Geneset-based cancer survival analysis from gene expression data: On the non-uniform distribution of p-values under the null hypothesis

Esteban Czwan[1,2], Benedikt Brors[2], David Kipling[1]
[1]School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, United Kingdom
[2]Department of Theoretical Bioinformatics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

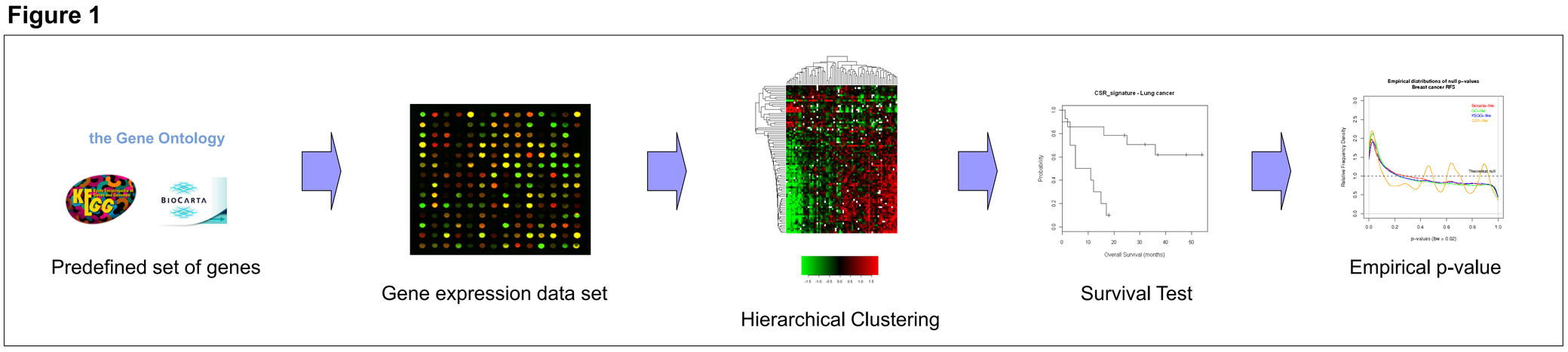
Introduction

In statistical tests from microarray gene expression data, the distribution of p-values under the null hypothesis is often not uniform. This can jeopardize conclusions as it renders traditional significance thresholds meaningless [1]. Yet usually researchers wrongly assume the uniform distribution of null p-values in gene expression experiments, including geneset-based studies in clinical oncology where the gene expression signature of all genes related to biological process is tested as a prognostic estimator. In these particular studies this false assumption may lead to the assertion that the biology of a gene set is associated with cancer prognosis when such a claim is actually impossible to confirm. To assess the (non-) uniformity of null p-values in this type of studies and its possible implications, an automated geneset-based method was developed.

