@CSoneson

Differential expression analysis in scRNA-seq data

Charlotte Soneson

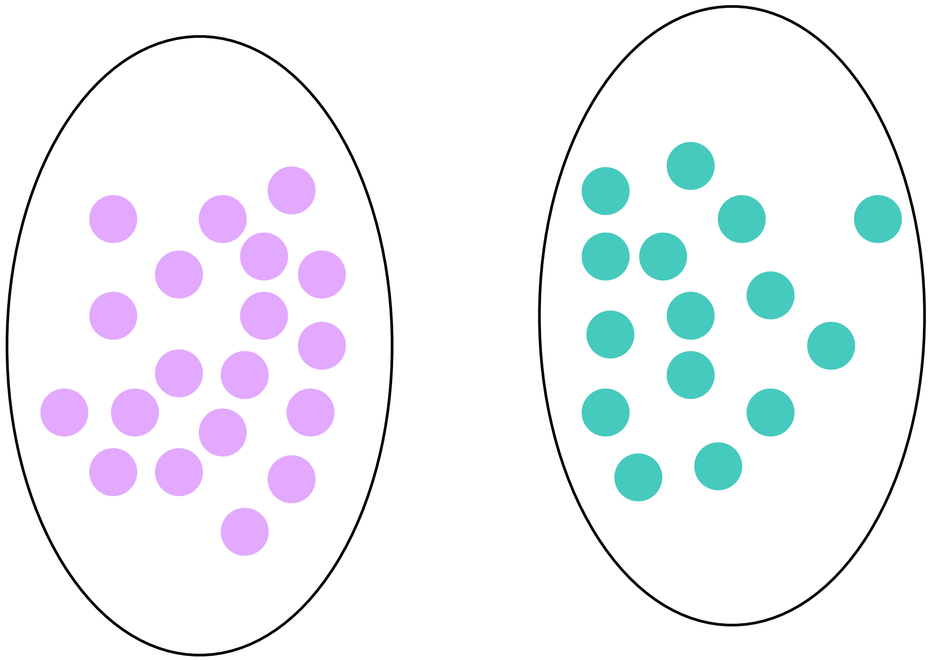
Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

SIB Swiss Institute of Bioinformatics

FMI Friedrich Miescher Institute for Biomedical Research

What do we mean by “differential expression analysis”?

Comparison of cell types (often within a single sample), to find “marker genes”

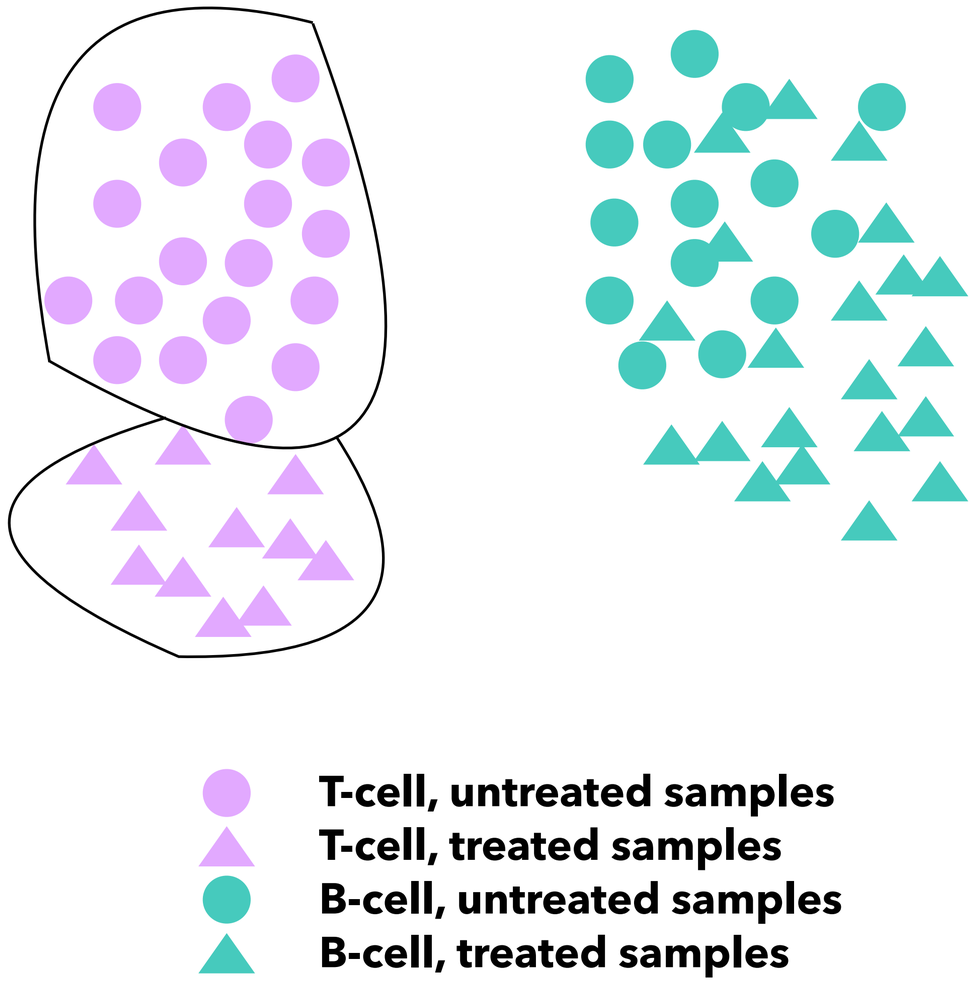


FOCUS OF TODAY

What do we mean by “differential expression analysis”?

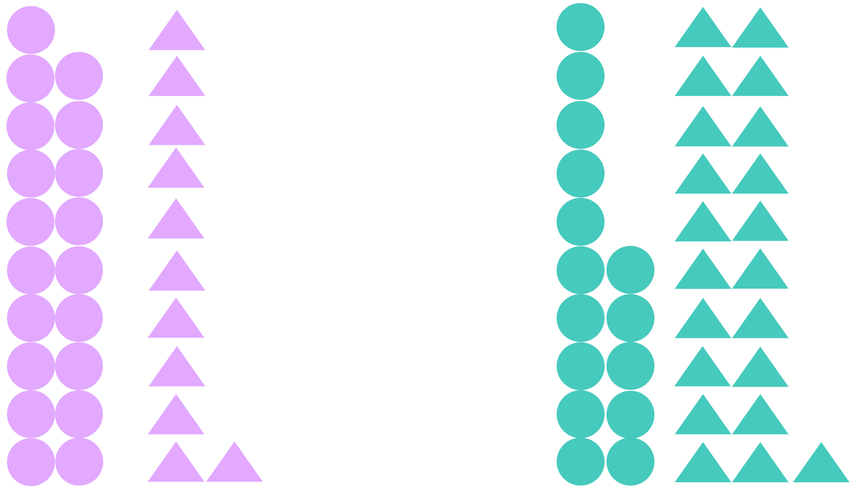
“Differential state analysis” - comparison of gene expression *within* a cell type, *between* samples/conditions (with replicates!)

TOMORROW!

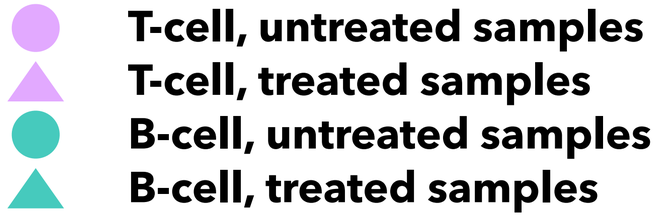


Differential abundance analysis

Comparison of cell type composition between samples/ conditions (with replicates!)

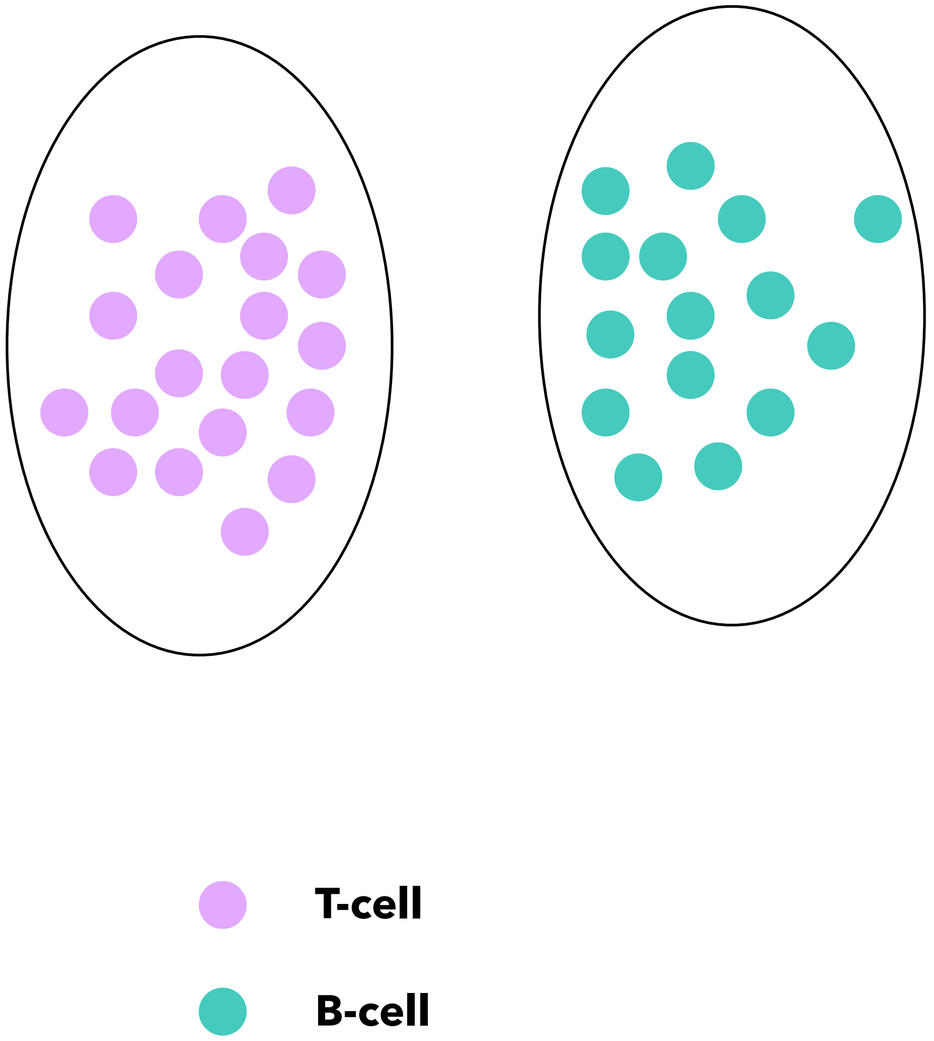


TOMORROW!



