Human Exon 1.0 ST array.

Analysis was performed using XRAY software, version 2.5 Excel Add-In (Biotique Systems, Inc.). Input files were normalized with full quantile normalization. For each input file and for each probe expression value, the ith percentile probe value was replaced with the average of all ith percentile points across all arrays. Probes with GC count less than 6 and greater than 17 were excluded from the analysis. The expression score for a probeset was defined to be the median of its probe expression scores and probe-sets with fewer than 3 probes (that pass all of the tests defined above) were excluded from further analysis. For this analysis only 'Core' ("highly reliable" annotation) probe-sets were analyzed. Probe-set p-value was derived via Fisher p-value combination, and probe-sets not significantly expressed above background (p-value > 0.05) were removed from analysis.

Low-variance probe-sets were excluded from the analysis via a Chi-Squared test. A probe-set is considered to be low-variance if its transformed variance is to the left of the 95 percent confidence interval of the Chi-Squared distribution with (N-1) degrees of freedom.

Mixed Model, Nested Analysis of Variance was used to identify genes with tissue specific gene expression or alternative splicing. The data generated above are fitted to the linear model.

\[ Y[i,j,l] = M + d[i] + e[j] + c[l(i)] + ed[i,j] + ec[j,l(i)] \]

where M is a global mean, d(i) is the effect attributable to tissue state i, e(j) is the effect of exon j, and e and ed are interaction effects. c, which is the hybridization (or chip) effect, is a random factor and all other factors are fixed. Note that the CEL file effect, c, is nested inside tissue state. Genes with significant D (tissue) effect are said to show significant tissue based gene expression difference. Genes with significant Exon-Tissue interaction (ED effect) are candidates for tissue specific alternative splicing (p-value < 0.001).

The splicing index (SI) was calculated as previously reported\(^1,2\). A T-test was performed with the average SI of Day 7 replicates and the average SI of Day 14 replicates, treating the average SI of each junction or exon probeset as individual data points. The p-value cutoff was set at 0.001, and the fold change cutoff was set at Log\(_2\) 0.5. The probesets that reached the p-value and fold change cutoffs were sorted by genomic location in Excel, and the best candidates for alternative splicing were further tested by RT-PCR.

Exon Junction Array Design

The research junction array contains probesets for all exons and exon-exon junctions observed within transcripts in the input data. As input the design uses transcript annotations from RefSeq (NCBI36), Ensembl (version 38), and ExonWalk. All transcripts were mapped to the March 2006 version of the human genome (NCBI 36, hg18). The array is designed primarily to interrogate well annotated exons and splicing events from known genes. ExonWalk is a program that merges cDNA evidence together to predict full length isoforms, including alternative transcripts. It is designed to incorporate the richness of transcript variation present in cDNA sequences but limit some of the noise present in EST libraries by including several rules. ExonWalk requires that every exon and junction be either: 1) Present in cDNA libraries of another organism, 2) Have multiple cDNA GenBank entries supporting it, or 3) Be evolving like
a coding exon as determined by Exoniphy. More information on ExonWalk is available through the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=exonWalk).

Probe selection regions (PSRs) were created for the exon probesets in a manner that is analogous to the Human Exon 1.0 ST Array (http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf). A exon was divided into multiple PSRs if there is evidence for alternative splice site usage. We targeted 8 perfect match probes for each PSR. All PSRs greater than or equal to 25bp were represented by at least one probe. All exon-exon junctions observed in the input transcripts (both alternative and constitutive junctions) are interrogated with a probeset containing 8 perfect match probes that are tiled in one-base increments from the -4 to the +4 position (relative to the joining event).

Similar to the Exon Array, the probes on the array are all 25-mers and are designed for sense strand target.. The research junction array contains the same control and background probes as the Human Exon 1.0 ST Array such that data processing and array QC methods can be shared by both array types.

**Junction Array Design Statistics:**

<table>
<thead>
<tr>
<th></th>
<th><strong>Human Research Junction Array (HJAY)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>NCBI 36, hg18 (Mar 2006)</td>
</tr>
<tr>
<td>Transcript Clusters (genes)</td>
<td>35,123</td>
</tr>
<tr>
<td>Total Transcripts</td>
<td>335,663</td>
</tr>
<tr>
<td>Observed Junctions</td>
<td>252,741</td>
</tr>
<tr>
<td>PSRs (exons)</td>
<td>298,285</td>
</tr>
<tr>
<td>Probes per Junction</td>
<td>8</td>
</tr>
<tr>
<td>Probes per PSR (exon)</td>
<td>8</td>
</tr>
</tbody>
</table>

**Analysis methods:**

Target prepared from five biological replicates (independent differentiations) of samples taken at day 7 and day 14 were hybridized to the research junction array. Array data was processed using the Affymetrix Power Tools. Principal component analysis of the data was carried out using Partek Genomics Suite. During this exploratory analysis, it was discovered that one replicate of the day 14 samples was a clear outlier from the other samples of that differentiation stage (figure below). This sample was removed from all subsequent analyses.

The research junction array data was analyzed using the Splicing Index approach$^1$. Each exon or junction probeset was treated as an independent unit. Intensities from individual probes were summarized into a single probeset value using PLIER. Gene-level signal was estimated by calculating the median intensity of all exon-level probesets within the transcript cluster. Splicing Index analysis was carried out as previously described$^1$. In short, each probeset is normalized to the expression level of the gene by dividing each probeset intensity by the gene-level signal to generate the gene-level-normalized intensity. The magnitude of the splicing change indicated by
each probeset was calculated by taking a log ratio of the median gene-level-normalized intensity of day 7 samples relative to day 14.

\[ \text{Splicing Index} = \frac{\text{Median}[(\text{Probeset}_{\text{Day7}}/\text{Gene Level}_{\text{Day7}})]}{\text{Median}[(\text{Probeset}_{\text{Day14}}/\text{Gene Level}_{\text{Day14}})]} \]

The statistical significance of the splicing change indicated by each probeset was determined using a student t-test that compares the gene-level-normalized intensities of the five replicates of Day 7 to the four replicates of Day 14. As previously described, the array data was filtered to remove probesets and/or genes that were not expressed above background prior to the splicing index analysis. Splicing Index results were filtered to include only probesets with t-test p-values less than 0.001 and splicing index magnitudes greater than 0.5.

We then looked for splicing events where more than one independent probeset passed the filters and thus were deemed significant. For example, in the case of a simple cassette exon, there are four probesets that monitor that splicing event. Three probesets are designed specifically to detect inclusion of the exon (two include junction probesets on either side of the exon, and the exon probeset itself). In addition there is one junction probeset that detects skipping of the exon. Furthermore, the include and skip probesets should act reciprocally. For example, if a cassette exon is included at a higher rate on day 14, we would expect the gene-level-normalized intensities of the exon probeset and the two include junction probesets to be higher, while the skip junction probeset is lower. For our analysis, we focused on splicing events that demonstrated this reciprocal behavior of the skip junction probeset and one or more of the probesets specific to the inclusion of the exon. Including this requirement for reciprocal changes dramatically reduces the number of false positive predictions that are commonly induced by changes in the expression level of the gene itself.
Principal Component Analysis of Research Junction Array Data

Principal component analysis was carried out using Partek Genomics Suite. One of the Day 14 replicates was found to be an outlier and was removed from subsequent analyses.

REFERENCES