SUPPLEMENTARY INFORMATION FOR:

Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*

Arun K. Ramani[1,2,5,6], John A. Calarco[1,2,3,5,6], Qun Pan[1], Sepand Mavandadi[1], Ying Wang[3], Andrew C. Nelson[4], Leo J. Lee[1], Quaid Morris[1,2], Benjamin J. Blencowe[1,2], Mei Zhen[2,3,6], and Andrew G. Fraser[1,2,6]

[1]Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1

[2]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada M5S 1A8

[3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario, Canada M5G 1X5

[4]Department of Physiology, Development and Neuroscience, Anatomy Building, Downing Street, University of Cambridge, Cambridge, UK, CB2 3DY

[5]These authors contributed equally to this work

[6]Correspondence should be addressed to:

Andrew G. Fraser, John A. Calarco, Arun K. Ramani
The Donnelly Centre,
160 College Street
Toronto, ON, Canada, M5S 3E1
andyfraser.utoronto@gmail.com (416) 978-2712
john.calarco@utoronto.ca
ramaniak@gmail.com

Mei Zhen,
Samuel Lunenfeld Research Institute,
Mount Sinai Hospital,
600 University Avenue,
Toronto, Ontario,
Canada M5G 1X5
zhen@lunenfeld.ca (416) 586-4800 ext. 1592

Microarray design

Alternative splicing events were identified from alignments of *C. elegans* mRNA/EST sequences (UniGene Build #26) to *C. elegans* genomic sequence, essentially as previously described (Pan et al. 2005; Pan et al. 2004). In total, 499 cassette type AS events were identified. For each AS event, 3 exon probes and 3 exon junction probes were designed to profile the AS event on the microarray, essentially as previously described (Pan et al. 2004). To facilitate the profiling of unannotated AS events in the *C. elegans* genome, an exon and splice junction tiling array was designed, in which each internal exon in a gene was treated as a potential alternative exon. To maximize the number of new AS events that could be profiled, we included probes for all known or predicted internal exons for ~50% of *C. elegans* genes, focusing on genes that are preferentially expressed in the nervous system. Using the same exon body and junction probe design as for the 499 EST/cDNA sequence-supported events, we were able to include probes covering 55,759 potential events in 8,649 genes. *C. elegans* genes and exon coordinates were first downloaded from Wormbase, and exon sequences were then extracted and probes were designed as described above.

RNA extraction, purification, and cDNA synthesis and

For RNA-Seq analysis, total RNA was prepared from mixed stage or synchronized whole animals using Trizol solution (Invitrogen) according to the manufacturer's protocol, and subsequently cleaned using RNeasy columns (QIAGEN) followed by treatment with 10 units of Dnase I (Roche) in 1x One-Phor-All buffer (Amersham) for 30 minutes. Poly A+ RNA was then purified from total RNA using Oligotex midi kits (QIAGEN) according to the manufacturers protocol. cDNA was then produced using SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) and purified using a PCR Purification kit (QIAGEN). Sequence data for the resulting cDNA was then obtained using an Illumina Genome analyzer as described (Wilhelm et al. 2008) For microarray analysis, total RNA was prepared from synchronized larval (L1 through L4) and adult stage animals using Tri reagent (Sigma Aldrich) according to manufacturers recommendations. Poly A+ RNA was isolated and used to synthesize cDNA as described previously (Ramani et al., 2009).

Mapping and analysis of RNA-Seq data

Sequencing reads were mapped to the WS200 version of the *C. elegans* genome using MAQ version 0.7.1 (Mapping and Assembly with Qualities; Li et al. 2008). Only uniquely mapped reads at a quality threshold of 30 or greater were retained. These were further filtered to remove reads that mapped with insertions or had >2 mismatches, and had a quality score less than or equal to 30 (i.e. 1 in 1000 mistake). The remaining 'unmapped' reads were then analyzed as described below.

