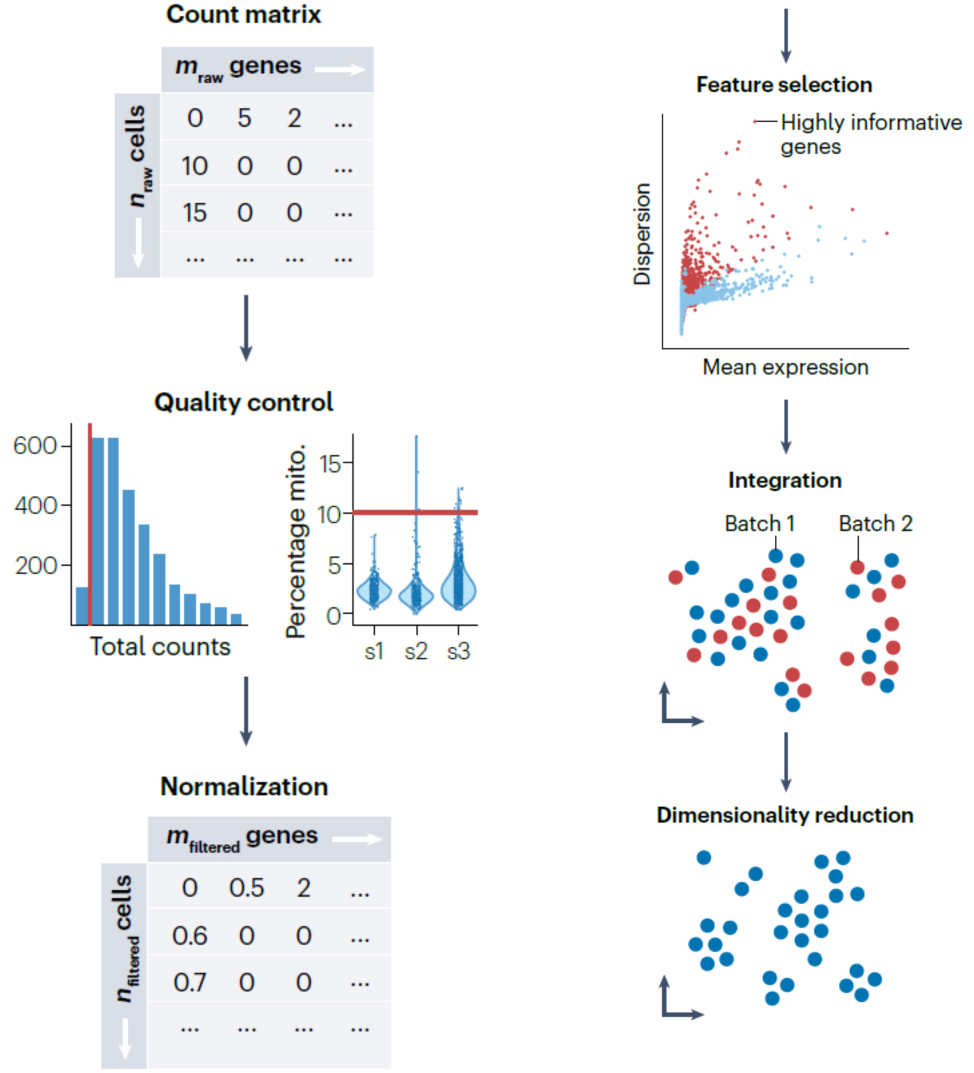
# scRNAseq course

**Dimensionality reduction** and batch effect correction

# A general single cell analysis workflow

| | |
|---|---|
| Quality control | • Gene counts<br>• UMI counts<br>• % mito gene, % ribo genes |
| Normalization | • Normalization<br>• Scaling |
| Dimensionality reduction | • HVG selection<br>• PCA and PCs selection<br>• UMAP or t-SNE |
| Data correction and integration | • Data integration<br>• Batch effect detection and correction |
| Clustering | • Clustering<br>• Clusters' gene markers identification<br>• Clusters annotation |



