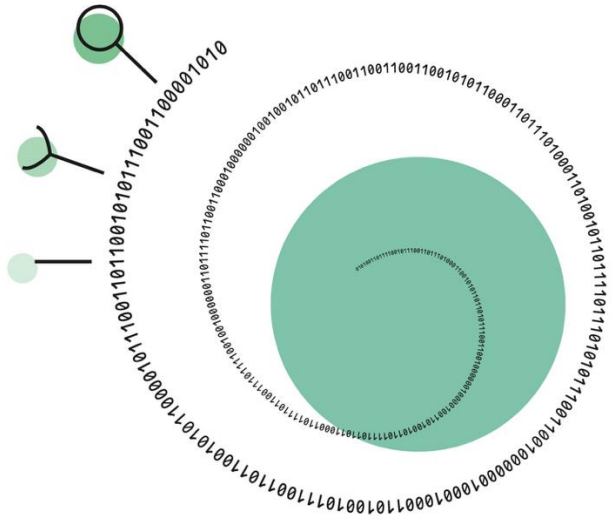
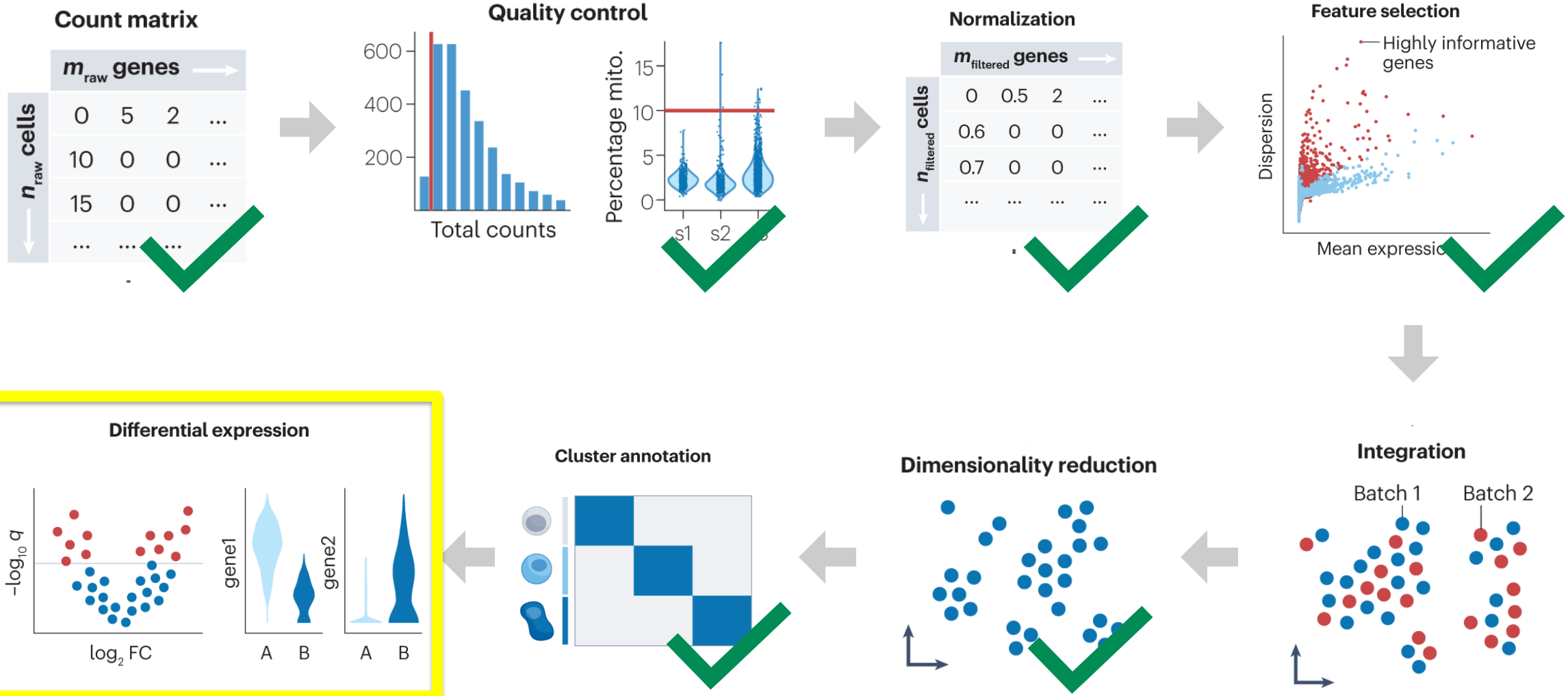


Single Dell Data Analysis Course

Differential abundance and gene expression analysis

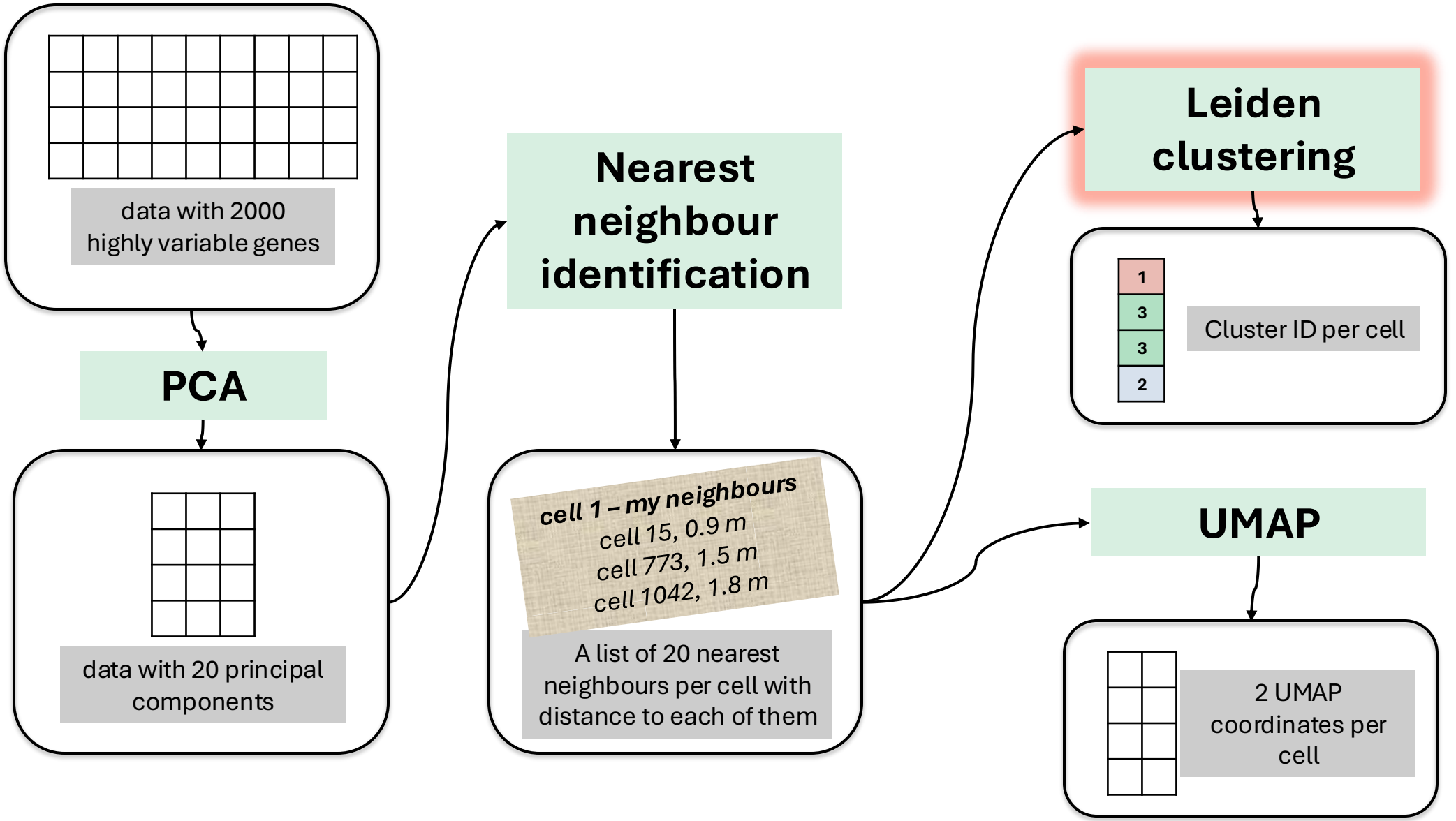
Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Today

