Data integration and batch correction in single-cell data

Paulo Czarnewski

ELIXIR Single-Cell Omics community co-lead

National Bioinformatics Infrastructure Sweden (ELIXIR-SE)

Sources of variation in single-cell

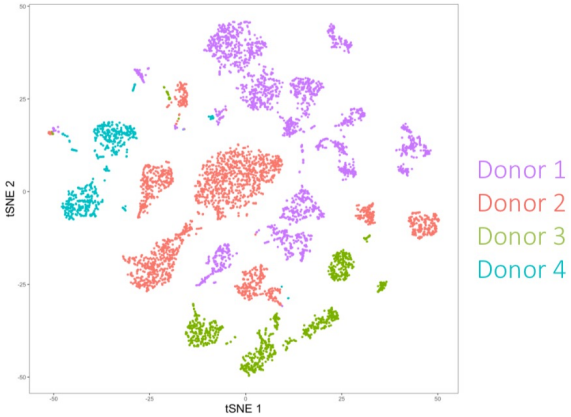
Biological:
- Cell Type Heterogeneity
- Genetics
- Cell State/Microenvir.
- GExpr Stochasticity
- Cell Cycle Dynamics
- Transcriptional Bursts
- Oscillations

...

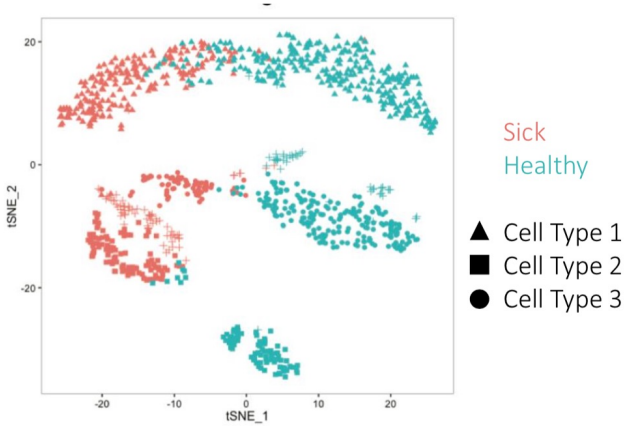
Technical:
- Capture Efficiency
- Amplification Bias
- PCR artifacts
- Contamination
- Cell Doublets
- Cell Damage
- Sampling (Jackpot Effects)

...

Sources of variation in single-cell: Donors

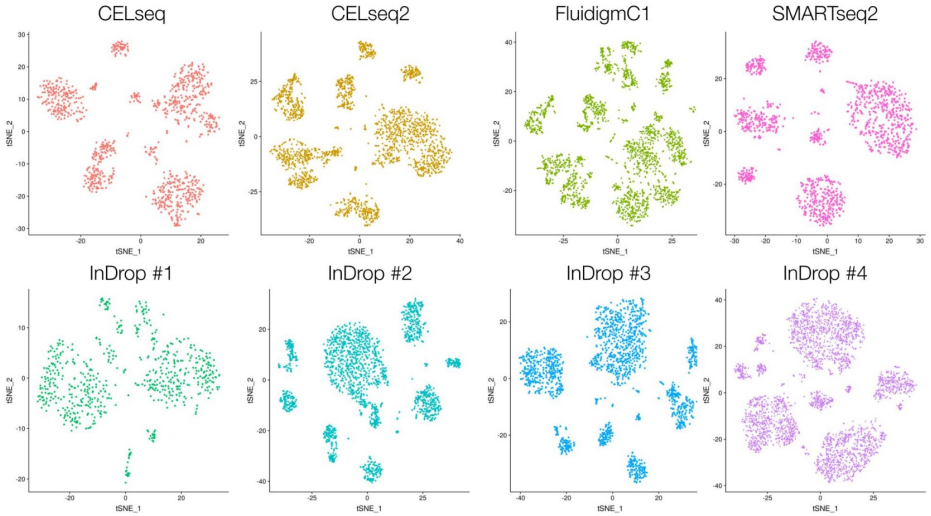


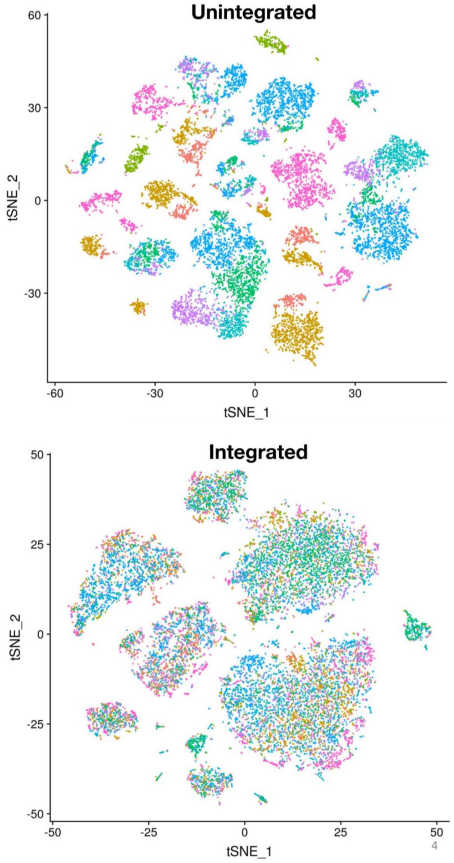
Same tissue from different donors



Cross condition comparisons

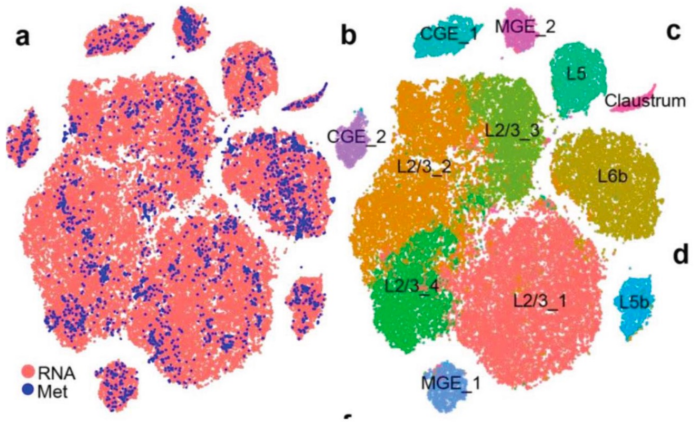
Sources of variation in single-cell: Technology



