@csoneson@fosstodon.org

Advanced transcriptomics inference (II)

Charlotte Soneson

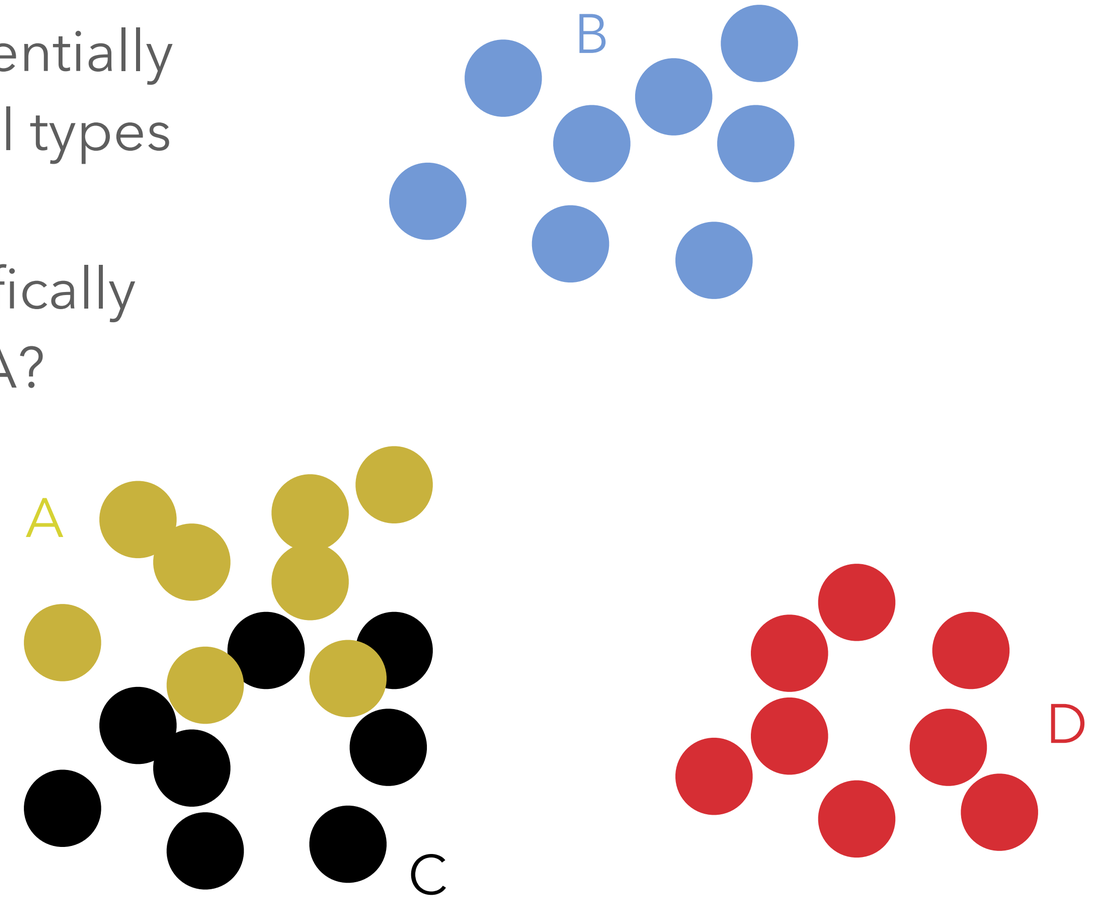
Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

CSAMA 2024

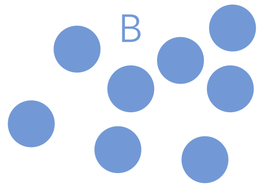
Differential analysis types for scRNA-seq

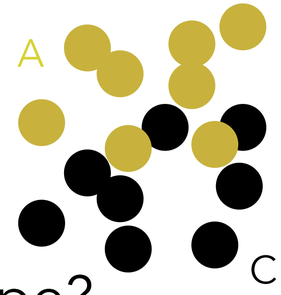
Marker gene detection

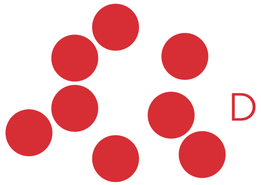
- Which genes are differentially expressed between cell types A and B?
- Which genes are specifically expressed in cell type A?



What do we mean by a "marker gene"?







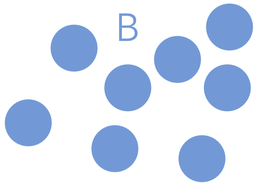
A gene that is

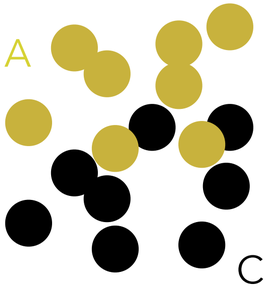
- upregulated in cell type A compared to *all* other cell types?
- upregulated in cell type A compared to *at least one* other cell type?
- upregulated in cell type A compared to "most" other cell types?
- upregulated in cell type A compared to the complement of cell type A?
- only ever seen in cell type A?
- ...?

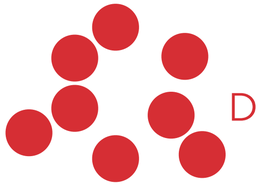
Marker gene detection for a specific cell type is strongly influenced by which other cells are present in your data set, as well as by the cluster resolution!

See `scran::findMarkers()` for implementations of different options

Double-dipping

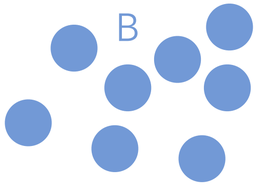


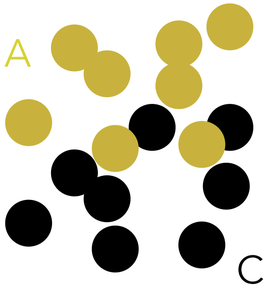


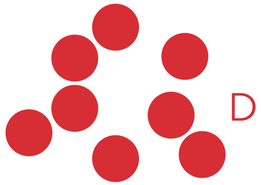


- Typically, cell types/clusters are first inferred from the data, after which expression values in each pair of clusters are compared to find marker genes (using the *same* data that was used to define the clusters).
- This approach is prone to overestimating the significance of cell type expression differences.
- Try it yourselves - generate random expression values, cluster cells into two clusters, see if you can find any genes that are significantly DE between these clusters!