Identifying reads mapping to splice sites

To map the remaining unmapped reads to splice junctions we created a non-redundant set of sequences 66 nucleotides long corresponding to all possible known and predicted splice junctions (annotated adjacent and non-adjacent exons based on WS200 (Harris et al. 2010) were used). This exon junction database (EJDB) was created by combining 33 nucleotides from the 3′ end of the upstream exon with 33 nucleotides from the 5′ end of the downstream exon. In total we have 548,374 junctions from 20,534 annotated multi-exon genes in Wormbase. The reads were aligned to the EJDB using BLAT (Kent 2002) and reads were identified as mapping to a junction if there was at least a four nucleotide overlap over either exonic half of the corresponding junction. Reads that had multiple hits to different junctions were eliminated. Junction sequences formed by combining non-adjacent exons and having reads mapping to them uniquely were determined to be alternative splice sites.

We classified AS events as either 'cassette-type' events, alternate initiation events, alternate termination events and 'ambiguous' based on specific rules. An event is assigned to be a cassette event if there are reads mapping to all three possible junctions and the ratio of the number of reads mapping to the two adjacent junctions is within 30 fold of each other (based on 5% false positive from the 'true-positive' data). If one of the adjacent junction count is '0' and there are at least 30 reads mapping to the other adjacent junction, it is classified as an alternate start (if the upstream junction is missing) or stop (if the downstream junction is missing). We also required in every case for the %In values to be at least 1% and less than 99%. If any of these criteria are not met, the event is classified as 'ambiguous'.

Identifying novel splice sites

We first trained our algorithm on WS180 version of the genome, assuming this was the current version and the WS200 version of the genome was the future 'true' version. We identified the set of events in WS200 that were not annotated in WS180 and thus considered 'novel'. Splice junctions involving unannotated splice sites were detected from this set of reads if they mapped in two perfect blocks of genomic sequence separated by at least 25 nucleotides (most introns were at least 25 nucleotides apart) and if they met additional criteria determined from analyzing known splice junctions. Genomic features in these events were used to derive a set of rules required to identify junctions: A. candidate junctions were flanked by canonical intronic splice site dinucleotide sequences (GT and AG) B. junctions had on average at least 5 supporting mapped reads. C. the ratio of the median number of reads mapping to the adjacent intron to the median number of reads mapping to the exon was < 0.1 (this indicates a positional change from an intron to an exon).

Reads that did not map to the genome or to 'EJDB' were mapped back to the WS200 version of the genome using BLAT. Here we accept reads to have positively mapped to the genome if they map uniquely to the genome in exactly two blocks and each of the blocks have at least 5 nucleotides perfectly matching the region. We then applied the rules learnt above to identify reads that span two distinct regions of a transcript.

cDNA labeling, Microarray hybridization and analysis

cDNA samples were labeled with cy3 and cy5 fluors using a ULS labeling kit (Kreatech) according to manufacturer's recommendations. Each cDNA sample was independently labeled with cy3 or cy5, and used in dye-swap replicate experiments during microarray hybridization. Microarray hybridizations were performed as described previously (Calarco et al. 2009) using an HS 4800 Pro hybridization station (Tecan). After hybridization and washing, microarray slides were scanned using an Agilent microarray scanner. The raw probe intensity values from the microarrays were preprocessed by Agilent Feature Extraction software (version 9.5) using a customized protocol, followed by variance stabilizing normalization (VSN) performed by the vsn package within Bioconductor (version 2.1). Normalized probe intensity values were then analyzed by the Probe Affinity Transcript Abundance (PATA) algorithm (Mavandadi et al. *manuscript in preparation*) to provide quantitative predictions of alternative exon usage across the developmental stages tested. Briefly, PATA takes into account sequence features in probes that contribute to cross-hybridization artifacts. After correcting for these potential biases, the algorithm then estimates the relative abundances of the isoforms including or skipping the alternative exon for each AS event profiled on the microarray. Percent inclusion (%In) values are then calculated as the abundance of the included isoform divided by the sum of the abundances of the included and skipped isoforms.

