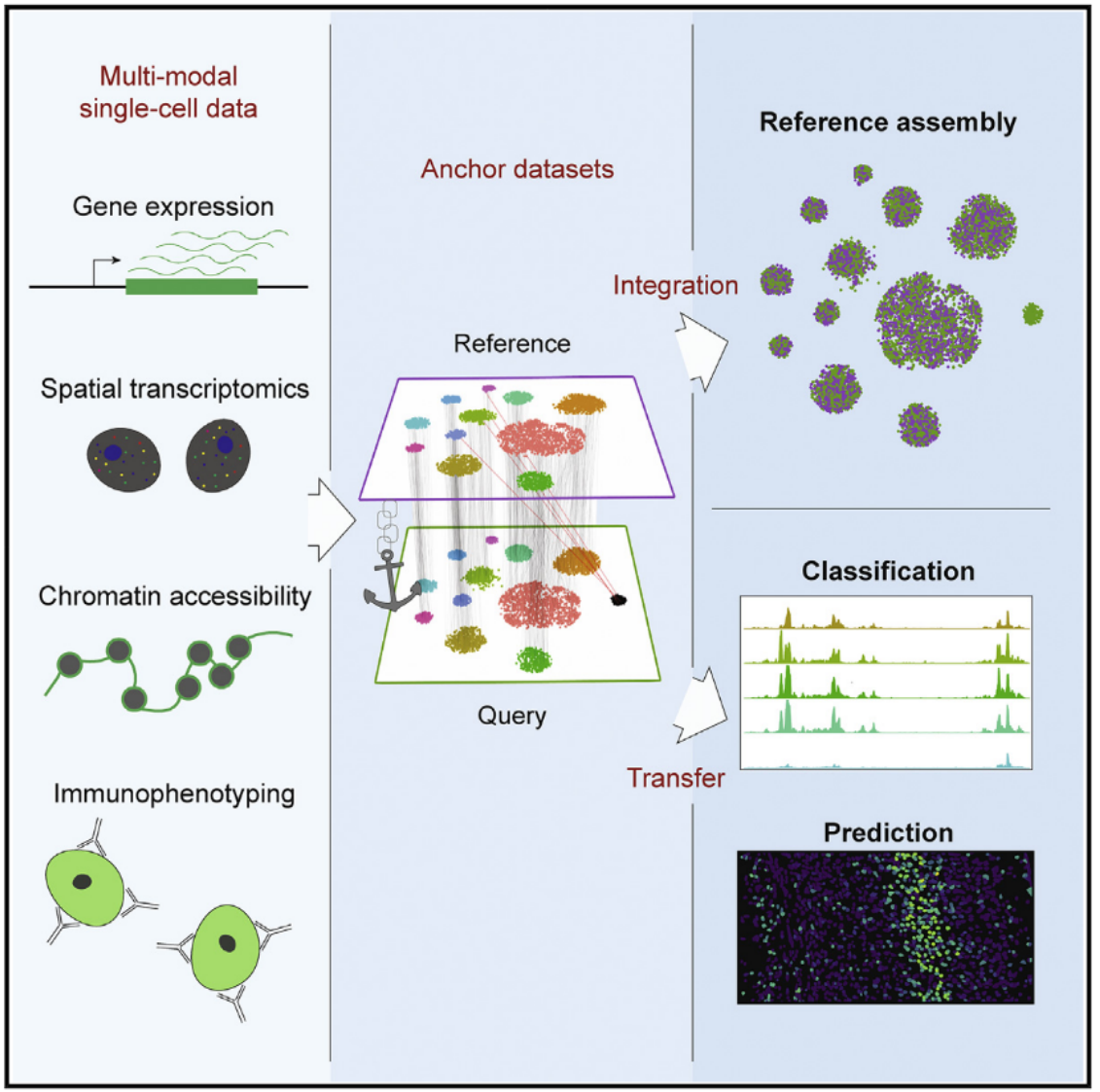
Single Cell OMICs Integration

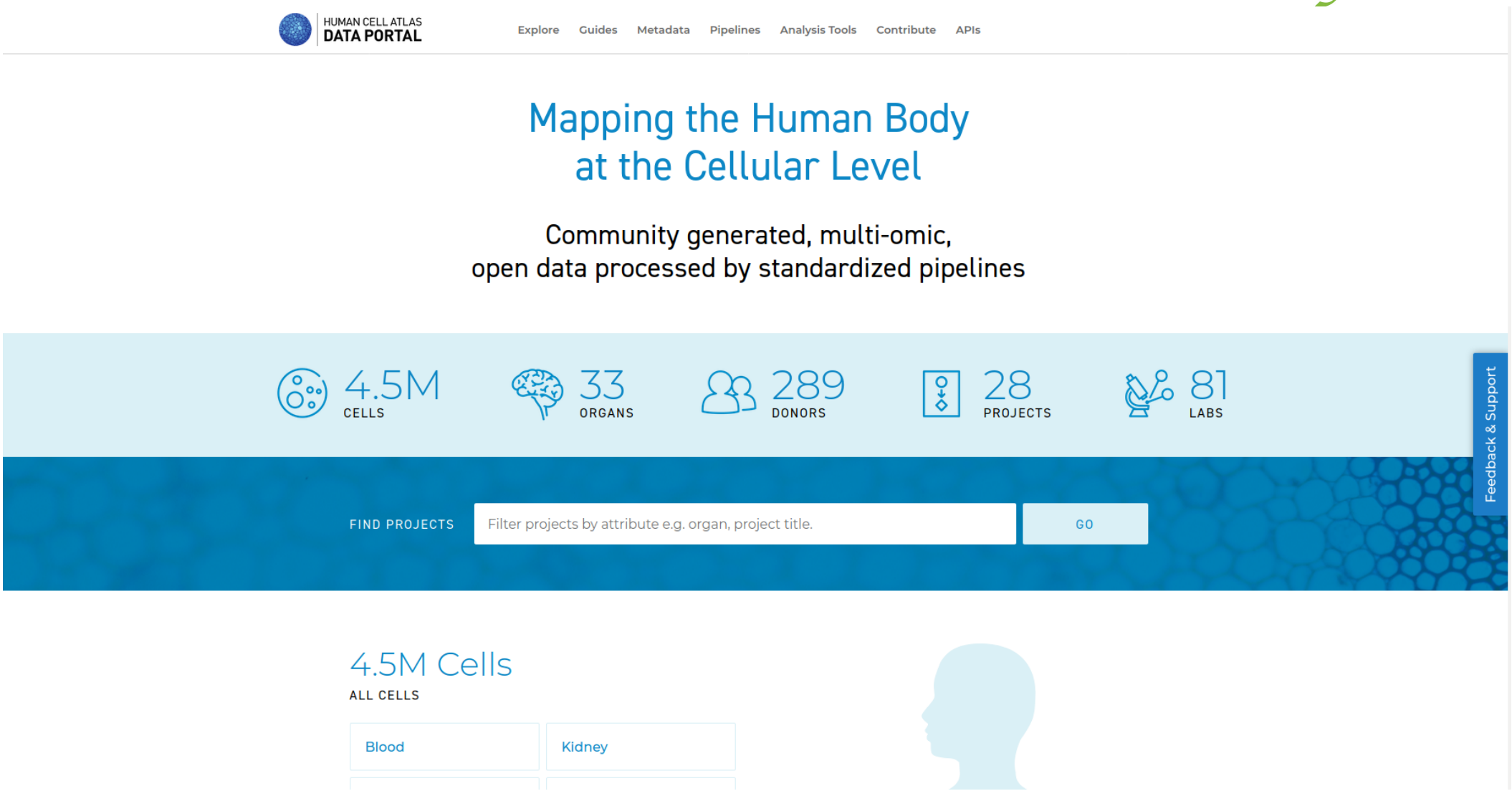
OMICs Integration and Systems Biology course

Nikolay Oskolkov, NBIS SciLifeLab

Lund, 5.10.2020

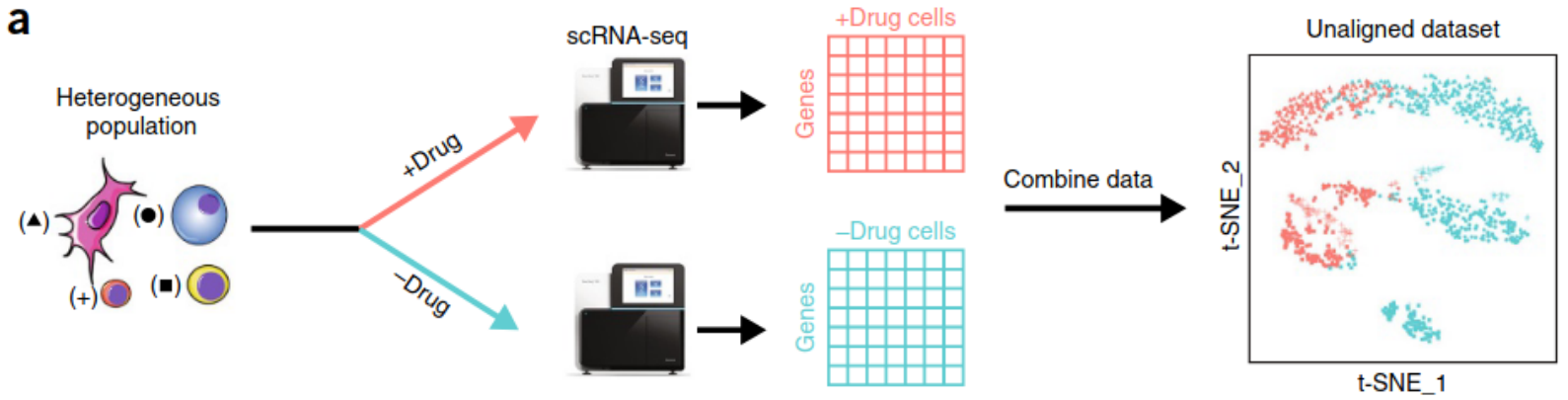


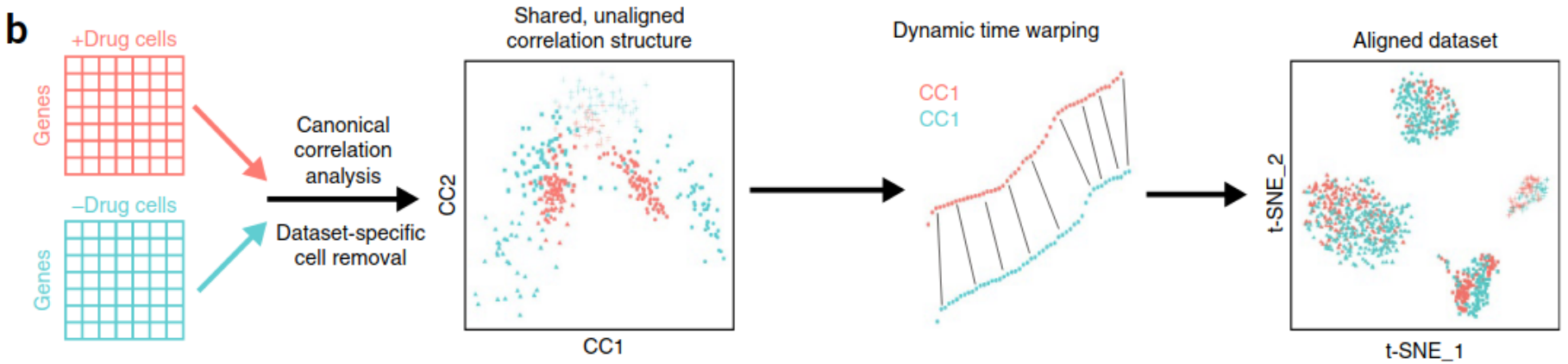
Human Cell Atlas (HCA) Consortium

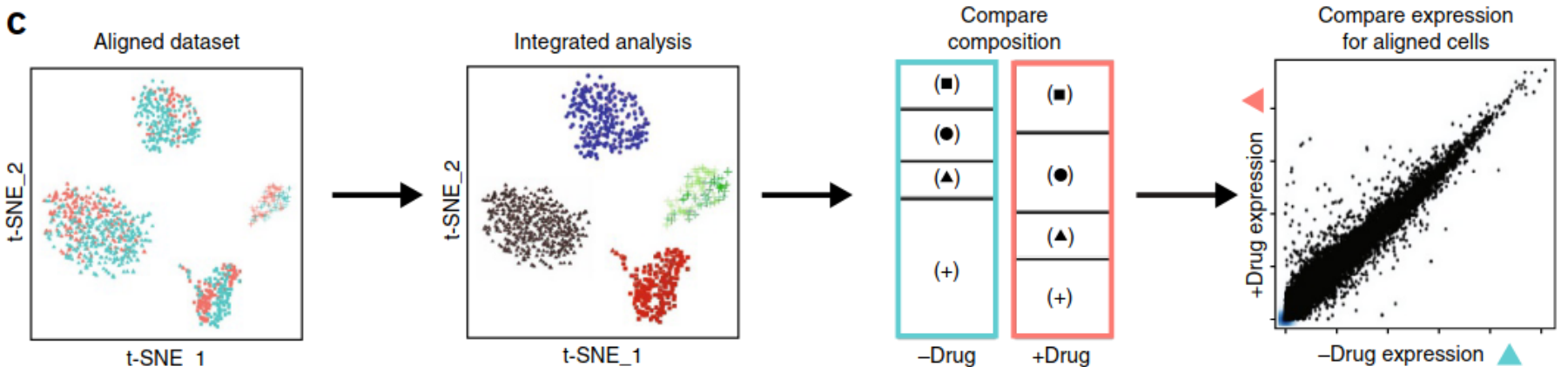


HCA ambition: create a comprehensive Atlas of human cells from all organs / tissues
Data harmonization / integration is one of major chellenges of HCA

