Identification, Improved Modeling and Integration of Signals to Predict Constitutive and Alternative Splicing

by

Gene W. Yeo

BSc Chemical Engineering

BA Economics

University of Illinois, Urbana-Champaign, 1998

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[February 2005]

NOVEMBER 4, 2004

Signature of Author: _____

Department of Brain and Cognitive Sciences

November 4, 2004

Certified by: _____

Tomaso Poggio

Professor, Brain and Cognitive Sciences, Thesis co-supervisor

Certified by: _____

Christopher Burge

Associate Professor, Biology, Thesis co-supervisor

Accepted by: _____

Mriganka Sur

Professor of Brain and Cognitive Sciences, Department Head

Identification, Improved Modeling and Integration of Signals to Predict Constitutive and Alternative Splicing

by

Gene W. Yeo

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES ON NOVEMBER 4, 2004 IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY IN BRAIN AND COGNITIVE SCIENCES

ABSTRACT

The regulation of pre-messenger RNA splicing by the spliceosomal machinery via interactions between *cis*-regulatory elements and splicing *trans*-factors to generate a specific mRNA i.e. constitutive splicing, or sometimes many distinct mRNA isoforms i.e. alternative splicing, is still a poorly understood process. Progress into illuminating this process is further exacerbated by the variation of splicing in the multitude of tissues and cell types present, as well as the variation of *cis* and *trans* elements in different organisms, and the possibility that some alternative splicing events present in expressed sequence tag (EST) databases may constitute biochemical 'noise' or transient evolutionary fluctuations.

Several studies, mainly computational in nature, addressing different questions regarding constitutive and alternative splicing are described here, ranging from improved modeling of splicing signals, studying the variation of alternative splicing in various tissues, analyzing evolutionary differences of *cis* and *trans* elements of splicing in various vertebrates, and utilizing attributes indicative of alternative splicing events conserved in human and mouse to identify novel alternatively spliced exons.

In particular: (i) A general approach for improved modeling of short sequence motifs, based on the Maximum Entropy principle, that incorporates local adjacent and non-adjacent position dependencies is introduced, and applied to understanding splice site signals. The splice site recognition algorithm, MaxENTScan, performs better than previous models that utilize as input similar length sequences; (ii) The first large-scale bioinformatics study is conducted that identifies similarities and differences in candidate *cis*-regulatory elements and *trans*-acting splicing factors in vertebrate genomes, resulting in the manipulation of intronic elements that enables fish genes to be spliced properly in mammalian cells; (iii) A computational analysis using EST data, genome sequence data, and microarray expression data of tissue-specific alternative splicing is conducted, which distinguishes human brain, testis and liver as having unusually high levels of AS, highlights differences in the types of AS occurring commonly in different tissues, and identifies candidate *cis*-regulatory elements and *trans*-factors likely to play important roles in tissue-specific AS in human cells; (iv) The identification of a set of discriminatory sequence features and their integration into a statistical machine-learning algorithm, ACEScan, which distinguishes exons subject to evolutionarily conserved alternative splicing from constitutively spliced or lineage-specifically-spliced exons is described; (v) The genome-wide search for and experimental validation of exon-skipping events using the combination of two silencing *cis*-elements, UAGG and GGGG.

Thesis Co-supervisor: Christopher B. Burge
Title: Associate Professor of Biology
Thesis Co-supervisor: Tomaso A. Poggio
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

I gratefully acknowledge my mentors Chris Burge and Tomaso Poggio, and my thesis committee members Phillip Sharp and Martha Constantine-Paton for their patience, guidance and advice.

I acknowledge my past and current colleagues in the Poggio lab (Alex Rakhlin, Gabriel Kreiman, Neha Soni, Ryan Rifkin, Shayan Murkherjee); in the Burge lab (Benjamin Lewis, Brad Friedman, Cydney Nielson, Dirk Holste, Eric van Nostrand, George Huo, Lee Lim, Luba Katz, Michael Stadler, Mehdi Yahyanejad, Michael Rolish, Namjin Chung, Noam Shomron, Rong Kong, Uwe Ohler, Vivian Tung, Zefeng Wang), and in the Sharp lab (Anthony Leung, Chonghui Cheng, Will Fairbrother) for the intellectual swashbuckling, fun collaborations and the camaraderie that I truly enjoyed.

I acknowledge Brayppa Venkatesh (IMCB, Singapore), Paula Grabowski (University of Pittsburg), Thomas Cooper (Baylor College of Medicine), Michael Koeris from Alex Rich's laboratory, and Richard Gatti (UCLA) for providing the opportunities for fruitful collaborations.

I also gratefully acknowledge Sydney Brenner for getting me hooked on Science, and for reminding me that science is good fun.

I acknowledge the Lee Kuan Yew Foundation for the Lee Kuan Yew Fellowship that supported me throughout my time at MIT, and Loh Wai Kiew and Mdm Ho Ching for their support.

Most importantly, I thank my parents, Philip and Jane Yeo, and my sister, Elaine, as well as Julia Powers for their encouragement and support throughout my time here. They have been the pillars of my existence!

Biographical Note

Gene W. Yeo

EDUCATION:

University of Illinois, Urbana-Champaign, Urbana-Champaign, IL
Bachelor of Science, Chemical Engineering (Highest distinction), Jan, 1998

University of Illinois, Urbana-Champaign, Urbana-Champaign, IL
Bachelor of Arts, Economics (High distinction), Jan, 1998

AWARDS:

Lee Kuan Yew Scholarship for graduate study (2000-2005) awarded by the Lee Kuan Yew Foundation in Singapore.
Sword of Honor, Top Naval Officer Cadet, Republic of Singapore Navy, 1999.
James Scholar, College of Liberal Arts and Sciences (Academic years 1996, 1998)
Senior 100 Honor Award (1996, 1997) for demonstrated contribution to the University through both extracurricular and scholastic achievement.
Hauser Chemical Engineering Scholarship for research (1996).
A.T. Widiger Chemical Engineering Scholarship (1996) for outstanding scholarship in Chemical Engineering and vision of the larger role the profession plays in society.
Dean's list, College of Engineering (1994).
Dean's list, College of Liberal Arts and Sciences (1995-1997).
Chemical Engineering Alumni Award (1997) for leadership, scholastic achievement and campus activities.

PROFESSIONAL MEMBERSHIPS:

The Honor Society of Phi Kappa Phi; The American Institute of Chemical Engineers; The Phil Lambdas Upsilon Honorary Chemical Society; The Tau Beta Pi National Engineering Honor Society; The Phi Beta Kappa Honor Society; The Golden Key Honor Society.

PUBLICATIONS:

Yeo, G, Holste, D, Van Nostrand, E, Poggio, T, Burge, C.B. Identification and analysis of alternative conserved exons in human and mouse. Submitted. 2004

Yeo, G, Hoon S, Venkatesh, B and Burge, C.B. Variation in the splicing regulatory elements and their organization in vertebrate genomes. Proceedings of the National Academy of Science, United States of America, 2004.

Yeo, G, Holste D, Kreiman, G, and Burge, C.B. Variation in alternative splicing across human tissues. **5**:R74. Genome Biology, 2004.

Han, K, **Yeo, G**, Burge, C.B and Grabowski, P. Exon-skipping by combinations of UAGG and GGGG motifs. Submitted. 2004.

Yeo, G and Burge, C. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, Journal of Computational Biology, **11**, No. 2-3, p 377-395, 2004.

Yeo, G and Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin, Germany, April 10-13, 2003.

Eng, L, Coutinho G, Nahas, S, Yeo, G, Tanouye, R, Drk, T, Burge, C.B, and Gatti, R.A. Non-classical splicing mutations in the coding and non-coding regions of the ATM gene: a comparison of cDNA with maximum entropy estimates of splice junction strengths, Human Mutation, 23(1), 67-76, 2004.

Rifkin, R, Yeo, G and Poggio, T. Regularized Least-squares Classification. Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and System Sciences, Vol. 190, ISO Press, Amsterdam 2003. Edited by Suykens, Horvath, Basu, Micchelli and Vandewalle.

Yeo, G and Poggio, T. Multiclass classification of small round blue cell tumors, CBCL Paper #204, AI MEMO #2001-018, MIT, 2002.

First or co-first author papers denoted in **bold.**

CONFERENCE PRESENTATIONS: Yeo, G, Holste D, Van Nostrand, E, Poggio, T and Burge, C.B. Predictive discrimination of conserved skipping events in human and mouse. Bioinformatics Workshop. Ninth annual meeting of the RNA society, June 1-6, Madison, Wisconsin, 2004.

Yeo, G, Hoon, S and Burge C. Variation in sequence and organization of splicing regulatory elements in vertebrate genes, Eukaryotic mRNA processing meeting, Aug 20-24, Cold Spring Harbor Laboratory, 2003.

Yeo, G, Hoon, S and Burge C. Genomics of vertebrate splicing

regulatory elements, Intelligent systems for molecular biology conference, June 29-July 3, Brisbane, Australia, 2003. Best Poster award.

Yeo, G, and Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Seventh annual international conference on research in computational molecular biology, April 10-13, Berlin, Germany, 2003.

PROFESSIONAL EXPERIENCES:

Bioinformatics intern, Millennium Predictive Medicine (MPMX), May-August, 2001
Developed and critiqued millennium in-house web-tools for classification and feature selection for marker selection in large-scale microarray data.
Integrated microarray data and clinical data using Bayesian networks.

Research technician, Research and Development, Chiron, Emeryville, July-August, 2000
Generated myc-tagged recombinant human gene-plasmid construct used in mammalian cell transfection. Generated GST-fusion proteins for use in the development of antibodies to human genes associated in wnt signaling pathway.

Research technician, Affymetrix, Santa Clara, May-Aug 1997
Operation of Applied Biosystems 392 DNA/RNA synthesizer to synthesize nucleotides on slides in optimizing photo-resist conditions for chip manufacture. Used modified confocal fluorescent microscope and silanated acetoxy/sodalime slides.

Research assistant, TECH semiconductors, Singapore, 1994
Development of graphical and mathematical model to predict total organic carbon spikes in deionized water treatment plant.

TEACHING EXPERIENCES:

Teaching assistant, Course 9.02, Brain laboratory (by Sonal Jahveri, Jim diCarlo and Christopher Moore, Dept of Brain Cognitive Sciences), Spring 2004.
Assisted the teaching of rat brain dissections using a microtome, and histological and immunochemical staining of brain sections.
Aided students in preparing technical lab reports.
Graded lab reports and evaluated student oral presentations.

Teaching assistant, Course 9.00, Introduction to Psychology (by Jeremy Wolfe, Dept of Brain and Cognitive Sciences, MIT), Fall 2002.
Conducted recitations and graded homeworks and examinations.

Teaching assistant, Course 9.35, Vision (by Bart Anderson, Dept of Brain and Cognitive Sciences, MIT), Spring 2002.
Conducted recitations and graded homeworks and examinations.

Lecturer, Bioinformatics Essentials Graduate Certificate, Northeastern University. 2001.
Structured course syllabus and co-taught a 12 week course in microarray data analysis, including classification, feature selection, experimental design, Bayesian networks and microarray technology.

Undergraduate teaching assistant, Mass Transfer Course, University of Illinois, Urbana-Champaign, 1997.
Conducted discussion sections on diffusion, mass transfer coefficients, heat and mass transfer, distillation and absorption, membrane separations, and chemical reactions in mass transfer.

RESEARCH EXPERIENCE:

Research Assistant, Venkatesh, B and Brenner, S. Lab at Institute of Molecular and Cell Biology, Singapore, May-June 2000.

Undergraduate research assistant, Ceramics group at Beckman Institute of Advanced Science and Technology, University of Illinois, Urbana-Champaign, June 1996-Dec 1997
Conducted studies on Thin Film Drying stresses using an automated ellipsometer
Conducted studies on the effect of electric fields on 20 micro gold-coated glass spheres using an atomic force microscope.

Contents

Chapter 1

Introduction

Computational approaches to addressing questions in molecular biology have in recent years benefited greatly from the accumulation of copious quantities of messenger RNA (mRNA) microarray data, sequence transcripts in the form of expressed sequence tags (EST) and complementary DNA (cDNA) sequences, as well as the rapid sequencing, assembly and annotation of multiple complete genomes. The field of splicing has also become infused by computational methods, which integrate some of the above data, and addresses relationships between evolution, sequence, tissue specificity, gene expression, and structure.

This thesis addresses several open questions in splicing, describing: (i) improved modeling of classical splicing *cis*-regulatory elements, such as splice sites, and applications of splice site models to predicting splicing phenotypes that lead to a particular disease; (ii) the identification of features predictive of constitutive splicing that differ between fish and mammals, leading to engineered changes that affect cross-species splicing phenotypes; (iii) the documentation of differences in both presence/absence of particular *cis*-elements and expression of *trans*-factors with regards to alternative splicing in human tissues; (iv) the prediction and experimental validation of evolutionarily conserved alternatively spliced exons using sequence features integrated into a regularized large-margin classifier; and (v) genome-wide searches for exon-skipping

events predicted by combinations of UAGG and GGGG motifs. The techniques described here encompass a variety of statistical and computational methods, ranging from large-margin classifiers such as support vector machines, and regularized least-squares classification, to efficient extraction of enriched sequence motifs from DNA sequences. In addition, experimental techniques such as reverse-transcriptase polymerase chain reaction (RT-PCR) and sequencing are used extensively to validate predicted alternatively spliced exons. In essence, the biological context of the thesis covers evolutionary consequences of conserved alternative splicing, the relationship between *cis* and *trans* regulators of splicing in particular environments, and has implications for the function of predicted evolutionarily conserved alternative exons in the brain, in particular in development and neurogenesis, as well as in genes involved with RNA and nucleic acid binding.

1.1 Organization

My thesis is mainly a compilation of five large self-contained pieces of work. As each study addresses a different aspect of splicing, the detailed background and motivation for each study is incorporated into the chapter in which the study is described in detail. Chapter 2 deals with splicing regulatory elements in two parts. The first part of chapter 2 consists of improved modeling of the most classical splicing signals i.e. the splice sites in human, and the applications of the splice site models in predicting the splicing phenotype of non-classical splicing mutations. These works have recently been published [1, 2]. In addition, applications of the improved models of splice sites in finding enriched elements in pseudo-exons, and in splicing simulation is discussed in

brief (in press, Wang, Z et al., *Cell*, 2004). The second part of chapter 2 consists of identifying exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs) in various vertebrate genomes, and finding variation in these elements and their corresponding associated classes of splicing *trans*-factors in fish versus mammals. This work has recently been published [3]. Chapter 3 consists of a rigorous approach to addressing the differences in alternative splicing patterns, overrepresented exonic motifs, and expression of *trans*-factors in multiple human tissues using a combination of available transcript data and microarray expression data. A novel metric for quantifying dissimilarity between alternative isoforms is also described. This work has recently been published [4]. Chapter 4 deals with the prediction of functional evolutionarily conserved alternative exons. The first part of chapter 4 describes the identification of features predictive of alternatively spliced exons conserved in human and mouse, and their integration into a regularized classifier to scan for novel alternative exons in the human and mouse genomes. The algorithm, called ACEScan (Alternative-Conserved Exon Scan) has been applied to multiple vertebrate genomes using large-scale multi-species alignments. Experimental validation of predicted candidates demonstrates the high sensitivity of this method. This work has been submitted (Yeo et al, 2004). Part two of chapter 4 describes collaborative work involving the identification of two motifs, UAGG in the exonic region, and GGGG in the first 10 bases of introns, that can cause exon-skipping, and the identification of the *trans*-factors that are involved. This work has been submitted (Han, K et al, 2004). Chapter 5 gives a brief overview of the current technologies for large-scale profiling of alternative splicing, and the design and

implementation of a splicing-specific microarray. Chapter 6 summarizes key findings and describes the future outlook for this work.

References

1. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.** *J Comput Biol* 2004, **11**(2-3):377-394.
2. Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, Dork T, Burge C, Gatti RA: **Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths.** *Hum Mutat* 2004, **23**(1):67-76.
3. Yeo G, Hoon S, Venkatesh B, Burge CB: **Variation in sequence and organization of splicing regulatory elements in vertebrate genes.** *Proc Natl Acad Sci U S A* 2004, **101**(44):15700-15705.
4. Yeo G, Holste D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues.** *Genome Biol* 2004, **5**(10):R74.

Chapter 2

Modeling and identifying splicing elements

The majority of protein-coding genes in higher eukaryotes consist of islands of coding regions i.e. exons, interrupted by oceans of non-coding, poorly conserved sequences i.e, introns, which are replete with repeat elements and potential regulatory signals. Pre-messenger RNA undergoes the post-transcriptional process of splicing, whereby introns are removed and exons are juxtaposed and ligated together to form messenger RNA (mRNA), before translation occurs to produce proteins. Disruption of the splicing process often leads to nonsense-containing transcripts, which are often degraded by a process known as nonsense-mediated mRNA decay (NMD). Escapees from this quality control mechanisms may result in mis-folded and misbehaved proteins, leading to genetic or acquired diseases.

2.1 Classical Splicing Signals

Understanding the molecular code for splicing requires first modeling the classical signals, namely the donor or 5' splice site (5'ss) and acceptor or 3' splice site (3'ss), before incorporating auxiliary elements that modulate splice site choice. Auxiliary *cis*-regulatory

elements, typically short sequences on the order of 6-10 bases long, located within exons and nearby flanking intronic regions serve to recruit *trans*-factors that enhance or repress splicing. For example, exonic splicing enhancers are thought to recruit serine-arginine rich (SR) proteins that interact with the spliceosome and promote the use of nearby splice sites. As such, it is crucial to model the splice signals and to estimate as accurately as possible the relative strengths of the splice sites as possible, without confounding the models by integrating signals from deep within the adjacent introns and/or exons.

In Section 2.2, I apply a probabilistic approach that incorporates non-adjacent dependencies to model splice sites.

2.2 Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals

2.2.1 Abstract

We propose a framework for modeling sequence motifs based on the Maximum Entropy principle (MEP). We recommend approximating short sequence motif distributions with the Maximum Entropy Distribution (MED) consistent with low-order marginal constraints estimated from available data, which may include dependencies between non-adjacent as well as adjacent positions. Many Maximum Entropy models (MEMs) are specified by simply changing the set of constraints, and are utilized to discriminate between signals and decoys. Classification performance using different MEMs gives insight into the relative importance of dependencies between different positions. We apply our framework to large datasets of RNA splicing signals. Our best models outperform previous probabilistic models in the discrimination of human 5' (donor) and 3' (acceptor) splice sites from decoys. Finally, we discuss mechanistically-motivated ways of comparing models.

2.2.2 Introduction

Given a set of aligned sequences representing instances of a particular sequence motif, what model should be used to distinguish additional motif occurrences from similar sequences? This problem occurs commonly in computational biology with examples of DNA, RNA and protein sequence motifs. For example, it is important identify signal peptides in protein sequences and to to recognize true sites of RNA splicing from 'decoy' splice sites in primary transcript sequences. A number of statistical models have been developed to approximate distributions over sets of aligned sequences. For example, Markov Models (MMs) and Hidden Markov Models (HMMs) are commonly used in bioinformatics[12], with applications in gene-finding and protein domain mod-

eling [18].

We propose that the most unbiased approximation for modeling short sequence motifs is the Maximum Entropy Distribution (MED) consistent with a set of constraints estimated from available data. This approach has the attractive property that it assumes nothing more about the distribution than that it is consistent with features of the empirical distribution which can be reliably estimated from known signal sequences. In this paper we consider low-order marginal distributions as constraints, but other types of constraints can also be accomodated. Such models have been exploited in natural language processing [3], amino acid sequence analysis [5] and as a weighting scheme for database searches with profiles [19].

We introduce our approach, define "constraints" and Maximum Entropy models (MEM), and describe the use of Brown's iterative scaling [4] procedure of iterative scaling to obtain the MED consistent with a given set of constraints in Section 2.2.3. We also describe the use of Brown's iterative scaling [4] procedure of iterative scaling to obtain the MED consistent with a given set of constraints. In addition, we introduce a greedy-search information maximization strategy to rank constraints. This approach is applied to splice site recognition[7], an important problem in genetics and biochemistry, for which an abundance of high quality data are available. We focus on effectively modeling the 9 base sequence motif at the 5' splice site (5'ss), and the $\sim$23 base sequence motif at the 3' splice site (3'ss) of human introns, and not on the general problem of gene prediction. However, better modeling of the splice signals should lead to improved gene prediction and can be used to predict the splicing phenotypes of mutations that alter or create splice sites. The constraints for a MEM can also be ranked in importance. Finally, we propose a straightforward mechanistically-motivated way of comparing splice site models in terms of local optimality.

2.2.3 Methods

Maximum Entropy Method Let X be a sequence of λ random variables $X = \{X_1, X_2, ..., X_\lambda\}$ which take values from the alphabet $\{A, C, G, T\}$. Let lower-case

$x = \{x_1, x_2, ...x_\lambda\}$ represent a specific DNA sequence. Let $p(X)$ be the joint probability distribution $p(X_1 = x_1, X_2 = x_2, ..., X_\lambda = x_\lambda)$, and upper case $P(X = x)$ denote the probability of a state in this distribution (i.e. there are 4^λ possible states).

The principle of maximum entropy was first proposed by Jaynes [16] and states that of all the possible distributions in the hypothesis space that satisfy a set of constraints (component distributions, expected values or bounds on these values), the distribution that is the best approximation of the true distribution given what is known (and assuming nothing more) is the one with the largest Shannon entropy, H, given by the expression

$$H(\hat{p}) = -\sum \hat{p}(x) log_2(\hat{p}(x)) \tag{1}$$

where the sum is taken over all possible sequences, x. We will use logarithms to base 2, so that the entropy is measured in bits. Shannon entropy is a measure of the average uncertainty in the random variable X, i.e. the average number of bits needed to describe the outcome of the random variable. The set of constraints should therefore be chosen carefully and must represent statistics about the distribution that can be reliably estimated. It is possible to specify a set of constraints which are "inconsistent" in that they cannot be simultaneously satisfied (e.g. $\{P(A, A) = 3/4, P(T, T) = 1/2\}$). However, all constraint sets used here will be subsets of the marginal frequencies of the "empirical distribution" on sequences of length λ, and will therefore be consistent. The uniqueness of the MED for a consistent set of constraints was proved by Ireland and Kullback [15].

The principle of minimum cross-entropy or minimum relative entropy (MRE), first introduced by Kullback, is a generalization of the MEP that applies in cases when a background distribution q is known in addition to the set of constraints. Of the distributions that satisfy the constraints, the MRE distribution is the one with the lowest relative entropy (or KL-divergence), D, relative to this background distribution:

$$D_{KL}(\hat{p}) = \sum \hat{p}(x) log \frac{\hat{p}(x)}{q(x)} \tag{2}$$

Minimizing $D_{KL}(\hat{p})$ is equivalent to maximizing $H(\hat{p})$ when the prior q is a uniform distribution on the sequences of length λ. Shore and Johnson (1980) proved that maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima [26]. This implies that if we believe that the constraints are correct and well estimated (and no other information is assumed), then the MED is the best approximation of the true distribution.

Marginal Constraints For convenience, we consider two categories of constraints: "complete" constraints, which specify sets of position dependencies and "specific" constraints, which are constraints on (oligo-)nucleotide frequencies at a subset of positions.

"Complete" Constraints

Omitting the hats over the variables for convenience, let S_X be the set of all lower-order marginal distributions of the full distribution, $p(X = \{X_1, X_2, ..., X_\lambda\})$. A lower-order marginal distribution is a joint distribution over a proper subset of X. For example, for $\lambda = 3$,

$$S_X = \{p(X_1), p(X_2), p(X_3), p(X_1, X_2), p(X_2, X_3), p(X_1, X_3)\} \tag{3}$$

Define $S_s^m \subseteq S_X$, where superscript m refers to the *marginal-order* of the marginal distributions and the subscript s refers to the *skips* of the marginal distribution. In Equation 3, the first three elements are $1st$-order marginals (i.e. $m = 1$), and the last three elements are $2nd$-order marginals (i.e. $m = 2$): $p(X_1, X_2)$ and $p(X_2, X_3)$ are the $2nd$-order marginals with skip 0 ($s = 0$), and $p(X_1, X_3)$ is the 2nd-order marginal with skip 1 ($s = 1$). They are illustrated in our notation below:

$$S_0^1 = \{p(X_1), p(X_2), p(X_3)\}$$

$$S_0^2 = \{S_0^1, p(X_1, X_2), p(X_2, X_3)\}$$

$$S_1^2 = \{S_0^1, p(X_1, X_3)\}$$

$$S_X = S_{0,1}^2 = \{S_0^1, p(X_1, X_2)p(X_2, X_3), p(X_1, X_3)\}$$

For convenience, we include S_0^1 in S_s^m whenever the marginal order, $m > 1$. For an aligned set of sequences of length λ, the $1st$-order constraints (S_0^1) are the empirical frequencies of each nucleotide (A,C,G,T) at each position, and the Maximum Entropy Distribution consistent with these constraints is the Weight Matrix Model (WMM), i.e. all positions independent of each other [7]. On the other hand, if 2nd-order nearest-neighbor constraints (i.e. S_0^2) are used, the solution is a Inhomogeneous 1st-order Markov model (I1MM) (Appendix A). Consequently, different sets of constraints specify many different models. The performance of a model tells us about the importance of the set of constraints that was used.

"Specific" Constraints

"Specific" constraints are observed frequency values for a particular member of a set of "complete" constraints. Continuing with the example above, the list of 16 "specific" constraints for $p(X_1, X_3)$ are: $\{A \cdot A, A \cdot C, A \cdot G, A \cdot T, ..., T \cdot A, T \cdot C, T \cdot G, T \cdot T\}$, where $A \cdot A$ is the observed frequency of occurrence of the pattern ANA ($N = A, C, G$ or T).

Maximum Entropy Models A Maximum Entropy Model (MEM) is specified with a set of complete constraints, and consists of two distributions, namely, the signal model ($p^+(X)$) and the decoy probability distribution ($p^-(X)$), both of which are the MEDs generated by iterative-scaling (Section 2.2.3) over constraints from a set of aligned signals and a set of aligned decoys of the same sequence length, λ, respectively. Given a new sequence, the MEM can be used to distinguish true signals from decoys based on the likelihood ratio, L,

$$L(X = x) = \frac{P^+(X = x)}{P^-(X = x)} \tag{4}$$

where $P^+(X = x)$ and $P^-(X = x)$ are the probability of occurrence of sequence x from the distributions of signals(+) and decoys(-), respectively. Following the Neyman-Pearson lemma, sequences for which $L(X = x) \geq C$, where C is a threshold that achieves the desired true-positive rate α, are predicted to be true signals.

Iterative Scaling to Calculate MED In simple cases, the MED consistent with a set of constraints can be determined analytically using the method of Lagrange multipliers, but analytical solutions are not practical in most real world examples. Instead, the technique of iterative scaling is used. This technique was introduced by Lewis [21] and Brown [4], who showed that the procedure described below converges to the MED consistent with the given lower order marginal distributions. There is no limitation on the number or type of component distributions that can be employed [4]. Brown showed that at each step of the iteration, the approximation to the MED improves, using Equation(2) as a measure of closeness of the approximating distribution to the true distribution, but the proof of convergence is not rigorous (see [15] for a rigorous proof of convergence).

The iteration procedure begins with a uniform distribution with terms $P^0(X) = 4^{-\lambda}$, so all sequences of length λ are equally likely. Next, we specify a set of complete constraints and a corresponding list of specific constraints. Represent each member of the ordered list of specific constraints as Q_i, where i is the order in the list. The next step is to sequentially impose the specific constraints, Q_i, that the approximating distribution must satisfy. The terms relevant to the constraint at the j^{th} step of iteration have the form:

$$P^j = P^{j-1}\frac{Q_i}{\hat{Q}_i^{j-1}} \tag{5}$$

where P^{j-1} is a term at the $(j-1)$th step in the iteration while P^j is the corresponding term at the jth step, Q_i is the ith constraint in the list of "specific constraints" and $\hat{Q}_i^{j-1}$ is the value of the marginal corresponding to the ith constraint determined from the distribution p at the $j-1$th step. To illustrate, we return to our example in Equation 3 and apply constraint $Q_i = A \cdot A$ at the jth step:

$$P^j(X = ANA) = P^{j-1}(X = ANA)\frac{Q_i}{\hat{Q}_i^{j-1}} \tag{6}$$

where

$$\hat{Q}_i^{j-1} = \sum_{N \in \{A,C,G,T\}} P^{j-1}(X = ANA) \tag{7}$$

All terms not included in this sum (i.e. triplets not matching ANA) are iterated as follows:

$$P^j(X = VNW) = P^{j-1}(X = VNW)\frac{1 - Q_i}{1 - \hat{Q}_i^{j-1}} \tag{8}$$

for VNW such that $V \neq A$ or $W \neq A$, $N \in \{A, C, G, T\}$.

Note that enforcing satisfaction of a constraint at step j may cause a previous constraint to be unsatisfied until the previous constraint is applied again. This process is iterated until convergence or until a sufficiently accurate approximation is obtained.

Ranking Position Dependencies As the iterations proceed, the entropy, H (Equation 1) of successive distributions $p(X)$ decreases from the maximum value $log_2(4^\lambda)$ to that of the MED. This makes intuitive sense- as more constraints are applied, the distribution contains more information, hence lower entropy. For our purposes, we say the entropy has converged when the difference in entropy between iterations becomes very small (e.g. $\mid \Delta H \mid \leq 10^{-7}$). A KL-divergence criterion gives similar results. We have found that convergence typically requires about 10-20 complete iterations of the constraints for a cutoff of $\mid \Delta H \mid \leq 10^{-7}$.

Applying different constraints reduces the entropy of the distribution by different amounts. Therefore, we can control the rate of convergence by changing the order in which the constraints are applied. We perform a greedy search to rank constraints by the amount that they reduce the entropy of the solution as described below.

Greedy-search Entropy-reduction Strategy

A first list ("bag of constraints") is initialized to contain all specific constraints. A second list, the "ranked list", is initially empty. At each iteration:

1. Initialize a uniform distribution.

2. Determine the MED consistent with all constraints from the "ranked list".

3. Apply the first constraint from the "bag of constraints". Determine the reduction in H relative to the distribution determined in step 2, ΔH_i. Repeat all the constraints separately, recording ΔH_i for each constraint.

4. Place the constraint with the largest ΔH_i in the "ranked list".

5. Repeat steps 1 to 4 until all constraints in the "bag of constraints" have been placed in the "ranked list".

It is important to emphasize that the ranking of a constraint depends on the constraints ranked before, so that this algorithm is not guaranteed to determine the optimal subset of k constraints for $2 \leq k \leq N-1$, where N is the total number of constraints. Another possible criterion for ranking (instead of ΔH_i) is ΔKL_i defined as the reduction in relative entropy (Equation 2). Constraints can also be ranked in larger groups, instead of one at a time, thus speeding up the process.

2.2.4 Splice Site Recognition

The success of gene finding algorithms such as Genscan [8], HMMgene [17] and Genie [20] is critically dependent on finding the signals that mark exon-intron boundaries, which are recognized in cells by the nuclear pre-mRNA splicing machinery. The two strongest contributing signals are the donor or 5' splice site (5'ss) and the acceptor or 3' splice site (3'ss), which demarcate the beginning and end of each intron, respectively.

In [28], a number of algorithms that predict human splice sites were compared, indicating, as might be expected, that algorithms which use global and/or local coding information and splice signals (HMMgene and NetGene2) perform better than algorithms that only use the splice signals themselves (NNSPLICE, SpliceView and GeneID-3). Here, we focus on modeling the discrete splicing signals of specific length, with the understanding that once these have been optimally modeled, they could be incorporated into more complex exon or gene models if desired.

A number of models have been developed that can be estimated from reasonably

sized sets of sequences[7]. Weight Matrix models (WMMs) assume independence between positions. Although this assumption is frequently violated in molecular sequence motifs [6], WMMs are widely used because of their simplicity and the small number of sequences required for parameter estimation (SpliceView and GeneID-3 score splice sites based on Weight Matrix models [25]). Inhomogeneous first-order Markov models (I1MMs) account for nearest-neighbor dependencies which are often present in sequences and usually can discriminate sites more accurately than WMMs. However, I1MMs ignore dependencies between non-adjacent positions, which may also be present. Higher-order Markov models account for more distant neighboring dependencies, but the number of parameters that have to be estimated and hence the required number of training samples increases exponentially with Markov order.
Decision tree approaches, such as the Maximal Dependence Decomposition (MDD) [7] used in Genscan and GeneSplicer [24] reduce the parameter estimation problem by partitioning the space of signals such that each leaf of the tree contains a sufficiently-sized subset of the sites and the strongest dependencies between positions are modeled at the earliest branching points when the most data are available. Cai and colleagues applied probabilistic tree networks and found that simple first-order Markov models are surprisingly effective for modeling splice sites[10]. Arita and colleagues utilize the Bahadur expansion to approximate training of Boltzmann machines to model all pair-wise correlations in splice sites and found no improvement compared to first-order Markov models for 5'ss, but better performance for the 3'ss [1]. Our work is related to the latter two approaches in that we introduce a general family of models in which Markov models appear as natural members. It is worth noting that in addition to (non-)adjacent pairwise dependencies, MEMs can accommodate third-order or higher-order dependencies.

Construction of Transcript Data To avoid using computationally predicted genes, available human cDNAs were systematically aligned to their respective genomic loci by using a gene annotation script called GENOA (L.P. Lim and C.B.B. unpublished). To simplify the analysis, genes identified by this script as alternatively

spliced were excluded. We used a total of 1821 non-redundant transcripts that could be unambiguously aligned across the entire coding region, spanning a total of 12,715 introns (hence 12,715 5'ss and 12,715 3'ss). Our training and test data sets comprise disjoint subsets of these data. We use sequences at positions {-3 to +6} of the 5'ss (i.e. last 3 bases of the exon and the first 6 bases of succeeding intron), which have the GT consensus at positions {+1,+2}, and the sequences at positions {-20 to +3} of the 3'ss with the AG consensus at positions {-2,-1} (see Table 1). These splice sites are recognized by the major class or U2-type spliceosome that is universal in eukaryotes. We excluded 5'ss that have the GC consensus and 5'ss or 3'ss that matched the consensus patterns for splicing by the minor class or U12-type spliceosome. Decoy splice sites are sequences in the exons and introns of these genes that match a minimal consensus but are not true splice sites e.g. decoy 5' splices sites are non-splice sites matching the pattern N_3GTN_4 and decoy 3'ss are non-splice sites that match the pattern $N_{18}AGN_3$ [9].

2.2.5 Results and Discussion

Models of the 5' splice site The various models tested are listed in Table 2. The text abbreviations are in the first column, where "me" stands for maximum entropy, "s" stands for skip and "x" stands for the maximum skip; the first number is the marginal order and the second is the skip number or maximum skip number. Figure 1 and Table 2 together illustrate the improvement in performance resulting from use of more complex constraints. From the ROC analysis (Figure 1 and Appendix A.0.11), it is clear that me2s0 (equivalent to a I1MM), does much better than the me1s0 (equivalent to a WMM), as has been observed previously [7], indicating that nearest-neighbor contributions are important in human 5'ss. Our best model according to ROC analysis and maximum correlation coefficient analysis (Appendix A.0.10) for the 5' site is the me2x5 model, which takes into account all pair-wise dependencies. The MDD model used in Genscan [8] performs slightly better than the me2s0/I1MM model. Analysis using maximum 'approximate correlation' (see Appendix A.0.10)

rather than maximum correlation coefficient gave similar results.

We observe that the me2x5 model shows significant improvement over the me1s0/WMM model: the false positive rate at 90% sensitivity was reduced by approximately a factor of 2. The correlation coefficients are not large, which likely reflects properties of the human pre-mRNA splicing mechanism, in which 5'ss recognition relies heavily on other signals, such as enhancers and silencers, distinct from the splice signal itself [14] [2].

Ranked Constraints

The top 20 2nd-order constraints determined for models me2s0 and me2x5 using the greedy-search algorithm are listed in Tables 3 and 4. Figure 2A illustrates the faster increase in information content of the model when the constraints were applied in ranked order (Table 3), versus a random ordering of constraints. Furthermore, higher performance is achieved with ranked constraints versus a similar number of randomly ordered constraints (Figure 2B). Of course, when all the constraints are used, there is no difference in performance. Clearly, certain pairs of positions contain more information useful for discrimination. Also, the information content of the distribution is related to the performance of the model, i.e. the performance increases with increasing information content of the model. It is useful that the rankings of the dependencies are not just on the level of positions, but also at the level of (oligo)nucleotide sequence, a feature not seen in [10]. Some of these effects could reflect preferences of trans-acting factors which may bind cooperatively to different 5'ss positions.

Models of the 3' splice site The 3'ss sequence motifs is much longer than the 5'ss, ~23 bases. For notational simplicity, we define the index of each position in the sequence starting from 1 to 21, excluding the invariant AG dinucleotide. To avoid the impractical task of storing and iterating over $4^{21} \approx 4 \times 10^{12}$ possible sequences, we may first break up the sequences into 3 consecutive non-overlapping fragments of length 7 each (fragments 1 to 3: positions 1 to 7, 8 to 14 and 15 to 21 respectively), build individual MEDs for the 3 fragment subsets (see Equation 9), and score new

sequences by a product of their likelihood ratios (Equation 4).

$$P'(X) = P(X_1, ..., X_7)P(X_8, ..., X_{14})P(X_{15}, ..., X_{21}) \tag{9}$$

However, using Equation 9 ignores dependencies between segments. The resulting loss in performance is illustrated in Figure 3 (compare me2s0 and mm1 curves). Again, the me1s0 is equivalent to a WMM. To retain the dependencies of the nucleotides between the segments while avoiding computer memory issues, we propose the following approach. Six other fragments are modeled (fragments 4 and 5: positions 5 to 11 and 12 to 18 respectively; fragments 6 to 9: positions 5 to 7, 8 to 11, 12 to 14, 15 to 18 respectively). We then multiply the likelihood ratios for fragments 1 to 5 and divided by the likelihood ratios of fragments 6 to 9. For dependencies within 7 bases, this approach "covers" all the positions.

$$P_{overlap}(X) = \frac{P'(X)P''(X)}{P(X_5, ..., X_{11})P(X_{12}, ..., X_{18})} \tag{10}$$

where

$$P''(X) = P(X_5, X_6, X_7)P(X_8, X_9, X_{10}, X_{11})P(X_{12}, X_{13}, X_{14})P(X_{15}, X_{16}, X_{17}, X_{18})$$

The performance of this "overlapping" Maximum Entropy model is illustrated in Figure 3 (labeled modified me2s0), and performs similarly to the corresponding Markov model. Models me3s0 and me4s0 were modified analogously. Previous researchers have found that nearest-neighbor dependencies were sufficient to specify good models for 3'ss sites ([7],[10]). In fact, we found that a 2nd-order Markov model of the 3'ss site performs better than a 1st-order Markov model, but that a 3rd-order model performs worse than a 1st-order Markov model, presumably because of parameter estimation and/or sample size issues for 3rd-order transition probabilities of the form $P(L|IKJ)$ where I, J and/or K are purines (low frequency in most 3'ss positions). This observation motivates our procedure for segmenting the signal into 9 fragments, which use only 2nd-order constraints and neglects some long-range dependencies (such as between positions 1 and 21). It is possible to segment the signal in a

way that captures such long-range dependencies (not shown). However we found that adding dependencies beyond 2-nucleotide separations does not significantly change the performance (Table 5 and Figure 4).

Clustering Splice Site Sequences The MDD model [7] [8] demonstrated that appropriate subdivision of the data can lead to improved discrimination. Here, we ask whether MEMs can be improved by first clustering the data into subsets. First, we generated a symmetric dissimilarity matrix D, where d_{ij} is the number of mismatches between splice site sequences i and j in the list of training set sequences. Next, we implemented hierarchical clustering on D using Ward's method. Results for our set of 5'ss are shown in Figure 5 and Figure 6.

Interestingly, we observe that the highest contributors to the information content (excluding the GT consensus) in cluster 1 come from the 3rd, 4th, 5th and 6th bases in the intron, whereas the last two bases in the exon contribute the most in cluster 2, indicating that clusters 1 and 2 represent "right-handed" and "left-handed" versions of the 5'ss motif respectively. These two classes of 5'ss might be recognized by different sets of trans-factors e.g. U6 snRNP would generally interact more strongly with "right-handed" 5'ss, while U5 snRNP should interact preferentially with "left-handed" 5'ss [9]. We can combine separately trained models in the following manner:

$$P_{combined}(X) = P(X|M_1)P(M_1) + P(X|M_2)P(M_2) \tag{11}$$

where $P_{combined}(X)$ is the probability of generating sequence X under the combined model, $P(X|M_1)$ and $P(X|M_2)$ are the conditional probabilities of generating X given the model constructed using cluster 1 and cluster 2 sequences, respectively, and $P(M_1)$ and $P(M_2)$ are the prior probabilities of cluster 1 and 2, respectively. The performance of combined 5'ss models are illustrated in Figure 7. Separating the sequences into the 2 clusters and modeling them separately with WMMs and then combining the models performs significantly better than using a WMM derived from all the sequences. However, modeling the separate clusters with me2x5 and I1MM models does not show significant improvements compared to modeling the

entire cluster. Apparently, the more complicated models are able to capture cluster-specific information using the entire set of sequences. Figure 8 shows the motifs for 3'ss clusters which appear to separate into T-rich versus C/T-rich pyrimidine tracts. Combined 3'ss models showed a similar effect as with the 5'ss models (data not shown).

2.2.6 Applications of Splice Site Models

The specificity of pre-mRNA splicing hinges on highly conserved base pairing between the 5' splice site (5'ss) and U1/U6 small nuclear RNAs as well as interactions with U1C protein [11] and U5 snRNA [23]. It is unclear whether decoy splice sites are recognized by the splicing machinery. A study showing that intronic 5' decoy sites are activated when cells are heat shocked demonstrates that intronic decoys may be functional under special conditions [22]. Therefore, decoys could potentially be real splice sites, but may be blocked by the presence of RNA secondary structures [29], or have suboptimal location relative to splicing enhancers and repressors [14] [13]. Nevertheless, a good computational model should generally assign higher scores (i.e. log-likelihood ratios) to real 5'ss and lower scores to decoys, when all other factors are equal.

Proximal 5'ss decoys in introns We have used several measures to compare the performance of different models, all of which involve comparing the sensitivity of the models for a given false positive rate (Appendices A.0.10 and A.0.11). This essentially sets a global threshold, C (see Section 2.2.3) in deciding whether a sequence is or is not a true splice site. However, the splice site recognition machinery does not appear to use a global setting- in some cases weak splice sites are used when positioned in close proximity to splicing enhancers. This suggests a local decision rule for splice site detection, i.e. the most important factor may be whether the true splice site has higher score than decoys in its proximity.

We compared models by scoring possible 5'ss in a dataset of $\sim 12,600$ human introns. Better models should result in a larger number of introns with no higher-scoring de-

coys downstream of the real 5'ss. Figure 9 shows that our best 5'ss model, me2x5, results in the greatest number of introns which have no higher-scoring decoys downstream of the real 5'ss, i.e. 69 introns more than the MDD model, and 639 more than the WMM. Moreover, the me2x5 model gives the lowest number of introns that have a first higher-scoring decoy (fhsd) in the intron within 250 bases from the upstream real 5'ss - me2x5 predicted 75 fewer such introns than MDD, and 686 fewer introns than the WMM. The three models result in approximately the same number of introns where the fhsd occurred further than 250 bases from the real 5'ss. On inspection of the length distribution of these introns, we observed that the median length for these introns were $\sim 2,770 - 2,900$ bases, whereas the rest of the introns had a median length of $\sim 650 - 750$ bases, suggesting that global optimality of splice site motifs is less important in long introns.

Ranking and Competing 5'ss The top 20 highest-scoring 5'ss sequences ranked by the me2x5 model are listed in Table 6, with their corresponding ranks by the MDD, 1IMM and WMM models and, in the last column, the "odds ratio" defined as the frequency of occurrence of the sequence as a splice site divided by its occurrence as a decoy. Different models result in significatly different rankings of the signals. Figure 10 shows that the top scoring sequences are well correlated between models, but the lower scoring sequences vary much more.

Predicing Splicing Mutations in the ATM gene Ataxia-telangiectasia (A-T) is an autosomal recessive neurological disorder caused by mutations in the *ATM* gene. Recently, our Maximum Entropy 5'ss and 3'ss models have been utilized to predict the consequences of genomic mutations in the *ATM* gene that perturb splicing with promising results [30].

2.2.7 Conclusions

We recommend using the Maximum Entropy Distribution as the least biased approximation for the distribution of short sequence motifs consistent with reliably

estimated constraints. We show that this approach grants us the flexibility of generating many different models simply by utilizing different sets of constraints. Our greedy-search strategy ranks constraints at the resolution of paired nucleotides at specific positions. This can be useful for determining correlations with binding factors. We demonstrate on a simple example that using the constraints in order of their ranking increases the rate of convergence to the MED, increases the information content of the distribution and improves performance much faster than using randomly ordered constraints. The ranking of these constraints may reflect biological dependencies between nucleotides at different positions in the motif. Our best models using simply dinucleotide marginal distributions outperform previous models, e.g. WMMs and IMMs. These models themselves are MEDs when position-specific frequencies or nearest-neighbor dinucleotide frequencies are used as constraints. MEMs are relatively easy to use, e.g. the 5'ss model is stored as a 16,384-long vector in lexicographic order. We have developed a 3'ss "overlapping" Maximum Entropy model using an approach which combines multiple sub-models that performs better than models utilizing only nearest-neighbor dependencies. We show that the MED takes into account possible sub-clustering information in the data. We use a straightforward biologically-motivated way to compare models in terms of local optimality. Importantly, the MED framework described can be applied to other problems in molecular biology where large datasets are available, including classification and prediction of DNA, RNA and protein sequence motifs.

2.2.8 Acknowledgements

We thank Philip Sharp and Tomaso Poggio for helpful discussions, and Uwe Ohler and the anonymous reviewers for comments on the manuscript. This work was supported by grants from the NIH and NSF (C. B. B.) and the Lee Kuan Yew Scholarship from the government of Singapore (G. Y.).

2.2.9 Appendix

A Inhomogeneous Markov Models

A *kth*-order Inhomogeneous Markov Model can be generated as follows:

$$p_{kMM}(X) = p(X_1, ..., X_k) \prod_{i=k+1}^{\lambda} p(X_i | X_{i-1}, ..., X_{i-k}). \tag{12}$$

where $X = \{X_1, X_2, ...X_\lambda\}$, k is the order and $p(X_i|X_{i-1}, ..., X_{i-k})$ is the conditional probability of generating a nucleotide at position i given the previous k nucleotides. As before, conditional probabilities and marginals are estimated from the corresponding frequencies of occurrences of nucleotide combinations at the specified positions.

It is important to note that the maximum entropy distributions using nearest-neighbor constraints of marginal-order $(k+1)$ are equivalent to kth-order Markov Models. In every case, the performance of the MED for constraints S_0^k was equivalent to that of a $(k-1)th$ order Markov model. Thus the class of Markov models is a subset of the class of solutions specified by MEM.

A.0.10 Performance Measures

Table 7 illustrates a confusion matrix, which contains information about how well a model performs given an independent test set with real splice sites (positives) and decoys (negatives). N is the total number of test sequences, i.e. $N = TP + FP + FN + TN$. Standard Measures of accuracy such as Correlation Coefficient (CC), Approximate Correlation (AC), Sensitivity (Sn) and False Positive Rate (FPR) are defined below:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{(TP + FN)(TN + FP)(TP + FP)(TN + FN))^{\frac{1}{2}}}$$

$$AC = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}) - 1$$

$$Sn = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

A.0.11 ROC analysis

Receiver Operating Curve (ROC) Analysis [27] is an effective way of assessing the performance of models when used in a binary hypothesis test. In our case, a sequence x is predicted as a splice site if the likelihood ratio, L, is greater than a threshold, C (Equation 4). A ROC curve is a graphical representation of Sn (on the y-axis) versus false positive rate (FPR) (on the x-axis) as a function of C, and has the following useful properties:

1. It shows the tradeoff between sensitivity and false positive rate (increases in sensitivity are generally accompanied by an increase in false positives).

2. The closer the curve follows the left-hand border and then the top border of the ROC plot, the more accurate the model. The area under the curve is a measure of test accuracy.

3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the model.

Analogous to the ROC analysis, we can plot the other standard measures as described above against changing values of the threshold, C. The maximum point on the curves will indicate the best setting for C and gives a performance measure which can be used to compare models. Hence we can define CC_{max} to be the maximum correlation coefficient i.e. the highest point on the curve, and $C_{CC_{max}}$ is the threshold where CC_{max} is obtained. AC_{max} and $C_{AC_{max}}$ can be defined similarly.

References

[1] M. Arita, K. Tsuda, and K. Asai. Modeling splicing sites with pairwise correlations. *Bioinformatics*, 18(2):S27–S34, 2002.

[2] B.J. Lam, K.J. Hertel. A general role for splicing enhancers in exon definition. *RNA*, 10:1233–41, 2002.

[3] A. Berger, S. Pietra, and V. Pietra. A maximum entropy approach to natural language processing . *Computational Linguistics*, 22(1):39–71, 1996.

[4] D. Brown. A Note on approximations to discrete probability distributions. *Information and Control*, 2:386–392, 1959.

[5] E. Buehler and L. Ungar. Maximum entropy methods for biological sequence modeling . *Workshop on Data Mining in Bioinformatics, BIOKDD*, 2001.

[6] M. Bulyk, P. Johnson, and G. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–61, 2002.

[7] C. Burge. Chapter 8. Modeling dependencies in pre-mRNA splicing signals. *S.L. Salzberg, D.B. Searls, S. Kasif (Eds.), Computational Methods in Molecular Biology, Elsevier Science*, pages 129–164, 1998.

[8] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[9] C. Burge, T. Tuschl, and P. Sharp. Chapter 20. Splicing of precursors to mRNAs by the spliceosomes. *R. Gesteland and T. Cech and J. Atkins (Eds.), The RNA World , Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York*, pages 525–560, 1999.

[10] D. Cai, A. Delcher, B. Kao, and S. Kasif. Modeling splice sites with bayes networks . *Bioinformatics*, 16(2):152–158, 2000.

[11] H. Du and M. Rosbash. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature*, 419(6902):86–90, 2002.

[12] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids . *Cambridge University Press*, 1998.

[13] W. Fairbrother and L. Chasin. Human genomic sequences that inhibit splicing. *Molecular and Cellular Biology*, 20(18):6816–6825, 2000.

[14] W. Fairbrother, R. Yeh, P. Sharp, and C. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, 2002.

[15] C. Ireland and S. Kullback. Contingency tables with given marginals . *Biometrika*, 55(1):179–188, 1968.

[16] E. Jaynes. Information theory and statistical mechanics I . *Physics Review*, 106:620–630, 1957.

[17] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. *T. Gaasterland et al(Eds.), Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, Cambridge, UK*, pages 179–186, 1997.

[18] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–31, 1994.

[19] A. Krogh and G. Mitchison. Maximum entropy weighting of aligned sequences of proteins or DNA. *In Proc. Third Int. Conf. Intelligent Systems for Molecular Biology (ISMB), Eds. C. Rawlings et al. AAAI Press*, pages 215–221, 1995.

[20] D. Kulp, D. Haussler, M. Reese, and F. Eeckman. A generalized hidden markov model for the recognition of human genes in DNA. *D.J. States et al(Eds.), Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Cambridge, UK*, pages 134–142, 1996.

[21] P. Lewis. Approximating probability distributions to reduce storage requirements. *Information and Control*, 2:214–225, 1959.

[22] E. Miriami, J. Sperling, and R. Sperling. Heat shock affects 5' splice site selection, cleavage and ligation of CAD pre-mRNA in hamster cells, but not its packaging in hnRNP particles. *Nucleic Acids Research*, 22:3084–3091, 1994.

[23] A. Newman. The role of U5 snRNP in pre-mRNA splicing. *EMBO*, 16(19):5797–800, 1997.

[24] M. Pertea, X. Lin, and S. Salzberg. GeneSplicer: a new computational method for splice site prediction . *Nucleic Acids Research*, 29(5):1185–1190, 2001.

[25] M. Shapiro and P. Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implicatinos in gene expression. *Nucleic Acids Research*, 15(17):7155–7174, 1987.

[26] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-Entropy . *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[27] J. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[28] T. Thanaraj. Positional characterisation of false positives from computational prediction of human splice sites. *Nucleic Acids Research*, 28(3):744–754, 2000.

[29] L. Varani, M. Hasegawa, M. Spillantini, M. Smith, J. Murrell, B. Ghetti, A. Klug, M. Goedert, and G. Varani. Structure of tau exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc Natl Acad Sci USA*, 96(14):8229–34, 1999.

[30] L. Eng, G. Coutinho, S. Nahas, G. Yeo, R. Tanouye, T. Dork, C.B. Burge, and R.A. Gatti Non-classical splicing mutations in the coding and non-coding regions of the ATM gene: maximum entropy estimates of splice junction strengths. submitted. 2003.

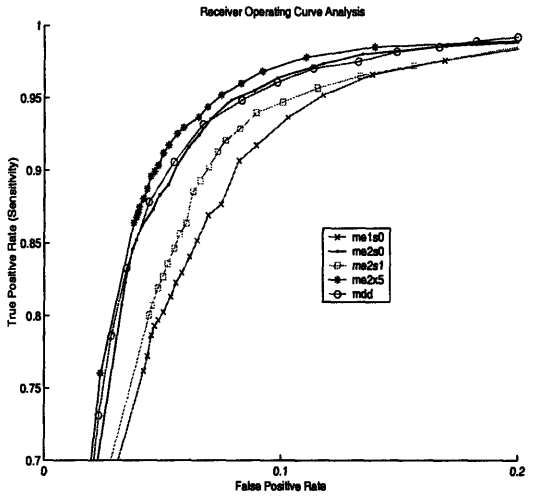


Figure 1: 5'ss: ROC curves for me2x5, me2s0, me2s1, me1s0 and MDD. The curves for the other models are not plotted, but can be inferred from Table 2.

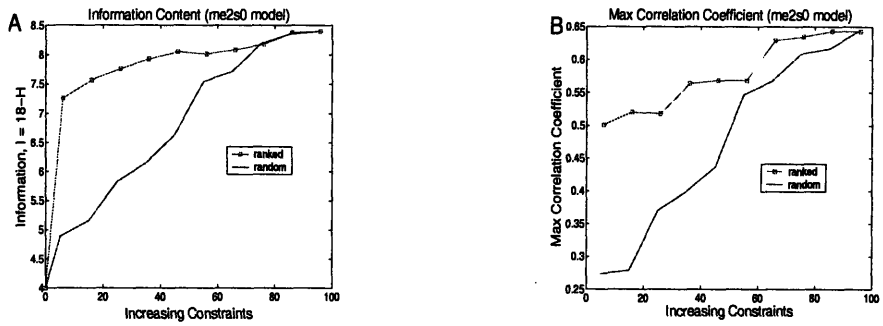


Figure 2: (A): Information content (I=18-H) of me2s0 model as constraints are added. If the constraints are ranked, the information content increases at a higher rate than if randomly ranked constraints are used. The x-axis corresponds to the model using the top N constraints. (B): Maximum Correlation Coefficient as a function of constraints. Ranked constraints added sequentially led to better performance with fewer constraints, compared to a random ordering of constraints. The model is me2s0 (excluding 1st order marginals).

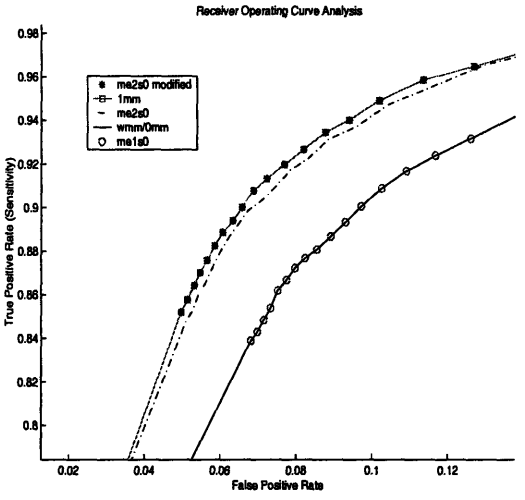


Figure 3: 3'ss: ROC curves for me2s0,me2s0 modified, me1s0, 0mm (0IMM) and 1mm (1IMM) models of the 3'ss. The curve labeled me2s0 was constructed by segmenting the 21 base long sequence set into 3 consecutive non-overlapping fragments of length 7 each. The curve labeled me2s0 (modified) was constructed as described in the text, and is equivalent to the 1st-order Markov model (I1MM).

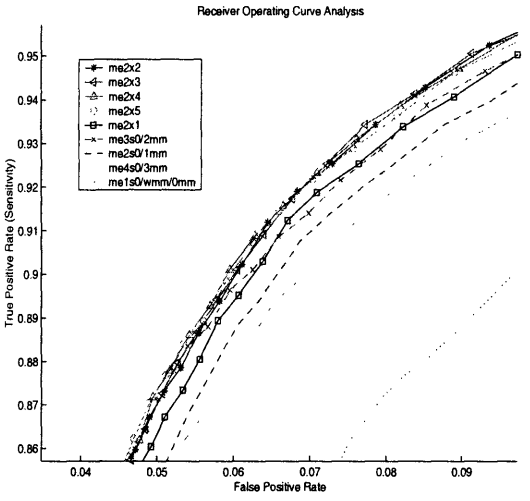


Figure 4: 3'ss: ROC curves for modified me2x5, me2x4, me2x3, me2x2, me2x1, me4s0, me3s0, me2s0 and me1s0 models of the 3'ss.

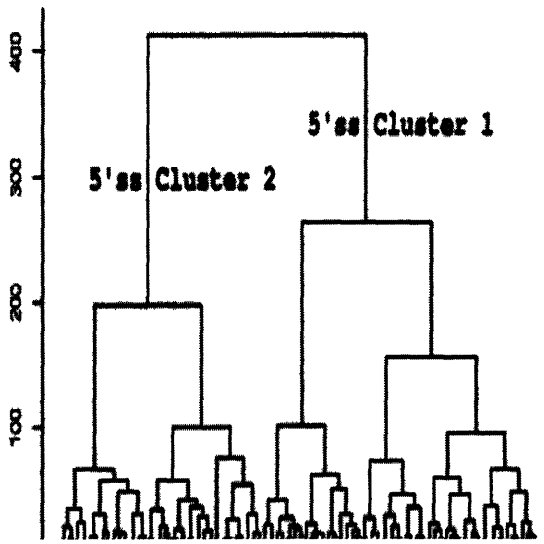


Figure 5: Truncated Dendrogram for 5'ss sequences (hierarchical clustering using ward's method). The two major clusters contain 7,260 and 5,367 sequences, respectively.

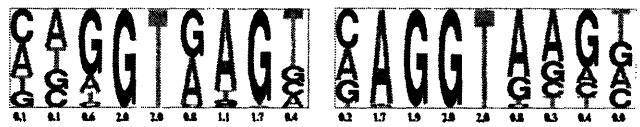


Figure 6: Sequence motifs for 5'ss cluster 1 (left) and 2 (right) created with the Pictogram program: http://genes.mit.edu/pictogram.html. The height of each letter is proportional to the frequency of the corresponding base at the given position, and bases are listed in descending order of frequency from top to bottom. The information content (relative entropy) for each position relative to a uniform background distribution is also shown.

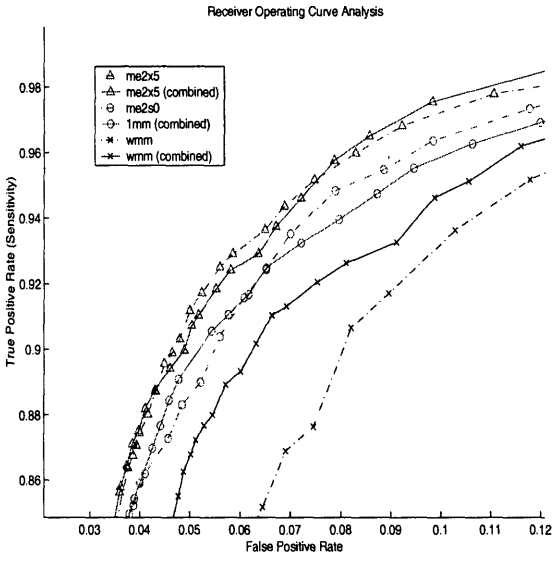


Figure 7: 5'ss: ROC curves for me2x5, 1IMM, WMM and me2x5 (combined), 1IMM (combined) and WMM (combined).

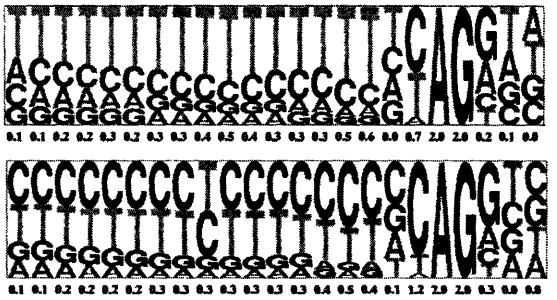


Figure 8: Sequence motifs for 3'ss cluster 1 (top) and 2 (bottom)

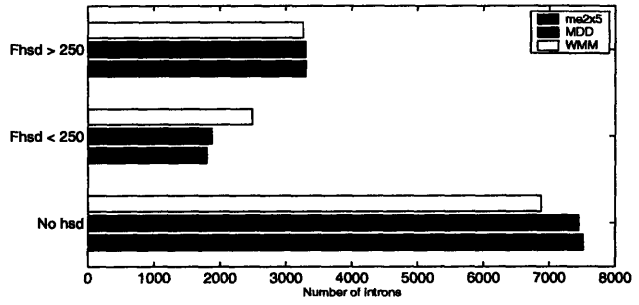


Figure 9: Bar Chart showing the number of introns that have no higher scoring decoy (hsd) than the real upstream 5'ss, and the number of introns that have a first higher scoring decoy (Fhsd) within 250 bases from the real 5'ss or greater than 250 bases from the real 5'ss.

