

ORIGINAL ARTICLE

Four parameters increase the sensitivity and specificity of the exon array analysis and disclose 25 novel aberrantly spliced exons in myotonic dystrophy

This article has been corrected since advance Online Publication, and a corrigendum is also printed in this issue

Yoshihiro Yamashita^{1,6}, Tohru Matsuura^{1,2,6}, Jun Shinmi¹, Yoshinobu Amakusa¹, Akio Masuda¹, Mikako Ito¹, Masanobu Kinoshita³, Hirokazu Furuya⁴, Koji Abe², Tohru Ibi⁵, Ko Sahashi⁵ and Kinji Ohno¹

Myotonic dystrophy type 1 (DM1) is an RNA gain-of-function disorder in which abnormally expanded CTG repeats of *DMPK* sequester a splicing *trans*-factor *MBNL1* and upregulate another splicing *trans*-factor *CUGBP1*. To identify a diverse array of aberrantly spliced genes, we performed the exon array analysis of DM1 muscles. We analyzed 72 exons by RT-PCR and found that 27 were aberrantly spliced, whereas 45 were not. Among these, 25 were novel and especially splicing aberrations of *LDB3* exon 4 and *TTN* exon 45 were unique to DM1. Retrospective analysis revealed that four parameters efficiently detect aberrantly spliced exons: (i) the signal intensity is high; (ii) the ratio of probe sets with reliable signal intensities (that is, detection above background P -value = 0.000) is high within a gene; (iii) the splice index (*SI*) is high; and (iv) *SI* is deviated from *SI*s of the other exons that can be estimated by calculating the deviation value (*DV*). Application of the four parameters gave rise to a sensitivity of 77.8% and a specificity of 95.6% in our data set. We propose that calculation of *DV*, which is unique to our analysis, is of particular importance in analyzing the exon array data.

Journal of Human Genetics (2012) **57**, 368–374; doi:10.1038/jhg.2012.37; published online 19 April 2012

Keywords: CUGBP1; Exon array; *MBNL1*; myotonic dystrophy

INTRODUCTION

Alternative splicing regulates developmental stage-specific and tissue-specific gene expressions and markedly expands the proteome diversity with a limited number of genes. High-throughput sequencing of total mRNAs expressed in cells has revealed that 98% or more of multiexon genes are alternatively spliced,¹ with an average of seven alternative splicing per multiexon gene.² Alternative splicing is achieved by exonic/intronic splicing enhancers/silencers (ESE, ISE, ESS, ISS) in combination with spatial and temporal expression of *trans*-acting splicing factors, such as serine/arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins.^{3,4} Aberrations of alternative splicing are mediated by either mutations disrupting splicing *cis*-elements or dysregulation of splicing *trans*-factors.^{5,6}

Myotonic dystrophy is an autosomal dominant multisystem disorder affecting the skeletal muscles, eye, heart, endocrine system and central nervous system. The clinical symptoms include muscle weakness and wasting, myotonia, cataract, insulin resistance, hypogonadism, cardiac conduction defects, frontal balding and intellectual

disabilities.⁷ Myotonic dystrophy is caused by abnormally expanded CTG repeats in the 3' untranslated region of the *DMPK* gene encoding the dystrophia myotonica protein kinase on chromosome 19q13 (myotonic dystrophy type 1, DM1)^{8–10} or by abnormally expanded CCTG repeats in intron 1 of the *ZNF9* gene encoding the zinc finger protein 9 on chromosome 3q21 (myotonic dystrophy type 2, DM2).¹¹ In DM1, normal individuals have 5–30 repeats; mildly affected individuals have 50–80 repeats; and severely affected individuals have 2000 or more repeats of CTG.^{12,13} In DM2, the size of expanded repeats is extremely variable, ranging from 75 to 11 000 repeats, with a mean of 5000 CCTG repeats.^{11,14}

In DM1 and DM2, expanded CTG or CCTG repeats in the non-coding regions sequester a splicing *trans*-factor muscleblind encoded by *MBNL1* to intranuclear RNA foci harboring mutant RNA.¹⁵ In addition, in DM1 cells, another splicing *trans*-factor CUG-binding protein encoded by *CUGBP1* is hyperphosphorylated by protein kinase C and is stabilized.^{16–18} Dysregulation of the two splicing *trans*-factors then causes aberrant splicing of their target genes. A total

¹Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Nagoya, Japan; ²Department of Neurology, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan; ³Department of Frontier Health Sciences, Graduate School of Human Health Sciences, Tokyo Metropolitan University, Tokyo, Japan; ⁴Neuro-Muscular Center, National Oomuta Hospital, Fukuoka, Japan and ⁵Department of Neurology, Aichi Medical University School of Medicine, Aichi, Japan

⁶These authors contributed equally to this work.

Correspondence: Professor K Ohno, Division of Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya 466-8550, Japan.

E-mail: ohnk@med.nagoya-u.ac.jp

Received 24 November 2011; revised 10 February 2012; accepted 23 March 2012; published online 19 April 2012

Table 1 A total of 28 aberrantly spliced exons and introns identified to date in skeletal and cardiac muscles in myotonic dystrophy