Seurat Omics Integration: CCA + DTW







PLS from DIABLO and CCA from Seurat

PLS from DIABLO Article

Denote Q normalized, centered and scaled datasets $X^{(1)}$ $(N \times P_1)$, $X^{(2)}(N \times P_2)$,..., $X^{(Q)}(N \times P_Q)$ measuring the expression levels of $P_1, \ldots, P_Q$ 'omics variables on the same N samples'. sGCCA solves the optimization function for each dimension $h = 1, \ldots, H$:

$$\max_{a_h^{(1)},\ldots,a_h^{(Q)}} \sum_{i,j=1, i\neq j}^{Q} c_{i,j}\ \mathrm{cov}(X_h^{(i)} a_h^{(i)}, X_h^{(j)} a_h^{(j)}), \quad (1)$$

$$\text{s.t. } ||a_h^{(q)}||_2 = 1 \text{ and } ||a_h^{(q)}||_1 \leq \lambda^{(q)} \text{ for all } 1 \leq q \leq Q$$

where $a_h^{(q)}$ is the variable coefficient or loading vector on dimension h associated to the residual matrix $X_h^{(q)}$ of the dataset $X^{(q)}$. $C = \{c_{i,j}\}_{i,j}$ is a $(Q \times Q)$ design matrix that specifies whether datasets should be connected. Elements in C can be set to zeros when datasets are not connected and ones where datasets are fully connected, as we further describe in Section 2.2. In addition in (1), $\lambda^{(q)}$ is a non-negative parameter that controls the amount of shrinkage and thus the number of non-zero coefficients in $a_h^{(q)}$. Similar to the LASSO (Tibshirani, 1996) and other ℓ_1 penalized multivariate models developed for single omics analysis (Lê Cao *et al.*, 2011), the penalization enables the selection of a subset of variables with non-zero coefficients that define each component score $t_h^{(q)} = X_h^{(q)} a_h^{(q)}$. The result is the identification of variables that are highly correlated *between* and *within* omics datasets.

CCA from Seurat Article

Two set canonical correlation. The first step in the alignment utilizes a variation on canonical correlation analysis (CCA) to find projections of both data sets such that the correlation between the two projections is maximized. Formally, CCA finds projection vectors u and v such that the correlation between the two indices $u^T X$ and $v^T Y$ is maximized[28].

$$\max_{u,v} u^T X^T Y v \text{ subject to } u^T X^T X u \leq 1,\ v^T Y^T Y v \leq 1 \quad (1)$$

To apply this in the context of scRNA-seq, let $X_{g,c}$ be a gene expression matrix of genes $g_1, g_2, \ldots, g_n$ by cells $c_1, c_2, \ldots, c_m$ and $Y_{g,d}$ be a gene expression matrix of the same genes $g_1, g_2, \ldots, g_n$ by cells $d_1, d_2, \ldots, d_p$. In many scRNA-seq experiments, the number of genes of interest that are shared between the two data sets is often much smaller than the total number of cells that were measured ($n << m + p$). Consequently, the vectors u and v that are returned from CCA as described in equation (1) will not be unique.

One potential solution to this is to regularize or penalize the CCA procedure to promote sparsity. However, this would assign many cells zero loadings in the resulting projections and result in a complete loss of information for a significant proportion of cells. Therefore, we treat the covariance matrix within each data set as diagonal, a solution that has demonstrated promising results in other high-dimensional problems[59,60]. We substitute the identity matrix for $X^T X$ and $Y^T Y$ to arrive at equation (2).

$$\max_{u,v} u^T X^T Y v \text{ subject to } \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \quad (2)$$

To construct our canonical correlation vectors, we standardized X and Y to have a mean of 0 and variance of 1.

$$\forall_c \sum X[,c]/n = 0,\ var(X[,c]) = 1 \text{ and } \forall_d \sum X[,d]/n = 0,\ var(X[,d]) = 1$$

We then are able to solve for the canonical correlation vectors u and v using singular value decomposition (SVD) as follows:

Let

$$K = X^T Y$$

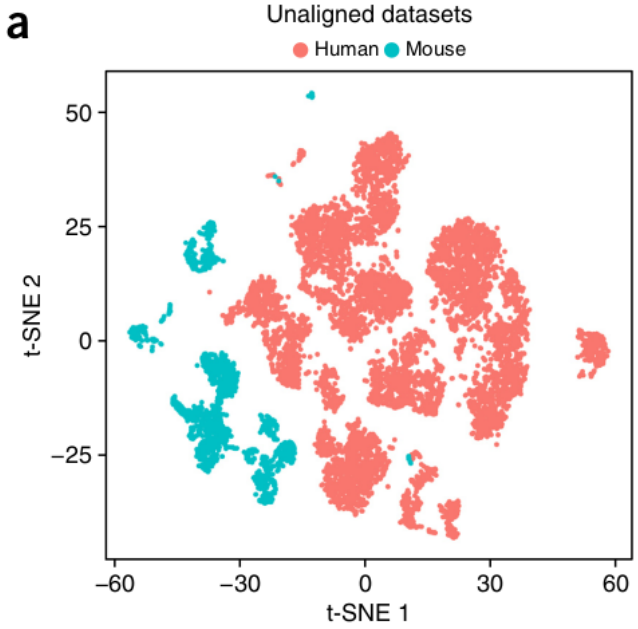
K can be decomposed using SVD as

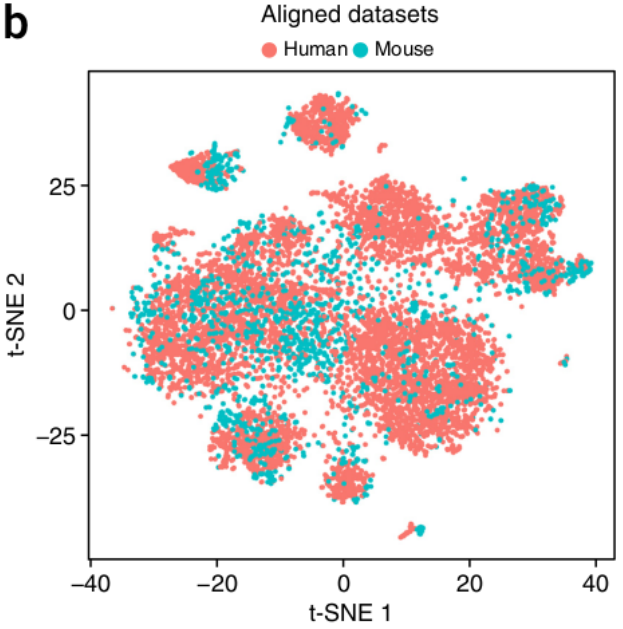
$$K = \Gamma \Lambda \Delta^T \quad (3)$$

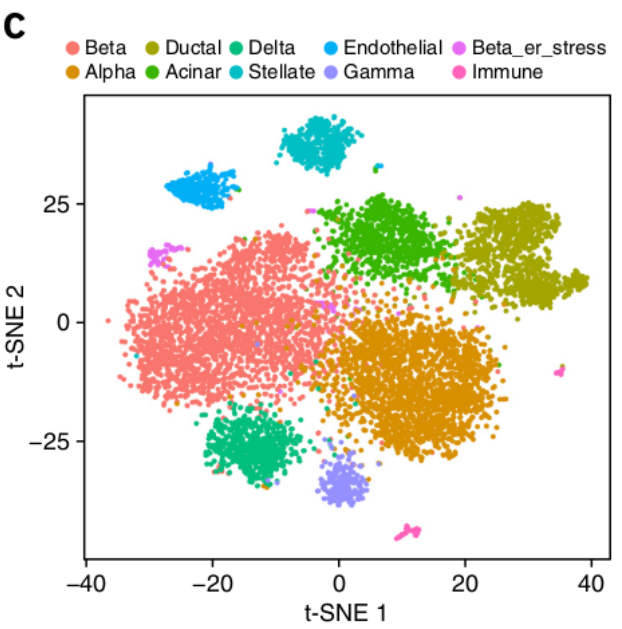
where

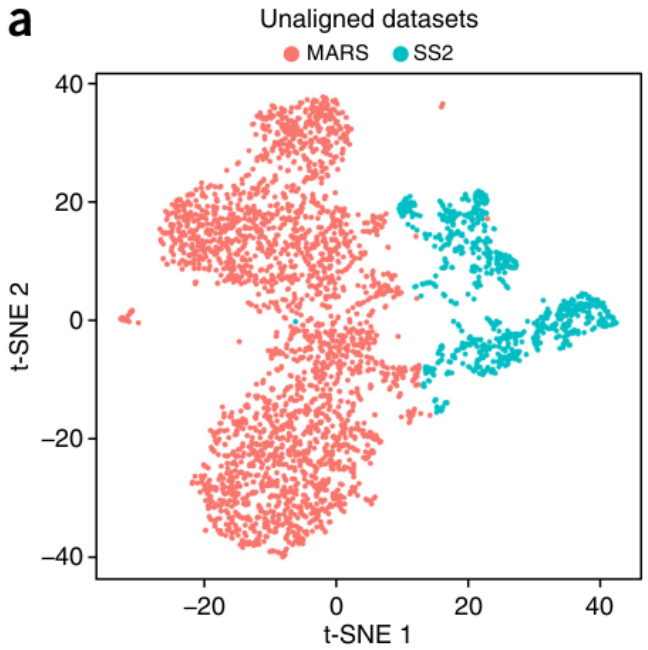
$$\Gamma = (\gamma_1, \ldots, \gamma_k)$$

