Reviewer's report

Title: Modeling complex genetic and environmental influences on comorbid Bipolar Disorder with Tobacco Use Disorder

Version: 2 Date: 13 November 2009

Reviewer: Marian Hamshere

Reviewer's report:

'Modeling complex genetic and environmental influences on comorbid Bipolar Disorder with Tobacco Use Disorder'

Richard C McEachin, Nancy L Saccone, Scott F Saccone, Yelena D Kleyman-Smith, Tiara Kar, Rajesh K Kare, Alex S Ade, Maureen A Sartor, James C Cavalcoli and Melvin G McInnis

The authors present a well written manuscript describing the application of a number of bioinformatics scripts employed to investigate potential genes linked to Bipolar Disorder (BD) and Tobacco Use Disorder (TUD). The methods and results are interesting.

Major Compulsory Revisions

- I’m not very familiar with TUD, but I wonder how certain are you that your data for smoking is actually TUD? Can you distinguish TUD from smoking?

- In the meta analysis, do you consider including smoking as a fixed effect to be appropriate? Is the only difference between studies the power to detect the outcome of interest? Would smoking be better treated as a random effect? If so, there is a nice DerSimonian-Laird random effects function in R (meta.DSL) which may be of use to you.

- I realise how difficult it is to write up this work, however, the manuscript is in danger of looking like a list of results output from multiple different bioinformatics tools. I am aware that there are an ever-expanding number of bioinformatics tools available for use online. For readers who are not familiar with every bioinformatics tool that you use, it would help if you could comment more on how these packages performed on your data, and whether you would recommend them to someone else. For example,
  - How do you interpret that given similar input data, MiMI, STRING and MetaCore identified networks of 41, 7 and 69 genes, respectively?
  - Why do you believe the MetaCore results (which support your hypothesis), as opposed to MiMI and STRING (which both agree, but don’t support your hypothesis)?
  - Are these methods suitable for all disease data sets? E.g. are they suitable when there are no candidate genes identified with a strong reliability?
  - Is it possible to perform the GeneGo analysis on biological pathways, as
opposed to disease terms? This would be an interesting addition to the manuscript.

- Have you acted on your bioinformatics results and performed any candidate gene analyses on your selected networks?

**Minor Essential Revisions**

- Please can you explain why the weight in the meta analysis for the Grant study is so much greater than the Carney study? What weighting was used?
- In the meta analysis summary table, you included some details as to where some of the control samples originated, do you have these data on the cases as well?
- Supp Table 5 would benefit from a detailed description of the parameters, e.g. Q, H and I^2.
- When selecting your gene lists with Gene2MeSH, do genes have an equal chance of co-occurring in a list for BD and TUD? Could there be a bias towards the disease that has undergone a greater amount of research, or the disease with more genes involved? Selecting genes that are required to have been studied for three TUD Gene2MeSH keywords, does this require a certain level of replication, more so than when selecting genes for BD? May be you could discuss this in your manuscript?
- When you read the papers to identify whether a gene was associated with disease, did you consider the associated p-value of a single SNP? Was this p-value adjusted for multiple testing in the original publication? Did you consider the power of the study?
- How did TPH1 and DRD4 fail to meet the requirement for further analysis? Were they studied sufficiently and not significant, or not studied enough?
- What was the maximum path length output by MetaCore?
- For the MetaCore analysis, was nicotine added to the network after the set of genes had been selected, or before?
- What is the “modified” Fishers Exact test?
- In your Common Elements analysis you have a positive control analysis. Do you have any negative controls analyses?
- “GAD is under continuous development”… how reliable is it now?
- It would be good to include software version information if available. If the online software you use are updated regularly, the date of use should be provided. My apologies if I’ve missed this in your manuscript.
- Please can you discuss further the observation that the genes over-represented in GAD, differ to those from ConceptGene? All that has been said so far is that it might be a false positive.
- It would be nice to have some more discussion in the Common Elements vs Network Analysis section. What do you mean that the “results are consistent and expand upon the PDG-ACE and GRAIL analyses”? 
In table 1, all three p-values are the same to 9 decimal places, is this correct?

This review has been complied following a useful discussion with Alex Richards (Cardiff University).

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.