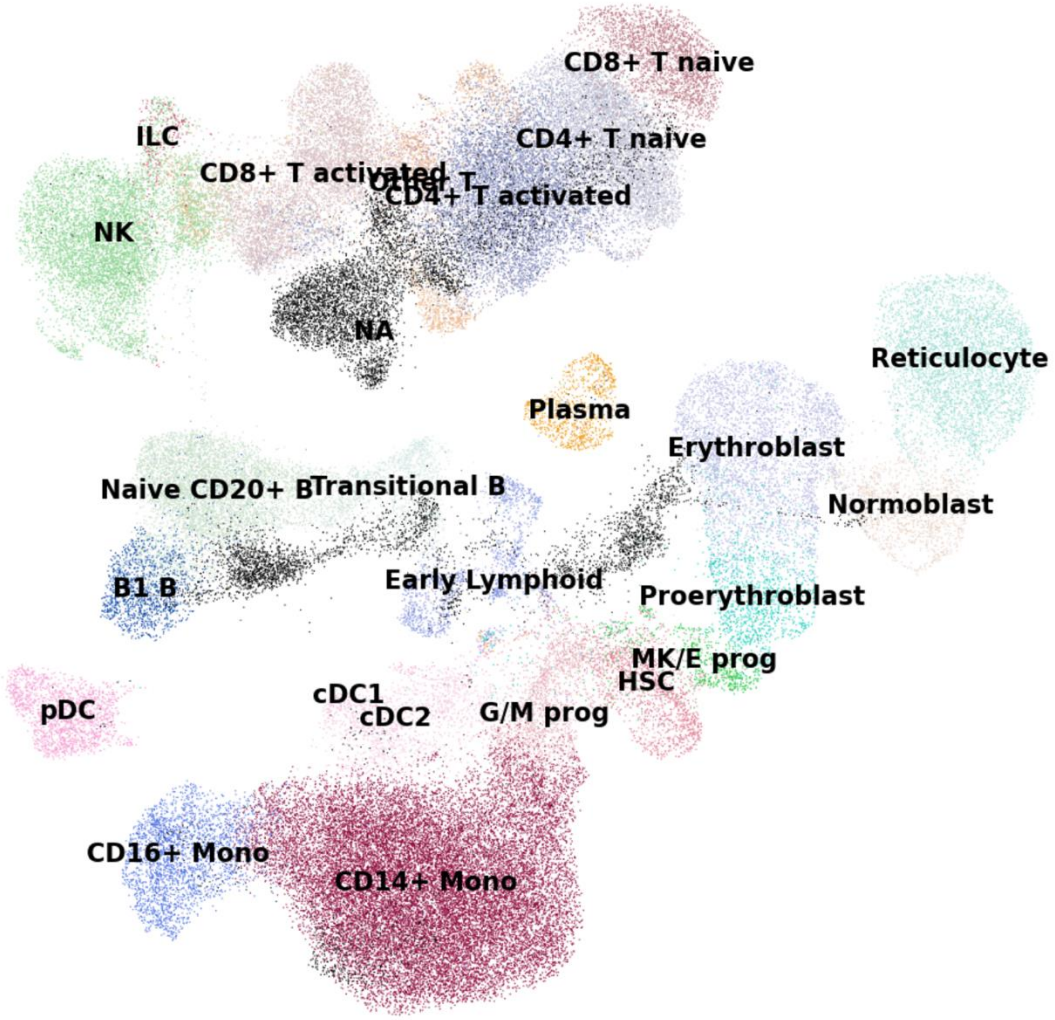


Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

Recap: Where we stand after a whole lot of processing

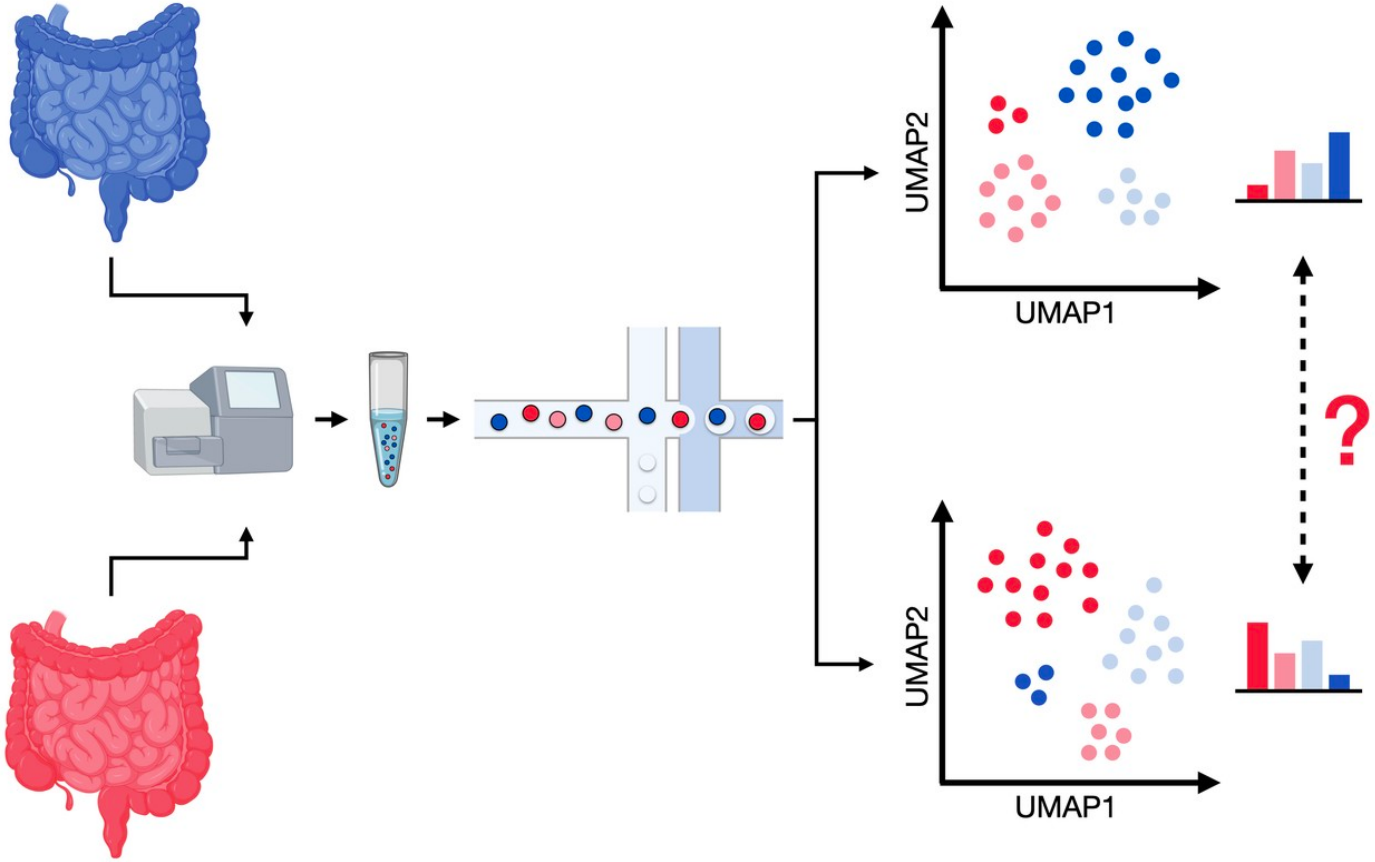


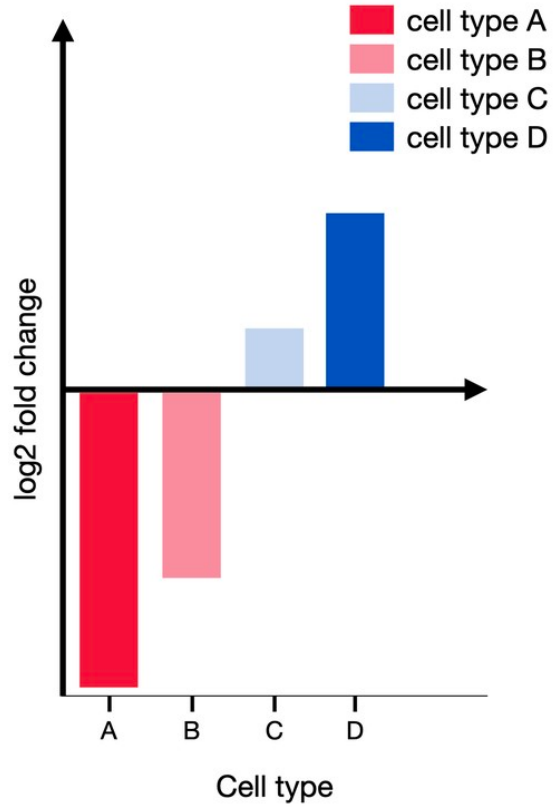
Finally, an annotated dataset.



https://www.sc-best-practices.org/cellular_structure/annotation.html

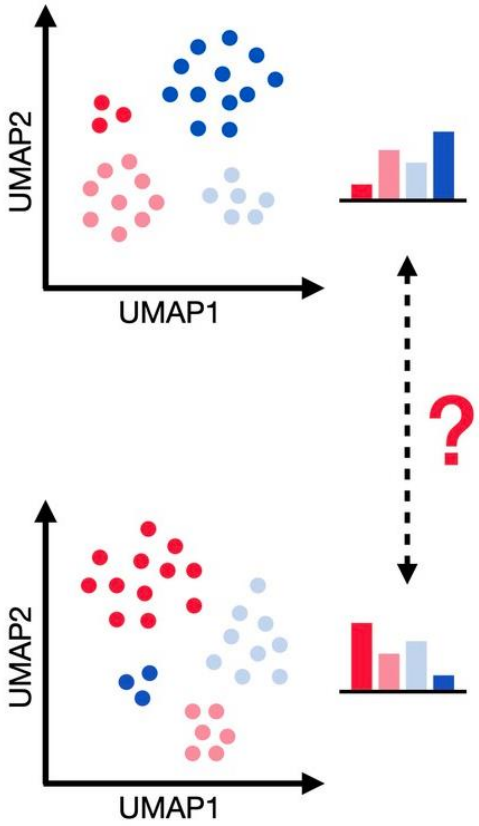
What to do with an annotated dataset:
1) Compositional analysis





https://www.sc-best-practices.org/conditions/compositional.html

Cell type fraction data is compositional.

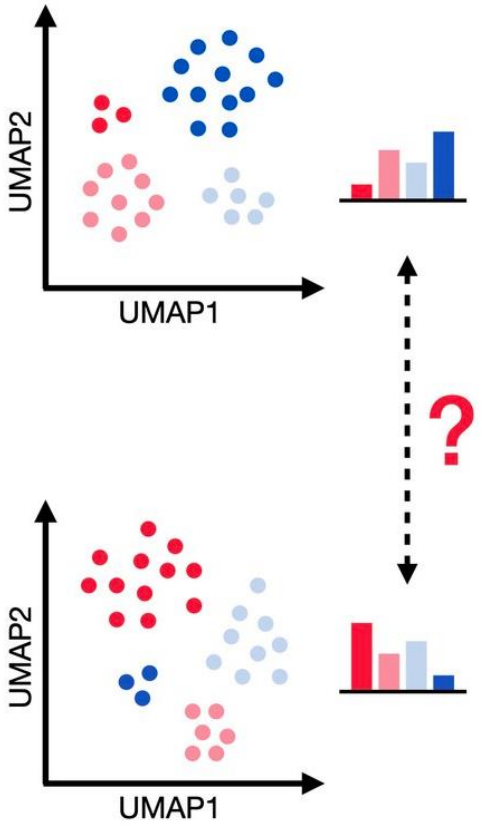


Type A	5%
Type B	30%
Type C	25%
Type D	40%

Sums to 100%

https://www.sc-best-practices.org/conditions/compositional.html

Cell type fraction data is compositional.



Type A	5%
Type B	30%
Type C	25%
Type D	40%

Sums to 100%

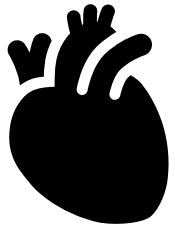
If one cell type becomes more abundant,

the fractional contribution of the other cell types goes down

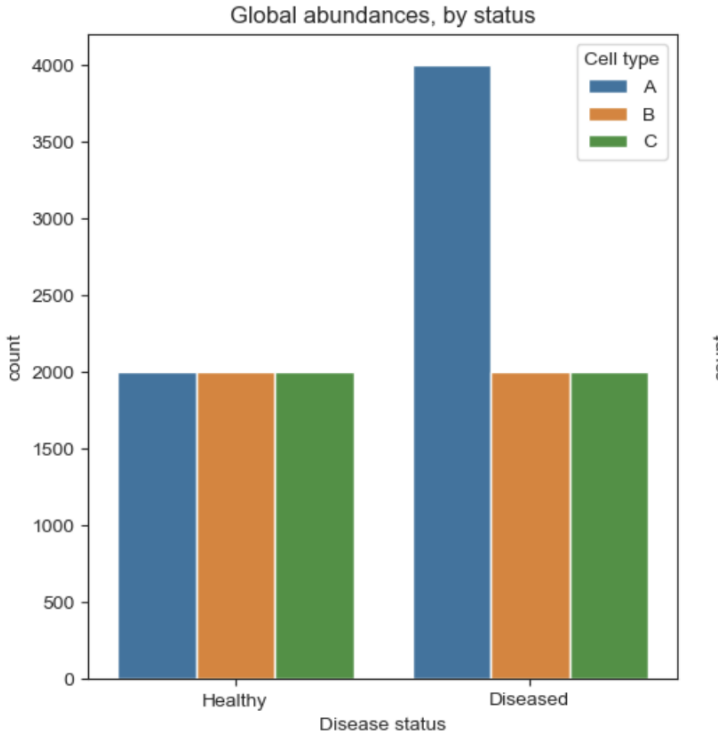
even though their absolute numbers may be unchanged.

https://www.sc-best-practices.org/conditions/compositional.html

Cell type count data **from a sample** is compositional.

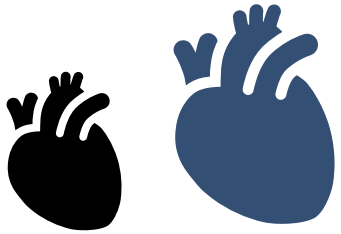


Whole organ

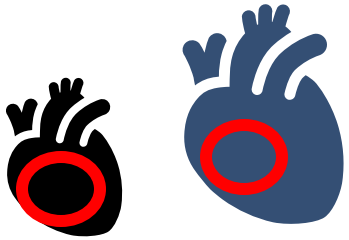


https://www.sc-best-practices.org/conditions/compositional.html

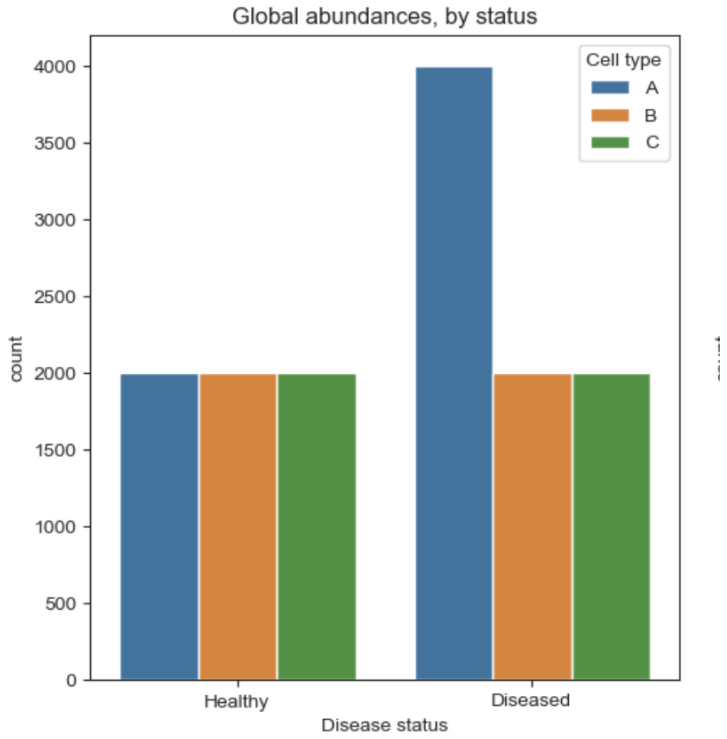
Cell type count data **from a sample** is compositional.

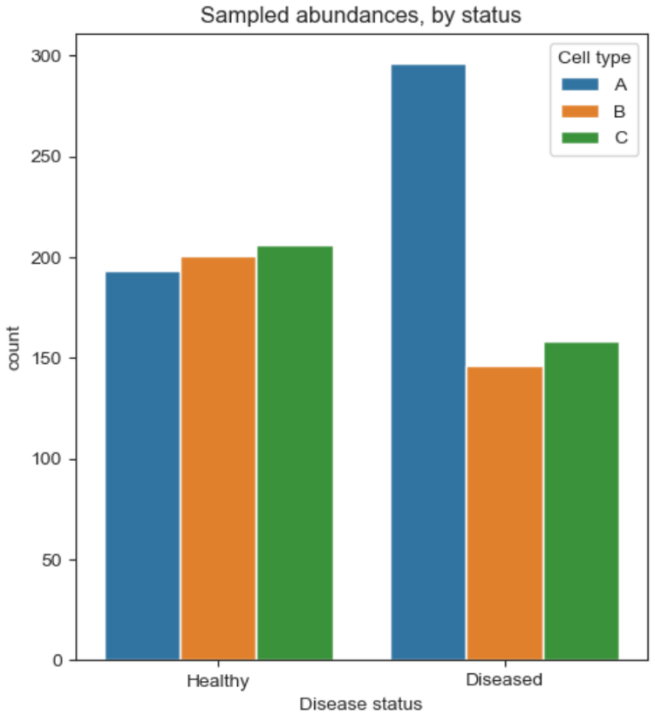


Whole organ



small sample from organ





https://www.sc-best-practices.org/conditions/compositional.html

Compositional data requires special statistical methods.