Gene ^a	Affected exon/intron ^b
ATP2A1 (<i>SERCA1</i>) ^{21,31}	Exon 22
ATP2A2 (<i>SERCA2</i>) ³¹	Intron 19
BIN1 ³²	Exon 11
CAPN3 ²¹	Exon 16
CLCN1 ³³	Intron 2
CLCN1 ^{34,35}	Exons 6b/7a
DMD ³⁶	Exon 71
DMD ³⁶	Exon 78
DTNA ³⁷	Exons 11A and 12
FHOD1 (<i>FHOS</i>) ²¹	Exon 11a
FNJ ³⁸	Exon 33
GFPT1 (<i>GFAT1</i>) ²¹	Exon 10
INSR ³⁹	Exon 11
KCNAB1 ⁴⁰	Exons 2b/2c
LDB3 (<i>ZASP</i>) ²¹	Exon 11 (189-nt exon 7)
MBNL1 ²¹	Exon 7 (54-nt exon 6)
MBNL2 ²¹	Exon 7 (54 nt, no exonic annotation)
MEF2C ⁴¹	Exons 4 and 5
MTMR1 ⁴²	Exons 2.1 and 2.2
MYOM1 ⁴³	Exon 17a
MYH14 ⁴⁴	Exon6
NRAP ²¹	Exon 12
PDLIM3 (<i>ALP</i>) ²¹	Exons 5a/5b
RYR1 ³¹	Exon 70
TNNT2 ⁴⁵	Exon 5
TNNT3 ⁴⁶	Fetal exon
TTN ²¹	Exons Zr4 and Zr5 (138-nt exon 11 and 138-nt exon 12)
TTN ²¹	Exon Mex5 (303-nt exon 315)

^aGene symbols by the HUGO Gene Nomenclature Committee. Symbols in parentheses represent alternative symbols that are used in the reference.

^bExon numbers in parentheses represent annotations of the NCBI Build 36.3.

of 28 exons/introns of 22 genes have been identified to date in the skeletal and cardiac muscles in myotonic dystrophy (Table 1).

The Affymetrix GeneChip Human Exon 1.0 ST array contains ~1.4 million probe sets comprised of ~5.4 million probes. The exon array is designed to measure the expression level of each exon and to enable quantitative analysis of alternative splicing. In designing the exon array, probe selection region(s) (PSR) are placed within each exon throughout the genome. Each PSR has a unique probe set ID and carries four probes. A group of PSRs placed on a single exon is given a unique exon cluster ID. In most instances, each exon cluster ID, which represents an individual exon, carries a single PSR. A group of exon clusters spanning a single gene has a unique transcript cluster ID. The exon array thus carries 30–40 probes along the entire length of each gene. Data analysis of the exon array, however, is more complicated than that of the expression array as in the HuEx1.0 ST exon array, (i) each probe set is comprised of only four probes, (ii) each probe does not have a corresponding mismatched probe and (iii) each probe cannot be optimally designed due to a short span of the PSR.

In an effort to elucidate a diverse array of alternatively spliced genes in myotonic dystrophy, we performed the exon array analysis with skeletal muscles of three DM1 patients and three normal controls. In the course of the analysis, we tested alternative splicing of 72 exons by RT-PCR and found that 27 were alternatively spliced, whereas 45 were not. We sought for parameters that best discriminate true and false positives, and found that four parameters discriminate the true and false positives with a sensitivity of 77.8% and a specificity of 95.6%.

MATERIALS AND METHODS

Patient samples

Skeletal muscles were previously biopsied for diagnostic purposes. Clinical features of the patients are summarized in Supplementary Table S1. Two control muscles were biopsied muscle specimens that showed no pathological abnormalities. One control muscle RNA was the Human Skeletal Muscle PCR-Ready cDNA from Life Technologies (Carlsbad, CA, USA). All experiments were performed under the IRB approvals of the Nagoya University Graduate School of Medicine and the Aichi Medical University. The samples were used for the current studies after appropriate informed consents were given. High-molecular weight DNA was extracted by the conventional proteinase K and phenol chloroform method. We determined the CTG repeat numbers at the 3' UTR of the *DMPK* gene by Southern blotting and found that skeletal muscles of patients 1, 2 and white blood cells of patient 3 carried 3430, 4500 and 1500 CTG repeats, respectively. Our analysis underscored a notion that skeletal muscles have larger numbers of repeats compared to leukocytes.¹⁹

RNA preparation and array hybridization

Total RNA was extracted by the RNeasy Mini Kit (Qiagen, Hilden, Germany). We confirmed that the RNA integrity numbers were all above 7.0. Hybridization and signal acquisition of the HuEx1.0 ST exon array (Affymetrix, Santa Clara, CA, USA) were performed according to the manufacturer's instructions. The signal intensities were normalized by the RMA method using the Expression Console 1.1 (Affymetrix).

Exclusion of genes with undependable signals using four criteria

Before we analyzed our exon array data, we excluded genes and probe sets with undependable signals using the following criteria. First, the gene must be comprised of four or more exons. Second, the smaller detection above background (DABG) *P*-value in either controls or DM1 muscles is ≤0.01 for a probe set to be analyzed. DABG is a detection metric generated by comparing perfectly matched probes to a distribution of background probes. Affymetrix expression arrays used a mismatched probe to measure the background signal for a specific probe, whereas Affymetrix exon arrays use shared background probes to estimate the background signals. As exon skipping results in low signals that give rise to high DABG *P*-values, we did not discard probe sets with unreliable signals in either controls or DM1, but not in both. Third, three or more dependable probe sets with DABG *P*-value ≤0.01 should be included in a gene to be analyzed, and such probe sets should comprise 15% or more of all the probe sets on the gene. Fourth, the average signal intensities of either the controls or DM1 should be no less than 150. Among the 336 293 exonic probe sets in our data set, 103 543 probe sets met these criteria.

Unique exon cluster IDs and unique transcript cluster IDs

In order to provide our unique exon cluster IDs and unique transcript cluster IDs, we analyzed annotations of the NCBI human gene database build 36.3 by writing and running Perl programs on the PrimePower HPC2500/Solaris 9 supercomputer (Fujitsu Ltd, Tokyo, Japan). We analyzed the exon array signals on Microsoft Excel by making VBA programs. Partitioning of parameters to distinguish true and false positives was performed by the JMP statistical software Ver. 8.0.1 (SAS Institute, Cary, NC, USA) with its default settings.

