OPEN PEER REVIEW

Open Peer Review (OPR)

As a new feature for progressing towards transparency, we decided to open a new window for all of our editors as well as authors titled "**Open Peer Review**". We hope by this new facility, our reviewers will be more motivated and authors will be more satisfied with the review process. We believe that publishing our peer review reports could make a transparent and clear environment for all our efforts within a journal, but not all reviewers tend to publish their comments.

What is "Open Peer Review" process?

An "Open Peer Review" process is making the details of all review process (including reviewers, associate editors, and EICs comments) as "Public" as it is agreed by EIC, Authors, and reviewers.

Advantages of "Open Peer Review" process

- More transparency, constructiveness and tactful comments of the peer review process: leads to an increase in the quality of reviews
- More motivations for all involved roles in the review process
- Authors' satisfactions from the review process: Increases honesty between authors and reviewers
- Education of both authors and new students
- Prevents reviewers from following their individual agendas and leading to the detection of reviewers' conflicts of interests

You can find out more at:
https://brieflands.com/briefland/knowledgebase/category/tree.html#opr.html

Hepatitis Monthly
Article DOI: 10.5812/hepatmon-122362
Published in: Hepatitis Monthly: 22(1); e122362

Peer Review Report for "Identification of Shared Gene Signatures in Different Stages of Nonalcoholic Fatty Liver Disease Using Integrated Microarray Datasets"

Author(s): B. Fatemeh Nobakht M. Gh, Yazdan Hasani Nourian, Masoud Arabfard

Review Timeline:

Submit Date: 9 Jan 2022
Revised Date: 22 May 2022
Accepted Date: 8 Jun 2022

Revision (0)

Here, you can see the **Reviewers**, **Associate Editors** and **EICs'** comments from the beginning to the end of the revision process.

REFEREE: Reviewer | **Revision** (0)

20 Feb 2022

This is an interesting study that has been obtained from the collection of appropriate data and advanced analysis methods. The article is generally well written and structured. However, the article has shortcomings that need further explanation of data processing and analysis. I have a few suggestions that the authors might consider, but none of them would prevent moving forward.

The data processing results, which include their normalization and clustering, should be presented in the supplementary file, and the method for removing data noise should be explained. On the other hand, the reason why datasets have been reviewed separately and not merged together should also be expressed.

Provide a table of DEGs in the article or the supplementary file.

It is better to classify gene modules in biological processes (BP), cellular component (CC), and molecular function (MF).

Drawing effective subnetworks using suitable packages in the Cytoscape should be considered, and the list of genes in these modules should be considered for pathway

enrichment analysis.
Although the authors have mentioned that this is a computational article and lacks an experimental section, in any case, validation of some highly differentially expressed genes and hub genes is recommended by qRT-PCR. Given these cases, the article needs revision.

REFEREE: Associate Editor | **Revision** (0) 14 Apr 2022

In this manuscript Authors, analyze a gene expression data set obtained from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) accession # GSE48452, # GSE63067 and # GSE89632. In these data set a group from the several geographical regions performed gene expression profiling of liver tissue from normal individuals, NAFLD and NASH using the Affymetrix Human Genome U133 Plus 2.0 Array. In fact, Global gene expression patterns patients with human steatosis and several NASH together with their respective control patterns were analyzed to define the non-alcoholic fatty liver disease (NAFLD) progression molecular characteristics and to define NASH early markers from steatosis. The Authors used the published gene expression data and performed a series of bioinformatics analysis and claim that the majority of differentially regulated genes and several hub genes in steatosis patients are related to the inflammation and oxidative stress pathways. Although the identification of gene expression signatures able to predict disease conditions such NAFLD and NASH patients is of great interest, the study presents with many flaws. Major criticisms and concerns are as follows:

1- GEO contains several datasets for gene expression data for NAFLD and NASH. What's the criteria for selection of 3 datasets (GSE48452, GSE63067, GSE89632)?

2- Genes should be clustered using a k-means clustering algorithm or using principal component analysis.

3- I suggest you extensively read the documentation of Heatmap at Bioconductor.
4- Weighted gene co-expression network analysis (WGCNA) should be used to cluster the specific up-regulated genes in each subgroup. WGCNA has been proven to be an effective method for detecting multiple co-expression modules and can be used to find clusters (modules) of highly related genes. The optimal Power value should be reported through the Power value scatter plot, and the distance between genes should be calculated. In addition, the average method and the dynamic method should be used for hierarchical clustering analysis. The clustering diagram and the module classification of genes should be established, and similar modules should be merged.

4- There are several Datasets when you perform a search with NAFLD. Some studies analysed the aforementioned GSE datasets. Could some of the findings presented in this manuscript be validated in some other cohort of patients? Also, for obtaining reliable and comprehensive results, RNA-seq data should be analysed.

