Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays

Jason M. Johnson,* John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M. Loerch, Christopher D. Armour, Ralph Santos, Eric E. Schadt, Roland Stoughton, Daniel D. Shoemaker*

Alternative pre–messenger RNA (pre-mRNA) splicing plays important roles in development, physiology, and disease, and more than half of human genes are alternatively spliced. To understand the biological roles and regulation of alternative splicing across different tissues and stages of development, systematic methods are needed. Here, we demonstrate the use of microarrays to monitor splicing at every exon-exon junction in more than 10,000 multi-exon human genes in 52 tissues and cell lines. These genome-wide data provide experimental evidence and tissue distributions for thousands of known and novel alternative splicing events. Adding to previous studies, the results indicate that at least 74% of human multi-exon genes are alternatively spliced.

Alternative pre-mRNA splicing is expected to make an important contribution to the complexity of the human proteome and at least half of human genes are alternatively spliced (*1*, *2*). Alternative splicing (AS) has been implicated in a wide and rapidly expanding set of physiological and pathophysiological processes, and it has been estimated that 15% of point mutations that cause human genetic disease affect splicing (*3*). Full-length cDNA sequencing projects supply gold-standard transcript definitions and are making substantial progress toward characterization of mammalian transcriptomes (*4*, *5*). However, these sequencing-based approaches are labor-intensive and expensive, and characterization of transcripts across all disease states, tissues, and stages of development remains a distant goal. Expressed sequence tags (ESTs) provide evidence of a vast number of alternative isoforms, but systematic studies of AS using ESTs are hampered by protocol differences, transcript end bias, and library coverage limitations (*6*). Recent experiments have illustrated the principle that microarray or fiber-optic array probes can monitor splicing events (*7*–*12*), and the idea of using probes positioned at exon-exon junctions to monitor splicing was suggested as early as 2000 (*13*).

Here, a set of five microarrays containing ~125,000 different 36-nucleotide (nt) junction probes were used to monitor the exon-exon connections of 10,000 multi-exon genes across 52 diverse samples. Splicing predictions from the array data were used to guide reverse transcriptase–polymerase chain reaction (RT-PCR) and sequencing validation ef-

Rosetta Inpharmatics LLC, Merck & Co., Inc., 12040 115th Avenue N.E., Kirkland, WA 98034, USA.

*To whom correspondence should be addressed. E-mail: jason_johnson@merck.com (J.M.J.); shoemakd@stanfordalumni.org (D.D.S.)

