
Single cell integration

Nathalie Lehmann

Introduction

So far: worked on **1 individual** matrix
Generally: more than **1** sample

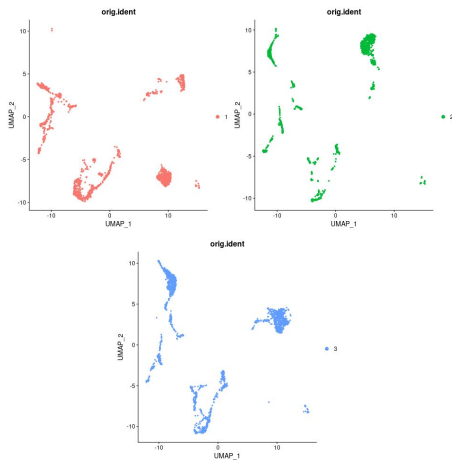
Introduction

So far: worked on **1 individual** matrix

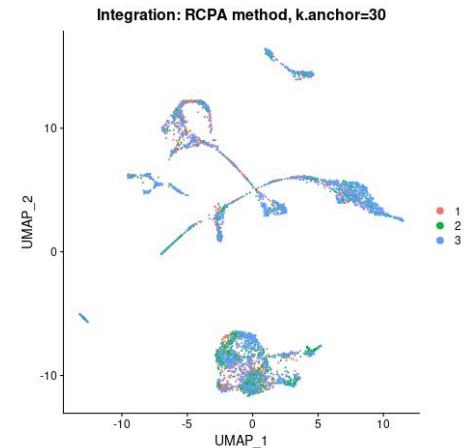
Generally: more than 1 sample

But should we study them

individually



all samples
together ?



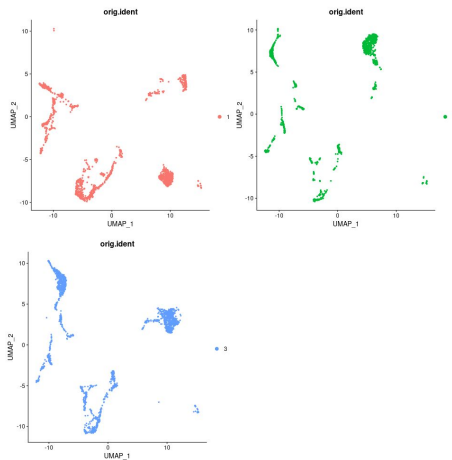
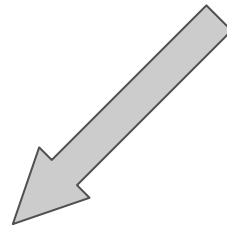
Introduction

So far: worked on **1 individual** matrix

Generally: more than 1 sample

But should we study them

individually



- Quick way to have a first look at data

- Repetitive
- Makes more sense to bring replicates together.
- Makes more sense to bring together similar samples (same experiment, organ...)

Introduction

So far: worked on **1 individual** matrix

Generally: more than **1 sample**

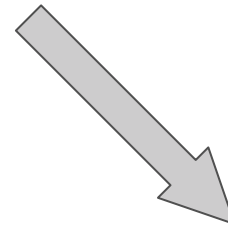
But should we study them



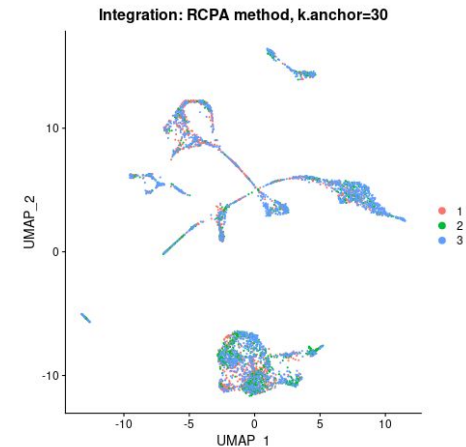
- Allows to work across multiple samples.
- Particularly important for cell populations visualization and identification
- Many cells : helps identifying rare populations



- Overcorrection ?



all samples together ?

