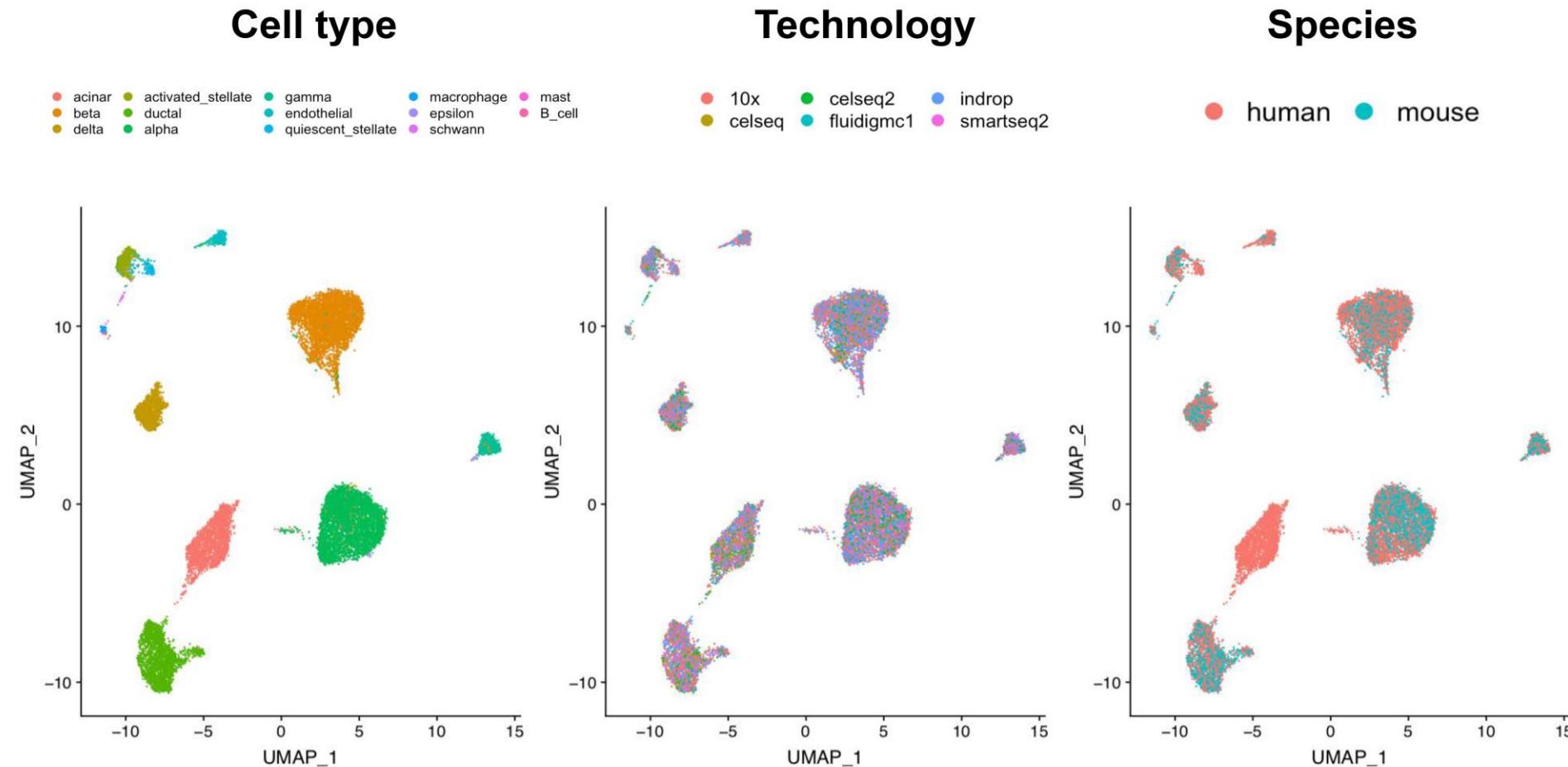
$$B = X[, a] - Y[, a]$$

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

$$C = BW^T$$

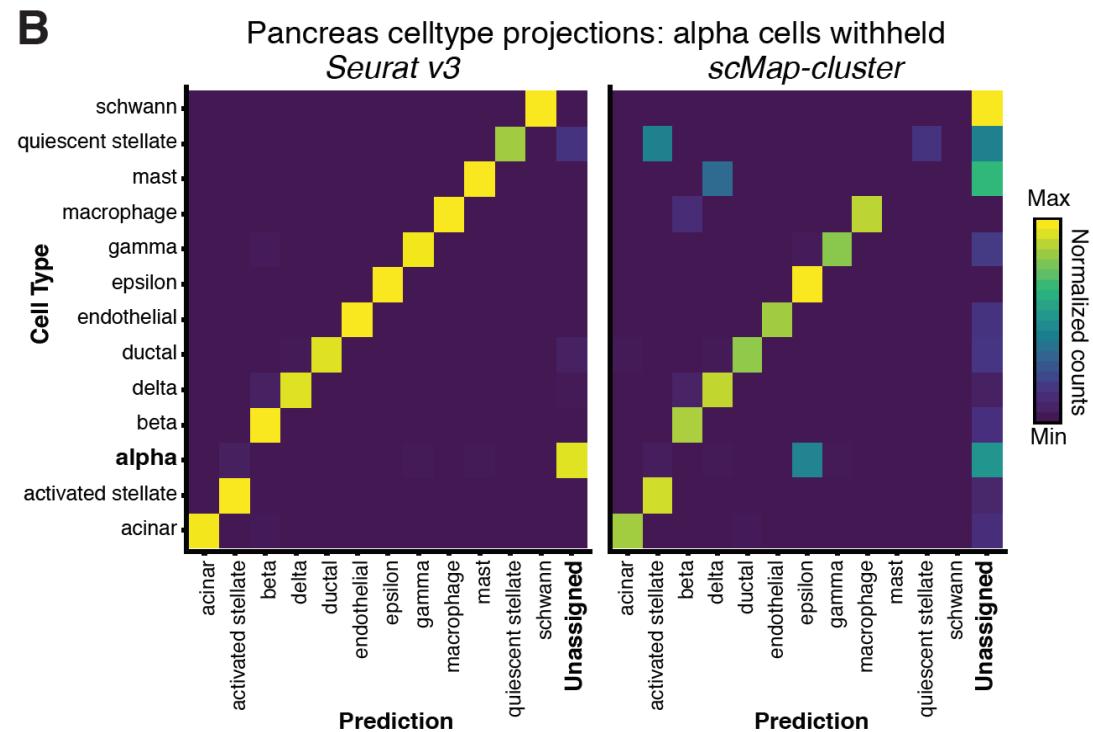
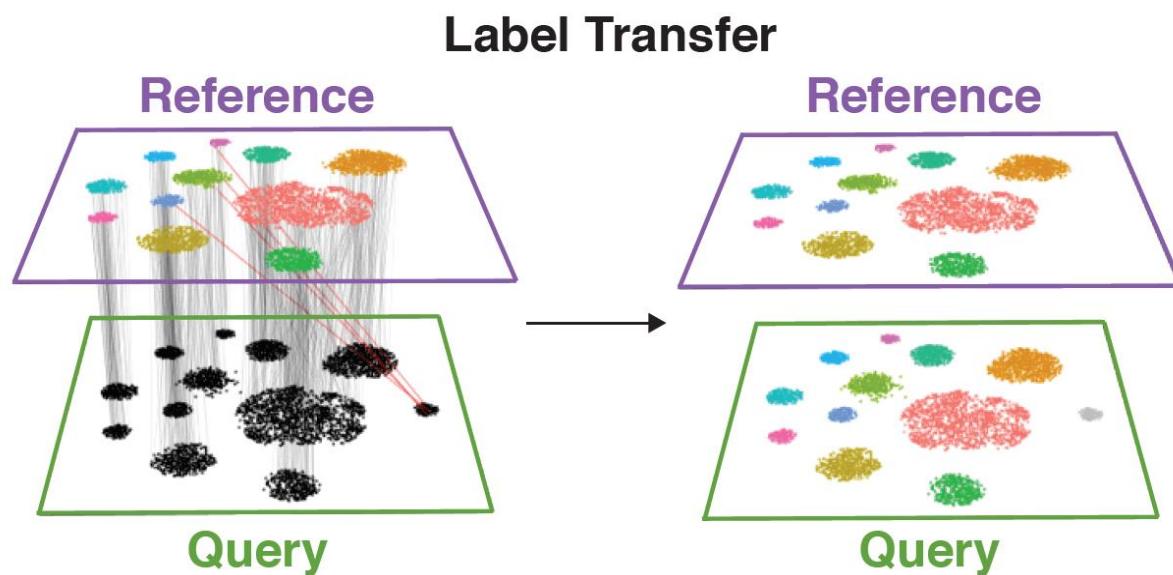
$$\hat{Y} = Y - C$$

CCA + anchors (Seurat v3)



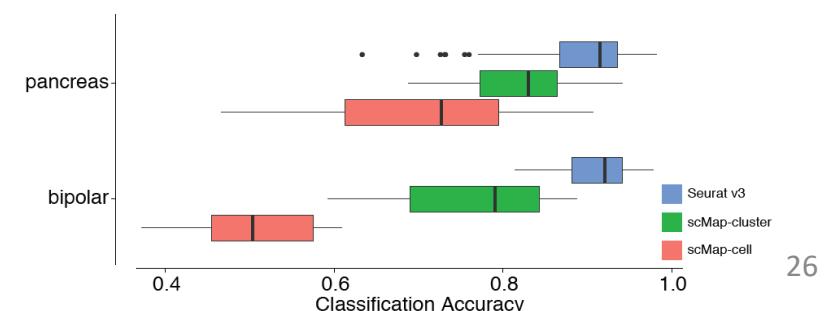
Retinal bipolar datasets: 51K cells, 6 technologies, 2 Species

Label transfer (classification)

