Author's response to reviews

**Title:** Modeling gene-by-environment interaction in comorbid depression with alcohol use disorders via an integrated bioinformatics approach

**Authors:**

Richard C McEachin (mceachin@umich.edu)
Benjamin J Keller (kellerb@emich.edu)
Erika FH Saunders (esaunder@umich.edu)
Melvin G McInnis (mmcinnis@umich.edu)

**Version:** 2  **Date:** 15 May 2008

**Author's response to reviews:** see over
Dear Editor,

Please find our revised manuscript "Modeling Gene-By-Environment Interaction in Comorbid Depression with Alcohol Use Disorders via an Integrated Bioinformatics Approach". We greatly appreciate the input of Drs. Ritchie and Reif, and have revised the manuscript accordingly. Please see our point-by-point responses:

Reviewer: Marylyn Ritchie

Major Compulsory Revisions

1) On page 6, where the 263 citations resulted from the PubMed query are discussed and are consistent with a statistically significant association between depression and AUD, did the authors manually curate or validate these 263 manuscripts to ensure that the context in which they share depression and AUD are consistent with the association you are looking for? Can you be certain that some of the papers are not reporting negative associations between the two disorders? Similarly can you be sure that the papers are looking at the association of these two disorders and it is not simply the report of different aspects of some other studies? It was not clear if these manuscripts were followed up after the query and this information is essential to comprehend what the query results really suggest.

Revisions: At Dr. Ritchie's suggestion, we manually validated these 263 manuscripts using the resources available in the University of Michigan's digital holdings. We included a discussion of negative associations, as well as co-occurrence of queried terms in earlier publications, versus comorbidity as studied in later publications.

We also used this as an opportunity to calculate Positive Predictive Value (PPV) for the citations returned from this query and included these results and interpretation in the text.

2) On page 8, the authors need to define their use of the word “interaction”. They discuss that the combinations of over-represented keywords provides evidence of interaction across a locus pair. Is this meant to be a statistical interactions, biological interaction, or simply co-occurrence? The word interaction has different meanings to different readers so it is important to define.

Revisions: We added our definition of "interaction" to the text, informed by Hartman, et al., "Principles for the buffering of genetic variation." Science 2001, 291:1001-1004. Notably, we explain that this definition explicitly requires statistically significant over-representation of biomedical keywords at a locus pair defined by a genotype/phenotype analysis, but it is not limited to biological interaction (e.g., epistasis, protein-protein
binding). This definition is particularly useful in PDG-ACE analyses because it allows us to consider a broad range of hypotheses on the genetic etiology of complex disease (from well-defined to completely novel) while only considering the strongest evidence.

3) - On page 10, the authors indicate that the 21 keywords provide statistically significant evidence of interaction between TNF and MTHFR. Did the authors test this approach on data where we know there is a true interaction? I think that this needs to be explained further. I am not sure that the co-occurrence of these keywords (even with permutation testing to demonstrate they occur more often than you would expect by chance) indicate “interaction”. Perhaps the word “co-occurrence” would be more appropriate. Again, this goes to my other point about the word interaction. For me, interaction implies a statistical interaction. I do not think that the co-occurrence of keywords allows one to infer a statistical interaction. Perhaps if these results were curated and a meta-analysis was performed, then one could make assessments of evidence for statistical interaction.

Revisions: We added text to briefly explain the testing of PDG-ACE: positive control locus pairs ("true" interactions) and negative control locus pairs (randomly selected locus pairs). The discussion of our definition of "interaction" is above.

4) - Were the results on page 12 shown in Table 2 validated or curated?
- I don’t know much about the genes described here. Is the result of this manuscript a novel discovery and a new hypothesis describing how this pathway functions to lead to susceptibility for depression and AUD? Or is this a demonstration of a novel bioinformatics approach where it is shown to discover a pathway that we know a lot about? When I was reading page 13, I started thinking this and was not sure. It would be nice to show that this approach works on something we know and then discover new things once we know the approach works. Perhaps this has already been done.

