Author's response to reviews

Title: A Systematic Comparison and Evaluation of High Density Exon Arrays and RNA-seq Technology used to Unravel the Peripheral Blood Transcriptome of Sickle Cell Disease.

Authors:

Nalini Raghavachari (nraghavachari@nhlbi.nih.gov)
Jennifer Barb (barbj@mail.nih.gov)
Yanqin Yang (Yanqin.yang@mail.nih.gov)
Poching Liu (pcliu@nhlbi.nih.gov)
Kimberly Woodhouse (kwoodhou@nhlbi.nih.gov)
Daniel Levy (levyd@nhlbi.nih.gov)
Chris O'donnell (odonnellc@nhlbi.nih.gov)
Peter J Munson (munson@mail.nih.gov)
Gregory Kato (gkato@nhlbi.nih.gov)

Version: 8 Date: 22 March 2012

Author's response to reviews: see over
We thank the reviewer again for the careful review of our revised manuscript.

Reviewer's report:
The paper has been improved substantially during revision. Below are some additional comments.

Major Compulsory Revisions
1) The comparison of sensitivities between RNA-Seq and exon array is based on cutoff values: 1.0 RPKM for RNA-Seq and 44.5 RMA for exon array, which gives 6% more genes being detected by exon array than RNA-Seq. I found this comparison unconvincing: the difference is marginal and the cutoff values are too arbitrary.

We agree that the cutoffs used for detection above background appear arbitrary and that the 6% excess of genes detected by exon arrays is at best a marginal difference. However, our major point is that using 30-fold less mRNA in the microarray study, we still get comparable or even greater detection efficiency. We used a very conservative and reasonable background cutoff level for microarray, on the RMA scale, corresponding to the level at which 50% of the probesets show the Affymetrix DABG (Detection Above Background) value of p<=0.01. (Our RMA background cutoff was 4.5, not 44.5 as was incorrectly typed in the previous manuscript). For RNAseq we used a background cutoff of 1.0 RPKM, a level used as a cutoff by one of the pioneers of this method (Mortazavi et al Nature Methods 2008), where they imply that at a sequencing capacity of 40 M mapped reads, transcript detection was robust at 1.0 RPKM and above for a typical 2-kilobase (kb) mRNA with a P value <10-16.

Applying these limits to one sample for a control subject C3 (Figure 2), we were able to detect 6% more transcripts (12,310/11,662=1.06) above background in exon arrays than in RNA-seq. Our emphasis here is that at the time we conducted this study, even with a 30 fold less input material, we can detect as many or slightly more transcripts above background with microarray as with RNAseq. We do understand that with the evolution of newer sequencing technologies this sensitivity should increase on the sequencing platform, but we are reporting our experience using the technology of the time. We have revised our abstract in the revised manuscript and have mentioned only in the results section about this marginal 6% increase in detection of transcripts with the cut-offs used.

2) Similarly, the comparison of differential gene expression is also based on arbitrary cutoff values: 2-fold change in exon array and 4-fold change in RNA-Seq. At least the same fold change cutoff value should be used for the two platforms unless there is a strong reason not doing so.

There is indeed a strong reason for not using the same cutoff fold change for both methods. It is well known that microarray tends to underestimate the true fold change of expression. We again observe this fact in Figure 4 and Figure 6, with RNA-seq reporting generally twice the log fold change compared with Exon array. In order to perform an unbiased comparative analysis of the two platforms and knowing that the microarrays have a compressed fold change, we chose two different cutoff values to select differentially expressed transcripts, guided by the observed
compression in fold-change in the Exon array data. We selected transcripts using different criteria for each method, specifically, requiring at least a 2-fold change in microarray data or a greater than 4-fold change in RNA-seq data.

3) I found the discussion section too long and poorly organized. It should be substantially shortened to make it more readable. Some of the detailed discussions may go to the sections of results and methods. Subsections and bullet points may be used for better organization of the text.

As suggested by the reviewer we have tried to shorten and organize the discussion section in the revised manuscript.