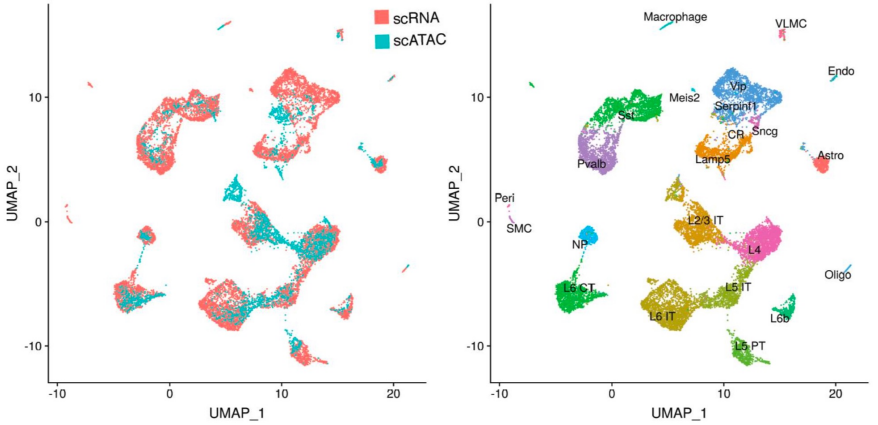


Sources of variation in single-cell: OMICs

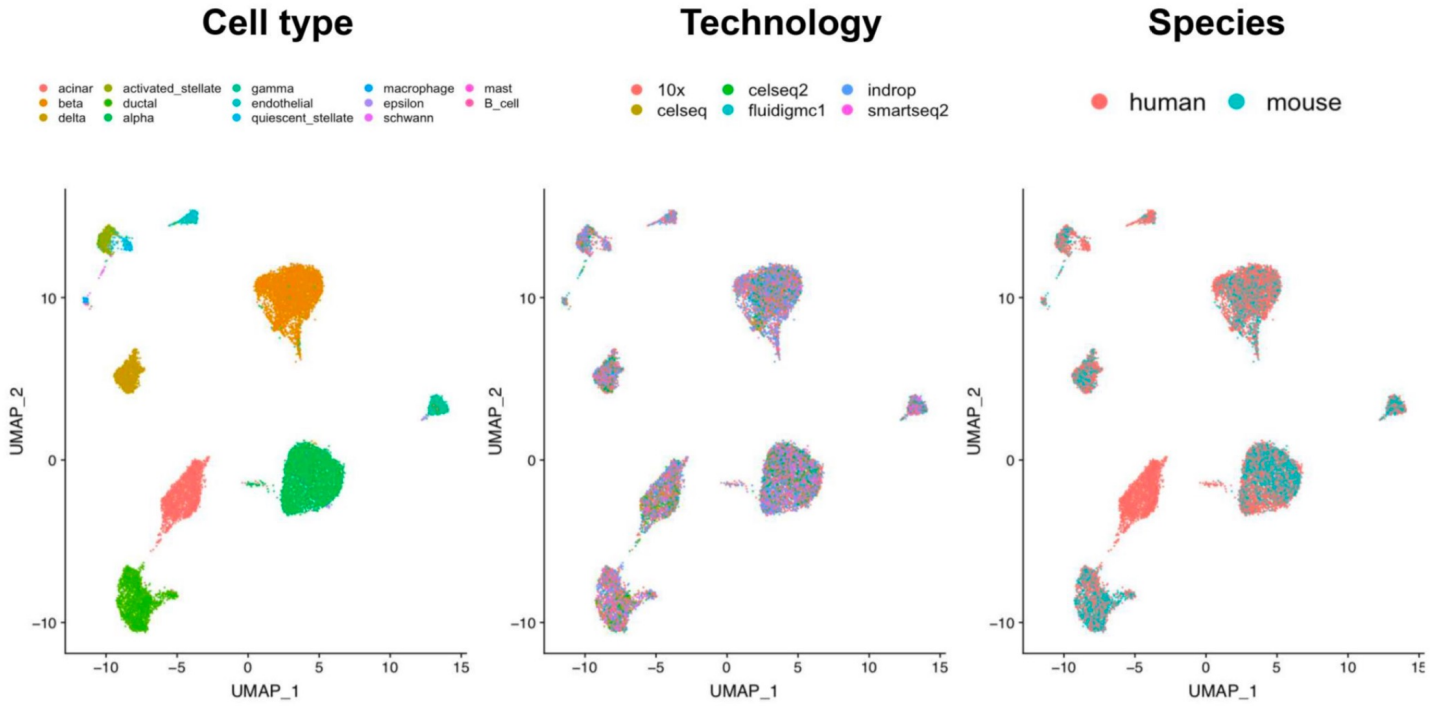
snRNAseq + snNMTseq



scRNAseq + scATACseq



Sources of variation in single-cell: Species

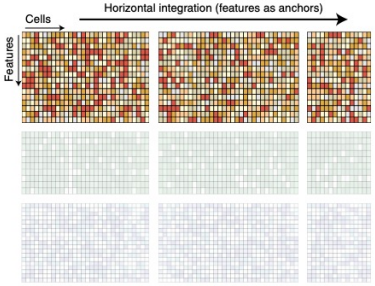


Retinal bipolar datasets: 51K cells, 6 technologies, 2 Species

Types of data integration

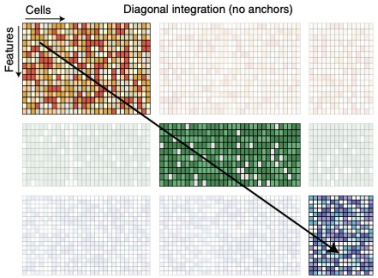
Horizontal

- Shared features
- Different samples



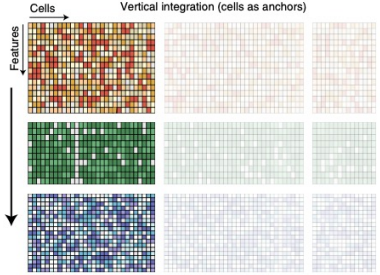
Diagonal

- Different features
- Different samples



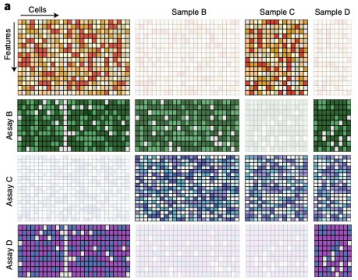
Vertical

- Different features
- Shared samples



Mosaic

- Some features shared
- Some samples shared



Argelaguet et al (2021) Nature Biotechnology

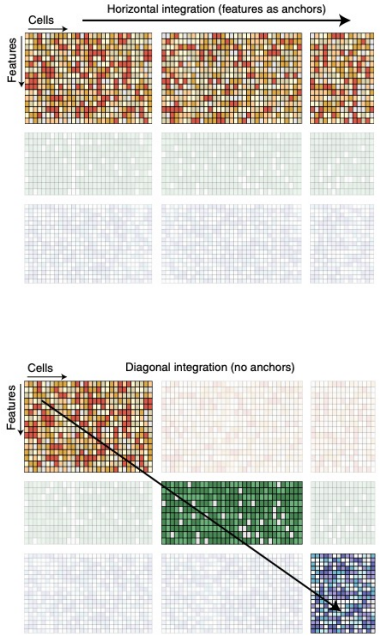
Types of data integration

Table 1 | Overview of common data integration methods classified according to their anchor choice

Integration task	Method	Ref.
Vertical (global)	CCA	112
Vertical (global)	JIVE	70
Vertical (global)	PLS	71
Vertical (global)	MCIA	113
Vertical (global)	MOFA+	65
Vertical (global)	scAI	114
Vertical (global)	iNMF	38
Vertical (global)	Seurat v4	11
Vertical (local)	Spearman's rank correlation coefficient	50
Vertical (local)	LMM	51
Horizontal	MNN	21
Horizontal	Seurat v3	22
Horizontal	LIGER	23
Horizontal	Harmony	24
Horizontal	Scanorama	29
Horizontal	BBKNN	25
Horizontal	scVI	26
Horizontal	scmap	28
Horizontal	conos	27
Diagonal	MATCHER	77
Diagonal	MMD-MMA	78
Diagonal	SCIM	115
Diagonal	UnionCom	116
Diagonal	coupledNMF	117