RESULTS

We provided our unique transcript cluster IDs and exon cluster IDs for the exon array based on the NCBI RefSeq database

We analyzed muscle specimens of three DM1 and three controls using the HuEx1.0 ST exon array. In the course of the analysis, we noticed that the exon array annotations provided by the manufacturer are based on comprehensive collation of several different gene databases and do not match to any single annotation database. We thus exploited the NCBI RefSeq annotation and provided our unique exon cluster IDs (Supplementary data). We also provided our unique transcript cluster IDs because in the manufacturer's annotations some exon clusters either upstream or downstream of the RefSeq-defined

gene region are given the same transcript cluster ID as intragenic exon clusters. An example of the *MBNL1* annotations is shown in Figure 1. Our data set was comprised of 336 293 probe sets that were grouped into 218 622 exonic clusters on 27 208 transcript clusters. For 1766 probe sets, we assigned duplicated exonic and transcript clusters, as two genes shared the same exonic regions. Our data set thus utilized 23.5% (336 293/1 432 144) of probe sets placed on the array.

The Affymetrix HuEx-1.0-st-v2 annotations release 32 carries 284 805 'core' probe sets. The 'core' probe sets represent the RefSeq transcripts and the full-length GenBank mRNAs. The 'core' probe sets are grouped into 192 554 exonic clusters on 19 231 transcript clusters. When the Affymetrix 'core' annotations are compared with the NCBI Build 36.3 database that we utilized, 17 372 (6.1%) of the 284 805 'core' probe sets are on non-exonic regions according to NCBI. Conversely, 69 278 (6.0%) of the 1 147 338 'non-core' probe sets are on exonic regions according to NCBI.

Deviation values (DVs) of splice indices (SIs) provide essential information to distinguish true and false positives

According to the manufacturer's suggestions, we first calculated the normalized intensity (*NI*) of each exon cluster by dividing the signal

intensity of a given exon cluster by a sum of all the exonic signals throughout the gene. We next calculated the *SI* by dividing NI_{DM1} by $NI_{control}$.²⁰ We also calculated *t*-test *P*-values of *SIs* between three controls and three patients.

$$NI(\text{exoncluster}_i) = \frac{\text{signal intensity of exon cluster}_i}{\text{expression level of a gene}}$$

$$SI(\text{exoncluster}_i) = NI(\text{exon cluster}_i)_{DM1}/NI(\text{exon cluster}_i)_{control}$$

Validation by RT-PCR of ~20 exons, however, revealed that only about a quarter of candidate exons were aberrantly spliced, whereas three quarters were not. In the course of analysis, we noticed that, in most cases, the *SIs* of the truly positive gene were all close to 1.0 throughout the gene, whereas those of the falsely positive gene were variable from probe set to probe set (Figure 2). In order to quantify how much the normalized *SI* of a particular exon is deviated from those of the other exons, we calculated the mean and standard deviation (*s.d.*) of *SIs* of the other exons. We then calculated the *DV* of the *SI* of an exon cluster of our interest.

$$DV(\text{exon cluster}_i) = [SI(\text{exon cluster}_i) - \text{mean}_{SI}]/SD_{SI}$$

