Author's response to reviews

Title: Expression-based Pathway Signature Analysis (EPSA): Mining publicly available microarray data for insight into human disease

Authors:

Jessica D Tenenbaum (jessie.tenenbaum@duke.edu)
Michael G Walker (mwalker@walkerbioscience.com)
Paul J Utz (piutz@stanford.edu)
Atul J Butte (abutte@stanford.edu)

Version: 2 Date: 21 August 2008

Author's response to reviews: see over
Aug 20, 2008

Melissa Norton, MD
Editor-in-Chief, *BMC Medical Genomics*

Dear Dr. Norton (and other editors):

This cover letter accompanies the revision of my manuscript entitled *Expression-based Pathway Signature Analysis (EPSA): Mining publicly available microarray data for insight into human disease*, manuscript number 7411472591952180. Please consider this manuscript for publication as a Technical advance in *BMC Medical Genomics*.

We appreciate the reviewers' comments and the chance to clarify sections of our paper. My responses to the comments are interspersed with the review as bulleted items in bold.

**Reviewer 1**

This analysis strategy is novel, and this problem is very important in the study of complex diseases.

Thank you for the positive comments.

**Major Compulsory Revisions**

1. The abstract should be structured into separate sections: Background, Results and Conclusions.
   - Corrected, with the addition of Methods per online instructions that Technical advances mirror research articles.

2a. The Methods section is not clearly described. The formula of the EPSA score should be given for better understanding.
   - *We thank reviewer for pointing out this key area of potential confusion. The EPSA score is simply the average Spearman rank coefficient. This point has been clarified in the text, rather than provide the commonly accepted formula for this metric.*

2b. How to choose the risk factor and how to perform survival analysis should be better described. Consequently, the part of C in Figure 1 should be enriched.
   - *We have added clarification text to figure and legend*

3. Are “pathway signatures” the significantly differential expression genes? In the first paragraph of the Results section, the authors should explain what “pathway signatures” refer to. The other similar problem is “positive control samples” in the “EPSA correctly identifies pathway mutations in cancer data” section.
   - *Both issues have been addressed in the text. The issue of pathway signatures in particular has been clarified in multiple places. Eg: “The term “pathway signature” refers to the specific pattern of differential gene expression that was observed upon experimental perturbation of a given pathway. It is represented by the log₂ of the fold change for the list of differentially expressed genes in the perturbed versus the non-activated state.”*
4. How many treated or untreated samples in training profiles? The authors should describe the number of samples clearly.
   - Now addressed in text- at least 3 replicates, specified when other than 3.
4a. Furthermore, why do you choose SAM software?
   - We chose SAM based on a number of criteria: has been demonstrated to perform well on microarray data and has been cited over 3000 times, is user-friendly, and is convenient with R plugins available.
5. The authors describe the mouse positive control in the legend of Figure 2. This information should be given in “Positive controls” section.
   - Addressed along with item 3 above
5a. Moreover, the “Compendia” section should be better described and integrated in the Results.
   - We have attempted to make the Results section clearer in this regard with additional explanation, and some additional references to Methods.

Minor Essential Revisions
1. The authors should check the references to make sure that they are cited correctly. For example, SAM firstly appears in the first paragraph of Results but it is not marked.
   - Corrected
2. In the last paragraph of page 5, “n=135” should be “n=135”.
   - I think the issue was italics of n? corrected.
   In the legend of Figure 5, all “n” should be Italic.
   - Corrected
In the last paragraph of page 5, “P-values” should be “p-values”. In the third paragraph of page 9, “p-values” should be “p-values”.
   - Corrected- used italics
3. In the legend of Figure 1, “log2(treated/untreated)” should be “log2(treated/untreated)”.
   - Corrected- made the 2 a subscript
4. It is better to keep consistent decimal for all the p values and q values if it is possible.
   - Changed to 2 decimal places where possible without losing information

Reviewer 2
This is an extremely useful endeavour. The method is conceptually simple but elegant…Overall, it was very well conceived and written study.
We thank the reviewer for the positive comments.

