Single cell integration

Nathalie Lehmann

EBAII 2024 - 20.11.2024

Inserm

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

INSTITUT PASTEUR

Introduction

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

Generally: more than 1 sample

Introduction

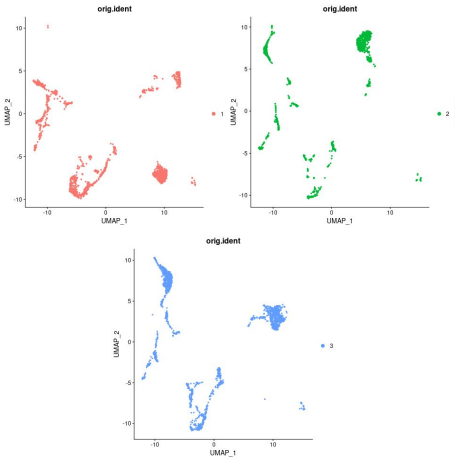
So far: worked on 1 **individual** matrix

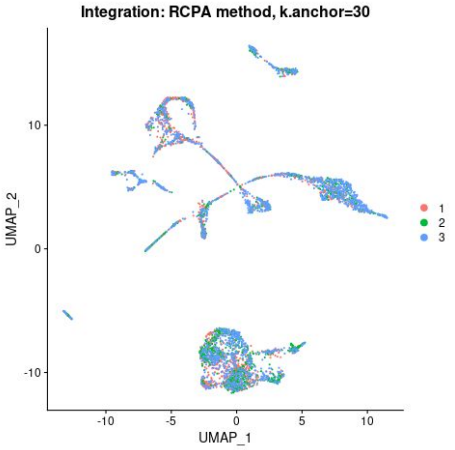
Generally: more than 1 sample

But should we study them

individually

all samples together ?





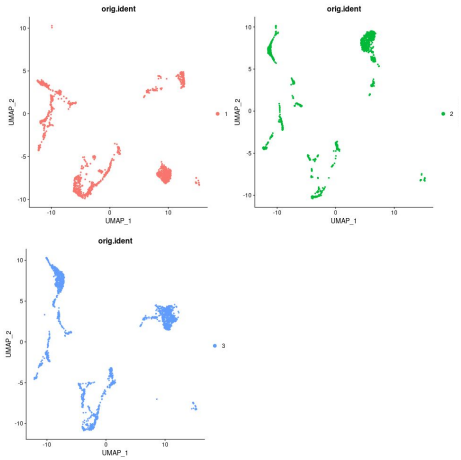
Introduction

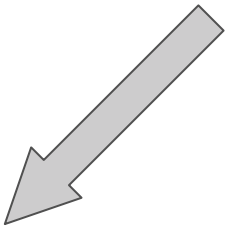
So far: worked on 1 **individual** matrix

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

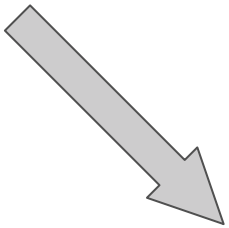
So far: worked on 1 **individual** matrix

Generally: more than 1 sample

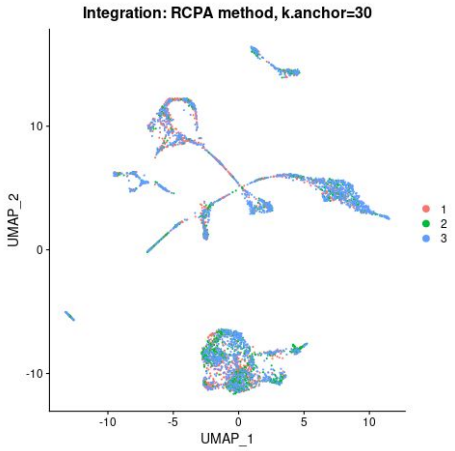
But should we study them

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

- Overcorrection ?



all samples together ?