# Normalization

# Normalization

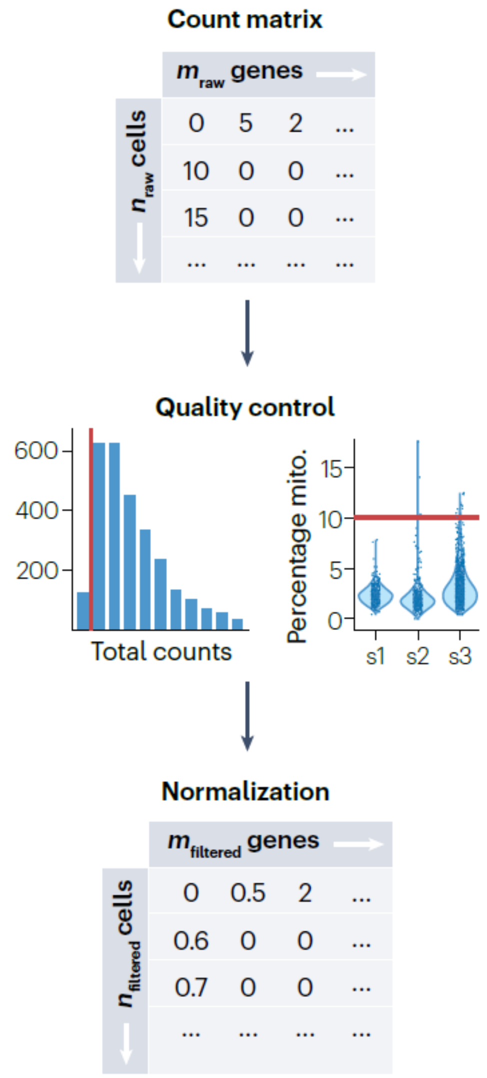
Heterogeneity in single-cell RNA sequencing data is known to be influenced by technical factors.
In some cases, these technical factors may confound our ability to measure true biological variation between samples, making it more challenging to address the research question at hand.

One of these confounding factors is sequencing depth.

Technical noise

- Low mRNA content per cell
- Variable mRNA capture
- Variable sequencing depth

**Count normalization** makes cellular profiles comparable.



Count matrix

| | $m_{raw}$ genes → | | | |
|---|---|---|---|---|
| $n_{raw}$ cells ↓ | 0 | 5 | 2 | ... |
| | 10 | 0 | 0 | ... |
| | 15 | 0 | 0 | ... |
| | ... | ... | ... | ... |

Normalization

| | $m_{filtered}$ genes → | | | |
|---|---|---|---|---|
| $n_{filtered}$ cells ↓ | 0 | 0.5 | 2 | ... |
| | 0.6 | 0 | 0 | ... |
| | 0.7 | 0 | 0 | ... |
| | ... | ... | ... | ... |

# Normalization

Main approaches

1. **log-normalize**
   - normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default).
   - After normalization, data matrices are log(x+1) transformed.

2. **SCTrasfrom**
   - UMI counts across cells in a dataset are modeled using a **generalized linear model (GLM).**
   - This method first fits independent regression models per gene.
   - Then it uses the relationship between model parameter values and gene mean to learn global trends in the data (combining information across genes) to perform regularization for all parameters.
   - Given the fitted model parameters, it transforms each observed UMI count into a **Pearson residual** which can be interpreted as the number of standard deviations an observed count is away from the expected mean.
     - Normalized expression values are Pearson residuals from regularized negative binomial regression.
     - Positive residual for a given gene in a given cell indicate that we observed more UMIs than expected given the gene’s average expression in the population and the cellular sequencing depth.

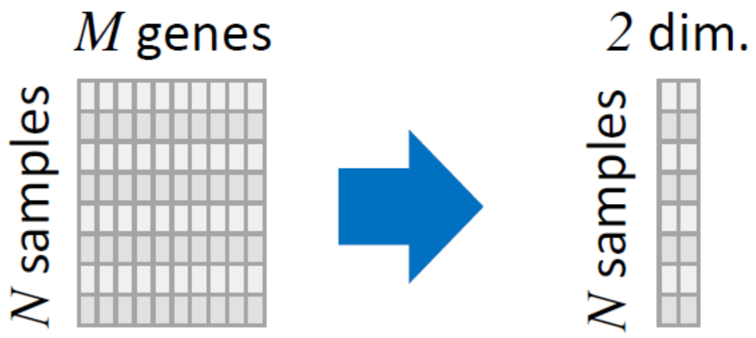
# Dimensionality Reduction

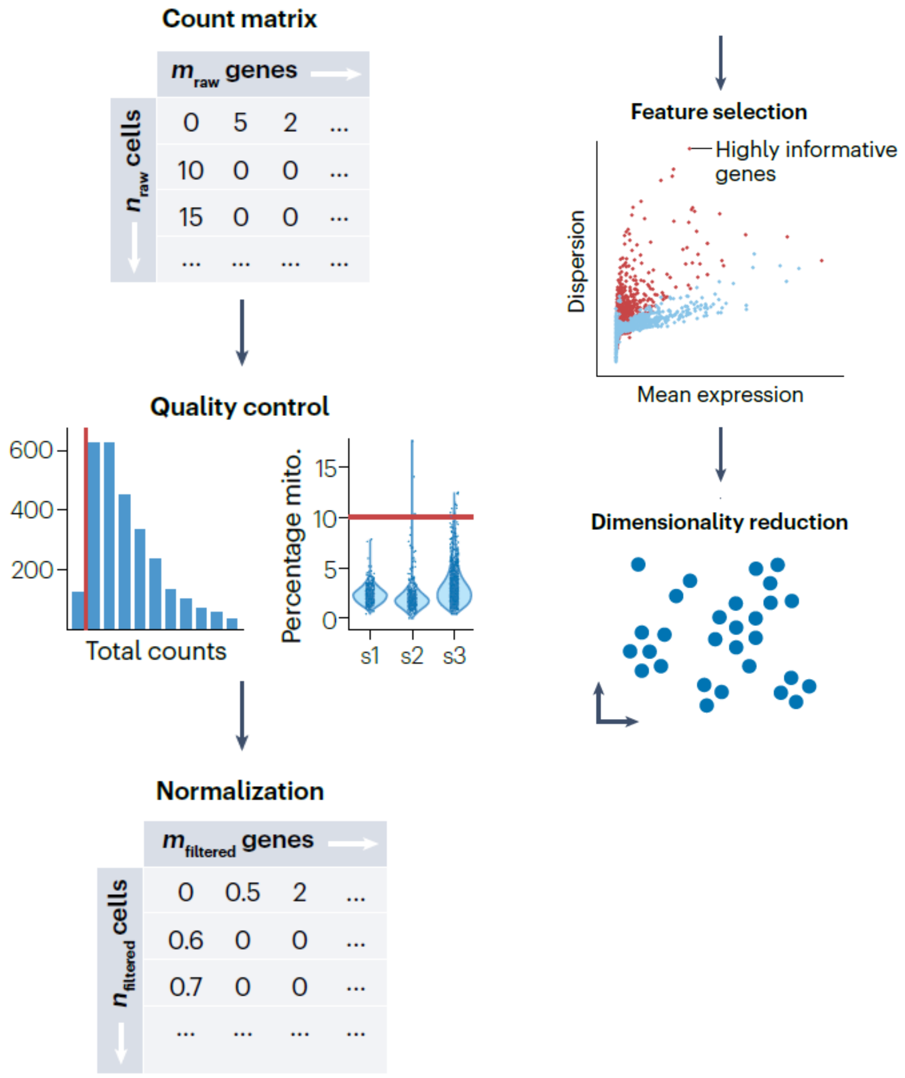
Feature selection, PCA and UMAP

# Dimensionality reduction

Approaches to reduce the dimensionality of the dataset in order to:

- Simplify complexity: reduce number of features (genes).
- Reduce the noise and identify the most relevant information in the data.
- Ease the computational burden on downstream analysis tools.
- Visualize the data.





Count matrix

| $n_{raw}$ cells \ $m_{raw}$ genes | | | |
|---|---|---|---|
| 0 | 5 | 2 | ... |
| 10 | 0 | 0 | ... |
| 15 | 0 | 0 | ... |
| ... | ... | ... | ... |

Normalization

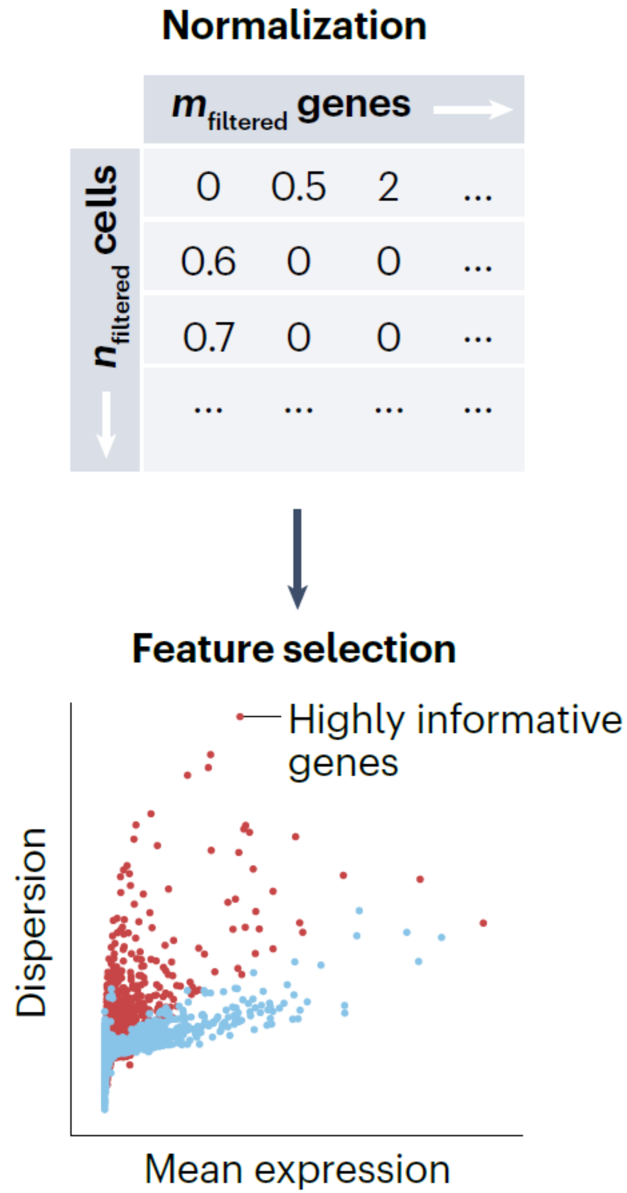
| $n_{filtered}$ cells \ $m_{filtered}$ genes | | | |
|---|---|---|---|
| 0 | 0.5 | 2 | ... |
| 0.6 | 0 | 0 | ... |
| 0.7 | 0 | 0 | ... |
| ... | ... | ... | ... |

