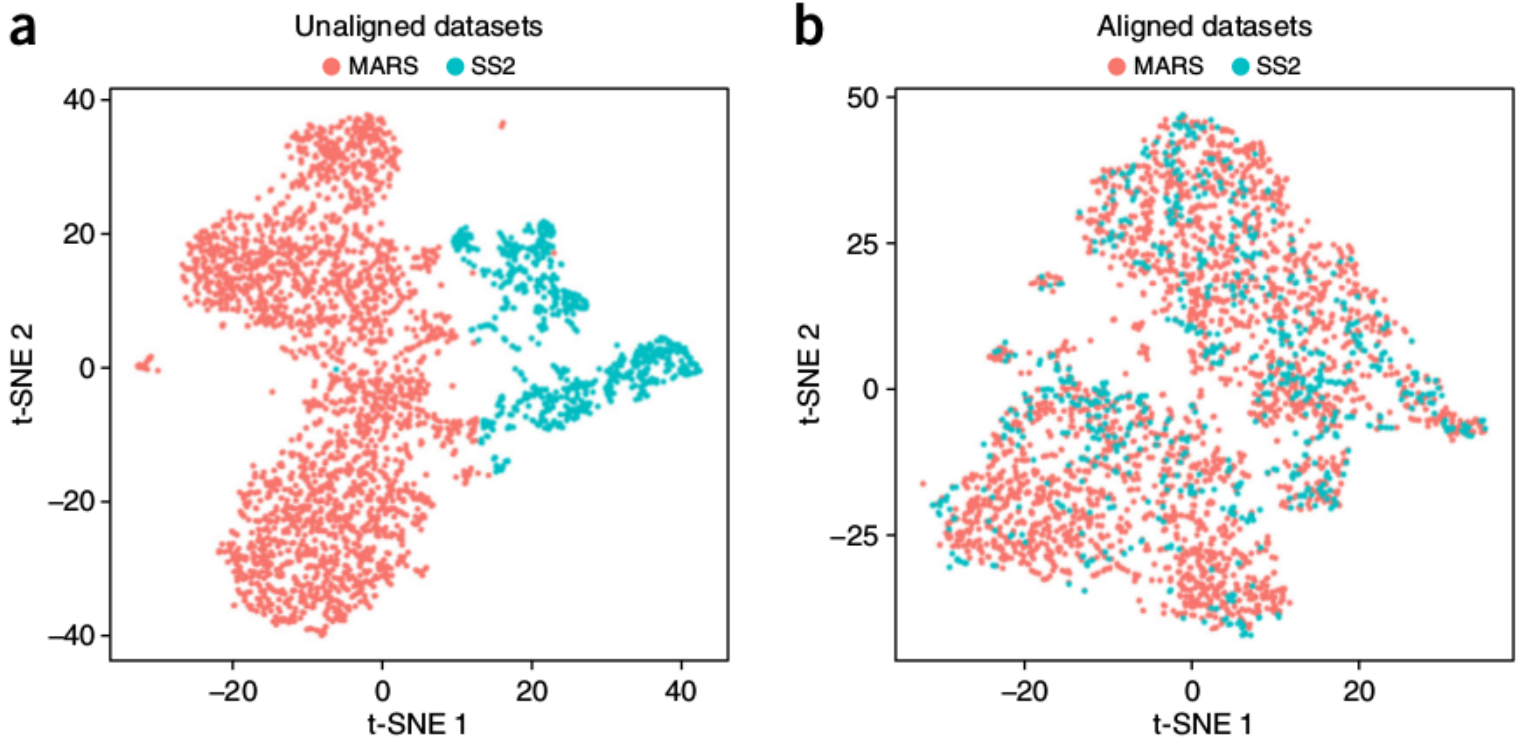
Integration and Batch Correction for Single Cell

Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden

scRNAseq course, 13.02.2024

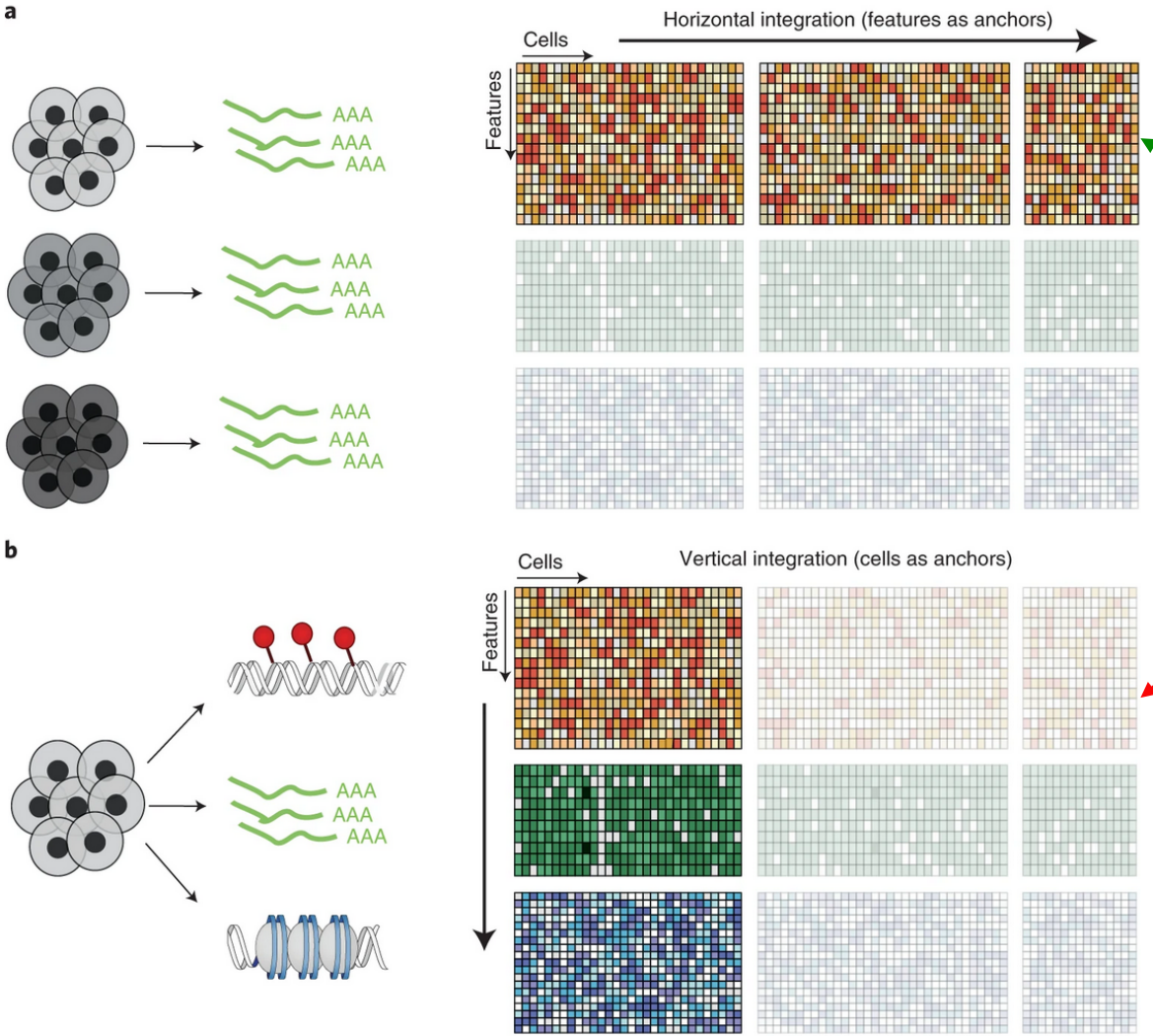


@NikolayOskolkov

GitHub
github.com/NikolayOskolkov

Image adapted from Butler et al. 2018

Integration across cells vs. features



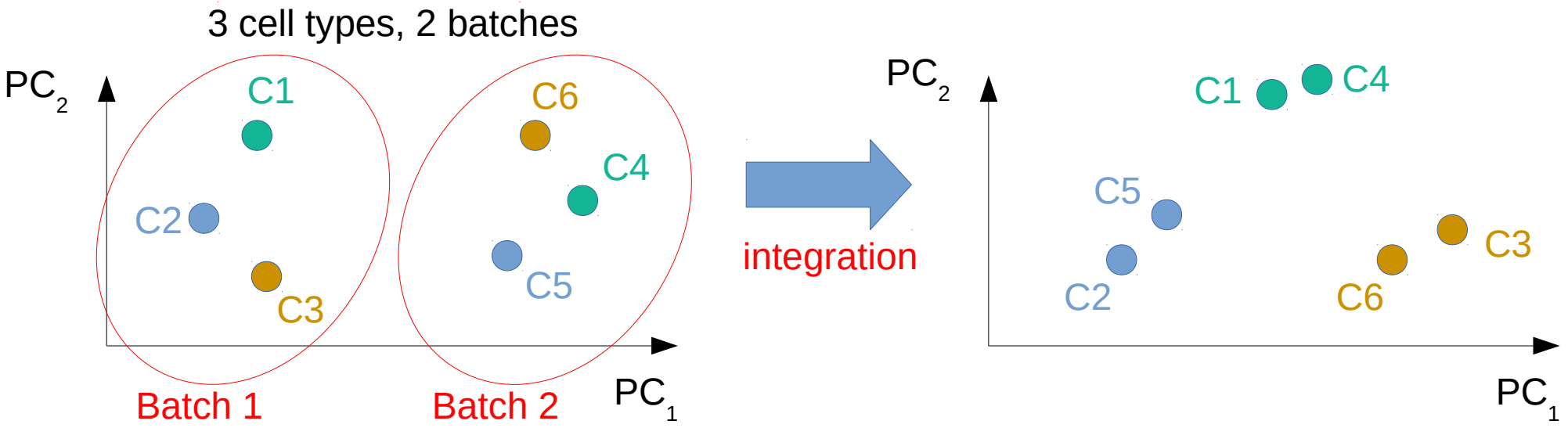
Similar to batch correction problem

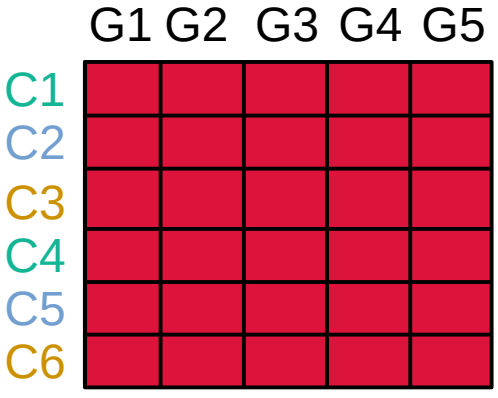
Conceptually difficult problem

Argelaguet et al. 2021

Integration across cells (horizontal integration)

Principles of batch correction





Gene expression

$$= \beta_1 * \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix} + \beta_2 * \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \end{bmatrix}$$

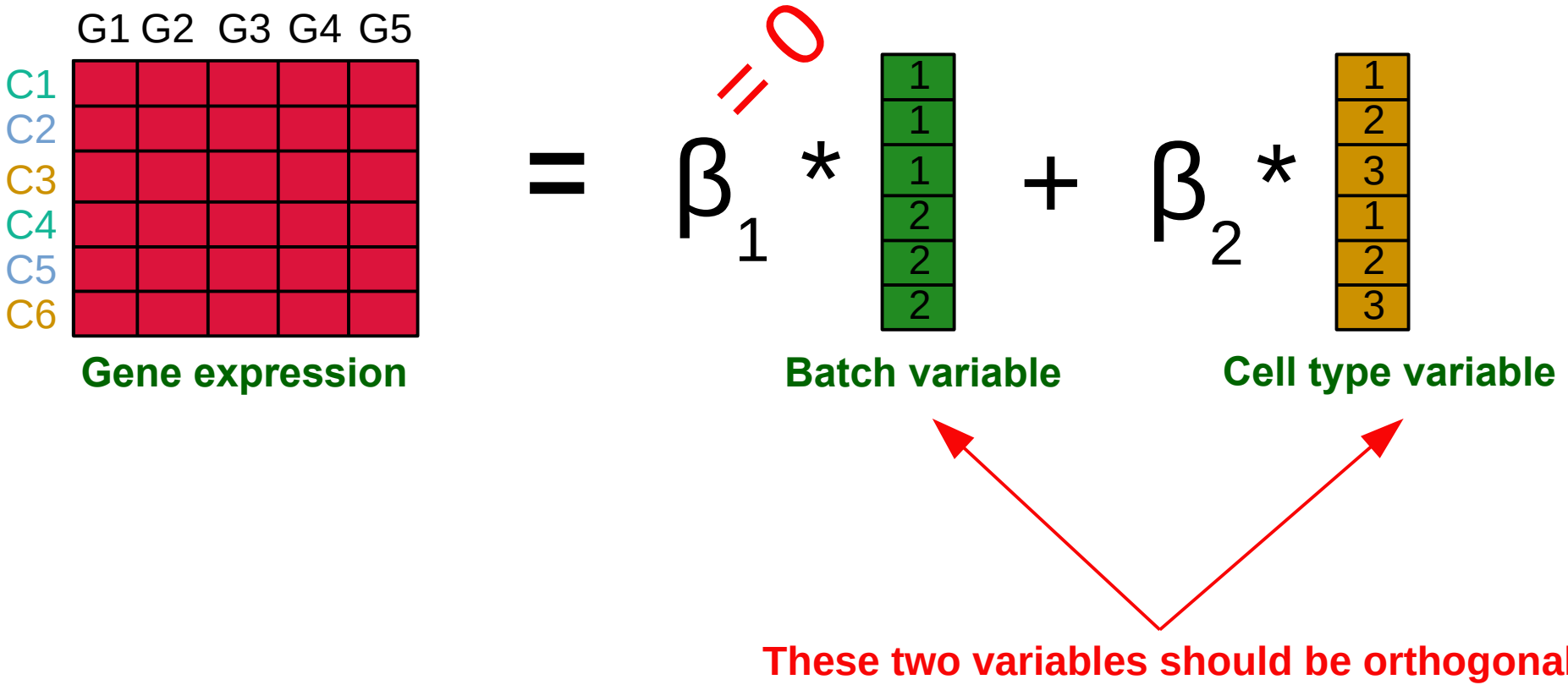
Get rid of this variable ($\beta_1 = 0$)

