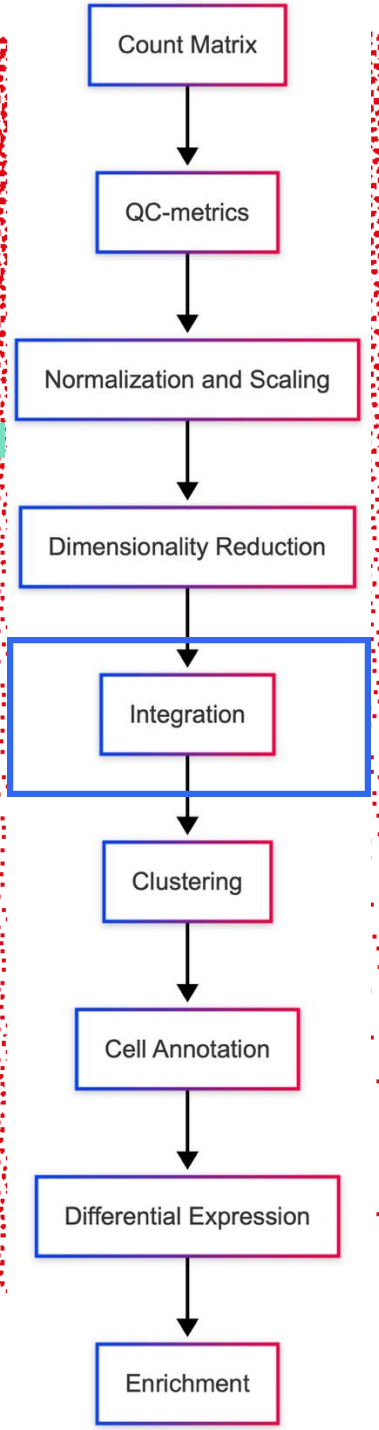
Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Integration analysis



Deepak Tanwar

July 02-04, 2025

Adapted from previous year courses

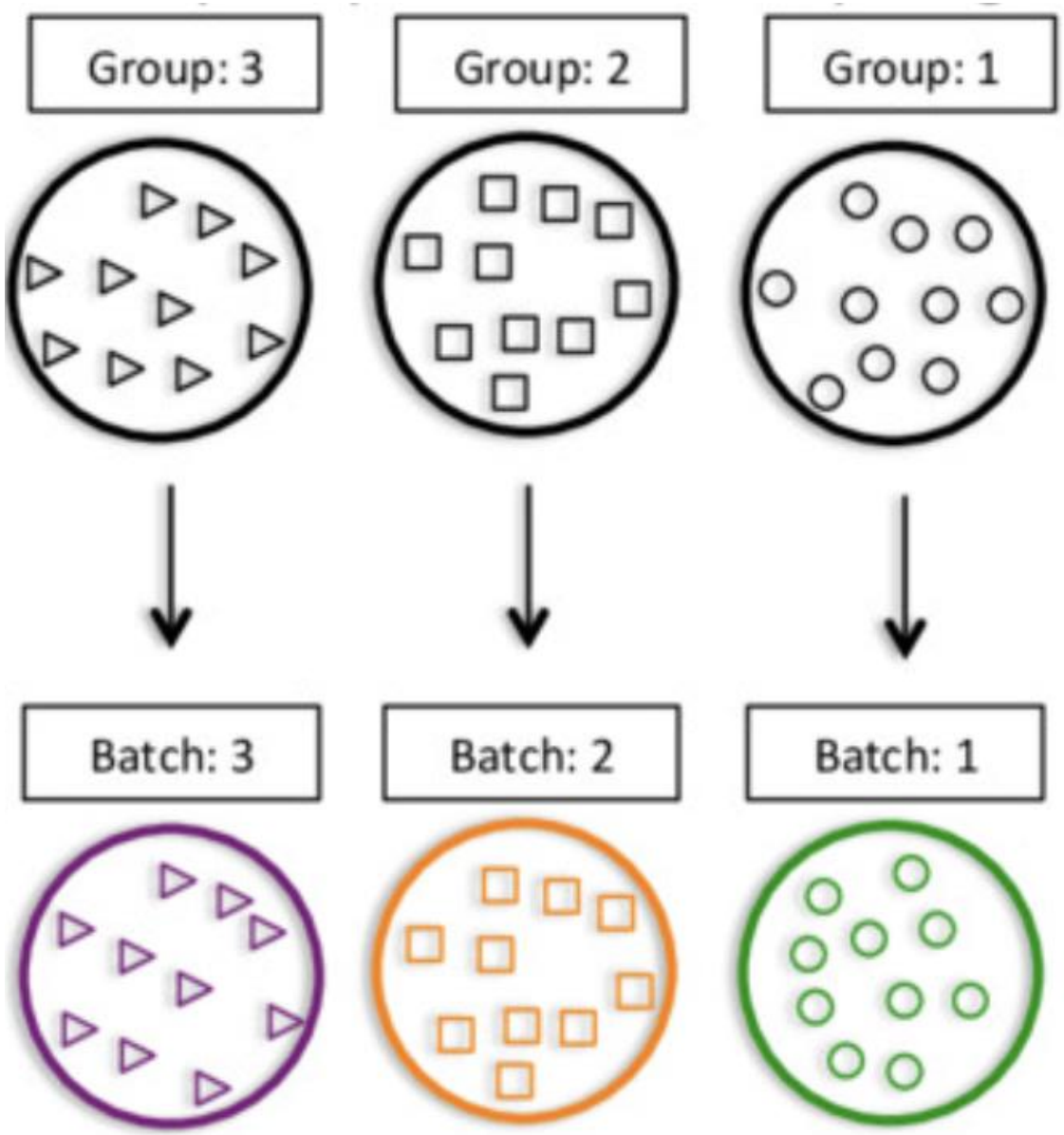
Learning objectives

Understand the importance of experimental design

Identify scenarios where integration is necessary for data analysis

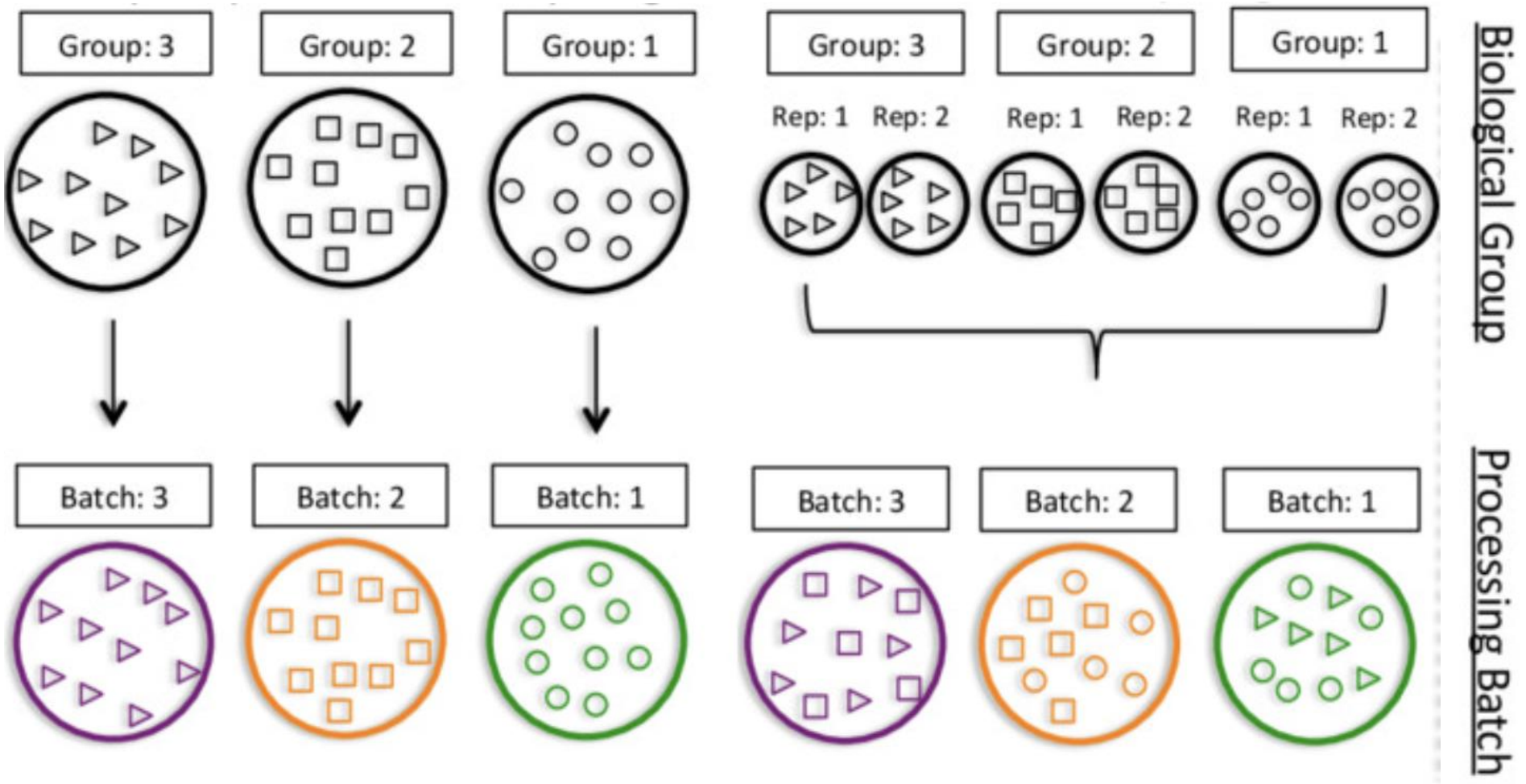
Apply canonical correlation analysis (CCA) for integrating datasets

Experimental design matters

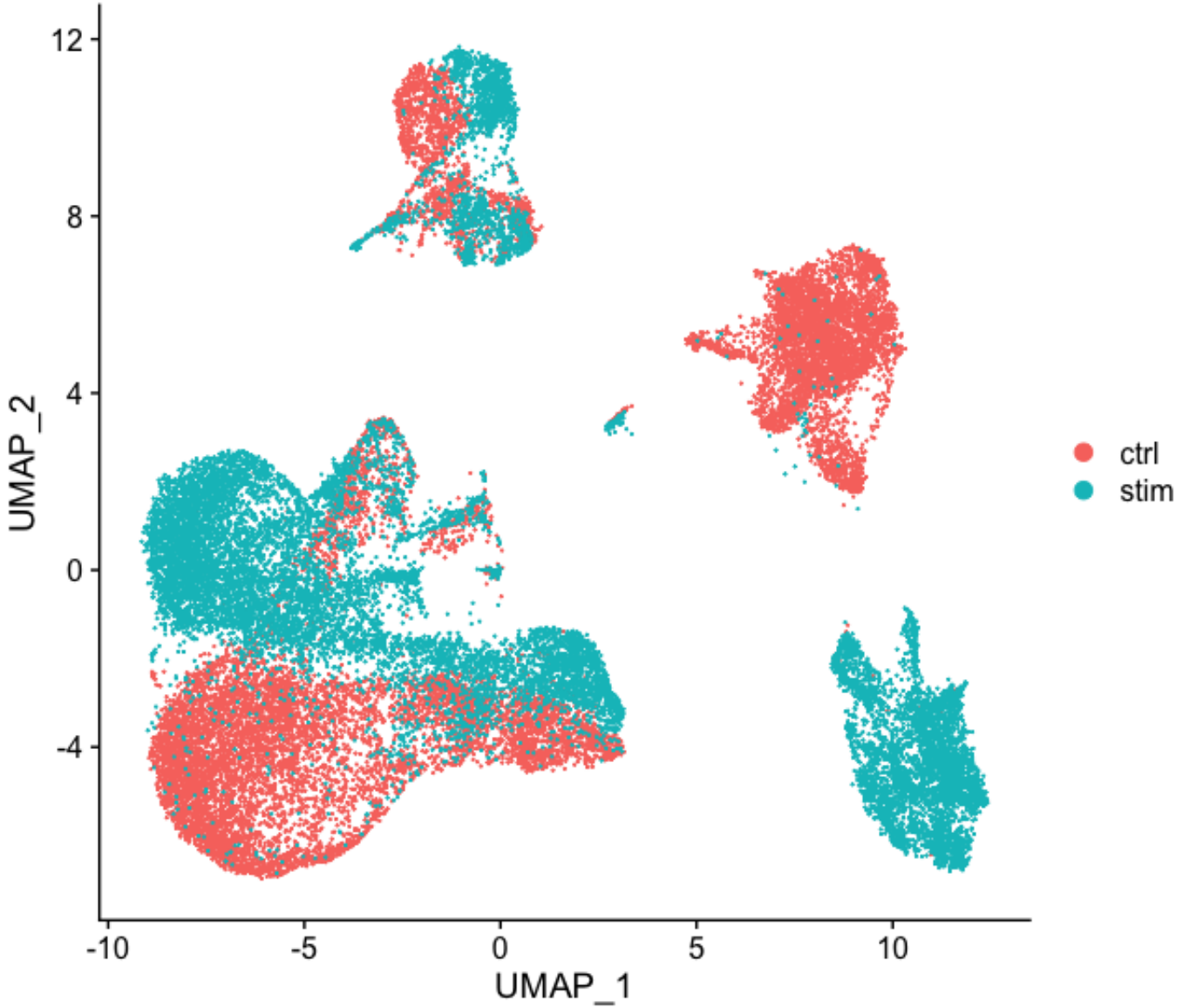


What changes would you make here to make the experimental design more optimal?

