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

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi://10.1186/1471-2350-10-149</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:8000930">http://nrs.harvard.edu/urn-3:HUL.InstRepos:8000930</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>


Differentially expressed alternatively spliced genes in Malignant Pleural Mesothelioma identified using massively parallel transcriptome sequencing

Lingsheng Dong1, Roderick V Jensen2, Assunta De Rienzo1, Gavin J Gordon1, Yanlong Xu3, David J Sugarbaker1 and Raphael Bueno*1

Address: 1The Thoracic Surgery Oncology Laboratory and the Division of Thoracic Surgery, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA, 2Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA and 3Enterprise Research IS and Informatics, Brigham and Women’s Hospital and Partners Healthcare Inc, Boston, MA, USA

Email: Lingsheng Dong - ladong@partners.org; Roderick V Jensen - rjensen@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

* Corresponding author

Abstract

Background: Analyses of Expressed Sequence Tags (ESTs) databases suggest that most human genes have multiple alternative splice variants. The alternative splicing of pre-mRNA is tightly regulated during development and in different tissue types. Changes in splicing patterns have been described in disease states. Recently, we used whole-transcriptome shotgun pyrosequencing to characterize 4 malignant pleural mesothelioma (MPM) tumors, 1 lung adenocarcinoma and 1 normal lung. We hypothesized that alternative splicing profiles might be detected in the sequencing data for the expressed genes in these samples.

Methods: We developed a software pipeline to map the transcriptome read sequences of the 4 MPM samples and 1 normal lung sample onto known exon junction sequences in the comprehensive AceView database of expressed sequences and to count how many reads map to each junction. 13,274,187 transcriptome reads generated by the Roche/454 sequencing platform for 5 samples were compared with 151,486 exon junctions from the AceView database. The exon junction expression index (EJEI) was calculated for each exon junction in each sample to measure the differential expression of alternative splicing events. Top ten exon junctions with the largest EJEI difference between the 4 mesothelioma and the normal lung sample were then examined for differential expression using Quantitative Real Time PCR (qRT-PCR) in the 5 sequenced samples. Two of the differentially expressed exon junctions (ACTG2.aAug05 and CDK4.aAug05) were further examined with qRT-PCR in additional 18 MPM and 18 normal lung specimens.

Results: We found 70,953 exon junctions covered by at least one sequence read in at least one of the 5 samples. All 10 identified most differentially expressed exon junctions were validated as present by RT-PCR, and 8 were differentially expressed exactly as predicted by the sequence analysis. The differential expression of the AceView exon junctions for the ACTG2 and CDK4 genes were also observed to be statistically significant in an additional 18 MPM and 18 normal lung samples examined using qRT-PCR. The differential expression of these two junctions was shown...
to successfully classify these mesothelioma and normal lung specimens with high sensitivity (89% and 78%, respectively).

**Conclusion:** Whole-transcriptome shotgun sequencing, combined with a downstream bioinformatics pipeline, provides powerful tools for the identification of differentially expressed exon junctions resulting from alternative splice variants. The alternatively spliced genes discovered in the study could serve as useful diagnostic markers as well as potential therapeutic targets for MPM.

**Background**

Alternative splicing of pre-mRNA is a post-transcriptional process that occurs in approximately 70% or more of all human genes presumably to increase the diversity of the transcriptome and proteome [1]. This process is tightly regulated during development and is different among different tissue types. Inherited and acquired changes in pre-mRNA splicing patterns have been reported in many human malignancies, including colorectal cancer [2], ovarian cancer [3], breast cancer [4], and lung cancer [5,6]. In some cases, specific alternative splicing variants have been proposed as potential clinical markers for cancer diagnosis [7,8] and prognosis [8] as well as function in the role of therapeutic targets [9].

To-date, several groups have reported their efforts to identify novel alternatively spliced genes by analyzing the public available Expressed Sequence Tag (EST) databases [10-12]. Microarray based methods have also been developed to characterize cancer related alternative splicing events using exon arrays and exon junction arrays [13]. Recently, we used whole-transcriptome shotgun pyrosequencing [14] to characterize the transcriptome profiles of individual MPM patient samples and controls obtained at surgery [15]. Approximately 1.6 Gb of sequence data was obtained from ~16 million sequencing reads of ~105 bp average length. We hypothesized that differentially expressed alternative splicing events might be observed in these sequencing datasets.

