Introduction to single-cell RNA-seq

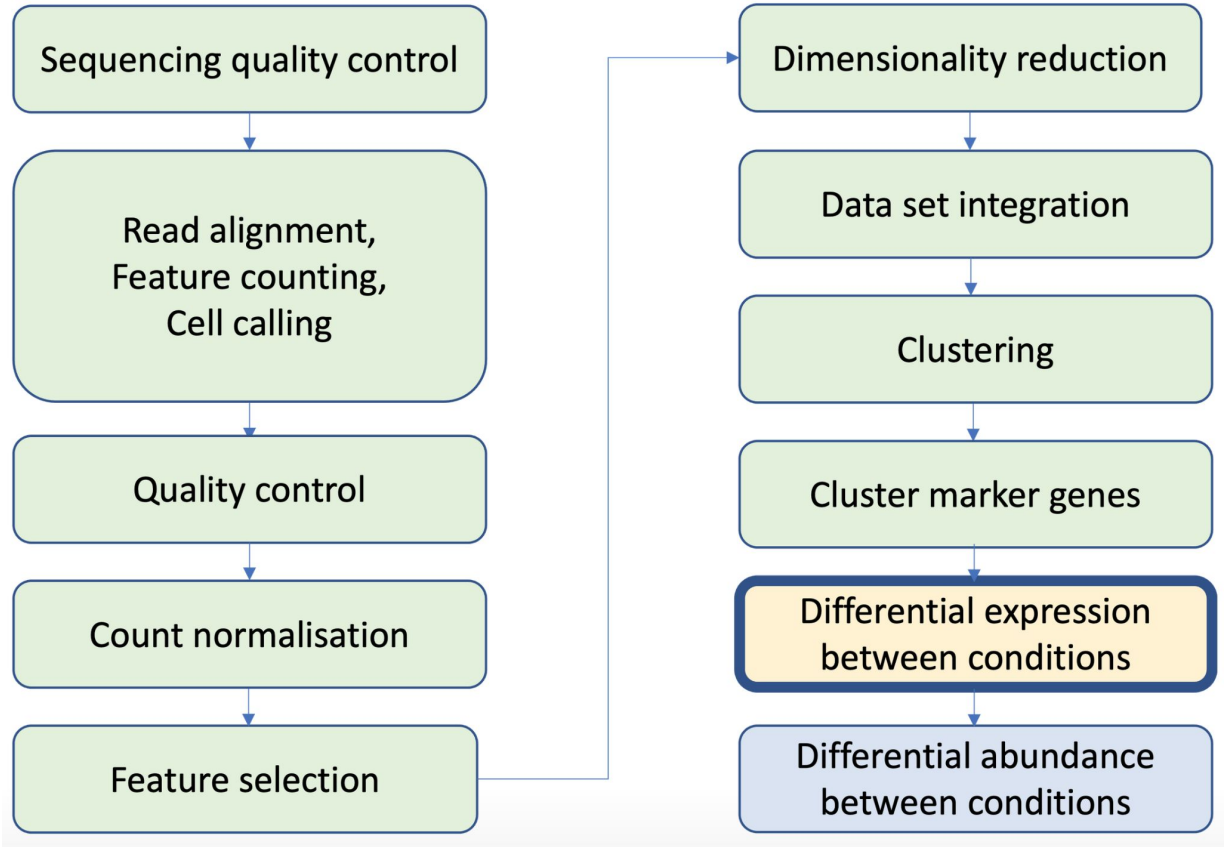
Differential Expression and Abundance

UNIVERSITY OF CAMBRIDGE

Bioinformatics Training

CANCER RESEARCH UK | CAMBRIDGE CENTRE

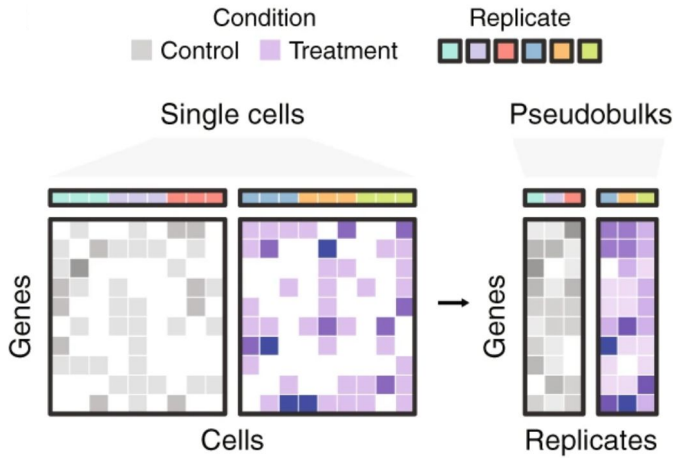
Analysis Workflow



Differential Expression - Pseudo-bulk Method

Test for significant changes in **gene expression** between conditions.

