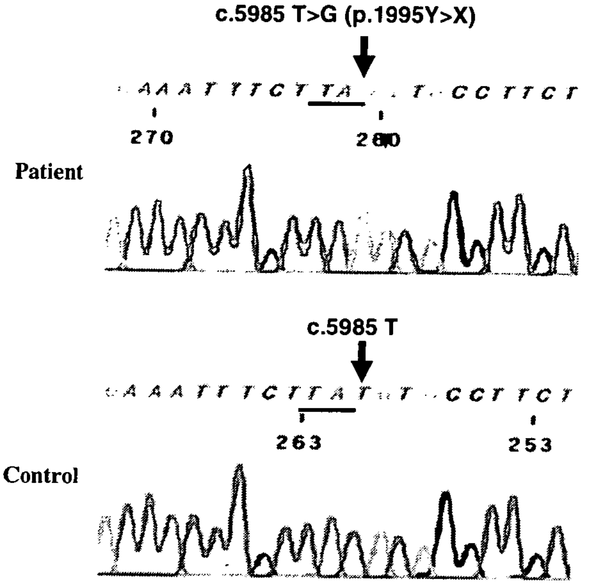

Fig. 1 Sequencing results from exon 42 are shown. The PCR products amplified from genomic DNA were directly sequenced. The 63rd nucleotide of exon 42 was a G in the index case (patient), whereas it was a T in the control sample (control). The nucleotide position corresponds to the 5,985th nucleotide of the dystrophin mRNA (c.5,985T > G). The mutation changed a tyrosine codon (TAT) to a stop codon (TAG) at the 1,995th amino acid residue of dystrophin (p.1,995Y > X)

dystrophin was consistent with the clinical diagnosis of DMD in this patient.

In order to confirm the molecular diagnosis, the gene product was examined at the mRNA level. When dystrophin mRNA extending from exons 40 to 45 was analyzed in the patient's lymphocytes by RT-nested PCR amplification, three separate products were obtained (Fig. 2a). The largest product consisted of the sequence of exons 40 to 45, and included the same nonsense mutation in exon 42 as observed in the genomic DNA. In the second largest product, 63 bp of the 5′ end of exon 42 was missing, whereas the sequences of the other exons were completely normal (dys-63). Interestingly, in the smallest product, the 3′ end of exon 41 was directly joined to the 5′ end of exon 43, which removed all 195 bp of exon 42 (dys-exon 42) (Fig. 2b). The latter two transcripts were considered natural products, because the exon boundaries were conserved and no other nucleotide changes were present in the sequenced exons. Dys-exon 42 was assumed to be a result of exon 42 skipping caused by the single nucleotide change. The exon 42 skipping observed in the lymphocytes of in the index case may have been due to the creation of a splicing silencer.

Examination of sequences near the mutation site disclosed that a novel AG dinucleotide, which is a conserved splice acceptor sequence, was introduced into the exon sequence by c.5,985T > G (Fig. 1). Therefore, the creation of the novel splice acceptor site was likely to cause the aberrant splicing that led to the production of dys-63. In order to confirm the activity of the nonsense mutation-created AG dinucleotide, experimental splicing analysis was conducted (Fig. 3). Either the wild-type or mutant exon 42 together with the flanking intron sequences were inserted into the preconstructed minigene to make hybrid minigenes and transcripts from the hybrid minigenes were analyzed by RT-PCR amplification. One PCR product containing the entire exon 42 sequence between the cassette exons A and B was obtained from the minigene encoding the wild-type exon 42 (Fig. 3b). On the other hand, two amplified products were obtained from the hybrid minigene containing exon 42 with the mutation: a major product corresponding to the normally spliced product and a minor, smaller product containing exon 42 without 63 bp of its 5′ end between exons A and B (Fig. 3c); this was the same as one of the aberrant splicing products (dys-63) identified in lymphocytes. The result indicated that the mutation-created splicing acceptor site was actually active in this hybrid minigene in HeLa cell. Therefore, dys-63 was confirmed to be a real splicing product that was transcribed from the mutated gene.

Dystrophin mRNAs obtained from lymphocytes were examined for their protein coding abilities. The authentically spliced product containing a premature stop codon in exon 42 was nonfunctional. On the other hand, dys-63 and dys-exon 42 maintained the translational reading frame and did not carry premature stop codons, and were therefore expected to produce truncated variants of dystrophin that lacked 21 and 65 amino acid residues in the rod domain, respectively. The index case, however, was diagnosed with DMD based on the lack of dystrophin in his skeletal muscle.

Considering that the dystrophin mRNA produced in muscle cells more accurately reflects the clinical phenotype than that produced in lymphocytes, muscle dystrophin mRNA from the patient was examined by RT-PCR amplification. Remarkably, the amplification of the region encompassing exons 40 to 45 produced a single PCR product (Fig. 2a). Sequencing of the product disclosed sequences of exons 40 to 45, including the nonsense mutation. It was concluded that authentic splicing was completely maintained in the skeletal muscle and no in-frame aberrant mRNA was produced in this tissue. This is compatible with the dystrophin deficiency in his muscle cells and the clinical phenotype of DMD.

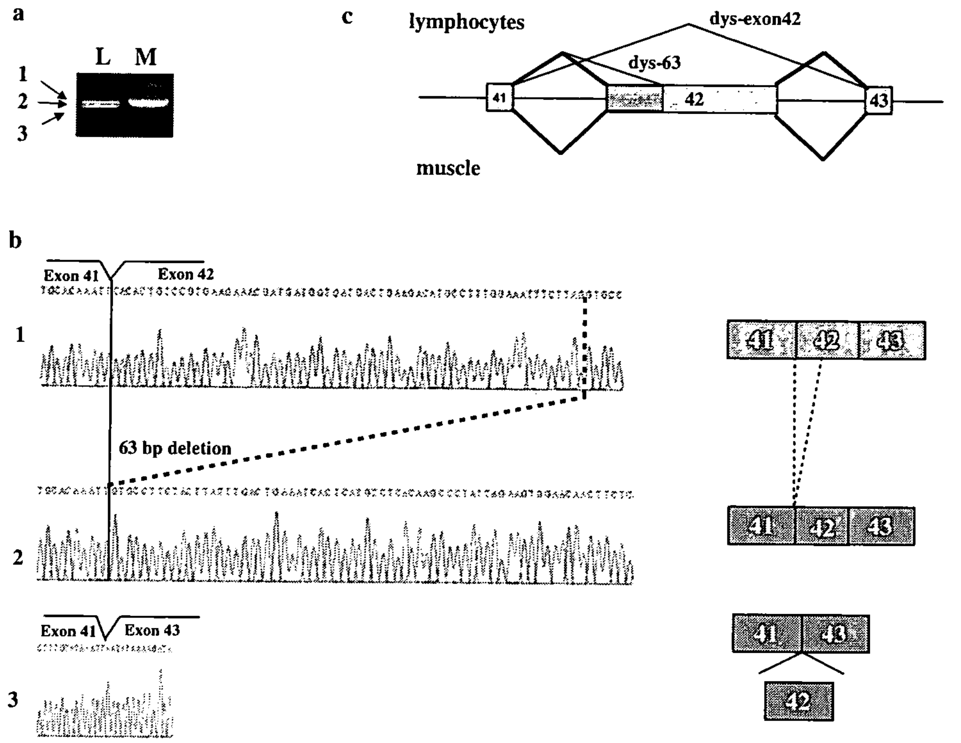

Fig. 2 Analysis of dystrophin mRNA. **a** The amplified products encompassing exons 40 to 45 are shown. Fragments encompassing exons 40 to 45 were amplified from cDNA prepared from the patient's lymphocytes and skeletal muscle. Three bands were visualized from the lymphocyte cDNA (L), whereas one clear band was visualized from the skeletal muscle (M). Numbers on the left side of the panel correspond to the numbers in panel **b** (*lower panel*). **b** The sequences of three different clones are shown. Each sequence has completely normal exons 40, 41, 43, 44, and 45. The sequence of the 3′ end of exon 41 (5′-AAATT-3′) is joined to the three different sequences in the three clones: CACAC (1), GTGCC (2), and AATAT (3). In the *top panel* (1), the normal exon structure from exon 40 to 45 is maintained, but the mutation is present. In the *middle panel* (2), 132 bp of the truncated exon 42 was followed by a completely normal exon 43. In the *bottom panel* (3), exon 41 joins directly to exon 43. The exon structure of each product is shown schematically on the right side. **c** The splicing patterns identified in the index case are represented schematically. The *diagonal lines* above and below the boxes indicate the splicing events that were observed in lymphocytes and skeletal muscle, respectively. The dys-63 and dys-exon 42 transcripts are aberrantly spliced gene products. *Boxes* and *horizontal lines* indicate exons and introns, respectively. The figure is not drawn to scale

Discussion

A novel single nucleotide change of c.5,985T > G in exon 42 of the dystrophin gene that changed a tyrosine codon to a stop codon (p.1,995Y > X) was identified in a Japanese boy diagnosed with DMD. Further molecular analysis revealed the mutation had a number of effects. In the patient's lymphocytes, the mutation caused three molecular events: (1) a premature stop codon was introduced into the authentically spliced mRNA product, (2) a mutation-created AG dinucleotide acted as a splice acceptor site, producing the aberrantly spliced dys-63 transcript, and (3) exon 42 skipping, producing the dys-exon 42 transcript (Fig. 2c). In skeletal muscle, however, only the authentically spliced product was observed. Although the patient's phenotype was expected to be mild due to the detection of in-frame dys-63 and dys-exon 42 in his lymphocytes, the patient had a typical DMD phenotype because all the dystrophin transcripts in his skeletal muscle carried the nonsense mutation.

In previous reports, the detection of aberrant splicing products in lymphocytes, which can be easily obtained, successfully led to the identification of the same transcripts in skeletal muscle (Barbieri et al. 1996; Shiga et al. 1997), thereby facilitating the molecular understanding of dystrophinopathy. Similar to previous reports (Adachi et al. 2003), however, our results showed different dystrophin mRNA splicing patterns in skeletal muscle cells and lymphocytes (Fig. 2a). This suggests that the regulators of splicing are not exactly the same in these tissues.

In vitro splicing analysis using a hybrid minigene clearly showed the nonsense mutation-created splice acceptor site was used by the spliceosome (Fig. 3). Using this hybrid minigene, a small amount of an aberrant splicing product that was produced using the novel

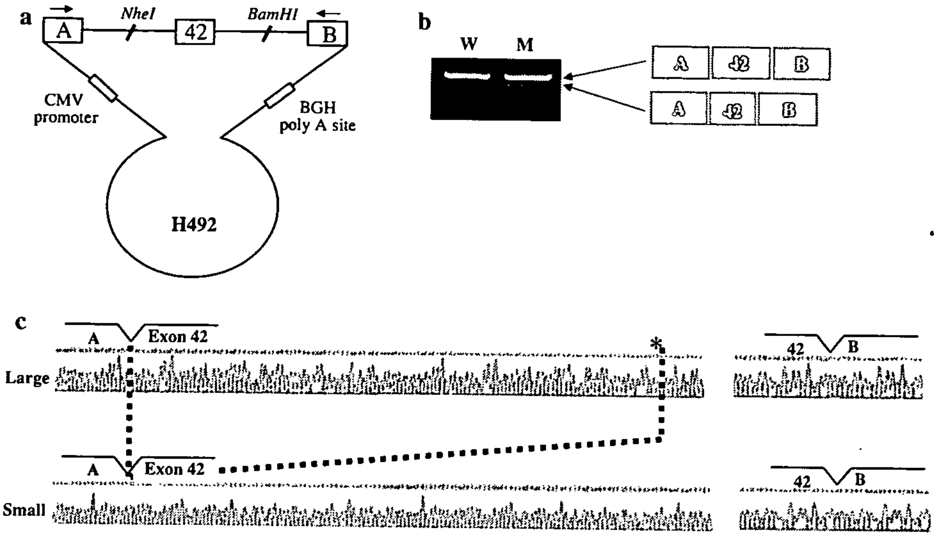

Fig. 3 Hybrid minigenes containing the indicated variants were tested in an in vitro splicing assay. **a** The hybrid minigene construct is schematically described. A minigene (H492) was constructed to encode two cassette exons (A and B) and an intervening sequence containing a multicloning site. The minigene contained a cytomegalovirus (CMV) enhancer-promoter and a bovine growth hormone gene (BGH) polyadenylation signal (*dark shaded boxes*) for complete synthesis of mRNA. The primers used in the RT-PCR assay are represented by arrows. **b** RT-PCR amplified products of hybrid minigene transcripts. A single transcript was generated from a minigene carrying the wild-type exon 42 sequence (W). From a minigene carrying the mutant exon 42, two transcripts were generated (M) and their nucleotide sequences are shown in panel **c**. A schematic description of the RT-PCR products is shown on the right. **c** Two transcripts from the mutant hybrid minigene. Nucleotide sequences at the junctions between exons are shown. The large product (*top*) consists of exon A, the complete exon 42, and exon B, whereas the small product lacked 63 bp of the 5′ end of exon 42 (*bottom*). c.5,985T > G is marked by an *asterisk*

splice acceptor site was obtained (Fig. 3). This indicates that the novel site can be recognized by the splicing machinery in HeLa cells. In contrast, the novel splice acceptor site was not used in the patient's skeletal muscle (Fig. 2). These differences in the use of the novel splice acceptor site suggest that trans-elements, such as nuclear proteins expressed in tissue-specific patterns, instead of cis-elements, such as splicing enhancer and silencer sequences, regulate the activation of the novel splice acceptor site. Future studies should clarify the trans-elements that determine whether or not the novel splice acceptor site is used.

Presently, there is no effective way to treat DMD. Recent DMD treatments have focused on converting the DMD phenotype to a BMD phenotype by changing dystrophin mRNAs from out-of-frame to in-frame. In our previous study, we showed that the induction of exon 19 skipping in a DMD patient carrying a deletion in exon 20 led to the production of in-frame dystrophin mRNA and dystrophin-positive skeletal muscle cells (Takeshima et al. 2006). Our present findings indicate that modulating the splicing of dystrophin mRNA in skeletal muscle to produce in-frame transcripts coding for truncated, semi-functional dystrophin is a potential target for treatment of this disease.

Acknowledgments We would like to thank Ms. A. Hosoda for her secretarial help. This work was supported by a grant-in-aid for Scientific Research from the Japan Society for the Promotion of Science; Health, and Labour Sciences Research Grants for Research on Psychiatric and Neurological Diseases and Mental Health; a research grant for Nervous and Mental Disorders from the Ministry of Health, Labour, and Welfare; and the Mitsubishi Foundation.

References

Adachi K, Takeshima Y, Wada H, Yagi M, Nakamura H, Matsuo M (2003) Heterogous dystrophin mRNAs produced by a novel splice acceptor site mutation in intermediate dystrophinopathy. Pediatr Res 53:125–131

Barbieri AM, Soriani N, Ferlini A, Michelato A, Ferrari M, Carrera P (1996) Seven novel additional small mutations and a new alternative splicing in the human dystrophin gene detected by heteroduplex analysis and restricted RT-PCR heteroduplex analysis of illegitimate transcripts. Eur J Hum Genet 4:183–187

Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S (2006) An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. Hum Mol Genet 15:999–1013

Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M (1997) Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. J Clin Invest 100:2204–2210

Takeshima Y, Yagi M, Wada H, Ishibashi K, Nishiyama A, Kakumoto M, Sakaeda T, Saura R, Okumura K, Matsuo M (2006) Intravenous infusion of an antisense oligonucleotide results in exon skipping in muscle dystrophin mRNA of Duchenne muscular dystrophy. Pediatr Res 59:690–694

Thi Tran HT, Takeshima Y, Surono A, Yagi M, Wada H, Matsuo M (2005) A G-to-A transition at the fifth position of intron 32 of the dystrophin gene inactivates a splice donor site both in vivo and in vitro. Mol Genet Metab 85:213–219

Tran VK, Takeshima Y, Zhang Z, Yagi M, Nishiyama A, Habara Y, Matsuo M (2006) Splicing analysis disclosed a determinant single nucleotide for exon skipping caused by a novel intra-exonic four-nucleotide deletion in the dystrophin gene. J Med Genet (in press)

Published online 28 August 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 18 5995–6003*
doi:10.1093/nar/gkm647

In vitro and *in silico* analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5′ splice sites

Kentaro Sahashi[1,2], Akio Masuda[1], Tohru Matsuura[1], Jun Shinmi[1], Zhujun Zhang[3], Yasuhiro Takeshima[3], Masafumi Matsuo[3], Gen Sobue[2] and Kinji Ohno[1,*]

[1]Division of Neurogenetics and Bioinformatics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, [2]Department of Neurology, Nagoya University Graduate School of Medicine, Nagoya and [3]Department of Pediatrics, Kobe University Graduate School of Medicine, Kobe, Japan

Received April 5, 2007; Revised August 6, 2007; Accepted August 6, 2007

ABSTRACT

We have found that two previously reported exonic mutations in the *PINK1* and *PARK7* genes affect pre-mRNA splicing. To develop an algorithm to predict underestimated splicing consequences of exonic mutations at the 5′ splice site, we constructed and analyzed 31 minigenes carrying exonic splicing mutations and their derivatives. We also examined 189 249 U2-dependent 5′ splice sites of the entire human genome and found that a new variable, the SD-Score, which represents a common logarithm of the frequency of a specific 5′ splice site, efficiently predicts the splicing consequences of these minigenes. We also employed the information contents (R_i) to improve the prediction accuracy. We validated our algorithm by analyzing 32 additional minigenes as well as 179 previously reported splicing mutations. The SD-Score algorithm predicted aberrant splicings in 198 of 204 sites (sensitivity = 97.1%) and normal splicings in 36 of 38 sites (specificity = 94.7%). Simulation of all possible exonic mutations at positions −3, −2 and −1 of the 189 249 sites predicts that 37.8, 88.8 and 96.8% of these mutations would affect pre-mRNA splicing, respectively. We propose that the SD-Score algorithm is a practical tool to predict splicing consequences of mutations affecting the 5′ splice site.

INTRODUCTION

In eukaryotes, splicing of the nuclear mRNA precursor (pre-mRNA) takes place mostly within the U2-dependent spliceosome, a complex of five uridine-rich small nuclear (sn) ribonucleoproteins (RNPs): U1, U2, U4, U5 and U6 snRNPs and numerous non-snRNP proteins. In the first step of spliceosome formation, U1 snRNP recognizes the 5′ splice site and regulates initiation of pre-mRNA splicing (1).

The 5′ splice site is composed of the last three nucleotides of an exon (positions −3, −2 and −1) and the first six nucleotides of an intron (positions + 1 to + 6). The consensus sequence of the U2-dependent 5′ splice sites is (C/A)AG|GT(A/G)AGT (2), where the vertical line (|) represents the exon–intron boundary, and the 'GT' dinucleotide at the 5′ end of an intron is invariable (Figure 1A) (3). In the latter stage of pre-mRNA splicing, U1 snRNA dissociates from the 5′ splice site, and U6 snRNA subsequently binds to nucleotides at positions +2, +5 and +6 (Figure 2A) (4–6).

In the course of identification of exonic splicing mutations in genetic forms of Parkinson's disease, we found two splicing mutations at the 5′ splice site that compromise binding to U1 snRNA. To clarify how exonic mutations at the 5′ splice site cause aberrant splicing, we analyzed 31 minigenes *in vitro* and examined 189 249 putative U2-dependent 5′ GT splice sites of the entire human genome *in silico*. We found that a new variable, the SD-Score, in combination with the information contents (R_i), which represents the amount of information in bits (7,8), can efficiently predict the splicing consequences of exonic mutations at the 5′ splice site. We validated our algorithm with 32 additional minigenes and with 179 previously reported splicing mutations, and found that the SD-Score algorithm has a sensitivity of 97.1% and a specificity of 94.7%. We believe that the SD-Score algorithm is a practical tool for predicting the splicing consequences at the 5′ splice site of mutations causing human disease.

*To whom correspondence should be addressed. Tel: +81-52-744-2446; Fax: +81-52-744-2449; Email: ohnok@med.nagoya-u.ac.jp

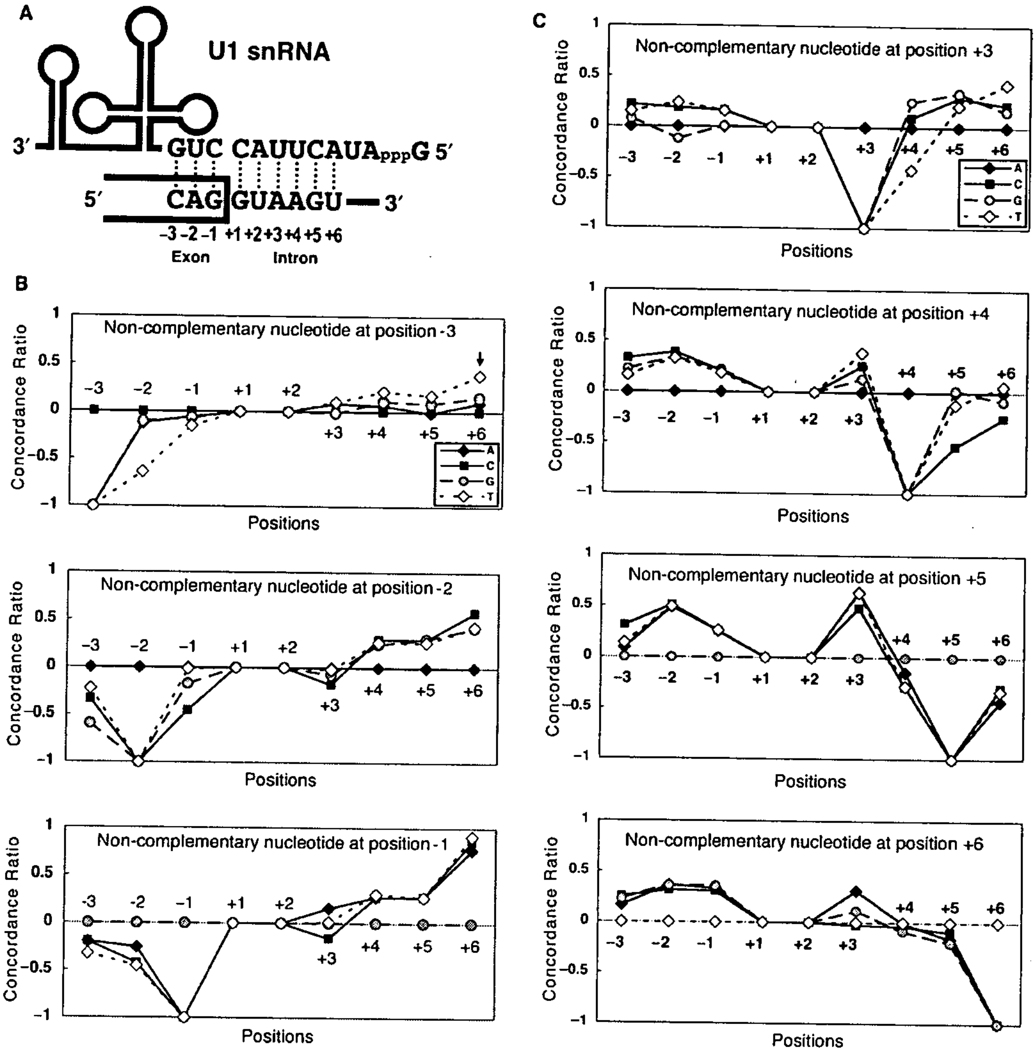

Figure 1. (A) The 5′ end of U1 snRNA recognizes three nucleotides at exonic positions −3, −2 and −1 and six nucleotides at intronic positions +1 to +6. Note that the consensus sequence, 5′-(C/A)AG|GT(A/G)AGT-3′, is complementary to U1 snRNA at most positions. (B and C) The 'two-point' analysis reveals that NCp-nucs at exonic positions −3, −2 and −1 are compensated for by Cp-nucs at positions +4, +5 and +6, and that NCp-nucs at intronic positions +4, +5 and +6 are compensated for by Cp-nucs at positions −3, −2, −1 and +3. For example, when a complementary C is used at position −3 (68 353 sites), the frequency of a complementary T at position +6 is 42.9% (29 307 of 68 353). In contrast, when a noncomplementary T is used at position −3 (22 667 sites), the frequency of a complementary T at position +6 is increased to 60.2% (13 636 of 22 667). The concordance ratio is calculated as (60.2–42.9)/42.9 = 0.403 (arrow). This means that when position −3 is a noncomplementary T, we observe a complementary T at position +6 40.3% more frequently than when position −3 is a complementary C. A positive concordance ratio at a specific position indicates that a Cp-nuc to U1 snRNA is preferentially used to compensate for an NCp-nuc at another position.

MATERIALS AND METHODS

Exon trapping vector

To examine splicing consequences of exonic mutations, we constructed minigenes in an exon-trapping vector, pSPL3 (a discontinued product of Invitrogen, Carlsbad, CA, USA), which was kindly provided by Dr Kazunori Imaizumi, Department of Anatomy, University of Miyazaki, Miyazaki, Japan. We introduced a cytomegalovirus promoter in place of the simian virus 40 promoter by means of a megaprimer-based, site-directed mutagenesis method (9) and also introduced a PacI recognition sequence at the multiple cloning site. Because the nonsense-mediated mRNA decay can degrade a specific splicing product with a premature termination codon and cause misinterpretation of splicing assays (10), we eliminated a premature stop codon and constructed pSPL3 vectors with three reading frames to insert any chimeric exons in-frame (Supplementary Figure 1).

Minigene constructs and mutagenesis

We used the polymerase chain reaction (PCR) to amplify an exon and the flanking introns of 200 bp of the genes of our interest using normal human genomic DNA extracted

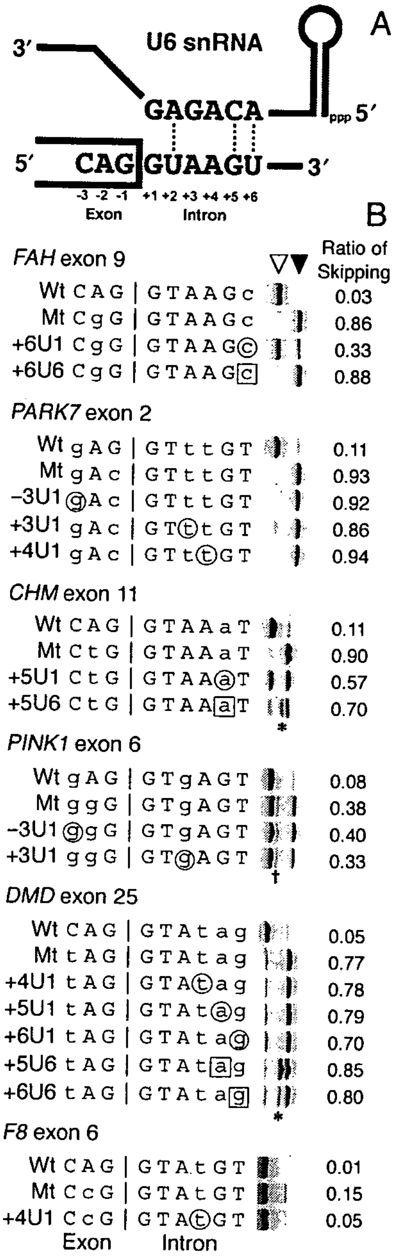

Figure 2. (**A**) The U6 snRNA base pairs with nucleotides at positions +2, +5 and +6. (**B**) RT–PCR analysis of minigene constructs transfected into HEK293 cells with artificial U1 or U6 snRNA. A single Cp-nuc is introduced into U1 or U6 snRNA while retaining the mismatch at the mutation. Wt, wild-type construct; Mt, mutation observed in a patient. For example, +6U1 indicates that a nucleotide on U1 snRNA corresponding to position +6 is substituted to match the 5′ splice site. Circles and squares represent nucleotides that become complementary to the artificial U1 and U6 snRNAs, respectively. The open arrowhead indicates a normally spliced fragment, whereas the closed arrowhead indicates an exon-skipped fragment. The rightmost column shows the densitometric ratio of the exon-skipped fragment. The asterisk indicates a mixture of fragments due to activated 5′ splice sites four and 13 nucleotides downstream of the native 5′ splice site at the 5′ exon of pSPL3. The dagger indicates a heteroduplex formed by normally spliced and exon-skipped products. Uppercase nucleotides are complementary to U1 snRNA, whereas lowercase nucleotides are not.

from HEK293 cells. NotI and PacI recognition sites were introduced to the 5′ and 3′ ends, respectively, of the PCR product. Each amplicon was inserted into one of the three **pSPL3** vectors so that the reading frame of the chimeric exon was retained.

For the *PARK7* and *DYSF* genes, wild-type pSPL3 constructs yielded a large proportion of exon-skipped products. We thus amplified a genomic segment spanning the mutation-harboring exon, the flanking introns and the neighboring exons, and then inserted the amplicon into the **pcDNA3.1(+)** mammalian expression vector (Invitrogen). Different splicing consequences between **pSPL3** and **pcDNA3.1(+)** constructs likely represent the complexity of splicing analysis.

The U1 snRNA gene with its own promoter was kindly provided by Dr Alan M. Weiner, Department of Biochemistry, University of Washington, Seattle, WA, USA. The U6 snRNA gene with the 5′ promoter region of 367 bp and a 3′ end region of 149 bp was amplified using normal human genomic DNA extracted from HEK293 cells and was inserted into the **pGEM-T Easy** Vector (Promega, Madison, WI, USA).

Naturally occurring and artificial mutations were introduced into the inserts with the QuikChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA). We confirmed by sequencing that there were no artifacts in any insert.

Transfection and RNA analysis

HEK293 cells were maintained in the Dulbecco's minimum essential medium (DMEM; Sigma–Aldrich, St Louis, MO, USA) with 10% fetal bovine serum (FBS; Sigma–Aldrich). At ~50% confluency (~5×10^5 cells) in a 6-well plate, 1 ml of fresh Opti-MEM I (Invitrogen) was substituted for DMEM, and 1 µg of a minigene with 3 µl of the FuGENE6 Transfection Reagent (Roche Diagnostics, Indianapolis, IN, USA) were then added. After 4 h, 2 ml of DMEM with 10% FBS was overlaid, and the cells were incubated overnight. The transfection medium was replaced with 2 ml of fresh DMEM with 10% FBS, and the transfected cells were incubated for 48 h before RNA extraction. When artificial U1 or U6 snRNA vector was used, 50 ng of a minigene and 950 ng of each snRNA vector were introduced. Total RNA was extracted using the GenElute Mammalian Total RNA Kit (Sigma–Aldrich). DNA was degraded on-column with the DNase I (Qiagen, Valencia, CA, USA). Twenty percent of the isolated RNA was used as a template for cDNA synthesis with the Oligo(dT)$^{12–18}$ primer (Invitrogen) and the SuperScript II Reverse Transcriptase (Invitrogen). Ten percent of the synthesized cDNA was used as a template for reverse transcriptase (RT)–PCR amplification with primers SD6 (5′-TCTGAGTCACCTGGACAACC-3′) and SA2 (5′-GTGAACTGCACTGTGACAAGCTGC-3′), both of which were on the **pSPL3** vector. For minigenes in **pcDNA3.1(+)**, gene-specific primers were employed. Amplification was performed for 30 to 35 cycles of denaturation at 94°C for 20 s, annealing at 52°C for 20 s and extension at 72°C for 45 s. We measured the signal intensities of the normal and aberrant fragments with the NIH Image 1.63 program. When the ratio of the aberrant product of the mutant construct was increased by 2.5-fold compared with that of the wild-type construct, we considered the mutant construct to have resulted in aberrant splicing. We tried several different thresholds and found that the threshold of 2.5-fold best represents the results of our visual inspections (data not shown).

For the *PINK1*, *CHM*, *BRCA1*, *DYSF*, *F8* and *DMD* genes, we cloned and sequenced all RT–PCR fragments to confirm that the expected normal and aberrant splicings indeed had taken place in these minigenes.

In silico analysis

We extracted all the nonredundant 5′ GT splice sites in the entire human genome using the CDS tags in the NCBI RefSeq Database Build 36.2. Each 5′ splice site on the genome is counted once, even if it is used multiple times in alternatively spliced transcripts. The analysis was performed with the PrimePower HPC2500/Solaris 9 supercomputer (Fujitsu Ltd., Tokyo, Japan). Using the JMP-IN Ver. 5.1.2 software (SAS Institute, Cary, NC, USA), we statistically determined a threshold for each variable using the default settings.

In humans, ~0.1–0.3% of introns are spliced by the minor U12-dependent spliceosome (2,11,12), and ~70% of the U12-dependent introns have GT-AG terminal dinucleotides (13). Previous *in silico* analyses of the human genome identified 275 (12), 469 (14) and 487 (13) GT-AG U12-dependent introns. We thus eliminated 487 U12-dependent 5′ GT splice sites from our analysis, according to the U12 Intron Database (http://genome.imim.es/cgi-bin/u12db/u12db.cgi). Our training and validation data sets (see 'Results' section) did not include any of the known U12-dependent splice sites.

RESULTS

Screening of exonic splicing mutations in genes causing Parkinson's disease

To identify exonic splicing mutations in genetic forms of Parkinson's disease, we analyzed 57 missense, nonsense and synonymous mutations deposited in the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/) (15) in the *SNCA*, *PARK2*, *PINK1* and *PARK7* genes using minigenes (Supplementary Table 1). We found that no mutation affected an exonic splicing enhancer (ESE) or silencer (ESS). Two mutations at the 5′ splice site, however, resulted in skipping of the mutation-harboring exon likely by compromising binding to U1 snRNA. One mutation was E417G in *PINK1* (16) and the other mutation was E64D in *PARK7* (17). To understand how complementary nucleotides (Cp-nucs) to U1 snRNA at the 5′ splice site compensate for noncomplementary nucleotides (NCp-nucs) at other positions, we introduced a single Cp-nuc to the mutant *PINK1* and *PARK7* minigenes while retaining the mutation (Supplementary Figure 2A and Supplementary Table 2). We employed these results to develop an algorithm to predict splicing consequences of mutations at the 5′ splice site.