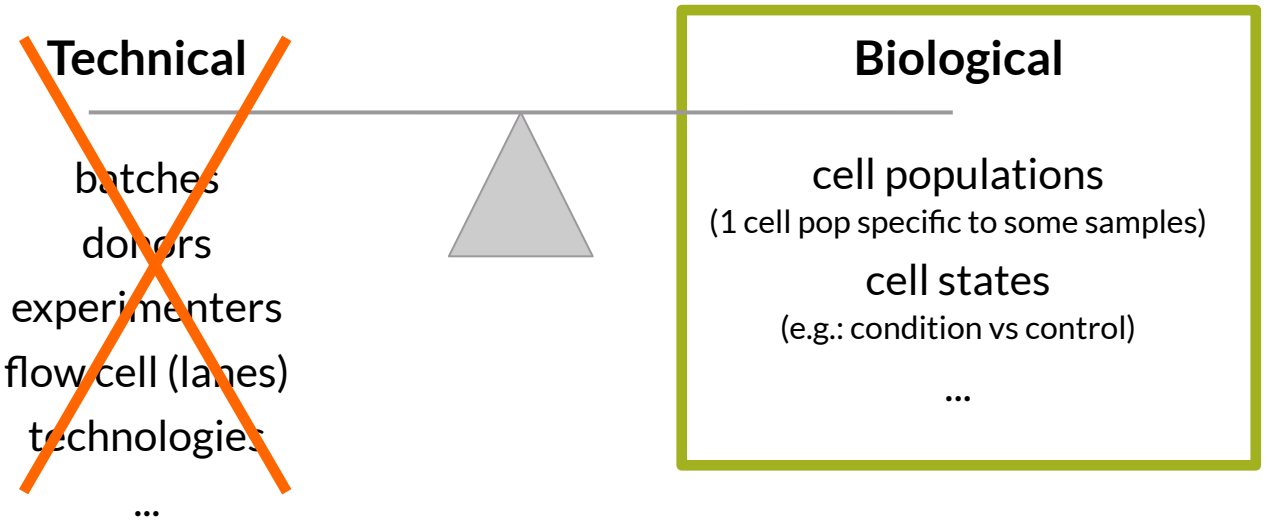


Question !

What do you think are the challenges in data integration ?

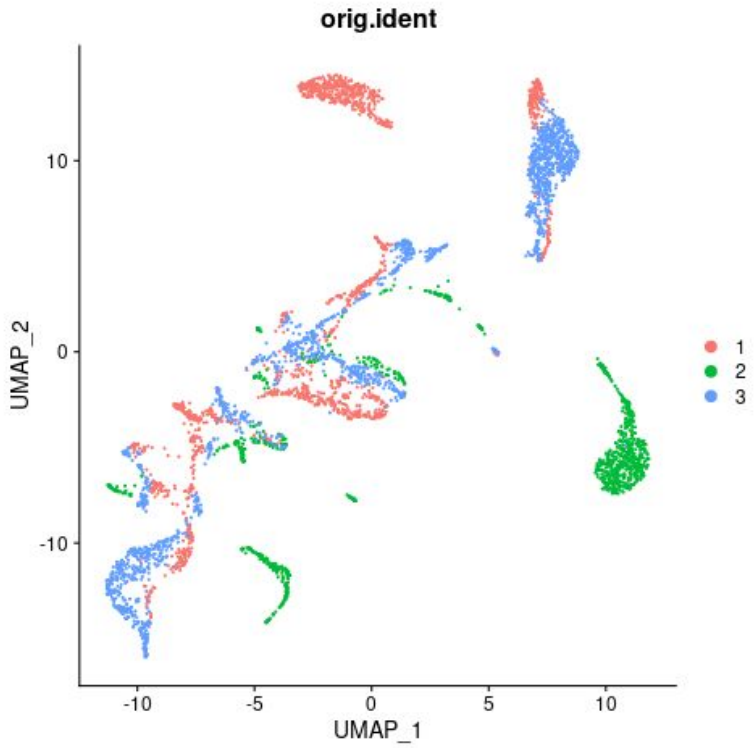
The challenge of data integration

2 sources of variability across samples



Question !

What do you think of this integration ?



Question !

What do you think of this integration ?