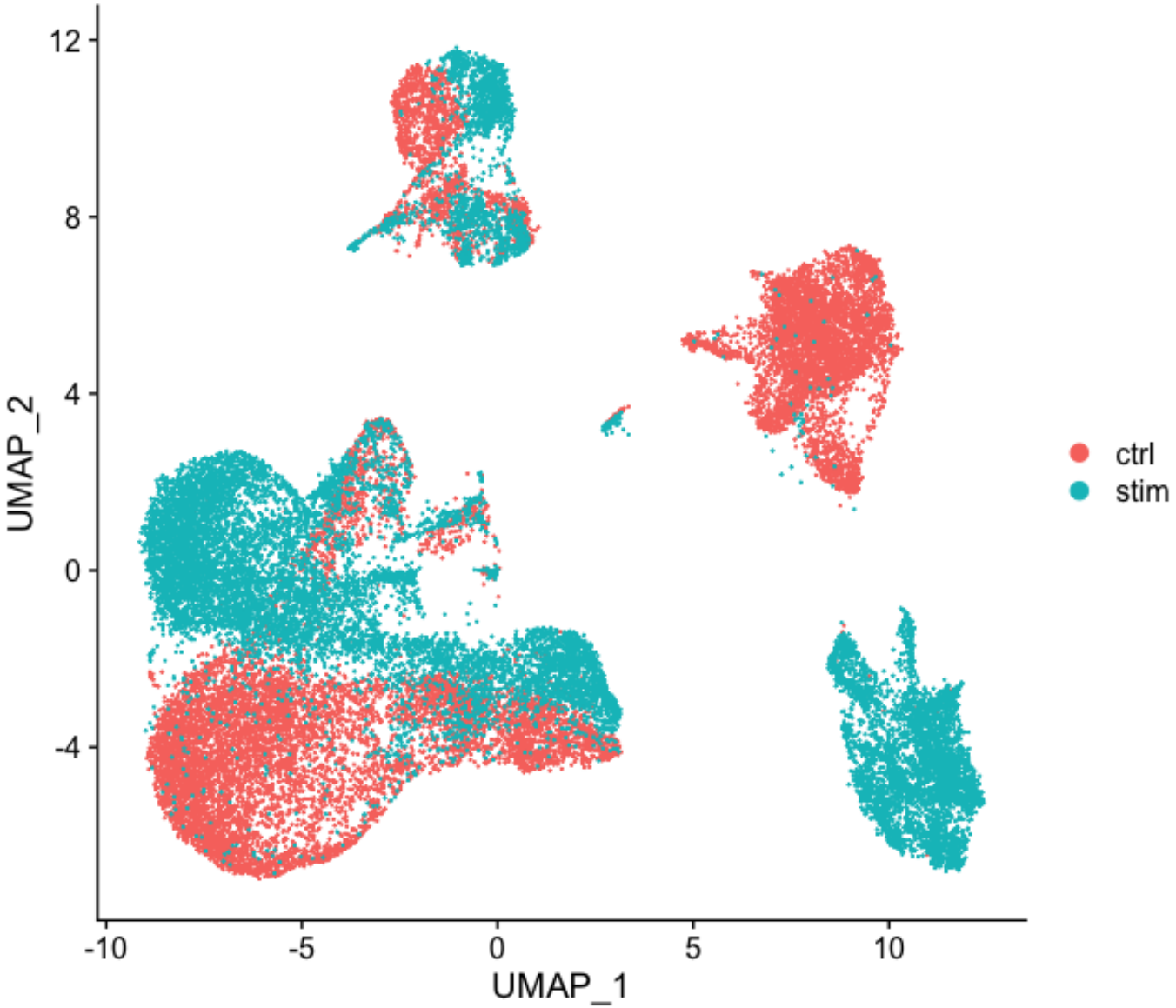
Experimental design matters



Exercise: Identify problem in this plot



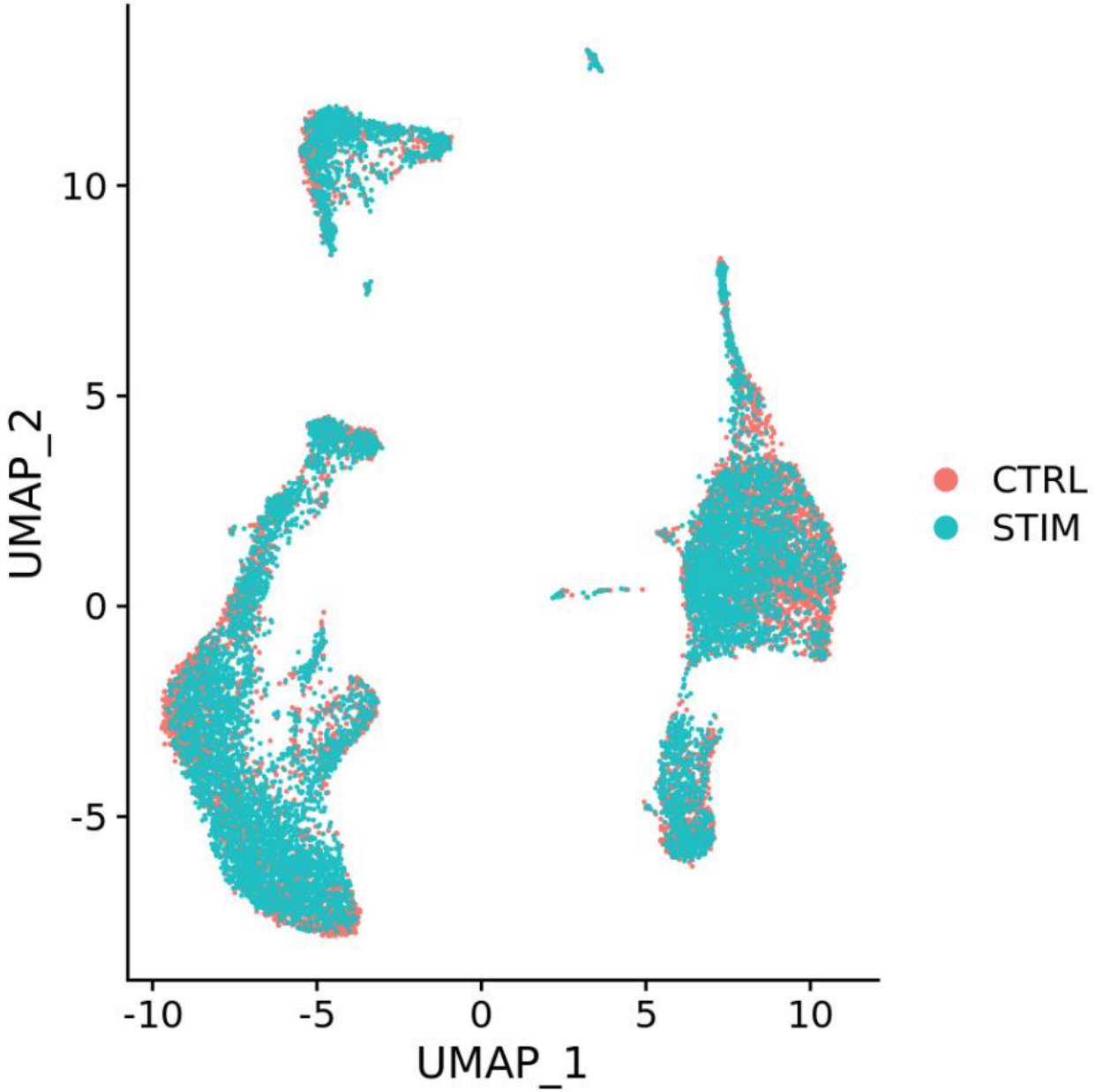
Exercise: Identify problem in this plot

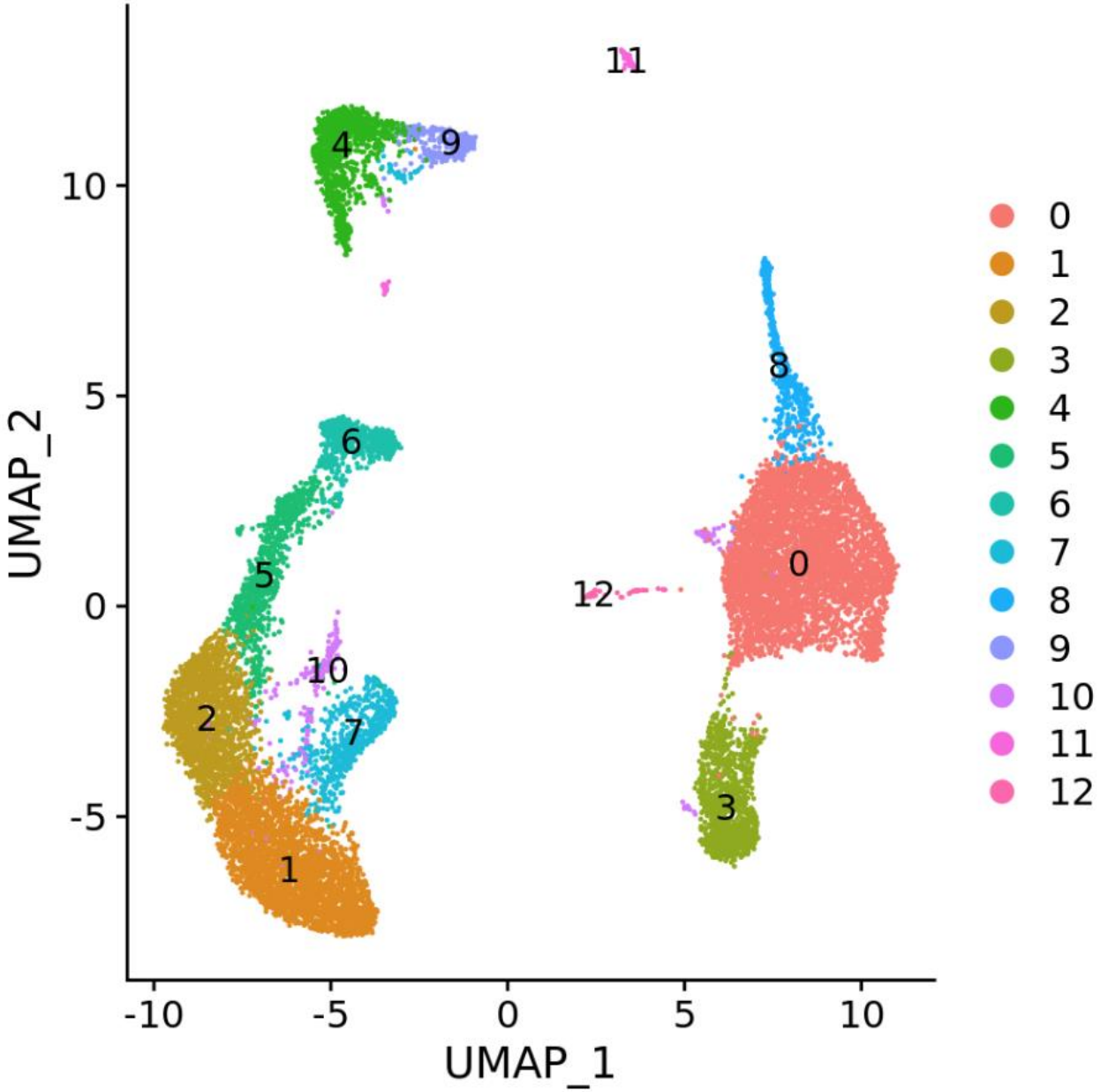


- **Explore the data:** do not just always perform