Significant prognostic gene and interconnection network enrichment (SPGINE) analysis stratifies breast cancers into three reproducible subclasses determined by novel genetic grading signatures

Vladimir A. Kuznetsov[1, 2], Luay Aswad[1, 2], Efthimios Motakis[1], Zhiqun Tang[1], Ghim Siong Ow, Anna V. Ivshina[1]

[1]*Bioinformatics Institute, 30 Biopolis street, Singapore 138671*
[2]*School of Computing Engineering, Nanyang Technological University, Singapore 639798*

The Cox-proportional hazard regression model could build the prognostic gene classifiers in network analysis of gene expression data. In this work, we tested the hypothesis that gene expression patterns selected via significant prognostic gene and interconnection network enrichment (SPGINE) analyses, integrating with survival data can select the functional cell cycle gene modules predicting genetically distinct cancer sub-classes, differentiating by tumor aggressiveness, chromosome instability and pluripotent cells gene expression patterns.

Meta-analysis of publicly available microarray profiles and clinical data from 1317 breast cancer (BC) patients from 7 cohorts were carried out in this study. Based on the Cox proportional hazard survival model, we developed a prediction method of post-surgery breast cancer patient risk stratification and SPGINE analysis, providing novel prognostic classifiers and perspective biomarkers selection.

This method links the gene expression profiles of subjects' primary tumors with their survival data. Firstly, our one-dimensional data-driven grouping approach was used for the selection of survival-significant genes, stratifying BC patients into two disease development risk groups. The method also used the pair enrichment test to select a sub-set of top-level synergistically significant genes according to the log-rank statistics and subsequently, hypergeometric test (at FDR<1%) specified based on the number of links of individual genes with the other survival significant gene partners.

SPGINE tested the H0-hypothesis that the number of gene partners for a survival significant gene occurred in the expression dataset at random. SPGINE automatically selected the 10-gene prognostic classifier, stratifying the patients of several studied cohorts into three reproducible prognostic subgroups (PSGs). Subsequently, by comparing global gene expression profiles between the PSGs, SPGINE selected a small subset of differentially expressed genes (13-gene DEG classifier) significantly discriminating the PSGs. The genes of our classifiers provided a highly-interconnected regulatory network, enriched with periodically changed mitosis/ cycle genes. Most genes of both signatures showed pro-oncogene survival patterns and were involved in mitosis, chromosome assembly/rearrangement, and stem cell-like self-renewal processes. Univariate and multivariate analyses showed prediction significance and reproducibility of the classification models across the cohorts. Importantly, the classification, provided by our prognostic classifiers were strongly correlated with histologic and genetic low- and high- genetic grading systems of breast cancers, reported in our previous studies (Ivshina et al., 2005; 2006; Kuznetsov et al. 2006; Aswad, et al, 2015), and predicted the risks of distant metastasis. Our SPGINE analysis provides reproducible cell-cycle genes related molecular classification of the breast cancer patients. These findings could help to better understand molecular basis of BC classifications and provide novel specific genetic determinants for personalized patient's prognosis and BC therapeutic intervention.