on differential discovery in scRNA-seq data

...using muscat & miloDE/lemur

Helena L Crowell, PhD

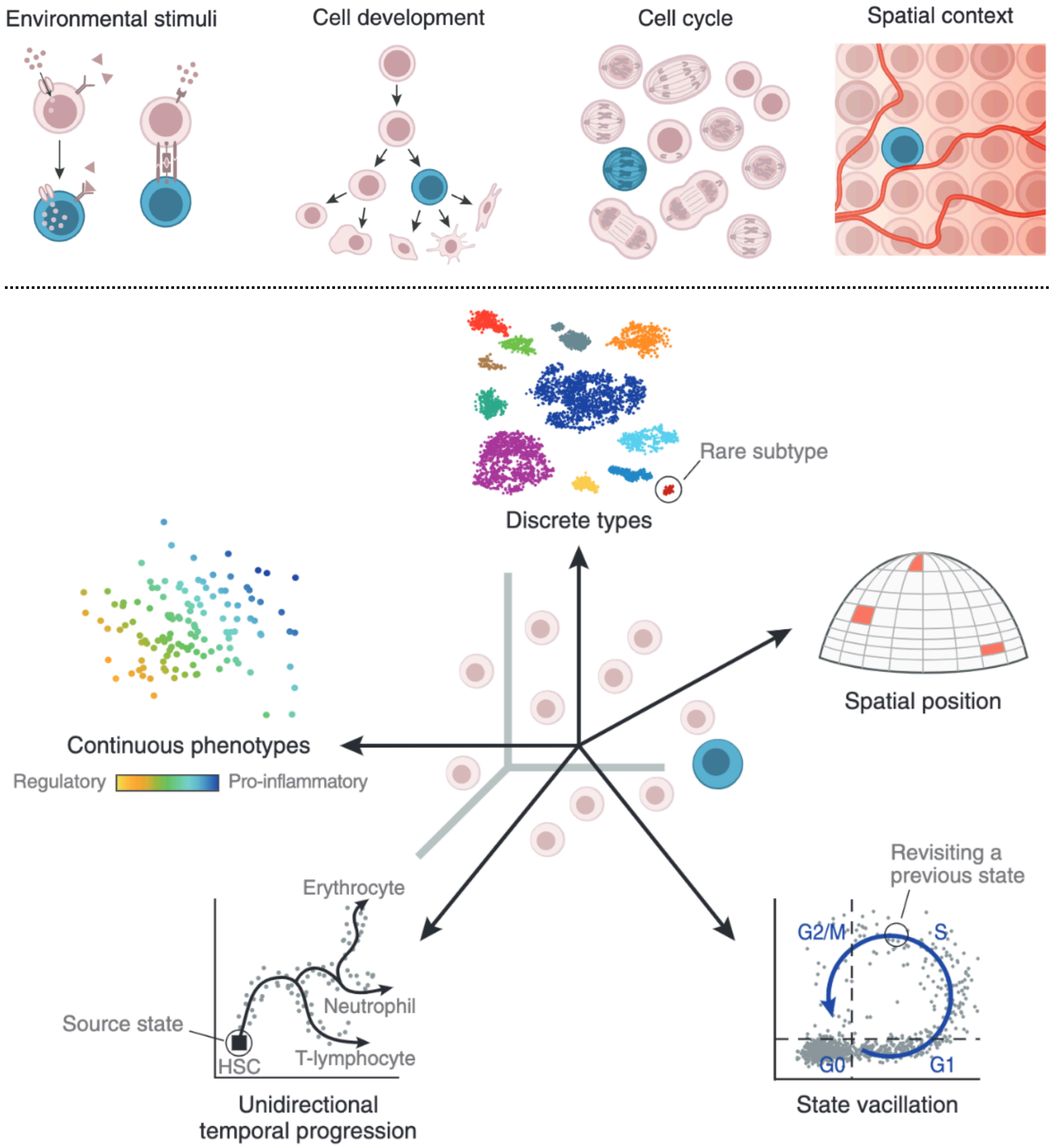
Autumn School for Single Cell-ers

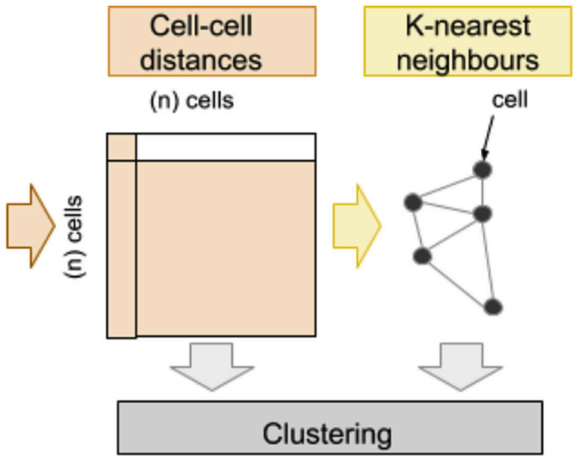
Oct 21, 2025 · GIMM, Oeiras, Portugal

CNAG – National Center
for Genomic Analysis
Barcelona, Spain

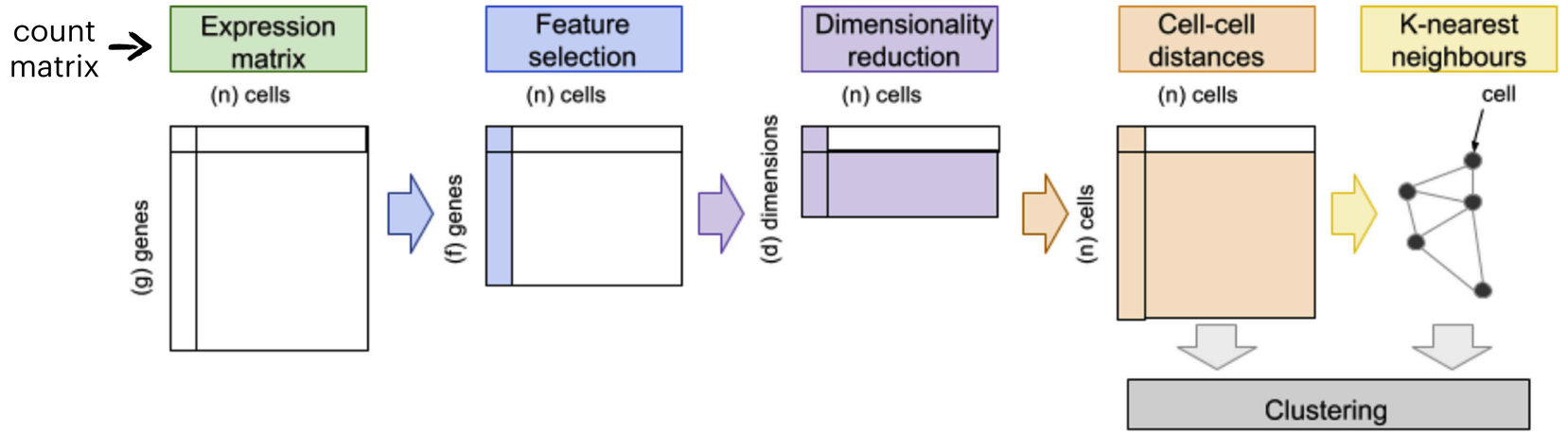
Swiss National
Science Foundation

scRNA-seq analysis at a glance

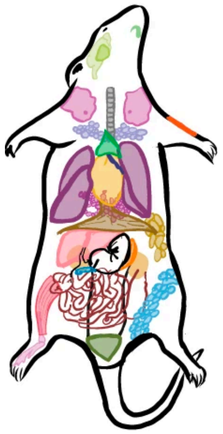




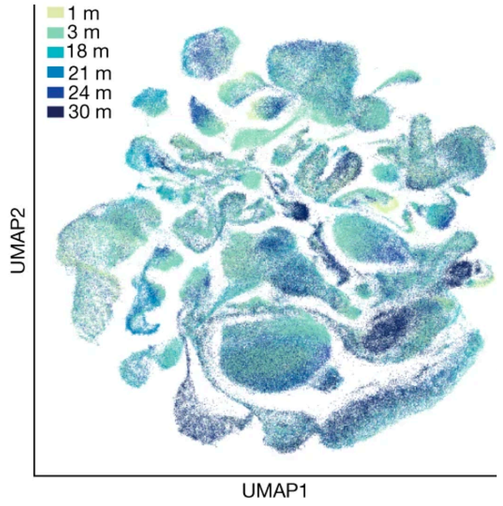
scRNA-seq analysis at a glance

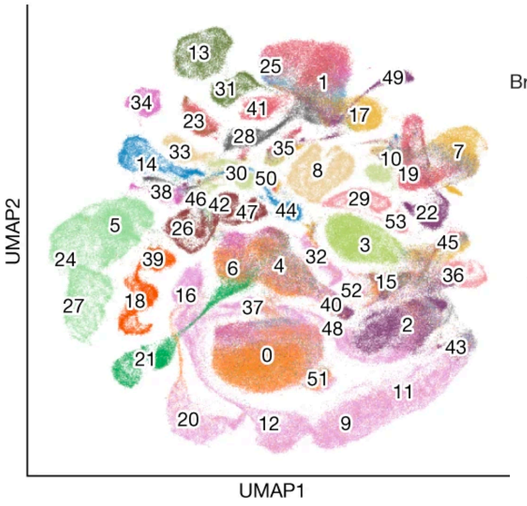


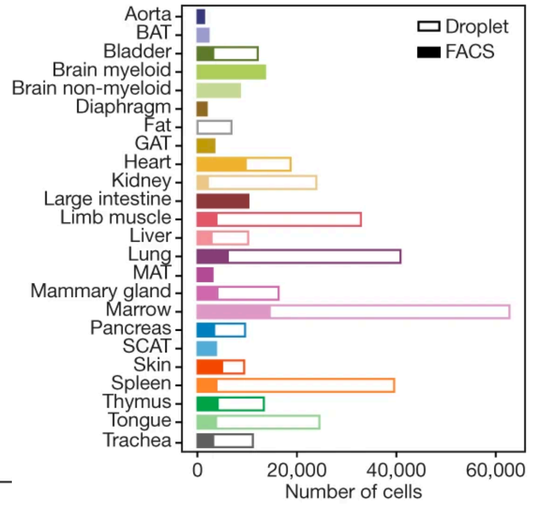
e.g., multiple tissues, timepoints, conditions, ... → *clustering* → *annotation*



The Tabula Muris Consortium (2020) *Nat.* 583, 590-595







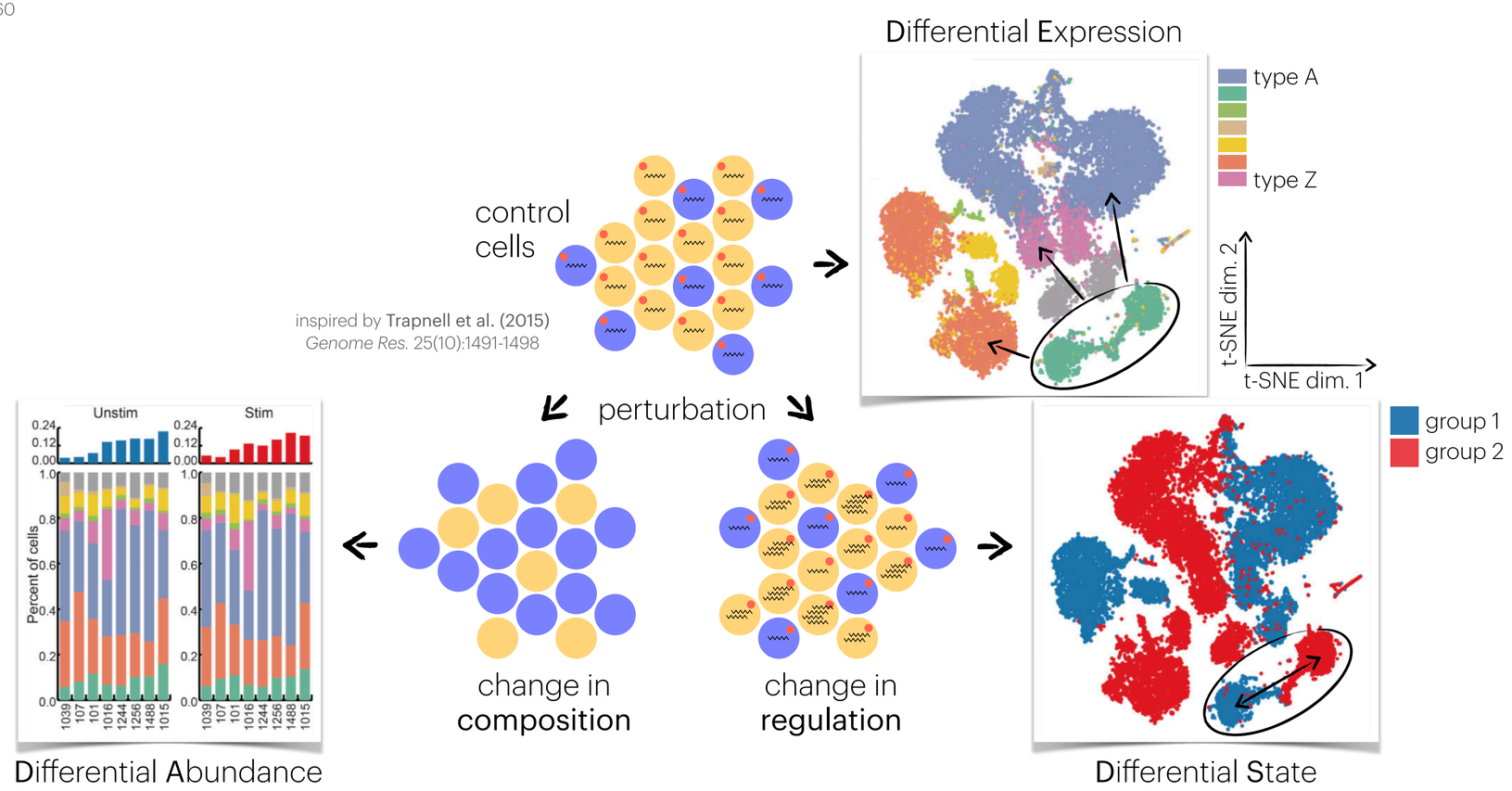
the (wrong but useful) model of **cell type & state**

The many facets of a cell's identity

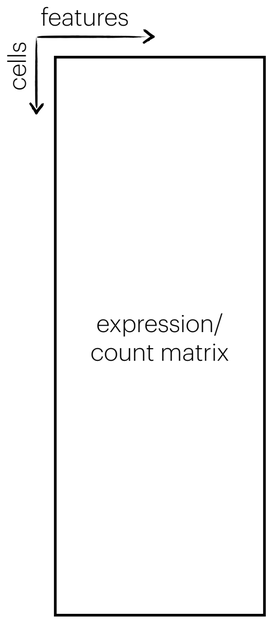
We define a cell's *identity* as the outcome of the instantaneous intersection of all factors that affect it. We refer to the more permanent aspects in a cell's identity as its *type* (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its *state*. Cell types are often organized in a hierarchical taxonomy, as types may be further

Wagner et al. (2016) *Nat. Biotechnol.* 34(11):1145-1160

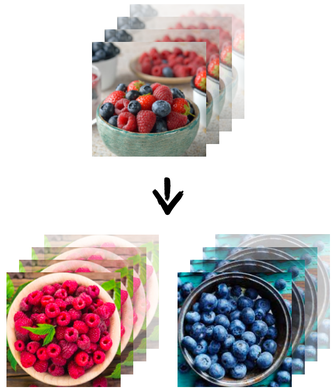
type = permanent/discrete
state = transient/continuous

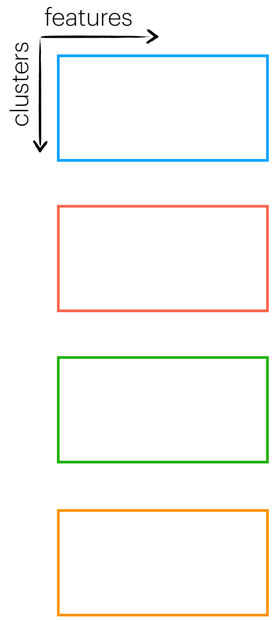


bookkeeping – DS analysis starts with...

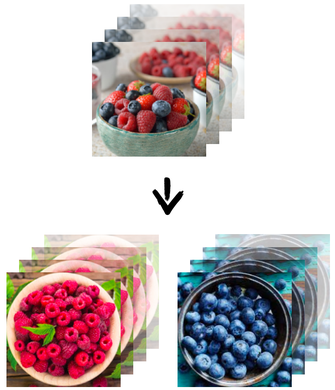


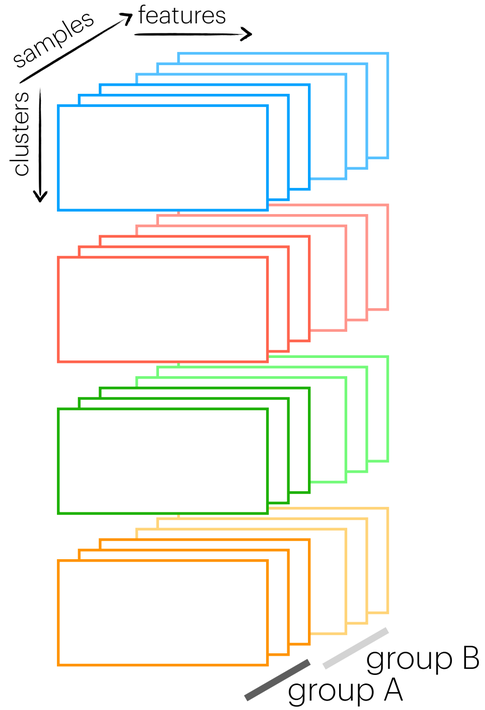
bookkeeping – DS analysis starts with...



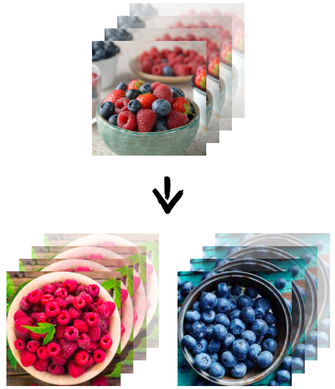


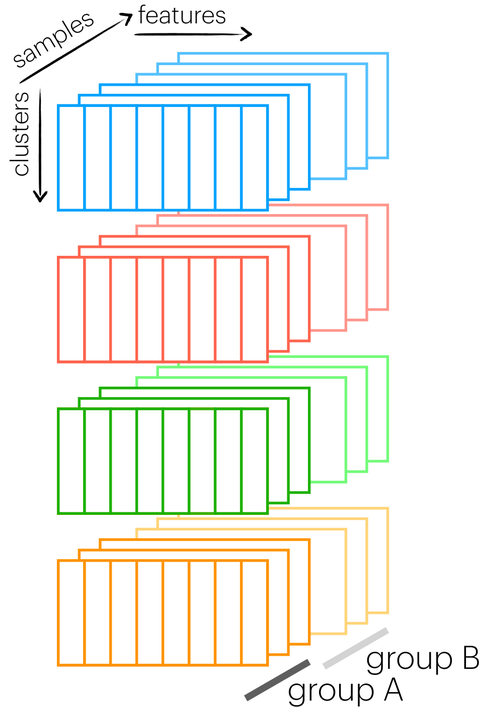
bookkeeping – DS analysis starts with...



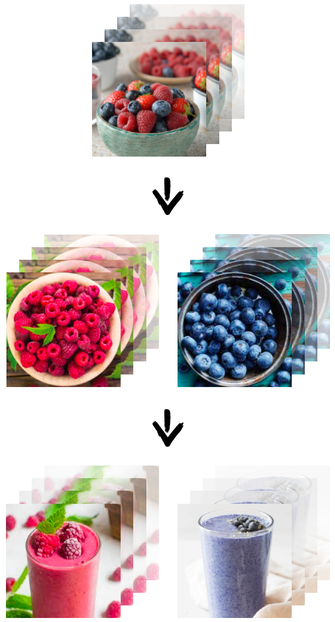


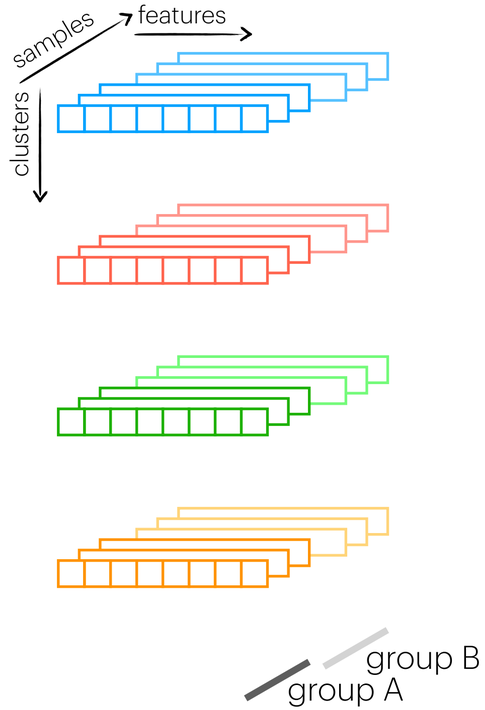
bookkeeping – DS analysis starts with...