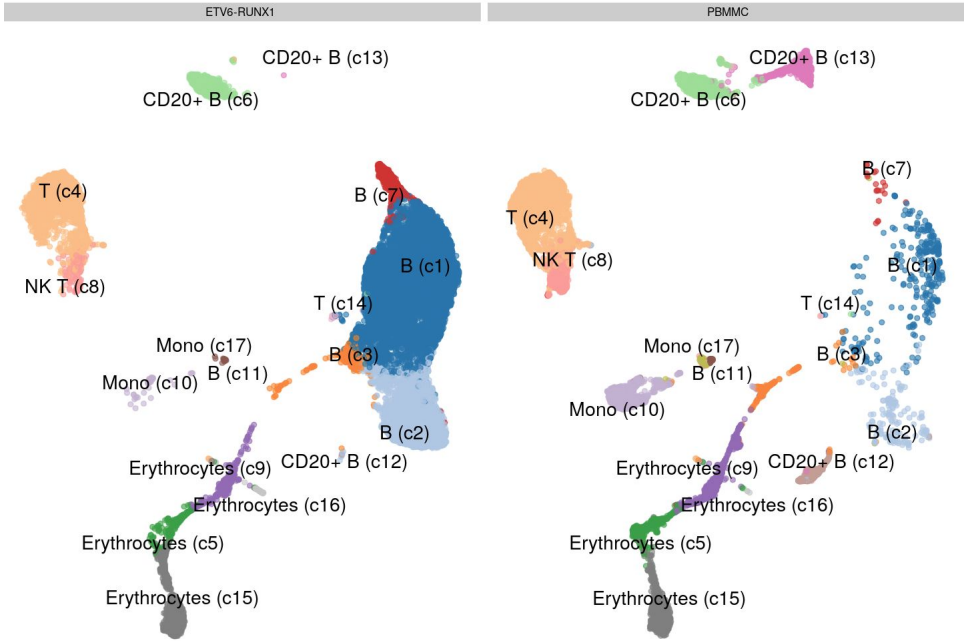
- Are any genes high- or down-regulated between *treated vs control* or *wild-type vs mutant* or *healthy vs disease*, etc.



- Create pseudo-bulk samples by summing raw counts across cells for each sample
- Apply standard bulk RNA-seq DE methods (*edgeR*, *DESeq2*, *limma*)

Benchmark study for differential expression methods in scRNA-seq:
Squair, J.W., Gautier, M., Kathe, C. *et al.* (2021) *Nature Communications* https://doi.org/10.1038/s41467-021-25960-2

Differential Expression - Pseudo-bulk Method



- Create pseudo-bulk samples by summing counts across cells for each **sample and cell type/cluster combination**
- Apply standard bulk RNA-seq DE methods (*edgeR*, *DESeq2*, *limma*)

Differential Expression - Pseudo-bulk Method

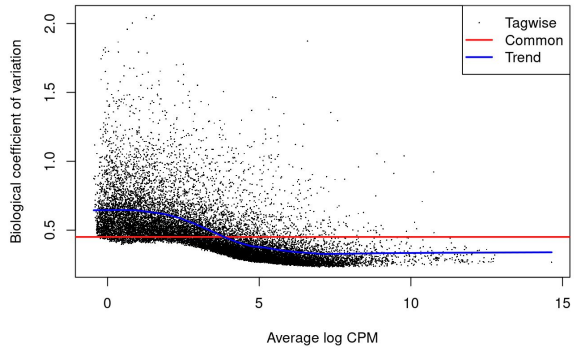
Cells are not biological replicates

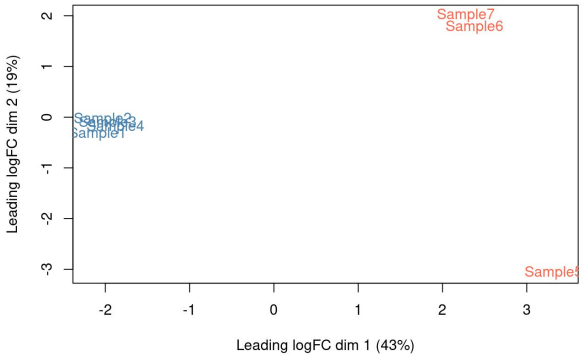
- Single cells within a sample are not independent of each other.
- Using cells as replicates amounts to studying variation inside an individual.
- We want to study variation across a population of individuals.

