Abstract

Alternatively spliced exons are observed to be either conserved across the species or are species specific. Conserved alternatively spliced events are believed to be functionally important. In this study, the aim is to determine a method which performs well in identifying events showing significant change in splicing among the two groups of tissues, and are also co-regulated across the species.

Four methods implementing ANOVA have been applied on the log-transformed and quantile normalized genome-wide human and mouse micro-array splicing index data to determine the conserved and significant exons. Negative controls were designed to assess the quality of performance of method. The gene enrichment of exons picked as significant by each method was determined through GO analysis.
Introduction

Alternative splicing is an important biological process observed in eukaryotes. Through alternative splicing, exons from a gene transcript are reconnected in multiple ways during RNA splicing resulting in different mRNAs [32]. These different mRNAs might translate into different protein isoforms, thus generating multiple proteins from a single gene. Some of these mRNAs are subjected to non-sense mediated decay (NMD), a mechanism which eliminates mRNAs containing premature termination codons thus, limiting the synthesis of abnormal proteins. Sequencing of the whole genome of organisms and sequencing their corresponding ESTs (expressed sequence tags) revealed that there is no direct correspondence between the size of the genome and the number of protein coding genes. Lack of alternative splicing mechanism in less complex organisms like prokaryotes and its occurrence in multicellular organisms like eukaryotes explains the importance of this phenomenon in such complex organisms with fewer number of protein coding genes. Among the total number of genes in human genome, alternatively spliced genes are estimated to be between 40% and 60% of all genes [1, 3, 34, 4] and up to three-quarters of multiexonic genes [16].

There are five different modes of alternative splicing mechanisms observed [7, 2, 30]. They are:

a) exon skipping or cassette exons: where, an exon is either retained or spliced out of the primary transcript.
b) mutually exclusive exons: one of the two exons is retained in the mRNA after splicing but not both.
c) alternative donor site exons: where, alternative 5’ splice junction (donor site) is used, changing 3’ boundary of the upstream exon.
d) acceptor site exons: where, alternative 3’ splice junction (acceptor site) is used, changing 5’ boundary of the downstream exon.
e) intron retention: where an intron is retained in the mRNA.

Splicing factors are observed to play a major role in the regulation of alternative splicing [14]. The finely tuned balance of these splicing factors determines how each molecule behaves and how a transcript is ultimately spliced [19]. Regulation of alternative splicing is observed to be crucial in the physiology of the heart, skeletal muscle, brain, and other tissues, and mis-regulation of AS is associated with human disease [15, 18, 23].
Since the discovery of alternative splicing in 1977 by Sambrook in adenoviruses [13], several methods have been developed to detect alternative splicing events, and to analyze their role and importance in the physiology of several organisms. Microarray based detection of alternative splicing events is one among the high-throughput technologies developed to detect and analyze the genome-wide alternative splicing events. Several types of microarrays like fiber-optic arrays, glass arrays, tiling arrays, exon-exon junction arrays, GeneChip Human Exon 1.0 ST array, splice-sensitive arrays and many more have been designed so far to detect alternative splicing events. Although micro-array technology has emerged as a powerful tool in the past few years in detecting genome-wide alternative splicing events, major problem lies in the poor reproducibility of results using other methods such as RT-PCR [12].

Alternative splicing events are observed to be either species specific or conserved across the species. Conserved alternative splicing events across the species are believed to be functionally important. Species specific alternative splicing events are often subjected to non-sense mediated decay. Comparisons of human and mouse transcript sequences have revealed that <20% of alternative splicing events have been conserved [5, 28, 27, 24, 11] during the ~80 to 90 million-year interval separating these species, suggesting an important role for alternative splicing in the evolution of mammalian species [9, 33]. In human and mouse, exons present in the majority of transcripts are observed to be most conserved, with less than 3% of alternatively spliced exons observed only in one species, whereas 75% of exons present in the minority of transcripts are observed to be species specific [5]. Conservation of a specific pattern of AS over the ≈ 90 million years since divergence of the mouse and human lineages provides strong evidence of biological function. Therefore, defining the set of AS events conserved between human and mouse is of primary interest in efforts to understand the biological importance of splicing regulation [11]. To detect the AS genes conserved across species, different types of microarrays such as species-specific arrays, single species arrays, multi-species arrays have been developed [20]. Genes conserved across the species are usually expected to show consistent expression patterns across species, but it has been shown that some of these conserved genes show expression divergence [31].