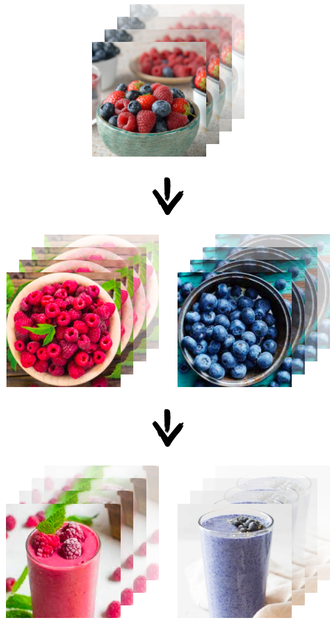


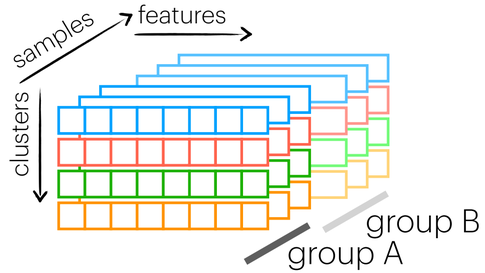
bookkeeping – DS analysis starts with...

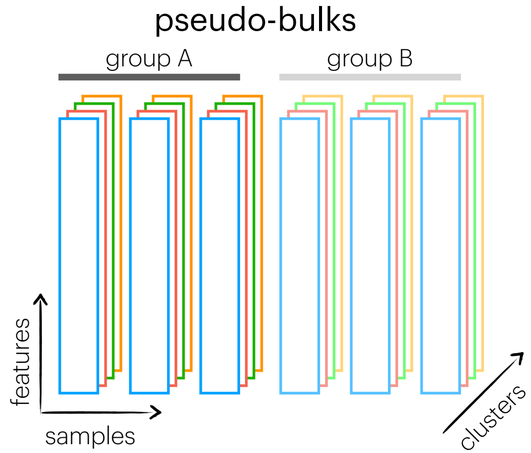




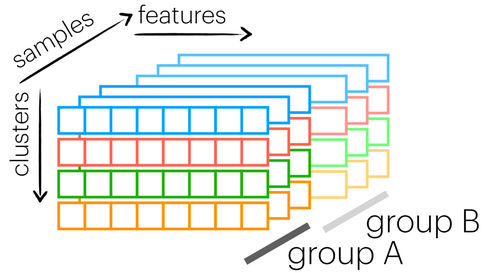
bookkeeping – DS analysis starts with...

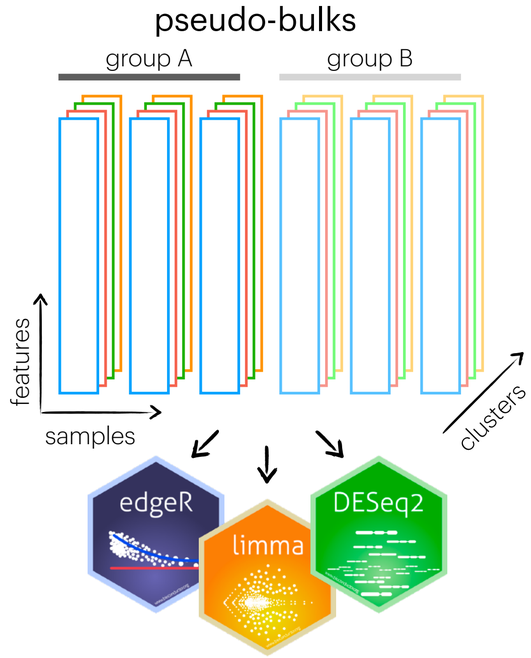




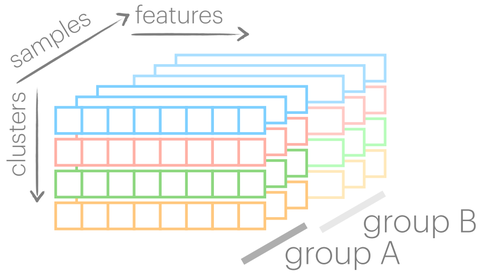


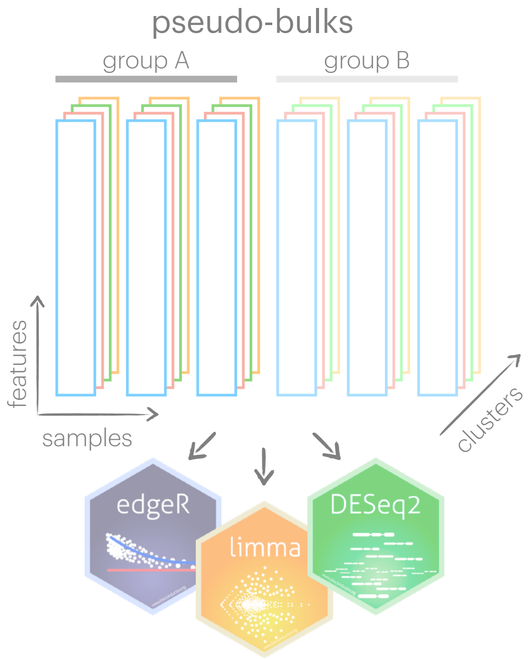
bookkeeping – DS analysis starts with...

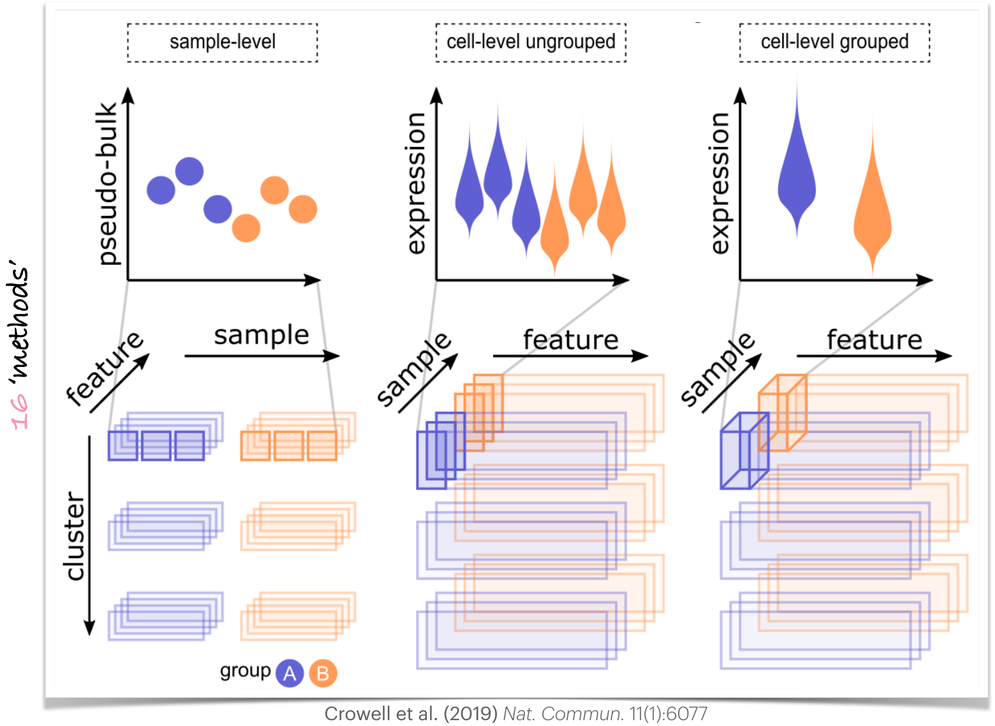




bookkeeping – DS analysis starts with...

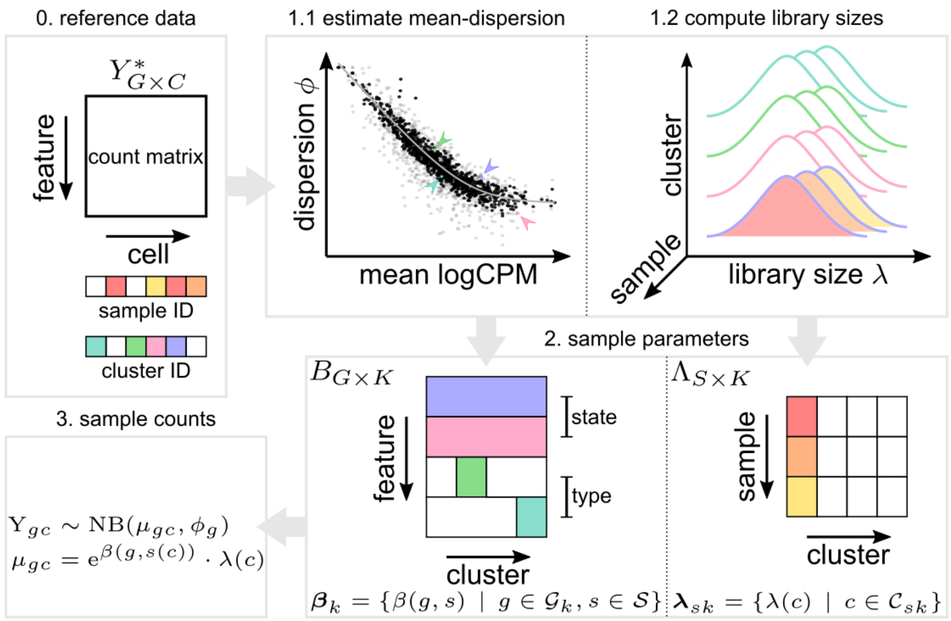






Crowell et al. (2019) *Nat. Commun.* 11(1):6077

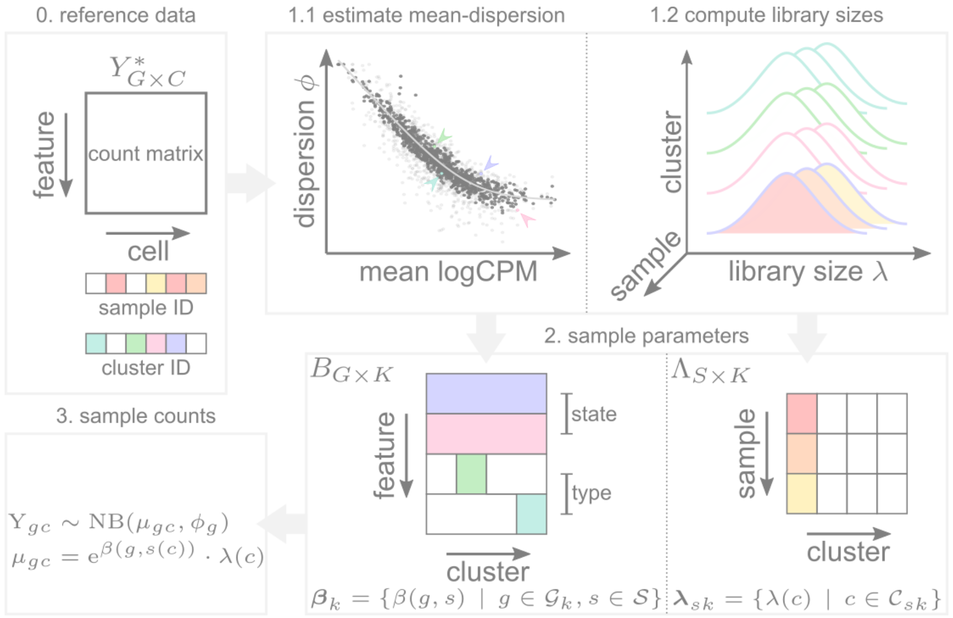
simulation framework to judge method performance