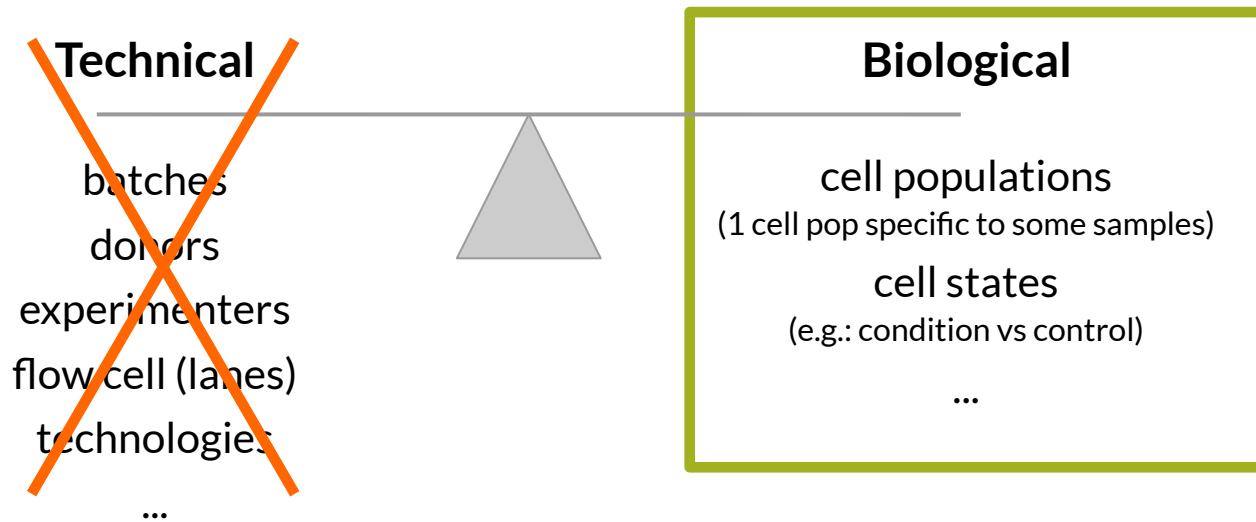


Question !

**What do you think are the challenges
in data integration ?**

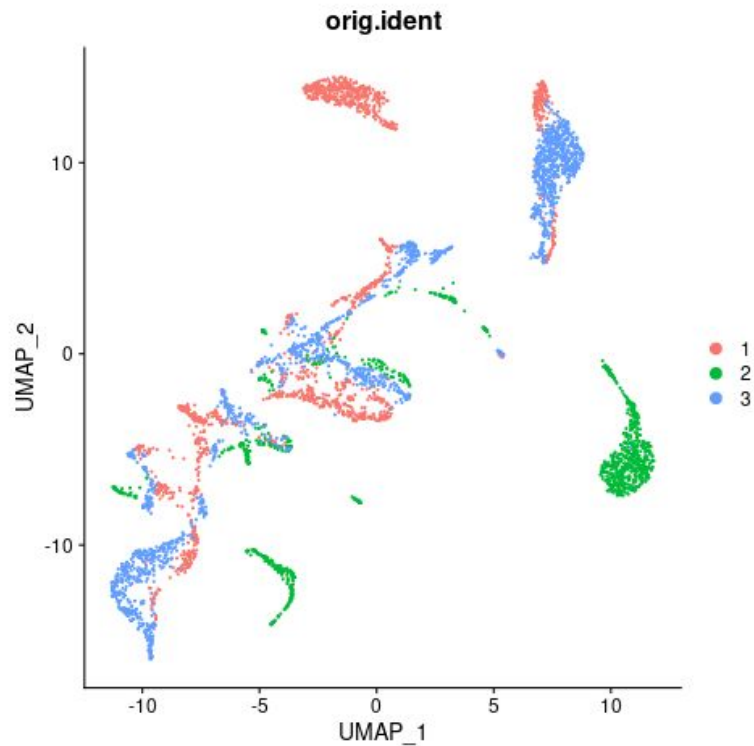
The challenge of data integration

2 sources of variability across samples



Question !

What do you think of this integration ?



Question !

What do you think of this integration ?



