A META-ANALYSIS OF THYROID CANCER GENE EXPRESSION PROFILING STUDIES IDENTIFIES IMPORTANT DIAGNOSTIC BIOMARKERS

Obi L. Griffith[1,2], Adrienne Melck[3], Steven J.M. Jones[1,2] and Sam M. Wiseman[3,4,5]
1.Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada; 2.Department of Medical Genetics, Faculty of Medicine, UBC, Vancouver, BC, Canada; 3.Department of Surgery, Faculty of Medicine, UBC, Vancouver, BC, Canada; 4.Genetic Pathology Evaluation Center, Prostate Research Center of Vancouver General Hospital & BC Cancer Agency, Vancouver, BC, Canada; 5.Department of Surgery, St. Paul's Hospital, Vancouver, BC, Canada.

Introduction: An estimated 4-7% of the population will develop a clinically significant thyroid nodule during their lifetime. In as much as one third of these cases pre-operative diagnoses by needle biopsy are inconclusive. In many cases, a patient will undergo a diagnostic surgery for what ultimately proves to be a benign lesion. Thus, there is a clear need for improved diagnostic tests to distinguish malignant from benign thyroid tumours. The recent development of high throughput molecular analytic techniques should allow the rapid evaluation of new diagnostic markers. However, researchers are faced with an overwhelming number of potential markers from numerous thyroid cancer profiling and classification studies. To address this challenge, we have carried out a systematic and comprehensive meta-analysis of thyroid cancer biomarkers from 21 published studies.

Methods: For each of the 21 studies, the following information was recorded wherever possible: Unique identifier (probe/tag/accession); gene name/description; gene symbol; comparison conditions; sample numbers for each condition; fold change; direction of change; and Pubmed ID. Clone accessions, probe ids or SAGE tags were mapped to a common gene identifier (Entrez gene) using the DAVID annotation tool, Affymetrix annotation files, and the DiscoverySpace SAGE tag mapping tool respectively. A heuristic ranking system was devised that considered the number of comparisons in agreement, total number of samples, average fold change and direction of change. Significance was assessed by random permutation tests. An analysis using gene lists produced from re-analyzed raw image files (ensuring standard methods) for a subset of the studies was performed to assess our method.

Results: In all overlap analysis groups considered except for one, we identified genes that were reported in multiple studies at a significant level ($p<0.05$). Considering the 'cancer versus non-cancer' group as an example, a total of 755 genes were reported from 21 comparisons and of these, 107 genes were reported more than once with a consistent fold-change direction. This result was highly significant ($p<0.0001$). In 10000 permutations, the simulated data never produced an overlap greater than three whereas real data identified 12 genes with overlap of four, five or six. Comparison to a subset analysis of microarrays re-analysed directly from raw image files found some differences but a highly significant concordance with our method (p-value = 6.47E-68).

Conclusions: A common criticism of molecular profiling studies by microarray (or other expression technologies) is a lack of agreement between studies. Comparison of any two studies often produces a disappointingly low level of overlap. However, looking at a larger number of published studies, we find that the same genes are repeatedly reported and with a consistent direction of change. These genes may represent real biologic participants that through repeated efforts have overcome the issues of noise and error typically associated with such expression experiments. In some cases these markers have already undergone extensive validation and become important thyroid cancer markers. But, other high-ranking genes (identified in as many as six independent studies) have not even been investigated at the protein level. A comparison of our meta-review method (using published gene lists) to a meta-analysis of a smaller subset of studies (for which raw data were available) showed a strong level of concordance. Thus, we believe our approach represents a useful alternative for identifying consistent gene expression markers when raw data is unavailable (as is generally the case). Furthermore, we believe that this meta-analysis, and the candidate genes we have identified, may facilitate the development of a clinically relevant diagnostic marker panel.