SUPPLEMENTARY INFORMATION FOR:

Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing

Qun Pan[1], Ofer Shai[2], Leo J. Lee[2], Brendan J. Frey[1,3] and Benjamin J. Blencowe[1,3,4]

[1] Banting and Best Department of Medical Research, University of Toronto
[2] Department of Electrical and Computer Engineering, University of Toronto
[3] Department of Molecular Genetics, University of Toronto

[4] Correspondence:

B.J. Blencowe, PhD
Centre for Cellular and Biomolecular Research
Donnelly CCBR Building
160 College Street, Room 1016
Toronto, Ontario
M5S 3E1
Canada

Tel 416-978-3016
Fax 416-946-5545
Email b.blencowe@utoronto.ca

Methods

Mining exons and splice junctions from human cDNA and EST data

Human UniGene sequences (Build #208) and human genomic sequences (Build #36) were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). RefSeq data were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz). UniGene clusters that contain no RefSeq sequence were removed and RefSeq sequences that map to the same locus were grouped. After removing redundant UniGene clusters, 15,702 genes containing at least one RefSeq sequence remained for further analysis. For the plots generated in Figure 2, the number of exons per gene is defined as the number of exons identified in the longest RefSeq sequence for a gene. In total, 175,944 exons were mined from the 15,702 genes.

To identify all possible known splice junctions, mRNA/EST sequences were aligned to genomic sequences using BLAST[1] and SIM4[2]. For quality control, a known splice junction is identified only if it is flanked by the intronic dinucleotide sequence GT at 5'-end and AG at 3'-end. Splice sites located outside of the genomic coordinates corresponding to RefSeq transcripts were removed. All possible additional pairings of the 5' and 3' splice sites in the extracted set were simulated to generate hypothetical new junctions. In total, 257,257 known splice junctions and 2,459,306 possible new junctions were identified in the 15,702 genes.

Detection and analysis of splice junction sequences using Illumina mRNA-Seq data

For each splice site junction sequence to be searched with Illumina reads, a 64 nucleotide-long sequence was extracted with 32 nucleotides of overlap in each upstream and downstream exon. To generate negative control sets of junctions with similar statistics, the 5'-half and the 3'-half of the junction sequence were arranged in reverse order (i.e. the same sequences from the sense strand were represented in reverse order).

All Illumina sequence reads used in searches are 32-mers and were generated from cDNA from six tissues: brain, cerebral cortex, heart, liver, lung and skeletal muscle. To assess gene/mRNA coverage, reads were aligned to all exons in a gene using Blat[3] and a full 32-nucleotide alignment was required with up to two mismatches/indels. To account for different exon lengths, gene/mRNA coverage was determined by calculating the number of reads per 100 bases.

To determine the numbers of splice junctions detected by the sequence reads, the reads were aligned to all the known and possible new splice junctions as well as to the control junctions using BLAT. A splice junction is aligned to a sequence read if the full length of the read can be aligned to the junction sequence when allowing no more than two mismatches/indels, and with at least five nucleotides of the read overlapping either the upstream or downstream exon.

To account for biases in the negative controls, we trained both linear and nonlinear classifiers to discriminate between true splice junctions and those generated by

false/random hits, using logistic regression[4] and a decision tree[5] with the following 5 features:

1) Total number of sequence reads matching to the junction with at least 5 nt overlap on either side of the junction and at most 2 mismatches and/or indels
2) Distribution of mismatches in aligned sequence reads (provided by three relative frequencies, corresponding to 0, 1, and 2 mismatches)
3) Distribution of gaps in aligned sequence read (provided by two relative frequencies, corresponding to 0 and more than 0 gaps)
4) Total number of unique alignments to the junction
5) Minimum balance factor among all sequence reads. Given that a sequence read is aligned to a junction starting from position -x (ending at y=32-x), the balance factor for that read would be |x-y|. For example, a read that is aligned at position -20 (and ends at +12) would have a balance factor of 8 (|20-12|). A balance factor of 0 indicates alignment of 16 nt of sequence on either side of the junction.

The classifiers were trained using 10-fold cross-validation to estimate the achievable specificity and sensitivity and to establish an operating threshold (Figure S1). The classifier was then trained on the entire training set to obtain the best estimation for the parameters. Both logistic regression and decision trees performed well and achieved similar results. Only results from the linear regression classifier are shown.

The performance of the classifier was also compared to the use of stringent, preset alignment thresholds without applying the classifier, for example, requiring reads to have a 5- or 6-base overlap across a junction and a perfect match. When scoring the rate of true versus false positive junctions, the classifier performed significantly better than when applying these stringent preset thresholds.

Estimates for the numbers of known and new splice junction sequences

To estimate a lower-bound for the detection of known and new splice junctions, all the sequence reads that mapped to more than one splice junction (either known, possible new or control) are removed. The logistic regression classifiers were trained using this new set and 128,395 known junctions and 4,294 new junctions were predicted to be true.

Of the 4,294 new junctions, 439 were detected in exactly two tissues and were further analyzed for their tissue specificity. The number of new junctions detected only in a pair of tissues was normalized by the total number of new junctions detected in that pair of tissues. The relative proportions of junctions common to pairs of tissues, for all 15 possible combinations, are depicted in a Hinton graph (see Fig. 1D).

To estimate the total proportion of multiexon genes with one or more AS event, we combined information from EST/cDNA data and from the detection rates of new junctions at progressively higher sequence read coverage (measured as the number of read counts per 100 nt window across the longest RefSeq cDNA surveyed per gene). When sampling four of the highest ranges of read counts per 100 nts (256-512, 512-1024,

1024-2048, and >2048), which represent on average 238 genes per range, on average we detect one or more AS event in ~95% of multiexon genes.

To project the total number of AS events in the 15,702 RefSeq genes, the known splice junctions identified by mRNA/EST sequences were evaluated for AS. An AS event is counted if there is an exon-skipping event or if there is a different 5' or 3' splice site. The number of AS events in a gene is further determined by the sum of AS events detected in known junctions and the number of new junctions detected in all hypothetically possible junctions. A scale factor of 2 is applied to the number of new junctions to account for AS events missed in the six tissues, since the number of new junctions increases linearly as the number of tissues increases (see Figure 1B), and from analyzing the six tissues, ~50% of the known junctions can be detected. Using this calculation, there are 71,462 known AS events identified from mRNA/EST sequences and 8,588-20,198 new AS events detected using sequence reads. The AS frequency was then determined as the number of AS events per exon and plotted against different sequence read coverage per 100 bases (Figure 2B). To achieve a reasonable sequence coverage and to also limit counting of AS generated by possible splicing errors, the medians of AS frequency in the middle range of sequence read coverage (from 32 to 256 reads per 100 bases) are taken to give estimates of the total number of AS events and are multiplied by the total number of exons in the 15,702 genes. It is predicted that there are 87,972-131,958 AS events in these genes.

Comparisons of Illumina mRNA-Seq and quantitative AS microarray profiling data

Confidence-ranked %in values for ~5000 cassette type alternative exons were obtained using the GenASAP algorithm[6,7] from pre-processed microarray data generated by profiling 54 diverse human tissues (including the same six tissues as analyzed by Illumina sequencing) using a custom Agilent 244K microarray (unpublished data). %in values from Illumina data were obtained by counting the number of reads mapped to the three junctions formed by AS of a single cassette skipping event in the six tissues and calculating the ratio of the average count for reads matching the two splice junctions formed by exon inclusion, over the count for reads matching the skipped exon junction. The plots in Supplementary Fig. 2 show cassette AS events represented by 20 or more reads to any of these three junctions, in each of tissue being compared. For the comparison of tissue-regulated AS levels, all cassette exons represented by probes on the AS microarray and which displayed at least a 50% inclusion level change (based on differential %in levels determined from read counts), were selected for analysis.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
2. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**, 967-74 (1998).
3. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
4. Agresti, A. *Categorical Data Analysis*, (Wiley, New York, 2002).

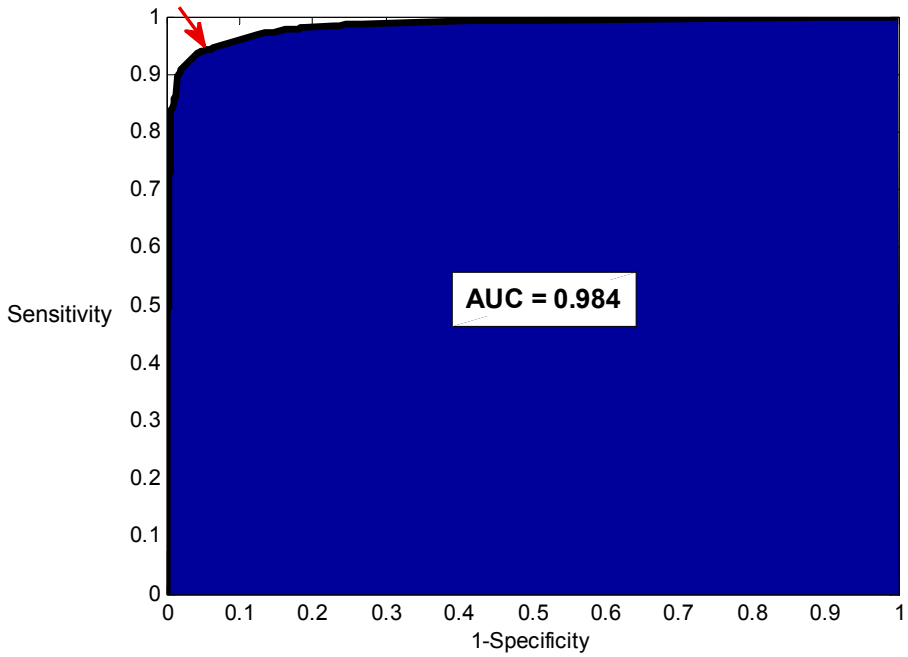
5. Breiman, L. *Classification and Regression Trees*, (Chapman and Hall, Boca Raton, 1993).
6. Pan, Q. et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* **16**, 929-41 (2004).
7. Shai, O., Morris, Q.D., Blencowe, B.J. & Frey, B.J. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics* **22**, 606-13 (2006).

Table S1

Gene Symbol	Gene Name	# of exons	# of known junctions	# of possible new junctions	# of detected known junctions	# of detected new junctions
TTN	titin	314	396	64667	313	35
MBP	myelin basic protein	4	113	2418	36	20
NEB	nebulin	149	185	13770	156	17
MYH7	myosin, heavy chain 7	40	58	1528	39	14
OBSCN	obscurin	106	117	6227	106	12
NRAP	nebulin-related anchoring protein	42	53	1046	47	12
RPL8	ribosomal protein L8	6	39	355	19	11
PKM2	pyruvate kinase, muscle	11	67	2352	14	10
FLNC	filamin C, gamma	48	49	1221	46	9
EEF2	eukaryotic translation elongation factor 2	15	39	810	19	9
LOC554235	hypothetical protein LOC554235	5	12	31	10	8
COL3A1	collagen, type III, alpha 1	50	69	2093	51	8
DYNC1H1	dynein, cytoplasmic 1, heavy chain 1	79	91	3589	76	8
MYH9	myosin, heavy chain 9, non-muscle	41	61	2229	41	8
TNNT3	troponin T type 3 (skeletal, fast)	11	28	241	20	8
PSAP	prosaposin	14	36	901	17	8
TNNT2	troponin T type 2	11	33	309	23	8
HUWE1	HECT, UBA and WWE domain containing 1	84	93	3873	79	7
COL6A2	collagen, type VI, alpha 2	26	47	857	28	7
COL1A1	collagen, type I, alpha 1	51	67	2025	51	7
A2M	alpha-2-macroglobulin	35	66	2067	37	7
SFTPB	surfactant, pulmonary-associated protein B	10	14	94	12	7
CD74	CD74 molecule, MHC, class II invariant chain	9	42	600	13	7
AQP7	aquaporin 7	8	13	49	6	6
HERC2	hect domain and RLD 2	93	97	4524	78	6
ANK2	ankyrin 2, neuronal	45	69	1623	59	6
PKD1	polycystic kidney disease 1	46	54	1281	49	6
PARP1	poly (ADP-ribose) polymerase family	23	36	556	22	6
NCOR2	nuclear receptor co-repressor 2	47	66	1480	58	6
SPTAN1	spectrin, alpha, non-erythrocytic 1	55	67	2065	56	6
PDE4DIP	phosphodiesterase 4D interacting protein	44	84	2275	61	6
TPM3	tropomyosin 3	10	43	502	17	6
FN1	fibronectin 1	46	78	2960	50	6
SNRPN	small nuclear ribonucleoprotein polypeptide N	13	45	504	26	6
ACADVL	acyl-Coenzyme A dehydrogenase	20	54	989	28	6
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	12	31	378	18	6
AP2M1	adaptor-related protein complex 2	11	40	576	18	6
ENO1	enolase 1	12	72	1461	8	6
GNAS	GNAS complex locus	13	60	926	20	6
C3	complement component 3	41	63	1537	40	6
ACTB	actin, beta	6	55	890	18	6

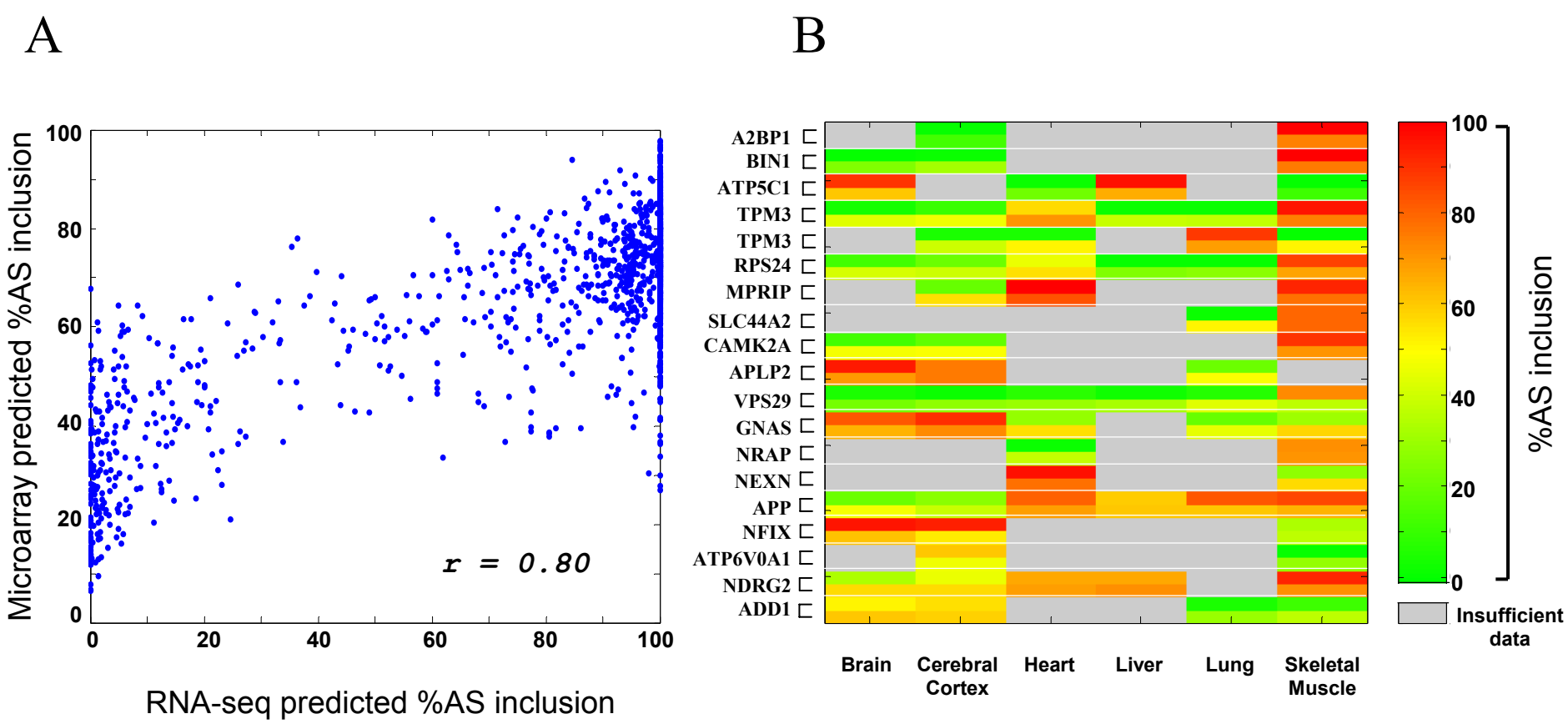
41 genes with more than five new splice junctions (lower bound estimate) detected using Illumina mRNA-Seq reads are listed. Most of these genes also have many additional, repetitive junction sequences detected in the mRNA-Seq data (not shown). Information provided for each gene includes the following: number of exons, numbers of known (i.e. EST/cDNA-supported junctions) junctions, number of hypothetically possible new junctions, number of known junctions detected by Illumina reads, and number of new junctions detected by Illumina reads predicted to represent true-positives (refer to Methods).

Figure S1



A Receiver Operating Characteristic (ROC) curve for the validation set of known and corresponding control junctions using a logistic regression classifier (see Methods). The ROC curve shows the sensitivity of the classifier versus 1-specificity as one varies the decision threshold. Sensitivity, also known as "recall", is the probability that a true event is detected, while specificity is the probability that a false event is rejected. Importantly, both measures are invariant to the number of positive and negative examples in the validation set, and therefore to large differences in the numbers of reads in our starting data. A random classifier would produce a ROC curve with area under curve (AUC) of ~0.5. An AUC of 0.984 indicates that our classifier has performed very well. The red arrow indicates the point of operation (94% sensitivity with 95% specificity) when the classifier is applied to known and hypothetically possible new junctions.

Figure S2


A
B
Microarray predicted %AS inclusion
RNA-seq predicted %AS inclusion
r = 0.80
0
20
40
60
80
100
A2BP1
BIN1
ATP5C1
TPM3
TPM3
RPS24
MPRIP
SLC44A2
CAMK2A
APLP2
VPS29
GNAS
NRAP
NEXN
APP
NFIX
ATP6V0A1
NDRG2
ADD1
Brain
Cerebral Cortex
Heart
Liver
Lung
Skeletal Muscle
%AS inclusion
Insufficient data

A. The ability of Illumina sequence data to generate quantitative measurements for percent exon inclusion (%in) levels for cassette alternative exons was assessed by directly comparing %in values generated using a previously described[6,7], validated quantitative AS microarray system. %in values from microarray profiling ~5000 human cassette alternative exons were compared with %in measurements calculated from the ratios of counts for Illumina reads that match the two splice junctions formed by exon inclusion and/or the splice junction sequence formed by exon skipping (see Methods). The correlation plot shows %in values for 1558 microarray-profiled AS events, for which there were 20 or more Illumina reads matching any one of the three splice junction sequences.

B. Illumina sequence read counts afford detection of tissue-regulated AS events. %in values are shown for cassette AS events analyzed in (A) represented by 20 or more reads per tissue that match at least one of the three junction sequences formed by inclusion and/or skipping of a cassette exon and which (based on the read counts) are predicted to have at least a 50% change in %in level between two or more of the six profiled tissues. The %in levels based on read counts (lower rows for each gene) are directly compared to %in values generated by quantitative AS microarray profiling the same tissues (upper rows for each gene). The colour scale on the right of the plot indicates %in levels. Gray boxes indicate that there were insufficient read counts to derive a %in measurement. Gene names are indicated on the left of the plot.