Supporting Text

Sequence Data Sets. Chromosome assemblies of the human genome (hg13) and the mouse genome (mm3) were obtained from the University of California, Santa Cruz, Genome Browser, http://genome.ucsc.edu. The transcript data we used included ≈94,000 human cDNA and ≈91,600 mouse cDNA sequences obtained from GenBank (release 134.0; flatfiles in categories gbpri, gbrod, and gbhtc) and ≈5×10^6 human ESTs and ≈3.5×10^6 mouse ESTs from dbEST (repository 032703). The GENOA genome annotation script (http://genes.mit.edu/genoa) was used for spliced alignment of cDNA sequences and ESTs to the human and mouse genomes. GENOA detected matches of significant blocks of identity between a repeat-masked cDNA sequence and genomic DNA using BLASTN (1). Matched pairs are then aligned by using the spliced alignment algorithm, MRNAVSGEN (http://genes.mit.edu/genoa). Subsequently, ESTs were aligned to cDNA-verified genomic regions by using SIM4 (2). For inclusion in the final GENOA annotation, all ESTs were required to overlap one or more cDNAs, and the first and the last segments of the spliced alignment were required to exceed 30 nucleotides in length with 90% sequence identity. In addition, the entire EST sequence alignment was required to extend >90% of the sequence length and have >90% sequence identity.

Overall, GENOA aligned ≈86,000 human cDNAs and ≈890,000 human ESTs and ≈27,000 mouse cDNAs and ≈483,000 mouse ESTs. Genes with multiple cDNA alignments were resolved into separate gene loci containing single genes and candidate regions with alternative exon–intron structures. 5'-terminal and 3'-terminal exons were separated from internal exons and excluded from further analyses, because they possess different splicing characteristics and sequence composition from internal exons. Exons were categorized as constitutive exons, alternative 3' splice site (3'ss) exons, alternative 5'ss exons, skipped exons, multiply alternatively spliced exons (e.g., exons observed to undergo exon skipping and alternative 5'ss usage), and exons containing retained introns. Genes with at least one identified alternative splicing (AS) event were categorized as AS genes; all other genes were considered constitutively spliced (CS) genes. An exon was defined as a skipped exon (SE) if it was included in one or more transcripts and excluded at least one

other transcript. Specifically, a transcript aligned such that the 3' end of the corresponding upstream exon and the 5' end of the corresponding downstream exon were juxtaposed was considered as evidence of exon skipping. Human and mouse SEs were identified independently by using transcript data specific to each organism. Human/mouse orthologous gene pairs were taken from ENSMART and ENSEMBL, version 16 (3). Reciprocal best BLAST hits were used to identify orthologous human–mouse exons within these orthologous genes. Spliced alignment of ESTs to cDNA-verified regions of assembled human and mouse genomic sequences was used to infer splicing patterns of exons.

Exon–Intron Sequence Regions and Feature Extraction. The following sequence features were extracted for each conserved human–mouse exon pair: exon length, upstream intron length, downstream intron length, 5'ss (donor site) and 3'ss (acceptor) scores, exon conservation (percent identity), upstream and downstream 150-base intron region conservation [CLUSTALW alignment score (4)], and a list of oligonucleotide occurrence counts, described below. Length features were transformed to logarithmic ($\log_{10}$) scale, and splice sites were scored by using a maximum entropy model (5). Exons were divided into four different regions: the last 150 bases of the upstream intron (or the entire intron for introns of <150 bases), the first 150 bases of the downstream intron (or the entire intron), the first 100 bases of the exon (or the entire exon), and the last 100 bases of the exon (or the entire exon). Occurrence counts for all oligonucleotides of length k for k ranging from 3 to 6 nt were calculated from the four regions described above. Counts were generated separately from unaligned and CLUSTALW-aligned regions. In either case, all overlapping k-mers contained completely in the given region were counted. k-mers that occurred less than twice in the $S_{H,M}$ and $S_{h,m}$ training sets were excluded from further analysis. For training of ACESCAN, k-mers were ranked by enrichment in $S_{H,M}$ versus $S_{h,m}$ exons and their flanking introns, as scored by using a χ^2 statistic for a 2×2 contingency table, with Yates correction factor (6). For each region in $S_{H,M}$ and $S_{h,m}$ exons (rows of contingency table), the number of occurrences of each k-mer and the number of occurrences of all remaining k-mers were determined (table columns). The oligonucleotide features were ranked, and the top N features were

extracted and concatenated into a (*M*+*N)*-dimensional vector, where *M* is the number of general sequence features used. The top-ranked oligonucleotide features used by ACESCAN included some that were 5 mers and some that were 4 mers (Table 1) but none that were 3 or 6 mers.

Known cis-Elements in High-Ranking Oligonucleotides. The motifs UGCAU and GCAUG were found to be overrepresented in the upstream and downstream introns, flanking exons subjected to conserved skipping (Fig. 2*B*). Similar sequences, e.g., the hexamer UGCAUG, are known to be involved in the regulation of splicing the *c-src*, fibronectin, nonmuscle myosin heavy chain, and calcitonin genes (7-10). The UCUCU pentamer, which is similar to sequences involved in splicing repression in the neural-specific N1 exon of the *c-src* transcript (11), was also identified as overrepresented in the introns upstream of $S_{H,M}$ exons and in the exons themselves. A number of other U-rich sequences were also overrepresented in upstream introns, consistent with previous observations (12, 13). The sequence UAGGG, which forms a portion of the consensus heterogeneous nuclear ribonucleoprotein (hnRNP) A1 binding site and can act in negative regulation of splicing (14), was also overrepresented in $S_{H,M}$ exons relative to unskipped exons. Motifs related to GUAGU, also overrepresented in $S_{H,M}$, have been validated as exonic splicing silencers (ESSs) in cultured human cells (15). On the other hand, two pentamers that were underrepresented in $S_{H,M}$ relative to $S_{h,m}$, CUGGA and AGAAG, resemble consensus ESEs (UGGA and GAGAAG, respectively) identified in previous analyses (16, 17). In fact, more detailed analyses suggest that a significantly higher fraction of oligonucleotides enriched in $S_{h,m}$ matched computationally predicted and experimentally validated ESEs (16) as compared with oligonucleotides enriched in $S_{H,M}$ (Table 2). A lower density of ESEs in skipped exons relative to constitutive exons is likely to reflect differing selective pressures, with constitutive exons selected for efficient inclusion, and skipped exons selected for less efficient inclusion, at least under some circumstances (e.g., in specific tissues or developmental stages).

Classification, Cross-Validation and Sampling. The regularized least-squares classifier (RLSC) was used to learn the features from $S_{H,M}$ and $S_{h,m}$. The RLSC has a quadratic loss

function and requires the solution of a single system of linear equations (18). Because of the unbalanced size of the two sets, i.e., there were ≈25-times more exon pairs in $S_{h,m}$ (negative examples) than in $S_{H,M}$ (positive examples), errors made on the positive examples cost a multiplicative factor of β times greater than the penalty for errors made on the negative examples. The binary-class RLSC classification problem was stated as

$$\min_f \left(\frac{1}{L}\right) \sum_{i=1}^{L} w_i (y_i - f(x_i))^2 + \lambda \|f\|_K^2, \qquad [\mathbf{1}]$$

where f and $||f||^2_K$ are the function and function norm induced in a reproducing kernel Hilbert-space respectively, L is the size of the training set, λ is the "tradeoff" between generalization and overfitting, and w_i is a misclassification penalty set to β if sample x_i had a positive label ($y_i = 1$) and otherwise set to 1. To address the potential for incorrect labeling of $S_{h,m}$ exons because of incomplete coverage by transcript data, the misclassification parameter β for positively labeled data were set to 5, higher than the value for negatively labeled data. Assuming a solution f^* of the form

$$f^*(u) = \sum_{i=1}^{L} c_i \mathrm{K}(u, x_i), \qquad [\mathbf{2}]$$

where $K(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{u},\boldsymbol{v} \rangle$, c_i are coefficients, K is the $L \times L$ kernel matrix satisfying $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and **W** is the diagonal matrix of penalties w_i, the problem was rewritten in matrix notation and the optimal **c**, defining $\boldsymbol{c} = [c_1 \ \ldots \ c_l]^{\mathrm{T}}$ was found, by substituting Eq. **2** into Eq. **1**:

$$(\mathrm{K} + \lambda \mathrm{L} \mathrm{W}^{-1})\mathrm{c} = y. \qquad [\mathbf{3}]$$

Fixing λ and β and solving for **c** by using the conjugate gradient method (implemented in MATLAB), test examples were assigned an output according to Eq. **2**. To solve Eq. **3** efficiently, K was expressed as $\mathbf{A}\mathbf{A}^{\mathbf{T}}$, where **A** was the $L \times d$ matrix of training examples with d features. By first computing $\alpha = \mathbf{A}^{\mathbf{T}}\mathbf{c}$, the outputs for unlabeled ("test") examples

were obtained by matrix multiplication of **B** and α, where **B** is the $n \times d$ matrix of n unlabeled examples.

Cross-validation was used: for each model, 80% of the exon pairs from $S_{H,M}$ and 80% of the pairs from $S_{h,m}$ were used to train the classifier, which then assigned outputs (predicted classifications) to the remaining 20% of unseen exon pairs. The performance of different models was averaged over 50 iterations of sampling training and test subsets. Area under the curve (AUC) values were obtained for each iteration, and the average AUC value was used to measure model performance (described below). Empirically, it was found that $\lambda = 0.01$ and $\beta = 5$ gave optimal performance. Empirically, it was also determined that the model labeled ***i*** in Fig. 5 obtained the highest AUC value, at a cutoff of approximately –0.5. The ACESCAN score for an exon pair was defined as the mean prediction output over 500 random samples of the training set. Similarly, when ACESCAN was used to score unseen ENSEMBL-annotated human–mouse exon pairs (i.e., exon pairs not in the training set), each pair was assigned an ACESCAN score calculated as the mean output from 50 random samples of the training data from $S_{H,M}$ and $S_{h,m}$. The approach of taking the average output from many different samplings of the training set corresponds closely to the use of "bagging" in statistical machine learning (19). The set of ACESCAN[+] exons will be made available on the Internet at http://genes.mit.edu/acescan at the time of publication.

Performance Measures. Receiver operating characteristic (ROC) curve analysis (20) was used to assess the performance of models in binary hypothesis testing. A ROC plot graphically represents the true positive rate (on the y axis) versus the false positive rate (x axis) as a function of the threshold used in prediction and displays the tradeoff between the sensitivity and the false positive rate (increases in sensitivity are generally accompanied by an increase in false positives). The integrated area under the ROC curve (AUC) was used to measure performance (higher AUC values correspond to improved classification performance).

Gene Ontology (GO) Analysis. GO identifiers (IDs) for each ENSEMBL-annotated gene were obtained from ENSMART (release version 19.1). Organizational principles ("molecular function" and "biological process") were obtained from http://www.geneontology.org. For each term (e.g., neurogenesis, GO ID 0007399), the fraction of genes containing predicted ACEs and not containing predicted ACEs relative to the genes under the overall principle (e.g., GO ID 0007399 was found under biological process) was compared by a χ^2 test of significance, with Yates correction factor (6). Adjusting for multiple hypothesis testing by using Bonferroni correction (6) (217 terms were compared with at least 10 genes belonging to the term for molecular function; 187 terms were compared for biological process), enriched terms were identified at a significance cutoff of $P < 0.05$.

Gene Expression Analysis. Affymetrix HG-U95A microarray gene expression from 47 human tissues and cell lines previously published by Su and colleagues (21) were obtained from the Gene Expression Atlas (http://expression.gnf.org). Mappings for Affymetrix probe identifiers were obtained from ENSMART (release 19.1). Average difference (AD) values at <20 were standardized to 20, as described in ref. 21. Genes expressed in a tissue or cell line at >2-times the standard deviation above the median expression across tissues or cell lines were defined as tissue-specifically expressed in that tissue or cell line. For each tissue, the fraction of genes containing predicted ACEs and not containing predicted ACEs relative to the set of all tissue-specifically expressed genes was compared by using a χ^2 test, with the Yates correction factor (6). Adjusting for multiple hypothesis testing by using Bonferroni correction, enriched tissues were again identified at a significance cutoff of $P < 0.05$.

SNP Analysis. We obtained 8,408 high-quality reference SNPs (33-mer centered on SNP) (22) and mapped them to exons scored by ACESCAN. The SNP density for a set of exons was calculated by dividing the total number of SNPs contained in the exons by the total length of all exons within the set.

Protein Domain Analysis. Human ENSEMBL transcripts and ENSEMBL annotated PFAM protein features (23) were obtained from the ENSMART database (ENSEMBL version 22.34). The start and end locations of each annotated protein feature with respect to the translated transcript were obtained and compared with the coordinates of the exons in the transcript. A protein feature was considered to overlap an exon if *W* bases or more of the exon was within the feature. *W* was adjusted from 5 to 30 bases in steps of 5 bases to test the robustness of the measurement. A χ^2 test was performed to determine whether ACESCAN[+] exons overlapped exons at a lower or higher rate compared with low-scoring ACESCAN[–] and S_h exons.

Experimental Validation. The Invitrogen Superscript III First-Strand synthesis system for RT-PCR (catalog no. 18080-051) was used to generate cDNAs from 3-4 μg of total RNA from human tissues (whole brain, fetal brain, heart, fetal liver, cerebellum, prostate, liver, lung, kidney, skeletal muscle, bone marrow, and testis) and mouse tissues (whole brain, testis, liver, lung, skeletal muscle, kidney, heart, and a pool from embryonic 5-, 11-, 15-, and 17-day tissues) from Clontech by using oligo(dT) primers. The Invitrogen *Taq*DNA polymerase kit (catalog no 18038-042) was used with primers designed by using the PRIMER3 program (24) targeted to exons flanking candidate ACEs. Forty cycles of PCR using an ABI 9700 thermocycler were conducted at denaturing temperature of 94°C for 30 s, annealing at 58°C for 30 s, and elongation at 72°C for 30–100 s, depending on the size of the predicted products. PCR products were resolved on a 2% agarose gel (Merck) at 116 volts in TBE buffer. Bands of the expected size were gel-purified with the QIAquick Gel Extraction kit (catalog no. 28704, Qiagen, Valencia, CA) according to the manufacturer's instructions. Each isolated band was amplified by additional rounds of PCR with the same primers before sequencing.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.

2. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. (1998) *Genome Res.* **8,** 967–974.

3. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al*. (2002) *Nucleic Acids Res.* **30,** 38–41.

4. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.

5. Yeo, G. & Burge, C. B. (2004) *J. Comput. Biol.* **11,** 377–394.

6. Glantz, S. A. (1997) *Primer of Biostatistics* (McGraw–Hill, New York).

7. Huh, G. S. & Hynes, R. O. (1994) *Genes Dev.* **8,** 1561–1574.

8. Modafferi, E. F. & Black, D. L. (1997) *Mol. Cell. Biol.* **17,** 6537–6545.

9. Black, D. L. (1992) *Cell* **69,** 795–807.

10. Brudno, M., Gelfand, M. S., Spengler, S., Zorn, M., Dubchak, I. & Conboy, J. G. (2001) *Nucleic Acids Res.* **29,** 2338–2348.

11. Chan, R. C. & Black, D. L. (1997) *Mol. Cell. Biol.* **17,** 2970.

12. Shibata, A., Hattori, M., Suda, H. & Sakaki, Y. (1996) *Gene* **175,** 203–208.

13. Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P. & Valcarcel, J. (2000) *Mol. Cell* **6,** 1089–1098.

14. Kashima, T. & Manley, J. L. (2003) *Nat. Genet.* **34,** 460–463.

15. Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. & Burge, C. B. (2004) *Cell* **119,** 831–845.

16. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002) *Science* **297,** 1007–1013.

17. Liu, H. X., Zhang, M. & Krainer, A. R. (1998) *Genes Dev.* **12,** 1998–2012.

18. Rifkin, R., Yeo, G. & Poggio, T. (2003) in *Advances in Learning Theory: Methods, Model and Applications*, ed. Suykens, H., Basu, Micchelli, Vandewalle (IOS, Amsterdam), Vol. 190, pp. 131–154.

19. Duda, R. O., Hart, P. E. & Stork, D. G. (2001) *Pattern Classification* (Wiley, New York).

20. Swets, J. A. (1988) *Science* **240,** 1285–1293.

21. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., *et al*. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4465–4470.

22. Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. (2004) *PLoS Biol.* **2,** E268.

23. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al*. (2004) *Nucleic Acids Res.* **32,** D138–D141.

24. Rozen, S. & Skaletsky, H. (2000) *Methods Mol. Biol.* **132,** 365–386.