Recapitulation of aberrant splicings due to previously reported exonic splicing mutations

We next recapitulated aberrant splicings of six previously reported exonic splicing mutations at position −2 in the *CHM*, *FAH*, *HMBS*, *UROS*, *BRCA1* and *CYP27A1* genes (Supplementary Figure 2B and Supplementary Table 2). All mutations except for *CYP27A1* caused the mutation-harboring exon to be skipped. The *CYP27A1* mutation is exceptional because it introduces a Cp-nuc rather than disrupting complementarity (18). We similarly introduced a Cp-nuc to the mutant constructs while retaining the mutation. These results were also used for developing a prediction algorithm of splicing mutations.

Site-directed mutagenesis of a single nucleotide of U1 snRNA and U6 snRNA

Because the 5′ splice site is recognized by both U1 and U6 snRNAs at different stages of pre-mRNA splicing (Figures 1A and 2A), we wondered which nucleotide of the 5′ splice site is most important for binding each snRNA. To this end, we introduced a single Cp-nuc to U1 or U6 snRNA while retaining the mismatch between U1 or U6 snRNA and the mutation.

Among 11 experiments with artificial U1 snRNAs, corrections of U1 snRNA corresponding to position +6 in *FAH* and to +5 in *CHM* ameliorated aberrant splicings, while the others were inefficient (Figure 2B). Among four experiments with artificial U6 snRNAs, only a correction corresponding to position +5 in *CHM* partially normalized aberrant splicing (Figure 2B). In contrast to manipulation of the splicing *cis*-elements, introduced artificial snRNAs are competed by endogenous snRNAs, and hence their effects tend to be compromised. In addition, substitution of these nucleotides might have modified the core secondary structure of U1 or U6 snRNA and made it nonfunctional. Nevertheless, it is interesting to note that corrections of U1 and U6 snRNAs ameliorate aberrant splicings in some mutants even in the presence of mismatch at the mutation. To our knowledge, no similar study has been performed in this scale (19), but the study size was still too small to draw a definite conclusion.

In silico analysis of human 5′ splice sites: the consensus sequence

To examine how the human 5′ splice sites are organized and why the identified exonic mutations resulted in aberrant splicings, we analyzed the 5′ splice sites of the entire human genome. According to the NCBI RefSeq Database, the human genome comprises 28 714 annotated genes with 192 643 5′ splice sites. Of these sites, 189 718 (98.5%) sites carry an invariant GT dinucleotide at positions +1 and +2, whereas 1859 (1.0%) and 311 (0.2%) sites have GC and AT dinucleotides, respectively. The remaining 755 (0.4%) sites carry other dinucleotides and likely include erroneous annotations. We excluded 487 U12-dependent 5′ GT splice sites (see 'Materials and Methods' section) and extracted nine nucleotide segments

Table 1. Nucleotide frequencies (%) at U2-dependent 189 249 human 5′ GT splice sites

Position	−3	−2	−1	+1	+2	+3	+4	+5	+6
A	33.4	63.5	10.0			59.5	69.4	8.9	17.9
C	36.1	10.9	2.8			2.9	7.7	5.7	15.0
G	18.5	11.6	80.3	100.0		34.6	11.8	77.6	19.4
T	12.0	14.0	6.8		100.0	3.0	11.1	7.8	47.7
Consensus sequence	C/A	A	G	G	T	A/G	A	G	T

Nucleotides that are complementary to U1 snRNA are underlined. In this study, we calculated the CV (21) with the equation $\mathrm{CV} = \sum_{l=-3}^{6}(F(n,\mathrm{l}) - 0.570)/5.772$, where F(*n*, l) is a ratio of a nucleotide '*n*' at position 'l'. Similarly, we calculated the R_i (7,8) with the equation $R_i = \sum_{l=-3}^{6}(2 + \log_2(\mathrm{F}(n,\mathrm{l})))$. We ignored the error function of R_i, because F(*n*, l) values are calculated using a large number of observations (189 249 sites), and hence the contribution of the error function should be negligible.

spanning positions −3 and +6 from the remaining 189 249 5′ GT splice sites.

Analysis of nucleotide frequencies at each position showed that the 'winner sequence' comprising the most frequently used nucleotides was CAG|GTAAGT (Table 1), which is entirely complementary to U1 snRNA (Figure 1A). The frequency of Cp-nuc was highest at position −1, followed in descending order by positions +5, +4, −2, +3, +6 and −3 (excluding positions +1 and +2, which are invariant in our analysis).

In silico analysis of human 5′ splice sites: the 'two-point' analysis

We next analyzed how an NCp-nuc to U1 snRNA at a specific position is compensated for by a Cp-nuc at the other positions. The 'two-point' analysis revealed that NCp-nucs at positions +4, +5 and +6 are compensated for by Cp-nucs at positions −3, −2, −1 and +3 (Figure 1C). Conversely, NCp-nucs at positions −3, −2 and −1 are associated with high concordance ratios at positions +4, +5 and +6 (Figure 1B). These results suggest that a stretch of Cp-nucs either in an exonic or an intronic region is essential for proper splicing, which also conforms to the notion that consecutive base pairings with U1 snRNA contribute to recognition of the 5′ splice site (20). It is also interesting to note that a Cp-nuc at position +6 most frequently compensates for an NCp-nucs at position −3, −2 and −1, although the frequency of a Cp-nuc at position +6 is only 47.7% in the human genome.

In our analysis, we assumed that only A at position +3 is a Cp-nuc. When we regarded both A and G as Cp-nucs at position +3, the concordance ratio at position +3 became always low (Supplementary Figure 3), likely because A or G is observed at position +3 in 94.1% of the human 5′ splice sites. This implies that the concordance ratio is less informative, when the frequency of a Cp-nuc is high.

In silico analysis of human 5′ splice sites: the SD-Score

The 'two-point' analysis disclosed interdependence between two nucleotides at the 5′ splice sites. However, we could not develop a scoring system using the 'two-point' analysis. We thus sought another quantitative measure to predict splicing consequences. We expected that the frequency of a specific 5′ splice site sequence in the human genome would represent the splicing signal intensity that we hoped to score. We analyzed the entire human genome and determined the frequency of each U2-dependent GT splice site sequence. The common logarithm of the frequency was calculated to give a new variable, the SD-Score (Supplementary Table 5). For example, the SD-Score for CAG|GTGAGG, which was observed at 2562 sites, is log (2562/189 249) = −1.868. The SD-Score of a splice site sequence that never appears in the human genome should be log (0/189 249) = −∞ but is defined as log (0.25/189 249) = −5.879 to simplify calculations. The correlation coefficients of the SD-Score with the R_i (7,8) and the consensus value (CV), which represents the similarity of a splice site sequence to the consensus splice site sequence (21), are 0.678 and 0.694, respectively, indicating that the SD-Score is similar to, but distinct from, the other variables.

Prediction of splicing consequences using the SD-Score algorithm

We next examined whether the SD-Score is indeed an effective scoring variable. To this end, we plotted the SD-Scores and the ΔSD-Scores of the 31 constructs of the *PARK7*, *PINK1*, *CHM*, *FAH*, *HMBS*, *UROS*, *BRCA1* and *CYP27A1* genes (Supplementary Table 2). The ΔSD-Score is calculated by subtracting the wild-type SD-Score from the mutant SD-Score. We found that a 5′ splice site with ΔSD-Score > −0.34 does not affect pre-mRNA splicing in 13 out of 13 sites, whereas a mutant site with ΔSD-Score < −0.34 and SD-Score < −2.9 causes aberrant splicing in 9 out of 10 sites. The 5′ splice sites with ΔSD-Score < −0.34 and SD-Score > −2.9 include a mixed population of three normal and five aberrant splicings. We thus employed the ΔR_i value, which is calculated by subtracting the wild-type R_i from the mutant R_i, and found that three out of three sites with $\Delta R_i > -1.45$ are normally spliced, whereas five out of five sites with $\Delta R_i < -1.45$ are aberrantly spliced. Therefore, these thresholds efficiently predict splicing consequences of our 31 minigene constructs (Figure 3).

Previously unrecognized exonic splicing mutations at positions −2 and −1

To validate the SD-Score algorithm, we employed previously unrecognized splicing mutations at positions

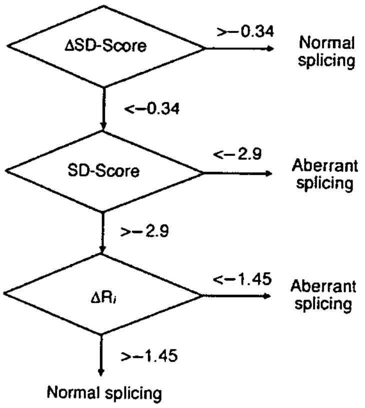

Figure 3. The diagram demonstrates the SD-Score algorithm to predict aberrant splicings due to mutations at the 5′ splice site. The algorithm is based on a training dataset of 31 minigenes and was validated with testing data sets of 32 additional minigenes and 179 naturally occurring splicing mutations (Supplementary Tables 2–4).

−2 and −1. To this end, we scrutinized 2477 exonic mutations from the Human Gene Mutation Database, and searched for mutations at positions −2 and −1. We then randomly selected three mutations in the *DYSF*, *F8* and *ABCD1* genes (Supplementary Table 2) whose splicing consequences have not been characterized. The SD-Score algorithm predicted that all the mutations would affect pre-mRNA splicings, and minigene analyses confirmed it (Supplementary Figure 2C and Supplementary Table 2).

We further introduced a single Cp-nuc to each mutant minigene while retaining the mutation. We thus constructed seven artificial minigenes, and six of these were spliced as predicted (Supplementary Figure 2C and Supplementary Table 2).

Previously unrecognized splicing mutations at position −3

We next sought exonic splicing mutations at position −3, which has been reported only in two mutations (22,23) according to the Human 5′ Splice Site Database (http://www.uni-duesseldorf.de/rna/). In the 2477 exonic mutations described above, we identified six mutations that disrupt a complementary C nucleotide at position −3 (Supplementary Table 2). The SD-Score algorithm predicted that four mutations in the *GLA*, *DMD* and *PARK2* genes would affect pre-mRNA splicing, whereas two mutations in the *ABCD1* and *NPC1* genes would not. We analyzed six mutant minigenes and found that all, except the *PARK2* mutant, were spliced as predicted (Supplementary Figures 2D and 5, and Supplementary Table 2). We confirmed in patient's lymphocytes that a C-to-T mutation at position −3 in *DMD* exon 5 indeed caused the same aberrant splicing as we observed with the minigene (data not shown). The aberrant splicing due to a mutation in *DMD* exon 5, however, was likely successfully predicted by the SD-Score algorithm, because additional mutagenesis at position −4, which did not create a novel cryptic site, failed to show exon skipping (Supplementary Figure 5).

We also constructed seven artificial minigenes, and the SD-Score algorithm successfully predicted the splicing consequences of all the minigenes (Supplementary Figure 2D and Supplementary Table 2).

Splicing mutations in the literature database

To further validate the SD-Score algorithm, we employed other exonic and intronic splicing mutations in the literature database (Supplementary Tables 3 and 4). We randomly examined 2, 9, 26, 45, 3, 83 and 11 splicing mutations at positions −3, −2, −1, +3, +4, +5 and +6, respectively. Our algorithm correctly predicted aberrant splicings in 174 of the 179 reported mutations and falsely predicted normal splicings in five mutations.

DISCUSSION

Clinical implications of exonic splicing mutations

Although our analysis failed to detect ESE- and ESS-affecting mutations in genetic forms of Parkinson's disease, we identified two exonic splicing mutations at the 5′ splice site: E417G in *PINK1* and E64D in *PARK7*. These mutations, as well as six other previously unrecognized exonic splicing mutations in the *DYSF*, *F8*, *ABCD1*, *GLA* and *DMD* genes (Supplementary Table 2), have been reported as synonymous, missense or nonsense mutations. Discrimination of splicing mutations from other types of mutations is essential for understanding human diseases, because different phenotypes and different therapeutic options should be considered for different disease mechanisms. For example, splicing abnormalities in the *IKBKAP* and *SMN2* genes can be normalized with kinetin (24) and sodium valproate (25), respectively.

Prediction of aberrant splicings using the SD-Score algorithm

To predict aberrant splicings due to mutations at the 5′ splice site, we developed the SD-Score algorithm using a training dataset and tested it using a validation dataset. Except for the *PINK1* and *PARK7* genes, we selected mutations without any bias in both minigenes and previously reported splicing mutations in the literature database. Of the 63 minigenes examined in the present study, six normally spliced and seven aberrantly spliced minigenes required the use of ΔR_i values for analysis. In contrast, of the 179 splicing mutations in the literature database, only four mutations required the ΔR_i values for analysis. Artificial minigenes that we constructed to understand interdependence between Cp-nucs and NCp-nucs carry two nonnative nucleotides, whereas naturally occurring mutations carry a single nonnative nucleotide. The SD-Score alone may not be powerful enough to predict the splicing consequences of mutants carrying two or more nonnative nucleotides at the 5′ splice site.

Recognition of an exon, however, is dependent not only on the 5′ splice site sequence, but also on other splicing *cis*-elements, including the branch point, the polypyrimidine tract, the 3′ splice site and ESEs/ESSs and intronic enhancers/silencers. Lack of information about the other

Table 2. Sensitivity and specificity of the SD-Score algorithm

Prediction	Aberrantly spliced^{a}	Normally spliced^{a}
Aberrant splicing^{b}	198 (24^{c}/174^{d})	2 (2^{c}/0^{d})
Normal splicing^{b}	6 (1^{c}/5^{d})	36 (36^{c}/0^{d})
Total	204 (25^{c}/179^{d})	38 (38^{c}/0^{d})

The table shows the ^{a}actual and ^{b}predicted splicing consequences of ^{c}63 minigenes and ^{d}179 splicing mutations at the 5′ splice site in the literature database. The overall sensitivity of the SD-Score algorithm is 97.1% (198 of 204) and the specificity is 94.7% (36 of 38). The specificity is dependent on only our minigene results, because no report has been made, in which a mutation at the 5′ splice site has no effect on pre-mRNA splicing.

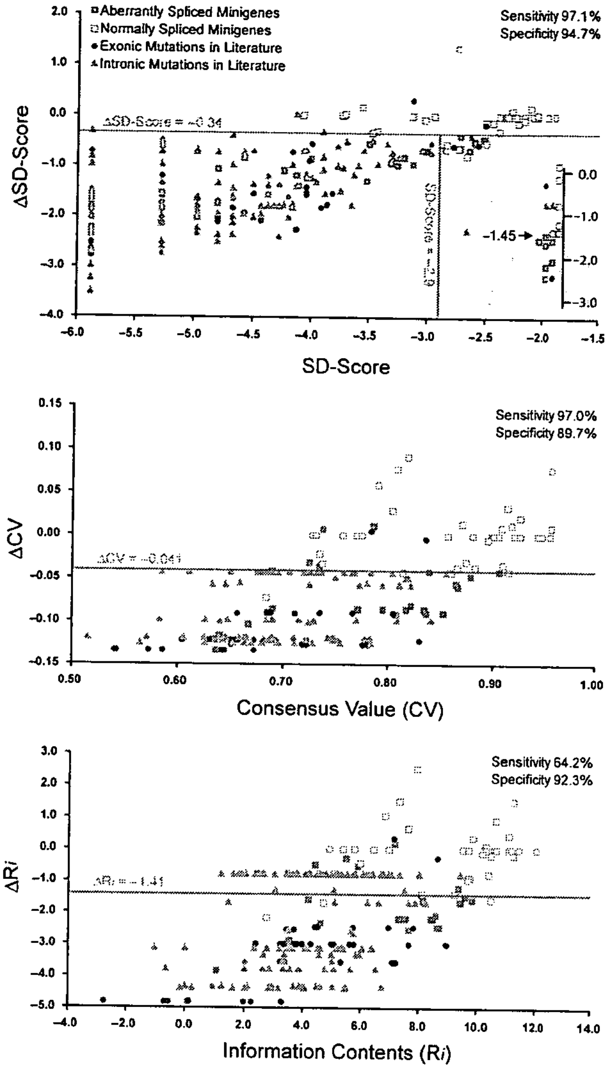

Figure 4. Scatter graphs of the SD-Scores (**A**), the CVs (**B**) and the R_is (**C**) of 63 minigenes and 179 splicing mutations in the literature database. Thirty-nine normally spliced and 24 aberrantly spliced minigenes and 37 exonic and 142 intronic splicing mutations are plotted on each graph. Thresholds for the CVs and R_is were determined with the JMP-IN statistical software to give the best discrimination between normal and aberrant splicings. For the SD-Score, we used 31 minigenes as a training data set and other 32 minigenes as a validation data set. We obtained the similar thresholds of the SD-Score, even when we included 63 minigenes in our training data set (data not shown).

splicing *cis*-elements and possible errors in the NCBI RefSeq annotations make the SD-Score algorithm less accurate. In addition, our training dataset comprises exclusively minigenes, and minigenes are not always spliced in the same way as their endogenous counterparts (Supplementary Figure 4). Moreover, the SD-Score algorithm is not trained to predict if any of exon-skipping, activation of a cryptic site, and intron retention occurs due to a mutation. The SD-Score algorithm, however, can efficiently predict splicing consequences of our datasets with a sensitivity of 97.1% and a specificity of 94.7% (Table 2).

Comparison with the free energy, CV and R_i

Roca and colleagues (26) reported that the free energy between the 5′ splice site and U1 snRNA can be used to predict the 5′ splice site strength. The SD-Score can correctly predict 24 out of the 26 active and inactive cryptic sites in their series (data not shown). The Pearson's correlation coefficient between the SD-Score and the free energy in their series was 0.792, implying that these two parameters are likely mutually dependent.

We used our training dataset of 31 minigenes and validation dataset of 32 additional minigenes in an attempt to develop similar algorithms for the CV (21) and the R_i (7,8). We found that either the CV or R_i alone is not as efficient as the SD-Score algorithm for predicting splicing mutations (Figure 4). Both the CVs and R_is are based on the sum of information of each position in a sequence. On the other hand, the SD-Score represents information of the entire sequence of the 5′ splice site, which should include mutual interdependence between multiple positions. The SD-Score, CV and R_i, however, are mutually complementary, and our algorithm indeed achieved a high sensitivity and a high specificity with the help of ΔR_i values.

We also attempted to create a similar algorithm for the 3′ splice site but were unsuccessful, likely because the 3′ splice site includes at least three splicing *cis*-elements, and because a limited number of splicing mutations have been identified at the 3′ splice site.

Underestimated exonic splicing mutations

Most exonic splicing mutations affecting the 5′ splice site have been reported at position −1. On the other hand, to our knowledge, only 2 and 14 exonic splicing mutations have been reported at position −3 and −2, respectively (Supplementary Tables 2 and 3). When we introduced *in silico* all possible mutations that substitute an NCp-nuc for a Cp-nuc at positions −3, −2 and −1 into the 189 249 5′ splice sites in the human genome, the SD-Score algorithm predicted that 37.8%, 88.8% and 96.8% of these mutations would affect pre-mRNA splicings, respectively (Table 3). These percentages, as well as those of

Table 3. Predicted ratios of exonic and intronic splicing mutations

Position Complementary nucleotide	−3 C	−2 A	−1 G	+1 G	+2 T	+3 A	+4 A	+5 G	+6 T
A	1.8	–	93.7	–	–	–	–	93.9	56.9
C	–	89.6	99.7	–	–	99.9	94.4	98.6	75.4
G	35.0	90.5	–	–	–	48.7	96.2	–	56.7
T	76.7	86.2	97.1	–	–	99.9	94.3	97.0	–
All mutations	37.8	88.8	96.8	–	–	82.8	95.0	96.5	63.0

Numbers indicate the percentages of generating splicing mutations according to the SD-Score algorithm. The mutations are weighed by the number of occurrences of the native 5′ splice site. For example, the CAG|GTGAGG sequence, which is observed at 2562 splice sites in the human genome, has a SD-Score of −1.868. A C-to-T mutation at position −3 should generate TAG|GTGAGG, which is observed at 145 splice sites and has a SD-Score of −3.116. The ΔSD-Score of the mutation is thus −1.247. This mutation is predicted to cause aberrant splicing and is counted as 2562 mutations instead of one, because the chance that this mutation occurs should be higher than those of rare 5′ splice sites. Only mutations that substitute an NCp-nuc for a Cp-nuc are considered in this analysis, and 2466918 mutations have been simulated.

intronic mutations at the 5′ splice site, are much higher than we expected. We hope that the SD-Score algorithm serves as a practical tool to predict splicing mutations at the 5′ splice site and sheds light on underestimated aberrant splicings in human diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online. Supplementary Table 5 is an Excel program to calculate the SD-Score algorithm.

ACKNOWLEDGEMENTS

We acknowledge Dr Kazunori Imaizumi, University of Miyazaki, Miyazaki, Japan and Dr Alan M. Weiner, University of Washington, Seattle, WA, USA for providing us with experimental materials and Dr Masao Okazaki, Jikei University, Tokyo, Japan for preparing publication materials. This work was supported by Grants-in-Aid for Scientific Research (B) and Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and by Research Grant for Nervous and Mental Disorders from the Ministry of Health, Labor, and Welfare of Japan, and was also supported in part by grants from the Naito Foundation and the Takeda Science Foundation. Funding to pay the Open Access publication charges for this article was provided by Grants-in-Aid for the Scientific Research on Priority Areas "System Genomics" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Reed,R. (2000) Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.*, **12**, 340–345.
2. Zhang,M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
3. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
4. Lesser,C.F. and Guthrie,C. (1993) Mutations in U6 snRNA that alter splice site specificity: implications for the active site. *Science*, **262**, 1982–1988.
5. Kandels-Lewis,S. and Seraphin,B. (1993) Involvement of U6 snRNA in 5′ splice site selection. *Science*, **262**, 2035–2039.
6. Kim,C.H. and Abelson,J. (1996) Site-specific crosslinks of yeast U6 snRNA to the pre-mRNA near the 5′ splice site. *RNA*, **2**, 995–1010.
7. Rogan,P.K. and Schneider,T.D. (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.*, **6**, 74–76.
8. Rogan,P.K., Faux,B.M. and Schneider,T.D. (1998) Information analysis of human splice site mutations. *Hum. Mutat.*, **12**, 153–171.
9. Ohno,K., Anlar,B., Özdirim,E., Brengman,J.M., DeBleecker,J.L. and Engel,A.G. (1998) Myasthenic syndromes in Turkish kinships due to mutations in the acetylcholine receptor. *Ann. Neurol.*, **44**, 234–241.
10. Ohno,K., Milone,M., Shen,X.M. and Engel,A.G. (2003) A frameshifting mutation in *CHRNE* unmasks skipping of the preceding exon. *Hum. Mol. Genet.*, **12**, 3055–3066.
11. Burge,C.B., Padgett,R.A. and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.
12. Levine,A. and Durbin,R. (2001) A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res*, **29**, 4006–4013.
13. Alioto,T.S. (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.*, **35**, D110–115.
14. Sheth,N., Roca,X., Hastings,M.L., Roeder,T., Krainer,A.R. and Sachidanandam,R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.
15. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
16. Hatano,Y., Li,Y., Sato,K., Asakawa,S., Yamamura,Y., Tomiyama,H., Yoshino,H., Asahina,M., Kobayashi,S. *et al.* (2004) Novel *PINK1* mutations in early-onset parkinsonism. *Ann. Neurol.*, **56**, 424–427.
17. Hering,R., Strauss,K.M., Tao,X., Bauer,A., Woitalla,D., Mietz,E.M., Petrovic,S., Bauer,P., Schaible,W. *et al.* (2004) Novel homozygous p.E64D mutation in *DJ1* in early onset Parkinson disease (PARK7). *Hum. Mutat.*, **24**, 321–329.
18. Chen,W., Kubota,S., Ujike,H., Ishihara,T. and Seyama,Y. (1998) A novel Arg362Ser mutation in the sterol 27-hydroxylase gene (*CYP27*): its effects on pre-mRNA splicing and enzyme activity. *Biochemistry (Mosc)*. **37**, 15050–15056.
19. Carmel,I., Tal,S., Vig,I. and Ast,G. (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA*, **10**, 828–840.
20. Kammler,S., Leurs,C., Freund,M., Krummheuer,J., Seidel,K., Tange,T.O., Lund,M.K., Kjems,J., Scheid,A. *et al.* (2001) The sequence complementarity between HIV-1 5′ splice site SD4 and U1 snRNA determines the steady-state level of an unstable env pre-mRNA. *RNA*, **7**, 421–434.
21. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
22. Williamson,D., Brown,K.P., Langdown,J.V. and Baglin,T.P. (1995) Haemoglobin Dhofar is linked to the codon 29C–>T (IVS-1 nt-3) splice mutation which causes beta+ thalassaemia. *Br. J. Haematol.*, **90**, 229–231.
23. Ries,S., Aslanidis,C., Fehringer,P., Carel,J.C., Gendrel,D. and Schmitz,G. (1996) A new mutation in the gene for lysosomal acid

lipase leads to Wolman disease in an African kindred. *J. Lipid Res.*, 37, 1761–1765.

24. Slaugenhaupt,S.A., Mull,J., Leyne,M., Cuajungco,M.P., Gill,S.P., Hims,M.M., Quintero,F., Axelrod,F.B. and Gusella,J.F. (2004) Rescue of a human mRNA splicing defect by the plant cytokinin kinetin. *Hum. Mol. Genet.*, **13**, 429–436.
25. Brichta,L., Holker,I., Haug,K., Klockgether,T. and Wirth,B. (2006) In vivo activation of SMN in spinal muscular atrophy carriers and patients treated with valproate. *Ann. Neurol.*, **59**, 970–975.
26. Roca,X., Sachidanandam,R. and Krainer,A.R. (2005) Determinants of the inherent strength of human 5′ splice sites. *RNA*, **11**, 683–698.

BMC Medical Genetics

BioMed Central

Research article

Two novel missense mutations in the myostatin gene identified in Japanese patients with Duchenne muscular dystrophy

Atsushi Nishiyama, Yasuhiro Takeshima, Kayoko Saiki, Akiko Narukage, Yoshinobu Oyazato, Mariko Yagi and Masafumi Matsuo*

Address: Department of Pediatrics, Kobe University Graduate School of Medicine, Kobe, 6500017, Japan

Email: Atsushi Nishiyama - 034d595m@y03.kobe-u.ac.jp; Yasuhiro Takeshima - takesima@med.kobe-u.ac.jp; Kayoko Saiki - ksaiki@med.kobe-u.ac.jp; Akiko Narukage - ped@med.kobe-u.ac.jp; Yoshinobu Oyazato - oyazato@med.kobe-u.ac.jp; Mariko Yagi - myagi@med.kobe-u.ac.jp; Masafumi Matsuo* - matsuo@kobe-u.ac.jp

* Corresponding author

Published: 12 April 2007

BMC Medical Genetics 2007, **8**:19 doi:10.1186/1471-2350-8-19

Received: 25 October 2006

Accepted: 12 April 2007

This article is available from: http://www.biomedcentral.com/1471-2350/8/19

Abstract

Background: Myostatin is a negative regulator of skeletal muscle growth. Truncating mutations in the myostatin gene have been reported to result in gross muscle hypertrophy. Duchenne muscular dystrophy (DMD), the most common lethal muscle wasting disease, is a result of an absence of muscle dystrophin. Although this disorder causes a rather uniform pattern of muscle wasting, afflicted patients display phenotypic variability. We hypothesized that genetic variation in myostatin is a modifier of the DMD phenotype.

Methods: We analyzed 102 Japanese DMD patients for mutations in the myostatin gene.

Results: Two polymorphisms that are commonly observed in Western countries, p.55A>T and p.153K>R, were not observed in these Japanese patients. An uncommon polymorphism of p.164E>K was uncovered in four cases; each patient was found to be heterozygous for this polymorphism, which had the highest frequency of the polymorphism observed in the Japanese patients. Remarkably, two patients were found to be heterozygous for one of two novel missense mutations (p.95D>H and p.156L>I). One DMD patient carrying a novel missense mutation of p.95D>H was not phenotypically different from the non-carriers. The other DMD patient was found to carry both a novel mutation (p.156L>I) and a known polymorphism (p.164E>K) in one allele, although his phenotype was not significantly modified. Any nucleotide change creating a target site for micro RNAs was not disclosed in the 3' untranslated region.

Conclusion: Our results indicate that heterozygous missense mutations including two novel mutations did not produce an apparent increase in muscle strength in Japanese DMD cases, even in a patient carrying two missense mutations.

Background

Duchenne muscular dystrophy (DMD), the most common inherited myopathy affecting approximately one in 3,500 males, is characterized by muscle dystrophin deficiency. Dystrophin deficiency is caused by translational reading frame shifts or nonsense mutations in the dystrophin gene [1]. DMD is a rapidly progressive disease occurring during childhood that causes affected individuals to lose their ability to walk by the age of 12 years old before they succumb during their twenties due to either respiratory or cardiac failure.

DMD is known to progress with a rather uniform pattern of muscle weakness, However, the existence of a modifying gene has been suggested due to the identification of unusually mild DMD phenotypes [2-4]. Some phenotypic variability has been explained by the precise locations of the mutations and their effects on the dystrophin-dystroglycan complex [5,6] or by the identification of aberrant splicing products from the dystrophin gene [7-10]. In some cases, however, the same dystrophin mutation has been reported to result in different phenotypes [11,12].

Although some phenotypic variability may arise due to environmental factors, such as diet or exercise, there are likely to be contributions from genetic components. In fact, differences in genetic backgrounds have been shown to influence the phenotypes of mice with a dystrophin-glycoprotein complex disorder caused by a mutation in the σ-sarcoglycan gene [13]. Therefore, it is highly plausible that unknown genetic factors modify the phenotype of DMD.

Myostatin, also known as growth and differentiation factor 8 (GDF8), is a muscle-specific secreted peptide that functions to limit muscle growth [14]. Several studies analyzing mutations of the myostatin gene have been conducted in the Western worlds [15-17]. To date, six polymorphisms and one intronic mutation have been identified in the myostatin gene. One of the identified polymorphisms (p.153K>R) has been associated with a hypertrophic response in muscles due to strength training [18]. Recently, an infant was identified with the first homozygous disruption of the myostatin gene, which resulted in the child being exceptionally muscular at birth and unusually strong with increased muscle mass at four years of age [19]. Remarkably, a single nucleotide change creating a potential illegitimate micro RNA target site in the 3' untranslated region of the sheep myostatin gene was disclosed to cause translation inhibition leading to the increase of muscularity [20].

Furthermore, disruption of endogenous myostatin by gene or RNA targetings was shown to result in anatomic, biochemical, and physiologic improvements in the dystrophic phenotype of *mdx* mice, a mouse model of DMD with a nonsense mutation in the dystrophin gene [21,22], including particularly prominent enlarged fiber diameters and greatly reduced fatty fibrosis. These results suggest that blocking endogenous myostatin is a potential strategy for treatment of DMD [23].

We hypothesized that genetic variation in the myostatin gene modifies the phenotype of DMD. Therefore, nucleotide changes in the myostatin gene were investigated in Japanese DMD patients, resulting in the identification of novel mutations.

Methods

Subjects

One hundred two DMD patients that were followed up at Kobe University Hospital were enrolled into this study. All of the mutations in the dystrophin genes were revealed to introduce premature stop codons in the dystrophin mRNA; 51 cases with mutations that induced a translational reading frame shift due to exon deletion or duplication, 31 cases with nonsense mutations, 12 cases with mutations of one or a few nucleotides deletion or insertion, and 8 cases with intron mutations that induced splicing error (data not shown). The subjects' ages ranged from 1 to 31 years old (average: 10 years old). Regular clinical check-ups, including determination of the serum creatine kinase (CK) levels, were performed at the outpatient clinic. The maximal voluntary isometric torque (MVIT) produced by the elbow flexor muscles and the knee extensor muscles was measured with a manual dynamometer (Microfet2 digital muscle tester, Value Medical Supplies, Hesperia, CA) with a precision of 0.1 Nm. A clear difference in the phenotypes was observed in the ages at which the patients became wheelchair bound, which occurred between the ages of 5 and 11 years old. Some patients, however, were able to walk independently after they were 12 years old even though they carried mutations that caused truncations of the dystrophin protein.

Protocols of this study were approved by the ethics committee of the Kobe University School of Medicine. Blood samples were taken after written informed consent was obtained.

Sequencing analysis of the myostatin gene

Genomic DNA samples were prepared from the peripheral blood of the patients via the standard phenol-chloroform extraction method and were used as templates for PCR amplification. All three of the myostatin exons were examined by PCR amplification and direct sequencing (Fig. 1). Exons 1 and 3 of the myostatin gene were PCR amplified as previously described [19]. Exon 1 was amplified as a 542-bp fragment, including 131 and 373 bp of the 5' untranslated and protein coding regions, respectively, as well as 38 bp of intron 1 (Fig. 1). Exon 3 was amplified as a 536-bp fragment including 43 bp of intron 2 and 381 and 112 bp of the protein coding and 3' untranslated regions, respectively (Fig. 1). Additionally a middle part of the 3' untranslated region was amplified as 396-bp fragment using a set of two primers (3UF: 5'-CATGTCATGCATCACAGAAAAGCAACTACT-3' and 3UR: 5'-

CAAAATCCCAATTTACAAAACAGAA-3'), since a single nucleotide change in this region of the sheep myostatin gene has been shown to create a microRNA target site, thereby leading translation inhibition [20]. Exon 2 was amplified using two primers (2F: 5'-ATTAATATGGAGGGGTTTTGTTAATGG-3', 2R: 5'-GCTTAGGGAATTTGTAGCTATTTTCCA-3') that resulted in a 537-bp fragment, including the 374 bp of exon 2, and 68 and 95 bp of introns 1 and 2, respectively.

Denaturation of the DNA was performed at 96°C for 5 min, followed by 35 cycles of denaturation at 96°C for 1 min, annealing at 60°C for 1 min, and elongation at 72°C for 1.5 min. The amplified products were analyzed on a 2% agarose gel and visualized by ethidium bromide staining. The PCR-amplified products were directly sequenced using a BigDye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA) and an automated DNA sequencer (ABI Prism 310 Genetic Analyzer; Perkin Elmer Applied Biosystems). For subcloning sequencing, the PCR-amplified products were cloned into the pT7 blue T vector (Novagen, Madison, WI) and sequenced. Sequencing results were compared with the wild-type sequence (Genbank: AC073120).

Results

All three coding regions and a part of the 3' untranslated region of the myostatin gene were successfully PCR amplified from genomic DNA; 102 DNA samples were subjected to direct sequencing. Sequencing results of the exon

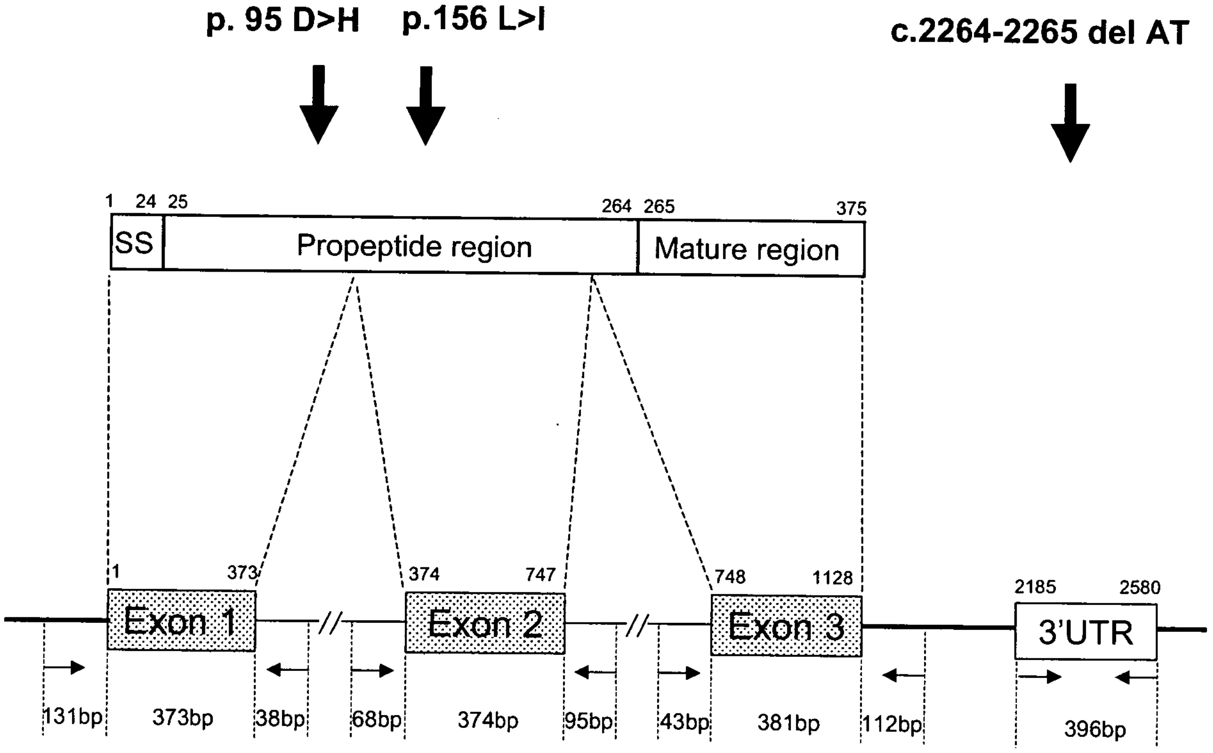

Figure 1

Myostatin and the human myostatin gene. The structure of myostatin, including the signal sequence (SS), and the regions of the propeptide and mature protein, is schematically described (Top). The numbers above the boxes indicate the amino-acid residue position. The vertical arrows indicate the locations of the two novel missense mutations and two nucleotides deletion identified in this study. The structure of the myostatin gene and the analyzed regions are schematically described (Bottom). Three coding regions (Exons 1, 2 and 3) and a part of the 3' untranslated region (3'UTR) of the myostatin gene were PCR amplified (boxes). The shaded and open boxes indicate the coding region and the sequenced region in the 3' untranslated region, respectively. Bold and thin horizontal lines indicate exons and introns, respectively. Horizontal arrows indicate the locations and directions of the primers used to amplify the regions. Numbers above the boxes indicate the nucleotide position according to the cDNA reference sequence in GenBank (accession no.: NM_005259), in which the "A" in the start codon is nucleotide #1. Numbers in the bottom indicate the size of each segment.

1-encompassing region disclosed completely normal sequences in all of the samples except for one nucleotide change in one sample. In this case (case 712), direct sequencing revealed overlapping G and C peaks at the 283rd position of the myostatin cDNA (c.283G>G/C) (Fig. 2). Because this nucleotide position is a G in the wild-type sequence, the presence of a C at this position was determined to be a mutation (c.283G>C).

c.283G>C changed the codon corresponding to the 95th amino-acid residue of myostatin from GAT to CAT, which substituted an Asp residue to a His residue (p.95D>H). This missense mutation was located at a conserved amino-acid residue in the propeptide region (Fig. 1) [24] and was predicted to affect the function of myostatin. Clinical examination of muscle strength, however, failed to reveal a clear difference between the DMD patient carrying this nucleotide change and the other DMD patients. At 8 years old, the patient could not stand up by himself, but was able to walk independently with a waddling gait.

In the exon 2-encompassing region, overlapping G and A peaks at the 490th nucleotide of the myostatin cDNA (c.490G>G/A) were uncovered in four samples. c.490G>A corresponded to a known polymorphism that changes a GAG codon for Glu to a AAG codon for Lys at the position

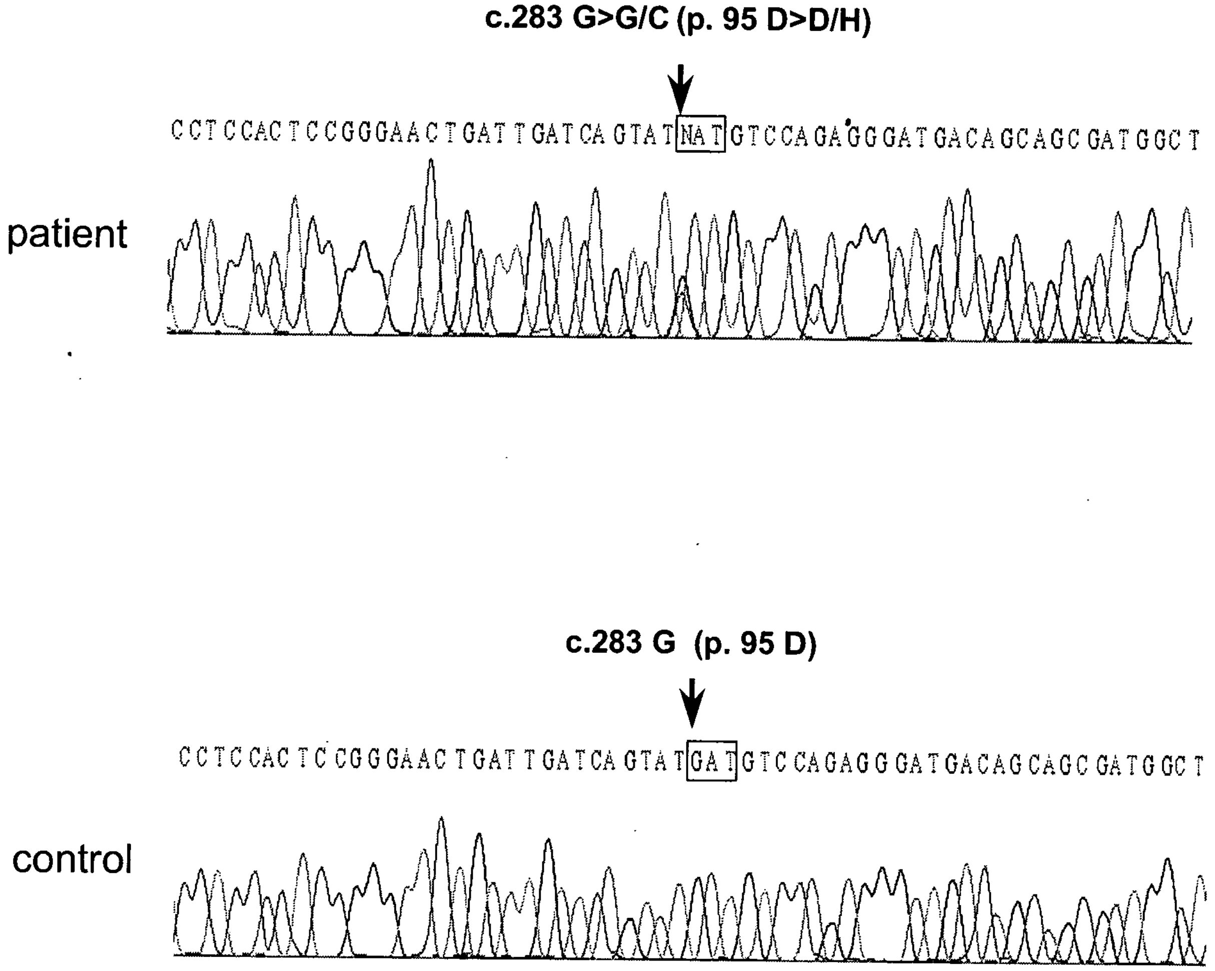

Figure 2
A novel mutation in exon 1 of the myostatin gene. A part of the sequencing results for exon 1 of the myostatin gene from a DMD patient (case 712) is shown (patient). Overlapping G and C peaks are present at the 283rd nucleotide of the myostatin cDNA (c.283G>G/C) (Top). The single nucleotide change from G to C at the 283rd nucleotide of the myostatin cDNA (c.283G>C) changed a GAT codon to a CAT codon at the position corresponding to the 95th amino-acid residue of myostatin (p.95D>H)(boxes). The wild type sequence is shown below (control).

corresponding the 164th amino-acid residue of myostatin (p.164E>K) [15]. The allele frequency of c.490G>A was 2.0% (4 of 204 alleles). Interestingly, in one of the four samples with c.490G>A (case 549), overlapping C and A peaks were also found at the 466th nucleotide of the myostatin cDNA (c.466C>C/A) (Fig. 3). Subsequent subcloning sequencing disclosed the presence of two clones; one was identical to the wild-type sequence (Fig. 3), whereas the second clone carried the two nucleotide changes observed with direct sequencing: c.466C>A and c.490G>A (Fig. 3). c.466C>A, which has not been previously described, changed a CTA codon for Leu to a ATA codon for Ile (p.156L>I). Because the novel missense mutation was located at a conserved residue, this mutation was predicted to disrupt myostatin function (Fig. 1), particularly when it was combined with the second amino acid substitution located only eight amino-acid residues away. Clinical examination of the patient carrying the allele encoding two nucleotide changes did not disclose a clear mild DMD phenotype in the 14 year old. He was wheel-chair bound at 7 years old. The phenotypes of the other three DMD cases carrying c.490G>A in one allele were clinically indistinguishable from the other DMD patients.

None of these Japanese DMD cases carried p.55A>T or p.153K>R, frequently observed polymorphisms in the United States, or the single nucleotide change in intron 2 that has been reported to result in gross muscular hypertrophy (Table 1) [15]. Moreover, examination of the exon 3-encompassing region (Fig. 1) did not disclose any nucleotide changes in the genomic samples. This was compatible with results obtained in previous studies.

Sequencing of the 396-bp fragment of the 3' untranslated region revealed no nucleotide change, especially in the stretch of ACGTTCCA (an underlined G is 2402nd nucleotide where the substitution of G with A has been reported to create the octamer motif for the micro RNA target site in the sheep myostatin gene [20]). Exceptionally, one DMD case (case 100) was found to have a deletion of AT dinucleotides at 2264 and 2265th position (c.2264-2265delAT)(Fig. 4). Though the possibility for the deletion to create the motif for the microRNA target site was searched in this deletion sequence, no candidate motif for the microRNA target site was pointed out [25]. Therefore this deletion seemed a polymorphism in the 3' untranslated region.

Discussion

Myostatin is a negative regulator of muscle growth that is attracting attention as a novel target for increasing muscle growth in cases of DMD [23]. In this study, we conducted extensive sequence analysis of the myostatin genes in 102 Japanese DMD patients. As a result, two novel missense mutations (p.95D>H and p.156L>I) were identified (Figs. 2 and 3). In addition, a known polymorphism (p.164E>K) was identified in four of the DMD patients. Because one of the DMD patients carried p.156L>I and p.164E>K in the identical allele (Fig. 3), the overall mutant allele frequency was 2.5% (5 of 204 alleles). No truncation mutations in the myostatin gene, however, were identified in our study. In particular, the intron mutation that introduces a premature stop codon in myostatin mRNA resulting in marked muscle hypertrophy [19,26] was not observed. Although a single nucleotide change in the 3' untranslated region of the sheep myostatin gene was shown to lead translational inhibition [20], any nucleotide change creating the octamer motif for the micro RNA target site was not disclosed in DMD patients. Exceptionally, one case was disclosed to harbor two nucleotides deletion in the 3' untranslated region (c.2264-2265delAT) in one myostatin gene (Fig. 4). It needs further study to clarify the meaning of this deletion.

Both of the novel mutations (p.95D>H and p.156L>I) were predicted to be pathogenic because they are located at conserved amino-acid residues at the propeptide region of myostatin [24]. The phenotypes of the DMD cases heterozygous for p.95D>H, p.164E>K, or p156L>I and p164E>K, however, were not significantly different from the phenotypes of the other DMD cases. It has been reported that a mother carrying a truncation mutation in intron 2 of the myostatin gene appeared muscular, whereas her son, who was homozygous for the same mutation, showed remarkable muscle hypertrophy [19]. Therefore, the muscle volume and strength of individuals that are heterozygous for myostatin mutations may not be markedly affected by these mutations. Considering that women with one missense polymorphism in the propeptide region of myostatin exhibited increase in muscle volume in response to strength training [18], it is supposed that the case with two amino acid substitutions can be phenotypically modified by providing proper muscle rehabilitation. Future studies will address this supposition.

In order to clarify the roles of the two novel mutations, it may be necessary to identify cases in which the mutations are homozygous. In previous studies, homozygous polymorphisms in the myostatin gene have been reported to cause no clear changes in muscle volume or strength [15,16,27]. In this study, some of the DMD patients had mild phenotypes, such as an ability to walk independently past the age 12 years old (data not shown). Although we hypothesized that in these cases the mild phenotypes were a result of a modifier of the DMD phenotype, these patients did not have mutations in their myostatin genes. Particularly we have reported that aberrant splicing products of the dystrophin gene are a modifier of DMD [8,10].

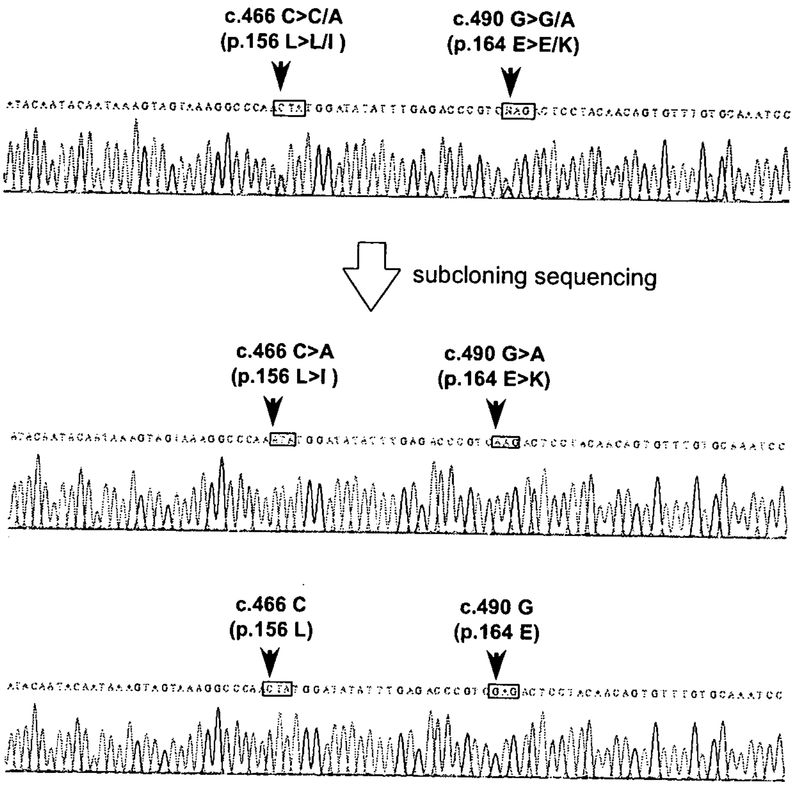

Figure 3
A novel mutation in exon 2 of the myostatin gene. A part of the sequencing results for exon 2 of the myostatin gene of one DMD case (case 549) is shown. Overlapping peaks were observed at two locations corresponding to c.466C>C/A and c.490G>G/A in one DNA sample (Top). Subcloning sequencing disclosed two different sequences: one had a completely normal sequence (Bottom 2), whereas the other one had two nucleotide changes (Bottom 1). The G to A change at the 490th nucleotide of the myostatin cDNA (c. 490G>A) matched with the previously described p.164E>K mutation (box). The other nucleotide change from C to A at the 466th nucleotide of the myostatin cDNA (c.466C>A) changes a CTA codon for Lys to a ATA codon for Ile at the position corresponding to the 156th amino-acid residue of myostatin (p.156L>I) (box).

The variability of the human myostatin gene has been studied in Western countries (Table 1) [15-17,27]. To date, six nucleotide changes (two are common and four are uncommon) have been identified in the myostatin

Table 1: Polymorphisms in the myostatin gene

mutation	Japan (n = 102)	USA Caucasian (n = 167*) (n = 95**)	USA African American (n = 96*) (n = 93**)	Italy (n = 450*) (n = 120**)	Belgium (n = 57)
p.55A>T	0	12 (het)	19 (het) 2(hom)	2 (het)	0
p.153K>R	0	7 (het)	24 (het) 3 (hom)	6 (het) 1 (hom)	1

n number of individuals studied (* p.55A>T, **p.153K>R)
Data are from the USA, Italy, and Belgium [15-17, 27].

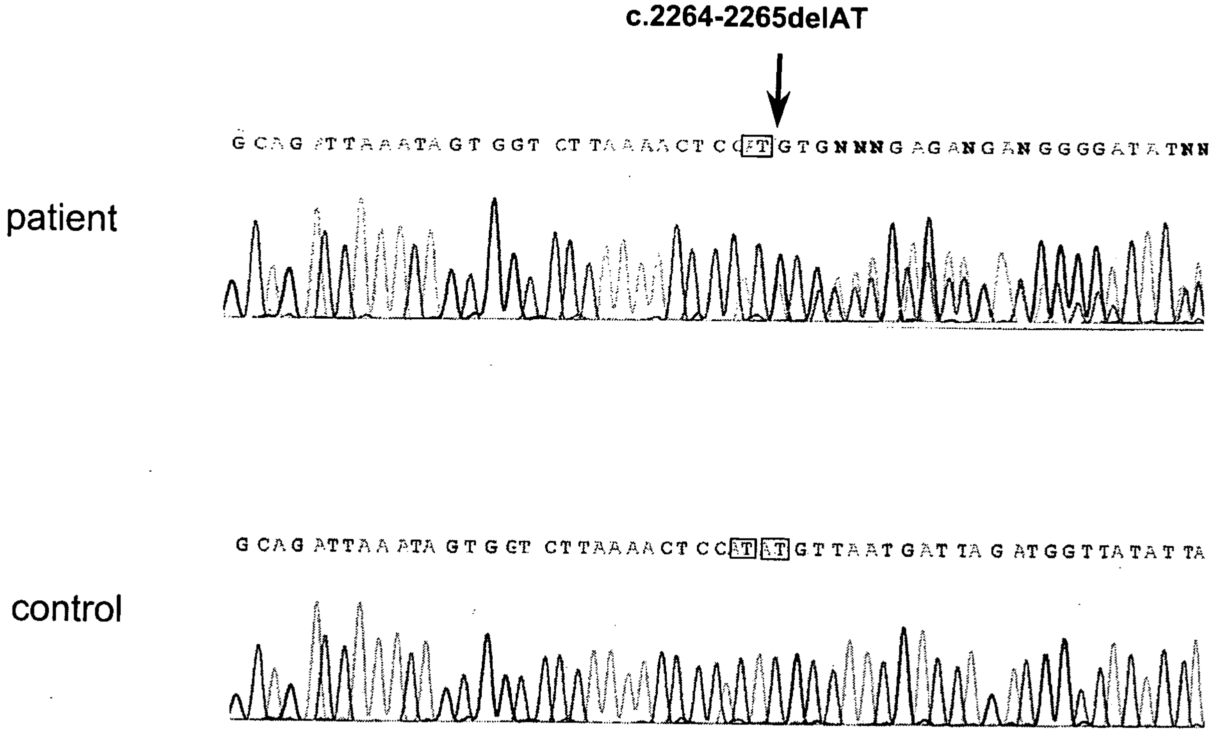

Figure 4

A novel deletion mutation in the 3' untranslated region of the myostatin gene. A part of the sequencing result for the 3' untranslated region of the myostatin gene is shown. In the control all of the sequences matched with that of the wild-type sequence (Genbank: AC073120), including the repeat of AT dinucleotides (boxes) (control). In the sequence of one DMD case (case 100) the repeat of AT dinucleotides was not present, removing AT dinucleotides at 2264 and 2265th nucleotide (c.2264-2265delAT) (patient).

gene (Table 1 and 2). Two polymorphisms, a G to A change at codon 55 in exon 1 (p.55A>T) and an A to G substitution in exon 2 (p.153K>R), were represented in 6.6% and 9.8% of the examined alleles in the USA [15]. In an Italian study [16], the p.55A>T and p.153K>R were identified in 0.2% and 3.3% of the examined alleles, respectively. In a study in Belgium, only one individual from 57 males was found to be heterozygous for p.153K>R [17]. None of the Japanese patients in this study, however, carried p.55A>T or p.153K>R (Table 1). These differences in the incidences of the polymorphisms

Table 2: Rare mutations in the myostatin gene in the Western world and Japan

mutation	Japan (n = 102)	USA (n = 189)	Italy (n = 120)	Belgium (n = 57)
p.95D>H	1	ND	ND	ND
p.156L>I	1	ND	ND	ND
p.164E>K	4	2	0	0
p.185R>T	0	ND	1	ND
p.198P>A	0	1	0	0
p.225I>T	0	2	0	ND
c.747+8G>A	0	1	0	ND
c.2264-2265 del AT	1	ND	ND	ND

ND not determined, *n* number of individuals studied
All numbers show individuals with mutation in heterozygote
Data are from the USA, Italy, and Belgium [15-17, 27].