Argelaguet et al (2021) Nature Biotechnology

Horizontal and Diagonal integration

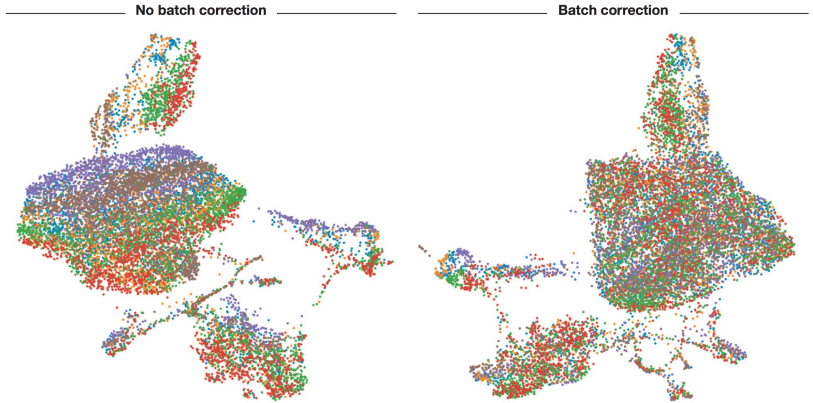


ComBat

Uses empirical Bayes regression on shared gene factors

Works well on simpler small-medium datasets

All datasets need to be similar in cell type composition



- limma::removeBatchEffect()
- seurat::ScaleData()
- sva::combat()
- batchelor::rescaleBatches()

Known issues with ComBat:

1. Do not account for differences in population composition
2. Assume batch effect is additive
3. Prone to overcorrection (in cases of partial confounding)
4. Will fail in large datasets with complex mixture of cell type

Luecken, Theis (2019) Molecular Systems biology

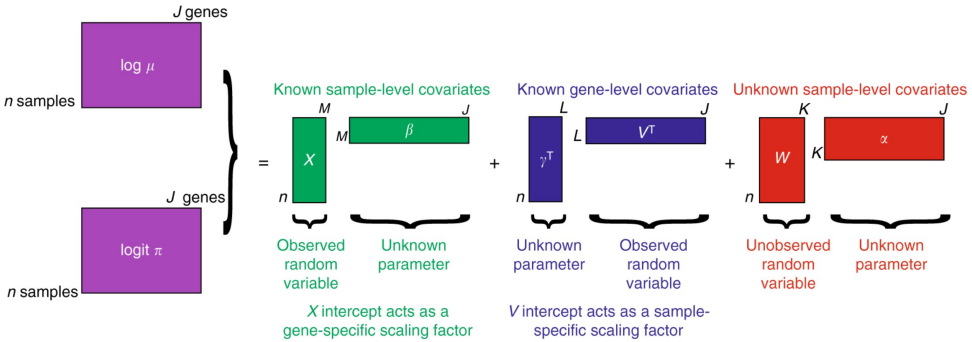
ZIMB-WaVE (factor analysis)

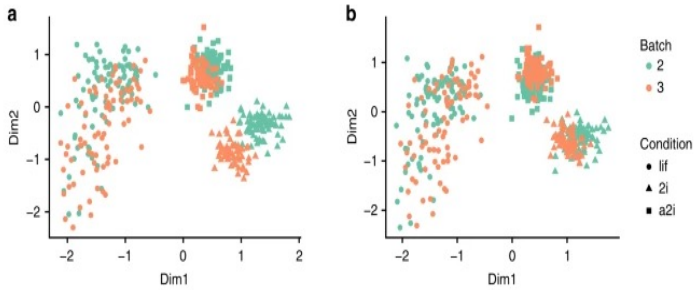
Zero-Inflated Negative Binomial-based Wanted Variation Extraction

Can accommodate both gene- and cell-level covariates

Known issues with ZIMB-WaVE:

1. Do not account for differences in population composition
2. Assume batch effect is additive
3. It will be slow on large datasets
4. Will fail in large datasets with complex mixture of cell type





Risso et al (2018) Nat Communicaiton

Mutual Nearest Neighbours (MNN)

Dimensionality reduction via multibatch PCA with all datasets

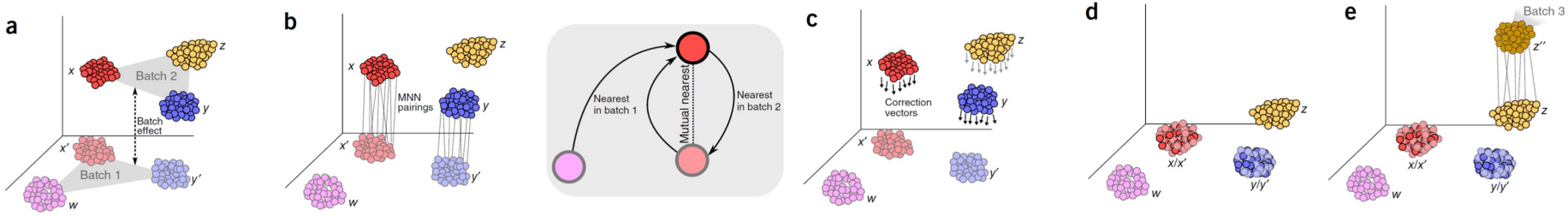
Find kNN within datasets and kMNN across datasets

Compute merging vectors for each cell

It scales well on large datasets

Known issues with MNN:

1. There is at least one cell population that is present in both batches
2. The batch effect is almost orthogonal to the biological subspace
3. Batch effect variation is much smaller than the biological effect variation between different cell types



Haghverdi et al (2018) Nat Biotechnol

Scanorama

Python

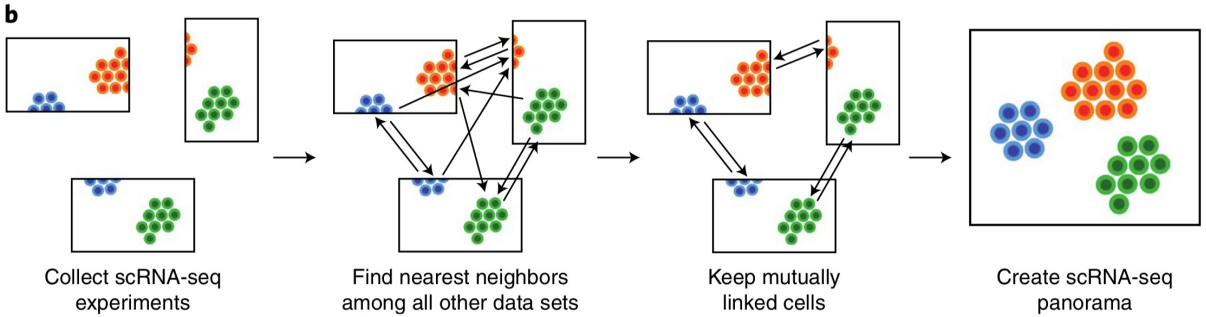
Dimensionality reduction via SVD with all datasets

Find K-NN within and across datasets

Compute merging vectors

Known issues with Scanorama:

1. There is at least one cell population that is present in both batches
2. The batch effect is almost orthogonal to the biological subspace
3. Batch effect variation is much smaller than the biological effect variation between different cell types



Hie et al (2019) Nat Biotechnol

Harmony

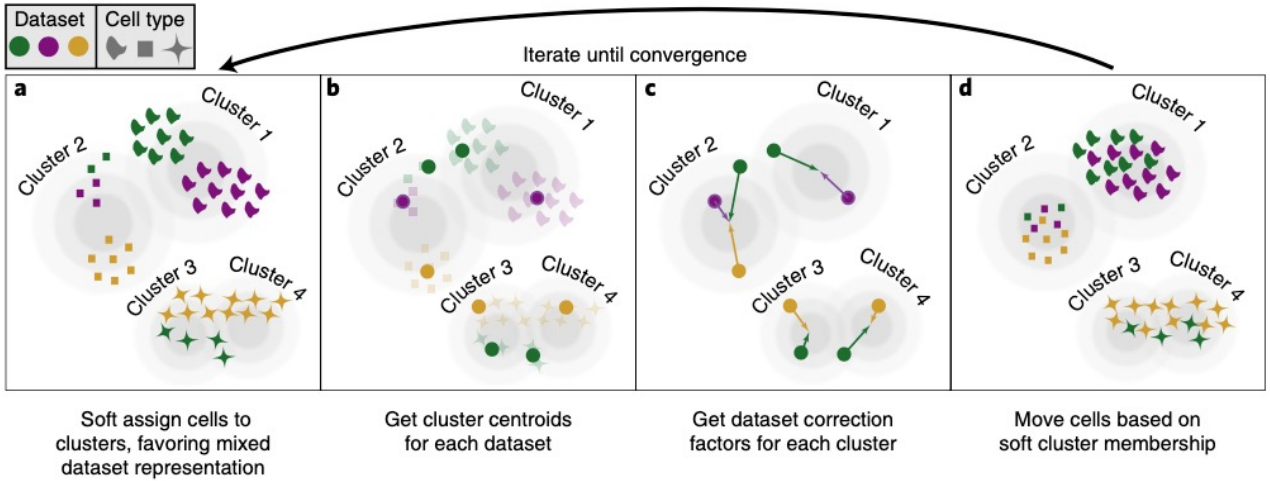
Employs fuzzy k-means clustering to speed-up computations