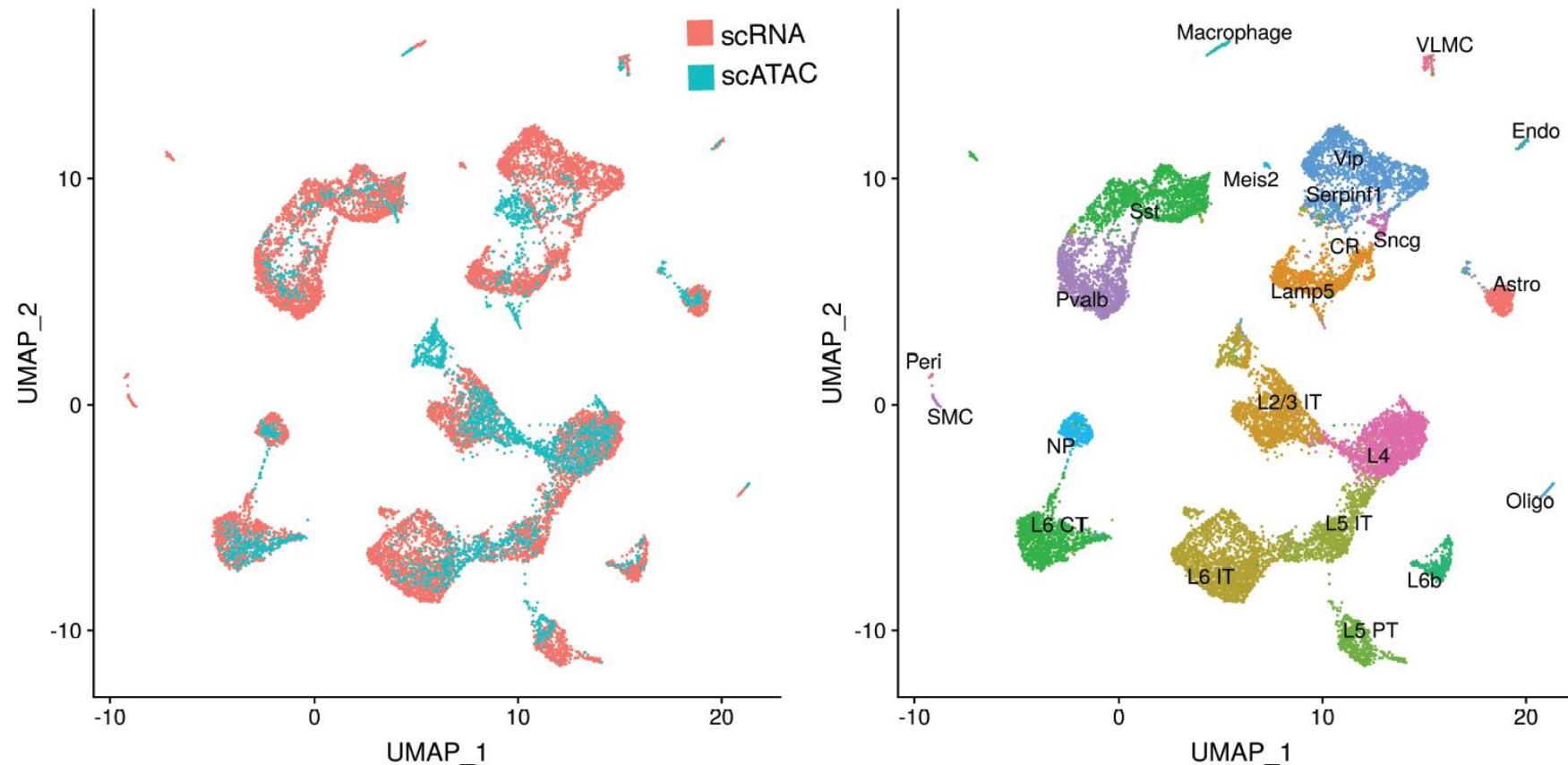


Weighted vote classifier

What is the classification of each cells nearest anchors?



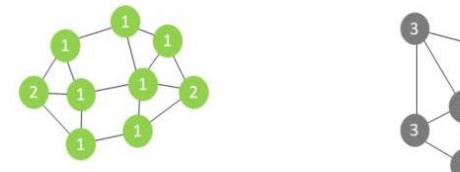
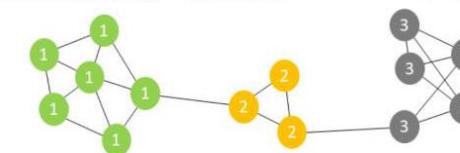
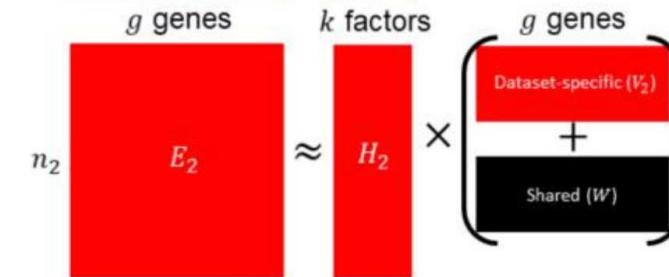
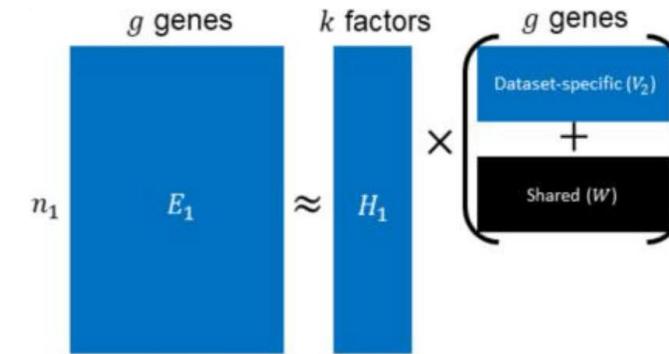
Integration across modalities



LIGER

Linked Inference of Genomic Experimental Relationships

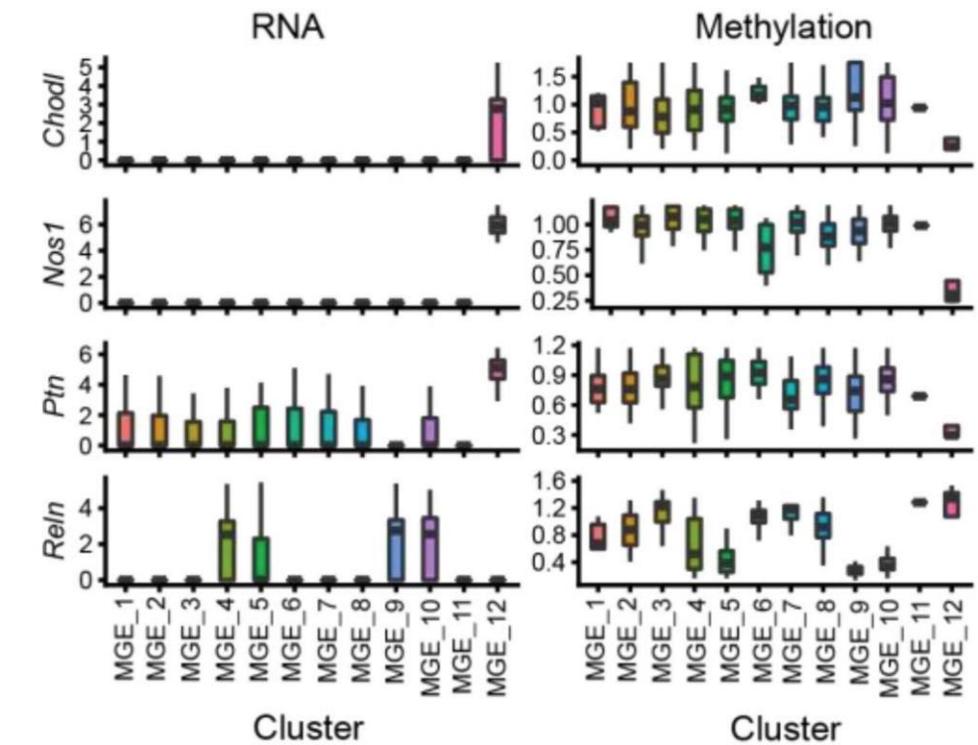
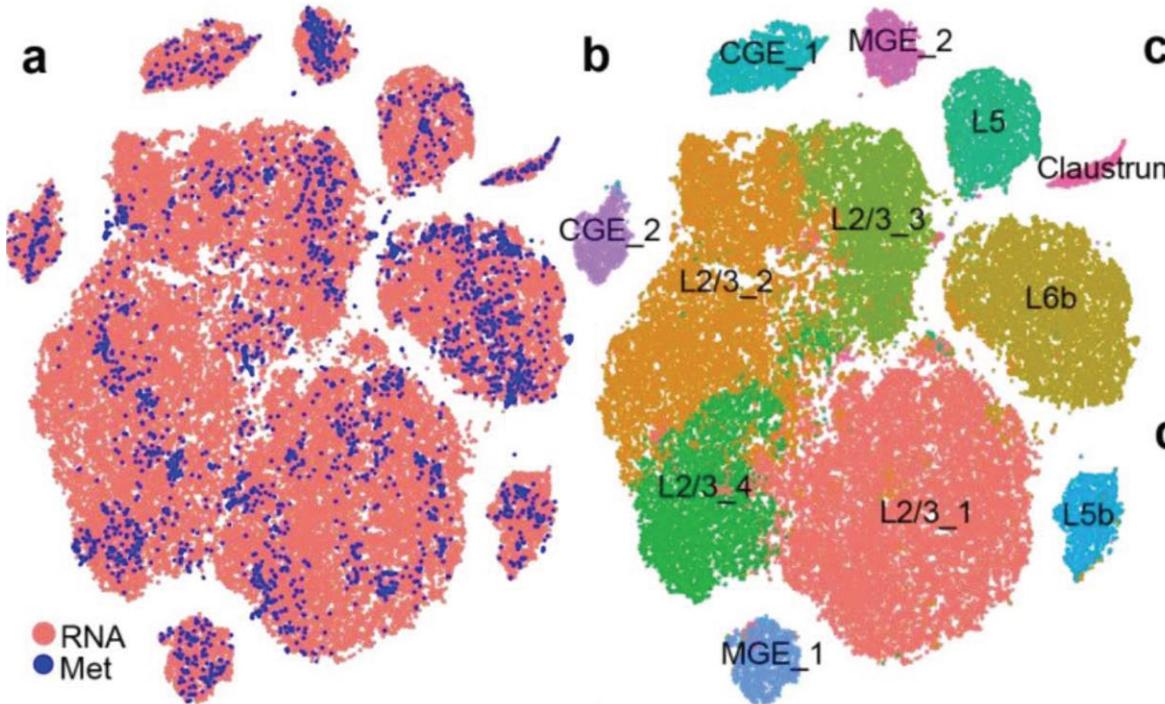
- 1) Integrative non-negative matrix factorization (iNMF) to learn a shared low-dimensional space
- 2) Perform joint clustering on the shared factor neighborhood graph
 - Factors are interpretable due to non-negative constraint
 - Finds set of dataset-specific factors and a set of shared factors



LIGER

Linked Inference of Genomic Experimental Relationships

- Joint clustering of gene expression and DNA methylation data



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:

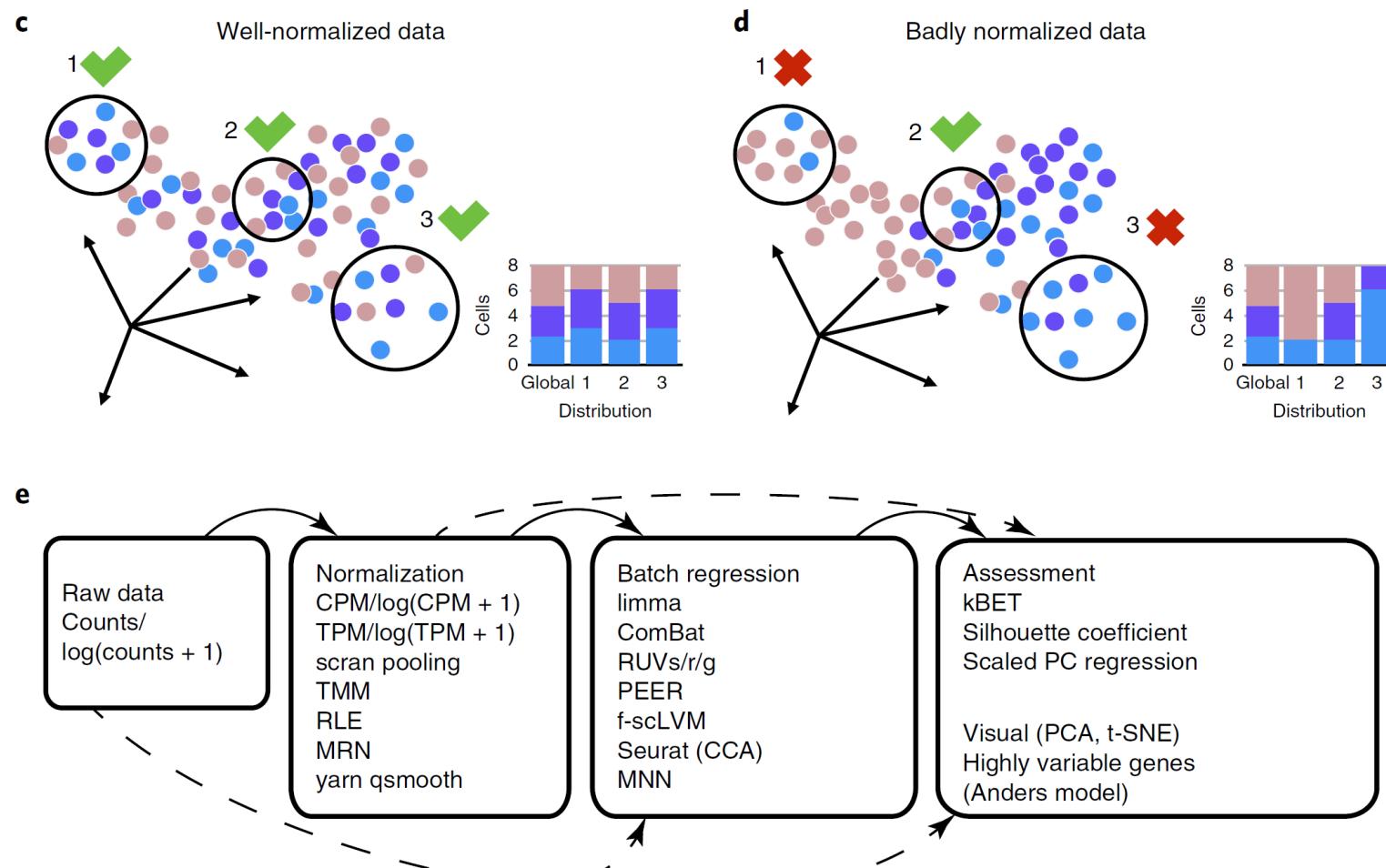
$$S = \frac{1}{N} \sum s(i)$$

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

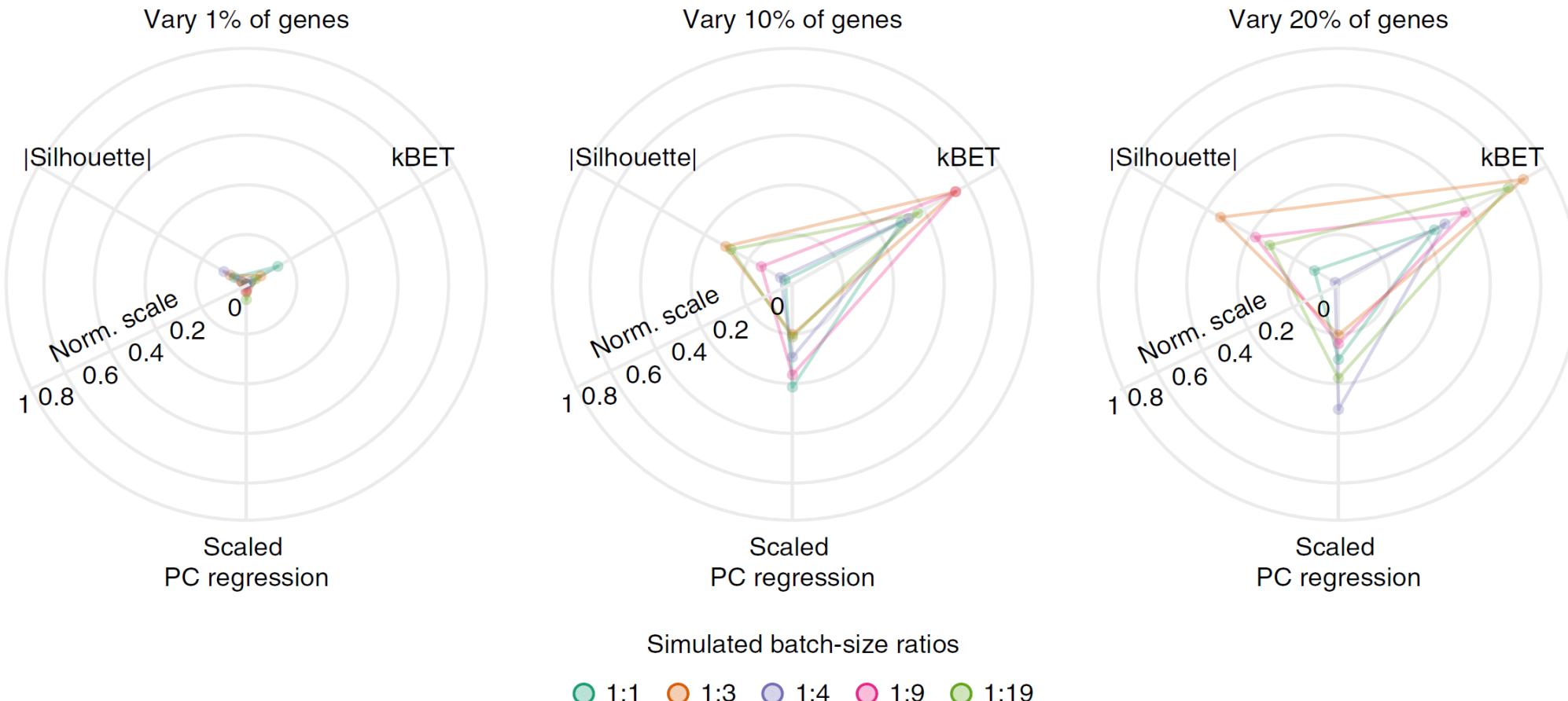
$$a(i) = \frac{1}{|C_i|} \sum_{\forall j} d(x_i, x_j)$$