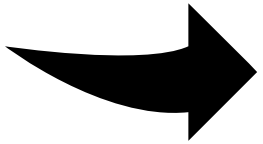
Compositional data is characterized by inherent negative correlations between its features

[If one goes up, another must go down]

Compositional data requires special statistical methods.

Compositional data is characterized by inherent negative correlations between its features

[If one goes up, another must go down]



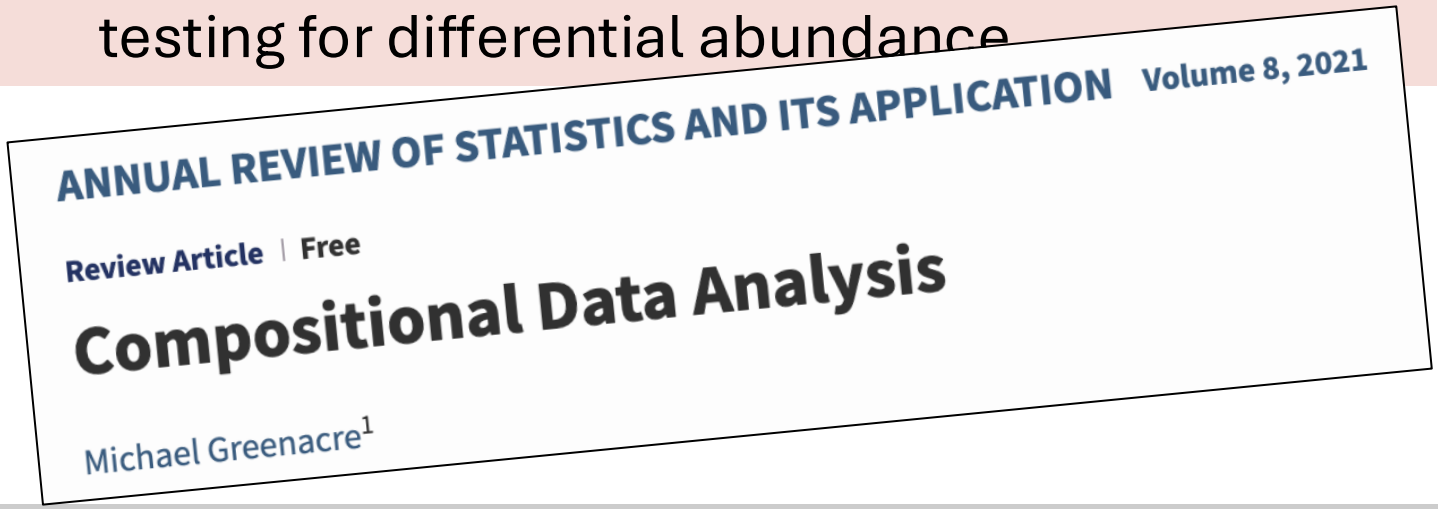
Statistical methods for non-compositional data (e.g. t-test, Wilcoxon's test…) may return false positive results when testing for differential abundance

Compositional data requires special statistical methods.

Compositional data is characterized by inherent negative correlations between its features

[If one goes up, another must go down]

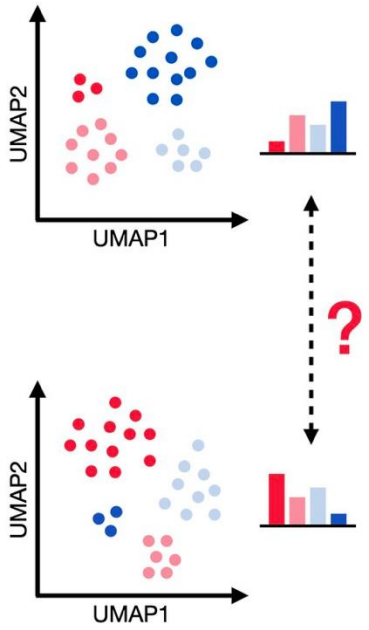
Statistical methods for non-compositional data (e.g. t-test, Wilcoxon's test…) may return false positive results when testing for differential abundance



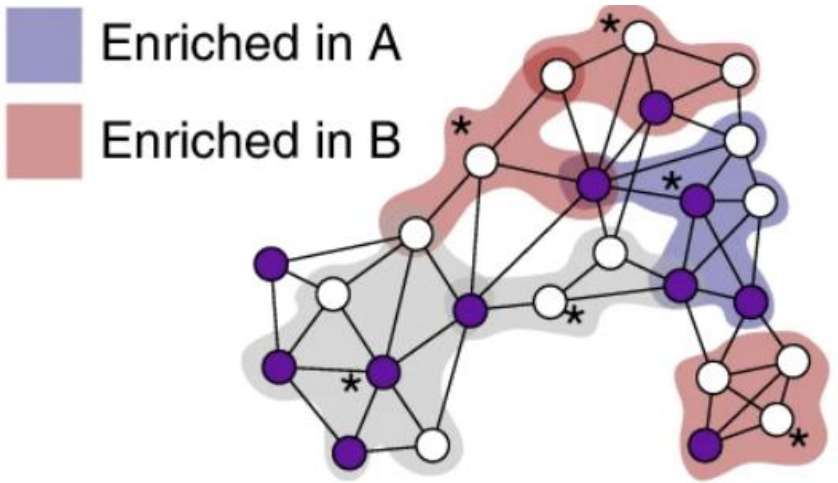
https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-124436

For differential abundance testing in sc data, use dedicated methods.

scCODA for labelled clusters



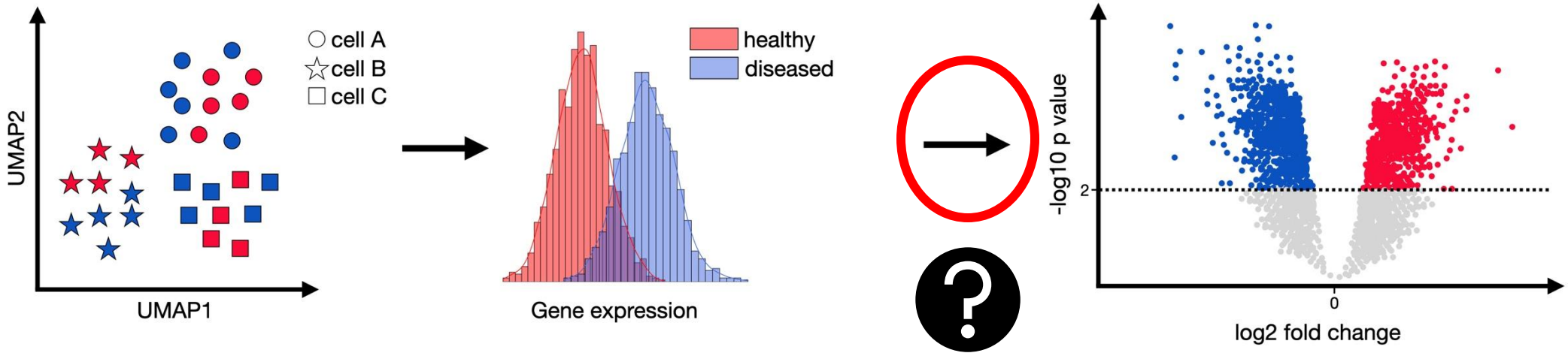
miloR for graph neighborhoods (does not need labels)



https://www.nature.com/articles/s41467-021-27150-6

https://www.nature.com/articles/s41587-021-01033-z

What to do with an annotated dataset:
2) Differential expression testing between conditions