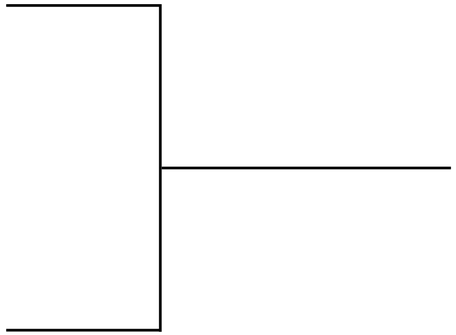
Comparing cell populations

Comparison of cell types (often within a single sample), to find “marker genes”



Comparing cell populations

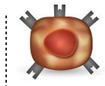
- Step 1: Get the cell populations
  - clustering
  - cell type assignment
  - known in advance (sorted cells)



Some **caution** is warranted, if we are using the same data to *define* the cell populations as to *compare* them.

- Step 2: Compare expression levels between populations

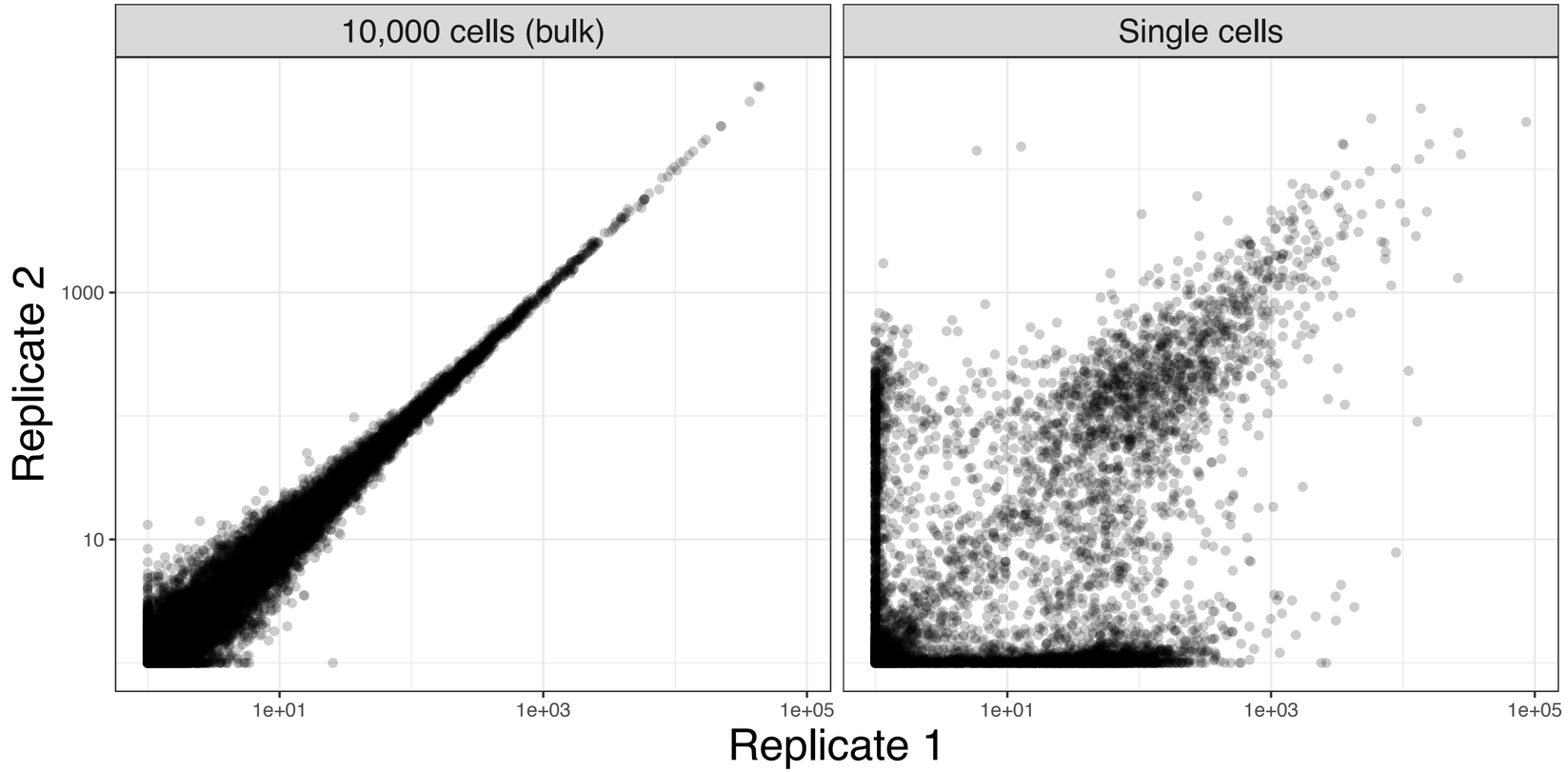
Differential expression analysis

																	
FLT3LG	0	2	0	1	4	0	0	0	4	6	4	0	1	1	0	0	0
NEAT1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
SCYL1	2	3	2	0	0	1	1	0	0	2	1	2	0	2	0	0	2
MALAT1	49	142	171	11	22	157	90	47	55	30	24	95	75	101	31	45	6
LTBP3	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
RPL13A	20	12	0	0	1	19	6	0	0	0	7	12	9	0	0	2	1
RCN3	0	0	0	1	0	1	1	1	0	0	0	2	0	0	0	0	0
RPS11	1	16	3	6	0	3	8	0	1	0	16	3	6	10	2	0	2

Setup is similar to bulk RNA-seq (gene-vs-observation matrix of counts)

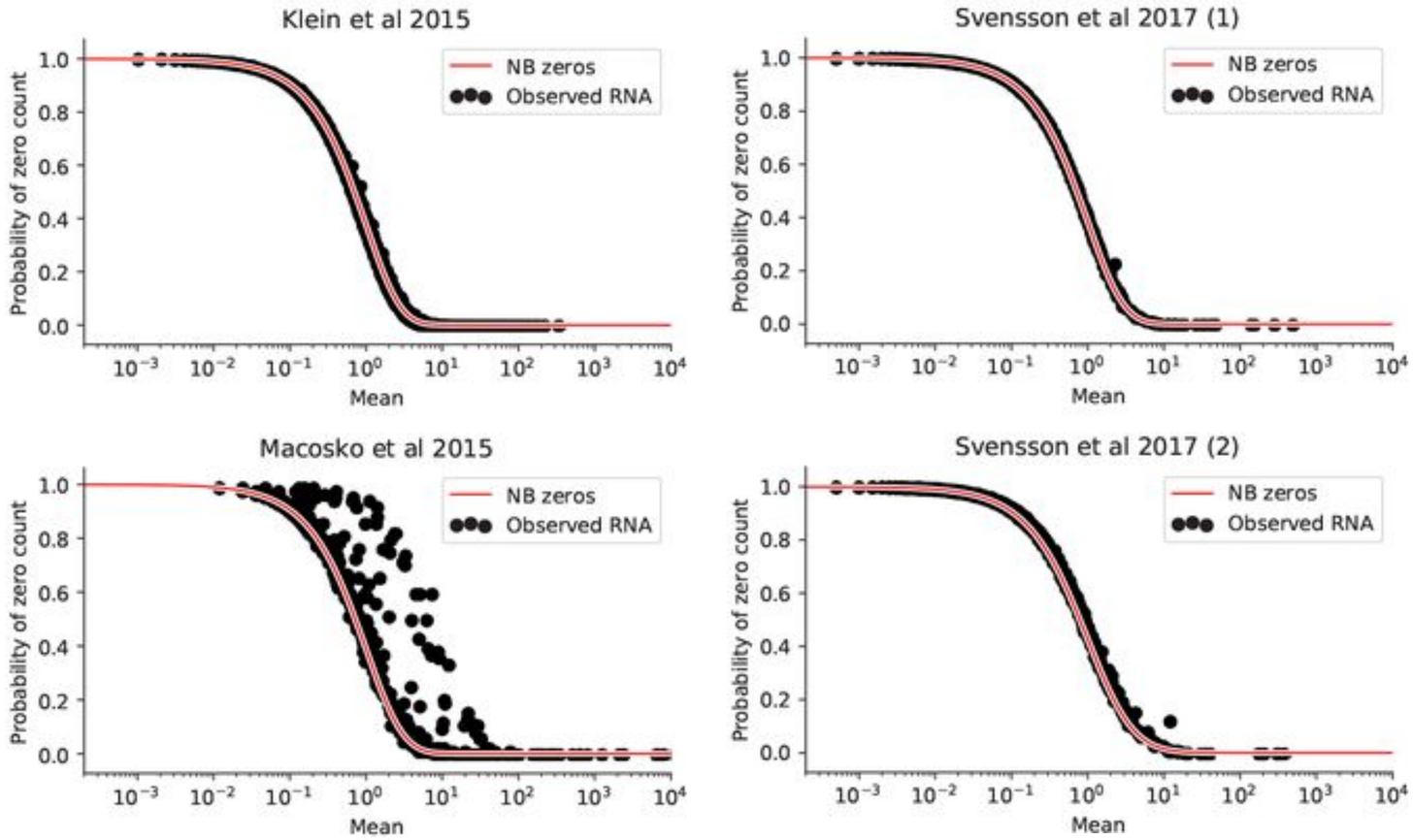
Can we use bulk methods?

