37. Gerke, V., and Moss, S. E. (2002) Annexins: from structure to function. *Physiol. Rev.* **82**, 331–371

38. Comera, C., and Russo-Marie, F. (1995) Glucocorticoid-induced annexin 1 secretion by monocytes and peritoneal leukocytes. *Br. J. Pharmacol.* **115**, 1043–1047

39. Cirino, G., Peers, S. H., Flower, R. J., Browning, J. L., and Pepinsky, R. B. (1989) Human recombinant lipocortin 1 has acute local anti-inflammatory properties in the rat paw edema test. *Proc. Natl. Acad. Sci. USA* **86**, 3428–3432

40. Yang, Y., Leech, M., Hutchinson, P., Holdsworth, S. R., and Morand, E. F. (1997) Antiinflammatory effect of lipocortin 1 in experimental arthritis. *Inflammation* **21**, 583–596

41. Dubois, T., Bisagni-Faure, A., Coste, J., Mavoungou, E., Menkes, C. J., Russo-Marie, F., and Rothhut, B. (1995) High levels of antibodies to annexins V and VI in patients with rheumatoid arthritis. *J. Rheumatol.* **22**, 1230–1234

42. Rodriguez-Garcia, M. I., Fernandez, J. A., Rodriguez, A., Fernandez, M. P., Gutierrez, C., and Torre-Alonso, J. C. (1996) Annexin V autoantibodies in rheumatoid arthritis. *Ann. Rheum. Dis.* **55**, 895–900

43. Kaufman, M., Leto, T., and Levy, R. (1996) Translocation of annexin I to plasma membranes and phagosomes in human neutrophils upon stimulation with opsonized zymosan: possible role in phagosome function. *Biochem. J.* **316**, 35–42

44. Gregory, C. Y., and Hall, M. O. (1992) The phagocytosis of ROS by RPE cells is inhibited by an antiserum to rat RPE cell plasma membranes. *Exp. Eye Res.* **54**, 843–851

45. Wistow, G., and Kim, H. (1991) Lens protein expression in mammals: taxon-specificity and the recruitment of crystallins. *J. Mol. Evol.* **32**, 262–269

46. Kim, R. Y., Gasser, R., and Wistow, G. J. (1992) mu-crystallin is a mammalian homologue of Agrobacterium ornithine cyclodeaminase and is expressed in human retina. *Proc. Natl. Acad. Sci. USA* **89**, 9292–9296

Received February 27, 2005; accepted June 21, 2005.

Table 1

Antibodies used in immunohistochemical studies and conditions of antigen retrieval treatments

Antigen	Retrieval Treatment	Primary Antibody		
		Host	Dilution	Supplier
Amyloid P Component	Pro K	Rabbit	200	Dako, Carpenteria, CA
Apolipoprotein E	-	Mouse	200	Biogenesis, Poole, UK
C5	Pro K	Rabbit	200	Dako, Carpenteria, CA
C5b-9	Pro K	Mouse	50	Dako, Carpenteria, CA
MCP	Autoclave	Rabbit	50	Santa Cruz, Santa Cruz, CA
Vitronectin	-	Mouse	100	Chemicon, Temecula, CA

Table 2

Macular status of aged monkeys

Grade	Examined Year			Total	Percentage
	2001	2003	2004		
Normal	**45**	**98**	**45**	**188**	**67.6%**
Mild	4	11	15	30	10.8%
Moderate	5	16	10	31	11.2%
Severe	6	17	6	29	10.4%
Total	60	142	76	278	100.0%

Two-hundred and seventy-eight aged female monkeys were examined by fundus scope and classified into 4 grades. Normal: macula with no detectable pigmentary abnormalities. Mild: fewer than 5 yellowish-white spots. Moderate: 5 to 20 spots. Severe: more than 20 spots.

Table 3

Protein components in monkey drusen

Protein	Accession No.	Protein	Accession No.
Actin, α 2	gi|4501883|	**Hemoglobin, β**	gi|4504349|
Albumin	gi|4502027|	*Hemoglobin, delta*	gi|70353|
Aldehyde dehydrogenase 3	gi|283971|	*Histone, H2A C*	gi|4504239|
Aldehyde dehydrogenase 5	gi|105247|	*Histone, H2A Z*	gi|4504255|
Aldolase A	gi|4557305|	*Histone, H2B F*	gi|10800140|
Alpha-1-antitrypsin	gi|1703025|	*Ig, α 2C*	gi|113585|
Alpha-1B-Glycoprotein	gi|46577680|	**Ig, gamma 2C**	gi|121043|
Annexin V	gi|4502107|	**Ig, lambda**	gi|87890|
Apolipoprotein E	gi|4557325|	Lactate dehydrogenase A	gi|5031857|
ATP synthase α chain, mitochondrial	gi|4757810|	Malate dehydrogenase 1	gi|5174539|
Calmodulin 2	gi|4502549|	Peptidylprolyl isomerase A isoform 1	gi|10863927|
Calreticulin	gi|4757900|	Phosphoglycerate kinase 1	gi|4505763|
cAMP-dependent protein kinase inhibitor, β	gi|14210480|	Phosphoinositide 3-kinase, T96	gi|7434348|
Cell adhesion protein SQM1	gi|105595|	Plectin 1	gi|14195007|
Ceruloplasmin	gi|4557485|	Prostatic binding protein	gi|4505621|
Clusterin	gi|4502905|	Protease inhibitor 4	gi|21361302|
Collagen, α 1(VII)	gi|627406|	Pyruvate dehydrogenase	gi|4885543|
Complement component 5	gi|4502507|	Pyruvate kinase, M1 isozyme	gi|20178296|
Complement component 9	gi|4502511|	Ran binding Protein 2	gi|1709217|
Creatine kinase B	gi|125294|	Recoverin	gi|4506459
Crystallin, β B2	gi|299263|	Retinol binding protein 3	gi|4506453|
Crystallin, β S	gi|345764|	Structural maintenance of chromosomes 1-like 1	gi|30581135|
Dystrobrevin, α isoform 8	gi|14916515|	Transferrin	gi|4557871|
Enolase 2	gi|5803011|	Triosephosphate isomerase 1	gi|4507645|
G3PDH	gi|7669492	*Tubulin, α 3*	gi|5174733|
Glucose phosphate isomerase	gi|18201905|	Ubiquitin and ribosomal protein L40	gi|4507761|
Glutamate-ammonia ligase	gi|2144562|	Ubiquitous mitochondrial creatine kinase	gi|10334859|
Haptoglobin	gi|4826762|	**Vimentin**	gi|14742600|
Haptoglobin-related protein	gi|123510|	**Vitronectin**	gi|72146|
Hemoglobin, α 2	gi|4504345|	14-3-3 protein β/α	gi|4507949|

The components consistent with those of AMD drusen are shown in bold letters. The components that belong to the gene families, other members of which are known to be constituents of drusen in AMD, are shown in italic letters. National Center for Biotechnology Information database accession and version numbers are listed.

Fig. 1

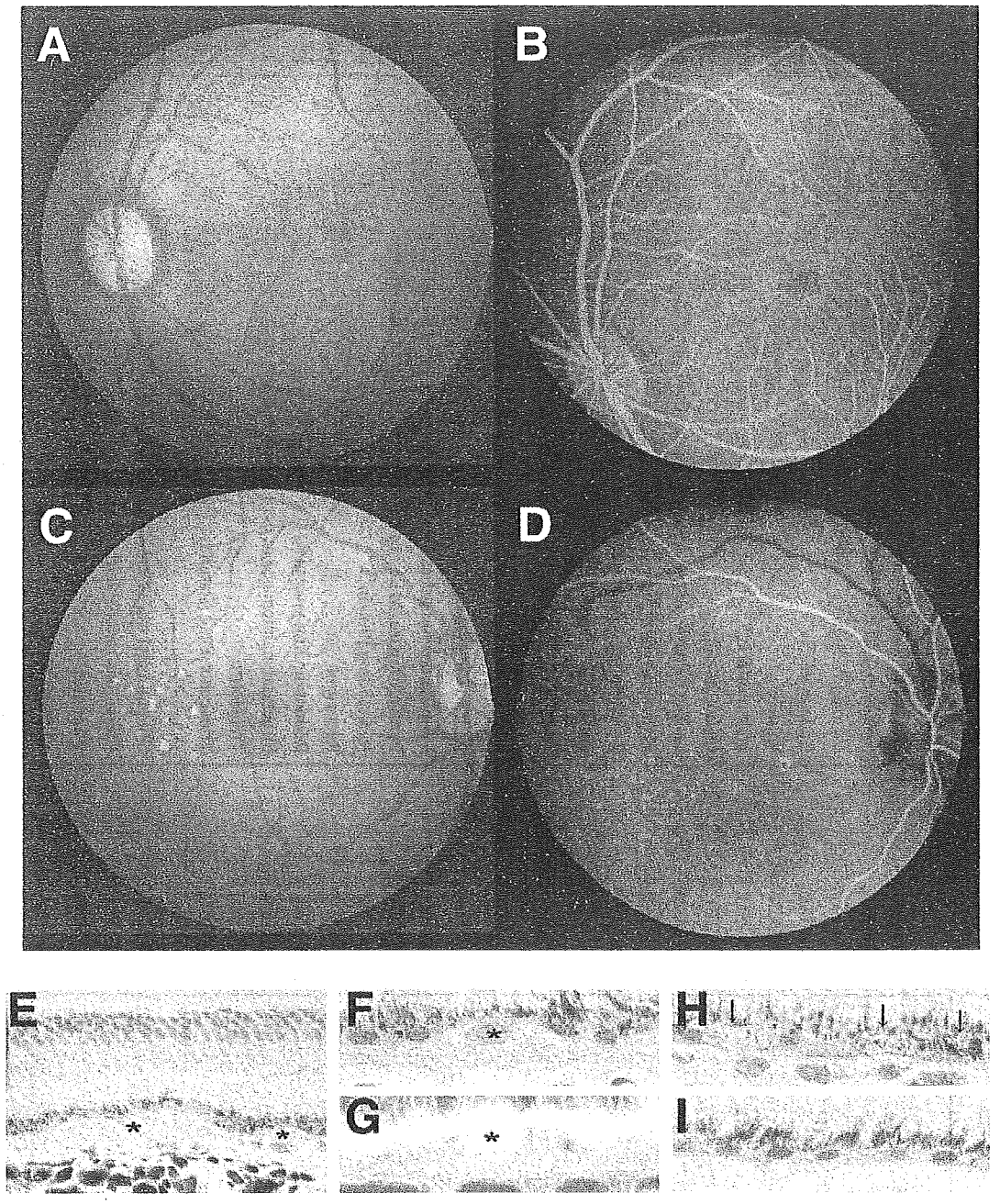

Figure 1. Drusen and degenerative changes of the RPE cells in late onset macular degeneration monkeys. Fundus photographs (*A*, *C*) and fluorescein angiography (FA) (*B*, *D*) of monkeys affected with late onset macular degeneration. Fine dots colored in yellowish-white could be observed in maculae (*A*, *C*). Hyperfluorescein dots could be detected by FA corresponding to these spots (*B*, *D*). Fundus photograph and FA of a 17-year-old monkey that showed vacuolation and hyper- or hypopigmentation of the RPE cells (*A*, *B*). The fundus photograph and FA of another 17-year-old monkey that showed drusen (*C*, *D*). No abnormalities were found in the optic disc nor the blood vessels. *E*) Various sized drusen accumulated between the RPE and choriocapillaris in the macular region (asterisks). Photoreceptor inner and outer segments appeared largely normal. *F*) Drusen that had an eosinophilic inclusion (asterisk). *G*) This spherical structure showed equivalent autofluorescence to that emitted by lipofuscin granules in the RPE (asterisk). *H*) Vacuolation and hyper- or hypopigmentation of the RPE cells (arrows). *I*) Intact region of the RPE in the same monkey as *H*.

Fig. 2

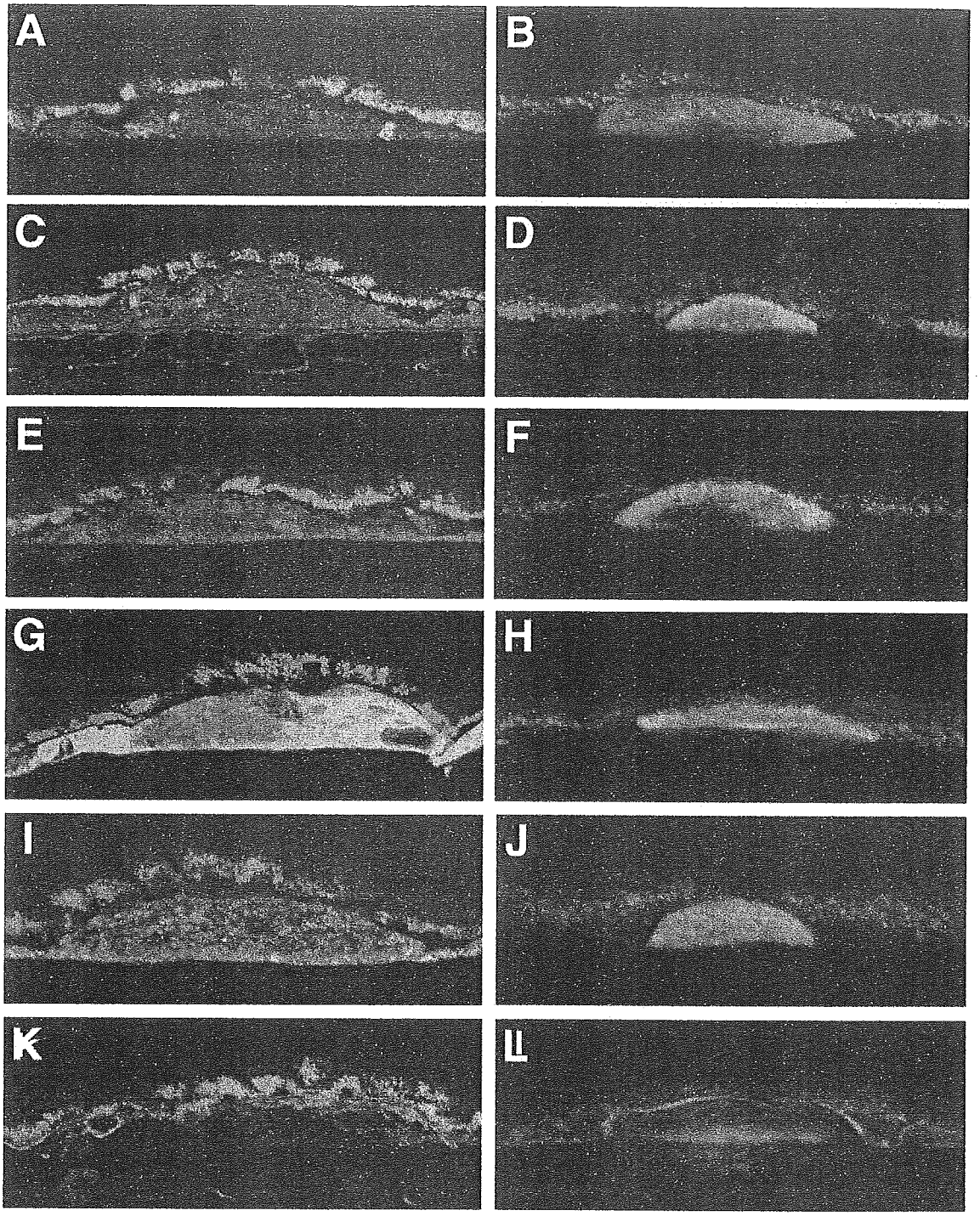

Figure 2. Drusen both in late onset and early onset macular degeneration monkeys are immun-reactive for protein components known in human AMD. Drusen in late onset (*A*, *C*, *E*, *G*, *I*, *K*) and early onset (*B*, *D*, *F*, *H*, *J*, *L*) macular degeneration were heterogeneously bound by antibodies directed against apolipoprotein E (*A*, *B*), amyloid P component (*C*, *D*), complement component C5 (*E*, *F*), the terminal C5b-9 complement complex (*G*, *H*), vitronectin (*I*, *J*), and membrane cofactor protein (*K*, *L*). Double-labeled images were generated by the green channel for each antigen and red channel for autofluorescence emitted by lipofuscin pigment in the RPE.

Fig. 3

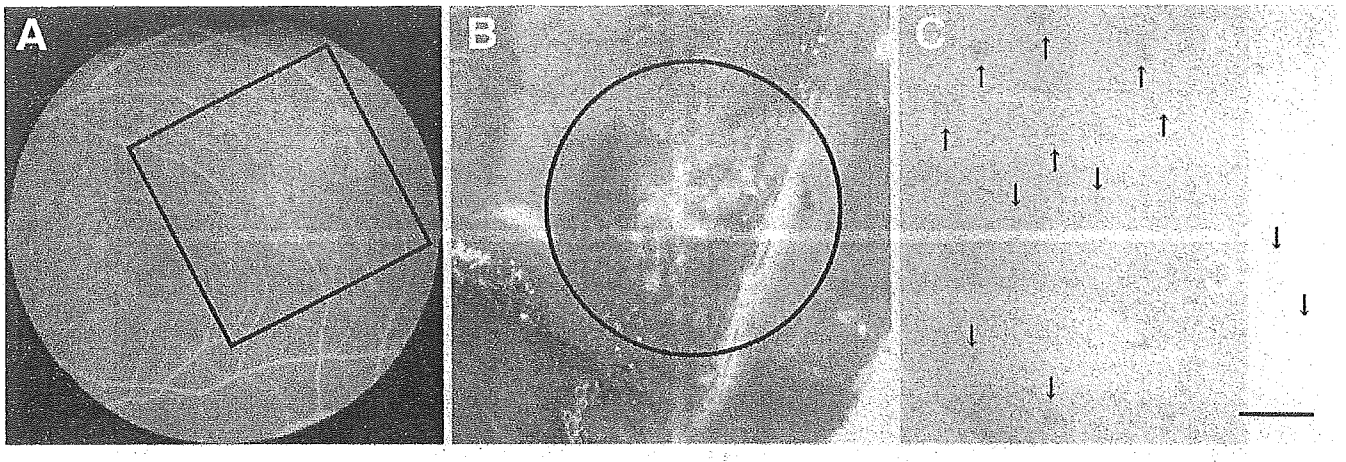

Figure 3. Isolation of drusen. ***A***) FA photograph of a monkey retina used for drusen isolation. A number of drusen that show hyperfluorescence could be observed in parafoveal region (indicated by a rectangle). ***B***) Drusen could be observed attached to surface of Bruch's membrane at magnifications between 20 and 30 diameters under a stereoscopic microscope (white materials in a circle). ***C***) Isolated drusen (arrows). Diameter of a circle in ***B*** = 3 mm. Bar in ***C*** = 1 mm.