Keep this variable

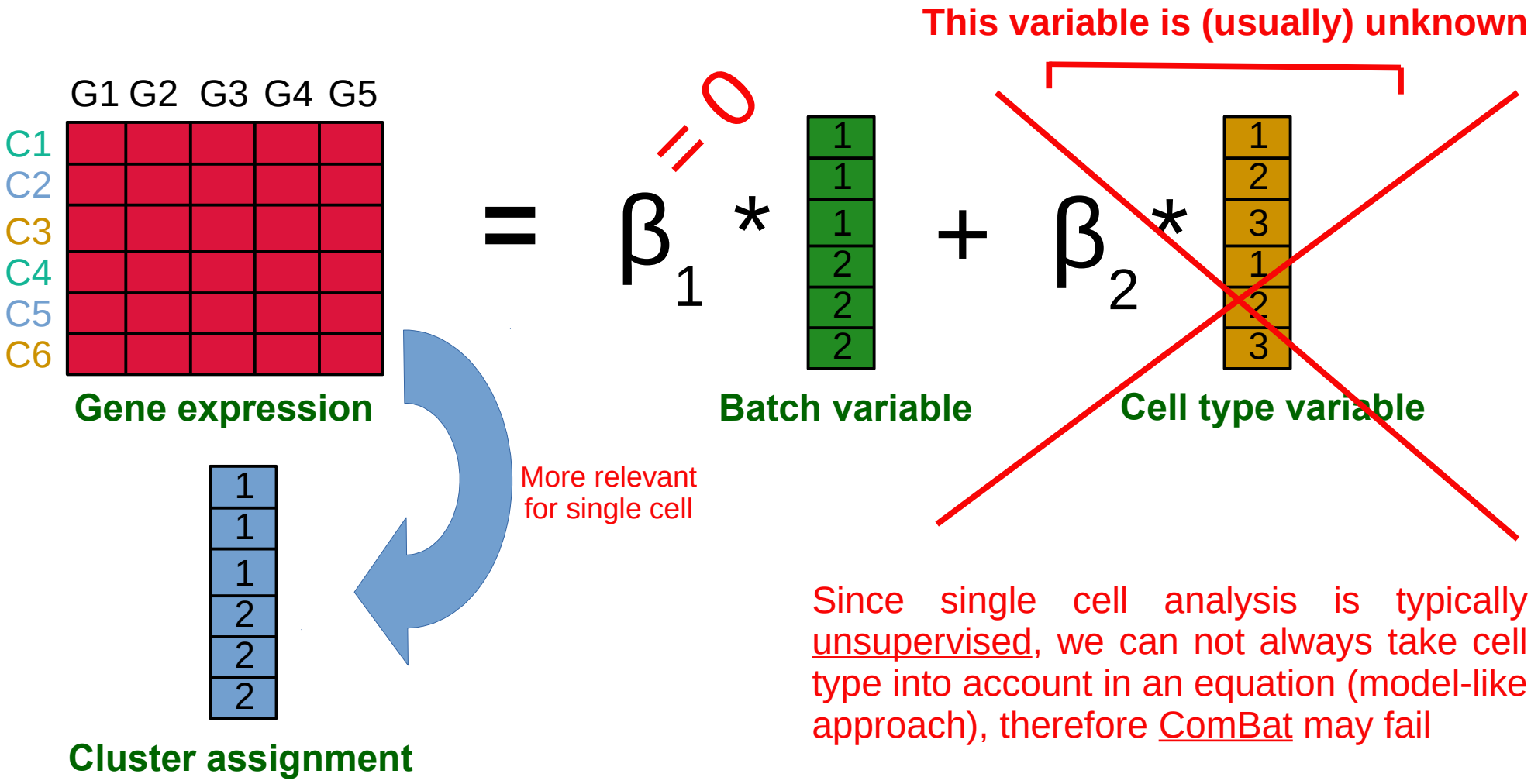
Batch variable

Cell type variable

Caveat: highly desired orthogonality of batch and variable of interest

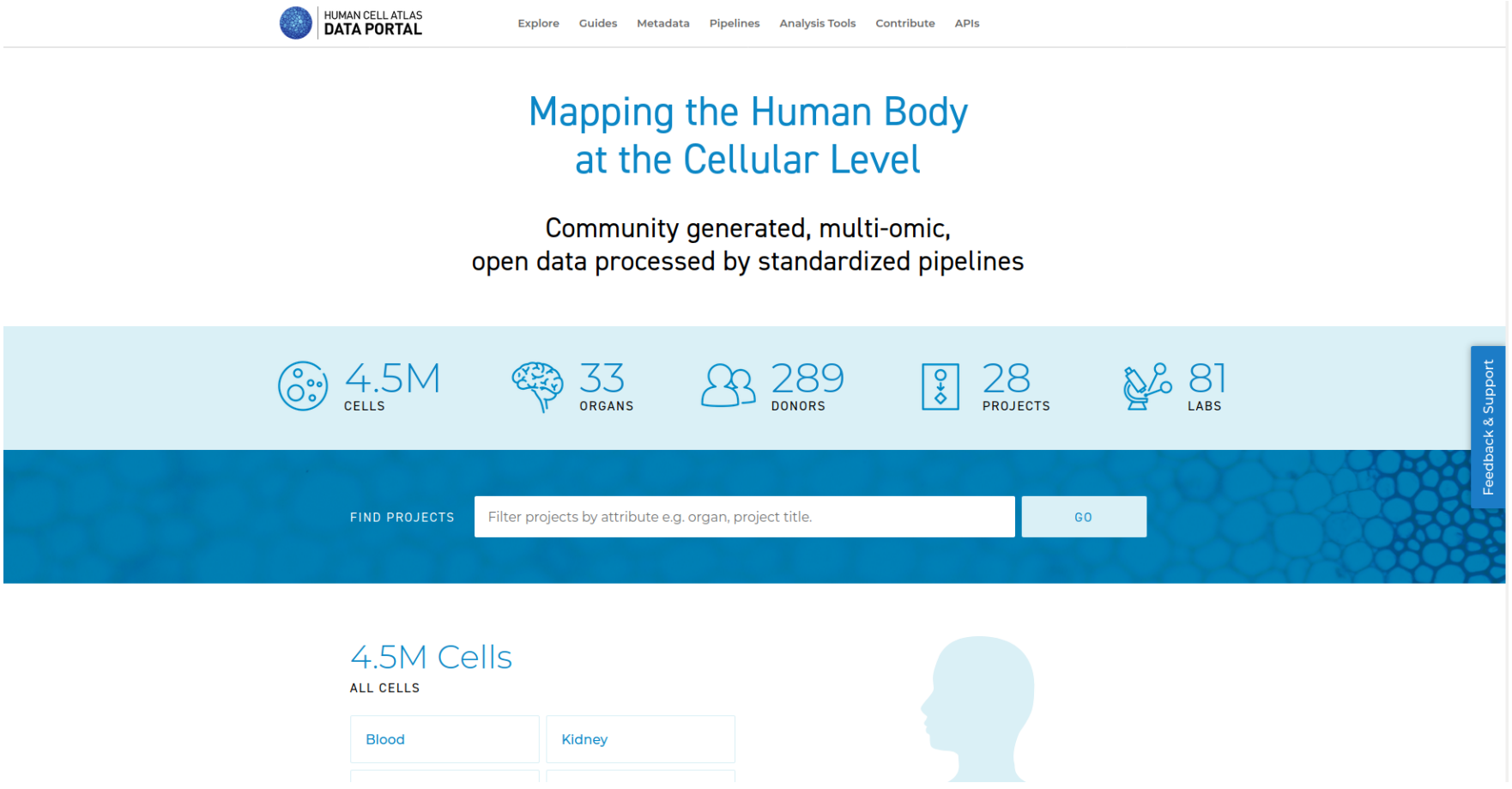


Peculiarities of batch correction in single cell



Since single cell analysis is typically unsupervised, we can not always take cell type into account in an equation (model-like approach), therefore ComBat may fail

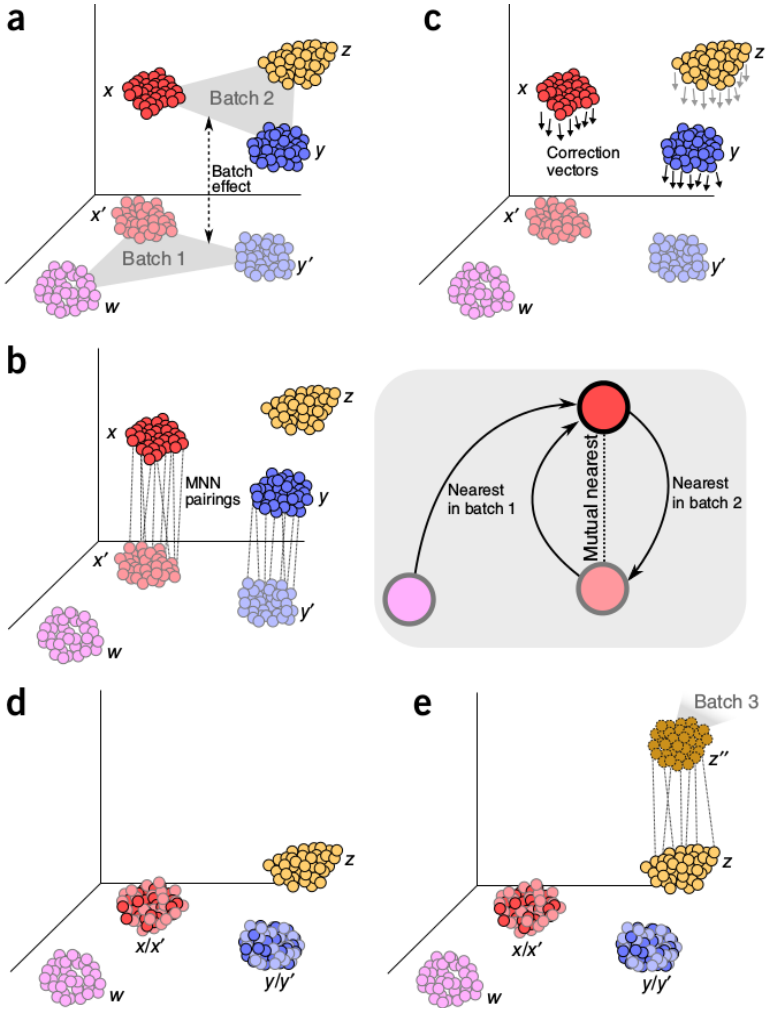
Human Cell Atlas (HCA) Consortium



HCA ambition: create a comprehensive Atlas of human cells from all organs / tissues

Data harmonization / integration is one of major chellenges of HCA

Mutual Nearest Neighbor (MNN)

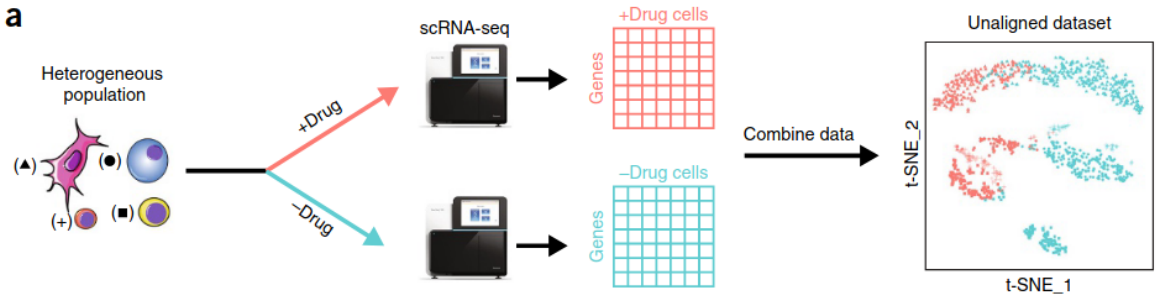


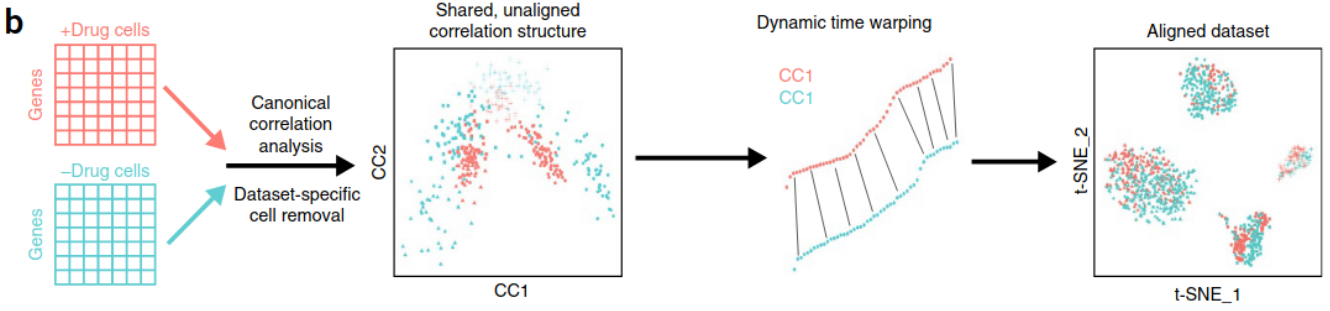
The next step involves identification of mutual nearest neighbors. Consider an scRNA-seq experiment consisting of two batches 1 and 2. For each cell i_1 in batch 1, we find the k cells in batch 2 with the smallest distances to i_1, i.e., its k nearest neighbors in batch 2. We do the same for each cell in batch 2 to find its k nearest neighbors in batch 1. 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 (**Fig. 1**). 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.

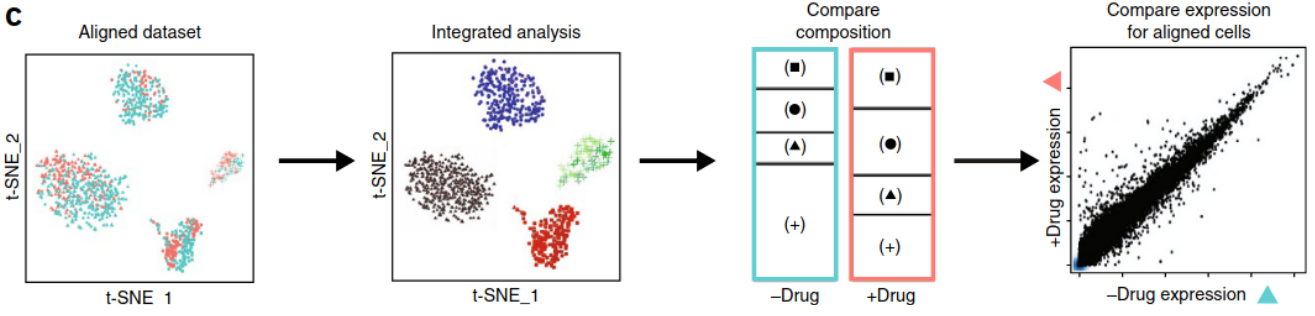
Our use of MNN pairs involves three assumptions: (i) there is at least one cell population that is present in both batches, (ii) the batch effect is almost orthogonal to the biological subspace, and (iii) the batch-effect variation is much smaller than the biological-effect variation between different cell types (more detailed discussion of these assumptions in **Supplementary Note 3**). The biological subspace

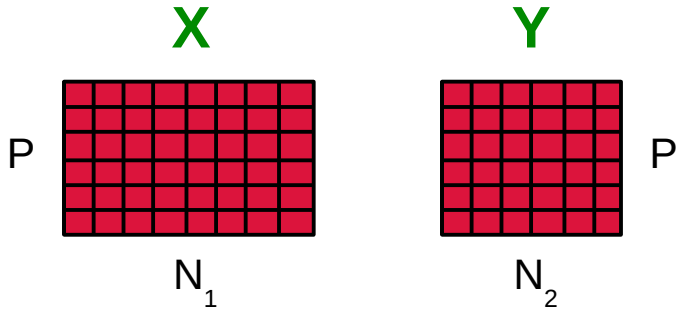
Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Nat Biotechnol. 2018 Jun;36(5):421-427

Seurat CCA + DTW integration





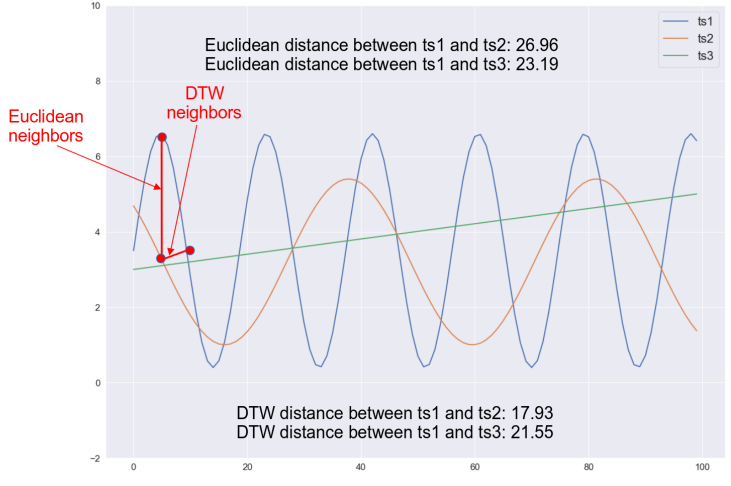




$$A = \mathrm{cov}(X, Y) = X * Y^T$$

$$A * u = \lambda * u$$

Dynamic Time Warping (DTW)



Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Nat Biotechnol. 2018; 36(5): 411-420

Harmony across cells integration

nature methods ARTICLES
https://doi.org/10.1038/s41592-019-0619-0

Fast, sensitive and accurate integration of single-cell data with Harmony

Ilya Korsunsky[1,2,3,4], Nghia Millard[1,2,3,4], Jean Fan[5], Kamil Slowikowski[1,2,3,4], Fan Zhang[1,2,3,4], Kevin Wei[2], Yuriy Baglaenko[1,2,3,4], Michael Brenner[2], Po-ru Loh[1,3,4] and Soumya Raychaudhuri[1,2,3,4,6*]

The emerging diversity of single-cell RNA-seq datasets allows for the full transcriptional characterization of cell types across a wide variety of biological and clinical conditions. However, it is challenging to analyze them together, particularly when datasets are assayed with different technologies, because biological and technical differences are interspersed. We present Harmony (https://github.com/immunogenomics/harmony), an algorithm that projects cells into a shared embedding in which cells group by cell type rather than dataset-specific conditions. Harmony simultaneously accounts for multiple experimental and biological factors. In six analyses, we demonstrate the superior performance of Harmony to previously published algorithms while requiring fewer computational resources. Harmony enables the integration of ~10⁶ cells on a personal computer. We apply Harmony to peripheral blood mononuclear cells from datasets with large experimental differences, five studies of pancreatic islet cells, mouse embryogenesis datasets and the integration of scRNA-seq with spatial transcriptomics data.

Recent technological advances[1] enable unbiased single-cell transcriptional profiling of thousands of cells in one experiment. Projects such as the Human Cell Atlas[2] (HCA) and Accelerating Medicines Partnership[3–5] exemplify the growing body of reference datasets of primary human tissues. While individual experiments incrementally expand our understanding of cell types, a comprehensive catalog of healthy and diseased cells will require the ability to integrate multiple datasets across donors, studies and technological platforms. Moreover, in translational research, joint analyses across tissues and clinical conditions will be essential to identify disease-expanded populations. Since meaningful biological variation in single-cell RNAseq datasets from different studies is often confounded by data source[6], investigators have developed unsupervised multi-dataset integration algorithms[7–10]. These methods embed cells from diverse experimental conditions and biological contexts into a common reduced dimensional embedding to enable shared cell-type identification across datasets.

Here, we introduce Harmony, an algorithm for robust, scalable and flexible multi-dataset integration to meet four key challenges of unsupervised scRNAseq joint embedding: scaling to large datasets, identification of both broad populations and fine-grained subpopulations, flexibility to accommodate complex experimental design, and the power to integrate across modalities. We apply Harmony to a diverse range of examples, including cell lines, peripheral blood mononuclear cells (PBMCs) assayed with different technologies, a meta-analysis of pancreatic islet cells from multiple donors and studies, longitudinal samples from mouse embryogenesis, and cross-modality integration of scRNA-seq data with spatial transcriptomics data. Harmony is available as an R package on github (https://github.com/immunogenomics/harmony), with functions for standalone and Seurat[7] pipeline analyses.

Results

Harmony iteratively learns a cell-specific linear correction function. Harmony begins with a low-dimensional embedding of cells, such as principal components analysis (PCA), (Supplementary Note 1 and Methods). Using this embedding, Harmony first groups cells into multi-dataset clusters (Fig. 1a). We use soft clustering to assign cells to potentially multiple clusters, to account for smooth transitions between cell states. These clusters serve as surrogate variables, rather than to identify discrete cell types. We developed a new soft *k*-means clustering algorithm that favors clusters with cells from multiple datasets (Methods). Clusters disproportionately containing cells from a small subset of datasets are penalized by an information theoretic metric. Harmony allows for multiple different penalties to accommodate multiple technical or biological factors, such as different batches and tissue sources. Soft clustering preserves discrete and continuous topologies while avoiding local minima that might result from maximizing representation too quickly across multiple datasets. After clustering, each dataset has cluster-specific centroids (Fig. 1b) that are used to compute cluster-specific linear correction factors (Fig. 1c). Since clusters correspond to cell types and states, cluster-specific correction factors correspond to individual cell-type and cell-state specific correction factors. In this way, Harmony learns a simple linear adjustment function that is sensitive to intrinsic cellular phenotypes. Finally, each cell is assigned a cluster-weighted average of these terms and corrected by its cell-specific linear factor (Fig. 1d). Since each cell may be in multiple clusters, each cell has a potentially unique correction factor. Harmony iterates these four steps until cell cluster assignments are stable.