Discretionary Revisions
1) The idea and approach are reminiscent of some of the work by Eran Segal (e.g., Segal et al. 2004. Nat Genet. 36(10):1090-8.). The authors might consider reading and commenting on those papers if they are not already aware of them.
   - We have most definitely come across this paper and many others by Segal et al., and have great respect for their undertaking. We have omitted discussion here because their analysis begins with “collecting 2,849 biologically meaningful gene sets”, which puts this approach in the category of knowledge-based methods- mentioned in our background section, but not discussed beyond this.

2) I am surprised that the authors use such a simple and overly stringent multiple testing correction method (Bonferroni) in some places and then relatively sophisticated statistics (FDR correction, survival analysis, cox proportional hazards, etc) elsewhere. A less conservative method such as Benjamini and Hochberg is nearly as simple to apply (try R multtest package) and might give
additional significant results. However, the authors may have had reason to be conservative. If so, this could be explained/justified.

- Essentially, where we had only a few multiple hypotheses, we used the stringent but straightforward Bonferroni approach. We applied the more sophisticated approaches where we had thousands of hypotheses such that Bonferroni had no hope (i.e. looking across thousands of genes) or where these methods are commonly used (survival).

3) A thought occurred that it would be interesting to compare the level of correlation between a pathway signature and a tumor/normal matched array dataset. Instead of seeing which signature is highly correlated or anti-correlated with a tumor population you could find signatures that are differentially correlated between normal and tumor states. This might identify important pathways that are activated or deactivated in the process of tumor progression. However, I am aware of the difficulty in finding such matched tumor/normal datasets. Perhaps the authors could comment on this as an area for future development.

- If I am understanding this point correctly, what the reviewer suggests is actually what we did do. Had we not used a tumor/normal match, it would have been impossible to compare across cell types. Finding appropriate matches was indeed a challenge in some cases.

**Minor Essential Revisions**

1) I would like to see a brief (1-2 sentences) explanation of what is meant by "perturbations to cells" earlier on in the manuscript. The specifics are detailed in the methods. But, a general explanation in the background would help the reader understand the method better. What kinds of perturbations? Why would we expect these to highlight pathways of relevance to human disease?

- Addressed in text in second paragraph of background: “perturbations of cells, such as stimulation with a small molecule, or alterations due to a transgenic mutation. If similarities can be detected between the changes in gene expression in diseased versus normal tissues, and those changes observed by perturbing known pathways, these observations may provide new information regarding pathways that are potentially affected in disease.”

2) Which SAM software package was used? Please provide details.

- samr package in R, version 1.20- specified in text

3) The authors state that they "...pinpoint the genes at play in the specific pathway of interest; without focusing on this subset of genes, other factors such as cell type or cell cycle stage could dominate any correlative signal detected." Can they explain/justify this further. How does focusing on specific pathway genes eliminate these other factors?

- We believe this point should be clearer given the clarification above, that we did in fact use ratios for the signatures (stimulated:unstimulated) AND for the disease tissue (tumor:normal).

4) The section titled 'EPSA correctly identifies pathway mutations in cancer data' and the corresponding figure (Figure 2A/2B) need additional clarification. How is E2F3 related to Rb null? This is not obvious to the uninitiated. Not being familiar with Bild et al makes these results hard to follow.

- Corrected- added “in the case of Rb null, knockouts of an inhibitor of E2F3.” Lees et al. 1993

5) I have difficulty with the conclusions drawn in the final sentence of the section titled 'EPSA predicts differential prognosis in cancer patients based on activation level of small-molecule pathways'.

Have the authors really shown that too much disruption (by the drug?) appears to
tip the balance towards poorer prognosis? They found that overall, a lower
degree of correlation was observed between the drug pathway (i.e., the pathway
induced by the drug in a separate study) and ovarian cancer patient profiles than
expected by chance. Then, stratifying by degree of correlation they found that
patients showing the least correlation with the drug signature had worse
outcomes and vice versa. So perhaps this pathway is dysregulated in ovarian
cancer. And, perhaps treatment with the drug would induce the beneficial drug
signature and improve the prognosis. However, since the ovarian cancer patients
have not been treated with the drug it is difficult to predict what the result will be
when they are. I suggest rewording this sentence to be more clear and cautious
in the interpretation. This result was one of the most interesting so it’s important
that you get it across clearly.