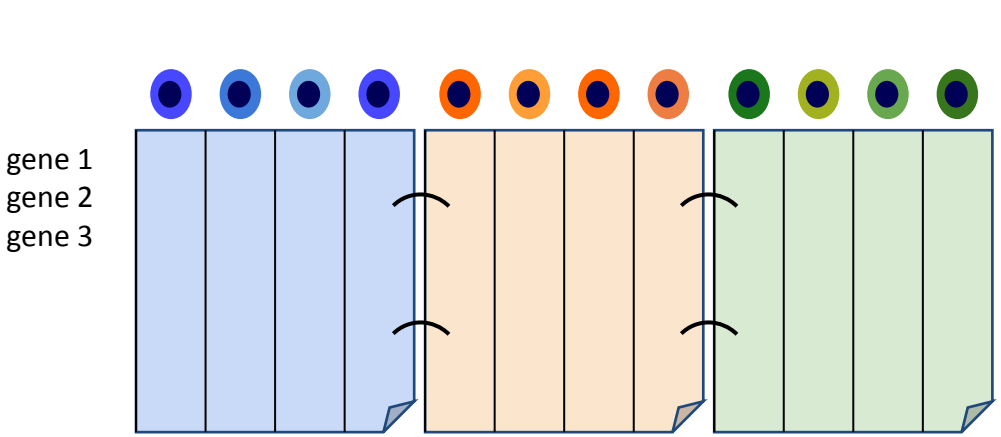
PBMCs

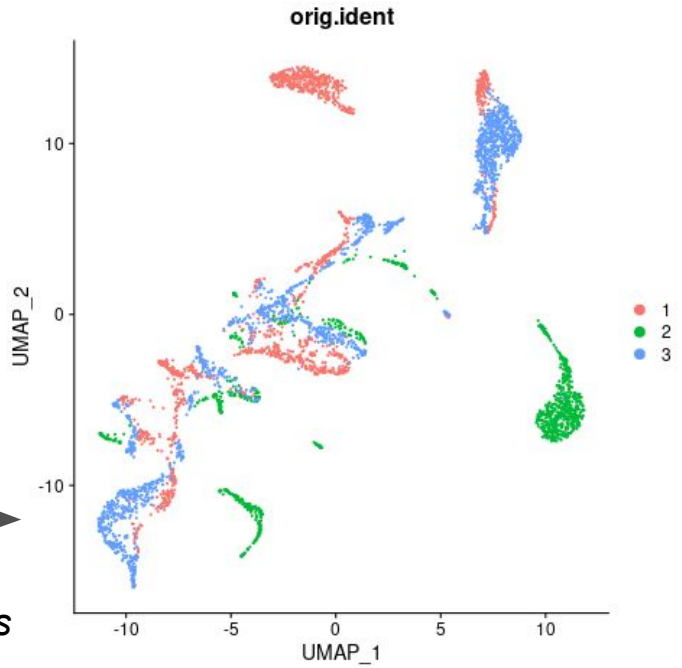
sample 1
sample 2

https://www.10xgenomics.com

Problem…

…simple matrix concatenation does not always work

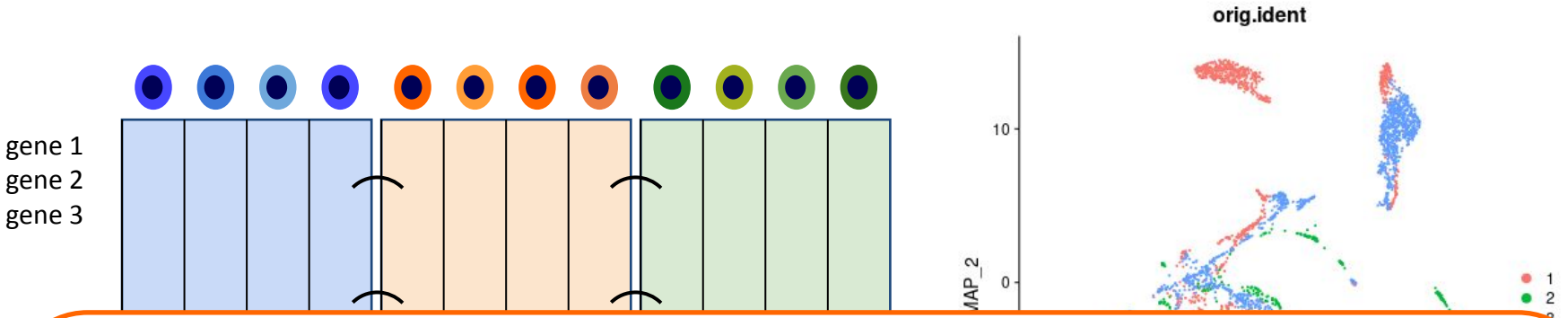




same model (PBMC), unaligned cells

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 ?