integration because you think there might be differences

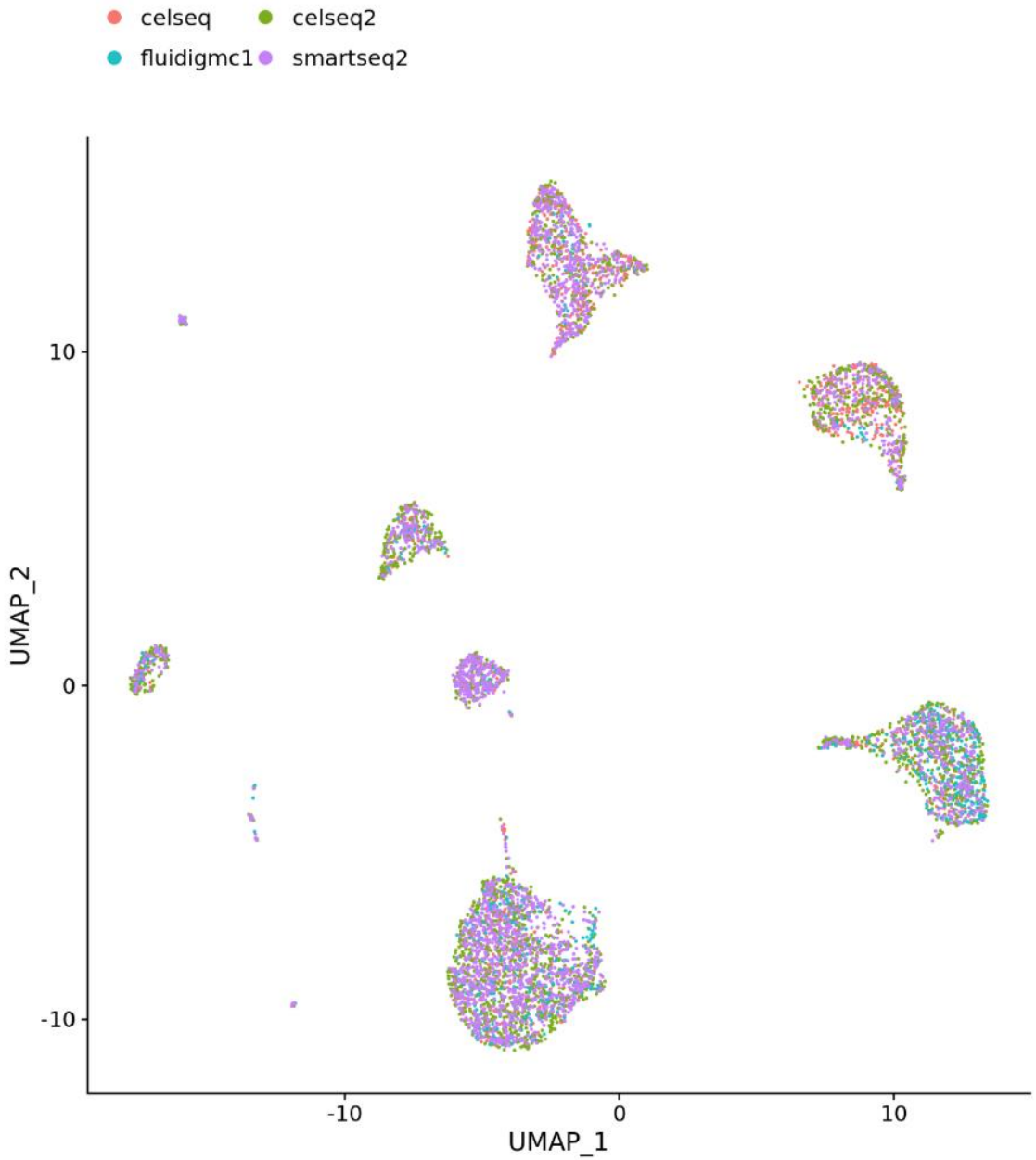
- If cells cluster by **sample, condition, batch, dataset, modality**, performing integration can help align cells across the groups to greatly improve the clustering and the downstream analyses.