Quantifying performance in cell-line data. We first assessed Harmony using three carefully controlled datasets, to evaluate

[1]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA. [2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. [6]Versus Arthritis Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. *e-mail: soumya@broadinstitute.org

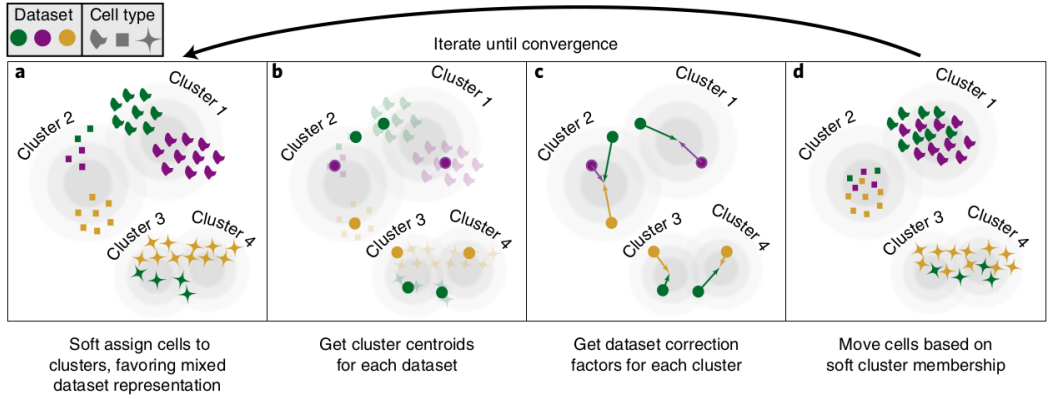


Fig. 1 | Overview of Harmony algorithm. PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects. **a**, Harmony uses fuzzy clustering to assign each cell to multiple clusters, while a penalty term ensures that the diversity of datasets within each cluster is maximized. **b**, Harmony calculates a global centroid for each cluster, as well as dataset-specific centroids for each cluster. **c**, Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. **d**, Finally, Harmony corrects each cell with a cell-specific factor: a linear combination of dataset correction factors weighted by the cell's soft cluster assignments made in step **a**. Harmony repeats steps **a** to **d** until convergence. The dependence between cluster assignment and dataset diminishes with each round. Datasets are represented with colors, cell types with different shapes.

Harmony tries to disrupt relation between the cluster assignment and batch variable by penalizing for diversity across batches in its fuzzy k-means algorithm

Objective function for maximum diversity clustering. The full objective function for Harmony's clustering builds on the previous section. In addition to soft assignment regularization, the function below penalizes clusters with low batch-diversity, for all defined batch variables. This penalty, derived in the following section, depends on the cluster and batch identities $\Omega(R, \phi_i) = \Sigma_{i,k} R_{ki} \log(O_{ki}/E_{ki})\phi_i$.

$$\min_{R,Y} \sum_{i,k} R_{ki}\|Z_i - Y_k\|^2 + \sigma R_{ki} \log R_{ki} + \sigma\theta R_{ki} \log\left(\frac{O_{ki}}{E_{ki}}\right)\phi_i \qquad (3)$$

Scanorama across cells integration

nature biotechnology

ANALYSIS

https://doi.org/10.1038/s41587-019-0113-3

Efficient integration of heterogeneous single-cell transcriptomes using Scanorama

Brian Hie[1], Bryan Bryson[2*] and Bonnie Berger[1,3*]

Integration of single-cell RNA sequencing (scRNA-seq) data from multiple experiments, laboratories and technologies can uncover biological insights, but current methods for scRNA-seq data integration are limited by a requirement for datasets to derive from functionally similar cells. We present Scanorama, an algorithm that identifies and merges the shared cell types among all pairs of datasets and accurately integrates heterogeneous collections of scRNA-seq data. We applied Scanorama to integrate and remove batch effects across 105,476 cells from 26 diverse scRNA-seq experiments representing 9 different technologies. Scanorama is sensitive to subtle temporal changes within the same cell lineage, successfully integrating functionally similar cells across time series data of CD14$^+$ monocytes at different stages of differentiation into macrophages. Finally, we show that Scanorama is orders of magnitude faster than existing techniques and can integrate a collection of 1,095,538 cells in just ~9 h.

Individual single-cell RNA sequencing (scRNA-seq) experiments have already been used to discover new cell states and reconstruct cellular differentiation trajectories[1–7]. Through global efforts such as the Human Cell Atlas[8], researchers are now generating large, comprehensive collections of scRNA-seq datasets that profile a diverse range of cellular functions, which promise to enable high-resolution insight into processes underlying fundamental biology and disease. Assembling large, unified reference datasets, however, may be compromised by differences due to experimental batch, sample donor or experimental technology. While recent approaches have shown that it is possible to integrate scRNA-seq studies across multiple experiments[9,10], these approaches automatically assume that all datasets share at least one cell type in common[9] or that the gene expression profiles share largely the same correlation structure across all datasets[10]. These methods are therefore prone to overcorrection, especially when integrating collections of datasets with considerable differences in cellular composition.

Here, we present Scanorama: a strategy for efficiently integrating multiple scRNA-seq datasets, even when they are composed of heterogeneous transcriptional phenotypes. Our approach is analogous to computer vision algorithms for panorama stitching that identify images with overlapping content and merge these into a larger panorama (Fig. 1a)[11]. Likewise, Scanorama automatically identifies scRNA-seq datasets containing cells with similar transcriptional profiles and can leverage those matches for batch correction and integration (Fig. 1b), without also merging datasets that do not overlap (Methods). Scanorama is robust to different dataset sizes and sources, preserves dataset-specific populations and does not require that all datasets share at least one cell population[9].

Our approach generalizes mutual nearest-neighbors matching, a technique that finds similar elements between two datasets, to instead find similar elements among many datasets. Originally developed for pattern matching in images[12], finding mutual nearest neighbors has also been used to identify common cell types between two scRNA-seq datasets at a time[9]. However, to align more than two datasets, existing methods[9,10] select one dataset as a reference and successively integrate all other datasets into the reference, one at a time, which may lead to suboptimal results depending on the order in which the datasets are considered (Supplementary Fig. 1). Although Scanorama takes a similar approach when aligning a collection of two datasets, on larger collections of data it is insensitive to order and less vulnerable to overcorrection because it finds matches between all pairs of datasets.

To optimize the process of searching for matching cells among all datasets, we introduce two key procedures. Instead of performing the nearest neighbor search in the high-dimensional gene space, we compress the gene expression profiles of each cell into a low-dimensional embedding using an efficient, randomized singular value decomposition (SVD)[13] of the cell-by-gene expression matrix, which also helps improve the method's robustness to noise. Further, we use an approximate nearest neighbor search based on hyperplane locality sensitive hashing[14] and random projection trees[15] to greatly reduce the nearest neighbor query time both asymptotically and in practice (Methods).

Notably, Scanorama can perform both scRNA-seq dataset integration and (optionally) batch correction. Integration methods (for example, Seurat CCA[10], a previous integrative method based on a canonical correlation analysis (CCA) strategy) find lower-dimensional representations of high-dimensional gene expression vectors such that the representations minimize confounding variation (for example, batch effects) with respect to some variation of interest (for example, biological differences among cell types). Batch correction methods (for example, scran MNN[9], a previous batch correction method based on a simpler accumulative mutual nearest-neighbors (MNN) strategy) also remove confounding variation in the original high-dimensional space. Scanorama always performs integration of low-dimensional embeddings but can also perform batch correction if required. Although incurring a greater computational cost, Scanorama makes batch correction feasible for large datasets, enabling a wider array of downstream analyses. For example, differential expression analysis can be performed on batch-corrected gene expression data but not on integrated low-dimensional representations.

[1]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA. [2]Department of Biological Engineering, MIT, Cambridge, MA, USA. [3]Department of Mathematics, MIT, Cambridge, MA, USA. *e-mail: bryand@mit.edu; bab@mit.edu

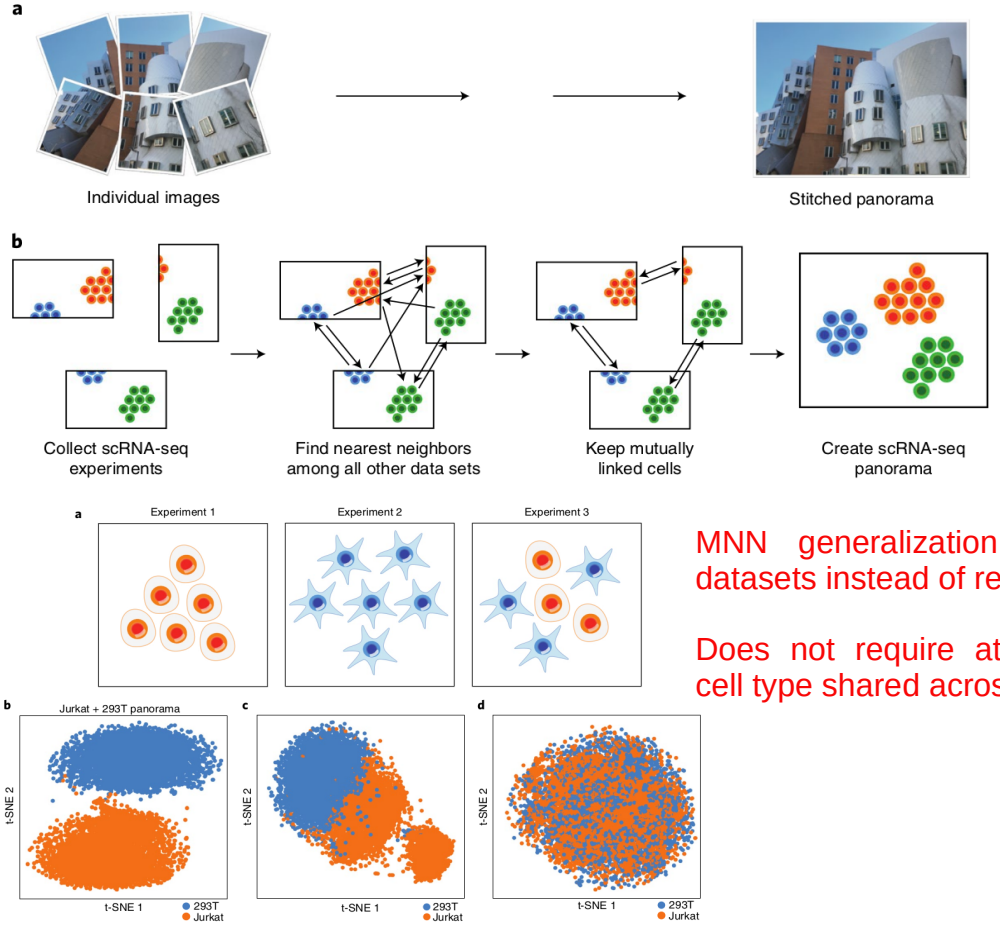
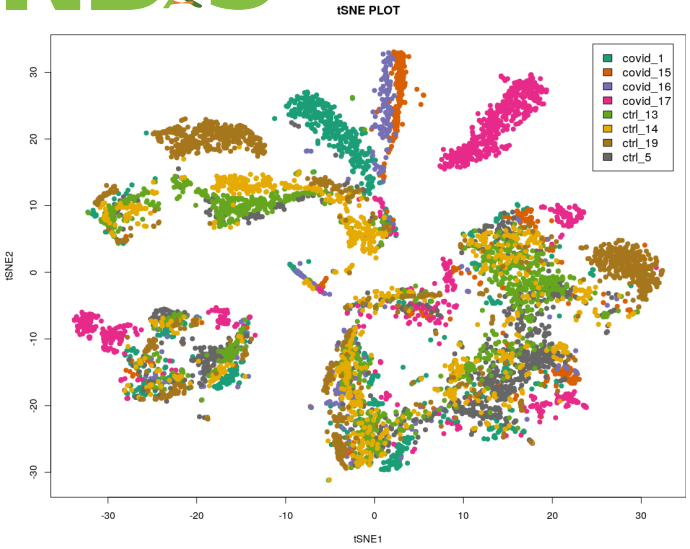


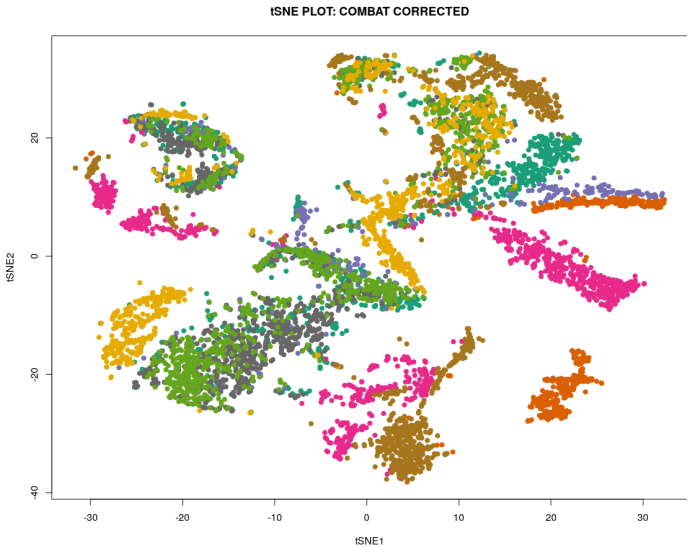
Fig. 2 | Scanorama correctly integrates a simple collection of datasets where other methods fail. a, We apply Scanorama to a collection of three datasets[11]: one entirely of Jurkat cells ($n = 3,257$ cells) (Experiment 1), one entirely of 293T cells ($n = 2,885$ cells) (Experiment 2) and a 50/50 mixture of Jurkat and 293T cells ($n = 3,388$ cells) (Experiment 3). **b**, Our method correctly identifies Jurkat cells (orange) and 293T cells (blue) as two separate clusters. **c**,**d**, Existing methods for scRNA-seq dataset integration are sensitive to the order in which they consider datasets (see Supplementary Fig. 1) and can incorrectly merge a Jurkat dataset and a 293T dataset together first before subsequently incorporating a 293T/Jurkat mixture, forming clusters that do not correspond to actual cell types: scran MNN corrected (**c**) and Seurat CCA integrated (**d**).

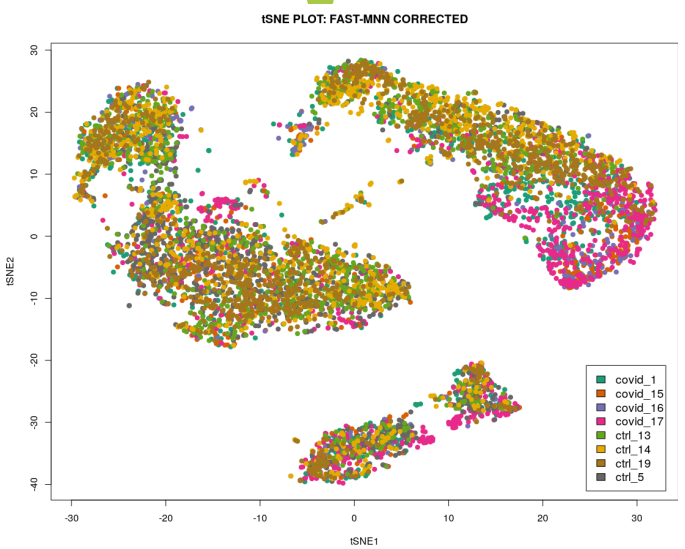
MNN generalization to many datasets instead of ref and query

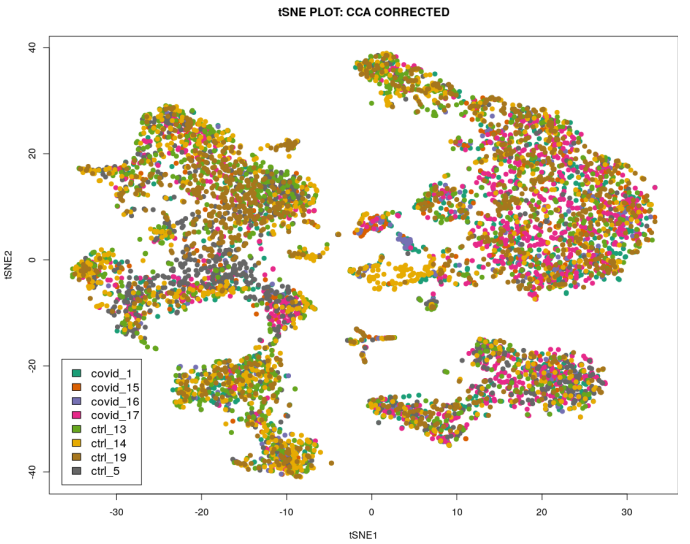
Does not require at least one cell type shared across datasets

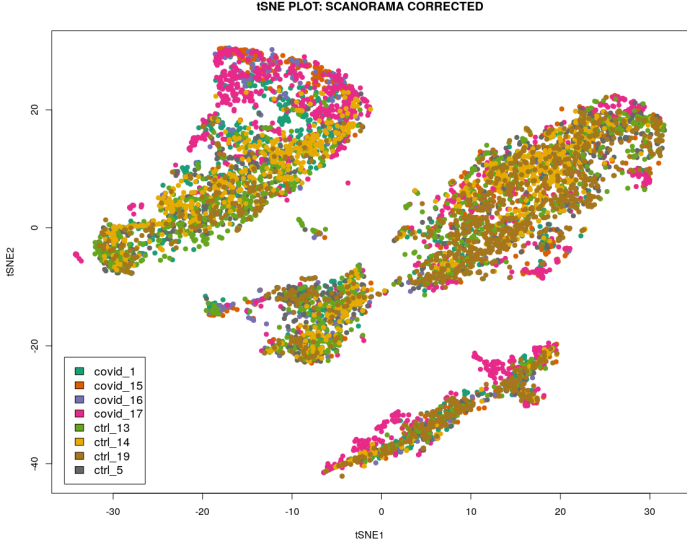
Comparing across cells integration methods

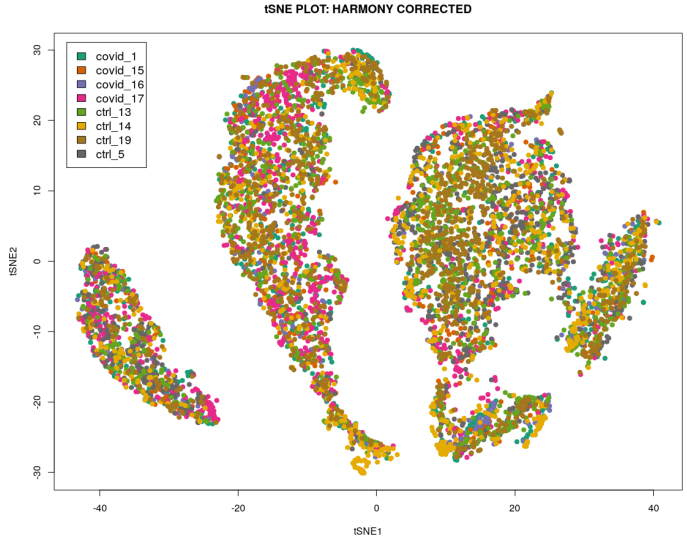












ComBat: low-complexity datasets (subtle batch, few batches, few cell populations)

A recent comparison of classical batch correction methods has revealed that ComBat (Johnson *et al*, 2006) performs well also for single-cell experiments of low-to-medium complexity (Buttner *et al*, 2019). ComBat consists of a linear model of gene expression where the batch contribution is taken into account both in the mean and the variance of the data (Fig 3). Irrespective of computational methods, the best method of batch correction is pre-empting the effect and avoiding it altogether by clever experimental design (Hicks *et al*, 2017). Batch effects can be avoided by pooling cells across experimental conditions and samples. Using strategies such as cell tagging (preprint: Gehring *et al*, 2018), or via genetic variation (Kang *et al*, 2018), it is possible to demultiplex cells that were pooled in the experiment.

