Differential expression in RNA-seq: A matter of depth

Sonia Tarazona[1,2], Fernando García-Alcalde[1], Joaquín Dopazo[1], Alberto Ferrer[2], Ana Conesa[1]

[1] *Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia*

[2] *Department of Applied Statistics, Operations Research and Quality, Univ. Politécnica Valencia*

Aim

RNA-seq technology has brought a great progress for omics sciences and has posed a new challenge in the bioinformatics field, since new algorithms and methods are needed to deal with the data generated by this new technique. But the methodology itself may introduce biases in the quantification of expression level, which is the number of reads mapping to a gene, transcript, exon, etc. (*counts*). In our work, we explore the influence of the **sequencing depth** (total number of reads mapped to the reference genome) and the **length** (number of bases) of the gene on the gene **expression level**, especially when computing differential expression. A **method** has been also developed to assess **differential expression** between two experimental conditions. It is called **NOISeq** and it is robust with respect to sequencing depth and length, do not need replicates and can detect genes with low expression level.

Some RNA-seq biases in public datasets

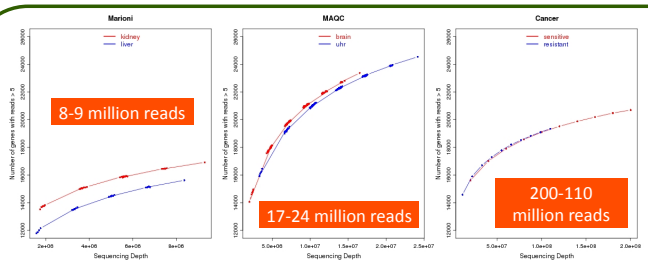


Fig. 1 ***SEQUENCING DEPTH****. Number of genes with more than 5 reads for several sequencing depths in three different public datasets [1,2,3]. The more it is sequenced the more it is detected. No plateau reached, even with 200 million reads.*

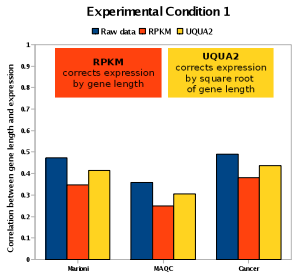


Fig. 2 ***LENGTH***
Correlation between gene expression and gene length in the three datasets [1,2,3]. Blue bar represents this correlation using the raw data. Red and yellow bars, correlation when applying normalization to correct length bias (RPKM [4] and Upper Quartile [2], respectively) . Normalization procedures partially reduce the medium correlation observed in raw data.

NOISeq

- No parametric assumptions
- No need of replicates
- Pairwise comparisons

1 For each gene, exon, transcript…

$$M = \log_2\left(x_1 / x_2\right)$$

$$D = |x_1 - x_2|$$

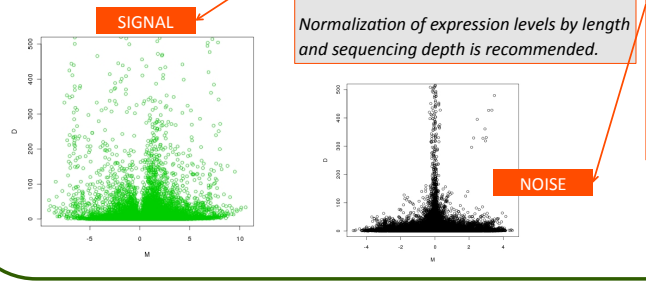
x_i = expression level in sample i

Normalization of expression levels by length and sequencing depth is recommended.

2 **Noise distribution**: M-D values comparing replicates within the same experimental condition:

- **NOISeq-real:** uses available replicates
- **NOISeq-sim:** simulates replicates from a multinomial distribution with probabilities proportional to gene counts in the samples.

3 **Probability of differential expression**: Computed by comparing M-D values of a gene against noise distribution.
A gene is declared as **differentially expressed** if its **probability > 0.8**.



SIGNAL
NOISE

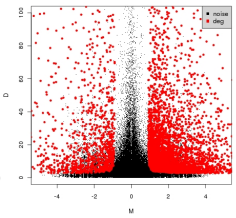


Fig. 3
M-D values in noise (black dots) and for differentially expressed genes (red dots).

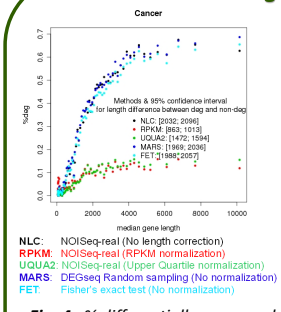
DEGseq methods

- **MARS**: MA-plot based method using random sampling to estimate noise level [5].
- **FET:** Fisher's exact test (non-parametric method).
- **LRT**: Likelihood ratio test based on Poisson model [1].

Assuming that M conditioned to A follows an approximate normal distribution

Our method **NOISeq** is to be compared to DEGseq methods

Gene length & Differential expression



NLC: NOISeq-real (No length correction)
RPKM: NOISeq-real (RPKM normalization)
UQUA2: NOISeq-real (Upper Quartile normalization)
MARS: DEGseq Random sampling (No normalization)
FET: Fisher's exact test (No normalization)

Fig. 4 *% differentially expressed genes detected for each different gene length with each method in dataset [3].*

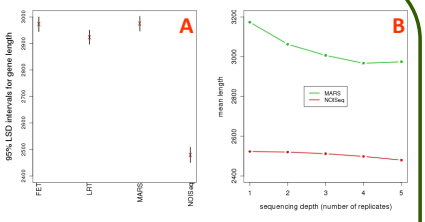


Fig. 5 *Dataset [1]*
A: 95% LSD intervals for average gene length of genes declared as differentially expressed by each method.
B: Mean gene length of differentially expressed genes for each sequencing depth.

Expression level & Differential expression

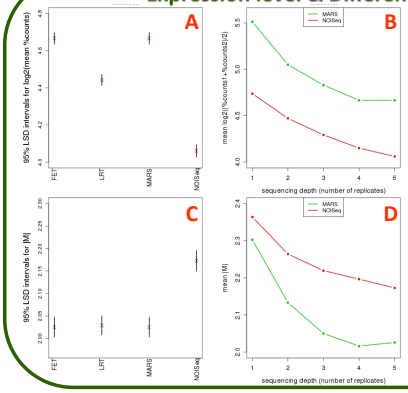


Fig. 6 *Dataset [1]*
A: 95% LSD intervals for average expression of genes declared as differentially expressed by each method.
B: Mean expression of differentially expressed genes according to sequencing depth.
C: 95% LSD intervals for average M value of genes declared as differentially expressed by each method.
D: Mean M of differentially expressed genes according to sequencing depth.

Conclusions

- Length and expression level for differentially expressed genes tend to decrease with higher sequencing depth.
- NOISeq detects shorter genes than methods in NOISeq and it does not depend so much on sequencing depth.
- NOISeq detects differential expression in genes with lower counts than methods in DEGseq, while differential expression value keeps higher.
- Normalization helps to correct the dependence of differential expression on sequencing depth and on gene length.

References

[1] Marioni *et al.* (2008) Genome Research, 18, 1509-1517.
[2] Bullard *et al.* (2010) BMC Bioinformatics, 11, 94+.
[3] Griffith *et al.* (2010) Nature Methods, 7: 843-847.
[4] Mortazavi *et al.*(2008) Nature Methods, 5: 621-628.
[5] Wang *et al.* (2010) Bioinformatics, 26, 136-138.