# Dimensionality reduction – Feature selection

The first step of reducing the dimensionality of scRNAseq datasets commonly is **feature selection.**

In this step, the dataset is filtered to keep only genes that are "informative" of the variability in the data.
Thus, **highly variable genes (HVGs)** are often used.

Genes are binned by their mean expression, and the genes with the highest variance to mean ratio are selected as HVGs in each bin.



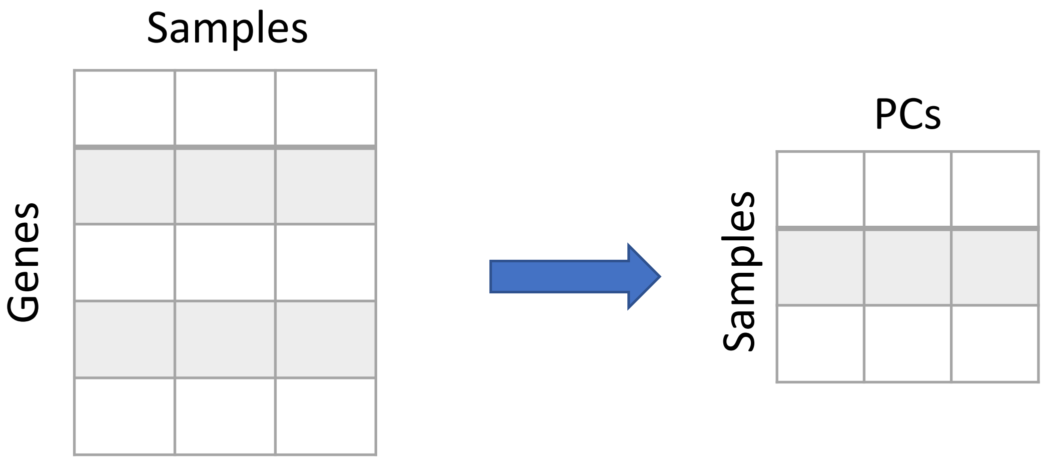
| | | | |
|---|---|---|---|
| 0 | 0.5 | 2 | ... |
| 0.6 | 0 | 0 | ... |
| 0.7 | 0 | 0 | ... |
| ... | ... | ... | ... |

# Dimensionality reduction – PCA

After feature selection, the dimensions of the data set can be further reduced by dimensionality reduction algorithms.
These algorithms embed the expression matrix into a low-dimensional space, which is designed to capture the underlying structure in the data in as few dimensions as possible.

**Principal component analysis (PCA)** is a linear algebraic method of dimensionality reduction.

PCA reduces the dimensionality of the dataset by transforming the original variables into a set of new, uncorrelated variables called principal components, that capture the highest variance possible.