Defining a set of annotated trancript variants in C. elegans

To define a "true positive" set of alternative mRNA processing events in *C. elegans* we first obtained all transcripts with EST/cDNA evidence from WormMart (Harris et al. 2010). We then used an algorithm based on the methodology and rules outlined in the *Drosophila melanogaster* Exon Database (Lee et al. 2004) to identify all annotated splice variants and isoforms. Based on these criteria, we placed the annotated events into seven categories: Exon skipping (cassette type), alternative 5′ splice site usage, alternative 3′ splice site usage, intron retention, alternative transcript start site usage, alternative terminal exon usage and mutually exclusive exons.

Filtering array events for analysis

By including all possible exon junctions from the 55,759 array events in the 'EJDB', we were able to identify the set of array events that had reads corresponding to the alternative junction. We used the intensities of the control probes on the array to determine the background probe intensity distribution at each stage. By setting a 5% false positive probe inclusion rate, we filtered events where the exons or junction probe intensity was less than this. We determined the expression level as the average of the intensities of the two constitutive exons. Once again, events were eliminated from a stage if this expression intensity was less than the determined background cutoff. We finally filtered out events that did not have reads mapping to all three junctions. Since the array was designed specifically for identifying cassette type AS events, we had to eliminate potential alternative starts and stops from the analysis. Since alternative starts and stops

have an extreme skew in the ratio of the two adjacent junctions (since one of them is used while the other is not), we determined the ratios of the two adjacent junctions in the cassette exons in the true positive set and applied a cutoff based on 5% false positive rate. We also determined the distribution of %In values based on RNA-Seq for these true-positive cassette exons and eliminated events in the array that had %In values that were below the lower 5% or higher than 95% of the events. By applying these filters we identified 704 events for further analysis. To identify temporally regulated events from the array, we compared the %In values of the events that were present in more than one stage and found that 624 of the 704 events were in at least 2 stages. Additionally, from the complete set of junctions mapped, we identified the set of exons that were part of at least 2 different junctions i.e. exons that were multiple donor splice sites or multiple acceptor splice sites.

Computational analysis of alternative exon properties

RNA-Seq percent inclusion value calculation

Percent inclusion values were calculated as the ratio of the average number of reads across the two adjacent junctions over the average of the number of reads across the adjacent and alternative junction.

$$\%In = \frac{(c1.a + a.c2)/2}{c1.c2 + (c1.a + a.c2)/2}$$

where, c1 and c2 refer to the two constitutive exons and a refers to the alternative exon, c1.a and a.c2 are the number of reads mapping to the two adjacent junction and c1.c2 is the number of reads mapping to the alternative junctions.

Frame preservation

Exons lengths were calculated and grouped in one of three categories (3n, 3n+1, or 3n+2) based on whether their lengths are perfect multiples of 3. Distribution of the three possible outcomes in each of the seven alternative splice categories determined above and in the novel alternative splice events was compared to the background evaluation of the three categories based on all the exons annotated in the WS200 version of the genome.

Sequence Conservation

Exon sequences were obtained from WormMart (WS200 version of the genome) and were aligned to the C. Briggsae genome (version WS200) using BLAT. For each exon, maximum percent identity in sequence aligned to the *C. briggsae* genome was calculated. This was compared against the average percent identity of the exons undergoing cassette type AS, either in the true-positive set or from the novel events.

Estimates of frequency of AS and developmentally-regulated AS

To estimate the frequency of AS in *C. elegans*, we first identified the number of genes possessing novel AS events involving annotated junctions (1354 genes with 2183 events) that belonged to either the cassette, alternative initiation, or alternative termination classes. From this subset we extrapolated to estimate all genes with AS (Total_AS_Genes) as:

$$Total\_AS\_Genes = \frac{\left(NG \Big/ \left(TP_{subsetID} \Big/ TP_{subset}\right)\right) + \left(TP_{subset}\right)}{TP_{subsetID} \Big/ TP_{tot}}$$

where, 'NG' is the 1,354 genes with 2,183 novel AS events (Fig 2A), 'TP_{tot}' is the total genes in the 'true-positive' set (2,065). While 'TP_{subset}' is the set of 'true-positive' genes having either a cassette, alternative initiation or alternative termination events (1,015), '$TP_{subsetID}$' is the subset of these (815) that we positively identify with our RNA-Seq data. Similarly we calculated the total AS events in the genome, by replacing gene numbers with AS event numbers.