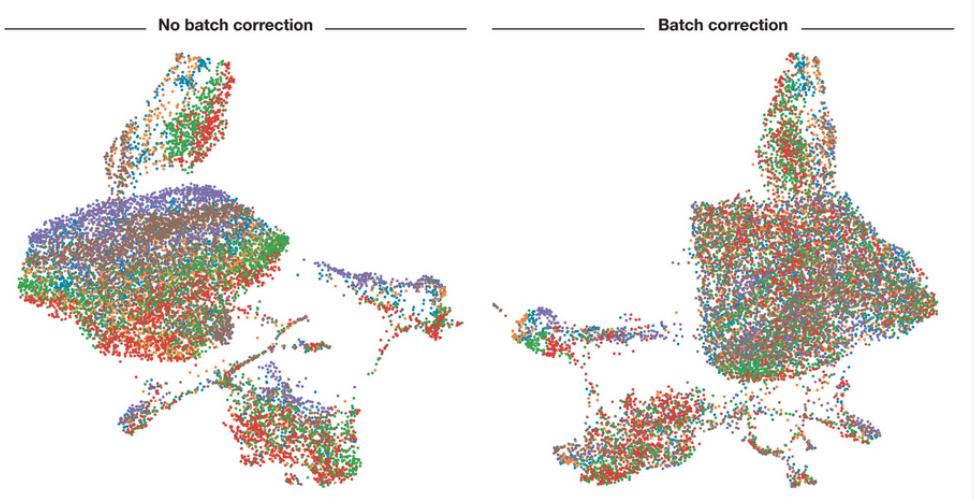


Figure 3 UMAP visualization before and after batch correction

Cells are coloured by sample of origin. Separation of batches is clearly visible before batch correction and less visible afterwards. Batch correction was performed using ComBat on mouse intestinal epithelium data from Haber *et al* (2017).

Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun 19;15(6):e8746. doi: 10.15252/msb.20188746

ANALYSIS NATURE METHODS

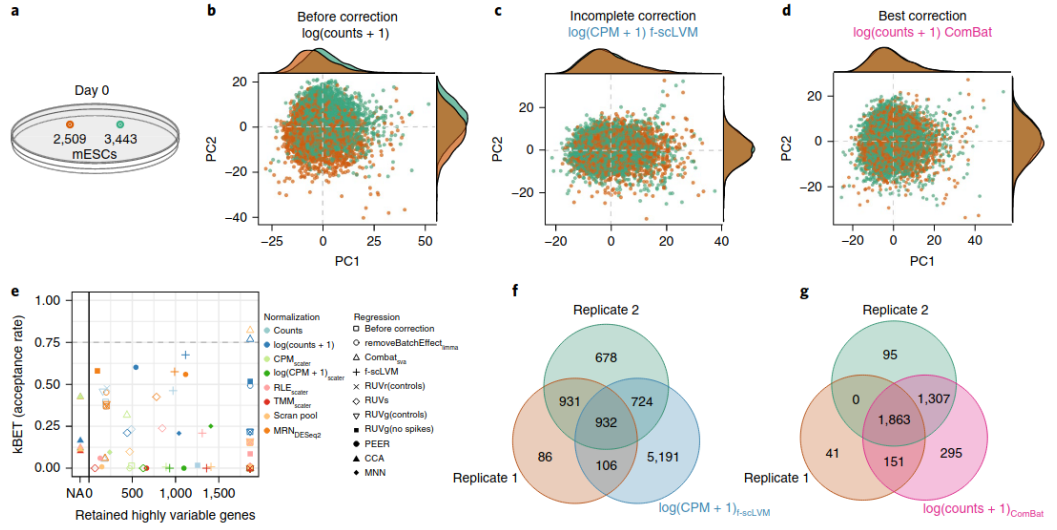


Fig. 3 | ComBat provides the best correction on mESC inDrop technical replicates. a, The inDrop protocol provides a large unique-molecular-identifier-count dataset with two technical replicates. **b**–**d**, PCA plots showing log-normalized counts (**b**), a biology-removing batch removal (f-scLVM on log-transformed CPM; **c**) and a biology-preserving batch removal (ComBat on log-transformed counts; **d**). Density plots on the axes show the frequency of replicates along the PCs. On the basis of visual inspection, the approaches in **c** and **d** appear to work equally well. **e**, Percentage of retained HVGs versus the mean acceptance rate (1 – rejection rate, from $n = 100$ kBET runs) for all combinations of normalizations and batch-regression approaches. Seurat's CCA alignment batch-corrects data only in a latent space generated by manifold learning, and thus we could not compute HVGs for it. **f**,**g**, HVGs per replicate before correction and for the whole dataset after batch correction. HVGs in each replicate are computed on log(counts + 1) values. f-scLVM (**f**) retained 932 HVGs but had a high false positive rate, whereas ComBat (**g**) captured all HVGs with a low false positive rate.

Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019 Jan;16(1):43-49. doi: 10.1038/s41592-018-0254-1

Harmony / Scanorama: complex datasets

Tran et al. Genome Biology (2020) 21:12
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

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*

Abstract

Background: Large-scale single-cell transcriptomic datasets generated using different technologies contain batch-specific systematic variations that present a challenge to batch-effect removal and data integration. With continued growth expected in scRNA-seq data, achieving effective batch integration with available computational resources is crucial. Here, we perform an in-depth benchmark study on available batch correction methods to determine the most suitable method for batch-effect removal.

Results: We compare 14 methods in terms of computational runtime, the ability to handle large datasets, and batch-effect correction efficacy while preserving cell type purity. Five scenarios are designed for the study: identical cell types with different technologies, non-identical cell types, multiple batches, big data, and simulated data. Performance is evaluated using four benchmarking metrics including kBET, LISI, ASW, and ARI. We also investigate the use of batch-corrected data to study differential gene expression.

Conclusion: Based on our results, Harmony, LIGER, and Seurat 3 are the recommended methods for batch integration. Due to its significantly shorter runtime, Harmony is recommended as the first method to try, with the other methods as viable alternatives.

Keywords: Single-cell RNA-seq, Batch correction, Batch effect, Integration, Differential gene expression

Introduction

Technological advances in the recent years have increased our ability to generate high-throughput single-cell gene expression data. Single-cell data is often compiled from multiple experiments with differences in capturing times, handling personnel, reagent lots, equipments, and even technology platforms. These differences lead to large variations or batch effects in the data, and can confound biological variations of interest during data integration. As such, effective batch-effect removal is essential. Batch

developed for microarray data batch correction such as ComBat [1] and limma [2] have been employed on single-cell RNA-seq (scRNA-seq) data. However, single-cell experiments suffer from "drop out" events due to the stochasticity of gene expression, or failure in RNA capture or amplification during sequencing [3]. This has prompted efforts to develop workflows to handle data with such characteristics [4–6].

A popular and successful approach, pioneered by Haghverdi et al. [5], identifies cell mappings between

nature methods ANALYSIS
https://doi.org/10.1038/s41592-021-01336-8

Check for updates

OPEN

Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken[1], M. Büttner[1], K. Chaichoompu[1], A. Danese[1], M. Interlandi[2], M. F. Mueller[1], D. C. Strobl[1], L. Zappia[1,3], M. Dugas[4], M. Colomé-Tatché[1,5,6] ✉ and Fabian J. Theis[1,3,5] ✉

Single-cell atlases often include samples that span locations, laboratories and conditions, leading to complex, nested batch effects in data. Thus, joint analysis of atlas datasets requires reliable data integration. To guide integration method choice, we benchmarked 68 method and preprocessing combinations on 85 batches of gene expression, chromatin accessibility and simulation data from 23 publications, altogether representing >1.2 million cells distributed in 13 atlas-level integration tasks. We evaluated methods according to scalability, usability and their ability to remove batch effects while retaining biological variation using 14 evaluation metrics. We show that highly variable gene selection improves the performance of data integration methods, whereas scaling pushes methods to prioritize batch removal over conservation of biological variation. Overall, scANVI, Scanorama, scVI and scGen perform well, particularly on complex integration tasks, while single-cell ATAC-sequencing integration performance is strongly affected by choice of feature space. Our freely available Python module and benchmarking pipeline can identify optimal data integration methods for new data, benchmark new methods and improve method development.

The complexity of single-cell omics datasets is increasing. Current datasets often include many samples[1], generated across multiple conditions[2], with the involvement of multiple laboratories[3]. Such complexity, which is common in reference atlas initiatives such as the Human Cell Atlas[4], creates inevitable batch effects. Therefore, the development of data integration methods that overcome the complex, nonlinear, nested batch effects in these data has become a priority: a grand challenge in single-cell RNA-seq data analysis[5,6].

Batch effects represent unwanted technical variation in data that result from handing cells in distinct batches. These effects can arise from variations in sequencing depth, sequencing lanes, read length, plates or flow cells, protocol, experimental laboratories, sample acquisition and handling, sample composition, reagents or media and/or sampling time. Furthermore, biological factors such as tissues, spatial locations, species, time points or inter-individual variation can also be regarded as a batch effect.

A single-cell data integration method aims to combine high-throughput sequencing datasets or samples to produce a self-consistent version of the data for downstream analysis[7]. Batch-integrated cellular profiles are represented as an integrated graph, a joint embedding or a corrected feature matrix.

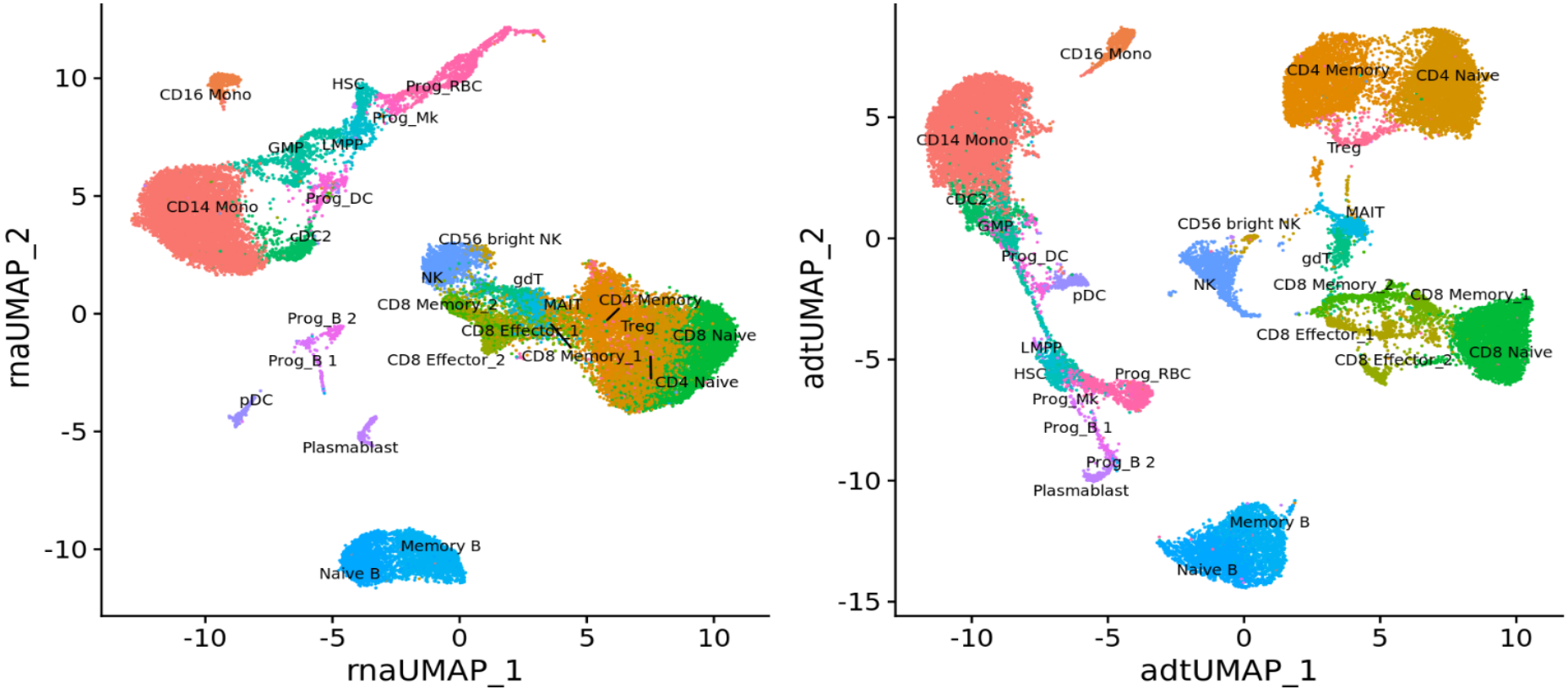
Currently, at least 49 integration methods for scRNA-seq data are available[8] (as of November 2020, Supplementary Table 1). In the absence of objective metrics, subjective opinions based on visualizations of integrated data will determine method evaluation. Benchmarking integration methods facilitates this process to provide an unbiased guide to method choice.

compare different output options such as corrected features or joint embeddings, finding that ComBat[11] or the linear, principal component analysis (PCA)-based, Harmony method[9] outperformed more complex, nonlinear, methods.

Here, we present a benchmarking study of data integration methods in complex integration tasks, such as tissue or organ atlases. Specifically, we benchmarked 16 popular data integration tools on 13 data integration tasks consisting of up to 23 batches and 1 million cells, for both scRNA- and single-cell ATAC-sequencing (scRNA-seq and scATAC-seq) data. We selected 12 single-cell data integration tools: mutual nearest neighbors (MNN)[12] and its extension FastMNN[12], Seurat v3 (CCA and RPCA)[13], scVI[14] and its extension to an annotation framework (scANVI[15]), Scanorama[16], batch-balanced *k* nearest neighbors (BBKNN)[17], LIGER[18], Conos[19], SAUCIE[20] and Harmony[21]; one bulk data integration tool (ComBat[22]); a method for clustering with batch removal (DESC[23]) and two perturbation modeling tools developed previously by one of the authors (trVAE[24] and scGen[25]). Moreover, we use 14 metrics to evaluate the integration methods on their ability to remove batch effects while conserving biological variation. We focus in particular on assessing the conservation of biological variation beyond cell identity labels via new integration metrics on trajectories or cell-cycle variation. We find that Scanorama and scVI perform well, particularly on complex integration tasks. If cell annotations are available, scGen and scANVI outperform most other methods across tasks, and Harmony and LIGER are effective for scATAC-seq data integration on window and peak feature spaces.

Integration across features (vertical integration)

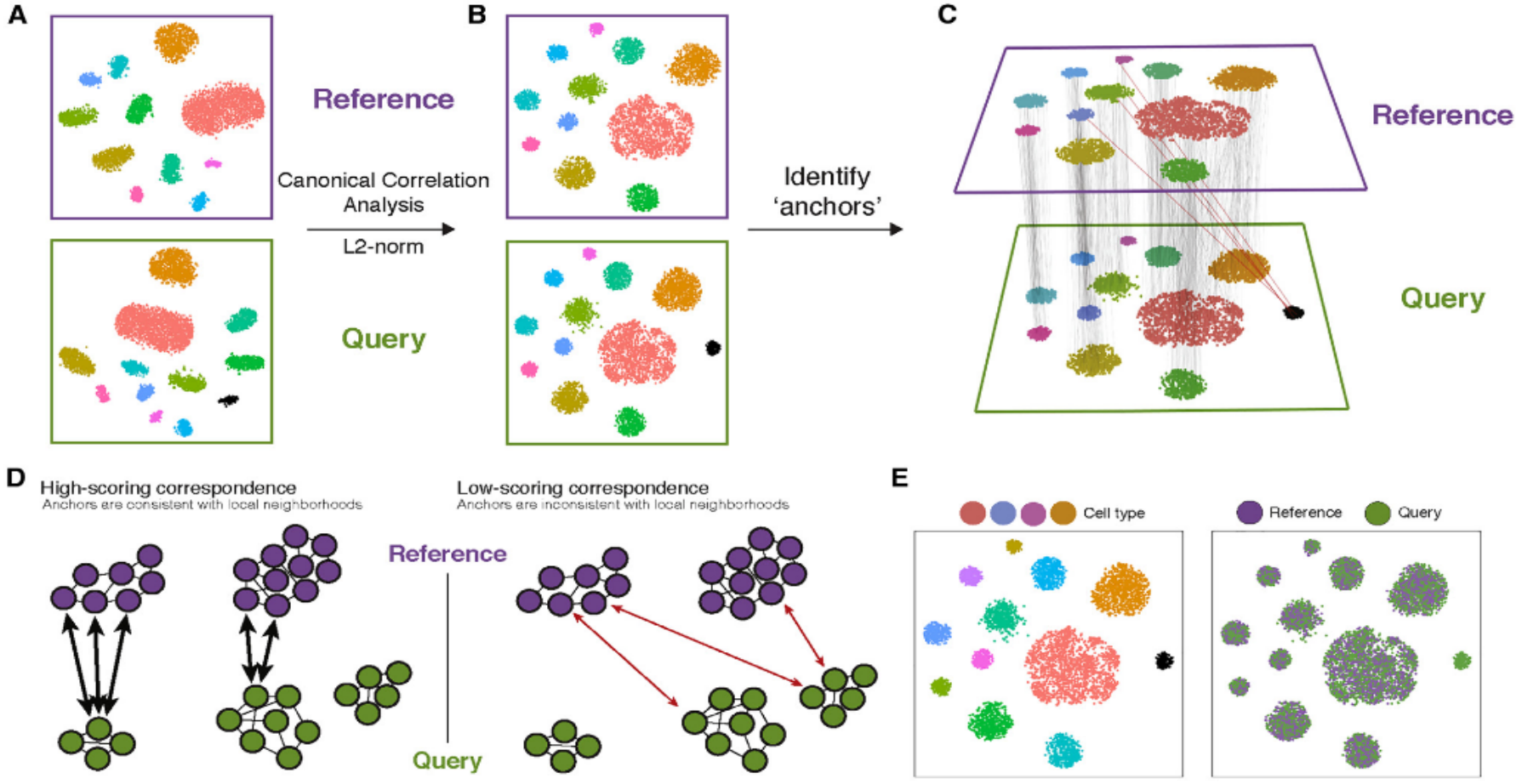
scRNAseq != Function (not always)