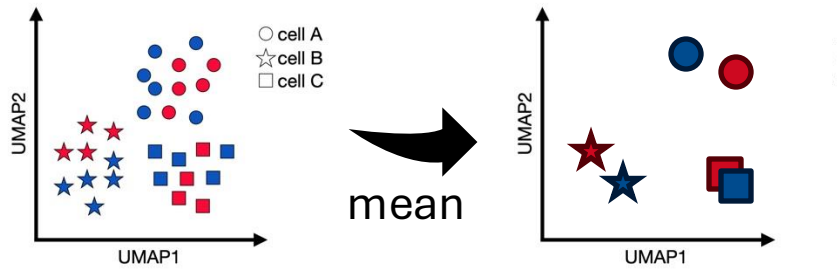


https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Overview: Approaches to sc DEG testing

pseudobulk methods

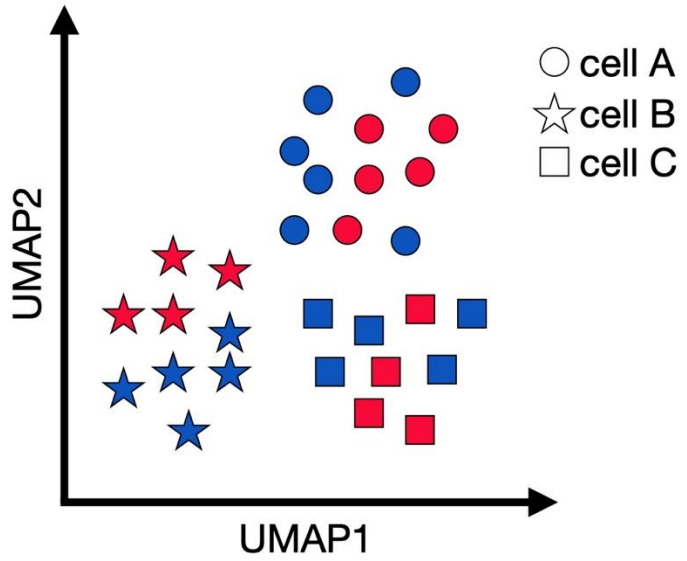
1. Preprocess, cluster and annotate the dataset
2. Aggregate counts by taking the mean per cell type and sample/patient



https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Differential Gene Expression Testing: Cells are not independent replicates

Aim: compare case and control, e.g. disease X vs. healthy



Naïve approach

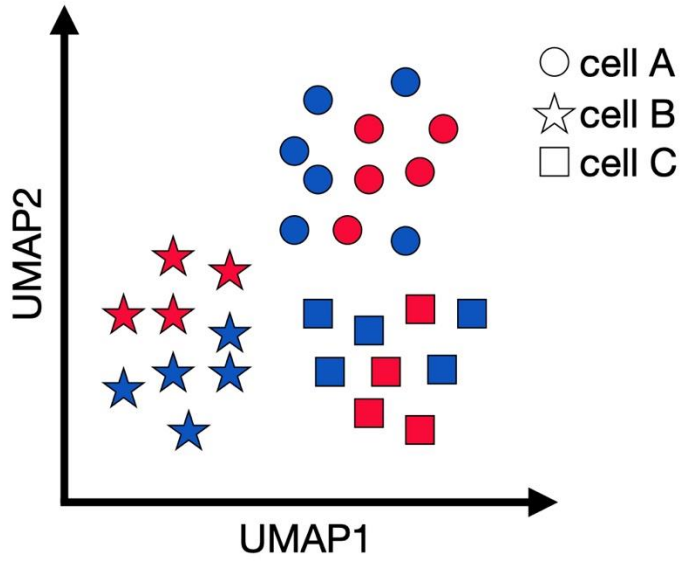
Statistical comparison (e.g. Wilcoxon via marker gene tests) between celltype A in case and control

Many highly significant results!

https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Differential Gene Expression Testing: Cells are not independent replicates

Aim: compare case and control, e.g. disease X vs. healthy



Naïve approach

Statistical comparison (e.g. Wilcoxon via marker gene tests) between celltype A in case and control

Many highly significant results!

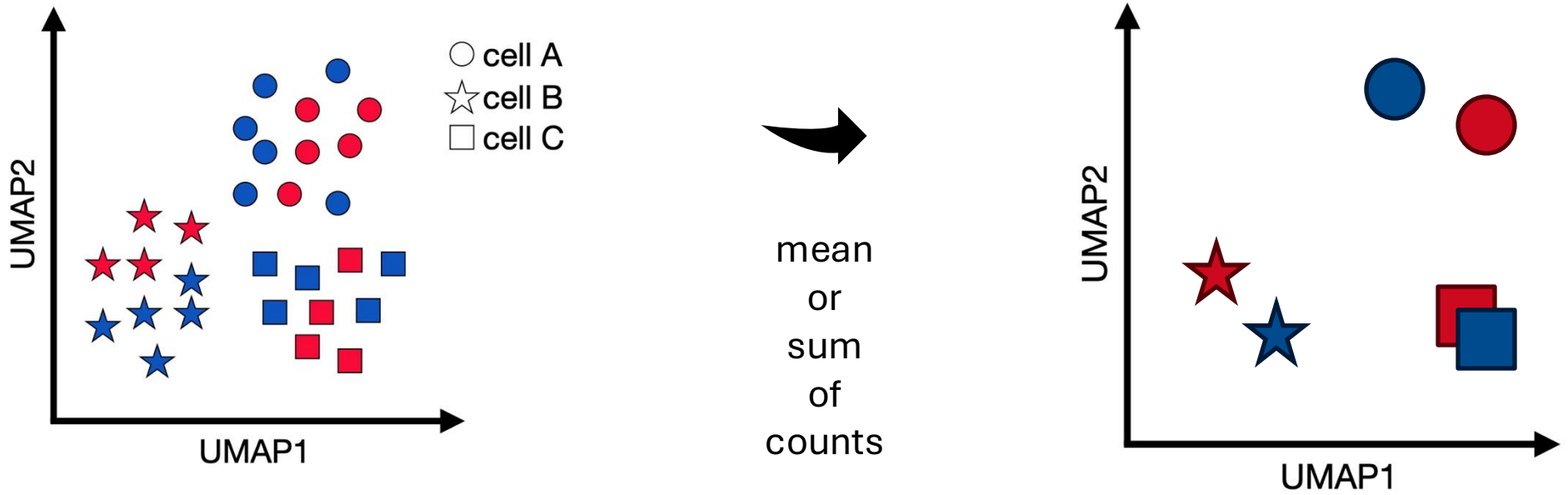
But: Treats each cell as an independent replicate, ignoring that cells from same patient are correlated!

https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Differential Expression – Marker Testing vs Condition comparison

Marker gene testing (within dataset)	Differential expression between conditions
Compares: "Cluster X" vs "all other cells"	Compares: "Condition A" vs "Condition B" within same cell type
Purpose: Identify cell type-defining features	Purpose: Identify condition-responsive genes
Tests assume cells are independent observations	- Requires biological replicates (multiple patients/samples) - Needs ability to account for donor-specific effects (paired samples)
FindMarkerGenes() and friends	**DEG framework like DESeq, edgeR and friend**

One valid strategy: Pseudobulking



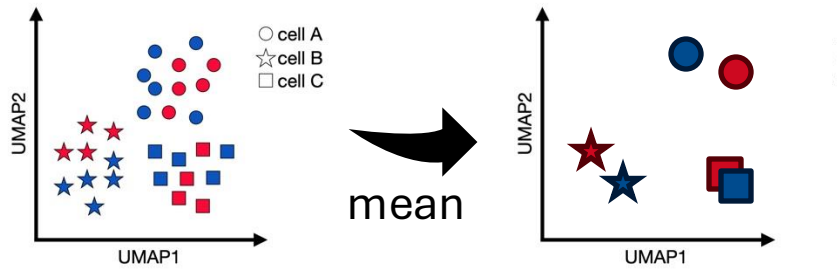
- One gene expression vector per sample and cell type
- Less sparse

https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Overview: Approaches to sc DEG testing

pseudobulk methods

1. Preprocess, cluster and annotate the dataset
2. Aggregate counts by taking the mean per cell type and sample/patient
3. Apply a differential expression method like for bulk data



https://www.sc-best-practices.org/conditions/differential_gene_expression.html

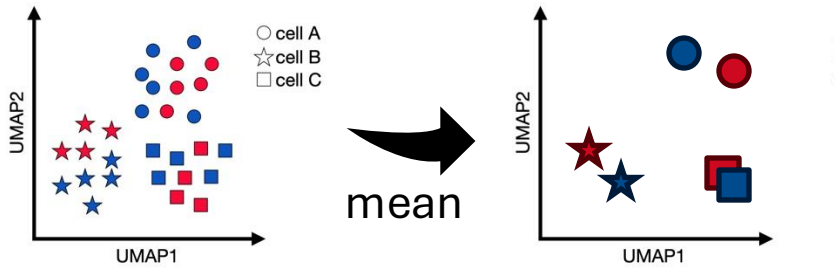
Overview: Approaches to sc DEG testing

pseudobulk methods

1. Preprocess, cluster and annotate the dataset
2. Aggregate counts by taking the mean per cell type and sample/patient
3. Apply a differential expression method like for bulk data