5- Please note that the hub genes extracted from topological analysis based on important centrality measures. I strongly recommended that "hub genes" should be reported.

Topological network analysis has performed for PPI networks and the most important hub genes can be identified. The network analyzer plugin of Cytoscape can used to analyze topological properties of the networks. Among several parameters, degree, betweenness centrality and closeness centrality had better outcomes and can used for topological analysis of the PPI networks.

6- Affymetrix gene expression data should be validated by an independent method such as real time qPCR in order to be sure of the obtained results.

7- Were the 3 groups (healthy, NAFLD and NASH) really matched based on clinical factors?

8- For increasing the credibility of DEGs Matching Authors should be used value distribution (by R tool) between normal-NAFLD, normal-NASH and NAFLD-NASH patients and then should be realize that they are completely match. Distribution plot is useful for determining if value data are median-centered across samples, and thus suitable for cross-comparison.

9- What guild line did you use, it was not important for you that the elements of study like inclusion and exclusion criteria, confounding factors and the way for controlling it, please mention those.

10- Funrich R is not powerful tool for enrichment analysis. To find the biological pathways involved and the gene ontology in the gene list, they should be also another tools such and Enrich R, DAVID Functional tool or BINGO plugin of Cytoscape.

11- KEGG pathways analysis should be reported.

12- The discussion is superficial. They have to discuss about the possible role of detected genes and pathways in NAFLD.

13- Manuscript is suffering from several scientific and English writing problems. Also, problem of punctuation, and different font sizes are observed in different parts of the text.

Seyed Moayed Alavian: EIC | **Revision** (**0**) 15 Apr 2022

Dear Authors,

Please follow the peer-review comments and submit a revision.

Regards
EIC

OPEN PEER REVIEW

Revision (1)

Reply to Reviewers

Ideally, the reviewing process can significantly improve the submitted manuscripts by allowing the authors to take into account the advice of reviewers. Author(s) must reply to all reviewers' comments in a separate Word file, point by point. A "**Reply to Reviewers**" document is submitted along with revised manuscript during submission of revised files, summarizing the changes that the authors made in response to the reviewers' comments. The responses to reviewers' comments specifies how the authors addressed each comment the reviewers made.

You can read the authors' responses to the reviewers' comments in the next page.

Manuscript Number: 122362

Manuscript Title: **Identification of shared Gene signatures in different stages of non-alcoholic fatty liver disease using integrated microarray datasets**

Dear Editor,

On behalf of my co-authors, I thank you very much for giving us an opportunity to revise our manuscript. We are also very grateful to the reviewers for their invaluable comments on our manuscript. We have revised our manuscript in accordance with the reviewers' comments. The revised portions in our manuscript are highlighted in green. The following are the point-by-point responses to each of the comments of the reviewers. Enclosed please find a highlighted copy. We believe that we have adequately responded to the reviewers' concerns and greatly improved the manuscript. We look forward to hearing from you soon.

Yours sincerely,

Corresponding Author

Reviewers' comments	**Response to reviewers**
Reviewer 1	
The data processing results, which include their normalization and clustering, should be presented in the supplementary file, and the method for removing data noise should be explained. On the other hand, the reason why datasets have been reviewed separately and not merged together should also be expressed.	Thank you for your comments. According to this comment, our manuscript has been revised by adding a paragraph about preprocessing step. Also, supplementary 1 contain normalization, clustering and PCA figures was added. The changes are highlighted in green in last paragraph of data processing section.
Provide a table of DEGs in the article or the supplementary file.	Thank you for your comment. We provided 3 supplementary files (Supplementary file 2, 3, and 4) for DEGs list. We mentioned it by green in the first paragraph of results section.
It is better to classify gene modules in biological processes (BP), cellular component (CC), and molecular function (MF).	We classified pathways and BPs in datasets using UPSET tool and added figure 6 for all DEGs in 3 datasets.

