Author’s response to reviews

Title: Identification of differentially expressed lncRNAs and mRNAs in luminal-B breast cancer by RNA-sequencing

Authors:
Cheng-Liang Yuan (13568235628@163.com)
Xiang-Mei Jiang (412584237@qq.com)
Ying Yi (ying8975@163.com)
Jian-Fei E (1172200180@qq.com)
Nai-Dan Zhang (znd1984@126.com)
Xue Luo (54393816@qq.com)
Ning Zou (deylab@163.com)
Wei Wei (33911035@qq.com)
Ying-Ying Liu (dysrmyyliu@163.com)

Version: 2 Date: 28 Oct 2019

Dear editors and reviewers,
Many thanks for your high efficiency of work and good suggestions. We have revised the manuscript according to the comments and the recommendations suggested by the editors and reviewers. We use ‘Track changes’ model to show the part that we revised in the manuscript.

The responses to the reviews' comments are as follows:

Reviewer reports:
supreet agarwal, PhD (Reviewer 2): In this manuscript, the authors analyzed 4 luminal B type breast cancer patients' tumors and adjacent normal tissues and identified several differentially expressed mRNA and lncRNAs.
Overall, the results of this study are weak and descriptive. The authors should confirm their findings in a large cohort by analyzing public databases. Further, no subsequent functional study was performed to validate the identified genes and lncRNAs.
Response: Thanks for your comments. As per your suggestion, we have validated our results (including
S100A7, CCL5, MIAT and WT1-AS) in TCGA database, which included 171 luminal B breast cancer tumor tissues and 36 normal tissues. The validation results revealed that, except for MIAT, the expression files of S100A7, CCL5 and WT1-AS were completely consistent with our RNA-sequencing results. These findings suggested the reliability of our RNA-sequencing results. As for subsequent functional study, we are collecting samples recently, and the key mRNAs and IncRNAs identified in this study will be validated in our following research with larger sample size.

Adriana Aguilar-Lemarroy (Reviewer 3): The manuscript is very interesting, and I consider it very valuable, however, there are some points that need to be reviewed:

- In the methods section, in patients and samples, the description of the patients is poorly detailed, authors should include if the patients were under any treatment or were free of treatment. In Table 1, one of the patients is HER2 positive, under which selection criteria they considered this sample as Luminal? The percentage found of each of the receptors should be included in Table 1, as also age of each patient. Additionally, how adjacent was the sample taken without a tumor? Why were tumor-free breast samples not included?

Response: Thanks for your comments. We have added the details of the description of the patients as follows: all patients involved in this study were free of treatment. The criteria we used to identify subtypes were defined at St. Gallen International Expert Consensus 2013. According to these criteria, the subtypes in question have been defined as: Luminal A, Luminal B (HER2 negative), Luminal B-like (HER2 positive), HER2 positive, and triple negative. The percentage of each receptors were included in Table 1, as well as age of each patient. The adjacent tissues were obtained at 5 cm distance from the tumor tissues, which could be considered as normal tissues.

- In the method sections, it is necessary to include several details such as:
  - what conditions and what software were used to filter the data and the quality of the readings?
  - what were the trimming data conditions? which phred was used?
  - the sequencing was done with paired-ends or single ends? and the depth?
  - It is important to mention what kind of normalization was used with String Tie.

Response: Thanks for your comments. According to the four questions proposed above, we added the details of sequencing and data process as follows: The sequencing was done with paired-ends and 10G depth. The FASTQ sequence data were obtained from the RNA-sequencing data by using Base Calling version 0.11.4 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read QC tool in FastQC version 0.11.4 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for the quality control of FASTQ data with Q > 30.Trimming of raw data was performed with cutadapt version 1.16 (http://cutadapt.readthedocs.io). Reads with low quality (adaptor sequences, sequences with a quality score < 20, and sequences with an N base rate of raw reads >10%) were removed to obtain the clean reads.
  - Did authors performed post hoc tests to get the padj? or Q value? and why authors did not take that value instead of the p?