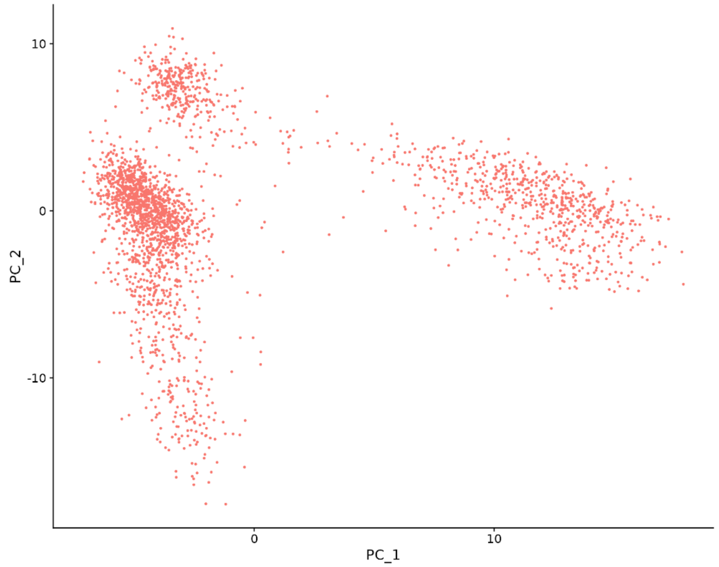


Samples

Genes

PCs

Samples



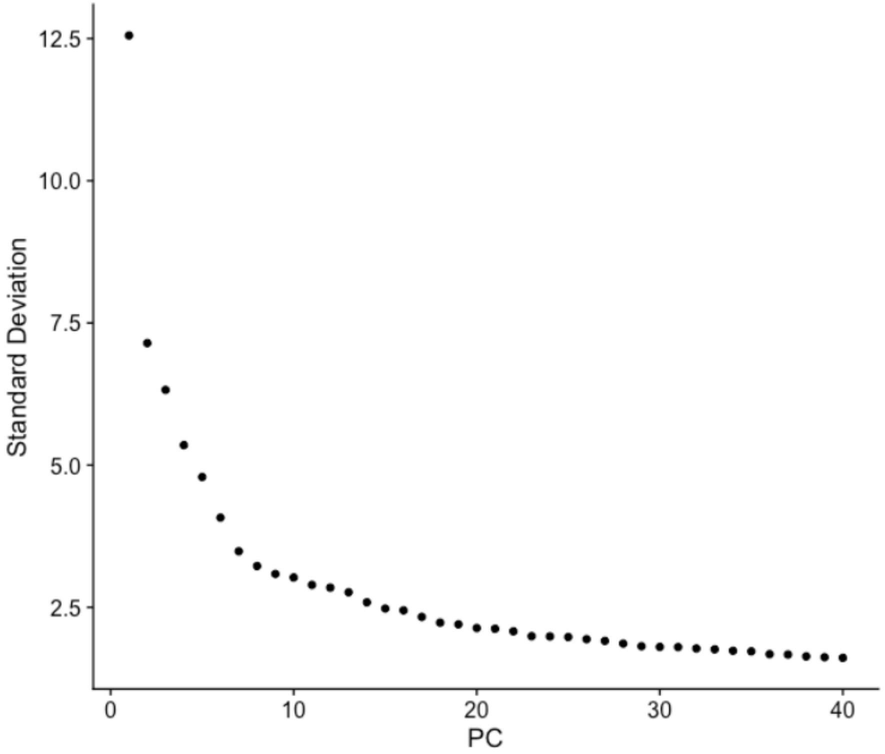
# Dimensionality reduction – PCA

Data is usually scaled prior to PCA (Z-score): scales the expression of each gene, so that the variance across cells is 1. This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate.

Typically, PCA summarizes a dataset via its top N principal components, where N can be determined by “elbow” heuristics.
The top principal components contain higher variance from the data.

Drawback:

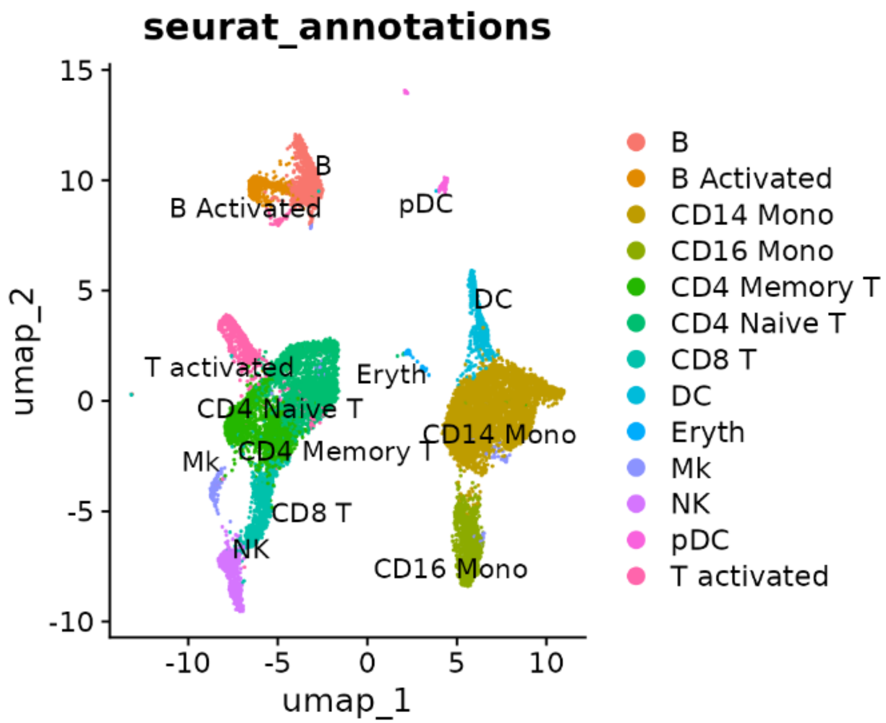
- PCA is restricted to linear dimensions and
- PCA assumes approximately normally distributed data.



