Reviewer’s report

Title: Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data

Version: 1 Date: 3 July 2008

Reviewer: Carol Rosenberg

Reviewer’s report:

In this paper, the authors note two drawbacks to the current application of gene expression data to cancer assessment and treatment: 1) gene signatures reported from different studies have very little overlap (especially in breast cancer, which is the focus of this report); and 2) gene signature are often difficult to interpret in context of the underlying biology (a more general problem).

They attempt to address these two issues by identifying gene signatures of breast cancer prognosis in context of known, cancer-relevant biological pathways. Their rationale was that if they identify gene signatures within well-defined pathways, then the issue of data over-fitting will be eliminated, the mechanism based gene signatures will be more biologically comprehensible and better-targeted treatments could be tested.

The approach is, first, to identify five different breast cancer gene expression profiling datasets. Next, each datasets was analyzed for 20 cancer-relevant pathways that were assembled from two commercially available pathway-based programs (Ingenuity and SuperArray). For each dataset and each of the 20 pathways assembled, expression data of genes involved in a specific pathway were extracted; an unsupervised two-way hierarchical clustering analysis was performed, and if this clustering resulted in distinct patient groups, patient outcome was compared by Kaplan-Meier analysis. Finally, the authors investigated whether or not each pathway was prognostic in each dataset.

After demonstrating proof of concept (using the Amsterdam 70-gene signature and a “breast cancer gene set” (264 well-known breast cancer genes) to show distinct survival outcomes in all 5 datasets, the authors tested the ability of each of the 20 pathways’ gene signatures to predict outcome. One pathway’s gene signature - the cell cycle pathway – predicted clinical outcome in all 5 datasets. Then, the PAM method was used to build a predictive gene signature based on the cell cycle pathway (as well as 2 others – the 70-gene signature and the 264 breast cancer genes). This cell-cycle classifier performed well as prognostic biomarkers in the training dataset and in two independent validation datasets. The Amsterdam 70-gene signature was less accurate, especially in the independent validation datasets.

In a very interesting section of the MS, the authors used 232 randomly selected genes to build a classifier using a similar approach. The classifier based on the
randomly selected genes performed exceptionally well in the training dataset, but did not show predictive power in the independent datasets.

The authors state that this is the first study to integrate gene expression data and pathway information to develop pathway-specific gene expression signatures for cancer prognosis. Their data demonstrate that cell cycle pathway-associated gene expression is associated with differential outcome in breast cancer patients. They conclude that cell cycle regulation may be the “single most important factor contributing to breast cancer progression”. While several cell cycle-regulated genes have been used (or have attempted to be used) individually as breast cancer outcome markers, the multi-gene signature applied in this study may provide more accurate prediction and are mechanistically implicated in breast cancer progression.

Comments: This is a clear, well-written and interesting paper. The limitations of gene expression signatures that the paper addresses are real, and the question posed is well-defined. The approach is solid and reasonable, and the methods are straightforward.

1) Some additional information as to how the authors identified the genes in each of the 20 pathways would be helpful (since these genes could easily change as knowledge increases).

2) It would also be helpful to add a bit more information about each of the 5 datasets used in this study. Even though these are published and publicly available (to some degree), a table or a few sentences addressing this information may be useful to a reader not familiar with these studies.

The results presented appear sound.

3) One question about the results is whether the cell-cycle pathway’s significance could be due to chance, since they tested 20 pathways and were using p < 0.05 as their significance level. The classifier’s success makes this less likely, however.

4) Another question about the results that the authors should address is the degree of overlap between the cell-cycle pathway classifier and the other signatures (especially the Amsterdam 70-gene signature and the 264 breast cancer genes). For example, if a large proportion of the genes from either signature are “cell cycle genes”, then it is no wonder why the cell cycle pathway is prognostic. While the approach to evaluating gene expression data would remain novel and important, the specific impact of the cell cycle pathway may be lessened.

The manuscript adheres to the relevant standards for reporting and data deposition.

5) It would be helpful to list the GEO accession numbers or other information about how the authors accessed the datasets that they used.

The discussion and conclusion are generally well balanced and adequately supported by the data.
6) The statements that the cell cycle may be the “single most important factor contributing to breast cancer progression” and that this approach challenges the existing methodology seem a bit premature, however).

7) It would be interesting if the authors would speculate how they might use this classifier in the future.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.