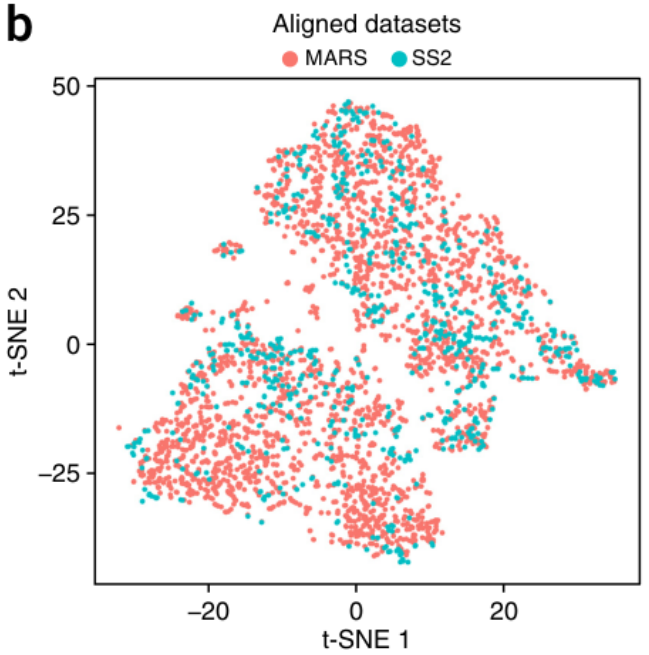
Seurat: CCA + DTW in Action

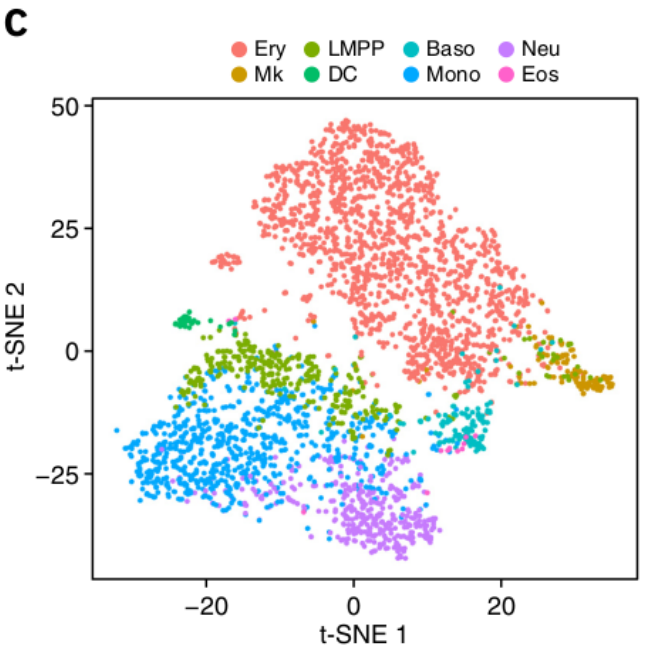






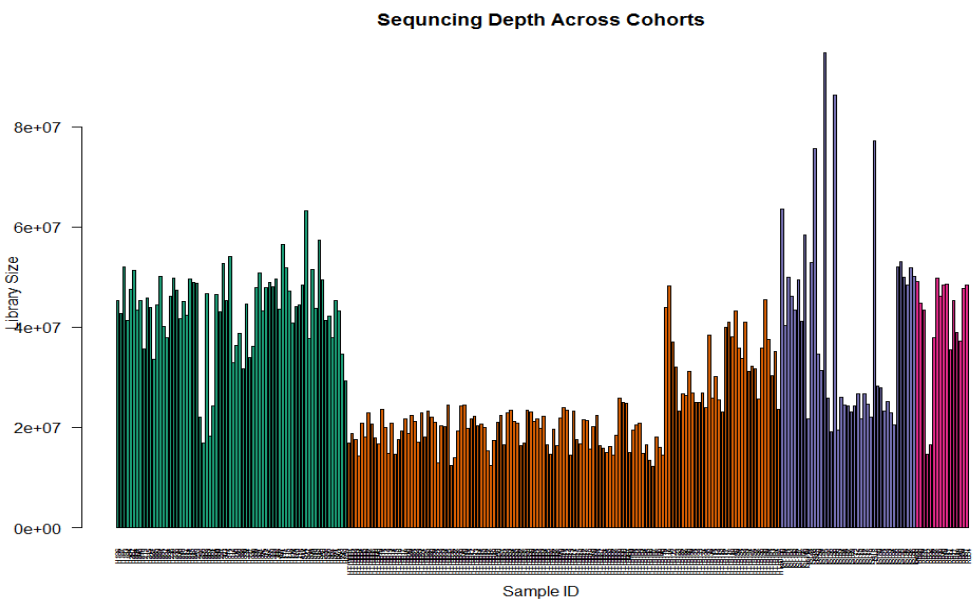




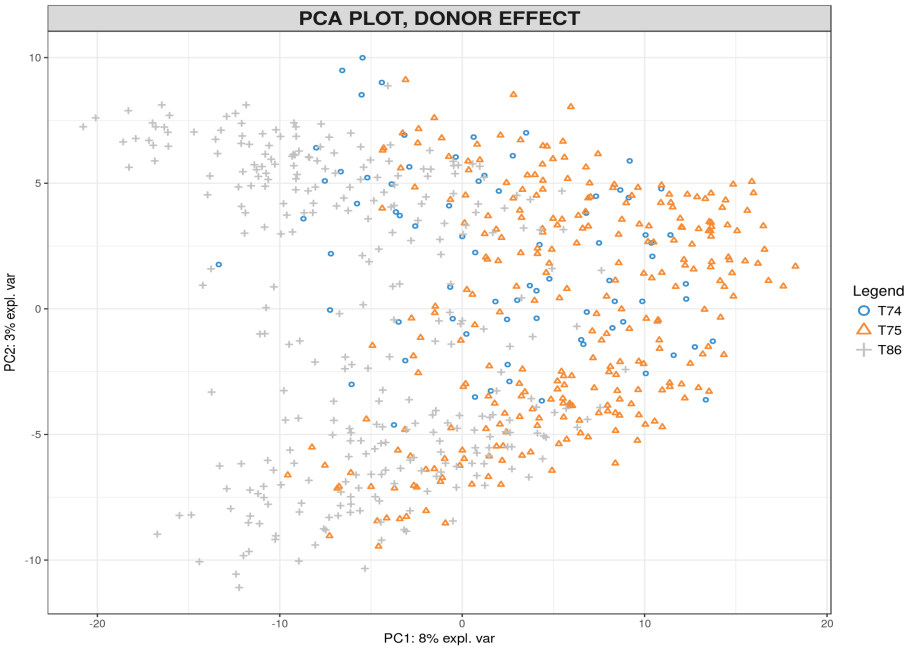


How to detect technical variation?

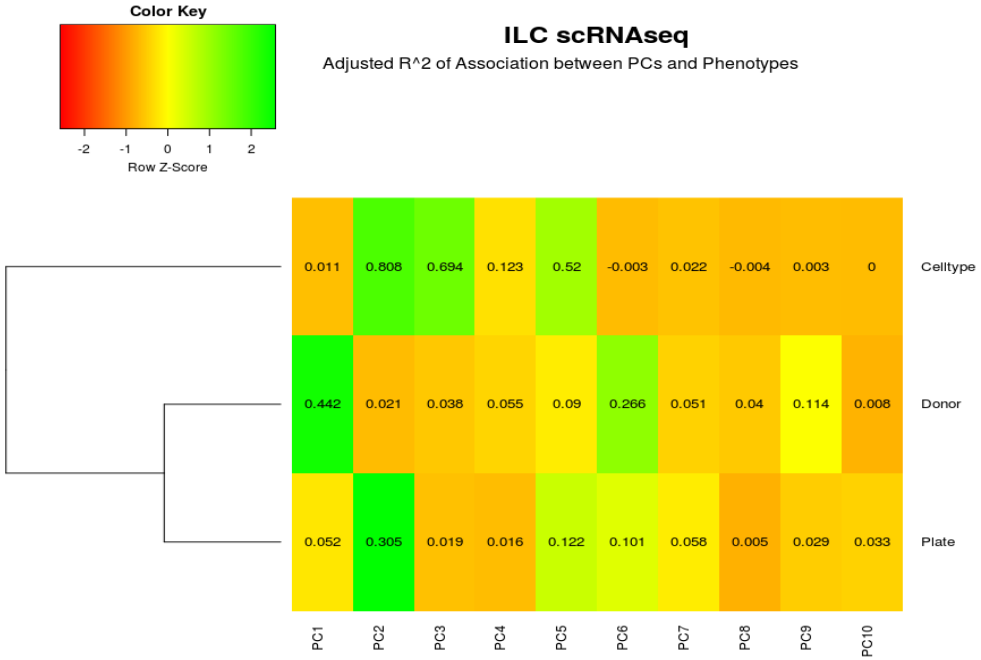
Difference in sequencing depth:



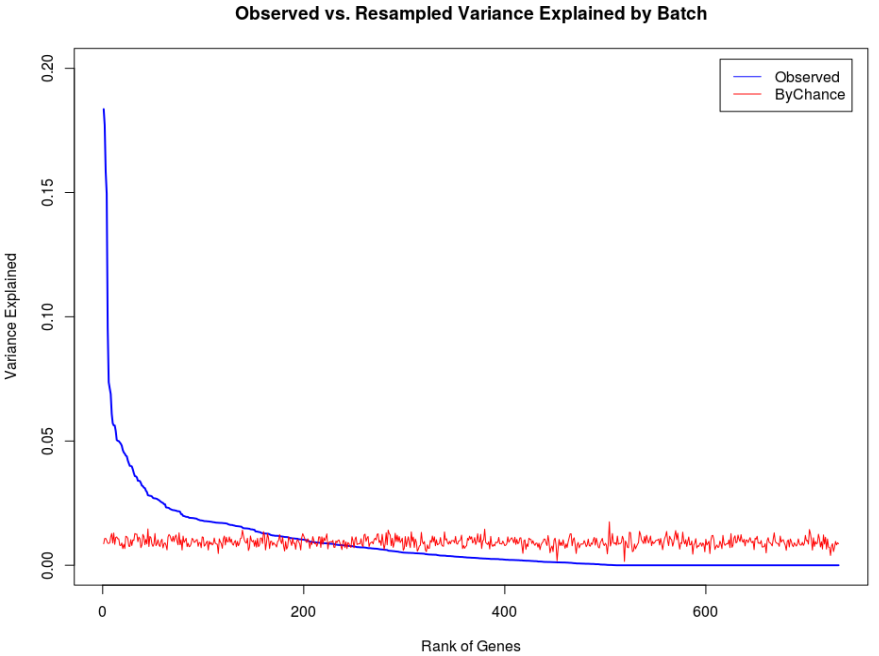
Genome-Wide Batch-effects:



Genome-Wide Batch-effects:



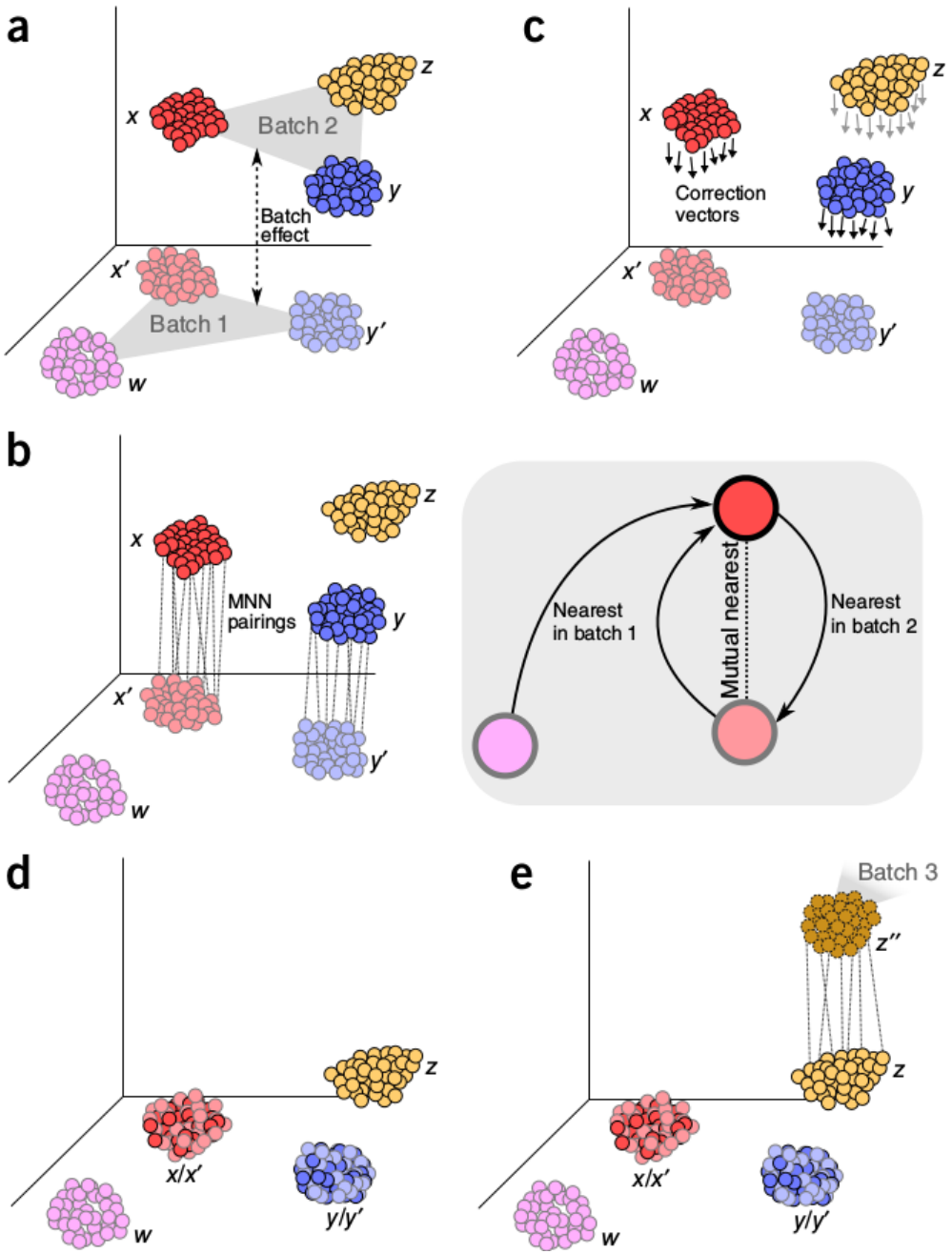
Per-Gene Batch-effects:



Single Cell Batch-Effect Correction: Mutual Nearest Neighbors (MNN)

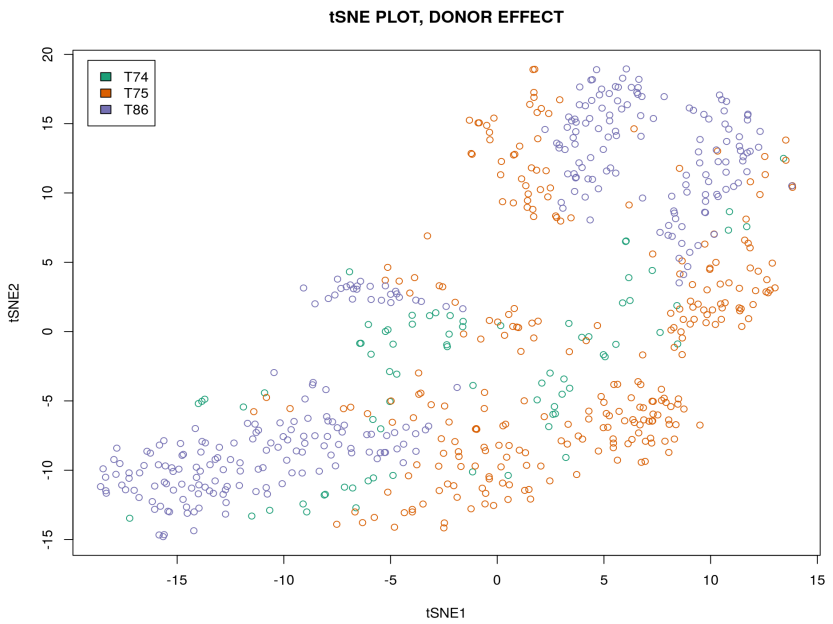
1) For each cell in batch 1 find a nearest neighbor in batch 2 and vice versa

2) Systematic difference in expression between MNN from batch 1 and 2 are to be removed

