**Reviewer’s report**

**Title:** Multi-level transcriptome sequencing identifies COL1A1 as a candidate marker in human heart failure progression

**Version:** 0  **Date:** 01 Jul 2019

**Reviewer:** Mark Ziemann

**Reviewer's report:**

In the article by Hua et al, transcriptome analysis (mRNA, lncRNA and smRNA) of cardiac tissue from heart failure (n=21) and healthy patients (n=9) was undertaken to better understand the pathogenesis of the disease. Using bioinformatics methods, differentially expressed genes were identified with each approach. Downstream enrichment analysis showed some association with cardiovascular disease and fibrosis. Network analysis identified hub genes linked to a higher than expected number of GO terms. At this point, the report begins to focus more narrowly on the regulation of fibrosis/ECM. A LASSO regression of the gene expression profiles with fibrotic content was conducted for n=18 patients. It is perhaps not surprising that genes with known fibrosis associations, such as COL1A1 were identified. COL1A1 expression was found to correlate with a faster rate of HF progression, as was COL1A1 protein by IHC. QPCR analysis in separate group of HF and control LV samples validated its over-expression in HF. ELISA analysis of plasma COL1A1 in 87 HF patients found its elevated expression correlated with poor outcomes. This paper presents much needed transcriptome data on HF in LV tissue, which to this date remains limited to a few small studies. The identification and validation of plasma COL1A1 is important to the field as there is some potential to apply this to complement existing physiological measures of HF. That being said, there are several weak points to this study that need to be addressed. Mostly these relate to lack of detail around methods, and poor word choice that conflates correlation with function.

Title: Implies causation. Must be amended

P4: Information about the healthy donors is missing from Table 1. This is essential to knowing whether they are compositionally similar to the HF group in terms of age, sex, smoking and other covariates. Moreover it is important to know whether the healthy heart samples were obtained from post-mortem or another source as time since death may confound statistical analysis (https://doi.org/10.1038/s41467-017-02772-x).

P4: The methods for library preparation are not described, these should include the kit vendor, version number and whether any modifications were made to the manufacturer's protocol. This information will help other researchers trying to repeat the analysis of the dataset.

P5 L13: "researches" → "researchers"

P5 L35: please provide version numbers for CutAdapt and all other software packages used.

P5 L44: Was a mapping quality threshold used for HISAT2 based read counting?

P6 L30: What is the rationale behind using Ballgown for mRNA and lncRNA DE analysis, and DESeq2 for miRNA? Why not simply use DESeq for all three types of count data? In performing DE analysis, were factors such as age, sex, smoking corrected for?
P6 L41: Downstream analysis is a weak part of the paper. Over-representation analysis methods like DAVID that use hypergeometric tests demonstrate relatively poor sensitivity as compared to those that use the full list of genes (eg: GSEA, CAMERA, etc) because the results are strongly influenced by the strictness of the DE cut-off. The use of non-adjusted p-values invalidates this analysis. The BH p-value adjustment method should be applied to all p-values for all the gene sets assessed. In addition, it seems the authors have fallen into the common pitfall of not using an appropriate background non-DE gene list to DAVID (https://doi.org/10.1186/s13059-015-0761-7). Running DAVID without a background list of genes expressed in LV will results in invalid findings and many false positives.

P6 L57: miEAA has two modes: Over-Representation Analysis (ORA) and miRNA enrichment analysis ((G)SEA). Which one was used here? Was there any FDR adjustment of the p-values done? What FDR threshold was used?

P9 L20: The wording of the 1st sentence in the results is somewhat unclear. The patients were not collected from the heart transplantation dataset, rather they were more likely recruited from the heart transplantation program of Fuwai hospital.

P9 L29: The age of HF patients is extraordinarily young in this study. In general, HF patients are mostly &gt;60 yrs old (doi: 10.1038/nrcardio.2010.165). Was there any particular rationale to profiling patients this young? This could be highlighted as a unique feature of this study in the discussion.

Table 2: The figures provided are a sum total of reads over the entire experiment, but this would be better shown as mean and SD.

Figure 4a: The enrichment needs to be plotted by FDR adjusted p-value.

P10 L59: What is the relevance of references 32-36 in this passage?

P11 L5: It is stated "Genes associated with multiple terms are presumed to be more critical in HF ..." However I cannot find any evidence that directly links essentiality of genes in disease with depth of GO term annotation.

P11 L39: "... these findings indicated the important roles of the DElncs and DEmiRs in the development of HF" The wording here implies causation which is highly problematic. With any type of gene expression analysis we are identifying correlations, and these should never be confused. Phrases such as this that appear throughout the paper (and in the title!) need to be amended.

P12 L7: "... miR-548ar-33 interacted with ..." The word "interaction" implies direct physical contact, perhaps "association" is better.

P13 L35 "recruited" → "obtained"

Fig 6c The coefficients should be ordered from lowest to highest. COL1A1 coefficient is extremely small compared to other genes like COLQ, what does this mean for the robustness of the correlation?

P13 L52 "implied" → "implicated"

P16 L5-9: This section is puzzling, as it states that plasma COL1A1 is anti-correlated with HF progression, which is the opposite of what is shown in Figure 7e.

P16 L28: Median plasma COL1A1 is reported to 6 significant figures, which implies that the instrument is accurate to that level of precision. A more reasonable number is 3 to 4 significant figures (here and in the abstract).

P17 L20: "... lncRNAs are more informative for labeling ..." I'm not sure this statement is reflected by the data. Author can qualify this by saying "According to PCA analysis, HF and control samples are better distinguished by lncRNA profiles than mRNA or miRNA."
P18 L5 "... the role of fibrotic genes in HF progression is still largely unknown." This seems rather contradictory. Fibrosis genes cause fibrosis; fibrosis causes myocardial stiffness; stiffness results in HF. Authors allude to this on P14 L24.

P20 L5: Authors state data is available upon request, however this is unsatisfactory. It is the norm for genomics based projects to submit the NGS profiling data to NCBI GEO or similar permanent data repository to enhance the reuse potential of these data. This is in accordance with the NIH policy (https://www.niaid.nih.gov/research/data-sharing-and-release-guidelines) and is compatible with privacy rules (ie: data is non-identifiable).

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

**Statement on potential review bias**

Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?
2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria’. If your reply is yes to any, please give details below.

I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal