Introduction to single-cell RNA-seq analysis - Normalisation

Stephane Ballereau

July 2021

Outline

- Motivation
- Initial methods
- Deconvolution
- sctransform

Motivation

Systematic differences in sequencing coverage between libraries occur because of:

- low input material,
- differences in cDNA capture
- differences in PCR amplification.

Normalisation removes such differences so that differences between cells are not technical but biological, allowing meaningful comparison of expression profiles between cells.

Normalisation and batch correction have different aims:

Normalisation addresses technical differences only, while batch correction considers both technical and biological differences.

Sources: chapters on normalisation in the OSCA book, the Hemberg group materials and sctransform.

Initial methods

- In scaling normalization, the "normalization factor" is an estimate of the library size relative to the other cells.
- Steps usually include:
  - computation of a cell-specific 'scaling' or 'size' factor
    - that represents the relative bias in that cell
  - division of all counts for the cell by that factor to remove that bias.
- Assumption: any cell specific bias will affect genes the same way.

Examples

CPM: convert raw counts to counts-per-million (CPM)

- for each cell
- by dividing counts by the library size then multiplying by 1.000.000.
- does not address compositional bias caused by highly expressed genes that are also differentially expressed between cells.

DESeq's size factor

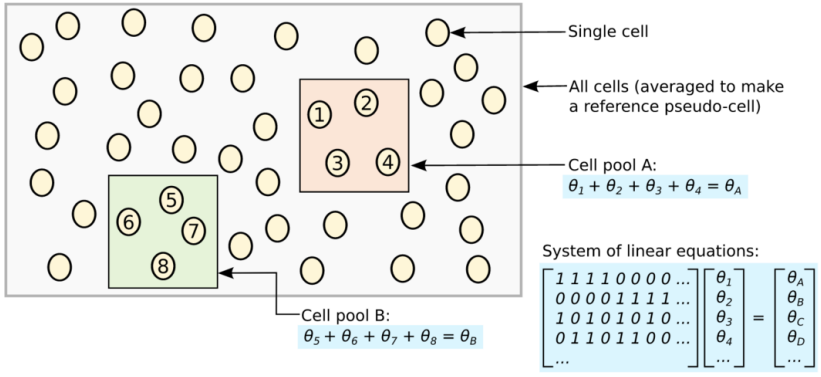
- For each gene, compute geometric mean across cells.
- For each cell
  - compute for each gene the ratio of its expression to its geometric mean,
  - derive the cell's size factor as the median ratio across genes.
- Not suitable for sparse scRNA-seq data as the geometric is computed on non-zero values only.

Deconvolution

Deconvolution strategy Lun et al 2016:

The deconvolution method consists of several key steps:

- Defining a pool of cells
- Summing expression values across all cells in the pool
- Normalizing the cell pool against an average reference, using the summed expression values
- Repeating this for many different pools of cells to construct a linear system
- Deconvolving the pool-based size factors to their cell-based counterparts (Fig. 3)



Steps:

- compute scaling factors,
- apply scaling factors

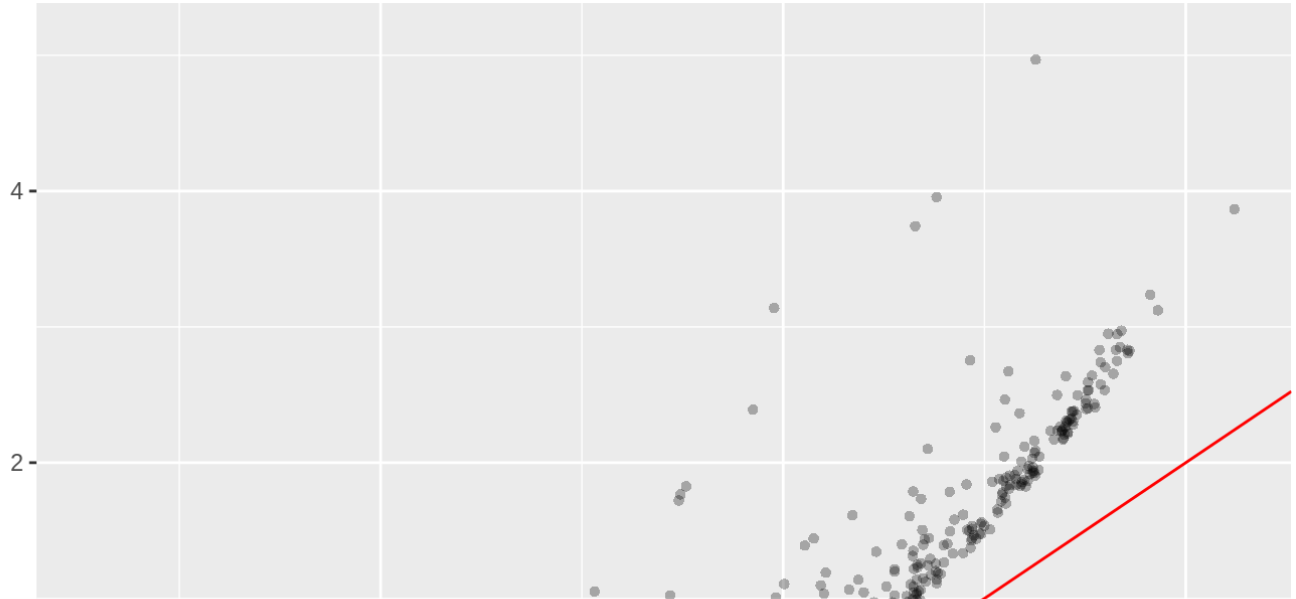
sctransform

With scaling normalisation a correlation remains between the mean and variation of expression (heteroskedasticity).

This affects downstream dimensionality reduction as the few main new dimensions are usually correlated with library size.

sctransform addresses the issue by:

- regressing library size out of raw counts
- providing residuals to use as normalized and variance-stabilized expression values



sctransform

Variables

- model the expression of each gene as a negative binomial random variable with a mean that depends on other variables
- which model the differences in sequencing depth between cells
- and used as independent variables in a regression model

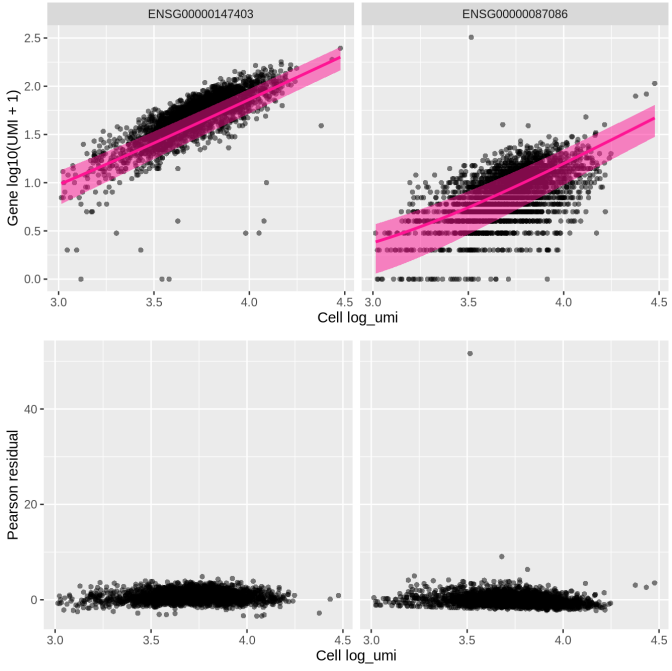
Regression

- fit model parameters for each gene
- combine data across genes using the relationship between gene mean and parameter values to fit parameters
- transform each observed UMI count into a Pearson residual
  - $\sim$ number of standard deviations away from the expected mean
- expect mean of 0 and stable variance across the range of expression

sctransform

Example of the transformation outcome for two genes:

- UMI counts and pearson residuals against library size
- with expected UMI counts in pink



Recap

Early methods developed for bulk RNA-seq are not appropriate for sparse scRNA-seq data.

The deconvolution method draws information from pools of cells to derive cell-based scaling factors.

The sctransform method uses sequencing depth and information across genes to stabilise expression variance across the expression range.