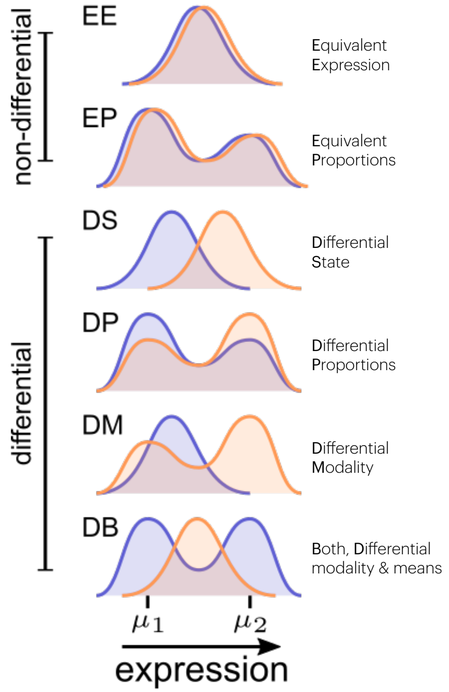


multiple clusters & samples

Crowell et al. (2020) *Nat. Commun.* 11(1):6077

simulation framework to judge method performance

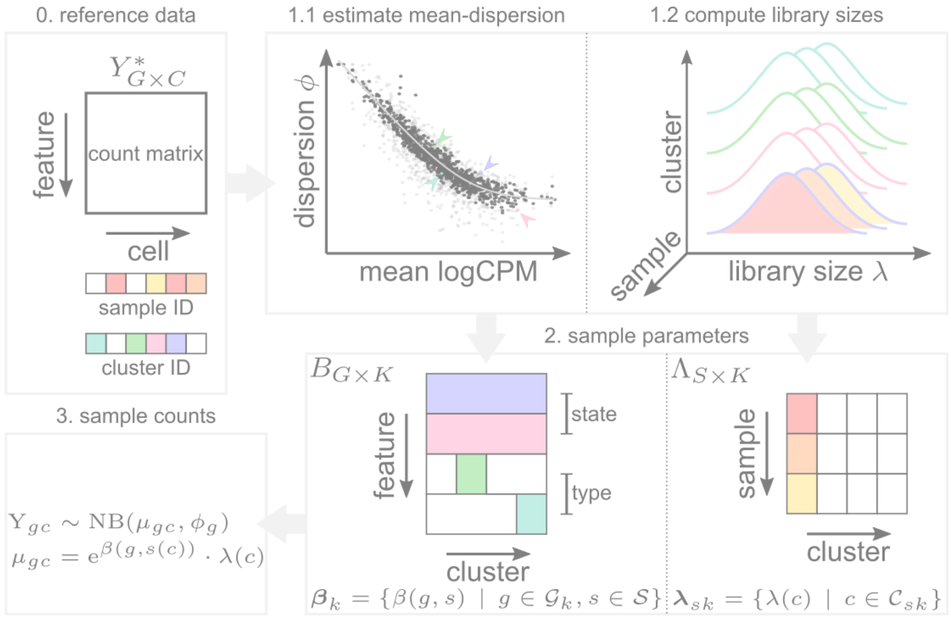


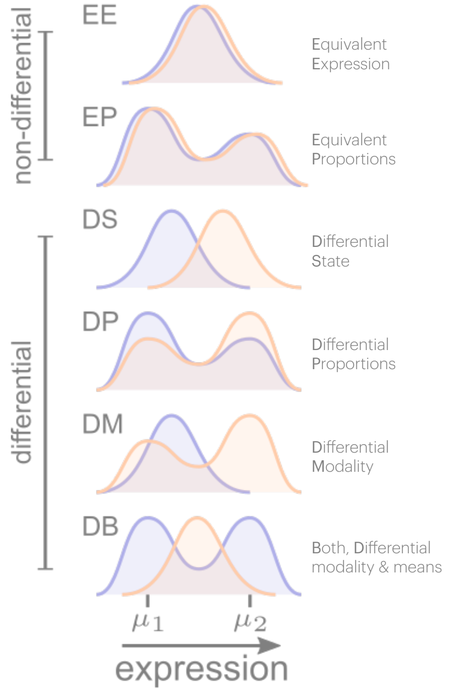


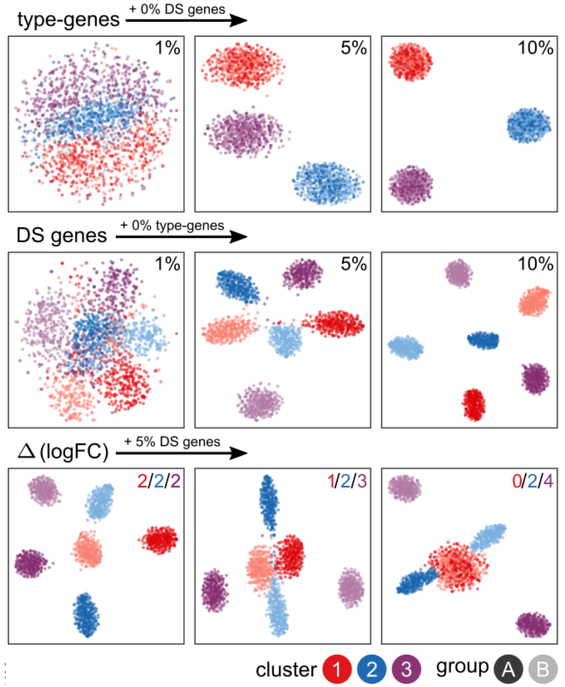
multiple clusters & samples + multiple groups

Crowell et al. (2020) *Nat. Commun.* 11(1):6077

simulation framework to judge method performance



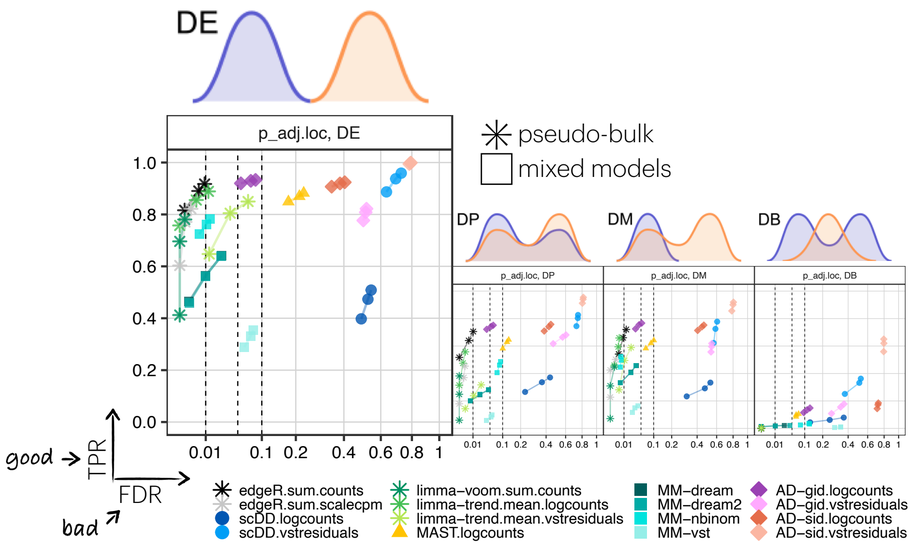




multiple clusters & samples + multiple groups → ground truth!

Crowell et al. (2020) *Nat. Commun.* 11(1):6077

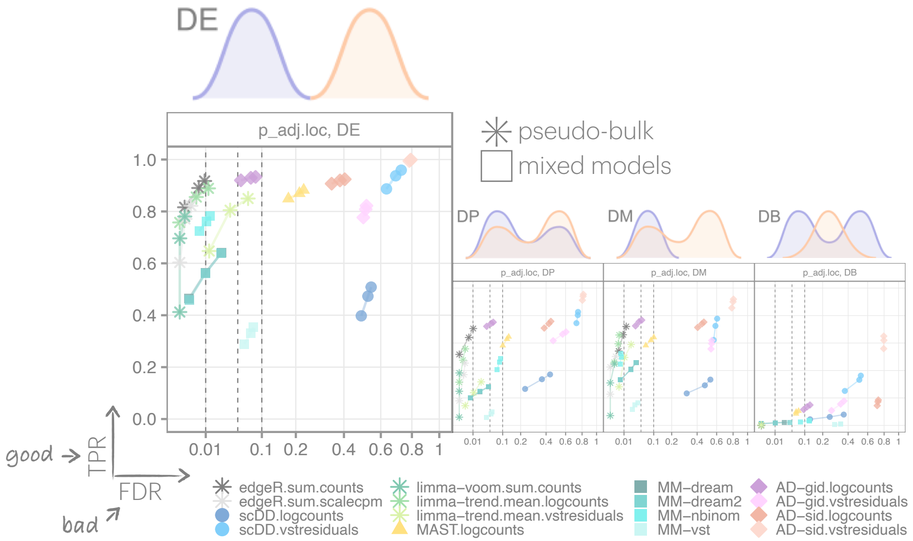
simulation framework to judge method performance



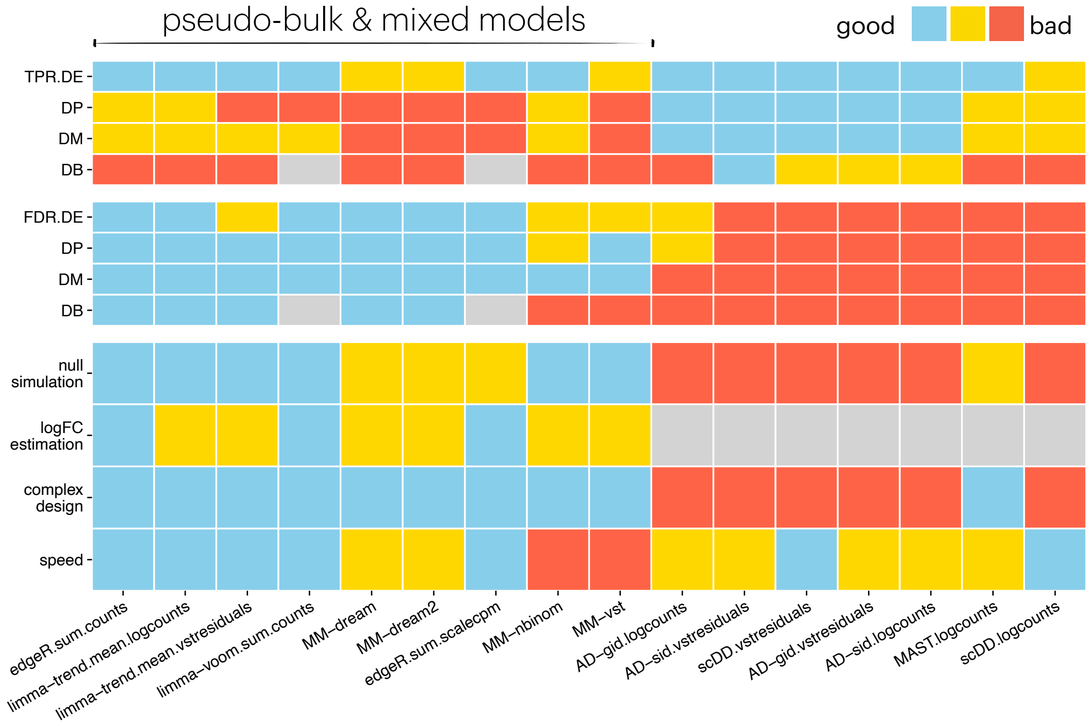
- points = TPR-FDR at FDR cutoffs 1, 5, 10%
  (16 methods x 5 replicates x 3 clusters x 4,000 genes)