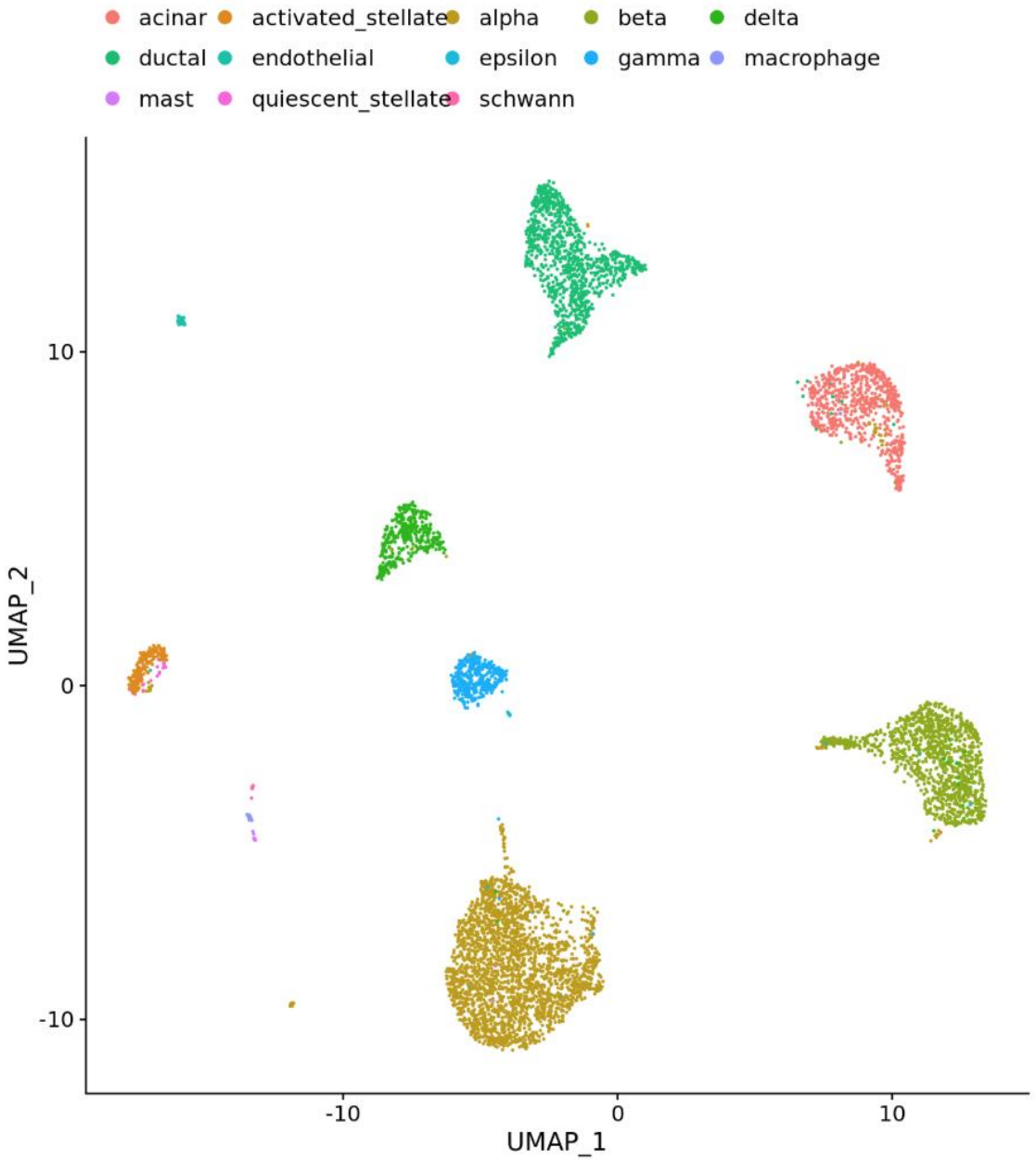
Example scenarios for integration: **conditions**





Example scenarios for integration: **datasets**





Relationship between growth of leaves and roots (not a simple one-to-one relationship)

Leaf length
Leaf width
Number of veins

Root depth
Root branching (how many branches)
Root mass

Relationship between growth of leaves and roots (not a simple one-to-one relationship)

Leaf length
Leaf width
Number of veins

Root depth
Root branching (how many branches)
Root mass

Goal: to find "best combined leaf characteristic" and the "best combined root characteristic" that are as strongly related as possible

Relationship between growth of leaves and roots (not a simple one-to-one relationship)

Leaf length
Leaf width
Number of veins

Root depth
Root branching (how many branches)
Root mass

Goal: to find "best combined leaf characteristic" and the "best combined root characteristic" that are as strongly related as possible

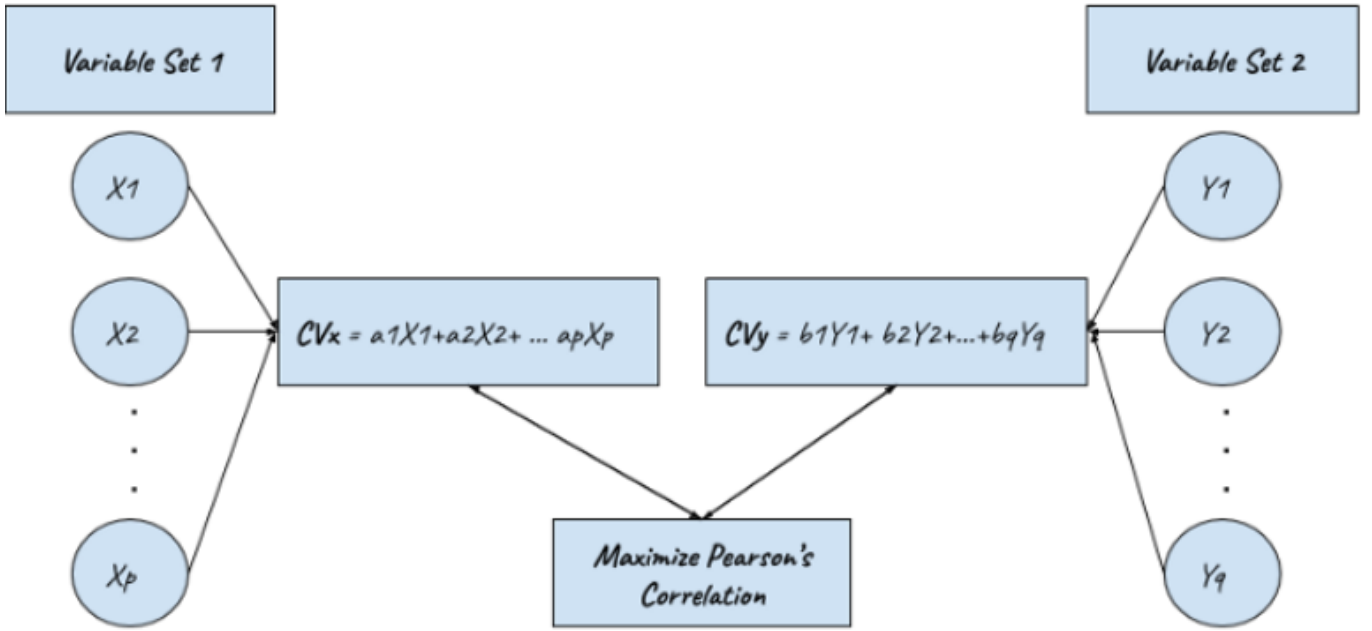
Approach: looks at all leaf measurements and figures out a special way to combine them into a single ”**Leaf Score**” and “**Root Score**” for each plant