Double-dipping - remedies





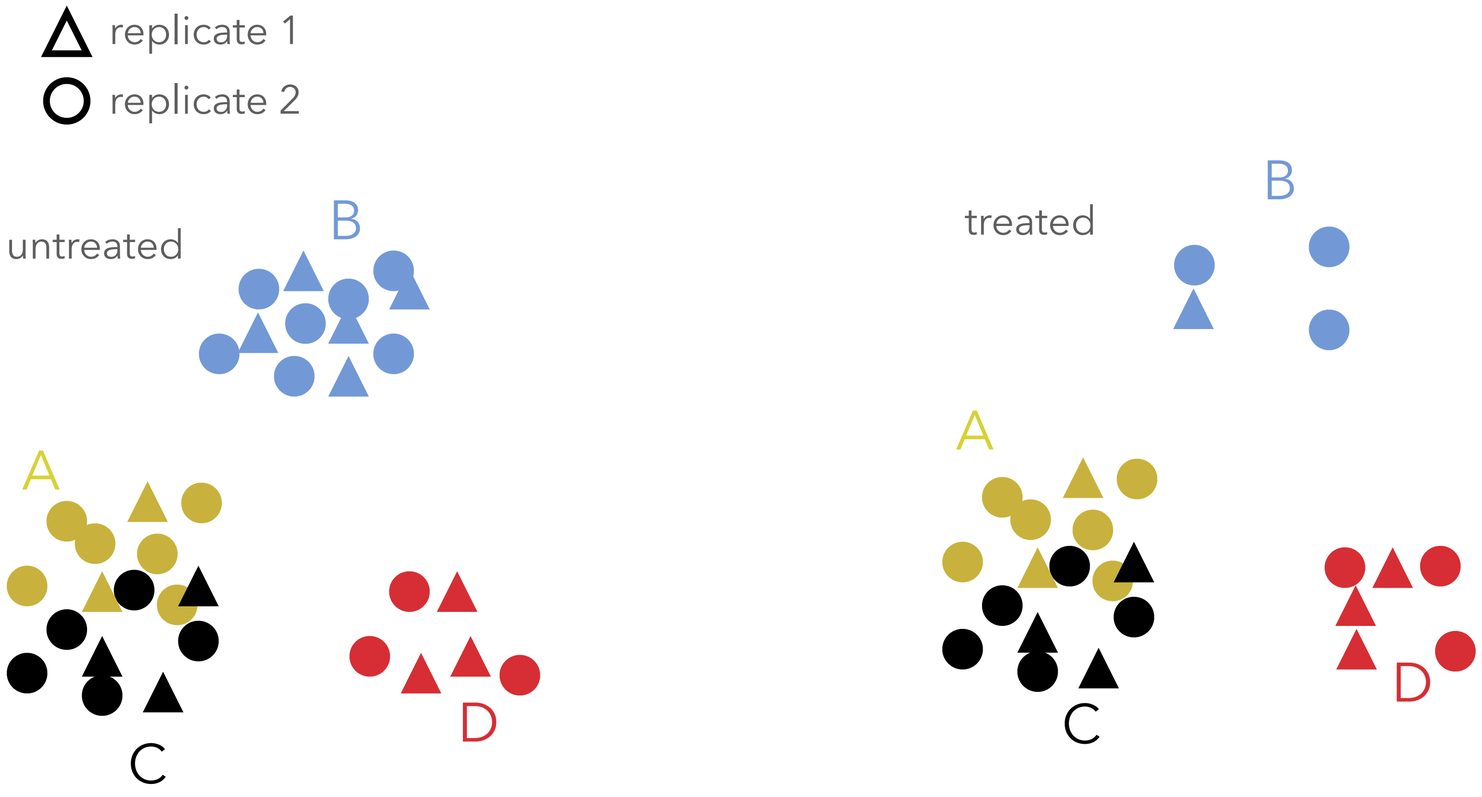


- Be careful when interpreting marker gene p-values.
- Infer cell types using independent data, or a subset of the genes (that are not tested afterwards). Note that splitting the *cells* into a training and a test set does not help - the 'test' cells still need to be assigned to a cluster based on their expression profiles.
- Approaches have been proposed to test for mean differences between clusters, *conditional* on the clusters having been found in the data (see e.g. https://arxiv.org/abs/2012.02936, https://arxiv.org/abs/2203.15267, https://www.lucylgao.com/clusterpval/).

Differential analysis types for scRNA-seq

Differential abundance analysis

- Are some cell types more/less abundant in one condition compared to another?



Cell type frequencies

	U1	U2	T1	T2
A	142	187	153	160
B	163	215	15	34
C	118	130	132	124
D	78	90	75	68

Differential abundance analysis - methods

- Conventional count-based methods (edgeR, DESeq2). See e.g. `diffcyt` (https://www.nature.com/articles/s42003-019-0415-5) or OSCA (http://bioconductor.org/books/3.15/OSCA.multisample/differential-abundance.html)
- Transformation + linear model. See e.g. `propeller` (https://academic.oup.com/bioinformatics/article/38/20/4720/6675456)

Cell type frequencies

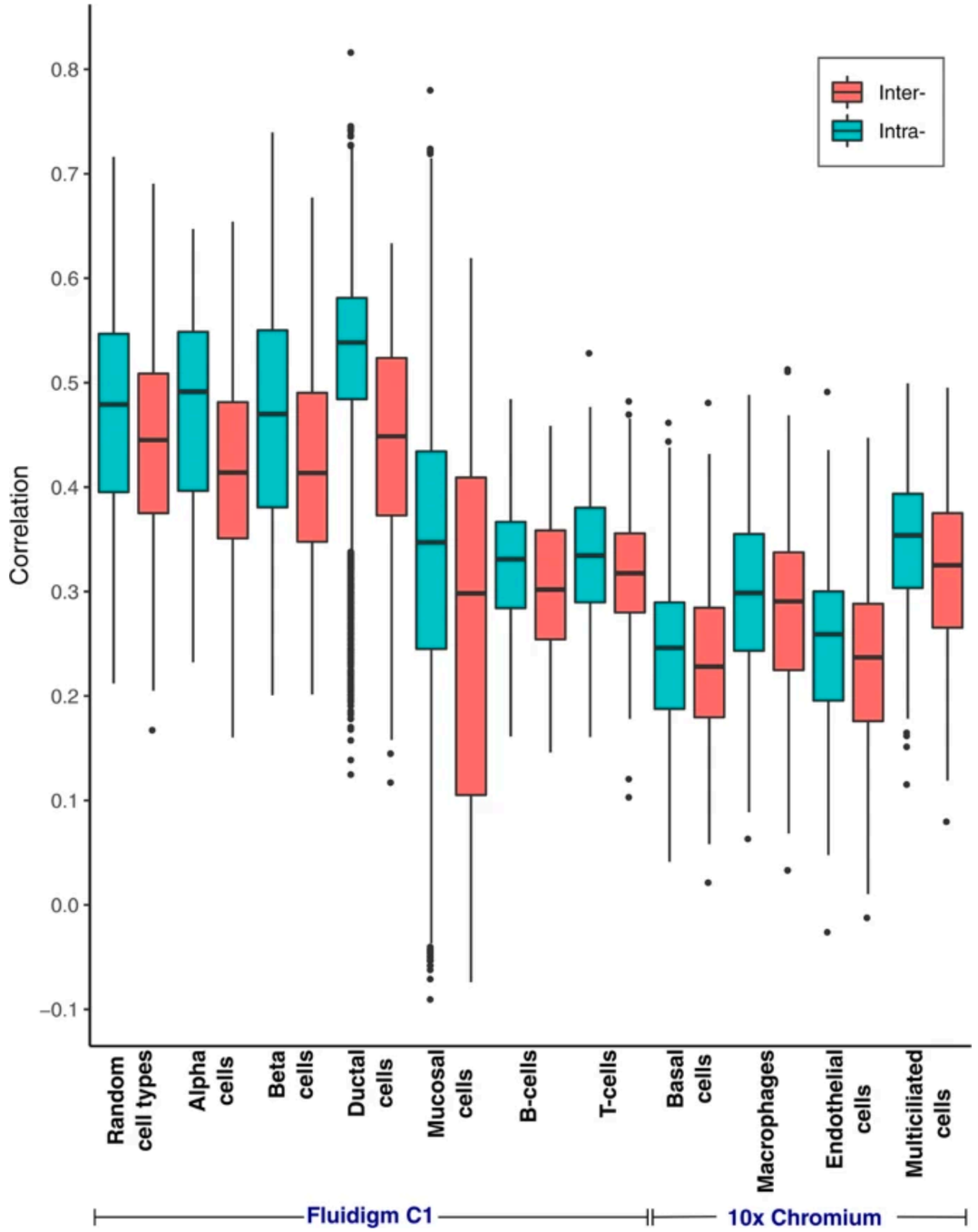
	U1	U2	T1	T2
A	142	187	153	160
B	163	215	15	34
C	118	130	132	124
D	78	90	75	68