In the current study, focus is to identify conserved alternative splicing exons across human and mouse using four different statistical methods applied on genome-wide human and mouse splicing index data obtained from splice-sensitive microarrays. The data consists of splicing index values for each exon obtained from muscle and brain tissues.
The aim of this study is to identify a method or methods which is or are efficient in picking up significant as well as conserved alternative splicing events across the two species human and mouse. Each method implements statistical method ANOVA applied using statistical software R [26], to determine a p-value for each exon. The p-value implies how well an exon shows difference in splicing across the two tissues muscle and brain. The p-value is transformed into \(-\log(10)\) p-value, and a FDR of 20%, 15%, and 10% is applied to determine the exons which pass that \(-\log(10)\) p-value score. The statistical method ANOVA is chosen to determine the p-value because of the ease through which the analysis could be expanded to multiple-species level.

**Methods**

In the current study, to analyze the conserved alternative splicing events across human and mouse, four methods are applied on the data. Each method takes different approach in combining the data. Only the exons showing consistent change in splicing in both the species are considered for this study. All these methods make use of statistical procedure ANalysis of VAriance (ANOVA) and were run on statistical software R [26].

**Significance of ANOVA** : From the human and mouse datasets used in this study, data from two tissues – brain and skeletal muscle is used for each exon. For this study, statistical procedure ANOVA is chosen because of the ease through which the analysis could be expanded to multiple-species level. Through ANOVA, a p-value is assigned to each exon based on its differential expression among the two tissues. Lower p-value implies that an exon shows significant difference in splicing in one tissue compared to the other.

Further description of the four methods applied in this study is given below :

**Method 1 : Independent Species ANOVA Analysis**

In this method, the statistical procedure ANOVA is applied initially on human and mouse orthologous exons data individually. The p-values are transformed into \(-\log(10)\) p-value. Product of the human and mouse negative log-transformed p-values is determined.
Method 2: Species Blind ANOVA Analysis

For this analysis, human and mouse brain data is grouped together and similarly human and mouse muscle data is grouped together. In this way an orthologous human and mouse exon has six data values for brain tissue and six values for muscle tissue. ANOVA is applied on this combined data.

Significance: In this analysis, the information of the species to which the data belongs to is hidden, hence the name “Species Blind Analysis”. Through this method, we expect to see more statistically significant results compared to “Independent Analysis”, since each exon now has more data and hence bigger sample size. Also, this method acts as a statistical control for “Species Aware Analysis”.

Method 3: Species Aware ANOVA Analysis

This method also makes use of the combined data, as described in “Species Blind Analysis”, but including the information of the species to which the data belongs to. The aim is to see if specification of this parameter “species” or “organism” yields more useful results about the difference in splicing not only with respect to brain and muscle tissue, but also with respect to the organism or species to which the data from brain and muscle belongs to.

Significance: The aim is to see if more meaningful results are obtained through this analysis compared to “Species Blind Analysis”, since, this analysis not only makes use of additional data but, the information of the species to which the data belongs to is also included. Since, the information of species is provided while determining pvalue for an exon via ANOVA, hence the name “Species Aware Analysis”.

Method 4: Meta-analysis

In this analysis, results obtained from “Independent Species ANOVA Analysis” are combined. Through “Independent Species Analysis” a pvalue is assigned to each human orthologous exon, and similarly a pvalue is assigned to each mouse orthologous exon via ANOVA. The pvalues of human and mouse orthologous exons are combined in this analysis. Here, the type of meta-analysis applied is “Symmuni” developed at Stuart lab [17]. Now, the human and mouse orthologous exon is assigned a new pvalue obtained through this meta-analysis.

Significance: Symmetric union of human and mouse orthologous exon pvalues is determined using “Symmuni”, with the aim to see if this analysis picks up greater number of significant exons compared to “Independent species Analysis”.
Data

In the current study, to analyze human and mouse alternative splicing events, following data sets were used.

a) Human and mouse gene expression data,
b) Human and mouse splicing index data and,
c) Human and mouse corresponding events data.

Human and Mouse Gene Expression Data

Human Gene Expression data consists of overall expression of 10642 genes obtained from brain (temporal lobe region), and skeletal muscle tissues. The data is obtained using Affymetrix microarrays at Affymetrix.

Mouse Gene Expression data consists of overall expression of 10609 genes obtained from brain, muscle, and several other tissues. The data is obtained using Affymetrix microarrays at UC Santa Cruz in Dr. Manuel Ares lab.

For the current study, genes which does not show significant expression in both brain and muscle are excluded. Further description is available in the section “Data used for the current study”.

Human and Mouse Splicing index Data

Human and mouse splicing index data is obtained using Affymetrix microarrays at Affymetrix, and Dr. Manuel Ares lab (UCSC) respectively. Each splicing index value corresponds to ratio of expression of each exon to overall gene intensity. Splicing index data is obtained from same tissue replicates as described for human and mouse overall gene expression data.

