Author's response to reviews

Title: Differentially Expressed Alternatively Spliced Genes in Malignant Pleural Mesothelioma Identified Using Massively Parallel Transcriptome Sequencing

Authors:

lingsheng Dong (ldong@partners.org)
Roderick V Jensen (rvjensen@vt.edu)
Assunta De Rienzo (aderienzo@partners.org)
Gavin J Gordon (ggordon@partners.org)
Yanlong Xu (yxu11@partners.org)
David J Sugarbaker (dsugarbaker@partners.org)
Raphael Bueno (rbueno@partners.org)

Version: 5 Date: 10 October 2009

Author's response to reviews: see over
Editor’s request:

“Please also ensure that your revised manuscript conforms to the journal style. It is important that your files are correctly formatted.”

Answer: Yes. We re-formatted the manuscript according to the requirements.

REVIEWER 1:

Minor Essential Revisions:

The authors have identified alternatively spliced genes in MPM using 454 Sequencing. Overall this was a well-written manuscript, however there were some concerns that the authors should consider:

1. Providing a better explanation regarding their sample size. It mentions that they analyzed 4 MPM samples with 1 normal lung sample a few times (n=5), however they also constantly reference their previous analysis using 4 MPM samples, 1 normal lung sample, and 1 lung adenocarcinoma sample (n=6).

   Answer: Corrected. In previous study, we sequenced 6 samples, but only 5 of them are related to the current project to identity the difference between MPM and normal lung. We revised the manuscript to make the sample size consistent and clear to the readers.

2. Providing a power calculation to show how they determine that 36 samples (18 MPM and 18 normal lung samples) were sufficient for analysis.

   Answer: Corrected. We used G*Power and calculated the power statistics in the result section.

3. Providing additional information on why they selected only the 3 top ranked exon junctions at higher levels in normal lung samples while selecting for the top 7 levels in the MPM samples, instead of just selecting for maybe the top 10 in each sample-type (so that they are not weighting just the number of junctions in the tumor samples).

   Answer: The candidate selections step was rather arbitrary. We weighted our list towards exon junctions over-expressed in tumor because there are more highly expressed exon junctions in MPM (76) than in normal lung (32).
4. Explaining why they removed the redundant exon junctions with the same or similar coordinates (under Methods / Annotation of AceView exon junctions and creation of virtual probes)? Especially since some genes have the potential to be very repetitive.

Answer: In Aceview database, there are often multiple transcripts for each gene, so exon junctions with same (similar) chromosome coordinates from different transcript of the same gene will have same mRNA sequence around it. In order to generate non-redundant virtual probe sequences, we eliminated the redundant exon junctions. This was explained in detail in the revised manuscript in the section: Method/annotation of Aceview exon junctions and creation of virtual probes.

5. Spelling out "GFF" abbreviation under Methods/AceView transcriptome database Section.

Answer: Corrected. GFF (General Feature Format) is a format for describing genes and other features associated with DNA, RNA and Protein sequences

6. In the Methods/Real time quantitative RT-PCR Section, the authors should think about extracting out the information regarding the CT Equation and place in an appendix for further explanation.

Answer: Corrected. An additional file (additional file 1) is created to explain the formula in detail.

7. Consider making an additional supplemental table including the resulting genes/exons and their levels so that the reader doesn't have to sort through the manuscript to look for them.

Answer: Corrected. We presented the Exon junction expression index in the candidate gene table (additional file 5).

REVIEWER 2:

This manuscript by Dong et al. describes the reanalysis of a previously generated (and published) dataset of more than 15 million ESTs generated for 4 malignant pleural mesothelioma (MPM) tumors, 1 lung adenocarcinoma and 1 normal lung. In this manuscript the authors investigate alternative splicing profiles among these samples. In general this is an interesting and useful study of alternative splicing using new sequencing technology. However, a number of issues left me unconvinced about the
reliability of the results. My major concern is that key data have not been sufficiently presented.

Major Compulsory Revisions:

1. The results are not presented in a comprehensive way, which makes it difficult to evaluate if the discussion is supported by the data. The absolute counts on which the authors based their discussion should be presented, at least for the highly expressed exon junctions (EJs). I would suggest a table with: (i) the number of reads mapping to the EJ in each sample; (ii) the number of reads mapping to any of the EJs of the gene (which are used to calculate the exon junction expression index, EJEI); (iii) total number of reads mapped to the gene in each sample.

