ELIXIR EXCELERATE course on single cell RNA-seq data analysis

Normalization

27.5.2019

Heli Pessa

University of Helsinki

&

SCellex

Normalization

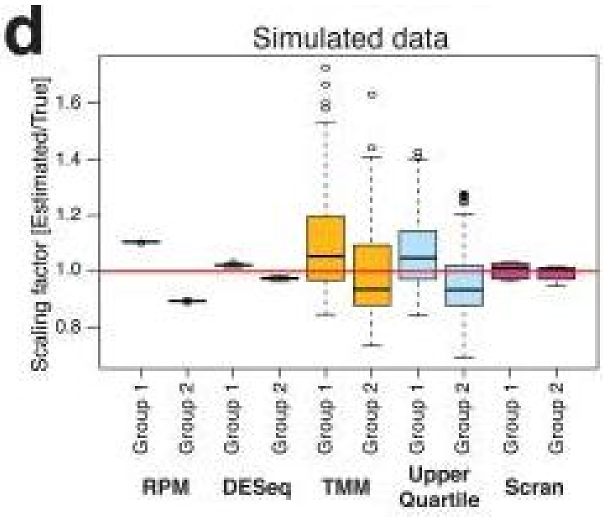
- Removing systematic non-biological variation
- Making count distributions comparable
- With 3’ tagged data, only cell-specific normalization is usually done
- In case of full-length data, normalization for gene length must also be done

Normalization aims

- Normalized expression of a gene should not correlate with the sequencing depth of the cell
- Variance of normalized gene expression should reflect biological variation across cells

scRNA-seq differs from bulk RNA-seq

- **Noise**
  - Low mRNA content per cell
  - Variable mRNA capture
  - Variable sequencing depth



Vallejos et al. (2017)

- **Different cell types in the same sample**
- **Bulk RNA-seq normalization methods (CPM, TMM, upper quartile) do not work well**

Estimation of technical variance

- **Spike-in RNA**
  - Used mainly in plate-based library methods
- **Whole data**
  - Assumption: most genes do not change expression

Normalization methods

- **Main approaches**
  1. Size factors
  2. Probabilistic methods
     - Zero-inflated negative binomial (ZINB) models

Size factor methods

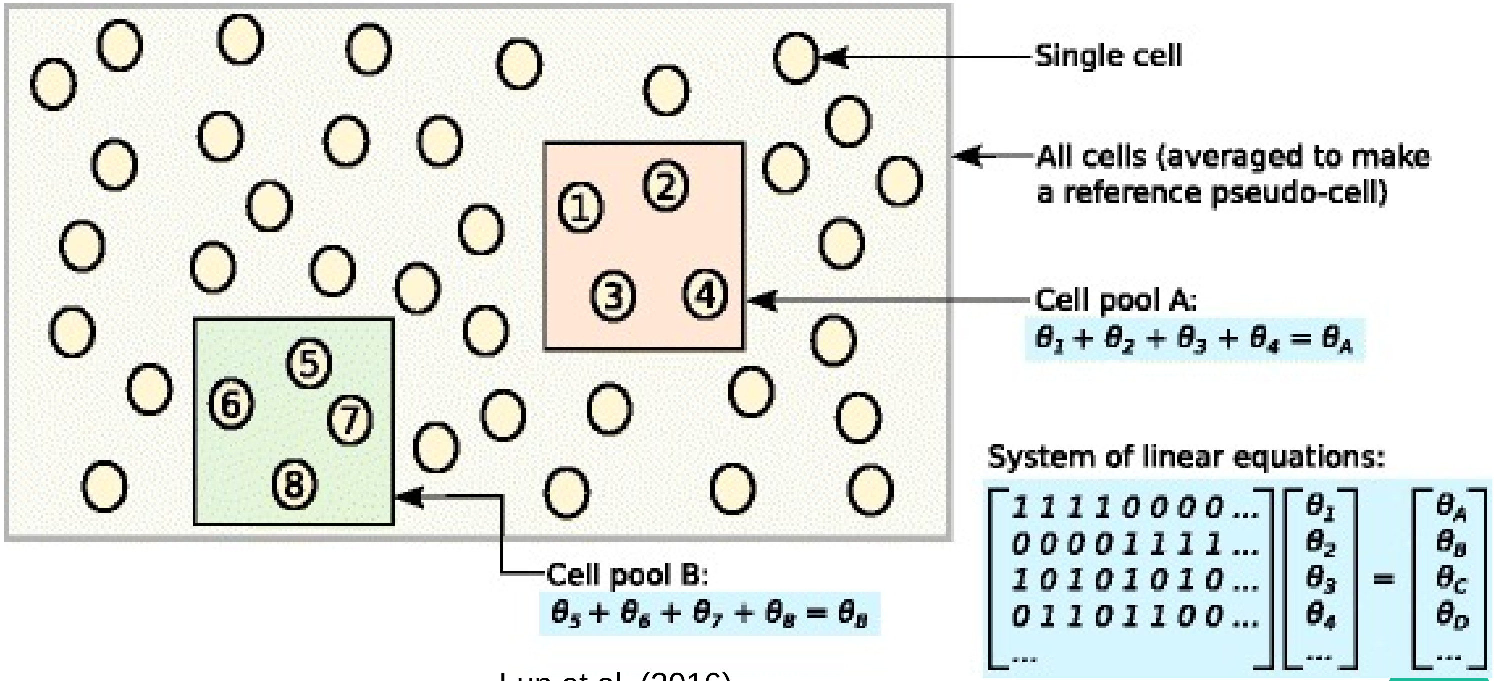
- **Bulk RNA-seq normalization methods are based on per-gene statistics – not suitable for zero-inflated data**
- **CPM, TPM**
  - Not sufficient for scRNA-seq data
- **TPKM, FPKM**
  - Full-length transcriptome only
- **DESeq**
  - Size factors may be zero

Size factor methods

- **Global scaling**
  - Assumption: RNA levels do not vary much between cells
  - Modified CPM normalization
  - Seurat, 10X Cell Ranger: log-normalization
-

Size factor methods

- **Deconvolution**
  - Pooling across cells, normalization to reference
  - Deconvolution of per-cell size factors
  - scran


Single cell
All cells (averaged to make a reference pseudo-cell)
1
2
3
4
Cell pool A:
$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$
5
6
7
8
Cell pool B:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
System of linear equations:
$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0 \ldots \\ 0\,0\,0\,0\,1\,1\,1\,1 \ldots \\ 1\,0\,1\,0\,1\,0\,1\,0 \ldots \\ 0\,1\,1\,0\,1\,1\,0\,0 \ldots \\ \ldots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \ldots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \ldots \end{bmatrix}$

Lun et al. (2016)

Size factor methods

- **BASICS**
  - Bayesian model for estimating cell-specific constants
  - (Originally) requires spike-ins

Feature selection

Selecting genes

- **Excluding invariable genes that do not contribute informative/interesting information**
  - Improved signal to noise ratio
  - Reduced computational requirements
- **Highly variable genes**
- **Correlated gene pairs/groups**
- **Top PCA loadings**

Highly variable genes

- **Genes which behave differently from a null model describing technical noise**
  - Mean-variance trend: genes with higher than expected variance
  - Coefficient of variation (Brennecke et al. 2013)
- **High dropout genes**
  - Number of zeros unexpectedly high compared to null model

Gene correlations

- **Principle: multiple genes will be differentially expressed between different cell types**
  - Assumes technical noise is random and independent for each cell
  - Batch effects violate