"Despite tremendous functional diversity, distinct populations of T cells such as effector, regulatory, γδ, and mucosal associated invariant T (MAIT), often cannot be effectively separated by scRNA-seq alone, even when using the most sensitive and cutting-edge technologies"

Hao et al. Cell 2021, 184(13): 3573-3587.e29. doi: 10.1016/j.cell.2021.04.048

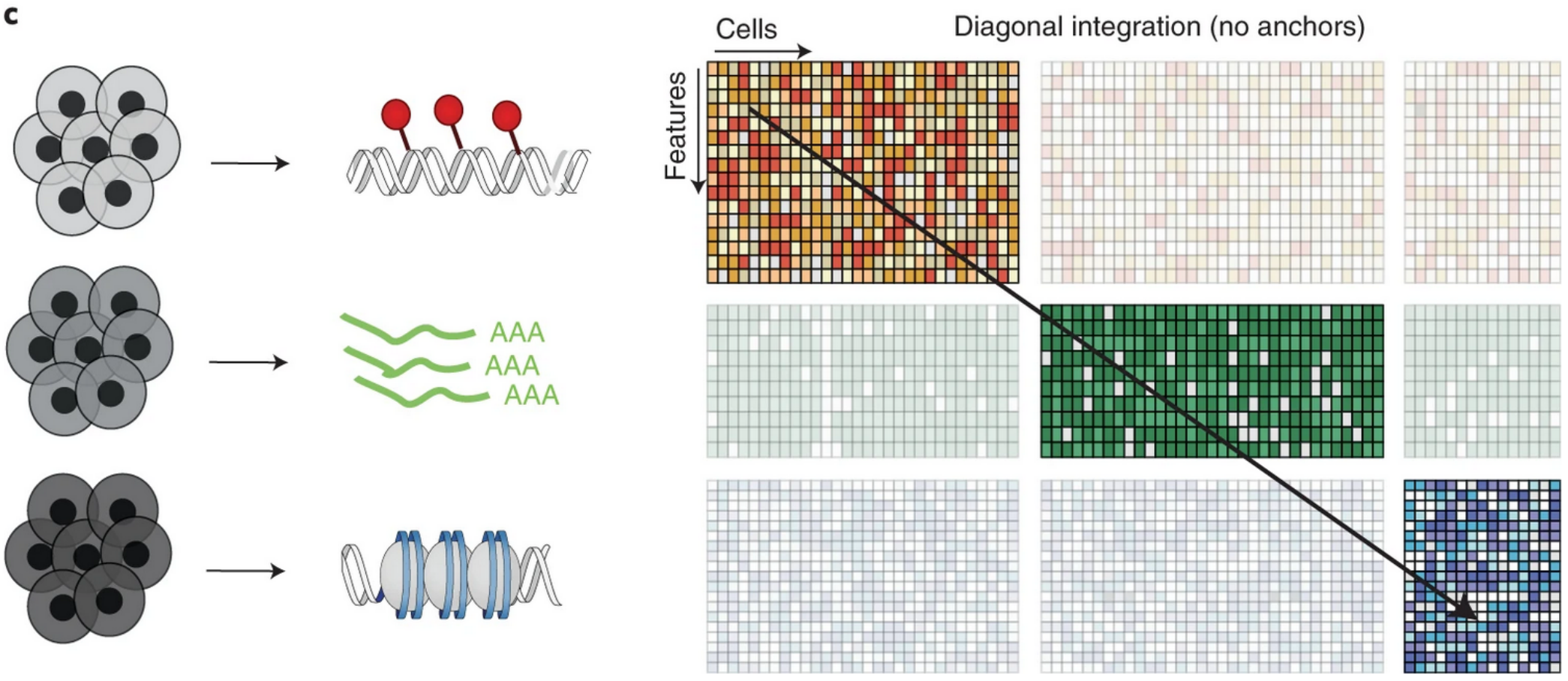
Seurat “transfer anchors across Omics”



Stuart T et al., Cell. 2019 Jun 13;177(7):1888-1902.e21.

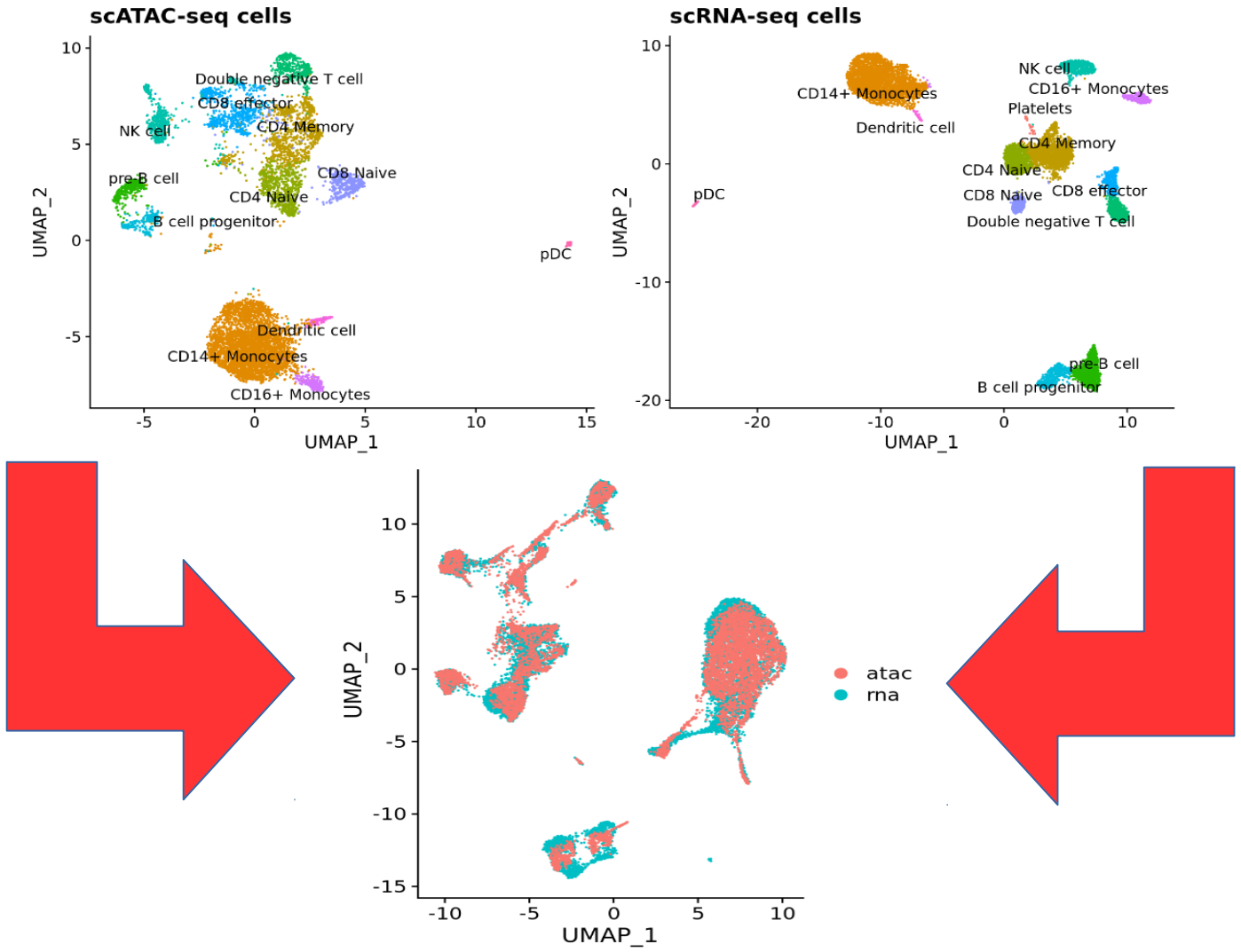
Not the same biological cells but scATAC peaks can be assgned to genes, so the feature names are the same

Seurat “transfer anchors across Omics” as an example of diagonal integration

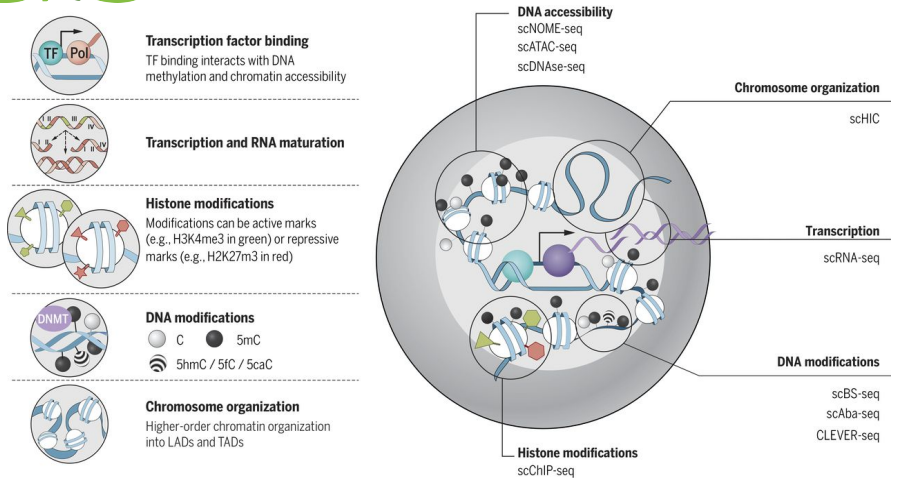


Argelaguet et al. 2021

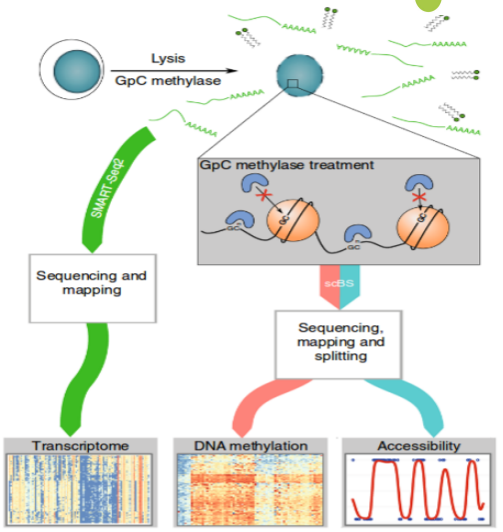
Seurat “transfer anchors across Omics”



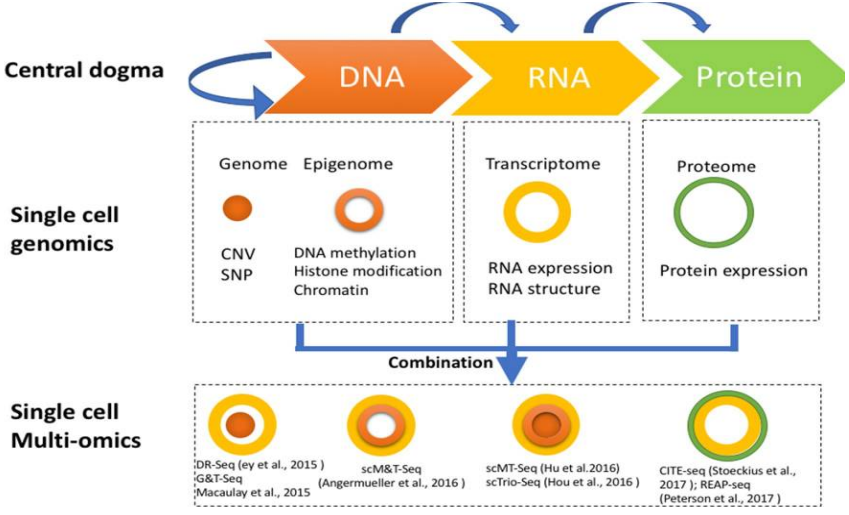
Multimodal scOmics technologies



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



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

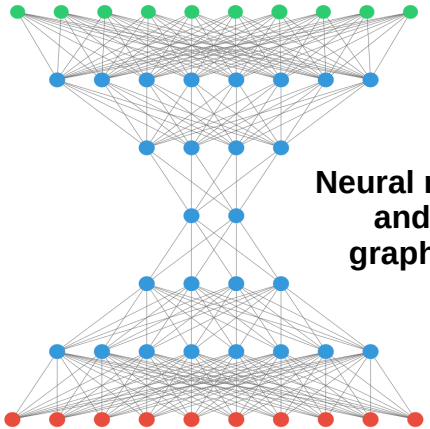


Hu et al., 2018, Frontier in Cell and Developmental Biology 6, 1-13

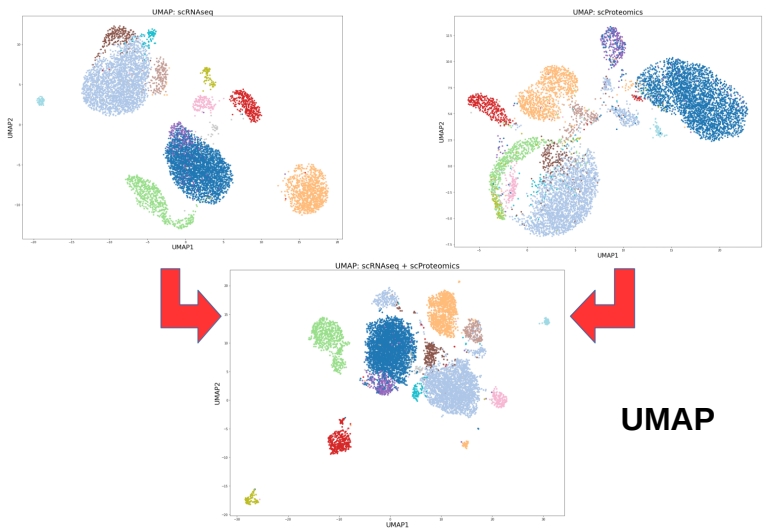
10X Genomics Multiome ATAC + Gene Expression

Overview of Omics integration methods

Convert to common space

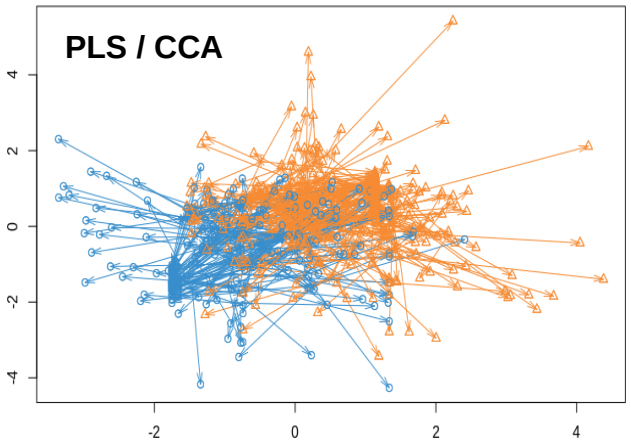


Neural nets and graphs

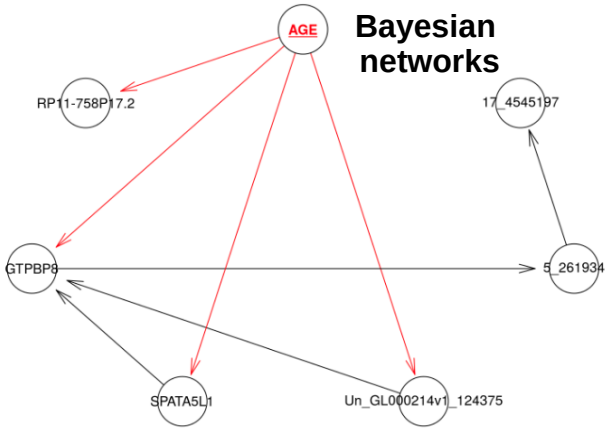


UMAP

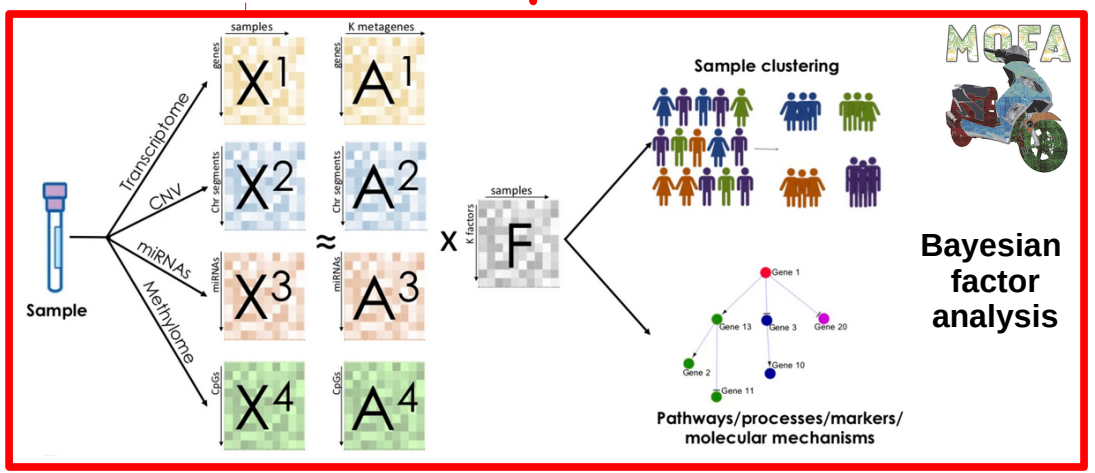
Extract common variation



Combine via Bayes rule



Bayesian networks



Bayesian factor analysis

MOFA (general) and MOFA+ (single cell)

Method

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet[1,†], Britta Velten[2,†], Damien Arnol[1], Sascha Dietrich[3], Thorsten Zenz[3,4,5], John C Marioni[1,6,7], Florian Buettner[1,8,*], Wolfgang Huber[2,**] & Oliver Stegle[1,2,***]

Abstract

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and *ex vivo* drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

Keywords data integration; dimensionality reduction; multi-omics; personalized medicine; single-cell omics
Subject Categories Computational Biology; Genome-Scale & Integrative Biology; Methods & Resources
DOI 10.15252/msb.20178124 | Received 27 November 2017 | Revised 28 May 2018 | Accepted 29 May 2018
Mol Syst Biol. (2018) 14: e8124

Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenome profiling (Hasin *et al*, 2017). Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omics profiling is increasingly applied across biological domains, including cancer biology (Gerstung *et al*, 2015; Iorio *et al*, 2016; Mertins *et al*, 2016; Cancer Genome Atlas Research Network, 2017), regulatory genomics (Chen *et al*, 2016), microbiology (Kim *et al*, 2016) or host-pathogen biology (Soderholm *et al*, 2016). Most recent technological advances have also enabled performing multi-omics analyses at the single-cell level (Macaulay *et al*, 2015; Angermueller *et al*, 2016; Guo *et al*, 2017; Clark *et al*, 2018; Colomé-Tatché & Theis, 2018). A common aim of such applications is to characterize heterogeneity between samples, as manifested in one or several of the data modalities (Ritchie *et al*, 2015). Multi-omics profiling is particularly appealing if the relevant axes of variation are not known *a priori*, and hence may be missed by studies that consider a single data modality or targeted approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus mapping, where large numbers of association tests are performed between individual genetic variants and gene expression levels (GTEx Consortium, 2015) or epigenetic marks (Chen *et al*, 2016). While eminently useful for variant annotation, such association studies are inherently *local* and do not provide a coherent global map of the molecular differences between samples. A second strategy is the use of kernel- or graph-based methods to combine different

Method | Open Access | Published: 11 May 2020

MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni & Oliver Stegle

Genome Biology **21**, Article number: 111 (2020) | Cite this article
30k Accesses | **127** Citations | **119** Altmetric | Metrics

Abstract

Technological advances have enabled the profiling of multiple molecular layers at single-cell resolution, assaying cells from multiple samples or conditions. Consequently, there is a growing need for computational strategies to analyze data from complex experimental designs that include multiple data modalities and multiple groups of samples. We present Multi-Omics Factor Analysis v2 (MOFA+), a statistical framework for the comprehensive and scalable integration of single-cell multi-modal data. MOFA+ reconstructs a low-dimensional representation of the data using computationally efficient variational inference and supports flexible sparsity constraints, allowing to jointly model variation across multiple sample groups and data modalities.