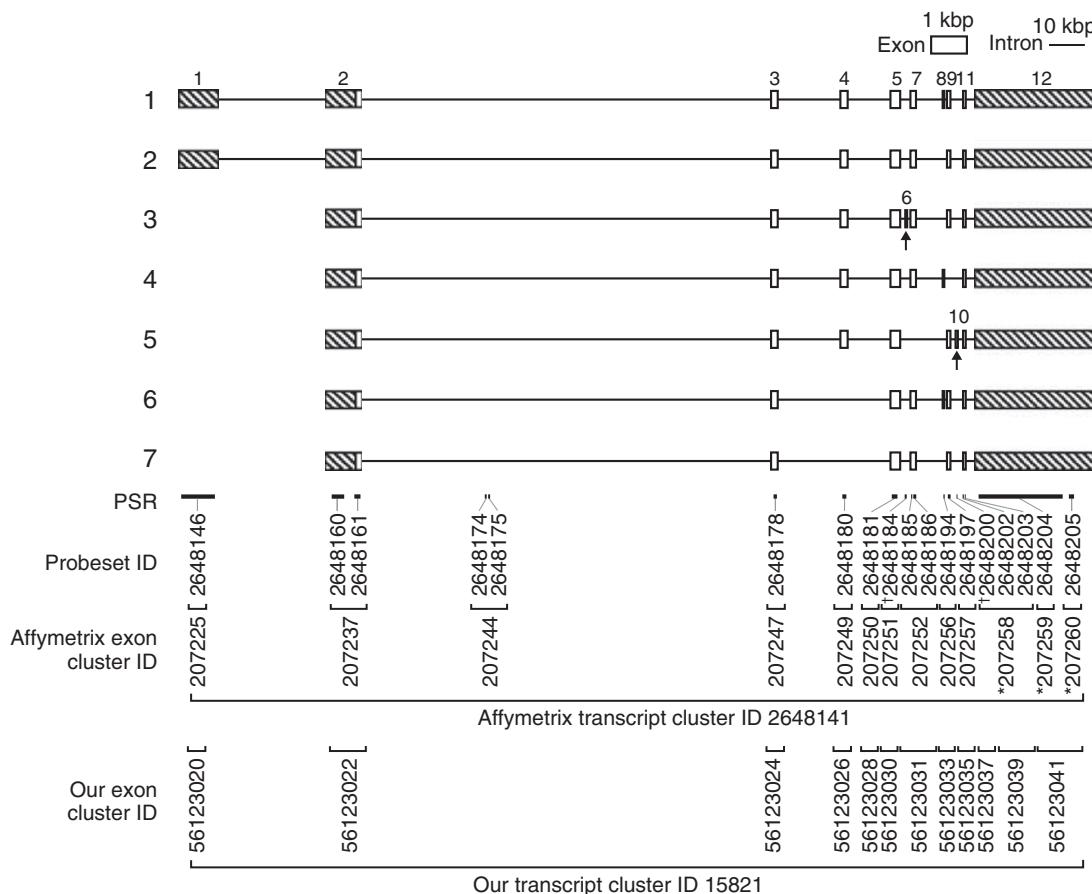


Figure 1 Comparison of Affymetrix annotations and our annotations. Seven alternative transcripts of *MBNL1* according to the NCBI Build 36.3 are drawn to individual exonic and intronic scales indicated at the top. The PSR bars represent the 'core' probe sets by Affymetrix. Each probe set is comprised of four probes (not shown). Each exon cluster corresponds to a single exon and carries one or more probe sets. A transcript cluster is comprised of exon clusters on the gene. Note that three 3' exon clusters (asterisks) by Affymetrix are discordant to the genomic structure by NCBI: the Affymetrix exon cluster ID 207258 corresponds to exons 10 and 11 by NCBI, and we provided two different exon cluster IDs of 56123037 and 56123039. The exon cluster IDs 207259 and 207260 correspond to exon 12 by NCBI, and provided a single exon cluster ID of 56123041. Our unique exon cluster IDs and transcript cluster IDs throughout the genome are in the Supplementary data. Arrows indicate aberrant exons 6 and 10 identified in the current studies, and daggers indicate the corresponding probe sets. Shaded areas represent non-coding regions.

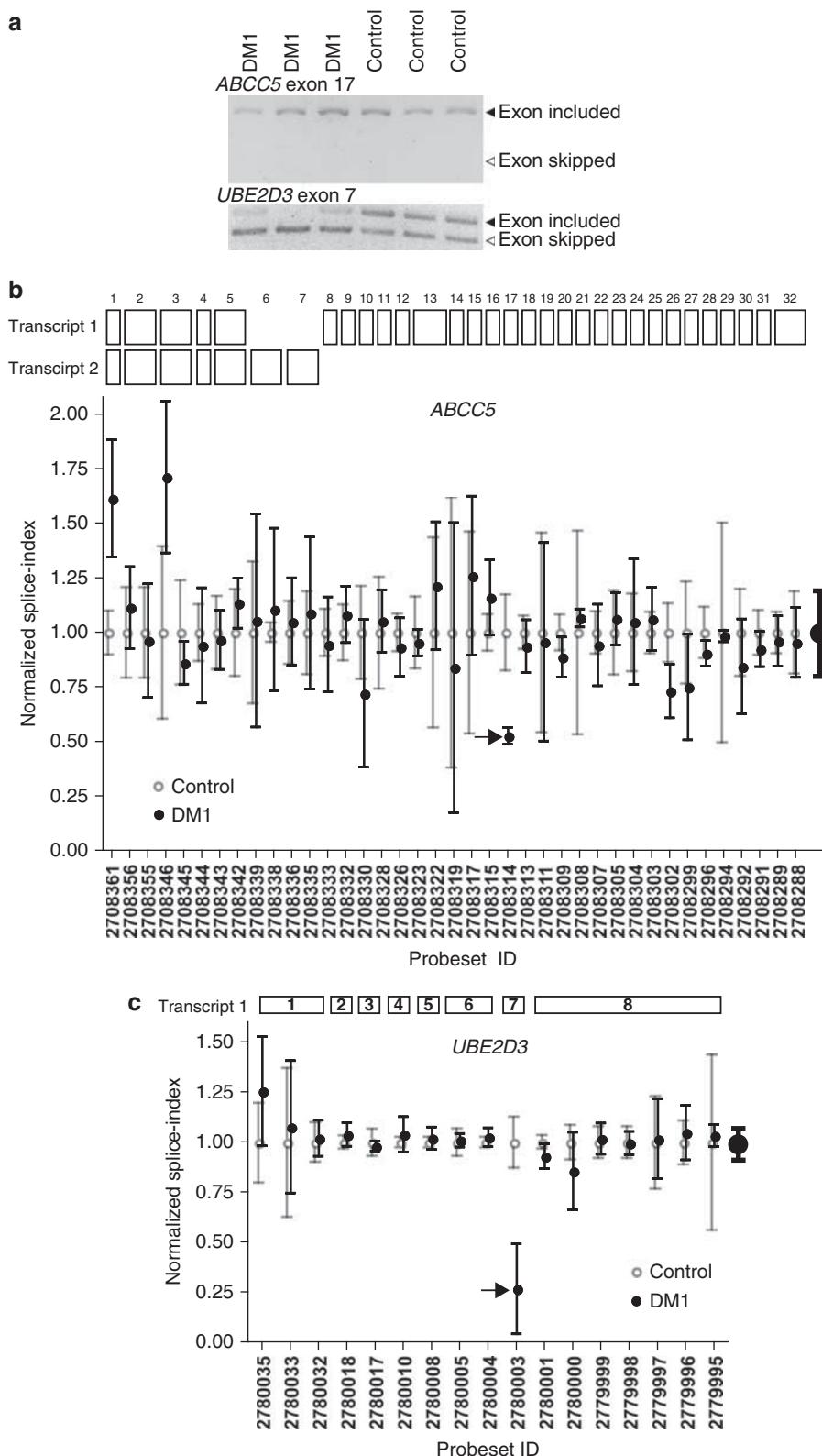


Figure 2 False predicted skipping of *ABCC5* exon 17, and correctly predicted skipping of *UBE2D3* exon 7. (a) RT-PCR analysis of three DM1 muscles and three control muscles. (b) Splice index (an exonic signals divided by a sum of all the exonic signals throughout a gene) of each probe set on *ABCC5* in DM1 is normalized for that in controls. The normalized splice indices in DM1 are variable from probe set to probe set, giving rise to the mean and s.d. of 1.00 ± 0.20 (bold symbols). An arrow points to exon 17. The deviation value of probe set ID 2708314 on exon 17 is -3.7 s.d. (c) Normalized splice indices of *UBE2D3* in DM1 are less variable compared with those of *ABCC5*, which gives rise to the mean and s.d. of 1.00 ± 0.08 (bold symbols). An arrow points to exon 7. The deviation value of probe set ID 2780003 on exon 7 is -9.7 s.d.

We indeed found that *SI*s of the truly positive genes were all close to 1.0 and alternative exons tend to give rise to high *DV* values as explained in Figure 4a and the relevant statements below.

We analyzed 72 exons and identified 27 aberrant exons in DM1

In order to seek for aberrantly spliced exons/introns in DM1, we arbitrarily set three thresholds of *DV* > 3.0, *SI* > 1.5 and *t*-test *P*-value < 0.1. The three criteria were satisfied in 256 exons. Among these, we arbitrarily chose 72 exons. As we started our analysis without knowing which parameters were efficiently able to predict true positives, there were no strict objective criteria how we chose these exons. We, however, looked into the following features when we chose candidate exons: (i) a conspicuous value in one or more of the three parameters; (ii) alternative spliced exons annotated in the NCBI database; or (iii) a possible pathogenic gene that can be causally associated with DM1. We also avoided previously known aberrant splicing except for *LDB3* exon 7 and *MBNL1* exon 6, as the two exons were of special interest to us. RT-PCR analysis of the 72 exons revealed that 27 exons were indeed aberrantly spliced (Supplementary Figure S1; Table 2), whereas 45 exons were not (Supplementary Table S2), which gave rise to a positive predictive value of 27/72 = 37.5% (Figure 3b). Most aberrant fragments were observed in normal and disease controls to variable extents, and we defined ‘aberrant’ splicing when the ratios of aberrant fragments in DM1 were more than those

Table 2 A total of 27 aberrantly spliced exons in DM1 identified in the current studies

Gene	Exon		<i>Deviation value</i> (<i>DV</i>)	<i>Splice index</i> (<i>SI</i>)	<i>t</i> -test <i>P</i> -value
	(size in bp)	Probe set ID			
<i>AKAP13</i>	16 (62) ^b	3606402	3.8	2.4 ^e	0.0132
<i>ATP5G2</i>	1 (273) ^b	3456337	108.2	2.6	0.0065
<i>FGD6</i>	2 (2425) ^c	3466416	16.5	2.0	0.0144
<i>ILF3</i>	18 (1352) ^b	3820705	16.1	2.5	0.0048
<i>LDB3</i>	4 (368) ^c	3255989	36.0	6.9	0.0889
<i>LDB3</i>	7 (189) ^{a,c}	3256033	34.4	6.6	0.0000
<i>MAP4K4</i>	17 (231) ^b	2496832	16.5	3.1	0.0063
<i>MBNL1</i>	6 (54) ^{a,b}	2648184	16.5	3.1	0.0063
<i>MBNL1</i>	10 (64) ^b	2648200	5.4	4.0 ^e	0.0041
<i>MBNL2</i>	8 (95) ^b	3497646	10.2	2.7	0.0008
<i>MSI2</i>	14 (73)	3728314	24.3	2.5	0.0550
<i>MXRA7</i>	4 (81) ^b	3771753	13.6	3.5	0.0464
<i>MYBPC1</i>	23 (54) ^b	3428645	46.9	4.8	0.0060
<i>MYBPC1</i>	31 (59) ^b	3428661	10.4	1.8	0.0808
<i>NCOR2</i>	10 (225) ^c	3476468	7.2	1.7	0.0163
<i>NDUFV3</i>	3 (1095) ^b	3922938	6.2	3.4 ^e	0.0034
<i>NEB</i>	116 (105) ^c	2581073	32.7	2.8	0.0004
<i>NEDD4L</i>	13 (132) ^c	3790056	3.5	3.0 ^e	0.0267
<i>NEDD4L</i>	14 (120) ^c	3790058	3.6	3.2 ^e	0.0184
<i>NEXN</i>	2 (42) ^c	2343241	12.9	1.9	0.0169
<i>NFIX</i>	7 (123) ^c	3822162	18.3	2.3	0.0087
<i>NR4A1</i>	4 (1642) ^b	3415256	12.1	2.9 ^e	0.0012
<i>PPHLN1</i>	7 (57) ^d	3412039	8.1	3.0 ^e	0.0001
<i>SOS1</i>	21 (45) ^c	2549106	5.0	2.0 ^e	0.0334
<i>TBC1D15</i>	7 (51) ^c	3422345	4.9	3.8 ^e	0.0149
<i>TTN</i>	45 (375) ^b	2589787	62.7	4.9	0.0484
<i>UBE2D3</i>	11 (50) ^b	2780003	8.9	3.3 ^e	0.0087

Exon numbers are according to the NCBI Build 36.3.

^aAberrant splicing in DM1 has been previously reported.

^bExons are known to be alternatively spliced according to (i) NCBI and ENSEMBL release 50.

^c(ii) ENSEMBL release 50.

^dor (iii) dbEST alone.

^eSignals in DM1 are weaker than those in controls, and the *SI* values are inversed.

in any normal controls. In 11 of the 27 exons, intensities of aberrant fragments in DM1 exceeded those of normal controls but not all of disease controls, indicating that the splicing aberrations are not specific to DM1. In the remaining 16 ‘specific’ exons (asterisks in Supplementary Figure S1), the ratios of aberrant fragments in DM1 were more than those in any disease controls. Especially, splicing aberrations in *LDB3* exon 4 and *TTN* exon 45 were almost exclusively observed in DM1, and were ‘unique’ to DM1. To summarize, among the 27 aberrantly spliced exons that we identified in the current studies, 25 were novel, 16 were ‘specific’ to DM1 and 2 were ‘unique’ to DM1 (*LDB3* exon 4 and *TTN* exon 45).

In an effort to understand the rarity of ‘specific’ and ‘unique’ aberrant splicing, we examined the disease specificity of four previously reported aberrant splicing in DM1.²¹ Limited availability of biopsied muscles hindered us from analyzing all the 28 previously reported splicing aberrations shown in Table 1. We found that aberrant splicing of *PDLIM3* exon 5 was ‘specific’; that of *CAPN3* exon 16 was ‘unique’; and those of *GFPT1* exon 10 and *NRAP* exon 12 were observed in normal and disease controls (Supplementary Figure S2). Thus, some of the previously reported splicing aberrations in DM1 (Table 1) are likely to represent muscle degeneration and/or regeneration.²²

Four parameters increased the sensitivity of exon array analysis

Using the 72 analyzed exons, we next asked which parameters were able to discriminate the true and false positives. We analyzed 10 parameters and found that *DV* were most discriminative and *SI* followed (Figure 4). Additionally, we observed significant differences in average signal intensities and in ratios of probe sets with DABG *P*-value = 0.000. A DABG *P*-value is attached to a signal intensity of each probe set and represents reliability of the signal intensity. The *t*-test *P*-values were lower in true positives, but without statistical significance.

Comparison of the true (Table 2) and false (Supplementary Table S2) positives using the recursive partitioning functionality of the JMP 8.0.1 statistical software indicated four thresholds. The partitioning functionality of JMP seeks for the best splitting point of the best factor, X_i , among a group of factors that best discriminate the response Y . The 10 factors indicated Figure 4a were analyzed to discriminate the true and false positives. First, the signal intensities of either controls or DM1 should be more than 270. Second, the ratio of probe sets with DABG *P*-value = 0.000 either in controls or DM1 should be more than 0.05. Third, *DV* should be more than 10.0 or *SI* should be more than 2.6. Application of the four thresholds excluded 6 out of 27 true positives and 43 out of 45 false positives, and gave rise to a sensitivity of 21/27 = 77.8% and a specificity of 43/45 = 95.6% (Figure 3b). If we exclude the threshold for the *DV* and include the threshold for *t*-test *P*-value of less than 0.05, which are commonly used in the analysis of exon arrays, the sensitivity becomes as low as 15/27 = 55.6%, whereas the specificity rather becomes 45/45 = 100%. Thus, the inclusion of *DVs* in the analysis increases a chance of identifying aberrantly spliced exons by 22.2%, although a chance of detecting false positives is rather increased by 4.4%.

Although the significance of four parameters is demonstrated in our data set (Figure 4a), the thresholds should be unique to our data set and different thresholds need to be applied to different data sets. To prove this, we analyzed four human exon arrays of GSE21795,²³ GSE28672,²⁴ GSE24581²⁵ and GSE21840²⁶ in the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>). Each data set was comprised of a pair of three to five samples, and aberrant and alternative splicing events of a total of 23 exons were validated by RT-PCR in the original papers. Although *SI*s and *DVs* of the 23

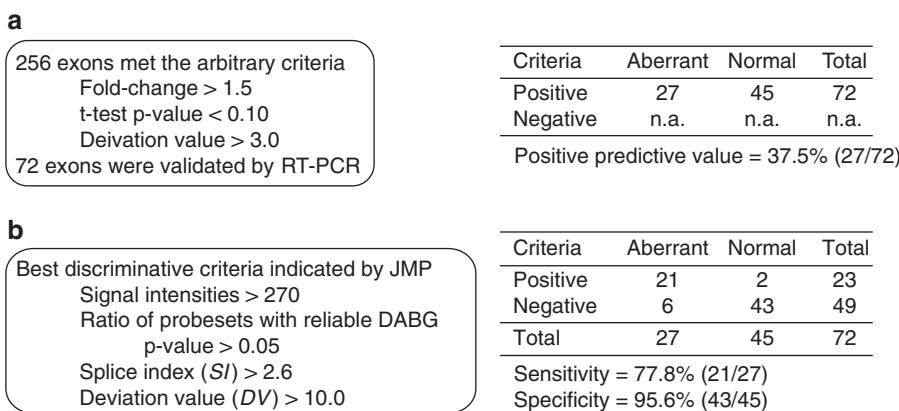


Figure 3 (a) Arbitrary criteria to search for aberrant splicing (left panel) and their results (right panel). (b) Four discriminative criteria indicated by JMP-IN (left panel) and their results (right panel).

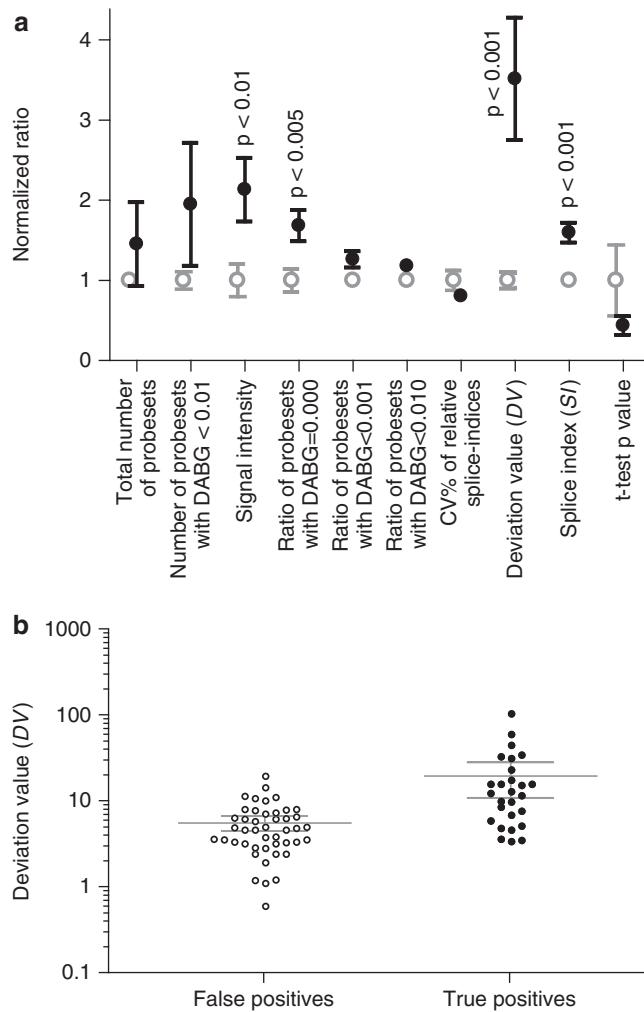


Figure 4 (a) Parameters that differentiate the true and false positives. Values are normalized to those of false positives, and the mean and s.e. are indicated. The true and false positives are indicated by solid and gray symbols, respectively. For SI, an inverse of the SI value is taken when the signal is decreased in DM1. Four parameters exhibit statistical significance with the Student's *t*-test. (b) Deviation values of 27 true and 45 false positives are plotted on a logarithmic scale. Gray lines indicate means and 95% confidence intervals.

exons were as high as 2.58 ± 1.05 and 8.3 ± 12.8 (mean and s.d.), respectively, 9 exons did not meet the criteria indicated in Figure 3b, which gave rise to a sensitivity of $14/23 = 60.9\%$. As no false positive results were documented in these papers, we could not calculate the specificity, and we could not estimate if calculation of four parameters indeed increases the sensitivity and specificity of the exon array analysis for these data sets.

DISCUSSION

A total of 25 novel aberrantly spliced exons in DM1

We identified 25 novel aberrantly spliced exons in DM1. Among these, aberrant splicing events of *LDB3* exon 4 and *TTN* exon 45 are 'unique' to DM1. Aberrations of the other 23 exons are observed in other muscle diseases with variable degrees. MBNL1 normally translocates from cytoplasm to nucleus in the postnatal period to induce adult-type splicing, and lack of muscleblind due to sequestration to RNA foci in myotonic dystrophy recapitulates fetal splicing patterns.^{21,27} Downregulation of MBNL1 and upregulation of CUGBP1 is likely to occur in rejuvenating muscle fibers, and is likely to result in altered splicing patterns that we observe in disease controls. Not all aberrantly spliced exons in DM1, however, are observed in disease controls. Pathological significance of aberrant splicing in disease controls thus remains to be elucidated.

Exon array analysis

In expression arrays, fold-changes and *t*-test *P*-values have been successfully employed to detect altered gene expressions. On the other hand, these parameters are not sufficient to detect aberrant splicing in the exon array data. We thus sought for additional parameters and found that four parameters are informative to discriminate true and false positives: (i) the DV, (ii) the normalized splice index, (iii) the signal intensity and (iv) the ratio of probe sets with DABG *P*-value = 0.000. Application of these four parameters has enabled us to achieve a sensitivity of 77.8% and a specificity of 95.6%. On the other hand, the *t*-test *P*-values are not significantly lower in true positives. This represents that the threshold of *t*-test *P*-value ≤ 0.10 is likely to be sufficient to exclude a large amount of false positives and that further stringent *P*-values would not help discriminate true and false positives.

In addition, our unique annotations of exon cluster IDs and transcript cluster IDs also make the DVs more dependable. This is because probe sets on rare transcripts or probe sets outside of the NCBI-defined gene region sometimes give rise to falsely strong signals

with dependable DABG *P*-values. Inclusion of these probe sets increases the standard deviation and decreases the DV of the aberrantly spliced exon.

Potential roles of novel aberrant splicing events in DM1

In this study, we identified 27 DM1-specific aberrant splicing, in which 25 have not been published yet. Among the 25 exons, aberrant splicing events of two exons were 'uniquely' observed in DM1: one is inclusion of the LIM domain binding 3 (*LDB3*) exon 4 and the other is inclusion of titin (*TTN*) exon 45. Interestingly, both encode structural proteins of muscle fiber.

LDB3, also known as Cypher/ZASP (Z-band alternatively spliced PDZ-motif protein), contains a PDZ domain at the N-terminus and one or three LIM domains at the C-terminus. *LDB3* is localized to the Z-line and interacts with α -actinin 2 through its PDZ-domain and with protein kinase C via its C-terminal LIM domains.²⁸ *LDB3* is likely to have an essential role in supporting Z-line structure and muscle function during contraction.²⁹ *LDB3* has several isoforms. As inclusion of exon 4 is preferentially observed in the fetal heart,³⁰ the aberrant inclusion of exon 4 in the skeletal muscles in DM1 would lead to dysfunction or morphological abnormalities of muscle fiber. Recently, phosphoglucomutase 1 (PGM1), an enzyme involved in glycolysis and gluconeogenesis, has been known to bind to the domain encoded by exon 4 of *LDB3*. *LDB3* mutations in exon 4 reduce the binding to PGM1 and develop dilated cardiomyopathy.³¹ On the other hand, the increased binding of PGM1 and *LDB3* through aberrant inclusion of exon 4 might be involved in the pathogenesis of muscle atrophy, weakness and histological abnormalities in DM1.

TTN encodes the largest protein in mammals and the third most abundant protein in muscle.³² An N-terminal Z-disc region and a C-terminal M-line region bind to the Z-line and M-line of the sarcomere, respectively, so that a single molecule extends half the length of a sarcomere. Titin is critically important for myofibril elasticity and structural integrity. Its elasticity lies specifically in the I-band region and contains two elements in series with different properties: the tandem immunoglobulin (Ig) and PEVK domains.³³ Different *TTN* isoforms contribute to differences in elasticity of different muscle types.³⁴ As exon 45 is located at the tandem Ig domains, aberrant inclusion of exon 45 in DM1 might lead to defective myofibril assembly and function.

ACKNOWLEDGEMENTS

This work was supported by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology as well as the Ministry of Health, Labor and Welfare of Japan.

- 1 Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
- 2 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
- 3 Smith, C. W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388 (2000).
- 4 Caceres, J. F. & Kornblith, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186–193 (2002).
- 5 O'Rourke, J. R. & Swanson, M. S. Mechanisms of RNA-mediated disease. *J. Biol. Chem.* **284**, 7419–7423 (2009).
- 6 Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).
- 7 Harper, P. S. & Monckton, D. G. in *Myology* 3rd edn (ed. Engel, A. G.) Vol. 2, 1039–1076 (McGraw-Hill, New York, 2004).
- 8 Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsifidis, C. et al. Cloning of the essential myotonic dystrophy region and mapping of the putative defect. *Nature* **355**, 548–551 (1992).
- 9 Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H. et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**, 799–808 (1992).
- 10 Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C. et al. Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355**, 547–548 (1992).
- 11 Liquori, C. L., Ricker, K., Moseley, M. L., Jacobsen, J. F., Kress, W., Naylor, S. L. et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* **293**, 864–867 (2001).
- 12 Gharehbaghi-Schnell, E. B., Finsterer, J., Korschineck, I., Mamoli, B. & Binder, B. R. Genotype-phenotype correlation in myotonic dystrophy. *Clin. Genet.* **53**, 20–26 (1998).
- 13 Consortium T. I. M. D. New nomenclature and DNA testing guidelines for myotonic dystrophy type 1 (DM1). *Neurology* **54**, 1218–1221 (2000).
- 14 Saito, T., Amakusa, Y., Kimura, T., Yahara, O., Aizawa, H., Ikeda, Y. et al. Myotonic dystrophy type 2 in Japan: ancestral origin distinct from Caucasian families. *Neurogenetics* **9**, 61–63 (2008).
- 15 Osborne, R. J., Lin, X., Welle, S., Sobczak, K., O'Rourke, J. R., Swanson, M. S. et al. Transcriptional and post-transcriptional impact of toxic RNA in myotonic dystrophy. *Hum. Mol. Genet.* **18**, 1471–1481 (2009).
- 16 Wang, G. S., Kearney, D. L., De Biasi, M., Taffet, G. & Cooper, T. A. Elevation of RNA-binding protein CUGBP1 is an early event in an inducible heart-specific mouse model of myotonic dystrophy. *J. Clin. Invest.* **117**, 2802–2811 (2007).
- 17 Kuyumcu-Martinez, N. M., Wang, G. S. & Cooper, T. A. Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. *Mol. Cell* **28**, 68–78 (2007).
- 18 Nezu, Y., Kino, Y., Sasagawa, N., Nishino, I. & Ishiura, S. Expression of MBNL and CELF mRNA transcripts in muscles with myotonic dystrophy. *Neuromuscul. Disord.* **17**, 306–312 (2007).
- 19 Thornton, C. A., Johnson, K. & Moxley, 3rd R. T. Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann. Neurol.* **35**, 104–107 (1994).
- 20 Srinivasan, K., Shiue, L., Hayes, J. D., Centers, R., Fitzwater, S., Loewen, R. et al. Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods* **37**, 345–359 (2005).
- 21 Lin, X., Miller, J. W., Mankodi, A., Kanadia, R. N., Yuan, Y., Moxley, R. T. et al. Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Hum. Mol. Genet.* **15**, 2087–2097 (2006).
- 22 Orengo, J. P., Ward, A. J. & Cooper, T. A. Alternative splicing dysregulation secondary to skeletal muscle regeneration. *Ann. Neurol.* **69**, 681–690 (2011).
- 23 Figuer, C., Klein, A. F., Hammer, C., Vassilopoulos, S., Ivarsson, Y., Toussaint, A. et al. Misregulated alternative splicing of BIN1 is associated with T tubule alterations and muscle weakness in myotonic dystrophy. *Nat. Med.* **17**, 720–725 (2011).
- 24 Sharma, A., Markey, M., Torres-Munoz, K., Varia, S., Kadakia, M., Bubulya, A. et al. Son maintains accurate splicing for a subset of human pre-mRNAs. *J. Cell Sci.* **124**, 4286–4298 (2011).
- 25 Chang, J. G., Yang, D. M., Chang, W. H., Chow, L. P., Chan, W. L., Lin, H. H. et al. Small molecule amiloride modulates oncogenic RNA alternative splicing to devalue human cancer cells. *PLoS One* **6**, e18643 (2011).
- 26 Dutertre, M., Sanchez, G., De Cian, M. C., Barbier, J., Dardenne, E., Gratadou, L. et al. Cotranscriptional exon skipping in the genotoxic stress response. *Nat. Struct. Mol. Biol.* **17**, 1358–1366 (2010).
- 27 Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B. et al. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc. Natl. Acad. Sci. USA* **105**, 20333–20338 (2008).
- 28 Zhou, Q., Ruiz-Lozano, P., Martone, M. E. & Chen, J. Cypher, a striated muscle-restricted PDZ and LIM domain-containing protein, binds to alpha-actinin-2 and protein kinase C. *J. Biol. Chem.* **274**, 19807–19813 (1999).
- 29 Zhou, Q., Chu, P. H., Huang, C., Cheng, C. F., Martone, M. E., Knoll, G. et al. Ablation of Cypher, a PDZ-LIM domain Z-line protein, causes a severe form of congenital myopathy. *J. Cell. Biol.* **155**, 605–612 (2001).
- 30 Huang, C., Zhou, Q., Liang, P., Hollander, M. S., Sheikh, F., Li, X. et al. Characterization and *in vivo* functional analysis of splice variants of cypher. *J. Biol. Chem.* **278**, 7360–7365 (2003).
- 31 Arimura, T., Inagaki, N., Hayashi, T., Shichi, D., Sato, A., Hinohara, K. et al. Impaired binding of ZASP/Cypher with phosphoglucomutase 1 is associated with dilated cardiomyopathy. *Cardiovasc. Res.* **83**, 80–88 (2009).
- 32 Gregorio, C. C., Granzier, H., Sorimachi, H. & Labeit, S. Muscle assembly: a titanic achievement? *Curr. Opin. Cell Biol.* **11**, 18–25 (1999).
- 33 Horowitz, R. The physiological role of titin in striated muscle. *Rev. Physiol. Biochem. Pharmacol.* **138**, 57–96 (1999).
- 34 Freiburg, A., Trombitas, K., Hell, W., Cazorla, O., Fougerousse, F., Centner, T. et al. Series of exon-skipping events in the elastic spring region of titin as the structural basis for myofibrillar elastic diversity. *Circ. Res.* **86**, 1114–1121 (2000).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)