Single cell sample integration

Remi Montagne

EBAII 2023 - 11/08/2023

aviesan

alliance nationale
pour les sciences de la vie et de la santé

Introduction

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

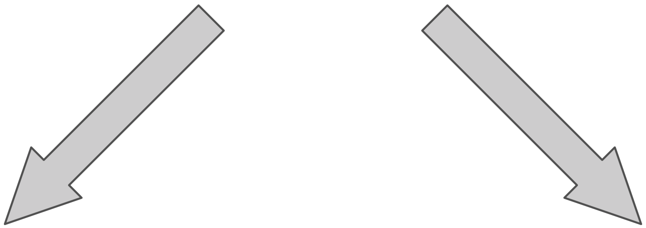
Generally: more than 1 sample

Introduction

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

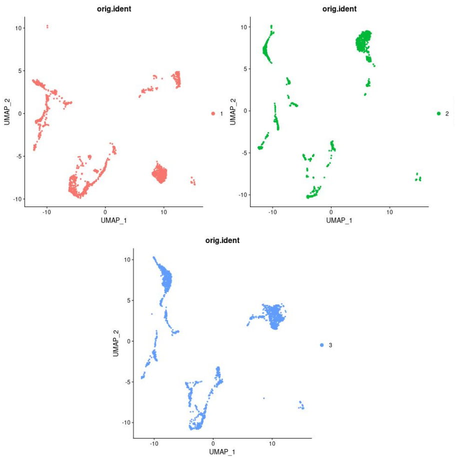
Generally: more than 1 sample

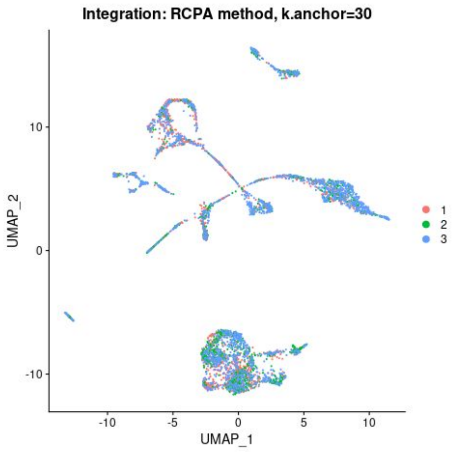
But should we study them



individually

all samples together ?





Integration: RCPA method, k.anchor=30

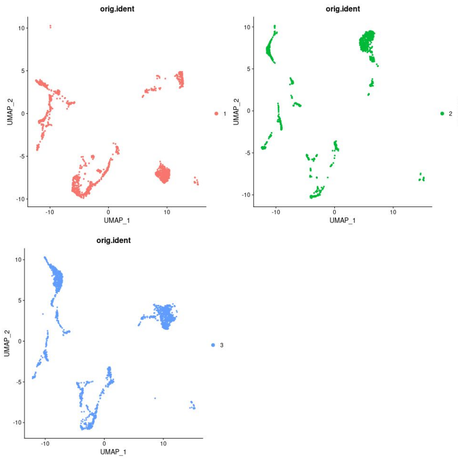
Introduction

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...)

aviesan

alliance nationale
pour les sciences de la vie et de la santé

Introduction

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

Generally: more than 1 sample

But should we study them

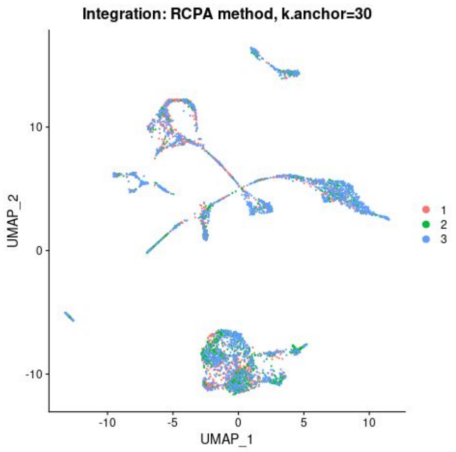
all samples together ?

🙂

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

🙁

- Overcorrection ?



Introduction

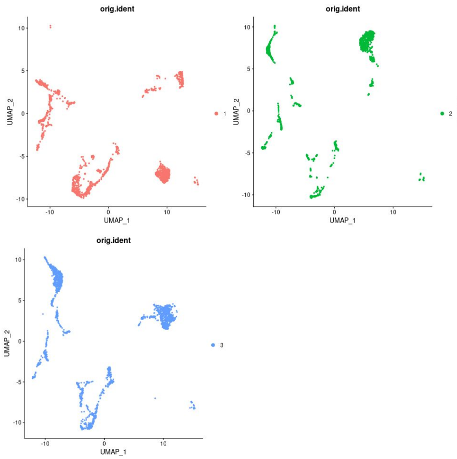
So far: worked on 1 **individual** matrix

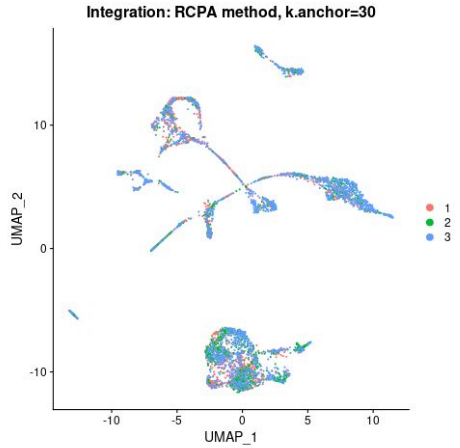
Generally: more than 1 sample

But should we study them

individually

all samples together ?