Differential Expression - Workflow

- Create **pseudo-bulk samples** → `aggregateAcrossCells()`
- **Filter** low-count samples/genes
  - Pseudo-bulks (samples) with very low number of cells (e.g. < 20)
  - Genes with very few counts (this is done internally with `edgeR::filterByExpr()`)
- Run **DE analysis** → `scran::pseudoBulkDGE()` (uses edgeR package)
  - Calculates **normalisation factors** to account for transcript composition differences across pseudo-bulk samples → `edgeR::calcNormFactors()`
  - Estimates **mean-dispersion** relationship across genes → `edgeR::estimateDisp()`
  - **Fits linear model** to the data → `edgeR::glmQLFit()`

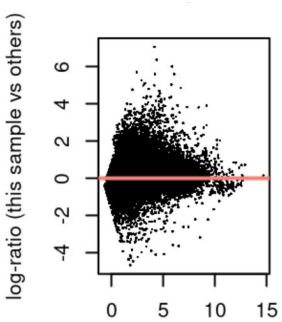
Differential Expression - Workflow





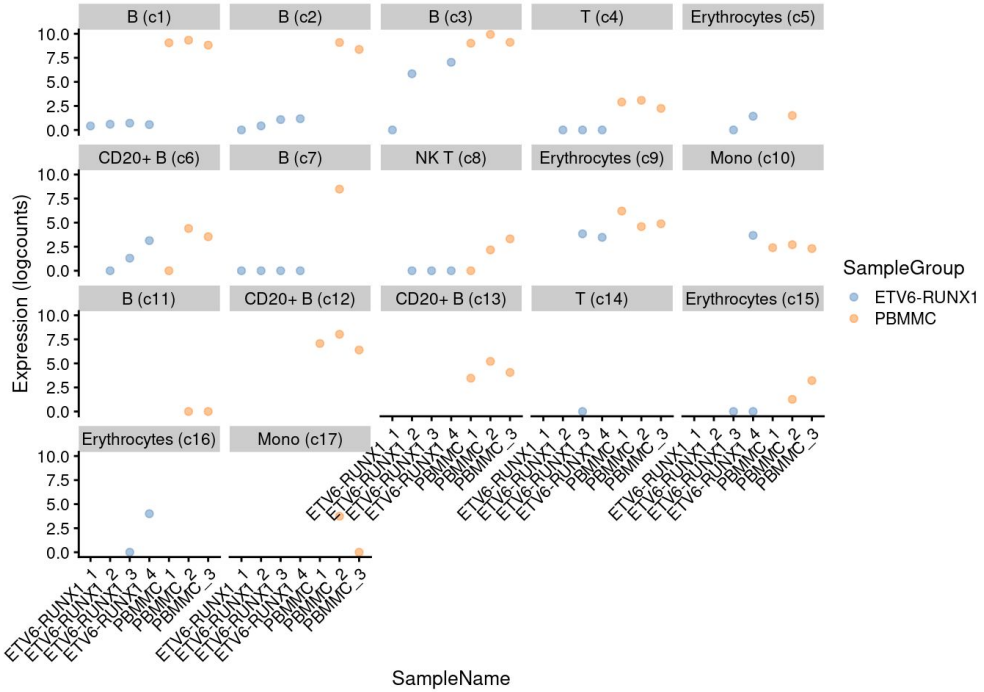
Once we have pseudo-bulks, the analysis is identical to standard bulk RNA-seq analysis

- Statistical models account for the mean-variance relationship observed in RNA-seq data
- Dimensionality reduction methods can be used to visualise how our samples cluster together
- Mean-difference plots show if library size normalisation was successful



