Author’s response to reviews

Title: Tumor classification and biomarker discovery based on the 5'isomiR expression level

Authors:

shengqin wang (wzsqwang@gmail.com)
Zhihong Zheng (zhihongzheng@zju.edu.cn)
Peichao Chen (chenpeichao@wzu.edu.cn)
Mingjiang Wu (wmj@wzu.edu.cn)

Version: 1 Date: 20 Dec 2018

Author’s response to reviews:

Dear Akila Sridhar, Linda Gummlich and Reviewer(s),

On behalf of my colleagues, I wish to offer our appreciation to you and the reviewer for all the work done on our paper “Tumor classification and biomarker discovery based on the 5'isomiR expression level” (Ms. No.: BCAN-D-18-02009). We now respond all comments in detail (in blue text and “#” marked) as follows and the comments made to us are in plain text.

Reviewer reports:

Hong-Qiang Wang (Reviewer 1): The authors constructed a combination of the genetic algorithms (GA) with support vector machine (SVM) algorithms to detect cancer-associated 5'isomiRs from TCGA isomiR expression data for tumor classification. This classified samples from 32 different tumor types with an average sensitivity of 91.5%. They detected many highly frequent 5'isomiRs with different 5’ loci from canonical miRNAs and showed that isomiRs play a significant role in the multiclass tumor classification. However, I have the following major concerns:

1. The authors said that they downloaded 33 tumor types which were used for evaluation. It is unclear the way they classified the samples? separate tumors and normal tissues, or classify different types of tumors.

# Thanks. Here, we only download the primary solid tumor samples, which have the infix “01” in the TCGA sample barcode, with the exception of the blood samples derived from acute myeloid leukemia (LAML; sample infix ‘03’) were included in the study. We classify the tumor based on the original annotation from TCGA, and we do not contain the normal tissues.
2. It seems to be not significant that they got an average sensitivity of 91.5%, given that previous work can classify more than 90% of samples from 31 tumor types with a set of 20 genes. What is the accuracy? and how about specificity? Why not try other previous methods except GA/SVM.

# Here, the another feature MCC (Matthew’s Correlation Coefficient) was provided for the performance of our work (Fig. 2B). We also tried GA/KNN and GA/SVM. The distribution of the expression level of isomiRs were quite uneven, we did not achieve good performance for GA/KNN after using kNN Imputation to impute the missing values. After compared with the result from GA/RF, which could effectively classify with an average sensitivity of 92% samples from 32 different tumor types, we decided to chose the GA/RF method finally. The goal of our method was to evaluate the effective reduction of discriminant isomiR features for multiclass TCGA tumor discrimination classification. we reduced the number of features by employing two strategies. In the first approach, we combined the isomiR with same 5’ loci to reduce the type of isomiRs. While in the second approach, the GA-based isomiR selection reduced the feature selection significantly.


# Thanks, we added it in the second paragraph of introduction: “The filter, wrapper and embedded methods are typically utilized for feature selection, though all of them are not good at deal with data which contained a large number of collinear variables. [20,21]. The isomiRs may belong to the same miRNA family, the same miRNA cluster, or some of them even have same seed region, leading to similar or related function and highly correlated expression. In the previous studies, the wrapper method could outperform embedded methods by combined the machine learning algorithm for classification [21].”

Wenyi Yan (Reviewer 2): The authors proposed a GA/SVM model for tumor classification from isomiR expression data. The method is interesting and useful. It is technically sound. The test results are promising. The paper is well written. But the following concerns need clarification:

1. Any specific reason to use/implement SVM after feature selection by GA? Did you compare the performance of SVM with other alternative algorithms such as random forest? And if they also work well with GA? The authors need to at least provide an explanation of why SVM was chosen.

# Here, we also tried GA/KNN and GA/SVM. The distribution of the expression level of isomiRs were quite unenven, we did not achieve good performance for GA/KNN after using kNN Imputation to impute the missing values. After compared with the result from GA/RF, which could effectively classify with an average sensitivity of 92% samples from 32 different tumor types, we decided to chose the GA/RF method finally.

2. The comparison with other similar models should be provided to show the improvement for tumor classification.
The goal of our method was to evaluate the effective reduction of discriminant isomiR features for multiclass TCGA tumor discrimination classification. We reduced the number of features by employing two strategies. In the first approach, we combined the isomiR with same 5' loci to reduce the type of isomiRs. While in the second approach, the GA-based isomiR selection reduced the feature selection significantly.

3. For the performance of the GA/SVM model, may be more metrics could provide meaningful results like MCC or F-score.

Thanks, the another feature MCC (Matthew’s Correlation Coefficient) was provided for the performance of our work (Fig. 2B).

Russell S. Schwartz (Reviewer 3): Wang et al. develop a classification framework for exploring the value of isomiRs, non-canonical miRNAs, in classifying cancers by tissue of origin. isomiRs are biologically interesting and the paper makes a good case from the prior literature for the value of exploring their relevance to cancers via their utility as biomarkers. The predictive power of isomiRs for this task was already established in prior work, and the specific contribution of the present work is to develop a feature selection approach to demonstrate that a small subset of isomiRs has predictive power close to that of the full set. The paper provides some good evidence that a small subset of isomiRs can still offer predictive power and thus likely have functional relevance independent of what one gets from canonical miRNAs alone.