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

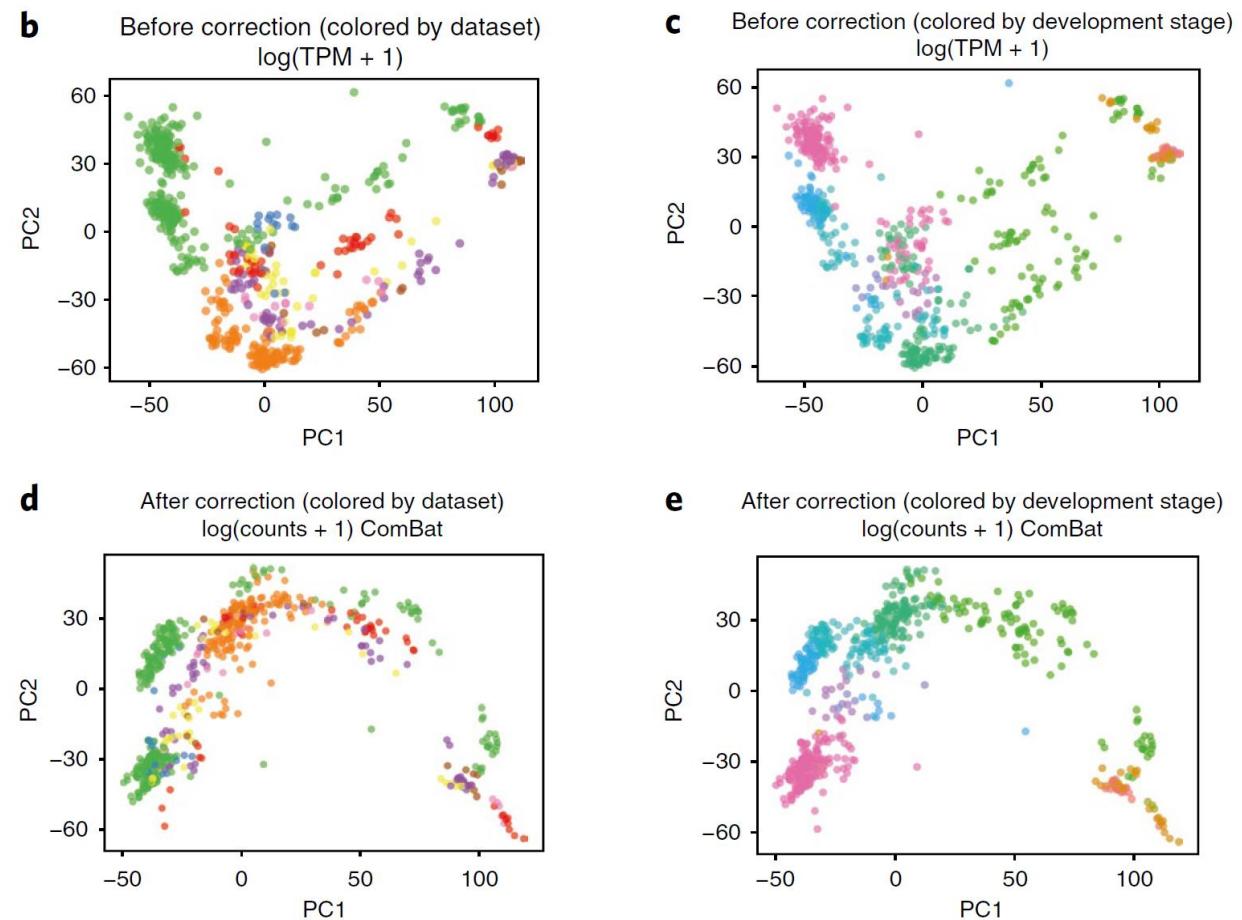
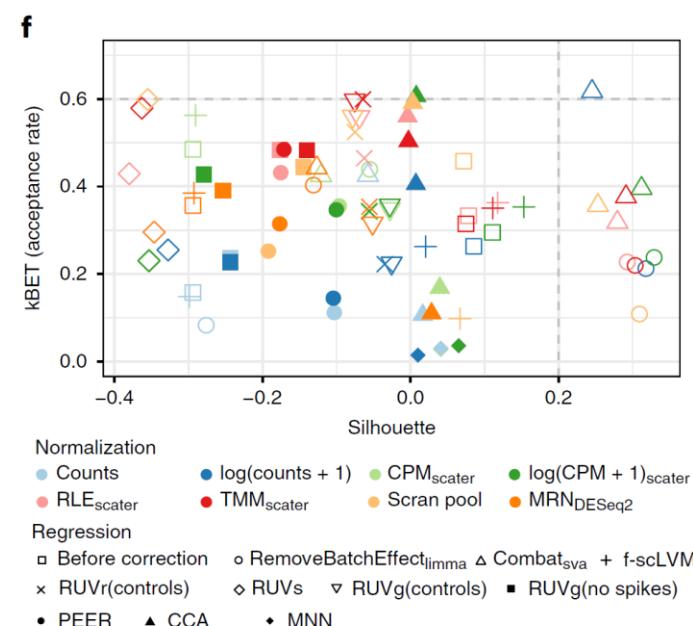
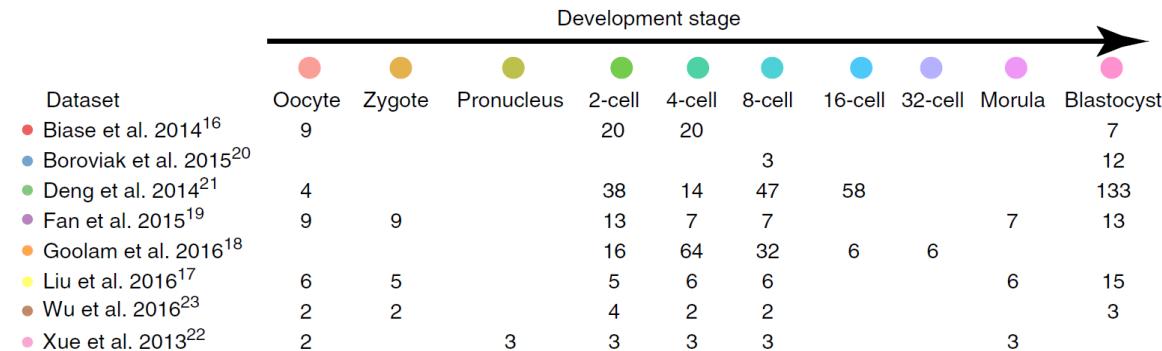
kBET: k -nearest-neighbor batch-effect test



kBET is more responsive than other batch tests on simulated data



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
- Joint dimension reduction can yield **interpretable factors** and aid in the identification of equivalent states, but is computationally expensive
- Graph-based methods alone can be **extremely fast**, but may struggle when technical differences are on a similar scale to biological differences

Sample multiplexing

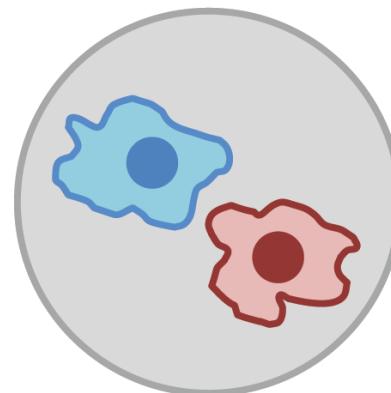
- To simultaneously measure cells combined from different samples/conditions/...
 - Pool many cells together in the same run
 - Mitigates technical effects
 - Able to identify conditions from output data

Multiplexing solves few problems

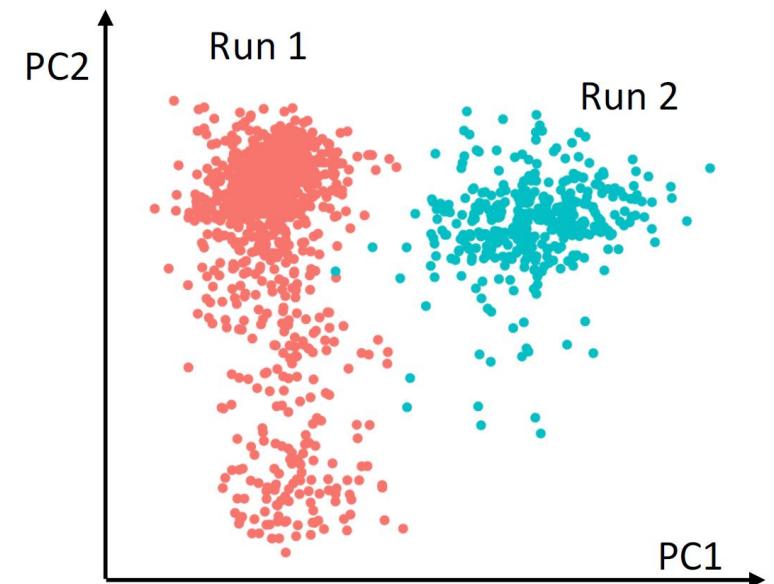
Cost



Doublets

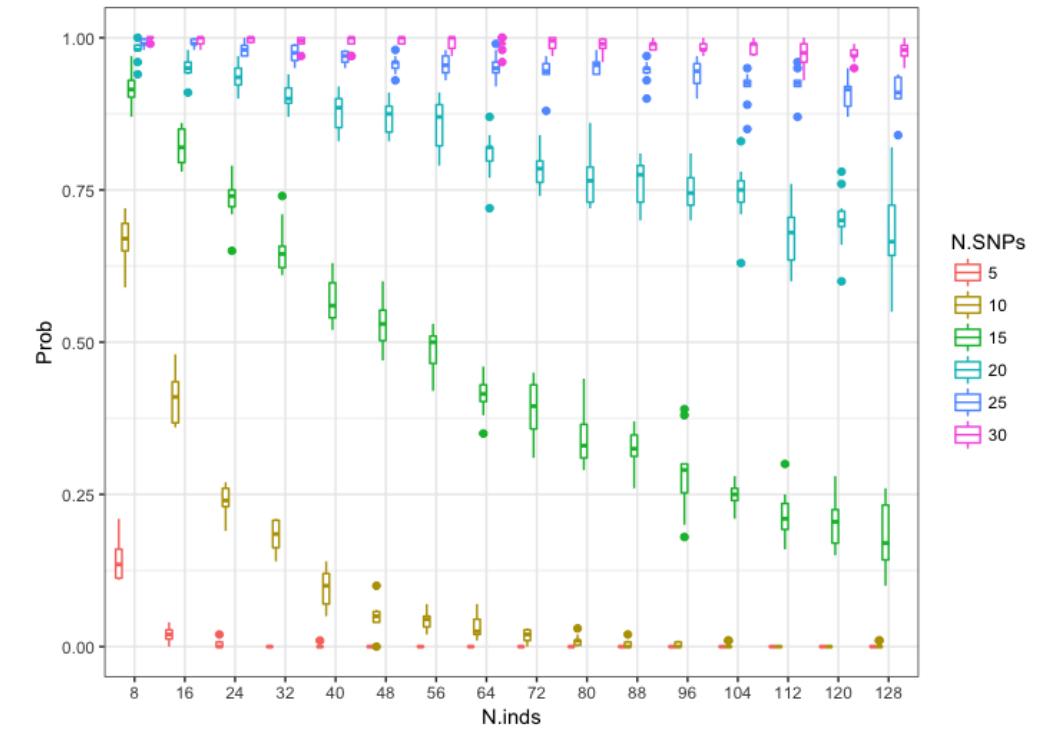
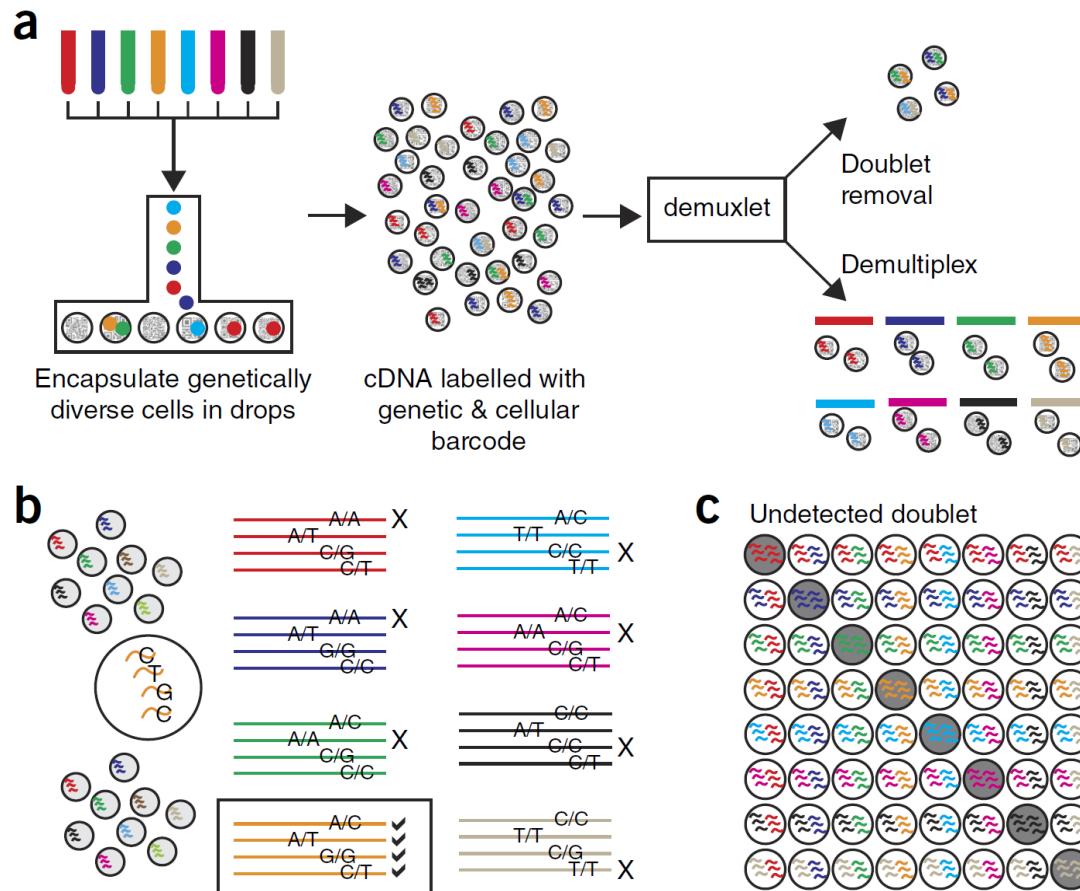