Answer: Corrected. We have added a table (additional file 5) to present the mapping results for the ten candidate exon junctions we selected to be validated using RT-PCR.

2. Some basic descriptive statistics are also missing, namely (i) the sum of the EJEI in each patient, whose sum was used to normalize EJEIs; (ii) the number of sample-specific variants; (iii) number of differentially expressed EJs.

Answer: Corrected. We have added three additional files (2 to 4) to present the distribution of the exon junctions.

3. A more careful dissection of the data would be important to identify systematic biases on their data. For instance, there have been some reports of a non-even distribution of 454-reads along the transcripts. Are the differentially expressed EJ evenly distributed along the gene?

Answer: We agree with the reviewer that the 454-reads are not evenly distributed along the transcript. However, the bias should equally exist as long as the transcript is expressed in each sample. To minimize the side-effect of this bias, instead of using absolute read count to identify differentially expressed exons, we calculated the ratio between the number of reads that cover a specific exon junction and the total number of the reads that map to any of the exon junctions of the gene. This is similar to the Splicing Index reported in ArrayAssist software for analyzing the exon level data on the Affymetrix Exon arrays. Here, we are counting the relative frequency of the exon junctions by themselves.
4. The authors should be able to provide the levels of expression (given by the total number of reads mapped to it) of the genes in which differentially expressed EJs were observed. How the levels of expression of the EJs relates to the levels of expression of the gene?

Answer: As suggested by the reviewer, we have three additional files (additional file 2 to 4) to present the total number of the reads (sum of exon junction reads) for genes related to the top differentially expressed exon junctions. We also attempted to use the total number of reads for each gene to represent the gene expression level. But because the 3’ bias mentioned in comment 3 (more 454 reads map to 3’ of the gene), we decided to use the sum of the reads mapping to all exon junctions for each gene. For this reason, we believe the expression levels of the EJs are not directly related to the gene expression level (represented by the total number of reads).

5. The authors state in the abstract that “15,789,974 transcriptome reads generated by the Roche/454 sequencing platform for all 6 samples were compared with 151,486 exon junctions from the AceView database”. Later in the text, they refer to only 5 of these samples. Were all the 15,789,974 reads mapped to the EJ virtual probes? The inclusion of the adenocarcinoma sample is an interesting approach to distinguish between MPM-specific and tumor-specific transcript variants. Have the authors assessed which variants were MPM-specific, tumor-specific or normal lung-specific?

Answer: We realized the confusion as it was also pointed out by another reviewer. During the analyses steps, we initially include the lung cancer sample, but later excluded the lung cancer, because we thought we would more likely to discovery the difference between mesothelioma and normal because we have 4 mesothelioma tumors. We have revised the manuscript and clarify the confusion.

6. I am not sure if the authors employed the best normalization to make the EJEI comparable among different exon junctions of different genes. Different genes can have very distinct number of EJs. Since the authors have previously corrected each EJ with the sum of reads mapped to all EJs in a given gene, I wonder how the EJEIs from genes with a different number of EJs can be compared.

Answer: We agree with reviewer’s point that each step of the normalization process may introduce system bias. Consequently, we tried to make the data comparable in each step. More specifically, to make exon junctions comparable among different genes, we calculated “the ratio of number of reads that cover a specific exon junction divided by the total number of the reads that map to any of
the exon junctions of the gene." In this way, the EJs with less total number EJs for the gene should be compensated by its less number of reads mapping to all the EJs. We have revised the manuscript to include this point in the method section.

7. I do not understand why the authors have not formally tested the differences between the EJEIs of the five samples. Instead, they have chosen the minimum normalized EJEI to identify highly expressed exon junctions in MPM relative to normal lung and the maximum normalized EJEI to identify highly expressed exon junctions in normal lung relative to MPM (pages 7 and 8). A statistician could be consulted to take advantage of all the information in the data.

Answer: We agree with reviewer’s comments and tried to acquire more information from the data. However, because of the small sample size (only 4 tumors and one normal), we were not able to find statistical difference between both differentially expressed exon junctions and the comparison of the results among different platforms. At a starting point, we used the most basic intuitive way to analyze it: directly checking the exon junction index difference between tumor and normal.