Find K-NN to cluster centroids across datasets

Compute merging vectors for each cell

Known issues with Harmony:

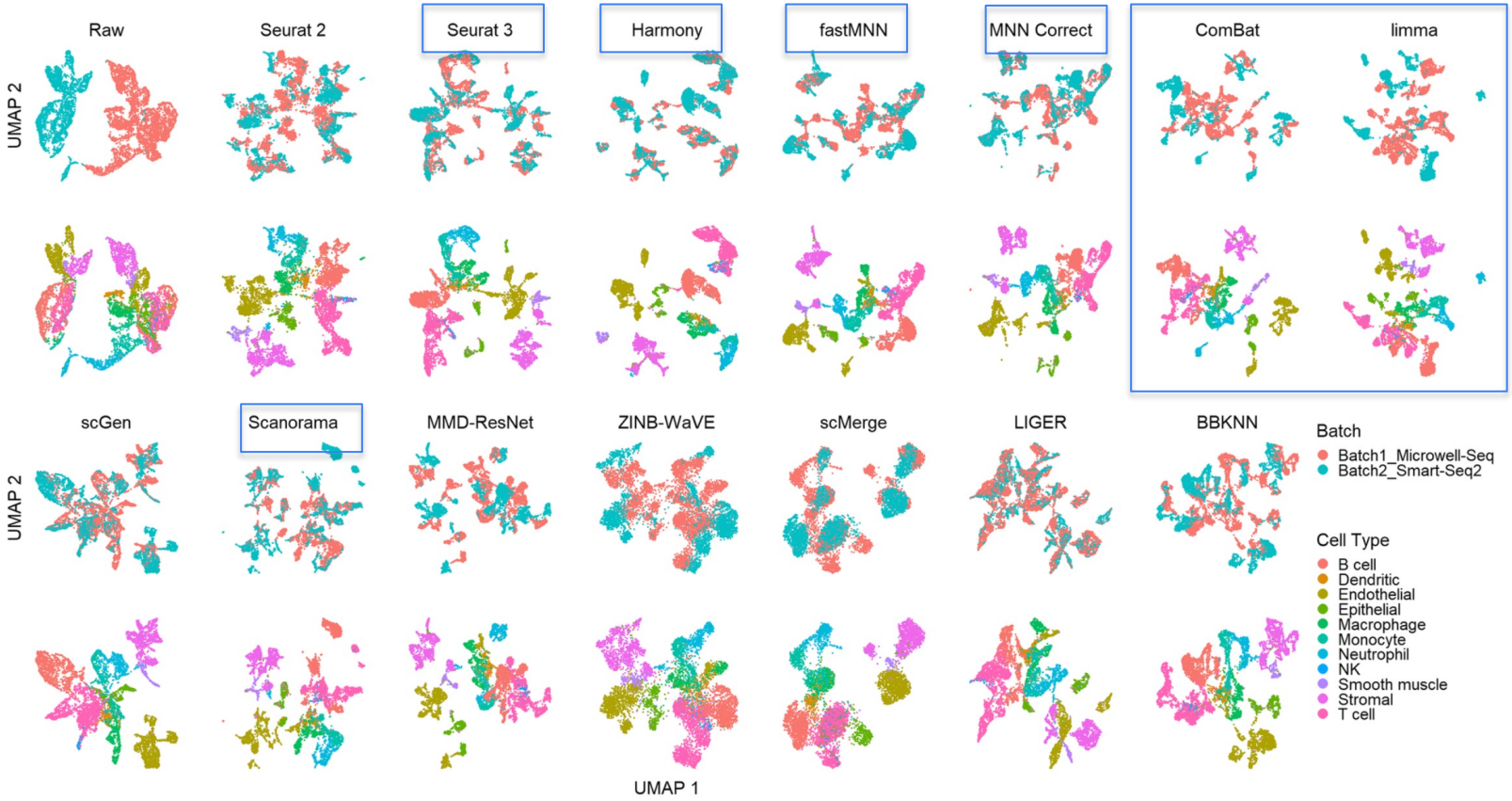
1. Need to find optimal number of clusters for your dataset
2. Need to define convergence criteria / thresholds