Human Splicing Index Data : Consists of data for 62406 exons obtained from brain, and skeletal muscle tissues.

Mouse Splicing Index Data : Consists of data for 6370 exons obtained from brain, skeletal muscle, and other tissue replicates.

Human & Mouse corresponding events Data

This dataset consists of human and mouse orthologous exon pairs, also referred to as “human and mouse mapped events”. These orthologus exon pairs are obtained through mapping mouse exon coordinates from the mouse mm8 genome assembly to
the human exon coordinates from the human hg18 genome assembly using the UCSC Genome Browser liftover utility. If each mouse exon overlapped with a human exon by at least 50%, the two exons were labeled as orthologous human and mouse exon pairs. Through this altogether 2520 human and mouse exon pairs were labeled as orthologous. Also, these events include only alternative splicing events (cassette exons, exons with alternative 3’ or 5’ splice sites, and mutually exclusive exons).

Data used for the current study

For the current study, data from brain and skeletal muscle tissues is used for both human and mouse since, human gene expression data is obtained from only those two tissues. Human brain data is from temporal lobe region, and in case of mouse data from cerebellum region of brain is is used for the current analysis. There are 3 replicates of cerebellum in mouse, hence 3 human brain replicates which cluster together were selected. Similarly, 3 skeletal muscle replicates which cluster together from human and mouse were chosen.

Selecting Genes
Gene expression data is log trasformed and quantile normalized. For each human and mouse gene expression values of 3 brain replicates and 3 muscle replicates which cluster together are chosen. Genes with expression greater than log2(15) compared to median of brain values, and median of muscle values are considered as significant genes. [NOTE : log2(15) is chosen as a background expression value, and genes showing expression above this value are considered as significant.]

Selecting Exons
Exons corresponding to the significant genes as described in section “Selecting Genes” are taken from the log-transformed, quantile normalized splicing index data. For each exon, data from same 3 brain replicates and from same 3 muscle replicates (considered for selecting genes) is extracted.

Selecting human and mouse orthologous Exons
The human and mouse orthologous exon datasets were extracted by comparing human and mouse exons corresponding to the significant genes, with human and mouse exons listed as orthologous in “Human and Mouse Corresponding Events Data”.
**Consistent Exons - Brain up-regulated and Muscle up-regulated exons**

Exons showing consistent change in splicing (change in same direction) in both species are labeled as consistent exons. These consistent exons are again classified into brain up-regulated and muscle up-regulated exons. Exons with high mean of brain splicing index values compared to muscle in both species are labeled as brain up-regulated exons, and similarly exons with high mean of muscle splicing index values compared to brain in both species are labeled as muscle up-regulated exons.

**Results & Discussion**

In the current study, four methods are applied on human and mouse orthologous consistent exons data. Each method take different approach to combine the data. All these methods make use of statistical procedure ANalysis of VAriance (ANOVA) and were run using statistical software R [26]. Splicing index data of human and mouse consist of expression values from brain and muscle tissue replicates. After normalization of data (see Data), only exons showing consistent change in splicing are considered. Genes corresponding to these consistent exons were determined, and most of the exons belong to different genes, very few exons are from the same gene. Evaluation of the performance of each method is done using several analyses.

I. Significant exons count.

II. Precision plots obtained from actual data vs negative controls.

III. Gene Ontology Analysis.

**Significant Exons count (at FDR of 20%, 15%, and 10%)**

For this analysis, human and mouse consistent exons data is randomly permuted, and 1000 such randomly permuted datasets are sampled using “Fisher-Yates-shuffling algorithm” [35]. At different p-value cutoffs, precision has been calculated from actual data vs 1000 randomly permuted datasets.

Where, Precision = estimated true positives / predicted true positives.

Predicted true positives = number of predicted true positives obtained from actual data.

and,

estimated true positives = (predicted true positives – avg. of false positives from permuted data).
Precision plots are generated in “R” using number of predicted true positives (on x-axis) vs precision (on y-axis) determined at different p-values. The precision plot is shown in Figure 1.

False Discovery Rate (FDR) = 1-precision. The p-value for each method is determined at different FDR thresholds. So, at a FDR threshold of 20%, one would expect to see 20% of false positives from the permuted data, or in other words 80% of true positives from the actual data. So, at a FDR threshold of 20%, precision is 80% and the p-value cutoff where the precision is 80% is determined. For each method we get a different p-value where precision is 80%.