Introduction

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

Generally: more than 1 sample

We will study them together at the same time

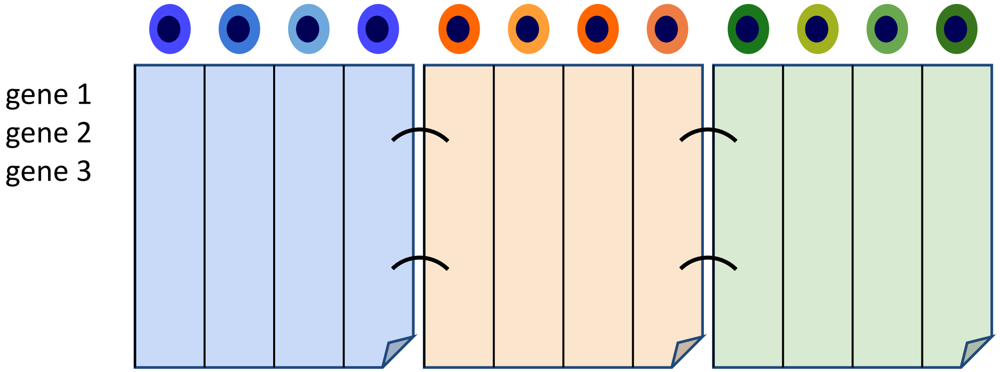
Introduction

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

Generally: more than 1 sample

We will study them together at the same time

Problem: simple matrix concatenation does not always work



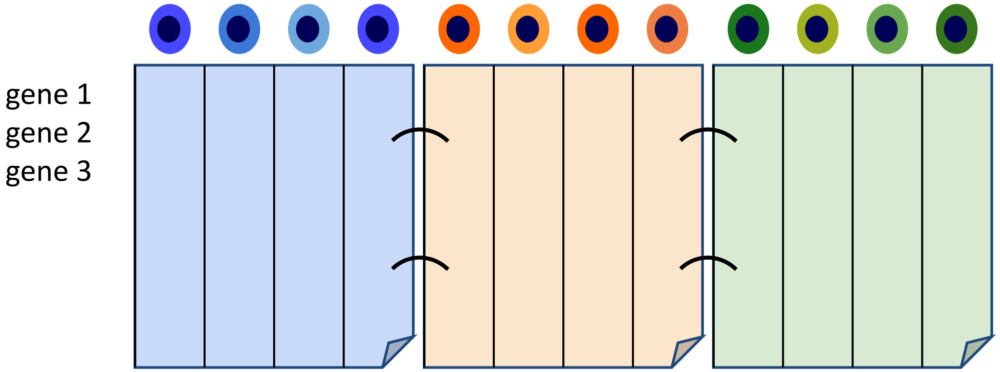
Introduction

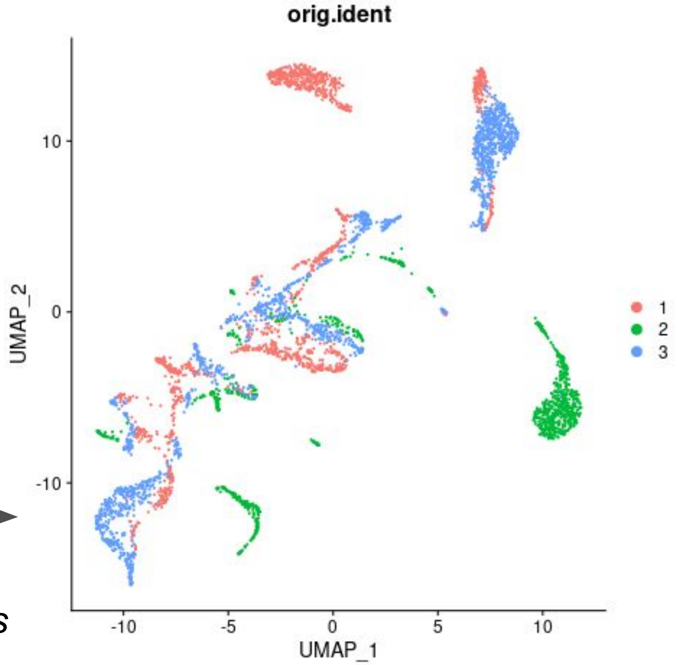
So far: worked on 1 **individual** matrix

Generally: more than 1 sample

We will study them together at the same time

Problem: simple matrix concatenation does not always work





same model (PBMC), unaligned cells

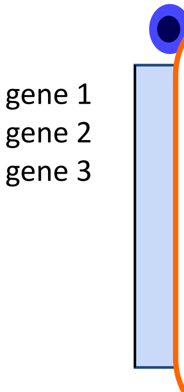
Introduction

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

Generally: more than 1 sample

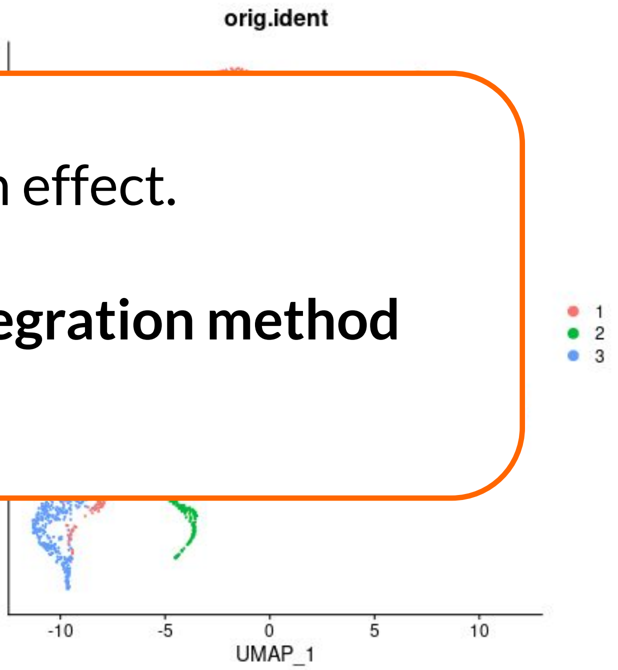
We will study them together at the same time

Problem: simple matrix concatenation does not always work



This is a problem of batch effect.

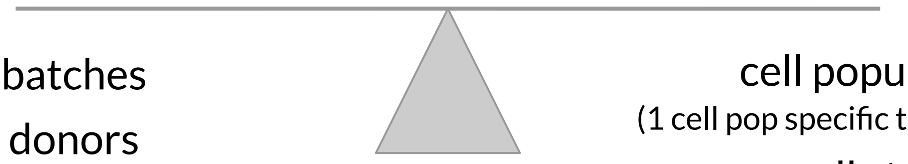
We need a more sophisticated **integration method**



Variability across samples

Variability across samples

2 sources of variability across samples



Technical

batches

donors

experimenters

flow cell (lanes)

technologies

...

Biological

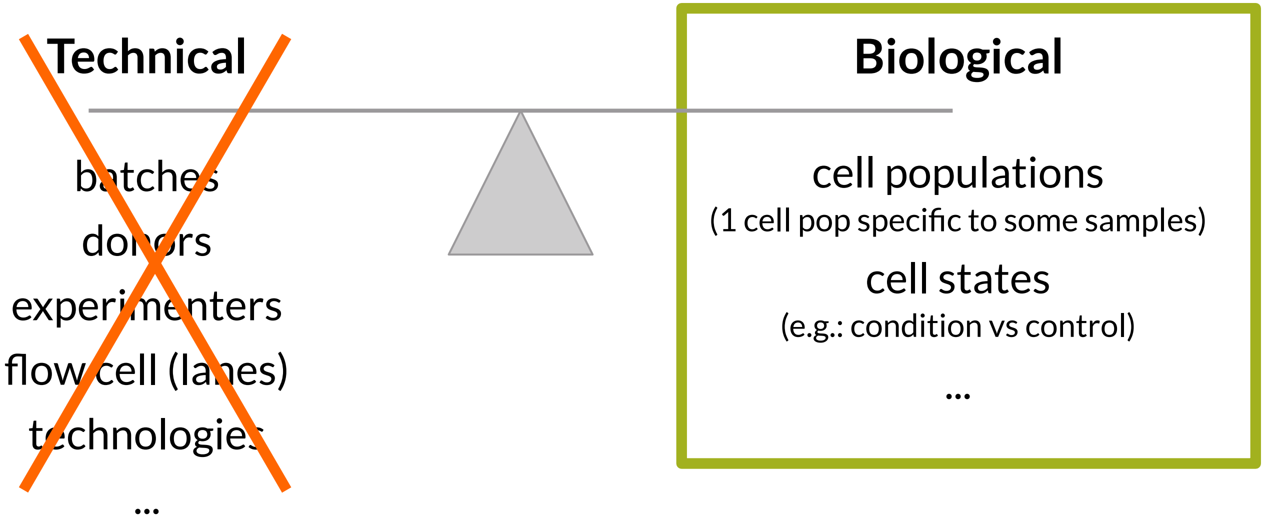
cell populations
(1 cell pop specific to some samples)

cell states
(e.g.: condition vs control)

...

Variability across samples

2 sources of variability across samples



~~Technical~~

~~batches~~

~~donors~~

~~experimenters~~

~~flow cell (lanes)~~

~~technologies~~

...

Biological

cell populations

(1 cell pop specific to some samples)

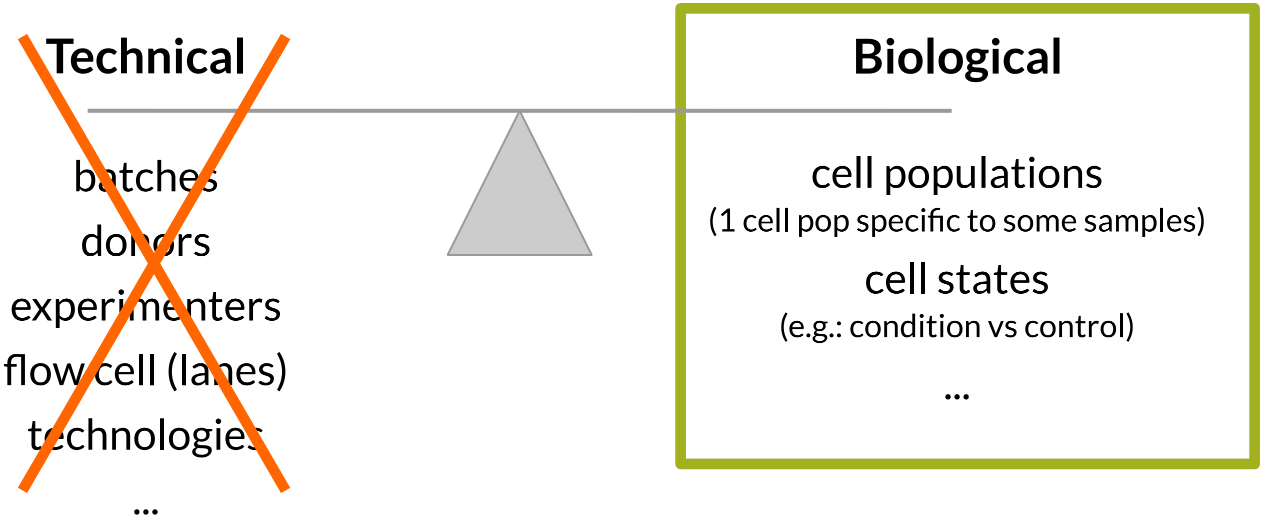
cell states

(e.g.: condition vs control)

...