PBMCs

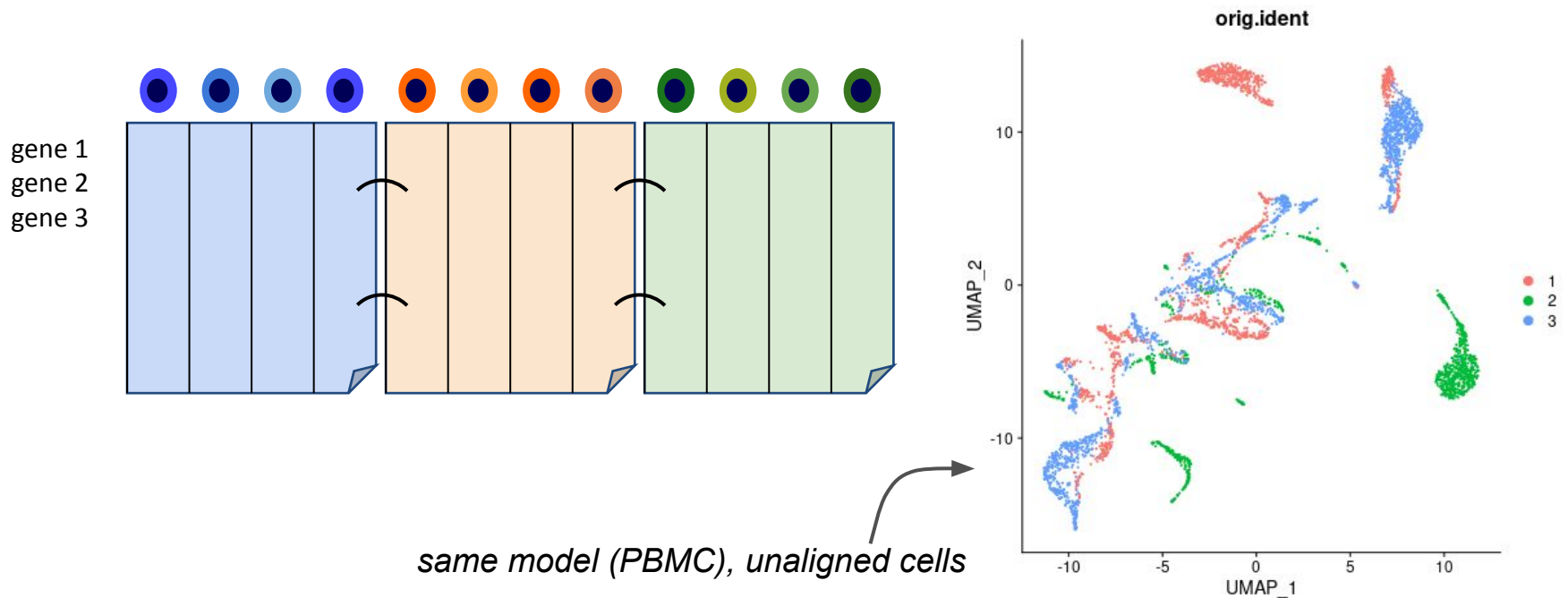
sample 1

sample 2

<https://www.10xgenomics.com>

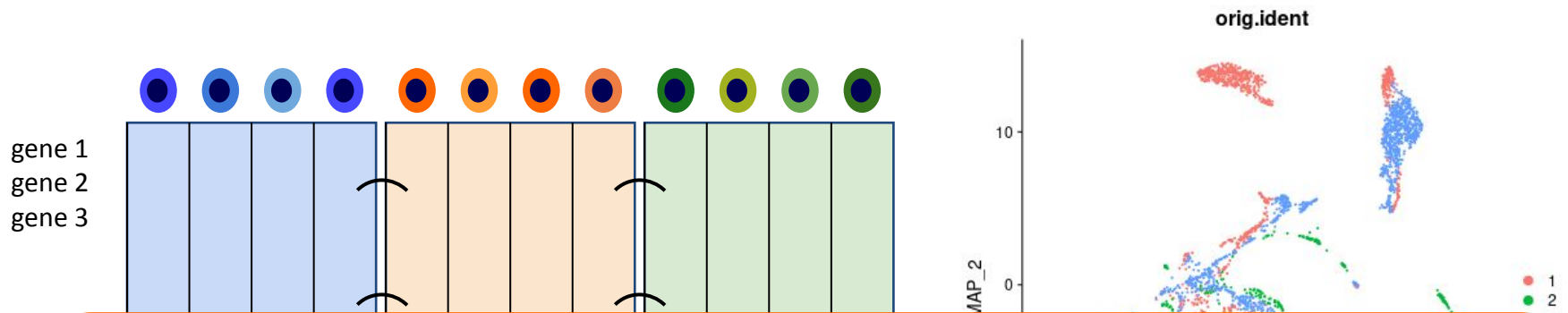
Problem...

...simple matrix concatenation does not always work



Problem...

