Theme-driven cancer survival analysis
Esteban Czwan¹,², Benedikt Brors², David Kipling¹
¹School of Medicine, Cardiff University, Heath Park, CF14 4XN Cardiff, United Kingdom
²Theoretical Bioinformatics, German Cancer Research Center, INF 280, 69120 Heidelberg, Germany

Background
Over the last few years there has been considerable success in defining gene expression signatures that have prognostic power. That is, groups of genes whose corresponding expression values are predictive of clinical outcome. These have obvious clinical utility from a prognostic perspective. However, closer inspection of the constituent genes in such signatures often reveals a heterogeneous selection of genes, with no obvious over-arching theme. This is disappointing, as it makes such signatures of little use in the development of new targets for therapy; the signature is often made up of genes that are the end-stage markers of an upstream (unknown) process, and do not readily lend themselves as direct targets for therapy. Instead, what would be far more use for a cancer biologist would be to know the identity of higher-level pathways or themes that are altered. A suitable druggable node can therefore be chosen from within that pathway [1]. With this in mind, we developed a “theme-driven” cancer survival methodology that screens hundreds of higher-level themes and tests associations with survival time [2].

Materials and methods
Two public gene expression datasets for breast and lung cancers were used as well as 1,400 genesets derived from Gene Ontology (GO) terms, Biocarta and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The core serum response (CSR) gene expression signature [3] was included for comparison. For each geneset, clustering was used to segregate samples into two groups, a survival test was performed to evaluate differences in survival time between both groups (Test 1), and a permutation test was done to assess the significance of the theme of the geneset (Test 2). A false discovery rate (FDR) was implemented to control for multiple hypothesis testing.

Results and conclusions
Our study presents two major findings on this “theme-driven” topic. First, an expansion of this approach that allows many hundreds of pre-defined genesets to be analysed to see if they have prognostic power in cancer survival datasets. We show this approach as applied to two well-studied public cancer datasets, and report novel, but biologically plausible, GO-based and KEGG-based genesets that are predictive of survival. However, perhaps the most crucial finding that comes out of our study is that current state-of-the-art statistical methods (such as those used by Chang et al.) are quite clearly not suited to such studies as they only perform Test 1 and disregard Test 2, which often leads to false positive correlations of the theme of a geneset and survival. Our approach provides a method to correct for this pitfall, and provides a novel route to identifying higher-level biological themes and pathways with prognostic power in clinical microarray datasets.

References