Variability across samples

2 sources of variability across samples



~~Technical~~

~~batches~~

~~donors~~

~~experimenters~~

~~flow cell (lanes)~~

~~technologies~~

...

Biological

cell populations

(1 cell pop specific to some samples)

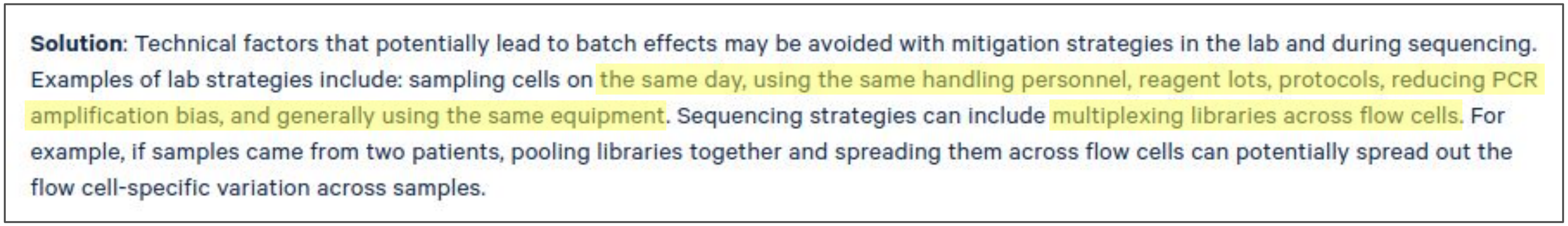
cell states

(e.g.: condition vs control)

...

→ Solutions:

Strategies to avoid factors causing batch effect in the lab

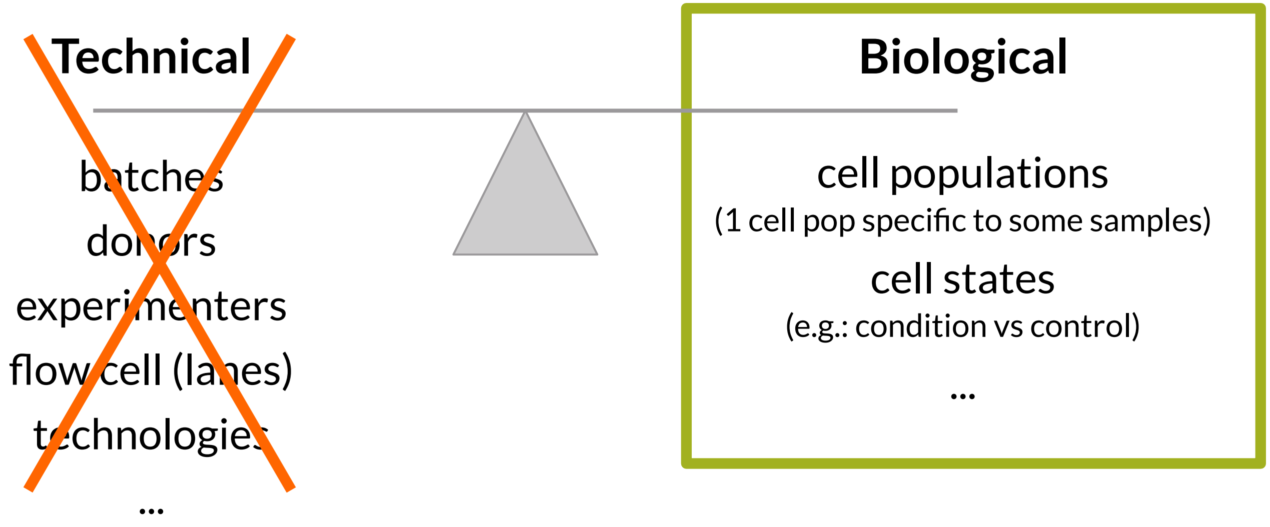


Solution: Technical factors that potentially lead to batch effects may be avoided with mitigation strategies in the lab and during sequencing. Examples of lab strategies include: sampling cells on the same day, using the same handling personnel, reagent lots, protocols, reducing PCR amplification bias, and generally using the same equipment. Sequencing strategies can include multiplexing libraries across flow cells. For example, if samples came from two patients, pooling libraries together and spreading them across flow cells can potentially spread out the flow cell-specific variation across samples.

https://www.10xgenomics.com/resources/analysis-guides/introduction-batch-effect-correction

Variability across samples

2 sources of variability across samples



~~**Technical**~~

~~batches~~

~~donors~~

~~experimenters~~

~~flow cell (lanes)~~

~~technologies~~

~~...~~

Biological

cell populations
(1 cell pop specific to some samples)

cell states
(e.g.: condition vs control)

...

→ Solutions:

Strategies to avoid factors causing batch effect in the lab

Computational data integration

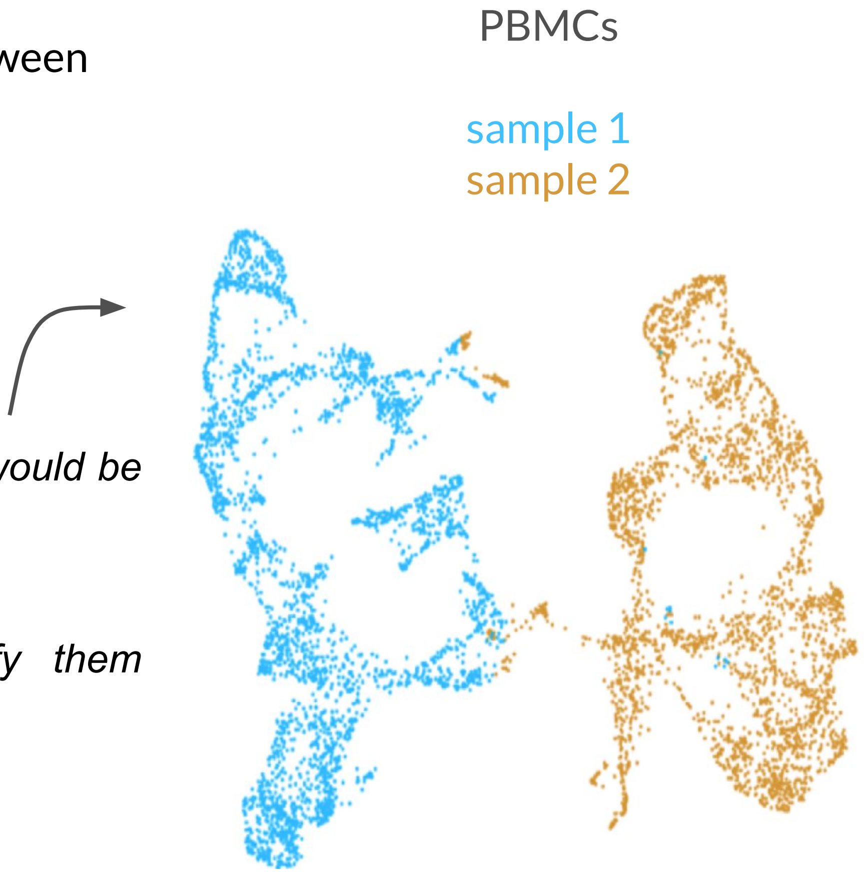
When to integrate

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

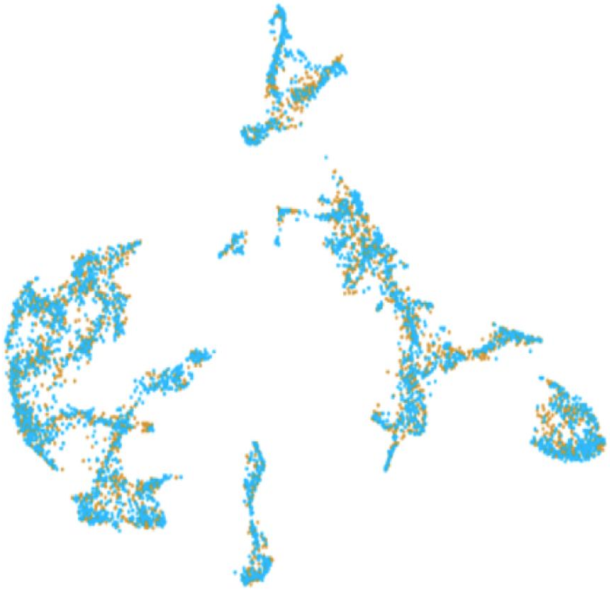
When to integrate

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

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

PBMCs

sample 1
sample 2



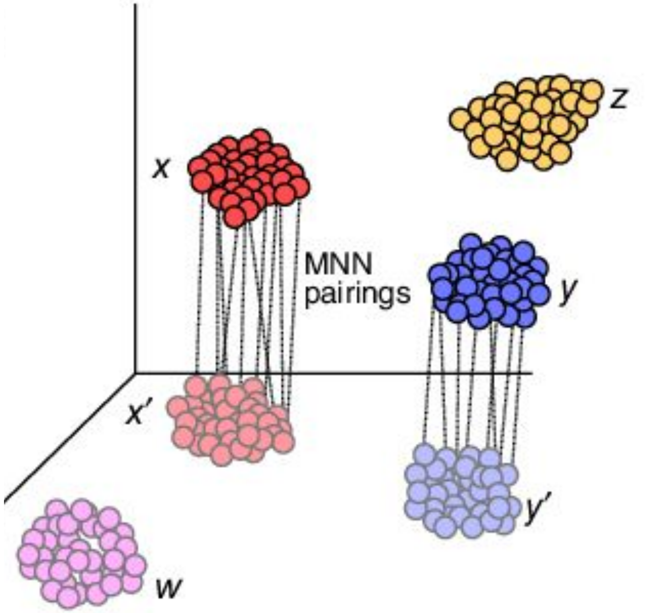
https://www.10xgenomics.com

When to integrate

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

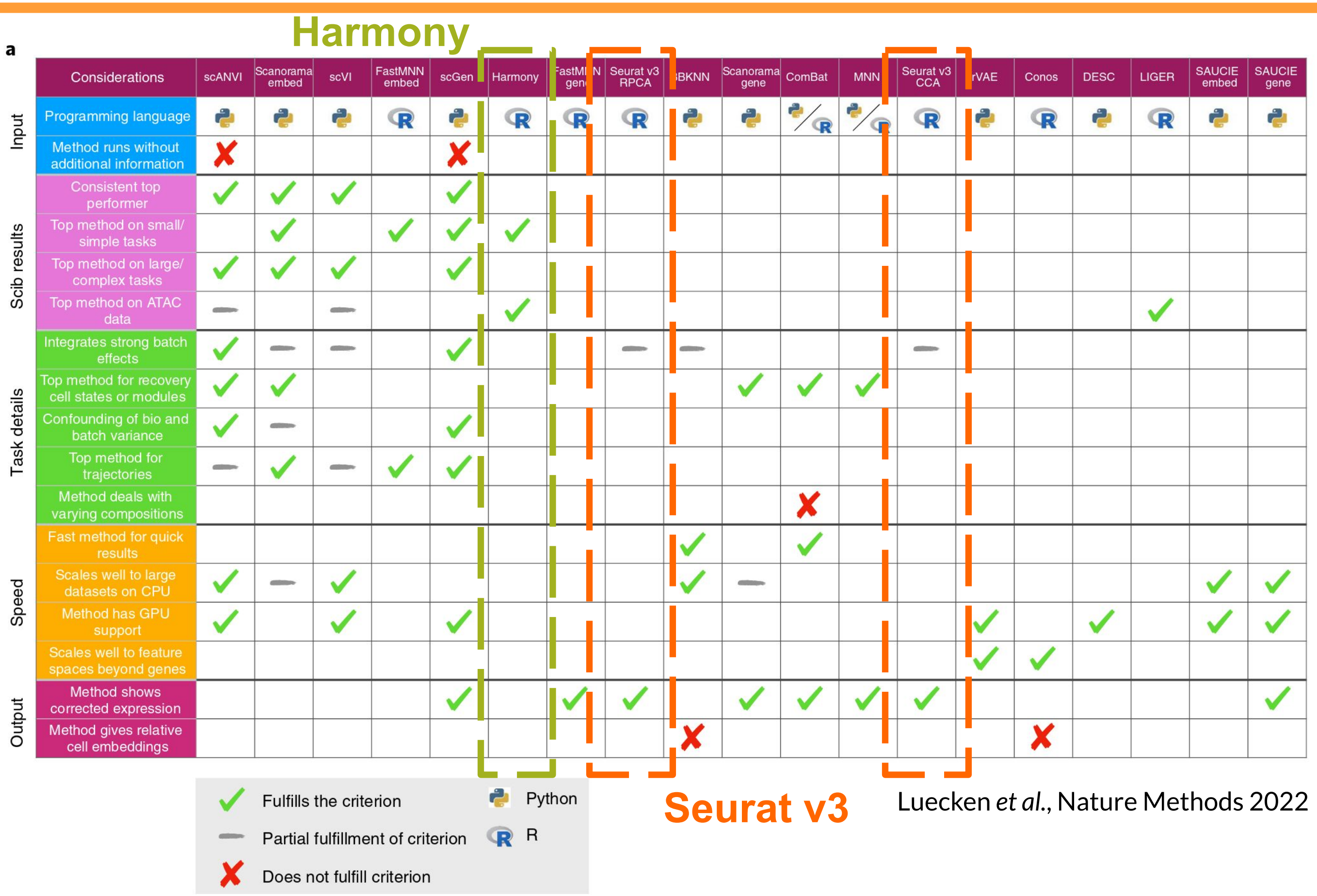
- Do not integrate otherwise:
  e.g.: replicates generated in the same time and exactly in the same manner may not need integration

- Integration corrects the data in samples to remove the batch effect and align the cell populations that are similar in each sample.



Many methods

Many methods



Harmony

a

	Considerations	scANVI	Scanorama embed	scVI	FastMNN embed	scGen	Harmony	FastMNN gene	Seurat v3 RPCA	BBKNN	Scanorama gene	ComBat	MNN	Seurat v3 CCA	trVAE	Conos	DESC	LIGER	SAUCIE embed	SAUCIE gene
Input	Programming language	Python	Python	Python	R	Python	R	R	R	Python	Python	Python / R	Python / R	R	Python	R	Python	R	Python	Python
	Method runs without additional information	✗				✗														
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											✗								
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									✗						✗				

✓ Fulfills the criterion

– Partial fulfillment of criterion

✗ 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

Many methods

Do not hesitate to test several methods

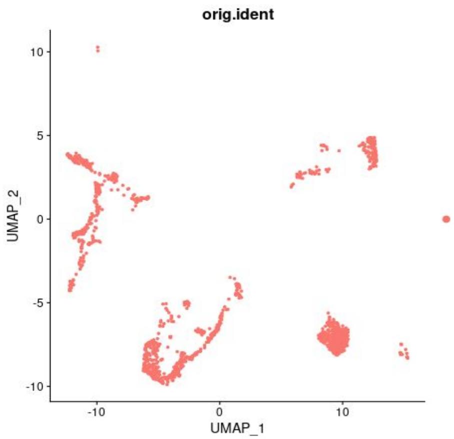
SCVI
CCA
LIGER
RPCA
Combat
SCVI
Gene
BBKNN
SAUCIE
embed
TRVAE
Seurat
DESC
FASTMNN
Conos
MNN
Scanorama
Seurat

