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 is increasingly being used for gene expression profiling. However, the properties of RNA-seq data have not been yet fully established, and additional research is needed for understanding how these data respond to differential expression analysis. In this work, we study how the **sequencing depth affects the detection of transcripts and their identification as differentially expressed**. We evaluate some differential expression algorithms and propose a novel approach, NOISeq. Our results reveal that most existing methodologies suffer from a strong dependency on sequencing depth for their differential expression calls, that results in an increasing number of false positives as the number of reads grows. **NOISeq** models the noise distribution from the actual data, so it can better adapt to the size of the data set, and is more effective in controlling the rate of false discoveries.

Gene detection & sequencing depth

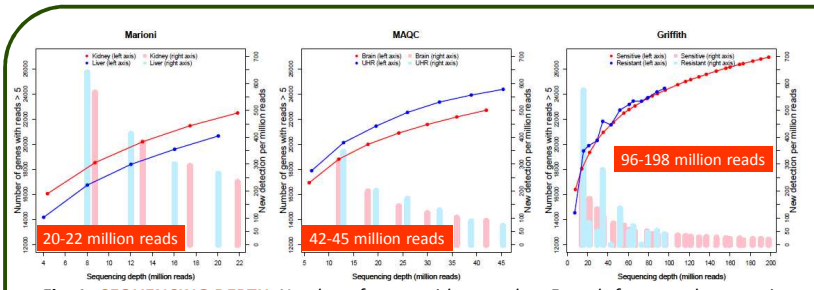


Fig. 1 ***SEQUENCING DEPTH****. Number of genes with more than 5 reads for several sequencing depths in three different public datasets [1,2,3]. The more sequenced the more detected. No plateau reached, even with 200 million reads.*

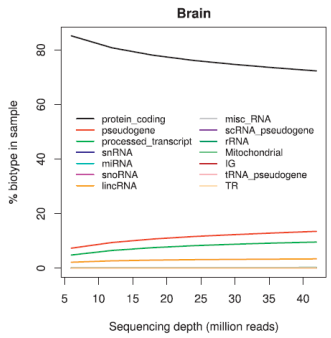


Fig. 2 ***BIOTYPE DETECTION*** *Percentage of gene biotype in the brain sample (dataset [2]) at different sequencing depths. The distribution of biotypes observed among detected features evolve with increasing sequencing depth, with the relative abundance of protein-coding transcripts steadily decreasing, whereas noncoding genes gain a proportional presence.*

NOISeq

- No parametric assumptions
- Can work without replicates
- Pairwise comparisons

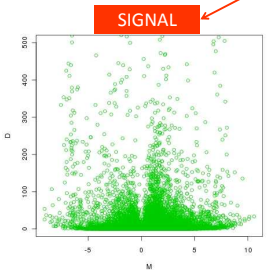
1 For each gene, exon, transcript…

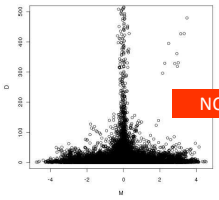
$$M = \log_2\left(\frac{x_1}{x_2}\right)$$

$$D = |x_1 - x_2|$$

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 (recommended)
- **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 **probability > 0.8**.

SIGNAL
NOISE

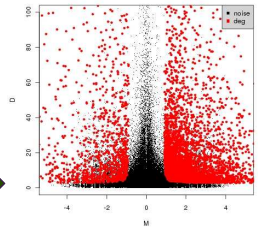


Fig. 3 *M-D values in noise (black dots) and for differentially expressed genes (red dots).*

Performance of differential expression methods

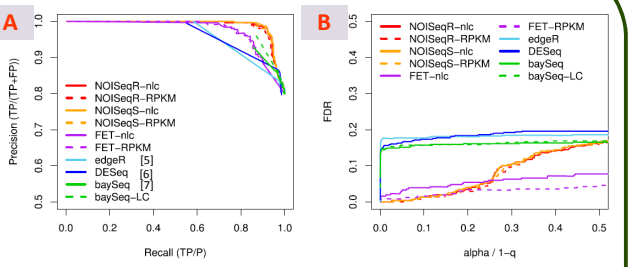
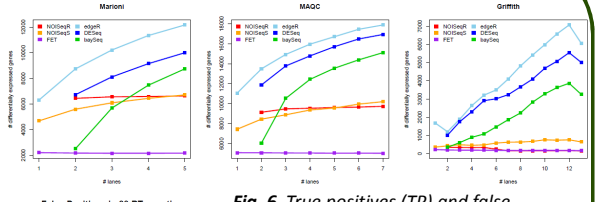


Fig. 4 *Precision-recall curves (A) and false discovery rates (B) for the differential expression methods compared on data set [2] using RT-PCR results as a gold-standard. NOISeq outperforms the other methods (A) while keeping a low false discovery rate (B).*
FET method is Fisher's Exact test. "nlc" means "no length correction" and "RPKM" is Reads Per Kilobase and per Million Reads [4].

Differential expression & Sequencing depth

Fig. 5 *Differentially expressed genes according to sequencing depth for each dataset and method (nlc). NOISeq and FET results are more robust to sequencing depth than the other methods.*



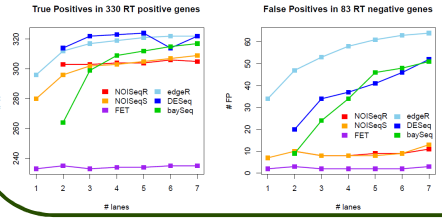


Fig. 6 *True positives (TP) and false positives (FP) at each sequencing depth. TP and FP were obtained applying the differential expression methods on dataset [2] and comparing the results to RT-PCR positive and negative genes. NOISeq gets a nice balanced rate of TP and FP. FET obtains a low FP rate, while the other methods present a high FP rate.*

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] Robinson *et al.* (2010) Bioinformatics, 26: 139-140.
[6] Anders and Huber (2010) Genome Biol, 11: R106.
[7] Hardcastle and Kelly (2010) BMC Bioinformatics, 11:422.

Conclusions

- The sequencing depth affects the detection and expression quantification of the transcripts.
- As more sequencing output is considered, the diversity and quantity of detected off-target RNA species increase.
- NOISeq method shows a good performance when comparing it to other differential expression methods: Fisher's Exact Test (FET), edgeR [5], DESeq [6] and baySeq [7].
- NOISeq and FET are more robust to sequencing depth than the other methods: the number of differentially expressed genes keeps similar at increasing sequencing depths.
- NOISeq maintains good True Positive and False Positive rates when increasing sequencing depth, while FET shows a poor detection rate and the other methods generate an increasing number of False Positives.