Nonparametric Simulation to Evaluate the Performance of Statistical Methods for DGE Analysis in lncRNA-seq

Alemu Assefa[1], Olivier Thas[1,2], Jo Vandesompele[3], Katrijn De Paepe[4]

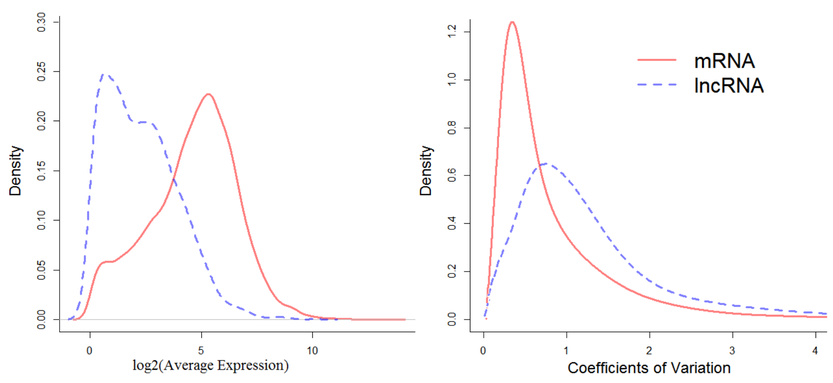
[1]Dept. of Mathematical Modeling, Statistics and Bioinformatics, UGent. [2]NIASR, University of Wollongong,
[3]Dept. of Pediatrics and medical Genetics, UGent. [4]Bain & Company, Belgiumt

Long Non-coding RNA (lncRNA)

- lncRNAs form a large and diverse class of transcribed RNA molecules, constituting up to 70% of the transcriptome with a length of $\geq$ 200 nucleotides,
- they do not encode proteins but are involved in fundamental gene regulatory mechanisms
- The discovery and study of lncRNAs is of major relevance to human biology and disease since they represent an extensive, largely unexplored, and functional component of the genome, for example for cancer therapeutic and diagnostic studies

Statistical Issues in lncRNA-seq

Sequencing counts of lncRNAs (lncRNA-seq) from Next Generation Sequencing (NGS) technologies (e.g. RNA-sequencing) often show unique characteristics compared to mRNA-seq data



In general, lncRNA-seq are characterized by low expression and high variability.

LOW Signal and HIGH Noise

Statistical tools

- Differential Gene Expression (DGE) testing is a statistical hypothesis testing on the difference in the abundance of genes/transcripts between two biological or experimental conditions. It is simultaneous test across many genes/transcripts
- We compared the most popular statistical tools introduced for testing DGE in RNA-seq data.

Class	Tools
Negative Binomial Models	edgeR, DESeq2, QuasiSeq
Log-Linear Modelling (Poisson)	PoissonSeq
Normal Linear Modelling	limma (voom)
Nonparametric Test	SAMseq

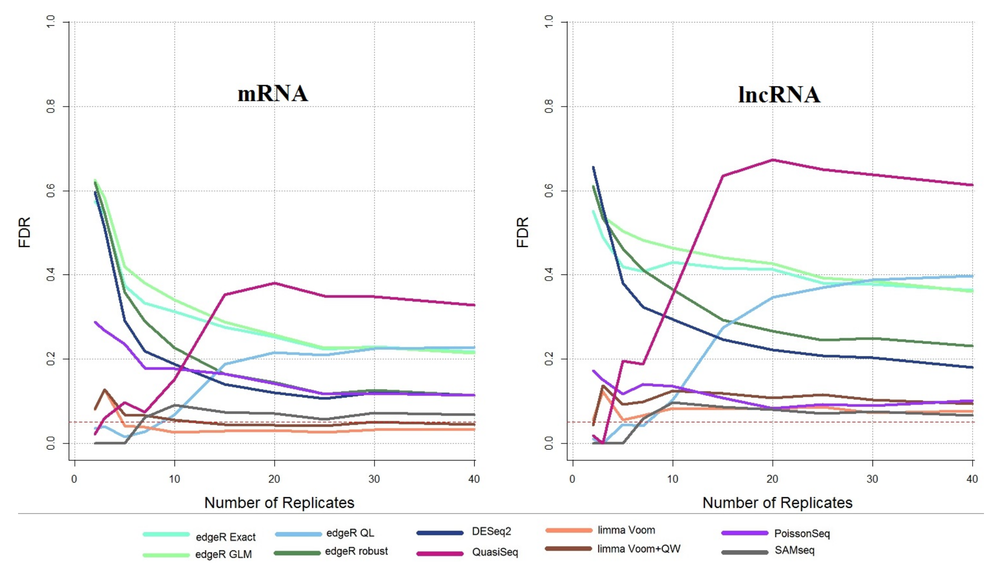
- Most of these tools are designed with an assumption that genes are expressed at sufficient level, for example genes with average expression greater than 1.
- However, majority of lncRNAs are expressed at low amount and this leads to the question that "Do these DGE tools perform at a *desired* level for lncRNAs?"

