Genome wide survival prediction and network analysis stratifies breast cancers into three reproducible subclasses determined by novel genetic grading signatures

Vladimir A. Kuznetsov

Bioinformatics Institute, Singapore 138671
vladimirk@bii.a-star.edu.sg

Zhiqun Tang

Bioinformatics Institute, Singapore 138671
tangzq@bii.a-star.edu.sg

Luay Aswad

Bioinformatics Institute, Singapore 138671
luaya@bii.a-star.edu.sg

Efthimios Motakis

Bioinformatics Institute, Singapore 138671
motakise@bii.a-star.edu.sg

Ghim Siong Ow

Bioinformatics Institute, Singapore 138671
owgs@ bii.a-star.edu.sg

Anna V. Ivshina

Bioinformatics Institute, Singapore 138671
annavi@bii.a-star.edu.sg

In this work, we tested a hypothesis that unsupervised prognosis analysis of gene expression patterns selected via prognostic significant genes and gene network analyses integrating with survival data, trained across different cancer cohorts, can reveal the molecular classifier(s) of the phenotypically distinct cancer sub-classes. Meta-analysis of microarray profiles and clinical data of the publicly available 1317 breast cancer (BC) patients from 7 cohorts were carried out. Based on the Cox proportional survival model, network analysis, sampling and feature selection algorithms, we developed a novel computational prediction method of disease risk stratification and the survival significant (prognostic) gene connectivity providing the prognostic biomarker selection. This method, called the prognosis stratification analysis (PSA), links the gene expression profiles of subjects' primary tumours

with survival and histo-pathologic data. Firstly, the one-dimension data driven grouping (1D DDg) approach was used for selection of the survival significant genes, grouping the patients onto two disease development risk groups. The method also used pair enrichment test, selecting the sub-set of top-level synergistically significant genes according to the log-rank statistics and hypergeometric test specified at FDR<1% the number of links of the genes with the other survival significant partners. In statistical sense, it tested the H0-hypothesis that the number of the gene partners for a survival significant gene is occurred at random. PSA used to identify a 10-gene survival prognostic biomarkers classifier (10-genes classifier), stratifying the patients in each studied cohort into three reproducible prognostic subgroups (PSGs). Comparing global gene expression profiles in the PSGs, PSA selected a small subset of differentially expressed genes (13-genes classifier) significantly discriminating the PSGs. Most genes of both signatures showed pro-oncogene survival patters and involved in mitosis, chromosome assembly/rearrangement, and stem cell 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 low- and high- genetic grading of breast cancers, reported in our previous studies ((Ivshina *et al.* 2006; Kuznetsov *et al.* 2006), and predicted risks of distant metastasis. Our network analysis, reproducible prognostic-based stratification and their associated biomarkers could help to understand molecular basis of BC classification and provide novel specific genetic determinants for personalized patient's prognosis and BC therapeutic intervention.

This study was supported by funding from the Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), Singapore.