Batch effects

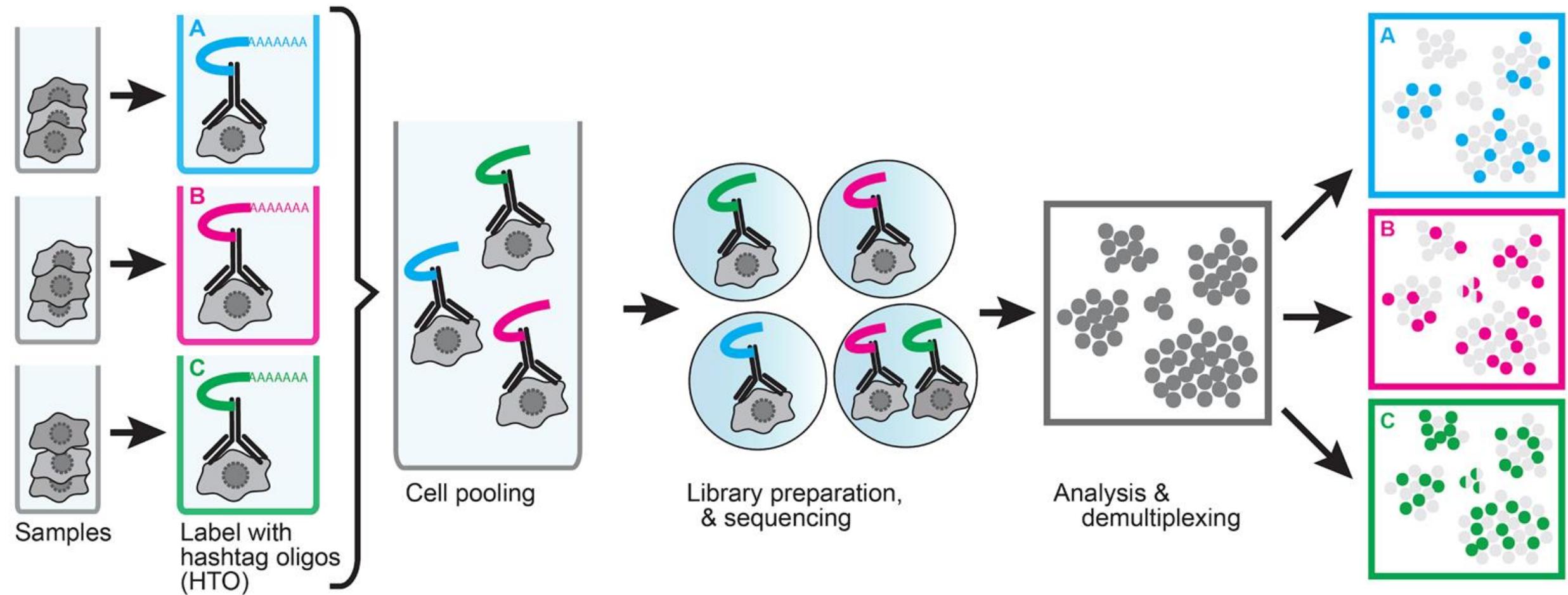


Demuxlet

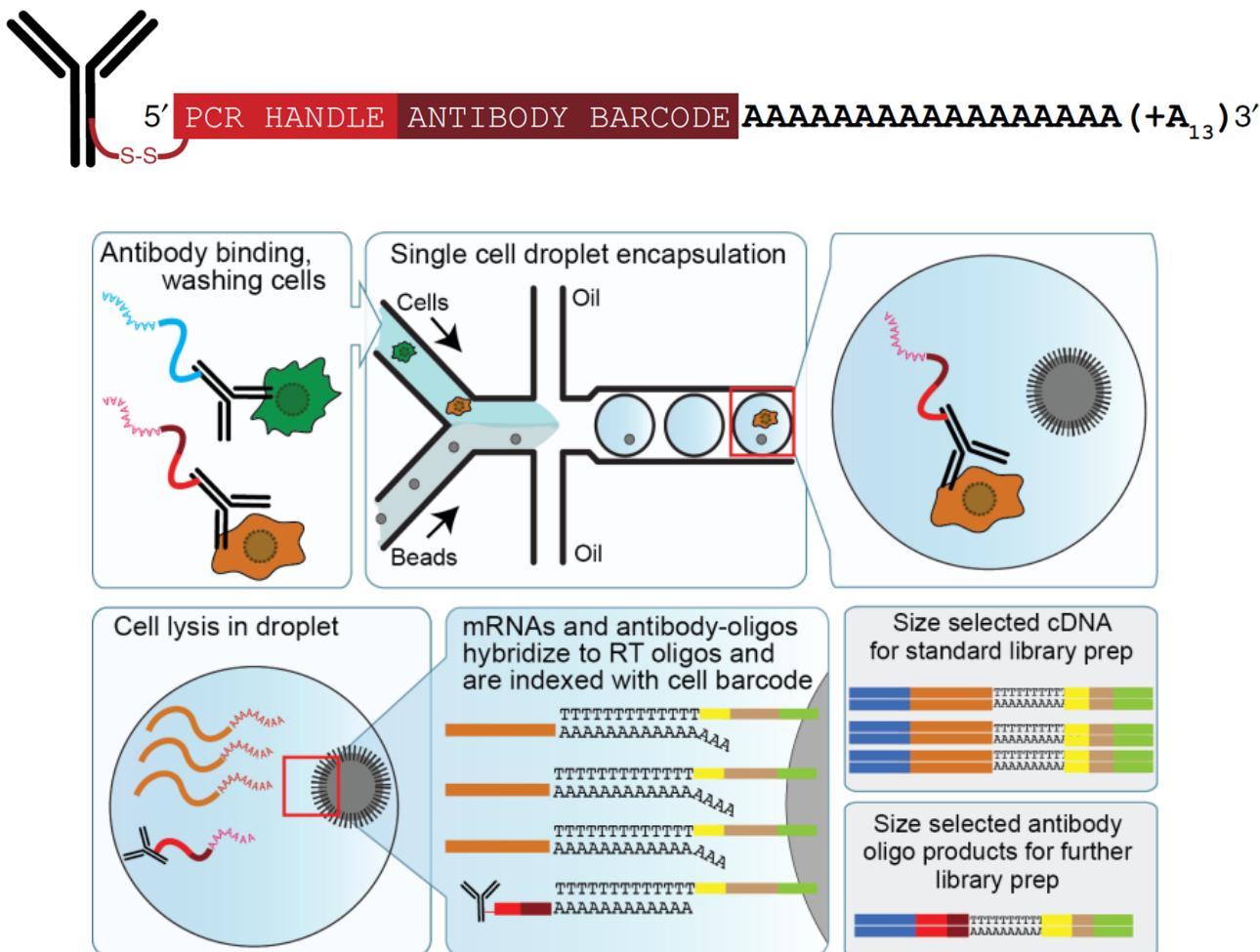
Natural SNPs



Cell hashing with barcoded antibodies

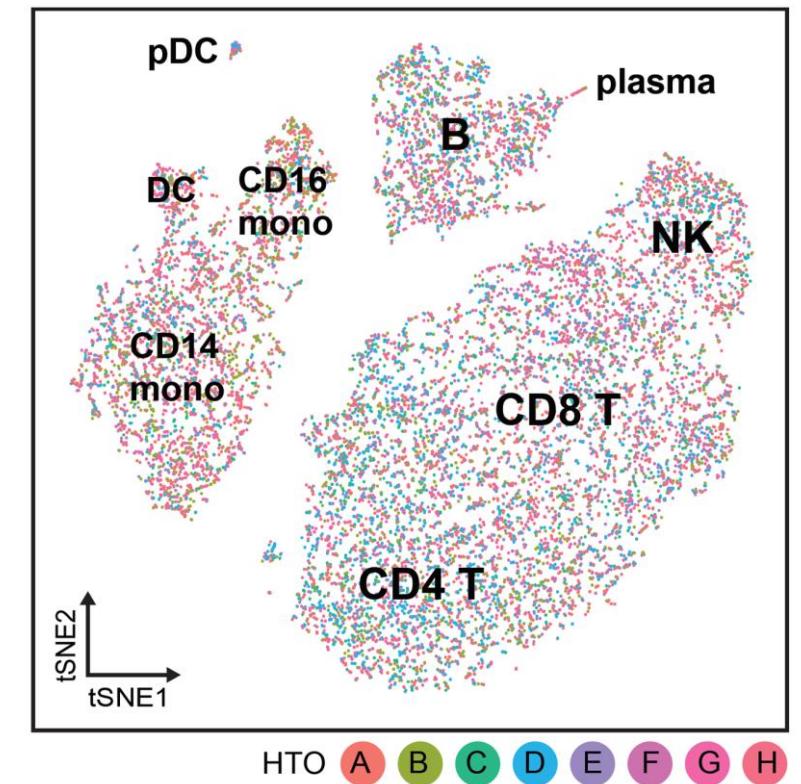
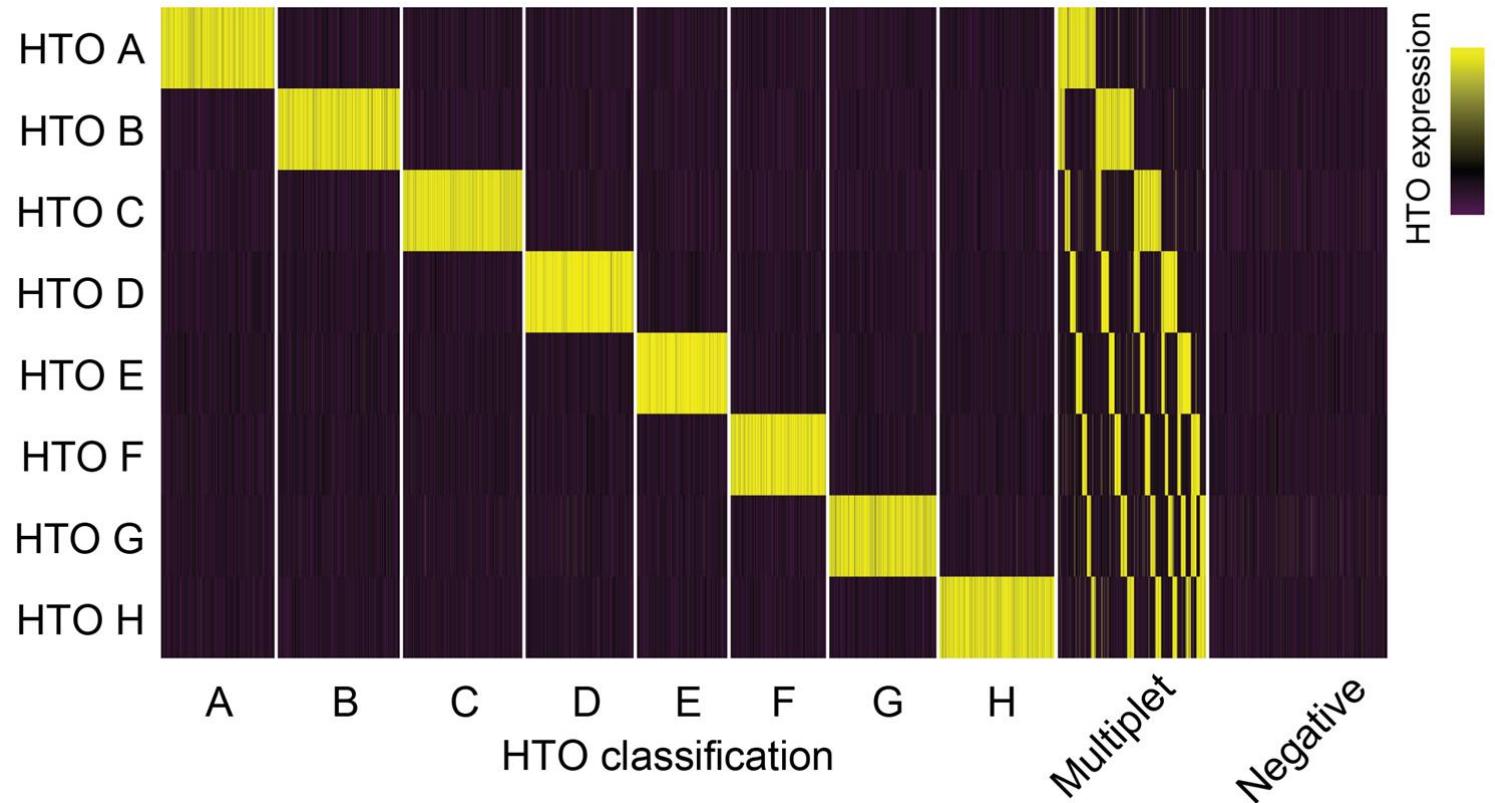


DNA-barcoded antibodies



Cell hashing with barcoded antibodies

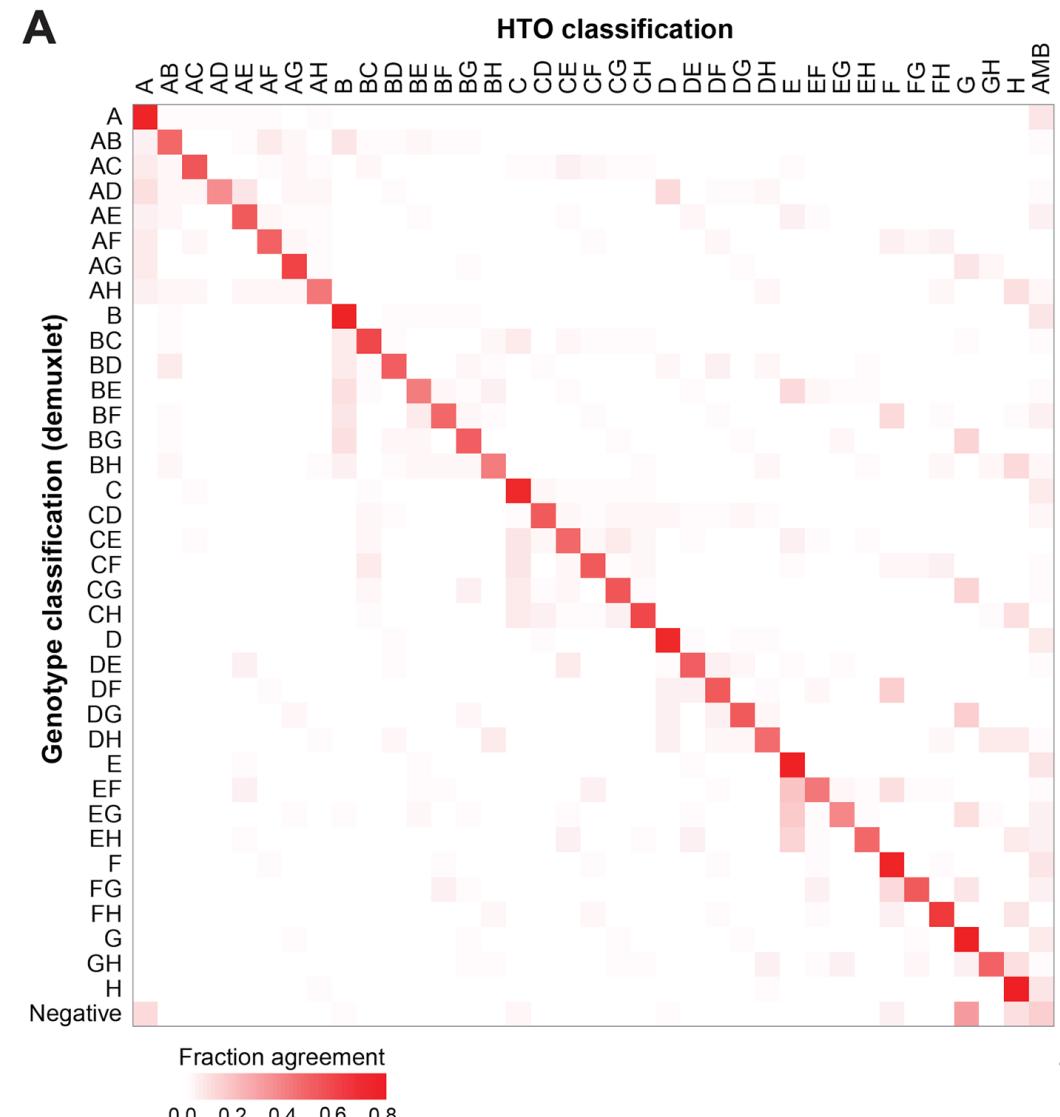
Equal concentration mixing of PBMCs from eight human donors A-H



Cell hashing with barcoded antibodies

Validation by comparison to *Demuxlet*

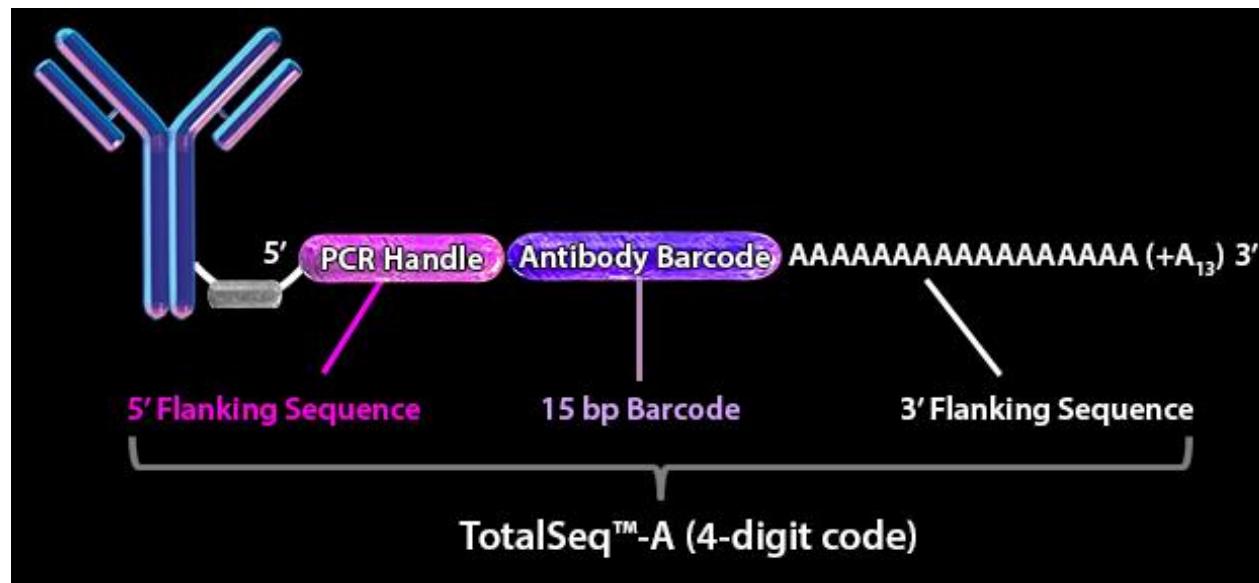
- Using genetic variations (SNPs) to determine the sources of cells (individuals)



Cell hashing with barcoded antibodies

Commercialized by BioLegend (TotalSeq™)

- Human: hashtags are made of two antibodies, CD298 and β2 microglobulin
- Mouse: hashtags are made of two antibodies, CD45 and H-2 MHC class I



Multiplexed quantification of proteins and transcripts in single cells

Vanessa M Peterson^{1,5}, Kelvin Xi Zhang^{2,5}, Namit Kumar¹, Jerelyn Wong³, Lixia Li¹, Douglas C Wilson³, Renee Moore⁴, Terrill K McClanahan³, Svetlana Sadekova³ & Joel A Klappenbach¹

We present a tool to measure gene and protein expression levels in single cells with DNA-labeled antibodies and droplet microfluidics. Using the RNA expression and protein sequencing assay (REAP-seq), we quantified proteins with 82 barcoded antibodies and >20,000 genes in a single workflow. We used REAP-seq to assess the costimulatory effects of a CD27 agonist on human CD8⁺ lymphocytes and to identify and characterize an unknown cell type.