Hie et al (2019) Nat Biotechnol

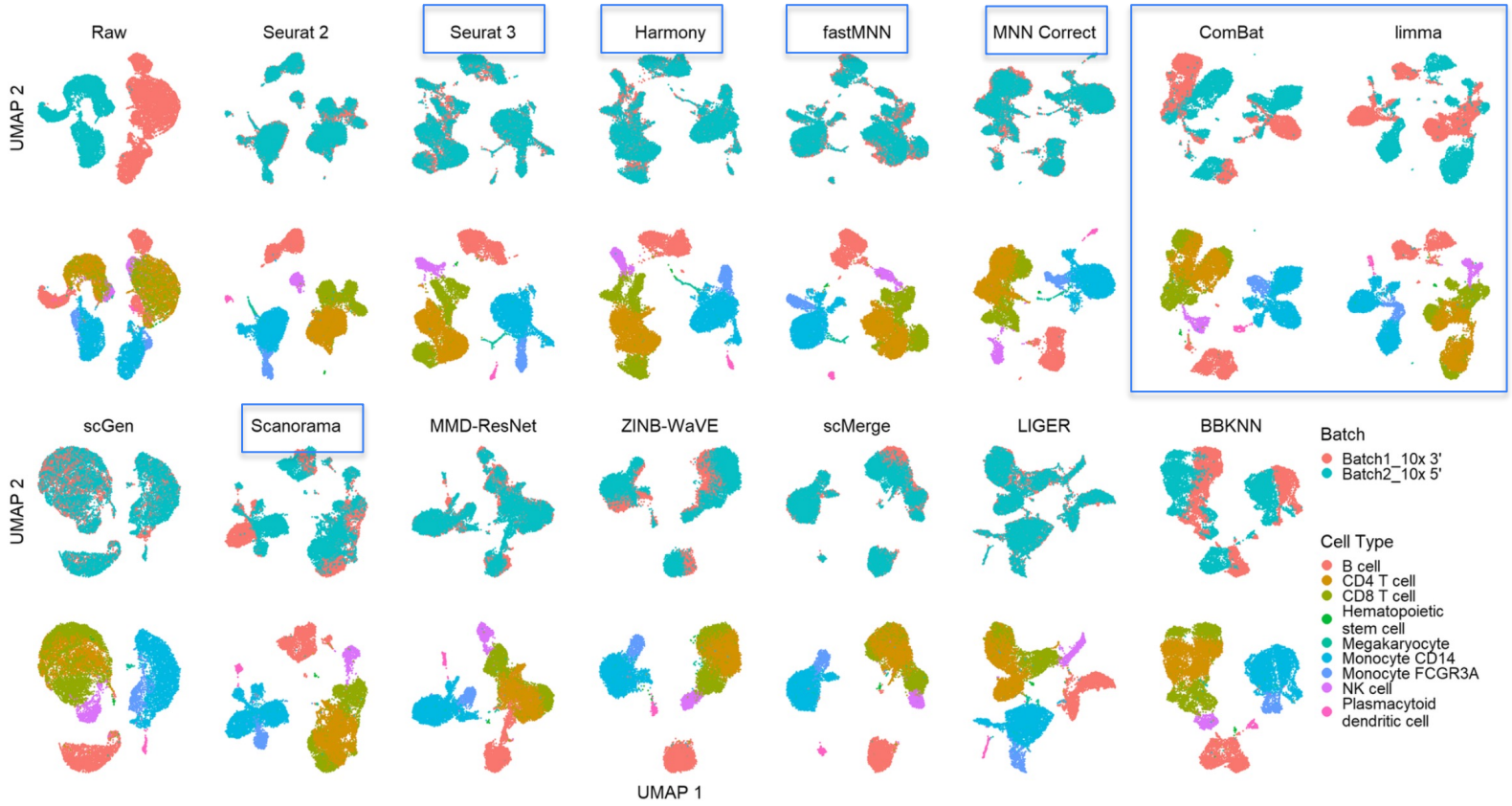
Integration performance assessment

Integration performance assessment



Tran et al (2020) Genome Biology

Integration performance assessment



Tran et al (2020) Genome Biology

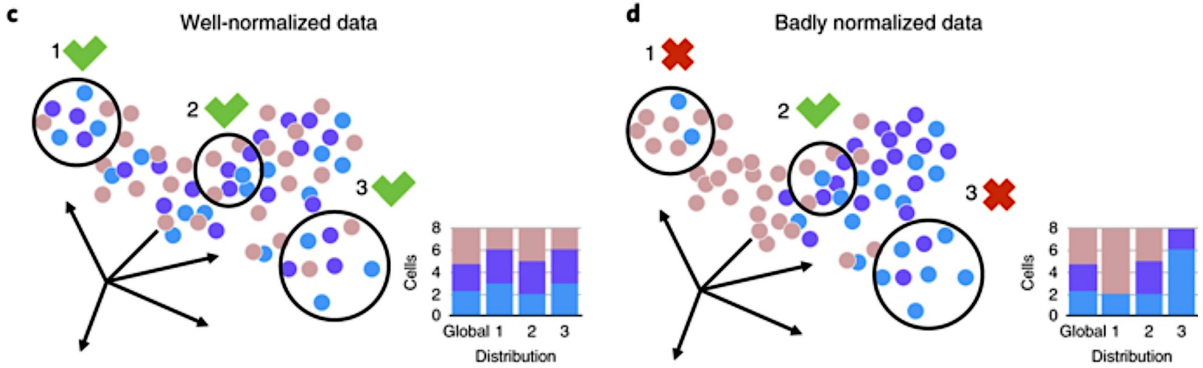
Integration performance assessment

Goal:

- The batch-originating variance is erased
- Meaningful heterogeneity is preserved
- No artefactual variance is introduced

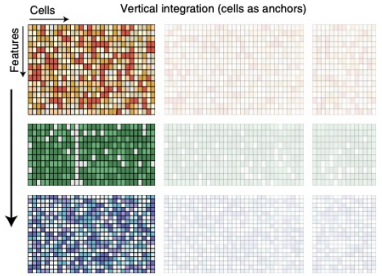
What it practically means:

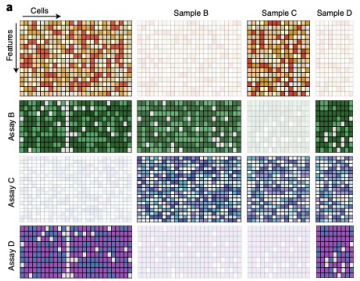
- Similar cell types are intermixed across batches
- We are not mixing distinct cell types (across or within batches)
- We do not separate similar cells within batches



Büttner et al (2020) Nat Methods

Vertical and Mosaic integration





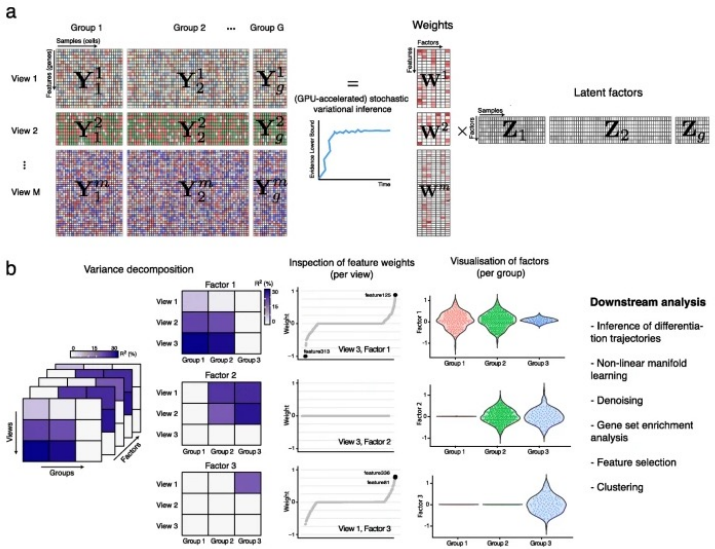
Factor analysis via NNMF

Accurate, flexible and robust methods

Known issues with NNMF for single-cell:

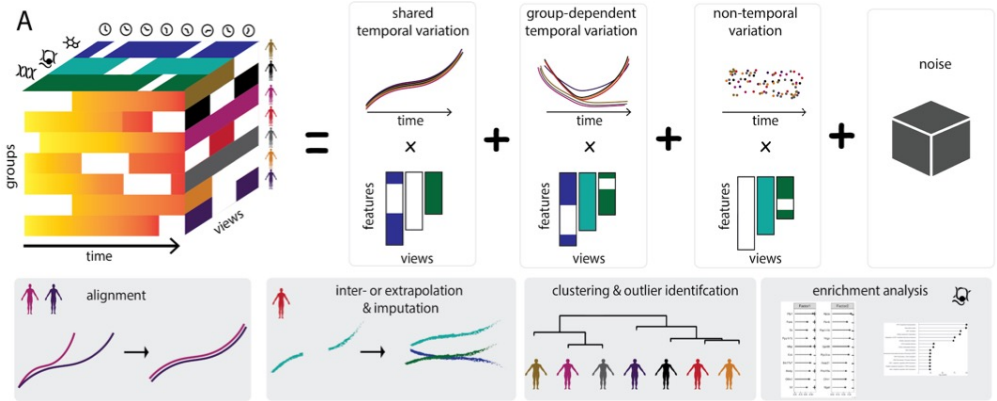
1. Will fail in large datasets with complex mixture of cell type

MOFA+



Argelaguet et al (2020) Genome Biology

MEPHISTO



Velten et al (2020) BioRxiv

Weighted Nearest Neighbours (WNN)

