Deep Transcriptome Profiling - A comparison of gene expression analysis programs for RNAseq

Willian A. da Silveira¹, E. Starr Hazard¹,², Dongjun Chung³, Gary Hardiman¹,³

¹MUSC Bioinformatics, Center for Genomics Medicine, ²Library Science and Informatics, ³Department of Public Health Sciences, Medical University of South Carolina (MUSC), Charleston SC

Background
- Different methods have evolved over the past five years to significantly determine differentially expressed genes from a comparison of the transcriptome under different physiological conditions (e.g. healthy vs diseased tissue).
- Selecting the best method is guided by the experimental design.

Methodology
- Next-generation RNA sequencing experiments were performed in HCC827 human lung adenocarcinoma cells expressing ZEB1 (a well-known Epithelial-to-Mesenchymal Transition stimulator) or empty vector and extracted from GEO (accession number GSE81167).
- The FASTQ files were downloaded and the quality of the FASTQ files were assessed using the FASTQC and preprocessing using Cutadapt and Trimmomatic.
- These processes increase the quality and reliability of the analyses and diminish the computational cost and execution time. We used the reference GRC37/hg19 human genome and the Tophat aligner. We then performed differential expression (DE) analysis using the TUXEDO pipeline, and after defining mRNA level counts with HTSeq, the DESeq2, edgeR and Limma/voom programs.

Results
- Using a Bonferroni-Hochberg adjusted p-value ≤ 0.1, TUXEDO uncovered 476 DE genes, DESeq2 found 186, edgeR 540 and Limma/voom 155.
- Considering only the lowest 10 adj p-values, no mRNAs were shared across all methods.
- For the complete list, DESeq2 provided the most overlap, i.e. 43.0% with TUXEDO, 58.0% with Limma/voom and 94.6% with edgeR.
- edgeR yielded the greatest list of DE genes. Tuxedo has a 59.6% of overlap with edgeR. DESeq2 has 94.6%, and Limma 81.2%, but only 48.0% of edgeR DE genes are unique compared to only 0.01% with DESeq2.
- Only 19.4% of the mRNAs from Limma/voom overlap the TUXEDO results, the lowest agreement.
- Even with this gene level discordance, the systems level analysis identified the same pattern of affected pathways.

Conclusion
- Between this methods, gene-level analysis of significance show a variability, but system level analysis revealed consistency.

Bibliography

If you are interested in including RNAseq analysis in your work, please, contact us: hardiman@musc.edu