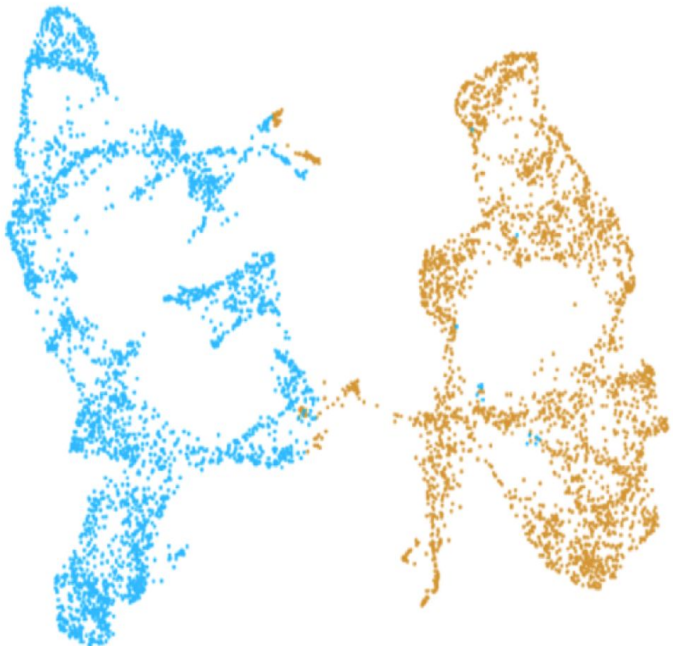
- **Integrate**
  when obvious batch effect between samples, typically seen on low dimension visualization

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

PBMCs

sample 1
sample 2

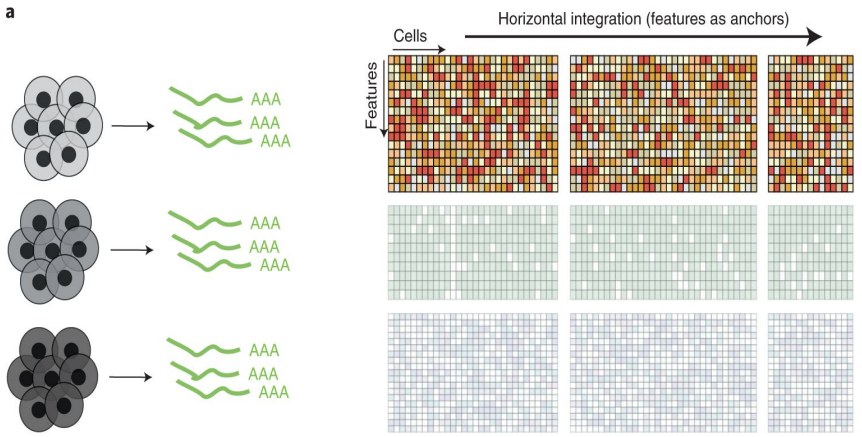


https://www.10xgenomics.com

Many methods

Different types of integrations

- Horizontal: different samples same modality

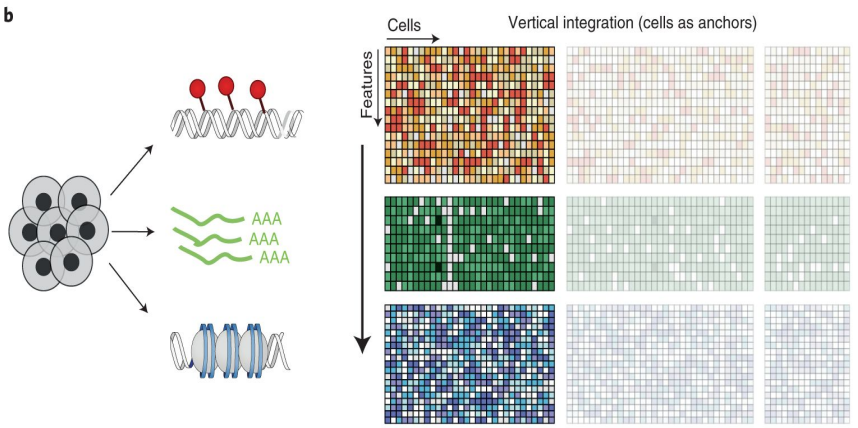


Luecken *et al.*, Nat Met 2021

Different types of integrations

- Horizontal: different samples
  same modality

- Vertical: same sample
  different modalities
  (multiomics)



