Author's response to reviews

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

Authors:

Jiangang Liu (liu_jiangang@lilly.com)
Andrew Campen (andrew.m.campen@gmail.com)
Shuguang Huang (shu444@gmail.com)
Sheng-Bin Peng (sbpeng@lilly.com)
Xiang Ye (xiang.ye@lilly.com)
Mathew Palakal (mpalakal@cs.iupui.edu)
A. Keith Dunker (kedunker@iupui.edu)
Yuni Xia (yxia@cs.iupui.edu)
Shuyu Li (li_shuyu_dan@lilly.com)

Version: 2 Date: 22 July 2008

Author's response to reviews: see over
The BioMed Central Editorial Team

July 20, 2008

Dear Editors:

Thank you very much for your response to our manuscript entitled “Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data” (Manuscript ID: 2320656962026561) and for the reports from two reviewers that have provided very insightful comments and recommendations. The following are a point-by-point summary of how we have addressed reviewers’ comments and highlight of changes to the manuscript. The revised text in the manuscript is colored in red.

**Reviewer 1:**

**Comment:**

When it comes to the stated goal to derive biologic importance from the data, however, the results are quite limited. The manuscript does not even attempt to demonstrate that the set of genes selected for classification are more stable then in previous publications. Still, even if that was demonstrated, it would be necessary to show that this is not the result of just using smaller candidate pools of genes, making an overlap between gene sets more likely by chance. But even such a finding would still not allow to equate "biologic" and "numerical" importance of the selected genes. The authors should thus adjust the introduction/discussion to more fairly reflect the progress being made towards "interpretability" of gene sets.

**Response:**

We agree with reviewer’s assessment. We did not attempt to evaluate the stability of the cell cycle gene signature since we have recognized that it would be difficult to do so. As the reviewer stated, even we could demonstrate it by building cell cycle gene signatures using different training datasets and then comparing these signatures, one could argue the
overlap is simply due to a greatly smaller gene set to start with. We have also avoided to interpret the biological significance of the signature at the individual gene level based on their coefficients in the signature model because the numerical parameters are very sensitive to technical variations. We only argued that our results suggested the cell cycle pathway overall may contribute significantly to breast cancer progression hence clinical outcomes and this interpretation is consistent with what has been known from literature information. We have added a paragraph in the discussion section (page 16, line 22 – page 17, line 3) to capture the above discussion points.

**Comment:**
The strategy used in this paper contains two major parts, selection of pathways based on classification power and subsequently development of a prognostic marker. Neither part is particularly novel (see e.g. [1,2] as one of several examples for part one, the second part, classifier development, is the standard application per se in the field). It would thus strengthen the paper if it included a more detailed discussion of the advantages of combining the two steps. This could, for example be a more detailed demonstration that the classifiers developed in this way are more reliable and/or generalizable than those developed without the pre-selection step.

**Response:**
We appreciate that the reviewer pointed out the published work in the same direction and have added two references (reference 41, 42). We have also added a paragraph in the discussion section (page 16, line 7-20) to discuss the point and emphasize that by building gene signature models after pre-selecting pathways may lead to more reliable and generally applicable classifiers.

**Comment:**
On a technical note, it was not clear if the p-values presented in table 2 are adjusted for multiple comparison.

**Response:**
The p-values are unadjusted. We understand that testing 20 pathways and using p < 0.05 as the cutoff could result in a significant pathway by chance. We only selected the pathway (cell cycle) that showed a significant correlation with survival in all 5 datasets, which makes its significance less likely false positive. Furthermore, as reviewer 2 pointed out (please see comment 3 from reviewer 2 below), we confirmed its predictive power by building a classifier using supervised methods.

Reviewer 2:
Comment:
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).

Response:
We have added a sentence in the result section to describe that the genes in each pathways were assembled manually from literature information as of February 2007 (page 9, line 20-21)

Comment:
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.

Response:
We have added a column “Study summary” in Table 1 (page 26) to summarize the 5 studies and the datasets we analyzed.

Comment:
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.

Response:
We agree with the reviewer that testing 20 pathways and using p < 0.05 as the cutoff could result in a significant pathway by chance. In our study, we only selected the pathway (cell cycle) that showed a significant correlation with survival in all 5 datasets, which makes its significance less likely false positive. Furthermore, as the reviewer pointed out, we confirmed its predictive power by building a classifier using supervised methods.

Comment:
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.

Response:
We agree with reviewer’s reasoning. Indeed, further examination of gene identities in the cell cycle pathway, the Amsterdam 70-gene signature, and the control 264 breast cancer gene set has revealed that the breast cancer gene set included 25 cell cycle genes that may significantly contribute to its predictive power for breast cancer prognosis. We have added a paragraph to address this issue in the discussion section (page 14, line 25 – page 15, line 2).

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

Response:
We have added a column in Table 1 (page 26) to list the URL addresses where we downloaded the 5 datasets.

Comment:
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).

Response:
We agree with reviewer’s assessment and have modified the text. Please see the abstract section, page 2, line 28-30 and the discussion section, page 14, line 5.

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

Response:
We have added a paragraph in the conclusion section (page 18, line 20-24) to offer our opinion on future directions of our study.

We believe that all points raised by the reviewers have been properly addressed. We hope that the editors and the reviewers find this revised version suitable for publication on *BMC Medical Genomics*.

Please do not hesitate to contact me if there are further questions.

Sincerely,
Shuyu Li, Ph.D.
Lilly Research Laboratories,
Eli Lilly and Company,
Lilly Corporate Center, Indianapolis, IN 46285
Tel: (317) 651-9923 Fax: (317) 433-6963
Email: li_shuyu_dan@lilly.com