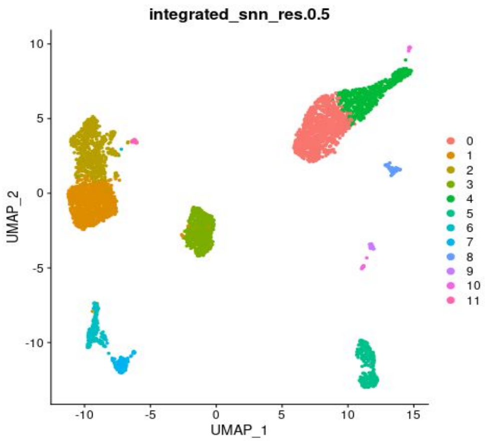
Luecken *et al.*, Nature Methods 2022

What is integration for

What is integration for

- It is intended for **visualization** and **clustering**



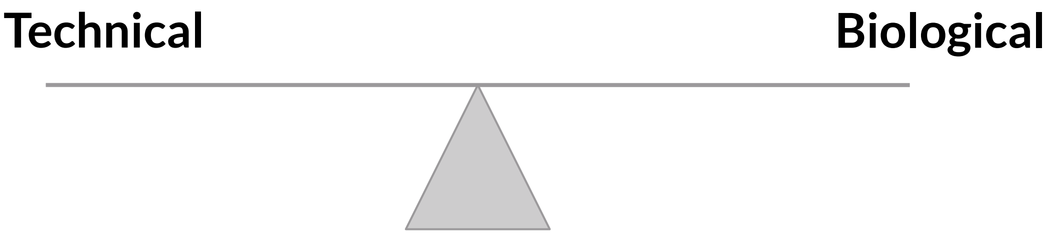


- For differential expression analysis, we go back to raw data

Stuart *et al.*, Cell 2019

Conclusion

A good integration method



Technical

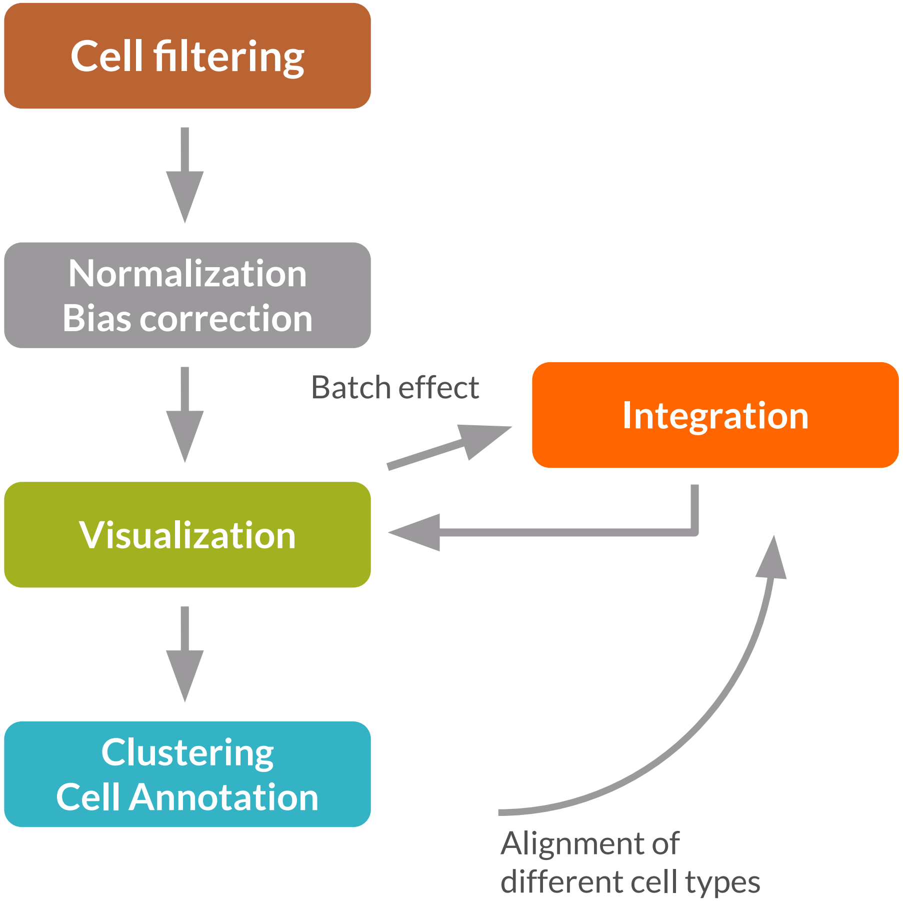
Biological

- 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...

Conclusion

Preparation of the data is not always a linear process

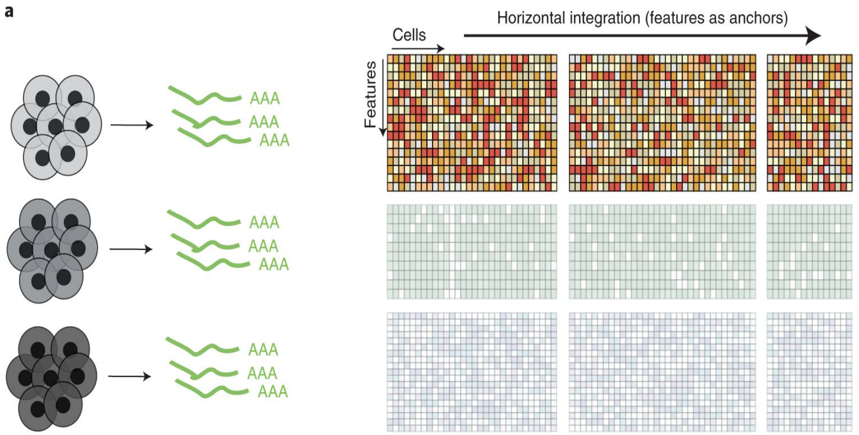


Conclusion

Different types of integrations

- Horizontal: different samples
  same modality

We saw horizontal integration



Luecken *et al.*, Nat Met 2021

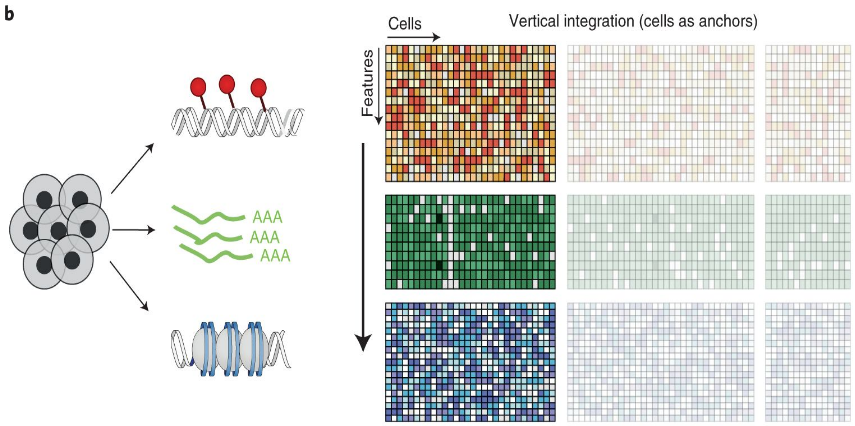
Conclusion

Different types of integrations

- Horizontal: different samples
  same modality

We saw horizontal integration

- Vertical: same sample
  different modalities
  (multiomics)



Luecken *et al.*, Nat Met 2021

Conclusion

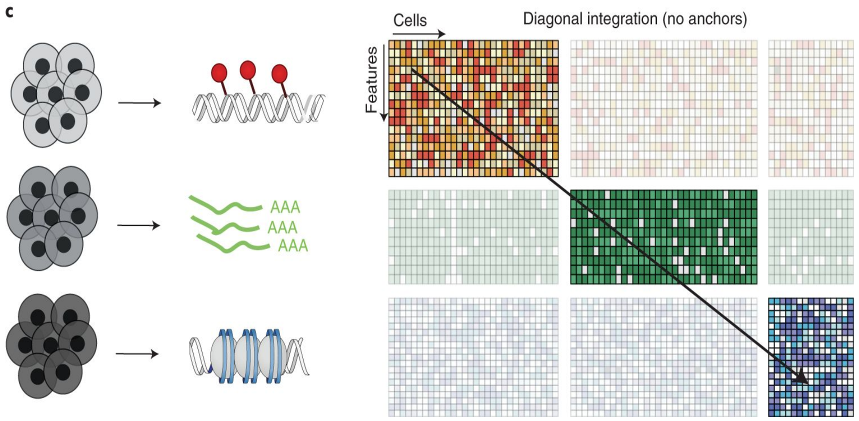
Different types of integrations

- Horizontal: different samples
  same modality

We saw horizontal integration

- Vertical: same sample
  different modalities
  (multiomics)