Crowell et al. (2020) *Nat. Commun.* 11(1):6077

simulation framework to judge method performance



- points = TPR-FDR at FDR cutoffs 1, 5, 10%
  (16 methods x 5 replicates x 3 clusters x 4,000 genes)



A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis

Murphy & Skene (2022) *Nat. Commun.* 13:7851

Confronting false discoveries in single-cell differential expression

Squair et al. (2021) *Nat. Commun.* 12:5692

Compare healthy and diabetic samples

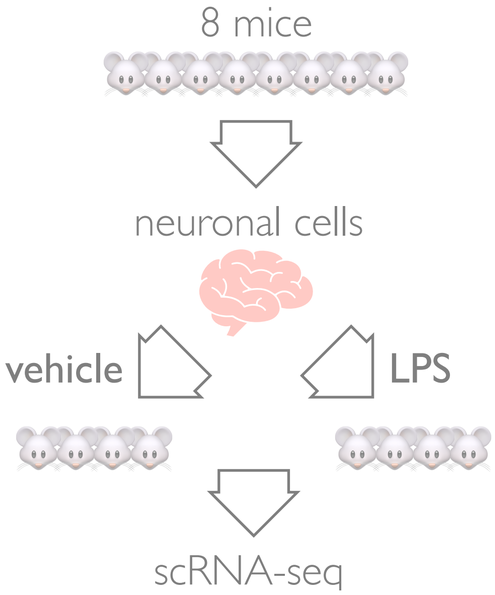
By integrating all samples together, we can now compare healthy and diabetic cells in matched cell states. To maximize statistical power, we want to use all cells - not just the sketched cells - to perform this analysis. As recommended by Soneson et al. and Crowell et al., we use an aggregation-based (pseudobulk) workflow. We aggregate all cells within the same cell type and sample using the `AggregateExpression` function. This returns a Seurat object where each 'cell' represents the pseudobulk profile of one cell type in one individual.

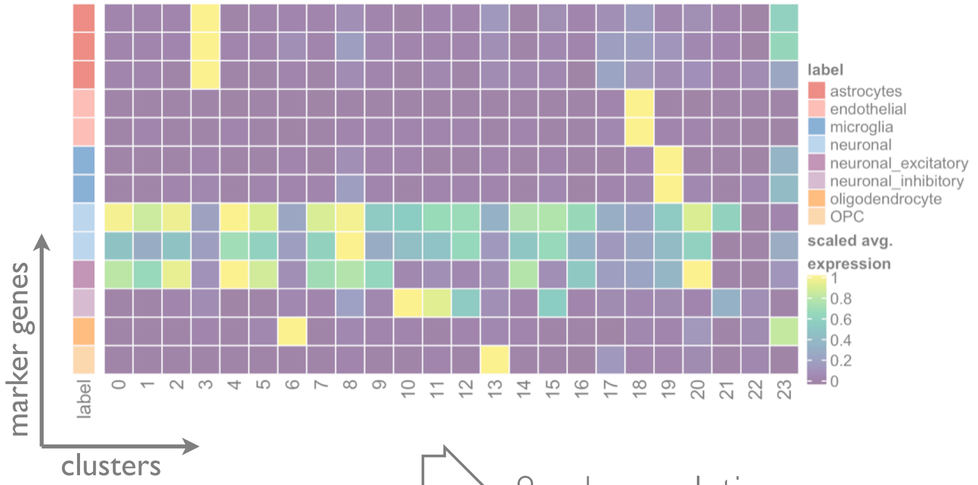
`Seurat` v5

Crowell et al. (2020) *Nat. Commun.* 11(1):6077

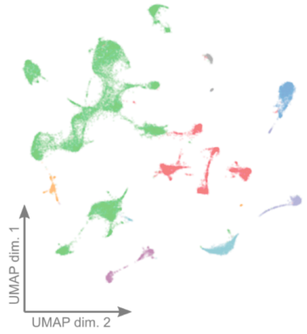
case study – vehicle vs. LPS treated mice

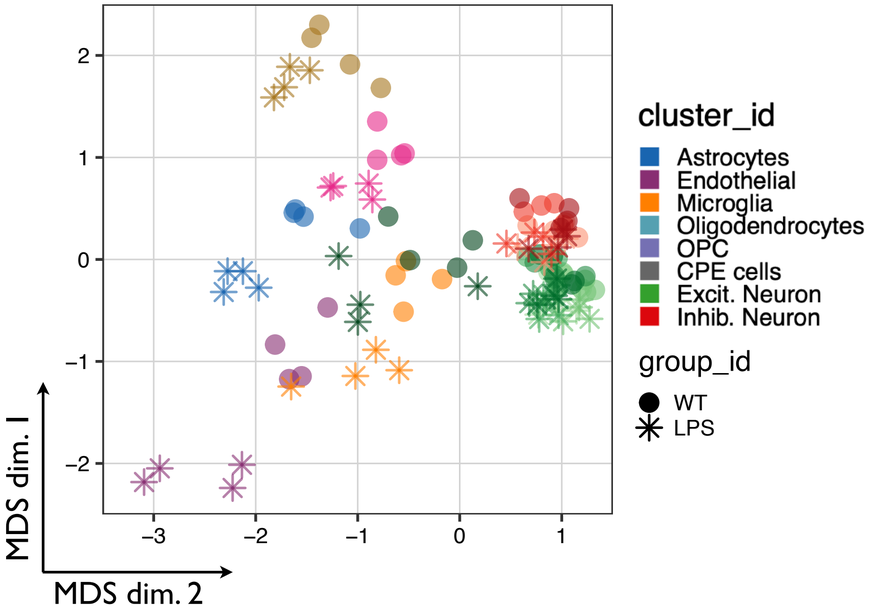
experiment → preprocessing → clustering → annotation





8 subpopulations



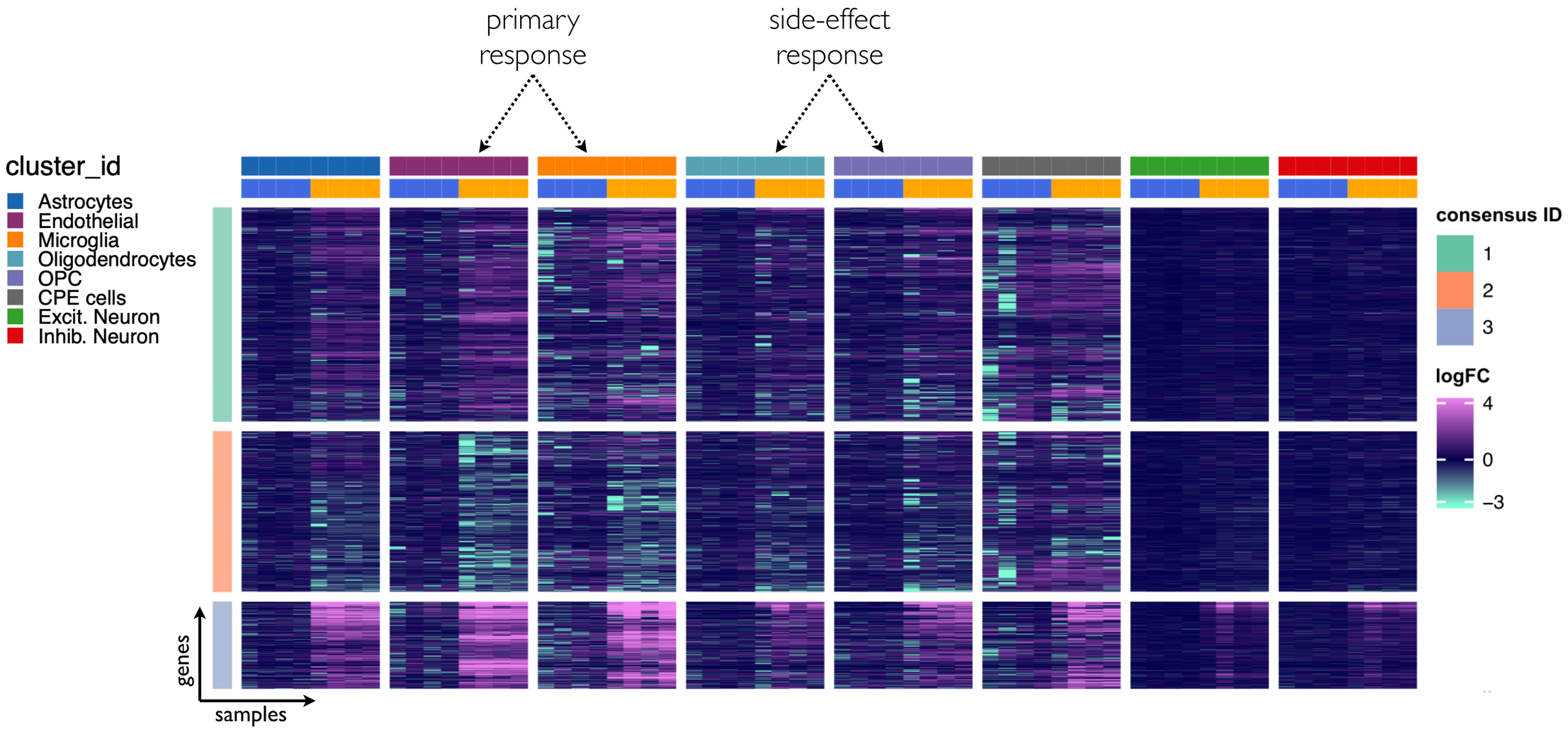


- aggregation: sum of counts
- points = cluster-sample combinations

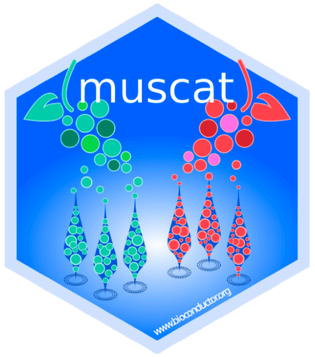
Crowell et al. (2020) *Nat. Commun.* 11(1):6077

case study – vehicle vs. LPS treated mice

experiment → preprocessing → clustering → annotation → **DS analysis**



multi-sample multi-group scRNA-seq analysis tools



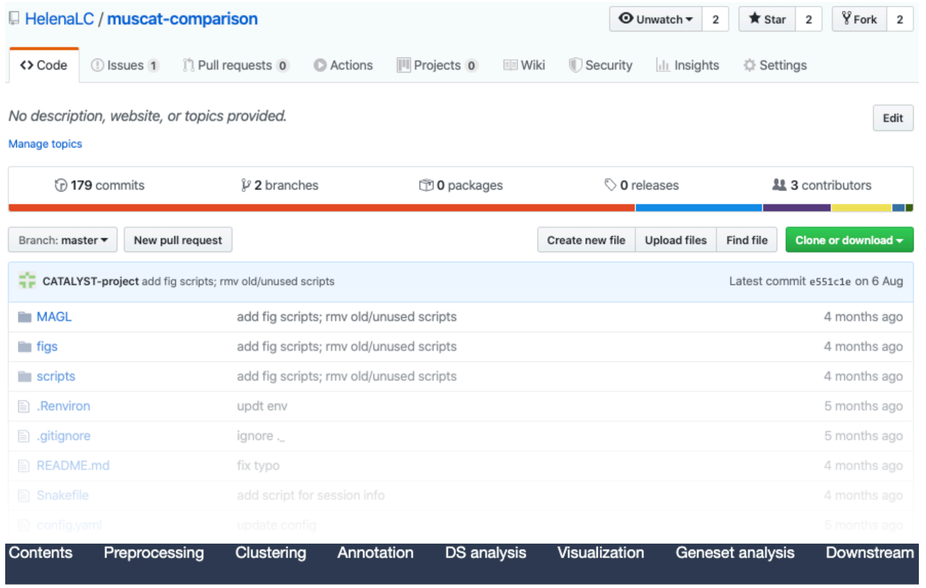
- Snakemake workflow (simulation study)
- browsable workflowr website (application to mouse cortex data)
- R/Bioconductor package
  - simulation framework
  - pseudobulk- & mixed model-based DS analysis methods
  - differential detection (DD)
  - bulk-based hypothesis weighting

muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data

Helena L. Crowell[1,2], Charlotte Soneson[1,2,3,6], Pierre-Luc Germain[1,4,6], Daniela Calini[5], Ludovic Collin[5], Catarina Raposo[5], Dheeraj Malhotra[5] & Mark D. Robinson[1,2]✉

Differential detection workflows for multi-sample single-cell RNA-seq data

Jeroen Gilis[1,2†], Laura Perin[3†], Milan Malfait[1], Helena L. Crowell[4], Koen Van den Berge[5], Alemu Takele Assefa[5], Bie Verbist[5], Davide Risso[3,6] and Lieven Clement[1*]



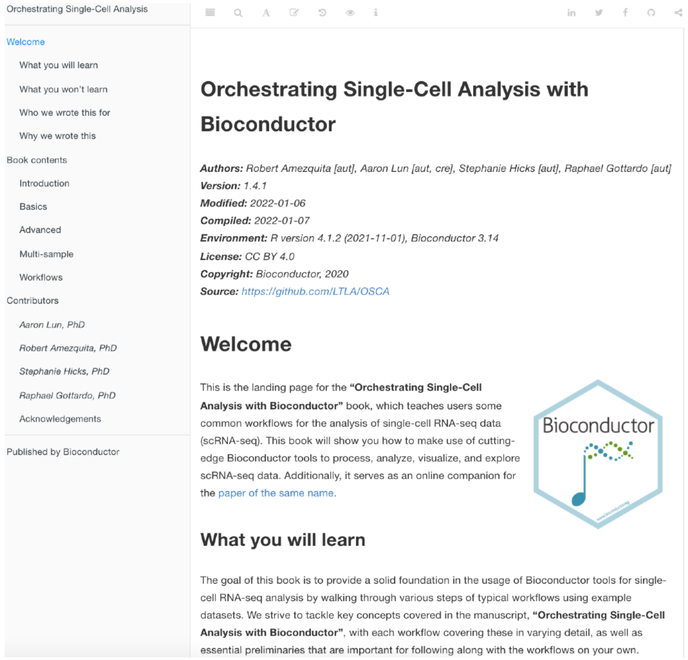
Contents

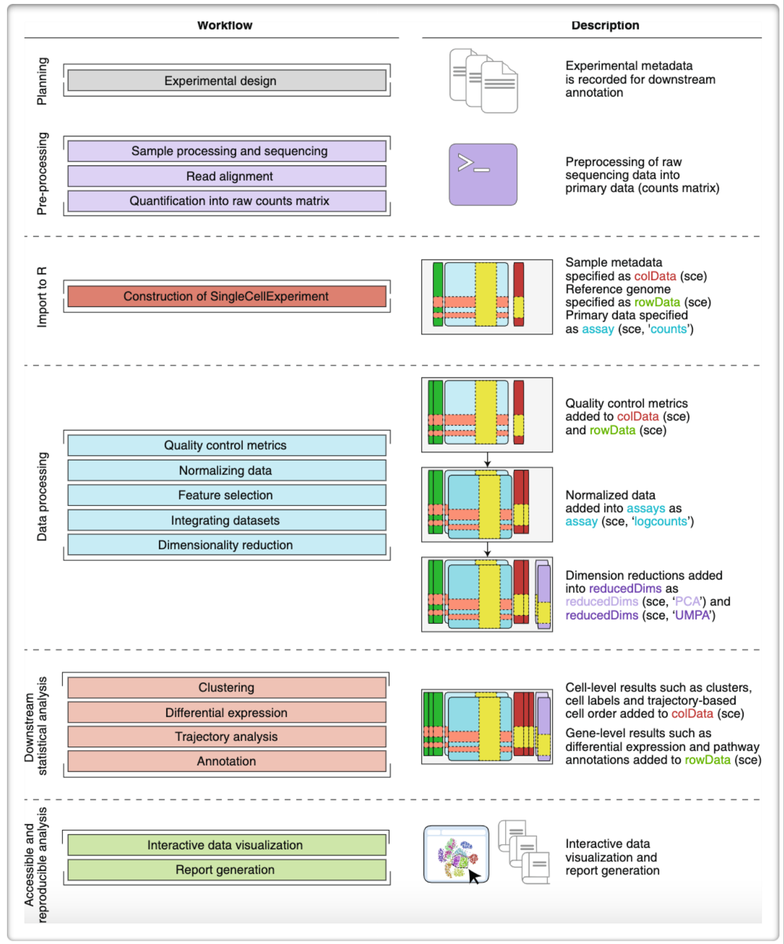
workflowr

- Preprocessing:
  - Calculation of QC metrics & identification of outliers using `scater`
  - Filtering of genes and cells
- Clustering:
  - Integration & clustering using a sequence of resolutions
  - t-SNE and UMAP dimension reductions colored by sample, group, and cluster ID
- Annotation:
  - Number of clusters by resolution
  - Number of cells by cluster-sample
  - Relative cluster abundances by sample
  - t-SNE colored by expression of known marker genes
  - Heatmap of mean known marker-gene expressions by cluster
  - Identification of cluster-markers using `scran`
  - Hetamap of mean `scran` marker-gene expressions by cluster
- DS analysis
  - Cluster annoation
  - Aggregation to pseudobulk counts
  - Pseudobulk-level MDS plot
  - Cluster-level DE analysis with `edgeR`
  - Results filtering & overview
  - Dimension reduction: t-SNE & UMAP

OSCA – Orchestrating Single-Cell Analysis with Bioconductor

- comprehensive **online book**
  "covering installation, sources of help, specialized topics pretaining to specific aspects of scRNA-seq analysis and complete workflows [...]"