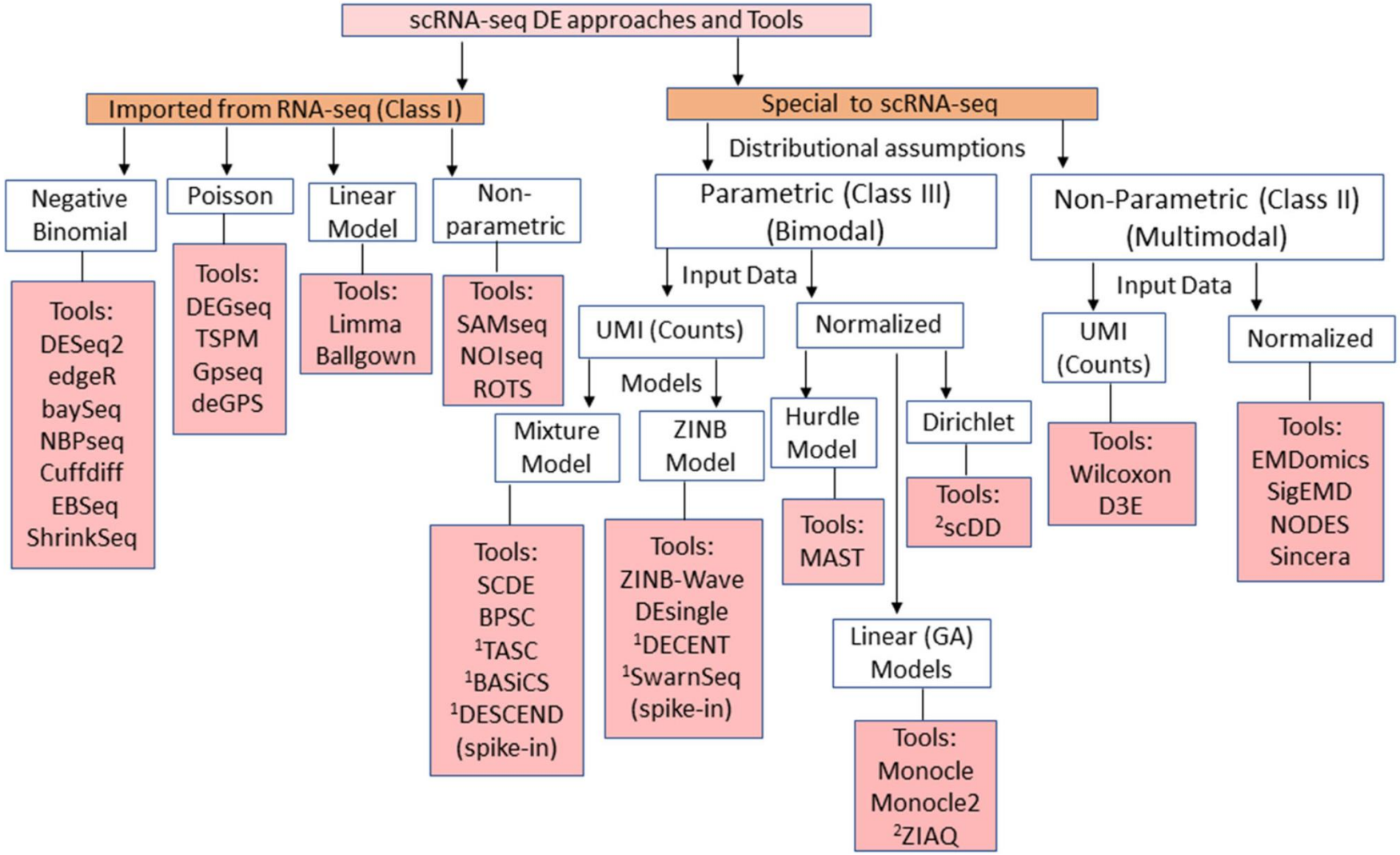
single-cell specific methods

- Typically use generalized mixed effects models
- Model specific single-cell data noise properties accurately



https://www.sc-best-practices.org/conditions/differential_gene_expression.html

Overview: Approaches to sc DEG testing



Das, S. et al. A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. Genes 2021, 12, 1947. https://doi.org/10.3390/genes12121947

Overview: Approaches to sc DEG testing

pseudobulk methods

single-cell specific methods

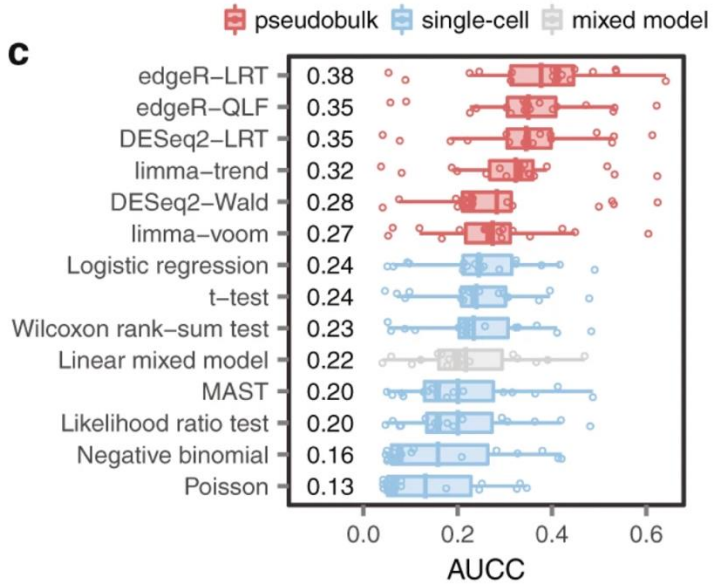
! Consensus and robustness across methods are low!

Squair, J.W., Gautier, M., Kathe, C. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2

Overview: Approaches to sc DEG testing

pseudobulk methods | **single-cell specific methods**

! Consensus and robustness across methods are low!



Pseudobulk methods perform favourably against single-cell specific methods.

Squair, J.W., Gautier, M., Kathe, C. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2

Example: pseudobulk/DESeq2
1) Create pseudobulk object

Example dataset: *in vitro* stimulated PBMCs from 8 Lupus patients before and after 6h-treatment with INF-β (16 samples in total)

label	cluster	cell_type	replicate
ctrl	9	CD14+ Monocytes	patient_1016
ctrl	9	CD14+ Monocytes	patient_1256
ctrl	3	CD4 T cells	patient_1488
ctrl	9	CD14+ Monocytes	patient_1256
ctrl	4	Dendritic cells	patient_1039

annotated single cell dataset with **raw counts** [thousands of rows]

calculate summed gene expression per patient and cell type

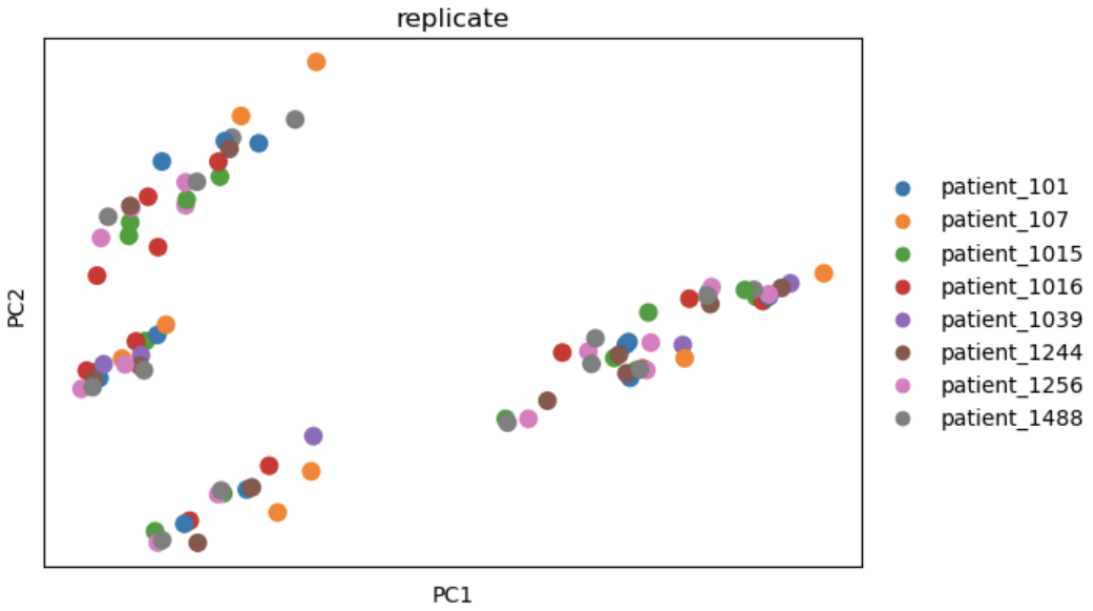
Pseudobulk dataset with one entry per patient_celltype [patient x cell types rows] [e.g. 16 x 7 = 112]