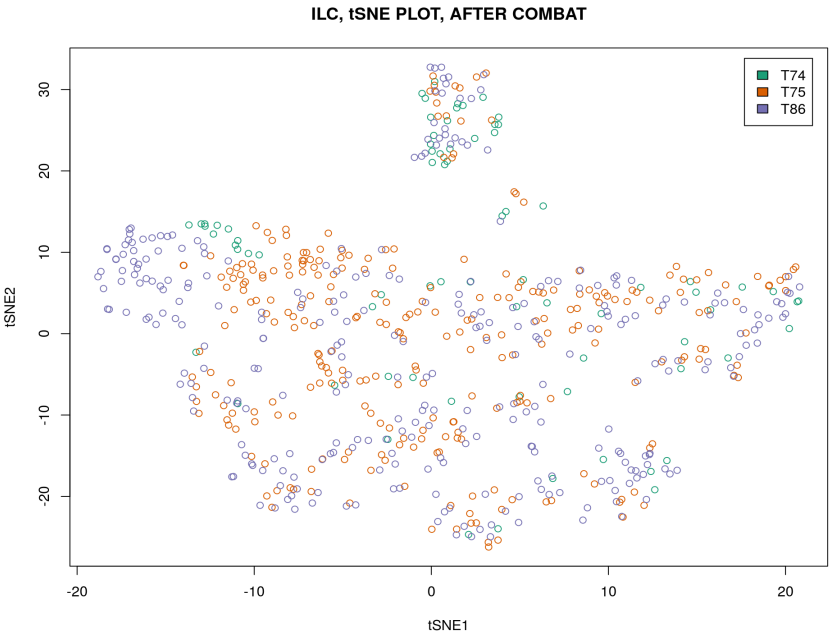


Compare Single Cell Batch Correction Methods

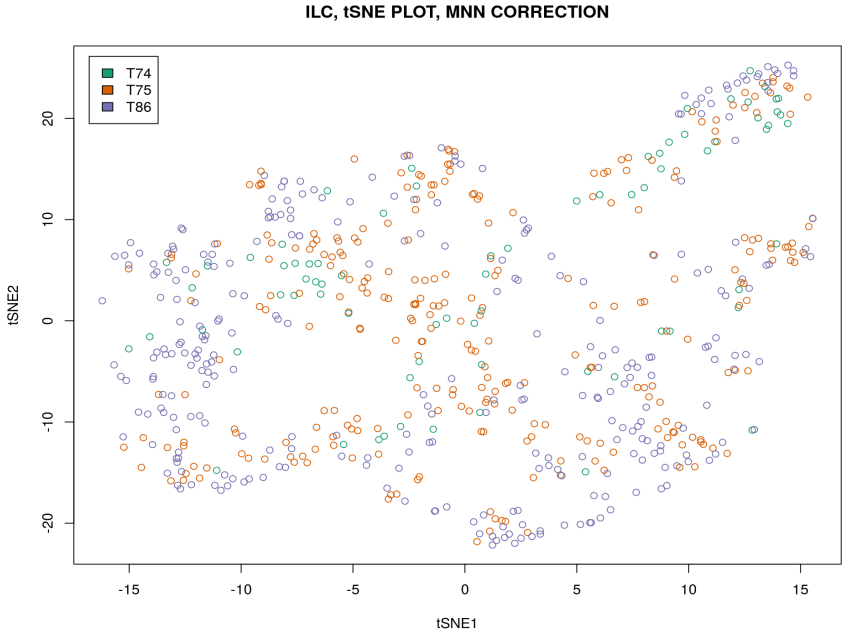
UNCORRECTED



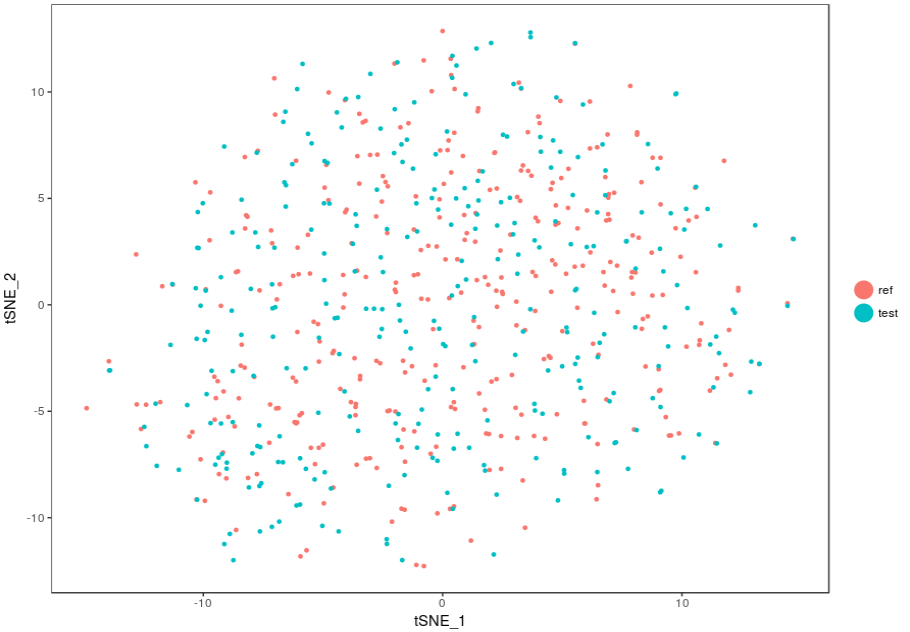
COMBAT



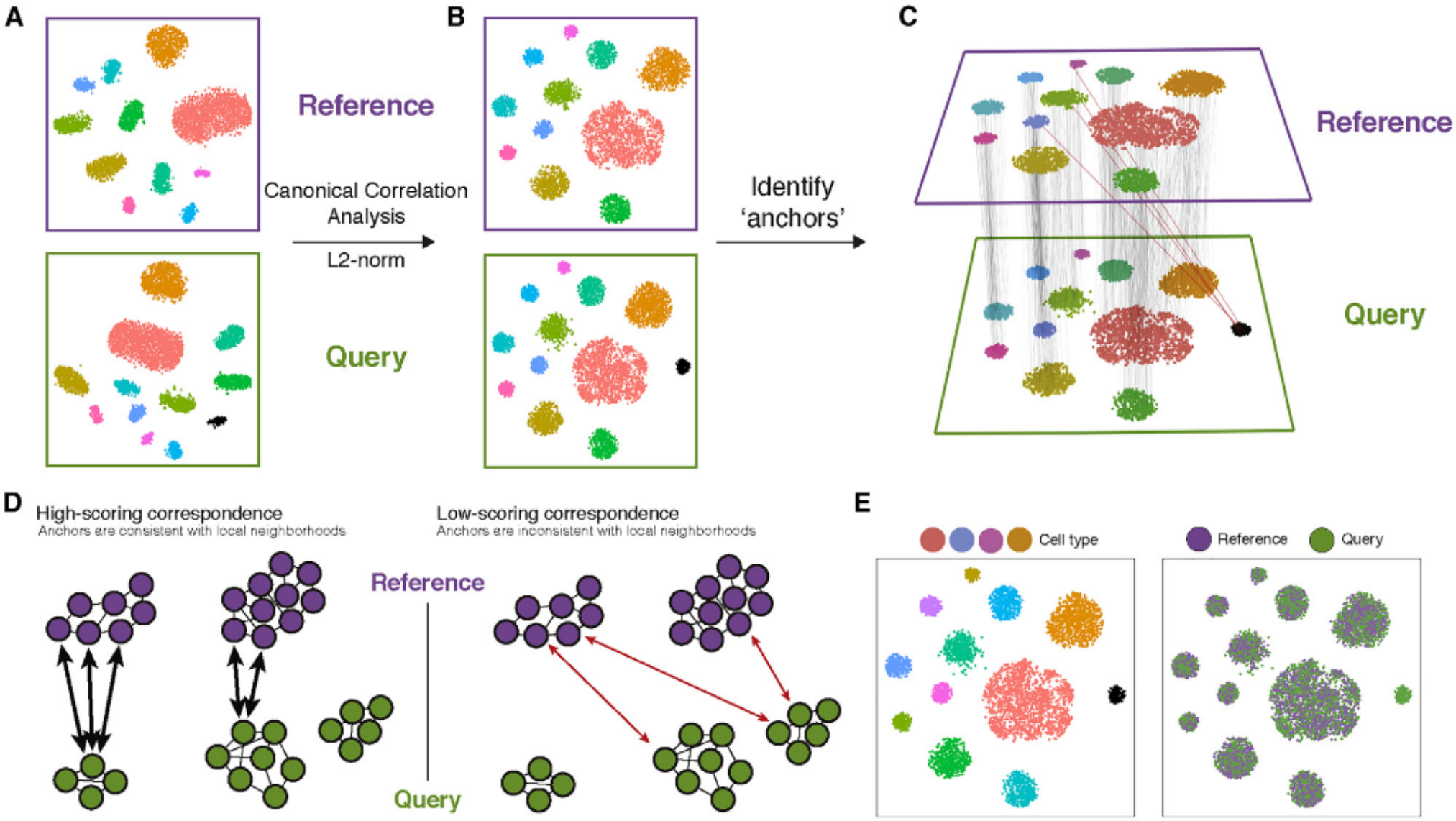
MNN



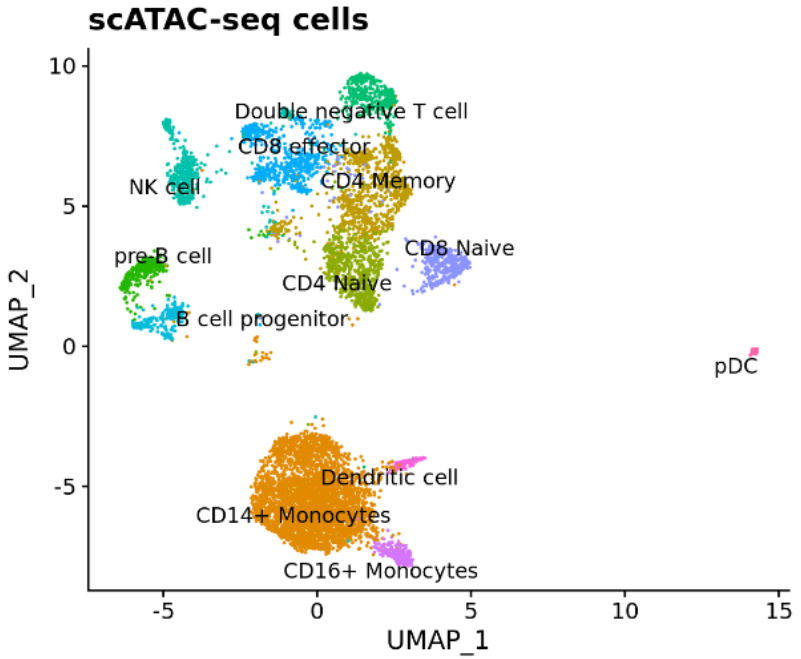
SEURAT

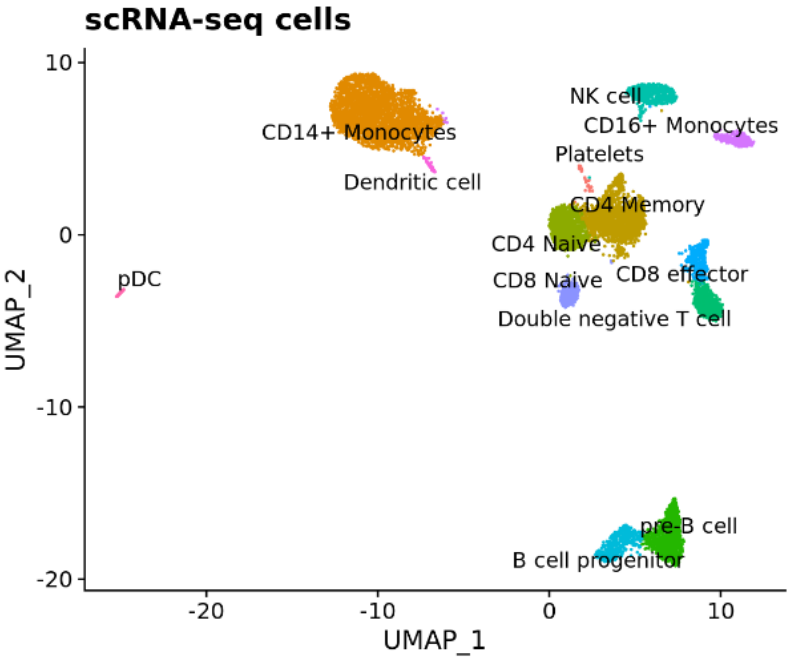


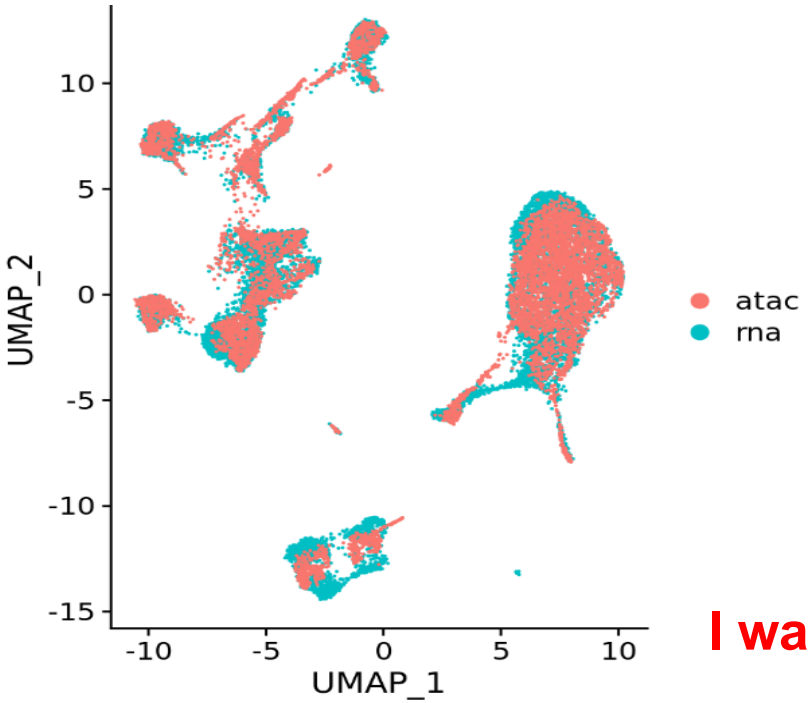
Seurat: Transfer Anchors Across Omics



Single Cell Multi-OMICs Integration: Seurat CCA







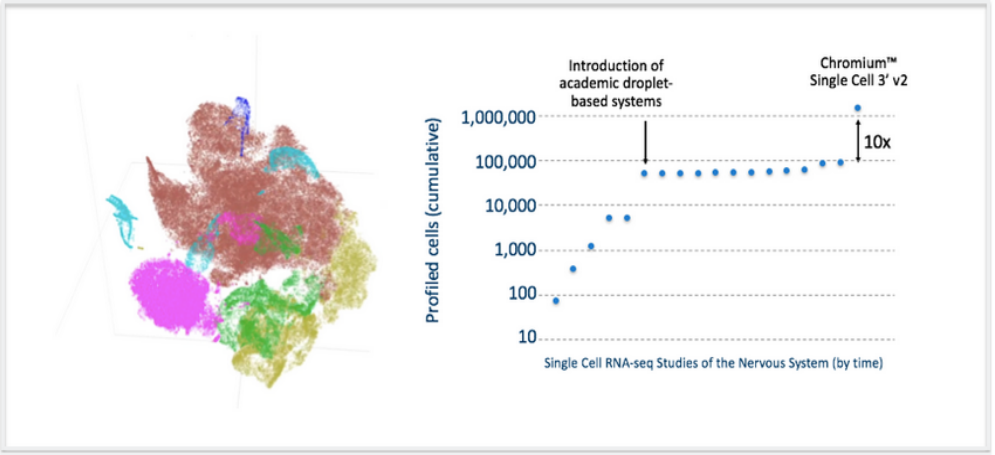
I want to see a new cell type

Big Data in Single Cell

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURC

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

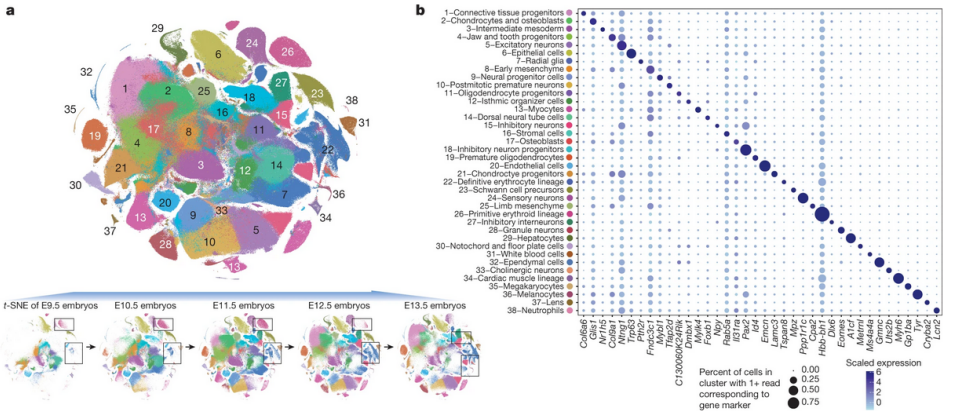
At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution.

Watch out Underfitting!
Paradise for Deep Learning!

MENU nature

Fig. 2: Identifying the major cell types of mouse organogenesis.

From: The single-cell transcriptional landscape of mammalian organogenesis



a, t-SNE visualization of 2,026,641 mouse embryo cells (after removing a putative doublet cluster), coloured by cluster identity (ID) from Louvain clustering (in **b**), and annotated on the basis of marker genes. The same t-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: n = 151,000 for E9.5; 370,279 for E10.5; 602,784 for E11.5; 468,088 for E12.5; 434,490 for E13.5). Primitive erythroid (transient) and definitive erythroid (expanding) clusters are boxed. **b**, Dot plot showing expression of one selected marker gene per cell type. The size of the dot encodes the percentage of cells within a cell type in