# Dimensionality reduction – UMAP

UMAP (**Uniform Manifold Approximation and Projection for Dimension Reduction**):

- It is based on topological structures in multidimensional space.
- It works better in plotting short and long-range interaction between cells.
- It preserves more the global structure of the data.



# Data Integration

Batch effect detection and correction

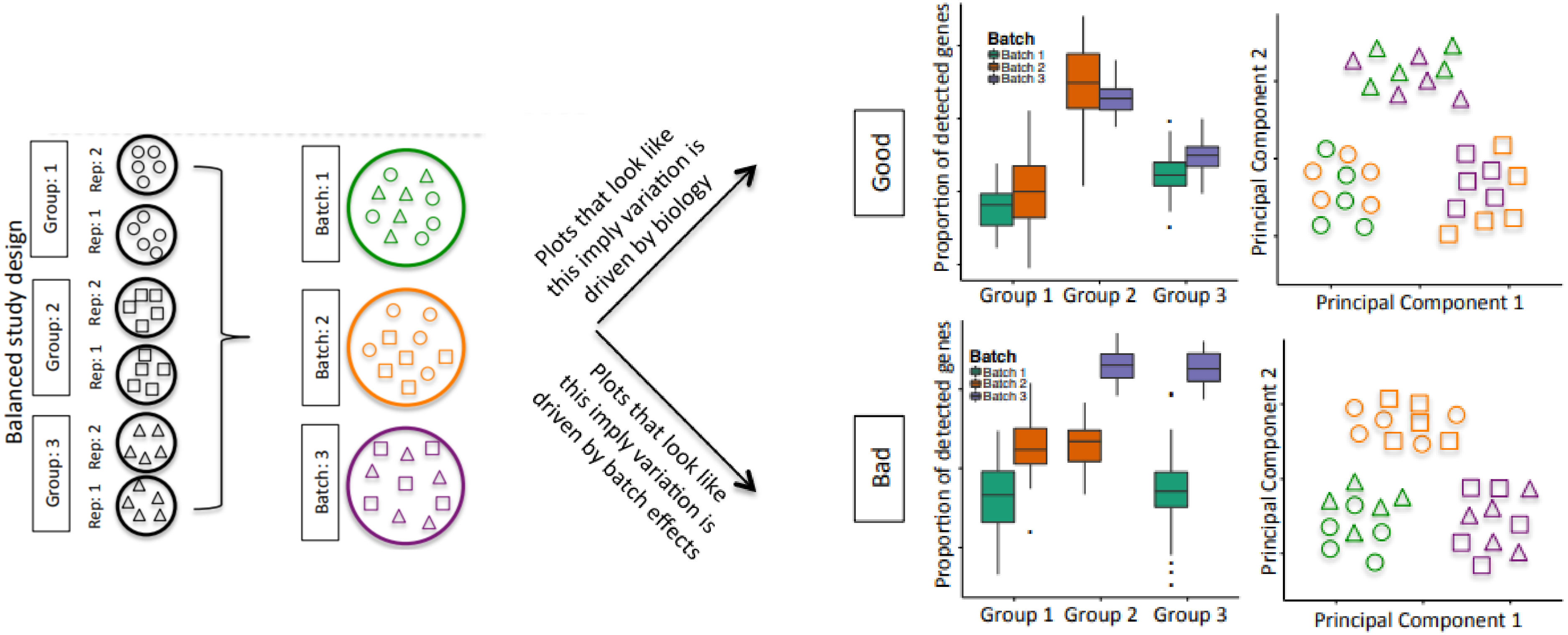
# Data integration and batch effect

**Data integration** → integration of data from multiple experiments with varied conditions. For example scRNAseq datasets from multiple experimental batches, donors, or conditions.

Data integration can give rise to technical variation, that is the batch effect.

**Batch effects** can occur when cells are handled in distinct groups. Batch effect may mask true biological variation between samples: cells are grouped based on the different batches, and we are not able to identify cell types shared between different batches and/or find cell type specific of a specific condition or treatment.