At FDR thresholds of 20%, 15%, and 10% with precision of 80%, 85%, and 90% respectively, the p-value for each method is determined. The number of exons that pass the p-value cutoff at a given FDR threshold are determined for each method. The exons that pass the p-value cutoff are considered significant. Among these significant exons, brain up-regulated, and muscle up-regulated exons were also determined. The results are shown in Table 1.

<table>
<thead>
<tr>
<th>Considering mouse cerebellum region</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR Threshold</td>
</tr>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Independent</td>
</tr>
<tr>
<td>Species Blind</td>
</tr>
<tr>
<td>Species Aware</td>
</tr>
<tr>
<td>Meta-analysis</td>
</tr>
</tbody>
</table>

Table 1: Shows of number of exons picked as significant by each method at FDR of 20%, 15%, and 10%. The significant exons are again separated into brain-upregulated, and muscle up-regulated exons. The numbers in each cell represents total number of significant exons, brain-upregulated, and muscle up-regulated exons respectively.
Precision Plots from actual data

Figure 1: Shows precision plots for each method at different p-value cutoffs. The plot represents the precision with which a method picks number of predicted true positives at a given p-value cutoff. A method picking more number of predicted true positives with a better precision is expected to be performing well.
Precision Plots from Negative Controls

To compare and validate the performance of methods, several negative controls were designed. These negative controls consists of data which does not make sense biologically. The idea is to look at the quality of performance of each method when applied on actual data vs when applied on the negative controls. The tissue labels for the data are unmodified. In few negative controls, mouse brain or muscle data is substituted with some other mouse tissue replicates data which cluster far apart from both brain and muscle splicing index data. For human data, only brain and muscle replicates are used, since these are the two tissue groups from which data is available for human.

In each negative control, exons showing consistent change in splicing are determined, and 1000 randomly permuted datasets are generated from these consistent exons data. The four negative controls are:

1) Human (muscle, brain) vs Mouse (brain, muscle):
   Human brain data is swapped with its skeletal muscle data and vice versa. This swapped human data, and mouse unmodified data is used for this negative control.

2) Human (brain, muscle) vs Mouse (testis, muscle):
   In this negative control, mouse brain data is replaced with its testicular data, and the human data is unmodified. Brain data is believed to be highly conserved whereas, testicular data is not. Hence, testicular data is used for replacement.

3) Human (brain, muscle) vs Mouse (brain, ovaries):
   In this negative control, mouse muscle data is replaced with mouse ovaries data, rest of the mouse data and the human data is kept unmodified.

4) Human (brain, muscle) vs Mouse (testis, ovaries):
   In this negative control, both the mouse brain and muscle data is replaced with mouse testis and ovaries data. Human data is kept unmodified.

Precision plots are generated for each of the negative controls, and are shown in Figure 2(a, b, c, d).
Figure 2(a) : Shows precision plots for each method for Negative Control 1 – where, human brain and muscle tissues data is swapped, and mouse data is unmodified.
Figure 2(b) : Shows precision plots for each method for Negative Control 2 – where, mouse brain data is substituted with mouse testicular data. In this negative control, Human (brain, skeletal muscle) vs mouse (testis, skeletal muscle) are considered.
Figure 2(c): Shows precision plots for each method for Negative Control 3 – where, mouse skeletal muscle data is substituted with mouse ovaries data. In this negative control, Human (brain, skeletal muscle) vs mouse (brain, ovaries) are considered.
Figure 2(d) : Shows precision plots for each method for Negative Control 4 – where, mouse brain data is substituted with mouse testis data, mouse skeletal muscle data is substituted with mouse ovaries data. In this negative control, Human (brain, skeletal muscle) vs mouse (testis, ovaries) are considered.
**Gene Ontology (GO) Analysis**

The enrichment of exons/genes in a particular phenotype or pathway is determined through this GO analysis. GO analysis of human and mouse orthologous exon sets is done using “sets overlap” program developed at Stuart lab. This program identifies enriched phenotypes associated with the exons set and assigns a log(10) pvalue. A better pvalue represents enrichment of the exons/gene set in that particular phenotype. The program “sets overlap” assings pvalue to an exon set by determining the overlap between the exon set of interest and the total set of brain and muscle upregulated exons/genes. The phenotype to which the exon set is best associated with, gets the high score with least log(10) pvalue and is ranked at the top among the several phenotypes to which the exon set is predicted to be associated with.

For the current GO analysis:

- The exons determined as significant at a FDR of 20% for each method are taken as test set. These significant exons are separated into brain up-regulated and muscle up-regulated exons.
- The exons showing consistent change in splicing across both species are considered for the background set.
- GO enrichment of the significant exons is determined using sets_overlap program developed at Stuart lab.
- Few GO terms picked commonly by all 4 methods are shown in Table 2(a, b).