BioTuring Solutions Resources

Explore 4,000,000 CELLS at ease with BIOTURING BROWSER

EXPLORER NOW

A next-generation platform to re-analyze published single-cell sequencing data

Single Cell Analysis

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September

by biomembers · August 30, 2019

Human Cell Atlas, single-cell data

We are glad to announce that we will upsize the current single-cell database in BioTuring Single-cell Browser to 5,500,000 cells this September. With this release, we will double the current number of publications indexed in BioTuring Single-cell Browser, and cross the number of cells hosted on available public single-cell data repositories like Human Cell Atlas (HCA) and Broad Institute's Single-cell Portal.

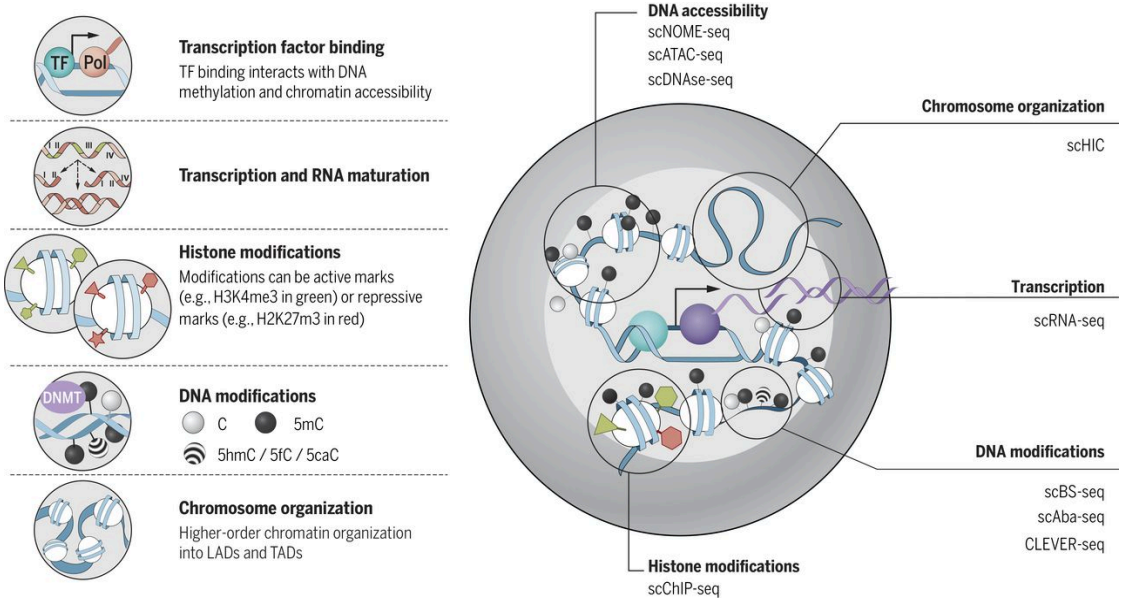
Search

RECENT POSTS

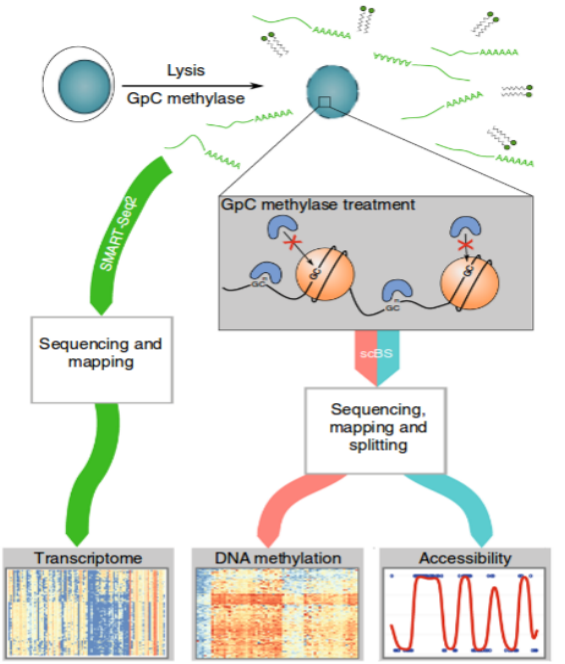
A new tool to interactively visualize single-cell objects (Seurat, Scanpy, SingleCellExperiments, ...)
September 26, 2019

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September
August 30, 2019

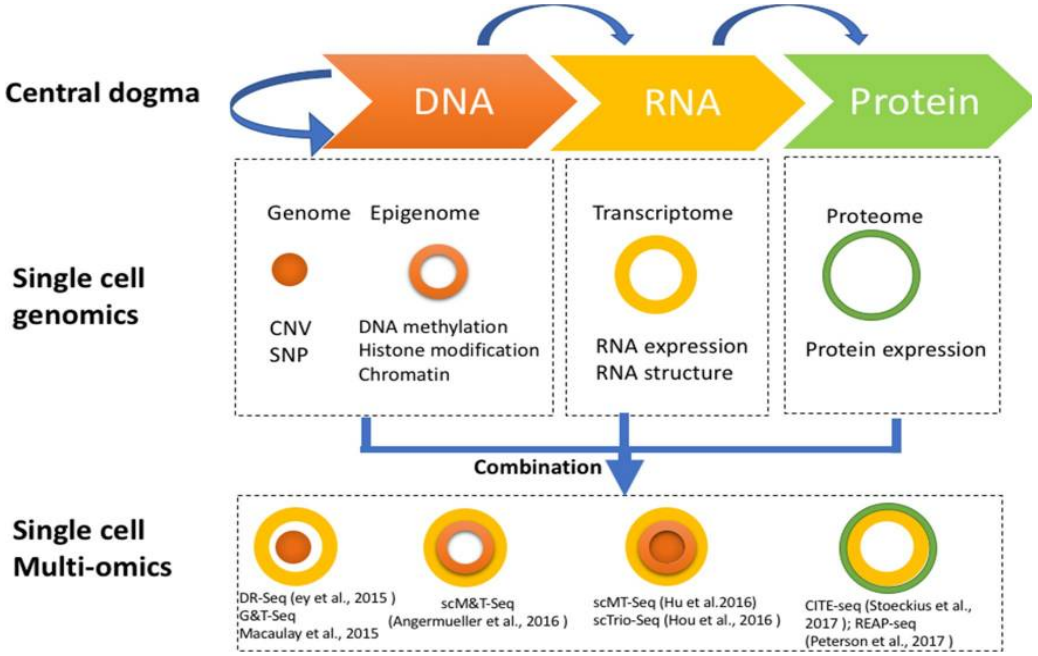
Single Cell Multi-OMICs Integration



Kelsey et al., 2017, Science 358, 69-75



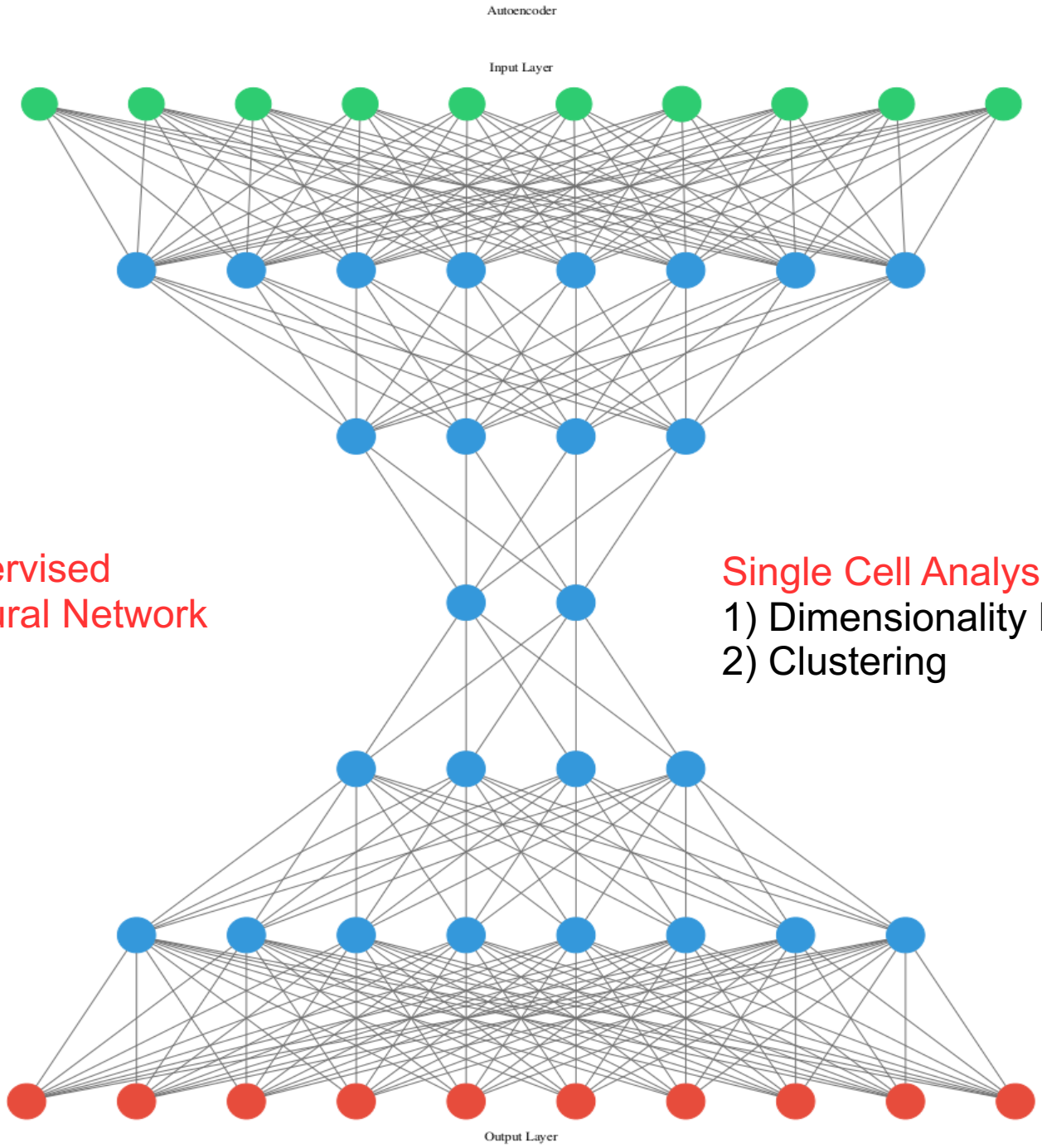
Clark et al., 2018, Nature Communications 9, 781
scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells



Hu et al., 2018, Frontier in Cell and Developmental Biology 6, 1-13

Ultimate Goal:
Model Behavior of Biological Cells

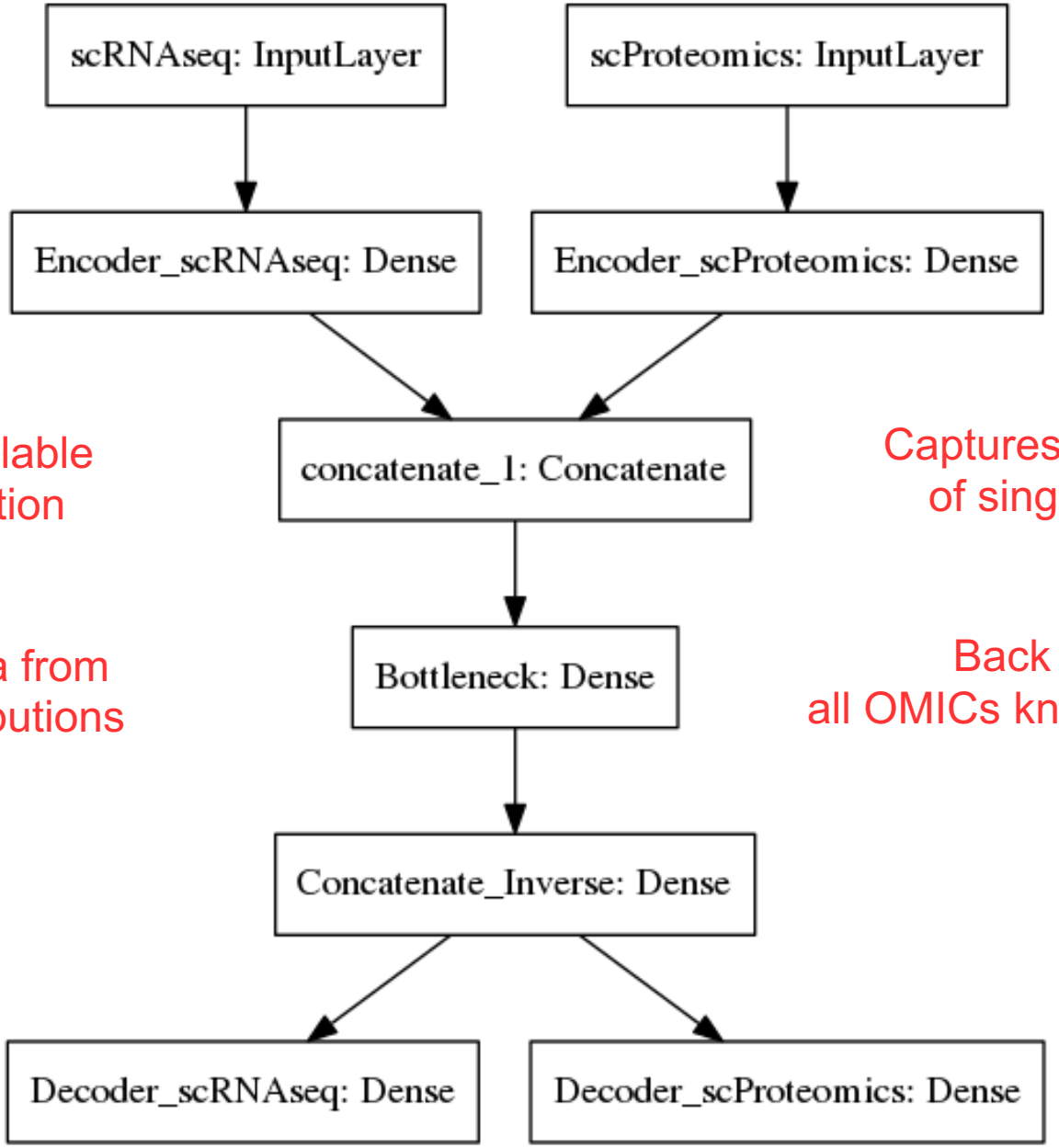
Why Autoencoder for Single Cell?



Unsupervised
Artificial Neural Network

Single Cell Analysis is Unsupervised
1) Dimensionality Reduction
2) Clustering

CITE-seq: Data Integration
Autoencoder



Flexible and scalable OMICs integration

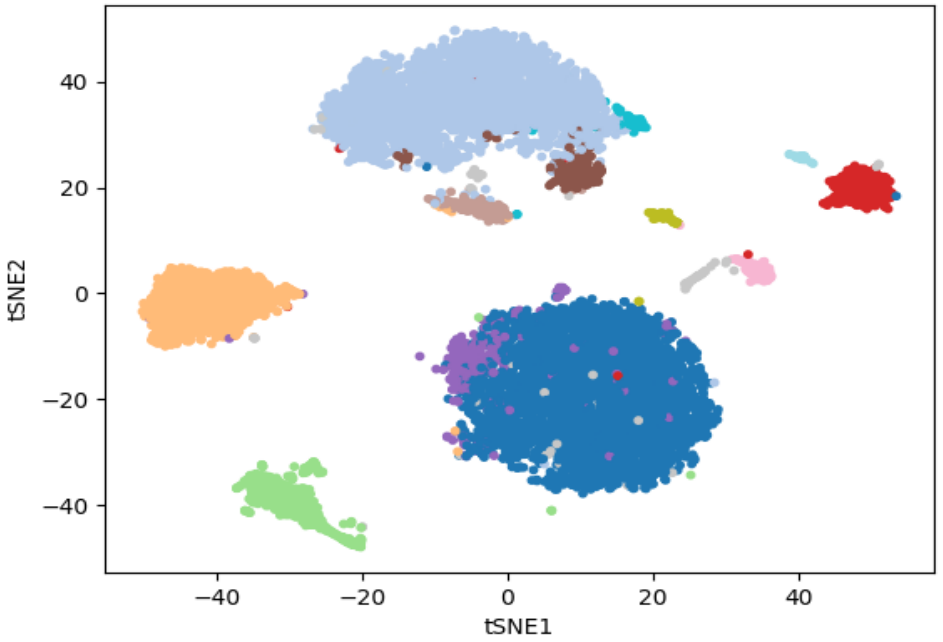
Captures non-linearity of single cell data

Combine data from different distributions

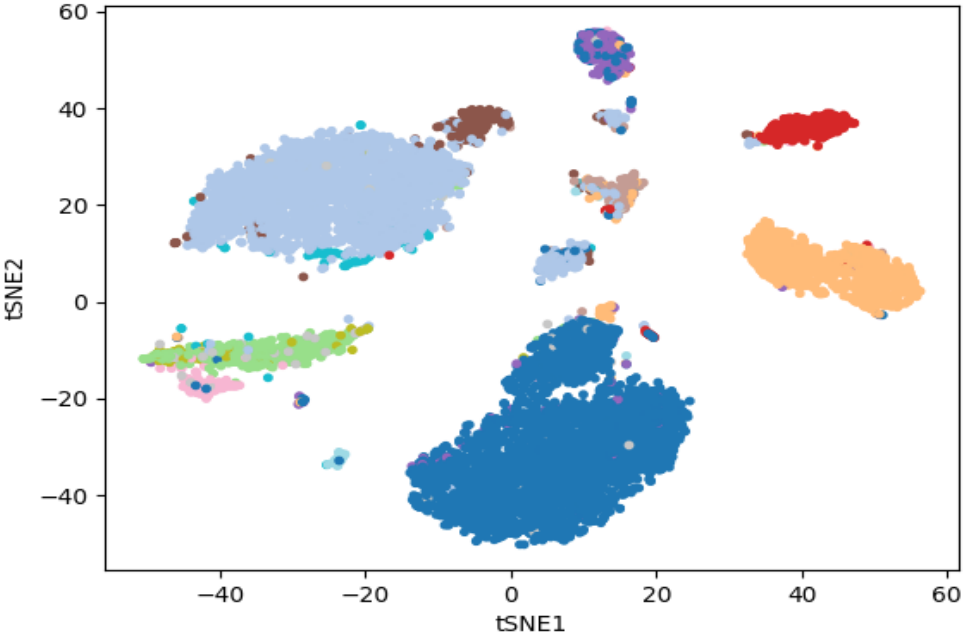
Back Propagation: all OMICs know about each other

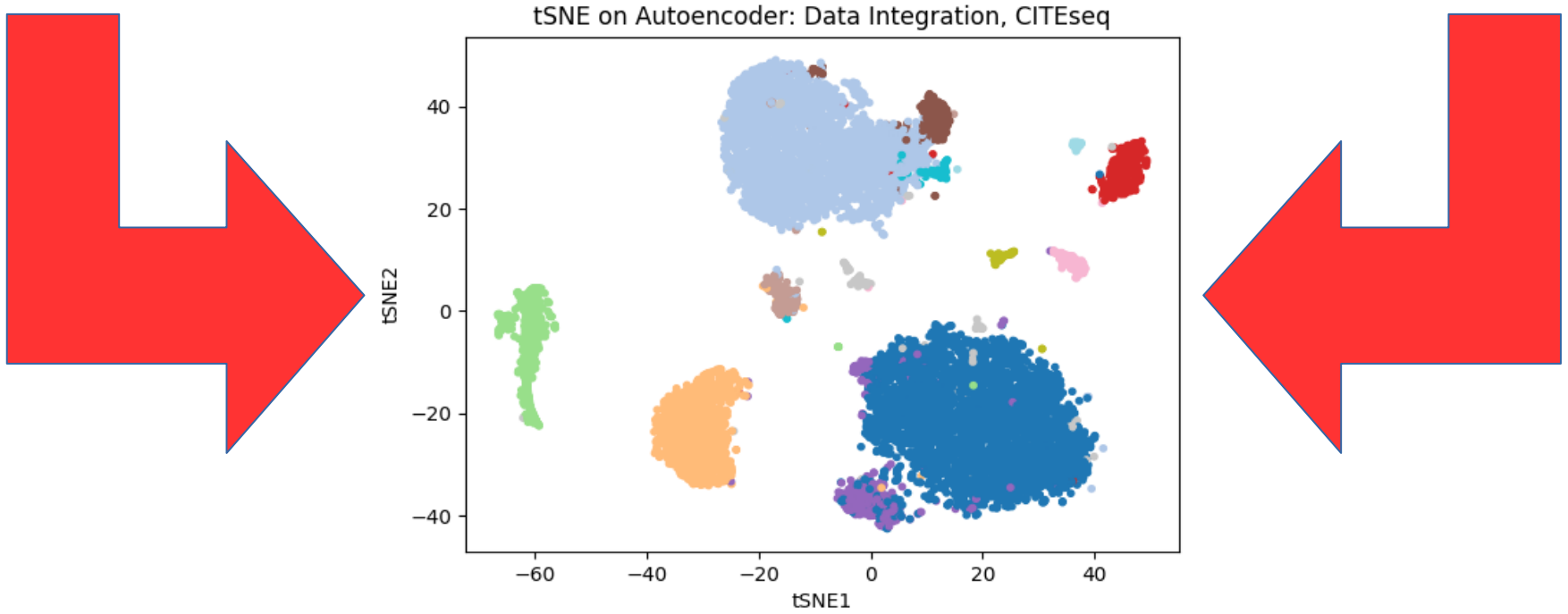
CITE-seq: scOmics Integration
scRNAseq + scProteomics, 8 617 cells