Relationship between growth of leaves and roots (not a simple one-to-one relationship)

Leaf length
Leaf width
Number of veins

Root depth
Root branching (how many branches)
Root mass

Method: Canonical Correlation Analysis



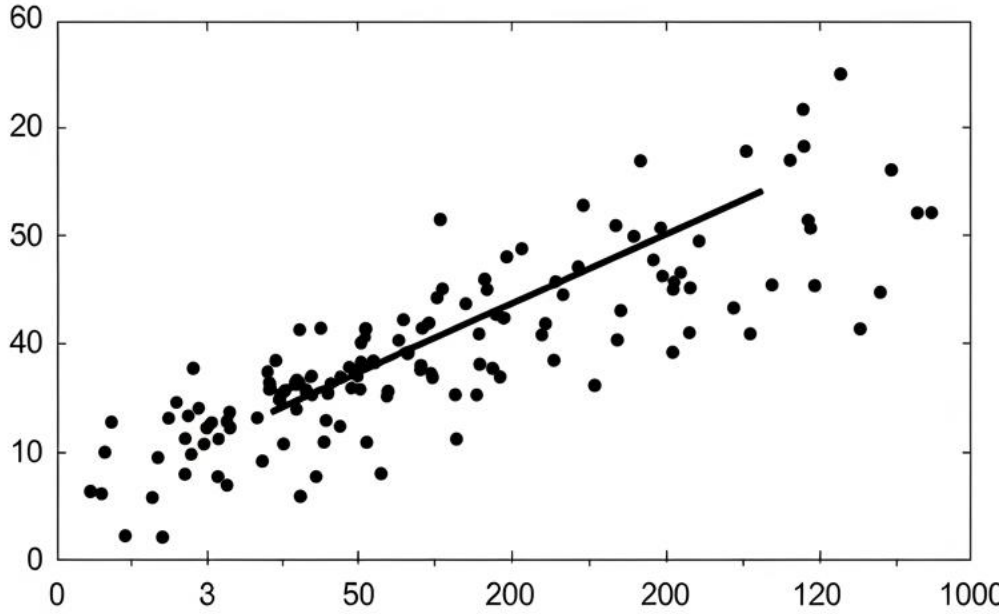
adding one more constraint: each new variate should be orthogonal and

Relationship between growth of leaves and roots (not a simple one-to-one relationship)

Leaf length
Leaf width
Number of veins

Root depth
Root branching (how many branches)
Root mass

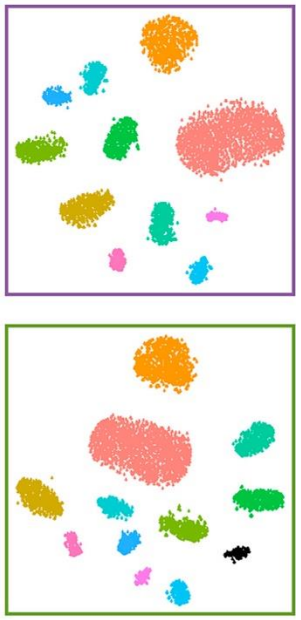
Method: Canonical Correlation Analysis



Integration using CCA: canonical correlation analysis

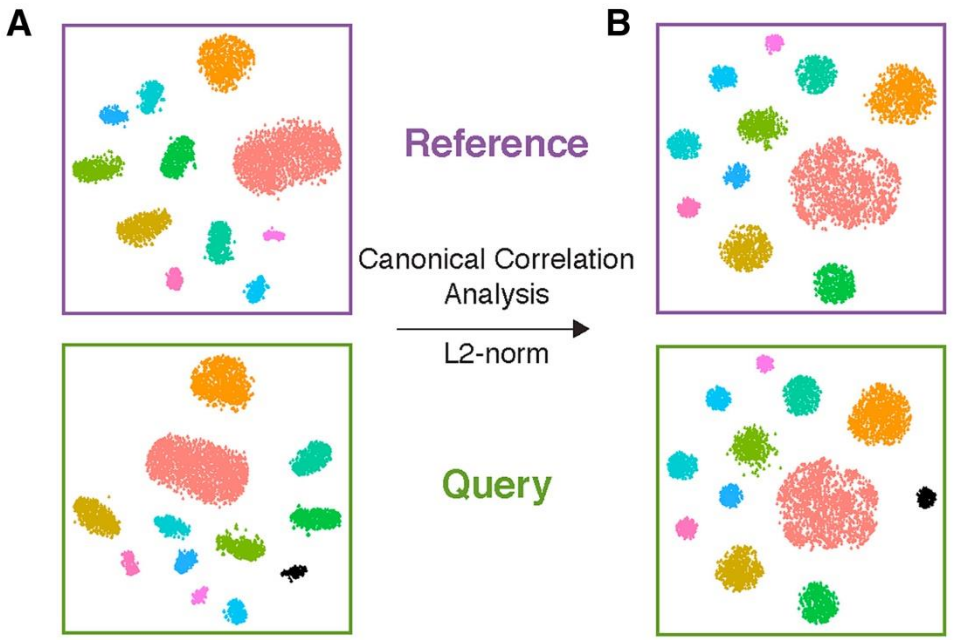
Datasets coming from 2 different platforms