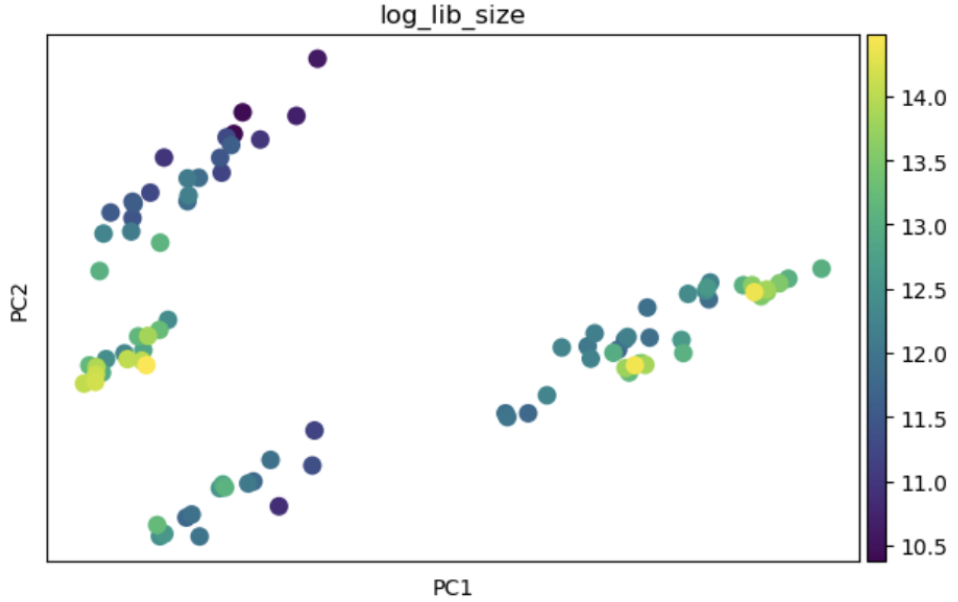
https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk

Example: pseudobulk/ DESeq2
2) Inspect major axes of variation

The statistical model used for differential gene expression must capture **major axes of variation** to return accurate differential gene expression results.

Lognormalize pseudobulk data → PCA → inspect covariates



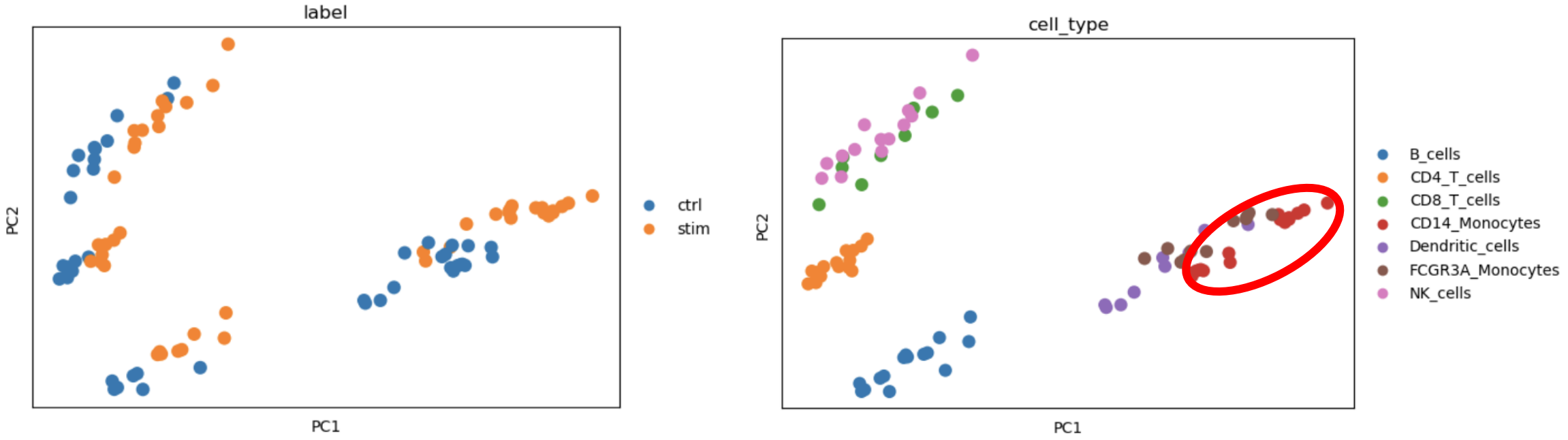


https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk

Example: pseudobulk/ DESeq2
2) Inspect major axes of variation

The statistical model used for differential gene expression must capture **major axes of variation** to return accurate differential gene expression results.

Lognormalize pseudobulk data → PCA → inspect covariates



https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk

Example: pseudobulk/ DESeq2
3) Run a differential expression test for a chosen cell type

Since we identified no major confounders during the exploratory analysis, we set up the simplest **design matrix** with the stimulation label as sole covariate.

[R-like pseudocode]

```
# Create DESeq2 dataset object
dds <- DESeqDataSetFromMatrix(
      countData = pseudobulk_counts,
      colData = sample_metadata,
      design = ~label)
```

design matrix

Sample	Intercept	condition	Interpretation
ctrl_1	1	0	baseline expression
ctrl_2	1	0	baseline expression
ctrl_3	1	0	baseline expression
stim_1	1	1	baseline + treatment
stim_2	1	1	baseline + treatment
stim_3	1	1	baseline + treatment

model

Gene expression =
$\beta_0 \times$ Intercept +
$\beta_1 \times$ condition

Example: pseudobulk/ DESeq2
3) Run a differential expression test for a chosen cell type

[R-like pseudocode]

```
# Create DESeq2 dataset object
dds <- DESeqDataSetFromMatrix(
      countData = pseudobulk_counts,
      colData = sample_metadata,
      design = ~label)

# Run the DESeq2 pipeline (normalization, dispersion estimation, statistical testing)
dds <- DESeq(dds)
```

Size factor estimation:
Normalizes for library size differences between samples

Dispersion estimation:
Models the relationship between mean expression and variance across genes

Statistical testing:
Fits negative binomial generalized linear models and performs Wald tests

Example: pseudobulk/ DESeq2
3) Run a differential expression test for a chosen cell type

design matrix

Sample	Intercept	condition	Interpretation
ctrl_1	1	0	baseline expression
ctrl_2	1	0	baseline expression
ctrl_3	1	0	baseline expression
stim_1	1	1	baseline + treatment
stim_2	1	1	baseline + treatment
stim_3	1	1	baseline + treatment

model

log(expected counts) =
β_0 × Intercept +
β_1 × condition

Is β_1 (the log2 fold change) different from 0?

Size factor estimation:
Normalizes for library size differences between samples

Dispersion estimation:
Models the relationship between mean expression and variance across genes

Statistical testing:
Fits negative binomial generalized linear models and performs Wald tests

Example: pseudobulk/ DESeq2
3) Run a differential expression test for a chosen cell type

[R-like pseudocode]

```
# Create DESeq2 dataset object
dds <- DESeqDataSetFromMatrix(
      countData = pseudobulk_counts,
      colData = sample_metadata,
      design = ~label)

# Run the DESeq2 pipeline (normalization, dispersion estimation, statistical testing)
dds <- DESeq(dds)

# Extract results for the comparison of interest
res <- results(dds, contrast = c("label", "stimulated", "control"))
```

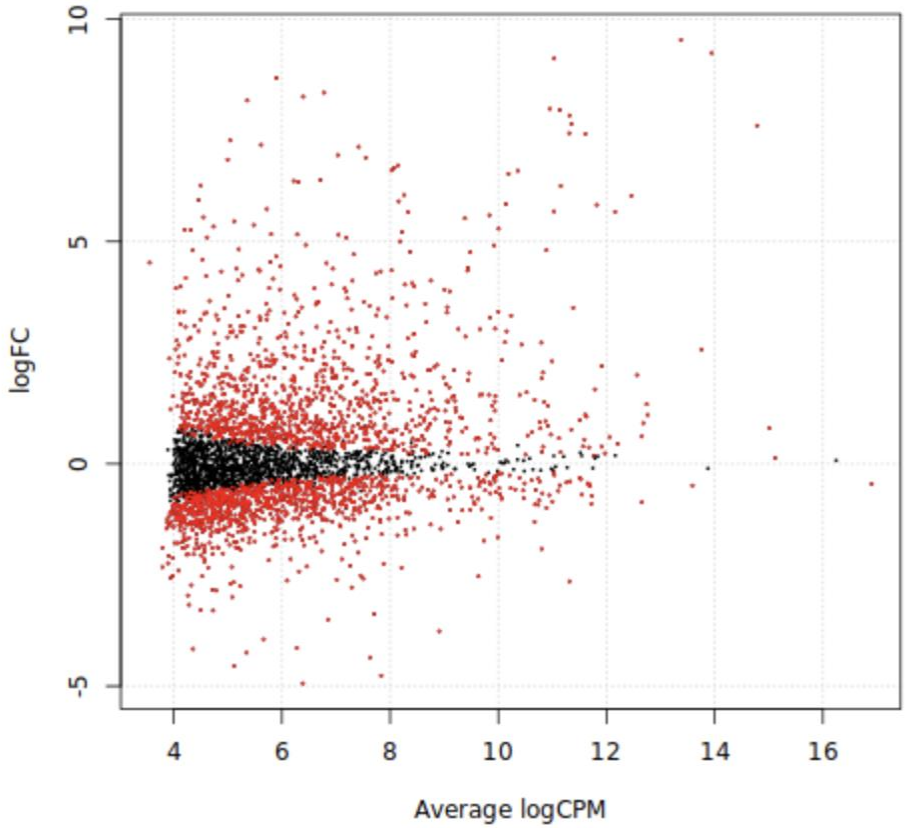
- Extracts the differential expression results for a specific comparison
- Returns log2 fold changes, p-values, and adjusted p-values for each gene
- contrast: Specifies which comparison to extract (condition, numerator, denominator)
- Here: "stimulated" vs "control" within the "label" column