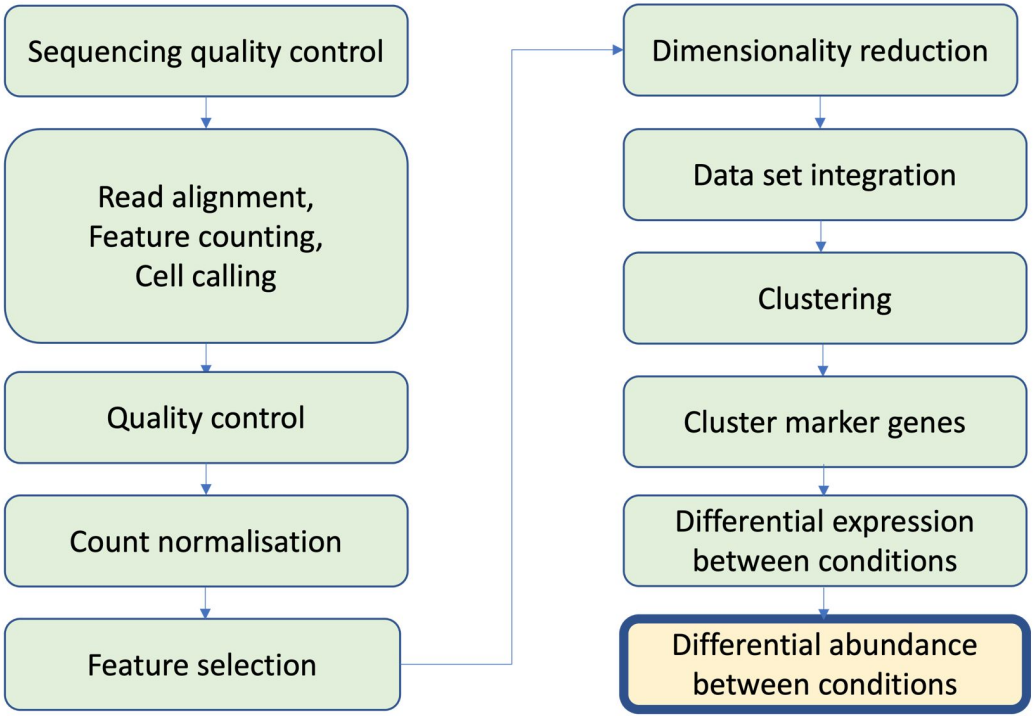
Differential Expression - Workflow

One difference from standard bulk analysis is that we have comparisons *per cell label* and so we need to decide which results we want to extract from our analysis.



Demo

Analysis Workflow

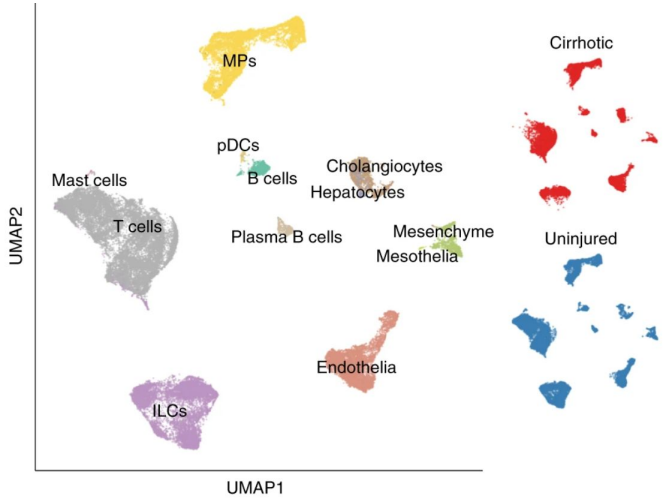


Differential Abundance

Test for significant changes in **cell abundance** across conditions.

- Are any cells enriched/depleted between *treated vs control* or *wild-type vs mutant* or *healthy vs disease*, etc.

A simple approach is to count how many cells there are in each cluster in each sample group and do a test to compare those counts.



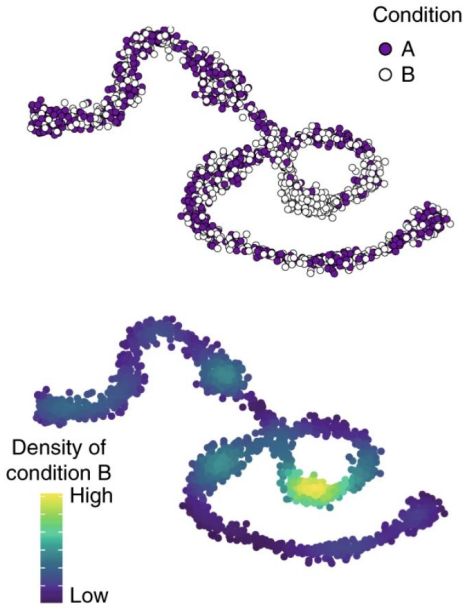
Differential Abundance

Test for significant changes in **cell abundance** across conditions.