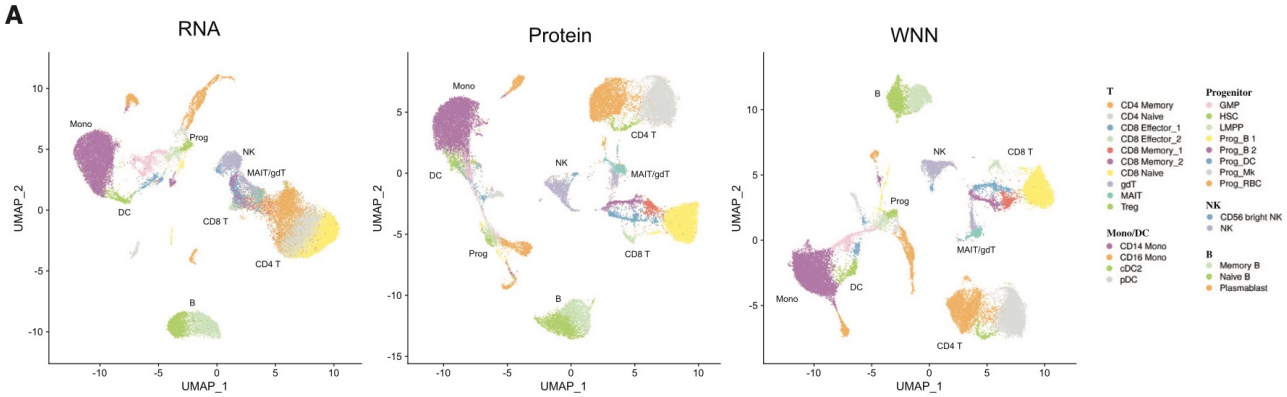
Computes KNN graphs within modalities

Graphs are then weighted and merged (just like in similarity network fusion: SNF)

Can use different similarities for each OMIC

Known issues with WNN:

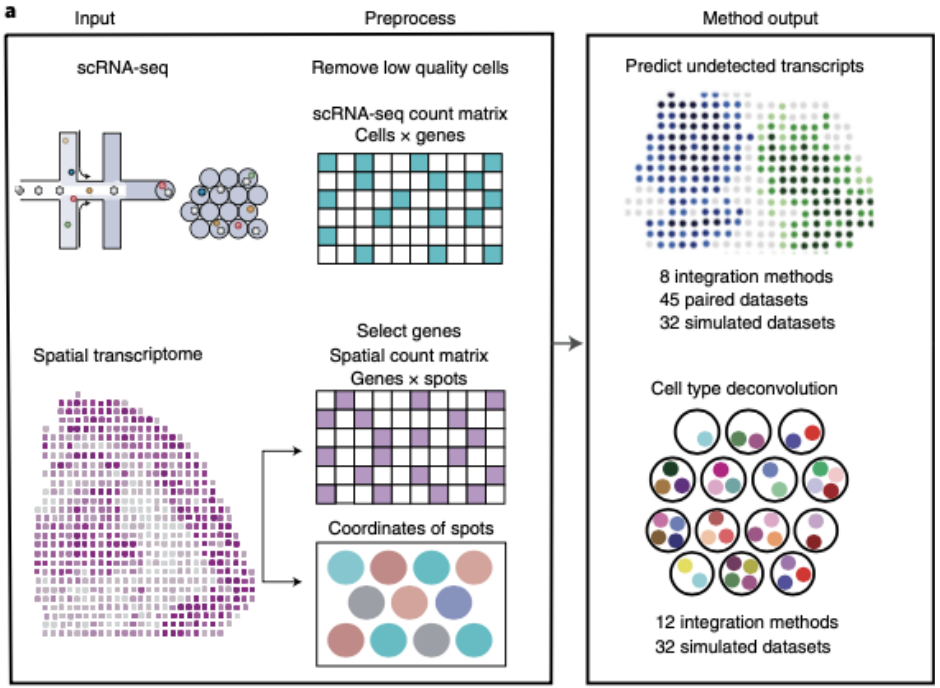
1. Graph weights are somewhat arbitrary



Hie et al (2019) Nat Biotechnol

Data integration between single-cell and spatial data

Types of integration tasks

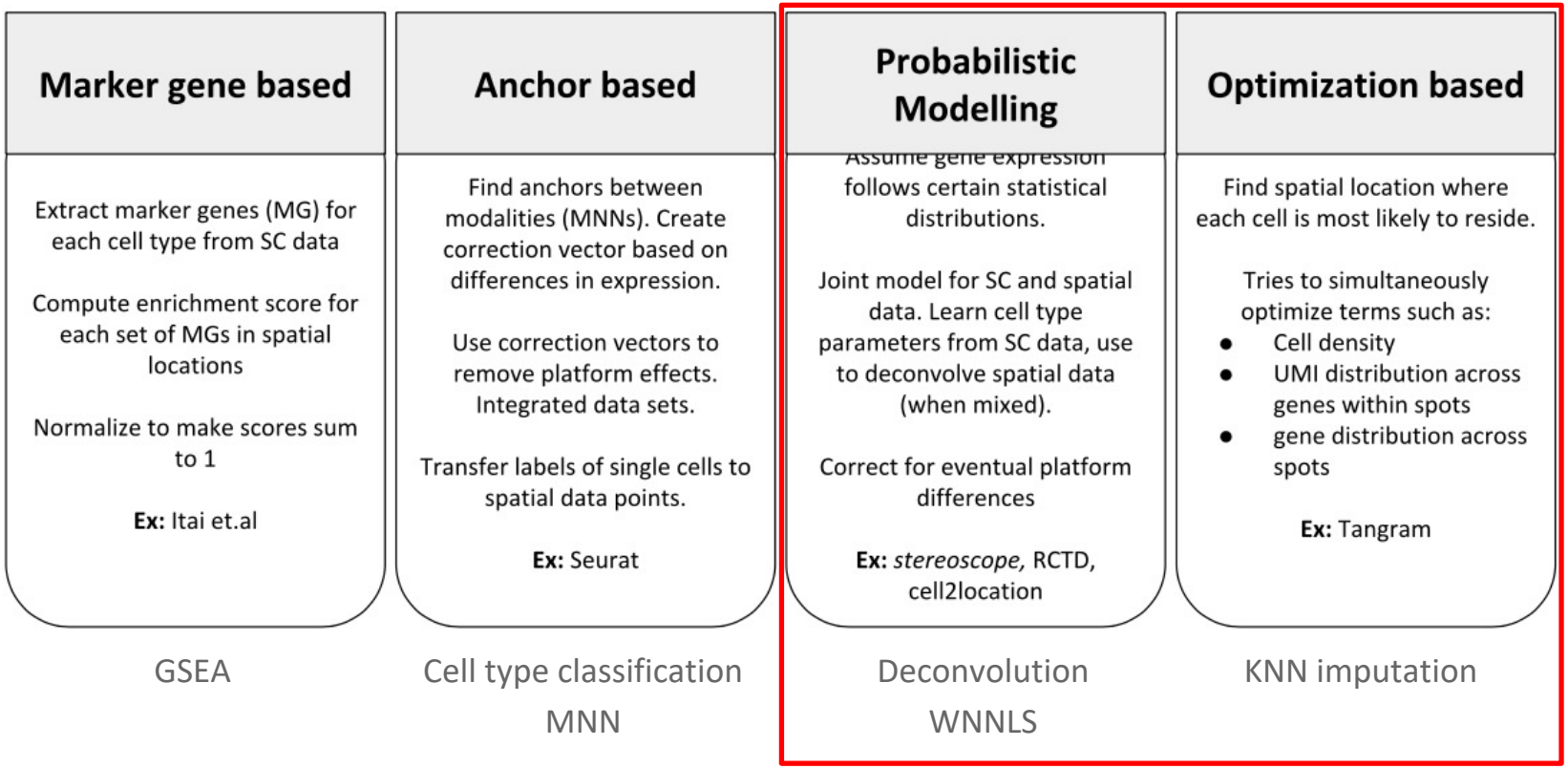


Gene **imputation**

Cell type **deconvolution**

Li et al (2022) Nat Methods

Types of integration tasks



Hie et al (2019) Nat Biotechnol

spaGE: imputation of genes using single-cell data

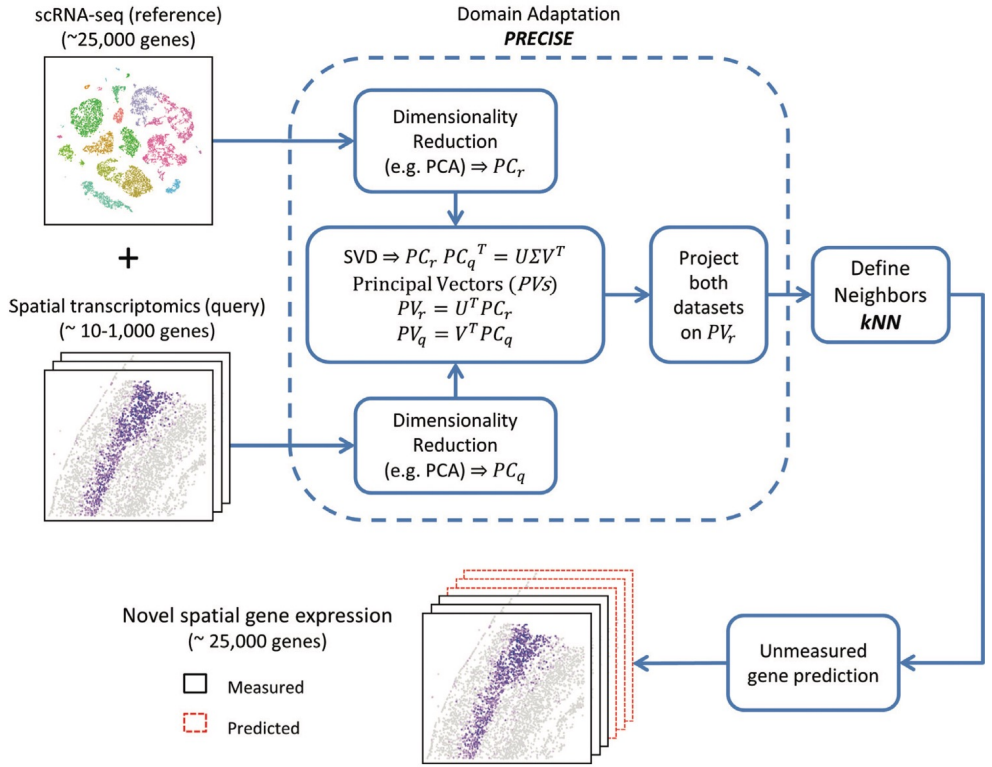
Single-cell genome-wide dataset (20K genes)

In situ probe-based spatial data (10-1K genes)

Integration of single-cell and spatial data (diagonal integration)

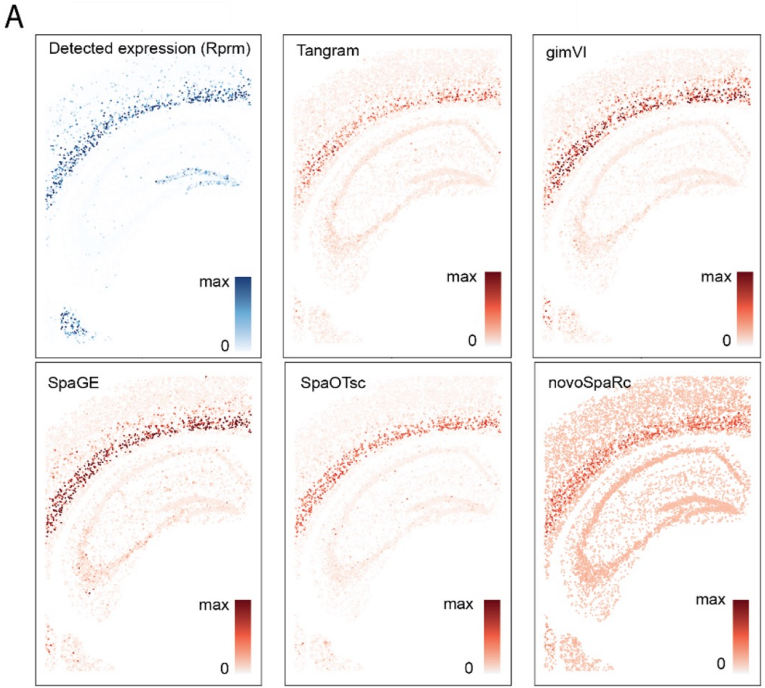
Define KNN in the integrated space across technologies (single cell vs spatial)

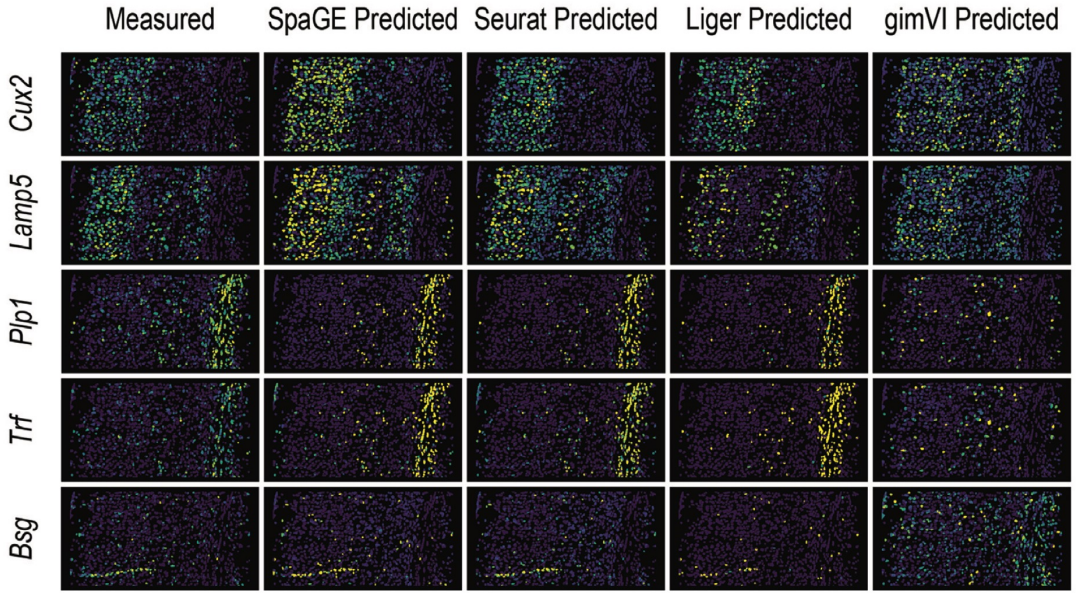
Predict gene expression value based on the expression of the single-cell



Abdelaal et al (2020) Nuclei Acid Recearch

Gene imputation method comparison





Salas et al (2023) BioRxiv

Abdelaal et al (2020) Nuclei Acid Recearch

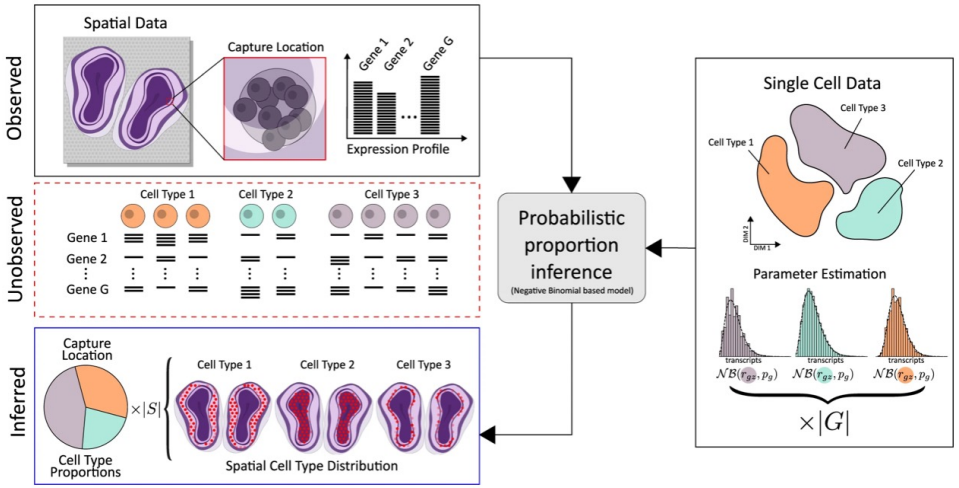
Deconvolution: Stereoscope

Creates a NB glm model on the single cell dataset

Creates a NB glm model on the bulk dataset, including a variable accounting for cell type proportions (w), given the gene-wise priors estimated from the single cell data

$$y_{gc} \sim \mathcal{NB}(s_c r_{gz_c}, p_g), \qquad s_c = \sum_{g \in G} y_{gc}.$$

$$x_{sg} \sim \mathcal{NB}\left(\beta_g \boldsymbol{v}_s^{\mathrm{T}} \boldsymbol{r}_g + \gamma_s \epsilon_g, p_g\right).$$