Response: Thanks for your comment. Actually, we have calculated the padj, and then fewer IncRNAs were obtained with the criteria of padj. For consistency, p-value was used as the filter criterion of DEmRNAs and DEIncRNAs.

- In the analysis of functional annotations authors need to describe what parameters they used for enrichment.

Response: Thanks for your comment. We attached the instructions of Metascape software as follows: When the Enrichment tab is activated, the program automatically takes the input gene list and displays all of the ontology categories, where the input genes fall under. If an ontology category contains zero input genes (i.e., when we forget to select input), it is not shown on the interface. We can check and uncheck boxes to specify the ontologies most appropriate for the project. We can then click
"Enrichment Analysis" to calculate p-values and enrichment factors for each ontology term, and only those that are deemed statistically significant (based on parameters in the control fields) are retained. To avoid the mistake of missing foreground genes, genes in the input list are automatically considered as background members as well. Background gene population is important for experiments conducted using a functionally-biased gene collection (such as kinases-only and secreted-protein-only gene collections). However, for most unbiased backgrounds, the precise list of background genes is often not known, and the cutoff to generate the input hit list is somewhat subjective, so we generally accept the practice of using the whole genome as our analysis background. An approximate unbiased gene population background may shift the absolute values of the p-values, but we expect it only has limited effects on the relative ranking of the enriched terms. This is also an accepted argument for comparing enrichments across different gene lists [Huang et al.].

- In protein-protein interaction networks it is necessary to include that confidence cut-off selected.
Response: Thanks for your comment. The protein-protein interaction networks were constructed with DEmRNAs that were top 100 up- and down-regulated scanned with the Biological General Repository for Interaction Datasets and then visualized with Cytoscape. There didn’t involve in confidence cut-off.

- I'm curious to know why authors included in their analysis of DElncRNA-DEmRNA co-expression networks all the genes and not only those found in the Cis-nearby-targeted DEmRNAs of DElncRNAs analysis?
Response: Thanks for your comment. As your suggestion, the construction of DElncRNA-DEmRNA co-expression networks included only those found in the Cis-nearby-targeted DEmRNAs of DElncRNAs analysis were performed, and then fewer mRNAs and IncRNAs were obtained. As a pilot study, we are currently focusing on the key DEmRNAs and DElncRNAs associated with luminal B breast cancer. Further research, including subsequent functional study, with larger sample size will be performed in the near future.

- In the figure 1c, I do not see the pink and light blue inner layer that authors described (that represents the distribution of up- and down-regulated DElncRNAs on different chromo-some).
Response: Sorry for our mistakes, and we have corrected that in our manuscript.

- I consider that the work is based on the analysis of very few samples, a larger number of samples would be necessary to strengthen the conclusions.
Response: Thanks very much for your comments. We have to admit that the sample size was small. Hence, we added a limitation in the last paragraph of discussion. To cover the shortage of our small sample size, we added the validation results (including S100A7, CCL5, MIAT and WT1-AS) based on TCGA database, which included 171 luminal B breast cancer tumor tissues and 36 normal tissues. The validation results revealed that, except for MIAT, the expression files of S100A7, CCL5 and WT1-AS were completely consistent with our RNA-sequencing results. These findings suggested the reliability of our RNA-sequencing results. Currently, we are collecting samples, and the key genes and IncRNAs identified in this study will be validated in our following research, including subsequent functional study, with larger sample size.

- Authors do not include if they have deposited their datasets in a publicly available repository, BMC Cancer strongly encourages that all datasets on which the conclusions of the paper rely should be available to readers.
Response: Thanks for your comment. We have uploaded all raw data to Gene Expression Omnibus (GEO) (GSE139274, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139274) and indicated that in the manuscript.
Thank you again for your great help and attention. I am looking forward to hearing from you about the final decision.
Best regards and wishes!
Yours sincerely,
Ning Zou