- Diagonal: different samples
  different modalities



Luecken *et al.*, Nat Met 2021

aviesan
alliance nationale
pour les sciences de la vie et de la santé

Acknowledgements

Parts of this course are inspired by

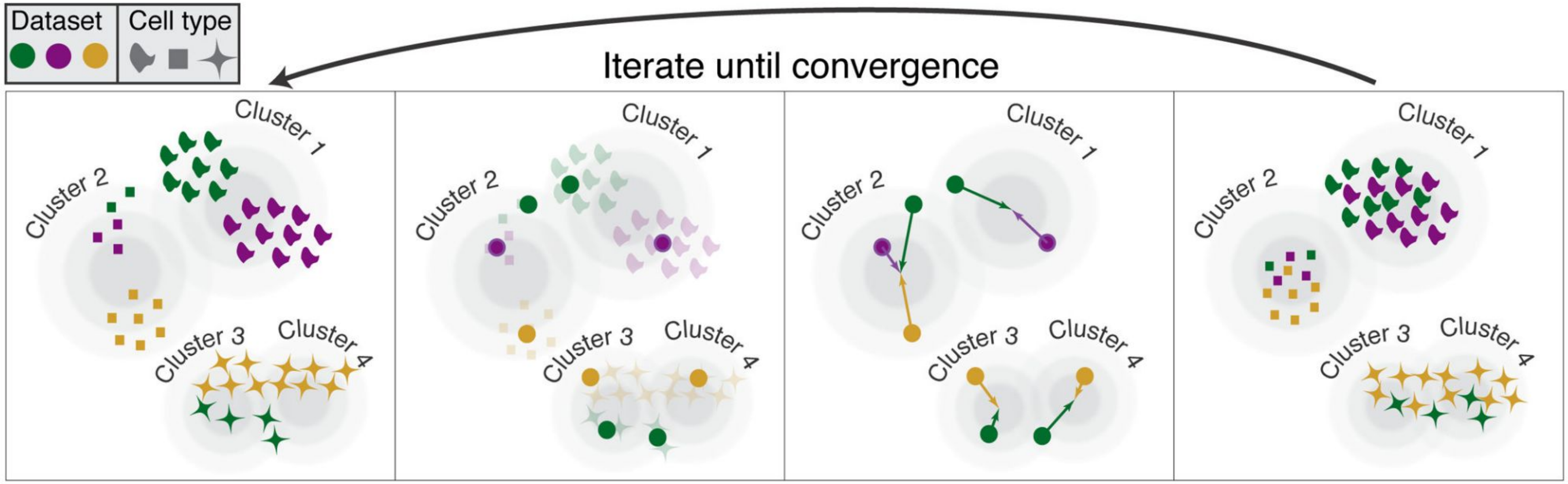
The *Swiss Institute of Bioinformatics* course Single Cell Transcriptomics

Integration with Harmony

Integration with Harmony

Many methods

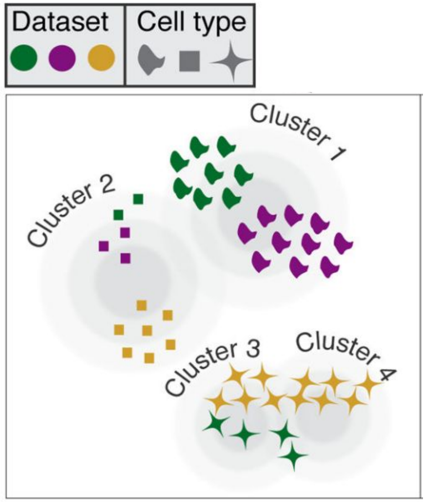
- Harmony integration: Iterative clustering in dimensionally reduced space



Korsunsky *et al.*, Nat Met 2020

Integration with Harmony

Principle

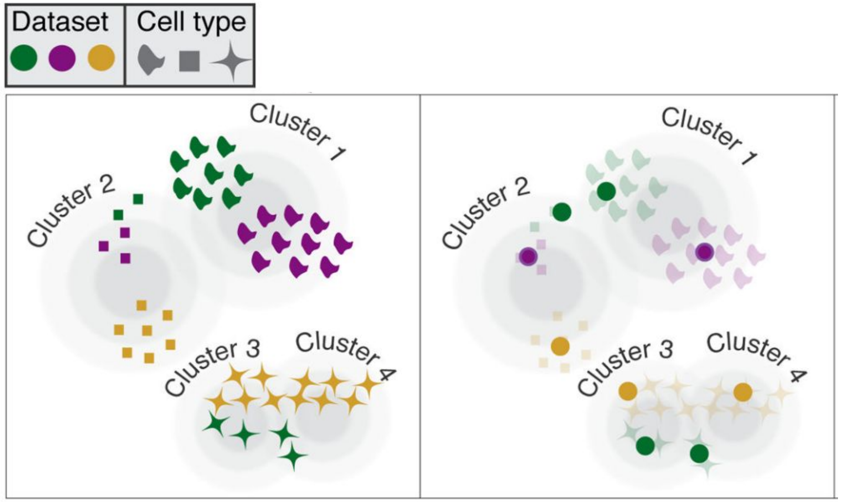


- Integration is **not** pairwise: correct all samples in the same time
- Find many small clusters
- Constraint: clusters must contain cell from several samples

Korsunsky *et al.*, Nat Met 2020

Integration with Harmony

Principle



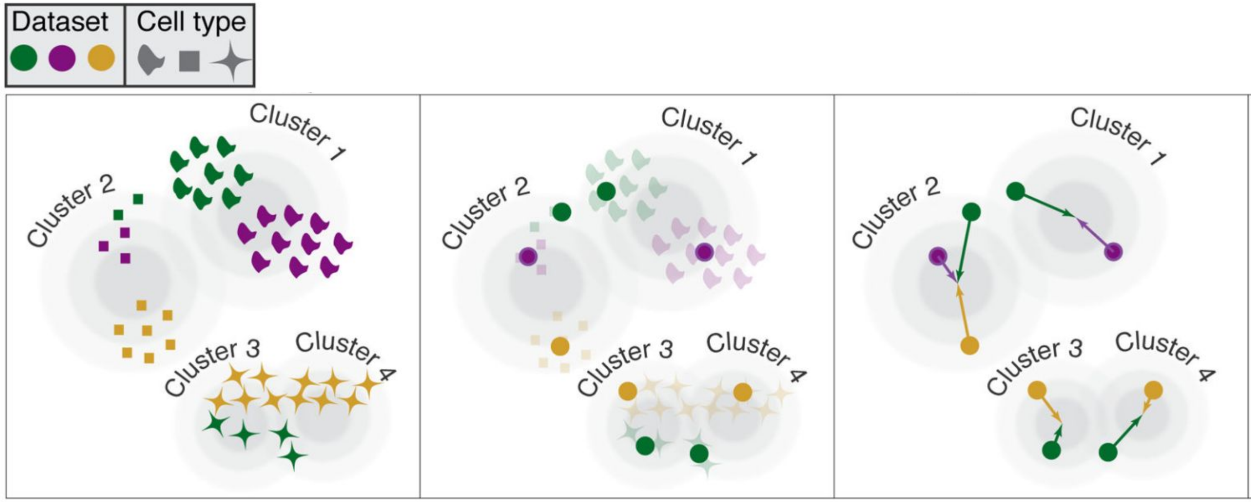
- Integration is **not** pairwise: correct all samples in the same time
- Find many small clusters
- Constraint: clusters must contain cell from several samples

- Get cluster centroids (= average position) of each sample.