In the present study, we demonstrate i) that cancer linked alternative splicing profiles can be identified in second generation transcriptome sequencing data *in silico* by mapping all sequence reads onto previously observed exon junction sequences, ii) that this approach can be used to identify expressed exons that are not present in the RefSeq mRNA database at NCBI, and iii) that specific exon junctions can be used to discriminate among MPM and normal samples.

**Methods**

**Sequencing data and patient samples**

Whole-transcriptome pyrosequencing [14] was previously described for 4 MPM tumors, 1 lung cancer and 1 normal lung using the Roche/454 sequencing system [15]. For this investigation, we used subset of those samples including all 4 MPM samples and one normal sample. Briefly, for each sample, >260 Mb of transcriptome sequence were obtained by shotgun, clonal pyrosequencing using Roche/454 technology. This subset consisted of 13,274,187 total sequencing reads of 105 bp average length. Overall, 98% of the reads matched to human RNA, DNA, and mitochondrial DNA sequences. In each sample, ~15,000 known RefSeq genes were detected by at least one read.

Additional 36 human discarded tissue specimens (n = 18 MPM, n = 18 normal lung) were acquired from our institution's Tumor Bank and used in compliance with a research protocol approved by the Institutional Review Board at Brigham and Women's Hospital.

**AceView transcriptome database**

AceView is a public database provided by the National Library of Medicine which provides a comprehensive annotation of transcripts and genes based on data from GeneBank, dbEST and RefSeq. This database was reported to offer a richer view of human transcriptome with 3 to 5 times more high-quality transcript forms than the publicly available databases (e.g., UCSC known genes, RefSeq or Ensemble) [16]. The AceView database of human transcribed sequences (‘Aug05’) and genome coordinate file (‘hg17, May 2004’) in GFF format (General Feature Format, is a format for describing genes and other features associated with DNA, RNA and Protein sequences) were obtained from the AceView website [16] and used to define the sequences and locations for 151,486 unique exon junctions.

**Annotation of AceView exon junctions and creation of virtual probes**

A customized Active State Perl script was used to index all possible exon junctions for each AceView transcript according to the genomic coordinates of the 3' end of the donor exon and the 5' end of acceptor exon that define the exon junction. In Aceview database, there are often multiple transcripts for each gene, so exon junctions with the same (or similar) chromosomal 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 with smaller than 5 bp differences. For each exon junction, a 200 bp transcript segment including 100 bp on each side of the exon junction point was extracted from the longest AceView transcript which contains the exon junction to create a virtual probe sequence for the exon junction. The name of the longest AceView transcript ID containing the junction and exon junction location in the transcript was used as a unique identifier for the exon junction and for the associated virtual probe sequence. For example, the exon junction 'ACTG2.aAug05.574' defines a junction that occurs in the ACTG2 gene in Aceview transcript 'a' in the 'Aug05' database at nucleotide position 574. These probes were also linked to relevant NCBI accession numbers, whenever possible.

Mapping of 454 transcriptome reads onto virtual probes

The alignment code was run on the Partners Healthcare System High Performance Computing Cluster based on the Microsoft HPC Server 2008 Operating system. NCBI blastall was used to map all the 454 transcriptome reads onto the virtual probe sequences with the following parameters: blastall -p blastn -a 2 -F "m D" -e 1e-10 -m 8. A customized Perl script was used to query the blast results to quantify the number of reads mapping to each virtual probe and cover the junction point with at least 5 bp on both side, at least 35 bp of continuous alignment, and at least 90% identity to the virtual probe sequence. Figure 1 illustrates the theoretical alignment of the ~100 bp long transcriptome reads with the different exon junctions spanned by virtual probes.

Filtering of mapped exon junctions and generation of an exon junction expression index (EJEI)

To filter out the low-expressed exon junctions across all samples, we required that each unique exon junction map to at least 5 sequencing reads in all patients combined. To filter out the low-expressed genes in any single sample, we also required that at least 4 sequencing reads map to any exon junction in a given gene in each sample. Because there are similar total numbers of reads among the 5 discovery samples [15], additional normalization to the total read count in each sample was unnecessary.

To correct for systematic differences in gene expression levels between the different samples, an exon junction expression index (EJEI) was calculated for each unique exon junction in each sample. The EJEI is 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. This is similar to the Splicing Index reported in ArrayAssist software [17] for analyzing the exon level data on the Affymetrix Exon arrays, however, here we are counting the relative frequency of the exon junctions themselves.