- We have modified the explanation in the text according to the reviewers suggestions,
  and thank him/her for the recommendation on how to make this important point
  more clear.

6) In the generation of Kaplan-Meier curves, patients are stratified into 'two extreme tertiles' for
comparison. This is a valid approach. However, I can not find where it is explained how these
tertiles are defined and how many patients (compared to the total number) are represented in each.
This is important for the reader to understand how generally applicable the associations are.

- Clarified each is 1/3 of patients.

7) I find the methods to be missing some minor details that would be necessary for replicating the
study. How were the microarray datasets processed (normalization, probe summarization, etc).

- Added to methods: Unless otherwise specified… data was processed as described by
  the original authors for each respective dataset.

8) Also, how was EPSA implemented? Was it a set of Perl/R scripts or a standalone executable?
Is there an implementation or R code available?

- R scripts, now specified in Methods

Reviewer 3

Discretionary Revisions

1. The clarifications made in the methods section (as well as in the results) helped me understand
what the random permutation approach is accomplishing and why the random correlations are
non-zero. However, I still feel the discussion regarding why choosing n genes at random is
different from
scrambling gene labels is confusing. There is a scheme where scrambling gene labels is equivalent
to the procedure of choosing n genes at random – specifically if you permuted the labels in the
same way for all of the samples. I understand now that you mean to draw a difference between
permuting only within the ligand and selecting a new set of n random labels, but it may be worth
further clarifying this point.

- We thank the reviewer for previous feedback that allowed us to make this point more
  clear. In the interest of simplicity, we rely on the opening statement above- that in its
  current form, the reviewer is able to understand what the permutation approach is
  accomplishing, and why the random correlations are non-zero.

2. Page 10, 1st paragraph – “Figure Figure 5B, C, and D”

- Corrected
Reviewer 4:

Major:
As I perceived the work presented here, the significance of it is not in the clinical implications of the correlations presented (since many had been reported in the publications associated with the analyzed data sets), but in the methodology used to detect correlations between different signatures. If this is true, then it would be necessary to evaluate the performance of the Spearman rank order correlation coefficient with respect to the performance of other measures employed in compendium analysis.

- While we agree that the significance of this work is not in the specific results presented, we respectfully disagree with the conclusion the reviewer has drawn. Rather, we believe that the significance of this work is not in the detection of correlations between the signatures but rather in 1. the creative re-use of seemingly unrelated data sets, 2. the simplicity of the approach while maintaining statistical rigor, and 3. the statistically significant results and hypothesis generation derived from comparisons across differing platforms and species. Evaluation of alternative methods could certainly be a direction of interest for future work.

Minor:
1. In Fig. 2A, panel A from Fig. 2 of the publication by Bild et al. is reproduced. Is the reproduction authorized by Bild et al.?
   - Yes- permissions were obtained free of charge from the Nature Publishing Group through Rightslink.
2. The abbreviations table is not complete, e.g. AIG, AfCS, RBC, GEO, HMEC.
   - Corrected
3. In the legend to Fig.3, it is not clear whether extreme tertiles are analyzed as in Fig.4.
   - Corrected- “Note that the curves shown for EPSA use only the extreme tertiles of subjects, but Cox proportional-hazards analysis was performed across all subjects in order to be directly comparable with the results in Bild et al.”
4. In supplementary Fig.1, most random signatures have positive R values while for random signatures one would expect roughly equal proportions with positive and negative R-values. This should be commented on.
   - As discussed in the context of our method of randomization (and in addressing Reviewer #3 above)- we would not expect there to be a roughly equal number of positive and negative R values- with biological relationships retained, we expect primarily positive correlations by random chance as these represent pathways necessary for survival, metabolism, etc.

Thank you again for considering this manuscript. We do hope for a favorable response.

Sincerely,

Jessica Tenenbaum, PhD
Project Leader
Duke Translational Medicine Institute
PO Box 17969
Durham, NC 27615
jessie.tenenbaum@duke.edu
+1 919-668-8811