- Are any cells enriched/depleted between *treated vs control* or *wild-type vs mutant* or *healthy vs disease*, etc.

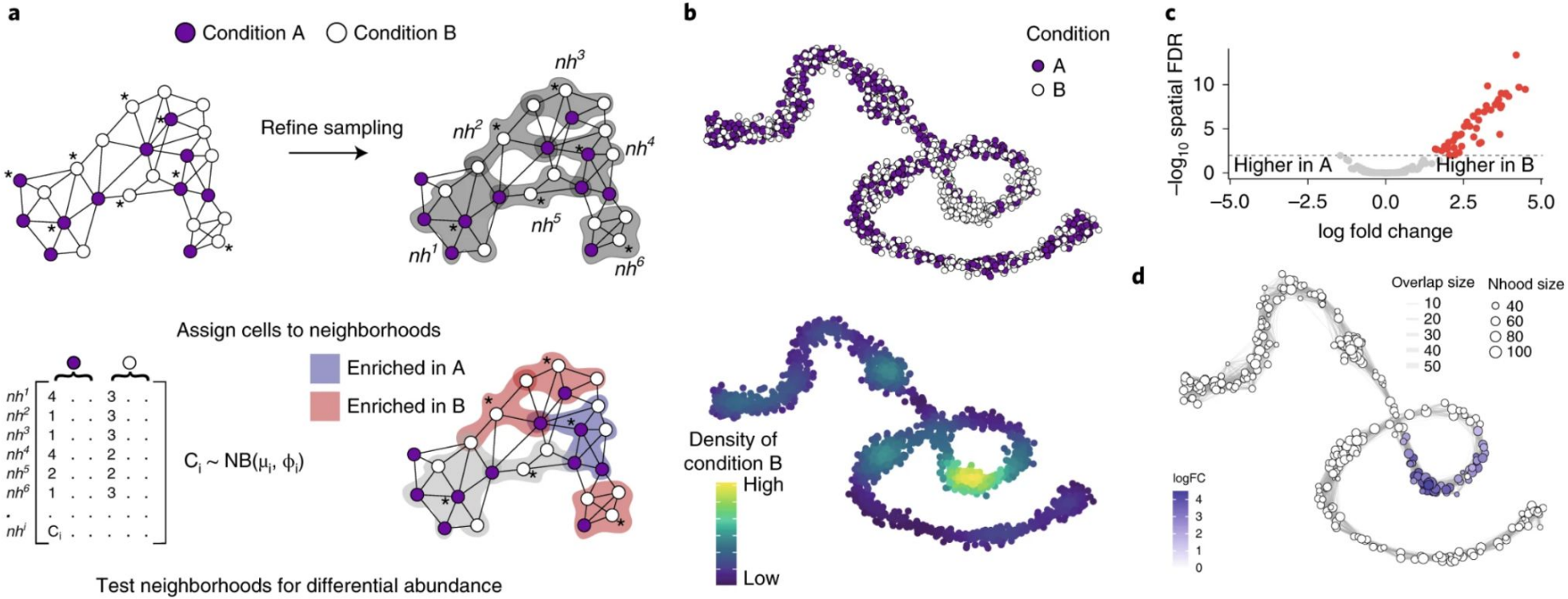
Methods that require **pre-defined clusters as input** are limited in the context of continuous differentiation, developmental or stimulation trajectories, non-discrete cell states.

Milo is a method that overcomes these limitations by performing differential abundance tests in local cell neighbourhoods