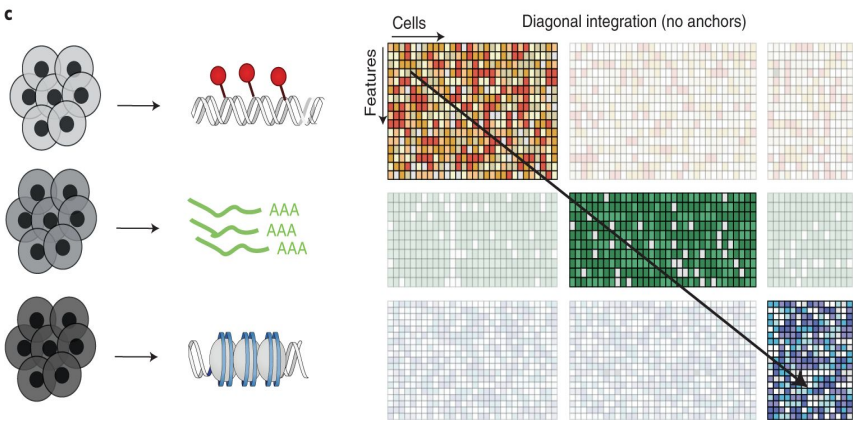
Luecken *et al.*, Nat Met 2021

Different types of integrations

- Horizontal: different samples
  same modality

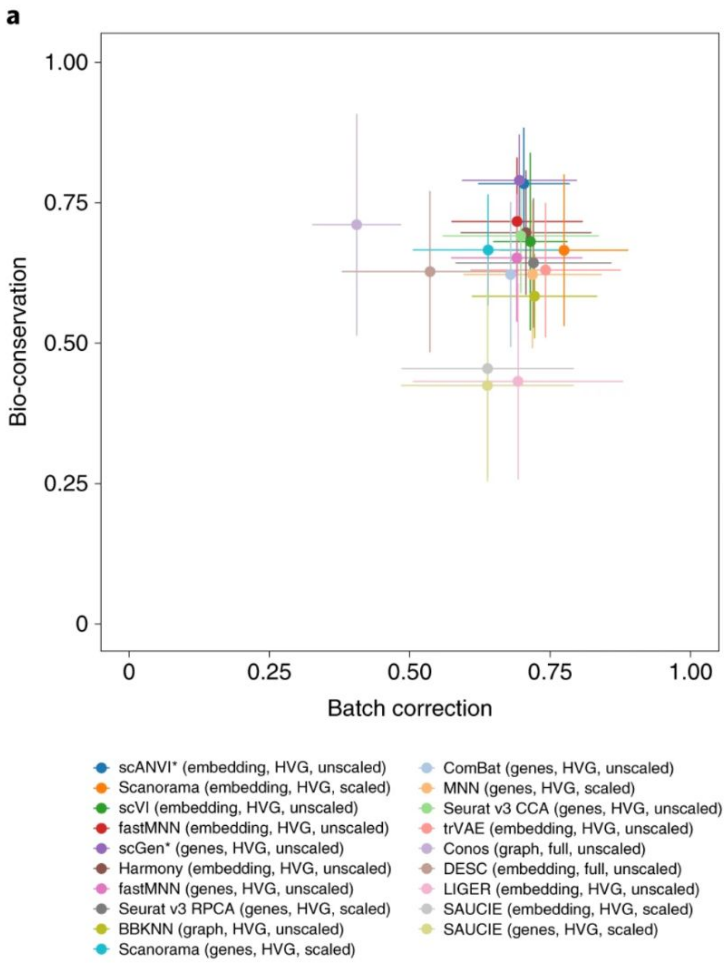
- Vertical: same sample
  different modalities
  (multiomics)