A common strategy to assess the presence of batch effects is to use dimensionality reduction strategies (PCA).



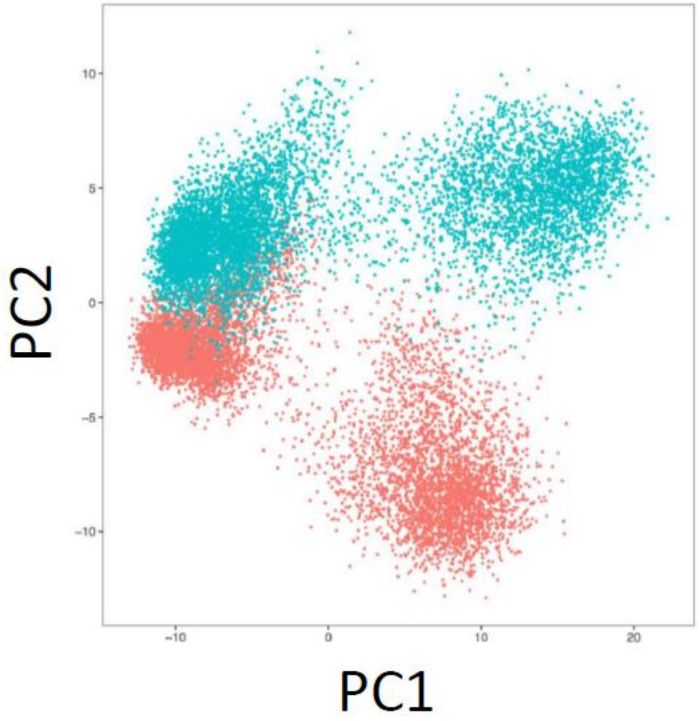
# Batch effect correction

## Batch effect detection

Batch effect detection: perform Principal Component Analysis (PCA) and inspect diagnostic plots to detect batch effect.

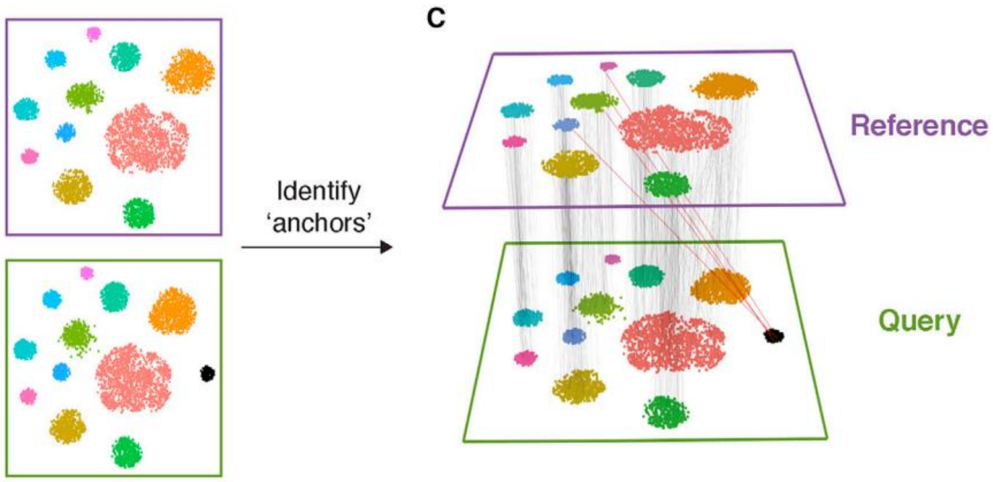
Common methods for batch effect correction:

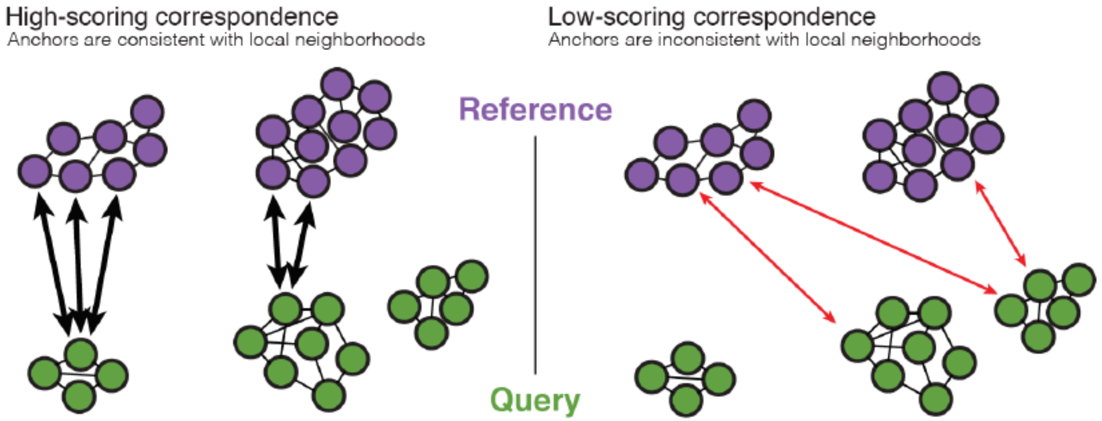
- CCA + anchors
- rPCA
- Harmony
- MNN correct
- LIGER
- Scanorma