Differential analysis types for scRNA-seq

Differential state analysis

- Are any genes differentially expressed between conditions within a given cell type?

Cells from the same individual are more highly correlated than cells from different individuals



https://www.nature.com/articles/s41467-021-21038-1

Differential state analysis - methods

- Perform a regular statistical test between all treated cells and all untreated cells.
  - Not recommended - the individual cells are not true biological replicates/ experimental units, and significance will be inflated!

- Mixed-effects model (random effect for individuals), accounting for the hierarchical correlation structure.
  - Can work well, but is computationally demanding.

- Pseudobulk creation + "regular" differential expression analysis.
  - Computationally efficient, ameliorates the sparsity in the single-cell data, masks within-sample heterogeneity.

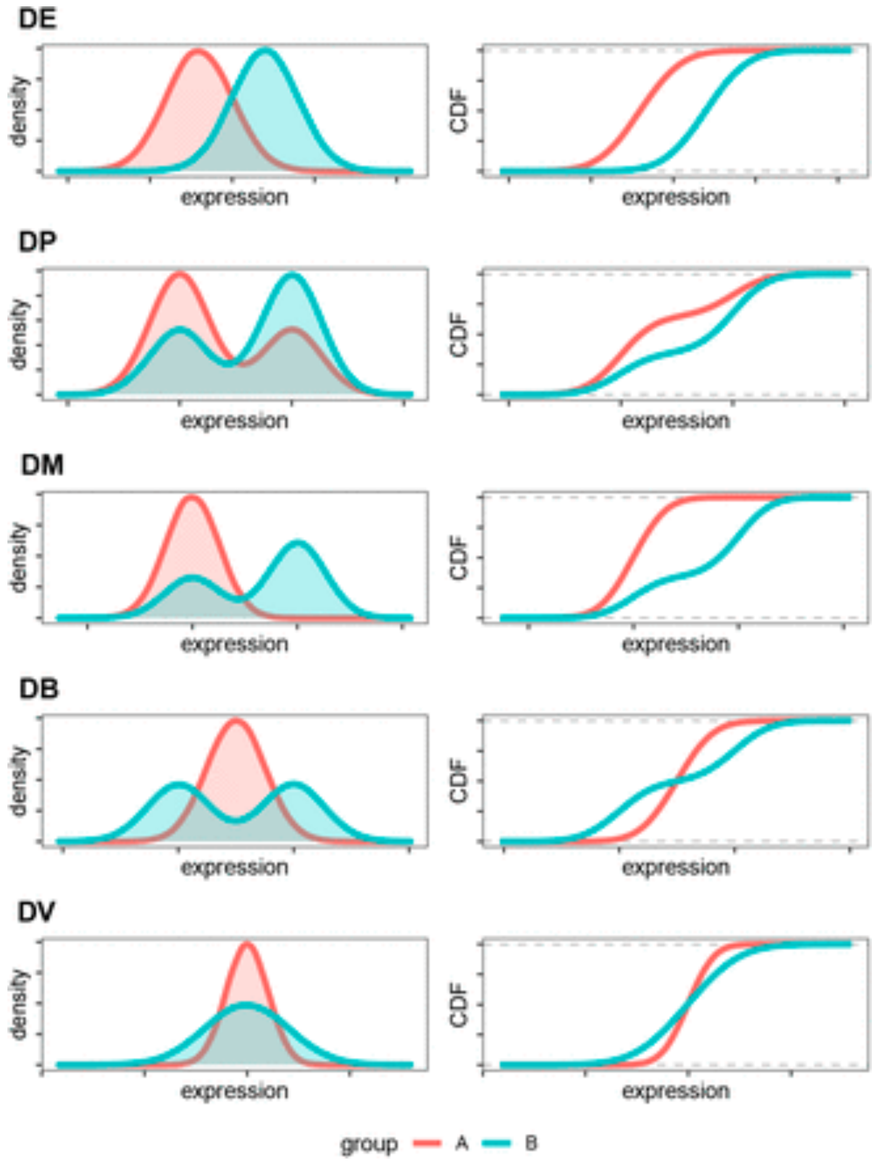
See e.g. the `muscat` package or `scuttle::aggregateAcrossCells()` for implementations of pseudobulk generation