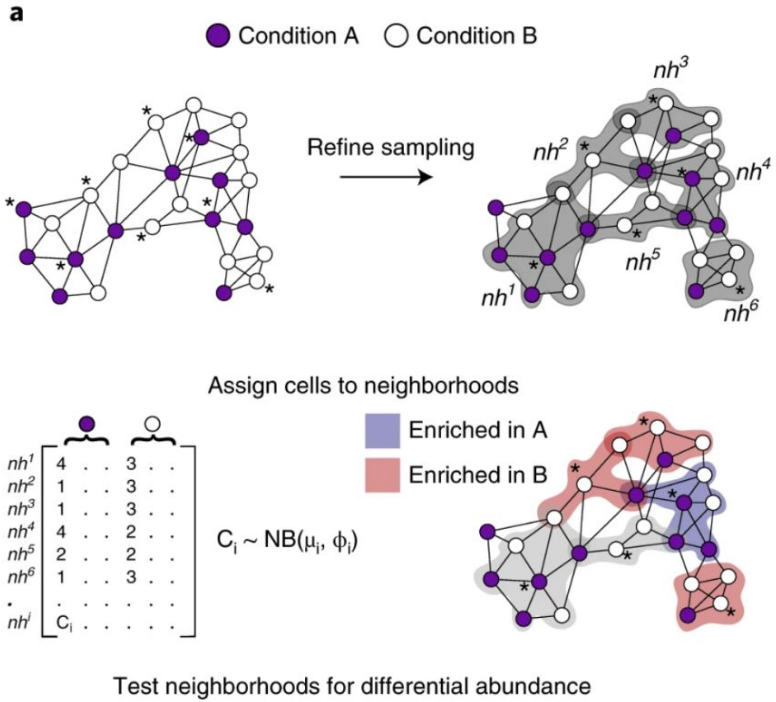
Paper on Milo method: Dann, E., Henderson, N.C., Teichmann, S.A. et al (2022) Nature Biotechnology. https://doi.org/10.1038/s41587-021-01033-z

Differential Abundance - Milo



Paper on Milo method: Dann, E., Henderson, N.C., Teichmann, S.A. et al (2022) Nature Biotechnology. https://doi.org/10.1038/s41587-021-01033-z

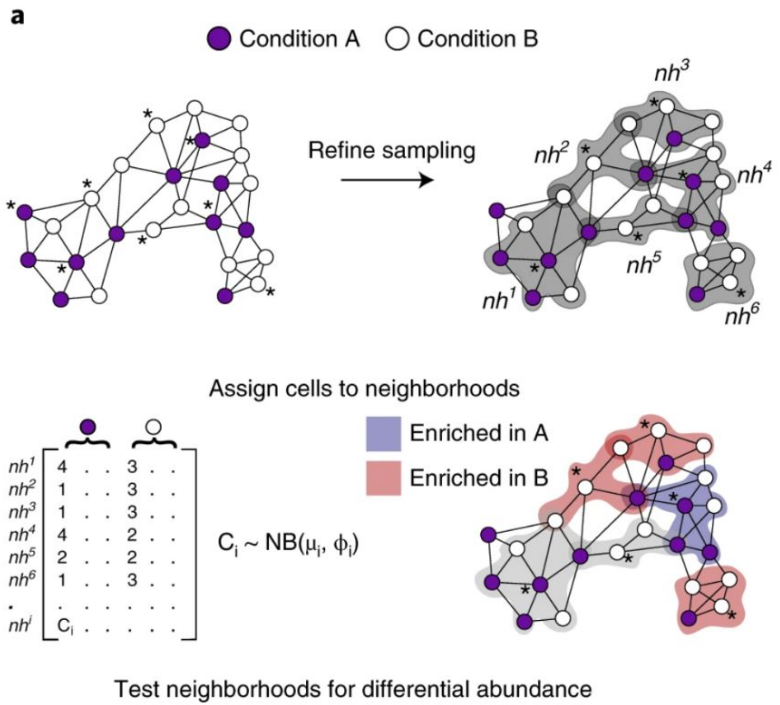
Differential Abundance - Milo



- Uses K-nearest neighbour graph to model cellular states as overlapping neighbourhoods.
- Spatial non-independence of the tests is accounted for with a weighted version of the Benjamini–Hochberg FDR method.
- Determines neighbourhoods and groupings independently of our defined clusters.
- Can be used for complex models.
- Fast and scalable.

Paper on Milo method: Dann, E., Henderson, N.C., Teichmann, S.A. et al (2022) Nature Biotechnology. https://doi.org/10.1038/s41587-021-01033-z

Differential Abundance - Milo



Workflow

- Construct KNN graph
  - use MNN-corrected matrix (or PCA for non-batched data)
  - calculates Euclidean distance between cells and its k nearest neighbours
- Define cell neighbourhoods by sub-sampling the graph to identify useful “index cells” (for computational efficiency)
- Counts cells in neighbourhoods
- Tests for DA in neighbourhoods (using a Negative Binomial linear model suitable for count data)
- Does a multiple testing correction (spatial FDR)
- Visualise the neighbourhood graph with our UMAP/t-SNE embedding

Paper on Milo method: Dann, E., Henderson, N.C., Teichmann, S.A. et al (2022) Nature Biotechnology. https://doi.org/10.1038/s41587-021-01033-z