...simple matrix concatenation does not always work



This is typically a problem of batch effect

We need a more sophisticated **integration method**

When to integrate

- **Do not integrate**
e.g.: replicates generated in the same time and exactly in the same manner may not need integration

PBMCs

sample 1
sample 2



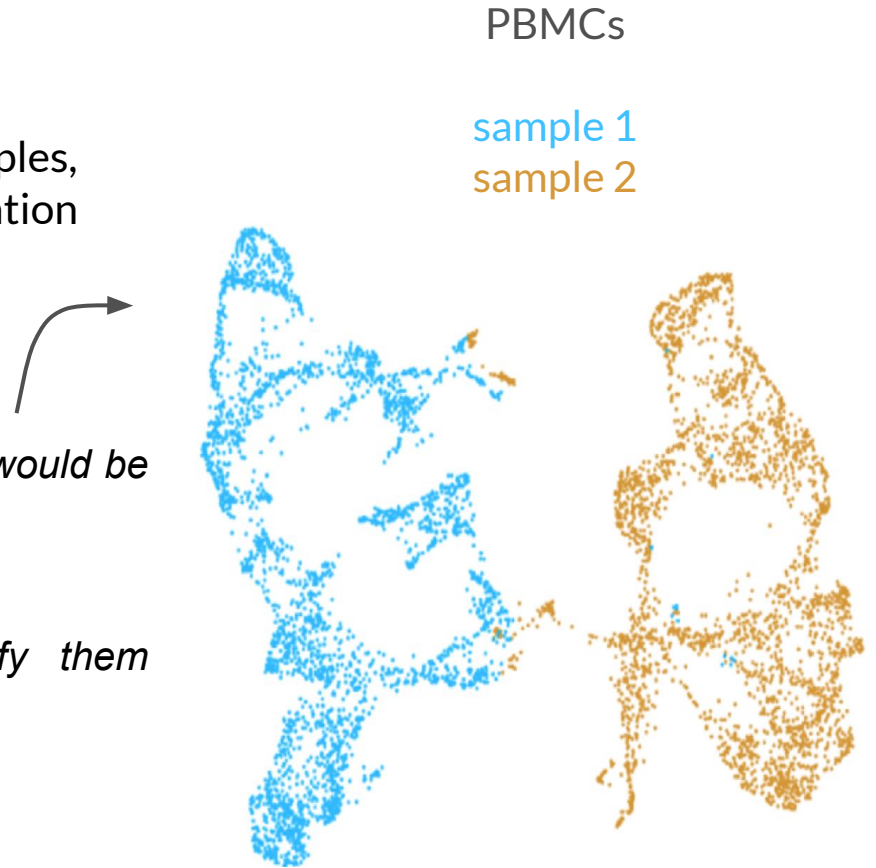
<https://www.10xgenomics.com>

When to integrate ?

- **Integrate**
when obvious batch effect between samples,
typically seen on low dimension visualization

In this example, the sample of origin would be a huge bias for clustering

The samples need integration to align cell types/clusters and then identify them correctly