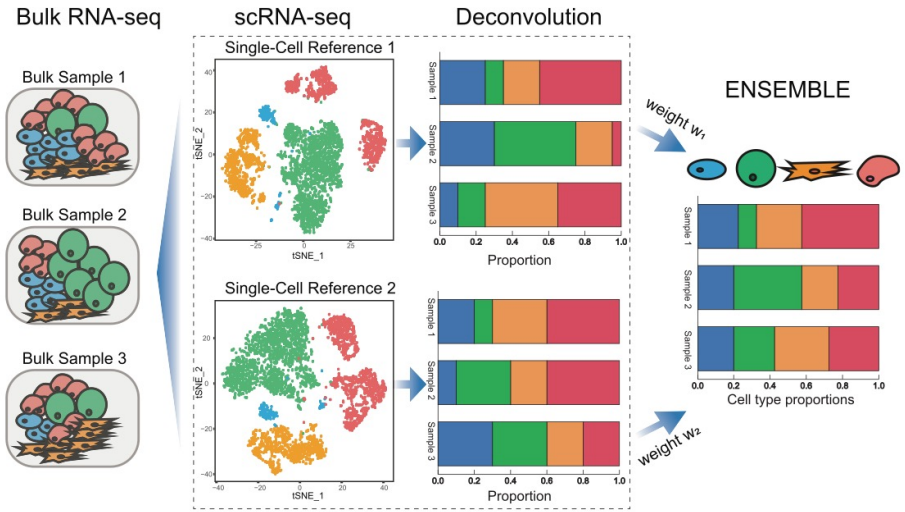
Andersson et al (2020) Communications Biology

Deconvolution: SCDC

Based on weighted non-negative least squares (WNNLS).

First, we create a basis-matrix contain the weighted gene expression across cell types in the single-cell dataset. Where P is the proportion of cell mixture (known).

Then, on the bulk data, SCDC try to minimize the distance between the basis matrix from single-cell and the bulk.

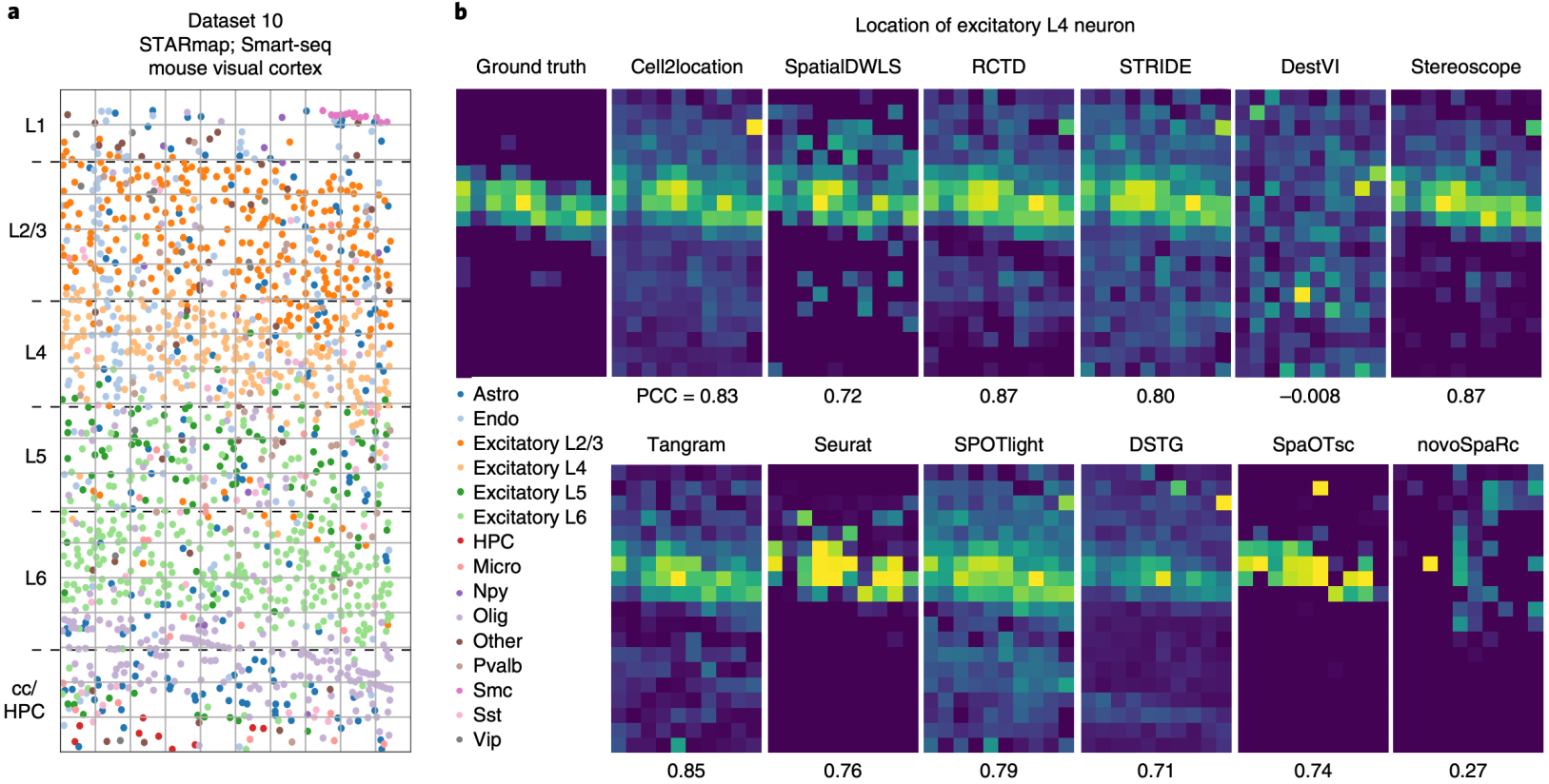


$$\hat{\mathbb{Y}} = \hat{\mathbb{B}}\hat{\mathbb{P}},$$

$$\mathbb{Y} \approx \mathbb{B}\mathbb{P},$$

Dong et al (2021) Briefings in Bioinformatics

Deconvolution method comparison



Li et al (2022) Nat Methods

Conclusions

Graph-based integration methods work well on very large datasets (> 100K samples/cells).

Matrix factorization are very flexible and accurate, but might not work in very large datasets.

Different integration techniques exist for different purposes, so please read up on benchmarking papers

Investigate which method is best on your case, and try a few (*most of them will agree to a certain extent*).

Thank you!

Paulo Czarnewski

ELIXIR Single-Cell Omics community co-lead

National Bioinformatics Infrastructure Sweden (ELIXIR-SE)