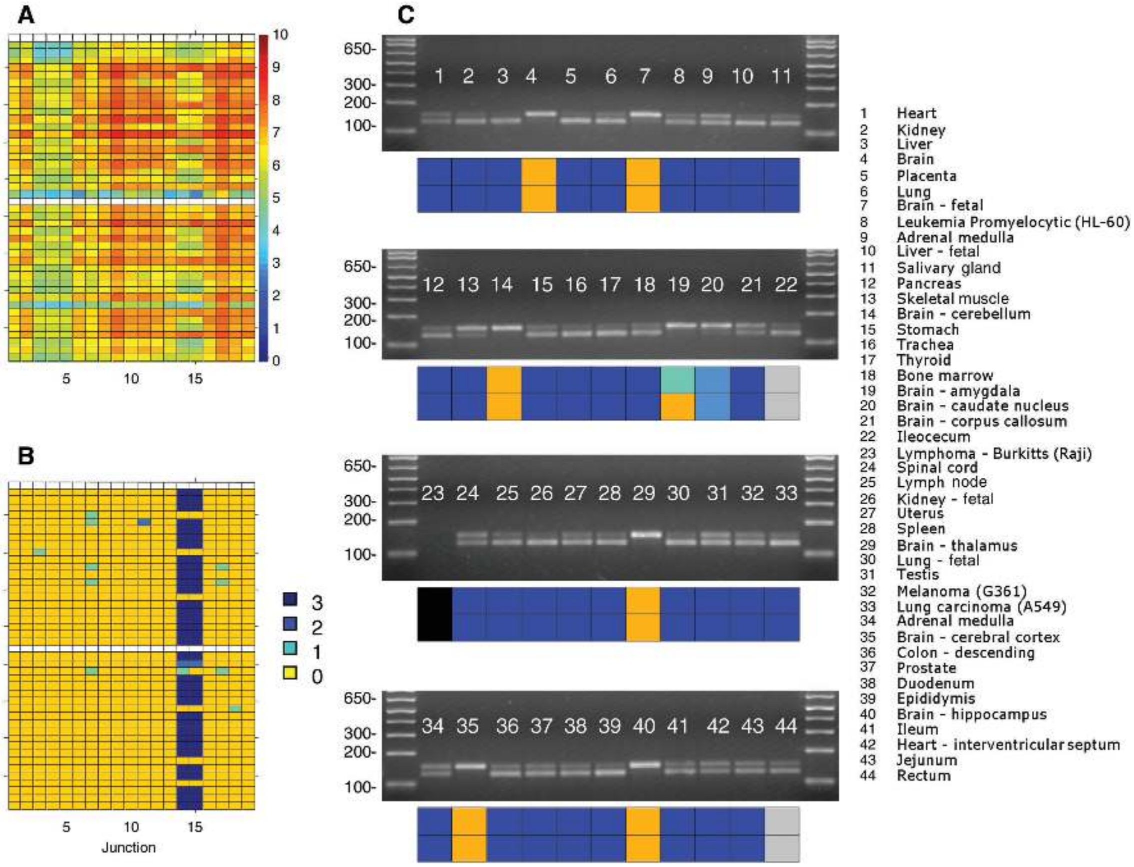

Fig. 1. Confirmation and tissue distribution of OCRL1 mRNA isoforms. (**A**) Each matrix point shows the natural log intensity (cy3) of one junction probe in one tissue. Probes complementary to the 19 junctions of the longer OCRL1 isoform (NM_000276.1) are ordered horizontally, 5′ to 3′. The four 5′-most exons did not map to the genomic contig and were excluded. Hybridization samples form the vertical axis of each matrix. (**B**) Predictions derived from the difference between modeled and observed intensities are color-coded by confidence. A score of 3 (dark blue) indicates the greatest deficit in observed intensity relative to the intensity expected by the model. (**C**) RT-PCR products from primers placed in exons flanking junctions 14 and 15 [exons 17 and 20 in previous studies (*15*)], with predictions from (B). Gray, samples not used in array experiments; black, samples where array data indicate OCRL is not expressed. Numbers to the left represent sequence lengths in basepairs (bp).

forts to specific transcript locations. This array-guided approach provides evidence of tissue-specific AS in thousands of genes and has allowed us to identify and sequence-verify splice variants not represented currently by ESTs or mRNAs. Because the expression level of every transcript is measured by two replicates for each exon-exon junction, these experiments also provide robust estimates of gene expression for each of the 10,000 genes in each sample (fig. S1). All expression data are freely available in the GEO database (*14*), accession number GSE740.

Hybridization data from the junction probes of each transcript were analyzed across 52 conditions to identify tissue-specific splicing differences. Intensities for junction probes were modeled as a function of probe response and tissue-specific expression level and empirically fit by assuming all exons in the transcript are present. Deviation between the observed and modeled intensities for a probe resulted in prediction of a tissue-specific splicing event. The predictions were scored from 0 to 3, where 3 indicates the greatest difference between predicted and observed intensities in both replicates. The result is a genome-wide set of tissue-specific predictions of AS.

Figure 1 shows example output for OCRL1, a gene with two known alternatively spliced mRNAs, one previously observed in retina and fetal brain (NM_000276) and a shorter form (NM_001587) observed in kidney that lacks a 24-nt exon (*15*). The data show decreased intensity for the two junction probes that flank this exon in many of the 52 conditions, consistent with an exon-skipping event. The tissue-specific accuracy of these predictions was confirmed by RT-PCR and sequencing. For every tissue with a prediction score of 3, a mixture of the two OCRL1 isoforms is observed (Fig. 1). Likewise, for tissues with prediction scores of 0 (all of which are brain subregions), the long form is predominant. The only neuronal samples with high prediction scores are from the corpus callosum and spinal cord, and both of these also contain a prominent shorter form. The example of OCRL1 shows that junction arrays can correctly detect tissue-specific regulation for alternative isoform mixtures.