Fig. 4

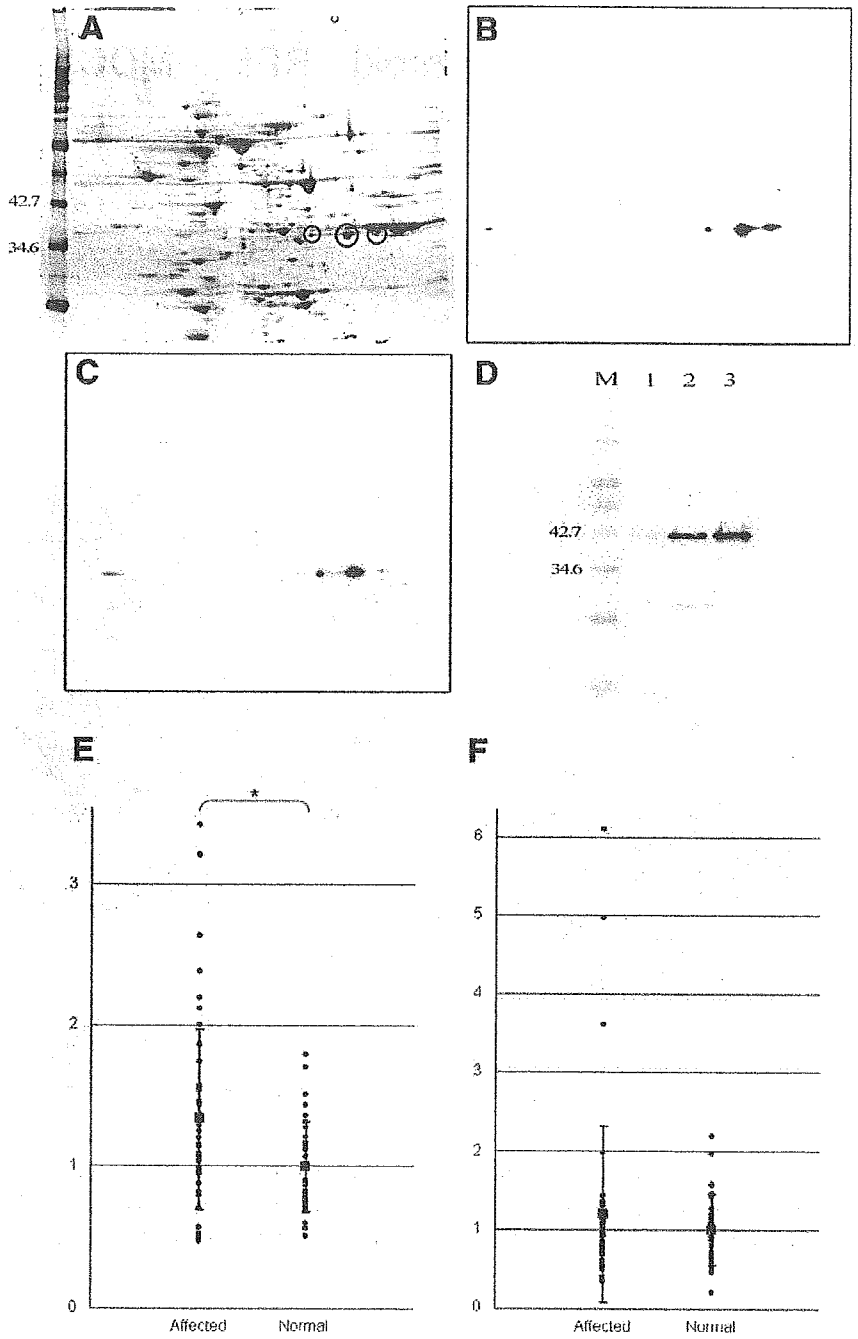

Figure 4. Identification of immunogenic molecules for autoantibodies in the affected monkey sera with late onset macular degeneration. *A*) 2-D electrophoretic image of retinal proteins visualized by SYPRO Ruby. *B*) Serum from same monkey in **Fig. 1***C* showed 3 immunoreactive spots in a row at ~38 kDa. Corresponding protein spots to chemiluminescent signals were excised (circles in *A*) and analyzed by LC-MS/MS. *C*) Chemiluminescent signals obtained by immunoreaction with anti-annexin II monoclonal antibodies completely matched those with the serum. *D*) The affinity purified recombinant annexin ran on SDS-PAGE gel at ~41 kDa (lane 1). Recombinant proteins reacted with both anti-annexin II monoclonal antibodies (lane 2) and autoantibodies in serum (lane 3). M, molecular size marker (kDa). Relative antibody titers against annexin II (*E*) or μ-crystallin (*F*) in sera from affected monkeys with late onset macular degeneration and age-matched control animals. Relative antibody titers of individual monkeys are indicated by ratio to mean of normal monkeys. **P* value <0.01.

Fig. 5

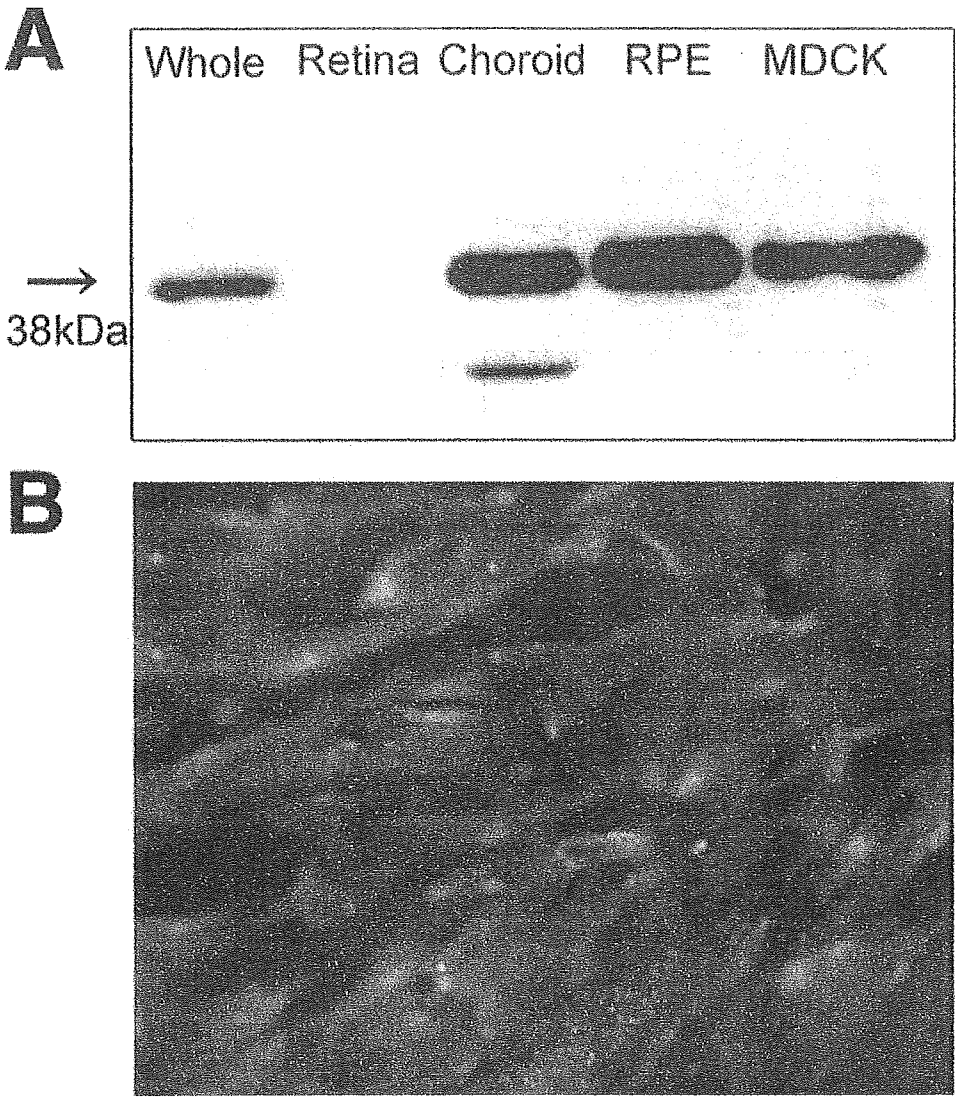

Figure 5. Expression of annexin II in the retina. *A*) Protein expression of annexin II (38 kDa) was confirmed in whole retina, choroid, and most intensively in cultured human RPE cells. Protein extract from MDCK cells was used for positive control. ***B***) Fluorescence microscopy demonstrated that RPE cells highly expressed annexin II in vitro.

J Hum Genet (2005) 50:382–394
DOI 10.1007/s10038-005-0261-9

ORIGINAL ARTICLE

Keiko Tadokoro · Mayu Yamazaki-Inoue
Maki Tachibana · Mina Fujishiro · Kazuaki Nagao
Masashi Toyoda · Miwako Ozaki · Masami Ono
Nobuhiro Miki · Toshiyuki Miyashita · Masao Yamada

Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in *DRPLA* affects subcellular localization of the products

Received: 12 May 2005 / Accepted: 17 May 2005 / Published online: 10 August 2005

Abstract Protein isoforms with or without a single amino acid residue make a subtle difference. It has been documented on a few genes that alternative splicing generated such isoforms; however, the fact has attracted little attention. We became aware of a subtle sequence difference in *DRPLA*, a polyglutamine disease gene for dentatorubral pallidoluysian atrophy. Some reported cDNA sequences lacked 3 nucleotides (nt) (CAG), which were positioned apart from the expandable and polymorphic CAG repeats and also coded for glutamine. We experimentally confirmed that the difference was indeed generated by alternative splicing utilizing two acceptors separated by 3 nt. In *DRPLA*, the expression ratio of two mRNA isoforms was almost constant among tissues, with the CAG-included form being major. The glutamine-included protein isoform was more predominantly localized in the nucleus. Database searching revealed that alternative splice acceptors, as well as donors, are frequently situated very close to each other. We experimentally confirmed two mRNA isoforms of 3 nt difference in more than 200 cases by RT-PCR and found interesting features associated with this phenomena. Inclusion of 3 nt tends to result in single amino acid inclusion despite the phase of translational frame. The expression ratio sometimes varied extensively among tissues.

Keywords Subtle alternative splicing · Splice acceptors · Polyglutamine diseases · DRPLA · Subcellular localization

K. Tadokoro · M. Yamazaki-Inoue · M. Tachibana
M. Fujishiro · K. Nagao · M. Toyoda · T. Miyashita
M. Yamada (✉)
National Research Institute for Child Health and Development,
2-10-1 Ohkura, Setagaya-ku, Tokyo 157-8535, Japan
E-mail: myamada@nch.go.jp
Tel.: +81-3-34160181
Fax: +81-3-54947035

M. Tachibana
Department of Pediatrics, Tokyo Medical University,
Tokyo, Japan

M. Fujishiro
Laboratory of Nucleic Acid Science, Nihon University,
Fujisawa, Japan

M. Ozaki
Laboratory for Memory and Learning, Brain Science Institute,
Riken, Wako, Japan

M. Ono · N. Miki
Institute of Clinical Endocrinology, Tokyo Women's
Medical University, Tokyo, Japan

Introduction

Alternative splicing of pre-mRNA is an important regulatory mechanism for quantitative control of gene expression and also for generating diverse protein species (Lopez 1998; Modrek and Lee 2002; Black 2003). It is highlighted, as the human-genome-sequencing project has revealed a limited number of genes (20,000–25,000) in the genome, which were far less than expected for functional complexity (International Human Genome Sequencing Consortium 2004). Recent surveys in databases revealed that as many as 40–60% of human genes are associated with alternative splicing (Black 2003). Three fundamental types of alternative splicing are noted: one is a choice in use of the entire potential exonic sequence (cassette exon or skipped exon) and the other two are selection of the site for a donor or acceptor in the potential continuous exonic sequence. Other types, mutually exclusive exon and intron retention, may be explained by the combination of the fundamental types. The choice for the first exon in alternative splicing is sometimes associated with differential regulation in expression, as regulatory elements are usually situated near the transcriptional start site. Alternative splicing

occurring in the coding region gives rise to generation of protein isoforms. Recent surveys also revealed that 70–88% of alternative splices cause a change in protein products (Black 2003). As one event of alternative splicing in a coding region usually generates a pair of protein isoforms, a multiple combination of alternative splicing events in a gene may generate an enormous number of protein isoforms, sometimes reaching to dozens or even hundreds.