Data characteristics are different - scRNA-seq data is much more sparse, with high variability



Is scRNA-seq data zero-inflated?

Droplet scRNA-seq negative control data (without biological heterogeneity) is consistent with a regular (non-zero-inflated) Negative Binomial distribution.

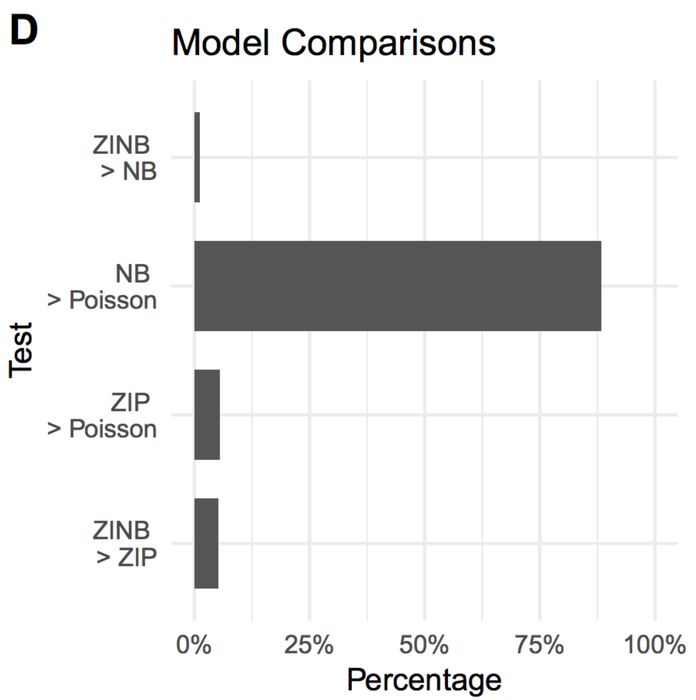


Svensson, bioRxiv doi: https://doi.org/10.1101/582064 (2019)

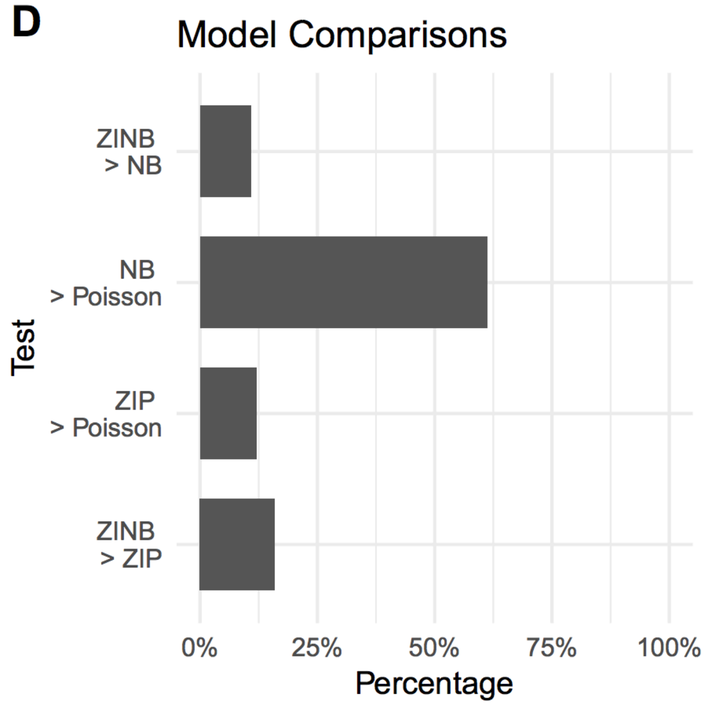
Is scRNA-seq data zero-inflated?

Zero-inflated Negative Binomial provides a moderate improvement over a regular Negative Binomial for UMI counts, more improvement for read counts.

UMI counts (STRT-Seq)



Read counts (SmartSeq)



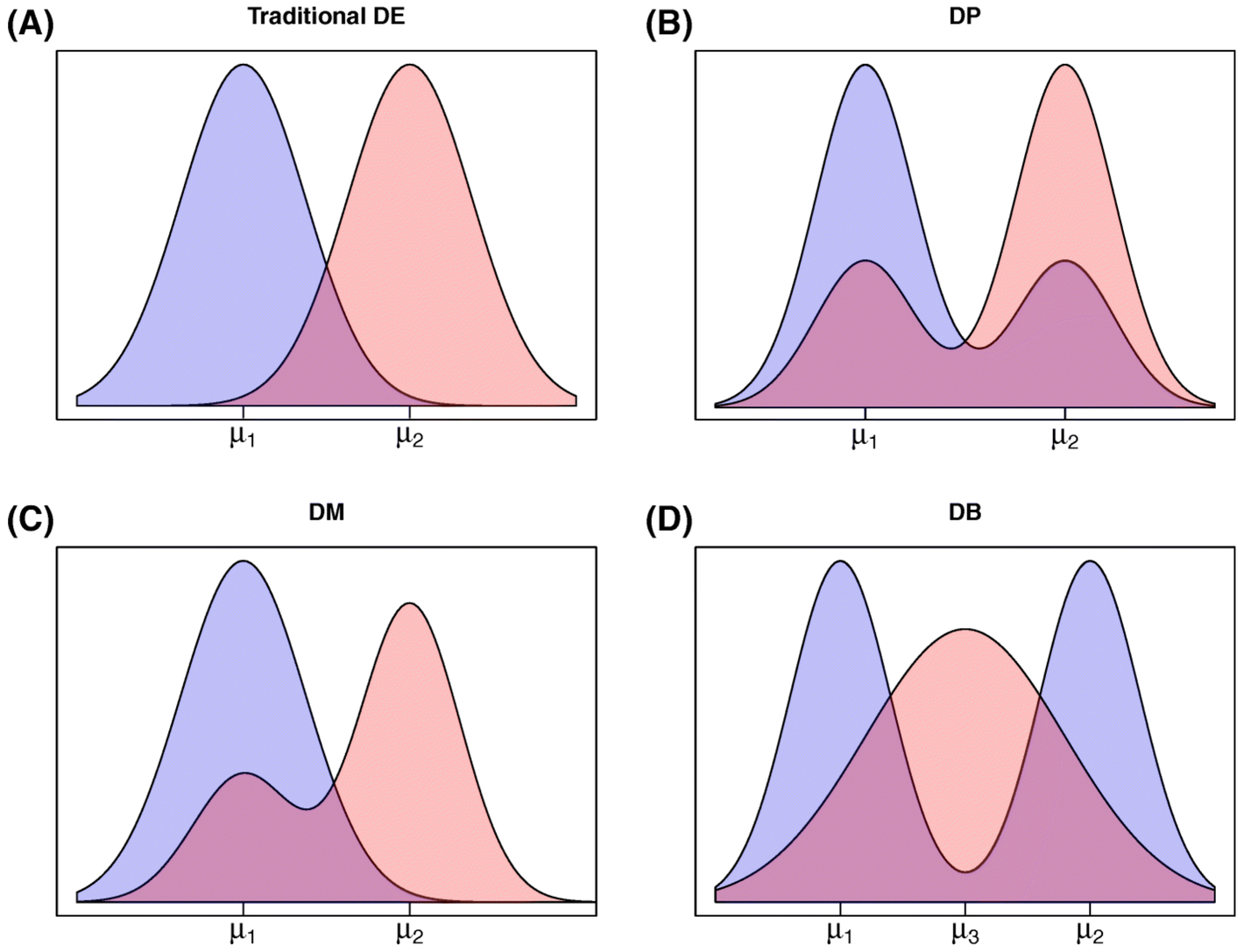
Vieth *et al*, Bioinformatics 33(21):3486-3488 (2017)

Can we use bulk methods?

- We have many more cells than the typical number of bulk RNA-seq samples, but often only from a single individual
  - what does it mean to treat the cells as “biological replicates”?
  - to what can we expect the results to generalize?

Can we use bulk methods?