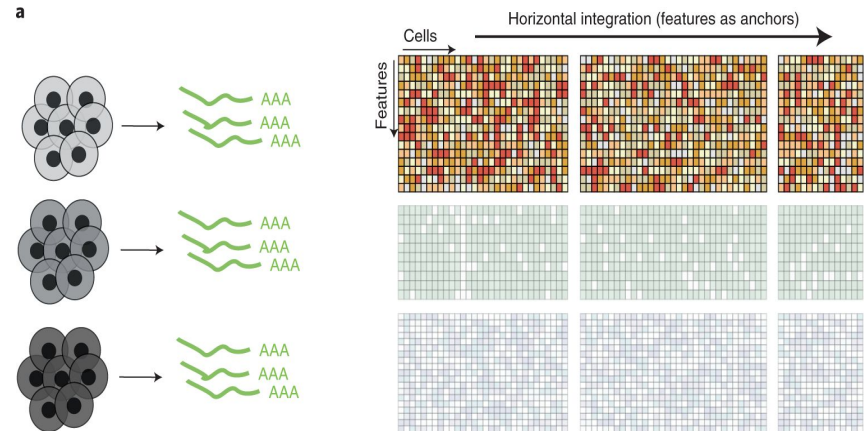


<https://www.10xgenomics.com>

Many methods

Different types of integrations

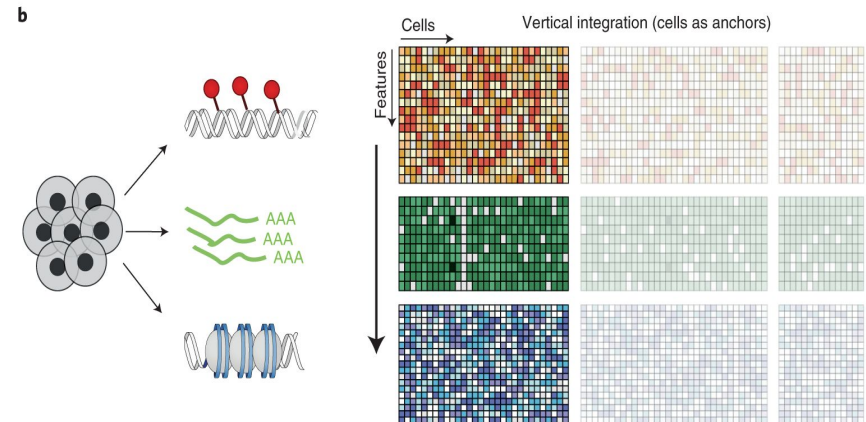
- Horizontal: different samples same modality



Different types of integrations

- Horizontal: different samples same modality

- Vertical: same sample different modalities (multiomics)

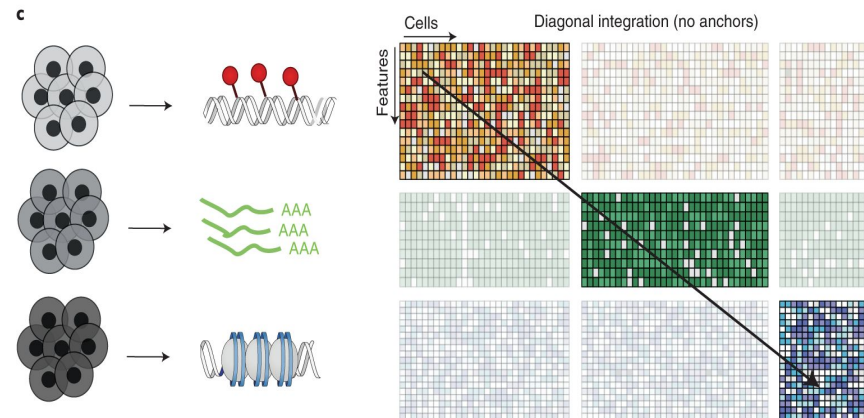


Different types of integrations

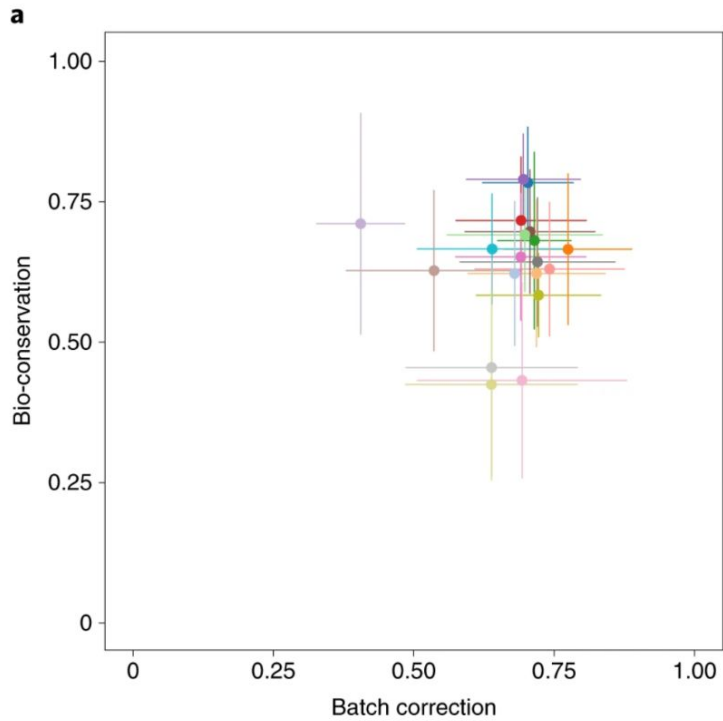
- Horizontal: different samples
same modality

- Vertical: same sample
different modalities
(multiomics)

