Author's response to reviews

Title: Integrative network analysis identifies key genes and pathways in the progression of Hepatitis C Virus induced Hepatocellular Carcinoma

Authors:

Siyuan Zheng (siyuan.zheng@vanderbilt.edu)
William P Tansey (william.p.tansey@vanderbilt.edu)
Scott W Hiebert (scott.hiebert@vanderbilt.edu)
Zhongming Zhao (zhongming.zhao@vanderbilt.edu)

Version: 2 Date: 29 July 2011

Author's response to reviews: see over
We very much appreciate the valuable comments of the reviewer and the Editor. We have taken each of the comments into account when revising the manuscript and/or have addressed the comments below. The revised text was highlighted in red.

Response to Reviewer

Minor Essential Revision:

1) It is surprising that there was no attempt to compare interaction modules identified in this report with available datasets of HCV-host protein interactions. HCV proteins have been reported to bind numerous cellular proteins. It would be interesting to determine if some of these cellular proteins that are directly targeted by HCV components also belong to the deregulated interaction modules described in this manuscript.

We thank the reviewer for this insightful comment. To address this, we downloaded the HCV and host protein interaction data from a manually curated database and compared with our data. Our result shows that the hub proteins (LCK, CDC2, for example) in our identified networks are often binding targets of HCV proteins. In the revised manuscript, we added a new table (Table 2) to provide this result, a short paragraph (page 5, paragraph 2) for data description in the Method section, a full paragraph (page 10, first paragraph) in the result section, a sentence in Abstract section, and also some discussion (page 15, paragraph 2). We felt this new data further support the quality of our unique network-based analysis.

2) In the “methods” section, authors “ranked the genes by their weights and selected the top 500 genes”. Thus, a list of 500 deregulated genes was extracted from each node-weighted protein interaction network corresponding to each stage of the disease. I understand that 4 lists of 500 genes were established (normal-to-cirrhosis; cirrhosis-to-dysplasia; dysplasia-to-early HCC; early-to-advanced HCC). Is that correct? This should be clarified in the text.

We thank the reviewer for pointing out this issue. Yes, we have 500 genes for each stage pair. In the revised manuscript, page 5, last paragraph, we added this information.

3) From each list of 500 seed genes, an interaction module was extracted using the algorithm described in Jia P. & al. Identified interaction modules are composed of only 40-60 genes (Table I). Thus, a list of 500 genes is shrunk to a highly-significant subnetwork that is about ten times smaller. But the algorithm developed by Jia P. & al is used to expand a network from seed genes. Even after reading the manuscript by Jia P. & al., it is unclear to me how their algorithm can reduce instead of increasing the size of the network. This should be better explained.

We thank the reviewer for this point. In general, the subnetwork expands when adding the most significant neighbor gene(s), one in each step, until $Z_{n+1} \leq Z_n \times (1 + r)$,
where $r^*$ is the increasing ratio of module score in each step. This allows us to expand the subnetwork by recruiting only nodes whose genetic signal (e.g. differential expression) is meaningful. In practice, not so many neighboring genes of each seed gene could increase the score by $r^*$. Since we started from each seed gene, our resulting networks are actually small but enriched with genetic signal.

Further, in the last step of our network identification, we refined our result network by removing network components with less than 5 nodes. Another major difference is that the input list of 500 genes was generated from the microarray gene expression data and may not be all present in the reference network. In this case, we used the genes overlapped with the reference network. Finally, dmGWAS network analysis (Jia P et al.) was specifically for genome-wide association studies (GWAS) datasets, while here we specifically applied an integrative network analysis based on multiple microarray gene expression datasets for disease progression network analysis. In the revised manuscript, we made some revision in text to clarify this point (page 6).

**Discretionary Revisions:**

*Table I. Horizontal lines should be added to clearly identify hub genes associated to each network.*

We have added horizontal lines in Table 1 and applied the same rule in Table 2.

*Figure 4 and supplementary figure 2 are almost identical (Figure 4 is showing only major core components). Supplementary figure 2 should be presented as Figure 1, and Figure 4 should be eliminated.*

We thank the reviewer for this point. After we carefully read the manuscript again, we would like to keep both figures since Figure 4 is to summarize the giant component of each stage specific network, which might be more informative to the readers. To respond the reviewer’s concern, we moved Figure 4 to Figure 1 to highlight the major component of each stage specific network. We still kept the detailed networks in Supplementary data for reader’s interest and further investigation. If the reviewer still likes us to move Suppl Figure 2 to Figure 1 and eliminate the original Figure 4, we are flexible to do it by another revision.