8. What was the rationale behind the use of the median value of relative exon junction expression levels in all 36 samples as a cut-off to classify samples as “tumor” or “normal”? This test did not provide convincing evidence that these exon junctions could be used as an unambiguous diagnostic tool.

Answer: Agreed. The cut-off value was quite arbitrary. Because we had equal number of normal and tumor samples, we assumed that the median value should be a reasonable cutoff to distinguish normal from tumor. To confirm the test as valid diagnostic tool, more samples are needed. We have revised the manuscript in the discussion section and mentioned the current limitation of this method.

9. This paper describes a reanalysis of a previously generated dataset, but it would be advisable to include a brief description of the samples and methods used to sequence the transcriptome. By doing this, the typical reader would not have to refer to the former article to get basic information. The reader would be interested, for instance, if the samples were normalized prior to sequencing.

Answer: Agreed. We have revised the manuscript in the method section to include a brief introduction of the data set.
10. The same suggestion holds for the results section. The authors could provide some basic descriptive figures, such as the number of reads obtained from each library.

Answer: Agreed. We have revised the manuscript in the discussion section to include some descriptive figures about the transcriptome sequencing reported earlier.

REVIEWER 3:
The paper is overall clearly written and interesting. The authors are using a dataset that has been previously published and analyzing it in a novel way. Detection of alternative splicing variants in tumors using next-generation sequencing is an important and interesting area which deserves further thought and study. They do a pretty thorough job in the discussion of pointing out the obvious shortcomings in their approach, although as there are several of these the work is a proof of concept as opposed to a thorough research study.

- Major Compulsory Revisions

None

- Minor Essential Revisions

1. Describe the distribution of EJEI observed in the dataset.

Answer: Corrected. We have added three additional files (2 to 4) to present the distribution of the exon junctions.

2. Describe where the numerical cutoff of EJEI was drawn (in terms of the EJEI used to define differential splicing)

Answer: Corrected. We used EJEI of 0.1 as the cutoff to select candidate exon junctions. All candidate exon junctions highly expressed tumor are now listed in additional file 3, and all candidate exon junctions highly expressed in normal are now listed in additional file 4.

3. Describe the extent of the expression difference suggested by EJEI and how well correlated it is to the RT-PCR data.
Answer: We agree with reviewer’s comments and created a new table consisting of both EJEI data and RT-PCR data together (additional file 5). We applied statistical analysis to correlate RT-PCR results with EJEI calculation. However, because of the small sample size, we were not able to find statistical difference between both differentially expressed exon junctions and the comparison of the results between the two different platforms. We have revised the manuscript in the discussion section and mentioned the current limitation of this method.

4. How much variance was observed for EJEI between the 5 "discovery" samples? Is this indicative of reliability of prediction?

Answer: We can see consistent difference between MPM tumor samples and normal lung samples comparing the EJEI (additional file 5) among the 5 “discovery” samples. Because these genes were chosen among tens of thousands of candidates, we expected to find certain amount of false discovery rate just like in Microarray data analyses.

5. Reduce use of jargon in MS – make clear what is meant by “misclassified samples” and give clear methods for which genes were chosen for RT-PCR.

Answer: We revised the manuscript to more clearly defined “misclassified”.

6. Clearly describe the limitations of only having one “control” dataset.

Answer: We now clearly point out the limitation of only having one “control” sample in the revised discussion section.

7. Discuss the relative merits and problems of using other sequencing technologies (such as Illumina or SOLiD) for the EJEI approach. Discuss the impact of increased sequence depth on the likely sensitivity and false positive rate.

Answer: As requested by the reviewer, we briefly discussed the possible application of our method in other next-generation sequencing platforms and the affect of increase sequencing depth.

- Discretionary Revisions
  1. Remove typos (e.g. “the4” and other obvious careless errors).

Answer: Corrected.

2. Describe the frequency of non-alternative-splicing derived false positives from EJEI (for example, the
frequency with which unspliced introns are seen that are likely the results of pre-mRNA sequencing). What next?

Answer: Corrected. We have revised the manuscript and discussed false positive and negative in the method section.