Beyond gene expression

Impact

The primary goal of the research is to add to the existing knowledge and deepen our understanding of a topic. The primary analysis of the RNA-Seq data is gene-based, however, in this thesis, we showed that a transcript-based analysis approach produces better biological interpretations. At the same time, it allows for novel analysis of the RNA-Seq data generating new hypotheses and results.

The study of comparison of different liver cell models through RNA-Seq data resulted in the generation of an exhaustive resource outlining the similarities and differences of the cell models to liver biopsies, as presented in chapter 2 (1). The cell models *ex-vivo* lose various *in-vivo* characteristics (2, 3). However, the unavailability of a comprehensive comparison restricted the assessment of the changes. A thorough study of the liver cell models through RNA-Seq data at the gene and transcript level illustrated the need for moving to a transcript-based approach. The biological changes in the cell models were highlighted with more precision using the transcript expression. The data generated through this comparison will be beneficial in selecting the cell models for specific research questions – what pathways, processes, and/or traits are of interest? Using the comparison data available, informed decisions can be made. This work can be used as a template to compare cell models from other tissues and cell types as well. A resource of cell models defining how similar or different they are to the *in vivo* systems would be helpful to perform better research.

The knowledge of different transcripts originating from the same or different genes making the same protein always existed (4, 5). However, while analyzing the RNA-Seq data the focus was always on evaluating the expression of individual genes or transcripts. The assimilation of this concept of the same protein from different transcripts to the RNA-Seq data analysis resulted in the creation of FuSe (6). Through this approach, the RNA-Seq data analysis provided more information on the dynamics of the biological system. However, various layers of regulation are involved in making a protein from the mRNA, the grouped expression calculation gives the preliminary protein expression estimates. This can be used as a starting point to accommodate other regulations and reach to the protein expression values. The work emphasized that the biological systems try to achieve homeostasis by producing different transcripts that code for the same proteins. The changes in the type and expression of these transcripts can be attributed to the internal or external environment. The identification of such transcripts that are involved in coding for the same proteins calls for studying their evolutionary relationships.

Using machine learning (ML) approaches with the transcript expression data novel potent transcript biomarkers were identified (7, 8). The study focused on various HCC cell models and presented a group of protein coding and non-coding transcripts as the potent biomarkers for detection of HCC. The inclusion of non-coding transcripts in the biomarker discovery results emphasized their underlying role in disease progression. The added advantage of using transcripts is that they are produced before the proteins in the system and can help in the early detection of the diseases. The identification of the transcripts as biomarkers can have a major impact on future of biomarker research. Novel and more potent biomarkers can then be identified for diseases where early diagnosis is a major hurdle.

Lastly, evaluating the expression of the protein complexes from the RNA-Seq data provided more functional assessment of the biological system. From an entity-based analyses (gene or transcript), we could then define the amount of work achievable in the biological system. Additionally, the information of the assembly of the protein complexes elevated our understanding of the multi-molecule machinery and opened new doors for investigating novel drug targets. The application of dynamic Bayesian networks to the temporal RNA-Seq data helped in generating hypothesis for assembly of protein complexes. Though preliminary, this work has the capacity to control and eliminate various diseases occurring due to mis-assembly of the protein complexes. Moreover, information on the assembly would highlight the evolutionary information. What genes moved apart and what came closer over years of evolution (9).

We demonstrated that the study of RNA-Seq data is no longer limited to evaluating genes or transcripts. It can be used to assess the amount of protein that can be formed in the biological system, present transcripts as potent disease biomarkers, can help elucidate protein complex assembly and more.

References

1. Gupta R, Schrooders Y, Hauser D, van Herwijnen M, Albrecht W, ter Braak B, et al. Comparing in vitro human liver models to in vivo human liver using RNA-Seq. Archives of Toxicology. 2021;95(2):573-89.
2. Sachinidis A, Albrecht W, Nell P, Cherianidou A, Hewitt NJ, Edlund K, et al. Road map for development of stem cell-based alternative test methods. Trends in molecular medicine. 2019.
3. Albrecht W, Kappenberg F, Brecklinghaus T, Stoeber R, Marchan R, Zhang M, et al. Prediction of human drug-induced liver injury (DILI) in relation to oral doses and blood concentrations. Archives of toxicology. 2019;93(6):1609-37.

4. Mariño-Ramírez L, Kann MG, Shoemaker BA, Landsman D. Histone structure and nucleosome stability. Expert review of proteomics. 2005;2(5):719-29.
5. Hegde A. Ubiquitin-proteasome system and plasticity. 2009.
6. Gupta R, Schrooders Y, Verheijen M, Roth A, Kleinjans J, Caiment F. FuSe: A tool to move RNA-Seq analyses from chromosomal/gene loci to functional grouping of mRNA transcripts. Bioinformatics. 2020;1:7.
7. Johnson NT, Dhroso A, Hughes KJ, Korkin D. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? RNA (New York, NY). 2018;24(9):1119-32.
8. Akter S, editor A Data Mining Approach for Biomarker Discovery Using Transcriptomics in Endometriosis. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018: IEEE.
9. Marsh JA, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson CV, et al. Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell. 2013;153(2):461-70.