Among diversified protein isoforms, a pair of isoforms with or without a single amino acid residue is the case with a minimum difference. Although it has been documented that such isoforms are generated by alternative splicing on a handful of genes (Manrow and Berger 1993; Condorelli et al. 1994; Vogan et al. 1996; Oberkofler et al. 1997; Lin et al. 2000) (Table 1), the phenomenon seems not to be justly recognized. The difference in amino acid as well as nucleotide sequences is so subtle that it may be disregarded or explained by experimental and/or human errors. We became aware of such a subtle difference in a previously reported cDNA sequence after comparison with sequences that other laboratories reported.

Dentatorubral pallidoluysian atrophy (DRPLA) is an autosomal dominant neurodegenerative disorder characterized by selective neuron loss in the cerebellar and pallidal outflow pathways (Kanazawa 1998). Patients show a combination of ataxia and extrapyramidal signs (chorea and athetosis) to varying degrees. We detected expansion of CAG repeats on chromosome 12p13 in the patients (Nagafuchi et al. 1994a) and then determined the entire cDNA sequence of the gene (Nagafuchi et al. 1994b). Several other neurodegenerative disorders, including Huntington's disease have been shown to be caused by expansion of CAG repeats in the coding region of respective genes (Cummings and Zoghbi 2000). As the CAG unit is situated in the translational frame and encodes glutamine, these disorders are collectively called polyglutamine diseases. It has often been observed through these diseases that the age of onset becomes younger with more severity in successive generations. This phenomenon, genetic anticipation, is now explained by the following two facts. The severity is correlated to the number of repeat iterations, and the extended repeats are further expanded when transmitted to the next generation. DRPLA is typical in genetic anticipation when the extended allele is transmitted paternally (Nagafuchi et al. 1994a; Cummings and Zoghbi 2000). As repeat expansion is a novel type of mutation associated with a subset of disorders and shows unique features, it attracts broad attention. The molecular mechanism underlying cell death has been demonstrated as induction of apoptosis at the final step, and several processes have been proposed for the route where the cohesive force of the polyglutamine tract is central (Okamura-Oho et al. 2003; Michalik and Van Broeckhoven 2003; Forman et al. 2004). In contrast, it is still uncertain how and why specific subsets of neurons are degenerated in respective polyglutamine diseases. To address this issue, we have been studying the normal functions of the *DRPLA* gene and its product.

In this report, we demonstrate an alternative splice in the *DRPLA* gene, which originated from a subtle sequence difference. We characterized the DRPLA protein isoforms with or without a single glutamine residue and found that such a subtle difference affected the protein nature. We extensively searched for similar phenomena in the literature and databases, experimentally examined cases, and revealed widespread occurrence of protein isoforms with or without a single amino acid residue in the human proteome due to alternative splicing. We propose the term of "subtle alternative splicing" for alternative splicing to generate a subtler difference, such as 3 nt, in mRNA and a single amino acid residue in protein.

Materials and methods

RT-PCR

RNA preparations from human, mouse, and rat tissues were purchased form several companies, including BD Biosciences Clontec, and Ambion and RNA from cultured cells and rodent tissue was prepared with a standard method, as described previously (Tadokoro et al. 1993; Miyashita et al. 1997). Fine, dissected, rat brain was performed with the aid of markers and ascertained by other means (Ozaki et al. 2004). One or two microgram of total RNA was subjected to direct cDNA synthesis using Superscript II RNaseH minus reverse transcriptase with the AP primer (an oligo-dT primer with an attached sequence) under the conditions recommended for the 3′-RACE system (Invitrogen, UK). An aliquot of synthesized cDNA was subjected to PCR amplification in a reaction containing 200 μM of each deoxynucleotide triphosphate, 0.5 μM of each primer, and 1 U of rTaq polymerase (Takara-bio). The PCR was carried out with a thermocycler in a cycling condition with an initial denaturation of 4 min at 94°C followed by 25–35 cycles of denaturation at 94°C for 1 min, annealing at Tm −10°C for 1 min, and extension at 72°C for 1 min, with a final extension at 72°C for 4 min (Tadokoro et al. 1993). The amounts of synthesized DNA from respective tissues were primarily adjusted to generate an equal level of *GAPDH* products, but the amounts were sometimes readjusted to facilitate the detection of two products with a 3 nt difference. Amplified PCR products were separated by electrophoresis through a 10% polyacrylamide or 3% SeaKem LE agarose gel (FMC BioProducts) and then detected by silver staining (Bio-Rad, Mississauga, ON, Canada) or ethidium bromide staining. Images were captured through a CCD camera, and the integrated optical density of detected bands was measured by the ImagePro-Plus and GelPro image analysis software (Media Cybernetics, Silver Spring, USA). In a mass screening, the ratio of two products with a 3 nt difference was represented in terms

Table 1 Representative genes with subtle alternative splicing utilizing two acceptor sites separated by 3 nucleotides (nt)

Genes[a]	Accession numbers			Exon	References
	cDNA with 3 nt	cDNA without 3 nt	Genome		
DRPLA	D31840	NM_001940 (U23851)	NT_000012 (U47924)	4–5	This study
GHRHR	–	NM_000823(L01406)	NT_000007(AC005155)	1–2	This study
BAIAP2	BC014020	NM_017450 (AB015019)	NT_000017(AC115099)	9–10	This study
PTMA	M14630	NM_002823(BC022433)	NT_000002(AC073476)	2–3	Manrow et al. (1993)
IGF1R	NM_000875 (X04434)	–	NC_0000156 (AY332722)	13–14	Condorelli et al. (1994)
PAX3	NM_181460 (AY251280)	BC008826	NT_000002 (AC010980)	2–3	Vogan et al. (1996)
PAX7	NM_002584 (X96743)	–	NT_000001 (AL021528)	2–3	Vogan et al. (1996)
LEP	NM_000230 (U43653)	D49487	NT_000007 (AC018635)	2–3	Oberkofler et al. (1997)
Dnmt1	AF162282	NM_010066 (X14805)	NT_039472 (CAAA01059910)	4–5	Lin et al. (2000)
CAST	NM_001750 (D16217)	NM_173061 (U58996)	NT_000005 (AC008906)	27–28	This study
MAN2B1	NM_000528 (U60266)	U05572	NT_000017 (AC010422)	7–8	This study
PSEN2	NM_000447 (L43964)	NM_012486 (U34349)	NT_000001 (AL391628)	10–11	This study
LAP1B	AK001780	NM_015602 (AK021613)	NT_000001 (AL353708)	2–3	ENSG00000143337
NOXO1	AB097667	NM_144603 (BC015917)	NT_037887	2–3	ENSG00000162042
CCL20	NM_004591 (D86955)	U64197	NT_005403	1–2	ENSG00000115009
SGNE1	BC005349	NM_003020 (Y00757)	NT_010194 (AJ290438)	3–4	ENSG00000166922
TGFA	NM_003236 (X70340)	BT006833	NT_022184	2–3	ENSG00000163235
	Nucleotide sequences[b]		Amino acid changes	Expression ratio in tissues[c] (longer/shorter)	
	Splice donor site	Splice acceptor site			
DRPLA	TGAGgtgggaa	gagtcttccctttttctacagCAGGAACTC	Gln/none	8:2–9:1	
GHRHR	GACCgtgagta	atcctgttcactgtttccagCAGGTATTG	Gln/none	2:8	
BAIAP2	ACAAgtaaggg	ttacctgtcccttgtcccagCAGCCGAGA	ThrAla/Thr	1:9–0:10	
PTMA	TGCTgtgagtg	atggcctgttttctgtcgagGAGAATGAG	Glu/none	0:10–1:9	
IGF1R	AAAAgtaaggc	tttcctcctctttgctgcagCAGGATATG	ThrGly/Arg	7:3–8:2	
PAX3	CAAGgtgagcg	gcccgcctgttctcttaaagCAGGTGACA	Gln/none	0:10–10:0	
PAX7	CAGAgtgagtg	tcccacctccacctctgaagCAGGTGGCG	Gln/none	0:10–9:1	
LEP	CACGgtaagga	tccctcttcctcctgcatagCAGTCAGTC	Gln/none	1:9–10:0	
Dnmt1	CTTTgtaagga	cactttctcttgtttttaagCAGTTGAAA	SerVal/Phe	4:6–6:4	
CAST	CTCGgtaagca	cagcattatttacttttcagCAGAGTGAC	Gln/none	9:1–10:0	
MAN2B1	GCAGgtcagtg	tccgtcgcctcccccagcagCAGGCAAAA	Gln/none	0:10–3:7	
PSEN2	ATGGgtgagta	ttctcttcctggacacccagAAGAAGACT	GluGlu/Glu	7:3	
LAP1B	CCAGgtaagaa	gttttctcttcctatattagCAGTGAGTG	AlaVal/Val	0:10–10:0	
NOXO1	CAAGgtgagtg	gccctggttcccgccctcagAAGACCCTC	Lys/none	0:10–9:1	
CCL20	GAAGgtaagtg	tcacttttttttttttttagCAGCAAGCA	AlaAla/Ala	4:6–9:1	
SGNE1	ACAGgtaacag	aaacctttggctgtttgcagCAGATGATG	AlaAsn/Asn	4:6–8:2	
TGFA	AGTGgtgagtg	tgcatctttcctctccccagCAGACCCGC	AlaAsn/Asn	5:5–9:1	

[a]All the genes listed in this table are human, except for mouse *Dnmt1*, as the structure of the corresponding intron–exon boundary in humans differs from that in the mouse. The phenomena for *PAX3* and *PAX7* were originally identified in mice, but the same phenomena were confirmed to occur in humans in this study

[b]Exon and intron for a longer isoform are indicated with upper and lower cases, respectively. Three nucleotides (nt) excluded in the shorter isoform by alternative splicing (optional 3 nt) are indicated in *bold*. The nucleotides encoding the indicated amino acid residues in the longer form are *underlined*

[c]The expression ratio of the longer to shorter isoform among tissues with a considerable expression level is indicated. When the ratio varied among tissues, the range is given. The figures are rounded out to the nearest whole number to bring the total to ten

of 11 scales from 0:10 to 10:0 (longer form/shorter form) after the figures were rounded out to the nearest whole number to bring the total to ten. When the minor form was detectable in images at 2–4% levels, it was indicated with f:10 or 10:f. The approximate estimation was also done by comparison with a standard picture, which was obtained by the mixture of *DRPLA* DNA fragments with a three nt difference in a defined ratio. The primer sequences for human *DRPLA* are described below, and those for other genes are shown in our Web site at http://www.nch.go.jp/genetics/subtlealtsp/.

DRPLA plasmid construction and splicing reaction

The *DRPLA* minigene pSP-DRPLA4-5 was constructed by insertion of a BglII-PstI fragment (1,893 bp from the 60th nt in exon 4 to the 870th nt in exon 5) derived from pMY1224 (Nagafuchi et al. 1994b) into the BamHI/PstI sites of the pSP65 vector (Promega, Madison, WI, USA). For the splicing reaction, pSP-DRPLA4-5 DNA was linearized by HindIII digestion and in vitro transcribed with SP6 RNA polymerase at 35°C for 3 h. Transcribed RNA was isolated with an RNeasy minikit (QIAGEN, Crawley, UK) and subjected to a splicing reaction with the HelaSplice Nuclear Extract (Promega, WA, USA) at 30°C for 16 h. Heat-inactivated nuclear extracts were used in a control to exclude contaminating sources. Treated RNA was converted to DNA with the R54 primer and amplified with the F42 and R51 primers in the RT-PCR condition described above. The sequences of each primer are as follows: R54: 5′-TGCTTCGGTTGTCCTGGTCGAT-3′; F42: 5′-TGAAAGTGAGGAGACCAATGCAC-3′; R51: 5′-GGAGGGAGACTGTGGCCGAG-3′. The PCR condition was adjusted at 58°C for annealing and 1.0 mM for the Mg concentration, and the product size was 70 bp or 73 bp. To construct *DRPLA* expression plasmid, a similar RT-PCR reaction was carried out with another set of primers—F41 and R54—to generate relatively large PCR products (167 bp or 170 bp). The sequence of the F41 primer was 5′-GGAGATCTCAGAGAGTGAAAGTG-3′. The PCR products were purified with QIAEX II silica gels (QIAGEN), digested with BglII and AvrII at their unique sites on the products, and then inserted into the corresponding sites of pMY1240. The BbsI-BstXI fragments (825 bp or 828 bp) covering the exon 4–5 junction in the resultant plasmid were used to replace the corresponding portion of pEGFP-DRPLA-Q14 or pEGFP-DRPLA-Q71, which were EGFP-tagged DRPLA expression plasmid under the CMV promoter (Miyashita et al. 1998). The Q14 was within the normal range while Q71 was in a disease range. It is noted that four Q are encoded by upstream CAGCAACAGCAA, thus the Q14 construct carries uninterrupted ten CAG repeats. The nucleotide sequences of the constructed plasmid were verified by sequencing. To construct the minigenes with a nucleotide substitution, in vitro mutagenesis was carried out on a PCR fragment, as described previously (Miyashita et al. 1997), and resulting fragments were used to substitute the BamHI/AvrII fragment of pSP-DRPLA4-5.

Cell manipulation

Human cervical carcinoma HeLa cells (ATCC CCL 2) were maintained in a Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin (Invitrogen) in a humidified atmosphere of 5% CO_2 at

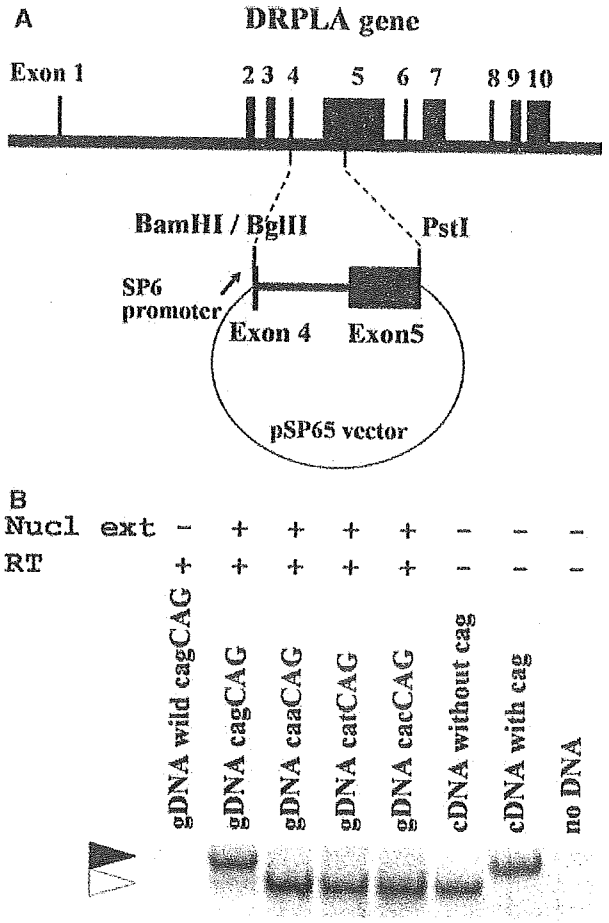

Fig. 1 The subtle difference in the *DRPLA* cDNA sequences caused by alternative splicing. **a** Schematic illustration of a *DRPLA* minigene construct with the genomic organization. pSP-DRPLA4-5 carried genomic DNA, which corresponded to a region from the middle of exon 4 through complete intron 4 to a halfway point in exon 5 under the SP6 promoter. **b** RT-PCR profiles showing generation of two RNA isoforms of 3 nucleotide (nt) difference after splicing reaction. RNA transcribed with the minigene and SP6 RNA polymerase was subjected to the reaction with nuclear extracts followed by RT-PCR. The products were analyzed by electrophoresis in parallel with PCR products generated with each form of cDNA as a reference. The products were confirmed to have expected sequences by sequencing. Heat-inactivated nuclear extracts did not generate spliced forms (*first lane*). After modifying the nucleotide at the −1 position to generate the longer isoform, only the shorter form was produced. g*DNA*, genomic DNA

37°C. Transient transfection was performed with the polycationic liposome method (LipofectAMINE PLUS Reagent, Invitrogen) according to the manufacturer's instructions. To assess the localization of EGFP-DRPLA fusion protein, cells were cultured in poly-d-lysine-coated glass-bottom plates (MatTek) and incubated in a medium containing 50 μM Z-VAD-FMK (Peptide Institute Inc.) after transfection (Miyashita et al. 1997). The cells were observed under a confocal microscope (FLUOVIEW, Olympus), and fluorescent-positive cells were counted in terms of the subcellular localization at the indicated period. The expression levels were ascertained by Western blotting, as described previously (Miyashita et al. 1997).

Database search and analyses

Previous reports on alternative splicing making a subtle difference were searched in Entrez Pub Med at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi. It was generally difficult to find because no common terms have been given for such cases. Nevertheless, we have detected a handful of reports although some others may remain undetected. For a subtle difference in amino acid sequences, we surveyed the ExPASy database (Expert Protein Analysis System) and found a file, humpvar.txt, at http://www.expasy.org/cgi-bin/lists?humpvar.txt, in which information on sequence variants is summarized from the Swiss-Prot human protein sequence entries. Among the list, we selected the cases of "one amino acid missing" and other subtle differences and then ascertained whether it could occur by subtle alternative splicing based on the standard genome and cDNA sequences at the National Center for Biotechnology Information (NCBI). A source data set of humans in the AltSplice database at http://www.ebi.ac.uk/asd/altsplice/ was downloaded and analyzed with database software, Kiri (KanriKogaku Kenkyusho Ltd., Japan). Although we used the prerelease Version 2, the release Version 1 was available after May 21, 2004. Throughout this study, we used genomic sequences to facilitate further analyses although RNA sequences may be more appropriate for splicing.

Results

Subtle alternative splicing in DRPLA

After detection of CAG repeat expansion in DRPLA pedigrees, we determined the entire cDNA sequence of the gene (accession number, D31840, Nagafuchi et al. 1994b). Since then, other laboratories have also reported cDNA sequences for the human *DRPLA* gene and orthologues (U23851, D38529 and others; Onodera et al. 1995; Love et al. 1995; Margolis et al. 1996). It is reasonable to find a difference in the number of iterations of the CAG repeats, which start at the 1,700th position of D31840, as it is polymorphic and expandable. Besides the repeats, the U23851 sequence lacked 3 nt occurring in other sequences (CAG at the 518th position of D31840), which resulted in the absence of the single glutamine residue at 94 in the DRPLA protein (also known as atrophin-1). After the entry of the sequence in the NCBI nucleotide database in 1995, we gradually considered the possibility of alternative splicing, as the inconsistent 3 nt was situated at the boundary of exon 4 and 5. To confirm this, we made a minigene system that contained genomic DNA covering a portion of exon 4, intron 4, and exon 5 under the control of SP6 promoter (Fig. 1). When in vitro transcribed products were subjected to react with nuclear extracts, two products with a 3 nt difference in size were generated. Sequencing verified that the products had the authentic cDNA sequences with or without the CAG nucleotides. This result clearly shows that alternative splicing takes place between exons 4 and 5. As the boundary around the distal end of intron 4 and the proximal start of exon 5 has a structure of cagCAGGAA, this result strongly suggests that two positions just after the cag and CAG serve as splice acceptors. (Hereafter, nucleotides in intron and exon are indicated with lower and upper cases, respectively, and the target nucleotides that are included or excluded in respective isoforms are indicated by bold upper case.) However, the result may be explained by other type of alternative splicing, for example, inclusion or exclusion of cassette exon consisting of only 3 nt. There are several cag sequences in intron 4, and even one

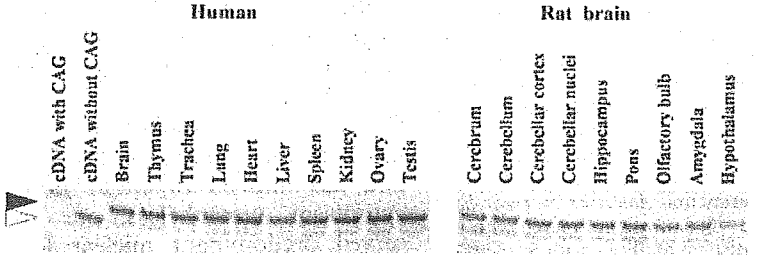

Fig. 2 Expression ratio of two *DRPLA* mRNA isoforms in various tissues. Total RNA from indicated tissues was analyzed by RT-PCR followed by electrophoresis. The PCR products (73 bp and 70 bp) generated with each form of cDNA as a template were loaded in parallel. As *DRPLA* expression levels in tissues varied extensively, the amount of transcribed RNA in RT-PCR was adjusted to give a similar level of PCR products. An almost constant expression ratio of the two transcripts was observed in repeated experiments, with the longer form being major, and representative results are illustrated

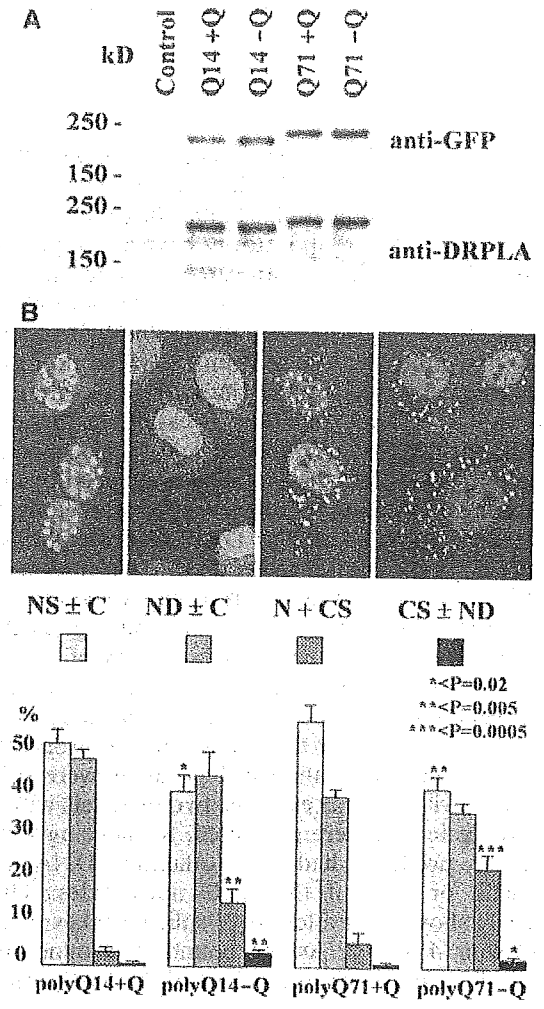

Fig. 3 Subcellular localization of DRPLA protein isoforms with or without a single glutamine residue. HeLa cells were transiently transfected, with one of the constructs carrying *DRPLA* cDNA (full length) tagged with EGFP at its N-terminus. *Q14* and *Q71* indicate the size of the polyglutamine chain (Q14 is within the normal range while Q71 is in a disease range). $+Q$ and $-Q$ indicate inclusion or exclusion of a single glutamine residue at the 94th position by subtle alternative splicing, respectively. **a** Expression levels were ascertained to be almost equal among various constructs by Western blotting with a mouse anti-GFP monoclonal antibody (B-2) or with a rabbit anti-DRPLA polyclonal antibody at 30 h after transfection. **b** Subcellular localization of the EGFP-DRPLA protein isoforms. We classified them into four categories: *NS± C* predominant nuclear localization as a large speckled form (or aggregates) with minor cytoplasmic distribution, *ND± C* predominant nuclear localization as small speckled or diffused forms with minor cytoplasmic distribution, *N+CS* almost equal distribution in the nucleus and cytoplasm, *CS± ND* predominant cytoplasmic localization with minor nuclear distribution. Typical images are illustrated in the *upper panels*. Respective patterns were counted with 100–200 cells at 30–40 h after transfection and repeated in four independent experiments. *Asterisks on error bars* show a significant difference of the −Q isoform compared with the corresponding +Q isoform by statistical analyses with Student's *t* test

agcag sequence. We substituted the g nucleotide at the −1 position to generate the longer transcript (CAG-included form) with one of the other nucleotides (i.e., caaCAGGAA). The minigene system with such a substitution only generated the shorter transcript. Thus, we conclude that the alternative splicing between DRPLA exons 4 and 5 takes place utilizing two acceptor sites separated by 3 nt.

Expression ratio of two mRNA isoforms of DRPLA

We previously reported with Northern blotting analyses that DRPLA was expressed in a wide variety of tissues (ubiquitous expression), but the extent varied considerably, with the brain, testis, and ovary being relatively high (Nagafuchi et al. 1994b). To distinguish two transcripts including or excluding the CAG sequence, the primer pair was set in exons 4 and 5 to generate relatively short RT-PCR products. The relative amounts of RT-PCR products varied considerably, which was almost consistent to our previous Northern blot results (data not shown). Thus, the amount of reverse-transcribed DNA from each tissue was adjusted so as to generate an almost equal level to easily detect the ratio of the two forms of transcripts. The expression ratio was almost constant among tissues, with the longer form being major (81%–91%) (Fig. 2, *left*). We pursued the possibility that either isoform of transcripts was prominent in the specific area of the brain, and we chose the rat because the size of the brain was large enough to be finely dissected, and the cagCAG structure at the boundary was conserved. The expression ratio of the two isoforms was almost constant among various sections of rat brain, with the longer form being major (79%–88%, Fig. 2, *right*). The result indicates that neither form alone nor the ratio of the two isoforms directly determines the brain area or the subset of neurons in DRPLA pathogenesis. We also analyzed mouse tissue during developmental stages and found consistent results (data not shown).

Characterization of DRPLA protein isoforms

Among the characters of DRPLA protein we were especially interested in the subcellular localization. Many previous reports indicate that DRPLA protein is a shuttle flying across the nuclear membrane (Ross 1997; Miyashita et al. 1998; Igarashi et al. 1998; Okamura-Oho et al. 1999; Miyashita et al. 1999; Ellerby et al. 1999; Yanagisawa et al. 2000; U et al. 2001; Okamura-Oho et al. 2003; Nucifora et al. 2003). DRPLA protein has both the nuclear localization and exclusion signals and has implied functions both in the nucleus and cytoplasm. Nonetheless, our group consistently observed predominant nuclear localization of DRPLA protein with very rare cytoplasmic localization in a variety of culture cells with expression experiments as well as for endogenous protein in staining with antibodies (Miyashita et al. 1998; Okamura-Oho et al. 1999; Miyashita et al. 1999; Yanagisawa et al. 2000; U et al.

Table 2 Distribution of distances between two adjacent splice donors as well as two adjacent splice acceptors according to the exon isoforms in the AltSplice data set[a]

Difference (nt)	3′ end of exon = donor sites (no. of pairs)					5′ end of exon = acceptor sites (no. of pairs)				
	A	B	C	D	E	F	G	H	I	J
2	17	4	2	3	2	36	9	4	5	3
3	30	16	11	8	5	600	310	258	199	176
3 nt, after cleaning up						(536	269	220	174	153)
4	108	33	18	13	8	146	57	31	36	21
5	28	13	8	5	3	89	41	27	24	20
6	35	11	9	4	4	71	26	15	12	10
7	14	7	5	2	2	30	7	4	5	3
8	14	6	5	5	4	17	6	2	3	1
9	48	23	18	15	13	35	10	4	3	2
10	19	9	5	4	1	17	5	3	4	3
11	21	12	8	6	5	16	4	0	1	0
12	47	27	22	13	11	51	19	18	12	12
13	13	6	2	2	1	18	6	3	5	2
14	10	2	1	0	0	21	8	3	4	1
15	26	9	7	5	4	50	19	10	7	3
16	13	5	3	4	2	18	5	4	2	2
17	16	8	8	4	4	24	5	3	3	2
18	35	11	7	6	4	61	17	15	11	9
19	16	6	6	5	5	28	8	3	3	1
20	9	2	2	1	1	29	12	9	9	7
21	15	5	3	2	2	44	17	11	8	8
22	14	7	4	5	2	24	7	3	4	1
23	6	3	1	2	0	15	3	0	2	0
24	18	8	7	3	3	41	17	11	7	7
25	11	2	1	0	0	18	5	3	4	3
Others	839	314	217	153	108	1496	586	410	323	239
Total	1422	549	380	270	194	2995	1209	854	696	536

[a]The data sets of human exon isoforms in the AltSplice database (http://www.ebi.ac.uk/asd/altsplice/, prerelease Version 2) were investigated. *A* From the exon file, pairs with an identical 5′ end and respective 3′ ends were selected and then it was confirmed whether the following intron shared the identical 3′ end (= downstream acceptor). *B* Among A, pairs were selected when the target exon as well as the following exon was covered with at least two transcripts for both forms. *C* Among B, pairs were selected when the ratio of the numbers of covering transcripts was in a range from 1:9 to 9:1. *D* Among A, pairs were selected when the target exon as well as the following exon was covered with at least three transcripts for both forms. *E* Among D, pairs were selected when the ratio of the numbers of covering transcripts was in a range from 1:9 to 9:1. *F* From the exon file, pairs with an identical 3′ end and respective 5′ ends were selected and then it was confirmed whether the preceding intron shared the identical 5′ end (= upstream donor). *G–J* Pairs were further selected as in B–E. The numbers in the *A and F columns* are slightly different from those described in the Event file in the AltSplice database

2001; Okamura-Oho et al. 2003). In contrast, some other groups, including Ross's group, who reported the sequence excluding the CAG sequence, more or less emphasized cytoplasmic localization (Ross 1997; Igarashi et al. 1998; Ellerby et al. 1999; Nucifora et al. 2003). We carefully examined subcellular localization in transfection experiments with four expression constructs with or without the glutamine residue coupled with extended polyglutamine in a disease range or short polyglutamine in a normal range after tagging with green fluorescent protein (GFP). Nuclear predominant localization was observed with the glutamine-included form. In contrast, a considerable fraction of the glutamine-excluded form localized in the cytoplasm although the majority was still in the nuclei (Fig. 3). The speckled structure in the cytoplasm may reflect a high tendency of agglutination of DRPLA protein although it was much influenced by the expression level, as observed in the nuclei. For another protein feature, phosphorylation by JNK (Okamura-Oho et al. 2003), two forms did not show a significant difference (data not shown).

Other cases of protein isoforms with or without a single amino acid residue

During the course of the above experiments, we experienced other cases of inclusion or exclusion of 3 nt in a cloning process of *GHRHR* encoding the growth hormone releasing hormone receptor (Miki et al. 1996) and also in the determination of exon–intron boundaries for *BAIAP2* encoding IRSp53 (Okamura-Oho et al. 2003; Miyahara et al. 2003). Literature surveys detected a few previous reports that clearly depicted isoforms with a subtle difference due to alternative splicing utilizing two acceptor sites separated by 3 nt (Manrow and Berger 1993; Condorelli et al. 1994; Vogan et al. 1996; Oberkofler et al. 1997; Lin et al. 2000; from *PTMA* to *Dnmt1* in Table 1). The *PAX3* and *PAX7* cases were originally reported with mice, and we confirmed that the human genome conserved the structure (aag**CAG**) at the intron–exon boundary and generated two mRNA isoforms, thus showing them with the human sequence in Table 1. In contrast, the human genome does not have a

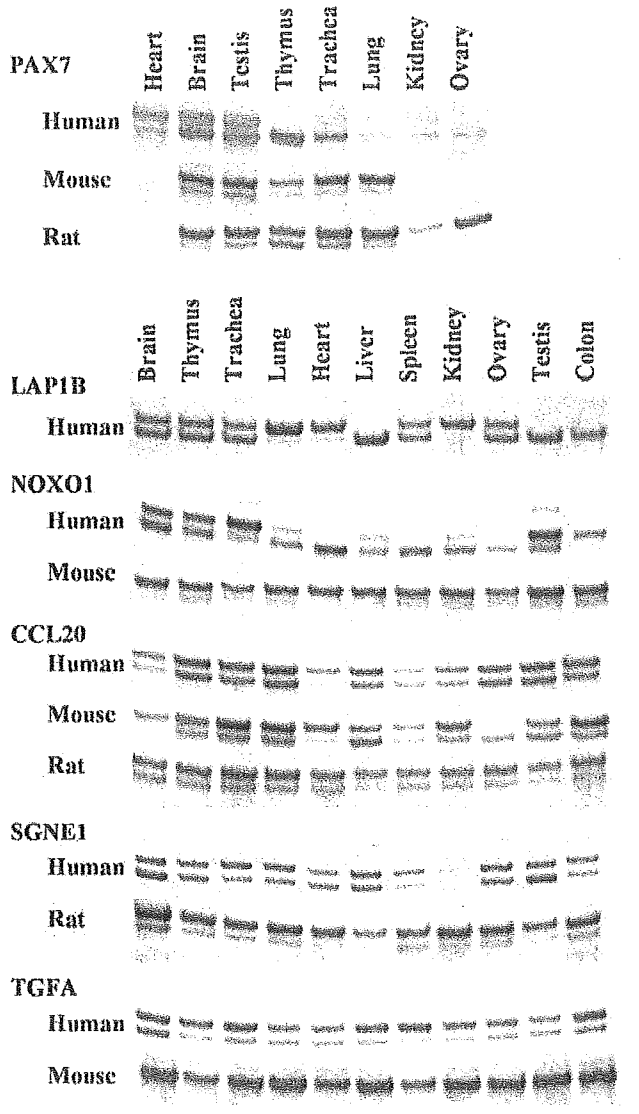

Fig. 4 Representative profiles of RT-PCR showing a variable ratio of two transcripts with 3 nucleotide (nt) difference by subtle alternative splicing among human tissues. Similar analyses of rodent's orthologues are also illustrated

corresponding sequence at the indicated intron–exon boundary of mice *Dnmt1*, thus only *Dnmt1* in Table 1 is shown with the mouse sequence. We then surveyed the ExPASy database for inconsistent human sequences with or without a single amino acid residue, where such cases are marked as "conflict." After consultation of genome databases to see whether it occurred at an exon boundary, we experimentally tested by RT-PCR and found that the cases of *CAST*,*MAN2B1*, and *PSEN2* were explained by this type of alternative splicing. Thus, subtle alternative splicing using two acceptor sites separated by 3 nt became 12 cases at this stage, and relevant sequences and features are illustrated in Table 1 (from *DRPLA* to *PSEN2*).

Candidates in the AltSplice database

Alternative splicing is one of the central issues in the current field of genetics, as the human-genome-sequencing project has revealed a limited number of genes in the genome. Several databases on alternative splicing have been constructed in which many transcripts, including expression sequence tags (ESTs), were aligned with each other and with genomic sequences to find a difference. Current databases on alternative splicing do not include most of the cases we point out above although such databases are valuable resources. During database searches, we found that a source data set was released to the public at one of the sites, the AltSplice database (http://www.ebi.ac.uk/asd/altsplice/, prerelease Version 2; Thanaraj et al. 2004), which was constructed with 18,632 known human genes and 614,877 transcripts and included 12,470 and 5,214 events for cassette exon and exon isoforms, respectively. When the retrieved data set was subjected to our analysis, we were amazed to find that alternative splice acceptors as well as donors were frequently situated close to each other. This phenomenon was not described in the database nor in the accompanied report (Thanaraj et al. 2004). The exon isoform events peaked at the position separated by 4 nt from the 3′ end of the exon (splice donor) and at the 3 nt position from the 5′ end of the exon (acceptor), sharply decreased along with the distance from the exon ends, and distributed widely with a slight high at the positions with a multiple of 3 (Table 2). As many as 600 alternative splicing events utilizing two acceptor sites separated by 3 nt were included in the data set. Some events shared the acceptor pairs with diverse donors, and some others were not confirmed with the standard genomic sequences at NCBI. After removing such events, the number of unique acceptor pairs became 536. We further selected 269 events, which were covered by at least two transcripts for both forms, and subjected them to RT-PCR to examine whether two forms of transcripts with a 3 nt difference were easily detectable. For the source of mRNA, commercially available mRNA preparations from ten human adult tissues as well as mRNA isolated from several cell lines were used. We evaluated the expression ratio with a score of 11 grades from 0:10 to 10:0 (longer form/shorter form) after the figures were rounded out to the nearest whole number to bring the total to ten. When the minor form was detectable in images at 2%–4% levels, the score was indicated with f:10 or 10:f.

According to the database, 220 out of the 269 events were represented by the numbers of covering transcripts within a range of 1:9 to 9:1 ratio. In more than 90% of these cases (202/220), two forms of 3 nt difference were easily detectable by RT-PCR, at least in a certain tissue showing a reasonable level of expression. In most cases, the ratio of two transcripts was almost constant among RNA sources. In contrast, only one transcript was apparently observed in most of the remaining 49 events that were represented by the numbers of covering tran-

Table 3 Summary of subtle alternative splicing utilizing two acceptors separated by 3 nucleotides (nt)[a]

Structure at boundary	Candidates in AltSp DB	Examined by RT-PCR	Confirmed two forms by RT-PCR	Expression ratio				Location in the coding region				Amino acid change		Others (aa-row)
				$L > S$	$L = S$	$L < S$	Ext. varied	In 5′ noncoding	In frame	Frame −1	Frame +1	Indel of single	Two/one change	
aagAAG	6	1	1	1							1	1		1
aagCAG	49	33v	25	8	2	12	3	3	9	5	8	22		10
aagGAG	2	1	1			1			1			1		1
aagTAG	11	6	6	1		5				2	4	4	2	1
cagAAG	67	22	15	13		1	1	1	3	2	9	13	1	4
cagCAG	206	123	110	60	5	34	11	20	34	19	37	86	4	24
cagGAG	17	5	3	3							3	3		1
cagTAG	28	13	12	8		3	1	4	1	3	4	8		2
gagAAG	3	1	1	1					1			1		1
gagCAG	5	2	1				1				1		1	
gagGAG	4	3	2			2		1	1			1		
gagTAG	1	1	1	1						1		1		
tagAAG	31	14	9	7		2		3			6	4	2	3
tagCAG	87	47	43	24	3	9	7	5	17	7	14	37	1	14
tagGAG	4	1	1	1					1			1		
tagTAG	15	4	4	2			2	1		3		3		3
Total	536	277	235	130	10	69	26	38	68	42	87	186	11	65
aagNAG	68	41	33	10	2	18	3	3	10	7	13	28	2	13
cagNAG	318	163	140	84	5	38	13	25	38	24	53	110	5	31
gagNAG	13	7	5	2	0	2	1	1	2	1	1	3	1	1
tagNAG	137	66	57	34	3	11	9	9	18	10	20	45	3	20
nagAAG	107	38	26	22	0	3	1	4	4	2	16	19	3	9
nagCAG	347	205	179	92	10	55	22	28	60	31	60	145	6	48
nagGAG	27	10	7	4	0	3	0	1	3	0	3	6	0	2
nagTAG	55	24	23	12	0	8	3	5	1	9	8	16	2	6

[a]The numbers of respective cases are shown. The examined cases include 11 original and previous cases but not mouse *Dnmt1*. In the expression ratio, $L > S$, $L = S$, and $L < S$ indicate that the longer form is more abundant than, almost equal to, and less than the shorter form, respectively. *Indel* insertion or deletion, *aa* amino acid. Supplementary information, including the respective genes, sequences, and the expression ratio, will be available on our Web site at http://www.nch.go.jp/genetics/subtlealtsp/

scripts outside the range of 1:9 to 9:1 ratio. The RT-PCR may not be suitable to validate such unbalanced cases. However, images sometimes showed the other form with the 3 nt difference at a very low level in certain tissue and also revealed a variable case (see below). Altogether, we confirmed by RT-PCR two isoforms of a 3 nt difference in 236 cases, including our original and previously reported cases, after removing redundancies.

The expression ratio of two forms of transcripts sometimes varied among RNA sources, and we assigned as an "extremely varied" case when the ratio fluctuated by four or more grades in the score (26 cases). In the ultimate cases, the quantitative major form was reversed, and even the ratio changed from the 0:10 to 10:0 grade. Several "extremely varied" cases are illustrated in Fig. 4, together with similar analyses with mouse and rat orthologues, if the structure of the intron–exon boundary was conserved. Human *PAX7* was expressed in a limited range of tissues, and the ratio was extremely varied, from 9:1 to almost 0:10, among tissues. Mouse and rat orthologues were expressed in a slightly different expression range of tissues, and the ratio was also different from that observed with the human counterpart. Five other human genes in Fig. 4 were ubiquitously expressed, but the ratio of two forms considerably varied among tissues. In contrast, rodent orthologues were expressed at an almost constant ratio, except for the mouse *Ccl20*, which was one of the two "extremely varied" cases that we detected in rodents to date. Genomic sequences of rat *Noxo1* are not available, and mouse *Sgne1* have a disrupted sequence (cagCTGATG). Orthologues of human *LAP1B* are probably mouse *MGC6357* and rat *Lap1b* and do not maintain the nagNAG sequence at the boundary. The ratio in human *TGFA* varied from 6:4 in the brain to 9:1 in the spleen, with the longer form being major, while both mouse (Fig. 4) and rat (data not shown) orthologues generated two forms in the f:10 ratio. Through these analyses, no correlation between predominance of one of the forms and the expression levels and also a particular tissue were detected.

The results are summarized in Table 3 where the events are classified in terms of the nucleotide species at the boundary (nagNAG) and include the expression ratio and resulting amino acid changes if inclusion or exclusion of 3 nt takes place in the coding region. Interesting features associated with this phenomenon are discussed in the following section.

Discussion

In this report, we have demonstrated that the subtle sequence difference in *DRPLA* cDNA was indeed due to alternative splicing utilizing two acceptor sites separated by 3 nt, and the resultant inclusion or exclusion of the single glutamine residue affected the subcellular localization of the product. With this study, together with a few previous studies on respective genes, we would call broad attention to the subtle differences in mRNA and protein structures by alternative splicing. Meantime, we found a variable data set on alternative splicing, experimentally examined by RT-PCR, and finally concluded that protein isoforms with or without a single amino acid residue are quite frequently generated. When we were preparing our manuscript, Hiller et al. (2004) reported widespread occurrence of alternative splicing at NAGNAG acceptors . Their conclusion was based on bioinformatics analyses. Our major conclusion is similar to theirs; however, we communicate our view as we have taken a different course.

Most databases on alternative splicing have been constructed by comparison of transcripts, mostly ESTs, with each other and also with genomic sequences. EST sequences were usually obtained by single-pass sequencing with a reverse-transcribed source of RNA isolated from a variety of tissues or cultured cells, and the accuracy was not usually validated. Thus, the EST sequences should be regarded as samples roughly representing the transcriptome although they are a powerful resource for various purposes. Scientists in the field of bioinformatics sometimes refer to the events revealed with EST as "experimentally confirmed" matters; however, they must be regarded as candidates and should be validated by any other means. In this point of view, we were very much interested in alternative splicing events making a 2 nt difference. The source of the AltSplice database contained as many as 36 events for two acceptor sites separated by 2 nt. If this type of alternative splicing really occurred, 2 nt were the minimal difference in mRNA isoforms although downstream amino acid sequences would differ considerably. We examined by RT-PCR all the nine cases that were covered by at least two transcripts for both forms in the data set (including *IKBKAP* covered by 12 and 4 transcripts and *NDUFB7* by 85 and 4 transcripts), as well as several other cases. All the cases examined apparently generated one form corresponding to the ag-included form, except for *GLYCTK* that generated the ag-excluded form, and no bands were detectable at the position where products of 2 nt difference were migrated. The results do not completely exclude the possibility of such a type of alternative splicing, as the events may occur at a more biased ratio or in a particular tissue that we did not examine. However, the RT-PCR results for the events with 2 nt difference are remarkably different from those with 3 nt difference where two forms were easily detectable at about 90%. As repeatedly pointed out, ESTs as well as alternative splicing databases based on ESTs contain the results of aberrant splicing (noise) (Sorek et al. 2004). Another irregularity with EST should be pointed out. The expression ratio of two forms that we experimentally determined by RT-PCR, excluding the extremely varied cases, does not necessarily coincide with the ratio of the numbers of covering transcripts in the database. This is partly accounted for by a relatively small number of transcripts covering the particular event for most cases. The number of EST is extensively biased due to the level of expression, the

source of transcripts, and the position within the gene, as most ESTs are generated from the 5′ or 3′ terminal of transcripts. Thus, experimental validation with other techniques is essential.

A total of 235 cases for which we experimentally confirmed two forms of 3 nt difference (after removing mouse *Dmnt1*) were subjected to further analyses and revealed the following features (Table 3). Preceding intron of all the cases is the GT-AG type for both proximal and distal acceptors, which generate longer and shorter forms, respectively. Thus, a nagNAG sequence appears at the intron–exon boundary of target exon, and the NAG sequence flanked by two acceptor sites is the target of inclusion or exclusion in the longer or shorter isoform (optional 3 nt). The frequencies of each nucleotide at the "n" and "N" positions in nagNAG show an interesting feature. A preferable nucleotide at the −3 position of exon is pyrimidine, according to the consensus splice site sequence, and a large data set in the AltSplice database (140,508 GT-AG type intron excluding redundancy) shows the frequencies of "a," "c," "g," and "t" at the position are 5.8%, 65.1%, 0.5%, and 28.6%, respectively. Compared with these frequencies, the nucleotide at the −3 position to generate the longer forms is somewhat unusual, with a slightly higher incidence of "a" and "g" (14.0% and 2.1%). Furthermore, the nucleotide at the −3 position to generate the shorter forms is extremely irregular, and appearance of "t" is significantly suppressed (9.8% versus 28.6%). This may be accounted for by the fact that the optional 3 nt is also served as the 5′ end of the exon. The large data set in the AltSplice database shows the frequencies of "A," "C," "G," and "T" at the +1 position of the exon are 25.7%, 14.4%, 48.6%, and 11.3%, respectively. Thus, the lower occurrence of TAG as the optional 3 nt may be explained by unfavorableness for the start of the exon. However, "C" is frequently situated at the position although it is also unfavorable for the start of the exon. For another possibility, mRNA having in-frame TAG may be degraded by a similar mechanism, as observed in genetic diseases where unexpected translational stop codon sometimes causes degradation of not only protein products but also mRNA. While our original cases imply a high occurrence of the optional 3 nt in-frame (8/12 cases), the data set from AltSplice shows the in-flame situation not so high (68/197 = 34.5%). Thus, the biased occurrence of nucleotide species at the "n" and "N" positions in nagNAG should be interpreted by other points.

The next issue is which adjacent acceptor is more frequently used? In general, the proximal acceptor (E acceptor in the report by Hiller et al. 2004) seems to be preferable, as the longer form is dominant in about 56% of cases. After classified in terms of nagNAG species, the distal acceptor (I acceptor) appears to be stronger when purine occurs at "n" and pyrimidine occurs at "N." For this consideration, it is more appropriate to include other cases where only one type of transcript is apparently produced with the nagNAG structure. The selection of the two adjacent sites in this type of alternative splicing is not explained by the simple scanning model and may be accomplished by interaction of a splice factor(s) through the sequence context. U2AF35 and Slu7 were demonstrated to be involved in the recognition of "ag" and also SPF45 in the selection of two adjacent "ag" separated by 16 nt in *Drosophila Sex-lethal* (Wu et al. 1999, Chua and Reed 2001; Lallena et al. 2002). However, no obvious rules come up at a glance, for example, in the length of pyrimidine-rich sequence and in the distance to a branch point, and no current programs for splice-site prediction detect the two sites with an appropriate value reflecting the ratio we experimentally determined. Moreover, we have revealed variable cases, which may become good markers to define the alternative slicing status in disease-associated changes and also to examine a role of isolated splice factors. These studies may be accelerated when the segments covering two adjacent acceptors are assembled on a microchip for detection of alternative splicing. Thus, the issue of site selection in this type of alternative splicing provides an enormous opportunity to improve the algorithm to find splice sites and also to elucidate the molecular mechanism of regulation in splicing.

The third feature associated with this type of alternative splicing is the location and amino acid changes. In about 15% of cases, the optional 3 nt localizes in the 5′ noncoding region and frequently occurs near the translational initiation site. In typical cases, the optional 3 nt is inserted just upstream of the initiation ATG, like tagCAGCC ATG in *STK38* and cagAAGCC ATG in *ORC1L*. These changes in mRNA may affect the efficiency of translation although we have yet only examined experimentally. For about 85% of cases, the optional 3 nt localizes in a coding region. The distribution of coding frames is not even, as previously pointed out (Hiller et al. 2004), but the biased distribution is milder in our study than the previous analysis. Despite the coding frame, inclusion and exclusion of optional 3 nt results in inclusion or exclusion of a single amino acid residue and exchange of one amino acid residue with two different amino acid residues (two/one type) is rare. This fact was already mentioned by Hiller et al. (2004) who proved this by counting the cases in a large data set and comparing with the events in an artificial null model. However, this is simply explained by the following facts: It is obvious when the optional 3 nt is situated in-frame. When the optional 3 nt is situated in coding frame −1 (intron phase 2 in the report by Hiller et al. 2004), the resultant frames are $N_1N_2NAGN_3$ and $N_1N_2N_3$ in the longer and shorter forms, respectively. The chance of an identical amino acid encoded by N_1N_2N at the proximal frame of the longer form and by $N_1N_2N_3$ in the shorter form is considerably high due to codon degeneracy, whatever nucleotide occupies N and N_3. Moreover, "G," especially "AG," frequently occupies the exon end as a mononucleotide and dinucleotide, and the large data set of the AltSplice database shows the frequencies at 81.3% and 55.6%, respectively. When "AG" occupies