<table>
<thead>
<tr>
<th>GO results for significant brain up-regulated exons at FDR 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td><strong>GO_term</strong></td>
</tr>
<tr>
<td>membrane organization</td>
</tr>
<tr>
<td>cell cycle process</td>
</tr>
<tr>
<td>regulation of cytokine production</td>
</tr>
</tbody>
</table>
positive regulation of cell proliferation  |  0.2997  |  **0.0278**  |  0.2577  |  0.6747  
DNA binding  |  **0.0059**  |  0.2155  |  **0.0209**  |  **0.0166**  

Table 2a: Shows enrichment of muscle up-regulated exons in a particular pathway. Few of the GO terms picked commonly by all 4 methods are shown here. GO terms with significant p-values (<0.05) are considered as significant. Those p-values are highlighted and italicized.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GO_term</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phosphoprotein phosphatase activity</td>
<td>0.2284</td>
<td>0.2440</td>
<td><strong>0.0366</strong></td>
<td>0.2369</td>
</tr>
<tr>
<td>GTPase activity</td>
<td>0.2284</td>
<td><strong>0.0231</strong></td>
<td>0.2200</td>
<td>0.2369</td>
</tr>
<tr>
<td>intracellular signaling cascade</td>
<td>0.1266</td>
<td>0.4717</td>
<td>0.1182</td>
<td><strong>0.0351</strong></td>
</tr>
</tbody>
</table>

Table 2b: Shows enrichment of muscle up-regulated exons in a particular pathway. Few of the GO terms picked commonly by all 4 methods are shown here. GO terms with significant p-values (<0.05) are considered as significant. Those p-values are highlighted and italicized.

**Discussion**

Different analysis have been applied on the consistent human and mouse orthologous exons data.

*Significant exons count*

Through Significant exons count analysis, it could be seen that performance of Independent analysis, Species-aware analysis, and Meta-analysis seems to be pretty close.
If the exons picked as significant by these 3 methods were same set of exons or different was determined, and most of the exons were found to be commonly picked as significant by the 3 methods. These observations further confirms that the performance of the Independent, Species-aware, and Meta-analysis methods is pretty close.

**Precision plots from the actual data vs the negative controls**

The precision plots from actual data show that the performance of the Independent, Species-aware, and Meta-analysis is close, which again confirms with the significant exons count results. Precision plots from negative controls show a significant drop in case of Independent method, and Meta-analysis when compared to precision plots from actual data.

In the case of negative control 1 where, human data is swapped, the performance of Species-aware method has dropped compared to Independent method, and Meta-analysis. This explains the sensitivity of Species-aware method when applied on biologically non-sense data.

On contrary, in case of negative control 2 (Human - brain, muscle vs Mouse - testis, muscle), and in case of negative control 4 (Human - brain, muscle vs Mouse - testis, ovaries), Species-aware method's performance seems to be enhanced compared to Independent and Meta-analysis methods. In case of negative control 3 (Human - brain, muscle vs Mouse - brain, ovaries), the performance of Independent, and Meta-analysis methods show significant drop compared to precision plots from actual data.

Comparing the precision plots from actual data vs negative controls, it can be said that Independent method, and Meta-analysis method seems to be performing well compared to the other two methods.

**GO Analysis**

For GO analysis, brain up-regulated and muscle up-regulated exons picked as significant by each method were used to determine their pathway enrichment. Since, the data for these exons is from brain and muscle, looked at brain and muscle related GO-terms from the GO results. None of the brain and muscle related GO-terms were found to have significant p-values (<0.05), hence they are not listed in the Tables 2a, 2b. Most of the pathways to which the exons seems to be enriched in are membrane organization, DNA binding, transcription related terms.
Conclusion

Through all these analyses, it can be said that Independent, Species-aware, and Meta-analysis methods performance seems to be close to each other and performance of Independent method, and Meta-analysis method seems to be somewhat better compared to the Species-blind and Species-aware methods. Although, the splicing index data is obtained from brain and muscle tissue replicates none of the exons were found to be associated with brain or muscle related GO terms with significant p-values. This observation and the results from the negative control precision plots indicate that the data does not look promising enough to yield any statistically/biologically significant results. Also, the sample size of the data (3 brain replicates, and 3 muscle replicates) does not seem to be good enough.
References

5. Modrek B and Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nature Genetics, 2003.
11. Yeo GW, Nostrand Van E, Holste D, Poggio T, and Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. PNAS, 2005.