To make the EJEIs comparable among different exon junctions of different genes, they were further normalized by the sum of the EJEI in all patients for the same exon junction. To identify highly expressed exon junctions in MPM relative to normal lung, we generated a ranked list of EJEIs by calculating the differences between the minimum normalized EJEI in the 4 MPM samples and the normalized EJEI for the same exon junction in the normal lung sample. To identify highly expressed exon junctions in normal lung relative to MPM, we similarly generated a ranked list of EJEIs by calculating the difference between the normalized EJEI in the normal lung sample and the maximum normalized EJEI for the same exon junction in the four MPM samples.

Quantitative Real Time RT-PCR

Total RNA (2 μg) was reverse-transcribed into cDNA using Taq-Man Reverse Transcription reagents (Applied Biosystems, Foster City, CA) and quantified using all recommended controls. Primers (synthesized by Invitrogen Life Technologies) were used at a final concentration of 800 nM in the reaction mixture and are shown in additional file 1. Primer sequences were blasted onto NCBI human
Refseq RNA database to confirm that the primers will not amplify other genes. PCR amplification and SYBR-Green-based detection of CDNA was performed using a Stratagene MX3000P machine and default thermal cycling parameters. No-template (i.e., negative) controls that contained water instead of template were run in multiple wells on every reaction plate. An automatically calculated melting point dissociation curve generated after every assay was examined to ensure the presence of a single PCR species and the lack of primer-dimer formation in each well. (The detail method is in Additional file 2).

**Statistical analysis**

Average expression levels of exon junctions were compared between MPM and normal samples using a two-tailed student's t-test. Classification accuracy was assessed using a 2 × 2 contingency table and Fisher's exact test. All comparisons were considered statistically significant if \( P < 0.05 \). All comparisons were conducted using S-PLUS software [18]. Post-hoc power analysis was conducted using G*Power [19].

**Results**

In total, we indexed 151,486 unique AceView exon junctions to create one virtual probe per unique exon junction. Mapping of the 13,274,187 transcriptome reads against these 151,486 virtual probes revealed 70,953 exon junctions that were covered by at least one sequence read in at least one of the 5 samples using our criteria. The distribution of these 70,953 exon junctions was summarized in Additional file 3.

We used ranked normalized EJEI values to identify the most differentially expressed exon junctions between the four MPM samples and the one normal lung sample. All the exon junctions with EJEI difference bigger than 0.1 were listed examined (see Additional file 4). We identified 76 highly expressed unique exon junctions in MPM and 32 uniquely expressed exon junctions in the normal lung sample (see Additional file 5). We chose to examine in the 5 discovery samples a total of 10 exon junctions: the top 3 ranked exon junctions expressed at relatively higher levels in normal lung and the top 7 ranked exon junctions expressed at relatively higher levels in the MPM samples using real time quantitative RTPCR (see table 1 and additional file 6 for the 10 candidate exon junctions). We arbitrarily weighted our list towards exon junctions over-expressed in MPM tumor because we found more highly expressed exon junctions in MPM (76) than in normal lung (32) and sought to increase our chances of identifying potentially diagnostic alternatively spliced transcripts.

We were able to validate that all predicted exon junctions were actually observed in the transcriptome when the samples were independently analyzed with qRT-PCR. Six of the 10 candidate exon junctions analyzed with qRT-PCR have a level of expression in the normal lung sample within the range of expression of the 4 MPM tumors. As far as the differential expression was concerned, we required that the individual expression level of each exon junction in at least 2 of the 4 MPM tumors (and the average for all 4 MPM tumors) be in the predicted direction relative to the level in the normal lung sample. We were able to validate the differential expression of 8 of the 10 exon junctions. Differential expression of exon junctions 'EMP2.shed3.aAug05' (over-expressed in normal) and 'DNAJB11.aAug05' (over-expressed in tumor) observed using 454 sequence data could not be confirmed by qRT-PCR. In both cases, expression of the exon junction was detected at similarly high levels in all samples using qRT-PCR (see table 1 and additional file 6 for detail comparison between qRT-PCR results and transcriptome sequencing results).