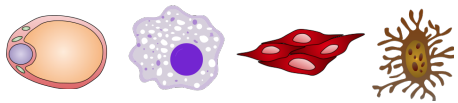
Background

Single-cell methods have provided unprecedented opportunities to assay cellular heterogeneity. This is particularly important for studying complex biological processes, including the immune system, embryonic development, and cancer [1,2,3,4].

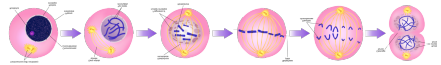
Download PDF

Sections | Figures | References

Factor Analysis: Common Axes of Variation



CELL TYPE



CELL CYCLE

L_1

L_2

P_1

N

P_2

N

P_3

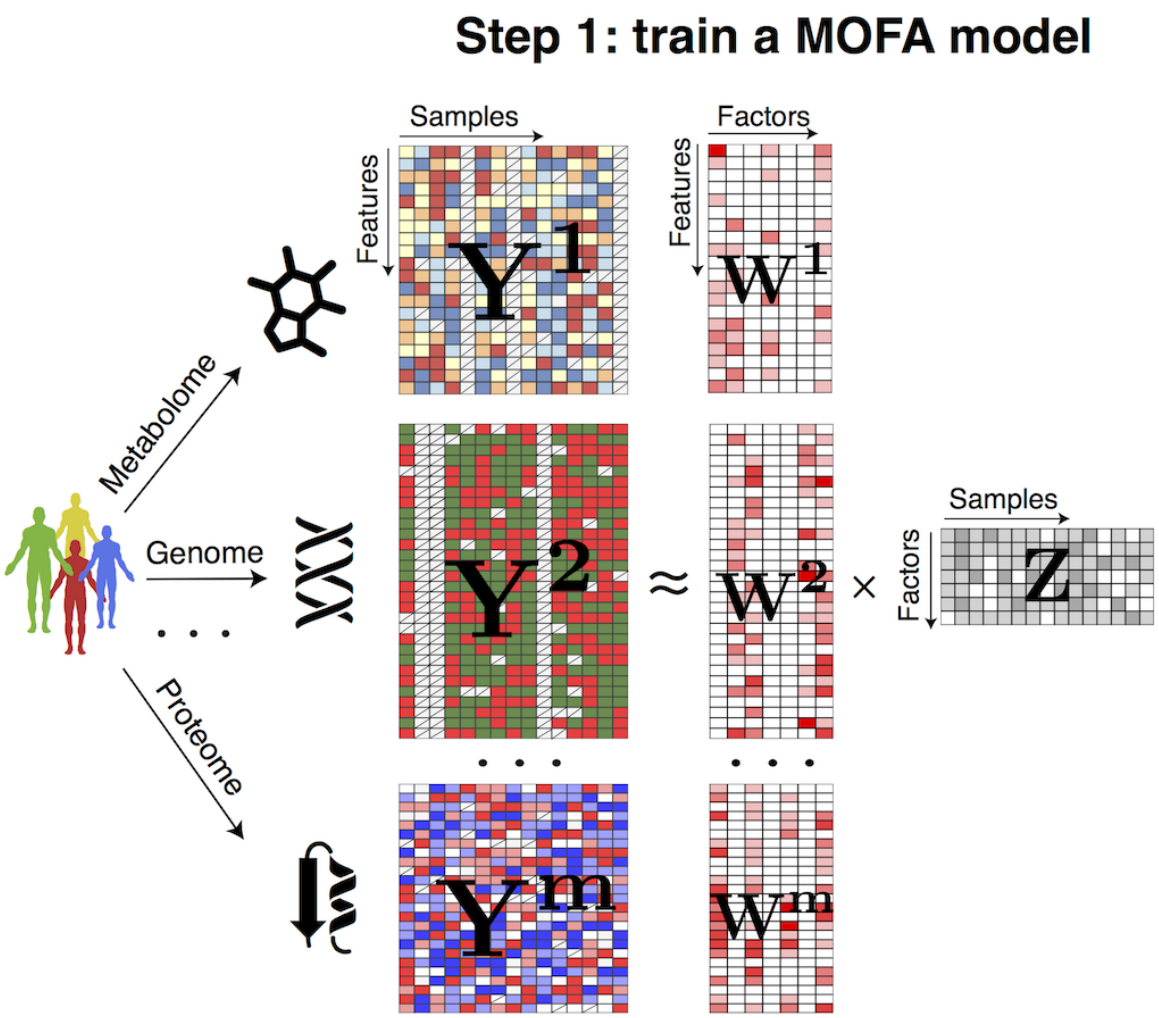
N

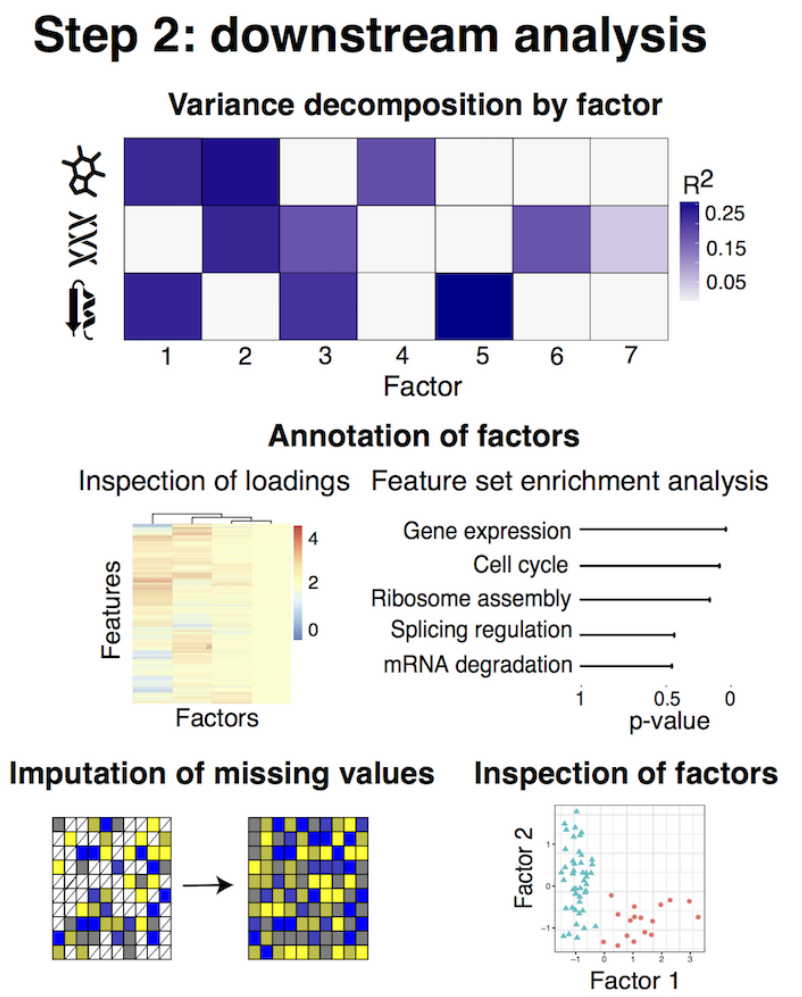
Gene expression

Methylation

Genetic mutation

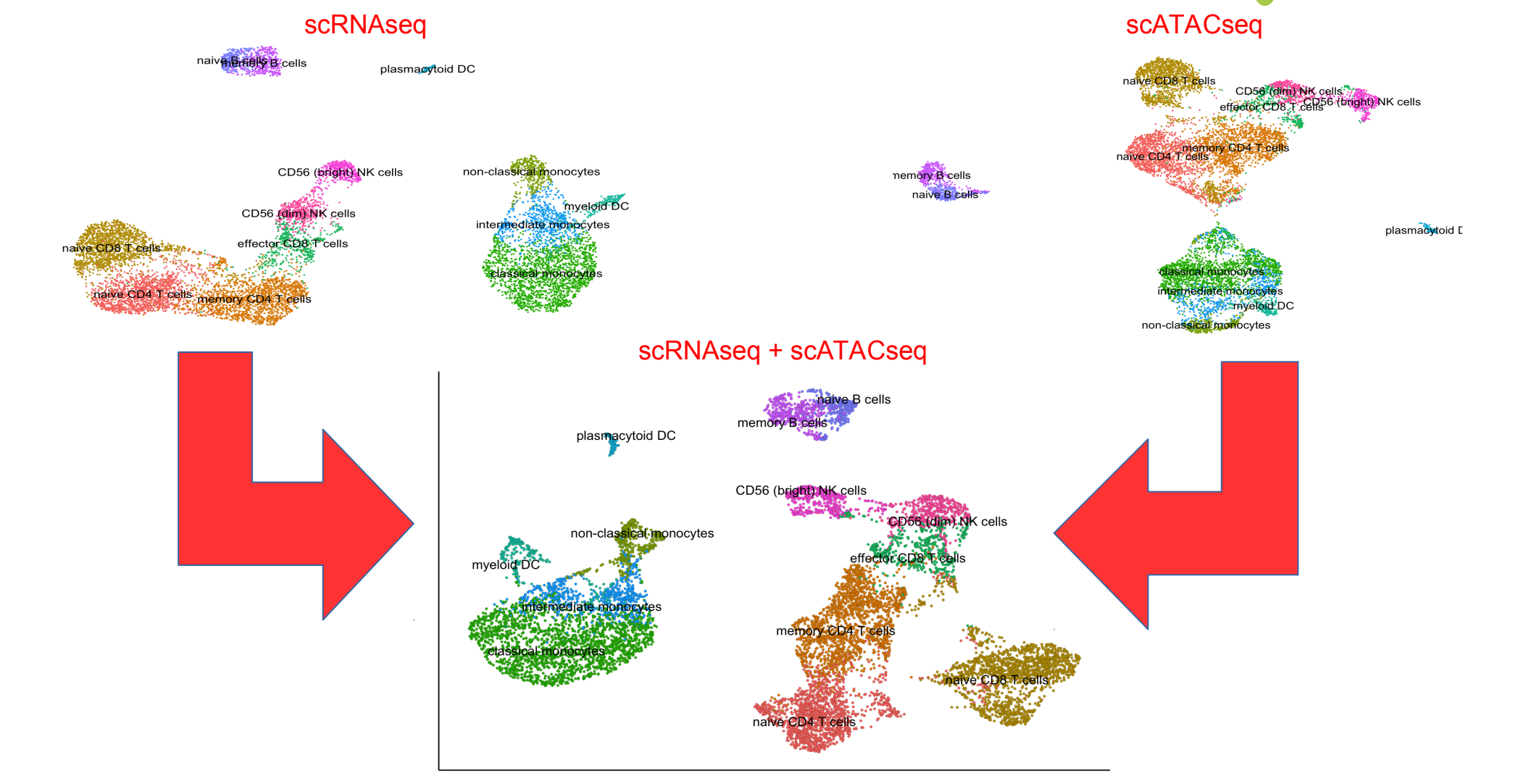
Multi-Omics Factor Analysis (MOFA)





R. Argelaguet et al., Mol. Syst. Biol. 2018

MOFA+ for scOmics Integration: 10X PBMC



MOFA application to scNMTseq

nature COMMUNICATIONS

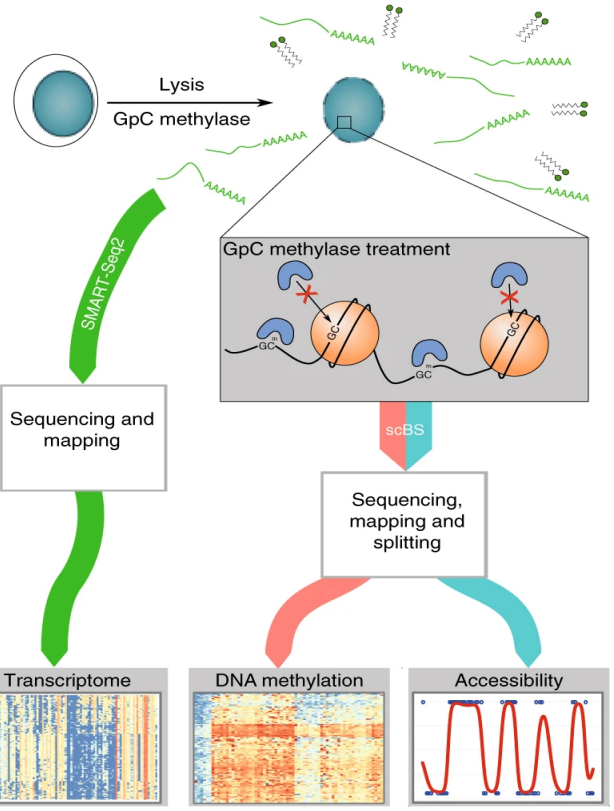
ARTICLE

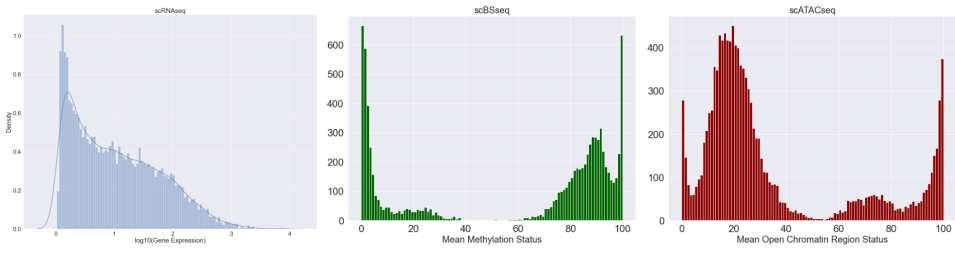
DOI: 10.1038/s41467-018-03149-4 OPEN

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

Stephen J. Clark[1], Ricard Argelaguet[2,3], ChantriolInt-Andreas Kapourani[4] Thomas M. Stubbs[1], Heather J. Lee[1,5,6], Celia Alda-Catalinas[1], Felix Krueger[7] Guido Sanguinetti[4], Gavin Kelsey[1,8] John C. Marioni[2,3,5] Oliver Stegle[2] Wolf Reik[1,5,8]

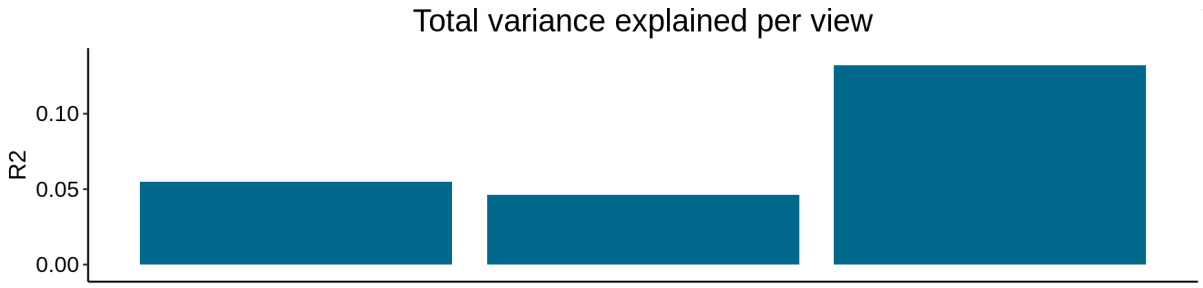
Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) uses a GpC methyltransferase to label open chromatin followed by bisulfite and RNA sequencing. We validate scNMT-seq by applying it to differentiating mouse embryonic stem cells, finding links between all three molecular layers and revealing dynamic coupling between epigenomic layers during differentiation.