Korsunsky *et al.*, Nat Met 2020

Integration with Harmony

Principle



- Integration is **not** pairwise: correct all samples in the same time
- Find many small clusters
- Constraint: clusters must contain cell from several samples

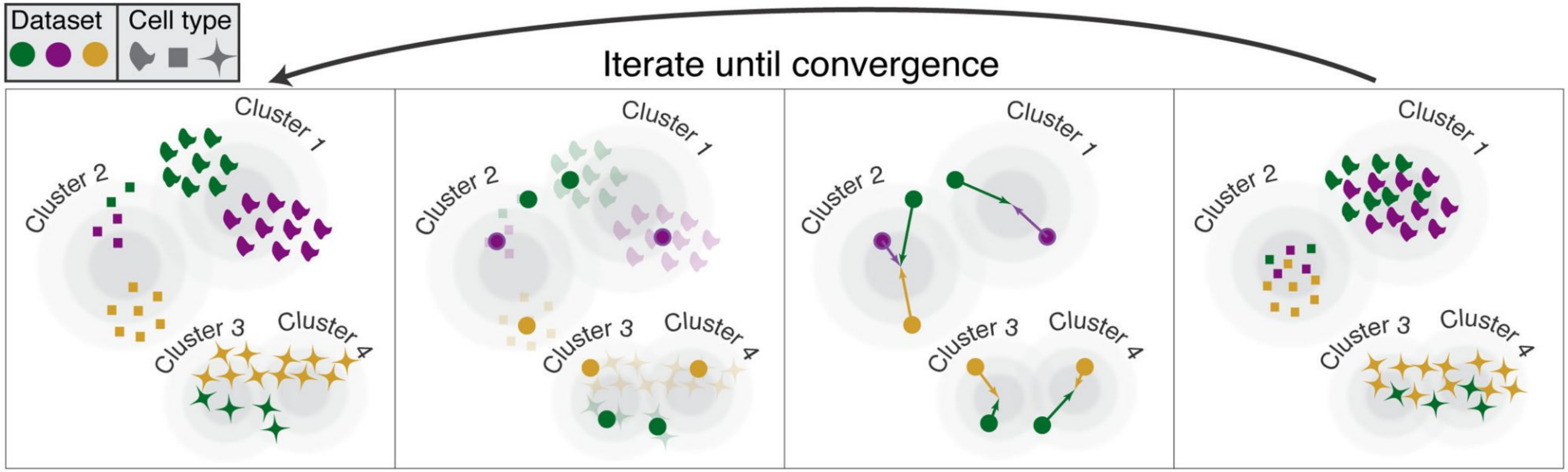
- Get cluster centroids (= average position) of each sample.

- Compute sample corrections for each cluster
- The aim is to get all centroids of the same cluster together

Korsunsky *et al.*, Nat Met 2020

Integration with Harmony

Principle



- Integration is **not** pairwise: correct all samples in the same time
- Find many small clusters
- Constraint: clusters must contain cell from several samples

- Get cluster centroids (= average position) of each sample.

- Compute sample corrections for each cluster
- The aim is to get all centroids of the same cluster together

- Apply corrections to cells

Korsunsky *et al.*, Nat Met 2020

Integration with Harmony

Principle

This methods relies on clustering but the clusters are only used for integration purpose.

They are not the clusters identified during cell characterization

not pairwise: correct all samples in the same time

- Find many small clusters
- Constraint: clusters must contain cell from several samples

centroids (= average position) of each sample.

corrections for each cluster

- The aim is to get all centroids of the same cluster together

to cells

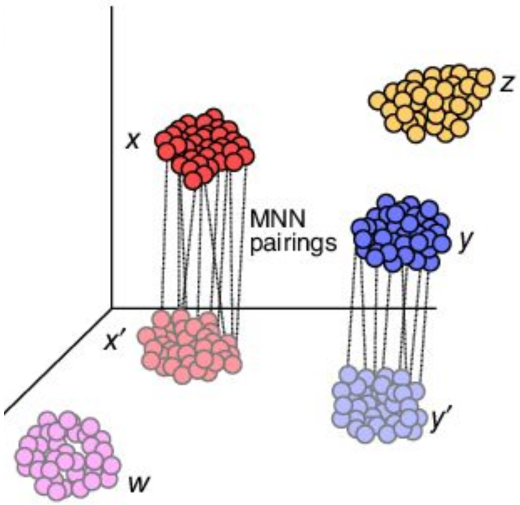
Korsunsky *et al.*, Nat Met 2020

Integration with Seurat

Integration with Seurat

Many methods

- Over 49 methods (Luecken et al., Nat Methods 2022)
- Seurat integration: group of **similarity-based** methods

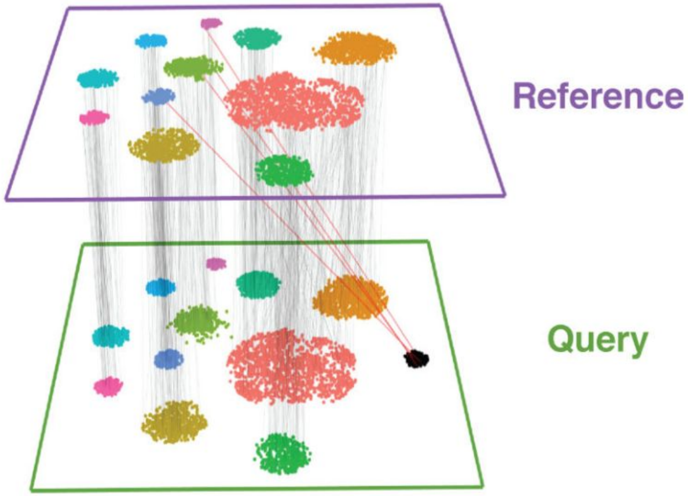


Stuart *et al.*, Cell 2019

Integration with Seurat

Principle

- Integration is always **pairwise:** correct a sample, **the query** to match the expression data of another sample, **the reference**.



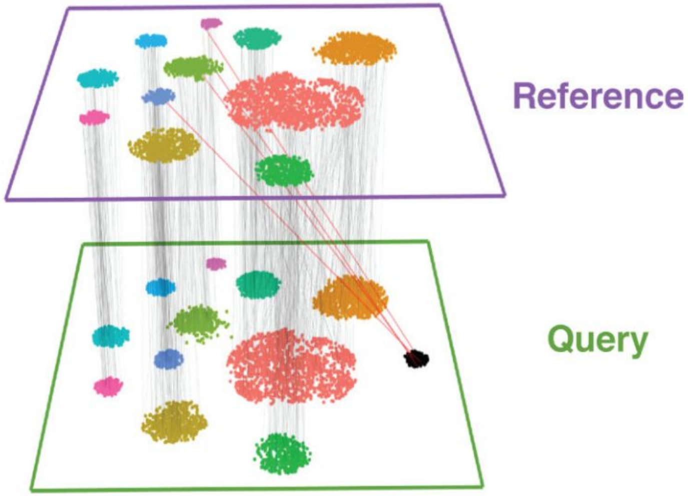
Stuart *et al.*, Cell 2019

Integration with Seurat

Principle

- Integration is always **pairwise:** correct a sample, **the query** to match the expression data of another sample, **the reference**.

- Seurat identifies pairs of close cells across both datasets, the **anchors** (Mutual Nearest Neighbors).



Stuart *et al.*, Cell 2019

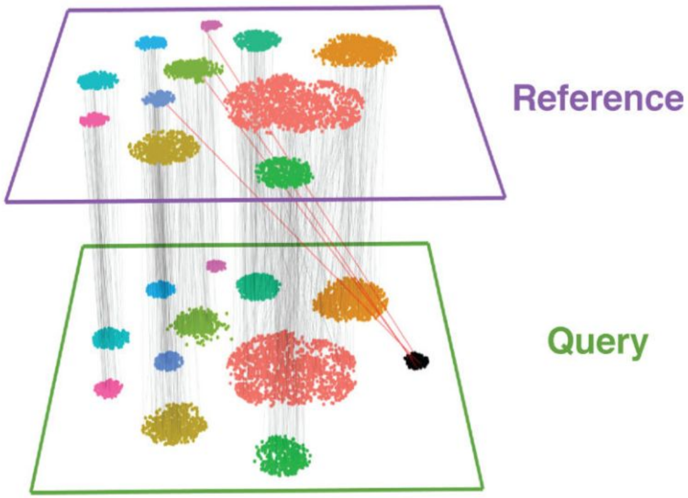
Integration with Seurat

Principle

- Integration is always **pairwise:** correct a sample, the query to match the expression data of another sample, the reference.

- Seurat identifies pairs of close cells across both datasets, the **anchors** (Mutual Nearest Neighbors).

- The difference between them is used to compute a **correction**.



Stuart *et al.*, Cell 2019

Integration with Seurat

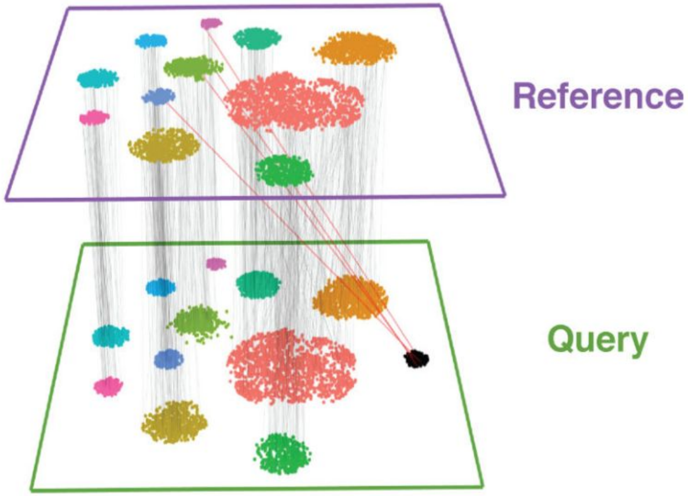
Principle

- Integration is always **pairwise:** correct a sample, the query to match the expression data of another sample, the reference.