Drawing effective subnetworks using suitable packages in the Cytoscape should be considered, and the list of genes in these modules should be considered for pathway enrichment analysis.	We used network analyser plugin in cytoscape tool for all groups in the all datasets.
Although the authors have mentioned that this is a computational article and lacks an experimental section, in any case, validation of some highly differentially expressed genes and hub genes is recommended by qRT-PCR.	Due to the nature of the study, which is of computational type, and the fact that the study does not include a specific budget for laboratory approval, this study was performed only computational. We have mentioned it in the discussion as a limitation.
Reviewer 2	
1- GEO contains several datasets for gene expression data for NAFLD and NASH. What's the criteria for selection of 3 datasets (GSE48452, GSE63067, GSE89632)?	We revised methods section according to your comment and added one paragraph about inclusion and exclusion criteria. The changes are highlighted in green.
2- Genes should be clustered using a k-means clustering algorithm or using principal component analysis. I suggest you extensively read the documentation of Heatmap at Bioconductor.	Given the gene aggregation for datasets has not occurred, this will not provide useful information about the status of genes because we have clustering gene separately for each dataset, for this reason, please disregard this comment.
3- Weighted gene co-expression network analysis (WGCNA) should be used to cluster the specific up-regulated genes in each subgroup. WGCNA has been proven to be an effective method for detecting multiple co-expression modules and can be used to find clusters (modules) of highly related genes. The optimal Power value should be reported through the Power value scatter plot, and the distance between genes should be calculated. In addition, the average method and the dynamic method should be used for hierarchical clustering analysis. The clustering diagram and the module classification of genes should be established, and similar	First of all, thanks for the great idea for the WGCNA study, but there are two points to this analysis that we will mention. 1- WGCNA analysis for this study can be done when we have used the method of fusion of all datasets. Now, considering that we have analyzed each database separately and then the results integrated, so it is not possible to perform co-expression analysis in this way. 2- Performing WGCNA analysis for each dataset is also possible according to the WGCNA analysis tutorial when the number of samples for the dataset is at least 20 samples. Because one of the datasets has less than 20 samples, it is not possible to perform an analysis for it. Therefore, we ask you to allow us to do this in your next manuscript.

modules should be merged.	
4- There are several Datasets when you perform a search with NAFLD. Some studies analysed the aforementioned GSE datasets. Could some of the findings presented in this manuscript be validated in some other cohort of patients? Also, for obtaining reliable and comprehensive results, RNA-seq data should be analysed.	The results of this analysis have been evaluated in valid databases and for each observation in the discussion; we have provided explanations as evidence. So we point out that some of the results expressed have been confirmed by other studies.
5- Please note that the hub genes extracted from topological analysis based on important centrality measures. I strongly recommended that "hub genes" should be reported. Topological network analysis has performed for PPI networks and the most important hub genes can be identified. The network analyzer plugin of Cytoscape can used to analyze topological properties of the networks. Among several parameters, degree, betweenness centrality and closeness centrality had better outcomes and can used for topological analysis of the PPI networks.	We used network analyser plugin in cytoscape tool for all groups in the all datasets. The figures 4 and 5 show the results of this analysis. Also we have mentioned it in the last paragraph of PPI network section.
6- Affymetrix gene expression data should be validated by an independent method such as real time qPCR in order to be sure of the obtained results.	Due to the nature of the study, which is of computational type, and the fact that the study does not include a specific budget for laboratory approval, this study was performed only computational. We have mentioned it in the discussion as a limitation.
7- Were the 3 groups (healthy, NAFLD and NASH) really matched based on clinical factors?	The groups were matched in each dataset, but not among datasets. Because the stage of disease among datasets was important. Of note, all datasets were chosen from adult patients but not exactly the same age. The characteristics of the people participating in each dataset were mentioned in the second paragraph of data processing section and in the Table 1.
8- For increasing the credibility of DEGs Matching Authors should be used value	We mentioned this by adding the supplementary file 1. In supplementary file 1, the distribution of

distribution (by R tool) between normal-NAFLD, normal-NASH and NAFLD-NASH patients and then should be realize that they are completely match. Distribution plot is useful for determining if value data are median-centered across samples, and thus suitable for cross-comparison.	sample values between groups and datasets can be seen.
9- What guild line did you use, it was not important for you that the elements of study like inclusion and exclusion criteria, confounding factors and the way for controlling it, please mention those.	We revised methods section according to your comment and added one paragraph about inclusion and exclusion criteria. The changes are highlighted in green.
10- Funrich R is not powerful tool for enrichment analysis. To find the biological pathways involved and the gene ontology in the gene list, they should be also another tools such and Enrich R, DAVID Functional tool or BINGO plugin of Cytoscape.	We used Enrichr tool to enrichment analysis and then updated the results in the table 3, 4, and 5 and discussion section.
11- KEGG pathways analysis should be reported.	The KEGG pathways list in the table 3, 4, and 5.
12- The discussion is superficial. They have to discuss about the possible role of detected genes and pathways in NAFLD.	We revised this section according to your comment. The changes are highlighted in green.
13- Manuscript is suffering from several scientific and English writing problems. Also, problem of punctuation, and different font sizes are observed in different parts of the text.	According to this comment, our manuscript has been revised by an English editor. The English edit certificate has been attached.

OPEN PEER REVIEW

Revision (1)

Here, you can see the **Reviewers**, **Associate Editors** and **EICs'** comments from the beginning to the end of the revision process.

REFEREE: Associate Editor | **Revision** (1) 8 Jun 2022

Dear Authors,
I am writing to confirm my acceptance of your revised article offer on 2022-05-21. Thank you for your patience with the review process.
Best Regards

Seyed Moayed Alavian: EIC | **Revision** (1) 8 Jun 2022

Dear Authors,

I am delighted to inform that your manuscript is accepted for publication in Hepatitis Monthly.

Regards,
EIC