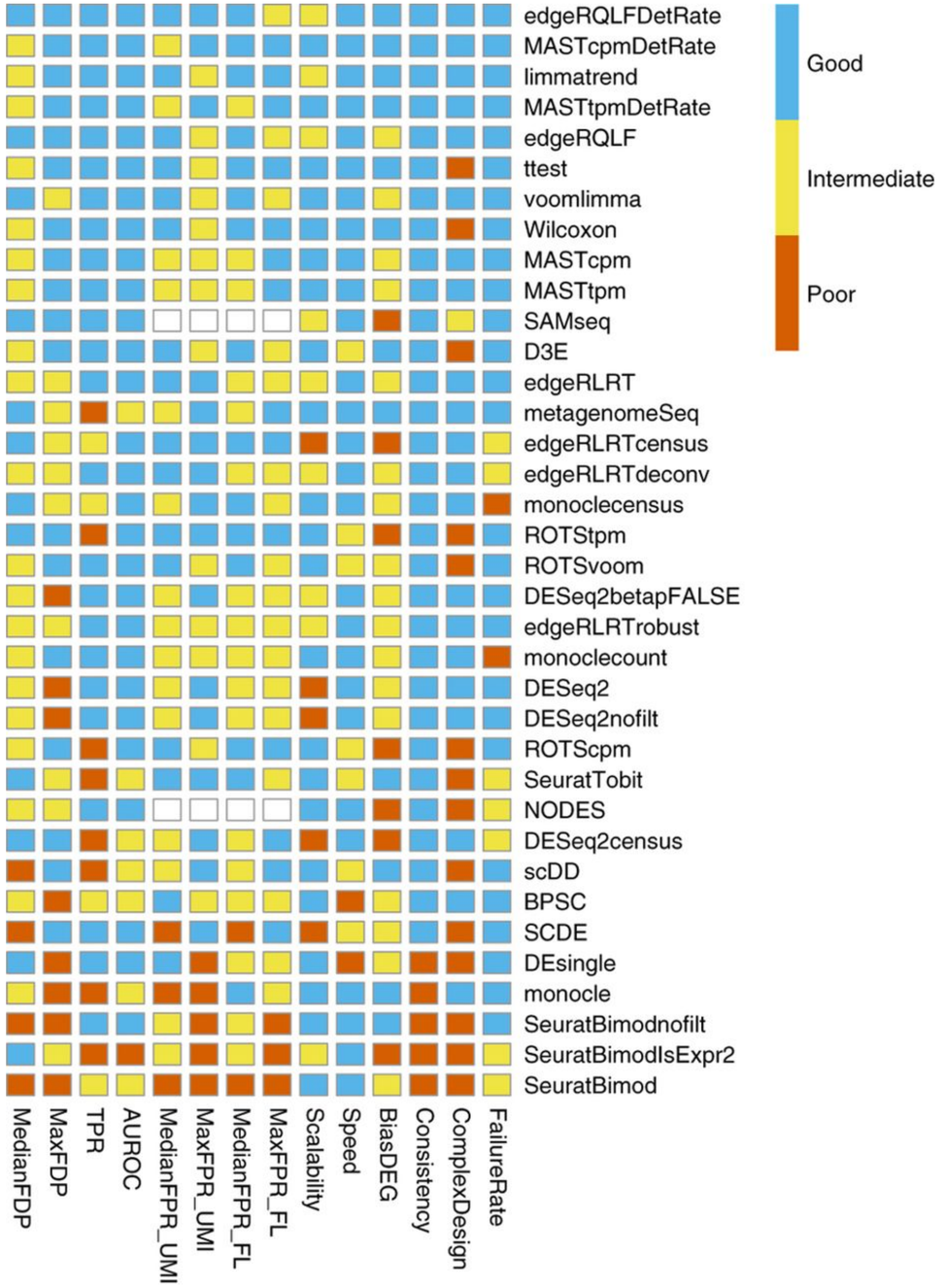
What do we want to compare? In bulk, almost always mean expression between different conditions, but could be other things (e.g., heterogeneity, proportion of cells that express a gene).



Korthauer *et al*, Genome Biology 17:222 (2016)

Comparing differential expression methods

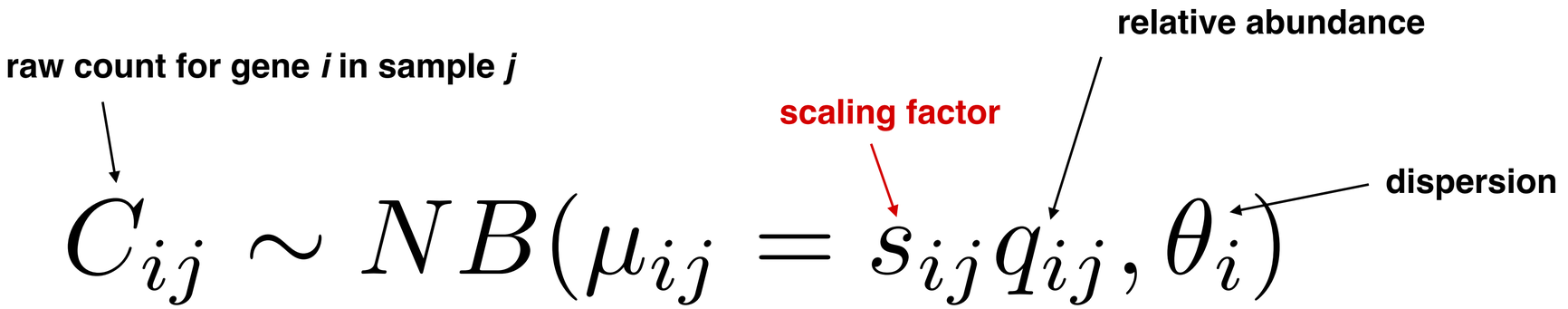
- Bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq
- Even the t-test and the Wilcoxon test work well (assuming that you have at least a few dozen cells to compare)
- Filtering out lowly expressed genes is quite important for good performance of bulk methods



Soneson & Robinson, Nature Methods 15:255-261 (2018); http://imlspenticton.uzh.ch:3838/scrnaseq_de_evaluation/

edgeR (QLF)

- Model the raw (UMI/read) counts with a Negative Binomial distribution, with offset accounting for sequencing depth/composition effects.
- Quasi-likelihood F-test.
- Empirical Bayes shrinkage to get robust dispersion estimates even with limited replication.
- Gene-wise null hypothesis: mean expression is the same across groups



Robinson *et al*, Bioinformatics 26:139-140 (2010); Lun *et al*, Methods in Molecular Biology 1418:391–416 (2016)

Extending bulk methods to zero-inflated data

- Idea:
  - For each observed zero count, estimate the probability that it is generated from the zero component (rather than the Negative Binomial).
  - Downweight the zeros from the zero component in the inference steps.
- Weights can be estimated e.g. with the `zinbwave` package.

Van den Berge *et al*, Genome Biology 19:24 (2018); Risso *et al*, Nature Communications 9:284 (2018)

MAST

- Model log(TPM+1) values
- Hurdle model - model the rate of expression as well as the mean expression (conditional on being expressed)
- Two-part model: logistic regression + linear model

$$logit\left(Pr\left(Z_{ig}=1\right)\right)=\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{D}}$$

$$\Pr\left(Y_{ig}=y\middle|Z_{ig}=1\right)=\mathrm{N}\left(\mathrm{X_i}\beta_{\mathrm{g}}^{\mathrm{C}},\ \sigma_g^2\right)$$

Finak *et al*, Genome Biology 16:278 (2015)

limma-trend

- Normalize and log-transform counts (often with a relatively large pseudocount of, e.g., 3)
- Apply limma (linear model with moderated variance)
    - modify the default empirical Bayes procedure to incorporate a mean-variance trend in the prior
- Gene-wise null hypothesis: mean expression is the same across groups

Law *et al*, Genome Biology 15:R29 (2014)

t-test

- Parametric two-group comparison.
- Gene-wise null hypothesis: mean expression in group 1 = mean expression in group 2.
- Typically allow different variance in the two groups (Welch t-test).
- Expression values should be pre-normalized and preferably approximately normally distributed within each group - typically applied to logcounts.
- Default test in `scran::findMarkers()`
- Also used in `Seurat::FindMarkers(…, test.use = "t")`

Wilcoxon (Mann-Whitney) test

- Non-parametric two-group comparison.
- Gene-wise null hypothesis: it’s equally likely that a randomly selected cell from group 1 will have higher or lower expression of the gene than a randomly selected cell from group 2.
- Expression values should be pre-normalized - typically applied to logcounts (monotonic transformations don't change outcome).
- Default test in `Seurat::FindMarkers()`

Bioc 3.10

- Also used in `scran::findMarkers(…, test.type = “wilcox”)`

Compare the proportion of zeros

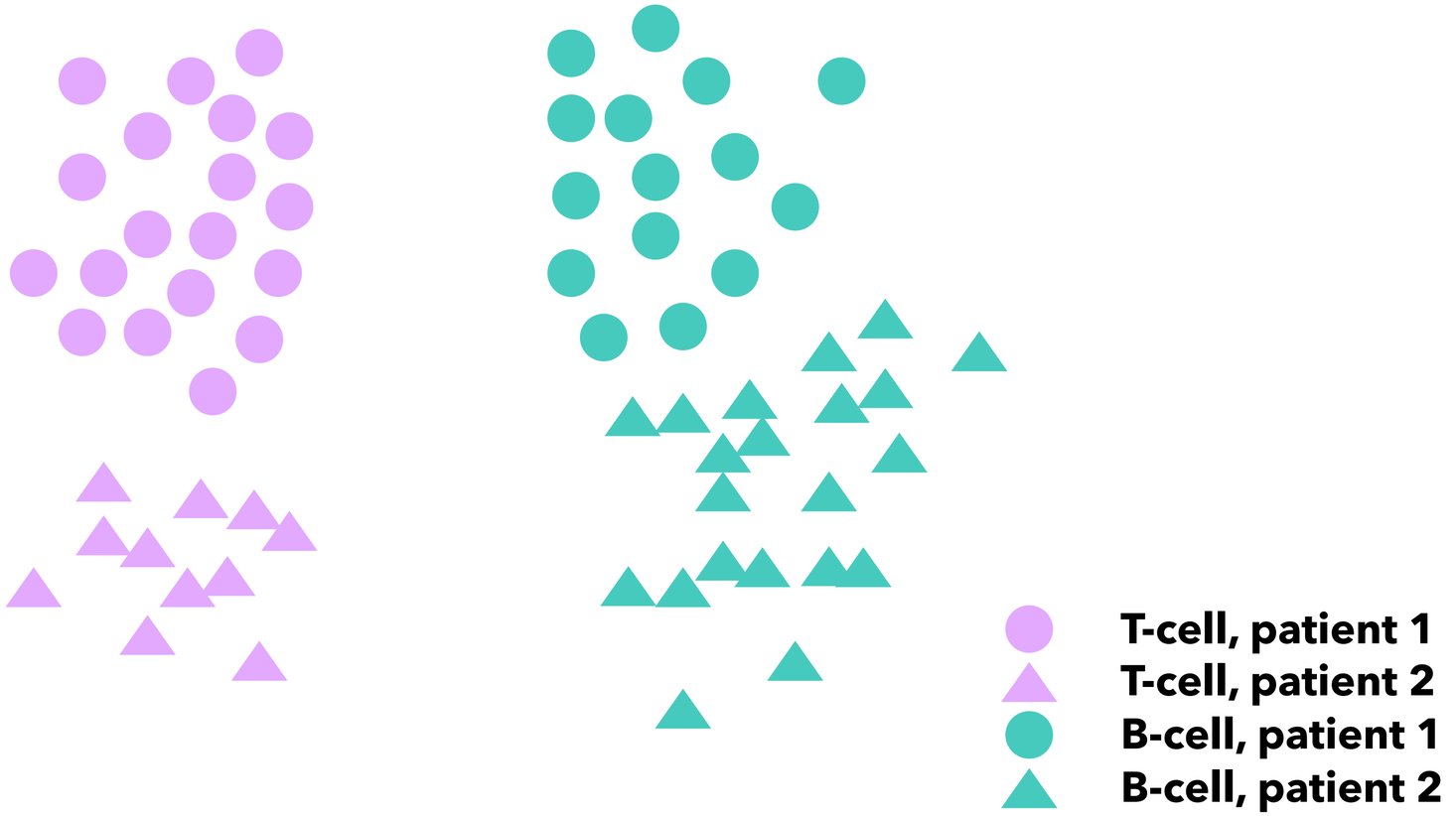
- Binomial test.
- Gene-wise null hypothesis: the probability of being expressed is the same in group 1 and group 2.