Among the 8 validated differentially expressed exon junctions, we selected ‘ACTG2.aAug05’ and ‘CDK4.aAug05’ for further evaluation in a set of 18 additional MPM samples and 18 normal lung samples using qRT-PCR. These two exon junctions were selected because they were associated with the greatest magnitude of differential expression in the predicted direction for each of the 5 discovery samples using both 454 read counts and qRT-PCR. We found that the average expression levels of both ‘ACTG2.aAug05’ and ‘CDK4.aAug05’ were statistically significantly different in the predicted direction between normal and tumor samples (\( P = 0.00054 \) and \( P = 0.00039 \), respectively, Figure 2A and 2B). ‘ACTG2.aAug05’ was associated with a greater range of expression levels across individual samples (\( \sim \)100-fold difference) than ‘CDK4.aAug05’ (\( \sim \)3-fold difference).

Using the median value of relative exon junction expression levels in all 36 samples as a cut-off, we found that ‘ACTG2.aAug05’ expression correctly classified 89% (16/18 tumor, 16/18 normal) of the samples (\( P = 5.2 \times 10^{-6} \), Figure 2C), whereas ‘CDK4.aAug05’ correctly classified 78% (14/18 tumor, 14/18 normal) of the samples (\( P = 2.2 \times 10^{-3} \), Figure 2D). All 36 samples were correctly classified by at least one exon junction. Using G*Power, post-hoc power analysis show sufficient sample size with Power (1- \( \beta \) err prob) = 1.000 for ACTG.aAug05 and Power (1- \( \beta \) err prob) = 0.997 for CDK.aAug05.

**Discussion**

There have been previous reports of alternatively spliced cancer-associated genes that could serve as diagnostic and prognostic markers as well as guide to potential therapeutic targets [10-12]. These studies used a variety of tools including analysis of publicly available databases and discovery using microarrays, both of which utilize high-
<table>
<thead>
<tr>
<th>Normal/Tumor</th>
<th>Gene</th>
<th>Typical AceView sequence</th>
<th>Exon junction location in typical AceView sequence</th>
<th>Normalized EJEI</th>
<th>RT-PCR EJEI result</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>EMP2</td>
<td>EMP2.shed3.aAug05</td>
<td>1003</td>
<td>Meso 1 0.000</td>
<td>Meso 2 0.000</td>
</tr>
<tr>
<td>normal</td>
<td>CYFIP1</td>
<td>CYFIP1.IAug05</td>
<td>49</td>
<td>Meso 1 0.000</td>
<td>Meso 2 0.000</td>
</tr>
<tr>
<td>normal</td>
<td>ACTG2</td>
<td>ACTG2.aAug05</td>
<td>574</td>
<td>Meso 1 0.062</td>
<td>Meso 2 0.000</td>
</tr>
<tr>
<td>tumor</td>
<td>C1QA</td>
<td>C1QA.aAug05</td>
<td>1449</td>
<td>Meso 1 0.221</td>
<td>Meso 2 0.264</td>
</tr>
<tr>
<td>tumor</td>
<td>MRPL51</td>
<td>MRPL51.bAug05</td>
<td>542</td>
<td>Meso 1 0.228</td>
<td>Meso 2 0.301</td>
</tr>
<tr>
<td>tumor</td>
<td>TXNRD1</td>
<td>TXNRD1.aAug05</td>
<td>1333</td>
<td>Meso 1 0.198</td>
<td>Meso 2 0.220</td>
</tr>
<tr>
<td>tumor</td>
<td>CDK4</td>
<td>CDK4.aAug05</td>
<td>1246</td>
<td>Meso 1 0.267</td>
<td>Meso 2 0.223</td>
</tr>
<tr>
<td>tumor</td>
<td>hfl-B5</td>
<td>hfl-B5.aAug05</td>
<td>1123</td>
<td>Meso 1 0.225</td>
<td>Meso 2 0.226</td>
</tr>
<tr>
<td>tumor</td>
<td>COL3A1</td>
<td>COL3A1.aAug05</td>
<td>1680</td>
<td>Meso 1 0.184</td>
<td>Meso 2 0.363</td>
</tr>
<tr>
<td>tumor</td>
<td>DNAJB11</td>
<td>DNAJB11.aAug05</td>
<td>1612</td>
<td>Meso 1 0.265</td>
<td>Meso 2 0.281</td>
</tr>
</tbody>
</table>