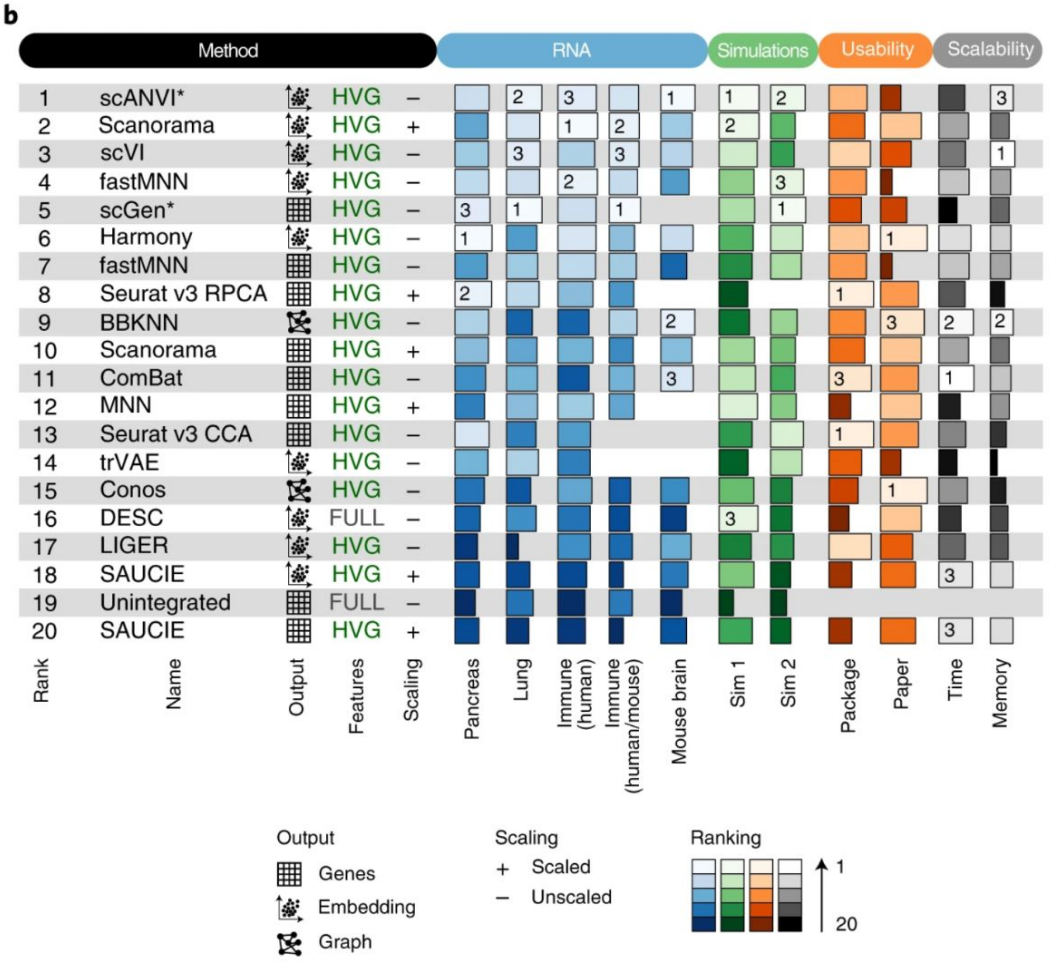
- Diagonal: different samples
  different modalities