on differential discovery in scRNA-seq data

...using muscat & miloDE/lemur

Helena L Crowell, PhD

Autumn School for Single Cell-ers

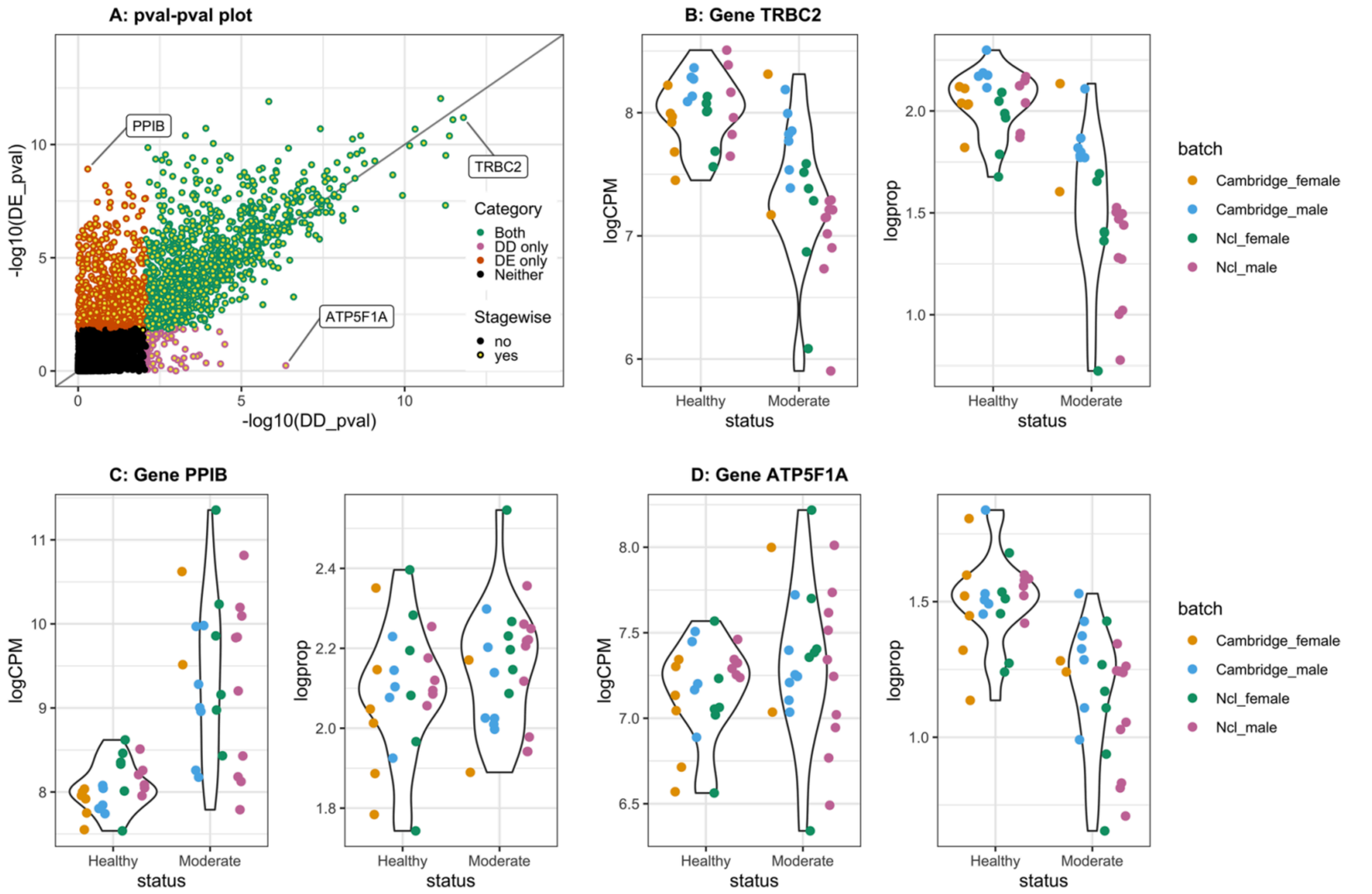
Oct 21, 2025 · GIMM, Oeiras, Portugal

CNAG – National Center
for Genomic Analysis
Barcelona, Spain

Swiss National
Science Foundation

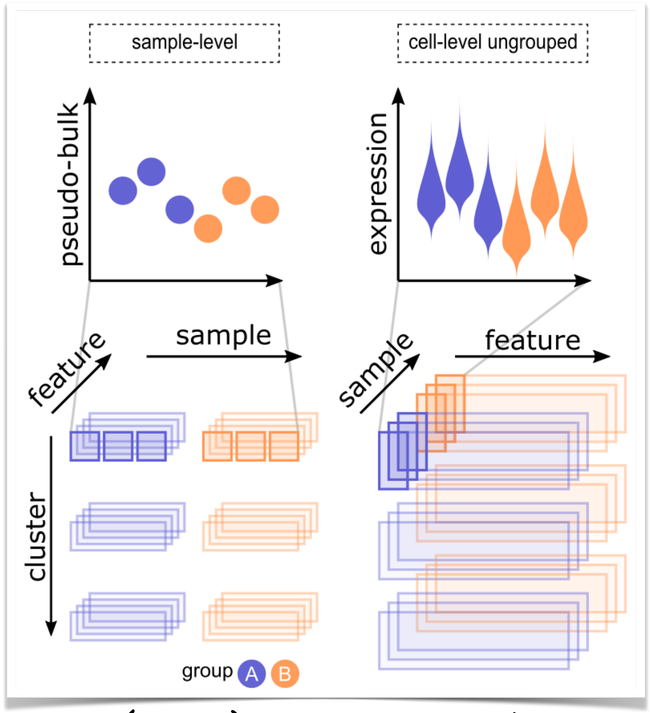
beyond expression – differential detection (DD)

One particularly interesting characteristic of gene expression not explicitly captured by the aforementioned frameworks is differential detection (DD), i.e. finding differences in the fraction of cells in which a gene is detected between groups. It has been reported that gene expression profiles may exhibit characteristic bimodal expression patterns, in which the expression of genes is either strongly positive or undetected within individual cells (e.g., Finak et al. [4]). At the single-cell level, such differences in gene detection may arise from technical artifacts or from the stochastic nature of gene expression [5], a phenomenon commonly referred to as transcriptional bursting [6]. At the tissue level, differential detection can be biologically meaningful as well. For instance, even if the overall expression level of a gene in a tissue is equal between two biological conditions, it is relevant to assess the fraction of cells in the tissue contributing to the observed expression levels. In addition, previous research has shown that assessing differences in detection may provide a more robust means of studying gene expression than assessing differences in the average gene expression.



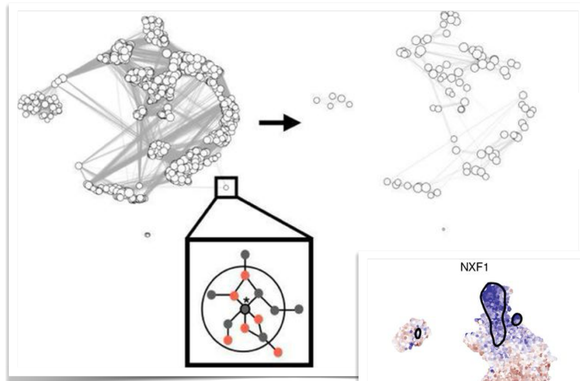
more on DA/S analysis – cluster-based vs. -free approaches