*Note: this is the brief version of additional file 6
throughput methods to characterize cancer related alternative gene splicing [13,20]. More recently, a high throughput RT-PCR method has been used to measure the differential expression of 3,327 alternative splicing events in 600 cancer-related genes for ovarian and breast cancer [3,21]. The present study provides one of the first attempts to show how whole transcriptome sequencing using massive parallel sequencing technology can be used to simultaneously profile as many as 151,486 alternative splicing events in cancer patient samples for all the human genes cataloged in the comprehensive AceView database of transcribed sequences. In our previously study [15], we used the Roche/454 next-generation sequencing system for whole transcriptome pyrosequencing of cDNA samples for 4 MPM tumors, 1 lung adenocarcinoma and 1 normal lung. We generated between 2.5 and 2.9 million shotgun sequencing reads for each sample of average length ~105 bp and focused on the identification of novel Single Nucleotide Polymorphisms (SNPs) in the expressed sequences. Herein, we hypothesized that these same transcriptome reads could be used to detect alternatively spliced genes in the patient samples by mapping all the read sequences onto known exon junction sequences as virtual probes. Using a subset of the same dataset [15], we have developed a software pipeline to quantify the expression levels of exon junctions by counting the number of reads that match to each exon junction to identify cancer related alternative splicing pattern.

In this study, we have identified several genes expressing alternatively spliced transcripts at different levels in the tumors and in the matching normal lungs. Many of these genes have been previously implicated in cancer. ACTG2 contains one 5’ un-translated exon and 8 coding exons

**Figure 2**

**Differential expression of specific exon junctions in MPM and normal lung samples.** Average expression levels of exon junctions in ACTG2 (A) and CDK4 (B) differ statistically significantly between MPM and normal lung samples. Individual expression levels of exon junctions in ACTG2 (C) and CDK4 (D) can be used to classify MPM and normal tissue samples using the median value as a cut-off. Error bars, SEM. Arrows indicate misclassified samples.
spanning 27 kb [22]. This gene has 7 known splicing variants shown in AceView website [16], but none of them has been directly related to cancer. An expression microarray analysis performed on a derivate breast cancer cell line resistant to cisplatin showed that ACTG2 expression increases in the chemotherapy resistant cell line compared to the normal indicating that it may be associated with cisplatin resistance [23]. In another study, ACTG2 expression was identified as cadmium-responsive [24]. The authors concluded that repressed expression of ACTG2 following cadmium exposure may contribute to the cell cycle arrest. CDK4 is a well-known for its role in cancer [25]. It is a cyclin-dependent kinase-4 involved in the cell cycle and can both start and stop the cell cycle in response to proliferative or anti-proliferative signals. It has 13 known splice variants according to AceView website [16]. Several reports have already linked CDK4 expression to mesothelioma [26,27]. However, this was the first study showing that different CDK4 splice variants have differential expression levels in MPM and matching normal lung.

The differentially expressed splice variants for ACGT2 and CDK4 were specifically chosen for further examination using qRT-PCR in additional 18 MPM and matched normal lung samples. This analysis suggested that the differentially expressed splice variants may provide reliable markers for disease and be used to classify the samples with high sensitivity and specificity.

Several of the other genes that appeared to exhibit differentially expressed splice variants in the present study have also been implicated in cancer and would be worthy of further study. CYFIP1 has been shown to be a novel tyrosine kinase substrate in a breast cancer model [28]. Interestingly, the differentially expressed transcript (CYFIP1.fAug05) is not included in NCBI Refseq sequence database. It has been suggested that COL3A1 could be a potential diagnostic marker for large B-cell lymphoma (DLBCL) as is shows statistically significant different expression between DLBCL and follicular lymphoma [29]. In addition, COL3A1 expression has been related to resistance to platinum drugs in ovarian cancer [30]. TXNRD1 is a key enzyme in the regulation of the intracellular redox environment [31]. Transcription of TXNRD1 involves alternative splicing, leading to a number of transcripts. In particular, expression of the TXNRD1_v3 transcript has been found in several cancer cell lines [32]. Recently, its locus has been associated with advanced colorectal adenoma by epidemiologic and animal studies [33].