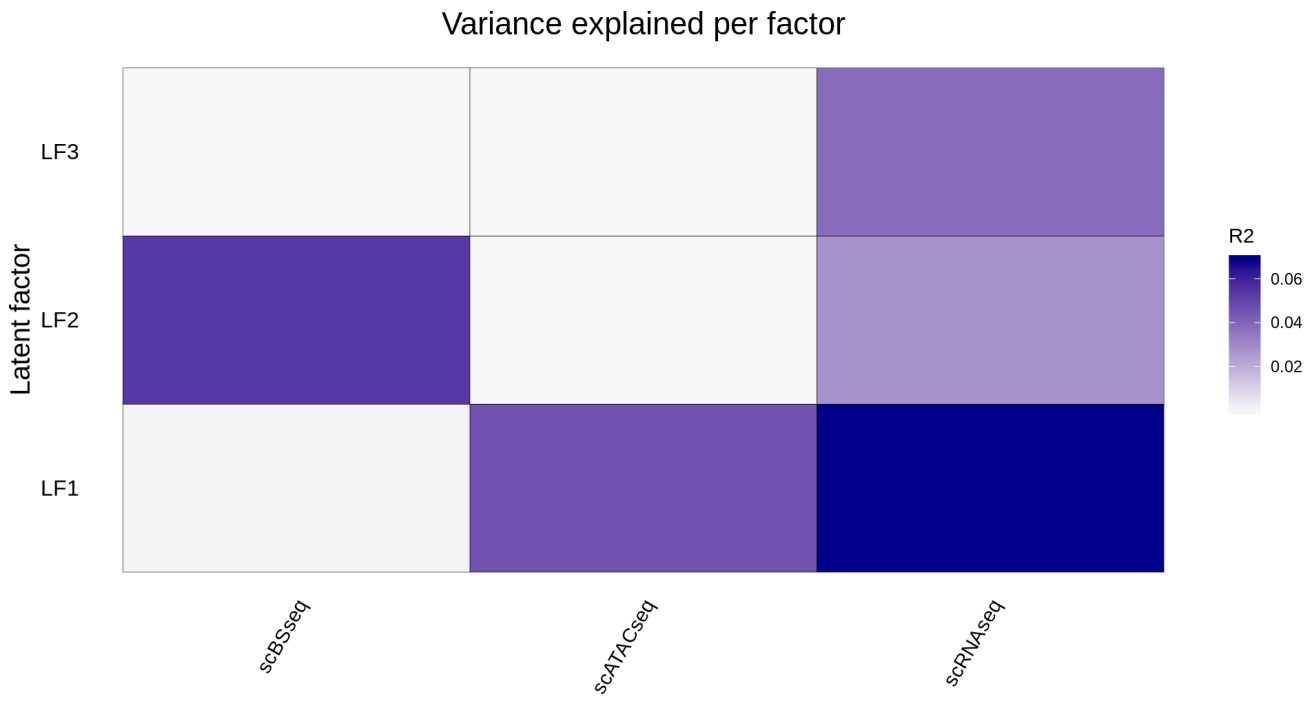


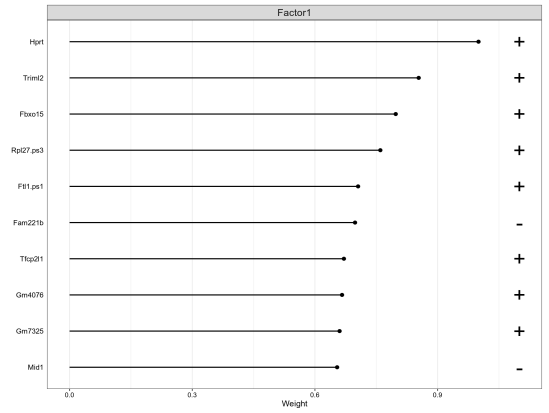


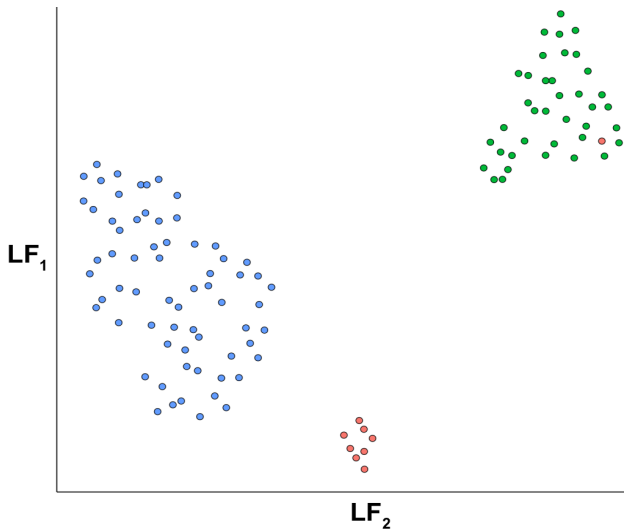
MOFA: Omics contributions to Latent Factors

$$R^2_{m,k} = 1 - \left(\sum_{n,d} y^m_{nd} - z_{nk} w^m_{kd} - \mu^m_d\right)^2 \Big/ \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$

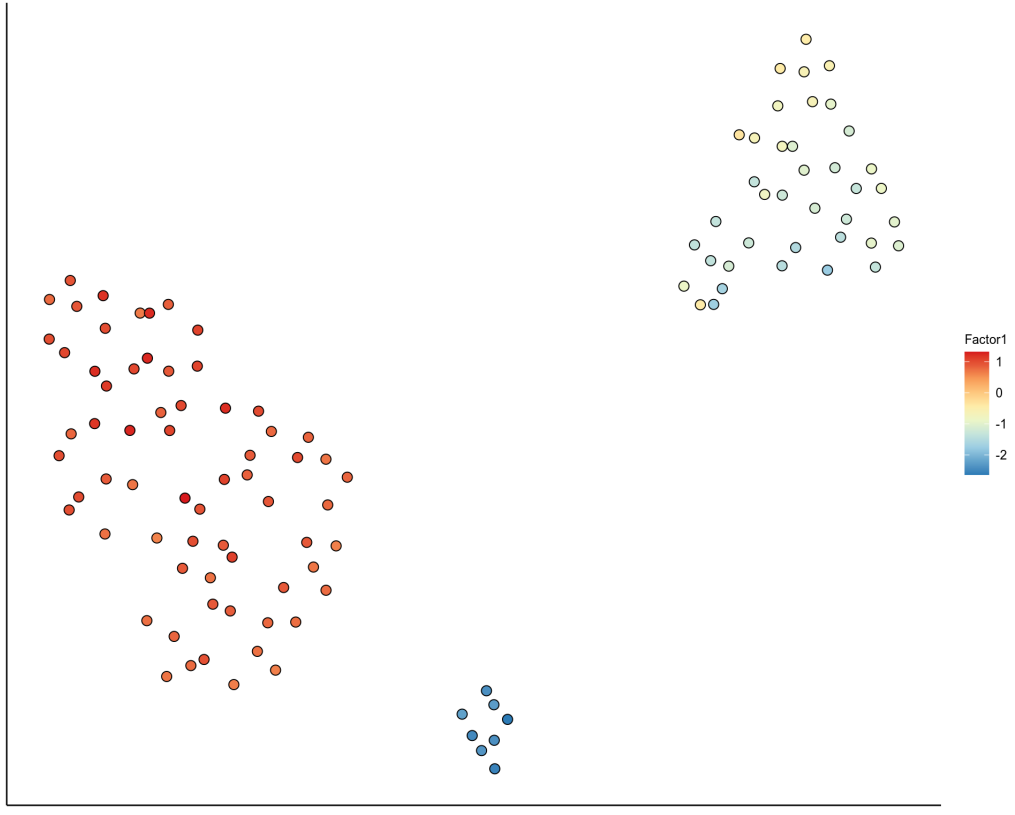


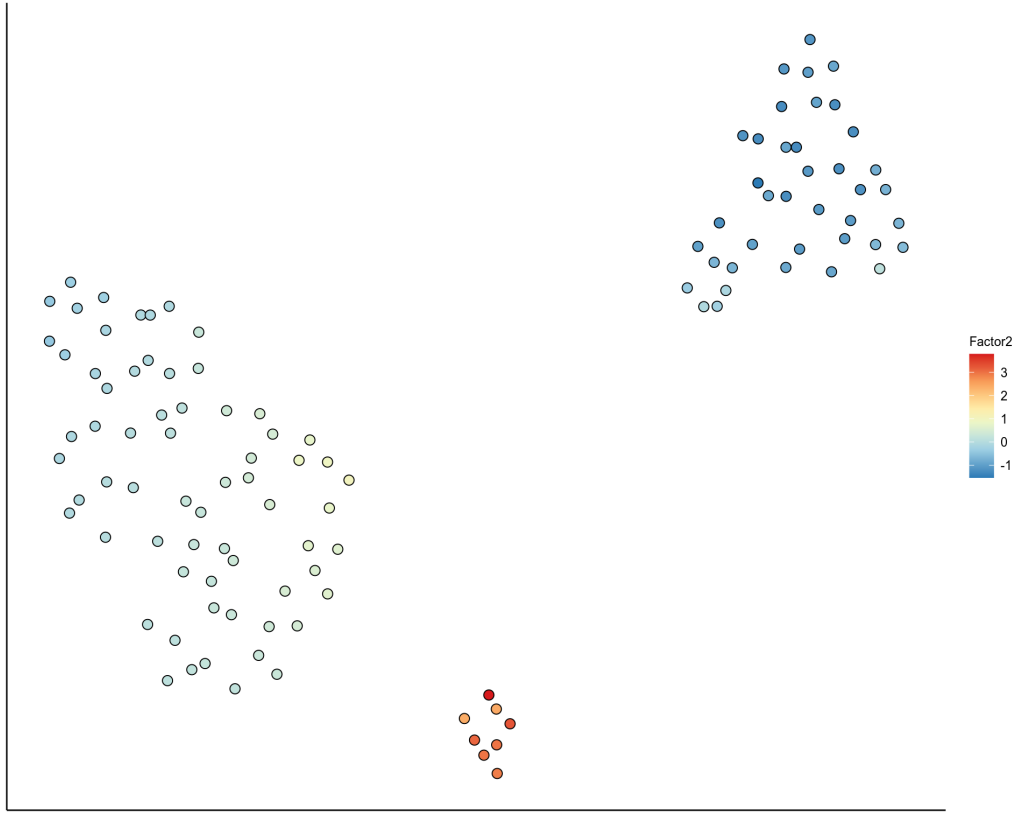






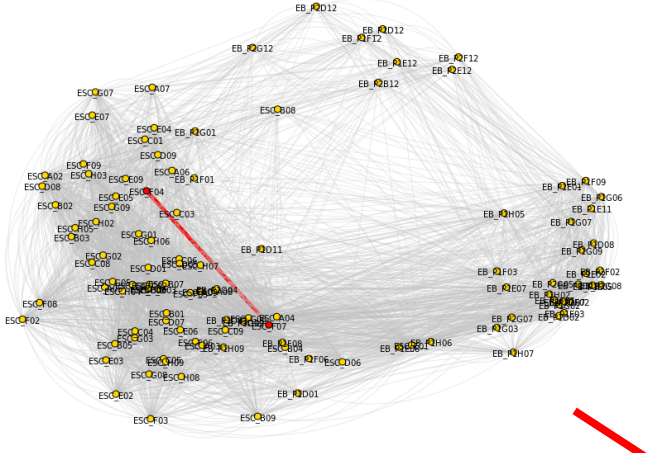
MOFA: Omics contributions to tSNE clusters





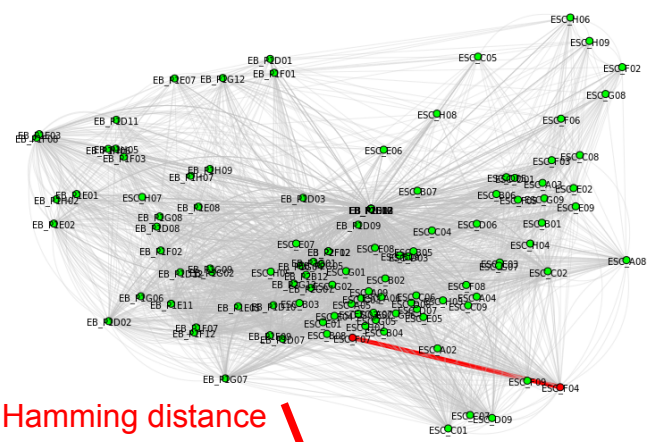
Graph Intersection Method

scRNAseq KNN Graph



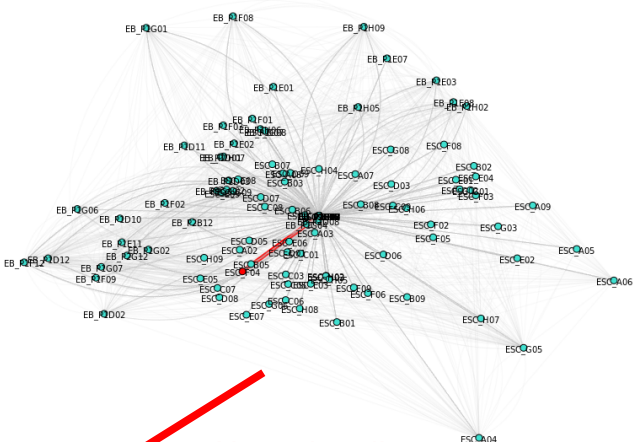
Euclidean distance

scBSseq KNN Graph



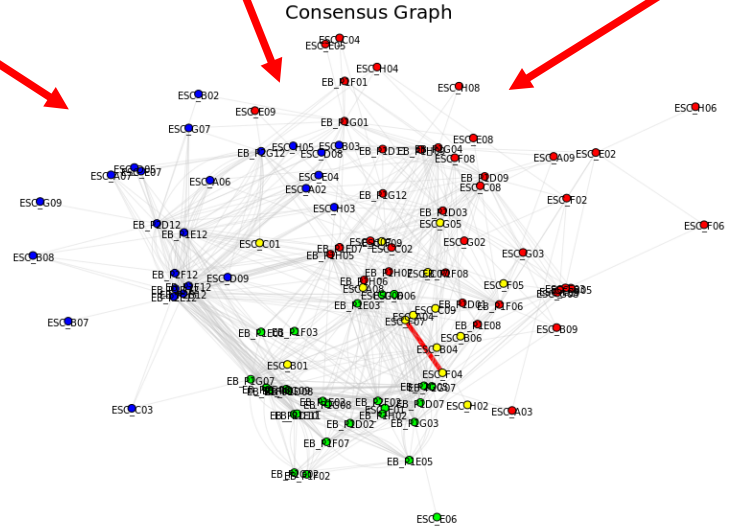
Hamming distance

scATACseq KNN Graph



Hamming distance

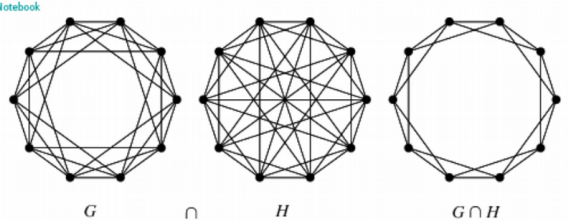
Consensus Graph



Keep edges consistently present across the Omics

Graph Intersection

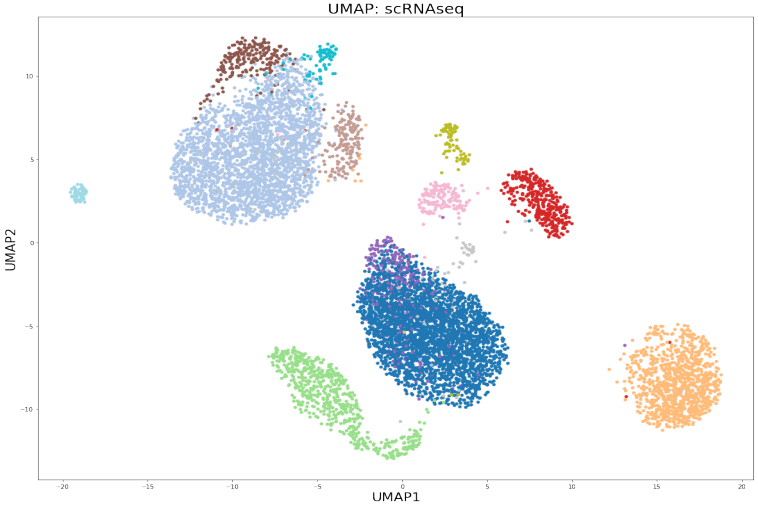
DOWNLOAD Wolfram Notebook

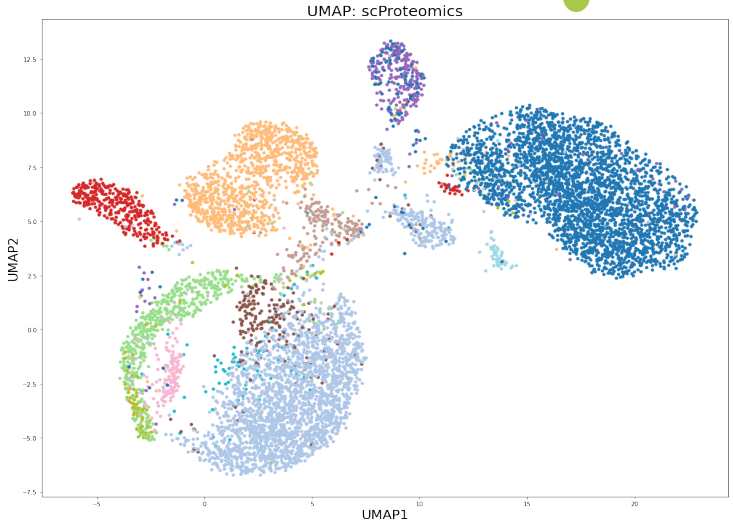


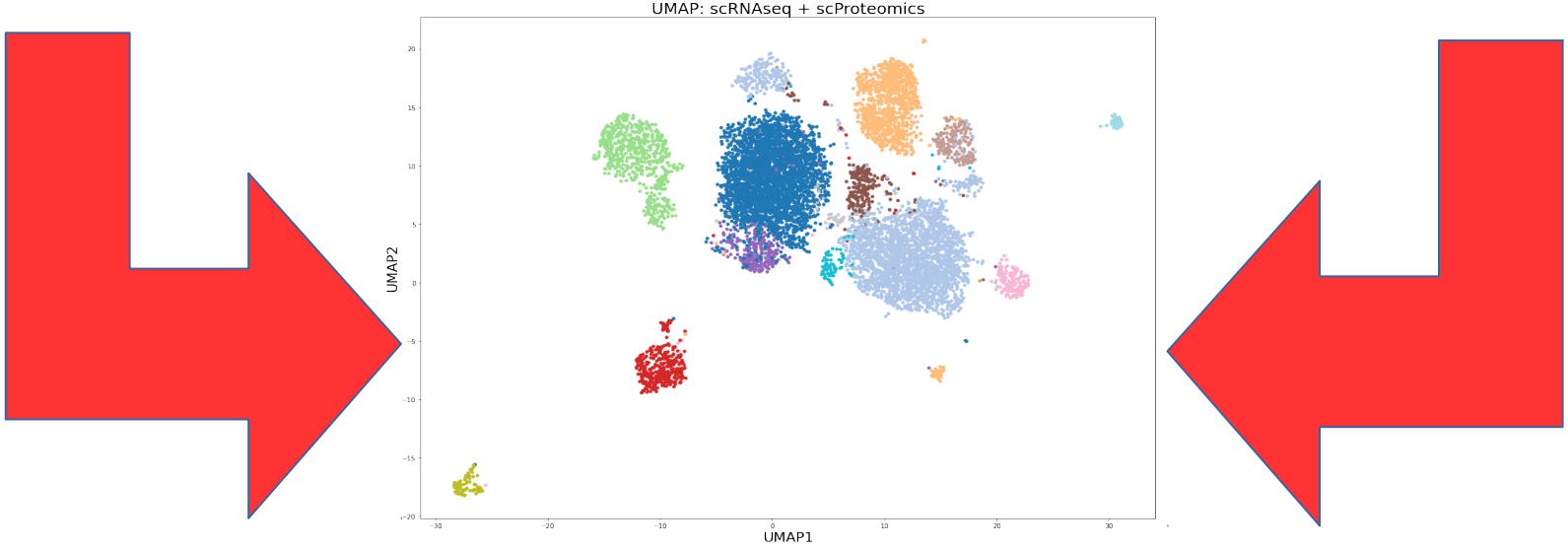
G ∩ H $G \cap H$

Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

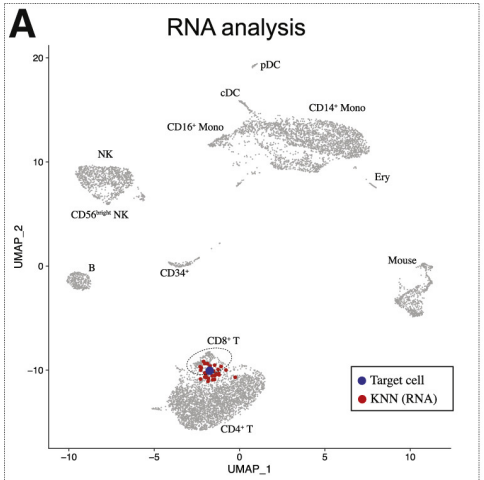
UMAP for data integration: CITEseq

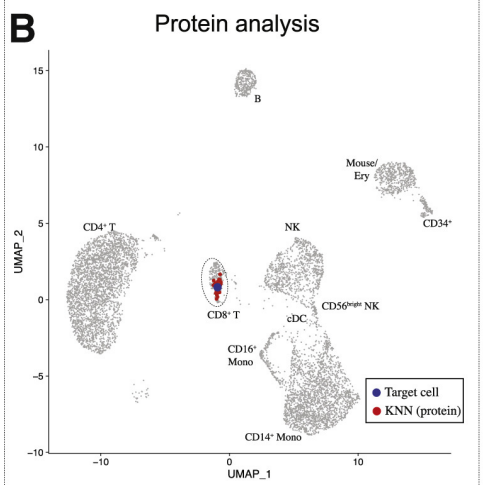






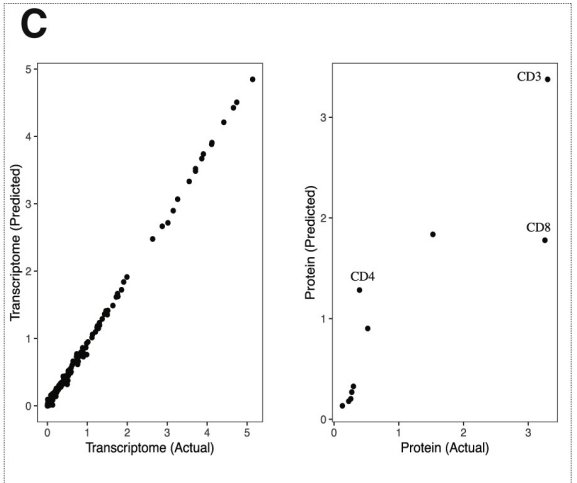
Seurat Weighted Nearest Neigbors (WNN)

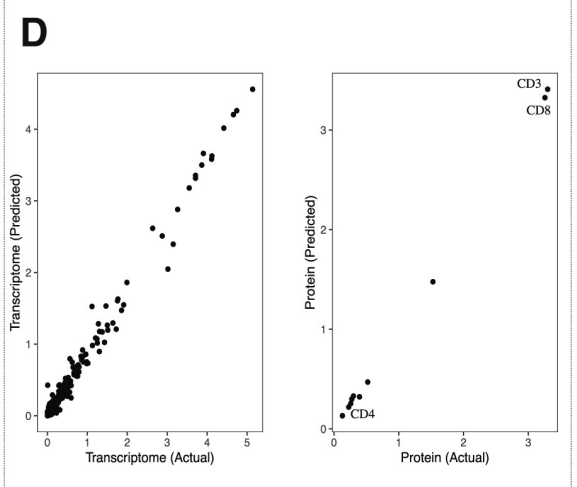


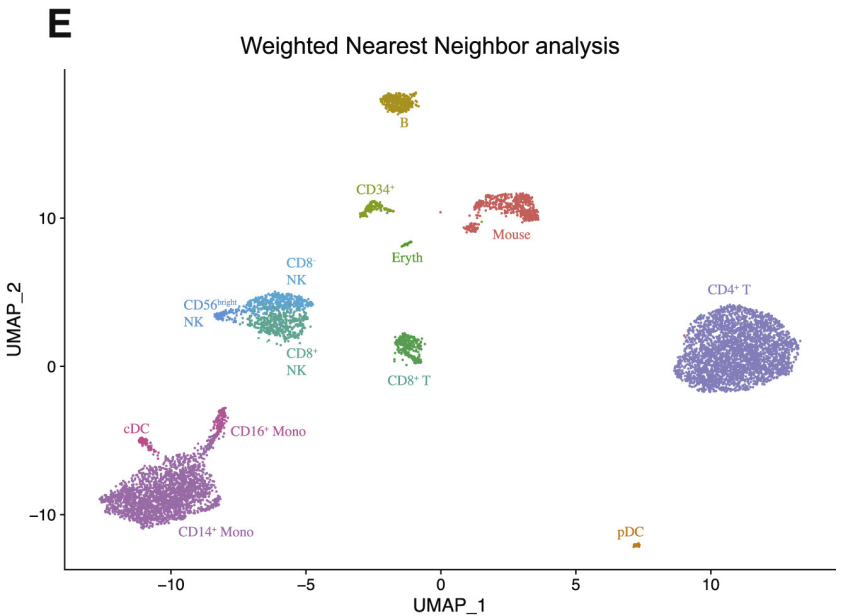


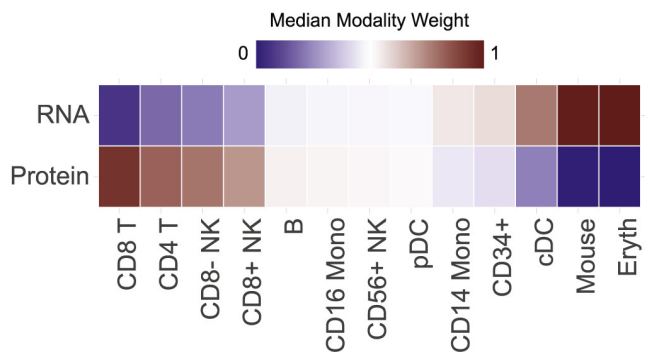
Predict cell from RNA neighbors

Predict cell from protein neighbors



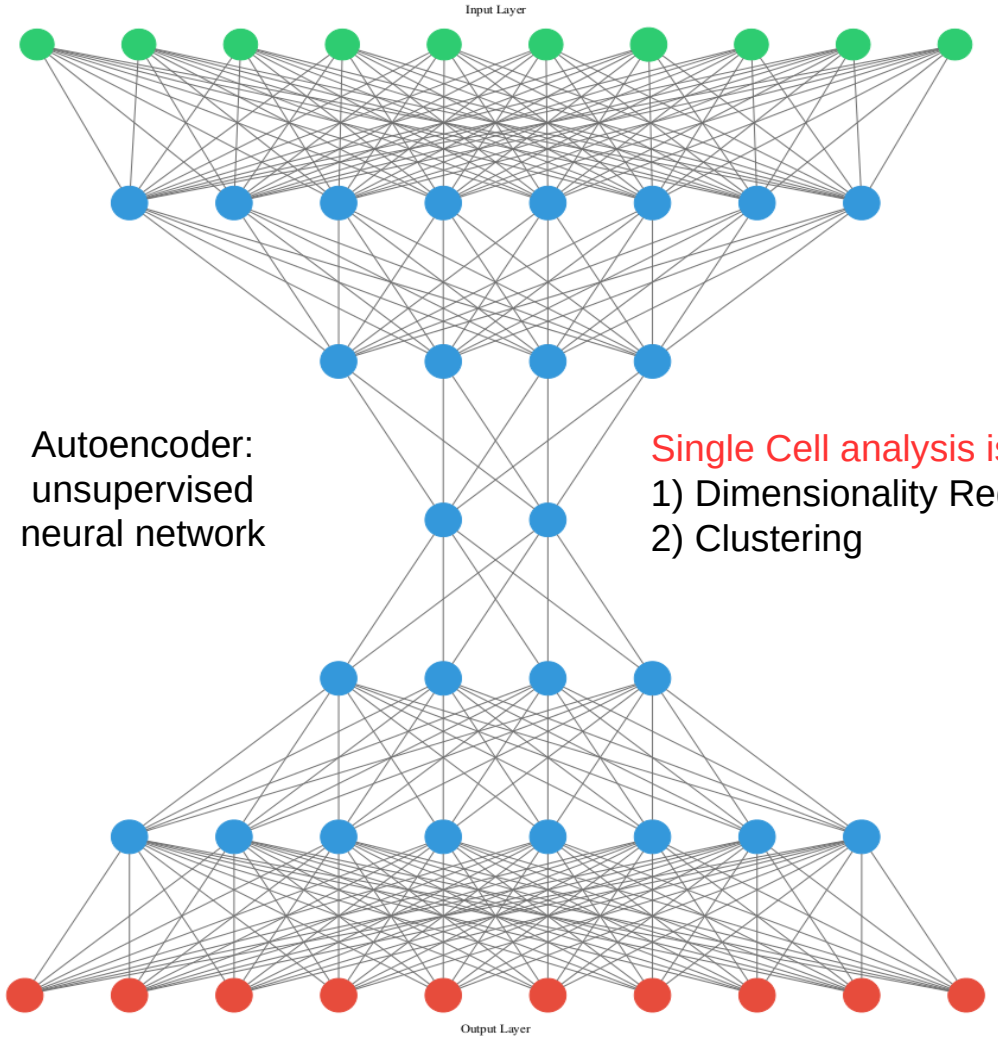






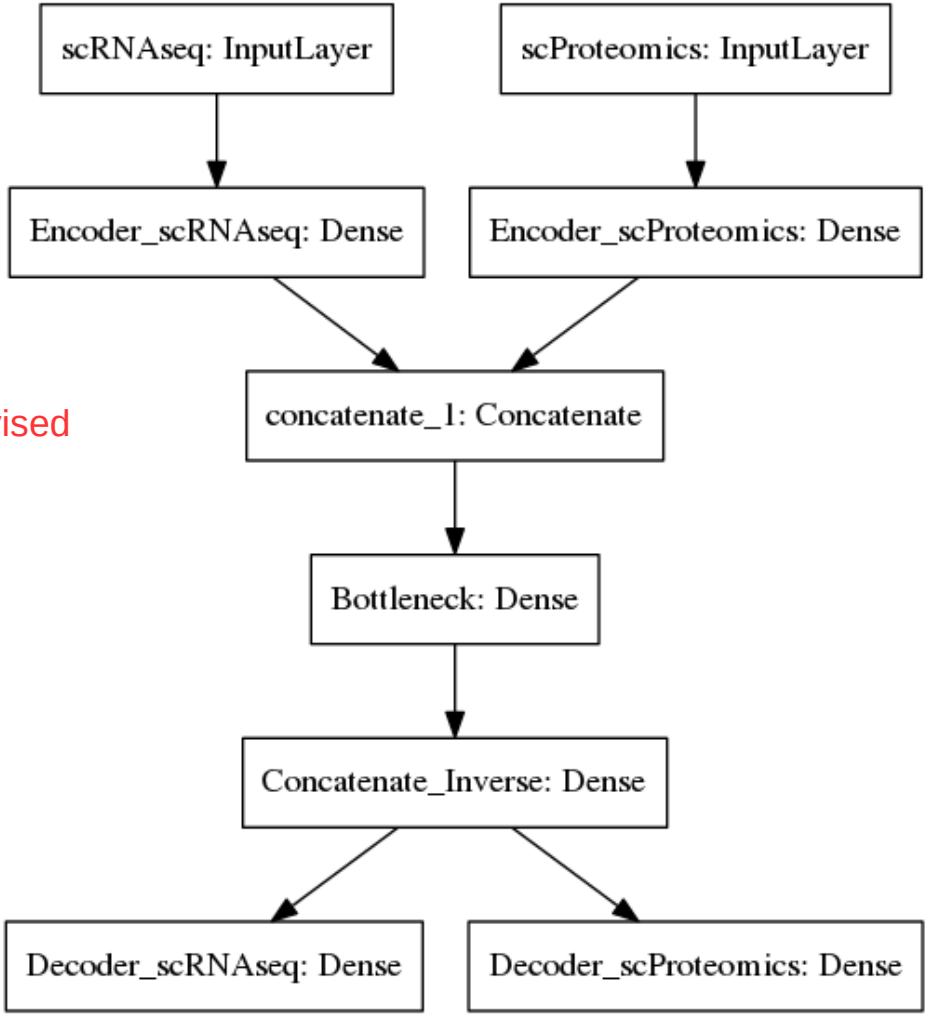
Hao et al. Cell 2021, 184(13): 3573-3587.e29. doi: 10.1016/j.cell.2021.04.048

Autoencoder for data integration

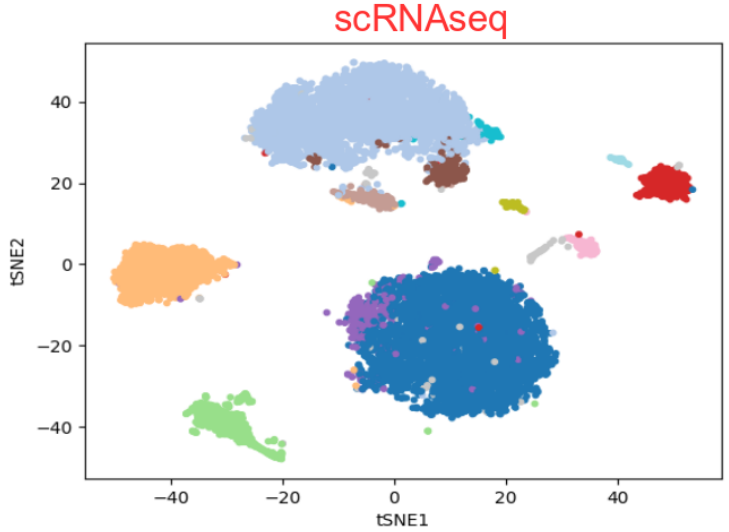


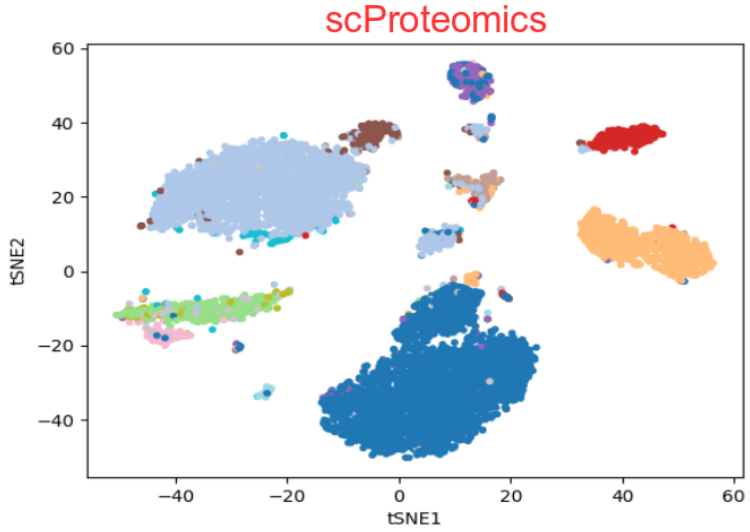
Autoencoder:
unsupervised
neural network

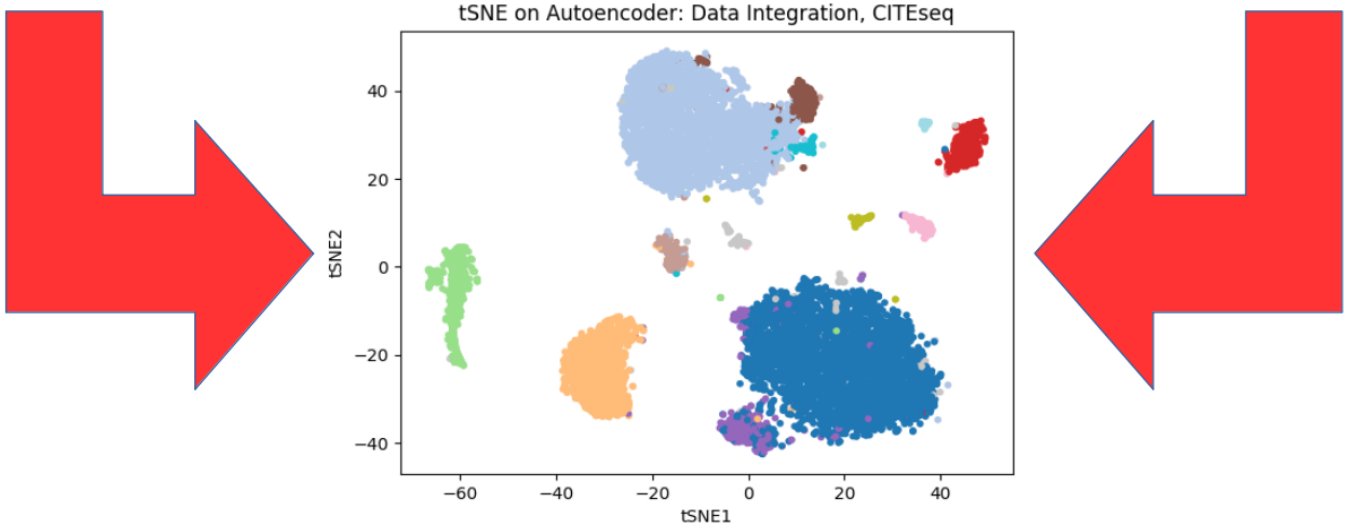
Single Cell analysis is unsupervised
1) Dimensionality Reduction
2) Clustering



Autoencoder for data integration: CITEseq







National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet