Information analysis of human splice site mutations

Peter K. Rogan,* † Brian M. Faux,*
and Thomas D. Schneider ‡

version = 1.73 of rfs.tex 1999 August 12

Abbreviated Title: Information at human splice site mutations

Human Mutation 12: 153-171 (1998)

ABSTRACT

Splice site nucleotide substitutions can be analyzed by comparing the individual information contents (R_i, bits) of the normal and variant splice junction sequences [Rogan and Schneider, 1995]. In the present study, we related splicing abnormalities to changes in R_i values of 111 previously reported splice site substitutions in 41 different genes. Mutant donor and acceptor sites have significantly less information than their normal counterparts. With one possible exception, primary mutant sites with < 2.4 bits were not spliced. Sites with R_i values ≥ 2.4 bits but less than the corresponding natural site usually decreased but did not abolish splicing. Substitutions that produced small changes in R_i probably do not impair splicing and are often polymorphisms. The R_i values of activated cryptic sites were generally comparable to or greater than those of the corresponding natural splice sites. Information analysis revealed pre-existing cryptic splice junctions that are used instead of the mutated natural site. Other cryptic sites were created or strengthened by sequence changes that simultaneously altered the natural site. Comparison between normal and mutant splice site R_i values distinguishes substitutions that impair splicing from those which do not, distinguishes null alleles from those that are partially functional, and detects activated cryptic splice sites.

keywords: information theory, mRNA splicing, donor, acceptor, cryptic, mutation, polymorphism, walker

*Department of Human Genetics, Allegheny University of the Health Sciences, Pittsburgh, PA

†Correspondence should be addressed to: Peter K. Rogan, Ph.D., Department of Human Genetics, MCP-Hahnemann School of Medicine, Allegheny University of the Health Sciences, 320 E. North Ave., Pittsburgh, PA 15212; tel:(412) 359-4988; fax:(412) 359-6488; email: progan@pgh.allegheny.edu

‡National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, P. O. Box B, Frederick, MD 21702-1201. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598, email address: toms@ncifcrf.gov.

INTRODUCTION

Mutations at splice sites make a significant contribution to human genetic disease, since approximately 15% of disease-causing point mutations affect pre-mRNA splicing [Krawczak et al., 1992]. Mutations in splice sites decrease recognition of the adjacent exon and consequently inhibit splicing of the adjacent intron [Talerico and Berget, 1990; Carothers et al., 1993]. Splice site mutations may result in exon skipping, activation of cryptic splice sites, creation of a pseudo-exon within an intron, or intron retention [Nakai and Sakamoto, 1994]: (1) Exon skipping, the most frequent outcome, is thought to result from failure of the normal and mutant splice sites to define an exon. (2) Most cryptic mutations activate splice sites of the same type and are typically located within a few hundred nucleotides of the natural site. This distance is probably limited by restrictions on the length of the resultant exon [Hawkins, 1988; Berget, 1995]. (3) Occasionally, mutations that are further away from the natural splice site create cryptic sites that are activated in the presence of a nearby cryptic splice site of opposite polarity, producing a novel non-coding exon within the intron. (4) Splice site mutations in very short or terminal introns can result in intron retention [Dominski and Kole, 1991]. In these instances, additional sequence elements may be required for normal splicing [Black, 1991; Black, 1992; Sterner and Berget, 1993].

Essential elements in donor and acceptor splice junctions have been defined by consensus sequences [Mount, 1982], by analysis of nucleotide frequencies at each position in a splice site [Senapathy et al., 1990] and by neural network prediction [Brunak et al., 1990]. Each of these methods have limitations. Although the GT and AG positions adjacent to donor and acceptor splice junctions are highly conserved, other positions are more variable [Mount, 1982; Stephens and Schneider, 1992]. The consensus sequence approximates the nucleotide frequencies at each position, and so it excludes the contributions of less frequent nucleotides present in a proportion of natural splice sites. Splice site sequences that deviate from the consensus do not necessarily produce significantly lower amounts of spliced mRNA [Rogan and Schneider, 1995]. Training a neural network requires sequences of both binding sites and sequences that are not bound [Stormo et al., 1982; Brunak et al., 1990]. Generally, non-bound sequences are taken to be those remaining after binding sites have been identified. However, these sequences do contain functional sites [Schneider, 1997*b*; Hengen et al., 1997], so neural networks may be inappropriately trained on overlapping data sets.

In contrast, information-theory based models of donor and acceptor splice sites require only functional sites and show which nucleotides are permissible at both highly-conserved and variable positions of these sites [Stephens and Schneider, 1992]. Information is the only measure of sequence conservation which is additive [Shannon, 1948]. The information content (R_i, in bits) of a member of a sequence family describes the degree to which that member contributes to the conservation of the entire family [Schneider, 1997*a*; Schneider, 1997*b*]. R_i is the dot product of a weight matrix derived from the nucleotide frequencies at each position of a splice site sequence database and the vector of a particular sequence. Individual information is related to thermodynamic entropy and therefore to the free energy of binding [Schneider, 1994; Schneider, 1997*a*]. Since splice sites are recognized prior to intron excision [Berget, 1995], the sequence of the splice site dictates the strength of the spliceosome-splice junction interaction and thus splice site use. It is our thesis that the strength of this interaction is related to the information content of the splice junction.

A group of sites with similar sequence and function can be described and quantified by their corresponding distribution of individual information contents. The mean of this distribution of R_i values is 7.92 ± 0.09 bits for the 10 nucleotide long splice donor sites and 9.35 ± 0.12 bits for the 28 nucleotide long acceptor sequences [Stephens and Schneider, 1992; Schneider, 1997*a*], representing the average amount of information required for splicing, $R_{sequence}$ [Schneider et al., 1986; Schneider, 1995; Schneider, 1994]. Strong splice sites have R_i values $\gg R_{sequence}$; weak sites have R_i values $\ll R_{sequence}$. Non-functional sites have R_i values less than or equal to zero [Schneider, 1994; Schneider, 1997*a*]. Since mutations at splice sites lessen or abolish splicing at those sites, we investigated whether the R_i values of mutant splice sites were related to defects in mRNA processing and whether mutant, cryptic and the corresponding natural splice sites could be ordered based on their respective R_i values.

MATERIALS AND METHODS

Individual information analysis

Information content is defined as the number of choices needed to describe a sequence pattern, using a logarithmic scale in bits [Schneider et al., 1986; Schneider, 1995]. A set of either donor or acceptor splice junction recognition sites are aligned and the frequencies of bases at each position are determined. The weight matrix used to model the splice junctions is computed from

$$R_{iw}(b, l) = 2 - (-\log_2 f(b, l) + e(n(l))) \qquad \text{(bits per base)} \tag{1}$$

where $f(b, l)$ is the frequency of each base b at position l in the aligned binding site sequences and $e(n(l))$ is a sample size correction factor [Schneider et al., 1986] for the n sequences at position l used to create $f(b, l)$ [Schneider, 1997*a*]. The matrix, $R_{iw}(b, l)$, is a 2 dimensional array in which row b corresponds to one of the 4 nucleotides in DNA and column l is the position along the aligned set of splice junction recognition sites. This individual information matrix represents the sequence conservation of each nucleotide, measured in bits of information. $R_{iw}(b, l)$ can be used to rank-order the sites, to search for new sites, to compare sites with one another, to compare sites to other quantitative data such as DNA-protein binding strength, and to detect errors in databases [Schneider, 1997*a*; Schneider, 1997*b*].

The individual information of a sequence j is the dot product between the sequence and the weight matrix:

$$R_i(j) = \sum_{l} \sum_{b=a}^{t} s(b, l, j) R_{iw}(b, l) \qquad \text{(bits per site)} \tag{2}$$

where $s(b, l, j)$ is a binary matrix for the jth sequence, in which cells have a value of 1 for base b at position l and a value of 0 elsewhere.

The mean of the distribution of R_i values of natural sites is $R_{sequence}$ [Schneider, 1997*b*; Schneider, 1997*a*]. The distribution of R_i values is approximately Gaussian, however the lower and upper bounds are zero bits and the R_i value of the consensus sequence.

The null R_i distribution was determined by creating a random 10,000 nucleotide sequence with a Markov chain process that maintained the same mono- and dinucleotide composition as the human splice junction database [Stephens and Schneider, 1992]. The means of the splice donor and acceptor null distributions were respectively -14.20 ± 6.88 and -14.67 ± 7.15 bits. The probability of observing either a donor or acceptor site with $R_i > 0$ in this random sequence was 0.02 (Z = 2.0).

The effects of nucleotide substitutions can be evaluated by comparing the individual information of the common and variant alleles. The minimum fold change in binding affinity of two sites is $2^{\Delta R_i}$, where ΔR_i is the difference between their respective individual information contents [Schneider, 1997*a*].

Computational tools have been developed to investigate and display individual information. The $R_{iw}(b, l)$ matrices were first computed from a set of 1799 splice donor and 1744 acceptor sequences [Stephens and Schneider, 1992]. To scan for potential sites or to determine the effects of a sequence change on the normal and neighboring sites, the individual information content of the donor or acceptor motif is computed for every site-length window in the sequence. To assess the effects of various substitutions on a specific donor or acceptor site, R_i was computed for the normal and variant sites with the program Scan and displayed with MakeWalker, DNAPlot and Lister (Schneider, 1997*b*; http://www-lecb.ncifcrf.gov/~toms/walker).

The Scan program uses the $R_{iw}(b, l)$ matrix to evaluate the individual information (R_i) at each position in a sequence. For each evaluation, it also computes the number of standard deviations away from $R_{sequence}$ (Z score), and the one-tailed probability (p) of observing a normal splice site with that value of R_i. Sequences with R_i values that are either significantly greater or less than $R_{sequence}$ have low probabilities of belonging to the natural population of sites.

A walker graphically shows the contributions of each position to a binding site. In the display (generated by Makewalker or Lister), favorable contacts between the spliceosome and a test sequence are indicated by letters that extend upwards, while positions that are predicted to make unfavorable contacts are shown by inverted letters. Makewalker is interactive and shows one walker at a time, while Lister displays multiple walkers aligned with sequences and annotated by coding regions (*e.g.* Figs 1-4).

Selection of mutations

Human splice site mutations were chosen from published reports for which corresponding genomic sequence data were available. Only a subset of reported mutations could be analyzed, as sufficient intron sequences were often unavailable (< 26 nucleotides for acceptor sites, < 7 nucleotides for donor sites). To investigate the relationship between R_i value and splice site use, studies that evaluated expression of the mutant mRNA were selected whenever possible. A sequence interval (> 100 nucleotides) surrounding the splice junction was scanned to detect potential cryptic splice sites in the vicinity of the natural site. Larger sequence windows were used for cryptic sites known to occur further away from the natural site (*e.g.* Table 2, #24).

Two mutations could not be analyzed because there were discrepancies at corresponding splice site sequences from different reports. A mutation in the IVS 10 acceptor of the hexosaminidase B gene could not be analyzed because the natural acceptor site had negative information content in one of the sequences [Neote et al., 1988; Proia, 1988]. A similar inconsistency was found in two different versions of the IVS 5 acceptor sequence of the protein kinase C gene [Foster et al., 1985; Soria et al., 1993].

Statistical analyses

Natural and variant sites with $R_i > 0$ were compared with $R_{sequence}$ [Stephens and Schneider, 1992] by using the Z statistic and associated probability of observing a site with a particular R_i value [Schneider, 1997*a*].

Primary mutations for either donor or acceptor sites were analyzed by determining the average differences in R_i values ($\overline{\Delta R_i}$) of natural versus mutant sequences. Significance was evaluated using a paired t-test. Mutations in which cryptic splicing was either predicted or demonstrated experimentally were excluded to avoid biasing estimation of $\overline{\Delta R_i}$, since cryptic splicing can alter natural splice site use in the absence of a change in the information content of that site.

The observed distributions of the locations of cryptic donor and acceptor sites were compared with a model that assumes that these sites are equally likely to occur upstream or downstream of the natural site. Significance was evaluated with the binomial distribution.

Relationship of information content to splice site use

Different mutation reports measured splice site use directly by either cDNA sequencing, reverse transcription-PCR, primer-extension, S1 nuclease analyses, or allele-specific hybridization. Direct comparisons of natural and mutant splicing patterns were not always available. In some instances, the effect of the mutation was measured indirectly using either Northern hybridization (Table 1, #46, 47, 49; Table 3, #4), antigen immunoprecipitation or protein levels (Table 1, #18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 49; Table 2, #40, 41, 42, 43; Table 3, #2, 3, 4) or measurements of enzymatic activity (Table 1, #18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35; Table 2, #40, 41, 42, 43 ; Table 3, #2, 3). Functional analyses of splicing were not reported for mutations #31, 32, 54 and 55 in Table 1, #14, 15, 23, 34-38, and 44, 45, 46 in Table 2, and #1, 7 and 8 (the natural site at 2621) in Table 3.

RESULTS

Several categories of mutations were distinguished by individual information analysis. A total of 111 nucleotide substitutions were evaluated. Fifty-seven mutations were nucleotide substitutions that solely altered use of the natural splice site and did not create cryptic splice sites (designated as primary splice site mutations, Table 1). Activated cryptic splice sites were predicted for 46 different mutations, 33 of which were corroborated experimentally (Table 2). Eight nucleotide substitutions were predicted not to alter splicing (Table 3).

I. Primary mutations in splice junction recognition sequences

Differences in information content of natural and mutant splice sites. Many of the primary splice junction mutants that showed complete exon skipping (residual splicing: −) had R_i values ≤ 0 bits (Table 1, #2, 3, 11, 12, 15, 16, 17, 19, 35). However, there are primary mutant donor and acceptor sites that were not used that have mostly small positive R_i values (Table 1, #4, 5, 14, 20, 21, 24, 36, 38, 40, 43, 47). This suggests that recognition of splice donor and acceptor sites requires more than zero bits. ⇐Table 1

Mutations that reduce or completely abolish splicing have significantly lower R_i values than the corresponding natural sites. The average difference in R_i between primary mutant and natural donor sites is $\overline{\Delta R_i} = -7.67 \pm 3.95$ bits ($n = 45$) and for acceptor sites it is $\overline{\Delta R_i} = -5.97 \pm 3.50$ bits ($n = 12$), and these differences are significant ($p < 0.0001$ for both $\overline{\Delta R_i}$ values). R_i values of primary acceptor mutations range from a minimum of -2.90 bits to a maximum of 11.75 bits, whereas donor mutations have a lower range from -14.25 to 6.87 bits.

We considered the possibility that the strength of a natural splice site, *i.e.* R_i value, might be related to its susceptibility to mutational inactivation. 15 of 24 (62%) natural sites in Table 1 with R_i values $> R_{sequence}$ were inactivated by mutation or had mutant R_i values ≤ 0, compared to 22 of 29 (76%) natural sites with R_i values $< R_{sequence}$. Inactivation of splicing is primarily determined by the specific nucleotide substitutions that occur at those sites, however weak natural splice sites may be more susceptible than strong sites to succumb to mutations that abolish splicing.

Amount of information required for splicing. The minimum quantity of information required for splicing, $R_{i,min}$, was defined by comparing the R_i values of inactivating to leaky primary mutations (cryptic splicing mutations were excluded because activation of cryptic sites may affect natural site use). $R_{i,min}$ is bounded by the maximum information content of a non-functional site and the minimum quantity of information required to produce normal transcripts.

The following minimally functional sites had small positive R_i values: A mutation at the exon 5 donor site ($5.7 \rightarrow 2.6$ bits) in the HEXA gene results in a low level (3%) of normal mRNA (Table 1, #41). Similarly, a mutation at the exon 4 acceptor site ($10.8 \rightarrow 2.7$ bits) in the APOE gene results in 5% of normal splicing (Table 2, #2) and a mutation at the IVS 14 donor site ($6.3 \rightarrow 2.7$ bits) in COL1A1 decreases (by 50-60%) but does not abolish normal splicing (Table 1, #9). Furthermore, a mutant 2.4 bit acceptor site in the IDS gene (Table 2, #30) is associated with a moderately abnormal phenotype (the other allele is null), consistent with production of some normal mRNA. Finally, a mutation at the IVS 6 acceptor in COL1A2 reduces the R_i value of the splice site from 5.4 to 2.4 bits and results in a mild form of Ehlers-Danlos (type VII) syndrome due to 50% exon skipping (Table 1, #13; Fig. 1). ⇐Fig 1 Splicing at this site is completely impaired *in vitro* at 39 °C and restored at 30 °C. The temperature sensitivity of this mutation indicates that this 2.4 bit sequence is weakly bound by the spliceosome.

By contrast, mutations at the exon 1 donor splice site in the CAT gene (Table 1, #4; $6.0 \rightarrow 2.5$ bits), in IVS 33 of COL1A2 (Table 1, #14; $6.7 \rightarrow 3.2$ bits) completely abolish mRNA splicing. The R_i value of this COL1A2 mutation is inconsistent with the result found for mutation #13, since the mutation with lower information content would be expected to be inactive. This difference may not be significant depending on the (unknown) precision of the $R_{iw}(b,l)$ matrix, however it seems more likely that residual splicing at the mutated site in mutation #14 may not have been detected. Residual splicing was observed at several mutant splice sites with R_i values greater than 2.4 bits and less than 3.2 bits (Table 1, # 9, 41 and 52). These splice junction mutations define a range of values for $R_{i,min}$ of either donor or acceptor sites. Although the confidence interval around $R_{i,min}$ is unknown, donor and acceptor splice sites with $R_i > 2.4$ bits are rarely found in a set of random sequences with human dinucleotide composition ($p = 0.008$). To simplify comparisons between $R_{i,min}$ and other R_i values, we use $R_{i,min} \approx 2.4$ bits.

Leaky splicing To determine whether the information present in a mutant site was related to splice site use, the R_i values of mutated splice sites that inactivated splicing were compared with R_i values of leaky splice sites. Completely inactivated sites generally had R_i values less than $R_{i,min}$ (*e.g.* Table 1, #46), whereas mutations with R_i values greater than $R_{i,min}$ reduced but generally did not abolish splicing. For example, a G→C point mutation in the exon 2 donor site of the LFA1 gene (Table 1, #44) decreased R_i from 8.6 to 4.2 bits and this mutation is leaky, *i.e.* 3% of the normal spliced product is detected from this allele [Kishimoto et al., 1989]. Likewise, a patient with mild cholesterol storage disease was homozygous for a donor site mutation in the LIPA gene (8.8 → 5.7 bits; Table 1, #45; Fig. 2). ⇐Fig 2
Mutations #1, 6, 7, 9, 10, 13, 18, 22, 26, 34, 41, 44, 45, 50, 52, and 56 (Table 1) and #2, 3, 4, 7, 9, 16, 21, 23, 27, 28, 30, 32 and 41 (Table 2), which have R_i values $\geq R_{i,min}$, are leaky at the respective natural splice sites. The average decrease in R_i values is smaller for primary mutations that result in reduced levels of normally spliced mRNA; $\overline{\Delta R_i}$ is -2.92 ± 0.98 bits for donor sites ($n = 12$; *versus* -7.67 for all donor sites) and -4.25 ± 2.20 for acceptor sites ($n = 4$; *versus* -5.97 for all acceptor sites). When cryptic splice site mutations that result in residual splicing at the natural site are considered in addition, the change is negligible: $\overline{\Delta R_i} = -3.00 \pm 0.98$ bits ($n = 14$) for donor sites and $\overline{\Delta R_i} = -4.68 \pm 3.29$ bits ($n = 15$) for acceptor sites.

Quantitative relationship The quantitative relationship between splice site use and information content is illustrated by the polymorphic alleles in IVS 8 of the CFTR gene (Table 1, #6; Fig. 3). The frequency of exon 9 skipping is inversely related to the length ⇐Fig 3
of the polypyrimidine tract of the upstream acceptor site [Chu et al., 1993; Chillon et al., 1995; Rave-Harel et al., 1997]. This is not surprising since the length of a homopolymeric polypyrimidine tract has also been related to splice site strength [Dominski and Kole, 1991]. The 4.1 bit difference between the R_i values of the shortest and longest alleles accounts for the lower amount of spliced mRNA from the shorter allele and is probably related to the phenotype of congenital bilateral absence of the vas deferens in male homozygotes. A 4.1 bit reduction in information would correspond to at least a 17 fold ($2^{\Delta R_i} = 2^{4.1}$) decrease in splicing, assuming minimal conversion of information to energy dissipated [Schneider, 1991*b*; Schneider, 1994]. This corresponds closely to the relative amounts of mRNA produced by the shortest (5T) and longest (9T) alleles [Chillon et al., 1995].

Only two exceptional mutations were found in which $R_i \gg R_{i,min}$, although these sites were reportedly not used (Table 1, #5 [11.6 bits], #43 [5.7 bits]). The minimum predicted decreases of 3 and 11 fold, respectively, in binding affinity would not be expected to completely abolish splicing at these sites. Reduced amounts of splicing can occur at mutant splice sites with $R_i > R_{i,min}$, although a modest decrease in R_i at a splice site can apparently sometimes inactivate splicing.

II. Detection of cryptic splice sites

Categories of cryptic splice sites R_i analysis detected secondary cryptic splice sites that are activated by mutation in or adjacent to the natural primary splice site. This indicates that the R_i values of activated cryptic sites may be determined with an information model derived from natural splice sites [Stephens and Schneider, 1992]. Table 2 shows 33 experimentally-identified cryptic sites confirmed by information analysis of the respective genomic sequences (section A), and 13 mutations that were predicted by R_i analysis to exhibit cryptic splicing (section B). For example, a mutation at position 35066 of the adenosine deaminase gene (Table 2, #1) does not alter the R_i value of the natural splice site (at 35099), but creates a secondary cryptic site of similar strength at position 35067. There were 7 additional mutations in which a new cryptic site was either created or predicted without altering the R_i value of natural splice site (Table 2, #12, 14, 15, 26, 31, 40, 43). Activation of cryptic sites can also prevent splicing at natural sites by promoting exon skipping (*e.g.* in 79% of transcripts resulting from a mutation in the iduronate-2-sulfatase gene; Table 2, #26; [Jonsson et al., 1995]). Exon skipping mutations occurred predominantly at donor splice sites (7 of 8) and in each instance, a cryptic site was created upstream whose R_i value exceeded or was similar to that of the natural site. ⇐Table 2

Several types of cryptic splicing mutations were distinguished:

1. The most common category ($n = 17$) showed a concerted increase in information at the cryptic site ($\overline{\Delta R_i} = +6.17 \pm 2.94$ bits) accompanied by a reduction in the R_i value at the natural site ($\overline{\Delta R_i} = -5.92 \pm 3.09$ bits). All of these were acceptor sites (Table 2, #7, 23, 25, 27, 29, 30, 32, 34, 36, 37, 38, 39, 41, 42, 44, 45, 46). The distance between these cryptic and natural splice sites is, on average, 4.3 nucleotides, which would be expected for a mutation that simultaneously alters the R_i values of both sites. Detection of cryptic sites that overlap the natural site requires sequence analysis of the mRNA, since changes in the size and sequence of the processed transcript are subtle. Use of these cryptic sites would either alter the reading frame or insert or delete one or more codons (*e.g.* Fig. 4). ⇐Fig 4

2. Novel cryptic sites were created simultaneously with either missense mutations (Table 2, #4, 12, 14, 15) or silent coding substitutions (Table 2, #31). By creating a cryptic site, some of these coding sequence substitutions (Table 2, #35, 36, 37, 38, 40, 43) could also inactivate the natural splice junction or cause frame shifting instead of exon skipping. Cryptic sites that generate mRNAs with in-frame insertions or deletions can also be recognized by R_i analysis [Allikmets et al., 1998].

3. Mutations that decreased the R_i value of the natural site resulted in the use of pre-existing cryptic sites with R_i values in the normal range (Table 2, #2, 3, 9, 10, 11, 13, 17, 18, 19, 20, 21, 22, 24, 28, 33). Some residual splicing may occur at a mutated natural site when the sequence change produces mutant and cryptic sites with similar R_i values (*e.g.* Table 2, #7). Natural and cryptic sites compete with each other [Treisman

et al., 1983; Orkin et al., 1982] when the natural site exhibits either a moderate or no reduction in R_i.

Susceptibility to activation Of 31 experimentally-verified cryptic splicing mutations (Table 2A, excluding #5 and 6), there are 19 splice sites whose R_i values exceeded the cryptic site prior to its activation ($\overline{\Delta R_i} = 6.65 \pm 3.65$ bits). For the remaining 12 mutations (10 of which involve the same site in HBB), the inactive cryptic sites exceed the natural site by only an $\overline{\Delta R_i}$ of 1.66 ± 0.66 bits. Furthermore, the differences in R_i values between natural and cryptic sites prior to mutational activation are much smaller for donor sites ($\overline{\Delta R_i} = 1.25 \pm 4.68$, $n = 17$ for donors *vs.* $\overline{\Delta R_i} = 7.03 \pm 3.59$, $n = 15$ for acceptors). Likewise, cryptic donors were activated by an increase of $\overline{\Delta R_i} = 3.12 \pm 2.85$ bits ($n = 5$), whereas cryptic acceptor sites were activated by $\overline{\Delta R_i} = 5.86 \pm 3.27$ bits ($n = 10$). From these observations it would appear that donor sites may be more susceptible to the effects of neighboring cryptic sites.

Distance effects Cryptic sites activated by a mutation that weakens the natural site must reside within a few hundred nucleotides of the natural splice site, since the novel exon is restricted in length [Hawkins, 1988; Berget, 1995]. For example, a strong cryptic acceptor in intron 2 of the β-hemoglobin gene is activated by mutations at the exon 3 acceptor 271 nucleotides downstream (Table 2, #24). Mutation at a natural site can, however, activate sites that are further away when a cryptic exon is created. For example, mutation at the exon 3 acceptor of the CFTR gene activates a cryptic, non-coding exon in intron 3 (2,354 nucleotides downstream of exon 3 and 19,329 nucleotides upstream of exon 4; Table 2, #3).

Exceptions Although pre-existing or novel cryptic sites with R_i values less than that of the strongest local splice site were usually not recognized, there were exceptions. Infrequently, a weaker cryptic site can interfere with a natural site, even when the natural site is strengthened by the mutation (*e.g.* Table 2, #16). For example, activated cryptic sites with R_i values lower than those of the natural splice site after mutation may sometimes be used (Table 2, #1, 3, 4, 6, 9, 16, 23, 32). In at least one instance (Table 2, #1), a cryptic acceptor site upstream of the natural site is predominantly used despite the fact that both sites have similar R_i values, which suggests that the cryptic site is recognized first. Conversely, the R_i value of the exon 1 donor in the β-globin gene is less than that of an upstream cryptic site (Table 2, #12-15, 17-22), however this cryptic site is not activated unless it is strengthened or the donor is weakened. These exceptions suggest that besides direct competition between the cryptic and natural splice sites, other factors can influence splice site selection.

Another class of exceptional splice sites were those that generated alternatively processed transcripts. Active "cryptic" sites that resided in introns of the CSPB gene had R_i values in the normal range (Table 2, #5, 6) [Trapani et al., 1988; Klein et al., 1989]. They may represent alternative splice sites regulated by other sequence elements that can be present in the adjacent exons [Lavigueur et al., 1993; Sun et al., 1993*b*; Dirksen et al., 1994; Huh and Hynes, 1994; Humphrey et al., 1995] or polypyrimidine tracts [Sun et al., 1993*b*; Wang et al., 1995].

III. Non-deleterious splice site substitutions

Nucleotide substitutions that do not significantly alter the R_i value of a natural site are expected to produce functional rather than mutant sites [Rogan and Schneider, 1995]. Given that such substitutions are not likely to be deleterious, they may be polymorphic in the germline, as has been shown for a sequence change in an hMSH2 splice acceptor site [Leach et al., 1993]. We identified other nucleotide substitutions that did not significantly alter the R_i value (Table 3): ⇐Table 3

1. Reported mRNA analyses of substitutions #4 and 5 did not reveal splicing defects that altered either the size, structure or quantity of these transcripts although these changes had been suggested to affect splicing [Carstens et al., 1991; Higashi et al., 1988; Speiser et al., 1992; Owerbach et al., 1992; Barbat et al., 1995].

2. A C→G substitution 12 nucleotides upstream of the IVS 2 acceptor of the CYP21 gene (Table 3, #6) decreases in R_i value by only 1.56 bits and mRNA of normal size and quantity was present [Higashi et al., 1988]. Asymptomatic individuals with this sequence have been reported [Day et al., 1996; Schulze et al., 1995; Day et al., 1995] and a comparable ΔR_i results from a benign C→A polymorphism at the same position (Table 3, #5).

3. An A→G substitution at the exon 7 donor site of the OTC gene was suggested to cause exon skipping, however Northern analysis did not show either the size or quantity of mRNA to differ from controls and the change in R_i was negligible (Table 3, #4). Since OTC protein was not detected, this patient may harbor a mutation elsewhere.

Splicing patterns for several nucleotide substitutions #1, 2, 3, and 7 (Table 3) were not reported, however, based on information analysis, these changes would not be predicted to alter mRNA splicing. The substitutions either maintain or increase the information content of the natural splice site. The R_i values of the proposed cryptic sites for substitutions #1, 2, and 8 were either negative or unchanged, suggesting that they are not activated by these substitutions. A proposed cryptic site in exon 3 of the p53 gene (substitution #7) is significantly weaker than the natural acceptor site (by 6.14 bits) and has an R_i value only slightly larger than $R_{i,min}$. It would seem unlikely that this cryptic site is preferentially used.

DISCUSSION

The number of bits in a splice site is related to the amount of splicing at that site. Previously, we demonstrated that a polymorphic splice junction variant caused little change in information [Rogan and Schneider, 1995]. The present study extends this finding and shows that mutant splice sites often contain significantly less information than their corresponding natural sites. Further, cryptic splice sites are activated by increases in information or by decreases at the natural splice site, and the information at activated cryptic sites is often comparable to or exceeds the natural site.

Predicting the effects of mutations A required step of information analysis is to compute the total information over all positions in a site. This value must then be compared with that of other sites prior to concluding that a substitution that changes a positive to a negative weighting is deleterious (compare Tables 1 and 2 to Table 3). Functional splice sites can have nucleotides with negative weightings (*e.g.* Fig. 1, position 63) that are offset by strong contributions at other positions (*e.g.* Fig. 1, position 64), as we have shown for other binding sites (Figure 2 in Schneider, 1997*b*; Hengen et al., 1997). Statistical analyses of the distributions of point mutations in splice sites are useful [Krawczak et al., 1992] but can sometimes obscure these compensating effects. Within a binding site, the context of a mutation can be as important as the mutation itself.

The difference between the observed value of $R_{i,min}$ (~ 2.4 bits) and its expected value (zero bits) may have a biological basis. However, this difference could also be explained by errors in the database used to create the splice weight matrices [Schneider, 1997*b*], statistical limitations of the data and matrices, motifs that are different from the majority of sites [Hall and Padgett, 1994], or intrinsic limits to the precision of splice site recognition [Schneider, 1991*a*]. Although the standard deviation of $R_{sequence}$ can be determined [Stephens and Schneider, 1992], the confidence intervals on individual R_i values are unknown. These intervals are expected to be larger at the lower and upper bounds of the R_i distribution, where fewer functional splice sites are observed. The existence of a natural site with $R_i < R_{i,min}$ (2.2 bits; Table 2, #26) and an exon-skipping mutation with $R_i > R_{i,min}$ (3.2 bits; Table 1, #14) suggests that $R_{i,min}$ is not known precisely. The error ($|R_i - R_{i,min}|$) may be as little as 0.2 bits ($R_i = 2.2$ bits; Table 2, #26), but it might be as much as 2.4 bits ($R_i = 0$ bits; Schneider, 1997*a*).

Susceptibility to mutation Donor sites may be more susceptible to inactivation than acceptor sites. The R_i values of mutant donor sites are more likely than mutant acceptors to be less than $R_{i,min}$. Natural donors possess less information than acceptors [Stephens and Schneider, 1992] and the average decrease in information due to mutation at donor sites exceeds the reduction in R_i at acceptors. Information is also less densely distributed across acceptor splice sites (0.3 bits per nucleotide) than in donor sites (0.8 bits per nucleotide), so changes at acceptors often have a smaller effect on R_i. Significantly more primary mutations in donor sites ($n = 45$) than acceptor sites ($n = 12$) were found, as has been noted [Krawczak et al., 1992; Nakai and Sakamoto, 1994].

Cryptic splicing The R_i values of most novel cryptic donor sites exceeded or were similar to those of the corresponding natural sites. Although similar results were also inferred from Shapiro-Senapathy consensus values [Krawczak et al., 1992], information analysis detects fewer incorrect cryptic splice sites [O'Neill et al., 1998], more accurately discriminates true sites from non-sites, and visually depicts both changes (Fig. 4).

An exon is initially defined by recognizing the acceptor [Berget, 1995]. Cryptic acceptor sites occur either upstream ($n = 9$) or downstream ($n = 7$) of the natural site ($p = 0.4$), suggesting that they are not located by scanning [Stephens and Schneider, 1992]. The exon definition model predicts that the spliceosome then scans downstream until a strong donor site is located [Robberson et al., 1990; Niwa et al., 1992], so a novel cryptic donor site created downstream of an intact natural site should not be recognized unless the natural site is mutated. In all cases, a decrease in the information content of the natural donor site activated pre-existing cryptic sites downstream (Table 2A). Furthermore, cryptic donor sites were activated more frequently upstream of the natural site (15 of 20; $p = 0.02$). The idea that the splicing machinery selects for the strongest local acceptor splice site and scans for donors is supported by R_i analysis.

Nucleotide substitutions within 17 natural acceptor sites have been shown to create or strengthen adjacent cryptic sites that are thereby activated (see results: II. Detection of cryptic splice sites). Only acceptors were found, perhaps because the variable polypyrimidine tract potentiates spliceosome recognition at many positions, whereas donor sites have high information density and a non-repeating sequence pattern [Stephens and Schneider, 1992]. For this reason, weaker cryptic sites are often found near natural acceptor sites (*e.g* Fig. 4). Mutations involving the natural acceptor sometimes strengthen and activate these cryptic sites. The resulting aberrant exons may in some cases have been misidentified as natural splice products (*e.g.* Table 2B), since their length and sequence would differ by only a few nucleotides from the normal mRNA.

Conclusion We have shown that individual information theory can be used to rank normal and mutant splice junctions. As a consequence, silent polymorphisms can be distinguished from true mutations, changes in individual information are related to splice site use, and activated cryptic splice sites can be detected. These distinctions are possible because the information measure is related to the thermodynamic entropy, and therefore can be connected to the binding energy [Szilard, 1964; Schneider, 1991*a*; Schneider, 1991*b*; Schneider, 1994]. The information in the splice site should be related to the specific binding interaction between the spliceosome and the site [Berg and von Hippel, 1987; Berg and von Hippel, 1988*a*; Berg, 1988; Berg and von Hippel, 1988*b*]. However, the relationship is an inequality—the second law of thermodynamics [Schneider, 1991*b*; Schneider, 1994]—and can only be explored empirically at this stage. The correlation between information measures and measured thermodynamic parameters is expected to more precisely relate genotypes to phenotypes in genetic disorders.

ACKNOWLEDGEMENTS

We thank Greg Alvord for statistical consulting, and Kenn Kraemer and Howard Young for reading the manuscript. Grant support is acknowledged from the Public Health Service (CA74683) and the American Cancer Society (DHP-132) to P.K.R. We thank the Frederick Biomedical Supercomputing Center for access to computer resources and support services.

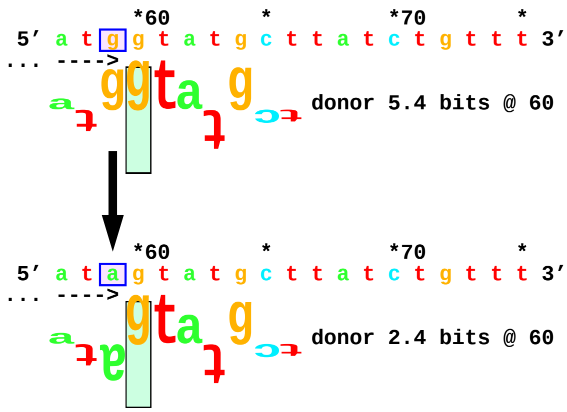

Figure 1: A primary splice junction mutation represented by sequence walkers. A G→A mutation 1 nucleotide upstream of the exon 6 donor of the COL1A2 [GenBank accession number M35391] gene results in 50% exon skipping and Ehlers-Danlos syndrome, Type VII (Table 1, #13). This substitution, which significantly reduced the R_i value, defines the lower threshold of information required for splice site recognition since it is temperature sensitive, being non-functional at 39 °C but functional at 30 °C. The splice sites are shown by walkers [Schneider, 1997*b*] in which the height of a letter is the contribution of that base to the total conservation of the site. The upper bound of the vertical rectangles is at +2 bits, and their lower bound is at −3 bits. Letters that are upside down and point downwards represent negative contributions. The upper walker shows the normal site; the lower one displays the mutant sequence. The black arrow shows the position of the mutation (boxed). The dashed arrow represents the coding region.

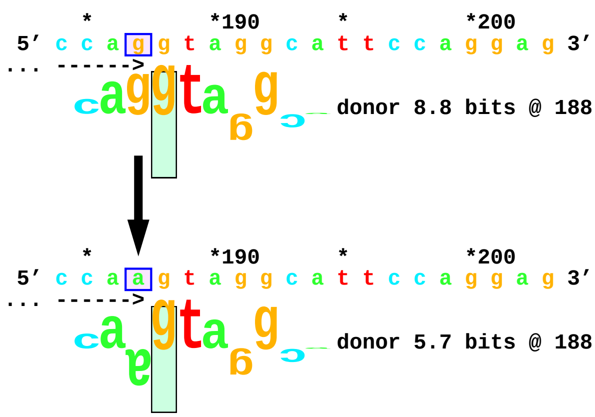

Figure 2: A leaky splice junction mutation.
A G→A mutation 1 nucleotide upstream of the exon 8 donor site of the lysosomal lipase gene [LIPA; U04292] results in mild cholesterol ester storage disease with 4-9% enzymatic activity (Table 1, #45). The reduction in information content is significant even though the R_i value is still much greater than $R_{i,min}$.

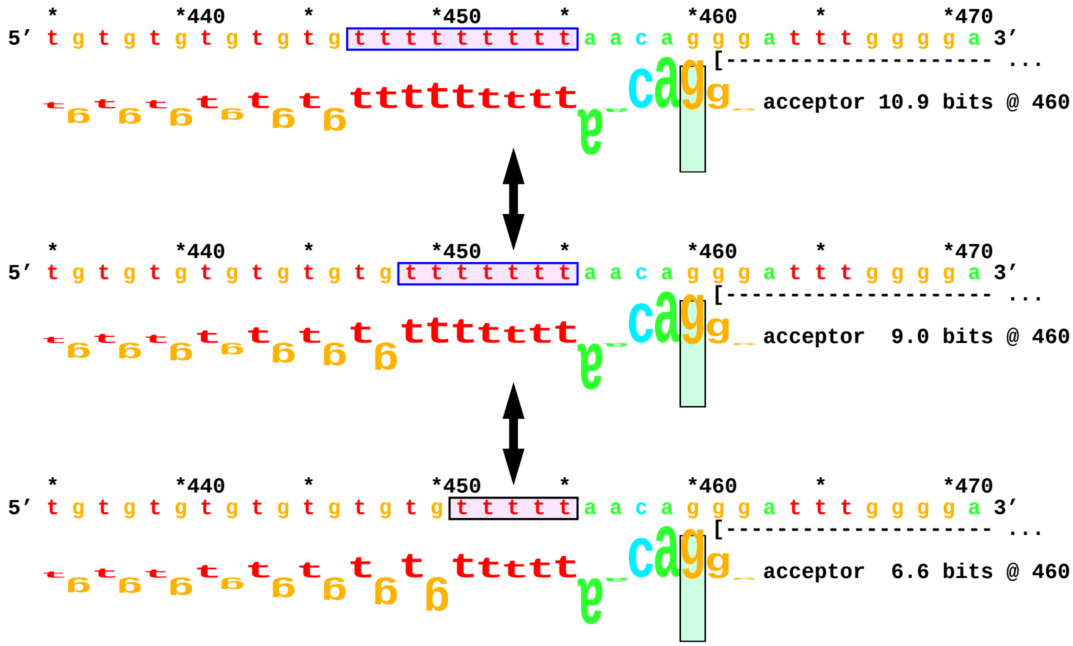

Figure 3: Polymorphic variation that affects splicing.

Splicing varies among 3 common alleles that differ in length in the polymorphic polythymidine tract of the IVS 8 acceptor of the gene encoding the cystic fibrosis transmembrane regulator [CFTR; M55114] (Table 1, #6). The shortest allele (bottom walker) shows 90% outsplicing of exon 9 and is associated with congenital absence of the vas deferens. Individuals with the two longer alleles have a normal phenotype, although the 7T allele produces less mRNA than the 9T allele. Exon 9 begins at the base indicated by the left bracket and dashes.

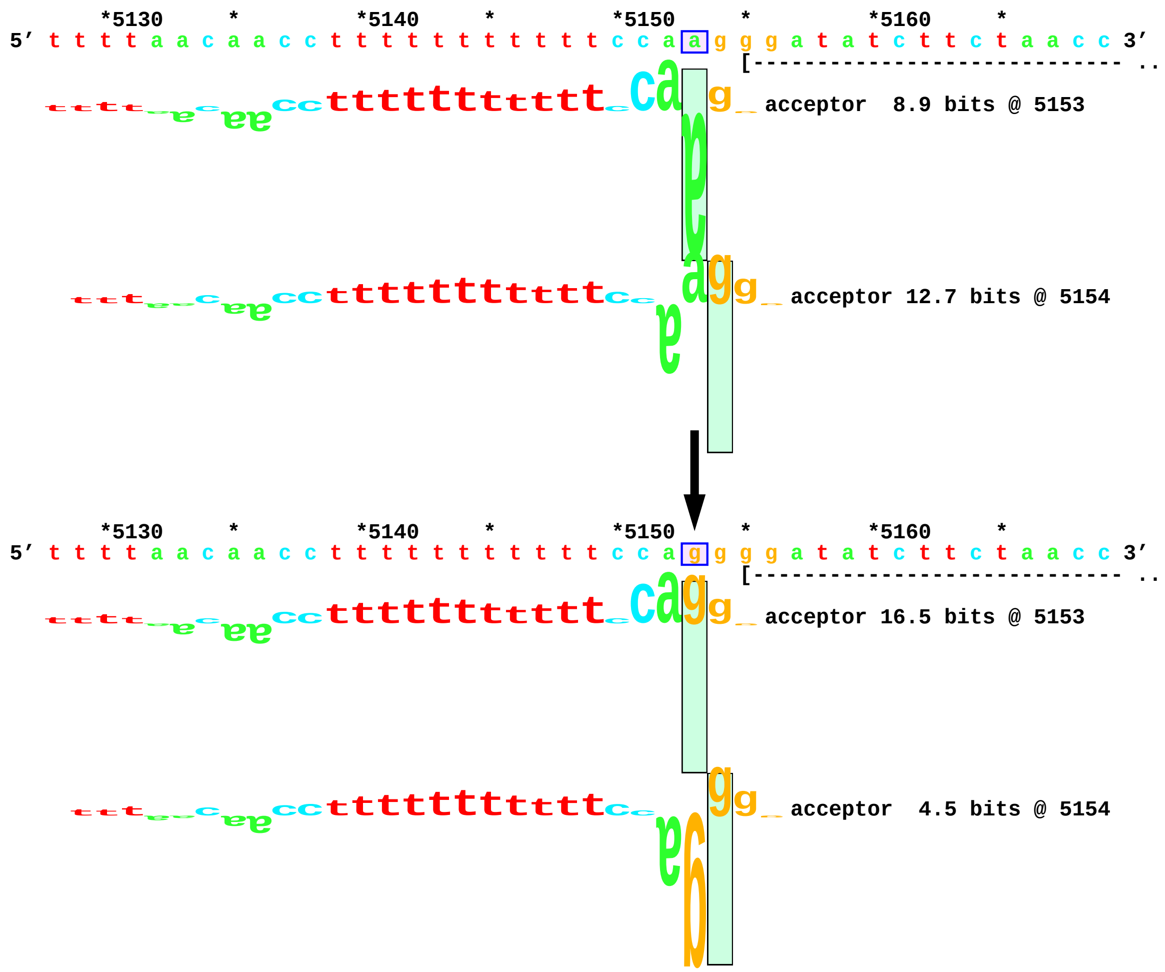

Figure 4: Cryptic site creation concurrent with mutation of the natural site.
An A→G mutation in intron 3 of the iduronidase synthetase gene [IDS; L35485] significantly decreases the information content of the IVS 3 acceptor while simultaneously creating a strong cryptic site at the position of the mutation, 1 nucleotide upstream from the natural splice junction (Table 2, #27). The upper two walkers show a pre-existing cryptic site at position 5153 and a natural site at 5154. The lower two walkers show the activated cryptic site at 5153 and the mutant site at 5154. For simplicity, only sites with greater than 4.3 bits are shown. In addition, a 4.2 bit site that is not used at position 5155, is reduced to 2.5 bits as a consequence of the mutation. The lower bound of the vertical rectangles is at −7 bits.

Table 1: Information analysis of primary splice site mutations

#	Gene [Accession]	Mutant allele, coordinate [1]	Natural site coord.	$R_{i,natural} \rightarrow R_{i,mutant}$ [2]	Residual splicing [3]	Reference(s)
1	ADA [M13792]	IVS 5, donor, T→A, 29944	29939	7.25 → 5.66	+	Santisteban et al., 1995
2	ADA [M13792]	IVS 8/Exon 9, acceptor, indel 32839-32851	32850	8.30 → -1.99	−	Arredondo-Vega et al., 1994
3	ADA [X02190]	IVS 2, donor, G→A, 108	108	12.52 → -0.28	−	Arredondo-Vega et al., 1994
4	CAT [X04088]	IVS 4, donor, G→A, 153	149	6.03 → 2.51	−	Wen et al., 1990; Kishimoto et al., 1992
5	CFTR [M55108]	IVS 2, acceptor, C→T, 182	184	13.24 → 11.58	−	Bienvenu et al., 1994
6	CFTR [M55114]	IVS 8, 5T (mutant acceptor) [4]	458	$R_i = 6.56$	+	Chu et al., 1993
		IVS 8, 7T (normal acceptor)	460	$R_i = 8.97$	+	Chillon et al., 1995
		IVS 8, 9T (normal acceptor)	462	$R_i = 10.62$	+	Rave-Harel et al., 1997
7	CFTR [M55127]	Exon 20, donor, G→C, 422	422	10.91 → 6.87	+	Jones et al., 1992
8	CFTR [M55118]	IVS 13, donor, G→A, 901	901	10.01 → -2.79	−	Audrezet et al., 1993
9	COL1A1 [M20789]	IVS 14, donor, G→A, 5218	5214	6.23 → 2.71	+	Bonadio et al., 1990, Sakuraba et al., 1992
10	COL1A1 [M20789]	Exon 6, donor, G→A, 3170	3171	8.42 → 5.36	+	Weil et al., 1989*a*; Sakuraba et al., 1992
11	COL1A2 [M35391]	IVS 6, donor, G→T, 60	60	5.41 → -2.39	−	Vasan et al., 1991; Watson et al., 1992; Lehmann et al., 1994
12	COL1A2 [M35391]	IVS 6, donor, C→T, 61	60	5.41 → -2.06	−	Weil et al., 1990; Ho et al., 1994

[1] The coordinate is the numerical location of the base in GenBank sequence [Schneider et al., 1982]. IVS and exon indicate the intronic or exonic location of the mutation.

[2] $R_{i,natural}$: the individual information value of the natural splice site; $R_{i,mutant}$: the individual information value of the mutated splice site

[3] + : mutation does not abolish natural splice site use (see methods); − : absence of normally spliced mRNA or functional protein; n.i.: no information reported

[4] #6; polymorphic alleles; 5T, 7T, 9T, refer to the length of the polythymidine tract

#	Gene [Accession]	Mutant allele, coordinate	Natural site coord.	$R_{i,natural} \rightarrow R_{i,mutant}$	Residual splicing	Reference(s)
13	COL1A2 [M35391]	IVS 6, donor, G→A, 59	60	5.41 → 2.35	+	Weil et al., 1989*b*
14	COL1A2 [M64229]	IVS 33, donor, G→A, 346	342	6.72 → 3.20	–	Ganguly et al., 1991
15	COL3A1 [M55603]	IVS 41, donor, G→A, 62	62	12.17 → -0.62	–	Cole et al., 1990
16	DMD [L05639]	IVS 26, donor, T→G, 307	306	7.46 → -0.73	–	Wilton et al., 1994
17	DMD [M86892]	IVS 68, donor, T →A, 259	258	4.52 → -3.26	–	Roberts et al., 1992; Roberts et al., 1993*b*
18	F9 [K02402]	Exon 3, donor, G→A, 9667	9668	7.59 → 4.54	+	Giannelli et al., 1991
19	F9 [K02402]	IVS 1, donor, del 3076-3085	3083	4.60 → -14.25	–	Giannelli et al., 1991
20	F9 [K02402]	IVS 1, donor, G→A, 3087	3083	4.60 → 1.08	–	Giannelli et al., 1991
21	F9 [K02402]	IVS 2, donor, del 9457-9460	9455	7.09 → 0.26	–	Giannelli et al., 1991
22	F9 [K02402]	IVS 2, donor, T→C, 9460	9455	7.09 → 5.67	+	Bottema et al., 1990
23	F9 [K02402]	IVS 3, acceptor, G→A, 13356	13356	5.26 → -2.32	–	Giannelli et al., 1991
24	F9 [K02402]	IVS 3, donor, T→C, 9669	9668	7.59 → 0.12	–	Giannelli et al., 1991
25	F9 [K02402]	IVS 3, donor, T→G, 9669	9668	7.59 → -0.61	–	Giannelli et al., 1991
26	F9 [K02402]	IVS 4, acceptor, del 20625-20628	20633	11.20 → 8.46	+	Giannelli et al., 1991
27	F9 [K02402]	IVS 5, donor, G→T, 20763	20763	2.42 → -5.37	–	Giannelli et al., 1991
28	F9 [K02402]	IVS 6, donor, G→A, 23531	23531	5.21 → -7.58	–	Giannelli et al., 1991
29	F9 [K02402]	IVS 6, donor, G→T, 23531	23531	5.21 → -2.58	–	Giannelli et al., 1991
30	F9 [K02402]	IVS 7, acceptor, G→A, 33786	33786	4.68 → -2.90	–	Giannelli et al., 1991
31	FGFR2 [M80635]	IVS A, acceptor, T→G, 67	69	13.97 → 9.65	n.i.	Schell et al., 1995
32	FGFR2 [M80635]	Exon B, acceptor, G→T, 70	69	13.97 → 11.75	n.i.	Schell et al., 1995
33	GBA/GCB [J03059]	IVS 2, donor, G→A, 1942	1942	12.73 → -0.07	–	He and Grabowski, 1992
34	GH [J03071]	IVS 3, donor, T→C, 5990	5985	5.06 → 3.64	+	Cogan et al., 1993
35	GH [J03071]	IVS 4, donor, G→C, 6242	6242	8.06 → -1.74	–	Cogan et al., 1993
36	GH [J03071]	IVS 4, donor, G→T, 6242	6242	8.06 → 0.25	–	Phillips 3rd and Cogan, 1994
37	GLA [X14448]	IVS 2, donor, T→G, 5270	5269	6.88 → -1.32	–	Eng et al., 1993
38	GLA [X14448]	IVS 6, donor, G→T, 10708	10708	9.21 → 1.41	–	Sakuraba et al., 1992
39	GYPB [M24135]	Exon 3, donor, C→A, 277 & IVS 3, G→T, 280	280	9.10 → -0.88	–	Kudo and Fukuda, 1989

#	Gene [Accession]	Mutant allele, coordinate	Natural site coord.	$R_{i,natural} \rightarrow R_{i,mutant}$	Residual splicing	Reference(s)
40	HBB [V00499]	IVS 1, acceptor, G→C, 375 [5]	375	9.40 → 2.11	–	Renda et al., 1992
41	HEXA [M16415]	Exon 5, donor, G→A, 127	127	5.70 → 2.64	+	Ozkara et al., 1995
42	HEXA [M16422]	IVS 12, donor, G→C, 107	107	9.78 → -0.01	– [6]	Ohno and Suzuki, 1988
43	HPRT [M26434]	IVS 8, donor, G→A, 40115	40111	9.26 → 5.74	–	Gibbs et al., 1990
44	LFA1 [S75381]	IVS 2, donor, G→C, 18	18	8.64 → 4.24	+	Kishimoto et al., 1992
45	LIPA [U04292]	Exon 8, donor, G→A, 187	188	8.75 → 5.69	+	Klima et al., 1993; Muntoni et al., 1995
46	LPL [S71696]	IVS 1, donor, G→C, 14	10	9.72 → -3.07	–	Chimienti et al., 1992
47	LPL [S71696]	IVS 2, acceptor, G→A, 40	40	7.66 → 0.08	–	Hata et al., 1990
48	NF1 [U17681]	IVS 18, donor, G→T, 1429	1429	10.20 → -2.60	–	Purandare et al., 1995
49	OTC [D00227]	IVS 7, donor, T→C, 78	77	6.52 → -0.94	–	Carstens et al., 1991
50	PBGD [M18799]	Exon 1, donor, G→T, 494	495	9.54 → 6.23	+	Grandchamp et al., 1989*a*
51	PBGD [M18799]	IVS 1, donor, G→A, 495	495	9.54 → -3.25	–	Grandchamp et al., 1989*b*
52	PFKM [S70308]	IVS 6, acceptor, A→C, 263	264	10.20 → 2.78	+ [7]	Tsujino et al., 1994
53	RB [M27853]	IVS 10, donor, G→T, 376	376	4.28 → -3.52	– [8]	Yandell et al., 1989
54	RB [M27860]	IVS 19, donor, T→C, 461	460	8.07 → 0.60	n.i.[8]	Horowitz et al., 1989
55	RB [M27862]	IVS 20, acceptor, A→G, 258	259	5.36 → 2.14	n.i.	Yandell et al., 1989
56	VWF [M25864]	IVS 50, donor, G→T, 915	913	9.41 → 5.31	+	Mertes et al., 1994
57	WT1 [X51630]	Exon 3, donor, del 1594-1619	1600	7.41 → -5.33	– [8]	Haber et al., 1990

[5] #40; natural cryptic site at coordinate 382 is strengthened by mutation, 3.00 → 4.99 bits
[6] #42: both exon skipping and intron retention observed
[7] appears to activate cryptic sites in exon 7 of 1.61 and 3.21 bits (at positions 20 and 27, respectively [Accession # M59724])
[8] #53, 54, 57: early onset; however, tumors were of somatic origin

Table 2: Information analysis of mutations that result in the use of secondary cryptic sites

#	Gene [Accession]	Mutation, coordinate	Natural site(s) coord.	Natural site(s) $R_{i,natural} \rightarrow R_{i,mutant}$ [1]	Cryptic site(s) coord.	Cryptic site(s) $R_{i,natural} \rightarrow R_{i,mutant}$ [1]	Residual natural splicing	Reference(s)
A. Experimentally-verified cryptic sites								
1	ADA [M13792]	IVS 10, acceptor, G→A, 35066	35099	9.99 → 9.99	35067	1.14 → 9.30	+	Santisteban et al., 1995
2	APOE [M10065]	IVS 3, acceptor, A→G, 3779	3780	10.81 → 2.65	3726	8.37 → 8.37	+	Cladaras et al., 1987
3	CFTR [M55109],	IVS 3, acceptor, G→T, 253	252	14.80 → 12.60	495 [2]	2.91 → 2.91	+	Will et al., 1994
	[L25269]	donor			677 [3] [4]	4.55 → 4.55		
4	CFTR [M55127]	IVS 20, donor, G→C, 422	423	10.91 → 6.87	451	1.15 → 1.15	+	Jones et al., 1992
5	CSPB [J03072]	IVS 1, acceptor	1289	6.42 → 6.42	1141	13.42 → 13.42	n.a.[5]	Trapani et al., 1988, Klein et al., 1989
6	CSPB [J03072]	IVS 2, donor	1438	10.95 → 10.95	1556	8.30 → 8.30	n.a.[5]	Trapani et al., 1988, Klein et al., 1989
7	DMD [L05648]	IVS 57 acceptor, G→C, 132	133	11.89 → 4.61	142	3.04 → 4.61	+	Roberts et al., 1993*a*
					131	4.51 → 4.09 [6] [7]		
					126	3.41 → 3.41 [7]		
8	F12 [M17466]	IVS 13, acceptor, G→A, 3622	3622	5.81 → -1.76	3623	-5.06 → 3.10	−	Schloesser et al., 1995
9	GAA [X55080],	IVS 1, acceptor, T→G, 142	154	13.78 → 11.83	437	9.51 → 9.51	+	Boerkoel et al., 1995, Huie et al., 1994
		IVS 1, cryptic acceptor			528	9.06 → 9.06		
	[X55079]	IVS 1, cryptic donor			1018	3.47 → 3.47		
10	GCK [M93280]	IVS 4, donor, del 4313-4327	4312	7.97 → -3.84	4288	5.55 → 5.55	−	Sun et al., 1993*a*

[1] When R_i values are the same, the substitution has not affected that site.
[2] cryptic acceptor site created as part of a cryptic exon which begins at 491 in L25269 and terminates at 677 in L25269.
[3] cryptic donor site activated in conjunction with the above cryptic acceptor.
[4] #3; polymorphic sequence: $R_i = 3.72$ bits for same cryptic donor site in [HSAC000111] at coordinate 51277
[5] #5, 6; n.a.: not applicable; natural cryptic site; no mutation occurs
[6] #7; reported cryptic site not present, predicted site at coord. 142 would also produce in-frame deletion (of 3 rather than 6 amino acids)
[7] #7, 24, 25; predicted by R_i analysis

#	Gene [Accession]	Mutation, coordinate	Natural site(s) coord.	Natural site(s) $R_{i,natural} \rightarrow R_{i,mutant}$	Cryptic site(s) coord.	Cryptic site(s) $R_{i,natural} \rightarrow R_{i,mutant}$	Residual natural splicing	Reference(s)
11	GH [J00148]	IVS 2, donor, G→A, 762	762	3.74 → -9.05	781	3.96 → 3.96	−	MacLeod et al., 1991
12	HBB [V00499]	Exon 1, donor, G→A, 232	246	5.66 → 5.66	230	7.61 → 8.01	+	Orkin et al., 1982
13	HBB [V00499]	Exon 1, donor, G→C, 245	246	5.66 → 1.62	230	7.61 → 7.61	−	Treisman et al., 1983, Vidaud et al., 1989
14	HBB [V00499]	Exon 1, donor, G→T, 235	246	5.66 → 5.66	230	7.61 → 8.83	n.i.	Orkin et al., 1982
15	HBB [V00499]	Exon 1, donor, T→A, 228	246	5.66 → 5.66	230	7.61 → 9.73	n.i.	Goldsmith et al., 1983
16	HBB [V00499]	IVS 1, acceptor, G→A, 355	376	9.40 → 9.69	355	1.44 → 4.89	+	Spritz et al., 1981
17	HBB [V00499]	IVS 1, donor, G→A, 250	246	5.66 → 2.14	230	7.61 → 7.61	−	Lapoumeroulie et al., 1987
18	HBB [V00499]	IVS 1, donor, G→C, 246	246	5.66 → -4.13	230	7.61 → 7.61	−	Vidaud et al., 1989
					208	7.63 → 7.63		
					258	2.53 → 2.53		
19	HBB [V00499]	IVS 1, donor, G→C, 250	246	5.66 → 1.71	230	7.61 → 7.61	−	Treisman et al., 1983
20	HBB [V00499]	IVS 1, donor, G→T, 250	246	5.66 → 1.75	230	7.61 → 7.61	+	Atweh et al., 1987
21	HBB [V00499]	IVS 1, donor, T→C, 251	246	5.66 → 4.24	230	7.61 → 7.61	+	Treisman et al., 1983
22	HBB [V00499]	IVS 1, donor, T→G, 247	246	5.66 → -2.54	230	7.61 → 7.61	−	Chibani et al., 1988
23	HBB [V00499]	IVS 1, acceptor, T→G, 361	375	9.40 → 7.72	361	-3.66 → 5.08	+ [8]	Metherall et al., 1986
24	HBB [V00499]	IVS 2, acceptor, A→G, 1447	1448	13.33 → 5.17	1177	9.69 → 9.69	−	Atweh et al., 1985
					1446	7.05 → 7.05 [7]		
25	HPRT [M26434]	IVS 8, acceptor, ATA→TTT, 41451-41453	41454	8.85 → 1.49	41471	2.66 → 4.65	−	Gibbs et al., 1989; Gibbs et al., 1990
					41457	2.75 → 7.73 [7]		
26	IDS [L35485]	Exon 3, donor, C→G, 2858	2882	2.19 → 2.19	2856	2.45 → 6.85	−	Jonsson et al., 1995

[8] No splicing data, but patient had β-thalassemia intermedia; the other allele is null. This implies that residual splicing occurs at the natural site.

#	Gene [Accession]	Mutation, coordinate	Natural site(s) coord.	$R_{i,natural} \to R_{i,mutant}$	Cryptic site(s) coord.	$R_{i,natural} \to R_{i,mutant}$	Residual natural splicing	Reference(s)
27	IDS [L35485]	IVS 3, acceptor, A→G, 5153	5154	12.70 → 4.54	5153	8.91 → 16.49	+	Bunge et al., 1993
28	IDS [L35485]	IVS 6, acceptor, A→G, 15750	15751	12.60 → 4.39	15802	5.91 → 5.91	+	Bunge et al., 1993
29	IDS [L35485]	IVS 7, acceptor, G→C, 19093	19093	4.35 → -2.94	19105	-0.15 → 1.40	– [9]	Bunge et al., 1993
30	IDS [L35485]	IVS 7, acceptor, T→G, 19086	19093	4.35 → 2.38	19086	2.55 → 11.30	+	Hopwood et al., 1993
31	IDS [L35485]	IVS 8, donor, C→T, 19209	19268	4.53 → 4.53	19208	3.27 → 10.74	+	Flomen et al., 1992
32	PAH [S76376]	IVS 10, acceptor, G→A, 76	86	5.22 → 4.78	77	-5.49 → 2.67	+ [10]	Dworniczak et al., 1991
33	ADA [M13792]	IVS 10, donor, G→A, 34484	34484	10.61 → -2.19	34488	3.22 → 3.22	–	Santisteban et al., 1993
B. Predicted cryptic splice sites [11]								
34	ALDOB [M15656]	IVS 6, acceptor, G→A, 196	196	9.62 → 2.05	197	-4.23 → 3.92	n.i.	Ali et al., 1994
35	CFTR [M55118]	IVS 13, donor, G→A, 901	901	10.01 → -2.79	905	3.10 →3.10	n.i.	Audrezet et al., 1993
36	CFTR [M55127]	IVS 19, acceptor, G→A, 266	266	10.24 → 2.67	267	-4.89 → 3.27	n.i.	Audrezet et al., 1993
37	CFTR [M55108]	IVS 2, acceptor, C→T, 182	184	13.24 → 11.58	186	5.10 → 5.74	n.i.	Bienvenu et al., 1994
38	COL2A1 [L10347]	IVS 20, acceptor, A→G, 17128	17129	13.30 → 5.19	17127	4.72 → 5.68	n.i.	Winterpacht et al., 1994 [12]
39	F8C [M88633]	IVS 5, acceptor, A→G, 385	386	13.80 → 5.63	385	-0.65 → 6.92	–	Naylor et al., 1991
40	F9 [K02402]	Exon 5, donor, G→A, 20701	20763	2.42 → 2.42	20699	5.15 → 5.56	+	Giannelli et al., 1991
41	F9 [K02402]	IVS 4, acceptor, A→G, 20632	20633	11.19 → 3.03	20632	-3.04 → 4.53	+	Giannelli et al., 1991

[9] cryptic splicing restores reading frame

[10] cryptic site use maintains reading frame; enzymatic activity is lost

[11] Prediction is based on decreased natural R_i accompanied by increase in R_i at a previously unrecognized cryptic site.

[12] # 38; report predicts a cryptic site at coordinate 17148; its $R_{i,natural} \to R_{i,mutant} = 0.00 \to -0.29$ bits

#	Gene [Accession]	Mutation, coordinate	Natural site(s)		Cryptic site(s)		Residual natural splicing	Reference(s)
			coord.	$R_{i,natural} \rightarrow R_{i,mutant}$	coord.	$R_{i,natural} \rightarrow R_{i,mutant}$		
42	F9 [K02402]	IVS 4, acceptor, G→C, 20633	20633	11.19 → 3.91	20635	0.21 → 6.19	−	Giannelli et al., 1991
43	F9 [K02402]	IVS 5, donor, A→G, 20775	20763	2.42 → 2.42	20775	-9.13 → 3.67	+	Giannelli et al., 1991
44	FGFR2 [M80635]	IVS A, acceptor, A→G, 68	69	13.97 → 5.81	68	0.45 → 8.03	n.i.	Schell et al., 1995
45	GLA [X14448]	IVS 5, acceptor, del 10507-10508	10509	12.10 → 3.03	10511	7.91 → 8.56	n.i.[13]	Eng et al., 1993
46	HPRT [M26434]	IVS 1, acceptor, A→T, 14777	14779	8.26 → 2.26	14784	4.40 → 6.55	n.i.	Gibbs et al., 1990

[13] Natural splicing at the other allele obscures measurement of residual splicing at the mutated site.

Table 3: Predicted non-deleterious splice site substitutions

#	Gene [Accession]	Nucleotide substitution, coordinate	Natural coord.	$R_{i,natural} \rightarrow R_{i,mutant}$	Cryptic coord. [1]	$R_{i,natural} \rightarrow R_{i,mutant}$	Residual natural splicing	Reference(s)
1	CFTR [M55126]	IVS 18, acceptor, T→C, 169	185	10.59 → 10.46	169	-27.56 → -26.10	n.i.	Audrezet et al., 1993
2	F9 [K02402]	Exon 5, donor, C→A, 20726	20763	2.42 → 2.42	20726	-16.78 → -19.78	+ [2]	Giannelli et al., 1991
3	F9 [K02402]	IVS 4, donor, A→G, 13477	13471	11.13 → 11.53	13475	4.13 → 3.73	+ [2]	Giannelli et al., 1991
4	OTC [D00227]	IVS 7, donor, A→G, 79	77	6.52 → 6.12	n.a.	n.a.	+	Carstens et al., 1991
5	CYP21 [M12792]	IVS 2, acceptor, C→A, 2333	2345	12.05 → 9.98 [3]	2333	0.70 → 8.54	+	Higashi et al., 1988
6	CYP21 [M12792]	IVS 2, acceptor, C→G, 2333	2345	12.05→ 10.49 [4]	2333	0.70 → 7.99	+ [5]	Higashi et al., 1988; Day et al., 1996
7	p53 [M14694]	Exon 7, acceptor, A→T, 14008	13999	8.55 → 8.55	14009	-5.00 → +2.41 [6]	n.i.	Hruban et al., 1994
8	SPB [M24461]	Exon 4, donor, C→GAA, 2588	2621	5.91 → 5.91	5754 [7]	1.42 → 1.42 [8]	n.i.	Nogee et al., 1994

[1] n.a.: not applicable

[2] Expression was inferred from clotting times and antigen bound.

[3] This variant was demonstrated to be a common polymorphism (Higashi et al., 1988).

[4] The report suggests that this site is not recognized; however, it contains more information than mut.#5.

[5] Relatives of individuals with this variant can be asymptomatic (Day et al., 1996)

[6] The cryptic splice site generated is significantly weaker than the natural site.

[7] Splicing was reported at this cryptic site in exon 8.

[8] The mutation in exon 4 is predicted by information analysis not to activate a cryptic site in exon 8.

REFERENCES

Ali M, Tuncman G, Cross NC, Vidailhet M, Bokesoy I, Gitzelmann R, Cox TM (1994): Null alleles of the aldolase B gene in patients with hereditary fructose intolerance. J Med Genet 31:499–503.

Allikmets R, Wasserman WW, Hutchinson A, Smallwood P, Nathans J, Rogan PK, Schneider TD, Dean M (1998): Organization of the ABCR gene: analysis of promoter and splice junction sequences. Gene 215:111–122. http://www.lecb.ncifcrf.gov/~toms/paper/abcr/.

Arredondo-Vega FX, Santisteban I, Kelly S, Schlossman CM, Umetsu DT, Hershfield MS (1994): Correct splicing despite mutation of the invariant first nucleotide of a 5′ splice site: a possible basis for disparate clinical phenotypes in siblings with adenosine deaminase deficiency. Am J Hum Genet 54:820–830.

Atweh GF, Anagnou NP, Shearin J, Forget BG, Kaufman RE (1985): Beta-thalassemia resulting from a single nucleotide substitution in an acceptor splice site. Nucleic Acids Res 13:777–790.

Atweh GF, Wong C, Reed R, Antonarakis SE, Zhu D, Ghosh PK, Maniatis T, Forget BG, KazazianJr HH (1987): A new mutation in IVS-1 of the human β globin gene causing β thalassemia due to abnormal splicing. Blood 70:147–151.

Audrezet MP, Mercier B, Guillermit H, Quere I, Verlingue C, Rault G, Ferec C (1993): Identification of 12 novel mutations in the CFTR gene. Hum Mol Genet 2:51–54.

Barbat B, Bogyo A, Raux-Demay MC, Kuttenn F, Boue J, Simon-Bouy B, Serre JL, Mornet E (1995): Screening of CYP21 gene mutations in 129 French patients affected by steroid 21-hydroxylase deficiency. Hum Mutat 5:126–130.

Berg OG, von Hippel PH (1987): Selection of DNA binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. 193:723–750.

Berg OG (1988): Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. J. Biomol. Struct. Dyn. 6:275–297.

Berg OG, von Hippel PH (1988*a*): Selection of DNA binding sites by regulatory proteins. Trend. Bioch. Sci. 13:207–211.

Berg OG, von Hippel PH (1988*b*): Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. J. Mol. Biol. 200:709–723.

Berget SM (1995): Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414.

Bienvenu T, Hubert D, Fonknechten N, Dusser D, Kaplan JC, Beldjord C (1994): Unexpected inactivation of acceptor consensus splice sequence by a -3 C to T transition in intron 2 of the CFTR gene. Hum Genet 94:65–68.

Black DL (1991): Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? Genes Dev 5:389–402.

Black DL (1992): Activation of c-src neuron-specific splicing by an unusual RNA element in vivo and in vitro. Cell 69:795–807.

Boerkoel CF, Exelbert R, Nicastri C, Nichols RC, Miller FW, Plotz PH, Raben N (1995): Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II. Am J Hum Genet 56:887–897.

Bonadio J, Ramirez F, Barr M (1990): An intron mutation in the human α 1(I) collagen gene alters the efficiency of pre-mRNA splicing and is associated with osteogenesis imperfecta type II. J Biol Chem 265:2262–2268.

Bottema CD, Ketterling RP, Yoon HS, Sommer SS (1990): The pattern of factor IX germ-line mutation in Asians is similar to that of Caucasians. Am J Hum Genet 47:835–841.

Brunak S, Engelbrecht J, Knudsen S (1990): Neural network detects errors in the assignment of mRNA splice sites. Nucl. Acids Res. 18:4797–4801.

Bunge S, Steglich C, Zuther C, Beck M, Morris CP, Schwinger E, Schinzel A, Hopwood JJ, Gal A (1993): Iduronate-2-sulfatase gene mutations in 16 patients with mucopolysaccharidosis type II (Hunter syndrome). Hum Mol Genet 2:1871–1875.

Carothers AM, Urlaub G, Grunberger D, Chasin LA (1993): Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. Mol Cell Biol 13:5085–5098.

Carstens RP, Fenton WA, Rosenberg LR (1991): Identification of RNA splicing errors resulting in human ornithine transcarbamylase deficiency. Am J Hum Genet 48:1105–1114.

Chibani J, Vidaud M, Duquesnoy P, Berge-Lefranc JL, Pirastu M, Ellouze F, Rosa J, Goossens M (1988): The peculiar spectrum of β-thalassemia genes in Tunisia. Hum Genet 78:190–192.

Chillon M, Casals T, Mercier B, Bassas L, Lissens W, Silber S, Romey MC, Ruiz-Romero J, Verlingue C, Claustres M, Nunes V, Férec C, Estivill X (1995): Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. N Engl J Med 332:1475–1480.

Chimienti G, Capurso A, Resta F, Pepe G (1992): A G-C change at the donor splice site of intron 1 causes lipoprotein lipase deficiency in a southern-Italian family. Biochem Biophys Res Commun 187:620–627.

Chu CS, Trapnell BC, Curristin S, Cutting GR, Crystal RG (1993): Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. Nat Genet 3:151–156.

Cladaras C, Hadzopoulou-Cladaras M, Felber BK, Pavlakis G, Zannis VI (1987): The molecular basis of a familial apoE deficiency. An acceptor splice site mutation in the third intron of the deficient apoE gene. J Biol Chem 262:2310–2315.

Cogan JD, Phillips3d JA, Sakati N, Frisch H, Schober E, Milner RD (1993): Heterogeneous growth hormone (GH) gene mutations in familial GH deficiency. J Clin Endocrinol Metab 76:1224–1228.

Cole WG, Chiodo AA, Lamande SR, Janeczko R, Ramirez F, Dahl HH, Chan D, Bateman JF (1990): A base substitution at a splice site in the COL3A1 gene causes exon skipping and generates abnormal type III procollagen in a patient with Ehlers-Danlos syndrome type IV. J Biol Chem 265:17070–17077.

Day DJ, Speiser PW, White PC, Barany F (1995): Detection of steroid 21-hydroxylase alleles using gene-specific PCR and a multiplexed ligation detection reaction. Genomics 29:152–162.

Day DJ, Speiser PW, Schulze E, Bettendorf M, Fitness J, Barany F, White PC (1996): Identification of non-amplifying CYP21 genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees. Hum Mol Genet 5:2039–2048.

Dirksen WP, Hampson RK, Sun Q, Rottman FM (1994): A purine-rich exon sequence enhances alternative splicing of bovine growth hormone pre-mRNA. J Biol Chem 269:6431–6436.

Dominski Z, Kole R (1991): Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11:6075–6083.

Dworniczak B, Aulehla-Scholz C, Kalaydjieva L, Bartholome K, Grudda K, Horst J (1991): Aberrant splicing of phenylalanine hydroxylase mRNA: the major cause for phenylketonuria in parts of southern Europe. Genomics 11:242–246.

Eng CM, Resnick-Silverman LA, Niehaus DJ, Astrin KH, Desnick RJ (1993): Nature and frequency of mutations in the α-galactosidase A gene that cause Fabry disease. Am J Hum Genet 53:1186–1197.

Flomen RH, Green PM, Bentley DR, Giannelli F, Green EP (1992): Detection of point mutations and a gross deletion in six Hunter syndrome patients. Genomics 13:543–550.

Foster DC, Yoshitake S, Davie EW (1985): The nucleotide sequence of the gene for human protein C. Proc Natl Acad Sci U S A 82:4673–4677.

Ganguly A, Baldwin CT, Strobel D, Conway D, Horton W, Prockop DJ (1991): Heterozygous mutation in the G^{+5} position of intron 33 of the pro-α 2(I) gene (COL1A2) that causes aberrant RNA splicing and lethal osteogenesis imperfecta. Use of carbodiimide methods that decrease the extent of DNA sequencing necessary to define an unusual mutation. J Biol Chem 266:12035–12040.

Giannelli F, Green PM, High KA, Sommer S, Lillicrap DP, Ludwig M, Olek K, Reitsma PH, Goossens M, Yoshioka A, Brownlee GG (1991): Haemophilia B: database of point mutations and short additions and deletions–second edition. Nucleic Acids Res 19:2193–2219.

Gibbs RA, Nguyen PN, McBride LJ, Koepf SM, Caskey CT (1989): Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of in vitro amplified cDNA. Proc Natl Acad Sci U S A 86:1919–1923.

Gibbs RA, Nguyen PN, Edwards A, Civitello AB, Caskey CT (1990): Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch-Nyhan families. Genomics 7:235–244.

Goldsmith ME, Humphries RK, Ley T, Cline A, Kantor JA, Nienhuis AW (1983): "Silent" nucleotide substitution in a β^{+}-thalassemia globin gene activates splice site in coding sequence RNA. Proc Natl Acad Sci U S A 80:2318–2322.

Grandchamp B, Picat C, Kauppinen R, Mignotte V, Peltonen L, Mustajoki P, Romeo PH, Goossens M, Nordmann Y (1989*a*): Molecular analysis of acute intermittent porphyria in a

Finnish family with normal erythrocyte porphobilinogen deaminase. Eur J Clin Invest 19:415–418.

Grandchamp B, Picat C, Mignotte V, Wilson J, Te Velde K, Sandkuyl L, Romeo P, Goossens M, Nordmann Y (1989*b*): Tissue-specific splicing mutation in acute intermittent porphyria. Proc Natl Acad Sci U S A 86:661–664.

Haber DA, Buckler AJ, Glaser T, Call KM, Pelletier J, Sohn RL, Douglass EC, Housman DE (1990): An internal deletion within an 11p13 zinc finger gene contributes to the development of Wilms' tumor. Cell 61:1257–1269.

Hall SL, Padgett RA (1994): Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J. Mol. Biol. 239:357–365.

Hata A, Emi M, Luc G, Basdevant A, Gambert P, Iverius PH, Lalouel JM (1990): Compound heterozygote for lipoprotein lipase deficiency: Ser-Thr244 and transition in 3′ splice site of intron 2 (AG-AA) in the lipoprotein lipase gene. Am J Hum Genet 47:721–726.

Hawkins JD (1988): A survey on intron and exon lengths. Nucl. Acids Res. 16:9893–9908.

He GS, Grabowski GA (1992): Gaucher disease: A G+1→A+1 IVS2 splice donor site mutation causing exon 2 skipping in the acid β-glucosidase mRNA. Am J Hum Genet 51:810–820.

Hengen PN, Bartram SL, Stewart LE, Schneider TD (1997): Information analysis of Fis binding sites. Nucl. Acids Res. 25 (24):4994–5002. http://www.lecb.ncifcrf.gov/~toms/paper/fisinfo/.

Higashi Y, Tanae A, Inoue H, Hiromasa T, Fujii-Kuriyama Y (1988): Aberrant splicing and missense mutations cause steroid 21-hydroxylase [P-450(C21)] deficiency in humans: possible gene conversion products. Proc Natl Acad Sci U S A 85:7486–7490.

Ho KK, Kong RY, Kuffner T, Hsu LH, Ma L, Cheah KS (1994): Further evidence that the failure to cleave the aminopropeptide of type I procollagen is the cause of Ehlers-Danlos syndrome type VII. Hum Mutat 3:358–364.

Hopwood JJ, Bunge S, Morris CP, Wilson PJ, Steglich C, Beck M, Schwinger E, Gal A (1993): Molecular basis of mucopolysaccharidosis type II: mutations in the iduronate-2-sulphatase gene. Hum Mutat 2:435–442.

Horowitz JM, Yandell DW, Park SH, Canning S, Whyte P, Buchkovich K, Harlow E, Weinberg RA, Dryja TP (1989): Point mutational inactivation of the retinoblastoma antioncogene. Science 243:937–940.

Hruban RH, van der Riet P, Erozan YS, Sidransky D (1994): Brief report: molecular biology and the early detection of carcinoma of the bladder–the case of Hubert H. Humphrey. N Engl J Med 330:1276–1278.

Huh GS, Hynes RO (1994): Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. Genes Dev 8:1561–1574.

Huie ML, Chen AS, Tsujino S, Shanske S, DiMauro S, Engel AG, Hirschhorn R (1994): Aberrant splicing in adult onset glycogen storage disease type II (GSDII): molecular identification of an IVS1 (-13T→G) mutation in a majority of patients and a novel IVS10 (+1GT→CT) mutation. Hum Mol Genet 3:2231–2236.

Humphrey MB, Bryan J, Cooper TA, Berget SM (1995): A 32-nucleotide exon-splicing enhancer regulates usage of competing $5'$ splice sites in a differential internal exon. Mol Cell Biol 15:3979–3988.

Jones CT, McIntosh I, Keston M, Ferguson A, Brock DJ (1992): Three novel mutations in the cystic fibrosis gene detected by chemical cleavage: analysis of variant splicing and a nonsense mutation. Hum Mol Genet 1:11–17.

Jonsson JJ, Aronovich EL, Braun SE, Whitley CB (1995): Molecular diagnosis of mucopolysaccharidosis type II (Hunter syndrome) by automated sequencing and computer-assisted interpretation: toward mutation mapping of the iduronate-2-sulfatase gene. Am J Hum Genet 56:597–607.

Kishimoto TK, O'Conner K, Springer TA (1989): Leukocyte adhesion deficiency. Aberrant splicing of a conserved integrin sequence causes a moderate deficiency phenotype. J Biol Chem 264:3588–3595.

Kishimoto Y, Murakami Y, Hayashi K, Takahara S, Sugimura T, Sekiya T (1992): Detection of a common mutation of the catalase gene in Japanese acatalasemic patients. Hum Genet 88:487–490.

Klein JL, Shows TB, Dupont B, Trapani JA (1989): Genomic organization and chromosomal assignment for a serine protease gene (CSPB) expressed by human cytotoxic lymphocytes. Genomics 5:110–117.

Klima H, Ullrich K, Aslanidis C, Fehringer P, Lackner KJ, Schmitz G (1993): A splice junction mutation causes deletion of a 72-base exon from the mRNA for lysosomal acid lipase in a patient with cholesteryl ester storage disease. J Clin Invest 92:2713–2718.

Krawczak M, Reiss J, Cooper DN (1992): The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 90:41–54.

Kudo S, Fukuda M (1989): Structural organization of glycophorin A and B genes: glycophorin B gene evolved by homologous recombination at Alu repeat sequences. Proc Natl Acad Sci U S A 86:4619–4623.

Lapoumeroulie C, Acuto S, Rouabhi F, Labie D, Krishnamoorthy R, Bank A (1987): Expression of a β thalassemia gene with abnormal splicing. Nucleic Acids Res 15:8195–8204.

Lavigueur A, LaBranche M, Kornblihtt AR, Chabot B (1993): A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. Genes Dev 7:2405–2417.

Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomäki P, Sistonen P, Aaltonen LA, Nyström-Lahti M, Guan XY, Zhang J, Meltzer PS, Yu JW, Kao FT, Chen DJ, Cerosaletti KM, Fournier REK, Todd S, Lewis T, Leach RJ, Naylor SL, Weissenbach J, Mecklin JP, Järvinen H, Petersen GM, Hamilton SR, Green J, Jass J, Watson P, Lynch HT, Trent JM, de la Chapelle A, Kinzler KW, Vogelstein B (1993): Mutations of a *mutS* homolog in hereditary nonpolyposis colorectal cancer. Cell 75:1215–1225.

Lehmann HW, Mundlos S, Winterpacht A, Brenner RE, Zabel B, Muller PK (1994): Ehlers-Danlos syndrome type VII: phenotype and genotype. Arch Dermatol Res 286:425–428.

MacLeod JN, Liebhaber SA, MacGillivray MH, Cooke NE (1991): Identification of a splice-site mutation in the human growth hormone- variant gene. Am J Hum Genet 48:1168–1174.

Mertes G, Ludwig M, Finkelnburg B, Krawczak M, Schwaab R, Brackmann HH, Olek K (1994): A G+3-to-T donor splice site mutation leads to skipping of exon 50 in von Willebrand factor mRNA. Genomics 24:190–191.

Metherall JE, Collins FS, Pan J, Weissman SM, Forget BG (1986): Beta zero thalassemia caused by a base substitution that creates an alternative splice acceptor site in an intron. EMBO J 5:2551–2557.

Mount SM (1982): A catalogue of splice junction sequences. Nucl. Acids Res. 10:459–472.

Muntoni S, Wiebusch H, Funke H, Ros E, Seedorf U, Assmann G (1995): Homozygosity for a splice junction mutation in exon 8 of the gene encoding lysosomal acid lipase in a Spanish kindred with cholesterol ester storage disease (CESD). Hum Genet 95:491–494.

Nakai K, Sakamoto H (1994): Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene 141:171–177.

Naylor JA, Green PM, Montandon AJ, Rizza CR, Giannelli F (1991): Detection of three novel mutations in two haemophilia a patients by rapid screening of whole essential region of factor VIII gene. Lancet 337:635–639.

Neote K, Bapat B, Dumbrille-Ross A, Troxel C, Schuster SM, Mahuran DJ, Gravel RA (1988): Characterization of the human HEXB gene encoding lysosomal β- hexosaminidase. Genomics 3:279–286.

Niwa M, MacDonald CC, Berget SM (1992): Are vertebrate exons scanned during splice-site selection? Nature 360:277–280.

Nogee LM, Garnier G, Dietz HC, Singer L, Murphy AM, deMello DE, Colten HR (1994): A mutation in the surfactant protein B gene responsible for fatal neonatal respiratory disease in multiple kindreds. J Clin Invest 93:1860–1863.

O'Neill JP, Rogan PK, Cariello N, Nicklas JA (1998): Mutations that alter RNA splicing of the human HPRT gene: A review of the spectrum. Rev. Mut. Res. in press:–.

Ohno K, Suzuki K (1988): A splicing defect due to an exon-intron junctional mutation results in abnormal β-hexosaminidase α chain mRNAs in Ashkenazi Jewish patients with Tay-Sachs disease. Biochem Biophys Res Commun 153:463–469.

Orkin SH, KazazianJr HH, Antonarakis SE, Ostrer H, Goff SC, Sexton JP (1982): Abnormal RNA processing due to the exon mutation of β E-globin gene. Nature 300:768–769.

Owerbach D, Ballard AL, Draznin MB (1992): Salt-wasting congenital adrenal hyperplasia: detection and characterization of mutations in the steroid 21-hydroxylase gene, CYP21, using the polymerase chain reaction. J Clin Endocrinol Metab 74:553–558.

Ozkara HA, Akerman BR, Ciliv G, Topcu M, Renda Y, Gravel RA (1995): Donor splice site mutation in intron 5 of the HEXA gene in a Turkish infant with Tay-Sachs disease. Hum Mutat 5:186–187.

Proia RL (1988): Gene encoding the human β-hexosaminidase β chain: extensive homology of intron placement in the α- and β-chain genes. Proc Natl Acad Sci U S A 85:1883–1887.

Purandare SM, Lanyon WG, Arngrimsson R, Connor JM (1995): Characterisation of a novel splice donor mutation affecting position +1 in intron 18 of the NF-1 gene. Hum Mol Genet 4:767–768.

Rave-Harel N, Kerem E, Nissim-Rafinia M, Madjar I, Goshen R, Augarten A, Rahat A, Hurwitz A, Darvasi A, Kerem B (1997): The molecular basis of partial penetrance of splicing mutations in cystic fibrosis. Am J Hum Genet 60:87–94.

Renda M, Maggio A, Warren TC, Kazazian HH (1992): Detection of an IVS-1 $3'$ end (G-C) β-thalassemia mutation in the AG invariant dinucleotide of the acceptor splice site in a Sicilian subject. Genomics 13:234–235.

Robberson BL, Cote GJ, Berget SM (1990): Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84–94.

Roberts RG, Bobrow M, Bentley DR (1992): Point mutations in the dystrophin gene. Proc Natl Acad Sci U S A 89:2331–2335.

Roberts RG, Bentley DR, Bobrow M (1993*a*): Infidelity in the structure of ectopic transcripts: a novel exon in lymphocyte dystrophin transcripts. Hum Mutat 2:293–299.

Roberts RG, Passos-Bueno MR, Bobrow M, Vainzof M, Zatz M (1993*b*): Point mutation in a Becker muscular dystrophy patient. Hum Mol Genet 2:75–77.

Rogan PK, Schneider TD (1995): Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. Hum Mutat 6:74–76.

Sakuraba H, Eng CM, Desnick RJ, Bishop DF (1992): Invariant exon skipping in the human α-galactosidase A pre-mRNA: A G+1 to T substitution in a $5'$-splice site causing Fabry disease. Genomics 12:643–650.

Santisteban I, Arredondo-Vega FX, Kelly S, Mary A, Fischer A, Hummell DS, Lawton A, Sorensen RU, Stiehm ER, Uribe L, Weinberg K, Hershfield MS (1993): Novel splicing, missense, and deletion mutations in seven adenosine deaminase-deficient patients with late/delayed onset of combined immunodeficiency disease. contribution of genotype to phenotype. J Clin Invest 92:2291–2302.

Santisteban I, Arredondo-Vega FX, Kelly S, Loubser M, Meydan N, Roifman C, Howell PL, Bowen T, Weinberg KI, Schroeder ML, Hershfield MS (1995): Three new adenosine deaminase mutations that define a splicing enhancer and cause severe and partial phenotypes: implications for evolution of a CpG hotspot and expression of a transduced ADA cDNA. Hum Mol Genet 4:2081–2087.

Schell U, Hehr A, Feldman GJ, Robin NH, Zackai EH, de Die-Smulders C, Viskochil DH, Stewart JM, Wolff G, Ohashi H, Price RA, Cohen Jr. MM, Muenke N (1995): Mutations in FGFR1 and FGFR2 cause familial and sporadic Pfeiffer syndrome. Hum Mol Genet 4:323–328.

Schloesser M, Hofferbert S, Bartz U, Lutze G, Lammle B, Engel W (1995): The novel acceptor splice site mutation 11396(G$\rightarrow$A) in the factor XII gene causes a truncated transcript in cross-reacting material negative patients. Hum Mol Genet 4:1235–1237.

Schneider TD, Stormo GD, Haemer JS, Gold L (1982): A design for computer nucleic-acid sequence storage, retrieval and manipulation. Nucl. Acids Res. 10:3013–3024.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986): Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188:415–431.

Schneider TD (1991*a*): Theory of molecular machines. I. Channel capacity of molecular machines. J. Theor. Biol. 148:83–123. http://www.lecb.ncifcrf.gov/~toms/paper/ccmm/.

Schneider TD (1991*b*): Theory of molecular machines. II. Energy dissipation from molecular machines. J. Theor. Biol. 148:125–137. http://www.lecb.ncifcrf.gov/~toms/paper/edmm/.

Schneider TD (1994): Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. Nanotechnology 5:1–18. http://www.lecb.ncifcrf.gov/~toms/paper/nano2/.

Schneider TD (1995):. Information Theory Primer. http://www.lecb.ncifcrf.gov/~toms/paper/primer/.

Schneider TD (1997*a*): Information content of individual genetic sequences. J. Theor. Biol. 189 (4):427–441. http://www.lecb.ncifcrf.gov/~toms/paper/ri/.

Schneider TD (1997*b*): Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. Nucl. Acids Res. 25:4408–4415. http://www.lecb.ncifcrf.gov/~toms/paper/walker/, erratum: NAR 26(4): 1135, 1998.

Schulze E, Scharer G, Rogatzki A, Priebe L, Lewicka S, Bettendorf M, Hoepffner W, Heinrich UE, Schwabe U (1995): Divergence between genotype and phenotype in relatives of patients with the intron 2 mutation of steroid-21-hydroxylase. Endocr Res 21:359–364.

Senapathy P, Shapiro MB, Harris NL (1990): Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Meth. Enzym. 183:252–278.

Shannon CE (1948): A mathematical theory of communication. Bell System Tech. J. 27:379–423, 623–656. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

Soria JM, Fontcuberta J, Chillon M, Borrell M, Estivill X, Sala N (1993): Acceptor splice site mutation in the invariant AG of intron 5 of the protein C gene, causing type I protein C deficiency. Hum Genet 92:506–508.

Speiser PW, Dupont J, Zhu D, Serrat J, Buegeleisen M, Tusie-Luna MT, Lesser M, New MI, White PC (1992): Disease expression and molecular genotype in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. J Clin Invest 90:584–595.

Spritz RA, Jagadeeswaran P, Choudary PV, Biro PA, Elder JT, deRiel JK, Manley JL, Gefter ML, Forget BG, Weissman SM (1981): Base substitution in an intervening sequence of a β^{+}-thalassemic human globin gene. Proc Natl Acad Sci U S A 78:2455–2459.

Stephens RM, Schneider TD (1992): Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. 228:1124–1136. http://www.lecb.ncifcrf.gov/~toms/paper/splice/.

Sterner DA, Berget SM (1993): In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. Mol Cell Biol 13:2677–2687.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982): Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli.* Nucl. Acids Res. 10:2997–3011.

Sun F, Knebelmann B, Pueyo ME, Zouali H, Lesage S, Vaxillaire M, Passa P, Cohen D, Velho G, Antignac C, Froguel P (1993*a*): Deletion of the donor splice site of intron 4 in the glucokinase gene causes maturity-onset diabetes of the young. J Clin Invest 92:1174–1180.

Sun Q, Mayeda A, Hampson RK, Krainer AR, Rottman FM (1993*b*): General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. Genes Dev 7:2598–2608.

Szilard L (1964): On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. Behavioral Science 9:301–310.

Talerico M, Berget SM (1990): Effect of $5'$ splice site mutations on splicing of the preceding intron. Mol Cell Biol 10:6299–6305.

Trapani JA, Klein JL, White PC, Dupont B (1988): Molecular cloning of an inducible serine esterase gene from human cytotoxic lymphocytes. Proc Natl Acad Sci U S A 85:6924–6928.

Treisman R, Orkin SH, Maniatis T (1983): Specific transcription and RNA splicing defects in five cloned β- thalassaemia genes. Nature 302:591–596.

Tsujino S, Servidei S, Tonin P, Shanske S, Azan G, DiMauro S (1994): Identification of three novel mutations in non-Ashkenazi Italian patients with muscle phosphofructokinase deficiency. Am J Hum Genet 54:812–819.

Vasan NS, Kuivaniemi H, Vogel BE, Minor RR, Wootton JA, Tromp G, Weksberg R, Prockop DJ (1991): A mutation in the pro α 2(I) gene (COL1A2) for type I procollagen in Ehlers-Danlos syndrome type VII: evidence suggesting that skipping of exon 6 in RNA splicing may be a common cause of the phenotype. Am J Hum Genet 48:305–317.

Vidaud M, Gattoni R, Stevenin J, Vidaud D, Amselem S, Chibani J, Rosa J, Goossens M (1989): A $5'$ splice-region G$\rightarrow$C mutation in exon 1 of the human β-globin gene inhibits pre-mRNA splicing: a mechanism for β^+-thalassemia. Proc Natl Acad Sci U S A 86:1041–1045.

Wang Z, Hoffmann HM, Grabowski PJ (1995): Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity. RNA 1:21–35.

Watson RB, Wallis GA, Holmes DF, Viljoen D, Byers PH, Kadler KE (1992): Ehlers Danlos syndrome type VIIB. Incomplete cleavage of abnormal type I procollagen by N-proteinase in vitro results in the formation of copolymers of collagen and partially cleaved pNcollagen that are near circular in cross-section. J Biol Chem 267:9093–9100.

Weil D, D'Alessio M, Ramirez F, de Wet W, Cole WG, Chan D, Bateman JF (1989*a*): A base substitution in the exon of a collagen gene causes alternative splicing and generates a structurally abnormal polypeptide in a patient with Ehlers-Danlos syndrome type VII. EMBO J 8:1705–1710.

Weil D, D'Alessio M, Ramirez F, Steinmann B, Wirtz MK, Glanville RW, Hollister DW (1989*b*): Temperature-dependent expression of a collagen splicing defect in the fibroblasts of a patient with Ehlers-Danlos syndrome type VII. J Biol Chem 264:16804–16809.

Weil D, D'Alessio M, Ramirez F, Eyre DR (1990): Structural and functional characterization of a splicing mutation in the pro-α 2(I) collagen gene of an Ehlers-Danlos type VII patient. J Biol Chem 265:16007–16011.

Wen JK, Osumi T, Hashimoto T, Ogata M (1990): Molecular analysis of human acatalasemia. Identification of a splicing mutation. J Mol Biol 211:383–393.

Will K, Dork T, Stuhrmann M, Meitinger T, Bertele-Harms R, Tummler B, Schmidtke J (1994): A novel exon in the cystic fibrosis transmembrane conductance regulator gene activated by the nonsense mutation E92X in airway epithelial cells of patients with cystic fibrosis. J Clin Invest 93:1852–1859.

Wilton SD, Chandler DC, Kakulas BA, Laing NG (1994): Identification of a point mutation and germinal mosaicism in a Duchenne muscular dystrophy family. Hum Mutat 3:133–140.

Winterpacht A, Schwarze U, Mundlos S, Menger H, Spranger J, Zabel B (1994): Alternative splicing as the result of a type II procollagen gene (COL2A1) mutation in a patient with Kniest dysplasia. Hum Mol Genet 3:1891–1893.

Yandell DW, Campbell TA, Dayton SH, Petersen R, Walton D, Little JB, McConkie-Rosell A, Buckley EG, Dryja TP (1989): Oncogenic point mutations in the human retinoblastoma gene: their application to genetic counseling. N Engl J Med 321:1689–1695.

Phillips 3rd JA, Cogan JD (1994): Genetic basis of endocrine disease. 6. Molecular basis of familial human growth hormone deficiency. J Clin Endocrinol Metab 78:11–16.