Bioc 3.10

- Accessible via `scran::findMarkers(…, test.type = "binom")`

Comparing cell populations in the presence of batch effects

First, make sure that clusters are properly defined!



Comparing cell populations in the presence of batch effects

- **Alt. 1**: Remove the batch effect first, perform tests on "corrected" data
  - Typically not recommended
  - Correction may not preserve magnitude (or direction) of gene expression changes between cell types
  - Correction may introduce "artificial agreement" between batches - expression values in one batch are adjusted to better match those in another batch

Comparing cell populations in the presence of batch effects

- **Alt. 2**: Include the (additive) batch effect as a predictor in the statistical model
  - Can handle the situation where some cell types are not present in all batches
  - All data is used for parameter estimation - can increase power
  - For linear models, places stronger assumptions on the data than e.g. the t-test (equal variance between groups)
  - Assumes that the batch effect is constant across cell types
  - Applicable via `scran::findMarkers(…, design = design)`

Comparing cell populations in the presence of batch effects

- **Alt. 3**: Perform separate test for each batch, then aggregate
  - Only possible if both clusters are present in at least one batch
  - Applicable via `scran::findMarkers(…, block = "batch")`
  - p-values are combined using Stouffer's Z method

Recap

- We have seen
  - how to compare two cell populations
  - using several different methods
  - both in the absence and presence of batch effects
- How can we use this to answer biological questions of interest?

What is a “marker gene”?

- Differential expression is always **comparative** - the results will depend on what we compare to!
    - If the data set consists only of T-cells, no generic T-cell markers will (or, at least, should) show up as differentially expressed between clusters
    - Important to keep in mind when comparing marker genes found in different studies, with potentially different composition

What is a “marker gene”?

- Typically we have more than two clusters in a data set
- For a given cluster, are we interested in “marker genes” that are:
  - DE compared to all cells outside of the cluster
  - DE compared to at least one other cluster
  - DE compared to *each* of the other clusters
  - DE compared to “most" of the other clusters

What is a “marker gene”?

- For a given cluster, are we interested in “marker genes” that are:
  - DE compared to all cells outside of the cluster

    ```
    Seurat::FindMarkers(…)
    ```

  - DE compared to at least one other cluster

    ```
    scran::findMarkers(…, pval.type = “any")
    ```

  - DE compared to *each* of the other clusters

    ```
    scran::findMarkers(…, pval.type = “all")
    ```

  - DE compared to “some” of the other clusters

    Bioc 3.10

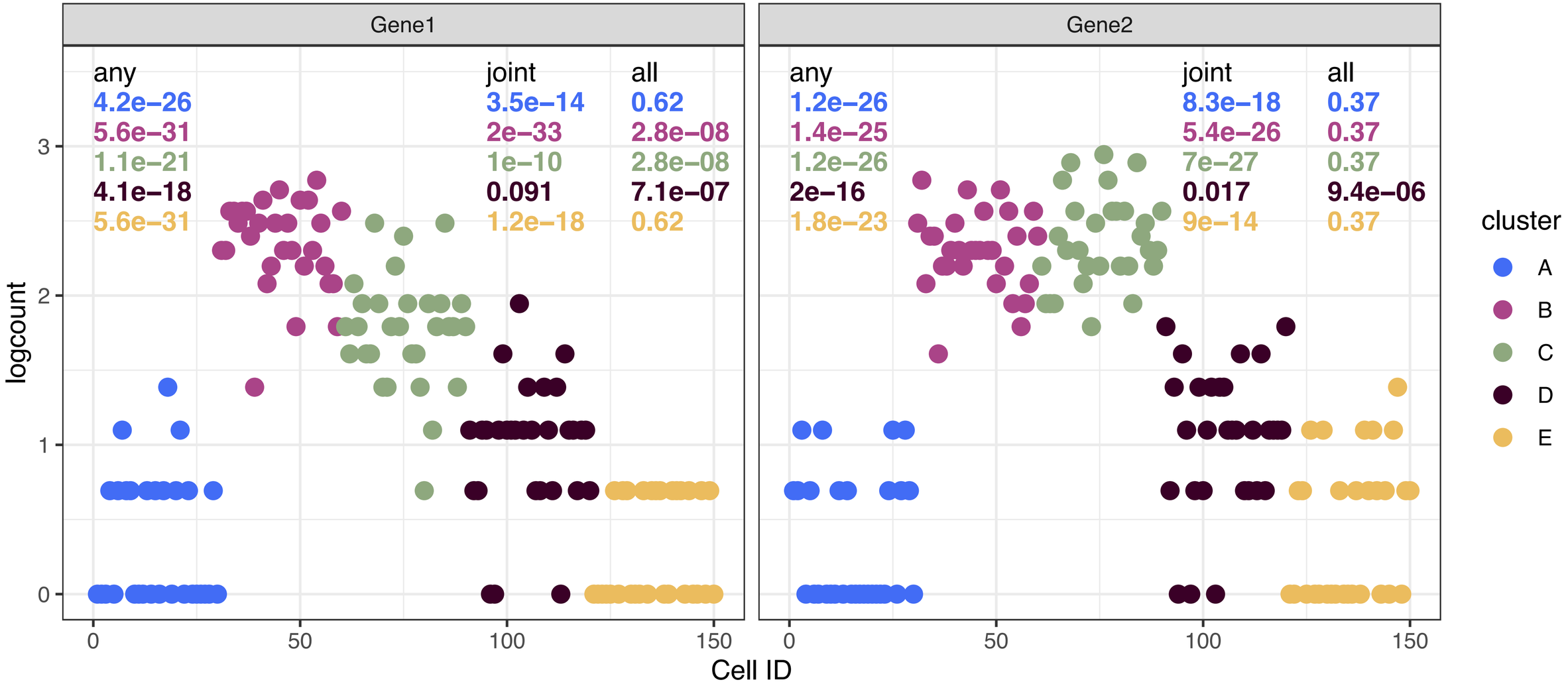
    ```
    scran::findMarkers(…, pval.type = “some")
    ```

What is a “marker gene”?

- Typically, upregulated marker genes are a bit easier to interpret
- `scran::findMarkers(…, direction = “up”)` only returns these (can also be set to “down”)
- Over/underclustering can have a big effect on the marker genes

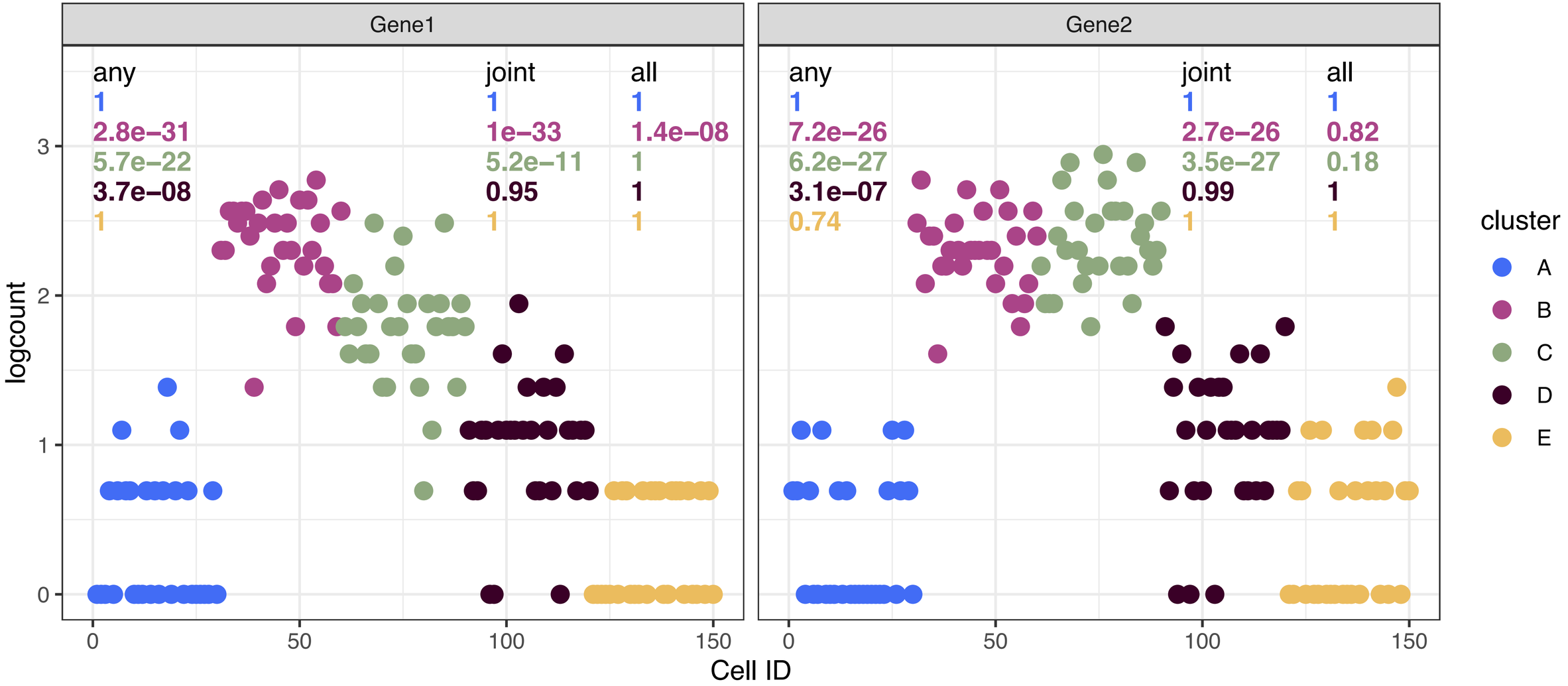
DE types

- any - DE wrt at least one other cluster
- joint - DE wrt all cells outside cluster
- all - DE wrt each other cluster