Methods

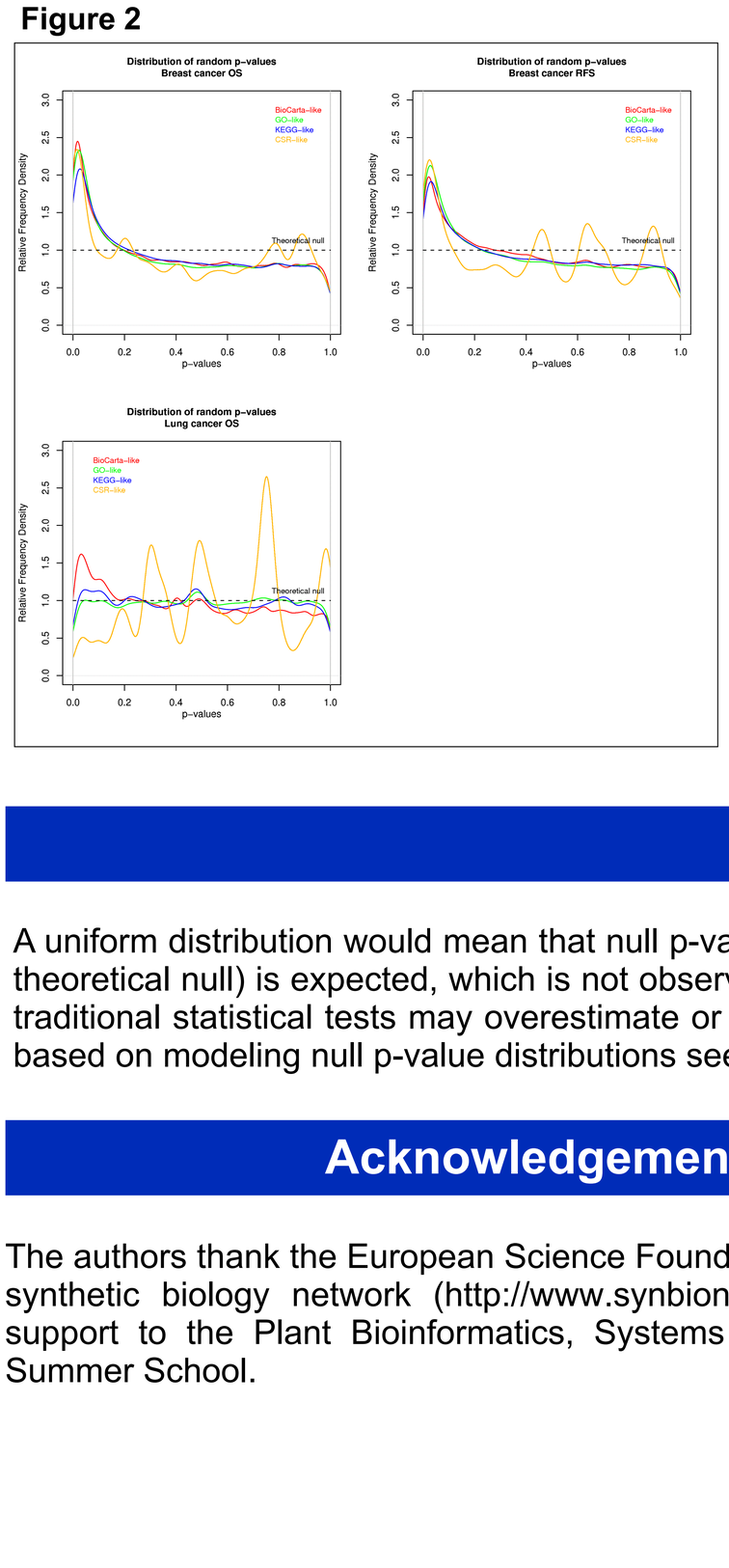
Two public gene expression data sets for breast [2] and lung [3] cancers were used as well as 850 gene sets derived from Gene Ontology (GO) terms, Biocarta and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The core serum response (CSR) gene expression signature [4] was included for comparison. For each gene set, hierarchical clustering was used to segregate samples into two groups, and log-rank test was performed to evaluate differences in survival between both groups (raw p-value p_1). A null distribution was empirically approximated for each type of gene set and for each survival estimate using a permutation approach based on 100,000 random sets of unrelated genes. Statistical significance was assessed by calculating an empirical p-value p_2 from the appropriate null distribution. Figure 1 describes the analysis flow.

Figure 1



Results

Figure 2



As shown in Figure 2, the four empirical null distributions (i.e. Biocarta-like, GO-like, KEGG-like, and CSR-like distributions) for breast overall survival (OS) and relapse free survival (RFS) as well as lung OS noticeably deviate from the expected uniform distribution. Table 1 shows a summary of results in terms of gene sets and their putative association with cancer prognosis. For breast OS and RFS, the numbers of significant gene sets drastically decrease from raw to empirical significance, implying overestimation of significance at the raw level.

The analysis of the CSR signature resulted in raw p-values comparable to those reported by Chang *et al.* (Table 2). However, our empirical results considerably differ from the raw results and, thus, from the original findings of Chang *et al*. After accounting for the non-uniform distribution of null p-values, CSR was found to be correlated to lung cancer OS but it was neither correlated to breast cancer OS nor to RFS.

Table 1: Summary of results

Survival estimate	# of gene sets ($p_1 < 0.01$)	# of gene sets ($p_2 < 0.01$)
Breast OS	50	3
Breast RFS	31	8
Lung OS	17	15

Table 2: CSR comparison

Survival estimate	Chang (p-value)	p_1	p_2
Breast OS	0.041	0.025	0.14362
Breast RFS	0.013	0.010	0.06125
Lung OS	0.0014	0.009	0.00249

Conclusion

A uniform distribution would mean that null p-values should be seen at the same frequency across the entire range ($0 < p < 1$), hence a flat line (i.e. theoretical null) is expected, which is not observed in our results (Figure 2). Consequently, since p-values under the null hypothesis are not uniform, traditional statistical tests may overestimate or underestimate significance. A method to account for this problem of geneset-based survival studies based on modeling null p-value distributions seems adequate and promising.

Acknowledgements

The authors thank the European Science Foundation and the SynBioNT synthetic biology network (http://www.synbiont.org/) for their overall support to the Plant Bioinformatics, Systems and Synthetic Biology Summer School.

References

[1] Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, **100**:9440-9445.

[2] Sørlie T, *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*, **98**:10869-10874.

[3] Garber ME, *et al.* (2001) Diversity of gene expression in adenocarcinoma of the*et al.* lung. *Proc Natl Acad Sci USA*, **98**:13784-13789.

[4] Chang HY, *et al.* (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol*, **2**:206-214.