Thank you for your positive review!

The overall work of the paper seems a fair way to explore biological relevance of isomiRs based on their predictive power, although it does raise some questions. The specific task of classifying cancers according to tissue of origin has some evident utility for exploring the role of miRNAs in cancer development, although it is not obviously the best task for the job. I might expect that distinguishing cancer from non-cancer or classifying cancer subtypes for individual tissues of origin would be a more useful and enlightening task on the functional relevance of the isomiRs in cancer. The paper could also do a better job of motivating why it is interesting in particular to reduce the predictive set to a small number of isomiRs.

Thanks. Here, we only download the primary solid tumor samples, which have the infix “01” in the TCGA sample barcode, with the exception of the blood samples derived from acute myeloid leukemia (LAML; sample infix ‘03’) were included in the study. We classify the tumor based on the original annotation from TCGA, and we do not contain the normal tissues. One of the reasons was that the original data only included few normal or non-cancer tissues. The reduction of discriminant isomiR features for multiclass TCGA tumor discrimination classification should be benefit for the Tumor related biomarker research and the detecting of the relationship among Tumor types. More other works should be done in the future.

The machine learning approach is reasonable, although not a significant intellectual contribution in itself. SVMs are a good generic classifier that could plausibly work well for this problem, although not the only or obviously best method for the problem. Genetic algorithms are
similarly a standard heuristic for hard optimization problems like the feature selection done here and again seem a reasonable if not greatly innovative choice for the algorithm. The paper might better justify why some other feature selection or regularization approaches were not chosen instead, e.g., LASSO or its variants. There were also some ad hoc parameter choices that might be justified, e.g., the use of 50 isomiRs per data set.

# The filter, wrapper and embedded methods are typically utilized for feature selection, though all of them are not good at deal with data which contained a large number of collinear variables. [20,21]. The isomiRs may belong to the same miRNA family, the same miRNA cluster, or some of them even have same seed region, leading to similar or related function and highly correlated expression. In the previous studies, the wrapper method could outperform embedded methods by combined the machine learning algorithm for classification [21]. As one of the embedded methods for feature selection, the LASSO should be good at finding feature sets, but not for achieving the best performance.

The methodology could use clarification on some technical points. I was unclear exactly what was meant by the phrase `combining all the miRNA isofomrms with same loci of 5' end together in TCGA isomiR expression data.` This statement could be made more precise. In general, the paper might have benefitted from a flow chart or pseudocode spelling out precisely what the steps of the overall method are.

# Here, we combined all the miRNA isoforms with same loci of 5' end together in TCGA isomiR expression data. And the miRNA isoforms with same loci of 5' end will be left only one in the reliable sets, which will dramatically reduce the type of isomiRs. The flow chart has been provided in Figure 1.

The application to TCGA data appears reasonable with appropriate preprocessing and some good follow-up on frequently occurring isomiRs in the strong predictors. The classification results are promising and do suggest that non-canonical miRNAs can offer some additional predictive power beyond what one gets with canonical miRNAs alone. I do think it might be worth exploring in a bit more detail why that is, though. That is, what kinds of mistakes does the method make with only canonical miRNAs that are corrected when one adds isomiRs? As noted above, I am also unconvinced this is the right predictive task on which to explore the relevance of isomiRs, since it would confound the problem of finding cancer-specific biomarkers with the problem of finding tissue-specific biomarkers. The other tasks suggested above (distinguishing tumor from normal or tumor subtyping) might be better suited at testing whether isomiRs are specifically relevant to identifying cancers or particular mechanisms of tumorigenesis. It would also be worth seeing how these results compare to predictive power one gets from gene or protein expression data.

# Thanks, our analysis only included tumor samples, and we cannot distinguish cancer-specific isomiRs from tissue-specific biomarkers. Actually, a group from Saarland university had utilized a tissue specificity index to define the distribution of miRNA across 61 tissue biopsies of two individuals, and people can check whether the detected isomiRs are the tissue-specific miRNA expression in their web-based repository[35].
The functional enrichment analysis is a good addition for understanding what these isomiRs might be doing, although the results that come up are fairly generic regulatory categories. Are these enrichment results based on enrichment relative to isomiRs or miRNA targets in general or relative to the whole genome? If the latter, I think it would be worth exploring the former, since it would be useful to know if there is something interesting about those specific isomiRs beyond the role of isomiRs or even miRNAs in general. There would also seem to be some uncertainty introduced by using predicted miRNA targets. One often has the gene expression for tumor and matched normal tissues for TCGA and might be able to refine these target sets based on which genes are actually dysregulated in the relevant cancers. The paper might also have benefitted from more follow-up on known biological relevance, if any, of at least the top few most frequently appearing isomiRs.

# Thanks, these enrichment results were based on enrichment relative to targets of the top 9 most frequently appearing isomiRs.

The paper is generally clear, well organized, and well written, aside from the request for an overall figure on the workflow noted above.

# Thank you for your positive review!

# Again, we would like to thank the reviewer and editors for their valuable suggestions and comments, which have led to the improvement of review articles.