The gene coding for β-amyloid precursor protein (APP) has three known alternative isoforms, presenting a more challenging mixture case. Figure 2 shows the array data and resulting AS predictions relative to the longest isoform of APP (NM_000484). Two APP isoforms, NM_000484 and X06989, are present in most nonneuronal samples, e.g., melanoma and lung carcinoma. Tissues 5 to 14, all of which are brain regions, have high prediction scores for junctions 7 and 8, suggesting that exon 8 of NM_000484 is skipped in these tissues. Fetal brain tissue also has a high prediction score for junction 6, suggesting that exon 7 is also underrepresented in APP transcripts in that tissue. These predictions were also confirmed by RT-PCR and sequencing.

Junction array predictions also occurred in regions of genes not previously known to undergo AS, including the 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase (HMGCR) gene, which encodes the target of the statin class of cholesterol-lowering drugs. Previous work suggested the existence of a lower molecular-weight version of HMGCR that may be resistant to statins (*16*). Figure 3 shows that junctions 12 and 13 have a pattern consistent with an exon-skipping event in a large number of the samples. RT-PCR of this

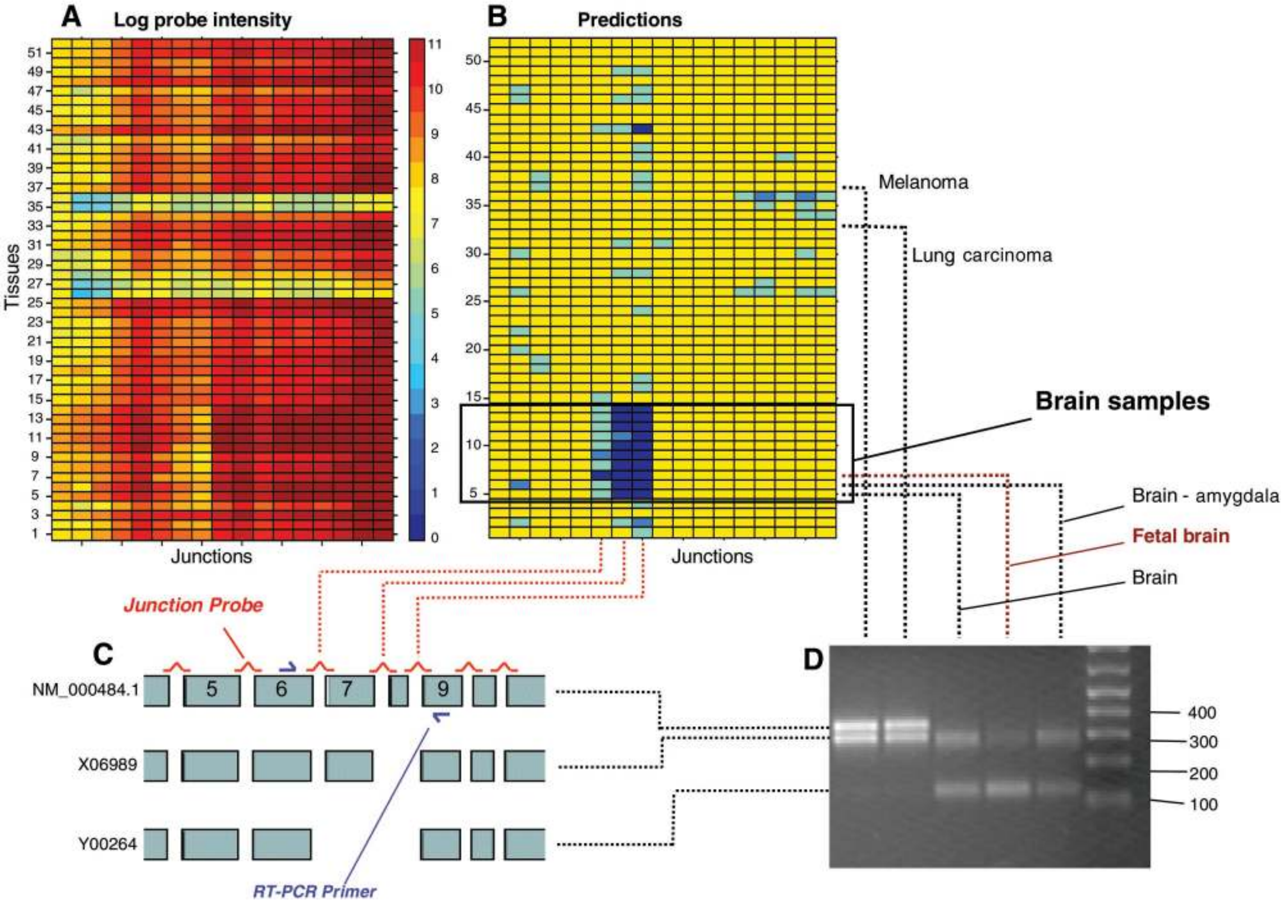

Fig. 2. Detection and validation of the known isoforms of APP. (**A**) Observed intensities and (**B**) AS prediction scores of junction probes designed to hybridize to NM_000484.1 with colors as in Fig. 1. Samples are ordered alphabetically and listed in table S3. (**C**) Exon-structure diagram of the three GenBank mRNA isoforms of APP. Red dotted lines link identical junctions. (**D**) RT-PCR results from a primer pair hybridizing to exons 6 and 9. Numbers to the right represent bp.

region showed two bands in most of the 44 tissues, one corresponding to the expected length and a shorter form. Sequencing revealed that the 255-nt fragment was missing exon 13, confirmed by isolating the full-length clone from a human liver library and later corroborated by deposition of GenBank clone 21707181. The novel isoform was present as a mixture with the known isoform in every human tissue tested, with the exception of peripheral leukocytes in which only the novel isoform band was visible. However, it is unlikely that the variant protein could catalyze the reaction of the longer isoform because it lacks a portion of the active site (fig. S2).

The OCRL1 and HMGCR examples illustrate a general observation that junction arrays detect many splicing events not detected by ESTs (*13*). One reason may be that lower-abundance transcripts are poorly represented by ESTs (*13*). To examine this possibility, we analyzed GenBank ESTs and mRNAs for AS events using previously described methods (*17*) and compared detection of alternatively spliced junctions by ESTs, mRNAs, and junction arrays as a function of gene-expression level. The results confirm that ESTs are biased against detecting AS events in genes with lower expression levels, showing a substantial increase in AS detections with expression level (Fig. 4A). In contrast, the fraction of junctions that are alternatively spliced in full-length mRNAs and junction arrays is largely independent of expression level, and the arrays are sensitive to detection of AS events for genes expressed at levels well below the mean. A second problem limiting EST-based analyses of AS is bias toward transcript termini. As shown in Fig. 4B, relatively few alternatively spliced junctions are detected by ESTs in regions of mRNAs distant from the 5′ or 3′ terminus. This most likely reflects undersampling in the centers of long transcripts due to nonuniformity in cDNA library construction and the use of end-sequence reads from these clones. This type of bias has been avoided here with a novel full-length RNA labeling protocol (*12*). New methods for sequencing ESTs, e.g., the ORESTES project (*18*), provide a partial remedy, but sampling inconsistency remains a major limitation for systematic studies of AS. Indeed, 20% of exon-exon junctions in RefSeq cDNAs are not sampled by an EST at all, and an additional 11% are represented by only one EST, precluding detection of AS. Results from the junction arrays suggest naturally occurring AS events are not concentrated at 3′ termini; however, the higher frequency of AS observed in ESTs and mRNAs in the 500 nt nearest the 5′ end is corroborated by the junction arrays, suggesting this is a true biological preference.

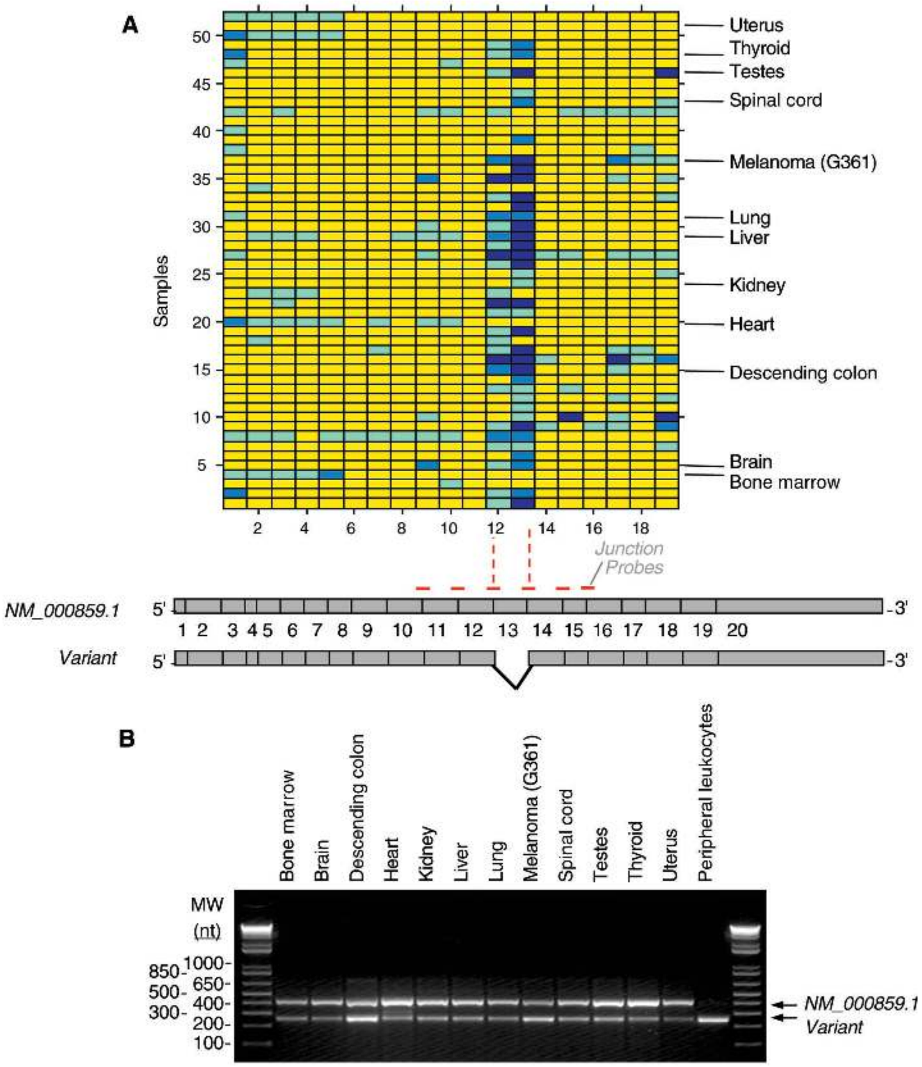

Fig. 3. Discovery of a novel isoform of HMGCR. (**A**) Junction array predictions of AS events at junctions 12 and 13. Probes representing each junction of NM_000859.1 are arranged in a 5′ to 3′ direction along the horizontal axis, with samples as in Fig. 2. (**B**) RT-PCR amplification of exons 12 to 15 from selected human tissues. The smaller, 255-nt fragment lacks exon 13.

We also investigated which genes and tissue samples contained the most AS events (table S1). Using Gene Ontology database (*19*) term frequencies, the set of 31 genes with the highest frequencies of AS was found to be significantly enriched for genes involved in cell communication (e.g., signal transduction through receptor tyrosine kinases) and enzyme regulation (e.g., small GTPase regulation) (*20*). The samples with the highest frequencies of AS events were cell lines; overall they had fewer genes expressed, but more AS events in those genes. Tissue samples were also clustered by the patterns of AS predictions in their expressed genes (fig. S3). Not surprisingly, similar tissues have similar patterns of AS, and the resulting clusters are similar to clusters of samples by gene expression levels. For example, liver pairs with fetal liver, stomach with duodenum, heart with skeletal muscle, and all neuronal tissues clustered together. Cell lines also form a separate cluster.

To validate junction array predictions of AS, we tested over 150 RT-PCR primer pairs, amplifying regions of 110 selected genes across a panel of samples similar to those in the array experiments. The RT-PCR validation was biased toward gene regions that contained array predictions and toward genes of therapeutic interest. Of 153 transcript regions evaluated, 134 did not contain AS events represented by mRNAs, and 91 did not have AS events represented by ESTs. Seventy-three alternative splices were sequence-verified in this validation exercise, 53 of them novel with respect to ESTs and mRNAs, including novel variants of many pharmaceutically relevant genes (table S2). Receiver operating characteristic (ROC) analysis (*21*) was used to assess the predictive power of the junction arrays as a discovery platform. Each primer pair was scored simply by the largest number of predictions observed for any single junction between the primers ($N_{\max}$), and various thresholds of $N_{\max}$ formed the ROC curve (fig. S4). The resulting classifier with $N_{\max} \geq 6$ achieves a specificity of 0.82 and a sensitivity of 0.49. In addition, 58% of primer pairs meeting this criterion produced sequence-validated alternative splice forms.

Using the validation success rate determined above, one can estimate the number of alternatively spliced genes identified by this 52-sample junction-array experiment that are not detected by ESTs. The junction arrays contain 3990 multi-exon genes with no EST evidence of AS at any of their exon-exon junctions. Among these genes, 1791 contain a junction meeting the $N_{\max} \geq 2$ array-prediction criterion, 798 of which would be validated by RT-PCR on the basis of the ROC analysis. Together with ESTs and mRNAs, this provides experimental evidence that 74% of human multi-exon genes are alternatively spliced. This could still be an underestimate, given that only 52 tissues were used and that RT-PCR validation is a conservative measure of array-detected AS events.

Detection of AS using junction arrays is limited in several ways. Like ESTs, junction arrays cannot determine whether two splicing events in one tissue are present in the same or separate transcripts. In addition, sequences of novel isoforms are not specified. Detection also requires differential expression; if two isoforms are present in the same proportion in every tissue, no predictions will result. Finally, cross-hybridization could cause false positives when sequence-similar genes have strong tissue-specific regulation. The resolution and sensitivity of this approach could be improved by adding probes in exons or adding probes to assay potential AS events (e.g., all single-exon skipping events).

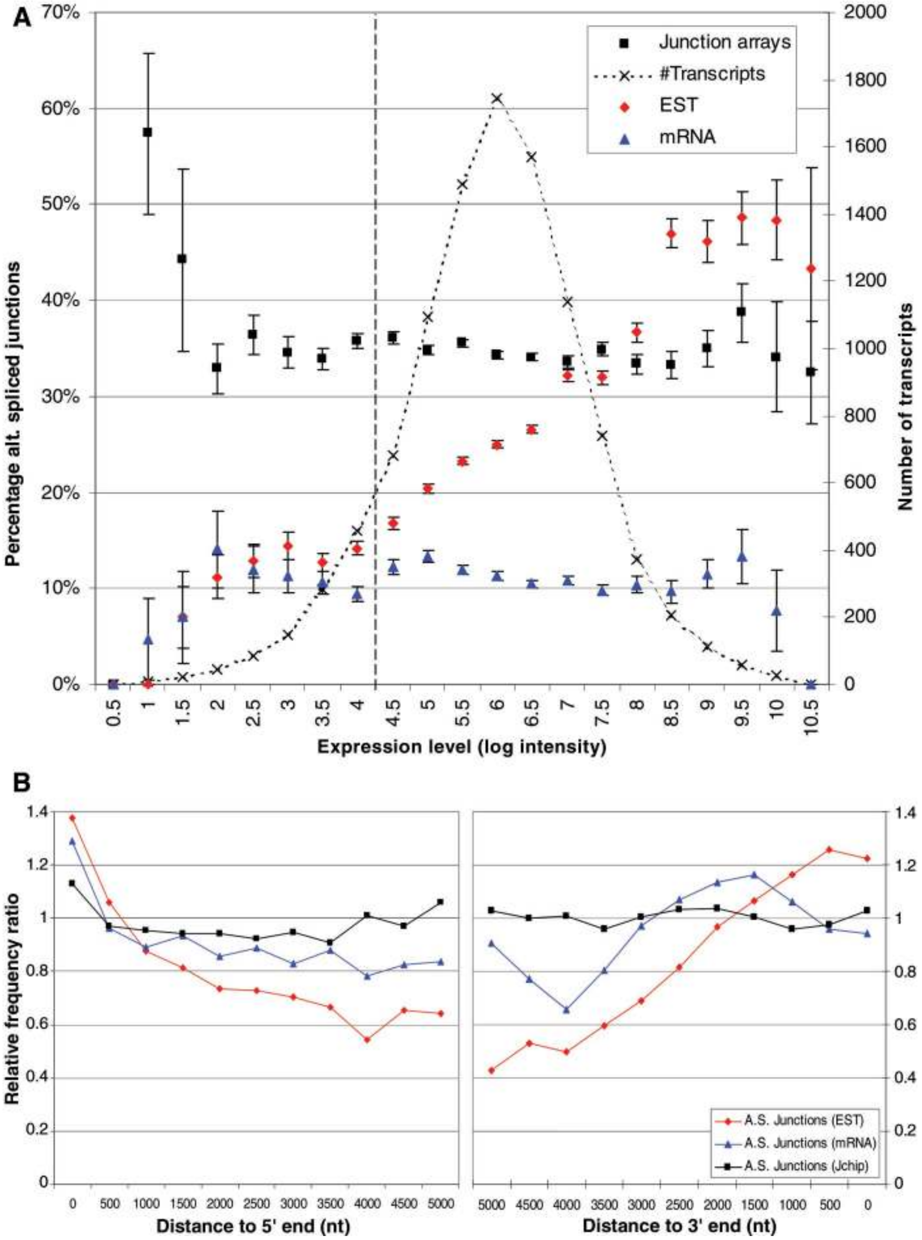

Fig. 4. (**A**) Percentage of exon-exon junctions found to be alternatively spliced as a function of transcript expression level for ESTs, mRNAs, and junction array data sets. Expression level is the natural log intensity of junction probes for each transcript averaged over all tissues. Most human transcripts have expression levels between 4 and 8 on this scale as shown by the histogram (dotted curve). The dashed vertical line marks 1 standard deviation above background, calculated from negative control spots. An array probe indicates AS at a junction if a prediction was observed in any one of the 52 tissues. EST and mRNA data sets indicate AS at a particular junction if they contain splicing events mutually exclusive at that junction (*20*). (**B**) Positional bias of AS detection for ESTs, mRNAs, and junction array data sets as a function of distance from the 5′ end (left) and 3′ end (right) of RefSeq transcripts. The relative frequency ratio (vertical axis) is the frequency of alternatively spliced junctions divided by the fraction of exon-exon junctions that occur in each bin. A ratio above 1 suggests that the frequency of AS in a particular distance bin is greater than expected by the average frequency for a given data set. The 5-kilobase distance range shown includes more than 93% of all exon-exon junctions in each panel.

References and Notes

1. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
2. B. Modrek, A. Resch, C. Grasso, C. Lee, *Nucleic Acids Res.* **29**, 2850 (2001).
3. M. Krawczak, J. Reiss, D. N. Cooper, *Hum. Genet.* **90**, 41 (1992).
4. R. L. Strausberg, E. A. Feingold, R. D. Klausner, F. S. Collins, *Science* **286**, 455 (1999).
5. M. Zavolan *et al.*, *Genome Res.* **13**, 1290 (2003).
6. B. Modrek, C. Lee, *Nature Genet.* **30**, 13 (2002).
7. J. M. Yeakley *et al.*, *Nature Biotechnol.* **20**, 353 (2002).
8. G. K. Hu *et al.*, *Genome Res.* **11**, 1237 (2001).
9. D. D. Shoemaker *et al.*, *Nature* **409**, 922 (2001).
10. T. A. Clark, C. W. Sugnet, M. Ares Jr., *Science* **296**, 907 (2002).
11. H. Wang *et al.*, *Bioinformatics* **19** (suppl. 1), I315 (2003).
12. J. Castle *et al.*, *Genome Biol.* **4**, R66 (2003).
13. D. L. Black, *Cell* **103**, 367 (2000).
14. R. Edgar, M. Domrachev, A. E. Lash, *Nucleic Acids Res.* **30**, 207 (2002).
15. R. L. Nussbaum, B. M. Orrison, P. A. Janne, L. Charnas, A. C. Chinault, *Hum. Genet.* **99**, 145 (1997).
16. N. Aboushadi, J. E. Shackelford, N. Jessani, A. Gentile, S. K. Krisans, *Biochemistry* **39**, 237 (2000).
17. Z. Kan, D. States, W. Gish, *Genome Res.* **12**, 1837 (2002).
18. A. A. Camargo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12103 (2001).
19. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
20. Materials and methods are available as supporting material on *Science* Online.
21. J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic Press Series in Cognition and Perception, New York, 1975).

22. We thank L. Lim, G. Cavet, B. Yerkovich, R. Raubertas, F. Roth, and D. Haynor for helpful discussions and manuscript reviews and S. Carlson, S. Dow, J. Guo, E. Apolonio, Z. Riley, M. McWharter, E. Coffey, M. Marton, and the Gene Expression Lab for technical and project support.

Supporting Online Material
www.sciencemag.org/cgi/content/full/302/5653/2141/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 to S3
References

5 August 2003; accepted 30 October 2003