scRNAseq



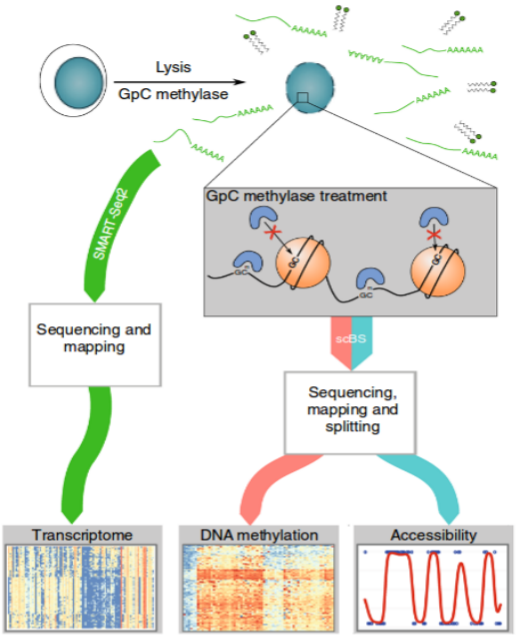
scProteomics



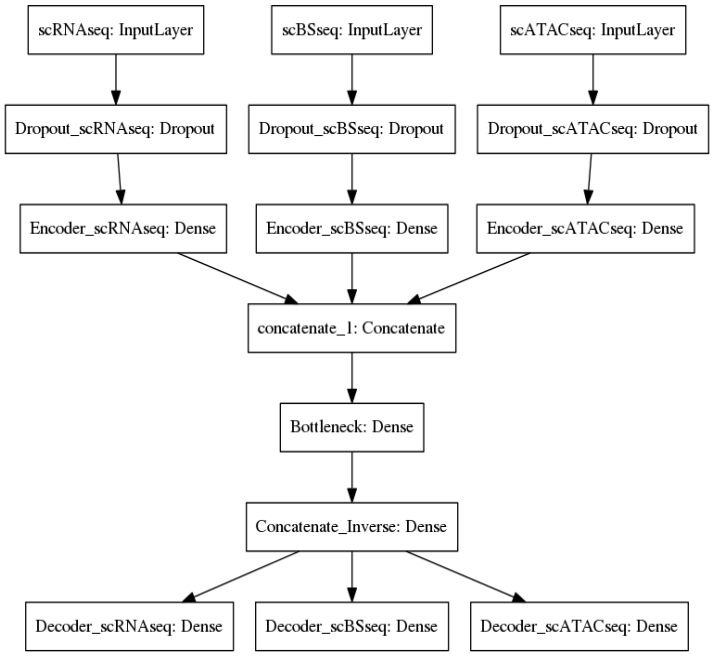


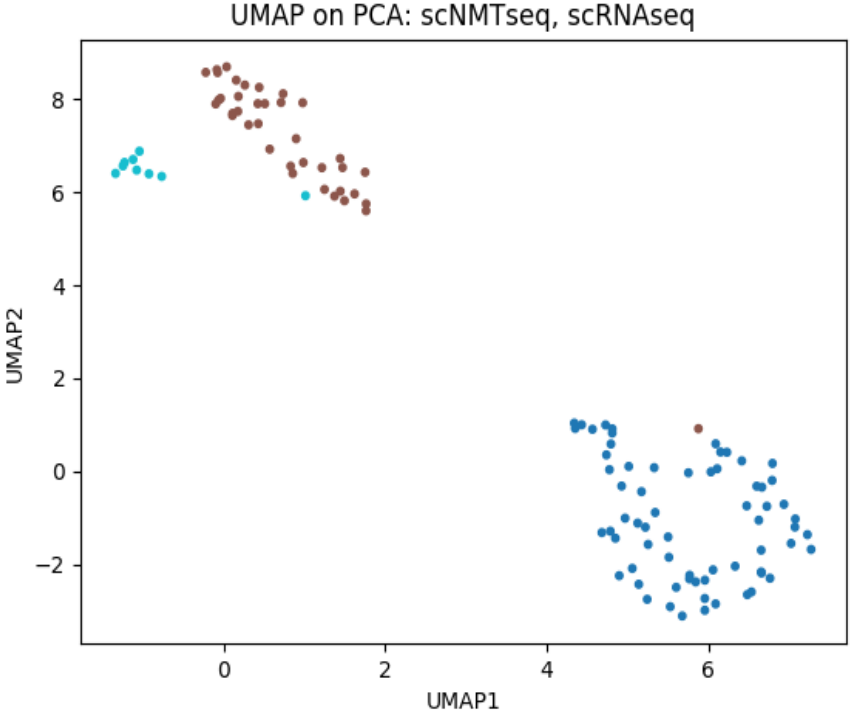
scNMT-seq: scOmics Integration

scRNAseq + scBSseq + scATACseq, 120 cells

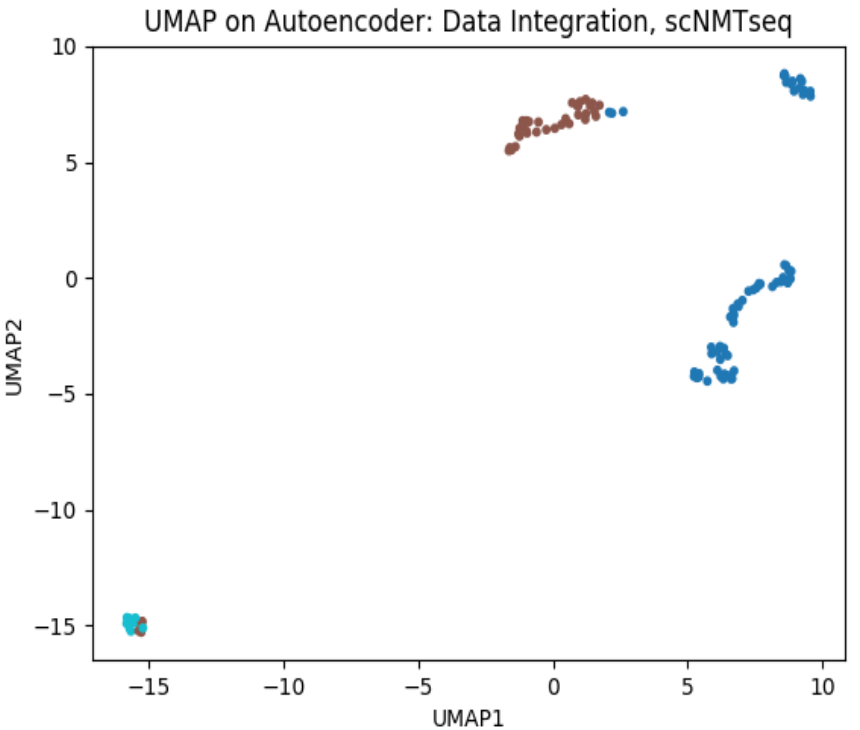


scNMTseq: Clark et al., 2018, Nature Communications 9, 781

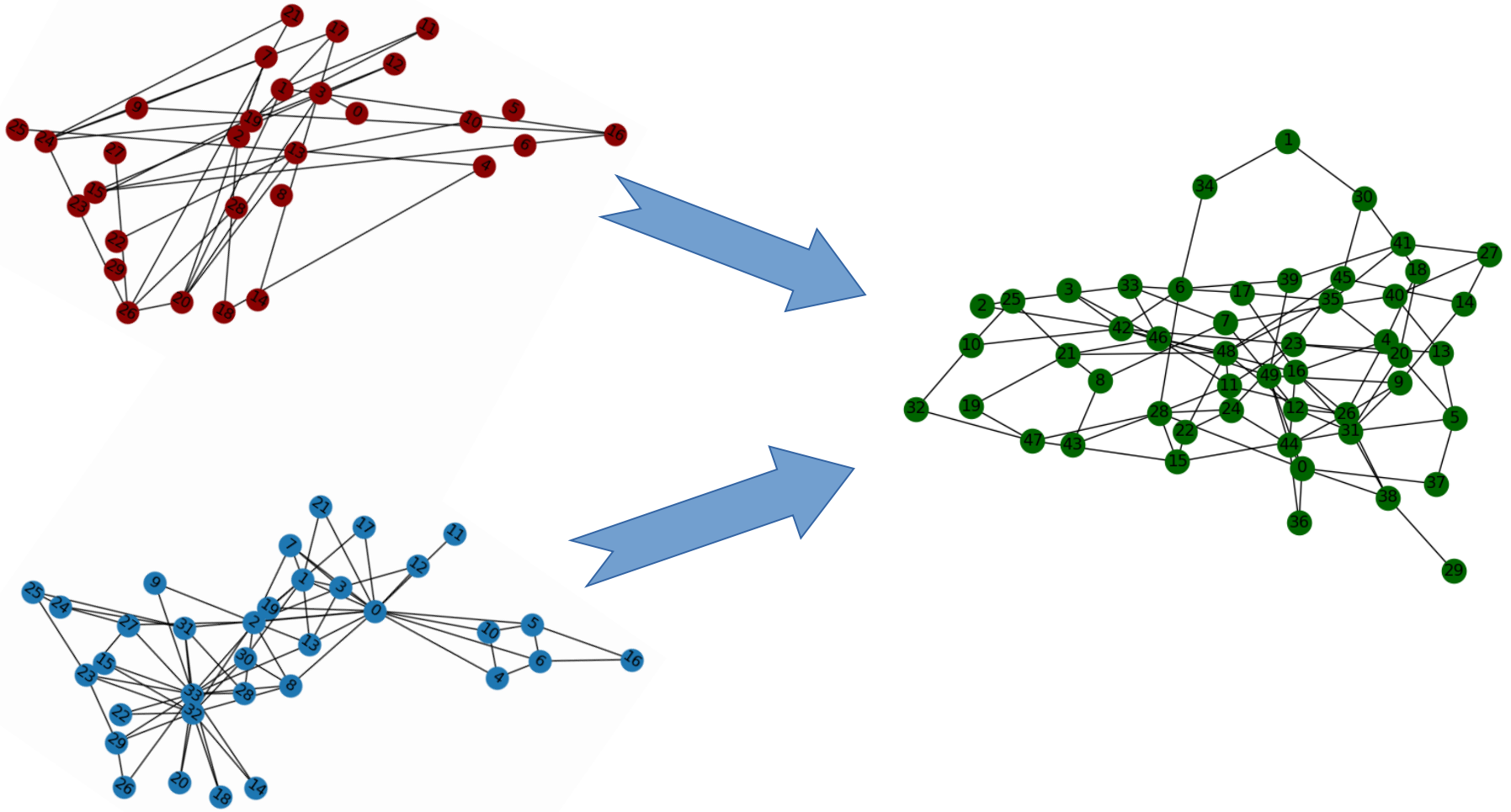




From Single To Multi-OMICs

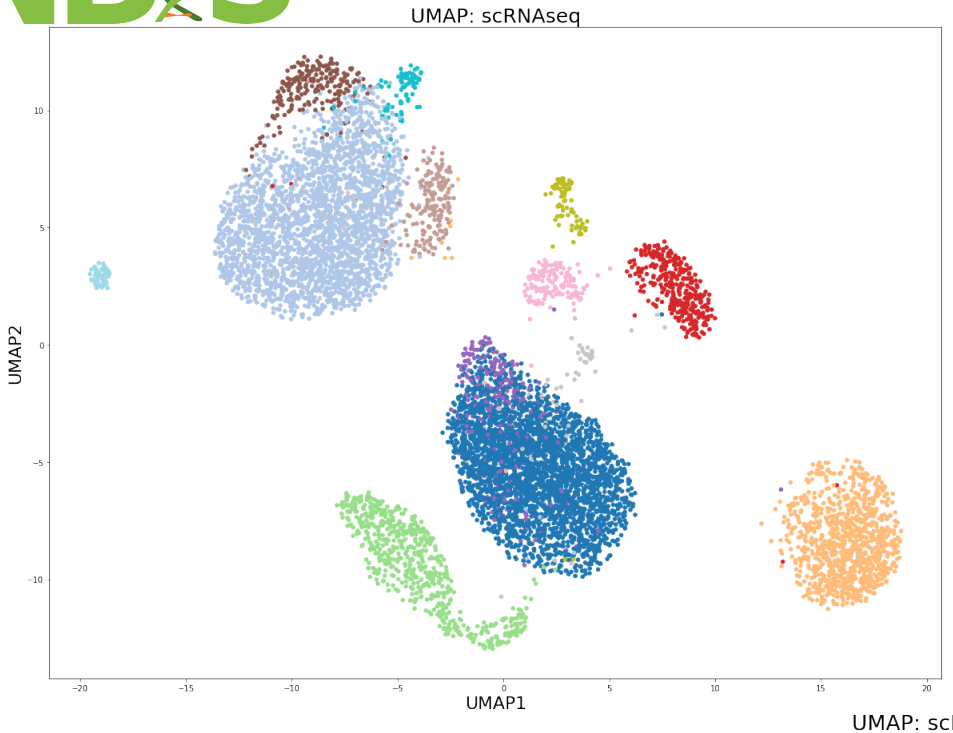


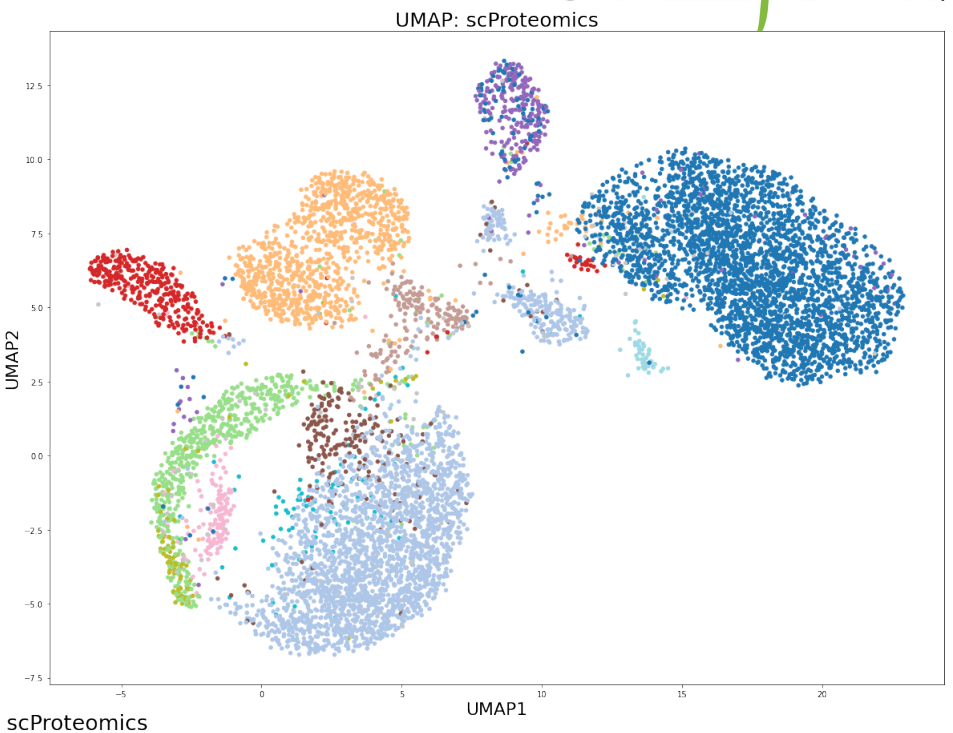
Graph-Based scOmics Integration With UMAP