A



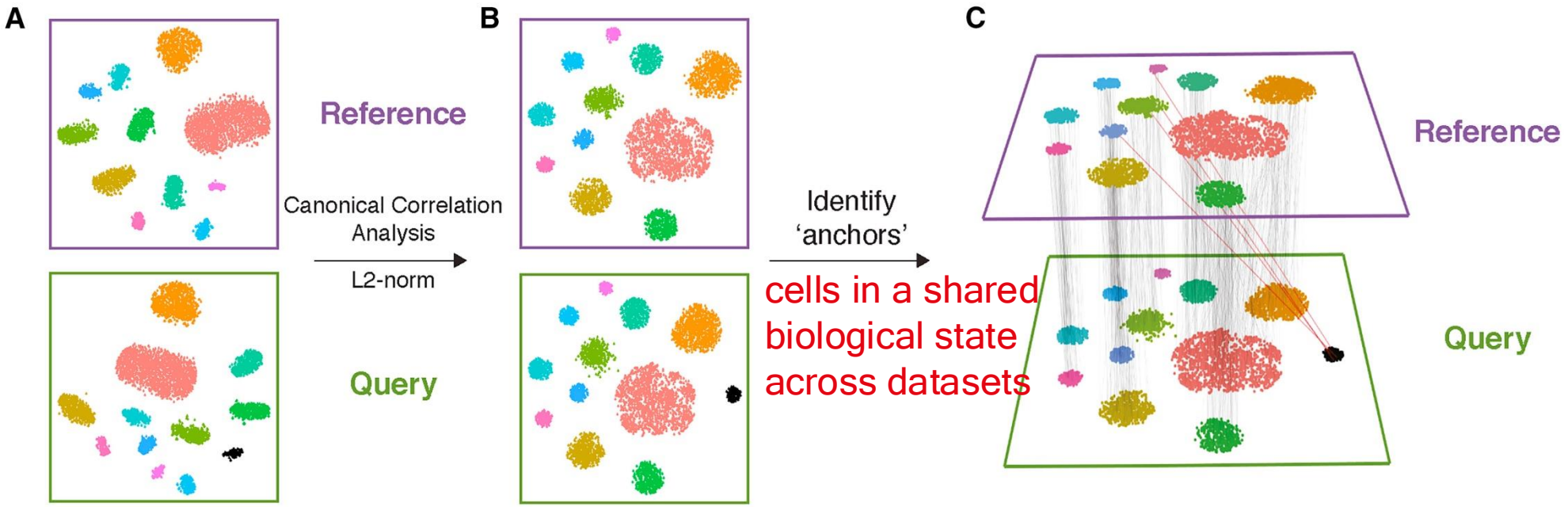
Integration using CCA: canonical correlation analysis

Datasets coming from 2 different platforms



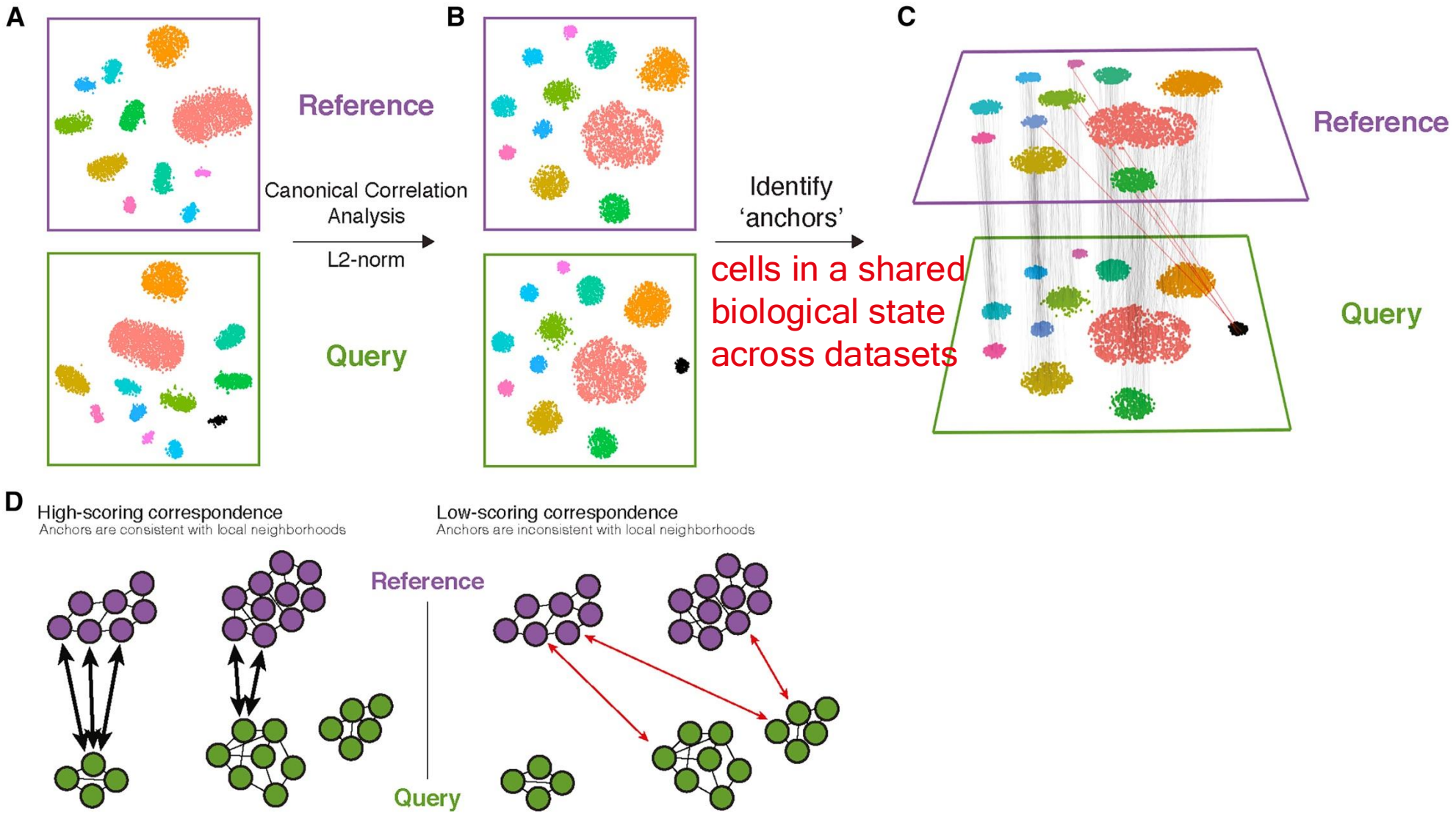
Integration using CCA: canonical correlation analysis