Revisions: The discussion on validation is in 1), above. We added text in the conclusions to explain that the novel parts of this analysis are a) the genetic etiology of comorbid AUD with depression, consistent with clinical experience, and b) the data mining approach.

Minor Essential Revisions

5) - In the methods section, “Queries against NCBI Databases” section, the authors comment that Entrez gene surveys of the whole genome are complementary to whole genome association. It would be nice to have another sentence or two following this one to explain what is meant by this statement before describing the advantages and disadvantages.

Revisions: We added text to explain how this approach complements WGA.
6) - Also in this section, it is not clear what you are querying Entrez gene on. This is described in more detail later in the manuscript, but something should be said in this paragraph to provide examples for how these queries are done.

Revisions: We added the Entrez Gene query in the text.

7) - On page 9, were no other genes returned from the Entrez gene query other than the three described? It was not clear if these were the best ones or the only ones.

Revisions: We made it explicit that only three genes met the initial query criteria.

8) - Table 3 is discussed in the manuscript after Table 4. Perhaps these table numbers should be switched.

Revisions: Table 3 is first mentioned, briefly, on page 9 while Table 4 is first mentioned on page 11.

Reviewer: David M. Reif

- Minor Essential Revisions

1. Figure 1 (Analysis Flow) does not seem to match the analysis progression described in the text. In particular, the middle layer in Figure 1 presents three hypotheses in parallel that do not interact directly with each other, whereas in the text, there was an ordered progression of these steps. The figure should be refined to show information flow as written in ‘Methods’ and ‘Results’.

Revisions: We revised Figure 1 accordingly.

2. The initials given preceding the email addresses on the title page for Erika FH Saunders are out of order.

Revisions: We corrected Dr. Saunders' initials.

3. In the first paragraph of ‘Methods’, the authors should mathematically define positive predictive value (PPV), since it is used frequently in the paper. I assume their definition is: $PPV = TP/(TP+FP) = TruePositives/(TruePositives+FalsePositives)$

Revisions: We added this definition of PPV in the text.

4. The last paragraph of ‘Conclusions’ should be extended to discuss how the increased volume of data in public repositories might affect this analysis approach and what issues exist with current data. For example: How might publication bias toward positive results affect this approach? Do only commonly-
studied genes accumulate enough evidence to be relevant? What additional data would have the greatest effect on the authors’ approach?

Revisions: We added text about publications bias in the discussion about queries against NCBI databases. We added text about publications bias, commonly studied genes, and the impact of additional data, especially WGA and microarray data, in the conclusions.

5. Since the title includes “gene-by-environment interaction”, more introductory and/or discussion text should be devoted the limitations of this integrated bioinformatics approach in characterizing the GxE interaction with respect to dose, exposure, heritability, and other common issues in those fields. It is beyond the scope of this paper to add a comprehensive review of these issues, but some discussion is necessary to put these results in context.

Revisions: We added text about dose, exposure, and heritability in the discussion.

- Discretionary Revisions

1. In the introduction, the authors should address the feasibility of using a natural language processing (NLP) approach in obtaining indications of association in PubMed queries. It may not be appropriate for this application, but the authors should acknowledge NLP as an alternative approach or perhaps as a future direction.

Revisions: We added a brief discussion of how NLP could be used in future analyses.

2. The authors could discuss the impact of analysis order on subsequent steps in their integrated informatics approach. For example, justification of why APOE was not included in formal analysis is provided; however, what if APOE was included in the PDG-ACE step rather than only in subsequent PubMed and GeneGo steps?

Revisions: We restarted the analysis, including APOE in every step, and added text to describe the results.

3. The authors should discuss availability of PDG-ACE software or if there are plans to distribute the tool.

Revisions: We added text to explain that PDG-ACE is available and added Dr. Keller's email address for contact.

Best regards,

Richard C. McEachin, PhD,
Senior Laboratory Research Specialist,
University of Michigan, Dept. of Psychiatry