Crowell et al. (2019) *Nat. Commun.* 11(1):6077



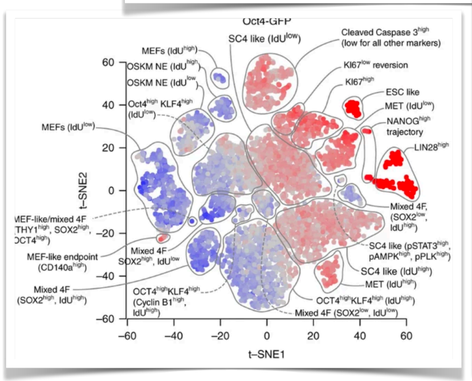
a priori **grouping** of cells

- clustering is not easy...
- ...but more interpretable



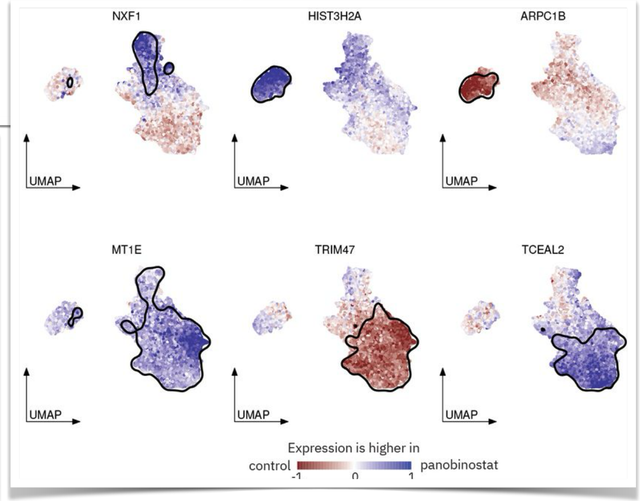
miloDE

Missarova et al. (2023) *bioRxiv*



cydar

Lun et al. (2017) *Nat. Methods* 14(7):707-709



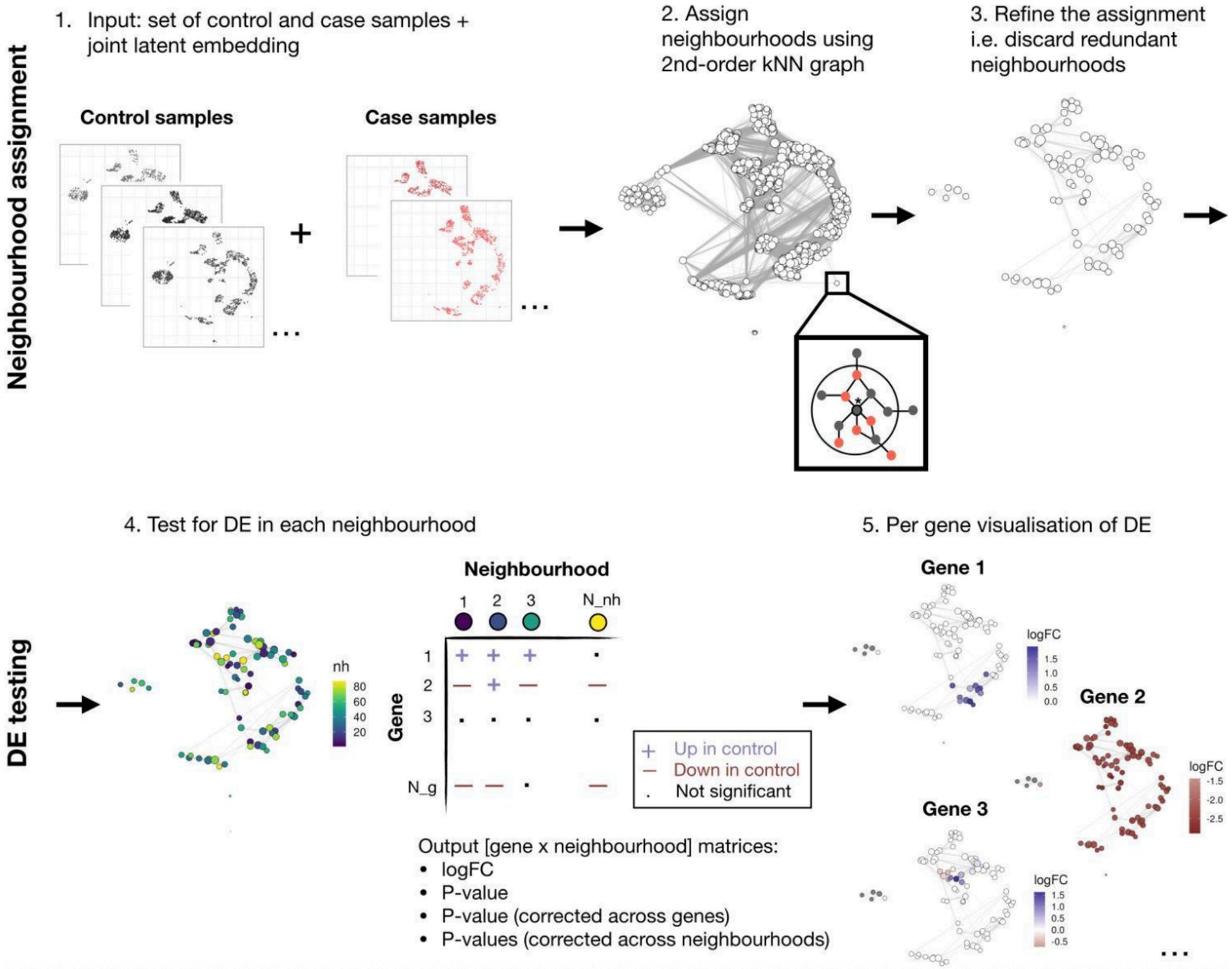
LEMUR

Ahlmann-Eltze & Hube (2023) *bioRxiv*

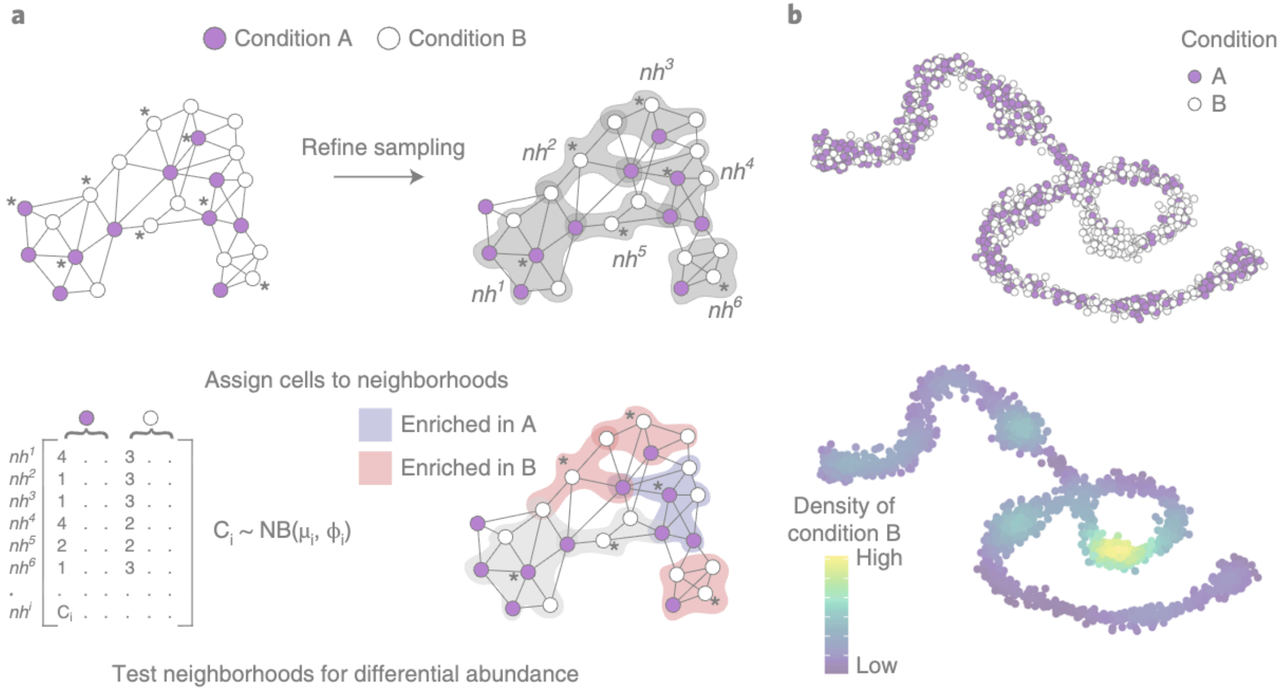
interesting **neighborhoods** are identified in a **data-driven** way

- no need for clustering...
- ...but need to make (biological) sense of results downstream

a closer look at **miloDE**



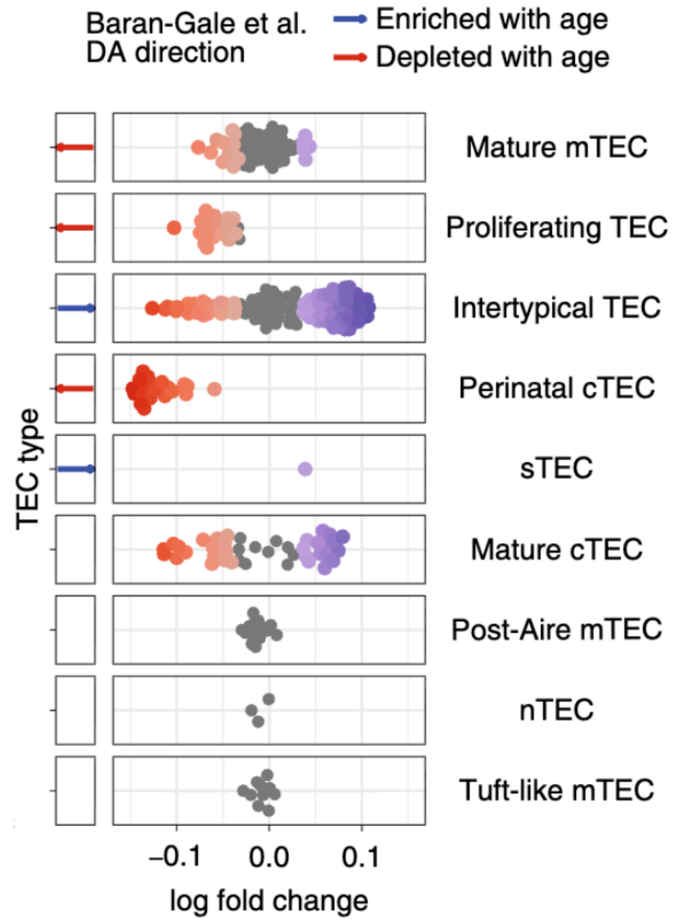
before miloDE, there was **milo(DA)** – *interpretability?*



stratification of cell-level results by subpopulation →

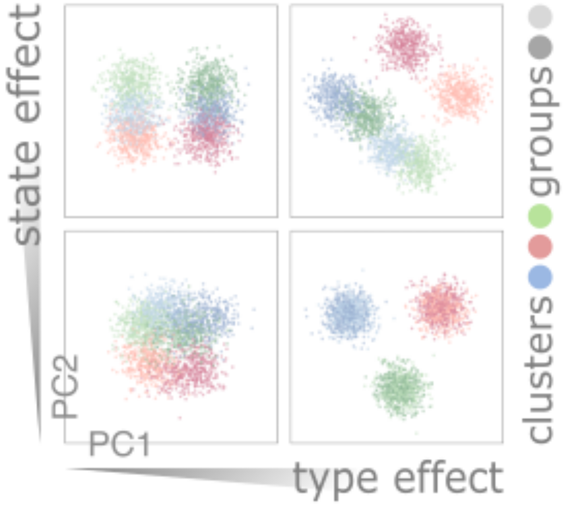
for miloDE: we have this x the number of genes!

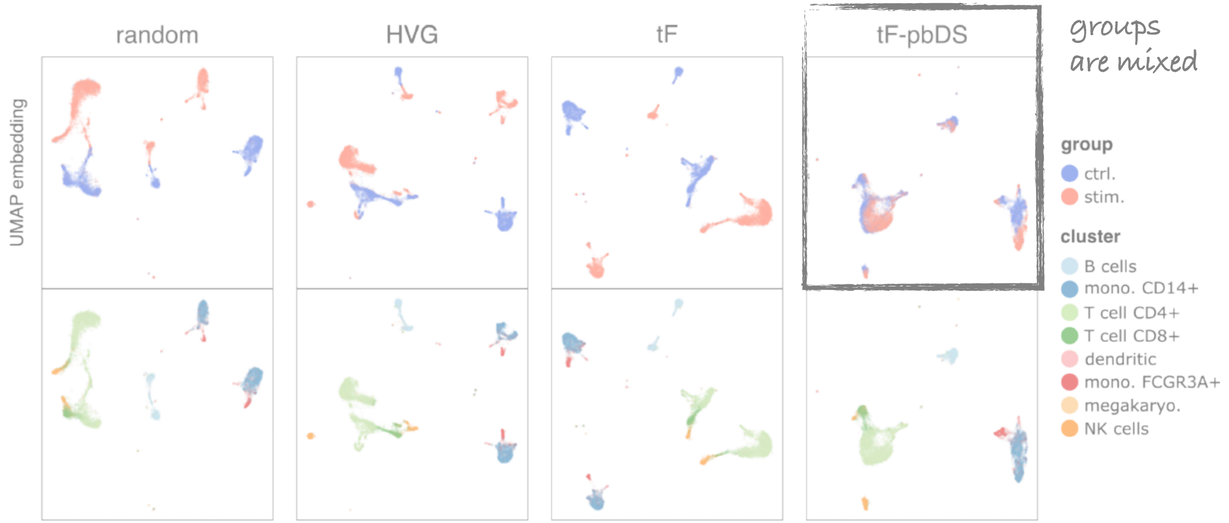
(point = cell, facet = cluster, color = logFC)



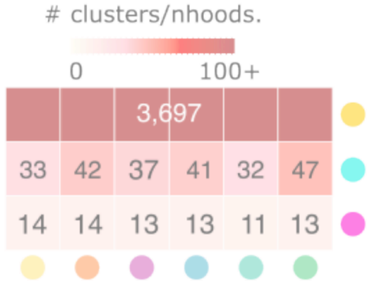
feature selection can disentangle cell type & state

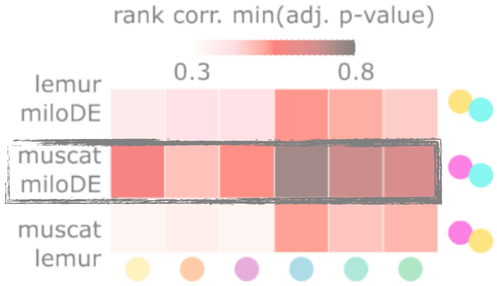
the features we put into PCs determine what we get out!
(in terms of clustering, etc.)



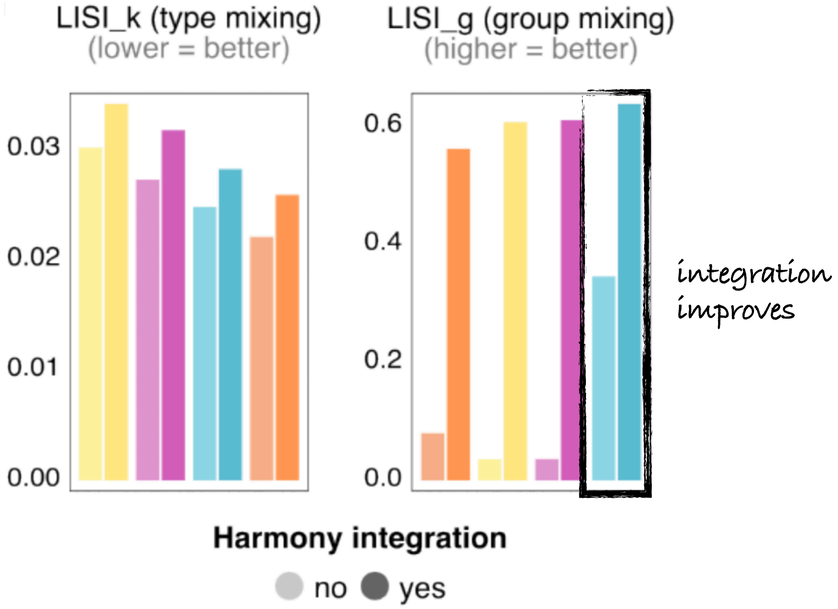


groups are mixed





cluster-free & -based DS methods comparable



integration improves

take home **message** / Gedankenexperiment

There is arguably an interplay between not only the type and state transcriptional programs governing cells, but also between the various types of differential analysis that are applied. While the readily-available DAA and DSA tools can in principle be considered orthogonal, they are often linked by how DE drives subpopulation definition[17]. As a toy example, consider a substance that strongly increases the expression of a gene γ in all treated (but not control) cells. If γ were used for clustering (or neighborhood generation), γ^- and γ^+ subpopulations should be obtained for each cell type; and, DAA analysis would detect γ^+ and γ^- populations as differentially abundant. In contrast, if γ were *not* contributing to the subpopulation definition, γ^+ and γ^- cells should be present in every cluster (whose markers are expressed independently from γ), and a DSA analysis should detect γ as higher expressed in treated cells (i.e., a differential state) for each cell type.

Wang, Crowell & Robinson (2025) *BMC Genomics*, in prep.