Luecken *et al.*, Nat Met 2021

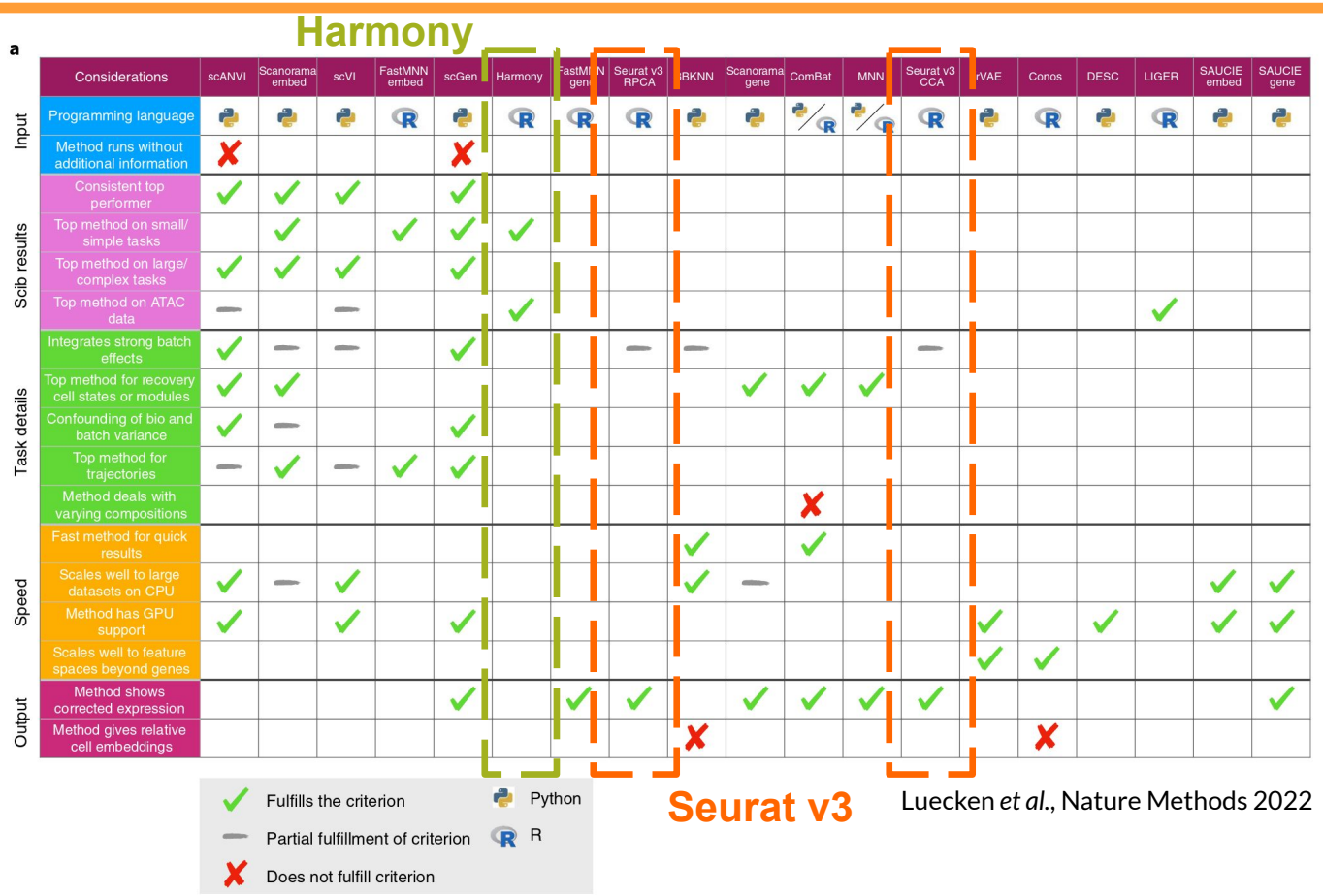
Many methods





Luecken *et al.*, Nature Methods 2022

Many methods



Luecken *et al.*, Nature Methods 2022

A few benchmarks, that do not agree with each other

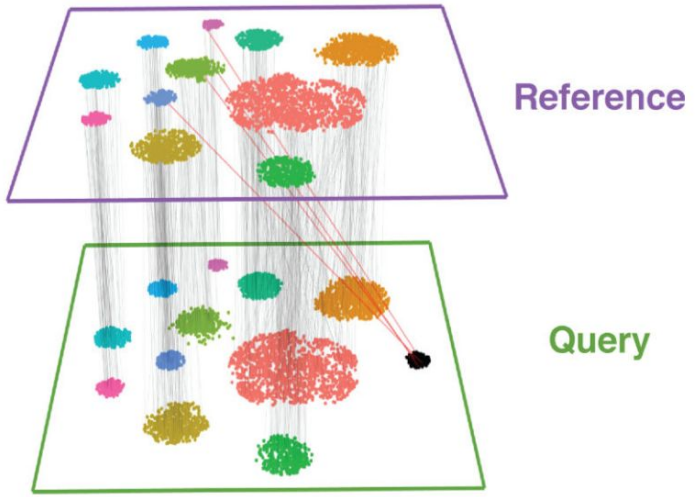
Büttner *et al.*, Nat. Methods. 2019
Chen *et al.*, Nat. Biotechnol 2020
Tran *et al.*, Genome Biol. 2020

