RNASeq analysis tutorial

Andrea Tangherloni
Department of Computing Sciences, Bocconi University
andrea.tangherloni@unibocconi.it

Università Bocconi
DEPARTMENT OF COMPUTING SCIENCES

Normalization

Normalization is the process **of scaling raw count** values to **account** for the **“uninteresting” factors**

RNA expression (“interesting”)

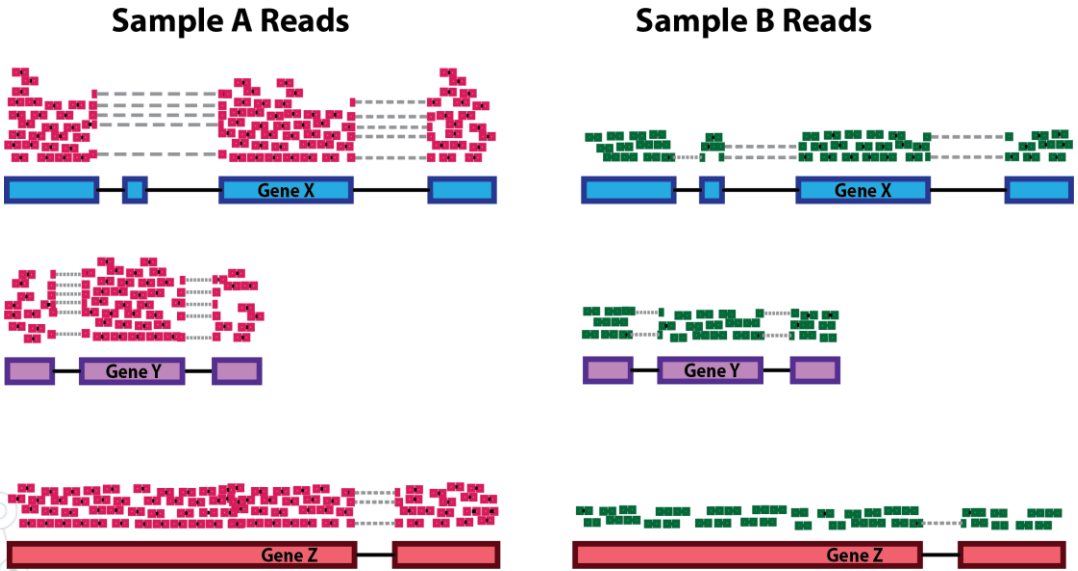
Other factors (“uninteresting”)

Normalization

- **Normalization** is essential for **differential expression analysis**
- It can also be fundamental for **exploratory analyses**
  - Data analysis
  - Data visualization
  - Whenever you are exploring or comparing counts between or within samples

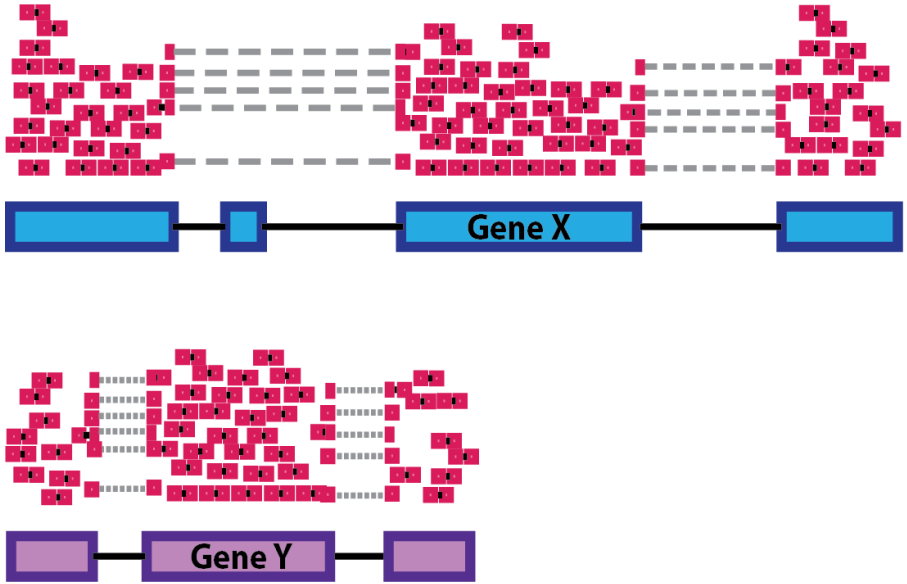
Sequencing depth

- Necessary for **comparing gene expression** between different **samples**



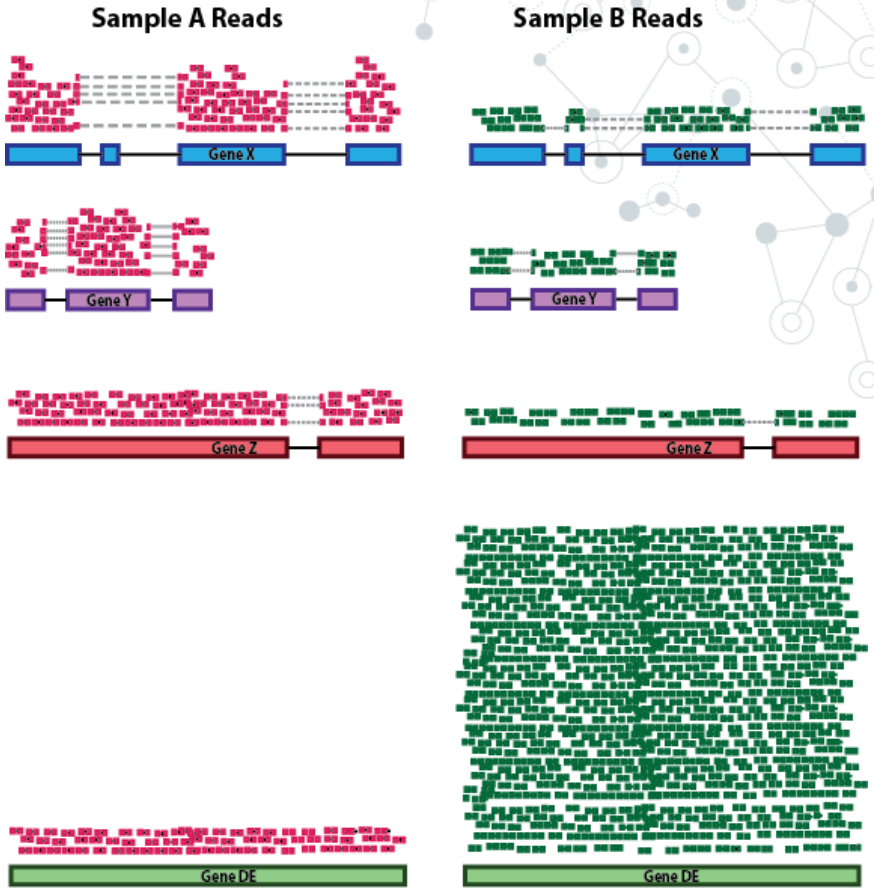
Gene length

- Necessary for **comparing expression** between different **genes** within the **same sample**



RNA composition

- Necessary for accurate **comparison** of **expression between samples**
- Particularly **important** when performing **differential expression analyses**



Common normalization methods

- There are **different normalization** methods that can be used to account for the previous issues
  - Counts per million (CMP)
  - Transcripts per kilobase million (TMP)
  - Reads/fragments per kilobase of exon per million reads/fragments mapped (RPKM/FPKM)
  - DESeq2's median of ratios
  - EdgeR's trimmed mean of M values (TMM)

Common normalization methods

Method	Description	Accounted factors	Recommendations for use
CPM	Counts scaled by total number of reads	Sequencing depth	Gene count comparisons between replicates of the same sample group; **NOT for within sample comparisons or DE analysis**
TPM	Counts per length of transcript (kb) per million reads mapped	Sequencing depth and gene length	Gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM	Similar to TPM	Sequencing depth and gene length	Gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2	Counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	Sequencing depth and RNA composition	Gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR	Uses a weighted trimmed mean of the log expression ratios between samples	Sequencing depth, RNA composition, and gene length	Gene count comparisons between and within samples and for **DE analysis**

DESeq2-normalized counts

- For **differential expression analysis**, we compare the **counts between sample groups** for the same gene
  - Gene length does not need to be accounted
  - Sequencing depth and RNA composition must be considered

- **DESeq2** uses the **median of ratios method**
  - Create a pseudo-reference sample
  - Calculate the ratio of each sample to the reference
  - Calculate the normalization factor for each sample
  - Calculate the normalized count values

Step 1

- For each gene, a **pseudo-reference sample** is created that is equal to the **geometric mean** across all samples

Gene	Sample A	Sample B	pseudo-reference sample (PRS)
gene_1	1489	906	sqrt(1489 * 906) = **1161.5**
gene_2	22	13	sqrt(22 * 13) = **17.7**
…	…	…	…

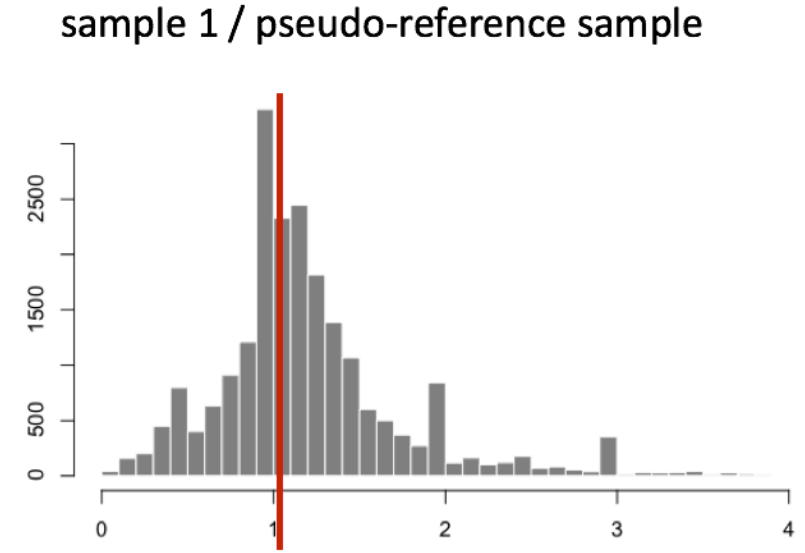
Step 2

- For every gene in a sample, **the ratios** (sample/ref) are calculated

Gene	Sample A	Sample B	PRS	ratio of sample A/ref	ratio of sample B/ref
gene_1	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
gene_2	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
…	…	…	…	…	…

Step 3

- The **median value** (column-wise) of **all ratios** for a given sample is used as the **normalization factor** (size factor) for that sample
- The median ratio
  - Sample A = 1.29
  - Sample B = 0.775



Step 3

- The **median of ratios** method assumes that **not ALL genes** are **differentially expressed**
  - The normalization factors should account for sequencing depth and RNA composition of the sample
  - Large outlier genes will not represent the median ratio values

- This method is **robust** to the **imbalance** in **up-/down-regulation** and **large numbers** of differentially expressed genes

Step 4

- This step is performed by dividing each **raw count value** in a given sample by that **sample's normalization factor**

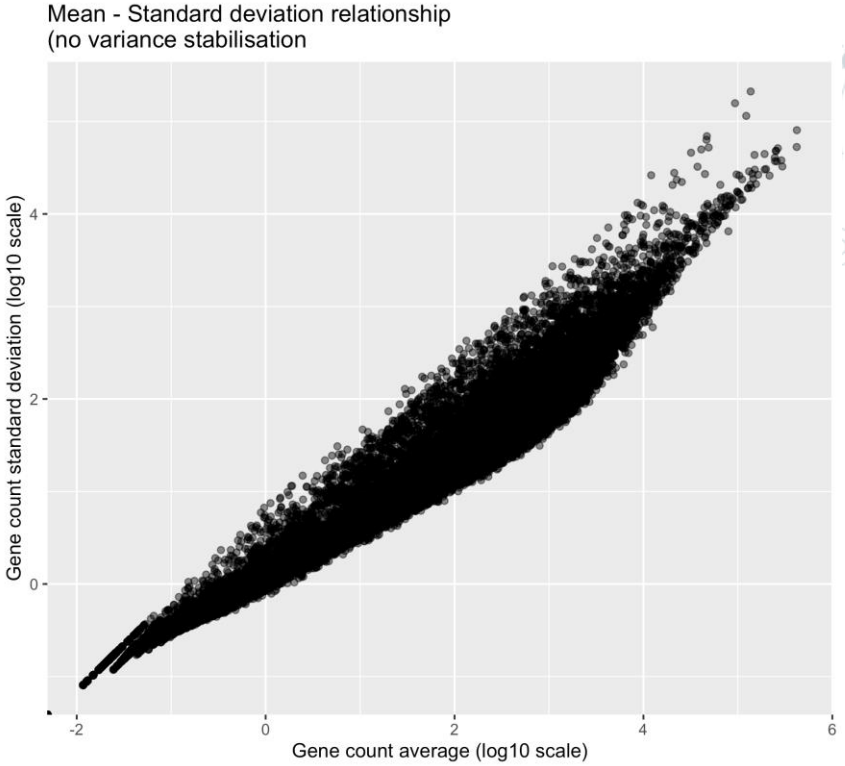
Gene	Sample A	Sample B	Normalized Sample A	Normalized Sample B
gene_1	1489	906	1489 / 1.29 = **1154.26**	906 / 0.775 = **1169.03**
gene_2	22	13	22 / 1.29 = **17.05**	13 / 0.775 = **16.77**
…	…	…	…	…

Variance stabilization

- **PCA** is used to **visualize** the **differences** (i.e., distances) between **samples** and how it relates to the **experimental design**
  - Samples from the same experimental condition should group together if the observed variability in the data relates to the experimental design
- **Genes** from **biological replicates** (i.e., samples of the same experimental condition) should **behave similarly** and result in **similar sample scores**

Variance stabilization

- In RNA-Seq data, the **gene variance** is **proportional** to the **gene mean**
  - The higher the gene mean is, the more variance it has
  - Genes with a low abundance (low counts) also suffer from a somehow inflated variance



Variance stabilization

- A **variance-stabilizing transformation** (VST) is a monotonous mapping such that for the transformed values
  - The variance is (approximately) independent of the mean