In this pilot study, we were not able to find statistical difference between differentially expressed exon junctions because of the small sample size (only 4 MPM and 1 normal lung sample), neither we were able to compare the results among different platforms. In addition, not all exon junctions predicted to be differentially expressed proved to be so to the same extent when examined with the qRT-PCR. This is likely due the limited number of specimens examined. Furthermore, the signal to noise ratio may have been over-amplified as not to miss any potential differentially expressed candidates. After all, the EJEI was designed to magnify the differential expression and is not in the same order of magnitude as the actual expression of the exon junctions. Other potential limitations may be due to sequencing artifacts, insufficient sequencing depth, SNPs near the EJ and incomplete database for possible exon junctions. These limitations may be avoided using other next-generation sequencing platforms, such as Helicos True Single Molecule Sequencing without amplification, Illumina or SOLiD [34], or by increasing the sequencing depth. Nevertheless, at least 2 of the top 10 exon junctions prioritized for analysis remained differentially expressed in most tested specimens supporting the utility of this approach.

The present study provides an example of a possible application of advanced sequencing technologies in cancer research. The current sequencing technologies are now capable of generating millions of shotgun transcriptome reads in a matter of days. For example, the latest Roche/454 GS FLX Titanium system generates over 1,000,000 reads of average length 400 bp in one 10-hour sequencing run providing orders of magnitude improvement in speed and cost over conventional Sanger-based sequencing. One of the great virtues of the shogun transcriptome sequencing process is that there is no need to impose any bias for known genes, exons, or splice-junctions as required for example with exon microarrays. As these technologies continue to improve their throughput and read lengths and lower their costs, they promise to revolutionize gene expression analysis by simultaneously providing information about expression levels, transcript variants, and SNPs.

The greater challenge for the successful application these technologies for our understanding of health and disease will be the analysis and interpretation of the data. Here we have introduced a data analysis pipeline to map 13,274,187 transcriptome reads from patient cDNA samples onto 151,486 known splice junctions cataloged in the comprehensive AceView database of transcribed sequences. However, in short order we can expect that the competing next generation sequencing technologies will be generating several orders of magnitude more transcriptome and genomic sequencing data for a wide variety of human diseases and cancer. Further advances in the bioinformatics analysis of this flood of data are clearly required, for example, to map sequencing reads directly to human genome to identify novel transcribed sequences.
and genes, alternative exons and splicing events, and possible gene fusions in patient samples

**Conclusion**
We demonstrated that whole-transcriptome shotgun sequencing and downstream bioinformatics pipeline can be powerful high-throughput tools for the identification of differentially expressed exon junctions resulting from alternative splicing variants. The alternative transcripts discovered in the study could be useful as diagnostic markers as well as potential therapeutic targets for MPM.

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
LD carried out the Perl programming, data analysis and PCR experiments. LD, RJ, GJG, ADR and RB drafted the manuscript. LD, RJ and RB conceived of and designed the study. RB and DJS provided access to the transcriptome dataset and provided critical manuscript review. YX provided computing support for the Partners High Performance Computing Cluster facilities. All authors read and approved the final manuscript.

**Additional material**

Additional file 1
PCR Primers used to quantify expression levels of candidate exon junctions
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S1.DOC]

Additional file 2
Methods to calculate exon junction expression index (EJEI) using RT-PCR
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S2.DOC]

Additional file 3
Distribution of the 70,953 exon junctions mapped by at least one transcriptome read in at least one patient sample among 5 discovery samples
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S3.XLS]

Additional file 4
76 highly expressed exon junction in MPM with EJEI difference > 0.1 and their mapped read numbers, EJEI, normalized EJEI and EJEI difference between normal lung and MPM samples
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S4.XLS]

Additional file 5
32 highly expressed exon junction in normal lung with EJEI difference > 0.1 and their mapped read numbers, EJEI, normalized EJEI and EJEI difference between normal lung and MPM samples
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S5.XLS]

Additional file 6
Mapping results for the 10 candidate differentially expressed exon junctions, their exon junction expression index (EJEI), EJEI difference between mesothelioma and normal samples and the RT-PCR results
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2350-10-149-S6.XLS]

**Acknowledgements**
This work was supported by grants to RB from the NCI (CA120528), the Vancouver Foundation and the International Mesothelioma Program at BWH. The authors would like to thank Danielle Thierry-Mieg and Jean Thierry-Mieg for their help with the AceView database.

**References**


Submit your manuscript here: http://www.biomedcentral.com/info/publishing_adv.asp

Publish with BioMed Central and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

BioMedcentral

Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2350/10/149/prepub