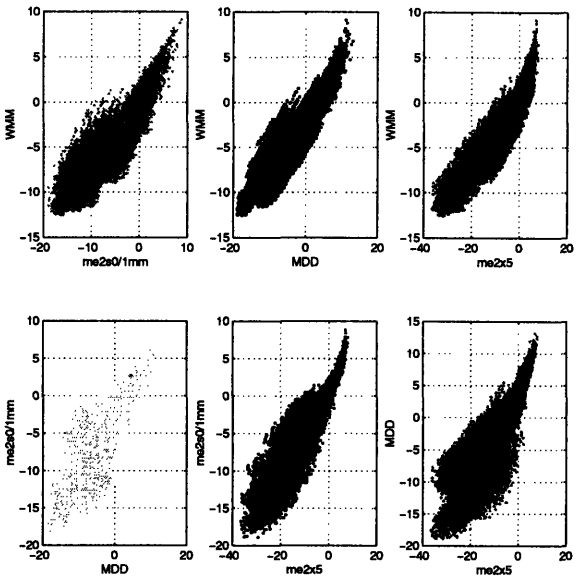


Figure 10: Scores of 5'ss sequences by different models plotted against each other. Model names are labeled in the x and y-axes.

	Real 5'ss	Decoy 5'ss	Real 3'ss	Decoy 3'ss
Train	8,415	179,438	8,465	180,957
Test	4,208	89,717	4,233	90,494
Total	12,623	269,155	12,698	271,451

Table 1: Number of sequences in 5'ss and 3'ss training and test sets.

Models	Constraints	CC
me2x5	$S^2_{1,2,3,4,5}$	0.6589
me2x4	$S^2_{1,2,3,4}$	0.6552
me2x3	$S^2_{1,2,3}$	0.6533
me5s0	S^5_0	0.6527
me2x2	$S^2_{1,2}$	0.6399
me4s0	S^4_0	0.6390
mdd	-	0.6493
me2s0	S^2_0	0.6425
me3s0	S^3_0	0.6422
me2s1	S^2_1	0.5971
me2s2	S^2_2	0.6010
me2s4	S^2_4	0.5861
me2s3	S^2_3	0.6031
me2s5	S^2_5	0.5924
me1s0	S^1_0	0.5911

Table 2: 5'ss Models ranked by ROC analysis(top to bottom), and the corresponding maximum Correlation Coefficients (CC).

Rank	ΔH_i	ΔKL_i
1	.AGgt....	.AGgt....
2	...gt.AG.	...gt.AG.
3	TA.gt....	TA.gt....
4	...gtTA..	...gtTA..
5	...gt..GT	...gt..GT
6	...gtGT..	...gtGT..
7	...gt.CG.	...gt.CG.
8	..TgtT...	..GgtG...
9	...gt.GG.	..AgtC...
10	...gtGG..	...gtCG..
11	...gtCG..	...gt..AC
12	.TAgt....	...gtCC..
13	..AgtC...	..CgtT...
14	...gt.AT.	CT.gt....
15	...gt.GT.	AG.gt....
16	...gt..CG	.TGgt....
17	...gt..AT	...gt.GT.
18	...gt..CT	.TAgt....
19	..GgtG...	..TgtC...
20	..CgtA...	...gtGG..

Table 3: Top 20 ranked constraints for me2s0. Lower letters refer to donor consensus positions. Capitalized letters are positional dependencies. All first order constraints were imposed as default.

Rank	ΔH_i	Sign
1	..Ggt..G.	-
2	...gt.AG.	+
3	.AGgt....	+
4	C..gt...C	+
5	...gtAA..	-
6	..GgtT...	+
7	..GgtC...	+
8	..GgtA...	-
9	...gtTA..	-
10	..Tgt..T.	-
11	..Tgt..A.	-
12	.G.gt..C.	-
13	...gtC.G.	+
14	.C.gt..C.	-
15	.T.gt..C.	-
16	..Cgt..A.	-
17	..Cgt..T.	-
18	..Agt..T.	-
19	..Agt..A.	-
20	..Cgt..G.	+

Table 4: Top 20 ranked constraints for me2x5 for 5'ss. + and - indicate whether the dinucleotide is more or less frequent than expected under independence assumption, respectively. All first order constraints were imposed by default.

Models	Constraints	CC
me2x2	$S^2_{1,2}$	0.6291
me2x3	$S^2_{1,2,3}$	0.6290
me2x4	$S^2_{1,2,3,4}$	0.6252
me2x5	$S^2_{1,2,3,4,5}$	0.6229
me2x1	S^2_1	0.6259
me3s0	S^3_0	0.6300
me2s0	S^2_0	0.6172
me4s0	S^4_0	0.6075
me1s0	S^1_0	0.5568

Table 5: 3'ss Models ranked by ROC analysis(top to bottom), and the corresponding maximum Correlation Coefficients (CC).

Sequence	me2x5	MDD	me2s0	WMM	odds ratio
ACGGTAAGT	1	2	5	26	184
TCGGTAAGT	2	3	12	114	77
ACGGTGAGT	3	17	18	90	11
GCGGTAAGT	4	14	10	56	3
ACGGTACGT	5	67	14	292	331
TCGGTGAGT	6	28	34	304	9
CAGGTAAGG	7	26	15	3	13
GAGGTAAGT	8	34	6	4	38
ATGGTAAGT	9	12	46	19	95
AAGGTAAGT	10	10	2	2	12
GACGTAAGT	11	41	136	86	10
CCGGTAAGT	12	1	7	17	22
CCGGTGAGT	13	7	22	68	18
CAGGTACGG	14	99	32	79	68
CAGGTAAGT	15	20	1	1	8
CAGGTAAGA	16	25	13	7	14
CGGGTAAGT	17	15	17	15	2
AAGGTACGT	18	54	8	46	233
AACGTAAGT	19	19	101	38	96
CAGGTGAGT	20	27	4	8	21

Table 6: Ranks of 5'ss sequences by different models.

	predicted positive	predicted negative
real positive	true positives, TP	false negatives, FN
real negative	false positives, FP	true negatives, TN

Table 7: Confusion Matrix

2.3 Applications for MaxENTScan and splice site models

The maximum entropy splice site models (MaxENTScan) (as well as a variety of other splice site models) are available for use online[1], taking as input sequences of specified lengths and returning a MaxENT score for each sequence. In addition, users can choose to build their own maximum entropy splice site models[2].

2.3.1 Prediction of non-classical splicing mutations

Splice site models can be utilized to predict the result of genomic mutations that disrupt existing splice sites, or create a competing proximal splice site, thereby causing a disease. An example of such an application of MaxENTScan to predict the result of splicing mutations in the ATM gene was recently published [1] and for conciseness will not be discussed here.

2.3.2 Finding enriched elements in pseudo-exons

It is well known that human introns contain many sequences that match the consensus splice site motifs as well as authentic sites, but are never used in splicing. Pairings of potential 3' splice sites and potential 5' splice sites, spaced by lengths typical of exons are referred to as 'pseudo-exons' [2]. As only about half of the information required for accurate recognition of exons and introns in human transcripts are contained in the

---

[1] http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
[2] http://genes.mit.edu/burgelab/maxent/Xmaxent.html

classical splice signals, it implies that auxiliary elements outside of the classical splice signals must play important roles in the promotion of authentic splice sites, or suppression of 'decoy' splice sites in pseudo-exons [3].

Recently, exonic splicing silencers (ESSs) were identified using a splicing reporter system used to screen a random sequence library for short sequences with splicing silencer activity in cultured human cells. The potential role of derived ESS motifs in suppression of pseudo-exons was studied by utilizing the MaxENTScan program to scan intronic sequences for pseudo-exons (50-250 bp in length) and comparing the frequencies of occurrence of the ESS sequences in the pseudo-exon dataset (PE) to a constitutive exon dataset (CE). Consistent with the expectation that ESSs would be favored in PEs rather than in constitutive exons, they were substantially underrepresented in CEs versus PEs (χ^2 test, $P << 2.2 \times 10^{-16}$). Furthermore, MaxENTScan was utilized to separate CEs intro 'strong' exons (CEs with both 3'ss and 5'ss in the top quartile of scores) and 'weak' exons (CEs with both splice sites in the bottom quartile of scores). Consistent with our expectation that exons with weaker splice sites are likely to be more prone to silencing by ESSs, the set of ESS decamers were found depleted in weak versus strong exons (χ^2 test, $P << 2.2 \times 10^{-16}$). These results are part of a larger study that has been recently accepted (Wang et al., *Cell,* in press, 2004).

2.3.3 Splicing simulation: ExonScan

The MaxENTScan algorithm has been incorporated into a first-generation splicing simulation algorithm called ExonScan, which integrates additional known or putative splicing regulatory sequences to predict the locations of internal exons in primary

transcript sequences. The results of large-scale splicing simulations for a dataset of 1820 human primary transcripts containing a total of 10,891 internal exons indicate that the version of ExonScan using models of 3'ss and 5'ss motifs alone identified 4,008 exons correctly at a cutoff where the number of false positive predictions and false negative predictions are equal, roughly 37% of all exons in the set. This work has been integrated into a larger piece of work (Wang et al., *Cell*, in press, 2004, Rolish et al., unpublished).

References

1. Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, Dork T, Burge C, Gatti RA: Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. *Hum Mutat* 2004, 23(1):67-76.
2. Sun H, Chasin LA: Multiple splicing defects in an intronic false exon. *Mol Cell Biol* 2000, 20(17):6414-6425.
3. Lim LP, Burge CB: A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* 2001, 98(20):11193-11198.

2.4 Splicing *cis*-regulatory and *trans*-acting elements

2.4.1 *Cis*-elements

Splicing *cis*-regulatory elements are typically short sequences [1-3] which can be divided into those that enhance the use of neighboring splice sites, called splicing enhancers, and those that suppress the recognition of splice sites, called silencers. They can be further categorized by their site of action, either in exons or introns. Accordingly there are four major classes of regulatory elements: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). Several groups of ESEs are known, such as purine-rich and AC-rich elements, as well as others [4, 5]. In comparison, ESS sequences are less studied. However, recently, computational screens for ESS motifs have been published [6, 7], one predicting three ESS motifs, and another predicting 974 octamers with ESS activity. A notable point is that the regulatory function of, and by extension, the category of a *cis*-element is entirely contextual. For example G-rich elements are a known ISE in mammalian introns [8], but when found localized (10 bases) next to the 5'ss of particular exons, will act as an ISS (Han et al., submitted), and have also been found enriched in a experimental screen for decamers enriched for exonic splicing silencer activity (Wang et al., submitted).

The successful computational screen and experimental verification of human ESEs by the RESCUE method [9] motivated application of this method to multiple available vertebrate genomes (Section **2.4.3**). Not surprisingly, candidate ESE motifs were found

to be highly conserved in vertebrates. In fact, this comparative genomic analysis might be useful in determining the most important or most active subset of ESEs identified in the original implementation of the RESCUE-ESE method [9]. As an illustration, single-nucleotide polymorphisms had a lower likelihood of interrupting ESEs conserved across vertebrates than those found in a single organism [10]. Interested users are able to annotate predicted RESCUE-ESEs in multiple organisms onto input sequences [11][1]. By using similarly motivated rules, the RESCUE method was then adapted to find ISEs in vertebrate genomes, and found interestingly, that dramatic differences in candidate ISEs exist between mammals and fish. Systematic studies of differences in the *cis*-elements involved with splicing regulation in different organisms is important in studying the evolution of the splicing process, in particular, the co-evolution of *cis*-elements and their associated *trans*-factors, in the context of an organism's genomic structure [12].

2.4.2 *Trans*-acting factors

Cis-regulatory elements generally function by recruiting protein *trans*-factors that interact favorably or unfavorably with components of the spliceosome [13, 14]. Most of the known ESEs are thought to be recognized by members of the serine-arginine rich (SR) protein family, which also interact with each other and with snRNP proteins to enhance the recognition of adjacent splice sites [4]. On the other hand, ESSs inhibit the use of adjacent splice sites, often via interactions with members of the heterogeneous nuclear ribonucleoproteins (hnRNP) family [15-21]. Large-scale proteomic studies of the factors

[1] http://genes.mit.edu/burgelab/rescue-ese/

involved with the splicing machinery have been conducted [22-24], and have led to associations of hnRNPs and SR or SR-related proteins with different stages of spliceosomal assembly.

Aside from constitutive splicing, alternative splicing (AS) is regulated precisely in different developmental stages and tissues [25, 26], and context-specific regulation is likely to be coordinated by multiple *cis*-regulatory signals [1, 2]. The tissue-specific expression of the various *trans*-factors involved with both alternative and constitutive splicing and the predominance of certain motifs in exons regulated in the corresponding tissues will ultimately give us insight into the context-specificity of regulation. For example, the N1 exon in the *c-src* gene, studied extensively by the Black laboratory, is repressed by the polypyrimidine-tract binding (PTB) protein in non—neuronal cells where it binds to UC-rich elements on both sides of the exon. In neurons, the PTB is replaced with a related neural PTB (nPTB) protein and PTB binding to upstream silencer elements is destabilized in an ATP-dependent process. The regulatory region downstream of the exon, requiring the UGCAUG element, exerts its positive effects under these conditions and causes inclusion of the exon (reviewed in [26]).

References

1. Ladd AN, Cooper TA: Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* 2002, 3(11):reviews0008.
2. Cartegni L, Chew SL, Krainer AR: Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002, 3(4):285-298.
3. Blencowe BJ: Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 2000, 25(3):106-110.

4. Graveley BR: Sorting out the complexity of SR protein functions. *Rna* 2000, 6(9):1197-1211.
5. Zheng ZM: Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci* 2004, 11(3):278-294.
6. Sironi M, Menozzi G, Riva L, Cagliani R, Comi GP, Bresolin N, Giorda R, Pozzoli U: Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res* 2004, 32(5):1783-1791.
7. Zhang XH, Chasin LA: Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 2004, 18(11):1241-1250.
8. McCullough AJ, Berget SM: G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol* 1997, 17(8):4562-4571.
9. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: Predictive identification of exonic splicing enhancers in human genes. *Science* 2002, 297(5583):1007-1013.
10. Fairbrother WG, Holste D, Burge CB, Sharp PA: Single Nucleotide Polymorphism-Based Validation of Exonic Splicing Enhancers. *PLoS Biol* 2004, 2(9):E268.
11. Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB: RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* 2004, 32(Web Server issue):W187-190.
12. Yeo G, Hoon S, Venkatesh B, Burge CB: Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* 2004, 101(44):15700-15705.
13. Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL: Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* 1994, 368(6467):119-124.
14. Wu JY, Maniatis T: Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* 1993, 75(6):1061-1070.
15. Caputi M, Mayeda A, Krainer AR, Zahler AM: hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. *Embo J* 1999, 18(14):4060-4067.
16. Chen CD, Kobayashi R, Helfman DM: Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes Dev* 1999, 13(5):593-606.
17. Del Gatto-Konczak F, Olive M, Gesnel MC, Breathnach R: hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. *Mol Cell Biol* 1999, 19(1):251-260.
18. Domsic JK, Wang Y, Mayeda A, Krainer AR, Stoltzfus CM: Human immunodeficiency virus type 1 hnRNP A/B-dependent exonic splicing silencer ESSV antagonizes binding of U2AF65 to viral polypyrimidine tracts. *Mol Cell Biol* 2003, 23(23):8762-8772.
19. Kashima T, Manley JL: A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet* 2003, 34(4):460-463.
20. Zahler AM, Damgaard CK, Kjems J, Caputi M: SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing

enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. *J Biol Chem* 2004, 279(11):10077-10084.

21. Si Z, Amendt BA, Stoltzfus CM: Splicing efficiency of human immunodeficiency virus type 1 tat RNA is determined by both a suboptimal 3' splice site and a 10 nucleotide exon splicing silencer element located within tat exon 2. *Nucleic Acids Res* 1997, 25(4):861-867.
22. Jurica MS, Moore MJ: Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* 2003, 12(1):5-14.
23. Rappsilber J, Ryder U, Lamond AI, Mann M: Large-scale proteomic analysis of the human spliceosome. *Genome Res* 2002, 12(8):1231-1245.
24. Zhou Z, Licklider LJ, Gygi SP, Reed R: Comprehensive proteomic analysis of the human spliceosome. *Nature* 2002, 419(6903):182-185.
25. Black DL, Grabowski PJ: Alternative pre-mRNA splicing and neuronal function. *Prog Mol Subcell Biol* 2003, 31:187-216.
26. Black DL: Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003, 72:291-336.

2.5 Variation in sequence and organization of splicing regulatory elements in vertebrate genes

2.5.1 Abstract

Although core mechanisms and machinery of pre-mRNA splicing are conserved from yeast to human, the details of intron recognition often differ between even closely related organisms. For example, genes from the pufferfish *Fugu rubripes* generally contain one or more introns that are not properly spliced in mouse cells. Exploiting available genome sequence data, a battery of sequence analysis techniques was used to reach several conclusions about the organization and evolution of splicing regulatory elements in vertebrate genes. The classical splice site and branch site signals are completely conserved across the vertebrates studied (human, mouse, pufferfish and zebrafish), and exonic splicing enhancers (ESEs) also appear broadly conserved in vertebrates. However, another class of splicing regulatory elements, the intronic splicing enhancers (ISEs), appears to differ substantially between mammals and fish, with G triples (GGG) very abundant in mammalian introns but comparatively rare in fish. Conversely, short repeats of AC and GT are predicted to function as ISEs in fish but are not enriched in mammalian introns. Consistent with this pattern, ESE-binding SR proteins are highly conserved across all vertebrates, while hnRNP proteins, which bind many intronic sequences, vary in domain structure and even presence/absence between mammals and fish. Exploiting differences in intronic sequence composition, a statistical model was developed to predict the splicing phenotype of *Fugu* introns in mammalian systems, and used to engineer spliceability of a *Fugu* intron in human cells by insertion of specific sequences, thereby rescuing splicing in human cells.

2.5.2 Introduction

The pufferfish, *Fugu rubripes*, with its sevenfold smaller genome than human, has proven to be an excellent resource for comparative genomics [1]. The *Fugu* genome

also has great potential for applications in genetics. The compactness of *Fugu* genes makes them ideal candidates for use in transgenesis, with the advantage over cDNA-derived constructs that they would be capable of producing all the isoforms of a particular gene under appropriate regulatory control. However, the potential for using *Fugu* genes as natural mini-genes for the production of transgenic mice has not been realized as initial efforts to express *Fugu* transgenes in mouse cells have failed due to incorrect transcript processing by the murine splicing machinery [2, 3]. However, *Fugu* genes studied to date are spliced and translated correctly in zebrafish, a fish whose genome size and gene organization are more similar to mammals than to *Fugu*.

These somewhat surprising results imply that substantial differences exist between fish and mammalian systems in exon-intron sequences and/or splicing factors. The relatively low information contents of the classical splice site signals in higher eukaryotes argues that additional transcript features are likely to be involved in recognition and splicing of many, if not all introns [4]. Exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs) and exonic or intronic splicing silencers can enhance or repress the use of 5′ or 3′ splice sites (5′ss, 3′ss), depending on their site and mode of action [5, 6, 7, 8]. ESEs have been the subject of many studies and most are known to be recognized by members of the serine-arginine-rich (SR) protein family [9, 10]. SR proteins bind to ESEs through their RNA-binding domains and promote splicing by recruiting spliceosomal components through protein-protein interactions via their arginine-serine-rich (RS) domains [9, 11, 12, 13]. The *trans*-factors that bind to intronic splicing regulatory elements have not been characterized as thoroughly, and both SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs) have been implicated in interactions with intronic *cis* elements.

Using the human [14], mouse[15] and *Fugu*[16] genome sequences, we applied and adapted the RESCUE approach for identification of splicing regulatory sequences [17],

and developed new methods to analyze similarities and differences in the sequences and organization of splicing regulatory elements in mammalian and fish genes. These methods revealed significant differences in predicted ISEs between mammalian and fish introns which appear to explain why certain *Fugu* introns are not faithfully processed by the mammalian splicing machinery.

2.5.3 Methods and Materials

Frequency Difference Plots The difference between the observed frequency of a pattern (enumerated as in Table 3, supporting information) occurring in windows of size 10-bp (exons of size >60 bp) or 30-bp (intronic region) and the mean frequency of the same pattern in 10 random permutations (shuffles) of the sequence in the window were determined as follows, with an offset of 3 bp between successive windows. The observed frequency of a pattern of length m bp in a window of size W bp, at position j in sequence i is defined as $f_{observed,i,j} = \frac{x_{i,j}}{W/m}$, where $x_{i,j}$ is the number of non-overlapping occurrences of the motif whose first positions fall within the window (i.e. excluding occurrences that overlap previously counted occurrences). The average shuffled frequency of the motif out of s total shuffles of the same window is defined as $f_{avgshuffled,i,j} = \frac{1}{s}\sum_{k=1}^{s}\frac{y_{i,j,k}}{W/m}$, where $y_{i,j,k}$ is the number of non-overlapping occurrences of the motif in the kth shuffled version of the same window of size W bp, at position j in sequence i. Therefore, the frequency difference (FD) of the motif at position j in sequence i is defined as $FD_{i,j} = f_{observed,i,j} - f_{avgshuffled,i,j}$. The mean FD value μ_j and variance σ_j^2 in a window of size W bp starting at position j over N sequences are calculated as $\mu_j = \frac{1}{N}\sum_{i=1}^{N} FD_{i,j}$ and the standard error of the mean (SEM), ε, is derived as $\varepsilon = \frac{\sigma}{\sqrt{N}}$, where $\sigma_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(FD_{i,j} - \mu_j)^2$.

Linear Discriminant Analysis and Intron Classification Linear discriminant functions g_1 and g_2 for n_1 *Fugu* introns and n_2 mouse introns, respectively, were defined as $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + b_i$, where $\mathbf{w}_i = \mathbf{\Sigma}^{-1}\mu_i$ and $b_i = -0.5\mu_i^t\mathbf{\Sigma}^{-1}\mu_i$, and $\mathbf{x}$ is the

vector of overlapping 3-mer counts computed from +5 to +65 of the intron and from -71 to -11 of the intron. Σ is the pooled covariance matrix from the individual covariance matrices: $\Sigma = ((n_1-1)\Sigma_1 + (n_2-1)\Sigma_2))/(n_1+n_2-2)$. The linear discriminant analysis (LDA) output [18], y, is defined as $y(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$. The intron length score, s_{len}, was defined as $s_{len}(l) = \log(f_{Fugu}(l)/f_{mouse}(l))$, where l is the length of the intron, and f_{Fugu} and f_{mouse} are the estimated frequencies of introns falling into the relevant intron length bins in the respective organisms (Fig. 13, supporting information). Scores were generated which combine the intron length scores and the LDA outputs for *Fugu* and mouse introns in the following way: $z(\mathbf{x}, l) = y(\mathbf{x}) + s_{len}(l)$, where $\mathbf{x}$ represents a 128-long vector of 3-mer counts from an intron, and l is the intron length.

2.5.4 Results

Splice Site Signals and Predicted ESEs are Conserved in Vertebrates. In order to identify potential splicing differences between different vertebrate organisms, three major classes of *cis*-acting elements were systematically analyzed: the canonical splice site/branch site motifs, and two classes of splicing enhancers. Using large datasets of annotated exon-intron structures, we found that the extended consensus sequences of the classical 5′ss and 3′ss sequence motifs are essentially the same in human, mouse, zebrafish and *Fugu* (these data are shown in Fig. 7A, which is published as supporting information on the PNAS web site, www.pnas.org). Putative branch point sequences identified using a motif finding algorithm also appear similar in sequence and are positionally conserved in orthologous mouse, human and *Fugu* introns, occurring most commonly 20 to 40 bases upstream of the 3′ss (Fig. 7B, supporting information). These data suggest that neither the branch point motif nor the 5′ss or 3′ss differ significantly between fish and mammals in the features required for recognition by the splicing machinery, and that the observed differences in splicing between these systems must lie elsewhere.

Both constitutive and alternative splicing events are often modulated by elements in exons known as exonic splicing enhancers (ESEs). In order to assess potential differences in ESE sequences between organisms, we applied the RESCUE-ESE approach that was used previously to identify ESEs in human genes [17] to large datasets of annotated mouse and *Fugu* genes (Table 2, supporting information; access to RESCUE-ESE hexamers for each of these organisms are available at http://genes.mit.edu/burgelab/rescue-ese/ [19]). Sets of candidate ESE sequences that satisfy the two RESCUE-ESE criteria of significant enrichment in exons relative to introns and significant enrichment in exons with weak (non-consensus) 5′ss or 3′ss sequences relative to exons with strong splice sites were identified. Previously, predicted human ESE hexamers were clustered into ten groups on the basis of sequence similarity [17], and then aligned to produce ten distinct ESE motifs (Fig. 1A). Comparing the candidate mouse and *Fugu* ESE hexamers with those identified in human exons, a great deal of overlap was observed, with many of the same hexamers identified independently in different organisms. For example, 90 out of the 100 hexamers comprising the purine-rich human 5C3D class were also predicted as ESEs in mouse, and 54 of these 100 hexamers were predicted as ESEs in *Fugu* exons (Fig. 1A). Of the ten clusters of human ESEs identified, only the smallest (cluster 5E) was not represented in mouse. Furthermore, seven of the ten human clusters were represented in *Fugu*, the exceptions being three of the most sparsely populated human ESE clusters. Thus, RESCUE-ESE analysis supports the presence in all three vertebrates of all of the large classes of ESEs identified in humans.

To further explore potential ESE-related differences between organisms, we analyzed the frequency difference (FD) plots of RESCUE-ESE hexamers along exons from each of the three vertebrates in sliding windows of ten bases in width. As shown for the 5C3D cluster (Fig. 1B), most clusters of RESCUE-ESE hexamers exhibit a concave ('smiley') distribution, with increased FD values in the vicinity of both the 5′ and 3′ splice sites. This distribution is likely to result from increased selection to con-

serve ESEs near splice sites, which would be consistent with previous studies showing that ESEs located closer to the 3′ss of exons have higher activity than those located more distally [20], and that ESE-disrupting single-nucleotide polymorphisms (SNPs) are under-represented in exons near splice sites [21]. For the majority of ESE classes, the shapes of the FD plots were similar in human, mouse and *Fugu* (Fig. 1A and Fig. 9, supporting information). The conservation of the splice site-biased distributions of many classes of predicted ESEs between human, mouse and *Fugu* argues for their functional importance in all three vertebrates.

Predicted ISEs Differ Between Mammals and Fish. In addition to exon sequences such as ESEs, intronic elements also commonly play a role in alternative and constitutive splicing [22]. To identify putative intronic splicing enhancers (ISEs) in vertebrate introns, we developed an approach called RESCUE-ISE (supporting information). Following a similar rationale to that used in our previous RESCUE-ESE method [17], RESCUE-ISE predicts as ISEs hexamers that share two properties: significant enrichment in introns relative to exons, and significant enrichment in introns with weak (non-consensus) 5′ss or 3′ss relative to introns with strong splice sites. Applying this method to large datasets of human and mouse introns identified the triplet motif GGG and a C-rich motif respectively in both mammals (Fig. 2). The GGG and C-rich hexamer clusters together comprised 96% (127 hexamers) of RESCUE-ISE predicted ISE hexamers in introns downstream of human 5′ss, and 89% (266 hexamers) of predicted ISE hexamers in introns upstream of human 3′ss. Similar clusters comprised comparably large proportions of RESCUE-ISE hexamers in mouse; the few remaining hexamers did not cluster into motifs that were similar between human and mouse.

Curiously, when the RESCUE-ISE approach was applied to datasets of *Fugu* introns, a very different set of ISE motifs was predicted (Fig. 2), including motifs containing repetitions of CA and GT dinucleotides, but no motifs similar to the GGG or C-rich

elements identified in mammals. To further explore this difference, a more detailed analysis of the predicted ISE motifs was undertaken in mammalian and fish introns, using the sea-squirt *Ciona* as an outgroup. From analysis of FD plots (Fig. 3), two trends were clear: (i), for GGG, an established mammalian ISE [23], there were pronounced peaks in the FD distribution in the vicinity of the 5′ and 3′ss in both human and mouse introns; and (ii) these peaks were much more dramatic in introns with weak (non-consensus) 5′ or 3′ss than they were in introns with strong splice sites (red versus blue curves in figure). These two features can be explained if the location of the peak reflects an optimal interaction distance between hypothetical splicing regulatory factors that bind to ISEs and components of the splicing machinery bound at the splice sites, and if ISEs in weak-ss introns are under increased selection to ensure efficient and accurate splicing [22]. We propose that these two features comprise a sequence signature that is characteristic for ISEs.

Consistent with the differences seen in terms of predicted ISE motifs, the FD plots for *Fugu* introns were substantially different than those for mammalian introns (Fig. 2). Specifically, GGG was not enriched at any distance relative to the 5′ or 3′ss of *Fugu* introns (all FD values near zero), and had a nearly flat distribution, consistent with the absence of function in splicing. Instead, the predicted *Fugu* ISE motifs ACAC and GTGT showed pronounced FD peaks near the 5′ and 3′ss of *Fugu* introns, respectively, which were comparable in magnitude to those seen for GGG in mammalian introns. Consistent with this pattern, the peaks were more dramatic in introns with weak 5′ and 3′ss. By contrast, the distributions of ACAC and GTGT near the 5′ and 3′ss of mammalian introns were essentially flat, with no discernable peaks and no difference between weak and strong introns. The introns of the non-vertebrate chordate *Ciona intestinalis* showed modest peaks of GGG near the 5′ and 3′ss but no clear peaks for ACAC or GTGT, and the GGG peaks in *Ciona* were higher for strong-ss rather than for weak-ss introns.

Exon and Intron Definition Mechanisms May Differ Between Mammals and Fish. The "exon definition" model of splicing postulates that the exon is the primary unit initially recognized by the splicing machinery, typically involving a complex formed across the exon containing factors that recognize the 3′ss, one or more ESEs and the 5′ss of an exon [24]. This mode of splicing appears to predominate in transcripts containing small or medium-sized exons flanked by long introns [25]. On the other hand, in splicing by the "intron definition" model, the intron is the primary unit initially recognized by the splicing machinery, with formation of a complex of factors recognizing the 5′ss, ISE(s), and the 3′ss of an intron [24]. This mode of splicing tends to predominate in transcripts containing short introns flanked by medium-sized or large exons [25]. To analyze the effects of flanking intron length on the distribution of putative ESEs and ISEs in vertebrates, introns were categorized by length as either short (<125 bp), intermediate (125-1000 bp), or long (>1000 bp).

In human and mouse, exons flanked by longer introns contained a significantly higher abundance of most classes of RESCUE-ESE hexamers than those flanked by intermediate-length introns which, in turn generally contained more such ESEs than exons flanked by short introns (Table 4 and Fig. 11, supporting information). Furthermore, short mammalian introns had higher relative frequencies of the candidate ISEs GGG and CCC near their splice sites than intermediate or long introns (Fig. 12, supporting information). Surprisingly, the relationship between ESE density and intron length was different in *Fugu* genes. In *Fugu*, there was no tendency for exons flanked by long introns to have higher densities of RESCUE-ESE hexamers - in fact, the opposite tendency was observed for several ESE classes (Table 4, supporting information). Furthermore, predicted ISE motifs ACAC and GTGT were more highly enriched in intermediate and long introns than in short introns (Fig. 12, supporting information). Our proposed model is summarized in Figure 4.

Differing Conservation of SR Protein and hnRNP Genes Between Mam-

mals and Fish. Conservation of *cis*-regulatory elements between organisms is expected to correlate with patterns of conservation of the corresponding *trans*-factors. To explore these relationships with respect to splicing in vertebrates, lists of human splicing factors identified previously through proteomic analysis by Zhou et al. [26] were utilized to identify mouse and *Fugu* orthologs from the EnsMart database using reciprocal best BLAST hits. Domains were then predicted using PFAM [27] and the results are shown in Tables 5-8 in supporting information. Core spliceosomal components such as snRNAs and proteins of the U1 snRNP, U2 snRNP and U4/U5/U6 tri-snRNP are highly conserved between mammals and fish (Table 5, supporting information, and data not shown). Additionally, clear orthologs with identical domain organization could be found in mouse and *Fugu* for all human SR proteins (Table 6, supporting information), nearly all of which are known to recognize ESEs, consistent with our analysis indicating that the major RESCUE-ESE classes are conserved between human, mouse and *Fugu*. However, greater variability was seen in the domain organization and even presence/absence of H-complex hnRNP proteins, many of which are known to bind ISEs or other intronic elements (Table 7 and 8, supporting information). For example, *Fugu* and zebrafish orthologs for hnRNP A2/B1 [28, 29] and hnRNP F were not identified, and fish orthologs for hnRNP H and hnRNP K were missing one or more of their RRM and/or KH domains, compared to human and mouse orthologs. In addition, *Fugu* orthologs for hnRNP RALY was not found and hnRNP I(PTB) was missing a RRM. Given that the *Fugu* and zebrafish genomes are not yet complete (95% covered in *Fugu* and 5.7x coverage in zebrafish) and genome annotations are still evolving, absence of a detectable ortholog from current assemblies does not necessarily imply that an orthologous gene does not exist. Nevertheless, current data suggests greater variability in hnRNP proteins between mammals and fish than was seen for SR proteins.

Discrimination of Mammalian and *Fugu* Introns. The results reported above suggest that the critical differences in splicing between *Fugu* and mammalian introns

may reside primarily in the abundance and locations of specific short oligonucleotides with ISE activity, with intron length-dependent effects also playing a role. To explore this idea, a model based on linear discriminant analysis (LDA) was developed which utilizes intron length and non-overlapping 3-mer counts (including GGG and CCC) as features to predict whether a given *Fugu* intron will be correctly spliced in mammalian cells (Fig. 13, supporting information). Introns of the *Fugu RCN1*, *HD* and *ARP3* genes [2, 3, 30] were scored with this model (Fig. 5). Comparing the scores of *Fugu* introns to their splicing phenotypes in mammalian cells, a correlation was observed, with the highest-scoring (most *Fugu*-like) introns generally failing to splice in mammalian cells, and introns with scores in the range observed for natural mouse introns almost always splicing correctly (Fig. 5). Thus our method recognizes intronic features that differ between *Fugu* and mammalian introns, and appears able to predict the spliceability of *Fugu* introns in mammalian cells. Independently of RESCUE-ISE, this method ranks G triples, C-rich motifs, and AC repeats as critical features that distinguish fish and mammalian introns.

Rescuing Splicing of *Fugu* Introns in Mammalian Systems. Our experience with the LDA model suggested that changing the sequence composition of a *Fugu* intron that was mis-spliced in mammalian cells by adding sequences that function as ISEs in mammalian introns might rescue the splicing phenotype. To test this idea, a *Fugu ARP3* construct (Fig. 14, supporting information) was transfected into human 293T cells and into a fish (minnow) cell line, PLHC-1 (supporting information). Following splicing, cDNA was synthesized by reverse transcription, and PCR using primers targeting exon 1 and exon 12 revealed 1.2 kb products in both cell lines. To assess the pattern of splicing, both 1.2 kb transcripts were cloned into pGEM-T vectors and sequenced. The presence of aberrant splicing was confirmed in the 293T cell line while the transcript from the PLHC-1 cell line was spliced correctly. In 293T cells, introns 4 and 9 were retained, exon 7 was skipped, and exon 5 was truncated by use of a cryptic 5′ss. Based on the LDA model, we attempted to rescue splicing

of *Fugu* ARP3 intron 4 by insertion of sequences similar to the G1 and G2 G-triples from intron 2 of the human alpha globin gene into the *Fugu* intron [23]. Insertion of these sequences reduces the score of the intron substantially, to a score range where tested *Fugu* introns have generally spliced correctly (Fig. 5). The wild type intron of length 88 was mutated using site-directed mutagenesis to generate two mutants with a single and double G-triplet located near the 5′ splice site resulting in mutant introns 99 bases and 107 bases long, respectively. These two mutant constructs were transfected into human 293T cells and cDNA was synthesized under the same conditions as before, and PCR with primers flanking the intron was used to assess the degree of splicing. A single G2 insert was sufficient to partially rescue splicing of intron 4 (Fig. 6, supplementary information). Insertion of both G1 and G2 increased the level of splicing to that seen in the PLHC-1 cell line. Thus, changing the ISE composition of a mis-spliced *Fugu* intron as guided by the LDA model restored levels of correct splicing in mammalian cells comparable to that seen in fish.

2.5.5 Discussion

Core components of the spliceosome are universally conserved in higher eukaryotes, but less is known about the conservation of the sequences and factors that regulate splicing. The observation that some *Fugu* introns are not properly spliced in mammalian cells suggests that substantive differences in splicing exist between mammals and fish. Here, we conducted a large-scale bioinformatic study of *cis*-elements and *trans*-factors that are important in splicing, comparing mammalian and fish genomes to identify similarities and differences between organisms.

Sequence motifs at the 5′ and 3′ splice sites were not significantly different between mammalian and fish genes, and predicted branch site motifs are also quite similar. Applying the RESCUE-ESE approach to identify candidate ESEs in human, mouse, and *Fugu* exons, substantial overlap in the sets of predicted ESE hexamers was found

(Fig. 1A). Previously, the ESE activity of representatives of ten candidate human ESE motifs predicted by RESCUE-ESE were confirmed using an *in vivo* splicing reporter assay, demonstrating high predictive accuracy for this method [17]. The validity of the cross-species RESCUE-ESE predictions are further supported by a recent study which found that the hexamers predicted here as ESEs in multiple vertebrates are significantly less likely to be disrupted by SNPs in human than those restricted to a single species [21]. Additional evidence of conserved function comes from FD plots, which document similar positional biases in RESCUE-ESE motifs along human, mouse and *Fugu* exons (Fig. 1B and Fig. 9, supporting information). High conservation of splice site and predicted ESE motifs across vertebrates was mirrored in patterns of splicing factor conservation. Orthologs for all ten human SR proteins were identified in mouse and *Fugu*, and domain structure was preserved.

To explore potential differences in intronic splicing enhancers (ISEs), we introduced RESCUE-ISE, a computational method to predict intronic splicing enhancers (ISEs). RESCUE-ISE and FD plot analysis identified GGG, a known mammalian ISE conserved in human and mouse [8], but did not identify any related motifs in *Fugu* or zebrafish introns (Fig. 2). In addition to GGG, a C-rich motif is also over-represented in introns near splice sites in human and mouse, but not in *Fugu* or zebrafish (Fig. 10, supporting information). Enrichment of CCC and GGG in human introns has also been observed previously, e.g., [31, 32], and references therein. McCullough and Berget showed that GGG elements in human introns can base pair to nucleotides 8 to 10 of U1 snRNA, recruiting U1 snRNP to the vicinity of the 5′ss [8]. Other splicing factors have also been implicated in binding to G-rich regions and influencing splicing, including hnRNPs A1, F, H and other members of the hnRNP H family [33, 34, 35, 36]. H complex hnRNP proteins, which often bind to exonic splicing silencers and intronic regulatory sequences, were less conserved between mammals and fish. Orthologs of hnRNP A1 and H were identified in all three vertebrates, but an ortholog for hnRNP F was not detected in the *Fugu* genome. Furthermore, the fish

orthologs of hnRNP H appears to lack an RRM present in both mammalian proteins. Other differences in hnRNP genes were also observed, including the apparent absence of hnRNPs A2/B1 and RALY from the *Fugu* genome. Two of these genes (hnRNP F and A2/B1) appear to be absent from the zebrafish genome as well, suggesting that these represent true gene losses in the fish lineage rather than genes missed due to the incompleteness of current genome assemblies or annotations. These differences in intron-binding factors between mammals and fish may explain why certain mammalian ISEs appear absent from fish.

Applying RESCUE-ISE to a dataset of *Fugu* introns identified short repeats of CA and GT dinucleotides as candidate ISEs in this organism (Fig. 2, motifs f3A, f5A). FD plots support a role for ACAC and GTGT sequences as enhancers of introns with weak 5′ss and weak 3′ss, respectively, in both *Fugu* and zebrafish (Fig. 3C,D). These elements have not been identified as ISEs involved in constitutive splicing in mammals. However, a recent study showed that hnRNP L binds specifically to CA repeats to enhance alternative splicing of an upstream exon in the human endothelial nitric oxide synthase gene [37], and an ortholog of hnRNP L is present in *Fugu*. GU repeat sequences were also recently shown to function as ISEs involved in tissue-specific alternative splicing of the human cardiac sodium calcium exchanger gene [38]. ETR-3 and the neuroblastoma apoptosis-related RNA-binding protein (NAPOR), an isoform expressed from the *CUGBP2* gene, bind to GU-rich sequences in certain mammalian introns and enhance alternative splicing [39, 40]. Orthologs of both genes are also present in *Fugu*. A search of the literature identified known mammalian splicing regulatory elements similar to candidate *Fugu* motifs f5D (TAG) [41] and f5E (T-rich) [42]. However, our search did not identify known elements similar to motif f5C, with consensus [A/T]TAC[A/T], whose potential role in splicing will require experimental tests. These observations suggest a model in which certain repetitive motifs used primarily to regulate alternative splicing in mammals have evolved a more prominent role in constitutive splicing in fish, despite substantial reduction in repeat content in

the *Fugu* genome.

In addition to the differences in the sequences of putative splicing regulatory elements described above, the organization of these elements also appears to differ between mammalian and fish genes. In mammalian genes, there is a compensatory relationship between ISEs and ESEs. Exons flanked by long introns are enriched in ESEs and deficient in nearby ISEs, whereas exons flanked by short introns are deficient in ESEs and enriched in nearby ISEs (Fig 4. and Fig. 11, supporting information). These observations are consistent with current splicing models for human transcripts, in which exons flanked by long introns are spliced by exon definition, which is generally dependent on ESEs, and short introns are recognized by an intron definition mechanism [25]. Sterner and colleagues observed that expanded human exons were efficiently included if flanking introns were at most 500 bp long, but were skipped if the introns were expanded [25], implying an upper bound of 500 bases for intron definition in mammals. The compaction of the *Fugu* genome has resulted in approximately 80% of introns being under 500 bases in length, presumably leading to a massive increase in intron definition. In contrast to what is seen in mammals, long *Fugu* introns have increased frequencies of putative ISE motifs relative to short *Fugu* introns, suggesting that even long *Fugu* introns may often be spliced by intron definition.

Our observations that putative ISE sequences differ substantially between mammalian and fish introns suggested that addition of mammalian ISEs to improperly spliced *Fugu* introns could rescue splicing in mammalian systems. Linear discriminant analysis was used to combine the sequence and architectural features that distinguish mammalian and fish introns. As an application, we inserted GGG sequences into intron 4 of the *Fugu* ARP3 gene. This modification was predicted by the LDA analysis to rescue splicing in mammals (FIg. 5) and, indeed, this modified intron was spliced in human cell lines at a comparable level to that of the wild-type intron in

a fish cell line (Fig 6). Thus, our computational analysis has applications for effective transfer of genetic information between vertebrates. This study also represents a paradigm for analyzing the evolution of gene expression regulation. Comparative genomic approaches similar to those described here should be applicable to other steps in gene expression, including transcription and translation, that are modulated by widespread *cis*-regulatory elements.

Acknowledgements

We gratefully acknowledge Paula Grabowski for critical reading of the manuscript, Tomaso Poggio and Phillip Sharp for advice. This material is based upon work supported by the National Science Foundation under Grant No. 0218506 (C. B. B., P. Sharp and T. Poggio). G.Y. was supported by the Lee Kuan Yew fellowship from Singapore. S.H and B.V. were supported by Singapore's Agency for Science, Technology and Research.

References

[1] Hedges, S.B. & Kumar, S. (2002) *Science* **297,** 1283-1285.

[2] Sathasivam, K., Baxendale, S. , Mangiarini, L. , Bertaux F. , Hetherington C. , Kanazawa I. , Lehrach H. & Bates G.P. (1997) *Human Molecular Genetics* **6,** 2141-2149.

[3] Miles, C.G. , Rankin, L. , Smith, S.I. , Niksic, M. , Elgar, G. & Hastie N.D. (2003) *Nucleic Acids Research* **31,** 2795-2802.

[4] Lim L.P. & Burge C.B., (2001) *Proc Natl Acad Sci USA* **98,** 11193-11198.

[5] Huh, G.S. & Hynes, R.O. (1993) *Mol Cell Biol* **13,** 5301-5314.

[6] Chan, R.C. & Black, D.L (1997) *Mol Cell Biol* **17,** 4667-76.

[7] Hedjran, F. , Yeakley, J.M. , Huh, G.S. , Hynes, R.O. & Rosenfeld, M.G. (1997) *Proc Natl Acad Sci* **94,** 12343-7.

[8] McCullough, A. & Berget, S. (2000) *Mol, Cellular Biol* **20,** 9225-9235.

[9] Graveley, B. (2000) *RNA* **6,** 1197-1211.

[10] Blencowe, B. (2000) *Trends Biochem Sci* **25,** 106-110.

[11] Graveley, B. & Maniatis, T. (1998) *Mol Cell* **1,** 765-771.

[12] Maniatis, T. & Tasic, B. (2002) *Nature* **418,** 236-243.

[13] Cartegni, L. , Chew, S.L. & Krainer, A.R. (2002) *Nature Review Genetics* **3,** 285-298.

[14] Lander, E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. et al. (2001) *Nature* **409,** 860-921.

[15] Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520-562.

[16] Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A. et al. (2002) *Science* **297,** 1301-1310.

[17] Fairbrother, W., Yeh, R., Sharp, P. & Burge, C.B. (2002) *Science* **297,** 1007-1013.

[18] Duda, R.O., Hart, P.E. & Stork, D.G. (2001) *Pattern Classification* (John Wiley & Sons)

[19] Fairbrother W.G., Yeo G.W., Yeh R., Goldstein P., Mawson M., Sharp P.A. & Burge C.B. (2004) *Nucleic Acids Res.* **32,** W187-90.

[20] Graveley B.R., Hertel K.J. & Maniatis T. (1998) *EMBO J.* **17,** 6747-6756.

[21] Fairbrother W.G., Holste, D., Burge, C.B. & Sharp, P.A. (2004) *PLoS Biol* **2,** 1-8.

[22] Ladd A. N. & Cooper T. A. (2002) *Genome Biol.* **3,** reviews0008.1-0008.16.

[23] McCullough, A. & Berget, S. (1997) *Mol, Cellular Biol* **17,** 4562-4571.

[24] Berget, S. (1995) *J Biol Chem* **270** 2411-2414.

[25] Sterner, D.A., Carlo, T. & Berget, S.M. (1996) *Proc Natl Acad Sci USA* **93,** 15081-15085.

[26] Zhou, Z., Licklider, L., Gygi, S. & Reed, R. (2002) *Nature* **419,** 182-185.

[27] Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M. & Sonnhammer E.L.(2002) *Nucleic Acids Research* **30,** 276-280.

[28] Kozu, T., Henrich, B. & Schafer, K.P. (1995) *Genomics* **25,** 365-371.

[29] Kamma, H., Horiguchi, H., Wan, L., Matsui, M., Fujiwara, M., Fujimoto, M., Yazawa, T. & Dreyfuss, G. (1999) *Exp Cell Res* **246,** 399-411.

[30] Venkatesh, B. & Brenner, S. (1998) *Gene* **211,** 169-175.

[31] Majewski, J. & Ott, J. (2002) *Genome Research* **12,** 1827-1836.

[32] Zhang, X.H., Heller, K.A., Hefter I., Leslie C.S. & Chasin L.A. (2003) *Genome Research* **13,** 2637-2650.

[33] Min, H., Chan, R.C. & Black, D.L. (1995) *Genes & Development* **9,** 2659-2671.

[34] Blanchette, M. & Chabot, B. (1999) *EMBO J* **18,** 1939-1952.

[35] Hastings, M.L., Wilson, C.M. & Munroe, S.H. (2001) *RNA* **7,** 859-874.

[36] Caputi, M. & Zahler, A.M. (2001) *The Journal of Biological Chemistry* **276,** 43850-43859.

[37] Hui, J., Stangl, K., Lane, W. & Bindereif, A. (2003) *Nature Structural Biol* **10,** 33-37.

[38] Gabellini, N. (2001) *Eur. J. Biochem.* **268,** 1076-1083.

[39] Zhang W., Liu H., Han K. & Grabowski P.J. (2002) *RNA* **8,** 671-85.

[40] Charlet N., Logan, P., Singh, G. & Cooper, T. (2002) *Mol. Cell.* **9,** 649-658.

[41] Kashima, T. & Manley, J.L. (2003) *Nature Genetics* **34,** 460-463.

[42] Del Gatto-Konczak F., Bourgeois, C.F., Le Guiner, C., Kister, L., Gesnel, M.C., Stevenin, J. & Breathnach, R. (2000) *Mol Cell Biol* **20,** 6287-99.

Table Legends

Table 1. **Conservation of splicing factors between human, mouse and *Fugu*.** Domains refer to predicted RNA recognition motifs (RRMs) and KH domains. Full tables listing accession numbers and Ensembl identifiers for all *trans*-factors analyzed are provided in Tables 5-8, supporting information.

Figure Legends

Figure 1. **Conservation of RESCUE-ESE sequences and distribution in vertebrates.** A. RESCUE-ESE [17] motifs, the number of predicted ESE hexamers in mouse and *Fugu* that overlap with RESCUE-ESE hexamers in human and the distribution of human RESCUE-ESE hexamers in sets of orthologous human, mouse and *Fugu* exons. +/− refers to significant increasing/decreasing frequency difference gradient towards the respective splice site. 0 indicates no gradient (computed similarly as described in Table 4, supporting information). * refers to conservation only in human and mouse, otherwise sign of gradient was conserved in all three organisms. B. As an example, the frequency difference plots for hexamers of RESCUE-ESE class 5C3D are shown as a function of distance from the 3′ss (left plot) or 5′ss (right plot) of orthologous exons in human, mouse and *Fugu*. Each point represents the start of a window of size 10 bases. Values are plotted at intervals every 6 bases. Black bars show standard error of the mean (Methods).

Figure 2. **RESCUE-predicted mammalian and *Fugu* ISE motifs.** GGG and C-rich motifs were predicted as ISEs in human and mouse introns at both splice sites. f5A-E are motifs enriched in *Fugu* introns near the 5′ss, and f3A-C are enriched near the 3′ss.

Figure 3. **Enrichment of predicted ISEs in introns near weak splice sites.** A. Frequency difference of GGG downstream of strong 5′ss and weak 5′ss, relative to locally permuted sequence windows of size 30 bases, starting from intron position +11. B. Frequency difference of GGG upstream of strong 3′ss and weak 3′ss, starting from intron position -41. C. Frequency difference of ACAC downstream of strong 5′ss and weak 5′ss, starting from intron position +11. D. Frequency difference of GTGT upstream of strong 3′ss and weak 3′ss, starting from intron position -41. Black bars show standard error of the mean (Methods). Values are plotted at intervals every 6

bases.

Figure 4. **Model of association between intron length and distribution of splicing regulatory elements** Green triangles represent the enrichment of RESCUE-predicted ESEs near the splice sites in human, mouse and *Fugu* exons. Red triangles represent the enrichment of RESCUE-predicted ISEs near the splice sites in human, mouse and *Fugu* introns. The height of the triangles illustrates the relative magnitude of enrichment of RESCUE-ESEs and RESCUE-ISEs. Intron sizes in base pairs (bp) are indicated above introns.

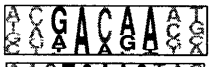
Figure 5. **Classification of vertebrate introns.** Distribution of model scores for independent sets of orthologous mouse and *Fugu* introns and splicing phenotypes for introns 1 to 5 the *Fugu RCN1* gene [3], introns 1 to 7 the *Fugu HD* gene [2], and introns 1 to 11 of the *Fugu ARP3* gene. Full details given in Fig 13C, supporting information.

Table 1.

Trans-factors	Mouse	*Fugu*
SR Proteins		
Domains same as human	10/10	10/10
Domains changed	0/10	0/10
Missing	0/10	0/10
hnRNPs		
Domains same as human	13/14	7/14
Domains changed	1/14	4/14
Missing	0/14	3/14

Figure 1

A

Human Cluster	RESCUE-ESE motif	No. of RESCUE-ESEs Human	Mouse	*Fugu*	Gradient 3'ss	5'ss
3B		9	6	0	+	+
3C		22	13	3	+*	+
3E		22	11	3	+	0
3F		27	15	5	+	–
3H		7	4	2	–	0
5A3G		21	19	5	+	+
5B3A		20	11	4	+	+
5C3D		100	90	54	+	+
5D		8	3	0	–	+
5E		5	0	0	0	0

B

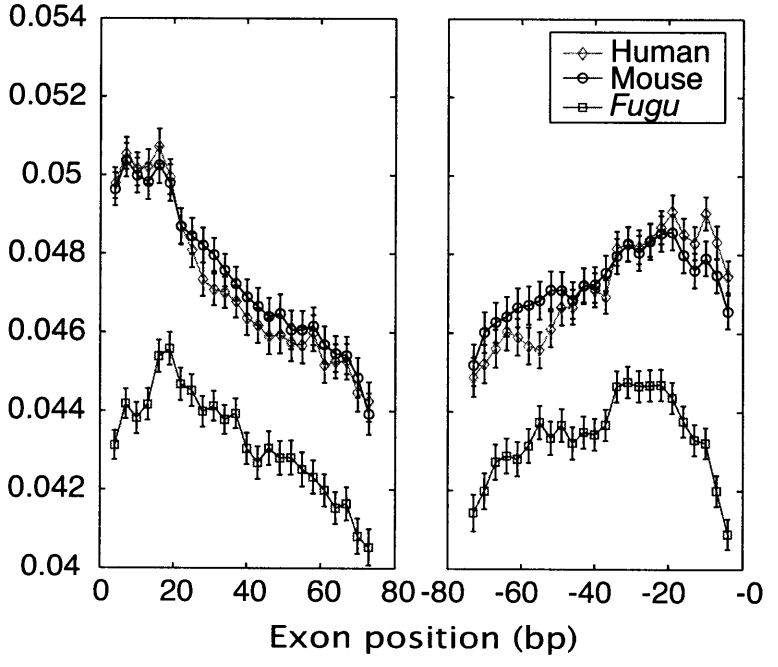


Figure 2

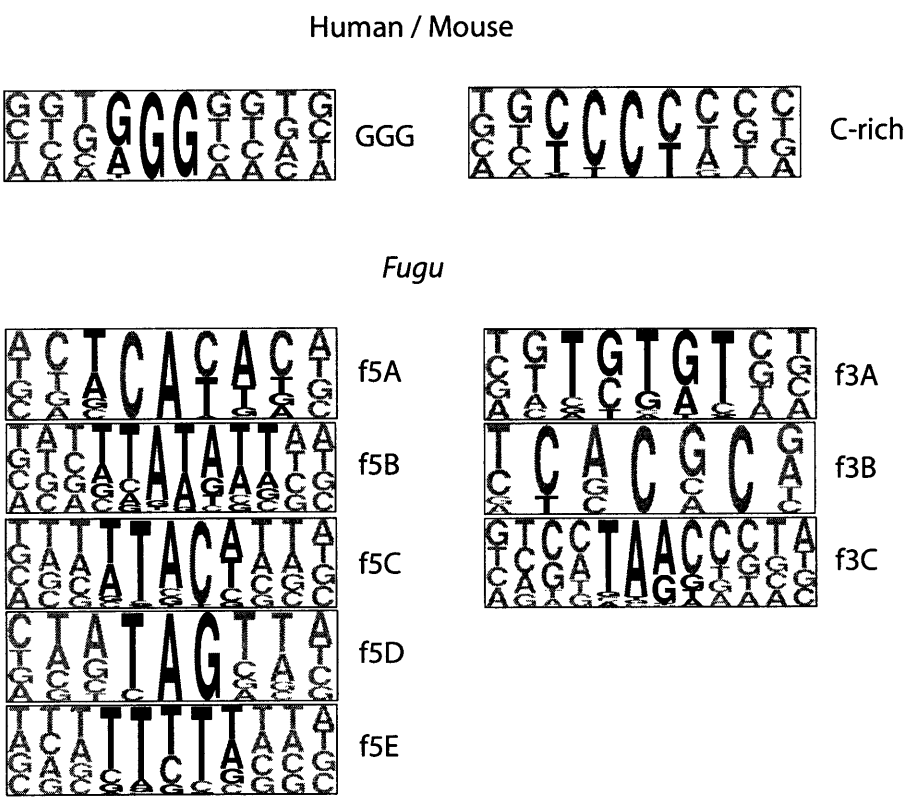


Figure 3

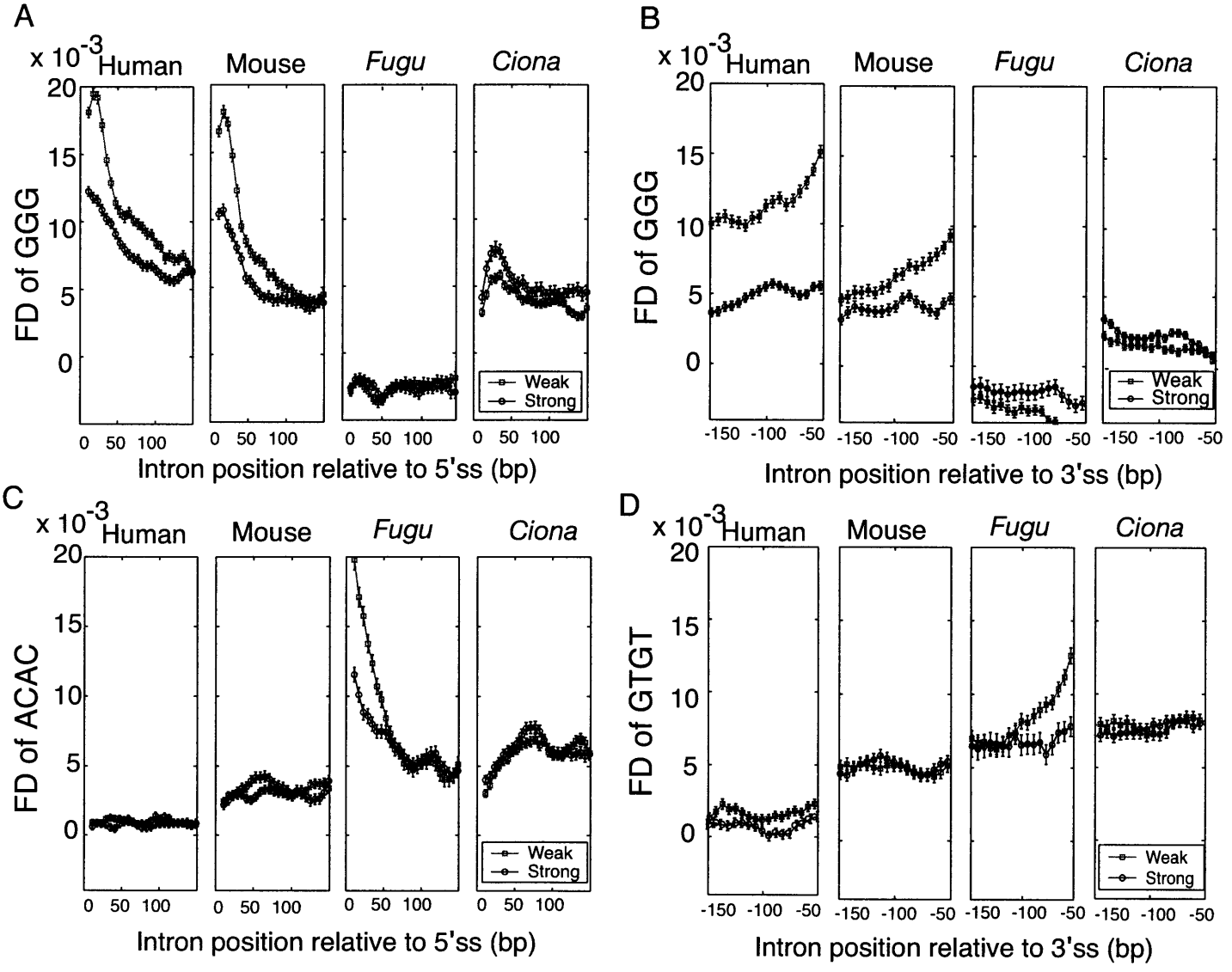

A
x 10-3
Human
Mouse
Fugu
Ciona
FD of GGG
Weak
Strong
Intron position relative to 5'ss (bp)
B
Intron position relative to 3'ss (bp)
C
FD of ACAC
D
FD of GTGT

Figure 4

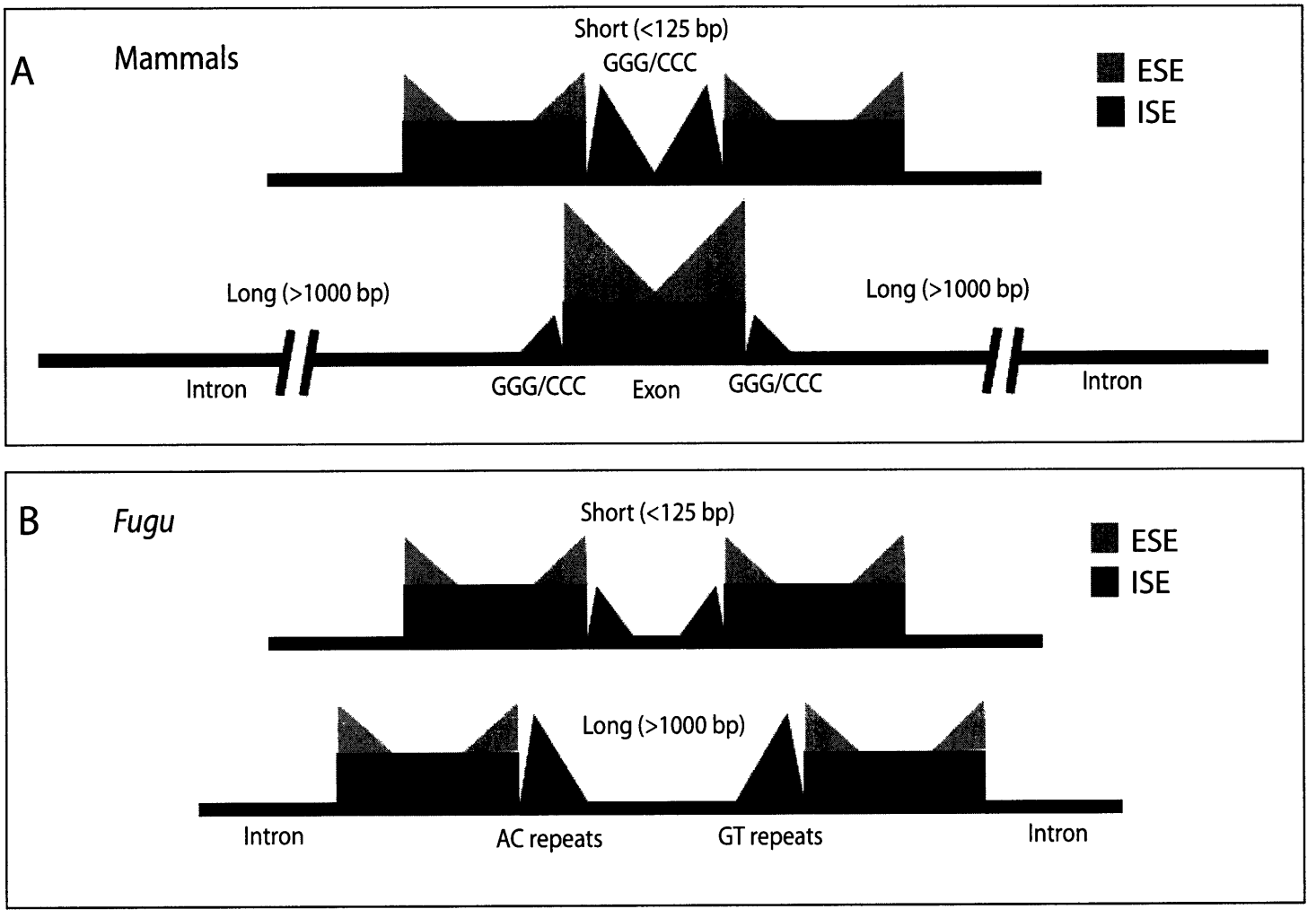
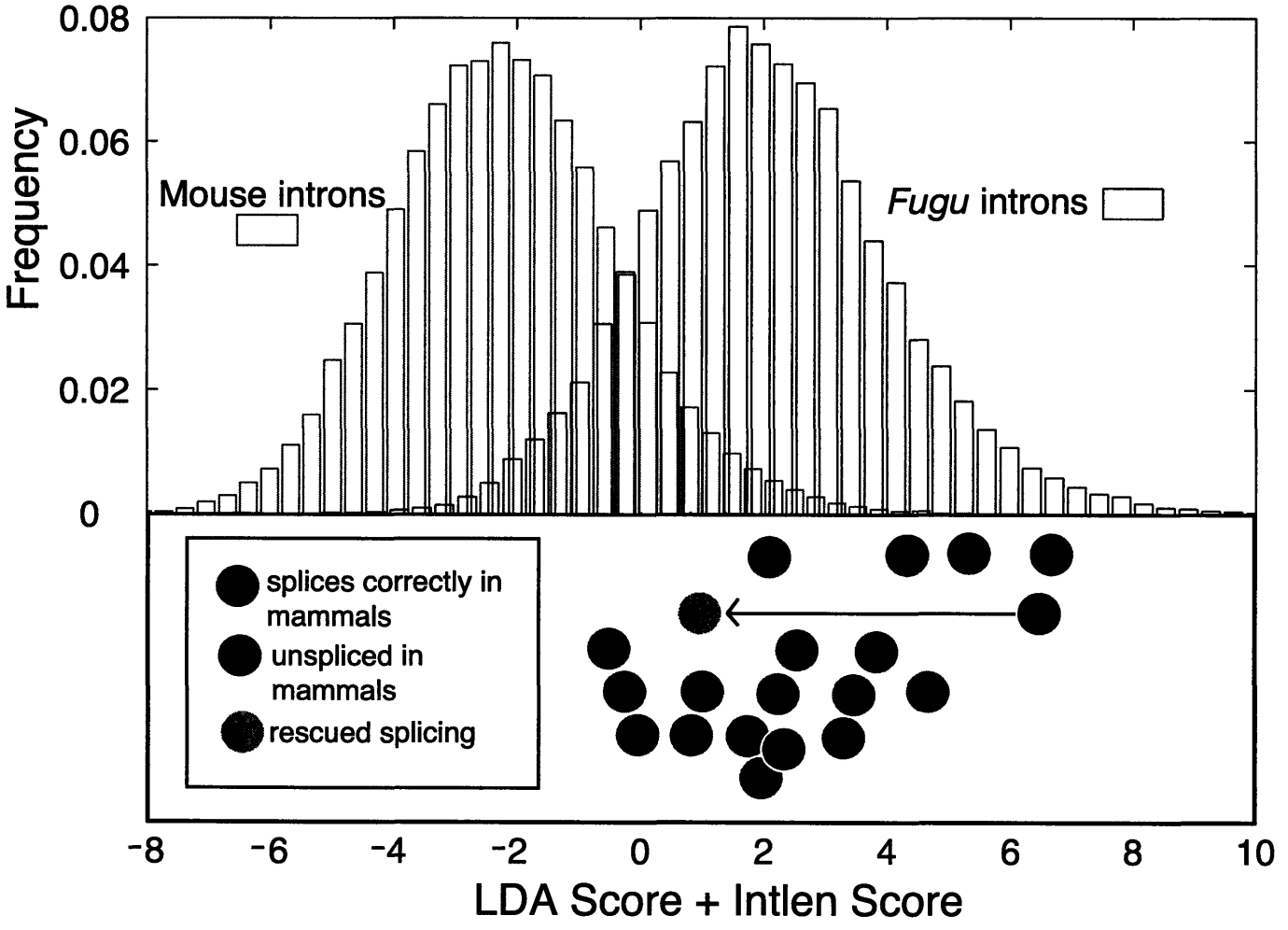

A
Mammals
Short (<125 bp)
GGG/CCC
ESE
ISE
Long (>1000 bp)
Long (>1000 bp)
Intron
GGG/CCC
Exon
GGG/CCC
Intron
B
Fugu
Short (<125 bp)
ESE
ISE
Long (>1000 bp)
Intron
AC repeats
GT repeats
Intron

Figure 5


0.08
0.06
0.04
0.02
0
Frequency
Mouse introns
Fugu introns
splices correctly in mammals
unspliced in mammals
rescued splicing
-8
-6
-4
-2
0
2
4
6
8
10
LDA Score + Intlen Score

Supporting Text

RESCUE-ESE and RESCUE-ISE

In order to identify oligonucleotides of size k (k-mers) that are over- (or under-) represented in one set of sequences relative to another, we assigned a difference score as follows. For sequence sets A and B totaling N_A bp and N_B bp, respectively, a k-mer occuring f_A times in set A and f_B times in set B is assigned a score ΔAB:

$$\Delta AB = \frac{(f_A - f_B)}{\sqrt{(1/N_A + 1/N_B)g(1-g)}} \quad (1)$$

where $g = \frac{(N_A f_A + N_B f_B)}{(N_A + N_B)}$. A statistical significance threshold of 2.5 standard deviations above (or below) the mean (corresponding to a P-value of ≈ 0.01) allows us to identify significantly over- or under-represented k-mers. Following the logic used previously to identify ESE sequences [1], we define predicted 5′ ISE oligonucleotides as k-mers that have the following two properties: (i) significant over-representation in introns versus exons; and (ii) significant over-representation in introns with weak 5′ splice sites (W5 introns) versus introns with strong 5′ splice sites (S5 introns). To meet the first criterion, a k-mer must satisfy $\Delta E5, I5 < -2.5$, where, as in the RESCUE-ESE approach, $E5$ and $I5$ are sets of exon and intron sequences that include sequences within 200 bases from the 5′ splice site (i.e. including the the entire sequence of exons $\leq$ 200 bases in length and the first 200 bases of longer exons, and analogously for introns). This convention is based on the reasoning that most ISEs that influence 5′ splice site choice are likely to be located near the 5′ss and it also helps to prevent very long introns (or exons) from exerting undue influence on the k-mer frequencies used. To meet the second criterion, a k-mer must satisfy $\Delta W5I, S5I > 2.5$, where $W5I$ is the set of introns whose 5′ splice site weight matrix model (WMM) scores are in the bottom 25% of all introns, and $S5I$ is the set of introns whose 5′ splice site WMM scores are in the top 25% of introns. Analogously, predicted 3′ ISE oligonucleotides are defined as k-mers satisfying $\Delta E3, I3 < -2.5$ and $\Delta W3I, S3I > 2.5$. Candidate ISE oligonucleotides are then clustered using the protocol devised in the RESCUE-ESE approach [1].

Spliced Gene Sets

Ensembl (http://www.ensembl.org/) core datasets for *Fugu rubripes* 11.2, *Danio rerio* 11.08, *Homo sapiens* 11.31 and *Mus musculus* 11.3 were used in this analysis [2]. We also downloaded 1,512 *Danio rerio* genes from Ensembl. *Ciona intestinalis* genomic and transcript sequence data were downloaded from the Department of Energy Joint Genome Institute website [3]. We aligned the 15,852 mRNAs to the genome using sim4 [4] to generate 55,404 exons and 82,064 introns.

Ortholog Identification

A list of known orthologous human, mouse and *Fugu* genes (21,804 entries) were obtained from EnsMart (http://www.ensembl.org/EnsMart/). The list was parsed to keep only genes that are present in all three species, a total of 8,520 sets of non-redundant orthologous genes. For an orthologous set of human, mouse and *Fugu* genes, exons in human were first aligned to exons in the orthologous mouse gene using BLAST (with thresholds of bit

score greater than 20 and percent identity greater than 75%), and were retained if the highest scoring pairs of exon are contiguous. This procedure was repeated by aligning the exons of the same human gene against the exons of the orthologous *Fugu* gene, and the subset of human exons that had observed orthology with both mouse and *Fugu* defined the final set of orthologous exons across human, mouse and *Fugu*. The 10,580 orthologous introns were defined as introns with flanking conserved orthologous exons defined as above.

Trans-factor Datasets

Accession numbers of known SR proteins, non-snRNP splicing proteins, H complex proteins and a subset of snRNP proteins were obained from Zhou et. al. [5]. Motifs were predicted by hmm-pfam using the PFAM database[6]. The orthologous Ensembl gene identifiers (version 16) and abbreviations are represented in Tables 5-8 of supporting information.

Classical Splice Signals

The sequence patterns in Fig. 7A (supporting information) are constructed using the Pictogram program (http://genes.mit.edu/pictogram.html). The frequencies of the four DNA nucleotides A, C, G and T are represented by the heights of the corresponding letters, with the letters shown in decreasing order of frequency from top to bottom. 5′ splice sites are represented by aligning sequences from positions -3 to +6 and 3′ splice sites were derived by aligning sequences from positions -21 to +3 of introns. The number of human, mouse, *Fugu* and *Ciona* 5′ and 3′ splice sites used to determine a consensus motif were 100,391 5′ss and 100,365 3′ss for human, 91,417 5′ss and 91,241 3′ss for mouse, 109,867 5′ss and 108,572 3′ss for *Fugu* and 69,934 3′ss and 70,755 5′ss for *Ciona* (obtained from spliced gene sets described above).

Distribution of Putative Branch signals

Branch signals were identified using the Gibbs sampling algorithm [7] as described previously in [8], demanding an A in position 6, using a dataset of introns between 60 and 500 bases long. Weight matrix models (WMMs) of the branch signals for human, mouse and *Fugu* were used to score the sequences that generated the consensus to find the thresholds for potential branch signals in a 100 bp region upstream of the 3′ss. The thresholds were set to be half a standard deviation below the mean (Thresholds in bits: human (6.64), mouse (6.58) and *Fugu* (6.80)). 7-mer motifs scoring higher than this cutoff in 30 base windows are defined to be possible branch signals. Background frequencies of A:0.3, C:0.2, G:0.2 and T:0.3 were determined for human and mouse and A:0.27, C:0.23, G:0.23 and T:0.27 for *Fugu*. We determined the frequency of occurrences of potential branch signals in 30 base windows in our orthologous intron dataset, excluding 10 bases upstream of the 3′ss (Fig. 7B, supporting information). 1,343, 1,154 and 956, 228 and 638 sequences comprise the branch signals for human, mouse, *Fugu*, zebrafish and *Ciona* respectively.

Clustering hexamers

A similar clustering procedure described previously was used to cluster hexamers [1]. Each pair of hexamers is assigned a dissimilarity distance defined as the number of shifts plus the number of mismatches in the best local

alignment of the two hexamers. The resulting dissimilarity matrix was used to cluster the hexamers using standard average linkage hierarchical clustering implemented in the R statistical package. Clusters were defined by suitable cutoffs and hexamers in each cluster of 4 or more members were aligned using ClustalW with default parameters. Pictogram (http://genes.mit.edu/pictogram.html) was used to generate each motif, with pseudocounts to pad edge positions not present in all positions.

Construction of the Arp3N1 plasmid

The *Fugu* Arp3 gene spanning from 1.4kb upstream of the transcription start to 2.7kb downstream of the polyadenylaition signal sequence was cloned into an EcoRI site in pBluescript II KS as described previously [9]. This plasmid was digested with SacI and cloned into another pBluescript II KS vector. This was followed by a partial digestion with *NotI* and *SacI*. The resulting fragment was cloned into the pEGFPN1 expression vector containing the highly inducible CMV promoter while removing the reporter gene in the process. This resulted in a 7.9kb Arp3N1 construct which is expressible in both mammalian and fish cell lines.

Cell Culture

293T (human embryonal kidney) and PLHC-1(top minnow hepatoma) were obtained from American Type Culture Collection. Both cell lines were cultured in Dulbecco's modified Eagles medium and supplemented with 10% fetal calf serum, anti-mycotic, penicillin and streptomycin. 293T was grown at 37 °C in 5% CO2 while PLHC-1 cells were grown at 25 °C in 5% CO2 .

Transient Transfection

293T cells were plated onto 6-well tissue culture collagen plates and grown to 95% confluency. PLHC-1 cells were plated onto 6-well tissue culture plates and grown to 95% confluency. Transfection was performed by complexing the Arp3N1 construct with the cationic lipids DMRIE-C (Invitrogen), according to manufacturer's instruction. 2μg of DNA per well was used. The transfection mix was then added to the 293T and PLHC-1 cell lines and incubated for 6 h at 37 °C and 25 °C respectively. Transfection was stopped by adding 2 ml cell media and 35 μl/ml fetal bovine serum and the cells were incubated for a further 36 hours.

Aberrant splicing products

Total RNA was isolated from cell cultures following induction using the TRIZOL reagent (Gibco-BRL). RNA was treated with Dnase I to remove any traces of genomic DNA. cDNA was synthesized using the SuperScriptTM One-Step System from Invitrogen. The Arp3 cDNA was amplified by PCR using three sets of gene specific primers, ARF1: 5′-ACACATGGC GGGCCGTCTAC-3′, ARR1: 5′-GGTTGTGGCGGCAGATGCTG-3′; ARF2: 5′-CCTCCTCTGAACACGCCAGAG-3′, ARR2: 5′-GGCGTTGATGCCTGTGTACTG3′; ARF3: 5′-CATCGTTGAGGACTGGGACCTG-3′, ARR3: 5′-GCGTGTTCAGAGGAGGCTCCG-3′. The region amplified by ARF1 and ARR1 spans exon 1 and exon 12 of the ARP3 transcript and would result in a 1.2kb PCR product if spliced correctly. The region amplified by ARF2 and ARR2 spans exon 5 and exon 8 of the ARP3 transcript and

would result in a 448bp product if spliced correctly. The region amplified by ARF3 and ARR3 spans exon 4 and exon 5 of the ARP3 transcript and would result in a 116 bp product if spliced correctly. PCR using primers (ARFR1) flanking exon 1 and exon 12 revealed a 1.2kb product. However upon using the internal primers ARFR2 that flank exon 5 and exon 8, it was revealed that the cDNA extracted from 293T cell line yielded a smaller product of about 200bp. To confirm the presence of aberrant splicing, both 1.2kb transcripts were cloned into a pGEM-T vector and sequenced. The sequence confirmed the presence of aberrant splicing in the 293T cell line while the transcript from the PLHC-1 cell line was spliced correctly. We found 3 different aberrant splicing phenotypes in the transcript from the 293T cells. There were 2 cases of intron inclusion (intron 4 and intron 9), 1 case of exon skipping (exon 7) and 1 case of exon truncation through the use of a 5′ cryptic splice site within exon 5.

Site-directed Mutagenesis

Insertion of G triplets into intron 4 of the *Fugu* ARP3 gene was accomplished by PCR-based site-directed mutagenesis using the PfuTurbo DNA Polymerease (Stratagene). The inserted sequence resembled the G1 and G2 triplets in intron 2 of the human alpha globin gene [10]. Primers were diluted to 125 ng/μl and constructs to 50 ng/μl. The primers designed to insert 1 pair of G-triplets were: CTGTTGGTGAGGACAgggtcgaggggCAACTCGCCCACCTGTTC (M2F), and GAACAGGTGGGCGAGTTGcccctcgacccTGTCCTCACCAACAG (M2R), where the inserted sequences are indicated in lower case. This resulted in the construct we term M2F8. The PCR reaction was carried out at a denaturing temperature of 95 °C for 30 sec, and 35 cycles of 95 °C for 30 sec, annealing at 55 °C for 1 min and extension at 68 °C for 20 min. The products were digested with *DpnI* at 37 °C for 1 hr to remove the parental DNA template. The second pair of G-triplets was inserted into the M2F8 construct by a second round of mutagenesis using the forward and reverse primers: GGGGCAACTCGCCCAgggccgggCCTGTTCTACCGGTG (M5F) and CACCGGTAGAACAGGcccggcccTGGGCGAGTTGCCCC (M5R), resulting in the construct we term M5F2 (Fig. 6). All mutated PCR products were subsequently transformed into *E. coli* and sequenced to confirm the presence of the correct inserted sequence.

References

[1] Fairbrother, W., Yeh, R., Sharp, P. & Burge, C.B. (2002) *Science* **297,** 1007-1013.

[2] 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 Research* **30,** 38-41.

[3] Dehal P., Satou Y., Campbell R.K., Chapman J., Degnan B., De Tomaso A., Davidson B., Di Gregorio A., Gelpke M., Goodstein D.M. et al. (2002) *Science* **298,** 2157-2167

[4] Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. & Miller, W. (1998) *Genome Research* **8,** 967-74.

[5] Zhou, Z., Licklider, L., Gygi, S. & Reed, R. (2002) *Nature* **419,** 182-185.

[6] Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M. & Sonnhammer E.L.(2002) *Nucleic Acids Research* **30,** 276-280.

[7] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. & Wootton, J.C. (1993) *Science* **262,** 208-214.

[8] Lim L.P. & Burge C.B., (2001) *Proc Natl Acad Sci USA* **98,** 11193-11198.

[9] Venkatesh, B., & Brenner, S. (1998) *Gene* **211,** 169-75.

[10] McCullough, A. & Berget, S. (1997) *Mol, Cellular Biol* **17,** 4562-4571.

PAGES (S) MISSING FROM ORIGINAL

PAGES (S) MISSING FROM ORIGINAL

Table 2. Splice site cutoffs and the sizes of sequences utilized.

	Splice	Mouse	Human	*Fugu*	*Ciona*
Score Cutoffs	3'ss	(6.99,11.53)	(6.99,11.56)	(7.04,11.29)	(6.24,9.70)
Intron number	3'ss	(20,912; 21,004)	(25,257; 25,358)	(35,449; 13,128)	(16,510; 17,000)
Exon number	3'ss	(23,920 ; 22,717)	(26,457; 22,921)	(39,747; 11,954)	
Score Cutoffs	5'ss	(6.32,9.60)	(6.32,9.48)	(5.68,8.64)	(5.74,9.47)
Intron number	5'ss	(22,806; 22,776)	(25,075; 24,918)	(27,144; 27,269)	(16,739; 7,093)
Exon number	5'ss	(27,165 ; 19,665)	(30,236; 21,613)	(37,435; 20,023)	

Weak and strong cutoffs for 5′ and 3′ splice sites in bits and the corresponding number of introns and exons partitioned into weak and strong sets. The first number in the brackets is the number of weak introns or exons, and the second number is the number of strong introns or exons, defined relative to the upstream 5′ or downstream 3′ splice sites.

Introns were required to be longer than 60 bases, resulting in a total of 100,402 human, 91,691 mouse, 109,038 *Fugu* and 69,934 *Ciona* introns). For each species, the splice site cutoffs were used to partition the introns into strong and weak 5′ss and strong and weak 3′ss introns. Strong and weak cutoffs for splice sites are determined by the 75th and 25th percentile of scores respectively, as described previously in [1]. The 5′ss background frequencies were obtained from sets of generated decoys (9-mers with a GT in the 4th and 5th positions): A:0.25, C:0.20, G:0.25, T:0.28 for human and mouse, A:0.25, C:0.25, G:0.24, T:0.26 for *Fugu* and A:0.32, C:0.14, G=0.18, T=0.35 for *Ciona*. The 3′ss background frequencies were obtained similarly (23-mers with an AG in the 19th and 20th positions): A:0.25, C:0.25, G:0.25, T:0.25 for human and mouse, A:0.25, C:0.25, G:0.25, T:0.25 for *Fugu* and A:0.3, C:0.17, G:0.17, T:0.35 for *Fugu*. From 105,380 human, 95,880 mouse and 113,628 *Fugu* exons, we partitioned the dataset into strong and weak exons as described above. Exonic regions are defined relative to the splice sites i.e. 5′ exon refers to the exonic region nearest to the 5′ss, and 3′ exon refers to the region nearest to the 3′ss.

References

[1] Fairbrother, W., Yeh, R-F., Sharp, P.A. & Burge, C.B. (2002) *Science* **297,** 1007-1013.

Table 3. Method of enumeration of non-overlapping occurrences of an oligonucleotide.

Sequence	Pattern(s)	Counts
GGGgcccGGGccc	GGG	2
GGGgCCCGGGCCC	GGG or CCC	4
GGGGCCcGGGccc	GGG or GCC	3
AATCAAacaATCAAA	AATCAA or ATCAAA	2

Consider sequences ggggcccgggccc and aatcaaacaatcaaa. Notice in the last two entries that we do not count overlapping patterns.

Table 4. Differential enrichment of RESCUE-ESEs in exons flanked by introns of different lengths.

ESE	Human 5′ S<M	Human 5′ M<L	Human 3′ S<M	Human 3′ M<L	Fugu 5′ S<M	Fugu 5′ M<L	Fugu 3′ S<M	Fugu 3′ M<L
3B	+	+	+	+	+	−	+	−
3C	NS	NS	−	NS	NS	+	+	+
3E	+	NS	+	+	NS	−	NS	−
3F	+	+	+	+	−	−	NS	−
3H	+	NS	+	+	−	NS	−	NS
5A3G	+	+	+	+	+	−	NS	NS
5B3A	+	+	NS	+	−	−	NS	−
5C3D	+	+	+	+	NS	−	NS	−
5D	+	+	NS	NS	NS	NS	NS	NS
5E	NS	−	NS	−	NS	NS	NS	−

Significance of difference in mean enrichment (frequency observed − frequency expected) of ESEs in exonic regions (+10 to +70 and −10 to −70) flanked by short (<125 bp) versus medium (125-1000 bp), and medium versus long (>1000bp) introns. + indicates significance (p <0.001) of enrichment in the direction indicated (e.g. a + in the human 5′ S<M column indicates that human exons flanked by short introns have fewer ESEs than exons flanked by medium introns) as measured by a Wilcoxon rank-sum test. − indicates significance in the opposite direction. NS indicates non-significance.

Table 5. Domain conservation of U1 snRNP specific proteins.

Accession No.	Protein Name	Human	Mouse	*Fugu*
P08621	U1-70kD	1 RRM	1RRM	No RRM
P09012	U1 A	2RRM Pro-rich	2RRM Pro-rich	2RRM
P09234	U1 C	ZF Pro-rich	ZF Pro-rich	ZF
P08621	U1-70kD	104852	030810	124896
P09012	U1 A	077312	040518	128942
P09234	U1 C	124562	024217	120439

Swiss-Prot accession numbers, predicted domains, and thelast 6 digits of the Ensembl gene identifiers for human (ENSG00000), mouse (ENSMUSG00000) and *Fugu* (SINFRUG00000) are listed.

Table 6. Domain conservation of serine-arginine (SR) proteins.

Accession No.	Protein Name	Human	Mouse	*Fugu*
Q08170	SRP75	2RRM, RS	2 RRM, RS	2 RRM, RS
Q05519	p54/SFRS11	1RRM, RS	1RRM, RS	1RRM, RS
Q13247	SRp55	2RRM, RS	2RRM, RS	2RRM, RS
Q13243	SRp40	2RRM, RS	2RRM, RS	2RRM, RS
Q07955	ASF/SF2	2RRM, RS	2RRM, RS	2RRM, RS
Q16629	9G8	1RRM, RS, ZF	1RRM, RS, ZF	1RRM, RS, ZF
Q01130	SC35	1RRM,RS,	1RRM, RS	1RRM, RS
Q13242	SRp30c	2RRM, RS	2RRM, RS	2RRM, RS
AF057159	hTra2	1RRM, RS	1RRM, RS	1RRM, RS
P23152	SRp20	1RRM, RS	1RRM, RS	1RRM, RS
Q08170	SRP75	116350	028911	144119
Q05519	p54/SFRS11	116754	039971	124246
Q13247	SRp55	124193	016921	148668
Q13243	SRp40	100650	021134	125175
Q07955	ASF/SF2	136450	018379	129244
Q16629	9G8	115875	024097	136446
Q01130	SC35	161547	034120	152188
Q13242	SRp30c	111786	029538	137486
AF057159	hTra2	136527	022858	140787
P23152	SRp20	112081	034437	138994

Swiss-Prot accession numbers, predicted domains, and thelast 6 digits of the Ensembl gene identifiers for human (ENSG00000), mouse (ENSMUSG00000) and *Fugu* (SINFRUG00000) are listed.

Table 7. Domain conservation of heterogeneous (H) complex proteins.

Accession No.	Protein Name	Human	Mouse	*Fugu*
Q13151	hnRNP A0	2RRM	2RRM	2RRM
P09651	hnRNP A1	2RRM	2RRM	2RRM
P22626	hnRNP A2/B1	2RRM	2RRM	No Ortholog
P51991	hnRNP A3	2RRM	2RRM	2RRM
P07910	hnRNP C1/C2	1RRM	1RRM	1RRM
Q14103	hnRNP D0	2RRM	2RRM	2RRM
NM_004966	hnRNP F	3RRM	3RRM	No Ortholog
L22009	hnRNP H	3RRM	3RRM	2RRM
P26599	hnRNP I/PTB	4RRM	4RRM	3RRM
Q07244	hnRNP K	3KH	3KH	1KH
P14866	hnRNP L	3RRM	1RRM	2RRM
043390	hnRNP R	3RRM	3RRM	3RRM
AL031668	hnRNP RALY	1RRM	1RRM	No Ortholog *
P35637	hnRNP FUS/hnRNP P2	1RRM	1RRM	1RRM
B54857	NF-AT 90K	2DSRM	2DSRM	2DSRM
A54587	NF-AT 45K	25A_synth	25A_synth	25A_synth
AF037488	GRY-RBP	3RRM	3RRM	--
P43243	Matrin3	2RRM	2RRM	1RRM
043684	hBUB3	WD40	WD40	WD40
Q15717	HuR	3RRM	3RRM	3RRM
Q92804	TAFII68	1RRM, ZF	1RRM, ZF	1RRM, ZF
P16991	YB1	CSD	CSD	CSD
P16989	DBPA	CSD	CSD	No Ortholog
P08107	HSP70	HSP70	HSP70	No Ortholog *
P11142	HSP71	HSP71	HSP71	HSP71
Q13151	hnRNP A0	177733	007836	145410
P09651	hnRNP A1	135486	036021	146321
P22626	hnRNP A2/B1	122566	004980	No Ortholog
P51991	hnRNP A3	176825	047468	124772
P07910	hnRNPC1/C2	092199	004563	144808
Q14103	hnRNP D0	138668	000568	129026
NM_004966	hnRNP F	169813	042079	No Ortholog
L22009	hnRNP H	169045	007850	129688
P26599	hnRNP I/PTB	011304	006498	125066
Q07244	hnRNP K	165119	021546	135555
P14866	hnRNP L	104824	015165	136448
043390	hnRNP R	125944	028666	151323
AL031668	hnRNP RALY	125970	027593	No Ortholog *
P35637	hnRNP FUS/hnRNP P2	089280	030795	125014
B54857	NF-AT 90K	129351	032178	144346
A54587	NF-AT 45K	143621	001016	133792
P43243	Matrin3	015479	037236	120558
043684	hBUB3	154473	015688	151235
Q15717	HuR	066044	040028	135392
Q92804	TAFII68	172660	020680	152729
P16991	YB1	065978	028639	137733
P16989	DBPA	060138	030189	No Ortholog
P08107	HSP70	096469	007033	No Ortholog *
P11142	HSP71	109971	015656	141042

Swiss-Prot accession numbers, predicted domains, and the last 6 digits of the Ensembl gene identifiers for human (ENSG00000), mouse (ENSMUSG00000) and *Fugu* (SINFRUG00000) are listed. Asterisks indicate genes having zebrafish orthologs in Ensembl.

Table 8. Abbreviations for domain descriptions.

Domain Name	Description	Pfam Accession Number
25A_synth	2'-5'-oligoadenylate synthase N-terminal region profile.	--
AT_hook	AT hook motif	PF02178
CSD	'Cold-shock' DNA-binding domain	PF00313
CPSFA	CPSF A subunit region	PF03178
DEAD	DEAD/DEAH box helicase	PF00270
DIM1	Mitosis protein DIM1	PF02966
DSRM	Double-stranded RNA binding motif	PF00035
EFG_C	Elongation factor G C-terminus	PF00679
EFG_IV	Elongation factor G, domain IV	PF03764
GYF	GYF Domain	PF02213
Hat	HAT (Half-A-TPR) repeat	PF02184
Helicase_C	Helicase conserved C-terminal domain	PF00271
HSP70	Hsp70 protein	PF00012
KH	K homology domain	PF00013
MIF4G	Middle domain of eukaryotic initiation factor 4G domain	PF02854
LRR	Leucine Rich Repeat	PF00560
Mov34	Mov34/MPN/PAD-1 family	PF01398
Myb_DNA_binding	Myb-like DNA-binding domain	PF00249
NLS_BP	Bipartite nuclear localization signal	--
NOP	Putative snoRNA binding domain	PF01798
Pro_isomerase	Cyclophilin type peptidyl-prolyl cis-trans isomerase	PF00160
Pro-rich	Proline Rich	--
PWI	PWI Domain	PF01480
RNA_pol_Rpb1_R	RNA polymerase Rpb1 C-terminal repeat	PF05001
RIBOSOMAL_L7A	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family	PF01248
RRM	RNA recognition motif	PF00076
RS	Arginine Serine rich domain	--
SART_1	SART-1 family	PF003343
Sec63	Sec63 domain	PF02889
SKIP_SNW	SKIP/SNW domain	PF02731
TPR	TPR Domain	PF00515
Tudor	Tudor Domain	PF00567
UCH	Ubiquitin carboxyl-terminal hydrolase	PF00443
WD40	WD domain, G-beta repeat	PF00400
WW	WW domain	PF00397
ZF	Zinc Finger domain	PF04071

PAGES (S) MISSING FROM ORIGINAL

Supporting Figure legends

Fig. 6 Rescue of *Fugu* ARP3 intron 4 in human 293T cells. A. Mutants of intron 4 were generated by insertion of G triples (see main text). B. RT-PCR of mRNA: Lane 1: PLHC-1 transfected with wildtype *Fugu Arp3N1* (PLHC-1); Lane 2: 293T transfected with wildtype *Fugu Arp3N1*(WT); Lane 3: 293T transfected with mutant containing a single G triple insert (M2F8); Lane 4: 293T transfected with mutant containing two G triple inserts (M5F2).

Fig. 7 A. Classical Splice Signals (5'ss, branch, 3'ss: left to right column respectively) of *Homo Sapiens, Mus musculus, Danio rerio, Fugu rubripes* and *Ciona intestinalis.* B. Distribution of putative branch signals upstream of 3'ss in human, mouse and *Fugu.* Each point represents the midpoint of a window of size 30 bp.

Fig. 8 Histograms of intron lengths (log_{10} bp). Distributions were modeled as mixtures of 2 normal distributions.

Fig. 9 Frequency difference plots of RESCUE-ESEs in human, mouse and *Fugu* exons.

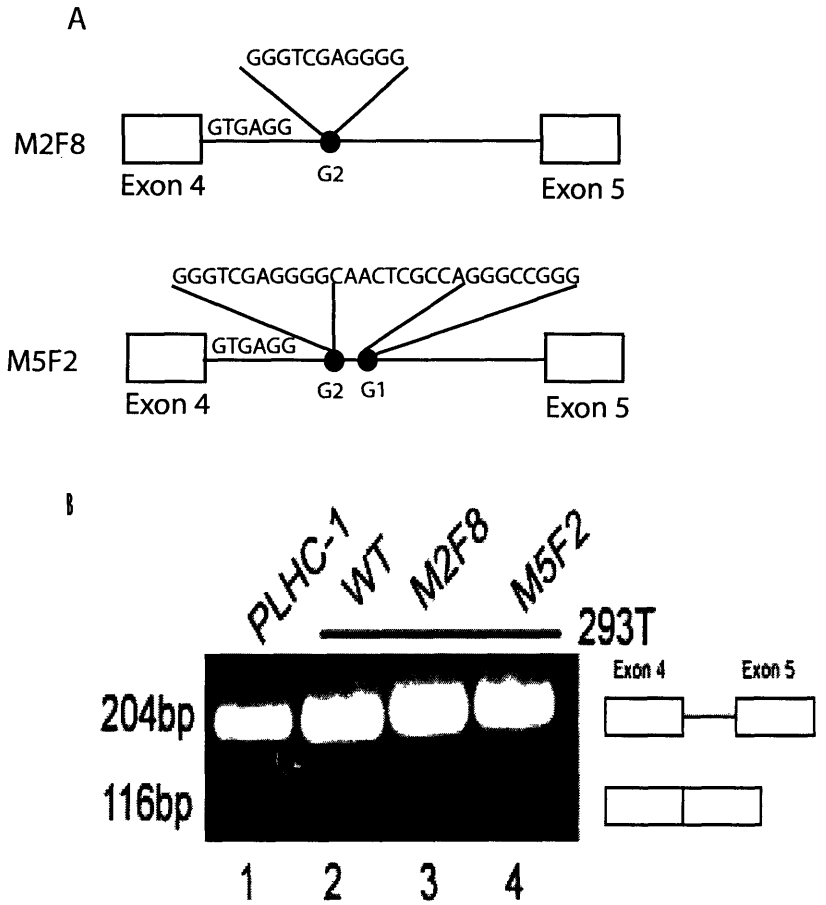
Fig. 10 Frequency difference plots of RESCUE-ISEs in five chordates.

Fig. 11 Frequency difference plots of RESCUE-ESEs in human exons flanked on both sides by introns of lengths <125 bp, $125 - 1000$ bp, or > 1000 bp.

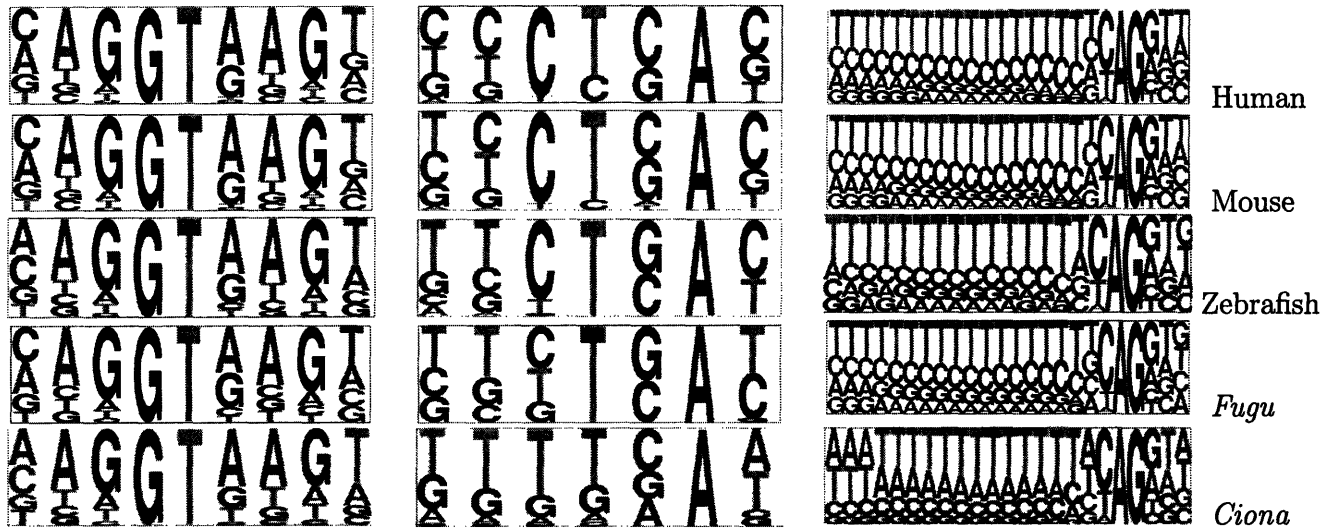
Fig. 12 Frequency difference plots of RESCUE-ISEs for human (GGG,CCC) and *Fugu* (ACAC,GTGT) in introns of three length groups as indicated.

Fig. 13 A. Binned frequencies of the lengths of *Fugu* and mouse introns. 64,313 *Fugu* introns and 74,908 mouse introns that are at least 2000 bases long were distributed into 50 bins. B. Binned scores from the Linear Discriminant Analysis (LDA) for an independent test set of 32,156 *Fugu* introns and 37,454 mouse introns (trained on independent training set of the same size; details in Methods and supporting text). C. Combined scores (LDA + intron length) for the independent test set of introns. 85% of *Fugu* introns the independent test set are classified as true *Fugu* introns, and 88% of mouse introns from the independent test set are predicted to be true mouse introns. Genes 1, 2 and 3 are the *Fugu* RCN1, HD and ARP3 gene. The location of scores for introns 1 to 5 of gene 1 (1.1-1.5), 1 to 7 of gene 2 (2.1-2.7) and 1 to 11 of gene 3 (3.1-3.11) with respect to the overall distributions are indicated by lines. A minus sign (-) indicates intron retention, a plus sign (+) indicates correct splicing and +/- indicates partial splicing of the corresponding *Fugu* introns in trangenic mice or mice cell lines, evident from the literature or by our aexperimental analyses. The table insert contains the corresponding intron lengths (bp) for the *Fugu* introns.

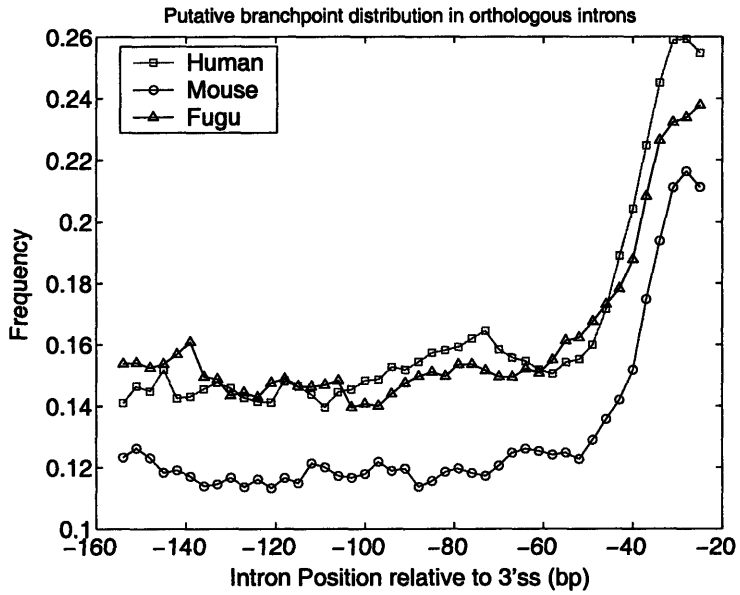
Fig. 14 Splicing Phenotype of *Fugu Arp3N1* expressed in Human 293T includes (a) unspliced introns 4 and 9 (b) truncated exon 5 (c) skipped exon 7.

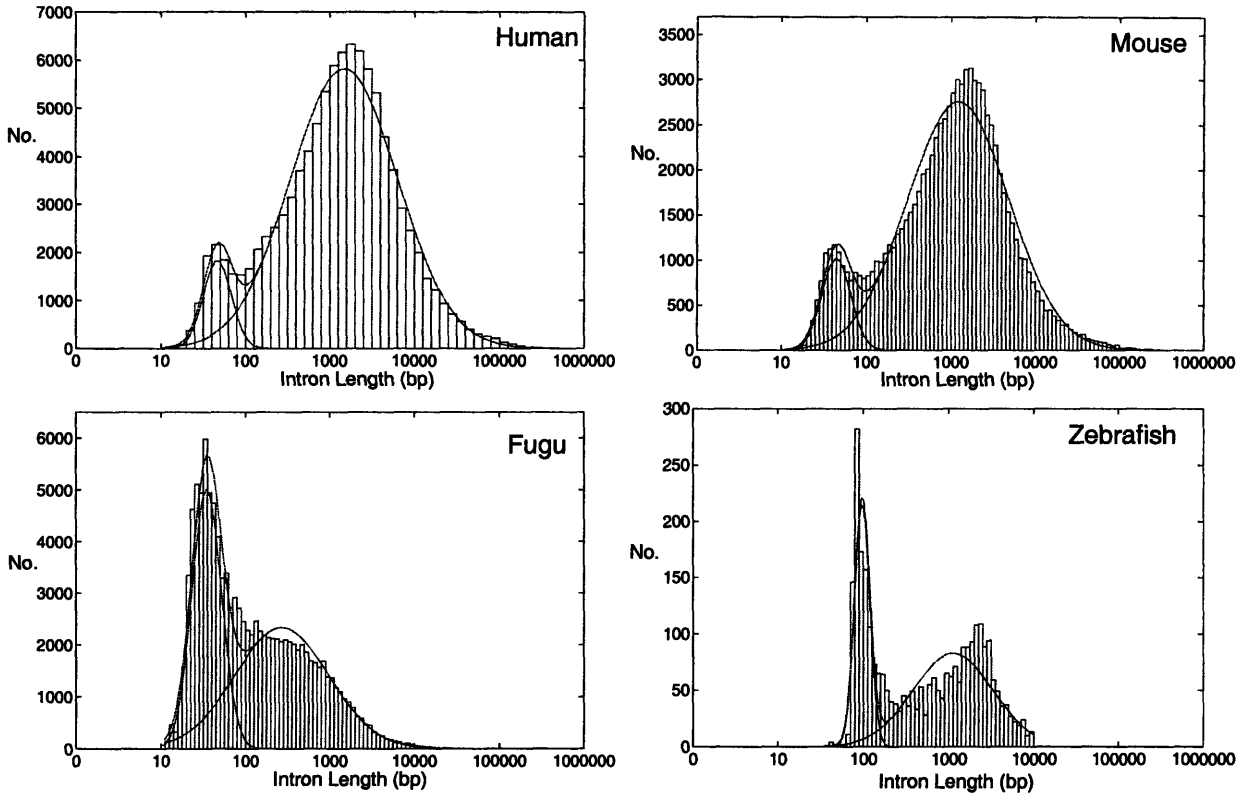

A
GGGTCGAGGGG
M2F8
GTGAGG
G2
Exon 4
Exon 5
GGGTCGAGGGGCAACTCGCCAGGGCCGGG
M5F2
GTGAGG
G2
G1
Exon 4
Exon 5
B
PLHC-1
WT
M2F8
M5F2
293T
Exon 4
Exon 5
204bp
116bp
1
2
3
4

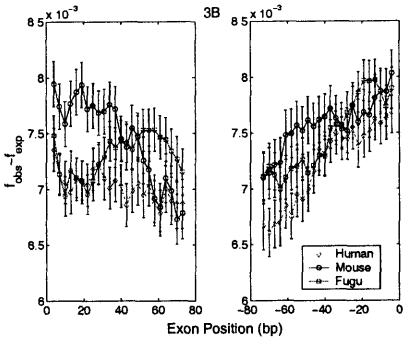
A

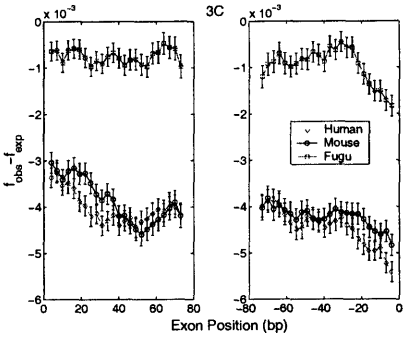

Human
Mouse
Zebrafish
Fugu
Ciona

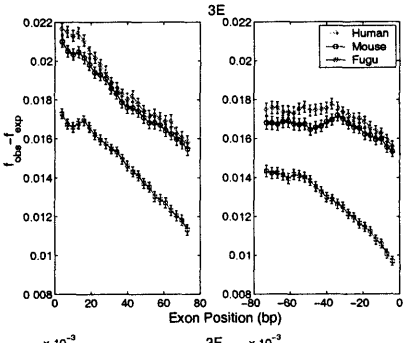
B

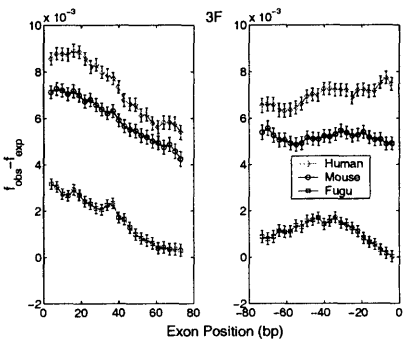

Putative branchpoint distribution in orthologous introns
Human
Mouse
Fugu
Frequency
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
−160
−140
−120
−100
−80
−60
−40
−20
Intron Position relative to 3'ss (bp)

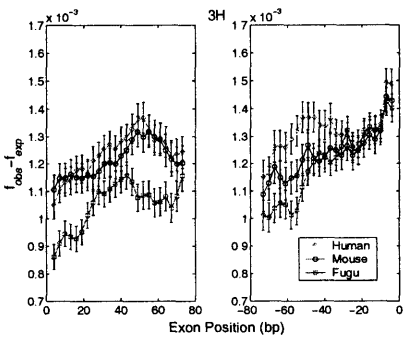

Human
No.
Intron Length (bp)
Mouse
No.
Intron Length (bp)
Fugu
No.
Intron Length (bp)
Zebrafish
No.
Intron Length (bp)

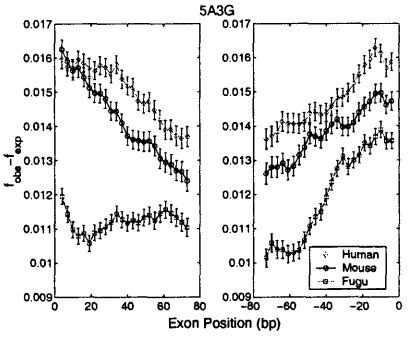

3B
Human
Mouse
Fugu
Exon Position (bp)

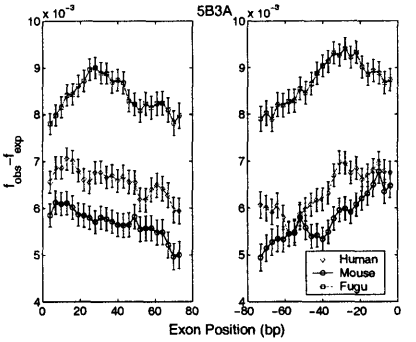

3C
Human
Mouse
Fugu
Exon Position (bp)

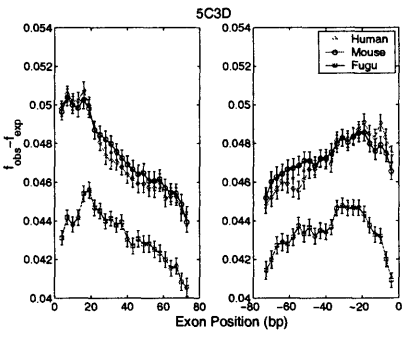

3E
Human
Mouse
Fugu
Exon Position (bp)

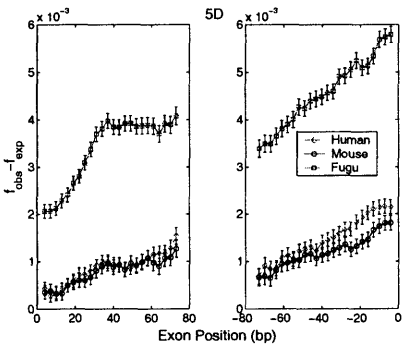

3F
Human
Mouse
Fugu
Exon Position (bp)

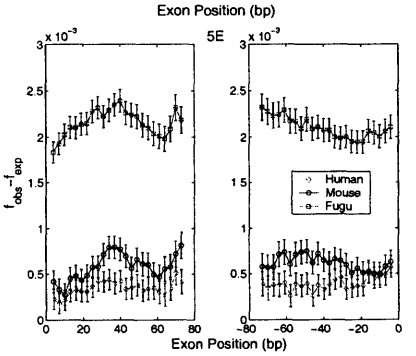

3H
Human
Mouse
Fugu
Exon Position (bp)

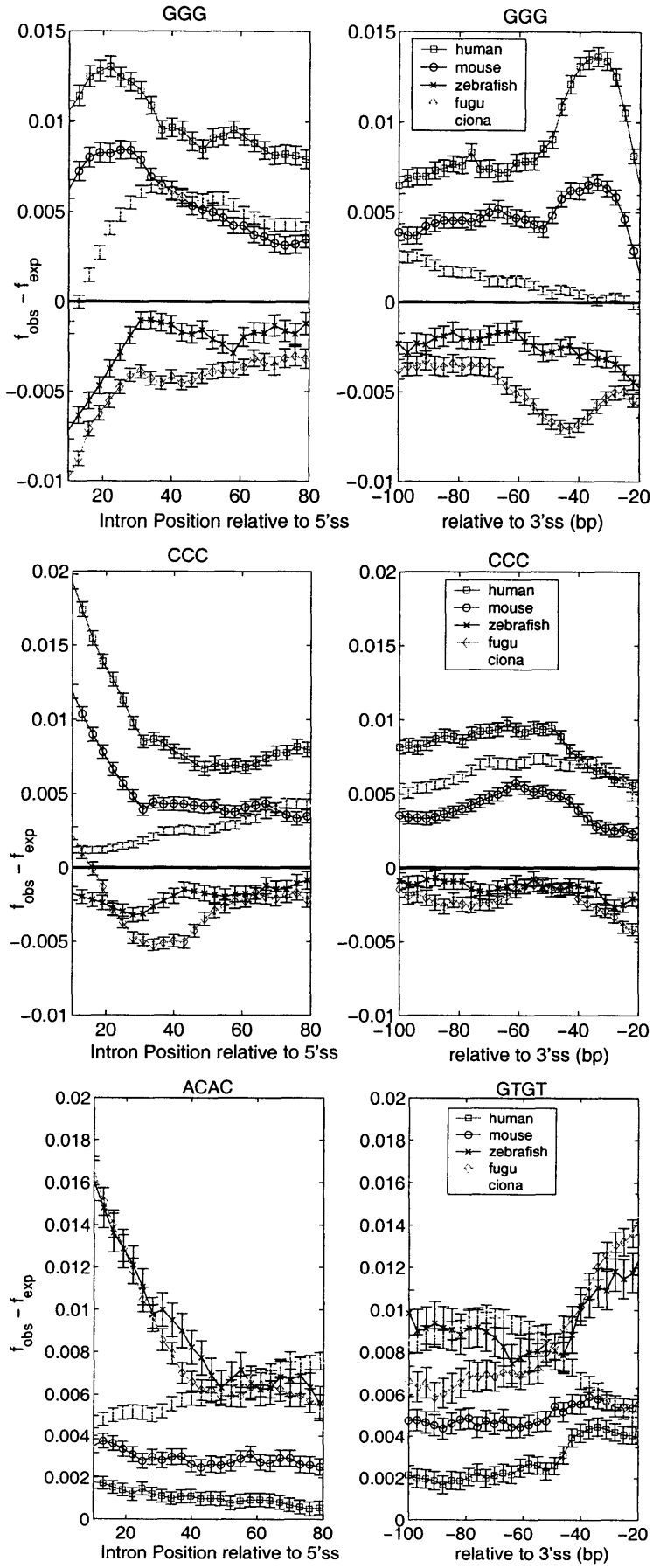

5A3G
Human
Mouse
Fugu
Exon Position (bp)

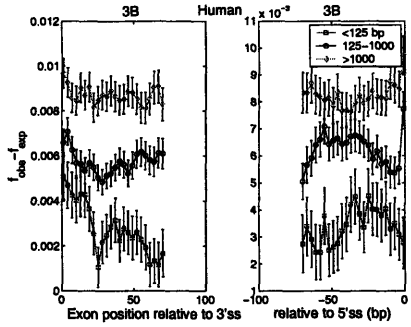

5B3A
Human
Mouse
Fugu
Exon Position (bp)

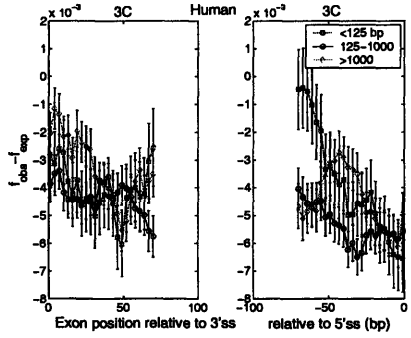

5C3D
Human
Mouse
Fugu
Exon Position (bp)

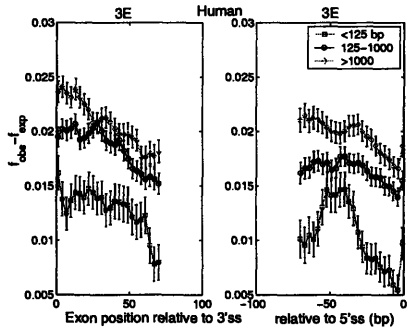

5D
Human
Mouse
Fugu
Exon Position (bp)

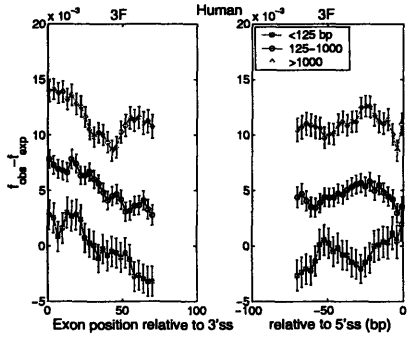

5E
Human
Mouse
Fugu
Exon Position (bp)

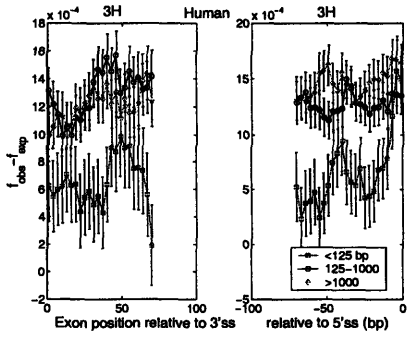

GGG
GGG
human
mouse
zebrafish
fugu
ciona
$f_{obs} - f_{exp}$
Intron Position relative to 5'ss
relative to 3'ss (bp)
CCC
CCC
ACAC
GTGT

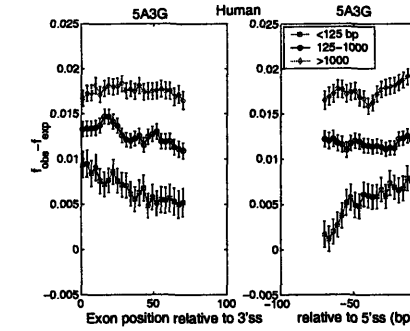

3B
Human
3B
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

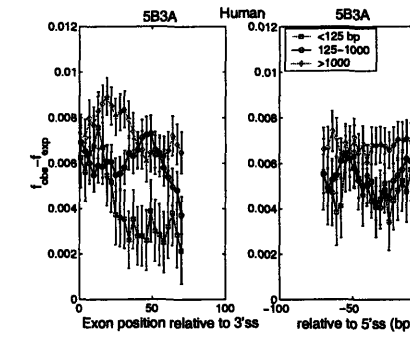

3C
Human
3C
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

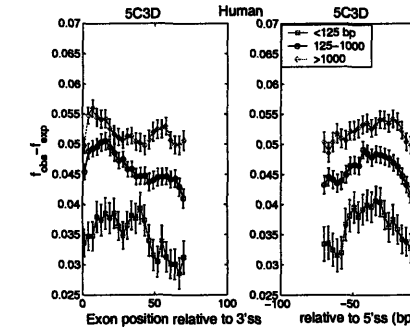

3E
Human
3E
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

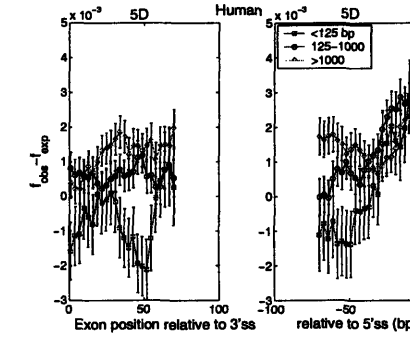

3F
Human
3F
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

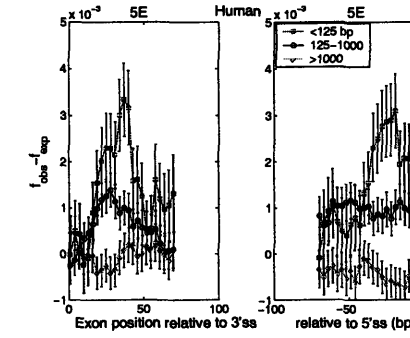

3H
Human
3H
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

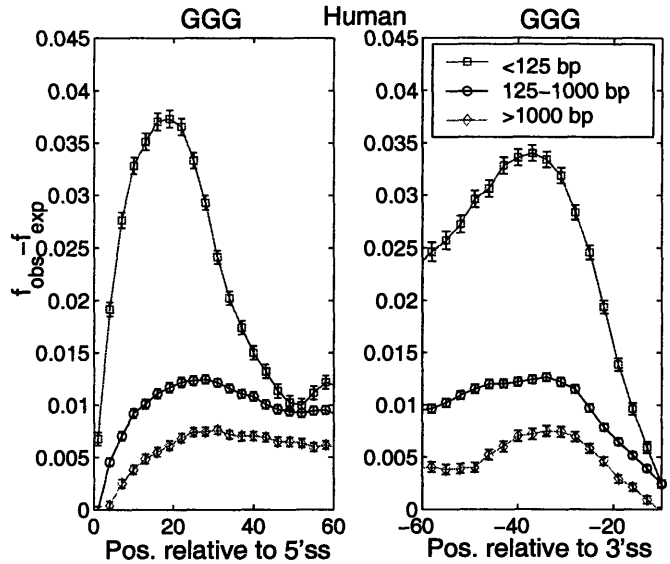

5A3G
Human
5A3G
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

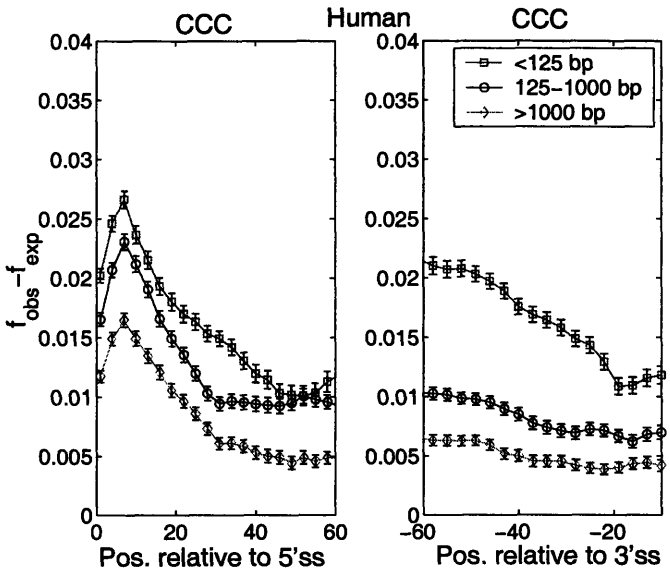

5B3A
Human
5B3A
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

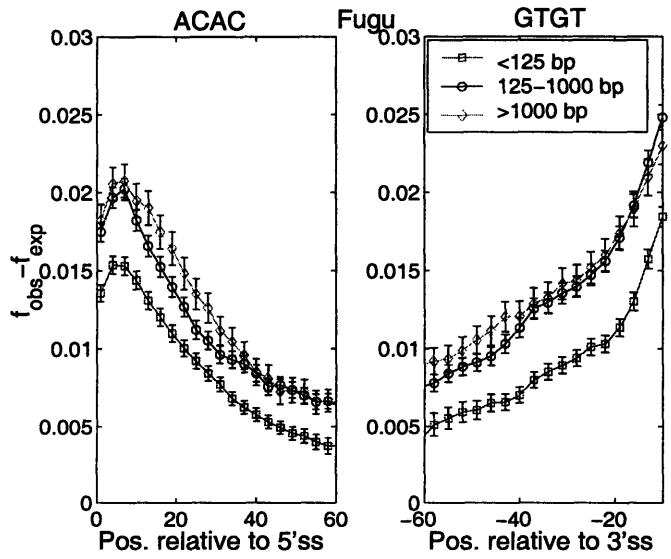

5C3D
Human
5C3D
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

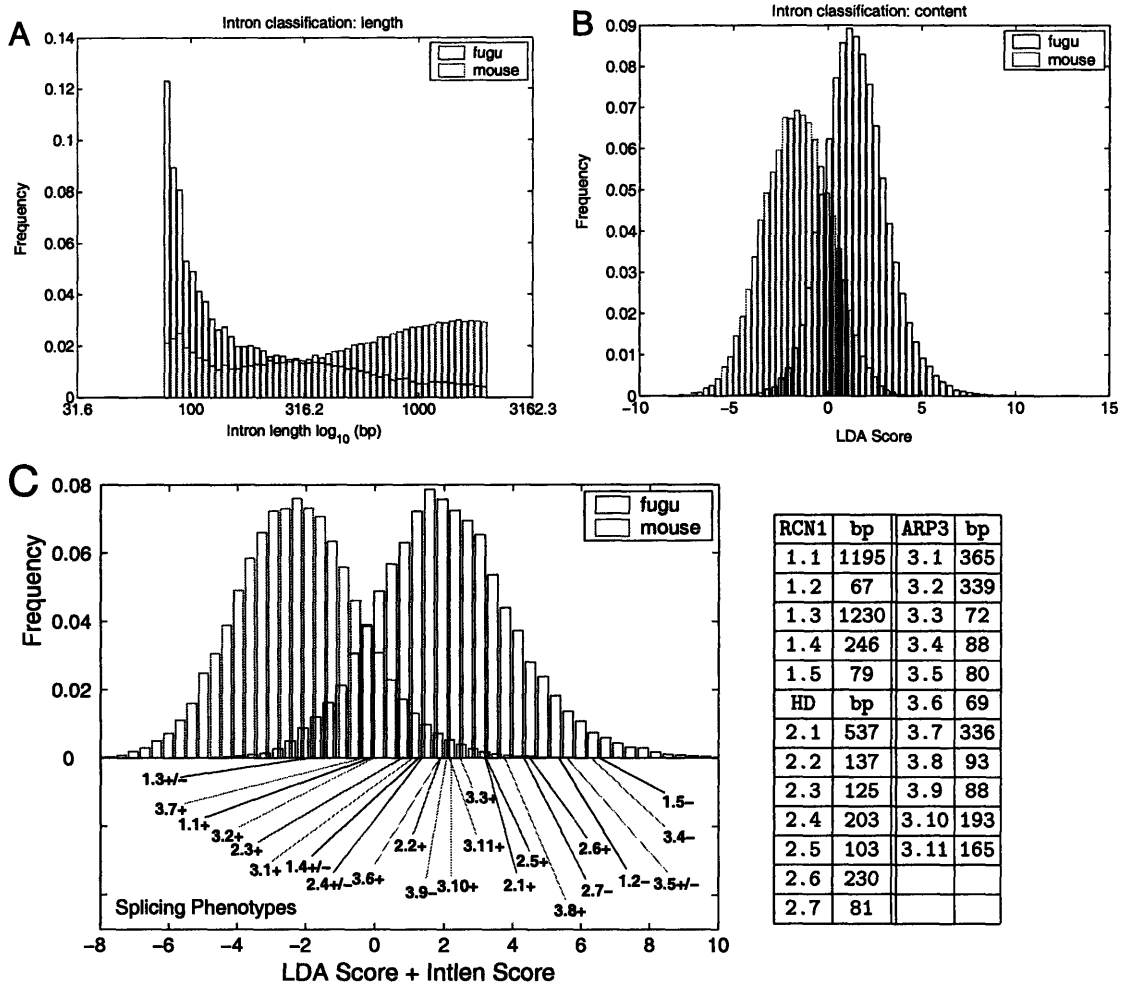

5D
Human
5D
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)


5E
Human
5E
<125 bp
125–1000
>1000
Exon position relative to 3'ss
relative to 5'ss (bp)

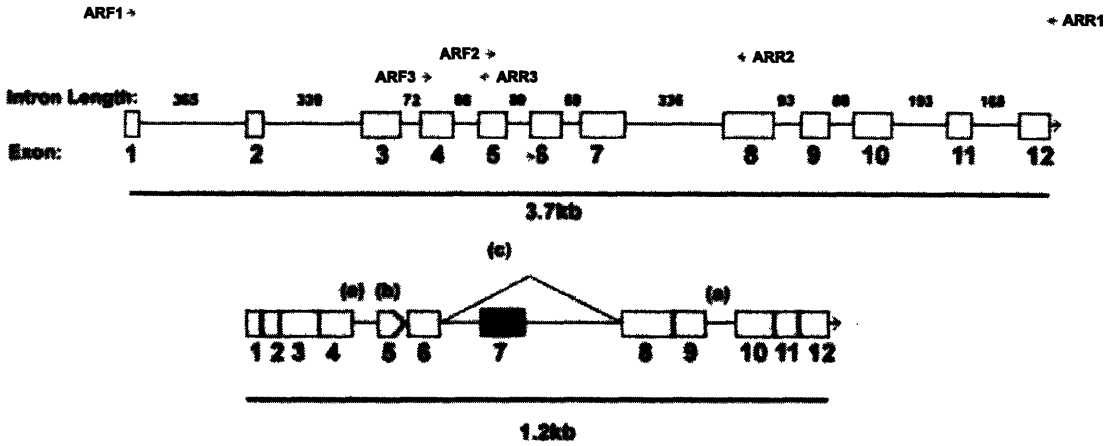

GGG
Human
GGG
<125 bp
125–1000 bp
>1000 bp
$f_{obs} - f_{exp}$
Pos. relative to 5'ss
Pos. relative to 3'ss


CCC
Human
CCC
<125 bp
125–1000 bp
>1000 bp
$f_{obs} - f_{exp}$
Pos. relative to 5'ss
Pos. relative to 3'ss


ACAC
Fugu
GTGT
<125 bp
125–1000 bp
>1000 bp
$f_{obs} - f_{exp}$
Pos. relative to 5'ss
Pos. relative to 3'ss



RCN1	bp	ARP3	bp
1.1	1195	3.1	365
1.2	67	3.2	339
1.3	1230	3.3	72
1.4	246	3.4	88
1.5	79	3.5	80
HD	bp	3.6	69
2.1	537	3.7	336
2.2	137	3.8	93
2.3	125	3.9	88
2.4	203	3.10	193
2.5	103	3.11	165
2.6	230		
2.7	81		


ARF1
ARR1
ARF2
ARR2
ARF3
ARR3
Intron Length:
Exon:
1
2
3
4
5
6
7
8
9
10
11
12
3.7kb
(c)
(a) (b)
(a)
1 2 3 4 5 6 7 8 9 10 11 12
1.2kb

Chapter 3

Alternative Splicing In Human Tissues

3.1 Abstract

Alternative pre-mRNA splicing (AS) is widely used to generate different protein isoforms in specific cell or tissue types. To compare AS events across human tissues, we analyzed the splicing patterns of genomically-aligned ESTs derived from libraries of cDNAs from different tissues. Controlling for differences in EST coverage between tissues, we found that the brain and testis had the highest levels of exon skipping. The most pronounced differences between tissues were seen for the frequencies of alternative 3' splice site and alternative 5' splice site usage, which were ~50% to 100% higher in the liver than in any other human tissue studied. Quantization of differences in splice junction usage, the brain, pancreas, liver, and the peripheral nervous system had the most distinctive patterns of AS. Analysis of available microarray expression data showed that the liver had the most divergent pattern of expression of SR protein and hnRNP protein genes compared to the other human tissues studied, possibly contributing to the unusually high frequency of alternative splice site usage seen in this tissue. Sequence motifs enriched in alternative exons expressed in the brain, testis and liver suggest specific splicing factors that may be important in AS regulation in these tissues. This study distinguishes the human brain, testis and liver as having unusually high levels of AS, highlights differences in the types of AS occurring commonly in different tissues, and identifies candidate *cis*-regulatory elements and *trans*-factors likely to play important roles in tissue-specific AS in human cells.

3.2 Background

The differentiation of a small number of cells in the developing embryo to the hundreds of cell and tissue types present in a human adult is associated with a multitude of changes in gene expression [1-3]. In addition to many differences between tissues in transcriptional and translational regulation of genes, alternative pre-mRNA splicing (AS) is also frequently used to regulate gene expression and to generate tissue-specific mRNA and protein isoforms [4-7]. Between one-third and two-thirds of human genes are estimated to undergo AS [8-12] and the disruption of specific AS events has been implicated in several human genetic diseases [13]. The diverse and important biological roles of alternative splicing have led to significant interest in understanding its regulation.

Insights into the regulation of AS have come predominantly from the molecular dissection of individual genes (reviewed in [4] and [13]). Prominent examples include the tissue-specific splicing of the *c-src* N1 exon [14], cancer-associated splicing of the *CD44* gene [15] and the alternative splicing cascade involved in *Drosophila melanogaster* sex determination [16]. Biochemical studies of these and other genes have described important classes of *trans*-acting splicing regulatory factors, implicating members of the ubiquitously expressed SR protein and heterogeneous nuclear ribonucleoprotein (hnRNP) families, and tissue-specific factors including members of the CELF [17] and NOVA [18] families of proteins, as well as other proteins and protein families, in control of specific splicing events. A number of *cis*-regulatory elements in exons or introns that play key regulatory roles have also been identified, using a variety of methods including site-directed mutagenesis, SELEX and computational approaches [19-23]. In addition, DNA microarrays and polymerase colony approaches have been developed for higher throughput analysis of alternative mRNA isoforms [24-27] and a cross-

linking/immunoprecipitation strategy (CLIP) has been developed for systematic detection of the RNAs bound by a given splicing factor [28]. These new methods suggest a path towards increasingly parallel experimental analysis of splicing regulation.

From another direction, the accumulation of large databases of cDNA and EST sequences has enabled large-scale computational studies, which have assessed the scope of AS occurring in the mammalian transcriptome [6, 9, 11, 29]. Other computational studies have analyzed the tissue specificity of AS events and identified sets of exons and genes that exhibit tissue-biased expression [30-32]. Yet a number of significant questions about tissue-specific alternative splicing have not yet been comprehensively addressed. Which tissues have the highest and lowest proportions of alternative splicing? Do tissues differ in their usage of different AS types, such as exon skipping, alternative 5' splice site (5'ss) choice or alternative 3' splice site (3'ss) choice? Which tissues are most distinct from other tissues in the spectrum of alternative mRNA isoforms they express? And to what extent do expression levels of known splicing factors explain AS patterns in different tissues?

Here, we describe an initial effort to answer these questions using a large-scale computational analysis of ESTs derived from about two dozen human tissues, which were aligned to the assembled human genome sequence to infer patterns of AS occurring in thousands of human genes. Our results distinguish specific tissues as having high levels and distinctive patterns of AS and identify pronounced differences between the proportions of alternative 5'ss and alternative 3'ss usage between tissues. Candidate *cis*-regulatory elements and *trans*-factors involved in tissue-specific AS are identified and discussed.

3.3 Results and Discussion

3.3.1 Variation in the levels of AS occurring in different human tissues

Alternative splicing events are commonly distinguished in terms of whether mRNA isoforms differ by inclusion/exclusion of an exon, in which case the involved exon is referred to as a 'skipped exon' (SE) or 'cassette exon', or whether isoforms differ in the usage of a 5'ss or 3'ss, giving rise to alternative 5'ss exons (A5Es) or alternative 3'ss exons (A3Es), respectively (depicted in Fig. 1). These descriptions are not necessarily mutually exclusive, e.g., an exon can have both an alternative 5'ss and alternative 3'ss, or have an alternative 5'ss or 3'ss and also be skipped in other isoforms. A fourth type of alternative splicing, 'intron retention', in which two isoforms differ by presence of an un-spliced intron in one transcript that is absent in the other, was not considered in this analysis because of the difficulty in distinguishing true intron retention events from contamination of the EST databases by pre-mRNA or genomic sequences. The presence of these and other artifacts in EST databases are important caveats to any analysis of EST data. Therefore, we employed stringent filters on the quality of EST to genomic alignments used in this analysis, accepting only about one-fifth of all EST alignments obtained (see Methods).

To determine whether differences occur in the proportions of these three types of AS events between human tissues, we assessed the frequencies of genes containing SEs, A3Es or A5Es for sixteen human tissues that had sufficiently high EST coverage. Since the availability of a larger number of ESTs derived from a gene increases the chance of observing alternative isoforms of that gene, the proportion of AS observed in a tissue will tend to increase with increasing EST coverage [11, 33]. Because the number of ESTs available in the dbEST database differs quite substantially between human tissues (e.g, brain ~8-fold higher than heart), in order

to compare the proportion of AS in different tissues in an unbiased way, we employed a sampling strategy that ensured that all genes/tissues studied were represented by equal numbers of ESTs.

It is important to point out that our analysis does not make use of the concept of a canonical transcript for each gene because it is not clear that such a transcript could be chosen objectively or that this concept is biologically meaningful. Instead, AS events are defined only through pairwise comparison of ESTs.

Our objective was to control for EST abundance differences between tissues while retaining sufficient power to detect a significant fraction of AS events. For each tissue we considered genes which had at least 20 aligned EST sequences derived from human cDNA libraries specific to that tissue ("tissue-derived" ESTs). For each such gene, a random sample of 20 of these ESTs was chosen (without replacement) to represent the splicing of the given gene in the given human tissue. For the gene and tissue combinations included in this analysis, the median number of EST sequences per gene was not dramatically different between tissues, ranging from 25-35 (Table S1). The sampled ESTs for each gene were then compared to each other to identify AS events occurring within the given tissue (Methods). The random sampling was repeated 20 times and the mean fraction of AS genes observed in these 20 trials was used to assess the fraction of AS genes for each tissue (Fig. 1A). Of course, different random subsets of a relatively large pool will have less overlap in the specific ESTs chosen (and therefore in the specific AS events detected) than for random subsets of a smaller pool of ESTs. And clearly increased numbers of ESTs give greater coverage of exons. However, there is no reason that the expected number of AS events detected per randomly sampled subset should depend on the size of the pool the subset was chosen from. While it is true that the error (standard deviation) of the measured AS frequency per gene should be lower when restricting to genes with a larger

minimum pool of ESTs, this restriction would not change the expected value. Unfortunately, the reduction in error of the AS frequency per gene is offset by an increase in the expected error of the tissue-level AS frequency resulting from the use of fewer genes. The inclusion of all genes with at least 20 tissue-derived ESTs represents a reasonable tradeoff between these factors.

The human brain had the highest fraction of AS genes in this analysis (Fig. 1A), with more than 40% of genes exhibiting one or more AS events, followed by the liver and testis. Previous EST-based analyses have identified high proportions of splicing in human brain and testis tissues [31, 32, 34]. These studies did not specifically control for the highly unequal representation of ESTs from different human tissues. Since larger numbers of ESTs increase the chance of observing a larger fraction of the expressed isoforms of a gene, the number of available ESTs has a direct impact on estimated proportions of AS, as seen previously in analyses comparing the levels of AS in different organisms [33]. Thus, the results obtained in this study confirm that the human brain and testis possess an unusually high level of AS, even in the absence of an EST-abundance advantage over other tissues. We also observe a high level of AS in the human liver, a tissue with much lower EST coverage where higher levels of AS have been previously reported in cancerous cells [35, 36]. The human muscle, uterus, breast, stomach and pancreas had the lowest levels of AS genes in this analysis (< 25% of genes). Lowering the minimum EST count for inclusion in this analysis from 20 to 10 ESTs, and sampling 10 (out of ≥ 10) ESTs to represent each gene in each tissue, did not alter the results qualitatively (data not shown).

3.3.2 Differences in the levels of exon skipping in different tissues

Alternatively spliced genes in this analysis exhibited on average between one and two distinct AS exons. Analyzing the different types of AS events separately, we found that the human brain

and testis had the highest levels of skipped exons (SEs), with > 20% of genes containing SEs (Fig. 1B). The high level of SEs observed in the brain is consistent with previous analyses [31, 32, 34]. At the other extreme, the human ovary, muscle, uterus and liver had the lowest levels of SEs (~10% of genes).

An example of a conserved exon skipping event observed in human and mouse brain is shown in Fig. 2A for the human fragile X mental retardation syndrome-related (*FXR1*) gene [37, 38]. In this event, skipping of the exon alters the reading frame of the downstream exon, presumably leading to production of a protein with an altered and truncated C-terminus. The exon sequence is perfectly conserved between the human and mouse genomes, as are the 5'ss and 3'ss sequences (Fig. 2A), suggesting that this AS event may play an important regulatory role [39-41].

3.3.3 Differences in the levels of alternative splice site usage in different tissues

Analyzing the proportions of AS events involving the usage of alternative 5'ss or 3'ss revealed a very different pattern (Fig. 1C,D). Notably, the fraction of genes containing A3Es was more than twice as high in the liver as in any other human tissue studied (Fig. 1D), and the level of A5Es was also about 40-50% higher in the liver than in any other tissue (Fig. 1C). The tissue with the second highest level of alternative splice site usage for both 5'ss and 3'ss was the brain. A similar group of human tissues – muscle, uterus, breast, pancreas and stomach – had the lowest level of A5Es or A3Es (< 5% of genes in each category). Thus, a picture emerges in which certain human tissues such as the muscle, uterus, breast, pancreas and stomach have low levels of AS of all types, while other tissues such as the brain and testis have relatively high levels of AS of all types, and the liver has very high levels of alternative splice site use of both the 5'ss and 3'ss, but exhibits only a low level of exon skipping. To our knowledge, this study

represents the first systematic analysis of the proportions of different types of AS events occurring in different tissues. Repeating the analyses by removing ESTs from disease-associated tissue libraries, using available library classifications [42], gave qualitatively similar results (Tables S2 and S3, and Fig. S1). These data show that ESTs derive from diseased tissues show modestly higher frequencies of exon-skipping, but the relative rankings of tissues remain similar. The fraction of genes showing alternative 5' and 3' exons do not show significant differences.

From the set of genes with at least 20 human liver-derived ESTs, this analysis identified a total of 114 genes with alternative 5'ss and/or 3'ss usage in the liver. Those genes in this set which were named, annotated and for which the consensus sequences of the alternative splice sites were conserved in the orthologous mouse gene (see Methods) are listed in Table 1. Clearly, conservation of the splice sites alone is necessary, but not sufficient by itself, to imply conservation of the AS event in the mouse genome. Many essential liver metabolic and detoxifying enzyme-encoding genes appear on this list, including enzymes involved in sugar metabolism (e.g., *ALDOB*, *IDH1*), protein and amino acid metabolism (e.g., *BHMT*, *CBP2*, *TDO2*, *PAH*, *GATM*), detoxification (e.g., *GSTA3*) or breakdown of drugs and toxins (e.g., *CYP3A4*, *CYP2C8*).

Sequences and splicing patterns for two of these genes for which orthologous mouse exons/genes and transcripts could be identified – the genes *BHMT* and *CYP2C8* - are shown in detail in Fig. 2B,C. In the event depicted for *BHMT*, the involved exons are highly conserved between the human and mouse orthologs (Fig. 2B), consistent with the possibility that the splicing event may play a (conserved) regulatory role. This AS event preserves the reading frame of downstream exons, so the two isoforms are both likely to produce functional proteins, differing by the insertion/deletion of 23 amino acids. In the event depicted for *CYP2C8*, usage of an alternative 3'ss removes 71 nucleotides, shifting the reading frame and leading to a premature

termination codon in the exon (Fig. 2C). In this AS event, the shorter alternative transcript is a potential substrate for nonsense-mediated decay [43, 44] and the AS event may be used to regulate the level of functional mRNA/protein produced.

3.3.4 **Differences in splicing factor expression between tissues**

To explore the differences in splicing factor expression in different tissues, available mRNA expression data was obtained from two different DNA microarray studies [45-47]. For this *trans*-factor analysis, we obtained a list of 20 splicing factors of the SR, SR-related and hnRNP protein families from proteomic analyses of the human spliceosome [75-77] - the specific genes studied are listed in supplementary information. The variation in splicing factor expression between pairs of tissues was studied by computing the Pearson (product-moment) correlation coefficient (r) between the 20-dimensional vectors of splicing factor expression values between all pairs of 26 human tissues, with 10 additional tissues to the 16 previously studied (Fig. 3). A low value of r between a pair of tissues indicates a low degree of concordance in the relative mRNA expression levels across this set of splicing factors, while a high value of r indicates strong concordance.

While most of the tissues examined showed a very high degree of correlation in the expression levels of the 20 splicing factors studied (typically with $r > 0.75$; Fig. 3), the human adult liver was clearly an outlier, with low concordance in splicing factor expression to most other tissues (typically $r < 0.6$ and often much lower). The unusual splicing factor expression in the human liver was seen consistently in data from two independent DNA microarray studies using different probe sets (compare two halves of Fig. 3). The low correlation observed between liver and other tissues in splicing factor expression is statistically significant even relative to arbitrary collections of 20 genes (Fig. S3). Examining the relative levels of specific splicing

factors in human adult liver versus other tissues, the relative level of SRp30c was consistently higher in liver and the relative levels of SRp40, hnRNP A2/B2 and Srp54 were consistently lower. A well-established paradigm in the field of RNA splicing is that usage of alternative splice sites is often controlled by the relative concentrations of specific SR proteins and hnRNP proteins [48-51]. The functional antagonism between particular SR and hnRNP proteins is often due to competition for binding of nearby sites on pre-mRNAs [48, 52, 53]. Therefore, it seems likely that the unusual patterns of expression seen in the human adult liver for these families of splicing factors may contribute to the high level of alternative splice site usage seen in this tissue. It is also interesting that splicing factor expression in human fetal liver is highly concordant with most other tissues, but has low concordance with adult liver (Fig. 3). This observation suggests that substantial changes in splicing factor expression may occur during human liver development, presumably leading to a host of changes in the splicing patterns of human liver-expressed genes. Currently available EST data were insufficient to allow systematic analysis of the patterns of AS in fetal liver relative to adult liver.

An important caveat to these results is that the DNA microarray data used in this analysis measure mRNA expression levels rather than protein levels or activities. The relation between the amount of mRNA expressed from a gene and the concentration of the corresponding protein has been examined previously in several studies in yeast as well as in human and mouse liver [54-57]. These studies have generally found that mRNA expression levels correlate positively with protein concentrations, but with fairly wide divergences seen for a significant fraction of genes.

3.3.5 Over-represented motifs in alternative exons in the human brain, testis and liver

The unusually high levels of alternative splicing seen in the human brain, testis and liver prompted us to identify candidate tissue-specific splicing motifs in the AS exons expressed in each of these tissues. Using a procedure similar to Brudno et al. [58], sequence motifs 4-6 bases in length that were significantly enriched in exons skipped in AS genes expressed in the human brain relative to constitutive exons expressed in the brain were identified. These sequences were then compared to each other and grouped into seven clusters, each of which shared one or two common 4-base motifs (Table 2). The motifs in cluster BR1 (CUCC, CCUC) resemble the consensus binding site for the polypyrimidine tract-binding protein (*PTB*), which acts as a repressor of splicing in many contexts [59-62]. A similar motif (CNCUCCUC) has been identified in exons expressed specifically in the human brain [31]. The motifs in cluster BR7 (containing UAGG) are similar to the high-affinity binding site UAGGG[A/U], identified for the splicing repressor protein hnRNP A1 by systematic evolution of ligands by exponential enrichment (SELEX) [63]. The consensus sequences for the remaining clusters, BR2-BR6 (GGGU, UGGG, GGGA, CUCA, UAGC, respectively), as well as BR7, all resembled motifs identified in a screen for exonic splicing silencers (ESSs) in cultured human cells (Z. Wang and C. B. B., unpublished data), suggesting that most or all of the motifs BR1-BR7 represent sequences directly involved in mediating exon skipping. For example, G-rich elements, which are known to act as intronic splicing enhancers [64, 65], may behave as splicing silencing elements in an exon sequence context.

A comparison of human testis-derived SEs to exons constitutively included in genes expressed in the testis identified only a single cluster of sequences, TE1, which shared the tetramer UAGG. Enrichment of this motif, common to the brain-specific cluster BR7, suggests a

role for regulation of exon skipping by hnRNP A1 – or a *trans*-factor with similar binding preferences - in the testis.

Alternative splice site usage gives rise to two types of exon segments – the 'core' segment common to both splice forms and the 'extended' portion that is present in only the longer isoform. Two clusters of sequence motifs enriched in the core sequences of alternative 5'ss exons expressed in liver relative to the core segments of A5Es resulting from alignments of non-liver-derived ESTs were identified, LI1 and LI2. Both are adenosine-rich, with consensus tetramers AAAC and UAAA, respectively. The former motif matches a candidate ESE motif identified previously using the computational/experimental RESCUE-ESE approach (motif 3F with consensus [AG]AA[AG]C) [20]. The enrichment of a probable ESE motif in exons exhibiting alternative splice site usage in the liver is consistent with a model that such splicing events are often controlled by the relative levels of SR proteins (which bind many ESEs) and hnRNP proteins. Insufficient data were available for the analysis of motifs in the extended portions of alternative 5'ss exons (which tend to be significantly shorter than the core regions) or for the analysis of alternative 3'ss exons.

3.3.6 A measure of dissimilarity between mRNA isoforms

To quantify the differences in splicing patterns between mRNAs or ESTs derived from a gene locus, a new measure called the splice junction difference ratio (*SJD*) was developed. For any pair of mRNAs/ESTs that align to overlapping portions of the same genomic locus, the *SJD* is defined as the proportion of splice junctions present in both transcripts that differ between them, including only those splice junctions that occur in regions of overlap between the transcripts (see Fig. 4). The *SJD* varies between zero and one, with a value of zero for any pair of transcripts that have identical splice junctions in the overlapping region (e.g., transcripts 2 and 5 in Fig. 4, or

for two identical transcripts), and has a value of 1.0 for two transcripts whose splice junctions are completely different in the regions where they overlap (e.g., transcripts 1 and 2 in Fig. 4). For instance, transcripts 2 and 3 in Fig. 4 differ in the 3'ss used in the second intron, yielding a *SJD* value of 2/4 = 0.5, while transcripts 2 and 4 differ by skipping/inclusion of an alternative exon, which affects a larger fraction of the introns in the two transcripts and therefore yields a higher *SJD* value of 3/5 = 0.6.

The splice junction difference ratio can be generalized to compare the splicing patterns between two sets of transcripts from a gene, e.g., to compare the splicing patterns of the sets of ESTs derived from two different tissues. In this case, the *SJD* is defined by counting the number of splice junctions that differ between all pairs of transcripts (*i*, *j*), with transcript *i* coming from set 1 (e.g., ESTs derived from transcripts expressed in the heart), and transcript *j* coming from set 2 (e.g., ESTs derived from transcripts in the lung), and dividing this number by the total number of splice junctions in all pairs of transcripts compared, again considering only those splice junctions that occur in regions of overlap between the transcript pairs considered. Note that this definition has the desirable property that pairs of transcripts that have larger numbers of overlapping splice junctions contribute more to the total than transcript pairs that overlap less. As an example of the splice junction difference between two sets of transcripts, consider the set S1, consisting of transcripts (1, 2) from Fig. 4, and set S2, consisting of transcripts (3, 4) from Fig. 4. Using the notation introduced in Fig. 4, $SJD(S1,S2) = d(S1,S2) / t(S1,S2) = [d(1,3)+d(1,4)+d(2,3)+d(2,4)]/[t(1,3)+t(1,4)+t(2,3)+t(2,4)] = [3+4+2+3]/[3+4+4+5] = 12/16 = 0.75$, reflecting a high level of dissimilarity between the isoforms in these sets, whereas the *SJD* falls to 0.57 for the more similar sets *S*1 = transcripts (1,2) versus *S*3 = transcripts (2,3). Note that in cases where multiple similar/identical transcripts occur in a given set, the *SJD* measure effectively weights the isoforms by their abundance, reflecting an average dissimilarity when

comparing randomly chosen pairs of transcripts from the two tissues. For example, the *SJD* computed for the set $S4 = (1,2,2,2,2)$, i.e. one transcript aligning as transcript 1 in Fig. 4 and four transcripts aligning as transcript 2, and the set $S5 = (2,2,2,2,3)$ is 23/95 = 0.24, substantially lower than the *SJD* value for sets $S1$ versus $S3$ above, reflecting the higher fraction of identically spliced transcripts between sets $S4$ and $S5$.

3.3.7 Comparison of splicing patterns between tissues

To globally compare patterns of splicing between two different human tissues, a tissue-level *SJD* value was computed, by comparing the splicing patterns of ESTs from all genes for which at least one EST was available from cDNA libraries representing both tissues. The "inter-tissue" *SJD* value is then defined as the ratio of the sum of $d(S_A,S_B)$ values for all such genes, divided by the sum of $t(S_A,S_B)$ values for all of these genes, where S_A and S_B refer to the set of ESTs for a gene from tissues A and B, respectively, and $d(S_A,S_B)$ and $t(S_A,S_B)$ are defined in terms of comparison of all pairs of ESTs from the two sets as described above. This analysis uses all available ESTs for each gene in each tissue (rather than samples of a fixed size). A large *SJD* value between a pair of tissues indicates that mRNA isoforms of genes expressed in the two tissues tend to be more dissimilar in their splicing patterns than is the case for two tissues with a smaller inter-tissue *SJD* value. This definition puts greater weight on those genes for which more ESTs are available.

Inter-tissue *SJD* values were then used to globally assess tissue-level differences in alternative splicing. A set of 25 human tissues for which at least 20,000 genomically aligned ESTs were available was compiled for this comparison (see Methods) and the *SJD* values were then computed between all pairs of tissues in this set (Fig. 5A). A clustering of human tissues on the basis of their inter-tissue *SJD* values (Fig. 5B) identified groups of tissues that cluster

together very closely (e.g. the ovary/thyroid/breast cluster, the heart/lymph cluster and the bone/b-cell cluster), while other tissues including the brain, pancreas, liver, peripheral nervous system (PNS) and placenta occur as out-groups. Calculating the mean *SJD* value for a given tissue when compared to the remaining 24 tissues (Fig. 5C) identified a set of human tissues including the ovary, thyroid, breast, heart, bone, b-cell, uterus, lymph and colon that have 'generic' splicing patterns which tend to be more similar to most other tissues. As expected, many of these tissues with generic splicing patterns overlap with the set of tissues that have low levels of AS (Fig. 1). On the other hand, another group of tissues including the human brain, pancreas, liver and PNS, have highly 'distinctive' splicing patterns that differ from most other tissues (Fig. 5C). Many of these tissues were identified as having high proportions of AS in Fig. 1. Taken together, these observations suggest that specific human tissues such as the brain, testis and liver, make more extensive use of AS in gene regulation and that these tissues have also diverged most from other tissues in the set of spliced isoforms they express. Although we are not aware of reliable, quantitative data on the relative abundance of different cell-types in these different tissues, a greater diversity of cell-types in specific tissues is likely to contribute to higher SJD values in these tissues.

3.4 Conclusions and Prospects

The systematic analysis of transcripts generated from the human genome is just beginning, but promises to deepen our understanding of how changes in the program of gene expression contribute to development and differentiation. Here, we have observed pronounced differences between human tissues in the set of alternative mRNA isoforms that they express, with the human brain, testis and liver exhibiting the highest levels of AS. Because our approach

normalizes the EST coverage per gene in each tissue, there is higher confidence that these differences accurately reflect differences in splicing patterns between tissues. Since human tissues are generally made up of a mixture of cell types, each of which may have its own unique pattern of gene expression and splicing, it will be important in the future to develop methods for systematic analysis of transcripts in different human cell types.

In our analysis of the levels of different AS types, the liver stood out as having substantially higher proportions of alternative splice site usage than other human tissues, while the brain and testis were found to have high proportions of exon skipping. The splicing patterns of liver- and brain-expressed genes were more likely to be distinct from the patterns seen in other tissues (Fig. 5). These differences are likely to result from tissue differences in splicing factor expression (or activity). Consistent with this hypothesis, the human liver was found to have a discrepant expression profile for SR protein and hnRNP protein encoding genes when compared across tissues (Fig. 3). The characteristic pattern of gene expression of these factors is likely to contribute to the higher proportion of alternative splice site usage seen in this tissue. On the other hand, our analysis of sequence motifs in alternative exons implicates the frequent occurrence of a number of putative exonic splicing silencer motifs in AS genes expressed in the human brain and testis, likely contributing to the high proportion of exon skipping observed in these tissues. The frequent use of exon skipping in gene regulation in the human brain and testis suggests the presence of specific developmental or regulatory responses involving changes in splicing factor expression or activity. The pronounced tissue-level differences in alternative splicing imply the importance of this regulatory mechanism in the biology of human tissues.

3.5 Methods

Data and resources

Chromosome assemblies of the human genome (hg13) were obtained from public databases [66]. Transcript databases included approximately 94,000 human cDNA sequences obtained from GenBank (release 134.0, gbpri and gbhtc categories), and approximately 5 million human expressed sequence tags (ESTs) from dbEST (repository 02202003). Human ESTs were designated according to their cDNA library source (in total about 800) into different tissue types. Pertinent information about cDNA libraries and the corresponding human tissue or cell line was extracted from dbEST and subsequently integrated with library information retrieved from the mammalian gene collection initiative (MGC) [67], the integrated molecular analysis of gene expression consortium (IMAGE) [68] and the cancer genome anatomy project (CGAP) [69]. Library information obtained from MGC, IMAGE and CGAP is provided as supplementary material in Table S4 at http://genes.mit.edu/burgelab/Supplementary/yeo_holste04/ as additional data file 4.

Genome annotation by spliced transcript alignments

The GENOA genome annotation script [70] was used to align spliced cDNA and EST sequences to the human genome. GENOA uses BLASTN to detect significant blocks of identity between repeat-masked cDNA sequences and genomic DNA, and then aligns cDNAs to the genomic loci identified by BLASTN using the spliced alignment algorithm MRNAVSGEN [70]. This algorithm is similar in concept to SIM4 [71] but was developed specifically to align high quality cDNAs rather than ESTs and thus requires higher alignment quality (at least ~93% identity) and consensus terminal dinucleotides at the ends of all introns (i.e. GT..AG or GC..AG). EST sequences were

aligned using SIM4 [71] to those genomic regions which had aligned cDNAs. Stringent alignment criteria were imposed: (1) ESTs were required to overlap cDNAs (so all of the genes studied were supported by at least one cDNA:genomic alignment); (2) the first and last aligned segments of ESTs were required to be at least 30 nucleotides in length, with 90% sequence identity; and (3) the entire EST sequence alignment was required to extend over at least 90% of the length of the EST with at least 90% sequence identity.

In total, GENOA aligned about 85,900 human cDNAs and about 890,300 ESTs to the human genome. The relatively low fraction of ESTs aligned (~18%) reflects the stringent alignment quality criteria that were imposed in order to be as confident as possible in the inferred splicing patterns. The aligned sequences resulted in about 17,800 gene regions with more than 1 transcript aligned that exhibited multi-exon structure. Of these, ~60% exhibited evidence of alternative splicing of internal exons. Our analysis did not examine differences in 3'-terminal and 5'-terminal exons, inclusion of which is frequently dictated by alternative polyadenylation or transcription start sites and therefore does not represent 'pure' AS [72, 73]. The EST alignments were then used to categorize all internal exons as: constitutive exons, alternative 3'ss exons (A3E), alternative 5'ss exons (A5E), skipped exons (SE), multiply alternatively spliced exons (e.g., exons that exhibited both skipping and alternative 5'ss usage), and exons that contained retained introns. An internal exon present in at least one transcript was identified as an SE if it was precisely skipped in one or more other transcripts, such that the boundaries of both the 5' and 3' flanking exons were the same in the transcripts that included and skipped the exon (e.g., exon E3 in Fig. 1). Similarly, an internal exon present in at least one transcript was identified as an A3E (A5E) if at least one other transcript contained an exon differing in length by the use of an alternative 3'ss (5'ss). The 'core' of an A3E (A5E) is defined as the shortest exon segment common to transcripts used to infer the A3E (A5E) event. The extension of an A3E (A5E) is the

exonic segment added to the core by the alternative 3'ss (5'ss). Pairs of inferred A3Es or A5Es differing by fewer than 6 nucleotides were excluded from further analysis, as in [9], because of the possibility that such small differences might sometimes result from EST sequencing or alignment errors. As the frequency of insertion-deletions errors greater than 3 bases in ESTs is vanishingly small (E. Birney, personal communication), a 6-base cutoff should exclude the vast majority of such errors. Alternatively spliced exons/genes identified in specific tissues are available for download from the GENOA web site [70].

Quantifying splice junction differences between alternative mRNA isoforms

To quantify the difference in splicing patterns between mRNAs or ESTs derived from a gene locus, the splice junction difference ratio (*SJD*) was calculated. For any pair of mRNAs/ESTs that have been aligned to overlapping portions a genomic locus, the *SJD* is defined as the fraction of the splice junctions that occur in overlapping portions of the two transcripts that differ in one or both splice sites. A sample calculation is given in Fig. 4. The *SJD* measure was calculated by taking the ratio of the number of "valid" splice junctions that differ between two sequences over the total number of splice junctions, when comparing a pair of ESTs across all splice junctions present in overlapping portions of the two transcripts. A splice junction was considered valid if: (1) the 5'ss and the 3'ss satisfied either the GT..AG or the GC..AG dinucleotide sequence at exon-intron junctions, and (2) if the splice junction was observed at least twice in different transcripts. The *SJD* measure can be generalized from a single overlapping EST pair to two groups of overlapping ESTs. When comparing two groups of EST sequences, the *SJD* was computed by first calculating the total number of different valid splice junctions for every compared pair of ESTs, and then dividing by the total number of splice junctions for every pair of ESTs.

Identification of candidate splicing regulatory motifs

Over-represented sequence motifs (*k*-mers) were identified by comparing the number of occurrences of *k*-mers (for *k* in the range of 4 to 6 nucleotides) in a test set of alternative exons versus a control set. In this analysis, monomeric tandem repeats (e.g., poly-A sequences) were excluded. The enrichment score of candidate *k*-mers in the test set versus the control set was evaluated by computing χ^2 (chi-squared) values with a Yates correction term [74], using an approach similar in spirit to that described by Brudno et al. [58]. We randomly sampled 500 subsets of the same size as the test set from the control set. The enrichment scores for *k*-mers over-represented in the sampled subset versus the remainder of the control set were computed as above. The estimated *P*-value for observing the given enrichment score (χ^2 value) associated with an over-represented sequence motif of length *k* was defined as the fraction of subsets that contained any *k*-mer with enrichment score (χ^2-value) higher than the tested motif. Correcting for multiple tests is not required since the P-value is defined relative to the most enriched k-mer for each sampled set. For the set of skipped exons from human brain and testis-derived EST sequences, the test sets comprised 1,265 and 517 exons skipped in brain and testis, respectively, and the control sets comprised 12,527 and 8,634 exons constitutively included in respectively human brain and testis-derived ESTs. Candidate sequence motifs of skipped exons from brain and testis-derived ESTs with associated *P*-values less than 0.002 were retained. For the set of A5E and A3E events from human liver-derived EST sequences, the test set comprised 44 A3Es and 45 A5Es, and the control set comprised 1,619 A3Es and 1,481 A5Es identified using ESTs from all tissues other than liver. In this analysis, A3Es and A5Es with extension sequences of less than 25 bases were excluded from and sequence longer than 150 bases were truncated to 150 bases, by retaining the exon sequence segment closest to the internal alternative splice junction.

Over-represented sequence motifs of A3Es and A5Es from liver-derived EST sequences with associated *P*-values less than 0.01 were retained.

Gene expression analysis of *trans*-acting splicing factors

Serine-arginine (SR), SR-related proteins, and heterogeneous nuclear ribonucleoproteins (hnRNPs) were derived from published proteomic analyses of the spliceosome [75-77]. Expression values for these genes were obtained from the "gene expression atlas" using the HG-U95A DNA microarray [45] and from a similar set of expression data using the HG-U133A DNA microarray [47]. Altogether twenty splicing factors, ASF/SF2, SRm300, SC35, SRp40, SRp55, SRp30c, 9G8, SRp54, SFRS10, SRp20, hnRNPs A1, A2/B2, C, D, G, H1, K, L, M, and RALY, were studied in 26 different tissues present in both microarray experiments (see Figure 5). The data from each gene chip - HG-U95A and HG-U133A - were analyzed separately. The average difference (AD) value of each probe was used as the indicator of expression level. In analyzing these microarray data, AD values smaller than 20 were standardized to 20, as performed in [45]. When two or more probes mapped to a single gene, the values from those probes were averaged. The Pearson correlation coefficient between the 20-dimensional vectors for all tissue pairs were calculated.

List of abbreviations

AS, alternative splicing or alternatively spliced; 5'ss, 5' splice site; 3'ss, 3' splice site; cDNA, complementary DNA; EST, expressed sequence tag; *SJD*, splice junction difference; SE, skipped exon; A5E, alternative 5'ss exon; A3E, alternative 3'ss exon; SR, serine-arginine; hnRNP, heterogeneous nuclear ribonucleoprotein.

Acknowledgements

We thank T. Poggio and P. Sharp for stimulating discussions and the anonymous reviewers for constructive suggestions. This work was supported by grants from the National Science Foundation and the National Institutes of Health to C. B. B. G.Y. was supported by the Lee Kuan Yew Fellowship from Singapore.

References

1. Lopez AJ: **Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation.** *Annu Rev Genet* 1998, **32**:279-305.
2. Grabowski PJ: **Genetic evidence for a Nova regulator of alternative splicing in the brain.** *Neuron* 2000, **25**(2):254-256.
3. Lodish H, Baltimore, D, Berk A, Zipursky, S.L., Matsudaira, P, Darnell, J: **Molecular Cell Biology**, Third edn. New York: Scientific American Books, Inc.; 1995.
4. Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem* 2003, **72**:291-336.
5. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 2002, **3**(4):285-298.
6. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world.** *Trends Genet* 2001, **17**(2):100-107.
7. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome.** *Genome Res* 2003, **13**(6B):1290-1300.
8. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9**(12):1288-1293.
9. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29**(13):2850-2859.
10. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30**(1):13-19.
11. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs.** *Genome Res* 2002, **12**(12):1837-1845.
12. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474**(1):83-86.
13. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease.** *Genes Dev* 2003, **17**(4):419-437.
14. Modafferi EF, Black DL: **A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon.** *Mol Cell Biol* 1997, **17**(11):6537-6545.
15. Naor D, Nedvetzki S, Golan I, Melnik L, Faitelson Y: **CD44 in cancer.** *Crit Rev Clin Lab Sci* 2002, **39**(6):527-579.
16. MacDougall C, Harbison D, Bownes M: **The developmental consequences of alternate splicing in sex determination and differentiation in Drosophila.** *Dev Biol* 1995, **172**(2):353-376.
17. Ladd AN, Charlet N, Cooper TA: **The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing.** *Mol Cell Biol* 2001, **21**(4):1285-1296.
18. Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY, Darnell RB: **Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability.** *Neuron* 2000, **25**(2):359-371.
19. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci U S A* 2001, **98**(20):11193-11198.

20. Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297**(5583):1007-1013.
21. Liu HX, Zhang M, Krainer AR: **Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins.** *Genes Dev* 1998, **12**(13):1998-2012.
22. Schaal TD, Maniatis T: **Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences.** *Mol Cell Biol* 1999, **19**(3):1705-1719.
23. Tian H, Kole R: **Strong RNA splicing enhancers identified by a modified method of cycled selection interact with SR protein.** *J Biol Chem* 2001, **276**(36):33833-33839.
24. Zhu J, Shendure J, Mitra RD, Church GM: **Single molecule profiling of alternative pre-mRNA splicing.** *Science* 2003, **301**(5634):836-838.
25. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**(5653):2141-2144.
26. Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y: **Predicting splice variant from DNA chip expression data.** *Genome Res* 2001, **11**(7):1237-1245.
27. Clark TA, Sugnet CW, Ares M, Jr.: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science* 2002, **296**(5569):907-910.
28. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: **CLIP identifies Nova-regulated RNA networks in the brain.** *Science* 2003, **302**(5648):1212-1215.
29. Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.** *Hum Mol Genet* 2002, **11**(4):451-464.
30. Gustincich S, Batalov S, Beisel KW, Bono H, Carninci P, Fletcher CF, Grimmond S, Hirokawa N, Jarvis ED, Jegla T *et al*: **Analysis of the mouse transcriptome for genes involved in the function of the nervous system.** *Genome Res* 2003, **13**(6B):1395-1401.
31. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ: **An alternative-exon database and its statistical analysis.** *DNA Cell Biol* 2000, **19**(12):739-756.
32. Xu Q, Modrek B, Lee C: **Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.** *Nucleic Acids Res* 2002, **30**(17):3754-3766.
33. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity.** *Nat Genet* 2002, **30**(1):29-30.
34. Lee CJ, Irizarry K: **Alternative splicing in the nervous system: an emerging source of diversity and regulation.** *Biol Psychiatry* 2003, **54**(8):771-776.
35. Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP: **Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer.** *Cancer Res* 2003, **63**(3):655-657.
36. Hui L, Zhang X, Wu X, Lin Z, Wang Q, Li Y, Hu G: **Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment.** *Oncogene* 2004.
37. Kirkpatrick LL, McIlwain KA, Nelson DL: **Alternative splicing in the murine and human FXR1 genes.** *Genomics* 1999, **59**(2):193-202.
38. Kirkpatrick LL, McIlwain KA, Nelson DL: **Comparative genomic sequence analysis of the FXR gene family: FMR1, FXR1, and FXR2.** *Genomics* 2001, **78**(3):169-177.
39. Sugnet CW, Kent, W.J., Ares JR, M., Haussler, D.: **Transcriptome and genome conservation of alternative splicing events in humans and mice.** In: *Pacific symposium on biocomputing: 2004; Hawaii*: World Scientific; 2004.

40. Sorek R, Ast G: **Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.** *Genome Res* 2003, **13**(7):1631-1637.
41. Kaufmann D, Kenner O, Nurnberg P, Vogel W, Bartelt B: **In NF1, CFTR, PER3, CARS and SYT7, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons.** *Eur J Hum Genet* 2004, **12**(2):139-149.
42. Megy K, Audic S, Claverie JM: **Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22.** *Genome Biol* 2003, **4**(2):P1.
43. Hillman RT, Green RE, Brenner SE: **An unappreciated role for RNA surveillance.** *Genome Biol* 2004, **5**(2):R8.
44. Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE: **Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes.** *Bioinformatics* 2003, **19 Suppl 1**:I118-I121.
45. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A *et al*: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99**(7):4465-4470.
46. **Gene Expression Atlas** [http://expression.gnf.org]
47. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101**(16):6062-6067.
48. Hanamura A, Caceres JF, Mayeda A, Franza BR, Jr., Krainer AR: **Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors.** *Rna* 1998, **4**(4):430-444.
49. Bai Y, Lee D, Yu T, Chasin LA: **Control of 3' splice site choice in vivo by ASF/SF2 and hnRNP A1.** *Nucleic Acids Res* 1999, **27**(4):1126-1134.
50. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hayward DG, Krainer AR: **Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1.** *Mol Cell Biol* 2000, **20**(22):8303-8318.
51. Kamma H, Portman DS, Dreyfuss G: **Cell type-specific expression of hnRNP proteins.** *Exp Cell Res* 1995, **221**(1):187-196.
52. Caputi M, Mayeda A, Krainer AR, Zahler AM: **hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing.** *Embo J* 1999, **18**(14):4060-4067.
53. Caputi M, Zahler AM: **SR proteins and hnRNP H regulate the splicing of the HIV-1 tev-specific exon 6D.** *Embo J* 2002, **21**(4):845-855.
54. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18**(3-4):533-537.
55. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**(11):7357-7368.
56. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R: **Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae.** *Mol Cell Proteomics* 2002, **1**(4):323-333.
57. Kawamoto S, Matsumoto Y, Mizuno K, Okubo K, Matsubara K: **Expression profiles of active genes in human and mouse livers.** *Gene* 1996, **174**(1):151-158.
58. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, Conboy JG: **Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing.** *Nucleic Acids Res* 2001, **29**(11):2338-2348.

59. Chou MY, Underwood JG, Nikolic J, Luu MH, Black DL: **Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing.** *Mol Cell* 2000, **5**(6):949-957.
60. Chan RC, Black DL: **The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream.** *Mol Cell Biol* 1997, **17**(8):4667-4676.
61. Grabowski PJ: **Splicing regulation in neurons: tinkering with cell-specific control.** *Cell* 1998, **92**(6):709-712.
62. Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CW: **Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay.** *Mol Cell* 2004, **13**(1):91-100.
63. Burd CG, Dreyfuss G: **RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing.** *Embo J* 1994, **13**(5):1197-1204.
64. Sirand-Pugnet P, Durosay P, Brody E, Marie J: **An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA.** *Nucleic Acids Res* 1995, **23**(17):3501-3507.
65. McCullough AJ, Berget SM: **G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.** *Mol Cell Biol* 1997, **17**(8):4562-4571.
66. **UCSC Genome Browser** [http://genome.ucsc.edu]
67. **Mammalian Gene Collection (MGC) Initiative** [http://mgc.nci.nih.gov]
68. **Integrated Molecular Analysis of Gene Expression (IMAGE)** [http://image.llnl.gov/image]
69. **Cancer Genome Anatomy Project (CGAP)** [http://cgap.nci.nih.gov]
70. **D. Holste, R.-F. Yeh, U. Ohler, G. Yeo, L.P. Lim, and C.B. Burge, Genome Annotation (GENOA) Program** [http://genes.mit.edu/genoa]
71. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8**(9):967-974.
72. Berget SM: **Exon recognition in vertebrate splicing.** *J Biol Chem* 1995, **270**(6):2411-2414.
73. Majewski J, Ott J: **Distribution and characterization of regulatory elements in the human genome.** *Genome Res* 2002, **12**(12):1827-1836.
74. Glantz SA: **Primer of Biostatistics,** Fourth edn. New York: McGraw-Hill; 1997.
75. Jurica MS, Moore MJ: **Pre-mRNA splicing: awash in a sea of proteins.** *Mol Cell* 2003, **12**(1):5-14.
76. Rappsilber J, Ryder U, Lamond AI, Mann M: **Large-scale proteomic analysis of the human spliceosome.** *Genome Res* 2002, **12**(8):1231-1245.
77. Zhou Z, Licklider LJ, Gygi SP, Reed R: **Comprehensive proteomic analysis of the human spliceosome.** *Nature* 2002, **419**(6903):182-185.

Figure legends

Figure 1. Levels of alternative splicing in sixteen human tissues with moderate or high EST sequence coverage. Horizontal bars show the average fraction (and estimated standard deviation) of alternatively spliced (AS) genes of each splice type for random samplings of N=20 ESTs per gene from each gene with ≥ 20 aligned EST sequences derived from a given human tissue. (A) Fraction of AS genes containing skipped exons, alternative 3'ss exons or alternative 5'ss exons; (B) fraction of AS genes containing skipped exons; (C) fraction of AS genes containing alternative 3'ss exons; and (D) fraction of AS genes containing alternative 5'ss exons.

Figure 2. Human tissue specific alternatively spliced genes. (A) Human fragile X mental retardation syndrome-related (*FXR1*) gene splicing detected in brain-derived EST sequences. *FXR1* exhibited two alternative mRNA isoforms differing by skipping/inclusion of exon E16. Exclusion of E16 creates a shift in the reading-frame predicted to result in an altered and shorter C-terminus. The exon-skipping event is conserved in the mouse ortholog of the human *FXR1* gene, and both isoforms were detected in ESTs derived from the mouse brain. (B) Human betaine-homocysteine S-methyltransferase (*BHMT*) gene splicing detected in liver-derived ESTs. *BHMT* exhibited two alternative isoforms differing by an alternative 5'ss exon usage in exon E4. Sequence comparisons indicated that the exon and splice site sequences involved in both alternative 5'ss exon events are conserved in the mouse ortholog of the human *BHMT* gene. (C) Human cytochrome P450 2C8 *(CYP2C8)* gene splicing. *CYP2C8* exhibited two alternative mRNA isoforms due to an alternative 3'ss in exon E4 (detected in ESTs derived from several tissues), where the exclusion of a 71 bases sequence created a premature termination codon in

exon E4b. Exons and splice sites involved in the AS event are conserved in the mouse ortholog of *CYP2C8*.

Figure 3. Correlation of mRNA expression levels of 20 known splicing factors across 26 human tissues (lower diagonal: Affymetrix HU-133A DNA microarray experiment [47]; upper diagonal: Affymetrix HU-95A DNA microarray experiment [45]); splicing factors listed in supplementary Table S5. Colored squares represent correlation coefficients of the mRNA expression patterns of 20 in each pair of tissues (see scale at top of figure).

Figure 4. Computation of the splice junction difference ratio (*SJD*). The *SJD* value for a pair of transcripts is computed as the number of splice junctions in each transcript that are not represented in the other transcript, divided by the total number of splice junctions in the two transcripts, in both cases considering only those splice junctions that occur in portions of the two transcripts that overlap. *SJD* value calculations for combinations of the transcripts listed above are also shown.

Figure 5. Comparison of alternative mRNA isoforms across twenty five human tissues. (A) Color-encoded representation of *SJD* values between pairs of tissues. (B) Hierarchical clustering of *SJD* values using average-linkage clustering. Groups of tissues in clusters with short branch lengths (e.g. thyroid/ovary, b-cell/bone) have highly similar patterns of AS. (C) Mean *SJD* values (versus other 24 tissues) for each tissue.

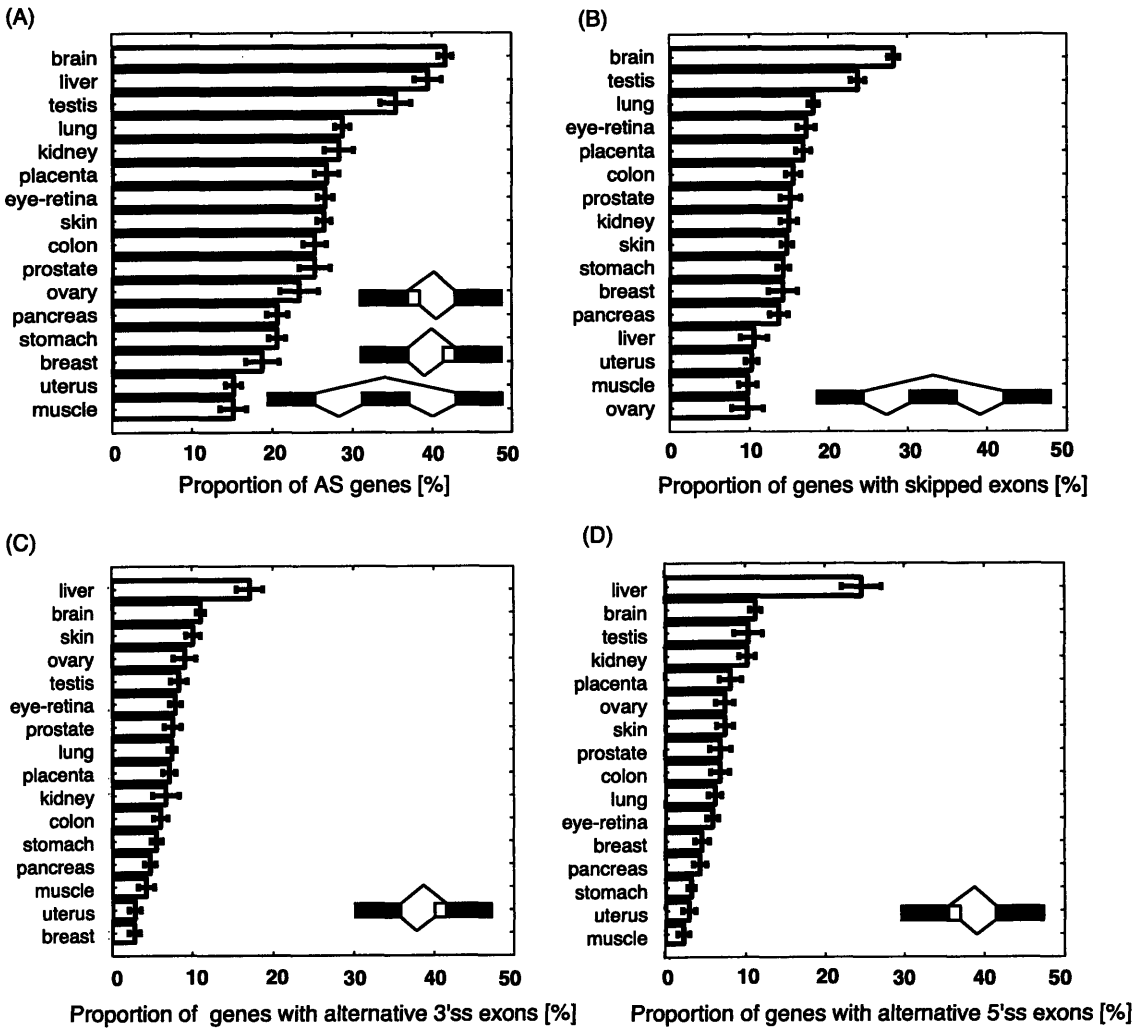
Type	Ensembl gene ID	Gene name	Exon numbers	Fold-change above median expression, HG-U95A	Fold-change above median expression, MG-U74A
A5E;A3E	091513	Serotransferrin Precursor, *TF*	8,9; 4	100	100
A5E;A3E	115414	Fibronectin Precursor, *FN1*	36; 31	10	-
A5E;A3E	117601	Antithrombin-III Precursor, *SERPINC1*	5; 4	100	100
A5E;A3E	136872	Fructose-Bisphosphate Aldolase, *ALDOB*	3,8; 4	100	10
A5E;A3E	140833	Haptoglobin-Related Protein Precursor, *HPR*	3	100	10
A5E;A3E	151790	Tryptophan 2,3-Dioxygenase, TDO2	3,5; 4	10	100
A5E;A3E	171759	Phenylalanine-4-Hydroxylase, *PAH*	6; 4,10	-	100
A5E	047457	Ceruloplasmin Precursor, *CP*	14,16	3	-
A5E	055957	Inter-Alpha-Trypsin Inhibitor Heavy Chain H1 Precursor, *ITIH1*	21	100	10
A5E	111275	Aldehyde Dehydrogenase, *ALDH2*	12	3	3
A5E	132386	Pigment Epithelium-derived Factor Precursor, *SERPINF1*	4	10	10
A5E	138356	Aldehyde Oxidase, *AOX1*	27,29	3	3
A5E	138413	Isocitrate Dehydrogenase, *IDH1*	3	1	-
A5E	145692	Betaine-Homocysteine S-Methyltransferase, *BHMT*	4	10	100
A5E	160868	Cytochrome P450, *CYP3A4*	5	10	10
A5E	171766	Glycine Amidinotransferase, *GATM*	8	3	3
A3E	080618	Carboxypeptidase, *CBP2*	10	-	-
A3E	080824	Heat Shock Protein HSP 90-Alpha, *HSPCA*	8	-	-
A3E	096087	Glutathione S-Transferase, *GSTA2*	4,6	10	10
A3E	106927	Protein Precursor, *AMBP*	5,9	100	100
A3E	110958	Telomerase-Binding Protein P23, *TEBP*	5	<1	1
A3E	134240	Hydroxymethylglutaryl-COA Synthase, *HMGCS2*	8	10	-
A3E	138115	Cytochrome P450, *CYP2C8*	4	100	10
A3E	145192	Alpha-2-HS-Glycoprotein Precursor, *AHSG*	6	100	100
A3E	163631	Serum Albumin Precursor, *ALB*	9	100	100
A3E	171557	Fibrinogen Gamma Chain Precursor, *FGG*	4	100	100
A3E	174156	Glutathione S-Transferase, *GSTA3*	4,6	10	10

Table 1. Human genes expressed in the liver with alternative 3'ss (A3E) or alternative 5'ss exons (A5E). Examples of human AS genes found to exhibit A3E and/or A5E splicing with both isoforms detected in liver-derived ESTs. AS types are listed in the first column, followed by the last six digits of the Ensembl gene number (ENSG00000#), the gene name and alternative exon numbers. The last two columns list expression levels in human liver and mouse liver, respectively, expressed in terms of the fold-change relative to the median expression level in other tissues (from the DNA microarray data of [45]).

AS type /Tissue (motif name)	Oligo-nucleotides	Occurrences	Consensus (% of exon containing)
SE/Brain (BR1)	**CUCC**UG	169	**CUCC** (45.3)
	CUCCU	323	
	CUCCC	264	
	CUCC	945	
	C**CUCC**C	137	
	CCUCC	363	
	CCUC	1021	**CCUC** (41.0)
	GCCUCC	136	
	GCCUC	375	
	GCCUCA	122	
	GGCCUC	118	
	UGCCUC	108	
SE/Brain (BR2)	**GGGUU**	97	**GGGU** (25.6)
	GGGU	411	
	AGGGU	116	
SE/Brain (BR3)	**UGGGA**	324	**UGGG** (47.2)
	UGGG	948	
	CUGGG	426	
	CCUGGG	171	
SE/Brain (BR4)	**GGGAUU**	58	**GGGA** (45.5)
	GGGAU	176	
	GGGA	840	
SE/Brain (BR5)	**CUCA**	925	**CUCA** (46.5)
	CUCAC	206	
	GCCUCA	122	
	GGCUCA	102	
	GCUCAC	79	
	CUCAGC	126	
SE/Brain (BR6)	**UAGC**	269	**UAGC** (18.0)
	UAGCU	106	
	GUAGC	96	
	GUAGCU	51	
	AGUAGC	47	
	UAGCUG	54	
SE/Brain (BR7)	**UAGG**	186	**UAGG** (13.8)
	UUAGG	63	
	UUAGGG	24	
SE/Testis (TE1)	**UAGG**	99	**UAGG** (16.6)
	UUAGG	33	
Core A5E/ Liver (LI1)	**AAAC**	42	**AAAC** (53.3)
	AAAAC	18	
Core A5E/ Liver (LI2)	**UAAA**	29	**UAAA** (40.0)
	UAAACC	5	

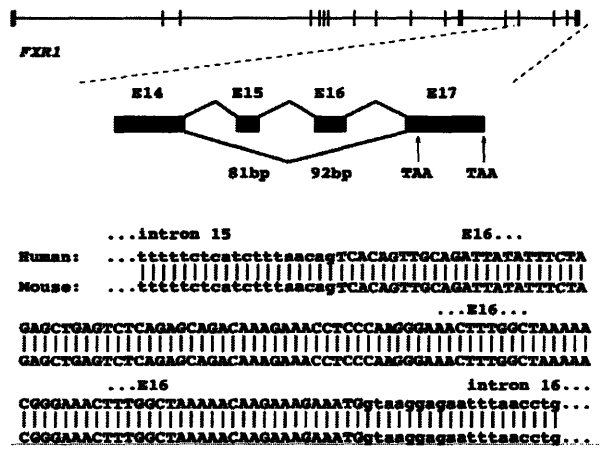
Table 2. Sequence motifs enriched in skipped exons and alternative 5'ss exons. Sequence motifs of length 4-6 bases significantly over-represented ($P < 0.002$) in SEs relative to constitutively spliced exons from brain or testis-derived ESTs are shown in the top and middle part, followed by the number of SE occurrences in these tissues. Sequence motifs are grouped/aligned by similarity, and shared tetramers are shown in bold and listed in the last column, followed by the fraction of SEs that contain the given tetramer. Sequence motifs significantly over-represented ($P < 0.01$) in core alternative 5'ss exons derived from human liver-derived ESTs are shown at the bottom part, followed by number of A5E occurrences and the fractions that contain the given tetramer. Statistical significance was evaluated as described in Methods.

Figure 1

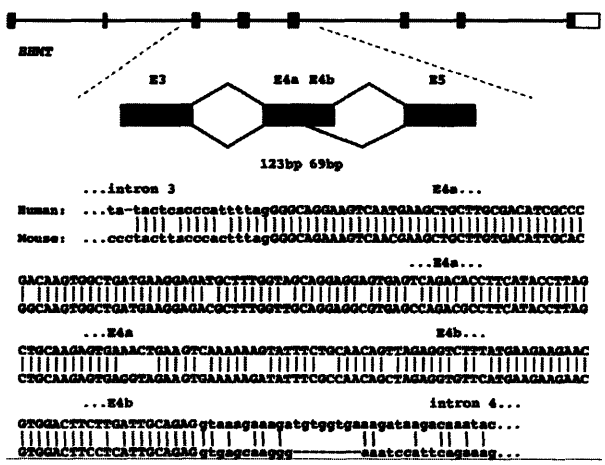

(A)
brain
liver
testis
lung
kidney
placenta
eye-retina
skin
colon
prostate
ovary
pancreas
stomach
breast
uterus
muscle
0 10 20 30 40 50
Proportion of AS genes [%]
(B)
brain
testis
lung
eye-retina
placenta
colon
prostate
kidney
skin
stomach
breast
pancreas
liver
uterus
muscle
ovary
0 10 20 30 40 50
Proportion of genes with skipped exons [%]
(C)
liver
brain
skin
ovary
testis
eye-retina
prostate
lung
placenta
kidney
colon
stomach
pancreas
muscle
uterus
breast
0 10 20 30 40 50
Proportion of genes with alternative 3'ss exons [%]
(D)
liver
brain
testis
kidney
placenta
ovary
skin
prostate
colon
lung
eye-retina
breast
pancreas
stomach
uterus
muscle
0 10 20 30 40 50
Proportion of genes with alternative 5'ss exons [%]

! "#$%&' ( )

*


FXR1
E14
E15
E16
E17
81bp
92bp
TAA
TAA
...intron 15
E16...
Human:
Mouse:
...E16...
...E16
intron 16...

+


SHNT
E3
E4a E4b
E5
123bp 69bp
...intron 3
E4a...
Human:
Mouse:
...E4a...
...E4a
E4b...
...E4b
intron 4...

'

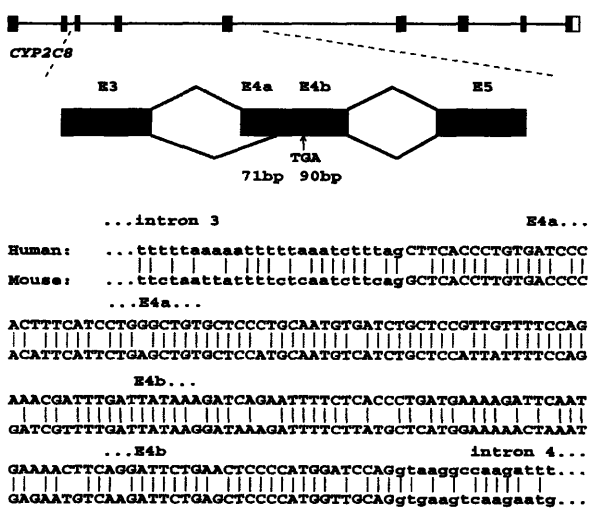

CYP2C8
E3
E4a
E4b
E5
TGA
71bp
90bp
...intron 3
E4a...
Human:
Mouse:
...E4a...
E4b...
...E4b
intron 4...

Figure 3

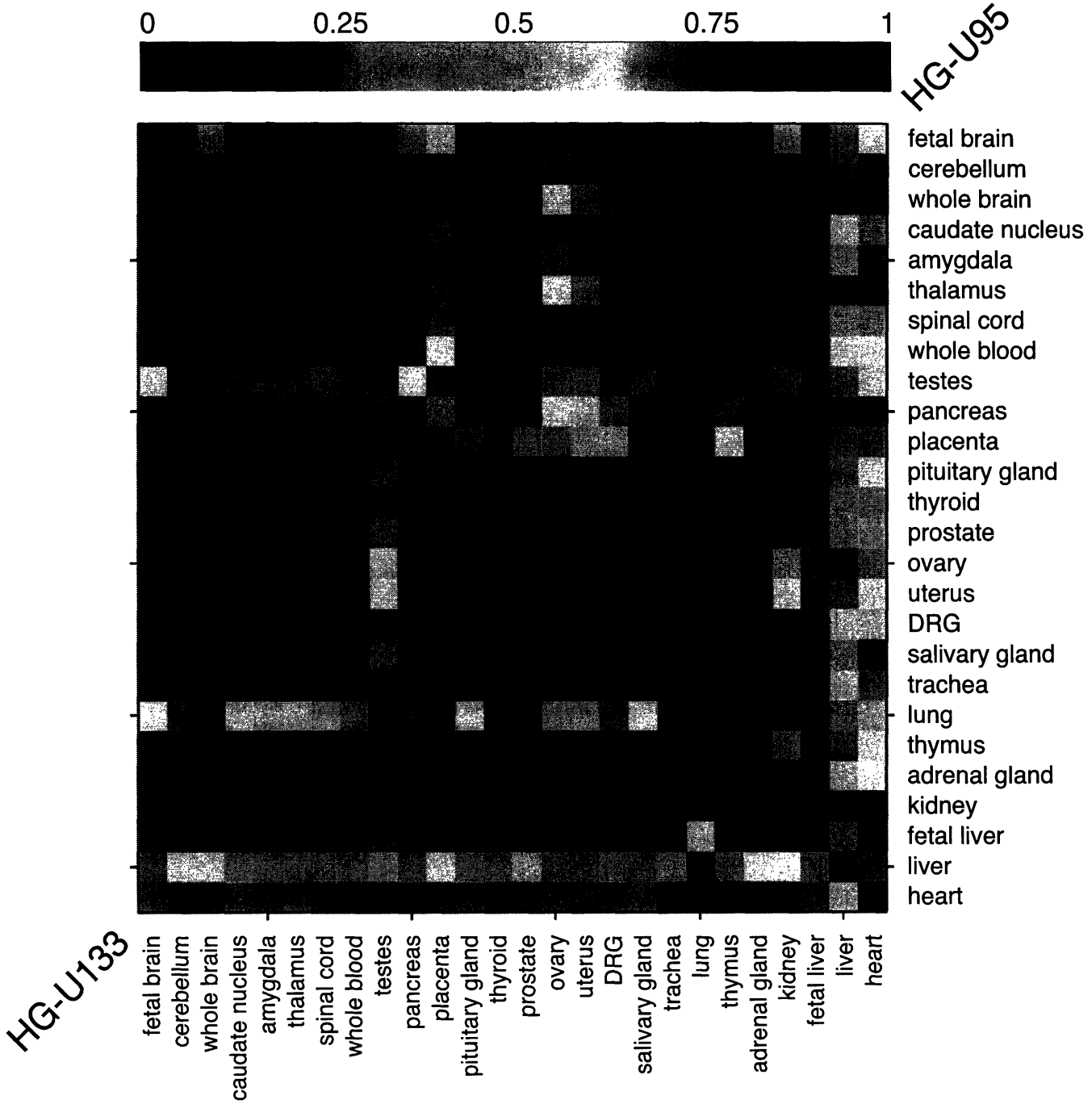
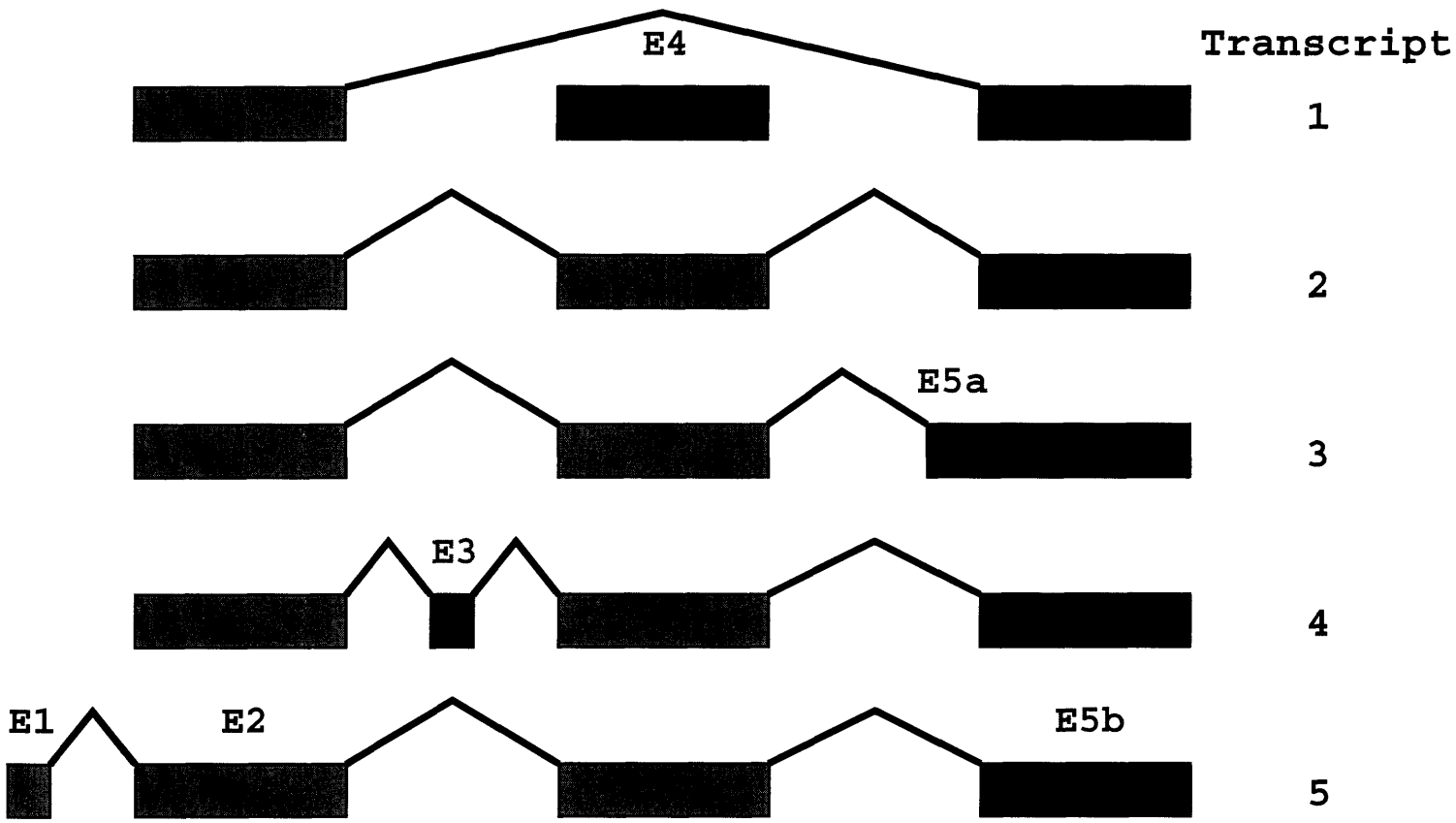

0
0.25
0.5
0.75
1
HG-U95
fetal brain
cerebellum
whole brain
caudate nucleus
amygdala
thalamus
spinal cord
whole blood
testes
pancreas
placenta
pituitary gland
thyroid
prostate
ovary
uterus
DRG
salivary gland
trachea
lung
thymus
adrenal gland
kidney
fetal liver
liver
heart
HG-U133
fetal brain
cerebellum
whole brain
caudate nucleus
amygdala
thalamus
spinal cord
whole blood
testes
pancreas
placenta
pituitary gland
thyroid
prostate
ovary
uterus
DRG
salivary gland
trachea
lung
thymus
adrenal gland
kidney
fetal liver
liver
heart

Figure 4



d(i,j): number of SJs that differ between transcripts *i,j*

t(i,j): total number of SJs in transcripts *i,j*

$SJD(i,j) = d(i,j) / t(i,j)$

transcripts *i*	transcripts *j*	*SJD(i,j)*
1	2	3/3=1
2	3	2/4=0.5
2	4	3/5=0.6
1	4	4/4=1
2	5	0/4=0

Figure 5

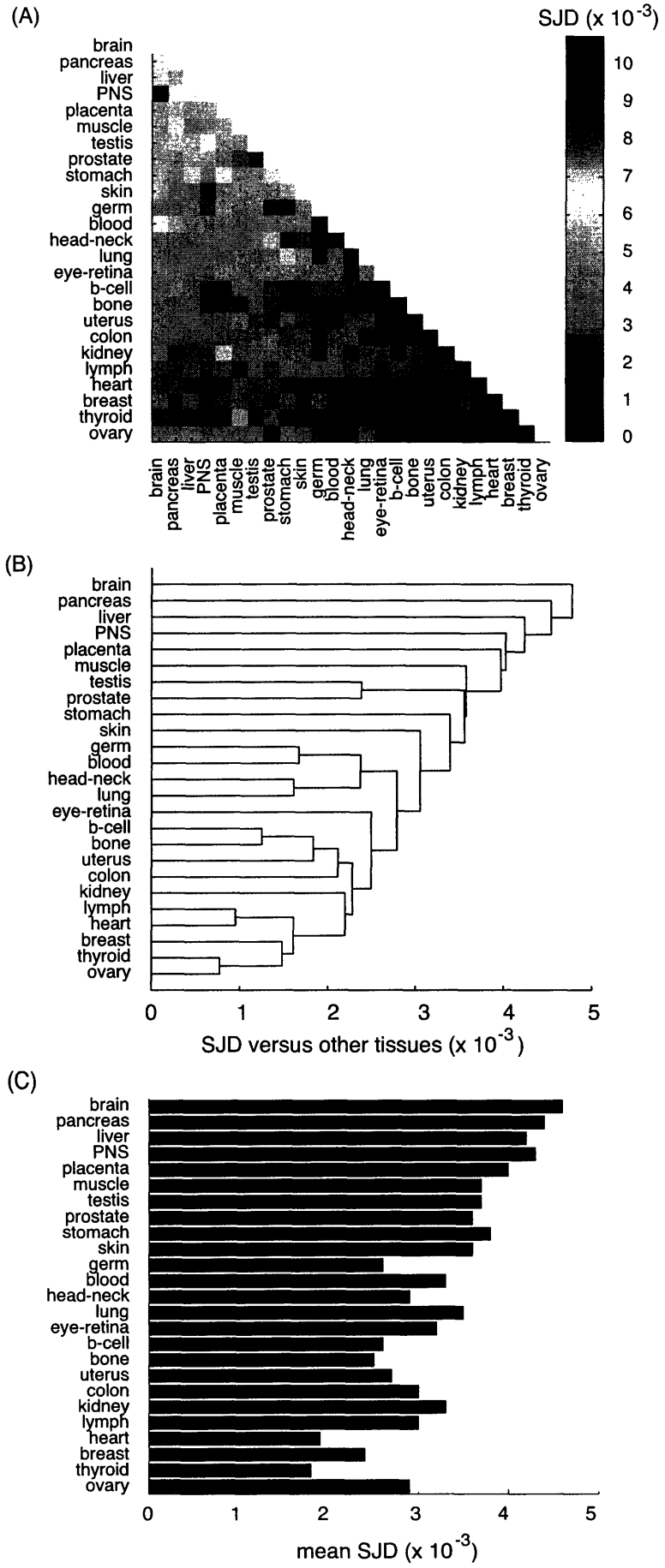

(A)
SJD (x 10-3)
brain
pancreas
liver
PNS
placenta
muscle
testis
prostate
stomach
skin
germ
blood
head-neck
lung
eye-retina
b-cell
bone
uterus
colon
kidney
lymph
heart
breast
thyroid
ovary
10
9
8
7
6
5
4
3
2
1
0
(B)
0
1
2
3
4
5
SJD versus other tissues (x 10-3)
(C)
mean SJD (x 10-3)

Figure S1. Sampling 10 ESTs from gene regions with at least 15 ESTs aligned the region. ESTs are derived from strictly normal cDNA libraries.

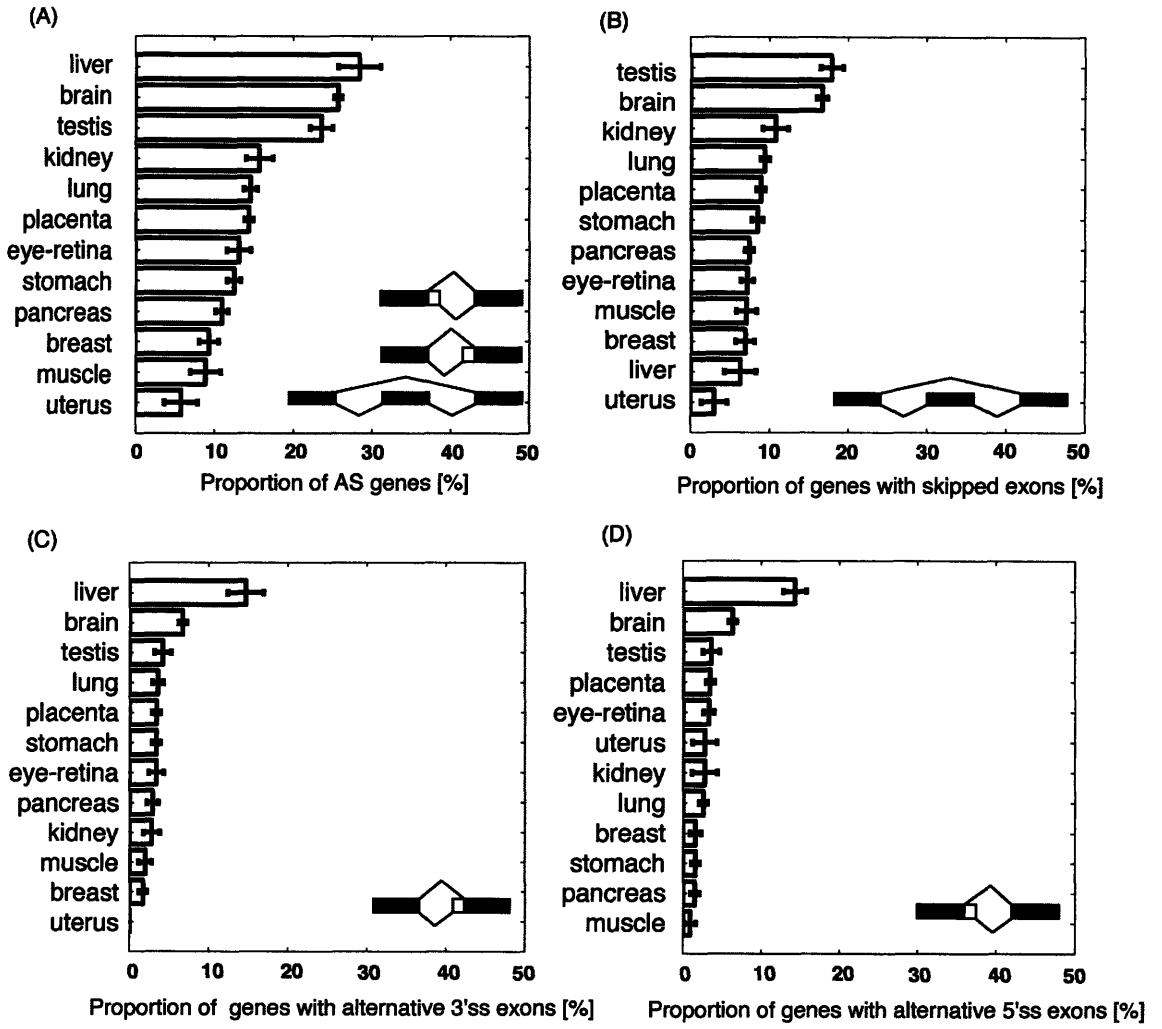
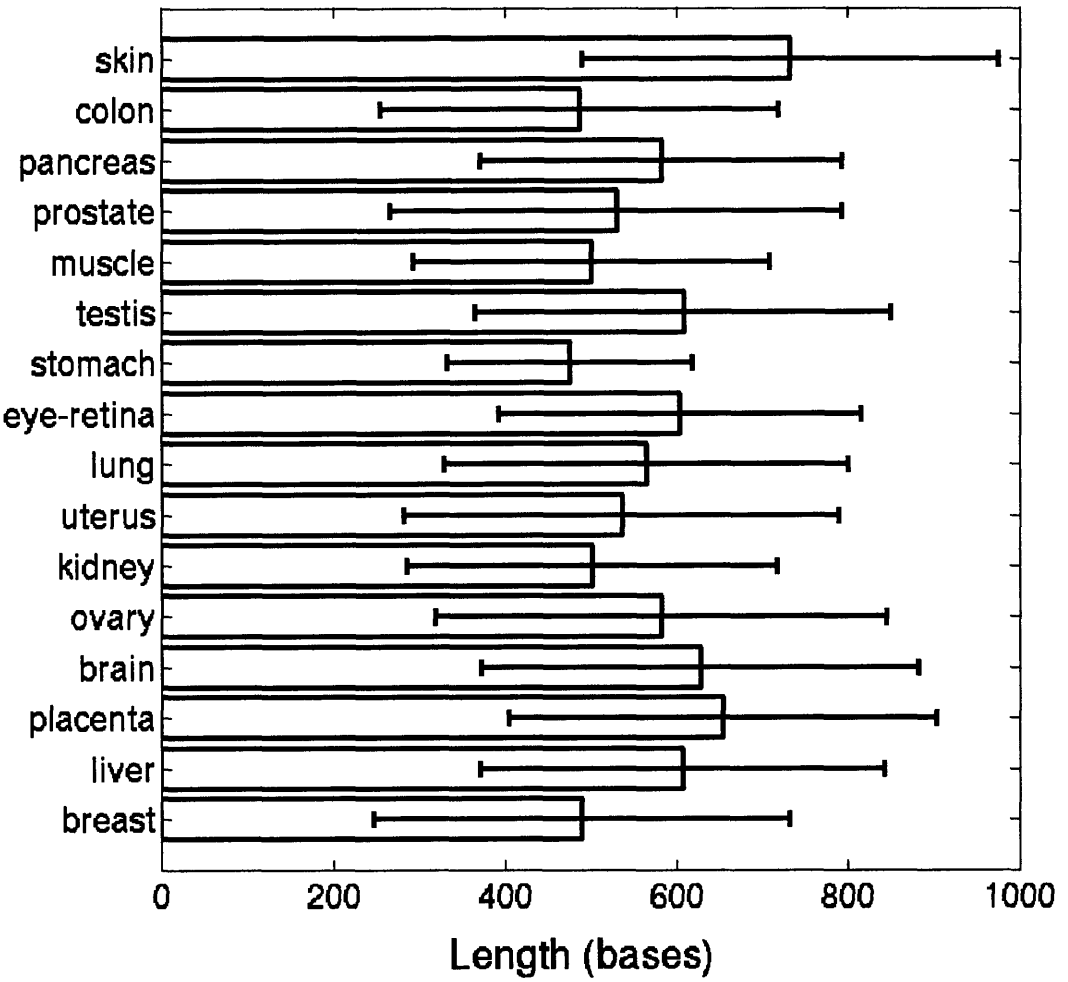
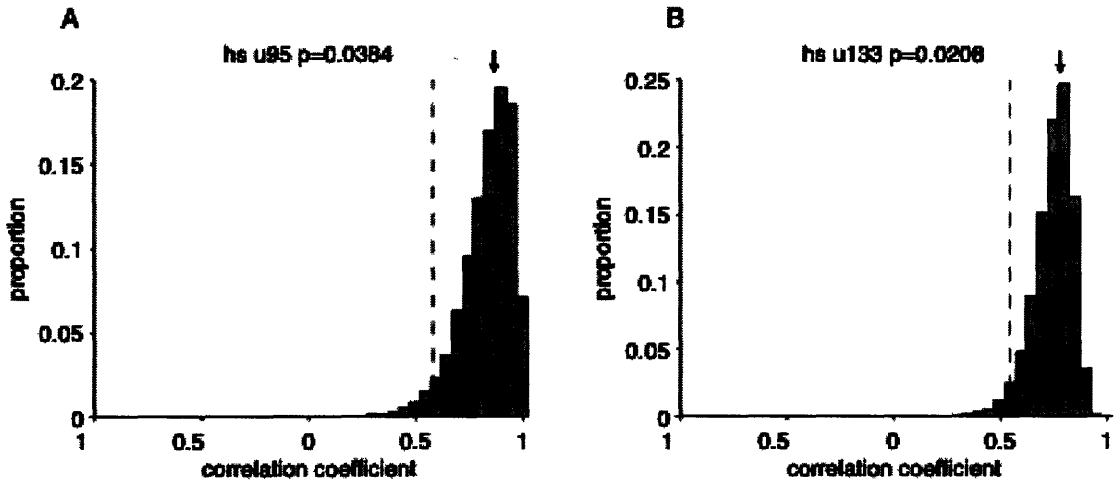


Figure S2. The average length in bases of ESTs stringently aligned to gene regions across tissues. Error bars are 1 standard deviation.



Supplementary Figure 3



Distribution of mean correlation coefficients for random sets of genes. For each iteration (total of 1000 iterations), we selected 20 random genes and computed the correlation coefficient between the expression levels in 26 different tissues as in Figure 3 for the U95A chip (A) and the U133A chip (B). The figure shows the distribution of the average correlation coefficient for each tissue. The arrow indicates the mean value. The dashed vertical line shows the mean correlation coefficient obtained in Figure 3 for the liver tissue. The *p*-value above each subplot shows the proportion of values that were less than the observed value for the liver (total number of coefficients in the distribution = 1000 iterations x 26 tissues).

Genes with >=1 EST per tissue			
Tissues	Mean	Median	Num. genes
kidney	3.69	2	6115
ovary	3.57	2	4771
breast	4.29	2	5877
uterus	4.89	2	6581
testis	4.29	2	7890
prostate	4.04	2	5466
colon	5.52	3	7343
eye_retina	5.13	2	6995
lung	6.72	3	7672
skin	5.92	3	6176
stomach	7.42	3	6181
brain	10.18	4	10437
placenta	6.82	3	6727
muscle	3.89	2	5243
pancreas	7.57	3	6658
liver	4.4	2	4326

Genes with >=30 ESTs per tissue			
Tissues	Mean	Median	Num. genes
kidney	42.5	36	20
ovary	38.82	35	33
breast	42.98	40	58
uterus	44.01	38	101
testis	45.95	39	64
prostate	46.82	36	57
colon	44.56	38	164
eye_retina	47.7	39	134
lung	51.16	42	276
skin	52.11	42	178
stomach	54.6	43	275
brain	57.73	46	797
placenta	70.31	41	212
muscle	74.29	51	56
pancreas	74.93	47	267
liver	81.44	50	72

Genes with >=20 EST per tissue			
Tissues	Mean	Median	Num. genes
kidney	28.16	25	82
ovary	29.01	25	87
breast	30.58	26	158
uterus	31.56	26	252
testis	32.49	27	161
prostate	33.15	28	140
colon	33.36	28	357
eye_retina	34.25	27	307
lung	36.8	29	582
skin	37.18	29	374
stomach	39.84	30	531
brain	42.7	31	1431
placenta	47.99	30	405
muscle	49.42	30	110
pancreas	51.02	31	502
liver	60.01	35	114

Genes with >=40 & <60 ESTs per tissue			
Tissues	Mean	Median	Num. genes
kidney	47.5	49	4
ovary	45.17	45	12
breast	48.13	48	23
uterus	47.83	47	36
testis	48.15	48	20
prostate	47.69	47	13
colon	47.52	46	54
eye_retina	47.21	46	43
lung	47.96	46	102
skin	48	48	62
stomach	47.87	48	100
brain	48.3	48	260
placenta	47.38	47	60
muscle	48.14	50	14
pancreas	48.31	47	91
liver	50.29	50	17

Table S1. Mean and median number of ESTs per gene, and the total number of genes inferred at different minimum number of ESTs required.

Normal tissues	Average number of genes	Average number of AS genes	Average number of genes with skipped exons	Average number of genes with alternative 5'ss exons	Average number of genes with alternative 3'ss exons
kidney	75.9	11.9	8.1	2.1	2.0
pancreas	288.7	32.1	21.3	4.5	8.8
testis	189.7	44.7	34.1	6.8	8.0
eye-retina	267.0	34.9	19.5	8.7	9.0
stomach	734.6	90.8	62.0	12.6	24.2
brain	1635.8	423.3	272.9	102.6	108.7
placenta	409.4	59.1	36.8	14.3	14.4
breast	173.7	15.9	12.1	2.5	2.6
muscle	177.1	15.6	12.5	1.5	3.5
uterus	71.3	4.2	2.2	2.0	0.0
liver	91.8	26.3	5.6	13.3	13.6
lung	430.7	62.9	41.1	11.6	16.1

Table S2. The average total number of genes, AS genes, genes containing skipped, exons, genes containing alternative 5'ss or 3'ss exons. Splicing patterns were inferred utilizing 10 ESTs from genes with at least 15 ESTs from a particular normal tissue cDNA library.

AS event	Normal tissue	Disease-associated tissue	Normal and disease
	ASG (CSG)	ASG (CSG)	ASG (CSG)
SE	195 (1201)	275 (1115)	261 (1128)
A5E	134 (1261)	119 (1271)	147 (1242)
A3E	124 (1271)	107 (1266)	107 (1282)
ASG	382 (1013)	435 (947)	435 (954)

AS event	Normal tissue	Disease-associated tissue	Normal and disease
	ASG / (CSG+ASG)	ASG / (CSG+ASG)	ASG / (CSG+ASG)
SE	14%	20%	19%
A5E	10%	9%	11%
A3E	9%	9%	8%
ASG	27%	32%	31%

Table S3. The number of alternatively spliced genes (ASGs) and constitutively spliced genes (CSGs), genes containing skipped exons (SEs), alternative 5'ss (A5E), and alternative 3'ss exons. Splicing patterns were inferred by using 16 ESTs per gene from normal cDNA libraries, 16 ESTs per gene from diseased cDNA libraries, and 8 ESTs from normal and 8 ESTs from diseased cDNA libraries. ESTs from the 16 tissue libraries similar to that in Figure 1 were combined in this analysis. The fraction of genes containing SEs using ESTs from normal tissue libraries was significantly smaller (p <1E-5) by a chi-square test (with Yates correction), as compared to using ESTs from disease associated libraries.

Ensembl ID	Splicing factor	HG-95A probe ID	HG-133A probe ID
136450	ASF/SF2	36098_at	211784_s_at
167978	SRm300	32761_at	207435_s_at
161547	SC35	36111_s_at	200753_x_at
100650	SRp40	40453_s_at	212266_s_at
124193	SRp55	35808_at	208804_s_at
111786	SRp30c	32573_at	201698_s_at
115875	9G8	32165_at	213649_at
116754	Srp54	32183_at	200685_at
136527	SFRS10	140_s_at	210180_s_at
112081	SRp20	351_f_at	208673_s_at
135486	hnRNP A1	31463_s_at	213356_x_at
122566	hnRNP A2/B2	36654_s_at	205292_s_at
092199	hnRNP C	32408_s_at	200751_s_at
138669	hnRNP D	38016_at	2000073_s_at
147274	hnRNP G	39731_at	213762_x_at
169045	hnRNP H1	41292_at	213472_at
165119	hnRNP K	39415_at	200775_s_at
104824	hnRNP L	35201_at	202072_at
099783	hnRNP M	37717_at	2000072_s_at
125970	hnRNP RALY	36125_s_at	201271_s_at

Table S5. Human splicing factors of SR, SR-related and hnRNP protein families. The last six digits of the Ensembl gene number (ENSG00000#) are in the first column, followed by the gene name and corresponding Affymetrix DNA microarray HG-95A and HG-133A probe identification numbers..

Yeo et al.

Chapter 4

Prediction of alternative exons

4.1 Predictive identification of alternative splicing events conserved in human and mouse

4.1.1 Abstract

Alternative pre-messenger RNA splicing affects a majority of human genes and plays important roles in development and disease. Alternative splicing (AS) events conserved since the divergence of human and mouse are likely of primary biological importance, but relatively few such events are known. Here we describe sequence features that distinguish exons subject to evolutionarily conserved AS, which we call 'alternative-conserved exons' (ACEs), from other orthologous human/mouse exons and integrate these features into an exon classification algorithm, ACEScan. Genome-wide analysis of annotated orthologous human-mouse exon pairs identified ~2,000 predicted ACEs. Alternative splicing was verified in both human and mouse tissues using an RT-PCR-sequencing protocol for 21 of 30 (70%) predicted ACEs tested, supporting the validity of a majority of ACEScan predictions. By contrast, AS was observed in mouse tissues for only 2 out of 15 (13%) tested exons that had EST or cDNA evidence of AS in human but were not predicted ACEs, and was never observed for eleven negative control exons

in human or mouse tissues. Predicted ACEs were much more likely to preserve reading frame, less likely to disrupt protein domains than other AS events, and were enriched in genes expressed in the brain and in genes involved in transcriptional regulation, RNA processing and development. Our results also imply that the vast majority of AS events represented in the human EST database are not conserved in mouse.

4.1.2 **Introduction**

The processing of human primary transcripts to produce the messenger RNAs (mRNAs) that will direct protein synthesis is often variable, producing multiple alternatively spliced (AS) mRNA products, most commonly by alternative inclusion or exclusion ('skipping') of individual exons (1-3). Alternative pre-mRNA splicing plays a major role in expanding protein diversity and regulating gene expression in higher eukaryotes (4, 5). Regulated AS is crucial in fruit fly development (3) and in the physiology of the heart, skeletal muscle, brain and other tissues, and mis-regulation of AS is associated with human disease (6-8).

Expressed sequence tag (EST) and cDNA sequence databases provide a rich source of information about splicing events occurring in the human and mouse transcriptomes. Considering the set of human ESTs and cDNAs which can be reliably aligned to a human gene locus overlapping a particular exon, this set can be subdivided into those transcripts which include and those which exclude or 'skip' the exon in question. Here, 'skipping' of an exon refers to the situation where a transcript aligns consecutively to an upstream exon and a downstream exon of the gene in question, omitting the given exon. This consideration can be applied to all of the exons in a human gene, and an analogous subdivision can be made of the mouse transcripts that align to exons of the orthologous mouse gene. Each orthologous

human/mouse exon pair can then be assigned to one of four categories, $S_{H,m}$, $S_{h,M}$, $S_{H,M}$, or $S_{h,m}$, depending on whether exon skipping has been observed only in human transcripts ($S_{H,m}$), only in mouse ($S_{h,M}$), in both human and mouse ($S_{H,M}$), or not observed in either species ($S_{h,m}$).

Using publicly available EST databases totaling over 5 million human and over 3 million mouse ESTs, and databases of ~94,000 and ~91,600 human and mouse cDNAs respectively, thousands of alternative exons can be inferred in each species. However, the overlap between these sets is relatively small, i.e., for only about 240 (~ 1 in 18) of the ~ 4,500 conserved human-mouse exons observed to be skipped in human was transcript evidence found supporting alternative usage (skipping) of the orthologous mouse exon, as discussed below (9-11). This observation raises the question of how many of the AS events observable in the human transcriptome are evolutionarily conserved, and therefore presumably contribute to organismal fitness, and how many are aberrant, disease- or allele-specific, or highly lineage-restricted events, which may or may not affect fitness. Although study of the latter types of events may lead to important insights and applications, a significant fraction of these events may constitute biochemical 'noise' or transient evolutionary fluctuations. On the other hand, conservation of a specific pattern of AS over the ~90 million years since divergence of the mouse and human lineages provides strong evidence of biological function. Therefore, defining the set of AS events conserved between human and mouse is of primary interest in efforts to understand the biological importance of splicing regulation.

Alternative inclusion/exclusion of exons is known to be influenced by a number of factors, such as intron length, exon length, splice site strength and pre-mRNA secondary structure (1, 3, 12). Certain *cis*-regulatory elements, including exonic and intronic splicing enhancers (ESEs and ISEs, respectively), as well as exonic and intronic splicing silencers (ESSs

and ISSs, respectively) can also control exon skipping by recruiting *trans*-acting splicing factors (4, 13). Computational studies have identified other sequence features that differ between skipped (also known as 'cassette') exons and constitutive exons in human and mouse genes, including increased conservation in the introns flanking exons skipped in both human and mouse (9, 10, 14-16). These observations motivated us to systematically identify, characterize and integrate sequence features into a classifier that could be used to identify exons subject to evolutionarily conserved exon skipping, here termed 'alternative-conserved exons' (ACEs).

4.1.3 Materials and Methods

Regularized least-squares classification

The regularized least-squares classifier (RLSC) was used to learn the features from $S_{H,M}$ and $S_{h,m}$ exons and to derive a real-valued output for unlabeled conserved exon pairs. The RLSC has a quadratic loss function and requires the solution of a single system of linear equations, $(\underline{K} + \lambda L\mathbf{W}^{-1})\,\mathbf{c} = \mathbf{y}$, in matrix notation. The goal is to obtain an optimal vector **c**, defined as $\boldsymbol{c} = [c_1 \ldots c_L]^T$, where L is the size of the training set, $\underline{K}$ is the L x L kernel matrix, λ is the "tradeoff" between generalization and over-fitting, **W** is the diagonal matrix of penalties w_i (equal to β for positive examples and equal to 1 for negative examples), and **y** is the column vector of labels (+1,-1). The algorithm, cross-validation, sampling, and performance measures are described in further detail in *Supporting Text*, supporting information.

Experimental validation

The Invitrogen Superscript III First-Strand synthesis system for RT-PCR (Cat. No. 18080-051) was used to generate cDNAs from normal human (fetal brain, fetal liver, cerebellum, heart, whole brain, prostate, liver, lung, kidney, bone marrow, skeletal muscle and testis) and normal

mouse (embryonic mix, whole brain, kidney, skeletal muscle, liver, lung, heart and testis) tissues using oligo(dT) primers. The Invitrogen Taq DNA polymerase kit (Cat. No. 18038-042) was utilized with primers targeted to exons flanking candidate ACEs (for further details see *Supporting Text*, supporting information). PCR products of the expected size were gel-purified using the QIAquick Gel Extraction Kit (Qiagen Cat. No. 28704).

4.1.4 **Results and Discussion**

Outline of strategy for identification of alternative-conserved exons

Our scheme for identifying ACEs consisted of three phases: 1) learning; 2) prediction; and 3) validation (Fig. 1). In the learning phase, a set of sequence features was identified, including exon and intron length, splice site strength, sequence conservation, and region-specific oligonucleotide composition, which differed between 'training sets' of 241 exons of the class $S_{H,M}$ and ~5,000 exons of the class $S_{h,m}$ defined above (Fig. 2). Additionally, for training purposes, exons of the $S_{h,m}$ class were chosen from genes containing at least one other exon with evidence for AS, as genes lacking AS may be under different degree of selection than AS genes (17). Next, these features were incorporated into a discriminant classifier, ACEScan, which was used in the prediction phase to predict which of ~96,000 annotated orthologous human/mouse exon pairs not previously known to exhibit conserved AS are in fact ACEs. Finally, in the validation phase, a subset of candidate exons with positive ACEScan scores (designated 'ACEScan-positive' or ACEScan[+] exons) was chosen for experimental testing, together with two sets of negative control exons with negative ACEScan scores ('ACEScan-negative' or ACEScan[-] exons): one set with previous transcript evidence for exon skipping in human (S_H category) and one set lacking such evidence (S_h category).

The following features were first incorporated into ACEScan: (i) exon length; (ii) upstream and (iii) downstream intron lengths; (iv) 5' splice site (5'ss) and (v) 3' splice site (3'ss) scores; (vi) nucleotide percent identity between orthologous human and mouse exons; and human-mouse intronic sequence conservation (vii) within the last 150 bases upstream and (viii) within the first 150 bases downstream of the exon. In general, exon pairs skipped in both human and mouse (set $S_{H,M}$) were observed to be shorter than unskipped exon pairs ($S_{h,m}$), were flanked by longer upstream and downstream introns, and possessed significantly weaker splice sites (Fig. 2). Strikingly, exon-pairs in $S_{H,M}$ have significantly higher sequence identity and higher flanking intronic conservation as compared to exon pairs in $S_{h,m}$ (Fig. 2). High levels of sequence conservation in the exons and flanking introns is suggestive of conservation of regulatory motifs or RNA structure. These observations are similar to and consistent with previous studies (10, 14-16).

Oligonucleotides useful in discrimination of alternative-conserved exons

Oligonucleotide features designed to score potential *cis*-regulatory elements consisted of the highest-ranking (most biased) over- and under-represented oligonucleotides of length k (k-mers) in different exon and intron regions. The regions considered were the first and last 100 bases of exons and the proximal 150 bases in the upstream and downstream introns flanking the exon, because of the high levels of sequence conservation in these regions and their proximity to the regulated splice junctions. Counts of conserved oligonucleotides in human-mouse nucleotide alignments of the 150 bases of upstream and downstream intronic sequence and in the entire exon were scored for enrichment in the set $S_{H,M}$ versus $S_{h,m}$. Inclusion of oligonucleotide counts

from aligned as well as unaligned sequences permits scoring of *cis*-regulatory elements that both do and do not require strict spatial constraints for function.

Oligonucleotides were ranked by their enrichment as measured by a χ^2 value. Several of the over-represented intronic elements were similar to known intronic regulatory elements (e.g. UGCAUG, UC-rich repeats; Table 1 and *Supporting Text*, which is supporting information). We propose that a significant fraction of the remaining elements may represent novel intronic regulatory sequences. A number of the over- and under-represented exonic elements were similar to known or predicted ESE or ESS motifs (18, 19). Their relative distribution (Table 2 and *Supporting Text*, which is supporting information) suggests that ACEs have a lower density of ESEs and a higher density of ESS sequences relative to constitutive exons. Both of these features would tend to facilitate exclusion by the splicing machinery. Previously, candidate ESEs were identified in part based on enrichment in constitutive exons with weak splice sites (19). The apparently reduced frequency of ESEs and increased frequency of ESS sequences in ACEs relative to constitutive exons might reflect differing degrees of selection, with constitutive exons presumably being under selection for efficient exon inclusion, while alternative exons are presumably selected for inefficient inclusion under at least some conditions (e.g., cell-type specificity or developmental stage-specific).

Integration and selection of features for accurate exon classification

The task of integrating the general features and oligonucleotide features described above into an algorithm that distinguishes exon pairs in $S_{H,M}$ (positively labeled) from those in $S_{h,m}$ (negatively labeled) was posed as a supervised binary classification problem. We adapted a regularized least-squares classifier, which finds the optimal separating hyperplane in a high-

dimensional space that distinguishes two classes of samples (20). As it was not known *a priori* which of the 8,245 general and oligonucleotide features were most important in the classification scheme, models utilizing different combinations of the eight general features and the region-specific oligonucleotide features were compared, and a feature selection protocol was used to reduce the number of parameters and to retain only the most relevant oligonucleotide features.

In order to determine the optimal features and parameters for the classifier, the training data were used to generate several models, by varying the choice of general features, the exon or intron regions from which oligonucleotide features were generated, and the number of most discriminative oligonucleotide features included. The model with the best performance utilized all of the general sequence features and 240 oligonucleotides of lengths 4 and 5 (shown in Fig. 5, supporting information). This model assigned correct labels to ~90 exon pairs for every 100 exon pairs drawn equally likely from $S_{H,M}$ and $S_{h,m}$. For an individual exon, the ACEScan score was defined as the mean of the classifier outputs over 50 random samplings of the training data. The distribution of ACEScan scores for the exon pairs in $S_{H,M}$ ranged from approximately –0.8 to 2.0 (arbitrary units), compared to a range of approximately –1.8 to 0 for most of the exons in $S_{h,m}$ (Fig. 1). At a cutoff score of zero, only ~2% of $S_{h,m}$ exons had positive ACEScan scores, compared to ~61% of the exons in $S_{H,M}$, suggesting that ACEScan[+] exon pairs are highly enriched for ACEs.

Experimental validation of conserved AS for 21 of 30 ACEScan[+] exon pairs

A combination of experimental tests and bioinformatic approaches was used to explore the features of ACEScan[+] and ACEScan[-] exon pairs. First, the splicing patterns of a set of 30 arbitrarily chosen ACEScan[+] exons were tested in a battery of human and mouse tissues by

reverse transcriptase PCR (RT-PCR) with primers targeted to flanking exons. ACEScan[+] exons were selected from four intervals: I1 (ACEScan score from 0.0-0.5); I2 (0.5-1.0); I3 (1.0-1.5); and I4 (greater than 1.5), spanning the range of scores of most $S_{H,M}$ exons. Panels of twelve normal human tissues and eight normal mouse tissues were assayed. In order to avoid the undesired detection of aberrant or disease-specific splicing, tumor or other diseased tissues were not utilized. The products of these 600 RT-PCR reactions (30 exons x 20 tissues) were analyzed by gel electrophoresis, and the identities of PCR products with expected sizes for mRNAs including or excluding the test exon were confirmed by sequencing (Fig. 3A). In all, 4 out of 9, 7 out of 8, 6 out of 8, and 4 out of 5 candidate ACEs in intervals I1, I2, I3 and I4, respectively, were observed to undergo skipping in both human and mouse, while for another two exons (both from interval I1) exon skipping was observed only in human tissues (Fig. 3; complete results shown in Table 3 supporting information). Thus, of 30 predicted ACEs interrogated by RT-PCR, 21 were observed to be skipped in both human and mouse tissues, and high rates of validation of AS were seen in all four score intervals. These data support the presence of conserved AS in a majority of ACEScan[+] exons. Although the 30 ACEScan[+] candidates had no previous transcript evidence for skipping, searches of the literature and low-stringency searches of the cDNA and EST databases (August 2004) identified possible evidence for a fraction of the AS events observed by RT-PCR, most often consisting of a single EST in only one species. In the examples studied, exon skipping was observed in many different combinations of human and mouse tissues, suggesting that many of the features utilized by ACEScan are characteristic of skipped exons generally, regardless of tissue-specificity. Variations in tissue specificity of AS were observed between human and mouse for several tested exons. However, a general tendency to conserve exon skipping in corresponding tissues was

apparent, e.g., 9 out of 10 predicted ACEs observed to be skipped in human whole brain or cerebellum were also skipped in mouse brain tissue (Table 3).

Low detection of conserved AS for ACEScan[-] exon pairs

As a negative control, eleven ACEScan[-] exon pairs from the set S_h were chosen from the five score intervals C1 (-0.5 to 0), C2 (-1.0 to –0.5), C3 (-1.5 to –1.0), C4 (-2.0 to –1.5) and C5 (less than –2.0) with at least one pair per interval. Using the same RT-PCR-sequencing assay and the same sets of human and mouse tissues, exon skipping was not observed for any of the eleven negative control exons in any of the 12 human or 8 mouse tissues studied (Table 3, supporting information). Thus, considering both the human and mouse exons tested, exon skipping was detected for 44 out of 60 ACEScan[+] exons (including 21 orthologous pairs), compared to 0 out of 22 ACEScan[-] exons, a highly significant difference ($P < 0.0001$, Fisher exact test). Of course, for either group of exons failure to detect exon skipping by our RT-PCR assay is not proof that exon skipping does not occur, and some exons not skipped in the tissues studied might be skipped in other untested tissues. However, low-stringency searches of the August 2004 human and mouse EST databases failed to detect any evidence of skipping of the 11 ACEScan[-] exons tested.

As a second type of negative control, an arbitrary set of 15 ACEScan[-] exon pairs were chosen from intervals C2 to C4, with the added requirement that transcript evidence of exon skipping was present for the human member of each exon pair. Using the same RT-PCR-sequencing assay in the same set of 8 mouse tissues as above, exon skipping was detected for only 2 out of the 15 mouse exons tested, suggesting that a substantial majority of these exon pairs are not ACEs. To explore the potential biological roles of the 13 remaining exons which

undergo possible human-specific AS, we examined the tissue sources of the transcripts that showed exon skipping: in 9 out of the 13 cases, these transcripts derived exclusively from cancer cell-lines or diseased tissues, suggesting that many of these exons may be skipped primarily in disease states rather than in normal human tissues. The difference in the rate of RT-PCR validation of exon skipping in mouse tissues for the ACEScan[+] exons tested (21 out of 30 = 70%), relative to the ACEScan[-] exons tested (2 out of 26 = ~8%), was also highly significant ($P < 0.002$, Fisher exact test), demonstrating the power of ACEScan to discriminate evolutionarily conserved AS exons from those which are either constitutively spliced or skipped in a species- (or disease-) specific manner.

Many Literature-derived AS events correspond to ACEScan[+] exons

The principle that important regulatory elements are usually evolutionarily conserved is well established, and forms the basis of a number of successful comparative genomics approaches for identifying such elements (21). To explore the extent to which this principle applies to AS events, we extracted known exon skipping events from the Manually Annotated Alternatively Spliced Events (MAASE) database (22), representing AS events that are curated from published works. A total of 29 exon skipping events in mouse were identified from this database, for which both the human and mouse orthologous exons were available. Strikingly, almost all of the extracted exons had ACEScan scores greater than –0.5 (28 out of 29) and 62% (18 out of 29) were ACEScan[+]. Thus, though small in scale, this analysis of published AS events suggests that a majority of 'interesting' exon skipping events (i.e., interesting enough to be described in the scientific literature) are ACEScan[+] and therefore that most such events are conserved between human and mouse (Table 4, supporting information).

About 11% of EST/cDNA-derived AS events are likely to be evolutionarily conserved

Of the ~4,300 exon-pairs with transcript evidence of skipping in human but not mouse (class $S_{H,m}$), only ~7% had positive ACEScan scores (Fig. 1). Together with the observation that ~61% of $S_{H,M}$ exons were ACEScan[+], this low fraction suggests that for only ~11% (= 0.07 / 0.61) of the $S_{H,m}$ exons is AS likely to be conserved in mouse. Thus, a surprising implication of these data is that the vast majority of the AS events inferable from human EST/cDNA-genomic alignments are not evolutionarily conserved in mouse. Instead, most of these events may represent aberrant, disease-specific, or allele-specific splicing (23), or events whose phylogenetic distribution is highly restricted.

Functional differences between ACEScan[+] and ACEScan[-] exons

To assess potential functional differences between ACEScan[+] and ACEScan[-] exons that either have or do not have EST or cDNA evidence of exon skipping in human, we analyzed the density of single nucleotide polymorphisms (SNPs) and the frequency of reading frame preservation and protein domain disruption for each of these three classes of exon. Selective pressure on nucleotide sequence was assayed by mapping stringently filtered reference SNPs onto exons that had been scored by ACEScan (Fig. 3B). This analysis found a ~50% higher density of SNPs in ACEScan[-] S_H exons than in ACEScan[+] exons (this difference is significant at $P < 10^{-5}$ by χ^2 test), suggesting that ACEs have been under much more stringent selection to conserve nucleotide sequence in recent human evolution than other exons. By contrast, ACEScan[-] S_H exons appear to have experienced a degree of selection that was more similar to constitutive exons than to ACEs.

Further evidence for the functional roles of many ACEScan[+] S_H exons came from the observation that a far higher fraction of these exons had lengths which were multiples of three (68%, comparable to that seen in the training set of $S_{H,M}$ exons) than was seen for ACEScan[-] S_H exons, for which only ~43% had lengths divisible by three, near background levels for constitutive internal exons (Fig. 3C). This difference is highly significant ($P < 10^{-15}$ by χ^2 test) and implies the existence of strong selection on the alternative protein products derived from alternative splicing of ACEScan[+] exons. Notably, divisibility of the exon length by three was not used in the predictions (only the general size of the exon, with shorter lengths favored over longer lengths).

The frequency of disruption or removal of a protein domain by AS has been studied by several groups (24-26). We found that only ~37% of ACEScan[+] exons overlapped open reading-frame regions encoding Interpro-annotated protein domains by 30 bases (10 codons) or more, a significantly lower fraction than for ACEScan[-] exons studied of either the S_H or S_h classes (Fig. 3D), both of which had similar frequencies of domain disruption (around 50%). Reducing the minimum overlap to 15 bases gave similar results (data not shown). This finding is generally consistent with the results of Kriventseva and coworkers, who observed that protein isoforms arising from AS are more likely to preserve protein domain structure than expected by chance (25). Taken together, the data shown in Fig. 3 consistently demonstrate that ACEScan[+] exons are under strong selection to conserve function, both at the nucleotide level (Fig. 3B), and at the level of the encoded alternative protein isoform (Figs. 3C, D). In contrast, ACEScan[-] exons show less evidence of selective constraints at the nucleotide level (Fig. 3B) and there is little if any evidence of additional constraints on the protein products derived from exon skipping

of ACEScan[-] exons, even when there is transcript evidence that such skipping occurs (Figs. 3C, D).

Applications of ACEScan at the gene level

Application of ACEScan to well-studied genes illustrates some of the strengths and limitations of our approach (*APP* and *GLUR-B* shown in Fig. 4; *PTB* and *CACNA1G* in Fig. 6, supporting information). Of the identifiable orthologous human/mouse exon pairs in these genes, known exon skipping events (marked by asterisks) all received positive ACEScan scores, implying that their skipping is likely to be conserved in mouse. Skipping of exons 7 and 8 of the β-amyloid precursor protein (*APP*) gene, implicated in Alzheimer's disease, was detected successfully in a recent large-scale microarray analysis of AS in human tissues (27). These exons, as well as exon 15 of the *APP* gene received positive ACEScan scores (Fig. 4A); all three of these exons are known to undergo exon skipping (28, 29). The *GLUR-B* gene, one of the four *GluR* subunits that assemble to form the AMPA glutamate receptor, contains two well-known skipped exons ('flip' and 'flop', exons 14 and 15), both of which received positive ACEScan scores, as well as an exon (number 13, marked with an 'E') that undergoes RNA editing (30). This edited exon and the downstream intron form an RNA hairpin and are highly conserved in sequence (30). Despite this high level of exonic and intronic sequence conservation, this exon received a negative ACEScan score (Fig. 4B), providing an example of the specificity of our method for AS exons. A web server has been set up at (http://genes.mit.edu/acescan) to provide access to all ACEScan plots for Ensembl-annotated orthologous human/mouse gene pairs.

Recently, Bejerano et al. reported 111 exonic "ultraconserved" regions (UCRs) longer than 200 bases with 100% sequence identity between the human and mouse genomes (31), most

of unknown function. Comparing these to our set of predicted ACEs, 33 of the 37 UCRs (~89%) that mapped to internal exons that could be scored by ACEScan received positive ACEScan scores, suggesting that a number of these elements correspond to ACEs.

Functional characteristics of ACEScan[+] genes

In total, 1,550 genes were identified, containing 2,041 ACEScan[+] exons, ~85% of which lacked prior transcript (EST/cDNA) evidence for exon skipping. Initial comparisons to the partially annotated rat genome showed a high correlation between human-mouse and human-rat ACEScan scores, as expected (data not shown). In order to determine whether genes that contain ACEScan[+] exons, which we refer to as ACEScan[+] genes, are biased towards particular biological activities, we compared these genes to the set of genes not found to contain any ACEScan[+] exons (ACEScan[-] genes) using Gene Ontology (GO) classifications (http://www.geneontology.org) as previously described (31, 32). The results showed that ACEScan[+] genes are enriched for transcription factors and aminopeptidase activity, and for the actin binding, RNA binding and nucleic acid binding GO molecular function categories (Fig. 3E). In terms of GO biological process categories, ACEScan[+] genes were more likely to be involved in transcriptional regulation, neurogenesis, and development, and less likely to be involved in transport than ACEScan[-] genes. Only slight biases in GO category representation were present in the training set of $S_{H,M}$ genes (Fig. 7, supporting information). Closer examination of the ACEScan[+] genes that encode RNA binding factors identified ACEScan[+] exons in genes encoding many of the heterogeneous nuclear ribonucleoproteins, a majority of which (including *PTB*) are candidates for nonsense-mediated mRNA decay (NMD) (Fig. 6 and Table 5, supporting information).

To explore the expression patterns of genes containing predicted ACEs, we used microarray data from the Gene Atlas survey of 47 diverse human tissues and cell lines (33). Overwhelmingly, ACEScan[+] genes were more likely to be differentially expressed in a spectrum of nervous system tissues, including spinal cord, fetal and adult whole brain, and in several brain regions, compared to ACEScan[-] genes (Fig. 8 , supporting information). Only two cell lines (both ovarian) of the 47 tissues/cell lines studied exhibited similar biases. These results imply an unusually high frequency of conserved AS events in the brain.

While this work was in progress, two other groups have demonstrated that conserved sequence features can be used to identify alternative exons in fruit fly (34) and human genes (14, 35). Our computational approach differs substantially in a number of ways: (i) ACEScan associates a real-valued score to orthologous human-mouse exon pairs, rather than associating a binary label to an exon-pair, which grants much greater flexibility in adjusting the algorithm's sensitivity/specificity compared to (14, 35); (ii) ACEScan does not use the length of the exon modulo three in its predictions (14, 35). This allows us to assess the degree of selection on ACEs to preserve protein reading frame (Fig. 3C) rather than assuming that reading frame must always be preserved, and it enables ACEScan to identify the subset of ACEs which create mRNAs that encode truncated proteins or which are subject to NMD, an emerging class of regulated AS events (36). Supporting the validity of this subset of predictions, approximately half of the ACEs validated by our RT-PCR-sequencing protocol had lengths that were not divisible by three (Fig. 3A; Table 3, supporting information); (iii) a much larger set of discriminatory features was utilized in ACEScan, including oligonucleotide features (compared to (14, 34)), many of which are likely to represent splicing regulatory elements, and inclusion of these features enhanced the performance of our algorithms (cf. 35). Experimental validation of

both predicted AS exons, as well as negative control exons, is crucial in providing estimates for the reliability and accuracy of any computational approach. A comparison of sensitivity and specificity based on experimental validation demonstrates that ACEScan has higher accuracy than previously published approaches (Table 6, supporting information compares computational differences and extent of validation). Finally, the accuracy and relatively large numbers of ACEs predicted by ACEScan allow us to identify functional and expression biases in the set of genes containing high-confidence ACEs.

Comparative genomics, machine-learning techniques and rigorous experimental validation have facilitated the accurate prediction of a core set of ~2,000 alternative-conserved exons. This much enlarged set of conserved alternative exons holds the potential for further elucidating the roles of AS in modulating the expression of mammalian genomes.

4.1.5 **Acknowledgments**

The authors thank P. Sharp and Z. Wang for helpful discussions. This work was supported by grants from the National Science Foundation and the National Institutes of Health (C.B.B.), and from the Lee-Kuan-Yew graduate fellowship from Singapore (G.Y.)

4.1.6 **Figure Legends**

Figure 1. Schematic overview of the learning and prediction stages of the ACEScan procedure. In 'Learning', sequence features that differed between sets $S_{H,M}$ (skipped and included in human and mouse) and $S_{h,m}$ (only included in human and mouse) were identified as described (*Supporting Text,* supporting information). Random subsets of $S_{H,M}$ and $S_{h,m}$ were used to train the ACEScan algorithm and cross-validation scores were calculated for the unseen subsets of $S_{H,M}$ and $S_{h,m}$. The cross-validated ACEScan score distributions for $S_{H,M}$ (red) and $S_{h,m}$ (black) are shown. For 'Prediction', spliced alignment of transcript sequences were used to assign Ensembl-annotated exons from ~10,000 human-mouse orthologous gene pairs (not necessarily alternatively spliced) to one of two sets: $S_{H,m}$ (included in some human transcripts and excluded in others, but included in all mouse transcripts) and $S_{h,m}$ (described above). ACEScan score distributions for $S_{H,m}$ (purple) and $S_{h,m}$ (blue) are shown.

Figure 2. Sequence features that differ between conserved alternative and constitutive human-mouse exons. (**A**) Features typical of exons of the $S_{H,M}$ (alternatively spliced) and $S_{h,m}$ (constitutive) training sets are depicted. $S_{H,M}$ exons had shorter median exon length (93 bp versus 126 bp, $P<10^{-22}$), longer upstream intron length ($P<0.005$), longer downstream intron length ($P<10^{-5}$), weaker 5' and 3' splice site scores ($P<10^{-5}$ and $P<0.02$, respectively, using MAXENTSCAN scores at http://genes.mit.edu), higher exon sequence conservation (percent identity; $P<10^{-46}$), and higher conservation (ClustalW alignment score) in the 150 bp intron regions immediately upstream and downstream of the exon ($P<10^{-63}$ and $P<10^{-66}$, respectively). For each feature, the Kolmogorov-Smirnov

(KS) test was used to test the null hypothesis of independent samples drawn from the same underlying population. Length and splice site score values are shown for human exons/introns; mouse values were similar. Average percent identity for alignments of flanking intron regions are shown in a 9-base sliding window for $S_{H,M}$ (red) and $S_{h,m}$ (dashed, black) exons. **(B)** Pentanucleotides utilized by ACEScan. Over- (under-) represented pentamers in exon or 150-base flanking intron regions of $S_{H,M}$ versus $S_{h,m}$ exons are shown in red (black). Pentamer frequencies were analyzed separately for ClustalW-aligned regions only ('aligned') or entire region ('unaligned'). Exon 5'-, 3'-ends refer to first/last100 bases of exon.

Figure 3. Validation of ACEScan[+] predictions using experimental and computational approaches. **(A)** Experimental validation via RT-PCR and sequencing of subsets of candidate ACEScan[+] exons and negative control ACEScan[-] exons in panels of normal human and mouse tissues with primers in flanking exons. Graphical representations of splicing patterns (inclusion/exclusion) and the number of exon-pairs observed to be excluded and included are designated in red and black respectively. The three randomly selected subsets tested were: (i) 30 ACEScan[+] exon pairs; and as negative controls, (ii) 15 ACEScan[-] S_H exon pairs (with EST/cDNA evidence for both inclusion and exclusion of the human exon, indicated by horizontal lines representing spliced transcripts); and (iii) 11 ACEScan[-] S_h exon pairs (with no transcript evidence for skipping in either human or mouse). **(B)** SNP density in ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons. The number of stringently filtered SNPs per 10,000 bases were computed for each exon set. **(C)** Fraction of $S_{H,M}$ exons, ACEScan[+] S_H exons, and

ACEScan[-] S_H exons that have lengths that were a multiple of 3, and the background fraction of frame-preserving constitutive exons. (**D**) Analysis of protein domain preservation of ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons that maintain reading-frame (i.e. length divisible by 3). Maximum exon size cutoffs (150, 110 and 108 bp for ACEScan[+], ACEScan[-] S_H, and ACEScan[-] S_h exons respectively) were utilized to avoid exon length biases. The median length of exons in each subset was 84 bases, with no significant difference in the distribution of sizes among the sets (by a Kruskal-Wallis non-parametric test). The minimum number of exonic bases overlapping the protein domain was set to 30 bases. (**E**) Gene Ontology (GO) "Molecular function" and "Biological process" categories which differed significantly ($P < 0.05$) in the representation between genes containing predicted ACEs (black bars) and genes not containing predicted ACEs (white bars) are shown. Statistical significance was assessed using χ^2 statistics with Bonferroni correction for multiple hypothesis testing. GO categories are ordered from right to left in order of increasingly significant bias towards genes containing predicted ACEs. Only one category (transport) was significantly biased towards genes without predicted ACEs.

Figure 4. ACEScan scores for internal exons of well-known alternatively spliced genes. Known alternative exons are indicated by asterisks; the known RNA edited exon of *GLUR-B* is indicated by the letter 'E'. The following known AS exons are illustrated: (**A**) Exons 7 (168 bp), 8 (57 bp) and 15 (54 bp) of the human amyloid beta protein precursor gene (*APP*, ENSG00000142192); (**B**) Exons 14 (115 bp) and 15 (249 bp) of the human glutamate receptor, AMPA 2 gene (*GLUR-B*, ENSG00000120251).

4.1.7 References

1. Black, D. L. & Grabowski, P. J. (2003) *Prog Mol Subcell Biol* 31, 187-216.
2. Maniatis, T. & Tasic, B. (2002) *Nature* 418, 236-43.
3. Lopez, A. J. (1998) *Annu Rev Genet* 32, 279-305.
4. Black, D. L. (2003) *Annu Rev Biochem* 72, 291-336.
5. Black, D. L. (2000) *Cell* 103, 367-70.
6. Caceres, J. F. & Kornblihtt, A. R. (2002) *Trends Genet* 18, 186-93.
7. Musunuru, K. (2003) *Trends Cardiovasc Med* 13, 188-95.
8. Faustino, N. A. & Cooper, T. A. (2003) *Genes Dev* 17, 419-37.
9. Thanaraj, T. A., Clark, F. & Muilu, J. (2003) *Nucleic Acids Res* 31, 2544-52.
10. Sorek, R. & Ast, G. (2003) *Genome Res* 13, 1631-7.
11. Nurtdinov, R. N., Artamonova, II, Mironov, A. A. & Gelfand, M. S. (2003) *Hum Mol Genet* 12, 1313-20.
12. Bell, M. V., Cowper, A. E., Lefranc, M. P., Bell, J. I. & Screaton, G. R. (1998) *Mol Cell Biol* 18, 5930-41.
13. Ladd, A. N. & Cooper, T. A. (2002) *Genome Biol* 3, reviews0008.
14. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. & Shamir, R. (2004) *Genome Res* 14, 1617-23.
15. Kaufmann, D., Kenner, O., Nurnberg, P., Vogel, W. & Bartelt, B. (2004) *Eur J Hum Genet* 12, 139-49.
16. Sugnet, C. W., Kent, W.J., Ares JR, M., Haussler, D. (2004) in *Pacific symposium on biocomputing*, ed. Altman, R. B., Dunker, A.K, Hunter, L., Jung, T.A, Klein, T.E. (World Scientific, Hawaii).
17. Iida, K. & Akashi, H. (2000) *Gene* 261, 93-105.
18. Liu, H. X., Zhang, M. & Krainer, A. R. (1998) *Genes Dev* 12, 1998-2012.
19. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002) *Science* 297, 1007-13.
20. Rifkin, R., Yeo, G. & Poggio, T. (2003) in *Advances in Learning Theory: Methods, Model and Applications*, ed. Suykens, H., Basu, Micchelli, Vandewalle (IOS Press, Amsterdam), Vol. 190.
21. Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M. & Frazer, K. A. (2000) *Science* 288, 136-40.
22. Zheng, C. L., Nair, T. M., Gribskov, M., Kwon, Y. S., Li, H. R. & Fu, X. D. (2004) *Pac Symp Biocomput*, 78-88.
23. Nembaware, V., Wolfe, K. H., Bettoni, F., Kelso, J. & Seoighe, C. (2004) *FEBS Lett* 577, 233-8.
24. Xing, Y., Xu, Q. & Lee, C. (2003) *FEBS Lett* 555, 572-8.
25. Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S. & Sunyaev, S. (2003) *Trends Genet* 19, 124-8.
26. Cline, M. S., Shigeta, R., Wheeler, R. L., Siani-Rose, M. A., Kulp, D. & Loraine, A. E. (2004) *Pac Symp Biocomput*, 17-28.
27. Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. & Shoemaker, D. D. (2003) *Science* 302, 2141-4.

28. Ponte, P., Gonzalez-DeWhitt, P., Schilling, J., Miller, J., Hsu, D., Greenberg, B., Davis, K., Wallace, W., Lieberburg, I. & Fuller, F. (1988) *Nature* 331, 525-7.
29. Konig, G., Monning, U., Czech, C., Prior, R., Banati, R., Schreiter-Gasser, U., Bauer, J., Masters, C. L. & Beyreuther, K. (1992) *J Biol Chem* 267, 10804-9.
30. Cha, J. H., Kinsman, S. L. & Johnston, M. V. (1994) *Brain Res Mol Brain Res* 22, 323-8.
31. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. & Haussler, D. (2004) *Science* 304, 1321-5.
32. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. (2003) *Cell* 115, 787-98.
33. 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., Hampton, G. M., Schultz, P. G. & Hogenesch, J. B. (2002) *Proc Natl Acad Sci U S A* 99, 4465-70.
34. Philipps, D. L., Park, J. W. & Graveley, B. R. (2004) *Rna* 10, 1838-1844.
35. Dror, G., Sorek, R. & Shamir, R. (2004) *Bioinformatics*.
36. Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. (2004) *Mol Cell* 13, 91-100.

FIGURE 1.

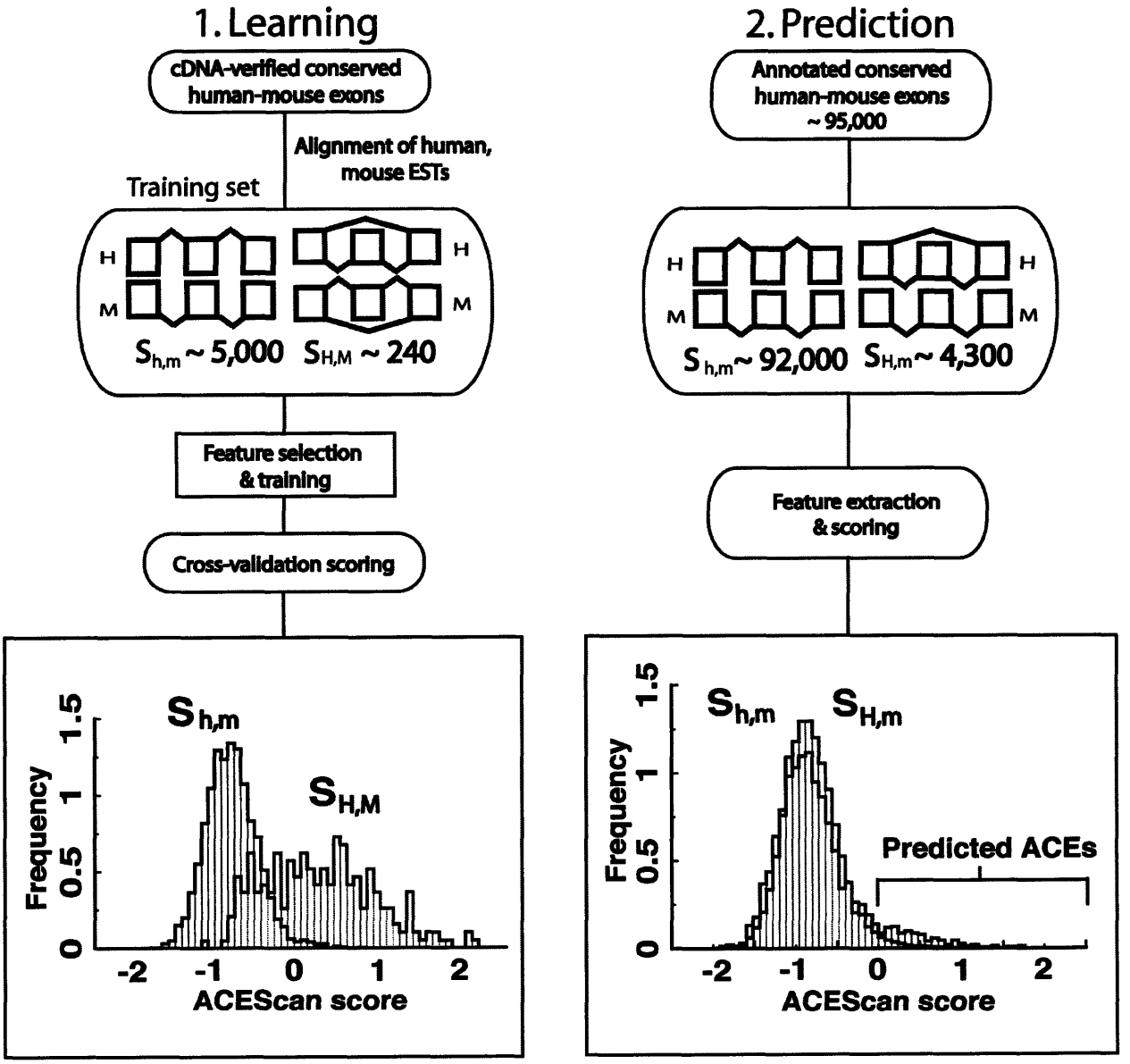
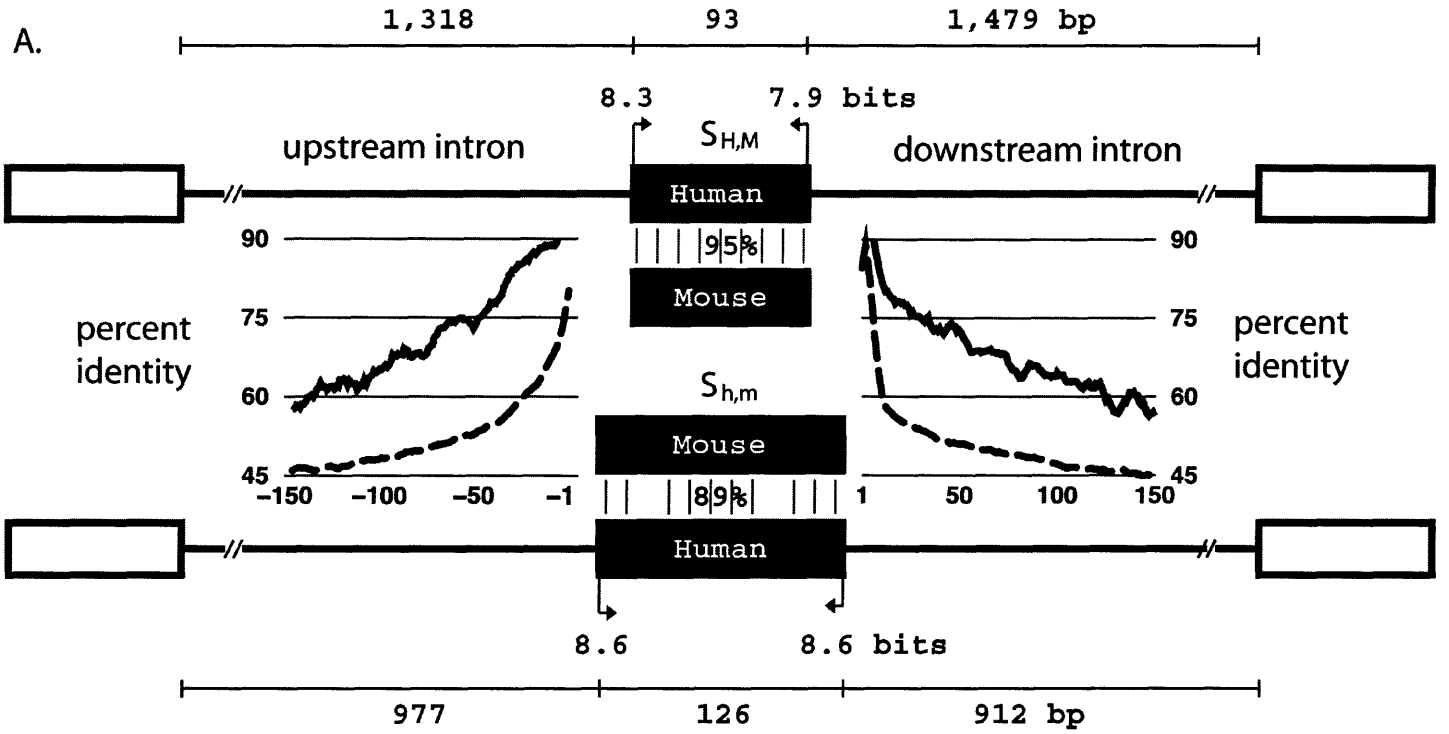

1. Learning
cDNA-verified conserved human-mouse exons
Alignment of human, mouse ESTs
Training set
H
M
H
M
$S_{h,m}$ ~ 5,000 $S_{H,M}$ ~ 240
Feature selection & training
Cross-validation scoring
$S_{h,m}$
$S_{H,M}$
Frequency
0
0.5
1
1.5
-2
-1
0
1
2
ACEScan score
2. Prediction
Annotated conserved human-mouse exons ~ 95,000
H
M
H
M
$S_{h,m}$ ~ 92,000 $S_{H,m}$ ~ 4,300
Feature extraction & scoring
$S_{h,m}$
$S_{H,m}$
Predicted ACEs
Frequency
0
0.5
1
1.5
-2
-1
0
1
2
ACEScan score

FIGURE 2.

A.



B.

Upstream intronic region

Unaligned	Aligned
GCCGC	UGCAU
UUUCC	GCAUG
UCUCU	CGGGG
UGUCU	ACACU
UUGUC	UUAAC
CUCUU	ACUAC
CUUUC	CCUAU
UCUUU	UCCAU
CGCCG	CAUAU
UCCUU	CUAUU
CCGCU	GUGAG
UGCAU	
CCUUU	
UUCUC	
UUUGC	
UUUUU	
GUUUC	
GGACA	
UGGAG	
GGCUG	
GACAG	

Exon region

Exon 5' end	Aligned	
CCUCC	UGUAG	UAGGG
CUCCC	GUAGU	CCUCG
CCCUC	ACUAG	CGGCG
CAAUC	UAGAA	UACGA
UCCCU	CCUCC	UCUCU
CCCCC	CGAAG	AGUAG
ACAUA	CCCGC	CCGCG
CGGCG	CUACG	AGUUA
CCCGC	UAACG	UUAAU
UAACC	UAAAC	UAGGA
UCUCU	UCUAG	ACUGA
CAGGG	UAGUG	UCCUU
CGUGG	UAACC	GAAAG
	CGCGG	CUCGC
	CGCCC	UAAUU
	UAGGC	CAAUC
	CACGA	CCGCU
	CGUAG	CCCCC
	CCCUA	GGCGG
	GUCGU	CUAGC
	UUACG	CUCUC
	AUCAA	ACCUG
	UGGAG	CUGGA
	CAUCA	GCUGG
		CUGGC

Downstream intronic region

Exon 3' end	Unaligned	Aligned
CCUCC	GCAUG	UGCAU
CCCUC	UGCAU	GCAUG
CAAUC	CAUGC	UCGCA
CUCCC	AUGCA	UGUGC
UCCCU	CUAAC	CAUUG
GUCCC	GUUUG	UGCCG
UCCUU	ACUAA	UCGCG
UCCCC	CACUA	CUAAC
GAAAG	CACUU	GUUUG
CCCGC	CAAAU	GUGAG
CGAAG	CAUGG	CAGGG
AAAAG	UAAGU	GGACA
UGAAA	UAAGA	AUGAU
UGGCC	UGAGU	GACAG
CAUCA	GUAAG	
	GUGAG	

FIGURE 3.

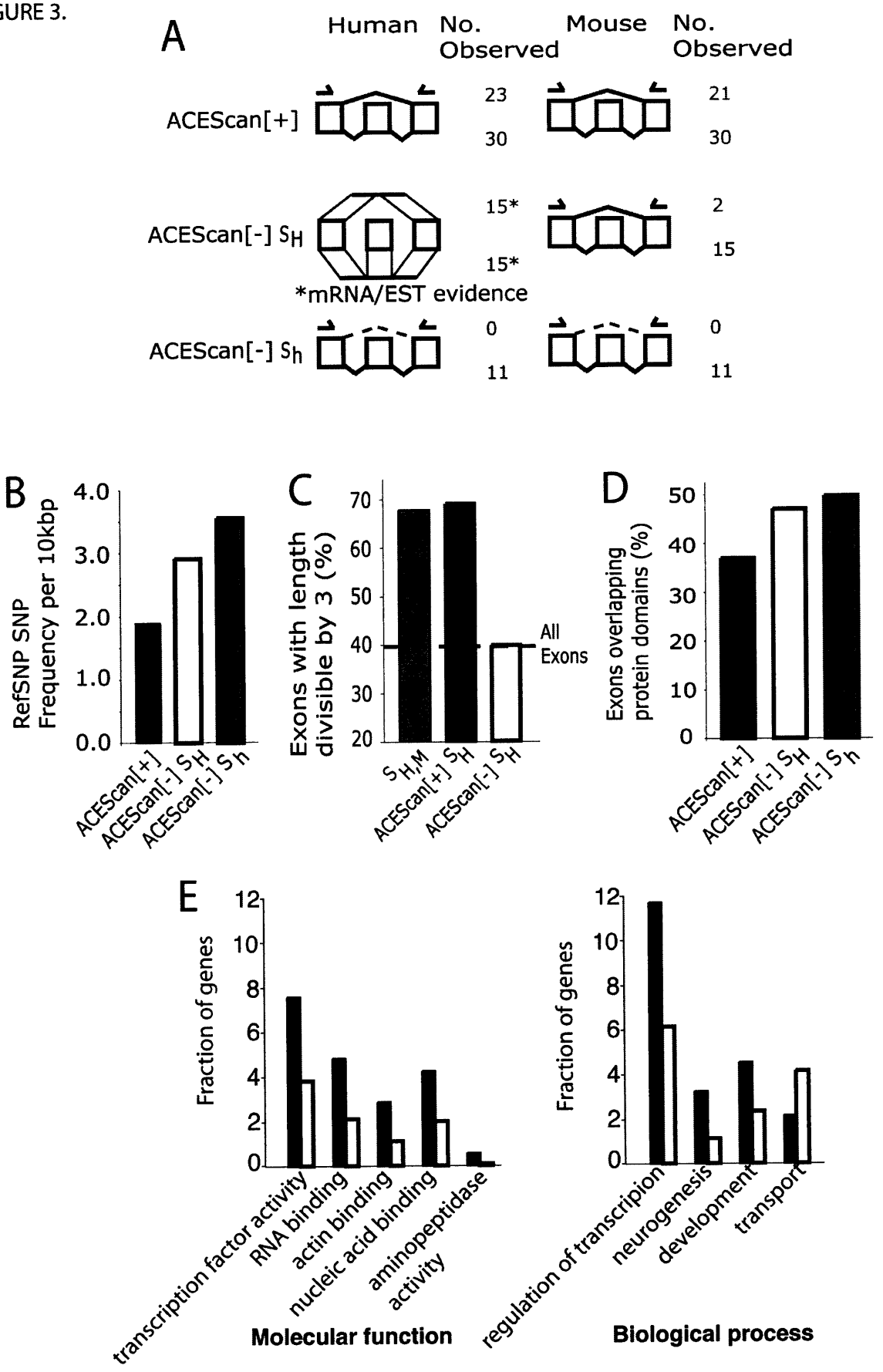
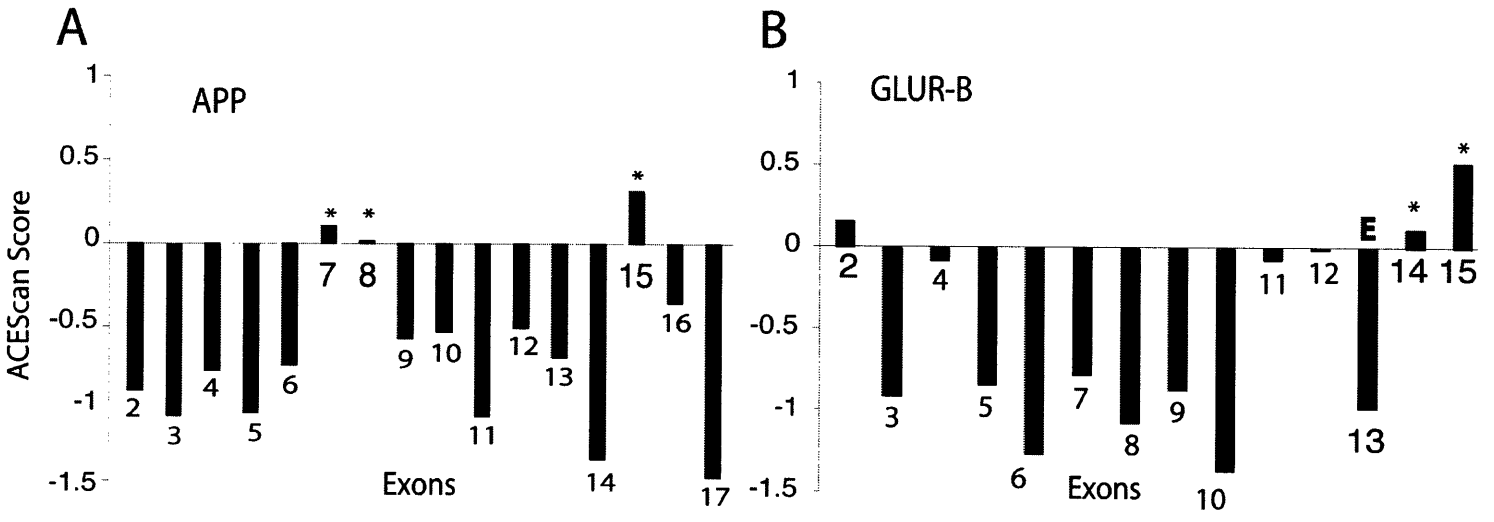

A
Human
No. Observed
Mouse
No. Observed
ACEScan[+]
23
30
21
30
ACEScan[-] SH
15*
15*
2
15
*mRNA/EST evidence
ACEScan[-] Sh
0
11
0
11
B
RefSNP SNP Frequency per 10kbp
4.0
3.0
2.0
1.0
0.0
ACEScan[+]
ACEScan[-] SH
ACEScan[-] Sh
C
Exons with length divisible by 3 (%)
70
60
50
40
30
20
All Exons
SH,M
ACEScan[+] SH
ACEScan[-] SH
D
Exons overlapping protein domains (%)
50
40
30
20
10
0
ACEScan[+]
ACEScan[-] SH
ACEScan[-] Sh
E
Fraction of genes
12
10
8
6
4
2
0
transcription factor activity
RNA binding
actin binding
nucleic acid binding
aminopeptidase activity
Molecular function
regulation of transcripion
neurogenesis
development
transport
Biological process

FIGURE 4.


A
APP
ACEScan Score
1
0.5
0
-0.5
-1
-1.5
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Exons
B
GLUR-B
1
0.5
0
-0.5
-1
-1.5
2
3
4
5
6
7
8
9
10
11
12
E
13
14
15
Exons

4.1.9 Supporting Information: Supporting Text, Figures and Tables

Sequence datasets

Chromosome assemblies of the human genome (hg13) and the mouse genome (mm3) were obtained from the UCSC Genome Browser, http://genome.ucsc.edu. Transcript data 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 expressed sequence tags (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, using the spliced alignment algorithm, MRNAVSGEN (http://genes.mit.edu/genoa/). Subsequently, ESTs were aligned to cDNA-verified genomic regions using SIM4 (2). For inclusion in the final GENOA annotation, all ESTs were required to overlap one or more cDNAs, and both 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 over 90% of the sequence length and have greater than 90% sequence identity.

Overall, GENOA aligned ~86,000 human cDNAs and ~890,000 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, as they possess different splicing characteristics as well as 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 both 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, 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 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 shorter than 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 nucleotides 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 using a χ^2-statistic for a 2 x 2 contingency table, with Yates correction factor{Glantz, 1997 #1300}. For each region in $S_{H,M}$ and $S_{h,m}$ (rows 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 5-mers and some 4-mers (Table 1), but no 3-mers or 6-mers.

Known *cis*-elements in high-ranking oligonucleotides

The motifs UGCAU and GCAUG were found to be over-represented in both the upstream and downstream introns flanking exons subjected to conserved skipping (Fig. 2B). Similar sequences, e.g., the hexamer UGCAUG, are known to be involved in regulation of splicing of the *c-src*, fibronectin, nonmuscle myosin heavy chain and calcitonin genes (6-9). The UCUCU pentamer, which is similar to sequences involved in splicing repression in the neural-specific N1 exon of the *c-src* transcript (10), was also identified as over-represented in the introns upstream of $S_{H,M}$ exons and in the exons themselves. A number of other U-rich sequences were also over-represented in upstream introns, consistent with previous observations (11, 12). The sequence UAGGG, which forms a portion of the consensus hnRNP A1 binding site, and can act in negative regulation of splicing (13), was also over-represented in $S_{H,M}$ exons relative to unskipped exons. Motifs related to GUAGU, also over-represented in $S_{H,M}$, have been validated as exonic splicing silencers (ESSs) in cultured human cells (Z. Wang, C. B. B., *Cell* in press). On the other hand, two pentamers which were under-represented in $S_{H,M}$ relative to $S_{h,m}$, CUGGA and AGAAG, resemble consensus ESEs (UGGA and GAGAAG, respectively) identified in previous analyses (14, 15). 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 (14) as compared to oligonucleotides enriched in $S_{H,M}$ (Table 2).

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 (16). Due to the unbalanced size of the two sets, i.e., there were about 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

$$\min_f (1/L) \Sigma_{i=1:L} w_i (y_i - f(x_i))^2 + \lambda \|f\|^2_K \qquad (1)$$

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 over-fitting and w_i is a misclassification penalty, set to β if sample x_i had a positive label (y_i=1), 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 was set to 5, higher than the value for negatively labeled data. Assuming a solution f^* of the form

$$f^*(\boldsymbol{u}) = \Sigma_{i=1:L} c_i K(\boldsymbol{u}, \boldsymbol{x_i}) \qquad (2)$$

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

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

Fixing λ and β and solving for **c** using the conjugate gradient method (implemented in Matlab), test examples were assigned an output according to Eq. (2). In order to solve Eq. (3) efficiently, $\underline{K}$ was expressed as $\mathbf{AA}^{\mathbf{T}}$, where **A** was the L x d matrix of training examples with d features. By first computing $\boldsymbol{\alpha}=\mathbf{A}^{\mathbf{T}}\mathbf{c}$, the outputs for unlabeled ('test') examples were obtained by matrix multiplication of **B** and $\boldsymbol{\alpha}$, where **B** is the n x 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 λ=0.01 and β=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 ~ –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 utilized 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 (17). The set of ACEScan[+] exons will be made available on the web at [http://genes.mit.edu/acescan] at the time of publication.

Performance measures

Receiver Operating Characteristic (ROC) curve analysis (18) 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 analysis

Gene Ontology (GO) identifiers (IDs) for each Ensembl-annotated gene were obtained from EnsMart (release version 19.1). Organizational principles (molecular function, 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 (19). Adjusting for multiple hypothesis testing using Bonferroni correction {Glantz, 1997 #1300}(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 (20) 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 lower than 20 were standardized to 20, as described (20). Genes expressed in a tissue or cell-line at greater than 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 using a χ^2 test, with the Yates correction factor {Glantz, 1997 #1300}. Adjusting for multiple hypothesis testing using Bonferroni correction, enriched tissues were again identified at a significance cutoff of $P<0.05$.

Single nucleotide polymorphism (SNP) analysis

8,408 high-quality reference SNPs (an 11-mer with 11-mer flanks on both sides) were obtained (21) and mapped 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 (22) 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 to 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 if high-scoring ACEScan exons overlapped exons at a lower or higher rate compared to low-scoring ACEScan exons that were designated as skipping (and non-skipping) events using transcript alignments from ESTs.

Experimental validation

The Invitrogen Superscript III First-Strand synthesis system for RT-PCR (Cat. No. 18080-051) was used to generate cDNAs from 3-4 ug 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 (BD Biosciences) using oligo(dT) primers. The Invitrogen Taq DNA polymerase kit (Cat. No. 18038-042) was utilized with primers designed using the primer3 program (23) targeted to exons flanking candidate ACEs. Forty cycles of PCR using an ABI 9700 thermocycler were conducted at denaturing temperature of 94° C for 30s, annealing at 58 ° C for 30 s, and elongation at 72° C for 30-100 seconds 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 using the QIAquick Gel Extraction Kit (Qiagen Cat. No. 28704) according to the manufacturer's instructions. Each isolated band was amplified by additional rounds of PCR with the same primers before sequencing.

Legends to Supporting Figures

Fig. 5. Performances of various models differing in the choice of features and the number of oligonucleotide features. A. The average area under the curve (AUC) values obtained from cross-validation with different models using varying numbers of top ranking oligonucleotide features are shown. B. Different models (***a*** to ***i***) are shown in the table, utilizing (denoted with a tick mark, otherwise a cross) different combinations of features, such as core features (5'ss, 3'ss scores, exon lengths, upstream and downstream flanking intron lengths), upstream and downstream intron alignment scores, exon alignment or similarity scores; and oligonucleotide features from aligned regions (exon alignment features, upstream and downstream intron alignment features), as well as unaligned regions (exon features, flanking intron features).

Fig. 6. ACEScan scores for all orthologous human-mouse exons identified by our automated procedure in example genes are shown. Known alternative exons are indicated by asterisks. The following known AS exons are illustrated: (**A**) Exons 14 (69 bp) and 26 (54 bp) of the human voltage-dependent T-type calcium channel alpha 1G subunit gene (*CACNA1G*, Ensembl gene identifier ENSG00000006283). Electrophysiological studies have shown that skipping of exon 26 of *CACNA1G*, which encodes α_{1g} (a human brain T Ca^{2+} channel α_1 subunit), affects the kinetics of deactivation and recovery from inactivation of the channel (24). This gene has three other cassette (skipped) exons, numbered 14, 34 and 35 (25), the latter two of which were not paired to annotated mouse exons by our automated exon orthology script, and therefore did not receive ACEScan scores. These latter examples remind us that incompleteness of annotation or orthology assignment places certain limits on the sensitivity of comparative methods like ACEScan. (**B**) Exon 11 (34 bp) of the human polypyrimidine tract-binding protein gene (*PTB*, ENSG00000011304). Skipping of this exon yields a premature stop codon in the downstream exon, generating a substrate for nonsense-mediated mRNA decay in an auto-regulatory negative feedback loop (26). This *PTB* exon provides an example of function for one of the ~32% of predicted ACEs which disrupt reading frame.

Fig. 7. The fraction of genes containing putative ACEs from $S_{H,M}$ used in training, the fraction of genes containing predicted ACEs, and the fraction of genes from all Ensembl-annotated genes annotated in the various Gene Ontology (GO) terms.

Fig. 8. Expression patterns of genes with predicted ACEs. Tissues exhibiting significant over- / under-representation of genes containing predicted ACEs ($P < 0.05$) in tissue-specifically expressed genes. Genes differentially expressed (greater than two-fold higher than median value across all tissues) in a HG-U95A microarray study (see main text) were considered tissue-specifically expressed. Tissues are ordered from right to left in order of increasingly significant bias towards genes containing predicted ACEs.

Table 1.

Rank	Upstream intron	Rank	Exon 5' end	Rank	Exon 3' end	Rank	Downstream intron	Rank	Aligned upstream intron	Rank	Aligned exon	Rank	Aligned downstream intron
12	tttc	11	cctcc	15	tccc	27	tgcat	10	tgcat	21	tgtag	2	gcatg
16	cttt	23	tccc	25	cctcc	80	gcatg	20	cgct	22	gtagt	6	tgcat
19	gccgc	30	cctc	39	ccctc	109	tcgca	28	gcat	26	actag	7	catg
47	tttt	46	ctccc	41	caatc	111	tccg	34	ccgc	31	tagaa	13	gcat
50	tttcc	61	ccctc	51	ctccc	136	cttt	37	gcatg	33	cctcc	17	catgc
63	ttgc	94	caatc	56	cctc	144	cgca	38	ctat	35	cgaag	18	tgca
65	tctct	119	tccct	67	ccct	165	tgtgc	43	ttgc	45	cccgc	24	ttgc
73	tctc	125	ctcc	74	aaag	168	cattg	48	cgtc	53	ctacg	29	acta
75	cgct	135	ccccc	96	tccct	171	tgccg	49	gtcg	55	taacg	58	atgca
78	ccgc	167	acata	110	gtccc	185	tcgcg	57	cgggg	64	taaac	60	ctaa
84	tccg	180	cggcg	113	tcctt	201	ccgc	59	acta	68	tctag	76	ctaac
97	tgtct	196	aaag	118	aggg	213	ctaac	72	caat	69	tagtg	77	gtttg
99	ttgtc	206	aggg	140	ctcc	217	gtttg	81	cggg	79	taacc	91	gccg
102	ctctt	207	cccc	152	tcccc	219	gacag	82	acgt	89	cgcgg	93	actaa
117	ctttc	215	cccgc	199	gaaa	184	ggaca	83	cacg	95	cgccc	122	ccaa
124	gccg	220	taacc	203	gaaag	194	atgat	86	atcc	101	taggc	123	cacta
126	tcttt	222	tctct	211	cccgc	198	agag	87	acact	103	cacga	139	cactt
132	tctt	223	cagggg	225	tagg	141	aggg	90	ttaac	106	cgtag	166	cggt
133	cgccg	212	tgat	234	cgaag	98	caggg	100	actac	128	cccta	175	caat
134	tcctt	230	cgtgg	236	aaaag	85	gagg	105	cctat	129	gtcgt	176	caaat
146	tcct	162	gacc	238	tgaaa	88	cagg	107	gccg	130	ttacg	205	catgg
148	ccgct	114	ctgg	172	catca	42	tgag	115	caac	137	taggg	221	atgc
149	tgcat			195	gtac	32	gtgag	120	cgtt	138	cctcg	228	cact
151	ttcc			155	tggcc			127	cgca	142	cggcg	150	agag
159	ccttt			70	ctgg			145	tccat	143	tacga	92	taagt
169	ctct							154	catat	153	tctct	40	taaga
182	ttctc							160	tgcg	158	agtag	36	gagt
187	ctat							161	ggcg	163	ccgcg	14	tgagt
200	tttg							178	cccg	164	agtta	9	taag
204	tttgc							186	ccga	170	ttaat	8	gtaa
224	ttttt							188	cgat	173	tagga	5	gtga
232	gtttc							190	catc	174	actga	4	tgag
208	ggac							214	ctatt	177	tccttt	3	gtaag
210	gagg							226	cgga	183	gaaag	1	gtgag
227	aggc							240	tccg	189	ctcgc		
229	gacag							237	cctg	191	taatt		
179	ggctg							239	ccag	193	caatc		
181	cagg							192	aaaa	197	ccgct		
156	tggag							147	cagg	209	ccccc		
121	ggaca							157	gagg	216	ggcgg		
131	gagc							116	gtgag	218	ctagc		
62	tggg							104	ctgg	231	ctctc		
44	ggag							66	tgag	233	ccgaa		
										235	atcaa		
										202	tggag		
										108	catca		
										112	acctg		
										71	ctgga		
										52	gctgg		
										54	ctggc		

Table 1. The highest ranked 240 oligonucleotide sequences utilized in ACEScan. These oligonucleotide features are ranked from most significantly differentially represented between $S_{H,M}$ and $S_{h,m}$ (1) to least significantly differentially represented (240) by the χ^2 statistic with Yates correction factor. Within each column, features are ordered by the level of enrichment in $S_{H,M}$ versus $S_{h,m}$, and represented according to pertinent sequence regions. Features in black (red) represent oligonucleotides enriched in $S_{H,M}$ ($S_{h,m}$).

Table 2.

Cutoff for enrichment	**3.84**	**5.02**	**6.64**
Fraction of oligonucleotides enriched in $S_{H,M}$ exons that overlap RESCUE-annotated ESE hexamers	0.36	0.39	0.39
Fraction of oligonucleotides enriched in $S_{h,m}$ exons that overlap RESCUE-annotated ESE hexamers	0.88	0.90	0.83
p-value	0	1.0E-13	1.0E-6

Table 2. Overlap of over and under-represented oligonucleotides in human $S_{H,M}$ versus $S_{h,m}$, and RESCUE-annotated ESEs. Oligonucleotides (4,5-mers) that were enriched in $S_{H,M}$ or $S_{h,m}$ at different cutoffs (3.84, 5.02, 6.64, corresponding to χ^2 values for different *p* value cutoffs) were considered to overlap a RESCUE-annotated ESE if a subsequence of length 4 or 5 bases of an oligonucleotide was an exact match to a continuous subsequence constructed from a RESCUE-annotated ESE hexamer (14). RESCUE-annotated ESEs are under-represented in the set of *k*-mers that are enriched in $S_{H,M}$, as opposed to the set that are enriched in $S_{h,m}$. The statistical significance was assessed by using the χ^2 test with Yates correction factor.

Table 3.

ID	Human Ensembl ID (ENSG 00000#)	Human Exon	Human Exon Length (bp)	Mouse Ensembl ID (ENSMUSG 00000#)	Mouse Exon	Gene Name	SEScan Score	Skipping in human tissues (RT-PCR)	Skipping in mouse tissues (RT-PCR)	Other evidence
S1	091129	24	36	020598	24	*NRCAM*	1.73	Ce	Br, Te, Emb	Ref. (27)
S2	078328	5	93	008658	5	*A2BP1*	1.67	-	Sk, Lu, Li, Te, Emb	Ref. (28)
S3	083312	20	76	009470	19	*TNPO1*	1.61	Br, Te, He, Ce, Sk, Bm, FBr	Br	
S4	079819	12	63	019978	12	*EPB41L2*	1.59	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Li, Sk, Te, Emb, He, Lu	
S5	137764	17	27	032364	16	*MAPKK 5*	1.48	-	Ki, Li	Human NM_002757.2 skips 17 & 18
S6	137764	18	33	032364	17	*MAPKK 5*	1.38	-	Ki, Li	Human NM_002757.2 skips 17 & 18
S7	175388	4	33	048320	4	*Q8N787*	1.26	Pr, Li, Lu, Ki, Sk, Bm, Te	Ki, Sk, Lu, Li, Te, Emb, He	
S8	156113	9	92	021780	2	*KCNMA1*	1.22	Te	Li, Lu	
S9	182872	4	74	031060	5	*RBM 10*	1.05	Li, Lu, Te	Br, Ki, Sk, Lu, Li, Te, Emb, He	
S10	144331	4	163	027016	3	*ZNF533*	1.03	Br, Pr, Te, He, Li, Ce, Sk, Bm, FLi, Ki, FBr	Br, Li	
S11	130558	4	220	026833	4	*OLFM1*	0.99	FBr	Br, Ki	
S12	155970	9	39	039478	8	NM_181723	0.95	Ce, FBr, Br	Br	
S13	172660	8	35	020680	8	*TAF15*	0.92	Br, Pr	Br, Ki, Sk, Lu	
S14	112062	9	80	024004	9	*MAPK14*	0.80	Pr, Li, Lu, Ki, Sk, Bm, Te	Sk, Li	
S15	169045	5	139	007850	5	*HNRNPH*	0.78	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Li, Te	
S16	079819	13	147	019978	13	*EPB41L2*	0.68	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Li, Sk, Te, Emb, He, Lu	
S17	114098	11	83	032468	13	NM_014154	0.67	Li, Lu, Ki	Sk, Lu	
S18	169057	2	125	031393	2	*MECP2*	0.36	-	Br, Li, Te, Ki, Lu, Bm,	Ref. (29)

									He	
S19	149970	9	147	025658	8	NM_014927	0.21	Br, Pr, Li, Lu, Ki, Sk, Bm, Te	Br, Ki, Te	
S20	122367	4	368	021798	5	*LDB3*	0.14	Te	Br, Ki, Li, Te, Emb, He	
S21	101977	5	117	031139	3	*MCF2*	0.10	Br	Br, Ki	
S22	060237	12	279	045962	Matched sequence in intron	*PRKWNK1*	0.26	Ki, Sk, Bm	Not Skipped	Mouse annotation unclear
S23	136531	3	92	026992	6	*Scn3a*	0.10	Pr, Te, Ki, FBr	Not Skipped	
S24	153944	2	41	034017	2	*MSI2*	1.75	Not Skipped	Not Skipped	
S25	121964	5	92	036890	9	NM_024659	1.30	Not Skipped	Not Skipped	
S26	155966	19	53	031189	5	*FMR2*	1.06	Not Skipped	Not Skipped	
S27	182197	3	108	038616	4	*EXT1*	0.69	Not Skipped	Not Skipped	mRNA AK130054 (Lu): Exons 2-6 skipped
S28	079739	7	116	025791	8	*PGM1*	0.36	Not Skipped	Not Skipped	
S29	164692	33	108	029661	37	*COL1A2*	0.14	Not Skipped	Not Skipped	
S30	117676	11	89	003644	11	*RPS6KA1*	0.14	Not Skipped	Not Skipped	
C1	069188	22	202	041592	16	*SDK2*	-0.27	Not Skipped	Not Skipped	
C2	073670	22	120	020926	22	*ADAM11*	-0.81	Not Skipped	Not Skipped	
C3	183773	6	97	022763	6	*NM_144704*	-1.41	Not Skipped	Not Skipped	
C4	143761	3	111	020440	3	*ARF1*	-1.62	Not Skipped	Not Skipped	
C5	146904	4	403	029859	5	*EPHA1*	-2.01	Not Skipped	Not Skipped	
C6	140859	6	112	031788	6	*KIFC3*	-0.99	Not Skipped	Not Skipped	
C7	018236	13	176	000107	14	*CNTN1*	-0.71	Not Skipped	Not Skipped	
C8	008405	12	104	020038	12	*CRY1*	-1.15	Not Skipped	Not Skipped	
C9	178035	8	91	006666	8	*IMPDH2*	-1.15	Not Skipped	Not Skipped	
C10	164070	7	245	025757	7	*OS94_HUMAN*	-1.66	Not Skipped	Not Skipped	
C11	130812	5	271	038742	6	*ANGPTL6*	-1.55	Not Skipped	Not Skipped	
E1	166164	10	108	031660	10	*BRD7*	-0.98	EST/mRNA evidence	Not Skipped	BX377621 (placenta cot 25-normalized) BI860548 (mammary adenocarcinoma cell

										line)
										AV747130 (Adult Pituitary)
E2	132849	32	84	028562	22	*NM_005799*	-0.66	EST/mRNA evidence	Not Skipped	mRNA AJ224748 (HeLa)
										Exon is not seen in any other mRNAs (5) or ESTs (2)
E3	100395	3	134	022394	3	*L3MBTL2*	-0.86	EST/mRNA evidence	Not Skipped	BU171558 (melanotic melanoma)
										Exon is included in all other mRNAs (8) and ESTs (>40)
E4	107897	3	121	026781	2	*ACBD5*	-0.73	EST/mRNA evidence	Br, Ki, Emb	ESTs (neuroblastoma, BM011542, BM011474); mRNAs (BC025309, neuroblastoma; AB082527, brain)
E5	120992	6	74	025903	6	*LYPLA1*	-0.65	EST/mRNA evidence	Not Skipped	BQ434220 (embryonic carcinoma), H04075 (placenta at birth), BE246387 (leukopheresis)
E6	076513	13	161	041870	13	*ANKRD13*	-1.24	EST/mRNA evidence	Not Skipped	AK095130 (substantia nigra)
E7	103876	12	102	030630	12	*FAH*	-1.26	EST/mRNA evidence	Not Skipped	S63549 (tyrosinemia patients), Exon included in 30 ESTs
E8	120137	3	177	018846	4	*PANK3*	-1.03	EST/mRNA evidence	Not Skipped	U46305 (Pancreatic cancer)
E9	124198	35	131	027682	35	*ARFGEF2*	-1.04	EST/mRNA evidence	Not Skipped	BU173319 (Retinablastoma)
E10	106443	5	160	029629	6	*PHF14*	-1.05	EST/mRNA evidence	Not Skipped	BF816107 (adenocarcinoma)
E11	170248	10	178	032504	10	*PDCD6IP*	-1.56	EST/mRNA evidence	Not Skipped	BQ961575 (leiomyosarcoma)
E12	107937	3	104	021149	3	*GTPBP4*	-1.58	EST/mRNA evidence	Not Skipped	AK097093 (spleen, skips exon but retains downstream intron)
E13	134899	12	145	026048	12	*ERCC5*	-1.59	EST/mRNA evidence	Not Skipped	AA191090 (hNT neuroteratocarcinoma neurons)
E14	150753	3	165	022234	3	*CCT5*	-1.83	EST/mRNA evidence	Not Skipped	AU143554 (retinoblastoma)
E15	132170	6	451	000440	7	*PPARG*	-1.54	EST/mRNA evidence	Br, Ki, Li, Emb, Te	BI524664 (NIH_MGC_122), CA426975 (subchondral bone)

Table 3. Candidate ACEs verified by RT-PCR experiments and subsequent sequencing, or by literature and transcript-based evidence. Samples S1 to S30 are candidate ACEs, samples C1 to C11 are low scoring ACEs (scored <0), and samples E1 to E15 are low scoring ACEs with human EST/mRNA evidence of skipping. (Abbreviations for tissues

are as follows. Ce: cerebellum, Br: whole brain, FBr: fetal brain, Sk: skeletal muscle, Li: liver, Lu: lung, He: heart, Bm: bone marrow, Te: testis, Pr: prostate, Ki: kidney, FLi: fetal liver, Emb: embryonic mix).

Table 4.

No.	MAASE Gene Name	MAASE ID	Ensembl mouse identifier (ENSMUSG0 00000#)	Exon #	Human exon length (bp)	Mouse exon length (bp)	ACEScan score
1	NOS1 - Nitric-oxide synthase, brain	267255	29361	15	102	102	0.12
2	MBP – Myelin basic protein	292857	41607	9	78	78	-0.41
3	MBP – Myelin basic protein	292857	41607	12	33	33	0.06
4	MBP – Myelin basic protein	292857	41607	13	120	123	-0.39
5	ACSL6 – Long-chain-fatty-acid--CoA ligase 6	325273	20333	13	78	78	0.10
6	ACSL6 – Long-chain-fatty-acid--CoA ligase 6	325273	20333	14	78	78	0.36
7	MCF2L - Guanine nucleotide exchange factor DBS	330913	31442	13	93	93	-0.39
8	MCF2L - Guanine nucleotide exchange factor DBS	330913	31442	14	74	75	-0.14
9	SPTAN1 - Spectrin alpha chain, brain	340288	57738	11	60	60	0.72
10	SPTAN1 - Spectrin alpha chain, brain	340288	57738	44	18	16	0.46
11	TEC – Tyrosine-protein kinase Tec	272000	29217	11	66	66	-0.61
12	TNNT2 - Troponin T, cardiac muscle isoforms	273074	26414	7	30	37	0.11
13	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	27	174	174	0.63
14	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	33	81	81	0.39
15	KCNMA1 - Calcium-activated potassium channel alpha subunit 1	275823	63142	39	29	29	1.45
16	CAST – Calpain inhibitor	287853	21585	9	66	57	-0.18
17	CAST – Calpain inhibitor	287853	21585	34	45	36	-0.33
18	DTNA - Dystrobrevin alpha	315011	24302	25	78	78	1.16
19	DTNA - Dystrobrevin alpha	315011	24302	33	21	21	1.58
20	DTNA - Dystrobrevin alpha	315011	24302	46	93	93	-0.38
21	MBNL2 - Muscleblind-like 2 Isoform 1	315631	22139	9	54	54	2.08
22	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	8	18	18	-0.1
23	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	9	21	15	0.22
24	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	24	41	41	0.54
25	TNNT3 - Troponin T, fast skeletal muscle isoforms	329096	61723	25	41	41	-0.42
26	TPM3 - Tropomyosin alpha 3 chain	272578	27940	9	76	76	0.38
27	TPM3 - Tropomyosin alpha 3 chain	272578	27940	10	76	76	0.16
28	TPM3 - Tropomyosin alpha 3 chain	272578	27940	14	79	83	-0.05
29	PHKA1 - Phosphorylase B kinase alpha regulatory chain, skeletal muscle	292503	34055	22	177	177	0.39

Table 4. ACEScan scores of mouse skipped exons from the MAASE database (30). MAASE identifier (ID), gene name, and exon sizes for the identified mouse and orthologous human exons. Exon pairs were then scored by ACEScan.

Table 5.

Gene Name	Human Ensembl ID	Human Exon No.	Exon length (bp)	Mouse Ensembl ID	Mouse Exon No.	Exon length (bp)	ACE-Scan Score	Evidence
HnRNP A1	135486	7	156	036021	3	159	0.81	Ref. (31)
HnRNP D0	138668	2	57	000568	2	57	0.21	Ref. (32)
HnRNP H	169045	5	139	007850	5	139	0.78	Table 3
HnRNP I/PTB	011304	11	34	006498	8	34	0.90	Ref. (26)
HnRNP K	165119	16	170	021546	14	170	0.33	-
HnRNP K	165119	6	44	021546	4	44	1.39	EST BI115223
HnRNP K	165119	8	72	021546	6	72	1.02	EST AK096385
HnRNP L	104824	6	73	015165	6	73	0.16	-
HnRNP M	099783	3	53	002291	3	53	1.00	4 Human, 3 Mouse ESTs
HnRNP R	125944	2	108	028666	3	119	0.28	-
HnRNP FUS	089280	6	35	030795	6	35	0.29	-

Table 5. Heterogeneous nuclear ribonucleoproteins with predicted ACEs. Proteins of the hnRNP family are involved with a host of important mRNA-related functions, such as nuclear export, subcellular localization, translation, and stability (33). In addition, some are known to regulate the splicing of their own or other transcripts (26, 31). The first column contains the Ensembl-identifier (last 6 digits) of the human and the orthologous mouse genes, the second column contains the exon number of the predicted exon in the longest transcript of the Ensembl-annotated genes, the third column contains the exon length, and the fourth column contains the corresponding ACEScan score and the last column contains literature or transcript-based evidence for exon skipping.

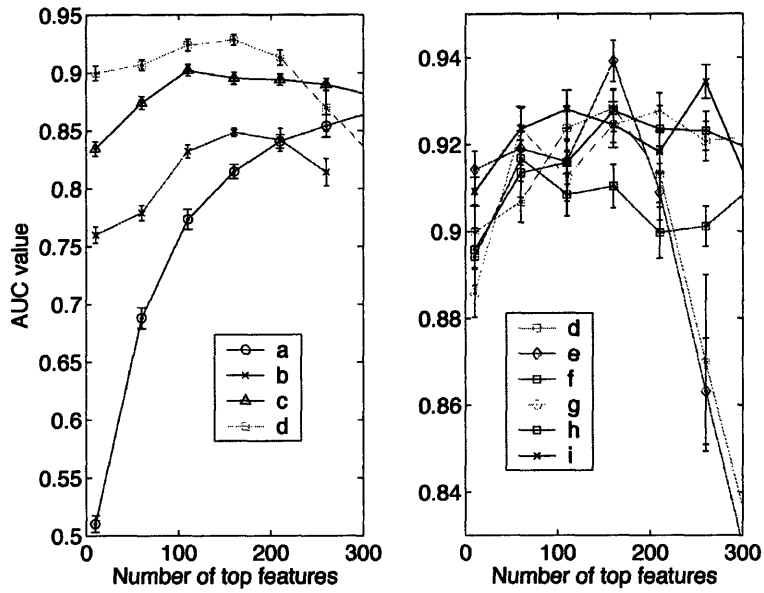
Table 6.

Differing Attributes	ACEScan	Sorek et al. (34)	Dror et al. (35)	Philipps et al. (36)
Goal	To identify conserved exon-skipping events in human and mouse	To identify skipped exons in human that are conserved in mouse	To identify skipped exons in human that are conserved in mouse	To identify various types of alternative splicing events in fruit fly
Computational				
General features	Exon size, intron size, exonic and intronic conservation	Exon size, exonic and intronic conservation. Length of best local alignments in 100 bases of flanking introns	Same as Sorek et al. (34)	Exonic and intronic conservation
Oligonucleotides	4- and 5-mers in aligned and unaligned regions in exons and 150 bases of flanking intronic regions	-	3-mers in exons and flanking 100 bases of unaligned introns.	-
Exon length divisible by 3	Not required	Required	Required	Not required
Splice Sites	MaxENT scores for 5'ss and 3'ss	-	5'ss information, polypyrimidine tract "intensity"	-
Type of score	Real-valued	Binary	Real-valued	Binary
Classifier	Regularized Least-Squares Classifier	Rule-based	Support Vector Machine	Rule-based
Experimental Validation of predicted exon-skipping events	Yes	Yes	-	Yes
Predictive accuracy by RT-PCR and sequencing	21/30 (70%) skipped in human and mouse	6/15 (40%) skipped in human only; 9/15 (60%) alternatively spliced (including alternative 5' and 3'ss) in human only	-	23/91 (~25%) alternatively spliced in *D. melanogaster;* 11/13 alternatively spliced in *D. pseudoobscura*
Negative Controls	0/11 (0%) ACEScan[-] exons skipped in human or mouse; 2/15 (13%) ACEScan[-] exons with human transcript evidence skipped in mouse	-	-	1/30 (3%) alternatively spliced in *D. melanogaster*
Tissues used	Normal (12 tissues in human and 8 tissues in mouse)	Normal and Tumor (14 tissues in human)	-	Normal (pooled embryo, larvae and adult)
Functional analysis at exon level	SNP density; reading-frame preservation;	-	-	-
Functional analysis at gene level	protein domain disruption; gene expression; gene functional characterization	-	-	-

Table 6. Comparison of ACEScan to other methods for predicting alternative splicing events. Different approaches are distinguished with respect to the features utilized in the computational approach, and the extent of experimental and computational validation performed.

Fig. 5.

A.



B.

Model index	A	b	c	d	e	f	G	h	i
Core features	x	√	√	√	√	√	√	√	√
Upstream intron alignment score	x	x	x	√	√	√	X	√	√
Downstream intron alignment score	x	x	x	√	√	√	√	x	√
Exon alignment score	x	x	√	x	√	x	√	√	√
Upstream intron alignment features	x	x	x	x	x	√	x	√	√
Downstream intron alignment features	x	x	x	x	x	√	√	x	√
Exon alignment features	x	x	√	x	x	x	√	√	√
Flanking intron features	√	√	x	√	√	x	x	x	√
Exon features	√	√	x	√	√	x	x	x	√

FIGURE 6.

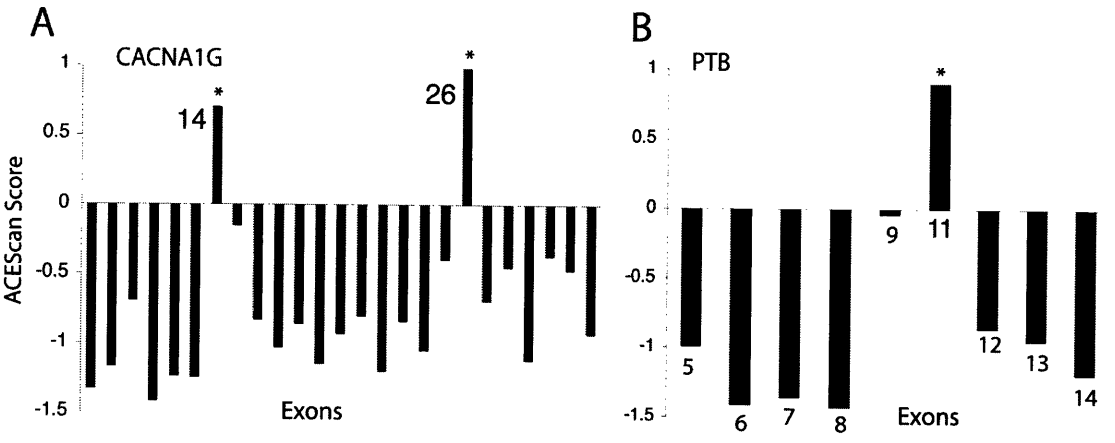

A
CACNA1G
ACEScan Score
1
0.5
0
-0.5
-1
-1.5
14
26
*
*
Exons
B
PTB
1
0.5
0
-0.5
-1
-1.5
*
5
6
7
8
9
11
12
13
14
Exons

Fig 7.

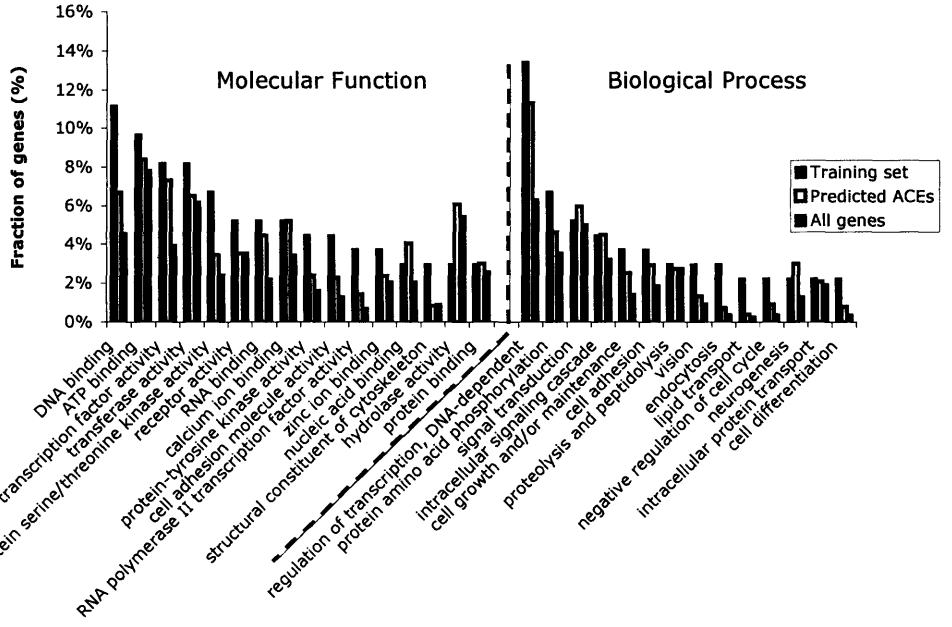
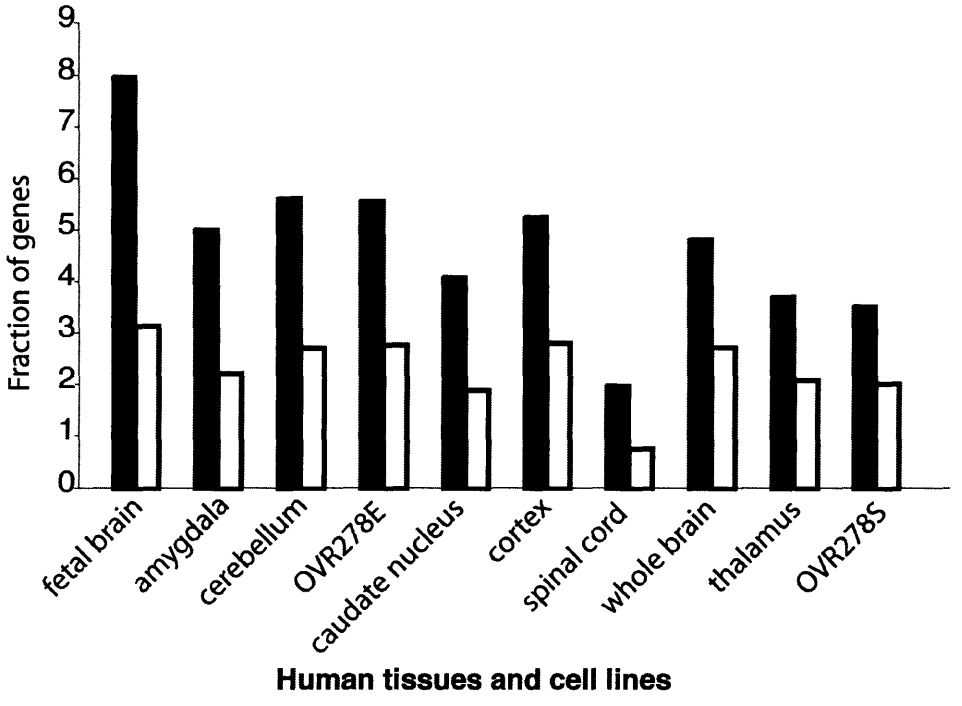

Fraction of genes (%)
16%
14%
12%
10%
8%
6%
4%
2%
0%
Molecular Function
Biological Process
Training set
Predicted ACEs
All genes
DNA binding
ATP binding
transcription factor activity
transferase activity
protein serine/threonine kinase activity
receptor activity
RNA binding
calcium ion binding
protein-tyrosine kinase activity
cell adhesion molecule activity
RNA polymerase II transcription factor activity
zinc ion binding
nucleic acid binding
structural constituent of cytoskeleton
hydrolase activity
protein binding
regulation of transcription, DNA-dependent
protein amino acid phosphorylation
intracellular signaling cascade
signal transduction
cell growth and/or maintenance
cell adhesion
proteolysis and peptidolysis
vision
endocytosis
lipid transport
negative regulation of cell cycle
neurogenesis
intracellular protein transport
cell differentiation

FIGURE 8.


Fraction of genes
9
8
7
6
5
4
3
2
1
0
fetal brain
amygdala
cerebellum
OVR278E
caudate nucleus
cortex
spinal cord
whole brain
thalamus
OVR278S
Human tissues and cell lines

References to Supplementary

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J Mol Biol* **215,** 403-10.
2. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. (1998) *Genome Res* **8,** 967-74.
3. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002) *Nucleic Acids Res* **30,** 38-41.
4. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res* **22,** 4673-80.
5. Yeo, G. & Burge, C. B. (2004) *J Comput Biol* **11,** 377-94.
6. Huh, G. S. & Hynes, R. O. (1994) *Genes Dev* **8,** 1561-74.
7. Modafferi, E. F. & Black, D. L. (1997) *Mol Cell Biol* **17,** 6537-45.
8. Black, D. L. (1992) *Cell* **69,** 795-807.
9. Brudno, M., Gelfand, M. S., Spengler, S., Zorn, M., Dubchak, I. & Conboy, J. G. (2001) *Nucleic Acids Res* **29,** 2338-48.
10. Chan, R. C. & Black, D. L. (1997) *Mol Cell Biol* **17,** 2970.
11. Shibata, A., Hattori, M., Suda, H. & Sakaki, Y. (1996) *Gene* **175,** 203-8.
12. Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P. & Valcarcel, J. (2000) *Mol Cell* **6,** 1089-98.
13. Kashima, T. & Manley, J. L. (2003) *Nat Genet* **34,** 460-3.
14. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002) *Science* **297,** 1007-13.
15. Liu, H. X., Zhang, M. & Krainer, A. R. (1998) *Genes Dev* **12,** 1998-2012.
16. Rifkin, R., Yeo, G. & Poggio, T. (2003) in *Advances in Learning Theory: Methods, Model and Applications*, ed. Suykens, H., Basu, Micchelli, Vandewalle (IOS Press, Amsterdam), Vol. 190.
17. Duda, R. O., Hart, P. E. & D.G., S. (2001) *Pattern Classification* (John Wiley & Sons, New York).
18. Swets, J. A. (1988) *Science* **240,** 1285-93.
19. Glantz, S. A. (1997) *Primer of Biostatistics* (McGraw-Hill, New York).
20. 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., Hampton, G. M., Schultz, P. G. & Hogenesch, J. B. (2002) *Proc Natl Acad Sci U S A* **99,** 4465-70.
21. Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. (2004) *PLoS Biol* **2,** E268.
22. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004) *Nucleic Acids Res* **32 Database issue,** D138-41.
23. Rozen, S. & Skaletsky, H. (2000) *Methods Mol Biol* **132,** 365-86.

24. Chemin, J., Monteil, A., Bourinet, E., Nargeot, J. & Lory, P. (2001) *Biophys J* **80,** 1238-50.
25. Mittman, S., Guo, J. & Agnew, W. S. (1999) *Neurosci Lett* **274,** 143-6.
26. Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. (2004) *Mol Cell* **13,** 91-100.
27. Wang, B., Williams, H., Du, J. S., Terrett, J. & Kenwrick, S. (1998) *Mol Cell Neurosci* **10,** 287-95.
28. Shibata, H., Huynh, D. P. & Pulst, S. M. (2000) *Hum Mol Genet* **9,** 1303-13.
29. Kriaucionis, S. & Bird, A. (2004) *Nucleic Acids Res* **32,** 1818-23.
30. Zheng, C. L., Nair, T. M., Gribskov, M., Kwon, Y. S., Li, H. R. & Fu, X. D. (2004) *Pac Symp Biocomput,* 78-88.
31. Blanchette, M. & Chabot, B. (1997) *Rna* **3,** 405-19.
32. Wagner, B. J., DeMaria, C. T., Sun, Y., Wilson, G. M. & Brewer, G. (1998) *Genomics* **48,** 195-202.
33. Dreyfuss, G., Kim, V. N. & Kataoka, N. (2002) *Nat Rev Mol Cell Biol* **3,** 195-205.
34. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. & Shamir, R. (2004) *Genome Res* **14,** 1617-23.
35. Dror, G., Sorek, R. & Shamir, R. (2004) *Bioinformatics.*
36. Philipps, D. L., Park, J. W. & Graveley, B. R. (2004) *Rna* **10,** 1838-1844.

4.2 Splicing silencing by combinations of UAGG and GGGG motifs

Before I begin this section, which describes a joint project with Kyoungha Han (KH) from Paula Grabowki's group at the Department of Biological Sciences, University of Pittsburgh/HHMI, it is important to credit the vast majority of the experimental portion of this work to KH. My main areas of contribution were mainly in the computational analysis of the use of the combination of motifs to predict exon-skipping events in human, and validation of the targets.

4.2.1 Abstract

Alternative pre-mRNA splicing is widely used to regulate gene expression by tuning the levels of tissue-specific mRNA isoforms. Yet, the molecular language that allows for intricate adjustments and the coordination of splicing patterns on a global scale is largely unknown. The CI cassette exon (exon 19) of the glutamate NMDA R1 receptor (*GRIN1*) transcript is used here as a model system to identify the sequence determinants for a brain region specific silencing mechanism. We identify a novel pattern of exonic UAGG and 5' splice site GGGG motifs that functions cooperatively to silence the CI cassette exon in mammalian cells. In this system, hnRNP A1 mediates silencing, whereas hnRNP H functions as an antagonist to silencing. This analysis was extended using bioinformatics to explore the wider role of the identified motif pattern in the human and mouse genomes. We find that, although uncommon, conserved patterns of UAGG and GGGG motifs serve generally as a predictive code to identify skipped exons genome-wide that otherwise bear no sequence relatedness. The identification of a similar arrangement of motifs in skipped exons of the hnRNP H family of splicing factors (*HNRPH1* and *HNRPH3* transcripts) has implications for their coordinate regulation at the level of splicing. These results provide a rationale to explain an essential feature of the tissue-specificity of the CI cassette exon - why this exon, which is equipped with strong splice sites, is prominently skipped in the hindbrain.

4.2.2 INTRODUCTION

Alternative pre-mRNA splicing is a major determinant of the protein functional diversity underlying human physiology, development and behavior (Lander et al. 2001). This process combines exonic sequences in various arrangements to generate two or more mRNA transcripts from a single gene. Splicing patterns are inherently flexible with variations observed in different cells, tissues and at different stages of development (Maniatis and Tasic 2002). Inducible changes in splicing pattern can also occur as a function of cell excitation in neuronal systems, T cell activation, heat shock, or cell cycle changes (Wang et al. 2001; Xie and Black 2001; Shin and Manley 2002; Shin et al. 2004). Thus, a central problem is to understand how the flexibility of splicing is controlled in different biological systems. A related issue is to understand how splicing errors, including alterations in splicing patterns, arise from inherited mutations or polymorphisms and contribute to human disease (Caceres and Kornblihtt 2002; Cartegni et al. 2002; Faustino and Cooper 2003).

Splicing decisions occur in the context of the spliceosome, a highly complex molecular machine containing the small nuclear ribonucleoprotein particles U1, U2 and U4/U5/U6 (snRNPs), and a host of protein factors (Makarov et al. 2002; Zhou et al. 2002; Jurica and Moore 2003). Spliceosome assembly occurs in a stepwise fashion to recognize the appropriate splice sites, to fashion the snRNP-based catalytic activity, and to couple the splicing process with transcription, 3' end formation, and nuclear export. Exon definition, or recognition of the exon as a unit, occurs early in spliceosome assembly, and its efficiency depends upon the strengths of the adjacent splice sites, as well as auxiliary splicing regulatory elements.

RNA control elements, which are distinct from the canonical splice sites, include the positive-acting exonic and intronic splicing enhancers (ESEs and ISEs), and the negative-acting exonic and intronic splicing silencers (ESSs and ISSs) (Cartegni et al. 2002; Fairbrother et al. 2002; Ladd and Cooper 2002). In order to achieve 100% inclusion of the exon in the processed mRNA, constitutive exons generally require some combination of ESEs in addition to the adjacent splice sites. Serine-arginine rich (SR) protein factors are important mediators of splicing enhancement in both constitutive and alternative splicing events. These

proteins recognize ESE motifs through their RNA binding domains, and recruit splicing factors via interactions with their RS domains (Tacke and Manley 1999; Blencowe 2000).

Alternative splicing affects the majority of human protein coding genes (Modrek et al. 2001; Johnson et al. 2003), but the molecular control mechanisms are poorly understood. Molecular dissection of a handful of prototypical alternatively spliced genes has shown that cassette exons are included at a frequency that depends on their complex arrangement of positive and negative RNA control elements. It is thought that combinatorial control, which involves the integrated actions of multiple RNA control elements and protein regulatory factors, is the basis of tissue specific patterns of splicing. Many protein factors of the SR protein and hnRNP protein families have been implicated in these mechanisms, and some of their expression patterns are tissue specific. The polypyrimidine tract binding protein (PTB/hnRNP I), for example, has important roles in mechanisms of negative control important for brain- and muscle-specific splicing events. Current evidence indicates that PTB/hnRNP I takes part in silencing by recognizing RNA elements containing UCUU and related motifs, and through protein oligomerization blocks recognition of the exon by the normal splicing machinery (Wagner and Garcia-Blanco 2001). The hnRNP A1 protein has also been implicated in a variety of cellular and viral splicing silencing mechanisms through its cooperative recognition of UAGGG[U/A] and related motifs (Chabot et al. 2003).

The CI cassette exon (exon 19) of the *GRIN1* transcript (NMDA-type glutamate receptor, NR1 subunit) is a valuable model to study mechanisms of regulation because of its striking patterns of tissue-specific splicing and developmental regulation in the rat brain (Wang and Grabowski 1996; Zhang et al. 2002). Note that the CI exon is referred to as E21 in these previous studies. The CI exon is prominently included in the forebrain, and prominently skipped in the hindbrain, but the control mechanisms underlying these patterns are poorly understood. The RNA binding protein NAPOR/CUGBP2 is thought to be a positive regulator of this exon since this factor promotes CI cassette exon inclusion in co-expression assays, and because its tissue-specific expression correlates with the spatial distribution of mRNA transcripts containing the CI exon in rat brain (Zhang et al. 2002). In mammals, NMDA-type glutamate receptors are assembled from *GRIN1* (NR1) and *GRIN2A* (NR2) subunits, where they play highly important roles impacting learning and memory

functions in the brain. Alternative splicing is used extensively for the generation of the brain-specific *GRIN1* transcripts, and CI exon inclusion affects the trafficking of NMDA receptors to the synapse (Ehlers et al. 1995; Mu et al. 2003).

In many cases tissue-specific exon inclusion is modulated by combinations of sequence motifs acting cooperatively or antagonistically to control splicing (Smith and Valcarcel 2000). An understanding of the essential ingredients for splicing silencing should allow skipped exons to be identified de novo from genomic sequence. Here molecular approaches are used to identify sequences responsible for silencing the CI cassette exon, and this analysis is extended using bioinformatics to explore the distribution of the identified motifs in the mammalian transcriptome. Paradoxically, the CI cassette exon undergoes predominant exon skipping in particular regions of the brain even though its adjacent splice sites match well to consensus patterns. In our previous study, a large portion of the downstream intron was shown to play a role in the silencing mechanism, but the factors involved in silencing were not defined (Zhang et al. 2002). Here we define a sensitive mechanism for silencing of the CI cassette exon that involves the interplay of exonic and 5' splice site motifs. We also show that different configurations of these silencing signals have predictive value for identification of skipped exons in the human and mouse genomes.

4.2.3 RESULTS

A 5' splice site GGGG and exonic UAGG motif are required in combination for silencing of a brain-region specific exon. The 5' splice site of the CI cassette exon is atypical due to the presence of an adjacent GGGG motif, which is conserved in human, rat and mouse *GRIN1* genes. GGGG motifs in the first 10 nucleotides of human introns are infrequent (see below). In the case of the CI cassette exon, the GGGG motif is immediately adjacent to the U1 snRNA complementary region of the 5' splice site, and the overall complementarity of the 5' splice site (6 base pairs) is in the normal range for mammals (6 to 7 base pairs), and includes all of the most highly conserved positions (-1 to +5).

The role of the GGGG motif in splicing silencing of the CI cassette exon was examined by generating site-directed mutations in nucleotides +6, +7, and +8 of the intron. These mutations were designed so as not to

disrupt the U1 snRNA complementary nucleotides, which include the last nucleotide of the CI exon and the first 5 nucleotides of the adjacent intron. Splicing assays involved transfecting splicing reporters into non-neuronal mouse myoblasts (C2C12 cells) followed by measurement of the levels of the exon-included and exon-skipped products by RT-PCR relative to the wild-type sequence.

Each mutation in the GGGG motif led to a dramatic increase in exon inclusion (Figure 1A). The strongest effects were observed when the ggg at +6 to +8 was converted to ccc (5m2) or aua (5m4), which resulted in approximately a four-fold increase in exon inclusion, compared to the wild-type sequence. Even a point mutation (5m9) resulted in a three-fold increase in exon inclusion. Thus, the GGGG motif plays an important role in the silencing mechanism. Additional sequence changes upstream and downstream of the GGGG motif had only modest effects on splicing. For example, mutations 5m1, 5m13, and 5m14 were designed to test potential RNA secondary structures involving the GGGG motif and complementary intron sequences. The modest changes in the splicing pattern resulting from these mutations, do not support a significant role in splicing for these putative structures.

Other than the GGGG motif at the 5' splice site, the sequence of this intronic region is devoid of guanosine rich sequences. Strikingly, introduction of a GGG at intron positions +40 to +42, (5m8) resulted in a 5-fold decrease in exon inclusion. Two additional mutations overlapping the 5m8 mutation that did not generate guanosine rich motifs had little or no effect on the splicing pattern (5m11 and 5m12). Thus, these results do not support an enhancing role for the original wild-type sequence, but imply that an additional guanosine-rich sequence can contribute to silencing.

The possibility that sequences within the CI cassette exon itself might contribute to the silencing mechanism was also explored. Either a scarcity of ESE sequences within the CI cassette exon might weaken exon definition, or the presence of exonic ESS sequences might enforce silencing. A model for the arrangement of ESE motifs in the CI cassette exon was based on the high affinity sequence recognition sites for the known SR splicing factors (Figure 1B, top). Mutations were then made in the ASF/SF2 (AGCCCGA, CACCCUG, CGUAGGU) and SC35 (CGACCCUA, GGCCUCCA, GUCCUCCA) motifs to test predictions

of this model anticipating that reduced exon inclusion should result from the disruption of functional ESE motifs.

The results of these experiments show that most of the mutations decreased exon inclusion consistent with ESE function (E1, E2, E3, E4, E5 and E6; Figure 1B). In contrast, a pair of double point mutations in a UAGG sequence beginning at position 93 of the exon generated a striking increase in exon inclusion, indicative of a silencing role for this sequence (E8 and E9; Figure 1B). Note that the overlapping ASF/SF2 motif is disrupted by the E9 mutation, but the E8 mutation generates a different ASF/SF2 motif. An additional six-nucleotide mutation (CAUCGU) that eliminates the ASF/SF2 motif at this position also resulted in a strong increase in exon inclusion (KH and PG, unpublished). These results show that the position 93 UAGG motif functions in C2C12 cells primarily as a silencer rather than as an ASF/SF2 motif. These results suggested the possible involvement of the splicing repressor hnRNP A1 based on the similarity of the UAGG motif to the hnRNP A1 high affinity binding sequence UAGGG[A/U] determined previously by SELEX experiments (Burd and Dreyfuss 1994).

A motif pattern for strong splicing silencing: analysis of quantity and position effects. The presence of two natural UAGG motifs in the CI cassette exon raised the question of how silencing might be affected by the number of exonic UAGGs. The number and position of UAGG motifs in the CI cassette exon were altered in the context of the wt0 splicing reporter (Figure 2). One set of mutations varied the position of the 5' splice site proximal UAGG by disrupting the original motif at position 93 of the exon, and by introducing a new UAGG motif at positions 11, 76, and 100 (splicing reporters E10, E11, E20). These position variations had surprisingly small effects on the pattern of splicing, and exon skipping predominated (Figure 2, lanes 1-4).

The effect of a single UAGG was examined at positions 11, 51, 76, 93 and 100 of the exon (splicing reporters E14, E8, E15, E13, and E21). The resulting splicing patterns showed predominant exon inclusion, within a range of 78 to 94% (lanes 5-10). Thus, even in the presence of an intact GGGG motif at the 5' splice site, the removal of one UAGG significantly reduces the rate of exon skipping, and again there is a modest effect of position.

From these experiments we conclude that splicing silencing in this context depends critically on the number of UAGG motifs in the exon, but less so on their relative position(s). Additional mutations were constructed to further test the hypothesis that the strength of splicing silencing is linked to the number of UAGGs in the exon. In one case, a third UAGG was introduced at position 11 of the exon (splicing reporter E18, lane 12), with the result that the level of exon inclusion decreased to 0.4%. Thus, the results of Figure 2 are consistent with a model in which multiple UAGG motifs function in a cooperative manner to modulate the splicing silencing of the CI cassette exon.

Exons lacking the two natural UAGG motifs in the presence and absence of the GGGG motif were also generated (splicing reporters, E17 and T8, respectively; lanes 11, 13). The resulting splicing pattern shows 83% exon inclusion in the presence of the GGGG motif, and this increases to ~100% inclusion when the motif is disrupted. These results show that the GGGG motif contributes to splicing silencing even in the absence of exonic UAGG motifs in agreement with our bioinformatics results below. Finally, the number of UAGG motifs was found to modulate the level of exon skipping in the absence of the GGGG motif (D0 and D8, lanes 14, 15). Thus, we conclude that strong splicing silencing of the CI cassette exon requires multiple UAGG motifs in the exon together with a GGGG motif adjacent to the 5' splice site.

Opposing roles of hnRNP A1 and H family proteins involved in silencing and anti-silencing effect of GGGG motif. What protein factors interact directly with the UAGG and GGGG motifs, and what are their roles in splicing silencing? GTP-labeled RNA substrates were subjected to UV crosslinking in HeLa nuclear extracts under in vitro splicing conditions. These experiments showed pronounced crosslinking to a protein doublet in the vicinity of 50 kDa for RNA substrates containing the intact GGGG motif (cs1 and 3h1, Figure 3A, lanes 1, 3). In contrast, a point mutation in the GGGG motif largely disrupts protein binding (cs3 and 3h3, lanes 2, 4). Because the apparent molecular weights of these proteins and the guanosine-rich binding specificity (Caputi and Zahler 2001) suggested hnRNP H/H' and F proteins, relevant antibodies were obtained for immunoprecipitation experiments. These results identify the bottom band of the doublet as hnRNP F (Figure 3A, lane 6), whereas the upper band corresponds to hnRNP H/H' (lane 8). Although the hnRNP F antibody is highly specific, the H/H' antibody crossreacts with hnRNP F, which is 95% identical to H/H' at

the protein sequence level. A control reaction (lane 10) shows the background level observed with preimmune serum.

Proteins that interact directly with the exonic UAGG motif were identified in a similar way, except that the RNA substrates contained a single radioactive label in the middle of the UAGG. Even with a single radioactive label, multiple proteins are observed to crosslink to the wild type substrate, wt3, under splicing conditions (Figure 3B, lane 4). To examine hnRNP A1 binding, the SELEX-derived consensus sequence, A1winner, was also tested in parallel. A low efficiency of UV crosslinking of hnRNP A1 has been observed previously (Burd and Dreyfuss 1994). The A1winner contains two UAGGGA sequences, and was found to crosslink predominantly to hnRNP H/H' and F, in comparison to A1 (lane 1 and data not shown). These results verify that A1 is immunoprecipitated as a ~35 kDa protein from the wt3 sample, as is the case for the A1winner (lanes 1-8). A control substrate, mt3, with a dinucleotide mutation in the UAGG showed little or no immunoprecipitation of crosslinked A1 (lanes 9-11). Thus, these results verify that hnRNP A1 is involved in the recognition of the exonic UAGG motif.

In order to establish the functional roles of hnRNPs F, H, and A1 in the silencing mechanism, each protein was co-expressed with splicing reporters containing the CI cassette exon, and effects on the splicing pattern were documented. For the wild-type splicing reporter containing an intact GGGG motif, overexpression of hnRNP F or H was found to enhance CI exon inclusion relative to the pcDNA control (Figure 3C, lanes 1-5). These effects were reduced, but not eliminated, in the presence of the 5m2 splicing reporter, which lacks the GGGG motif (lanes 6-10). Curiously, these results rule out a role in silencing of the CI exon for hnRNP F and H, and instead support an anti-silencing role for these factors.

Next we asked whether the silencing role of the GGGG motif is mediated through the effects of hnRNP A1, since the CI cassette 5' splice site is related to the A1 consensus binding motif (ACGgua**aggggaa** versus UAGGG[A/U]). These experiments also examined effects of the flanking introns, since our previous study demonstrated a role for the downstream intron in this silencing mechanism. Chimeric splicing reporters contained the CI cassette exon and various portions of the flanking introns inserted between exons 1 and 3 of the $GABA_A$ receptor g2 subunit (Figure 3D). When the complete downstream intron was present, co-

expression of hnRNP A1 decreased exon inclusion from 78.8% to 29.1%, nearly a 3-fold effect (Figure 3D, lanes 5, 6). The effect of hnRNP A1 depends upon the intact downstream intron, since the silencing effect was substantially reduced when most of the downstream intron was removed, (rGgCI-wt0 and rGgCI-up, lanes 1-4). The role of the GGGG motif was then examined in the context of the rGgCI-dn reporter by introducing mutations 5m2 and 5m4, which destroy the G cluster. The ability of hnRNP A1 to induce splicing silencing was reduced significantly by these mutations (rGgCI-dn5m2 and rGgCI-dn5m4, lanes 7-10). The observation that silencing is not completely abolished by the 5m2 and 5m4 mutations is consistent with the presence of multiple UAGG motifs in the exon and in regions of the downstream intron important for the hnRNP A1 effects.

Genome-wide analysis reveals association of silencing motifs with EST-confirmed exon skipping in human and mouse. We next wished to generalize this analysis by determining the extent to which the combination of UAGG and GGGG motifs is associated with exon skipping in the human and mouse genomes. This analysis involved computationally sorting a database of ~96,000 human and mouse orthologous exons into two datasets based on the presence or absence of the CI cassette motif pattern (Figure 4). Confirmed skipped exons were then mapped onto these datasets to determine the fraction of exon skipping in the sorted datasets. Confirmed skipped exons were identified by stringent cDNA and EST evidence (see Materials and Methods).

If the motif pattern indeed functions as a general splicing silencing signal, we would expect the frequency of exon skipping to be higher in the group of exons containing the UAGG and GGGG motif pattern, compared to those without. A perfect correlation with exon skipping was considered unlikely due to the abundance of ESE motifs, which in particular combinations might counteract the effects of silencing.

These results actually show that 18.8% of human exons of typical length ($\leq$ 250 base pairs) containing an exonic UAGG and 5' splice site GGGG arrangement are skipped exons, compared to 4.6% of exons in the database lacking these motifs (Figure 4). When exon length is not constrained the fraction of skipped exons with the motifs was slightly lower (15.8%), and still significant.

Variations of the CI cassette motif pattern were also analyzed for comparison. Notably this analysis showed that the occurrence of 5' splice site GGGG by itself is associated with exon skipping. That is, exons containing the GGGG motif in the first ten bases of the intron but entirely lacking UAGG and GGGG within the exon showed a higher rate of exon skipping (7.8%) compared to those without the GGGG intronic motif (4.6%). How critical is the proximity of the intronic GGGG to the 5' splice site? Comparing the fraction of exon skipping when the position of the intronic motif is moved slightly downstream to bases 11-20 of the intron rather than bases 1-10 reduced the fraction of skipped exons observed by approximately 2-fold. In these searches the exons identified were required to contain a conserved exonic UAGG in addition to the intronic GGGG motif. These data suggest that this type of silencing mechanism may involve direct competition with U1 or U6 snRNPs for binding to the 5' splice site.

Since UAGG is a known hnRNP A1 motif this analysis also searched for ≥ 1 UAGG in the exon and UAGG in the first ten bases of the adjacent downstream intron. The dataset containing this motif pattern showed a significant enrichment of confirmed skipped exons (7.0%) compared to those without (4.6%). Another pattern, ≥ 1 GGGG in the exon and UAGG in the first ten bases of the intron also showed enrichment for confirmed skipped exons (8.4%) compared to those without (4.6%). Searching more broadly for any combination of UAGG and/or GGGG motifs in the exon again showed enrichment for confirmed skipped exons when GGGG (6.9%) or UAGG (8.1%) was present in the first ten bases of the intron. Nonetheless the motif pattern most predictive of exon skipping is that originally identified for the CI cassette exon.

UAGG: under-represented in constitutive exons, over-represented in skipped exons. Under-representation of UAGG in constitutively spliced exons would be expected if this motif frequently plays a role in splicing silencing. For this analysis ~5000 known human cDNAs were downloaded from Ensembl (www.ensembl.org), and sequences that begin with AUG and end with UGA, UAG or UAA were shuffled 50 times using the program Codonshuffle. Condonshuffle randomizes the nucleotide sequence by swapping synonomous codons, preserving the encoded amino acid sequence, codon usage and base composition of the native mRNA (Katz and Burge 2003). Consequently, the program controls for constraints on the protein coding function of the mRNA, and for constraints on codon usage. It should be noted that in this analysis the

codon arrangements that allow for UAGG do not permit the UAG portion of the motif to be in-frame. Based on the codon shuffling analysis we observed a 1.5-fold reduced occurrence of UAGG in real coding sequences as compared to shuffled sequences. This effect is significant and indicates that for constitutive exons there is strong selection against UAGG sequences.

Next we asked if UAGG is indeed overrepresented in skipped human exons. As expected, internal UAGG and GGGG are significantly overrepresented in skipped exons as compared to constitutive exons in human ($c^2 = 436$ and 87, respectively; *P*-value $< 10^{-5}$). More rigorously, orthologous exons that are skipped in both human and mouse have a significant enrichment for UAGGC and UAGGG motifs that are conserved in sequence and position between mouse and human ($c^2 = 15$ and 13, respectively; *P*-value $< 10^{-4}$) compared to orthologous pairs of constitutive exons.

Identification of skipped exons with conserved UAGG and GGGG motif patterns across the human and mouse genomes. We next wished to identify exons unrelated to the CI cassette that might be silenced by a similar motif configuration. Consequently, we focused in more detail on the UAGG and GGGG motif pattern by searching for these motifs singly and in combination in the database of ~96,000 human and mouse orthologous exons. Exons containing a GGGG in the first 10 bases of the intron and one or more exonic UAGG(s) were identified in the human and mouse subsets of the database and as the intersection of these datasets. These data are presented as Venn diagrams, and specific examples selected from the Intersection dataset are shown to illustrate the pattern and conservation of the motifs (Human and Mouse subsets, and Intersection; Figure 5A).

As expected, the CI cassette exon of the *GRIN1* gene was found in all three of the overlap datasets. Of the 19 exons containing the motif pattern in the intersection dataset, 16 exons $\leq$ 250 bases in length were considered for further study based on the observation that skipping of longer exons is quite rare (Sorek et al. 2004). This dataset contained two well known splicing factors, hnRNP H1 and H3 (*HNRPH1* and *HNRPH3*). Although human hnRNP H1 contains 14 exons and H3 contains 10 exons, the UAGG and GGGG motif pattern was found associated with only one exon in these genes. As hnRNP H proteins are known to bind to

guanosine-rich sequences, the presence of a conserved GGGG motif in the 5' splice sites of these hnRNP H exons suggests the possible involvement of a splicing autoregulation.

The hnRNP H exons and additional candidates in the Intersection dataset (total of 12) were selected for experimental validation of exon skipping patterns by RT-PCR, and to investigate the tissue specificity of the splicing patterns in human tissues (Figure 5B and Table 1). The CI cassette exon was included in the analysis as a positive control (*GRIN1*). Skipping of the candidate exon for both the human hnRNP H1 and H3 genes was confirmed in several tissues. Candidate exons of *GRIPAP1*, *UTRN* and an uncharacterized hypothalamus protein gene were also confirmed to be skipped exons, and tissue-specific exon skipping was evident for *HNRPH1* exon 5, *HNRPH3* exon 3, *GRIPAP1* exon 2, *UTRN* exon 5. These tissue-specific patterns were not previously characterized. The results of Figure 5B were confirmed by DNA sequence analysis of the gel-purified products of the RT-PCR reactions. Although the candidate exon in the *ANXA8* gene was not experimentally validated in our analysis, EST and mRNA evidence confirms that the exon is skipped in cDNA libraries derived from choriocarcinomas (Table 1). The exon skipping pattern of *UTRN* exon 5 was of particular interest, since this pattern was clearly brain specific (UTRN_5 panel). Further analysis showed that this splicing pattern was uniform in forebrain and hindbrain regions, unlike the CI cassette exon (KH and PJG, unpublished). An important caveat of these experiments is that because our sampling of human tissues was not exhaustive, the true number of skipped exons could be significantly higher than the number confirmed by RT-PCR.

The mouse orthologs of hnRNP H1 exon 5 and hnRNP H3 exon 3 were chosen for further experimental confirmation of their exon skipping patterns (Figure 5C, panel: 1TAGG + GGGG exons). These splicing patterns were determined using RNA derived from mouse heart and brain tissue, as well as from the mouse C2C12 cell line. For each RNA sample, radioactive RT-PCR reactions were performed for a set of three serial dilutions of the input RNA. These serial dilutions show good consistency in the % exon inclusion values for each set of samples. Sequence alignments showed that the hnRNP H3 exon 3 of both human and mouse have an additional exonic GGGG motif not found in the orthologous hnRNP H1 exon 5 sequences (Figure 5C, bottom). Taken together with the results of Figure 4, which show the association of exonic GGGG motifs with

exon skipping, the presence of the additional GGGG motif is consistent with the higher rate of exon skipping observed for hnRNP H3 exon 3.

HnRNP H1 exon 8 and b-actin exon 2 served as control exons, since these exons do not contain UAGG or GGGG motifs (Figure 5C, panel: 0 TAGG, 0 GGGG exons). As expected, the 0 TAGG, 0 GGGG control exons were observed to be constitutive exons (100% exon inclusion).

The observation that a second UAGG is associated with increased strength of splicing silencing of the CI cassette exon prompted us to examine several exons with multiple UAGGs that were identified in the searches. From the dataset of 213 human exons containing UAGG and GGGG, 13 exons with ≥ 2 UAGGs were identified, and from the dataset of 200 mouse exons containing UAGG and GGGG, 12 exons with ≥ 2 UAGGs were identified (Table 1). Exons within these datasets that have lengths typical for internal coding exons (≤ 250 bases) were chosen for validation of their splicing patterns. RNA derived from mouse heart, brain and C2C12 cells confirmed the skipping patterns of *Hp1bp3* exon 2, *NCOA2* exon 13 and trace levels for *MEN1* exon 8 (Figure 5C). Additional cDNA evidence was found in the databases in support of these splicing patterns (Table 1). In the case of *Hp1bp3*, sequence alignments show that 2 TAGGs and the 5' splice site GGGG motif are conserved in the human and mouse orthologs, however, these exons are not included in the Intersection dataset because these motifs reside in the first human exon of the transcript. Sequence alignments for the more weakly skipped exons, *NCOA2* exon 13 and *MEN1* exon 8, show that one motif in the pattern is missing or imperfect in each set of orthologs (Figure 5C, bottom).

4.2.4 DISCUSSION

A sequence motif code for exon skipping. Here we use molecular approaches to define a novel pattern of UAGG and GGGG motifs required for silencing the *GRIN1* CI cassette exon, and show that skipped exons in the human and mouse genomes can be identified through bioinformatics searches that preserve the sequence and spatial configuration of the silencing motifs. We also illustrate, using the CI cassette model system, how the pattern of motifs determines the strength of exon silencing. While a single exonic UAGG or 5' splice site GGGG motif is associated with weak exon skipping (up to 22%), two or more UAGGs in the

exon together with the GGGG motif at the 5' splice site specifies predominant exon skipping (up to 96%). Bioinformatics searches show that the motif pattern is relatively uncommon, since only 0.2% of a large database of human and mouse exons (~96,000) harbor the UAGG and GGGG motifs in combination. Nonetheless, compared to exons lacking the motif pattern, a significantly higher frequency of exon skipping is associated with 16 exons ($\leq$ 250 nucleotides) in which the motif pattern is conserved in the human and mouse orthologs (Figure 4). An imperfect correlation between the presence of the motif pattern and confirmed exon skipping is not unexpected, since splicing enhancers may override the effects of UAGG and GGGG silencer motifs. This may be due not only to the arrangement of ESE and ISE motifs in and around a target exon, but also to tissue-specific variations in splicing factors.

Numerous ESE motifs have been functionally identified, but far less is known about sequence motifs that control silencing. Evidence for the role of hnRNP A1-regulated exonic UAGG motifs has been previously reported for a splicing silencing mechanism involving the K-SAM exon of human FGFR2 (Del Gatto et al. 1996), and related motifs have been reported in SMN2 exon 7 (UAGACA) (Kashima and Manley 2003), HIV Tat exon 2 (UAGACU) (Si et al. 1997; Bilodeau et al. 2001), CD44 exon v5 (UAGACA) (Matter et al. 2000), protein 4.1 exon 16 (Hou et al. 2002); c-src exon N1 (UAGGAGGAAGGU) (Rooke et al. 2003), and in the hnRNP A1 transcript itself (UAG, and UAGAGU) (Chabot et al. 1997; Chabot et al. 2003). Taken together with structural evidence that hnRNP A1 recognizes TAGG motifs directly (Ding et al. 1999), A1 is a likely mediator of these silencing events. Our computational analysis extends these previous studies by showing that the UAGG motif is significantly under-represented in constitutive exons and over-represented in skipped exons genome wide, consistent with its role as an exonic splicing silencer. Moreover, these results show that the GRIN1 CI cassette exon is subject to negative regulation by the co-expression of hnRNP A1. In contrast to the previous studies, however, the GGGG motif adjacent to the 5' splice site of the CI cassette exon is a novel and integral component of the silencing mechanism. GGGG motifs are notably absent from the 5' splice site regions of all of the previously studied exons.

The silencing role of the GGGG motif adjacent to the CI cassette exon contrasts with enhancing roles for guanosine rich intronic motifs as defined in other systems. The *c-src* transcript contains a complex intronic

enhancer downstream of the neuron-specific NI exon in which two GGGGG tracts are required for normal patterns of NI exon inclusion (Modafferi and Black 1997). G triplets are generally enriched in short mammalian introns (Lim and Burge 2001), and these sequences have been shown to enhance inclusion of an unusually small exon of cardiac troponin T (Carlo et al. 1996; Carlo et al. 2000), as well as additional exons of human alpha globin, and chicken b-tropomyosin (Sirand-Pugnet et al. 1995) transcripts. Moreover, a disease-related point mutation in an intronic guanosine cluster was recently found to disrupt the normal pattern of splicing of the human pyruvate dehydrogenase E1a transcript (Mine et al. 2003). Whereas the position of the GGGG motif is proximal to the U1 and U6 snRNA complementary regions of CI cassette 5' splice site, the guanosine rich intronic enhancers found in these previously studied transcripts are further downstream in the intron. Our computational analysis shows that the position of intronic GGGG motifs is generally important, since the frequency of exon skipping genome-wide is most closely associated with GGGG motifs when these are located in the first 10 nucleotides of the intron.

Implications for tissue-specific splicing of the CI cassette exon. A full understanding of CI cassette exon regulation will require explanations for the complex spatial and temporal variations observed in vivo. In a previous study, we presented evidence that NAPOR/CUGBP2 enhances CI exon inclusion in the rat forebrain. However, low levels of NAPOR/CUGBP2 cannot account for skipping of this exon in the cerebellum, since the CI cassette exon is inherently a strong exon. Initial support for this idea came from the observation that its splice sites match well to consensus sequences. Second, the experiments shown here demonstrate that the combination of exonic UAGG and 5' splice site GGGG motifs imposes silencing on an otherwise strong exon, since the exon can be converted to a constitutive exon in the absence of NAPOR/CUGBP2 when these motifs are destroyed by site-directed mutagenesis (splicing reporter T8; Figure 2). Thus, the mechanism described here provides a rationale for the tissue specific splicing patterns of the CI cassette exon in the brain (Figure 6). Our results clearly demonstrate that a wide range of control of CI cassette exon inclusion can be mediated by different arrangements of UAGG and GGGG motifs together with differential roles of hnRNP proteins. Such a model would allow for the intricate adjustments needed for

biological control of splicing in the nervous system. Notably, the UAGG and GGGG motifs are conserved in sequence and position in orthologous CI cassette exons of mouse, rat and human *GRIN1* genes.

The results shown here suggest antagonistic roles for hnRNP A1 and hnRNP H proteins in the regulation of CI cassette exon splicing. In this system, hnRNP A1 interacts directly with exonic UAGG, and its co-expression induces exon skipping, whereas hnRNP H and F interact with the intronic GGGG motif, and these factors induce exon inclusion. The downstream intron was previously shown to play a role in the silencing mechanism (Zhang et al. 2002), and this idea is reinforced in the present study. A provocative feature of this mechanism is the finding that the GGGG motif plays roles in both silencing and enhancement of the CI cassette exon. Although the intronic GGGG motif and hnRNP A1 are involved in silencing, hnRNP A1 does not crosslink directly to the GGGG motif (KH and PJG, unpublished). Thus, the GGGG motif may interact with a distinct protein factor, or it may play a structural role in the silencing mechanism. Interestingly, hnRNP H and F may principally function as anti-silencing factors in this mechanism by binding to the GGGG motif in a way that disrupts its normal silencing function. Although these proteins bind more weakly to UAGG motifs, they may have wider roles in counteracting the silencing function of the UAGG motifs. Previous studies have identified hnRNP F and H as factors that recognize guanosine-rich intronic enhancer motifs involved in the positive regulation of *c-src* N1 exon inclusion (Min et al. 1995; Chou et al. 1999). With these exceptions, a recent computational study of neuron-specific exons found that GGG motifs were generally lacking in the first 100 bases of the adjacent downstream intron (Brudno et al. 2001).

Overall, the results shown here are consistent with a model in which an extended silencing complex is assembled by the recognition of motifs in the exon, 5' splice site region and downstream intron. The involvement of hnRNP A1 in this mechanism is consistent with previous demonstrations of the cooperative binding of hnRNP A1 to pre-mRNA (Eperon et al. 2000; Zhu et al. 2001; Damgaard et al. 2002; Marchand et al. 2002). The ratio of hnRNP A1 transcripts to hnRNP F and H transcripts shows considerable variations in tissues in both human and mouse (Su et al. 2002), and we suggest that such variations may be involved in directing tissue specificity of exons that are regulated by UAGG and GGGG motifs.

Six ESE motifs within the CI cassette exon were also functionally identified in this study, and one of these, an ASF/SF2 motif, overlaps with the position 93 UAGG silencer (Figure 6). We observe that UAGG motifs are embedded in 32 ESE motifs reported in the ESEFinder database (Cartegni et al. 2003), but their effects on exon inclusion have not previously been reported. In the case of the CI cassette exon, such an arrangement of opposing splicing signals would predict that competition between ASF/SF2 and hnRNP A1 may provide additional options to fine-tune splicing patterns in different tissues or stages of development. This builds upon the well-established models for competition between hnRNP A1 and SR proteins in modulating 5' splice site selection (Fu et al. 1992; Mayeda and Krainer 1992; Mayeda et al. 1993; Caceres et al. 1994; Yang et al. 1994; Chabot et al. 2003).

Genome-wide analysis and implications for autoregulation of hnRNP H expression at the level of splicing. Since the CI cassette exon skipping pattern of the *GRIN1* transcript is brain-region specific, we wished to determine the characteristics of exons with a similar arrangement of these motifs in the human and mouse genomes. In one case, human utrophin exon 5 (*UTRN*), the exon-skipping pattern was found to be brain-specific. Other transcripts harboring skipped exons that were identified by bioinformatics searches, however, were found to be involved in a variety of cellular functions, such as RNA processing, chromatin structure/function, cell signaling and regulation of transcription. These include, hnRNP H1 and H3 (*HNRPH1; HNRPH3*), *GRIPAP1*, menin (*MEN1*), nuclear receptor co-activator 2 (*NCOA2*), heterochromatin protein 1 binding protein 3 (*Hp1bp3*), and an uncharacterized hypothalamus transcript (Table 1). Notably, a high proportion of the exon-skipping patterns identified were found to be tissue-specific.

It was surprising to find exon 5 of hnRNP H1 and exon 3 of hnRNP H3 in the list of exons with conserved UAGG and GGGG motifs, since hnRNP H proteins were found to crosslink specifically to the GGGG motif adjacent to the CI cassette exon. These exon-skipping patterns were confirmed by RT-PCR analysis in this study, and there is additional supporting cDNA and EST evidence in the databases. The RT-PCR analysis shows that these exon-skipping patterns are relatively weak, but this is consistent with a motif pattern containing a single exonic UAGG and 5' splice site GGGG motif. Skipping of exon 5 of hnRNP H1 or exon 3 of hnRNP H3 would result in a shift in the reading frame and introduction of a termination codon.

Thus, silencing of these exons at the level of splicing is expected to reduce protein expression due either to nonsense-mediated mRNA decay or premature termination of protein synthesis. The results shown here suggest a model in which hnRNP H protein expression can be positively autoregulated at the level of splicing. Under certain conditions, such a positive autoregulatory role for the hnRNP H proteins may provide a buffering effect against negative control by hnRNP A1. Autoregulation by a negative feedback loop was recently demonstrated for the splicing factor, PTB, which induces skipping of the eleventh exon of its cognate pre-mRNA (Wollerton et al. 2004). Similarly, hnRNP A1, SRp20, SC35, TIA1 and TIAR proteins are all involved in mechanisms that regulate the splicing patterns of their cognate transcripts (Blanchette and Chabot 1999; Sureau et al. 2001).

Prospects. If alternative splicing events are as prevalent as recent studies suggest (Modrek et al. 2001; Okazaki et al. 2002; Xu et al. 2002; Johnson et al. 2003), it will be important to understand on a global scale the biochemical language that determines tissue-specific patterns, and tunes these patterns in response to physiological stimuli (Grabowski 1998; Black 2000). Understanding a set of elements that cooperate in splicing silencing should allow the prediction of skipped exons from genomic sequence using bioinformatics searches for exons with a similar arrangement of elements. Here we identify UAGG and GGGG motifs that function in the silencing of the CI cassette exon, and which serve as patterns to recognize other skipped exons in the human and mouse genomes. With the exception of the hnRNP H1 and H3 exons, these groups of exons are otherwise unrelated in sequence outside of the UAGG and GGGG motifs. In the context of the CI cassette exon, multiple arrangements of these motifs are compatible with silencing function, and these functionally antagonize ESE motifs that are also shown here to regulate the CI cassette exon. Mechanisms of regulated exon skipping are well understood in only a handful of cases, and previous studies have not addressed the brain region-specific splicing switch that is characteristic of the CI cassette exon. Our results suggest that, in general, it might be a useful strategy to use motif pattern searches to identify co-regulated exons together with information about spatial constraints. The observation that UAGG and GGGG motif patterns are generally predictive of exon skipping may also have implications for interpreting the effects of mutations underlying

genetic disease. Future work will be needed to test the proposed autoregulatory functions of the hnRNP H proteins, and to more fully understand the complex biochemical language responsible for the regulation and coordination of splicing events genome wide.

4.2.5 MATERIALS AND METHODS

Plasmid construction and mutagenesis

All splicing reporter plasmids except for the experiments of Figure 3D were derived from the parent plasmid wt (previously called E21wt), in which the CI cassette exon is flanked by full-length introns and adjacent exons (Zhang et al. 2002). Site-directed mutations were introduced into the CI cassette exon or downstream intron using the QuikChange® Site-Directed Mutagenesis Kit (Stratagene), and mutations were confirmed by DNA sequencing. The splicing reporters wt and wt0 are identical except that wt has a point mutation at position 78 (C to G change) of the CI exon, which creates a XhoI site. Chimeric splicing reporters were derived from parent plasmid rGγ25 (Zhang et al. 1996), in which the CI cassette exon and 164 and 103 base pairs of the flanking introns (upstream and downstream, respectively) were introduced as a NotI-Bam HI fragment. The full-length upstream intron was introduced by replacing the XbaI-NotI fragment of of rGγCI-wto with the XbaI-NotI PCR product containing *GRIN1* exon 18 and 1092 base pairs of adjacent intron (plasmid, rGγCI-up). The full-length downstream intron was introduced by replacing the BamHI-EcoRI fragment of rGγCI-wto with the BamHI-EcoRI fragment containing *GRIN1* exon 20 and 1810 base pairs of adjacent upstream intron. All splicing reporter plasmids were constructed in a pBS vector followed by transfer into the vector, pBPSVPA+ vector (Nasim et al. 1990), in which expression is driven by the SV40 promoter. Expression plasmids for hnRNP proteins F, H, and A1 were generated by subcloning the complete open reading frames into the BamHI site of pcDNA4/HisMax vector (Invitrogen). Open reading frames were obtained from the following plasmids: hnRNP F from plasmid pFlag-F (Chou et al. 1999); hnRNP H/H' from pFlag-DSEF-1 (Bagga et al. 1998; Arhin et al. 2002); hnRNP A1 from plasmid Myc-A1 (Siomi and Dreyfuss 1995). All plasmid constructions were confirmed by DNA sequencing, and protein expression was verified by Western blot analysis.

Transient Expression and analysis of RNA splicing patterns

Cell growth, transfection and RT-PCR analysis was performed as described (Zhang et al. 2002). Briefly, transfections were performed in 60 mm plates at ~70% cell confluency using Lipofectamine. Transfections contained 3.5 μg of total plasmid DNA made up of splicing reporter plasmid with empty vector and/or protein expression plasmid at DNA ratios as specified. After 48 hours of expression, cells were harvested, and total RNA was purified, DNase treated, and ethanol precipitated. For analysis of splicing patterns, 1 μg of RNA was reverse transcribed with random primers, and $1/20^{th}$ of the reaction was then amplified for 20-24 PCR cycles in a 10 μl reaction containing 0.2 μM specific primers, 2 units Taq polymerase, 0.2 mM dNTPs, 1 μCi [a^{32}P]dCTP in reaction buffer. Primers used to amplify the CI cassette exon included and skipped mRNA products were specific for the flanking exons. Sequences in Ensembl were used to design primers for the experiments of Figure 4. Primer sequences are available upon request. For gel analysis, 25% v/v of each PCR reaction was resolved on 6% polyacrylamide, 5 M urea sequencing gels. Electrophoresis was performed for 1 hr at 30 W. Gel images and quantitation of results were obtained using a Fuji Medical Systems BAS-2500 phosphorimager and Science 2003 ImageGuage software.

Transcription and site-specific RNA labeling

Radioactive RNA substrates were prepared for UV crosslinking analysis as follows. RNAs containing the GGGG motif were prepared by in vitro transcription in 25 μl reactions containing T7 RNA polymerase, 0.4 mM each of ATP, UTP, CTP, and 0.3 mM GTP plus 25 μCi [a^{32}P]GTP, 0.5 mM GpppG and 0.1 μg DNA template in standard T7 reaction buffer. DNA templates were prepared by annealing complementary oligonucleotides with the top strand containing the T7 promoter sequence at its 5' end, followed by the RNA test sequence; bottom strands were complementary. RNAs were purifed after DNase treatment by Sephadex G25 chromatography, phenol extraction and ethanol precipitation. Site-specific labeling of RNA substrates containing the exonic UAGG motif was performed essentially as described (Moore and Query 2000). Transcription (non-radioactive) of the downstream RNA half was performed as above except that reactions

were larger (125 µl) and contained 2 mM guanosine instead of GpppG. After gel purification, the 5' end of the downstream half RNA was labeled by polynucleotide kinase with 25 pmol of the purified RNA and 25 pmol of [g^{32}P]ATP (6000 Ci/mmol). After removal of ATP by Sephadex G25 chromatography, the upstream and downstream RNA halves were annealed to a complementary DNA splint covering 16 bases on either side of the desired ligation position. Ligation reactions were performed in 10 µl reactions with 15 Weiss Units of T4 DNA ligase for 4 hours at 16°C, followed by DNase treatment and gel purification. The concentrations and integrity of the RNA preparations were verified by electrophoresis on 10% polyacrylamide/7M urea gels.

UV crosslinking and immunoprecipitation analysis

UV crosslinking reactions (12.5 µl) were performed under splicing conditions as described (Ashiya and Grabowski 1997) with 100,000 dpm radiolabeled RNA transcript and HeLa nuclear extract (4 mg/ml final concentration). Following UV treatment, samples were digested to completion with RNAse A (1 mg/ml, 20 min at 30°C), and held on ice for immunoprecipitation or SDS-PAGE analysis. For immunoprecipitation reactions, 25 µl of protein A beads (Sigma) were equilibrated in Buffer A (10 mM Tris/Hcl, pH 7.5, 100 mM NaCl, and 1% TritonX100), and antibody was bound to the beads for 1 hr on ice (5 µl of R7263 or R7264 for analysis of hnRNP F and H, respectively (Veraldi et al. 2001); or 1 µl of 9H10 for analysis of hnRNP A1). Equivalent amounts (w/v) of rabbit preimmune serum or purified mouse IgG were used for control reactions. Antibody beads were washed three times with Buffer A, and added to UV crosslinking reactions (25 µl) for 20 min on ice. Bound samples were washed four times with Buffer A, and centrifuged to separate pellet and supernatant. Each reaction component was boiled in SDS sample buffer, and resolved on discontinuous 12.5% polyacrylamide gels.

Generation of datasets and computational analysis

Human and mouse genes that were annotated as orthologs were obtained from Ensembl release 16 (www.ensembl.org). Human-mouse exons were aligned by BLAST (percent identity ≥85 and bit score ≥20), and genes were checked for consistency in terms of orthologous exon order. A total of ~94,000 conserved

human-mouse exons were retained for further analysis. In a separate analysis, ~14,600 internal exons in human were designated as skipped exons based on stringent alignments of cDNA and EST sequences to cDNA-verified genomic loci using the genome annotation script, GENOA (http://genes.mit.edu/genoa). Mapping these exons to the conserved human-mouse Ensembl set identified 4,455 skipped, internal human exons that are conserved in mouse. For the shuffling analysis, the first 30 bases and the last 60 bases of the original sequences were removed prior to shuffling to simulate removal of the first and last exons. Each sequence was shuffled 50 times using the Codon Shuffle program (Katz and Burge 2003). The fraction of occurrence of each oligonucleotide, e.g. UAGG, relative to the number of occurrences of all possible oligonucleotides of equal length, was compared to the fraction computed for the shuffled sets. The final fold under-representation was computed by taking the mean of the fractions computed over the shuffled sets, and dividing by the observed (true) fraction.

FIGURE LEGENDS

Figure 1. Exonic UAGG and 5' splice site GGGG motifs are required in combination for silencing of the CI cassette exon.

(A) A GGGG splicing silencer motif at the 5' splice site. Top: Sequence of the 5' splice site region (5' to 3') with exonic (uppercase) and intronic (lowercase) nucleotides is shown. Numbering is relative to the first nucleotide of the intron. Arrowhead, 5' splice site. A predicted SRp40 motif overlying the last seven bases of the exon is indicated. Engineered mutations and names of splicing reporters are indicated immediately below the affected nucleotides. Effect of mutations on the pattern of splicing is shown in a 5' to 3' arrangement (gel panel and graph). All splicing reporter plasmids have a three-exon structure in which CI is the middle exon (schematic: vertical bars, exons; horizontal lines, introns). Splicing reporter plasmids were expressed in vivo in mouse C2C12 cells, and splicing patterns assayed by radioactive RT-PCR of cellular RNA harvested from the cells. PCR primers are specific for the flanking exons. Results of multiple experiments are shown graphically as the average % exon included product (y axis) for each splicing reporter construct (x axis).

(B) Analysis of ESE motifs: an exonic UAGG splicing silencer motif overlaps an ASF/SF2 motif. Sequence of the CI exon (5' to 3') is shown, with engineered mutations (underscored) and names of splicing reporters indicated immediately below the affected nucleotides (bold). Nucleotide numbering is relative to the first nucleotide of the exon. Predicted ESE motifs for ASF/SF2 and SC35 are highlighted above the exonic sequence as indicated in brackets. The UAGG motif required for silencing (boxed) is indicated below the overlapping ASF/SF2 motif (asterisk). Effect of mutations on the in vivo pattern of splicing is shown in a 5' to 3' arrangement (gel panel and graph).

Figure 2. Effect of number and position of CI cassette exon splicing silencer motifs. Splicing reporters were constructed with variations in the number and position of UAGG and/or GGGG motifs. Three sets of schematics (boxed at left) illustrate the CI cassette exon and adjacent 5' splice site region with position(s) of exonic UAGG (vertical bars) and 5' splice site GGGG (vertical stripe) motifs. Splicing reporter names are

indicated at left. Vertical arrowhead, 5' splice site. Each splicing reporter was generated by site-directed mutagenesis from parent plasmid, wt0. Natural UAGG positions 51 and 93 represent the starting position of the motif relative to the first base of the exon. Engineered UAGG positions 11, 76, and 100 are also indicated. Sequence changes of the mutations are shown below. Representative splicing patterns are shown in gel panels at right. Values for the average % exon inclusion are shown for each splicing reporter.

Figure 3. Identification and functional roles of protein factors involved in the recognition of GGGG and UAGG motifs.

(A) Detection of protein binding to the 5' splice site GGGG motif by UV crosslinking in HeLa nuclear extract. Wild-type (cs1, 3h1) and mutant (cs3, 3h3) RNA substrates were internally labeled at guanosine nucleotides; mutations are underscored. Pattern of UV crosslinking is shown following RNase digestion and SDS-PAGE (lanes 1-4). Immunoprecipitation reactions (lanes 5-11) contained the 3h1 substrate together with antibody specific for hnRNP F or H/H'; control samples contained preimmune rabbit serum. The positions of hnRNP H/H' and F (arrowheads) and protein molecular weight standards (kDa) are indicated. The hnRNP F and H/H' antibodies were a gift of C. Milcarek.

(B) UV crosslinking of exonic position 93 UAGG motif in HeLa nuclear extract. RNA substrates were prepared with a single radiolabeled nucleotide as indicated by the asterisk; sequences are shown (bottom). The wild-type (wt3) and mutant (mt3) substrates are identical except for the underscored mutation. The A1winner substrate corresponds to the high affinity hnRNP A1 binding sequence previously identified by SELEX. Gel panel shows the pellet (P), supernatant (S), and input (I) of the immunoprecipitation reactions following SDS-PAGE. The position of hnRNP A1 is indicated (arrowhead). Monoclonal antibody, 9H10, was a gift of G. Dreyfuss.

(C) Exon inclusion is enhanced by co-expression of hnRNP F or H. Gel panel shows splicing pattern resulting from co-transfection of wild-type (wt) or mutant (5m2) splicing reporter with hnRNP F or H expression plasmid; splicing reporters are identical to those shown in Figure 1A. Control samples were transfected with empty vector; wedge indicates two levels (4 and 6 μg) of protein expression plasmid. Arrowhead, 5' splice

site. Immunoblot verification of transfected protein expression (bottom panel): nuclear extracts from transfected cells were separated by SDS-PAGE, transferred to nylon membrane, and developed with an antibody specific for the Xpress tag at the N-terminus of each pcDNA-protein sample. Graph shows fold effect on exon inclusion as calculated by normalizing the % exon inclusion value for each test sample to that of the pcDNA control (lane 1 for wt; lane 6 for 5m2). Raw % exon inclusion values are shown.

(D) Silencing effect of hnRNP A1 requires the intact 5' splice site GGGG motif and full-length downstream intron. Structures of chimeric splicing reporters are shown in which the CI cassette exon and intron flanks were introduced into an unrelated splicing reporter containing sequences from the $GABA_A$ receptor g2 transcript; rGgCI-wt0 (both introns truncated); -up (full-length upstream intron, truncated downstream intron); -dn (truncated upstream intron, full-length downstream intron). Numbers above indicate length of each intron segment in nucleotides. Arrowhead, 5' splice site. The splicing reporters rGgCI- dn5m2 and dn5m4 contain the full length downstream intron with 5' splice site mutations of Figure 1A. Gel panel shows splicing pattern resulting from co-transfection of splicing reporter with hnRNP A1 expression plasmid or vector control. Immunoblot verification of transfected protein expression (bottom panel), and graph showing fold effect on exon inclusion is as described in (C).

Figure 4. Computational analysis of UAGG and GGGG motif pattern in skipped exons of the human genome.

(Top) Schematic illustrates the computational sorting of a large dataset of human and mouse exons, based on the presence and absence of various combinations of splicing silencer motifs, followed by determination of the number of confirmed skipped exons in each of the sorted datasets. Whereas the exonic motif(s) was allowed at any position within the exon, the position of the 5' splice site motif was limited to the first 10 bases of the intron. Exon lengths were constrained to $\leq$ 250 nucleotides. (Bottom) Table shows for each motif pattern, the number of exons in the group (parentheses) and the percentage of confirmed skipped exons within that group (%). Searches were performed for the original motif combination found in the CI cassette exon and 5' splice site region, as well as for these individual motifs (motifs highlighted in bold). Reciprocal and mixed

combinations of these sequence motifs were also analyzed. (Middle) Graph illustrates the percentage of confirmed skipped exons for each motif pattern according to the data provided below. Exonic UAGG and 5' splice site GGGG motif pattern (§) and 5' splice site GGGG motif alone (◊).

Figure 5. Genome-wide analysis of UAGG and GGGG silencing motifs: identification of exons and validation of exon-skipping patterns.

(A) Bioinformatics searches. A database of 96,089 orthologous human and mouse exons was searched for TAGG located anywhere in the exon and GGGG in the first 10 bases of the intron. Venn diagrams indicate the number of exons with either or both sequence motifs in the human or mouse subsets of the database (Human and Mouse subset, respectively). The number of exons (19), in which UAGG and GGGG silencer motifs are conserved in orthologous human and mouse exons, is also shown (Intersection). The motif patterns are illustrated in the context of the exon (uppercase) and 5' splice site region (lowercase) for 12 examples from the Intersection dataset (human sequences are shown). Colon indicates 5' splice site. The conserved TAGG and GGGG motifs are highlighted in red to illustrate natural variations in their arrangements. Gene name (HUGO ID) and exon number within the gene are indicated at far right. In one case, the Genbank # is given for an uncharacterized transcript.

(B) RT-PCR confirmation of exon skipping patterns in human tissues. Twelve orthologous exons (≤ 250 nucleotides) were selected for experimental validation in a panel of 8 human tissues. These exons are derived from the Intersection dataset in which conserved TAGG and GGGG motifs are present in combination in the human and mouse orthologous exons. The 12 gel panels correspond to the examples listed at the bottom of part (A); additional cDNA and EST evidence for these skipping events are summarized in Table 1. Specific primer pairs were designed for each test exon to amplify the exon-included (<<) and exon skipped (<) products by RT-PCR. Gel panels show the products of reactions for one test exon resolved on agarose gels in the arrangement given in the inset. Gene name, exon number, and Ensembl number is provided above each gel panel. The far left and right lanes of each gel panel contain 1kb DNA molecular weight markers.

(C) RT-PCR analysis of splicing patterns in mouse tissues. Splicing patterns were determined by radioactive RT-PCR for selected mouse exons: Control reactions include b-actin exon 2 and hnRNP H1 exon 8, which were selected due to the absence of the silencing motifs in these exons (Panel: 0 TAGG, 0 GGGG exons). HnRNP H1 exon 5, and hnRNP H3 exon 3 are representative of the 1 TAGG and GGGG motif pattern (Panel: 1 TAGG + GGGG exons). *Hp1bp3* exon 2, *GRIN1* CI cassette exon, and *NCOA2* exon 13, and are examples of tissue-specific exon skipping associated with 2 TAGG and GGGG motif pattern (Panel: 2 TAGG + GGGG exons). *MEN1* exon 8 is also shown. Each gel panel shows splicing patterns tested in RNA samples from mouse heart and brain tissue and mouse C2C12 cells. Gene name, exon number, and Ensembl number are provided above each gel panel. Three Brackets represent the average % exon inclusion and standard deviation for each set of serial dilutions; raw values are given immediately below each lane. Sequence alignments of the corresponding human and mouse orthologs illustrate the patterns of silencer motifs.

Figure 6. GRINI CI cassette exon: splicing regulatory motifs and model for tissue-specificity.

(Top) Schematic of intron/exon structure and prominent splicing patterns observed in the forebrain (top) and hindbrain (bottom) of rat brain. (Bottom) Summary of splicing regulatory motifs functionally defined in this study is depicted on an expanded version of the CI cassette exon (yellow). ESE and hnRNP H/H' binding motifs are indicated above the exon. Also shown are nucleotides complementary to U1 snRNA and the interaction of the positive regulator NAPOR/CUGBP2 with the downstream intron [‡, as determined in (Zhang et al. 2002)]. UAGG and GGGG splicing silencing motifs defined in this study are highlighted in red. The working model for splicing silencing, based on the results shown here, proposes that exon skipping is mediated by multiple weak interactions of hnRNP A1 with two exonic UAGGs, and a 5' splice site GGGG motif. An intronic region downstream of the GGGG motif is also functionally involved, but the precise sequence motif was not defined. Interesting characteristics of this mechanism include the dual functionality of the GGGG motif at the 5' splice site, which functions primarily as a silencer in C2C12 cells, or as an anti-silencer in conjunction with high expression of hnRNP H or F protein. It is also of interest that the position 93 UAGG silencer is embedded in an ASF/SF2 motif. Tissue-specific exon inclusion is proposed to occur by

multiple pathways: by increased ratios of hnRNP H, hnRNP F or ASF/SF2 relative to hnRNP A1, by the expression of the positive regulator of CI exon inclusion, NAPOR/CUGBP2.

References

Arhin GK, Boots M, Bagga PS, Milcarek C, Wilusz J (2002) Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. Nucleic Acids Res 30(8): 1842-1850.

Ashiya M, Grabowski PJ (1997) A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. Rna 3(9): 996-1015.

Bagga PS, Arhin GK, Wilusz J (1998) DSEF-1 is a member of the hnRNP H family of RNA-binding proteins and stimulates pre-mRNA cleavage and polyadenylation in vitro. Nucleic Acids Res 26(23): 5343-5350.

Bilodeau PS, Domsic JK, Mayeda A, Krainer AR, Stoltzfus CM (2001) RNA splicing at human immunodeficiency virus type 1 3' splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element. J Virol 75(18): 8487-8497.

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103(3): 367-370.

Blanchette M, Chabot B (1999) Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. Embo J 18(7): 1939-1952.

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25(3): 106-110.

Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, et al. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res 29(11): 2338-2348.

Burd CG, Dreyfuss G (1994) RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. Embo J 13(5): 1197-1204.

Caceres JF, Kornblihtt AR (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet 18(4): 186-193.

Caceres JF, Stamm S, Helfman DM, Krainer AR (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science 265(5179): 1706-1709.

Caputi M, Zahler AM (2001) Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem 276(47): 43850-43859.

Carlo T, Sterner DA, Berget SM (1996) An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. Rna 2(4): 342-353.

Carlo T, Sierra R, Berget SM (2000) A 5' splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. Mol Cell Biol 20(11): 3988-3995.

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4): 285-298.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 31(13): 3568-3571.

Chabot B, Blanchette M, Lapierre I, La Branche H (1997) An intron element modulating 5' splice site selection in the hnRNP A1 pre-mRNA interacts with hnRNP A1. Mol Cell Biol 17(4): 1776-1786.

Chabot B, LeBel C, Hutchison S, Nasim FH, Simard MJ (2003) Heterogeneous nuclear ribonucleoprotein particle A/B proteins and the control of alternative splicing of the mammalian heterogeneous nuclear ribonucleoprotein particle A1 pre-mRNA. Prog Mol Subcell Biol 31: 59-88.
Chou MY, Rooke N, Turck CW, Black DL (1999) hnRNP H is a component of a splicing enhancer complex that activates a c-src alternative exon in neuronal cells. Mol Cell Biol 19(1): 69-77.
Damgaard CK, Tange TO, Kjems J (2002) hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. Rna 8(11): 1401-1415.
Del Gatto F, Gesnel MC, Breathnach R (1996) The exon sequence TAGG can inhibit splicing. Nucleic Acids Res 24(11): 2017-2021.
Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, et al. (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. Genes Dev 13(9): 1102-1115.
Ehlers MD, Tingley WG, Huganir RL (1995) Regulated subcellular distribution of the NR1 subunit of the NMDA receptor. Science 269(5231): 1734-1737.
Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, et al. (2000) Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. Mol Cell Biol 20(22): 8303-8318.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297(5583): 1007-1013.
Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17(4): 419-437.
Fu XD, Mayeda A, Maniatis T, Krainer AR (1992) General splicing factors SF2 and SC35 have equivalent activities in vitro, and both affect alternative 5' and 3' splice site selection. Proc Natl Acad Sci U S A 89(23): 11224-11228.
Grabowski PJ (1998) Splicing regulation in neurons: tinkering with cell-specific control. Cell 92(6): 709-712.
Hou VC, Lersch R, Gee SL, Ponthier JL, Lo AJ, et al. (2002) Decrease in hnRNP A/B expression during erythropoiesis mediates a pre-mRNA splicing switch. Embo J 21(22): 6195-6204.
Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302(5653): 2141-2144.
Jurica MS, Moore MJ (2003) Pre-mRNA splicing: awash in a sea of proteins. Mol Cell 12(1): 5-14.
Kashima T, Manley JL (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat Genet 34(4): 460-463.
Katz L, Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13(9): 2042-2051.
Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3(11): reviews0008.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.
Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A 98(20): 11193-11198.
Makarov EM, Makarova OV, Urlaub H, Gentzel M, Will CL, et al. (2002) Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome. Science 298(5601): 2205-2208.
Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418(6894): 236-243.
Marchand V, Mereau A, Jacquenet S, Thomas D, Mougin A, et al. (2002) A Janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. J Mol Biol 323(4): 629-652.
Matter N, Marx M, Weg-Remers S, Ponta H, Herrlich P, et al. (2000) Heterogeneous ribonucleoprotein A1 is part of an exon-specific splice-silencing complex controlled by oncogenic signaling pathways. J Biol Chem 275(45): 35353-35360.
Mayeda A, Krainer AR (1992) Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. Cell 68(2): 365-375.

Mayeda A, Helfman DM, Krainer AR (1993) Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF. Mol Cell Biol 13(5): 2993-3001.
Min H, Chan RC, Black DL (1995) The generally expressed hnRNP F is involved in a neural-specific pre-mRNA splicing event. Genes Dev 9(21): 2659-2671.
Mine M, Brivet M, Touati G, Grabowski P, Abitbol M, et al. (2003) Splicing error in E1alpha pyruvate dehydrogenase mRNA caused by novel intronic mutation responsible for lactic acidosis and mental retardation. J Biol Chem 278(14): 11768-11772.
Modafferi EF, Black DL (1997) A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. Mol Cell Biol 17(11): 6537-6545.
Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29(13): 2850-2859.
Moore MJ, Query CC (2000) Joining of RNAs by splinted ligation. Methods Enzymol 317: 109-123.
Mu Y, Otsuka T, Horton AC, Scott DB, Ehlers MD (2003) Activity-dependent mRNA splicing controls ER export and synaptic delivery of NMDA receptors. Neuron 40(3): 581-594.
Nasim FH, Spears PA, Hoffmann HM, Kuo HC, Grabowski PJ (1990) A Sequential splicing mechanism promotes selection of an optimal exon by repositioning a downstream 5' splice site in preprotachykinin pre-mRNA. Genes Dev 4(7): 1172-1184.
Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420(6915): 563-573.
Rooke N, Markovtsov V, Cagavi E, Black DL (2003) Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. Mol Cell Biol 23(6): 1874-1884.
Shin C, Manley JL (2002) The SR protein SRp38 represses splicing in M phase cells. Cell 111(3): 407-417.
Shin C, Feng Y, Manley JL (2004) Dephosphorylated SRp38 acts as a splicing repressor in response to heat shock. Nature 427(6974): 553-558.
Si Z, Amendt BA, Stoltzfus CM (1997) Splicing efficiency of human immunodeficiency virus type 1 tat RNA is determined by both a suboptimal 3' splice site and a 10 nucleotide exon splicing silencer element located within tat exon 2. Nucleic Acids Res 25(4): 861-867.
Siomi H, Dreyfuss G (1995) A nuclear localization domain in the hnRNP A1 protein. J Cell Biol 129(3): 551-560.
Sirand-Pugnet P, Durosay P, Brody E, Marie J (1995) An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. Nucleic Acids Res 23(17): 3501-3507.
Smith CW, Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 25(8): 381-388.
Sorek R, Shamir R, Ast G (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20(2): 68-71.
Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, et al. (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99(7): 4465-4470.
Sureau A, Gattoni R, Dooghe Y, Stevenin J, Soret J (2001) SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. Embo J 20(7): 1785-1796.
Tacke R, Manley JL (1999) Determinants of SR protein specificity. Curr Opin Cell Biol 11(3): 358-362.
Veraldi KL, Arhin GK, Martincic K, Chung-Ganster LH, Wilusz J, et al. (2001) hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. Mol Cell Biol 21(4): 1228-1238.
Wagner EJ, Garcia-Blanco MA (2001) Polypyrimidine tract binding protein antagonizes exon definition. Mol Cell Biol 21(10): 3281-3288.
Wang HY, Xu X, Ding JH, Bermingham JR, Jr., Fu XD (2001) SC35 plays a role in T cell development and alternative splicing of CD45. Mol Cell 7(2): 331-342.
Wang Z, Grabowski PJ (1996) Cell- and stage-specific splicing events resolved in specialized neurons of the rat cerebellum. Rna 2(12): 1241-1253.

Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CW (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell 13(1): 91-100.

Xie J, Black DL (2001) A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. Nature 410(6831): 936-939.

Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30(17): 3754-3766.

Yang X, Bani MR, Lu SJ, Rowan S, Ben-David Y, et al. (1994) The A1 and A1B proteins of heterogeneous nuclear ribonucleoparticles modulate 5' splice site selection in vivo. Proc Natl Acad Sci U S A 91(15): 6924-6928.

Zhang L, Ashiya M, Sherman TG, Grabowski PJ (1996) Essential nucleotides direct neuron-specific splicing of gamma 2 pre-mRNA. Rna 2(7): 682-698.

Zhang W, Liu H, Han K, Grabowski PJ (2002) Region-specific alternative splicing in the nervous system: implications for regulation by the RNA-binding protein NAPOR. Rna 8(5): 671-685.

Zhou Z, Licklider LJ, Gygi SP, Reed R (2002) Comprehensive proteomic analysis of the human spliceosome. Nature 419(6903): 182-185.

Zhu J, Mayeda A, Krainer AR (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol Cell 8(6): 1351-1361.

Table 1. Human and mouse orthologous exons containing TAGG and GGGG motif patterns

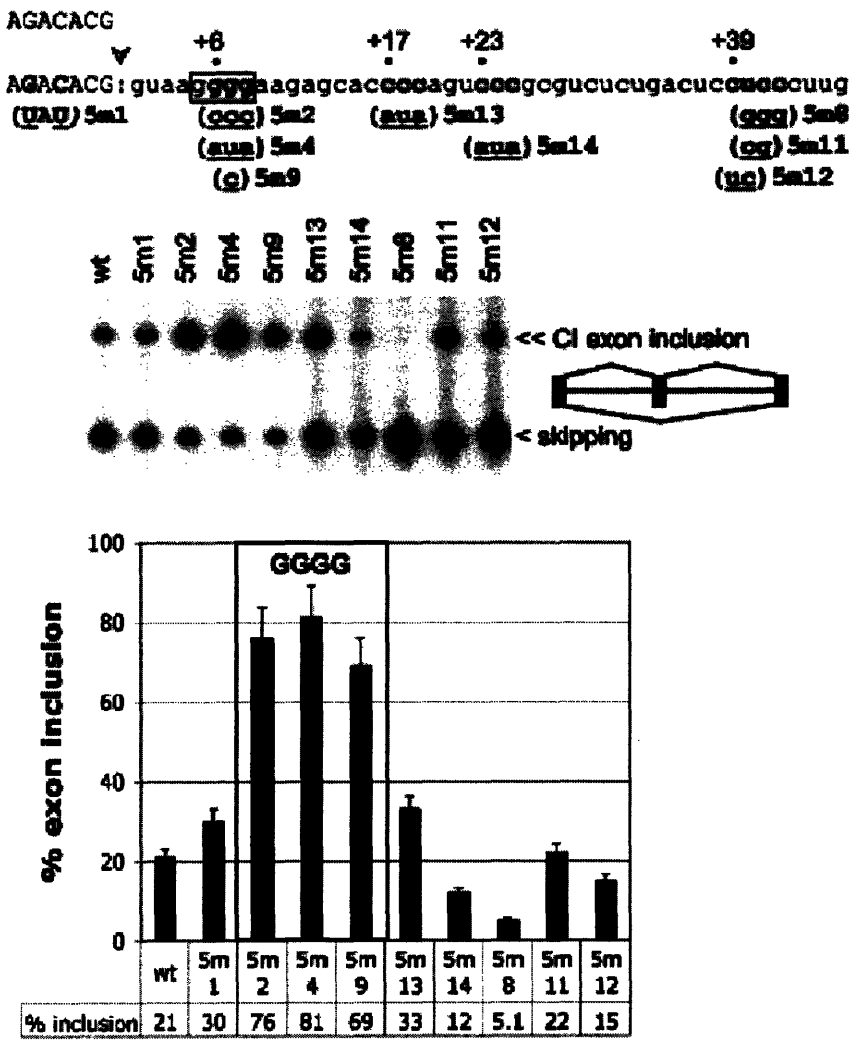
	Ensembl ID_Exon#: Human (ENSG00000-) **Mouse** (ENSMUSG00000-)	**HUGO ID or Genbank accession #**	**Exon length (bp)**	**# TAGG motifs**	**5' splice site sequence: Human Mouse**	**RT-PCR evidence for exon skipping (this study)**	**cDNA and/or EST evidence for exon skipping***
Intersection dataset							
1	158195_4 028868_4	**WASF2**	118	1 1	AGGgtga**ggggg**aa AGGgtga**ggggg**aa	negative	No info
2	169045_5 007850_5	*HNRPH1*	139	1 1	CAGgt**ggggg**atgg CAGgt**ggggg**atgg	skipped	AW579178, and many others
3	096746_3 020069_2	*HNRPH3* *NM_012207*	139	1 1	CAGgt**ggggg**atgg CAGgt**ggggg**atgg	skipped	BE747312 BM916242 BQ882744 AW878310
4	176884_19 026959_19	*GRIN1*	111	2 2	ACGgtaa**ggggg**a ACGgtaa**ggggg**aa	skipped	See below
5	136044_8 020263_8	*NM_018171,* *DIP13BETA*	147	1 1	CAGgta**ggggg**agt CAGgta**ggggg**atg	negative	No info
6	158865_8 030769_9	*NM_052944,* *KST1*	81	1 1	ACAgtaagt**gggg** ACAgtaagt**gggg**	negative	No info
7	168453_3 022096_3	*HR*	793	1 3	AAGgtaa**ggggg**c GAGgtaa**ggggg**t	ND	BX341278
8	068400_2 031153_13	*GRIPAP1*	96	1 1	AAGgta**ggggg**aac AAGgt**ggggg**catc	skipped	No info
9	136478_8 040548_8	*NM_018469* *Uncharacterized*	133	1 1	AGGgtaa**gggg**ct AGGgtaa**gggg**ct	skipped	No info
10	108592_17 020706_18	*FTSJ3*	100	1 1	CCGgtaaa**gggg**c CCGgtaa**gggg**ca	negative	No info
11	152818_5 019820_6	*UTRN*	93	1 1	CAGgt**ggggg**aaat CAGgt**ggggg**acct	skipped	No info
12	158887_4 005678_4	*MPZ* *ENST00000289928*	136	1 1	CAGgtaa**gggg**cg CAGgtaa**gggg**cg	negative	No info
13	181045_5 039908_6	*SCL26A11*	143	1 1	CAGgtga**gggg**cc CAGgtga**gggg**ac	negative	No info
14	147255_16 031111_15	*IGSF1*	288	1 1	CAGgtaa**gggg**aa CTGgtaa**gggg**at	ND	No info
15	173957_7 037336_6	*No description*	91	1 1	CAGgtat**gggg**tt CAGgtat**ggggg**t	ND	No info
16	106404_2 001739_2	*CLDN15*	165	1 1	CCGgtaact**gggg** CTGgtaat**ggggg**	ND	BU164601 AJ245738
17	150165_4 021950_5	*ANXA8* *NM_001630*	91	1 1	AAGgtaa**gggg**tg AAGgtaa**gggg**tt	ND	BC008813 BE902538 BE902353 BE900246
18	179593_2	*ALOX15B*	220	1	CAGgtga**ggggg**cg	ND	No info

	020891_2	*NM_001141*		1	CAGgtgaggggac		
19	165816_11 025082_12	*NA*	552	2 1	GAGgtgagtgggg TGAgtggggataa	ND	No info
Human subset							
h1	176884_19	*GRIN1*	111	2	ACGgtaaggggga	skipped	L13266, AF015730, L05666, L13267, AW900783
h2	097054_10	*ABCA4*	117	2	AGAgtaagggggg	negative	No info
h3	140396_13	*NCOA2*	207	2	CAGgtaaggggtc	skipped	No info
h4	099308_21	*O60307*	245	2	CTGgtaagtgggg	negative	No info
h5	135709_2	*Y513_HUMAN*	501	2	AGGgtaaggggcc	ND	No info
h6	165816_11	*ENST00000298715*	552	2	GAGgtgagtgggg	ND	No info
h7	130283_7	*LASS1*	637	2	GCGgtgagtgggg	ND	No info
h8	007565_3	*DAXX*	832	2	CAGgtagggggtt	ND	-
h9	185133_2	*PIB5PA*	1166	2	CCGgtgagggggc	ND	-
h10	111077_18	*TENC1*	1212	3	CAGgtgaggggca	ND	-
h11	142102_4	*Q8TEG9*	1418	2	CAGgtgaggggac	ND	-
h12	135835_5	*Q9HCF8*	1556	2	ATGgtaaggggct	ND	-
h13	138080_4	*EMILIN1*	1929	2	CTGgtgaggggac	ND	-
Mouse subset							
m1	026959_19	*GRIN1*	111	2	ACGgtaaggggaa	skipped	CD363997
m2	023938_18	*No description*	123	2	GAGgtcaggggcc	ND	No info
m3	024947_8	*MEN1_MOUSE*	165	2	CAGgtgagagggg	skipped	BC036287
m4	026791_8	*GRTR8_MOUSE*	171	2	CTGgtaaagggga	ND	BY347810, BY349516
m5	028759_2	*Hp1bp3*	198	3	GAGgtaggggctg	skipped	AK075725, AK043260
m6	005886_13	*NCOA2*	207	2	CAGgtaagggctc	skipped	BC053387
m7	007021_2	*NM_011522*	238	2	CAGgtgggcgggg	negative	No info
m8	015852_2	*NM_030707*	208	2	AAGgtaggggact	ND	No info
m9	024112_9	*CCAH_MOUSE*	440	2	CAGgtaggggtgt	ND	No info
m10	028782_26	*NM_173071*	620	2	GAGgtgaggggct	ND	No info
m11	022096_4	*HAIR_MOUSE*	781	2	GAGgtaagggggt	ND	No info
m12	052325_5	*MAPB_MOUSE*	6490	2	CAGgtaggtgggg	ND	No info

Entries 1-19 correspond to the Intersection dataset of Figure 5A, human (top), mouse (bottom); human subset, h1-h13, and mouse subset, m1-m12, are also shown. ND, not determined.
*(http://genome.ucsc.edu; http://www.ncbi.nlm.nih.gov/)

Figure 1

A. Cl cassette exon, 5' splice site region



B. ESE motifs

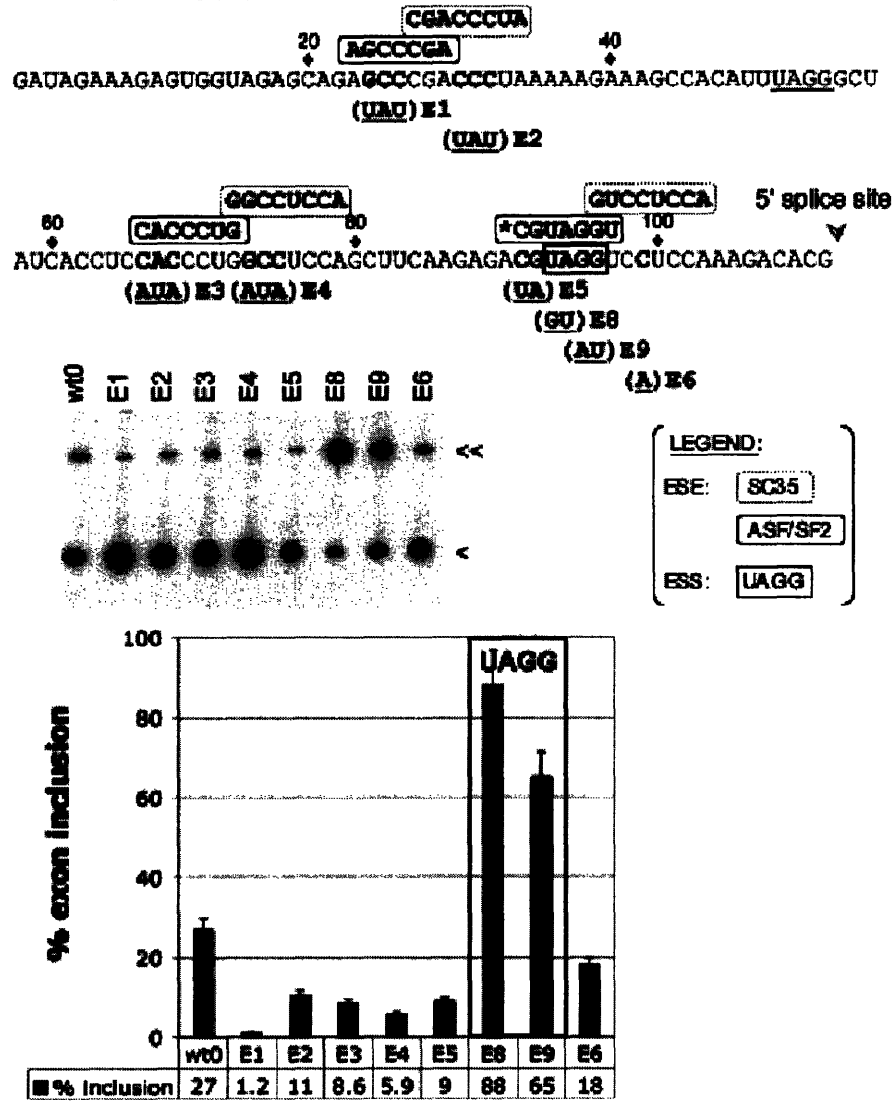
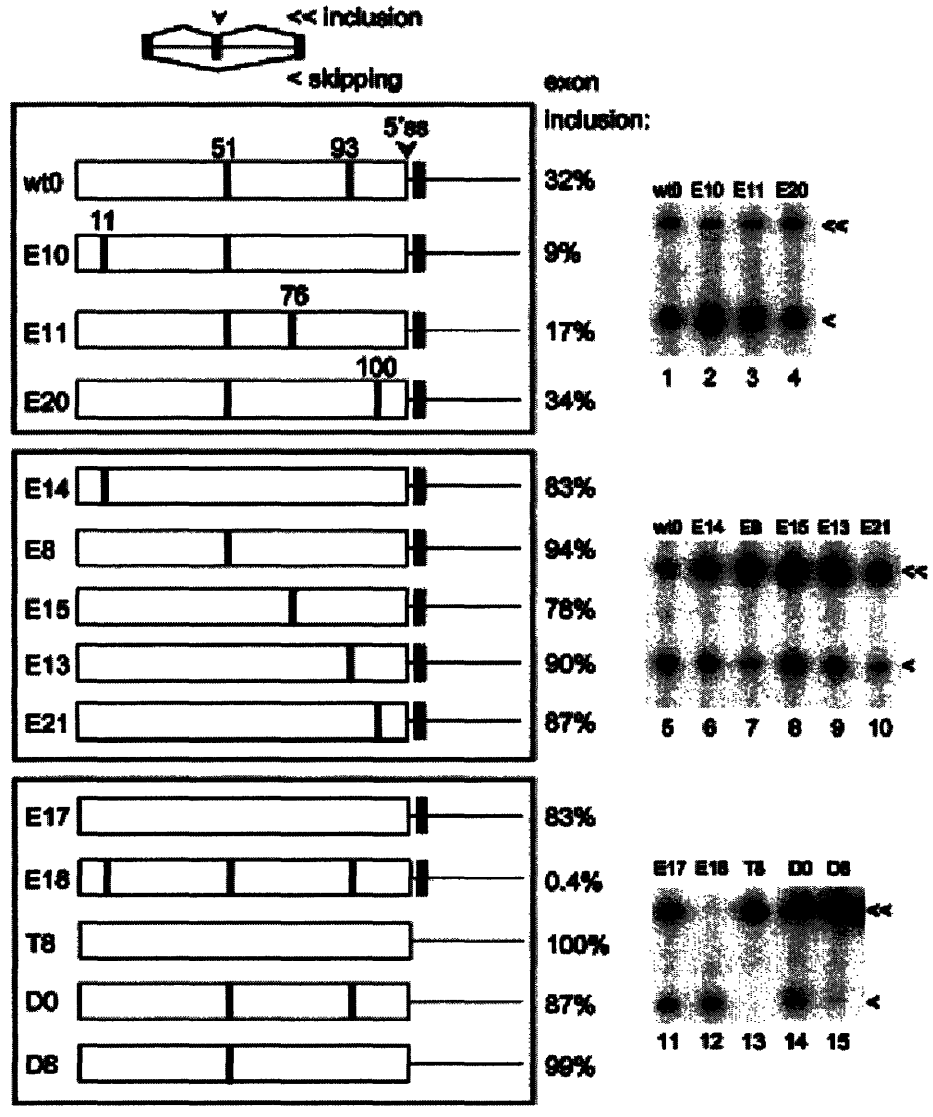


Figure 2



Exonic mutations:
(Position) wt > mutant
(11) GUGG > UAGG
(51) UAGG > AUGG
(76) CCAG > UAGG
(93) UAGG > GUGG
(100) UCCAA > UAGGC

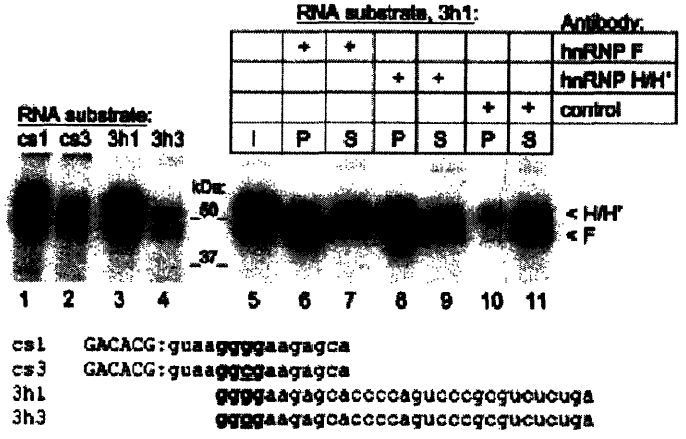
Strong silencing motif patterns:

UAGG_UAGG_UAGG∨GGGG

UAGG_UAGG∨GGGG

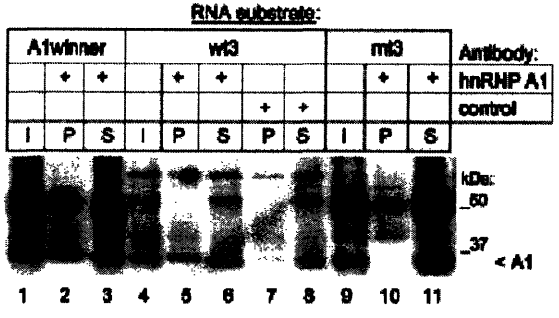
Figure 3

A. GGGG motif



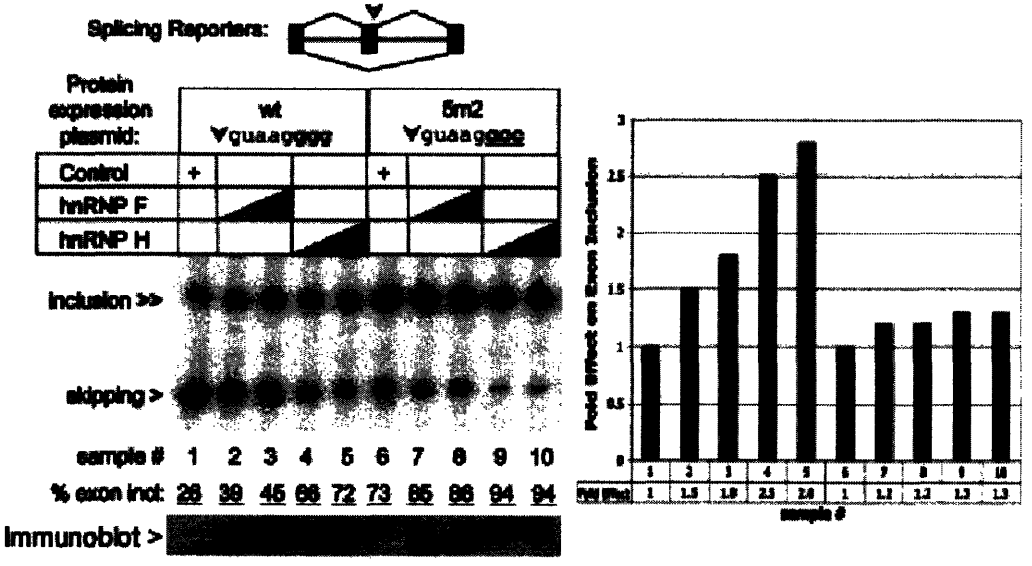

```
cs1   GACACG:guaaggggaagagca
cs3   GACACG:guaaggccgaagagca
3h1             ggggaagagcaccccaguccgcgucucuga
3h3             ggggaagagcaccccaguccgcgucucuga
```

B. UAGG motif




```
Alwinner: GAUGAUAGGGACUUAGGGUG
wt3: GGCCUCCAGCUUCAAGAGACGUA*GGUCCUCCAAAGACACG:guaaggggaa
     gagcaccccа
mt3:                        GU*GG
```

C. hnRNP F, H



D. hnRNP A1

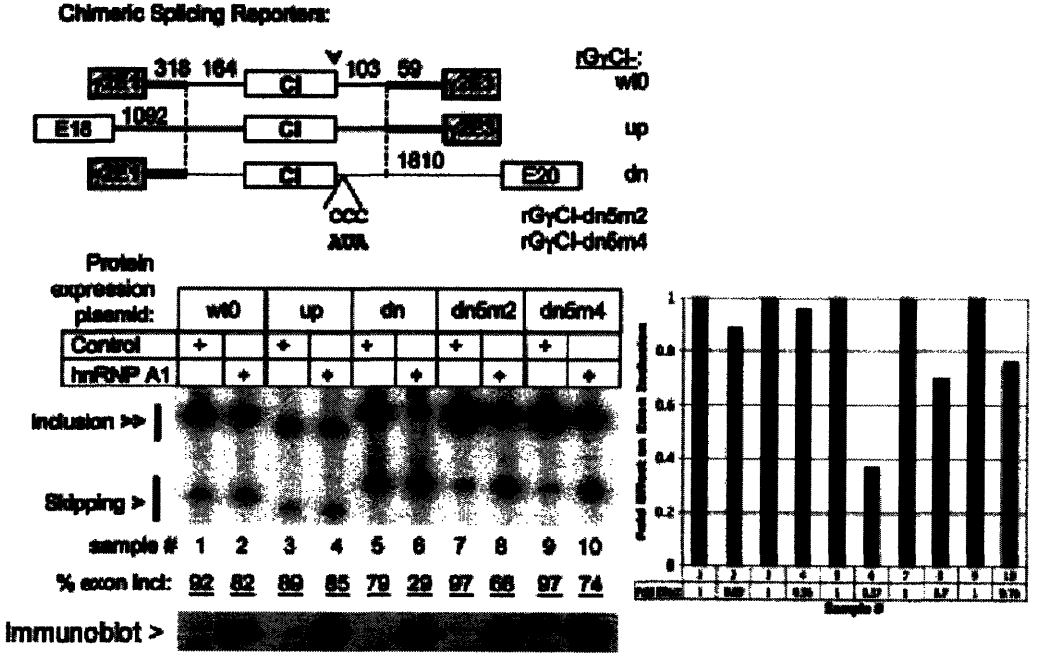
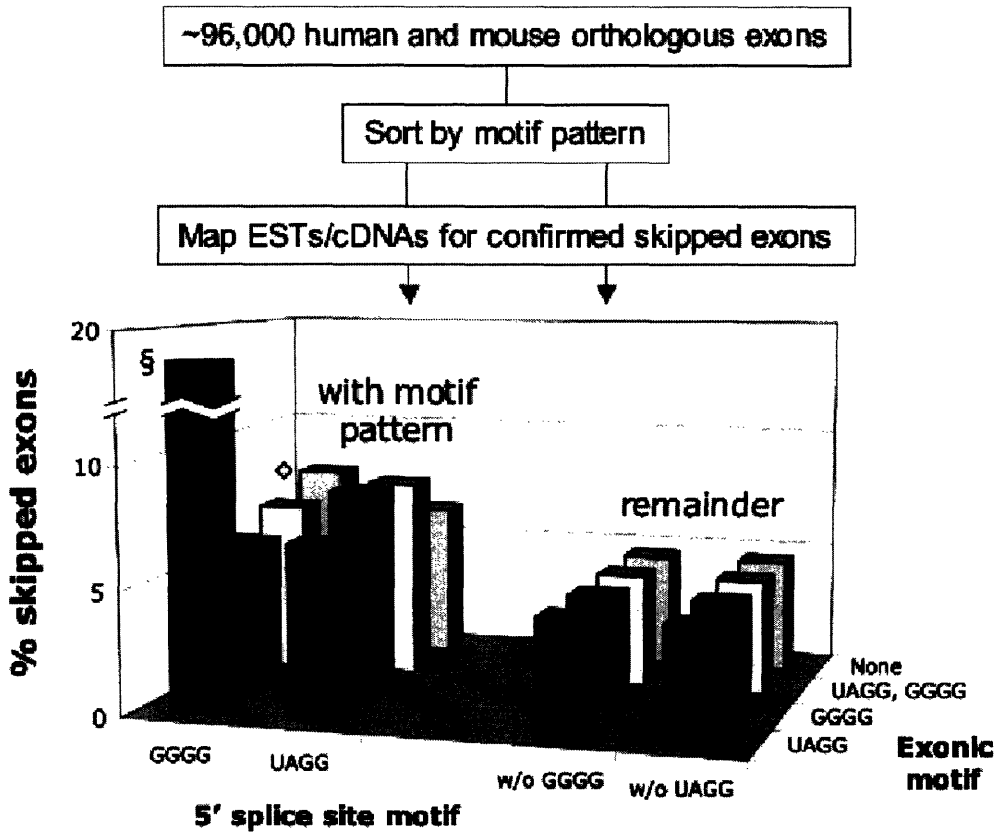


Figure 4

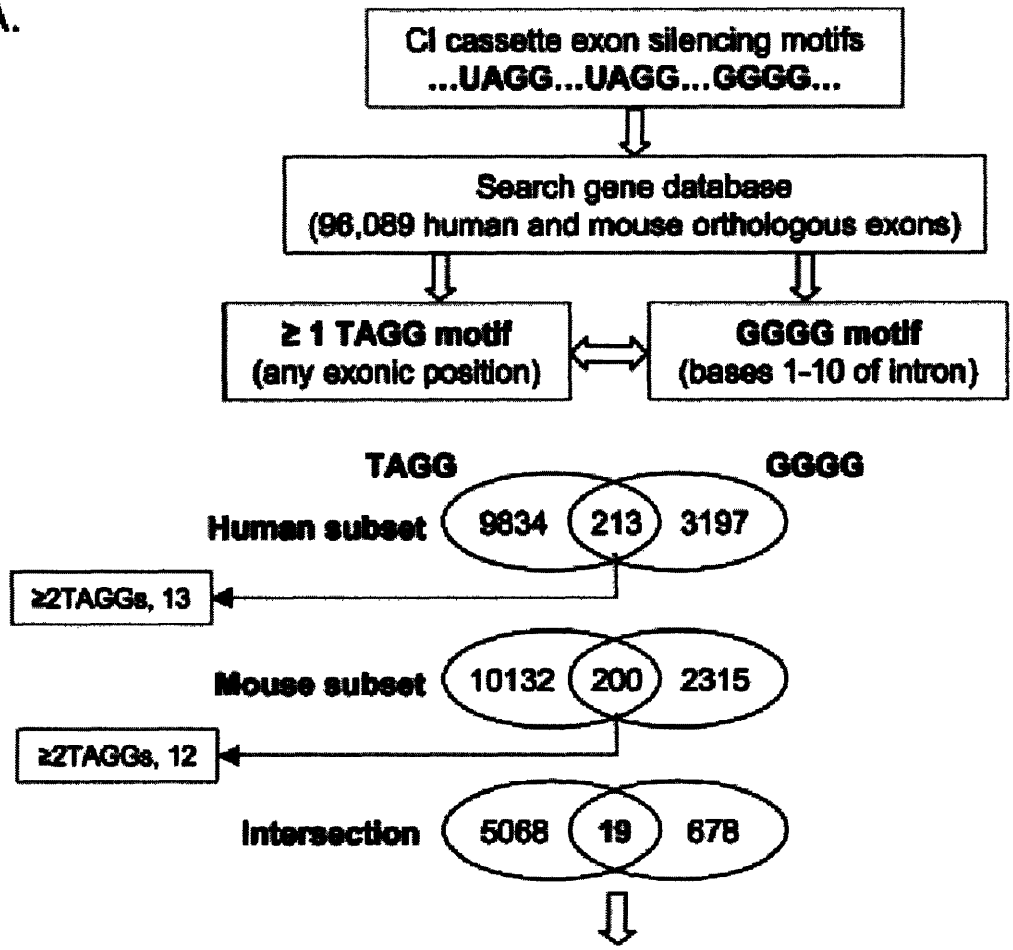


Exonic motif	5′ splice site motif			
	GGGG	UAGG	w/o GGGG	w/o UAGG
None □	*7.8% (589)	6.3% (2,742)	4.6% (89,602)	4.6% (87,449)
≥ 1 (UAGG, GGGG) □	6.9% (175)	8.1% (706)	4.6% (90,016)	4.6% (89,435)
≥ 1 GGGG ■	6.1% (163)	8.4% (617)	4.6% (90,028)	4.6% (89,574)
≥ 1 UAGG ■	§18.8 % (16)	7.0% (114)	4.6% (90,175)	4.6% (90,077)

Fraction of exons with motif pattern

Remainder w/o motif pattern

A.



Human examples from Intersection dataset

	EXON:intron	Gene_Exon#
TAGG found at ≤ 50 bases upstream of 5' splice site		
AGATGCTGCAGGACACCAAGGATATCATGAAAGAGAAGAGAAAGCA**TAGG**:gtgaggggaa		**WASP2_4**
GAAATAGCTGAAAAGGCTCTAAAGAAACACAAGGAAAGAA**TAGG**GCACAG:gt**gggg**atgg		**HNRPH1_5**
GAGATAGCAGAAAATGCTCTGGGGAAACACAAGGAAAGAA**TAGG**GCACAG:gt**gggg**atgg		**HNRPH3_3**
CCTCCACCCTGGCTTCCAGCTTCAAGAGGCG**TAGG**TCCTCCAAAGACACG:gtaa**gggggg**a		**GRIN1_19**
GAAAGCAAATGGCCATGATGGAGCCCATGA**TAGG**CTTTGCCCATGGACAG:gta**gggg**agt		**DIP13b_8**
CCTGCAGACGCTGATCATGCTTA**TAGG**AGCGCTCACCTTGATGGGCTACA:gtaagt**gggg**		**RKST1_8**
TAGGGCAGGAGAAGGAGCAGTTGACCCAGGAATTACAGGAGGCTCGGAAG:gta**gggg**aac		**GRIPAP1_2**
TAGG found at > 50 bases upstream of 5' splice site		
GGTCCTTTCTGATGACTCTCGAGACCAAAATGAATTTGACCAAAC**TAGGT**		
AAAGAGCCTCTTGTTGAAGCCCTGAAGGTTGGAGAAATTGGCAAAGAAGG:gtaa**gggg**ct		**NM_018469_8**
CGAAACATCGGATACTGGACCCCGAAGGCCTTGCTC**TAGG**TGCTGTTATT		
GCCTCTTCCAAAAAGGCCAAGAGAGACCTCATAGATAACTCCTTCAACCG:gtaaa**gggg**c		**FTSJ3_17**
GTGGAATTAGTGAATA**TAGG**GGAACTGACATTGTGGATGGAA		
ATCACAAACTGACTTTGGGGTTACTTTGGAGCATCATTTTGCACTGGCAG:gt**gggg**aaat		**UTRN_5**
TGCCAAC**TAGG**TACGGGGTCGTTCTGGGAGCTGTGA		
TCGGGGGTGTCCTCGGGGTGGTGCTGTTGCTGCTGCTGCTTTTCTACGTG		
GTTCGGTACTGCTGGCTACGCAGGCAGGCGGCCCTGCAGAGGAGGCTCAG:gtaa**gggg**cg		**MPI_4**
GG**TAGG**TGACGCCGTCCTGGGGCTGGTCTGCATGCTGCTGCTG		
CTGGTGCTGAAGCTGATGCGGGACCACGTGCCTCCCGTCCACCCCGAGAT		
GCCCCCTGGTGTGCGGCTCAGCCGTGGGCTGGTCTGGGCTGCCACGACAG:gtga**gggg**cc		**SCL26A11_5**

Figure 5B

B. **Human exon skipping patterns**
ID_Exon# (ENSG00000-)

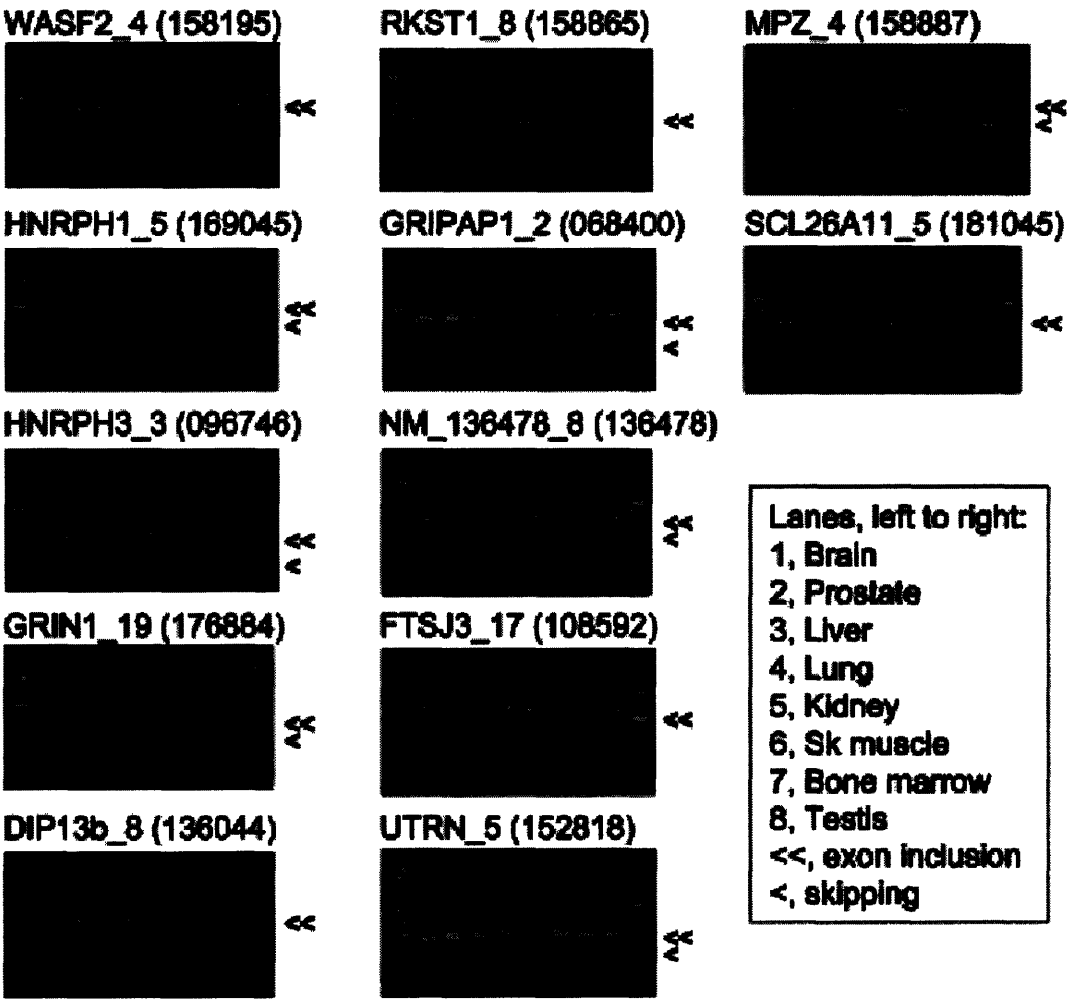


Figure 5C

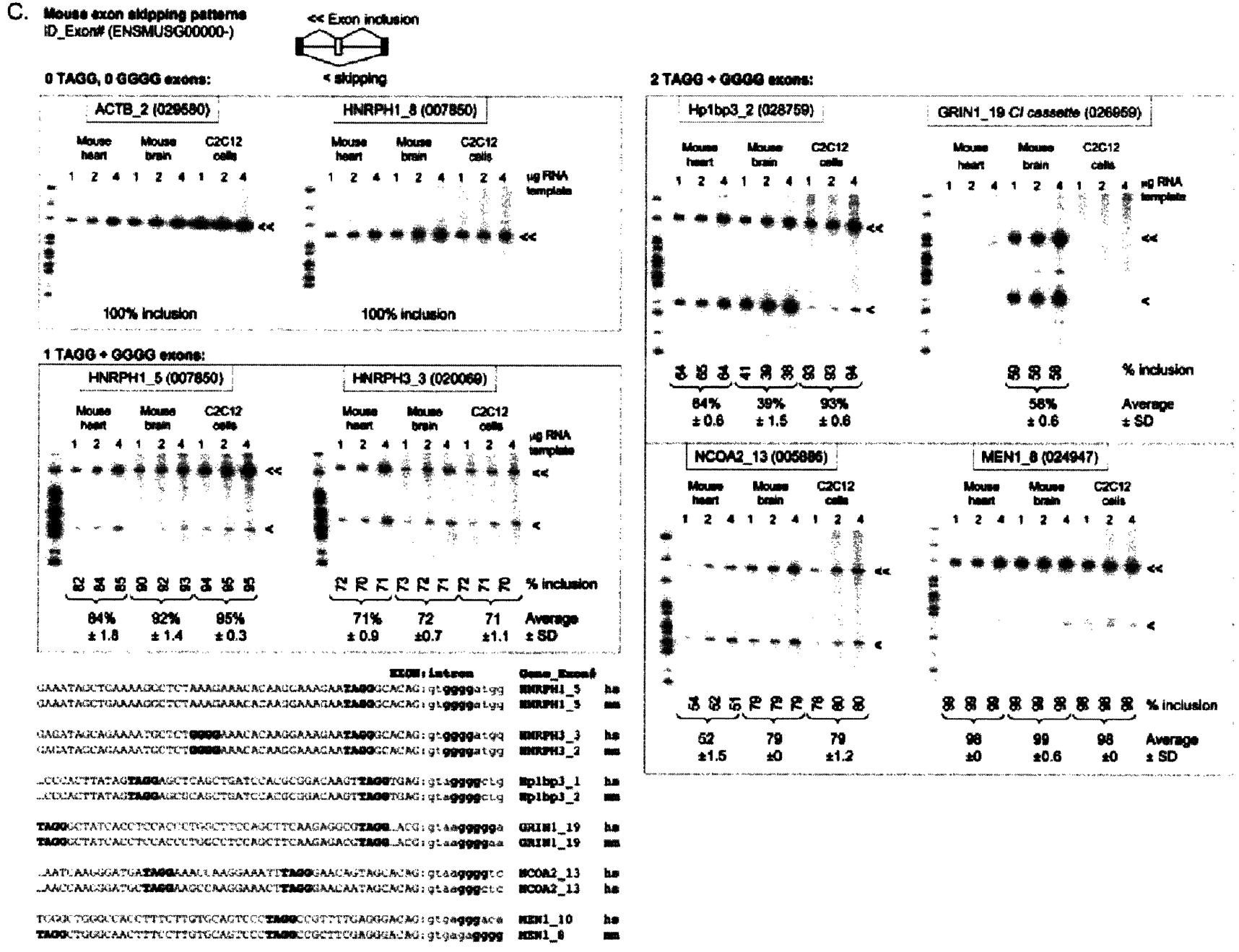
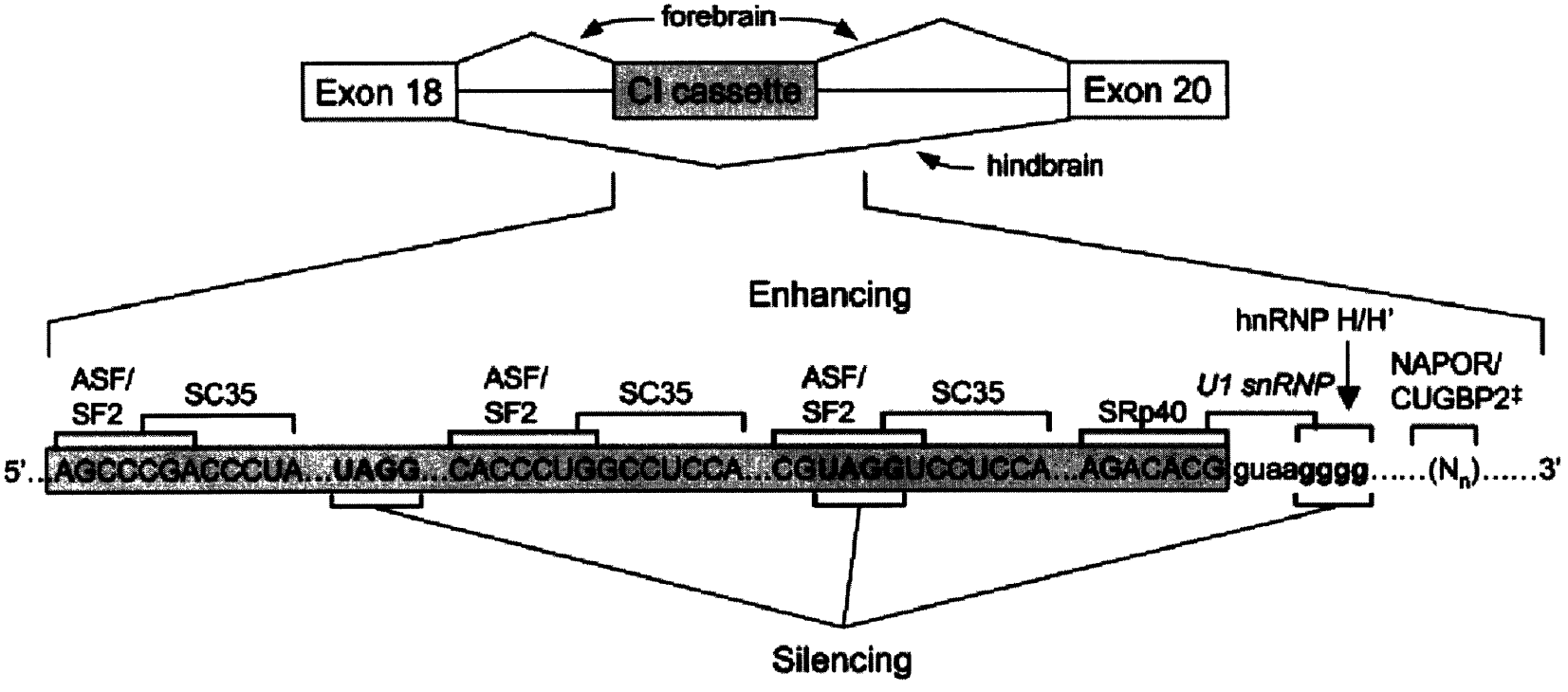

C.
Mouse exon skipping patterns
ID_Exon# (ENSMUSG00000-)
<< Exon inclusion
< skipping
0 TAGG, 0 GGGG exons:
ACTB_2 (029580)
HNRPH1_8 (007850)
Mouse heart
Mouse brain
C2C12 cells
µg RNA template
100% inclusion
100% inclusion
1 TAGG + GGGG exons:
HNRPH1_5 (007850)
HNRPH3_3 (020069)
% inclusion
84% ± 1.8
92% ± 1.4
95% ± 0.3
71% ± 0.9
72 ±0.7
71 ±1.1
Average ± SD
2 TAGG + GGGG exons:
Hp1bp3_2 (028759)
GRIN1_19 Cl cassette (026959)
NCOA2_13 (005886)
MEN1_8 (024947)
84% ± 0.6
39% ± 1.5
93% ± 0.6
58% ± 0.6
52 ±1.5
79 ±0
79 ±1.2
98 ±0
99 ±0.6
98 ±0
Average ± SD
Exon:Intron
Gene_Exon#
HNRPH1_5
HNRPH3_3
Hp1bp3_1
Hp1bp3_2
GRIN1_19
NCOA2_13
MEN1_10
MEN1_8

Figure 6

GRINI_CI Cassette Exon: Splicing Regulatory Motifs and Model for Tissue Specificity



Chapter 5.

Conclusions and Perspectives

5.1 Perspectives

In summary, the major accomplishments of this thesis are as follows: (i) the introduction of a general approach to incorporating adjacent and non-adjacent local dependencies in modeling sequence motifs, with applications to RNA splicing signals; (ii) the first large-scale comparative bioinformatics study of similarities and differences in the *cis* and *trans* elements involved with constitutive splicing across vertebrates, which resulted in identifying substantial differences in the regulation of splicing in *Fugu* versus mammalian introns, as well as developing the first intron classifier which can be used to design spliceability into *Fugu* introns in mammalian cell lines; (iii) a rigorous study of the variation in alternative splicing across various human tissues which incorporates the gene expression of *trans* factors, and the identification of tissue-specific *cis* elements; (iv) the development of a novel method for large-scale detection of conserved alternative splicing events in mammals; and (v) the identification of a combination of motifs that are predictive of exon-skipping in human and mouse.

Aside from the specific achievements described above, each of the works mentioned above has led or could lead to various independent lines of research. For example, in

identifying splicing *cis*-regulatory elements, such as ESEs and ISEs in multiple vertebrates, a list of conserved ESEs (*1*) and potential ISEs in vertebrates have been generated. A recent study by Fairbrother et al (*2*) utilized the predicted ESEs conserved in mammals and fish, that were generated in this thesis, and showed that single nucleotide polymorphisms (SNPs) in human genes avoid candidate ESEs that are conserved in multiple vertebrates with higher frequency than those identified in single genome studies, supporting the validity of our methods. In addition, the candidate vertebrate ISEs can be experimentally tested by tranfecting mutiple cell lines with constructs which contain the candidate intronic *cis* elements in a weak exon context.

In our survey of the variation of alternative splicing in multiple tissues, tissue-specific *cis*-regulatory elements for splicing have also been identified, which requires further experimental validation. Furthermore, the methods introduced for measuring dissimilarity between isoforms can be improved on and applied also to data from non EST-based experimental methods for quantifying alternative splicing, such as splicing-specific microarrays (*3*).

A promising direction of further research lies in the conserved alternative spliced exons identified, which are evolutionarily preserved and likely to be biologically important. A host of questions can be raised of this set of alternatively spliced exons. For example, is alternative splicing affected by changes in the external physiological conditions, such as stress (for example UV damage and cell starvation)? What is the impact on alternative splicing from administrating common pharmacological molecular interventions such as protease inhibitors and inflammation drugs to human cell lines? Can we identify

connections between cancer and alternative splicing by assaying cancer cell lines and tissues? Can we test the effects of microRNAs on the fate of alternative splicing regulation as several splicing factors are potentially regulated by microRNAs (*4*)? Can we address the connection between alternative splicing and nonsense mediated mRNA decay (NMD) (*5, 6*) by suppressing this mechanism chemically or using a human cell lines with specific gene knock downs? Can we identify species-specific alternative splicing, distinguished from ACEScan[+] exons? In addition, candidate conserved *cis*-regulatory elements that were utilized as important features can be experimentally verified by mutations to a construct containing the alternatively spliced exon, or by insertion of the element into a constitutive exon.

5.2 Current splicing-sensitive technologies

To address several of the questions above, large-scale sensitive methods and resources for assaying alternative splicing have to be tapped, which are briefly covered in this section. Microarray technologies allow the levels of many different RNA or DNA sequences to be assayed simultaneously, but existing methods are not sensitive or specific enough to study alternative splicing (AS). Long probes used in cDNA arrays are not designed to distinguish subtle nucleotide-level differences that can take place due to the alternative use of a splice site, resulting in the insertion of a dozen extra bases, for example, oligonucleotide-based arrays, including exon arrays, tiling arrays and exon-exon junction arrays (*7, 8*) allow the detection of mRNA isoforms, but complications arise due to target amplification methods being biased towards 3'UTR sequences, and thus require larger amounts of RNA to preserve the true ratios of isoforms. Nevertheless, oligonucleotide arrays of exon-exon junction probes can be useful for guiding reverse

transcriptase-polymerase chain reaction (RT-PCR) and sequence validation efforts to identify alternative splicing events (*3, 9*).

In addition to oligonucleotide exon-exon microarray technologies, Church and colleagues have introduced "digital polony exon profiling"(*10*), and Fu and colleagues have introduced a method called RASL (RNA-mediated annealing, selection and ligation) using fiber-optic arrays (*11*) to monitor alternative splicing events, both requiring significantly smaller amounts of RNA. These two methods (summarized in detail in the Table below) will be useful in measuring complicated combinatorial alternative splicing events quantitatively with small RNA samples.

Finally, unlike the above methods, differential analysis of transcripts with alternative splicing (DATAS) (*12*) has a major advantage over the other technologies in that it allows the systematic generation of libraries of alternatively spliced isoforms without prior knowledge of the event. This method is based on hybridizing mRNA from one condition/cell line to cDNA from another condition/cell-line, and digesting looped-out RNA sequences, which would be indicative of alternative splicing of the mRNA in one condition/cell line, but not the other. This method allows the comparison of full-length RNA transcripts between two biological samples to extract novel alternative exons and introns between these transcripts.

Table 1. Comparison of analytical methods for profiling alternative splicing events.

	Oligonucleotide exon-exon junction arrays (*3, 9, 13*)	**Fiber-optic arrays (RASL) (*11*)**	**Digital polony exon profiling (*10*)**	**DATAS (*12*)**
Summary of procedure	Oligonucleotide probes designed across exon-exon junctions. RNA is amplified and labeled using Cy5-Cy3 dyes and hybridized to the microarray. The intensity of pairs of junction probes in different tissues and cell lines indicate the presence of alternative splicing. Candidate alternatively spliced exons are reverse-transcribed, and sequence validated.	Self-assembled bead array is formed when microspheres are loaded onto the tip of eched fiber-optic bundles. Each bead contains a specific oligonucleotide sequence (address). A charge-coupled device camera at the opposite end of a fiber bundle records hybridization signals. Signals from multiple beads carrying identical addressees on the array are combined to derive final signal output.	Parallel solid-phase amplification of DNA molecules via PCR in an acryl-amide gel attached to the surface of a glass microscope slide forms spherical colonies of DNA, called polonies. The products are linked to the gel and serves as a template for probe hybridization and/or single base extensions. Combinations of spectrally distinct fluorophores and/or repeated cycles of probing allows the studying of combinatorial patterns of AS.	DATAS is designed to capture mRNA splice variants that distinguish different conditions. cDNA from one population is hybridized with mRNA from the second population. Magnetic streptavidin beads are used to isolate the biotinylated cDNA. Hybrids are treated with RNAse H to release mRNA that do not hybridize to the heteroduplex. Release mRNAs are isolated, reverse transcribed, cloned and collected into libraries.
Sequence knowledge of AS event	Oligonucleotide probes are designed specifically across exon-junctions.	DNA oligos complementary to alternative exons are linked to an address that can hybridize to its complementary strand on the bead array, and only juxtaposed oligos (by splicing) are ligated, and amplified using	Cy3 and Cy5 labeled exon probes are required for detection of AS.	Not required

		universal primers		
Advantages	Scale of method is currently the largest. Can potentially address complex combinatorial AS.	Small numbers of cells (10) required, suitable for neuronal and immune system characterization of AS.	Quantification of individual mRNA isoforms is sensitive and accurate as each polony arise from a single molecule. Single based extensions combined with polonies allows the correlation of single nucleotide changes (SNCs) with splicing defects. Can address complex combinatorial AS.	No sequence knowledge required.
Disadvantages	Relies on sequence knowledge of AS. Suffers from biases of conventional microarray platforms. Cannot be used to study small number of cells.	Relies on sequence knowledge of AS.	Currently, multi-plexing is limited, but can be extended by quantum-dot labeled probes. Relies on sequence knowledge of AS.	Have to be performed in pairs of groups (cancer versus normal). Difficult to use to study complex combinatorial AS.

References

1. W. G. Fairbrother *et al.*, *Nucleic Acids Res* **32**, W187 (Jul 1, 2004).
2. W. G. Fairbrother, D. Holste, C. B. Burge, P. A. Sharp, *PLoS Biol* **2**, E268 (Aug 31, 2004).
3. J. M. Johnson *et al.*, *Science* **302**, 2141 (Dec 19, 2003).
4. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, *Cell* **115**, 787 (Dec 26, 2003).
5. R. E. Green *et al.*, *Bioinformatics* **19 Suppl 1**, I118 (Jul, 2003).
6. J. T. Mendell, C. M. ap Rhys, H. C. Dietz, *Science* **298**, 419 (Oct 11, 2002).
7. G. K. Hu *et al.*, *Genome Res* **11**, 1237 (Jul, 2001).
8. D. D. Shoemaker *et al.*, *Nature* **409**, 922 (Feb 15, 2001).
9. J. Castle *et al.*, *Genome Biol* **4**, R66 (2003).
10. J. Zhu, J. Shendure, R. D. Mitra, G. M. Church, *Science* **301**, 836 (Aug 8, 2003).
11. J. M. Yeakley *et al.*, *Nat Biotechnol* **20**, 353 (Apr, 2002).
12. F. Schweighoffer *et al.*, *Pharmacogenomics* **1**, 187 (May, 2000).
13. T. A. Clark, C. W. Sugnet, M. Ares, Jr., *Science* **296**, 907 (May 3, 2002).