DE types - only upregulation

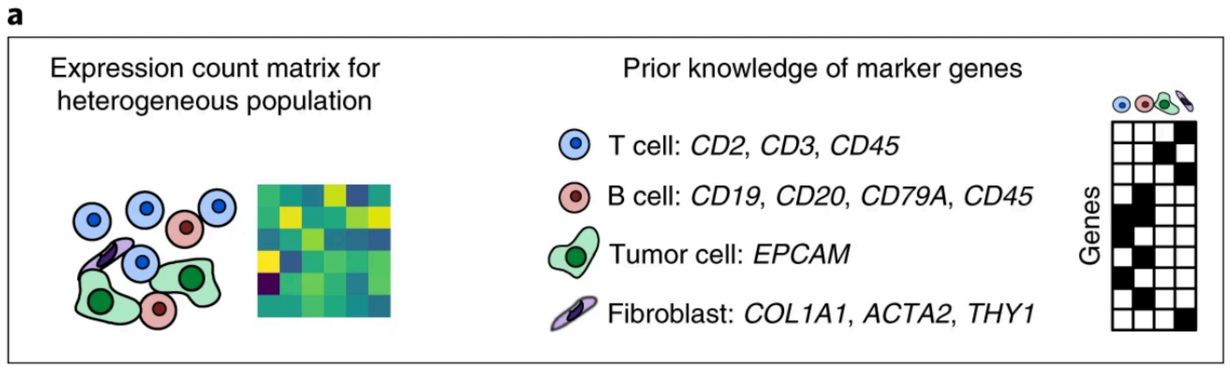
- any - DE wrt at least one other cluster
- joint - DE wrt all cells outside cluster
- all - DE wrt each other cluster



“Automatic” cell type assignment

- Idea: assign clusters or individual cells a (cell type) label, based on the (marker) genes that it expresses
- Do this automatically rather than manually, for reproducibility, objectivity, consistency, and in order to only have to do the laborious manual cell type identification once
- Focus on known signal (which may not always be the strongest signal in the data!)

CellAssign - input

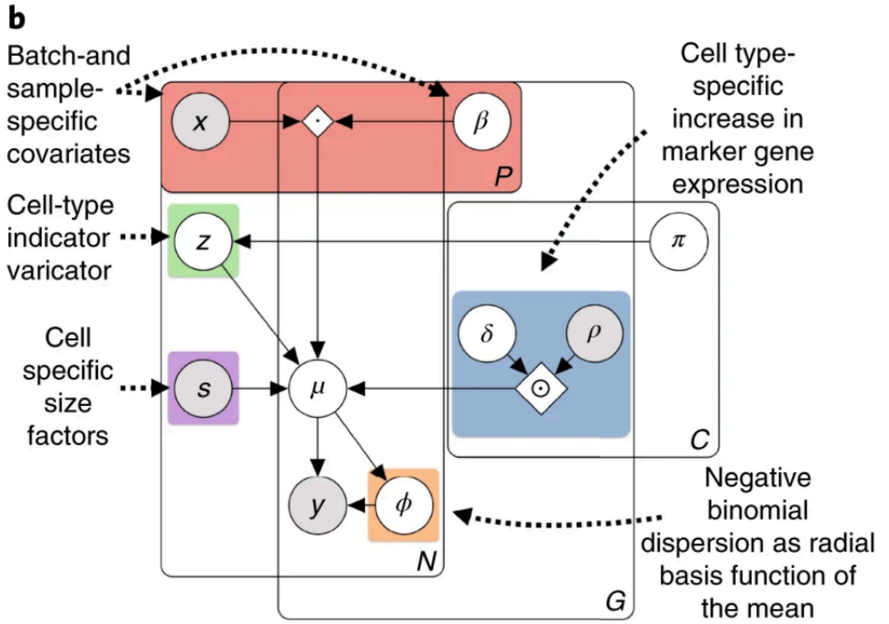


- Single-cell RNA-seq data set to annotate.
- Set of marker genes for each cell type/label.
- Markers are assumed to be more highly expressed in the cell types they define compared to other cells.
- Analyses are confined to the provided marker genes.

Zhang *et al*, Nature Methods (2019)

CellAssign - modeling

- Model observed raw counts (for marker genes) as a composite of cell type, library size, batch, …
- Based on a hierarchical framework, estimate the probability that each cell belongs to each of the annotated cell types (can be unassigned).
- Model parameters are estimated using an EM algorithm.



c

Variable	Distribution	Description
y_{ng}	Negative bionomial	Single-cell count
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
μ_{ngc}	Deterministic f^n	Modeled average expression
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	log-normal	Marker overexpression
ρ_{gc}	None	Marker/cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients
π_c	Dirichlet	Prior probability of cell type

Zhang *et al*, Nature Methods (2019)

CellAssign - (slightly) more details

$Y -$ cell-by-gene expression matrix

$z_n = c$ if cell n is of type c

Compute $p(z_n = c | Y, \hat{\Theta})$

$\hat{\Theta} -$ MAP estimates of model parameters

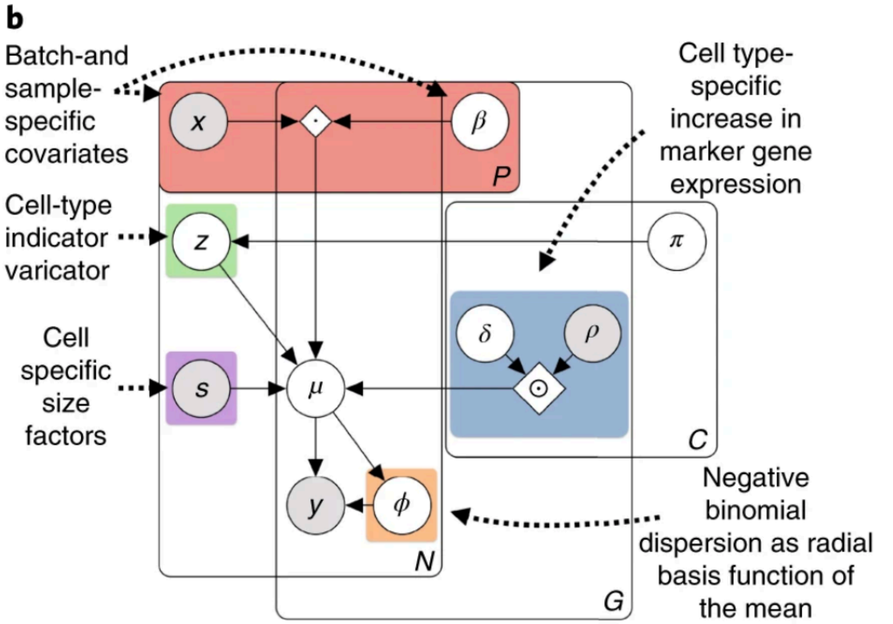
$\rho_{gc} = 1$ if gene g is a marker for cell type c

$$\mathbb{E}\left[y_{ng} | z_n = c\right] = \mu_{ngc}$$

where

$$\overbrace{\log \mu_{ngc}}^{\text{Log mean expression}} = \underbrace{\log s_n}_{\text{Cell size factor}} + \overbrace{\delta_{gc}\rho_{gc}}^{\text{Cell type specific}} + \underbrace{\beta_{g0}}_{\text{Base expression}} + \overbrace{\sum_{p=1}^{P} \beta_{gp} x_{pn}}^{\text{Other covariates (incl.batch)}}$$

with the constraint that $\delta_{gc} > 0$.

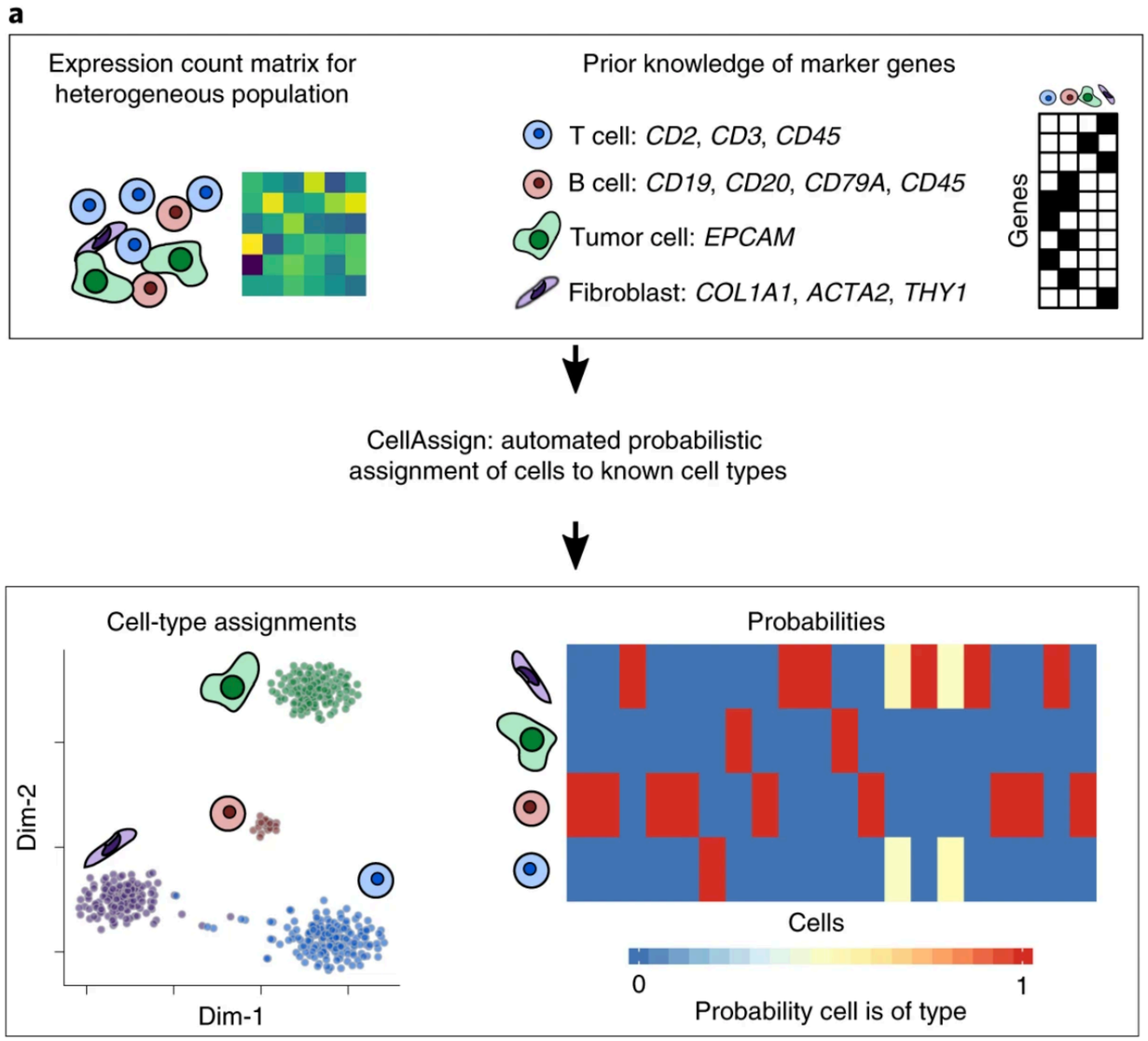


c

Variable	Distribution	Description
y_{ng}	Negative bionomial	Single-cell count
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
μ_{ngc}	Deterministic f^n	Modeled average expression
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	log-normal	Marker overexpression
ρ_{gc}	None	Marker/cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients
π_c	Dirichlet	Prior probability of cell type

Zhang *et al*, Nature Methods (2019)

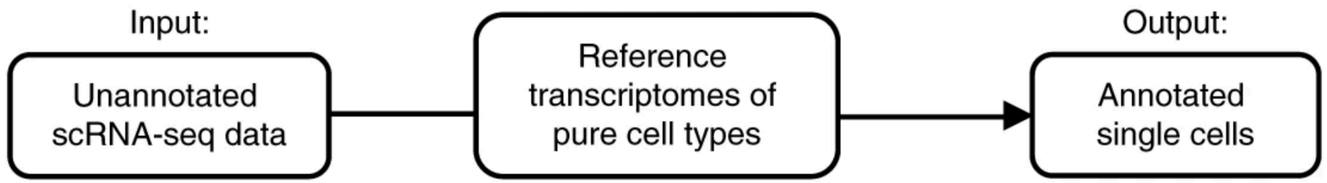
CellAssign - output



Zhang *et al*, Nature Methods (2019)

singleR - input

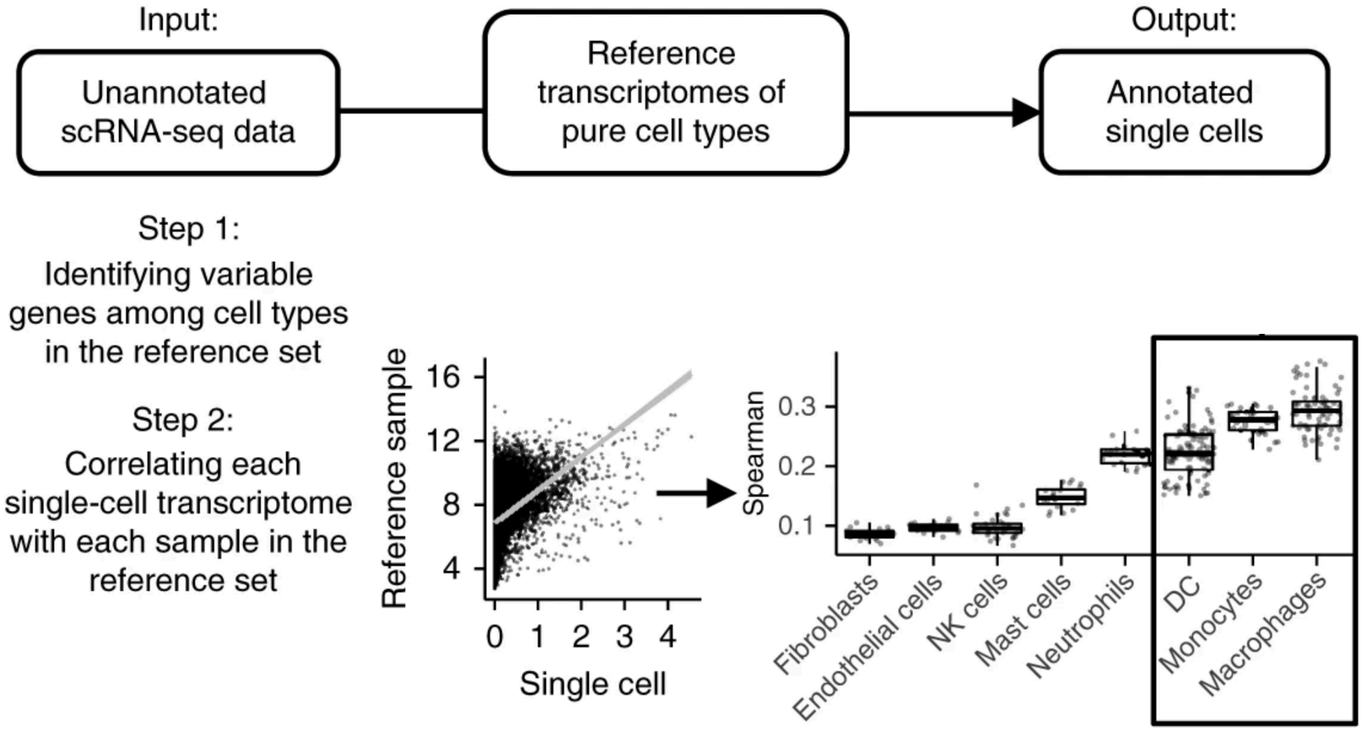
- Single-cell RNA-seq data set to annotate (either cell- or cluster-wise).
- Reference data set with pure cell types (multiple samples per cell type/label).
- Both bulk (“default”) and single-cell reference data sets can be accommodated.



Aran *et al*, Nature Immunology 20:163-172 (2019)

singleR - first assignment

- Define set of marker genes to use as the basis for calculations.
- For each cell, calculate Spearman correlation with all reference samples with a given label.
- Cell score = given quantile of these correlations.

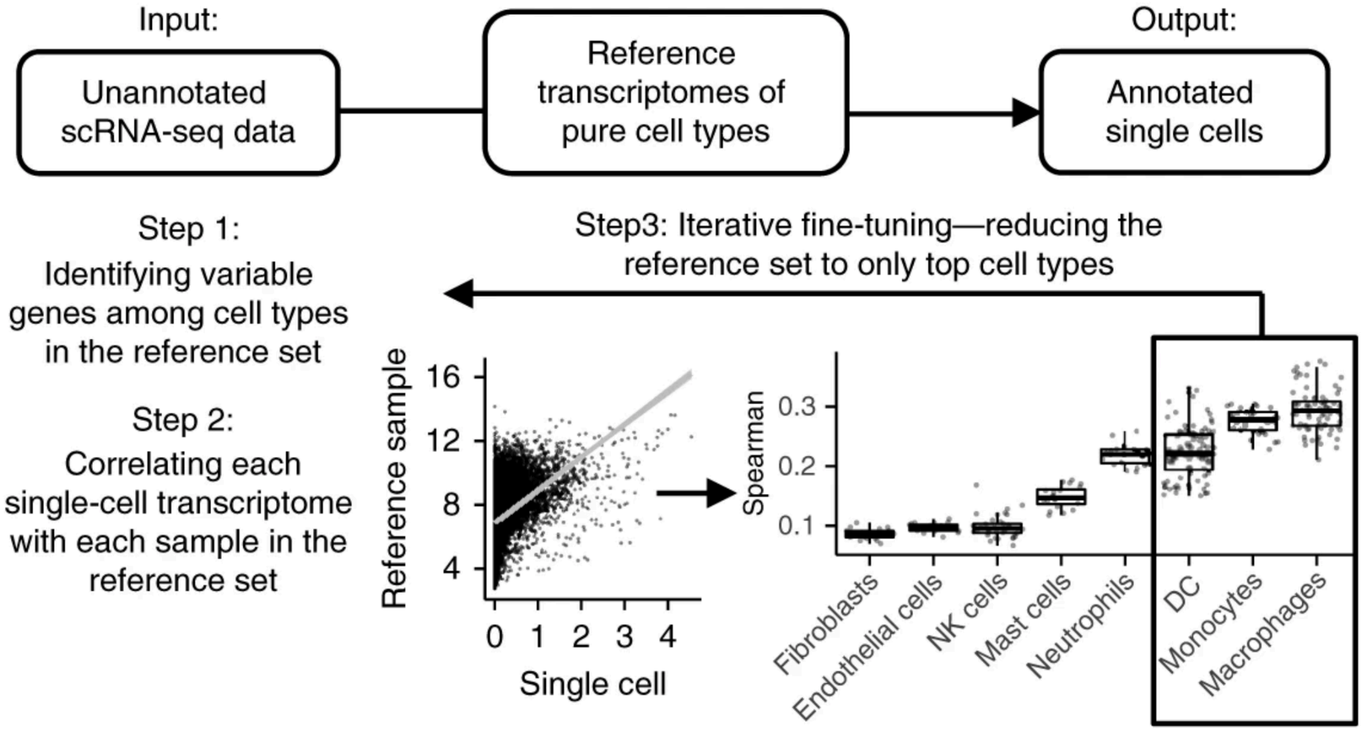


- Assign cell to label with highest score -> `first.labels`

Aran *et al*, Nature Immunology 20:163-172 (2019)

singleR - fine-tuning

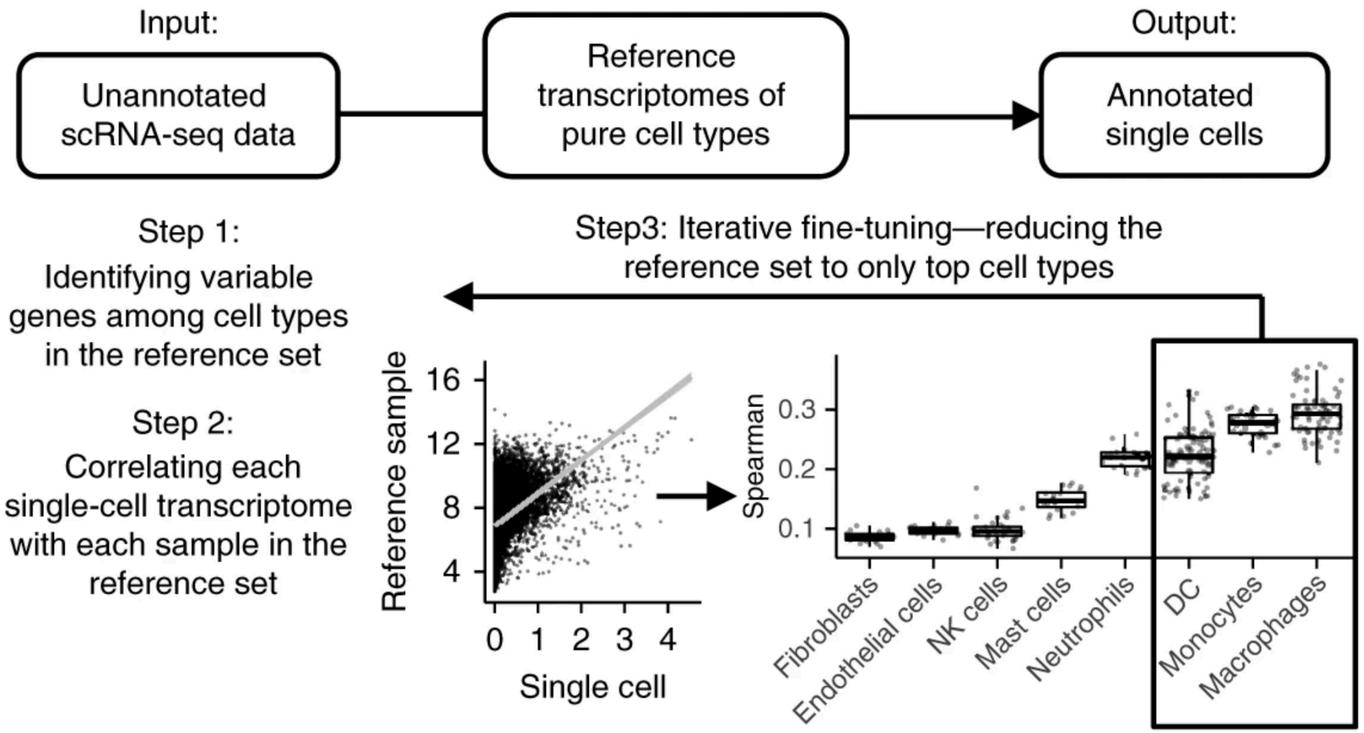
- Keep only labels with highest scores.
- Recalculate marker genes.
- Recalculate correlations and scores based on these genes.
- New assignments
  -> `labels`



Aran *et al*, Nature Immunology 20:163-172 (2019)

singleR - pruning

- Calculate the difference between the maximal score and the median score across all labels.
- Small difference - ambiguous assignment.
- For each label, find outliers (cells with small differences), and remove their label assignment.



- Remaining assignments -> `pruned.labels`

Aran *et al*, Nature Immunology 20:163-172 (2019)

singleR reference gene selection

- Each correlation is calculated based on a subset of the genes
- Several options:
  - “de” - differentially expressed genes between each pair of labels (largest difference in medians); final set is the union of all pairwise sets
  - “sd” - genes with largest standard deviation of label-wise medians
  - “all” - no feature selection
  - pre-defined set

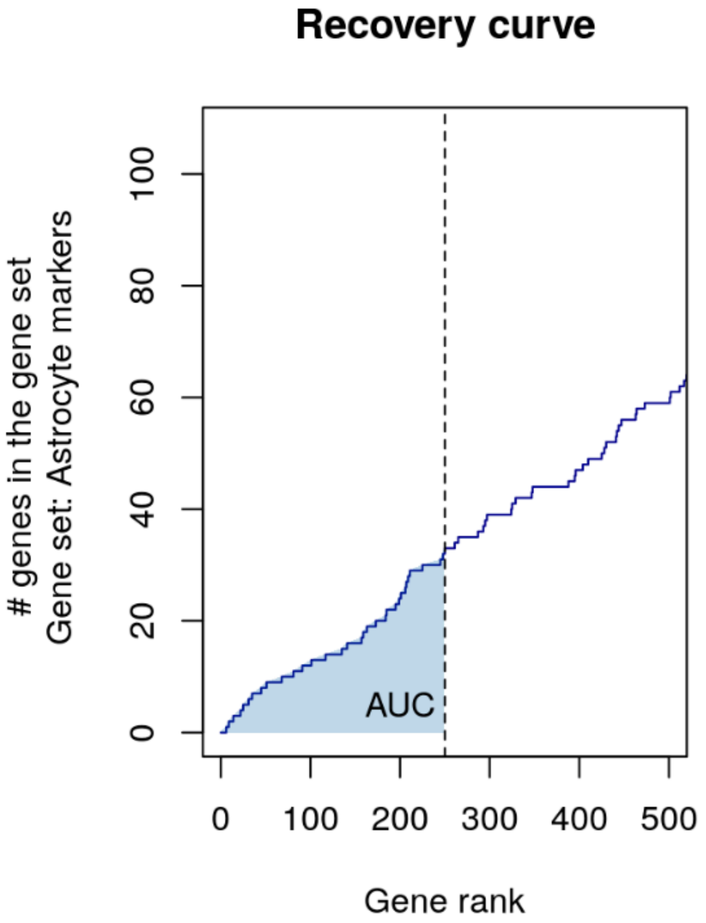
Designed for bulk references! For single-cell references, use other DE criterion (e.g., scran::pairwiseTTests()) or aggregate cells into pseudo-bulk samples

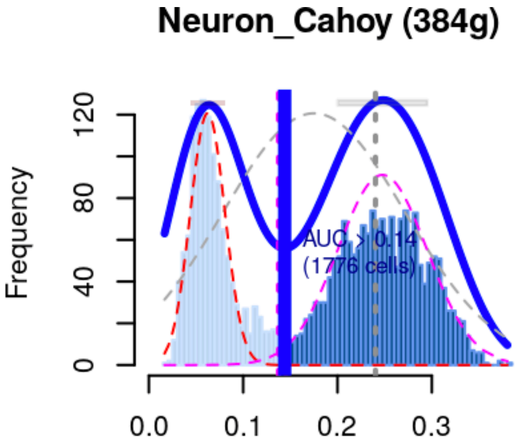
singleR built-in reference data sets

Data retrieval	Organism	Samples	Sample types	No. of main labels	No. of fine labels	Cell type focus
`HumanPrimaryCellAtlasData()`	human	713	microarrays of sorted cell populations	37	157	Non-specific
`BlueprintEncodeData()`	human	259	RNA-seq	24	43	Non-specific
`DatabaseImmuneCellExpressionData()`	human	1561	RNA-seq	5	15	Immune
`NovershternHematopoieticData()`	human	211	microarrays of sorted cell populations	17	38	Hematopoietic & Immune
`MonacoImmuneData()`	human	114	RNA-seq	11	29	Immune
`ImmGenData()`	mouse	830	microarrays of sorted cell populations	20	253	Hematopoietic & Immune
`MouseRNAseqData()`	mouse	358	RNA-seq	18	28	Non-specific

AUCell - cell annotation using gene sets

- In each cell, rank genes by expression
- Evaluate enrichment of genes in a gene set, using the AUC (area under the recovery curve)
- Outputs gene set "activity" score, which can be used to annotate cells, or as a summary representation of the data set (using a large number of gene sets as the “features")





Aibar *et al*, Nature Methods 14:1083–1086 (2017)

References

- Aibar *et al*: SCENIC: Single-cell regulatory network inference and clustering. Nature Methods 14(11):1083-1086 (2017)
- Amezquita *et al*: Orchestrating single-cell analysis with Bioconductor. bioRxiv doi:10.1101/590562 (2019)
- Aran *et al*: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology 20:163-172 (2019)
- Finak *et al*: MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology 16:278 (2015)
- Korthauer *et al*: A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology 17:222 (2016)
- Law *et al*: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15:R29 (2014)
- Love *et al*: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550 (2014)
- Lun *et al*: A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 5:2122 (2016)
- Lun *et al*: It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. Methods in Molecular Biology 1418:391–416 (2016)
- Risso *et al*: A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications 9:284 (2018)
- Robinson *et al*: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140 (2010)
- Soneson and Robinson: Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods 15:255-261 (2018)
- Svensson: Droplet scRNA-seq is not zero-inflated. bioRxiv doi:10.1101/582064 (2019)
- Van den Berge *et al*: Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biology 19:24 (2018)
- Vieth *et al*: powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics 33(21):3486-3488 (2017)
- Zhang *et al*: Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nature Methods 16:1007-1015 (2019)