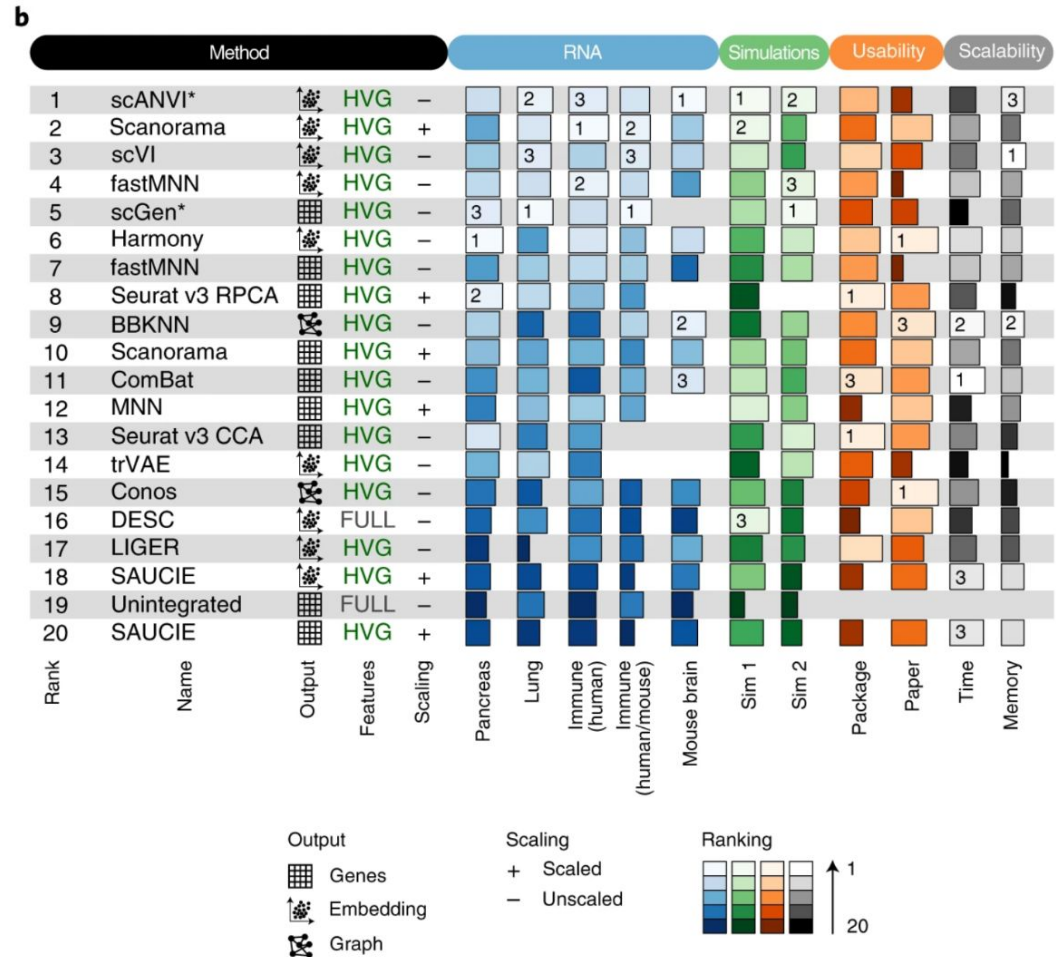
- Diagonal: different samples
different modalities



Many methods



- scANVI* (embedding, HVG, unscaled)
- Scanorama (embedding, HVG, scaled)
- scVI (embedding, HVG, unscaled)
- fastMNN (embedding, HVG, unscaled)
- scGen* (genes, HVG, unscaled)
- Harmony (embedding, HVG, unscaled)
- fastMNN (genes, HVG, unscaled)
- Seurat v3 RPCA (genes, HVG, scaled)
- BBKNN (graph, HVG, unscaled)
- Scanorama (genes, HVG, scaled)
- ComBat (genes, HVG, unscaled)
- MNN (genes, HVG, scaled)
- Seurat v3 CCA (genes, HVG, unscaled)
- trVAE (embedding, HVG, unscaled)
- Conos (graph, full, unscaled)
- DESC (embedding, full, unscaled)
- LIGER (embedding, HVG, unscaled)
- SAUCIE (embedding, HVG, scaled)
- SAUCIE (genes, HVG, scaled)