Minor Essential Revisions
1) The two level, one-way ANOVA for testing differential expression applied to exon array and RNA-Seq data is essentially equivalent to two sample t-test, which is a more widely used test for differential expression.

We have now referred to the oneway anova as a two sample t-test as suggested by the reviewer in the revised manuscript.
It now reads as follows:
A two sample t-test (sickle cell N₁=6, control N₂=3) was used on the normalized, transformed data to test for differential expression.

2) On page 9, the meanings of all the symbols in the ExonANOVA model need to be explained.

We have included the meaning for the symbols in the revised manuscript

3) The method used for FDR estimation needs to be explained.

We followed the Benjamini and Hochberg method for FDR correction. We have included this reference in the revised manuscript.

4) On page 10 it was mentioned that there are ~15.1M reads per sample. However, on page 21, the number is 1~1.5M mappable reads. Why is there such a big difference?

We apologize for this error. We observed that out of the ~ 15 million reads, 82-84% were mappable. This has been corrected in the revised manuscript. It reads as follows:
RNA-seq analysis generated an average of 83% mappable reads from the whole blood samples after poly(A)+ enrichment.

5) All the analyses and comparisons are based on the RNA-Seq data generated using the Illumina GA-II platform, with ~15.1M 36bp reads per sample. However, the sequencing technology improves rapidly. The latest Illumina HiSeq 2000 platform can generate ~80M 100bp reads per lane, which will totally change the
comparisons between RNA-Seq and exon arrays regarding cost, CoV and sensitivity. Further, since multiple samples can be sequenced in one lane using multiplexing, even the starting material can be much fewer for RNA-Seq. All of these need to be discussed.

We fully agree with the reviewer that the sequencing technology has improved quite a lot since GAII with respect to cost, input material and reads. We started this project 2 years back when GAII was widely used for RNA-seq.

We have added the following statement in the revised manuscript:
With the rapid evolution of NGS instruments and library preparation methods with multiplexing barcodes, longer read lengths and large number of paired end reads associated with reduced cost per lane is highly feasible in the near future.

6) It was mentioned on page 20 that a few hundred nonagrams of total RNA is required for RNA-Seq, but on page 22 the number is 1.5 micrograms.

We have modified this statement in the revised manuscript to read as below
Although current technology for transcriptome sequencing requires at least 100 ng total RNA (tens of thousands of cell equivalents), along with additional enrichment steps to select for poly(A)+ RNA and/or to reduce the content of ribosomal RNA (rRNA) prior to NGS library construction, to minimize the loss of input material researchers tend to start with a minimum of 1 microgram total RNA.

7) On page 20 it was mentioned that reads mapped to multiple genes and paired reads complicate the mapping process. The latest mapping softwares can handle long reads (>50bp) and paired reads very efficiently and multi-mapped reads can be reduced to a very small percentage.

We agree that with the substantial improvements in mapping software, this process has become efficient now and this has been added to the revised manuscript.

8) What is the blue solid line in Figure 2?

The blue line represents the median expression levels of transcripts on both the platforms. This has been added to the figure legend

9) In Figure 8, the region between the novel exon and exon 4 also show some changes between SCD and controls. Is it another novel exon or a longer version of exon 4?

We have carefully examined the region between the novel exon 4a and exon4 that show some changes in SCD and found it difficult to determine anything conclusive based on the small number reads in that region (55,067,610 -55,067,875). We have investigated that region and have illustrated in the figure below. We cannot definitively interpret the reads in that region, but it could be an extension of Exon 4, observed primarily in the SCD patients. Three gapped reads
suggest the possibility of splicing, joining an extended Exon 4 with the novel Exon 4a, but the evidence is not sufficient for a conclusion, we feel.

<table>
<thead>
<tr>
<th>Exon 4a</th>
<th>Exon 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(chrX:55,067,519-55,067,610)</td>
<td>(chrX:55,067,876-55,067,998)</td>
</tr>
</tbody>
</table>

control

SCD