- Seurat identifies pairs of close cells across both datasets, the **anchors** (Mutual Nearest Neighbors).

- The difference between them is used to compute a **correction**.

- The correction is used to **align** all the query cells on the reference cells.

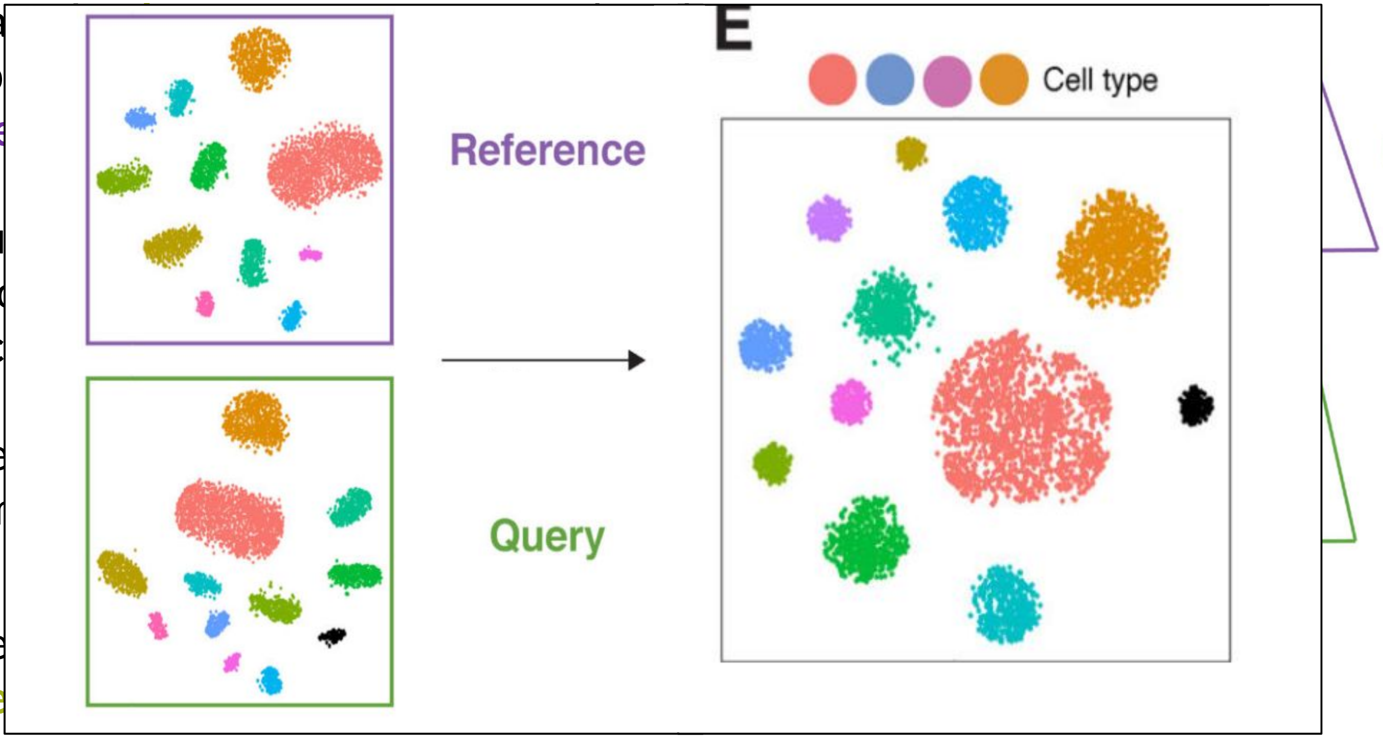


Stuart *et al.*, Cell 2019

Integration with Seurat

Principle

- Integration is always **pairwise:** correct
  a sa
  exp
  refe
- Seu
  acr
  anc
- The
  com
- The
  que



Reference

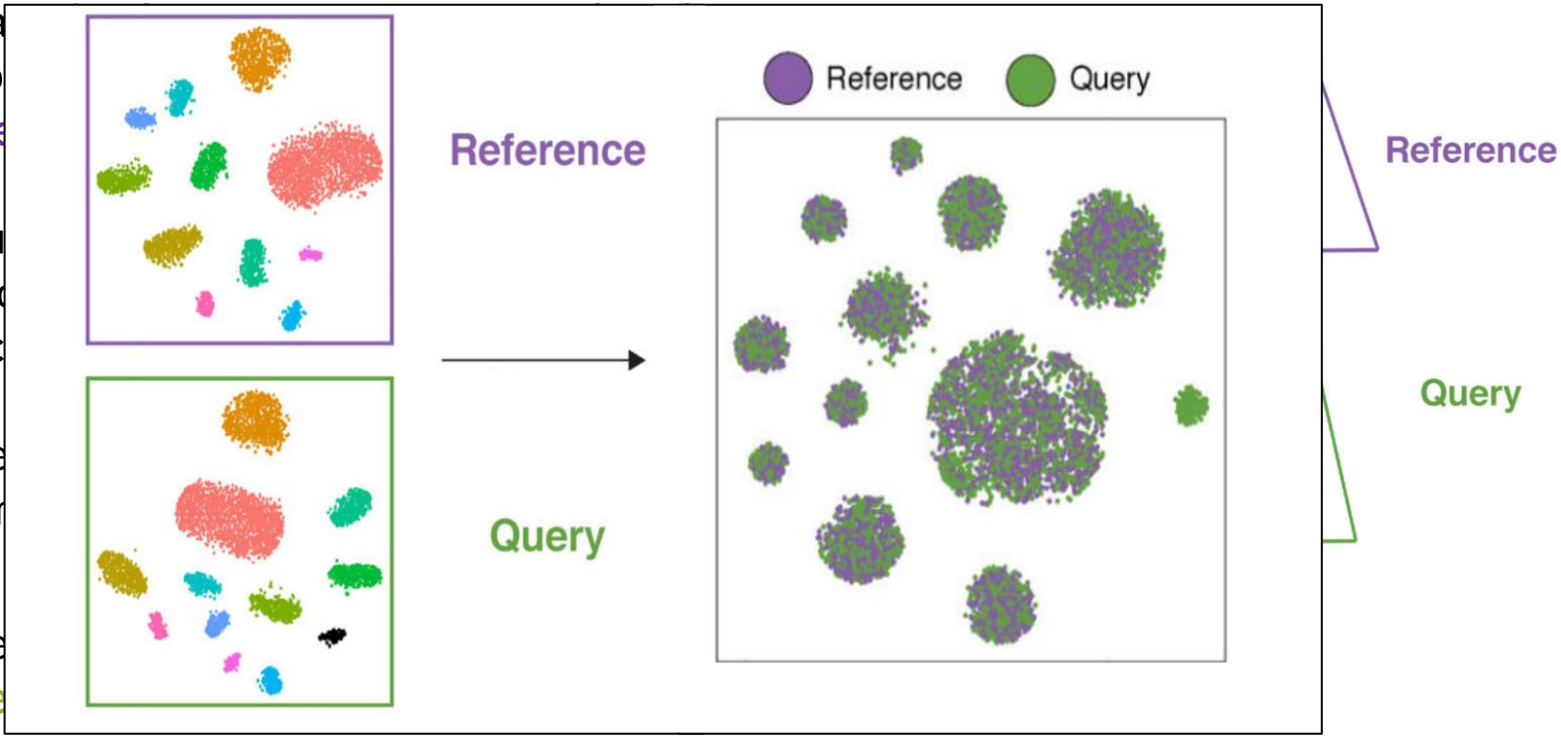
Query

Stuart *et al.*, Cell 2019

Integration with Seurat

Principle

- Integration is always **pairwise:** correct
  a sa
  exp
  refe
- Seu
  acro
  anc
- The
  com
- The
  que



Stuart *et al.*, Cell 2019