Inspect and visualize DEG results: Smear plot

CD14+ monocytes

	logFC	logCPM	F	PValue	FDR
HESX1	8.345536	6.773420	1281.013295	1.837373e-15	2.766927e-12
CD38	7.126846	7.420668	1243.793133	2.266164e-15	2.766927e-12
NT5C3A	5.657050	8.327003	1218.102628	2.628780e-15	2.766927e-12
SOCS1	4.388247	6.943768	1191.289806	3.079524e-15	2.766927e-12
GMPR	6.943484	7.031832	1159.601183	3.730018e-15	2.766927e-12

Filter for genes with FDR < 0.01 (here marked in red)



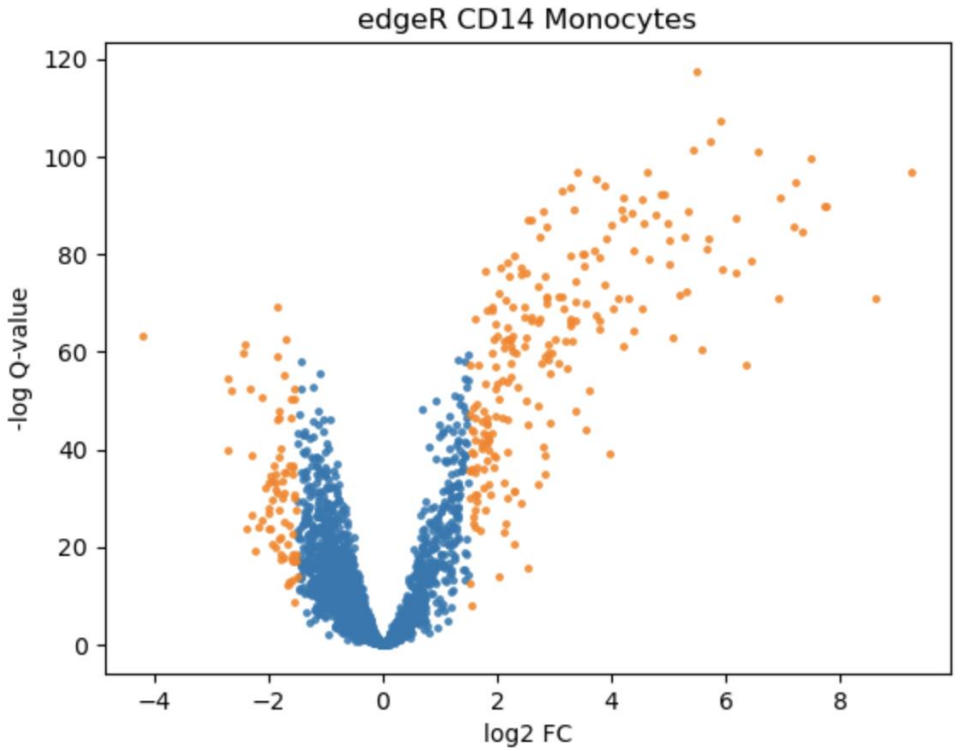
https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk

Inspect and visualize DEG results: Volcano plot

CD14+ monocytes

	logFC	logCPM	F	PValue	FDR
HESX1	8.345536	6.773420	1281.013295	1.837373e-15	2.766927e-12
CD38	7.126846	7.420668	1243.793133	2.266164e-15	2.766927e-12
NT5C3A	5.657050	8.327003	1218.102628	2.628780e-15	2.766927e-12
SOCS1	4.388247	6.943768	1191.289806	3.079524e-15	2.766927e-12
GMPR	6.943484	7.031832	1159.601183	3.730018e-15	2.766927e-12

Filter for genes with FDR < 0.01 And logFC>1.5, (here marked in orange)

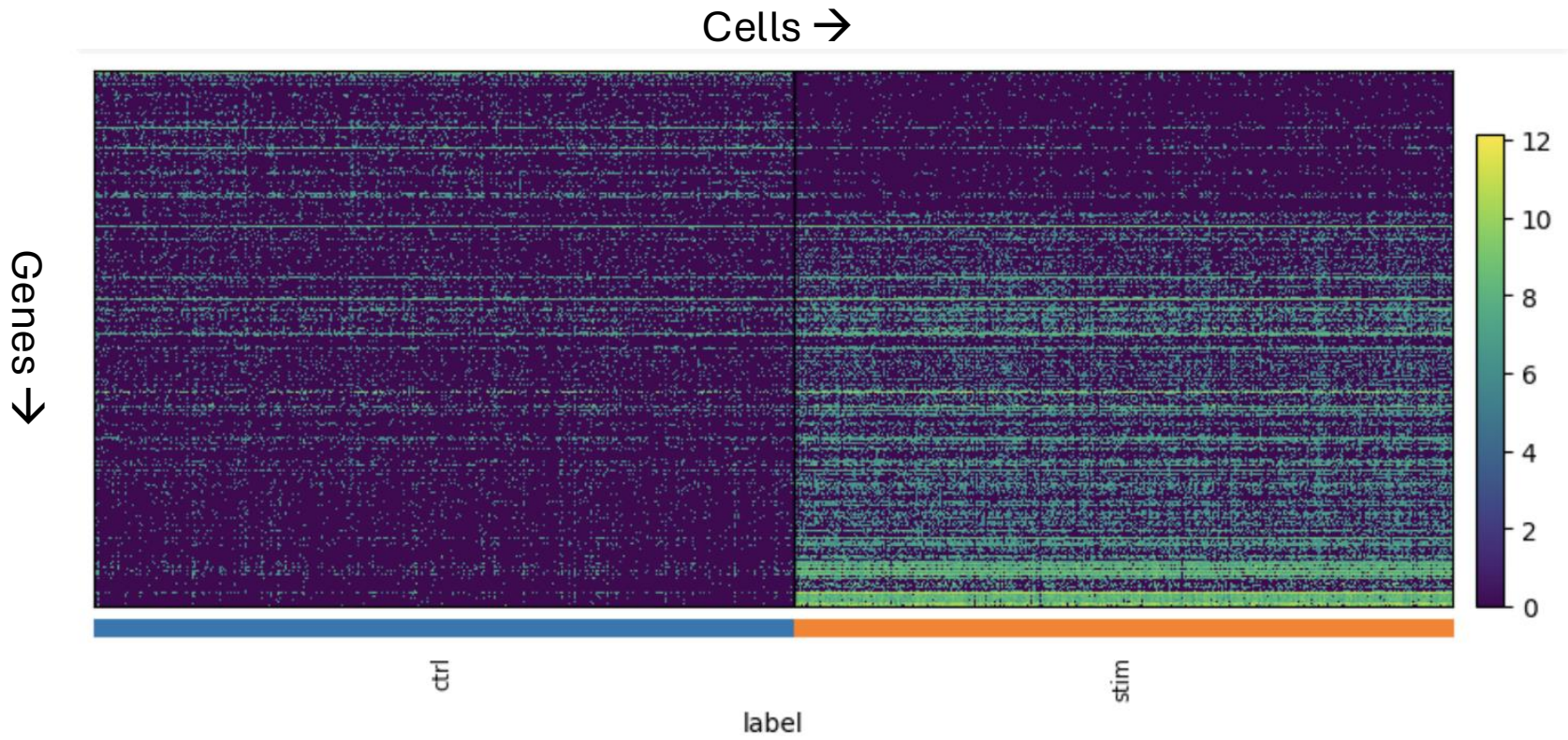


https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk

Inspect and visualize DEG results: Heatmap

CD14+ monocytes

Filter for genes with FDR < 0.01 And logFC>1.5



https://www.sc-best-practices.org/conditions/differential_gene_expression.html#pseudobulk