Note This assumption is not explicitly mentioned but all software packages come with a recommended default cut off to filter genes and various comparative studies demonstrated that the tools perform worse for low count genes due to higher noise.

Nonparametric Simulation

- Several comparative studies of DE tools used parametric assumption (e.g. Negative Binomial or Poisson distribution) to simulate gene expressions that resulted optimistic comparisons
- In our study we used **nonparametric simulation** with *SimSeq* R Bioconductor package
- This simulation starts with a source real RNA-seq data that has large number of replicates (also includes mRNAs and lncRNAs) and then
- by subsampling of samples and genes from the source RNA-seq data, realistic gene expressions are generated in such a way that the simulated counts reflect the underlining characteristics of the source data.
- Wide range of scenarios are simulated,
  - → Two gene biotypes: mRNA and lncRNA
  - → Proportion of truly DE genes: 0 - 30%
  - → Number of replicates per group: 2 - 40
  - → Two source real RNA-seq datasets (with different homogeneity of samples)

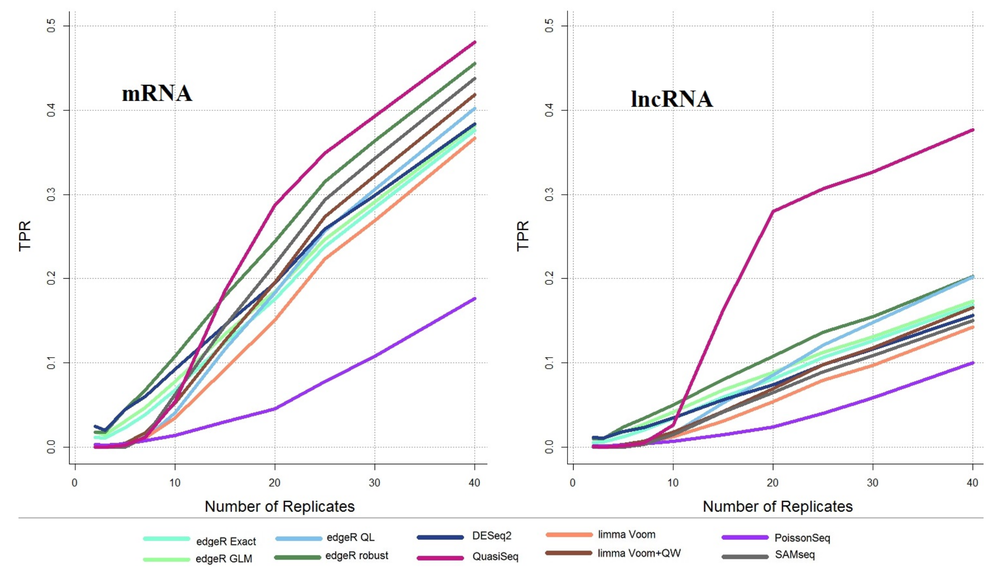
False Discovery Rate (FDR)



False Discovery Rate (FDR)

- FDR is the average proportion of false discoveries (rejection of true null hypothesis H_0: $\mu_{g1} = \mu_{g2}$) among all discoveries. $FDR = E\{\frac{FP}{FP+TP}\}$
- The true FDR is expected to be at most the desired level (e.g. 5%) but discrete distribution based models (edgeR, DESeq, QuasiSeq, and Poissonseq) showed higher FDR which is worse for lncRNAs than for mRNAs.

True Positive Rate (TPR)



True Positive Rate (TPR)

- TPR is the average proportion of truly DE genes that are correctly identified by the tool $TPR = E\{\frac{TP}{TP+FN}\}$
- The results again signifies that majority of the tools have much less power to correctly identify truly differentially expressed lncRNAs,

Conclusion

- Most tools showed high FDR for lncRNAs than for mRNAs
- None of the tools were able to achieve 50% power for a typical RNA-seq experiment designs with up to 40 number of replicates (especially for cancer study)
- Strong dependence on the true proportion of DE genes.
- Negative Binomial models (edgeR and DESeq) are powerful for designs with small sample size at a cost of high FDR
- Normal linear models (limma) and nonparametric tests (SAMSeq) control FDR much better with competitive power but for designs with sample size $\geq$ 10