Pseudobulk generation

- For each cell type/cluster, and each sample, sum the count vectors for the individual cells to create a single aggregated count vector -> gene `x` cell type-sample count matrix.
- Subsequent differential expression analysis is typically performed separately for each cell type.
- Remove columns generated from too few cells -> some comparisons may not be possible (e.g., if the treated samples don't contain a given cell type).
- From this point onwards, it's essentially a bulk differential analysis - we have leveraged the single-cell data to generate "pure" samples.

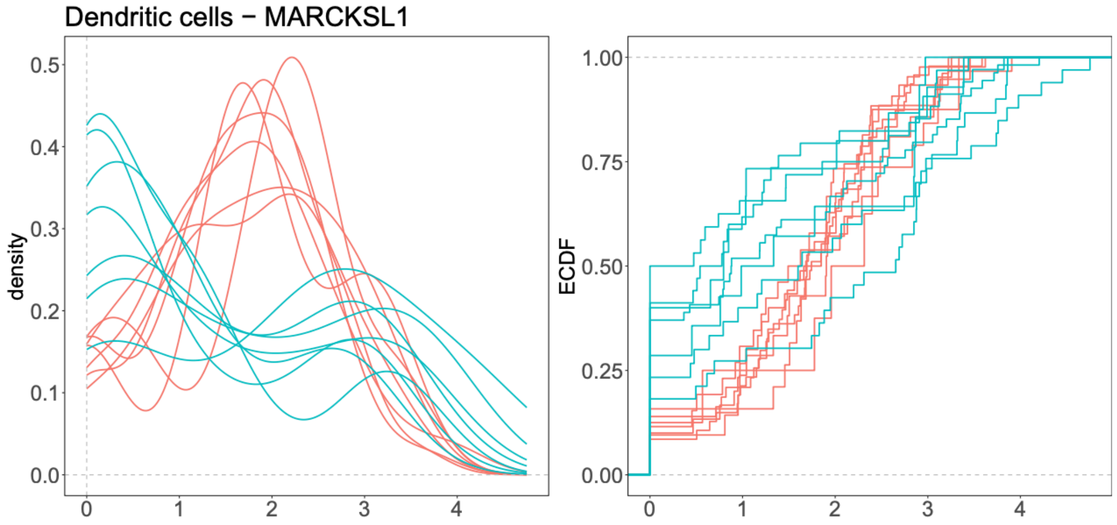
See e.g. the `muscat` package or `scuttle::aggregateAcrossCells()` for implementations of pseudobulk generation

Going beyond comparisons of means



https://www.biorxiv.org/content/10.1101/2020.11.24.394213v5.full

- "Classical" methods (e.g. Kolmogorov-Smirnov test) do not accommodate replicates.
- A generalization was proposed in the `distinct` package, by defining a test statistic from the difference between two groups of eCDFs at each point in a fine grid, and running a permutation test.



References

- http://bioconductor.org/books/3.17/OSCA.multisample/multi-sample-comparisons.html
- https://www.nature.com/articles/s41467-021-25960-2
- https://www.nature.com/articles/s41467-020-19894-4
- https://www.nature.com/articles/s41467-021-21038-1
  - https://www.nature.com/articles/s41467-022-35519-4
  - https://www.nature.com/articles/s41467-022-35520-x