To estimate the frequency of AS events that are temporally regulated, we averaged the fraction of events identified to change by at least 20% from our microarray or have a pvalue of < 0.01 from our RNA-Seq data between any two stages. We find that 38% of the array events are temporally regulated (275/624 events corrected for false positive rate determined from RT-PCR). Of the 1,573 AS events with sufficient read depth we identify 22% (349) with a pvalue < 0.01.

Motif analysis

We used SeedSearcher (http://www.psi.toronto.edu/~weijun/snap/tools.php), a hyper-geometric approach for identifying putative binding sites. The algorithm compares a set of input sequences against a set of background sequences to identify statistically enriched motifs of specified length (pentamers in our case). We used 50-nucleotide intron sequences upstream (or downstream) of temporal-regulated exons and compared these with the set of 50-nucleotide upstream (or downstream) intron sequence from all identified AS exons. This analysis therefore identifies motifs that are enriched specifically in temporally-regulated events rather than generic AS events.

References

Calarco, J.A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D. et al. 2009. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138:** 898-910.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R. et al. 2010. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res* **38:** D463-467.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Lee, B.T., Tan, T.W., and Ranganathan, S. 2004. DEDB: a database of Drosophila melanogaster exons in splicing graph form. *BMC Bioinformatics* **5:** 189.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851-1858.

Mavandadi, S., Calarco, J.A., Wang, X., Pan, Q., Blencowe, B.J., and Morris, Q. *manuscript in preparation*. A new method for alternative splicing microarray data analysis reveals differences between human and chimpanzee transcriptomes predicted to impact HIV-1 tropism.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R., and Blencowe, B.J. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet* **21:** 73-77.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D. et al. 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* **16:** 929-941.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453:** 1239-1243.

Figure S1: Analysis of how read depth affect junction identification

We compare the total junctions identified versus new junctions identified **A**. True positive junctions from Fig 2A, **B**. All adjacent junctions from WS200, and **C**. All non-adjacent junctions from WS200). In each case we increase the read depth by adding additional sequence reads and ask how many total reads we see and how many of these are new junctions and we notice that we saturate the potential junctions we can query. We also notice that while we saturate new junctions we can find, we can provide additional read support to junctions that were previously overlapped by single reads **D**. To estimate the accuracy of junction identification, we compared the accuracy of our algorithm by mapping the same set of reads to either a real database (blue diamonds) or to a random junction database (red diamonds). This shows that far fewer random junctions are mapped compared to real junction sequences **E** and that we have a very low false-positive rate misidentifying junction sequences (~2%).

Figure S2: Novel AS events identified by RNA-Seq

RT-PCR assays confirming six additional novel AS events identified by our RNA-Seq analysis. Primers are designed to anneal to flanking constitutive exons, leading to the amplification of two products. For each image, the slower migrating product represents the exon-included isoform, and the faster migrating product corresponds to the exon-skipped isoform.

Figure S3: AS events differentially regulated during development

RT-PCR assays validating AS events predicted by either our RNA-Seq or microarray analysis to undergo differential regulation between any two stages of development surveyed in our study. Products in each gel image are the same as described in Figure S2.

Figure S4: Motifs identified in the upstream region of temporally-regulated AS events

Using SeedSearcher, we identified pentamer motifs located in the 50-nucleotide sequence upstream of temporally-regulated AS events. These 12 motifs were found to be statistically enriched in these temporally-regulated AS events compared to all AS events.

Figure S5: Motifs identified in the downstream region of temporally-regulated AS events

Using SeedSearcher, we identified pentamer motifs located in the 50-nucleotide sequence downstream of temporally-regulated AS events. We identify 19 motifs that are statistically enriched in these temporally-regulated AS events compared to all AS events.