Many methods

Harmony

| | scANVI | Scanorama embed | scVI | FastMNN embed | scGen | Harmony | fastMNN gen | MNN | Seurat v3 RPCA | tBKN | Scanorama gene | ComBat | MNN | Seurat v3 CCA | tVAE | Conos | DESC | LIGER | SAUCIE embed | SAUCIE gene |
|--|--------|-----------------|--------|---------------|--------|---------|-------------|-----|----------------|--------|----------------|----------|----------|---------------|--------|-------|--------|-------|--------------|-------------|
| Input | Python | Python | Python | R | Python | R | R | R | R | Python | Python | Python/R | Python/R | R | Python | R | Python | R | Python | Python |
| Method runs without additional information | X | | | | X | | | | | | | | | | | | | | | |
| Scib results | ✓ | ✓ | ✓ | | ✓ | | | | | | | | | | | | | | | |
| Consistent top performer | ✓ | ✓ | ✓ | | ✓ | | | | | | | | | | | | | | | |
| Top method on small/simple tasks | | ✓ | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | |
| Top method on large/complex tasks | ✓ | ✓ | ✓ | | ✓ | | | | | | | | | | | | | | | |
| Top method on ATAC data | - | | - | | | ✓ | | | | | | | | | | | | ✓ | | |
| Task details | ✓ | - | - | | ✓ | | | | - | - | | | | - | | | | | | |
| Integrates strong batch effects | ✓ | | | | ✓ | | | | | | | | | | | | | | | |
| Top method for recovery cell states or modules | ✓ | ✓ | | | | | | | | | ✓ | ✓ | ✓ | | | | | | | |
| Confounding of bio and batch variance | ✓ | - | | | ✓ | | | | | | | | | | | | | | | |
| Top method for trajectories | - | ✓ | - | ✓ | ✓ | | | | | | | | | | | | | | | |
| Method deals with varying compositions | | | | | | | | | | | | X | | | | | | | | |
| Speed | | | | | | | | | | | | | | | | | | | | |
| Fast method for quick results | | | | | | | | | | ✓ | | ✓ | | | | | | | | |
| Scales well to large datasets on CPU | ✓ | - | ✓ | | | | | | | ✓ | | | | | | | | | ✓ | ✓ |
| Method has GPU support | ✓ | | ✓ | | ✓ | | | | | | | | | | ✓ | | ✓ | | ✓ | ✓ |
| Scales well to feature spaces beyond genes | | | | | | | | | | | | | | | ✓ | ✓ | | | | |
| Output | | | | | ✓ | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | ✓ |
| Method shows corrected expression | | | | | | | | | | | | | | | | | | | | |
| Method gives relative cell embeddings | | | | | | | | | X | | | | | | | X | | | | |

✓ Fulfills the criterion

— Partial fulfillment of criterion

X Does not fulfill criterion

Python

R

Seurat v3

Luecken et al., Nature Methods 2022

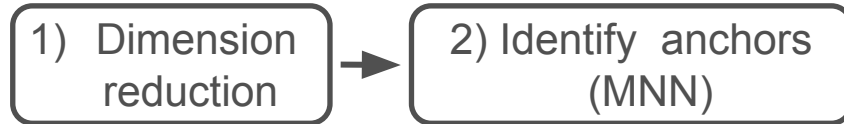
A few benchmarks, that do not agree with each other

Büttner et al., Nat. Methods. 2019
 Chen et al., Nat. Biotechnol 2020
 Tran et al., Genome Biol. 2020

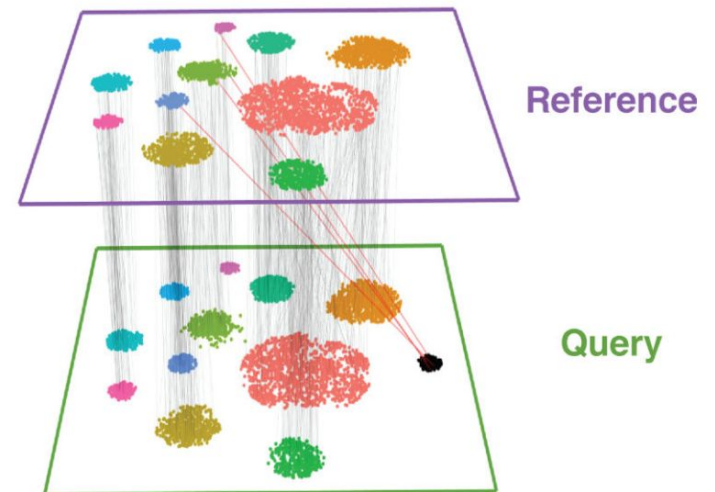
Integration with Seurat



Algorithm



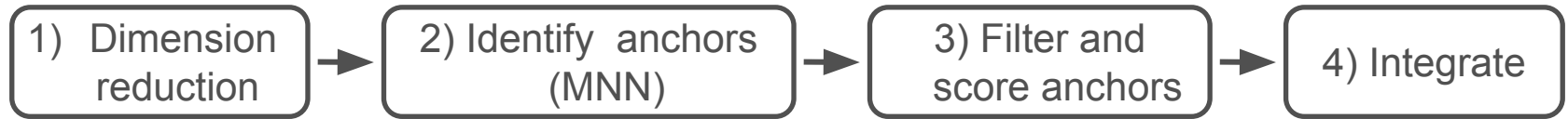
- MNN: Mutual Nearest Neighbors
- In **reference** and **query**, identify 2 cells that are close (neighbors) in terms of euclidean distance: **anchors**
- Identify many anchors