Recent increases in the throughput of single-cell (sc) RNA-seq^{1,2} experimentation has enabled its use in the identification and characterization of novel or rare cell types³, in addition to providing insights into the underlying mechanisms of cellular development⁴ and the response to therapeutic interventions⁵. However, proteins, not mRNAs, are the primary targets of drugs, and protein abundance cannot necessarily be inferred directly from mRNA abundance^{6–9}. An unbiased view of proteins is thus necessary to model cellular dynamics and response to environmental and therapeutic perturbations.

REAP-seq enables simultaneous measurement of proteins and mRNAs in single cells. Cells are labeled via methods similar to standard flow cytometry methods but with antibodies conjugated to DNA barcodes instead of fluorophores. This removes the limitations imposed by spectral overlap of fluorescent labels (-17) (ref. 10) or the available number of stable isotopes (-40) (ref. 11), in flow and mass cytometry. Using sequencing as a readout instead of qPCR¹², a DNA barcode of eight nucleotides provides up to 65,536 unique indices (B^n , where $B =$ any of the four bases GATC, and $n =$ length of the nucleotide sequence). In addition to the unique 8-bp barcode, the antibody DNA label consists of a poly (dA) sequence for priming to the cell barcode and a universal sequence for amplification (Supplementary Figs. 1–3 and Supplementary Discussion). Excess unbound antibody barcodes (AbBs) are washed from the labeled cells before they are processed using

the standard 10x Genomics single-cell (sc)RNA-seq platform³, which is a droplet-based system designed for 3' digital counting of mRNA in thousands of single cells.

REAP-seq leverages the DNA polymerase activity of the reverse transcriptase to simultaneously extend the primed AbB with the poly(dT) cell barcode and synthesize complementary DNA from mRNA in the same reaction. Exonuclease I is then used to degrade any excess unbound single-stranded oligonucleotides from the protein double-stranded (ds) DNA (-155 bp) products to prevent crosstalk between AbBs and cell barcodes from different cells (Supplementary Fig. 4). Dextran sulfate was added to AbB labeling buffer to reduce non-specific binding of negatively charged DNA barcodes to the cell surface and isotype controls (Mouse IgG1, Mouse IgG2a, Mouse IgG2b, Rat IgG1, Rat IgG2a) were used to determine the threshold of background noise (Supplementary Figs. 5 and 6).

To initially test REAP-seq, we stained peripheral blood mononuclear cells (PBMCs) with a mixture of 45 AbBs (Fig. 1 and Supplementary Tables 1 and 2) and then magnetically enriched for three populations of cells: CD3⁺ T cells, CD11b⁺ myeloid cells, and CD19⁺ B cells (Supplementary Fig. 7). Cell barcodes identified in both gene and protein expression matrices were filtered for cells with a mitochondrial read rate of $<20\%$ and >250 genes expressed (3,797 CD3⁺, 2,883 CD11b⁺, 1,533 CD19⁺ cells, and 7,271 PBMCs). We used the nonlinear dimensionality reduction method t-distributed stochastic neighbor embedding (t-SNE) to visualize the principal component analysis (PCA)-reduced data set in two-dimensional space¹³ where the cells were color-coded by cluster (Fig. 1a and Supplementary Fig. 7a). The cells were also colored by the magnetic beads used for isolation (CD3⁺, CD19⁺, CD11b⁺) (Supplementary Fig. 7b), which showed three easily discernible purified populations of cells, and was used as a positive control to assess the sensitivity and specificity of REAP-seq mRNA and protein measurements for canonical markers of these cell types (Supplementary Fig. 7c). Also as a control, scRNA-seq alone was run on PBMCs to ensure the protein assay has no effect on mRNA measurements (Supplementary Figs. 8 and 9).

Protein and mRNA expression of canonical markers for monocytes (CD11b, CD14, CD33), B cells (CD20, CD19), T cells (CD3, CD4, CD8), and natural killer (NK) cells (CD56, CD158e1) were projected on the mRNA-t-SNE plot to visualize expression across all PBMCs, and to assess the specificity and sensitivity of the protein and mRNA assays (Fig. 1b). For each marker, the Pearson correlation coefficient between mRNA and protein expression was calculated. The markers most highly correlated were HLA-DR ($R = 0.69$), CD20 ($R = 0.46$), and CD14 ($R = 0.51$), and these markers also had the highest levels of transcriptional expression (Supplementary Table 3). For CD4, the correlation between mRNA and protein was low, and we found it expressed both in monocytes and T cells, a finding we confirmed by flow cytometry, ruling out non-specific

¹Genetics & Pharmacogenomics, Department of Translational Medicine, Merck & Co., Inc., Boston, Massachusetts, USA. ²Informatics IT, Merck & Co., Inc., Boston, Massachusetts, USA. ³Department of Profiling and Expression, Merck & Co., Inc., Palo Alto, California, USA. ⁴Protein Sciences, Department of Biologics, Merck & Co., Inc., Boston, Massachusetts, USA. ⁵These authors contributed equally to this work. Correspondence should be addressed to V.M.P. (vanessa.peterson1@merck.com).

Received 14 June 2017; accepted 24 August 2017; published online 30 August 2017; doi:10.1038/nbt.3973

Simultaneous epitope and transcriptome measurement in single cells

Marlon Stoeckius¹ , Christoph Hafemeister¹ , William Stephenson¹ , Brian Houck-Loomis¹ , Pratip K Chattopadhyay² , Harold Swerdlow¹ , Rahul Satija^{1,3}  & Peter Smibert¹ 

High-throughput single-cell RNA sequencing has transformed our understanding of complex cell populations, but it does not provide phenotypic information such as cell-surface protein levels. Here, we describe cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), a method in which oligonucleotide-labeled antibodies are used to integrate cellular protein and transcriptome measurements into an efficient, single-cell readout. CITE-seq is compatible with existing single-cell sequencing approaches and scales readily with throughput increases.

The unbiased and high-throughput nature of modern single-cell RNA-seq (scRNA-seq) approaches has proven invaluable for describing heterogeneous cell populations^{1–3}. Prior to single-cell genomics, cellular states were routinely described using curated panels of fluorescently labeled antibodies directed at cell-surface proteins, which are often reliable indicators of cellular activity and function⁴. Recent studies^{5,6} have demonstrated the potential for coupling ‘index-sorting’ measurements from a cell sorter with single-cell transcriptomics; this process allows immunophenotypes to be mapped onto transcriptionally derived clusters. However, massively parallel approaches based on droplet microfluidics^{4–6}, microwells^{7,8} or combinatorial indexing^{9,10} are incompatible with cytometry and therefore cannot be augmented with protein information. Targeted methods to simultaneously measure transcripts and proteins in single cells are limited in scale or can only profile a few genes and proteins in parallel^{11–15} (Supplementary Table 1).

Here, we describe CITE-seq, a method that combines highly multiplexed protein marker detection with unbiased transcriptome profiling for thousands of single cells. We demonstrate that the method is readily adaptable to two high-throughput scRNA-seq applications and show that multimodal data analysis can achieve a more detailed characterization of cellular phenotypes than transcriptome measurements alone.

We sought to characterize the quantitative nature of the CITE-seq protein readout. Flow cytometry is the gold standard for

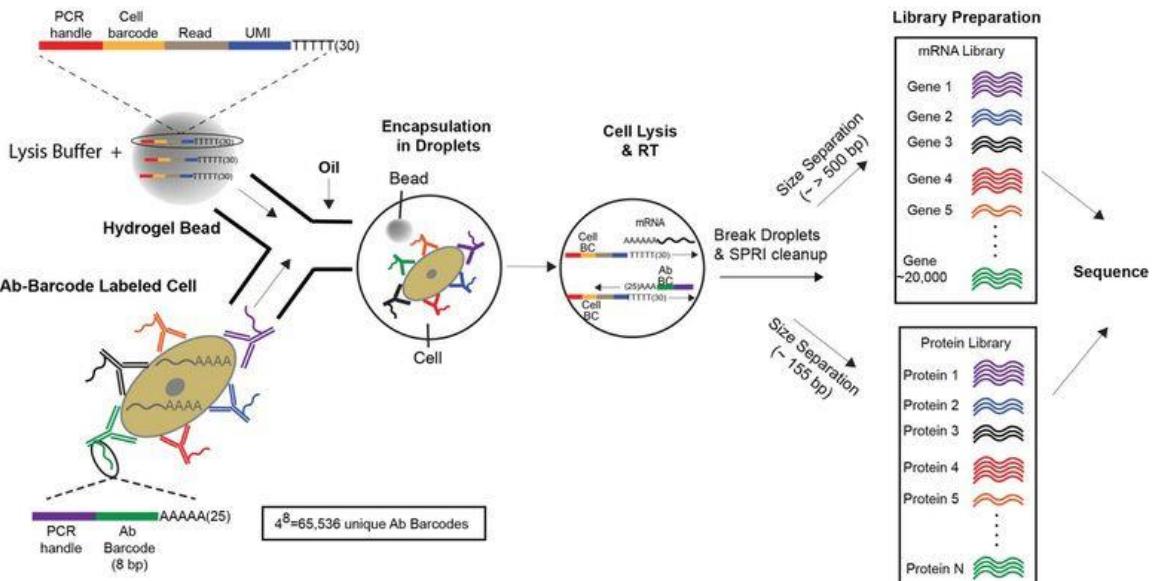
¹New York Genome Center, New York, New York, USA. ²New York University Medical Center, New York, New York, USA. ³New York University Center for Genomics and Systems Biology, New York, New York, USA. Correspondence should be addressed to M.S. (mstoeckus@nygenome.org).