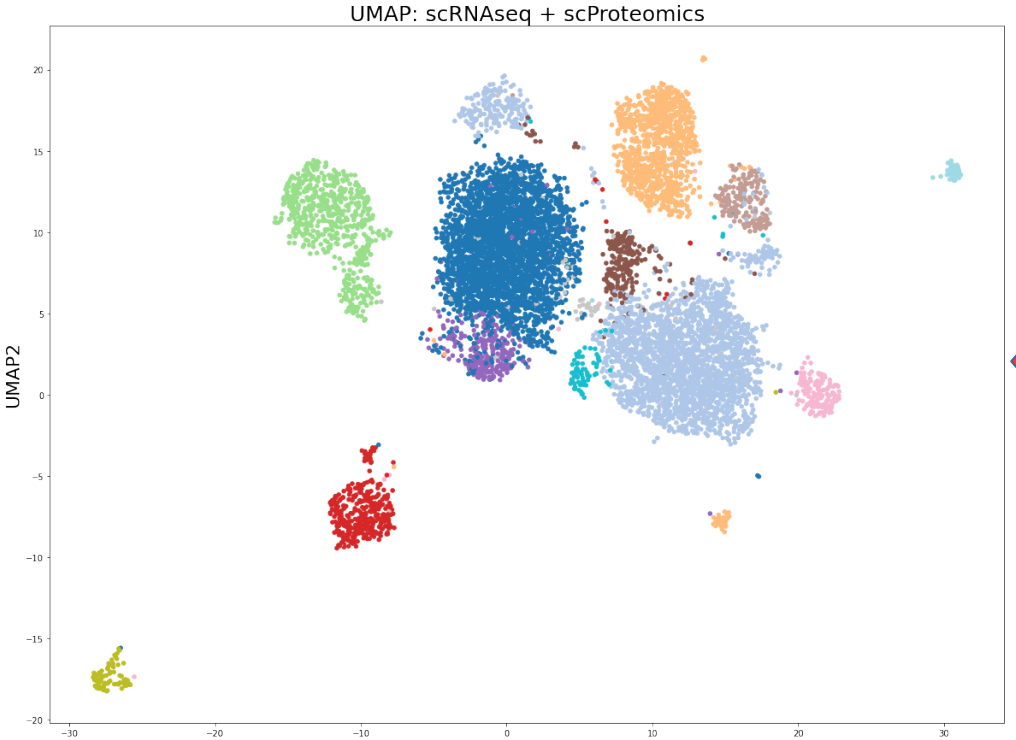


Idea: convert individual OMICs into a graph (non-parametric space), find intersection of the graphs and visualize the intersection

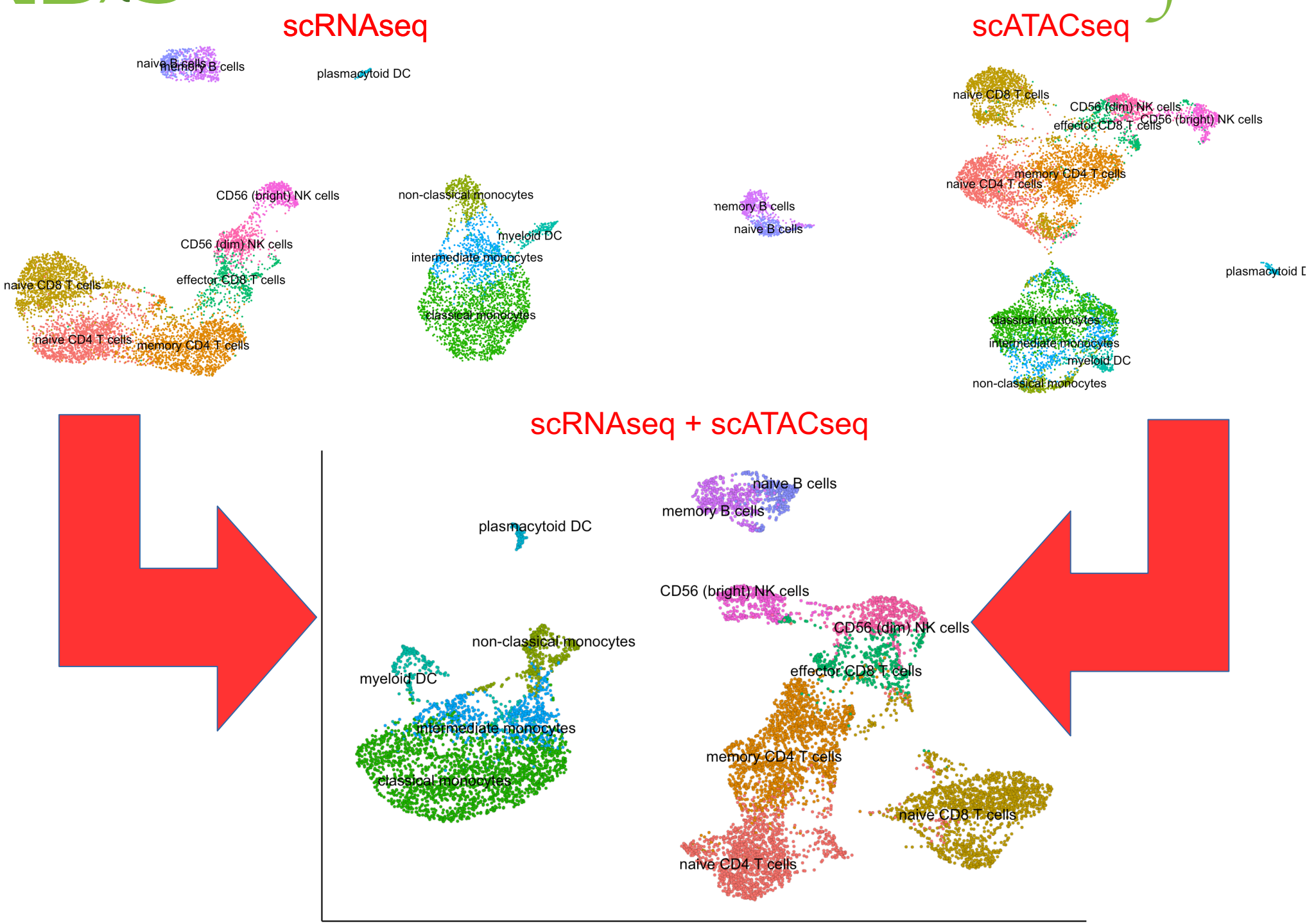
UMAP for scOmics Integration: CITEseq



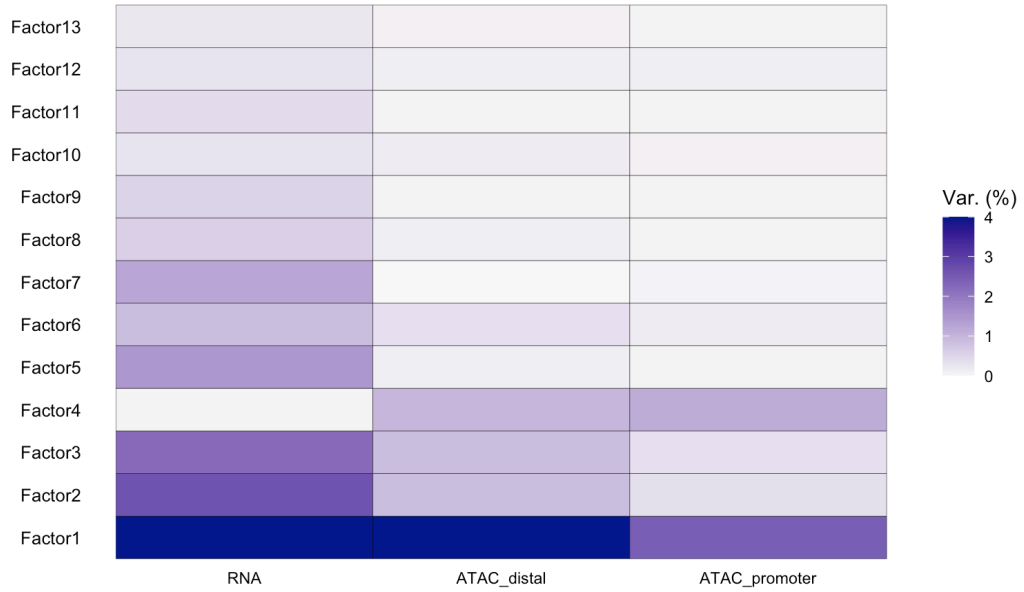


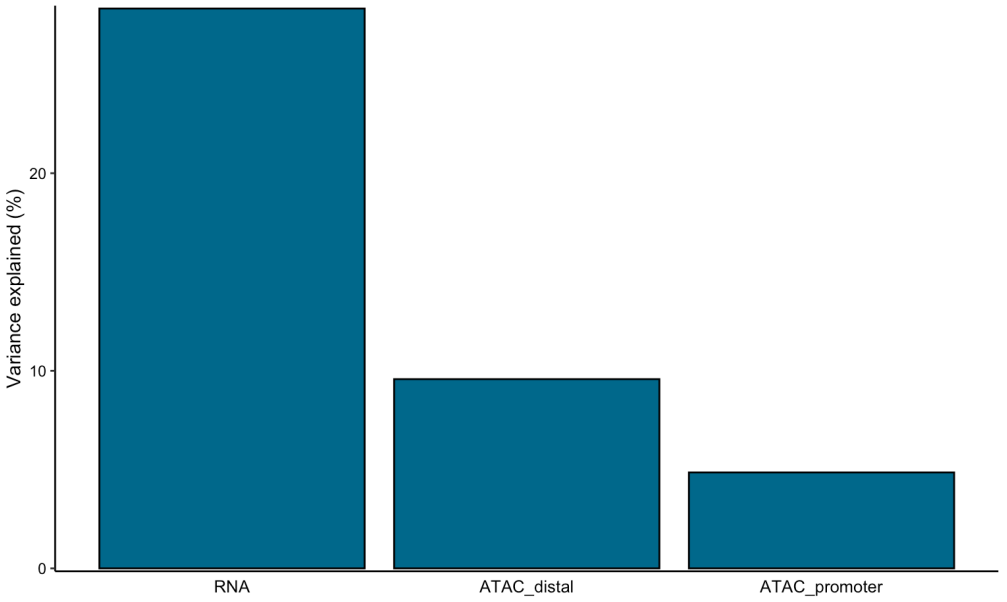


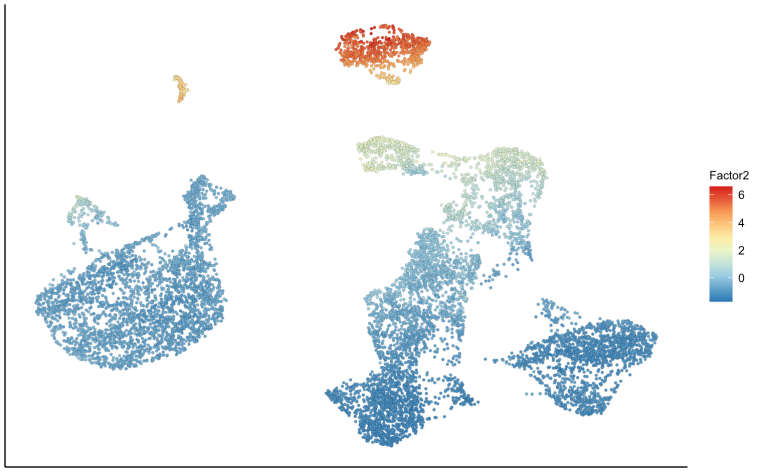
MOFA for scOmics Integration: 10X PBMC

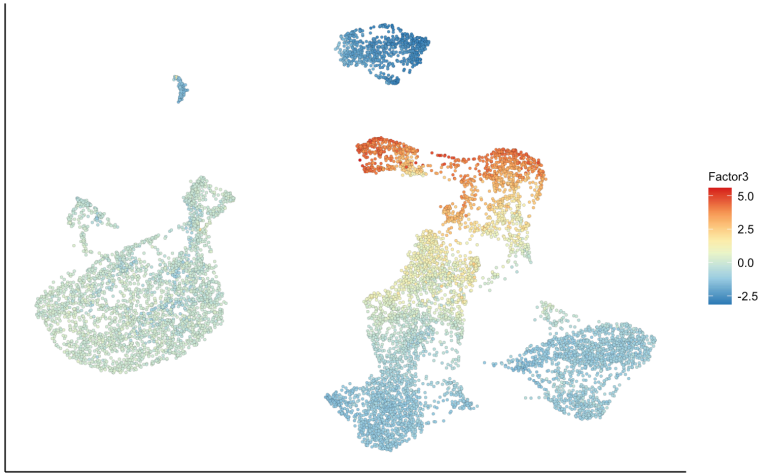


MOFA for scOmics Integration: 10X PBMC









National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet