Transcriptome and proteome analysis: a perspective on correlation

Opinion

Druggable proteins (such as enzymes, receptors, transporters and channel proteins, signal transduction proteins, oncogenes and transcription factors) are powerful drug targets for therapeutics of diverse diseases. Among the approximately 22,000 human proteins, novel drug targets are likely to emerge. Numerous bioinformatics tools are used to monitor the gene expression at mRNA level, including UniGene and Expressed Sequence Tag, EST, Serial Analysis of Gene Expression, SAGE, Digital Differential Display, DDD, Cancer Genome Anatomy Project X-Profiler and Digital Gene Expression Displayer. High throughput transcriptome analysis is often performed using Microarray, and Next-Generation RNA sequencing, NGSRNA-seq.

In recent years proteome analysis has been greatly aided by various protein expression tools, the Allan Brain Atlas, Multi Omics Profiling Expression Database, MOPED, the Human Protein Reference Database, HPRD, and the recently described Human Proteome Map and proteomics Db. The last two mentioned studies together account for approximately 84% of the total annotated protein-coding genes in humans. Together, these two studies identified >18,000 proteins encoded by both known genes and uncharacterized Open Reading Frames. Over 20,000 protein isoforms expressions were also characterized.

Numerous normal tissues, body fluids and cell lines were used in these two studies. These recently described protein expression databases greatly expand our ability to establish correlative evidence for mRNA and protein expression for most of the human proteome.

Based on mRNA expression using a high throughput transcriptome analysis, interpretations are often made about the functional relevance, pathways, interacting proteins and the possibility of drug therapy use. Protein expression is often inferred, but frequently a correlation of mRNA versus protein is missing in studies. The complex regulation of protein expression at the level of noncoding RNAs, DNA methylation, epigenetic changes, gene amplifications, copy number variations, mutations, post translational modifications such as acetylation, amidation, myristylation, phosphorylation, sumoylation, Ubiquitination etc., stability and degradation as well as the interacting proteins adds to the complexity of transcriptome-based interpretations.

For identifying candidate driver genes for therapeutic target discovery and functional elucidation, knowledge of mRNA and protein correlation is critical. The vast amount of data generated by microarray and the NGSRNA-sequencing technology can be reduced to a manageable number of putative targets using the correlation at the mRNA and the protein levels. In a recent study, Zhang et al., using The Cancer Genome Atlas (TCGA) tissues (n=97) demonstrated that the mRNA transcript abundance and gene amplifications did not often correlate. Among 3,764 genes analyzed, only 32% showed statistically significant correlation. Further, copy number alterations had no significant effects on the protein levels. Our own results as well as several other reports support such a lack of correlation.

During the early stages of microarray technology, vast data sets were generated with no standards. The establishment of Minimum Information About a Microarray Experiment, MIAME, greatly reduced the noise in subsequent array-based datasets. Similar guidelines for RNA and protein expression are needed to develop meaningful interpretations. Where possible, the investigators and authors should be expected to demonstrate such a correlation using the protein expression tools, which are becoming increasingly available. This would greatly help reduce the noise in the published literature, verify functional roles for the proteins and optimize the chances of identifying candidate driven genes for druggability.

Further, the development of single cell transcriptome and proteome analysis capability (in contrast to whole tissue-based analysis as is currently done) would allow for a single cell-based correlation between mRNA and proteins. This will allow for cellular heterogeneity in expression profiling in a tissue environment. An offshoot of the protein expression databases, such as creation of a separate human proteome database showing correlation with the transcriptome, would greatly aid in the therapeutic target discovery.

Acknowledgements

We thank Jeanine Narayanan for editorial assistance.

Conflict of interest

The author declares no conflict of interest.

References