Integration with Seurat

Algorithm

1) Dimension reduction → 2) Identify anchors (MNN)

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



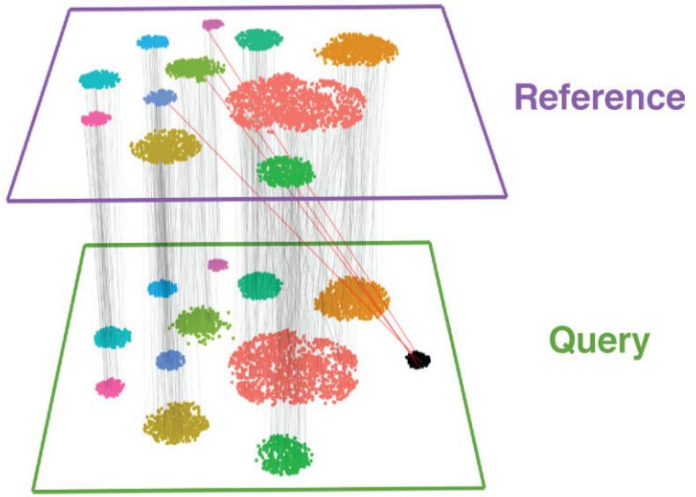
Stuart *et al.*, Cell 2019

Integration with Seurat

Algorithm

1) Dimension reduction → 2) Identify anchors (MNN) → 3) Filter and score anchors → 4) Integrate

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



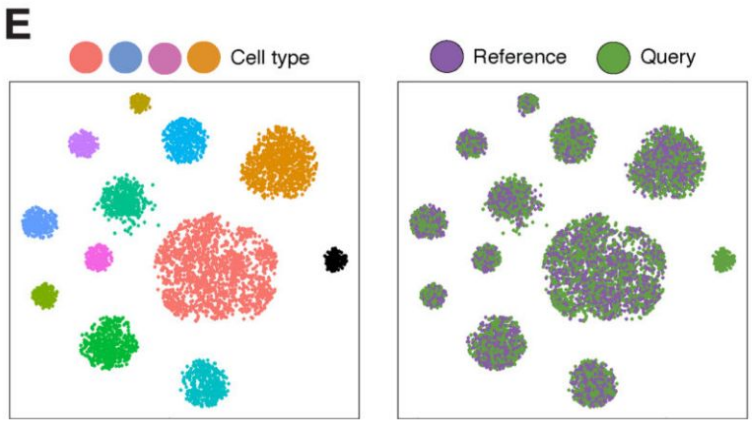
Stuart *et al.*, Cell 2019

Integration with Seurat

Algorithm

1) Dimension reduction → 2) Identify anchors (MNN) → 3) Filter and score anchors → 4) Integrate

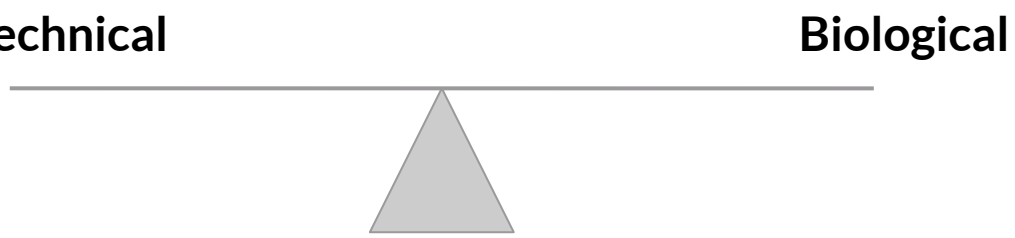
- Deduce correction from anchors
- Apply correction vector to all query cells.



Stuart *et al.*, Cell 2019

Conclusion

A good integration method



- 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 !)