RECEIVED 2 MARCH; ACCEPTED 7 JULY; PUBLISHED ONLINE 31 JULY 2017; DOI:10.1038/NMETH.4380

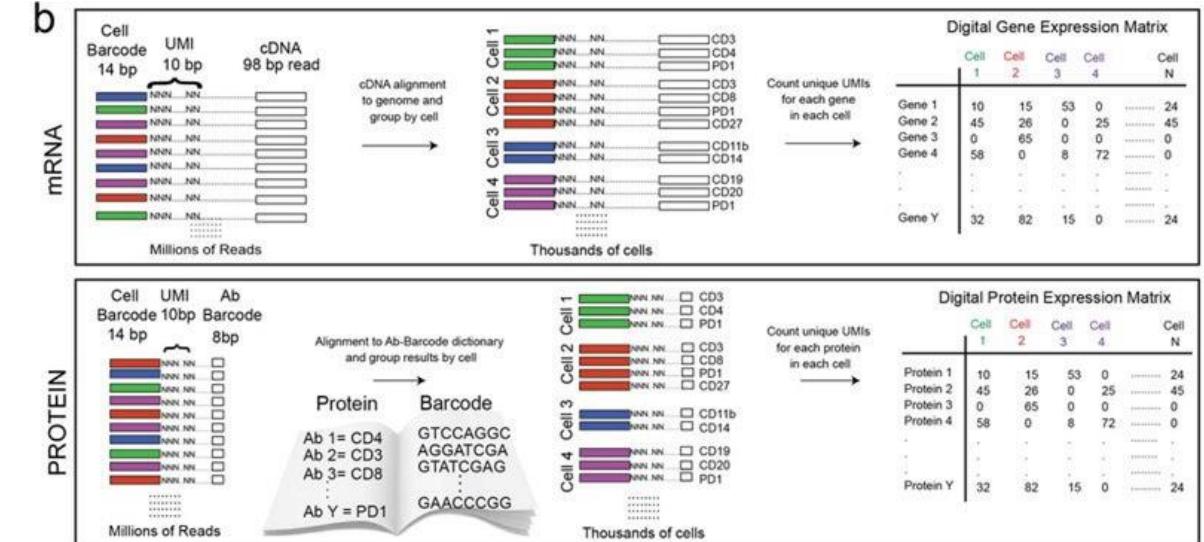
REAP-seq

RNA expression and protein sequencing assay

a

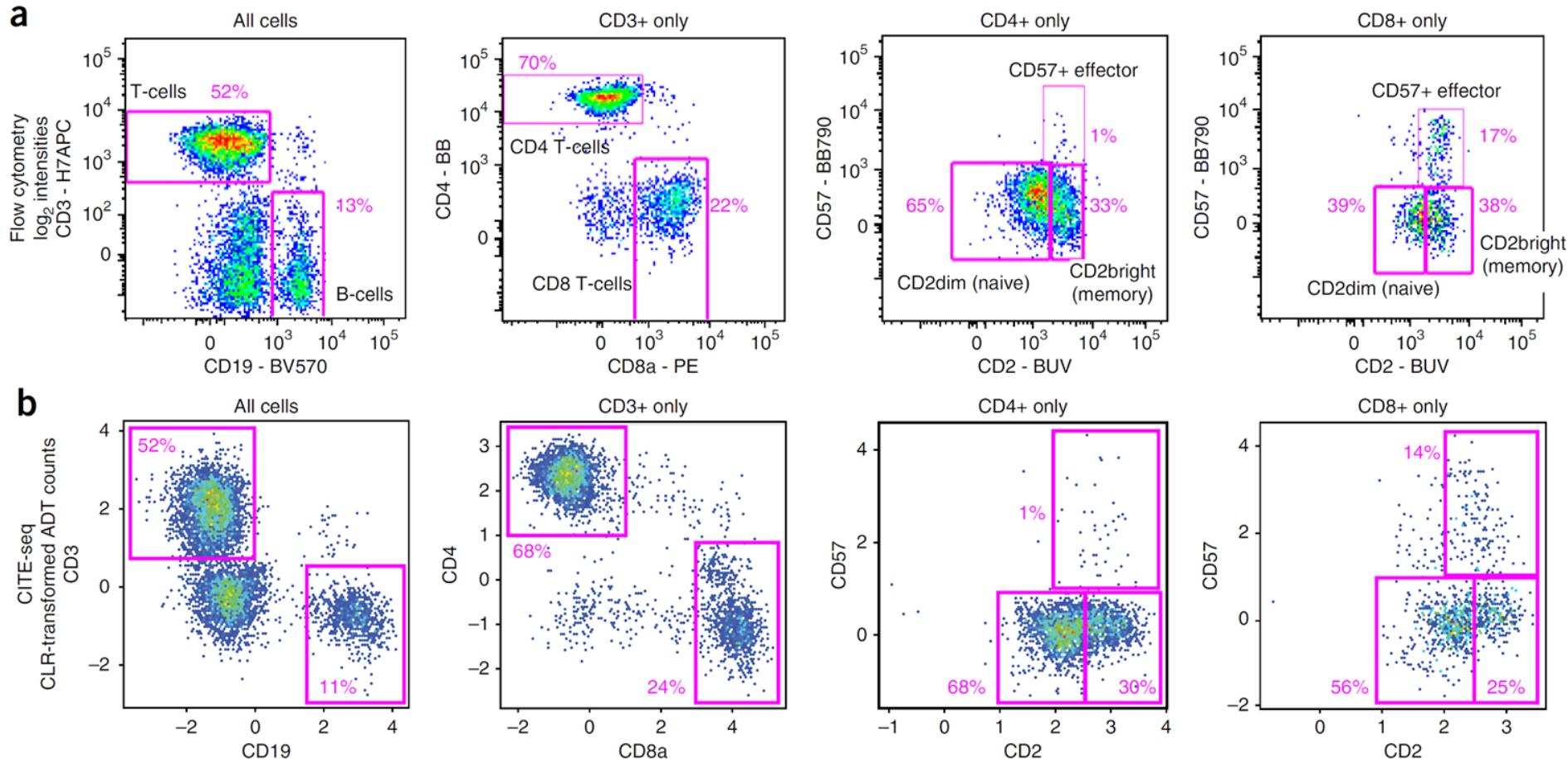


b



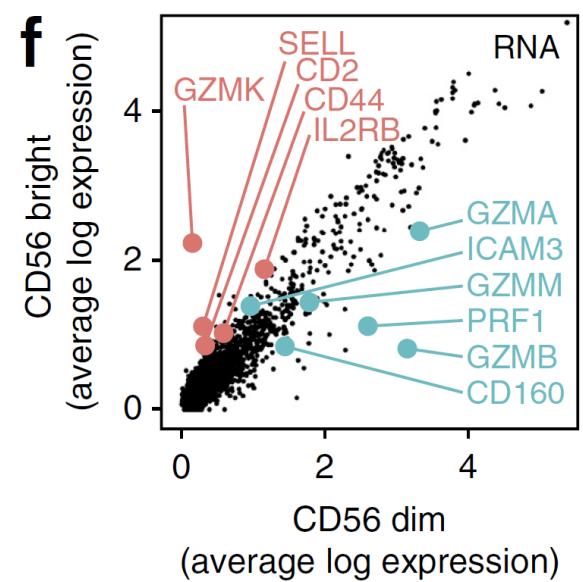
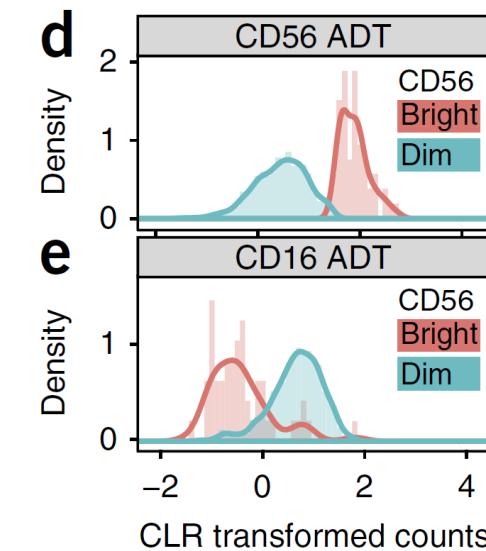
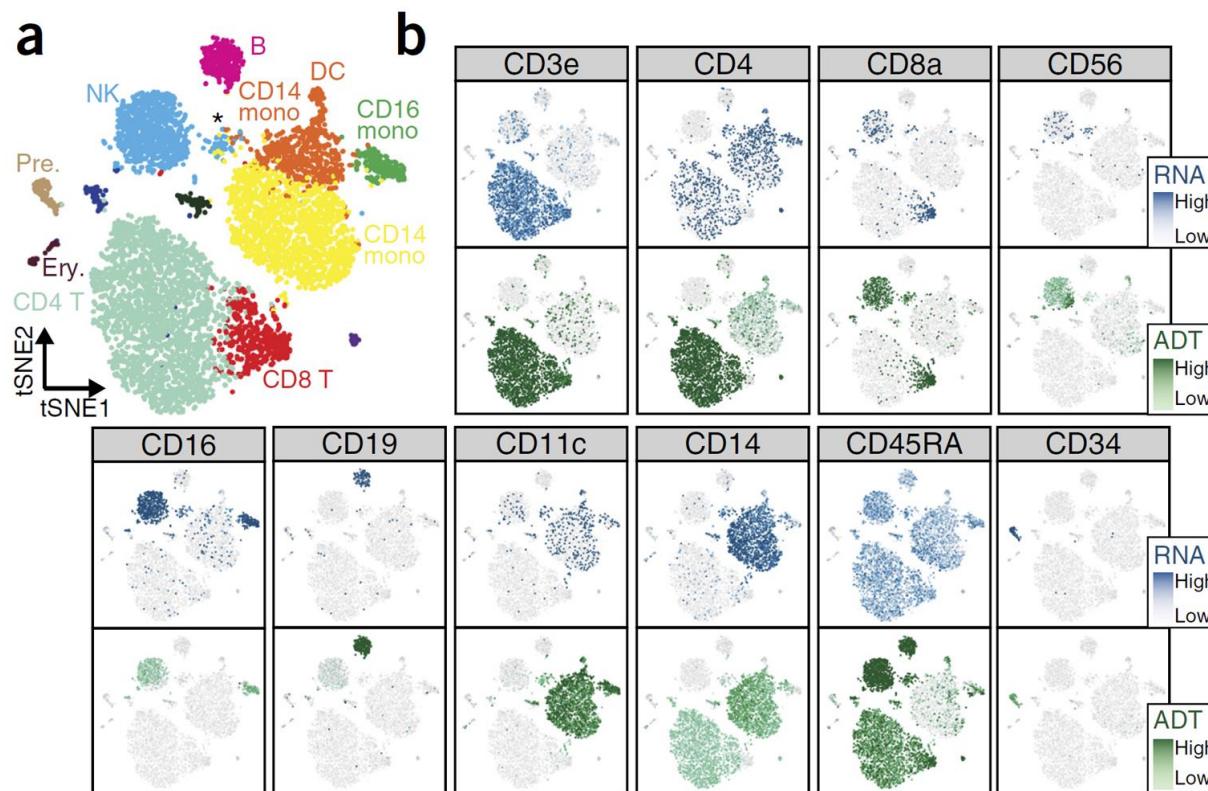
CITE-seq

Cellular Indexing of Transcriptomes and Epitopes by Sequencing



CITE-seq

Cellular Indexing of Transcriptomes and Epitopes by Sequencing



Summary

- Batch effects sometimes not avoidable
- Many batch correction/integration methods available, mainly using **joint dimension reduction**, or **joint clustering**, or a combination of both
- Performance assessment is challenging
- Sample multiplexing can help alleviate batch effects
- Simultaneous mRNA and protein profiling: REAP-seq and CITE-seq

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