Datasets coming from 2 different platforms



Integration using CCA: canonical correlation analysis

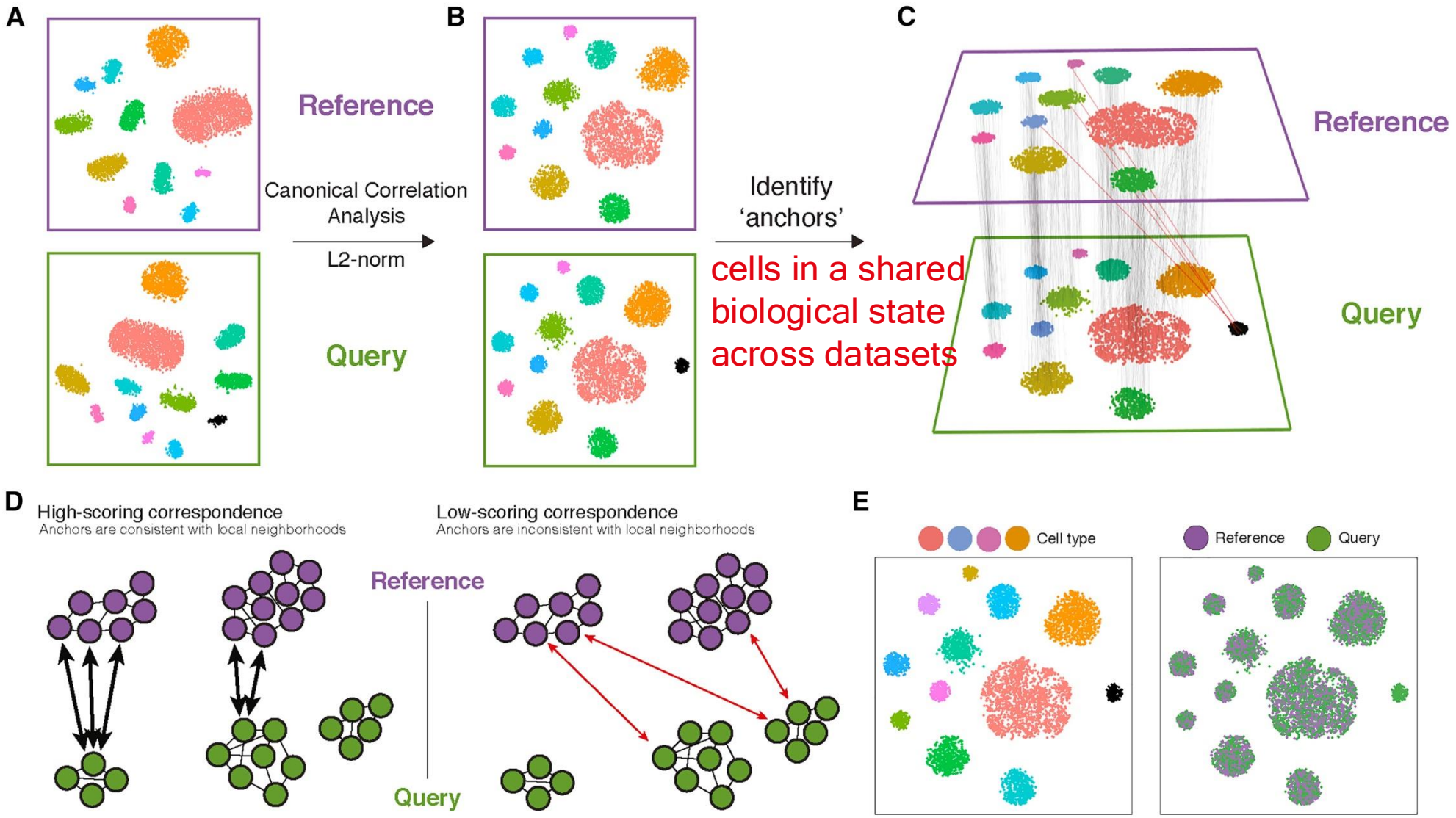
Datasets coming from 2 different platforms



scores to compute "correction" vectors for each query cell, transforming its expression so it can be jointly analyzed

Integration using CCA: canonical correlation analysis

Datasets coming from 2 different platforms



scores to compute "correction" vectors for each query cell, transforming its expression so it can be jointly analyzed

Is CCA in Seurat really a CCA?

The "Seurat CCA" is taking the projection vector from the traditional CCA directly as cell embeddings. But in fact, **the classical definition of CCA would imply projecting genes into a common space rather than cells.**

Based on our understanding, the math behind the "Seurat CCA" algorithm is technically closer to a dual PCA formulation.

In the original paper, the assumption that the covariance matrix of gene expression is diagonal, is not necessary.

Furthermore, considering the formulation to preserve the most similarity (dual PCA), the low-dimensional cell embeddings should multiply the singular value, which is currently missing in the "Seurat CCA" algorithm.

And finally, there is an intrinsic connection between MNN and "Seurat CCA" (extended dual PCA).

https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

Quiz

In which condition will you perform integration?

A) When cells cluster by sample
B) When cells cluster by condition
C) When cells cluster by batch
D) When cells cluster by dataset
E) None of the above
F) All of the above

Summary

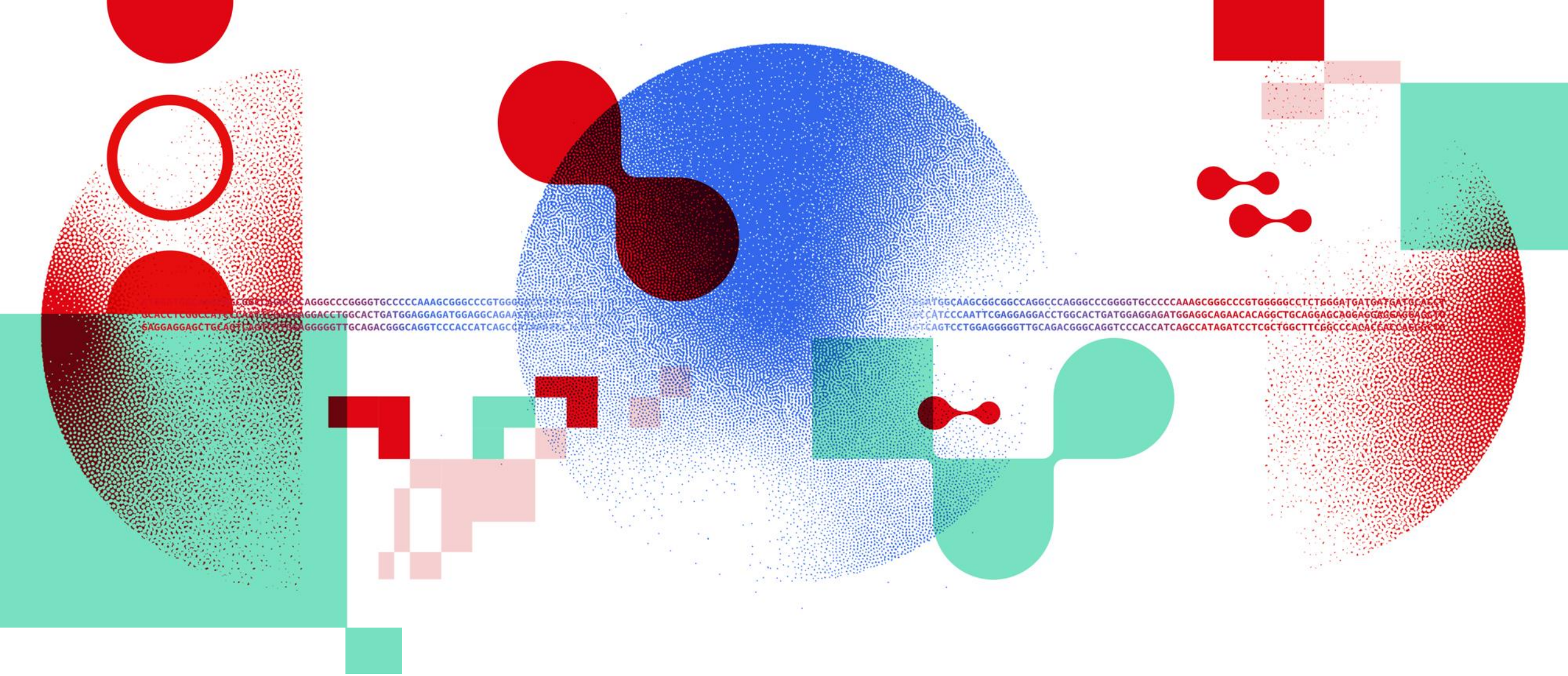
Experimental Design Matters: Optimize design to improve data quality and analysis

Integration Scenarios:

- Conditions: Compare different experimental conditions
- Datasets: Analyze data from different platforms together

Integration Using CCA:

- Align cells across groups to improve clustering and downstream analyses
- Compute correction vectors for each query cell to transform its expression for joint analysis



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss