Abstract

Driven by the development of innovative approaches to quantify gene expression levels across large numbers of samples, differential transcriptome analysis is emerging as a powerful strategy to interrogate the complex interplay of genes accountable for malignancies. The CSD method is a correlation-based method to systematically classify differential genetic associations, facilitating identification of dissimilar interactions driving pathogenesis. In this work, we have used the CSD framework for analyzing gene correlation for thyroid carcinoma (THCA) patients. THCA is the most common endocrine cancer type. These tumours frequently resist standard treatments and are thus associated with poor clinical outcome. By using publicly available samples from The Cancer Genome Atlas, the transcriptomic landscape was investigated by contrasting these to normal thyroid expression profiles. The CSD method successfully pinpointed several interesting gene pairs in networks enriched for processes linked to carcinogenic behaviour. Examination of gene interactions revealed relevant gene groups driving aberrant signaling and regulatory cascades. Looking into well connected network regions identified hubs coordinating destructive information processing, likely responsible for deteriorated mechanisms needed to combat tumor progression. Probing gene associations characterized by transition into abnormal character resulted in potential novel prognostic markers of thyroid carcinoma.

In the second part, robustness and potential method improvements to the CSD framework were assessed. Quality control investigation demonstrated that obtaining consistent analysis results required proper data pre-processing, including batch effect correction. A fundamental step in correlation-based methods for differential studies, is quantifying gene-pair relationships from gene expression data. Here, we explored three alternatives to the conventional inference algorithm. First, weighted topological overlap (wTO) with soft thresholding was applied. This provided a robust computation, also giving meaningful results in the case of low sample sizes and appeared to produce biologically meaningful modular structures. The second method was based on computing the mutual information (MI) as a more far-reaching similarity measurement. Although it was more dependent on larger sample sizes, it elucidated numerous novel relevant gene pairs not captured by Spearman or wTO. Motivated by achieving a computational reduced footprint allowing applicability to larger data sets, the last alternative involved a simplified version of CSD omitting variance estimation. While maybe offering some false positives, the relaxed condition will produce useful result sets even for very large transcriptomic data. For quality assessment, gene interactions identified by any of the similarity measures, were analyzed with regard to biological function and significance. Alternative similarity measures augment the outcomes of the original CSD method, and yield candidate genes which may contribute to deciphering the pathogenesis of THCA.