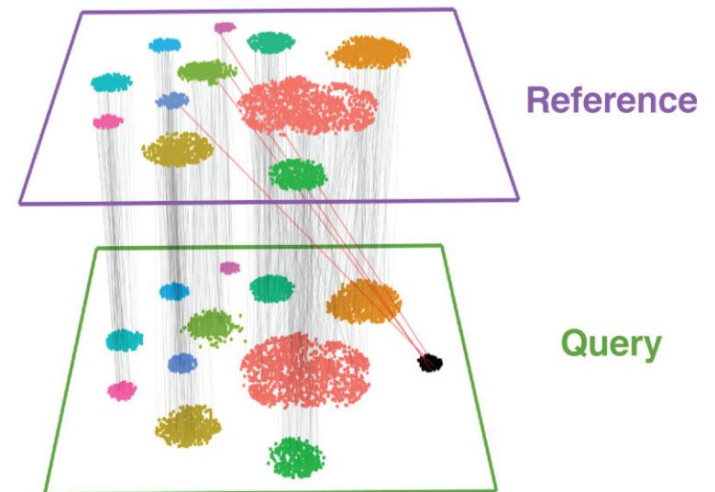
Integration with Seurat



Algorithm



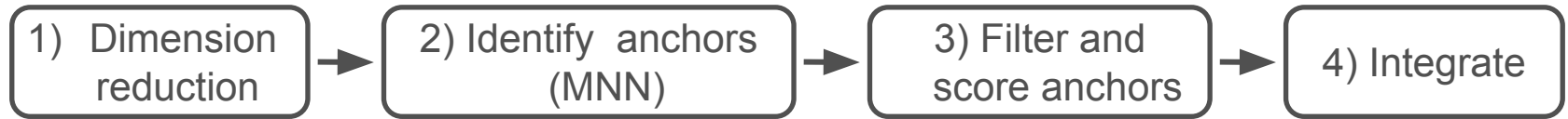
- Deduce correction from anchors
- Apply correction vector to all query cells.



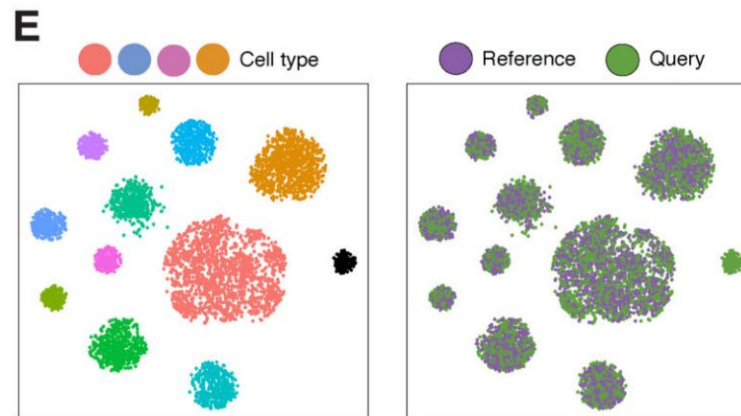
Integration with Seurat



Algorithm



- Deduce correction from anchors
- Apply correction vector to all query cells.

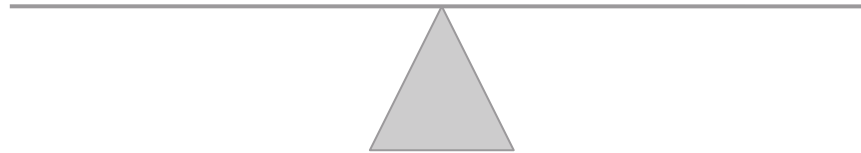


Conclusion

A good integration method

Technical

Biological



- Corrects for technical variability:

- samples
- donors
- experimenter
- technologies

- Preserves biological signal

- cell types across different samples, tissues
- cell trajectories
- differences (cell subtypes, cell states) between condition and control
- population (cell subtypes, cell states) unique to a condition...

Acknowledgements

Parts of this course are inspired by

The *Swiss Institute of Bioinformatics* course [Single Cell Transcriptomics](#)

Slides inspired from Rémi Montagne (Thanks !)