# Batch effect correction – CCA and anchors

Canonical correlation analysis (CCA) and anchors.
This method searches for corresponding cells across datasets.




High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query

- Perfom **CCA**: it is a form of PCA  that identifies the common sources of variation between the conditions/groups (using the 3000 most variant genes from each sample).
- Identify **anchors** or **mutual nearest neighbors (MNNs)** across datasets.
- Score anchors: assess the similarity between anchor pairs by the overlap in their local neighborhoods (incorrect anchors will have low scores).
- Use anchors and corresponding scores to transform the cell expression values (so cells in the same neighborhood should have similar correction values) allowing for the integration of the datasets .

# Batch effect correction – Reciprocal PCA (RPCA)

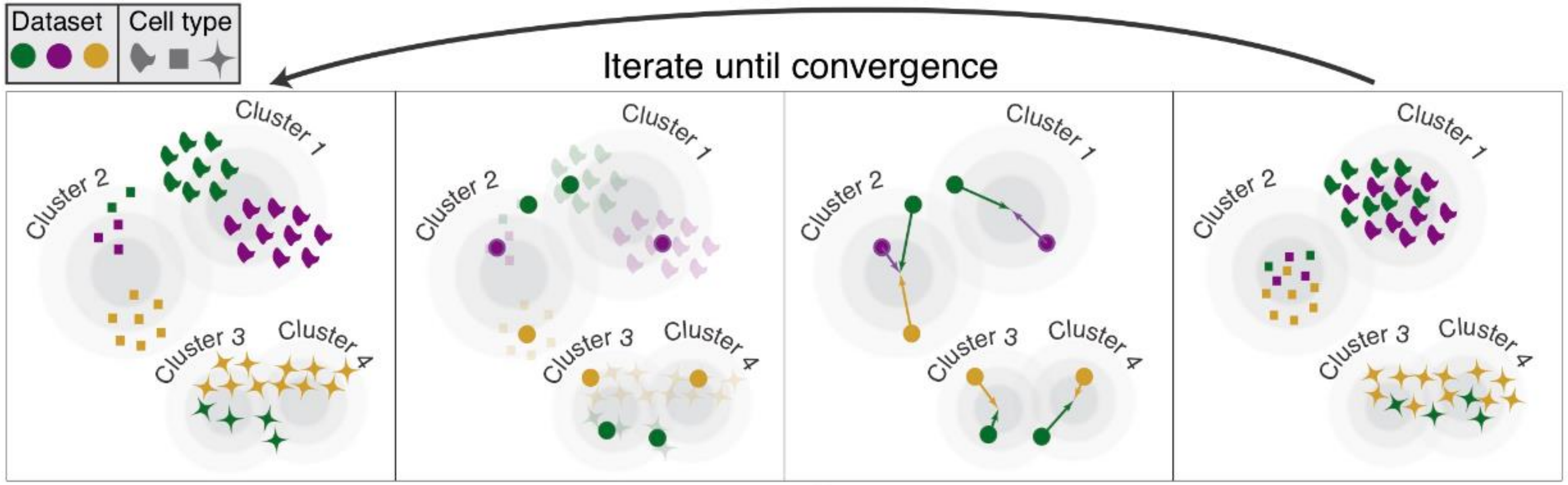
Utilize **reciprocal PCA (‘RPCA’)** to identify anchors, instead of CCA.

RPCA-based integration runs significantly faster, and also represents a more conservative approach where cells in different biological states are less likely to ‘align’ after integration.

RPCA it is recommended during integrative analysis where:

- A substantial fraction of cells in one dataset have no matching type in the other.
- Datasets originate from the same platform (i.e. multiple lanes of 10x genomics),
- There are a large number of datasets or cells to integrate (see here for more tips on integrating large datasets).

# Batch effect correction – Harmony



**A** Soft assign cells to clusters, favoring mixed dataset representation

**B** Get cluster centroids for each dataset

**C** Get dataset correction factors for each cluster

**D** Move cells based on soft cluster membership