J Hum Genet (2003) 48:410–414
DOI 10.1007/s10038-003-0047-x

ORIGINAL ARTICLE

**Atsushi Miyahara · Yuko Okamura-Oho
Toshiyuki Miyashita · Akinori Hoshika
Masao Yamada**

Genomic structure and alternative splicing of the insulin receptor tyrosine kinase substrate of 53-kDa protein

Received: 2 April 2003 / Accepted: 26 May 2003 / Published online: 16 July 2003

Abstract Insulin receptor tyrosine kinase substrate of 53-kDa protein (IRSp53) is now known to be a key factor in cytoskeleton reorganization. The human IRSp53 was identified as a binding partner with DRPLA protein, a product of the gene responsible for a neurodegenerative disorder, dentatorubral pallidoluysian atrophy, as well as a binding partner with brain-specific angiogenesis inhibitor 1. Previous studies identified at least four isoforms (L-, M-, S- and T-forms) in human, where 511 amino acid residues from the N-terminus were identical, followed by unique sequences of 9–41 amino acid residues. As each isoform had a distinct function, the unique sequences at the C-terminus had a vital role in its function. Here we report that these isoforms were indeed generated by alternative splicing, which was established by experimental and computational studies on human and rodent genomes. Previous biochemical reports suggested that rodents may lack one of the isoforms (L-form). This study solved this issue, as a nucleotide substitution occurred at a splice donor site followed by a large deletion in the rodent genome compared with human, which made the generation of the L-form impossible. This study also revealed overlapping of the *IRSp53* and *AATK* genes coded for by complementary strands.

Keywords Alternative splicing · Genomic organization · Insulin receptor tyrosine kinase substrate · IRSp53 · AATK

Accession numbers: AB104726, AB104729, AB105194, AB105195, AB105196

A. Miyahara · Y. Okamura-Oho · T. Miyashita · M. Yamada (✉)
Department of Genetics, National Research Institute for Child Health and Development,
3-35-31 Taishido, Setagaya-ku,
Tokyo 154-8567, Japan
E-mail: myamada@nch.go.jp
Tel.: +81-3-34160181
Fax: +81-3-34143208

A. Miyahara · A. Hoshika
Department of Pediatrics, Tokyo Medical University,
6-7-1 Nishishinjuku, Shinjuku-ku, Tokyo 160-0023, Japan

Introduction

An insulin receptor tyrosine kinase substrate of 53/58-kDa protein was originally identified in hamster cells through biochemical studies after insulin and/or IGF-I treatment (Yeh et al. 1996). It is phosphorylated upon stimulation with insulin and/or IGF-I, but differs from other members of the well-known insulin receptor substrate groups (namely, human *IRS1*, *IRS2* and *IRS4*) in terms of conserved amino-acid sequence motifs and other features (Hubbard and Till 2000). The human homologue was identified as a binding partner with *DRPLA* (Okamura-Oho et al. 1999), in which CAG triplet repeat expansion in the coding region causes a neurodegenerative disorders, dentatorubral pallidoluysian atrophy (Naito and Oyanagi 1982; Nagafuchi et al. 1994a, 1994b; Koide et al. 1994). The human homologue was also identified as a binding partner with a serpentine receptor, brain-specific angiogenesis inhibitor 1 (*BAI1*), and named as BAI1-associated protein 2 (*BAIAP2*) (Oda et al. 1999). The human homologue not only has a sequence similarity with hamster IRSp53/58, but also has been demonstrated to be phosphorylated upon stimulation with insulin and/or IGF-I (Okamura-Oho et al. 1999). IRSp53 is now highlighted as a key factor in the cytoskeleton reorganization: IRSp53 functions as an adaptor that binds Rho family GTPases (Rho, Rac and cdc42) and their effectors (mDia, WAVE2 and Mena), and mediates the activation of these molecules (Miki et al. 2000; Krugmann et al. 2001; Miki and Takenawa 2002). The cdc42 protein controls the formation of actin bundles in membrane ruffling and filopodia formation at the cellular periphery. IRSp53 is also known to localize at postsynaptic density of the central nerve system, which suggests a role in neurite outgrowth (Abbott et al. 1999; Soltau et al. 2002).

To date, at least four isoforms of IRSp53 have been identified in human. We identified IRSp53-L and IRSp53-S, consisting of 552 and 521 amino acid residues, respectively, as binding partners with DRPLA protein (Okamura-Oho et al. 1999). Oda and co-workers

(1999) identified two isoforms, named as BAIAP2-α and -β, which were composed of 521 and 520 amino acids, respectively. BAIAP2-α is identical to IRSp53-S, while BAIAP2-β is unique. Accordingly, we use IRSp53-T in this report rather than BAIAP2-β. The fourth isoform (IRS-58), with 534 amino acid residues, was identified during a cloning process of binding partners with cdc42 (Govind et al. 2001). As the relationship between protein isoforms of 53 or 58 kDa and mRNA isoforms is still uncertain, IRS-M is used in this report for the isoform. The four mRNA isoforms have been repeatedly confirmed in RT-PCR by both others and ourselves, as well as in many expression sequence tags (EST). The four IRSp53 transcripts generate respective protein isoforms sharing the identical 511 amino acid residues from the N-terminus and differing only in short peptide sequences at their C-terminus. Each isoform has distinct functions; for example, IRSp53-L and -S were phosphorylated with insulin but not with IGF-I in transfected cultured cells, while IGF-I phosphorylated only the T-form (Okamura-Oho et al. 2001). Thus, the unique short peptide sequences at the C-terminus have a vital role in its function probably through regulating accessibility to functional sites by intra-molecular binding. This is quite important as there are several discordance results in functional analyses with IRSp53 expression vectors. These isoforms are supposed to be generated by alternative splicing, but it has been proved yet. Here, we report that the four isoforms are indeed generated by alternative splicing by experimental and computational studies. This study on human and rodent genomes solved the issue of whether rodents lack one of the isoforms (L-form).

Materials and methods

DNA analyses

Human genomic DNA was isolated from peripheral leukocytes with the standard phenol/chloroform extraction method. Genomic DNA of mouse and Sprague-Dawley (SD) rat was prepared from tail tissues with the Dneasy Tissue Kit (QUIAGEN, Hilden, Germany). Polymerase chain reaction (PCR) was conducted as previously described (Tadokoro et al. 1992). Briefly, the reaction mixture consisted of 10 ng genomic DNA, 0.5 μM primers, 200 μM of each dNTP and 0.5 U Taq DNA polymerase (Takara, Shiga, Japan) in standard reaction buffer. The PCR conditions consisted of one cycle at 94 °C for 4 min, 30 cycles at 94 °C for 1 min for denaturation, 56 °C for 1 min for annealing and 72 °C for 4 min for extension, and then another step at 72 °C for 6 min to ensure complete extension. The following primer sets were used to amplify the exon 16-*AATK* region: mouse 5′-TGCAGTCCTGTGCCTTGCGA-3′ (forward) and 5′-AGAGATGCCCTCTGCAGGGTAGT-3′ (reverse); rat 5′-CAGGAATCCCTTCGCCAACGTC-3′ (forward) and 5′-AGATGCCCTCTGCAGGGTAGT-3′ (reverse). PCR products were purified with QIAXII (QIAGEN, Hilden, Germany), and subjected to a direct sequence analysis with CEQ2000 Dye Terminator Cycle Sequencing with the Quick Start Kit (Beckman Coulter, Fullerton, Calif., USA) and a Beckman CEQ2000 automatic sequencer (Beckman Coulter, Fullerton, Calif., USA). Sequence analyses were done with computer software, Genetyx-Win (Genetyx, Tokyo, Japan).

Results and discussion

Genomic organization of human *IRSp53*

When we started this study, few genomic sequences for *IRSp53* were detectable in public databases, which made it impossible to determine the genomic organization of the *IRSp53* gene only with computational analyses. Thus, we attempted to isolate genomic clones especially covering the C-terminal region with primer pairs designed based on the cDNA sequences. As the order of exons and their boundaries were unknown at that time, multiple combinations of primers were used to try to clone intronic sequences. Several sets of primers successfully generated DNA fragments by PCR and the nucleotide sequences were determined, some of which were deposited in a public database (see Comments). After more genomic sequences were deposited in public databases, along with the progress of human genome project, it became easier to identify genomic sequences covering the *IRSp53* gene by BLAST searches with the cDNA sequences as well as the genomic sequences we determined. The representative clones and sequences covering the *IRSp53* gene turned out to be PR11-149I9 (AC115099) and RP13-1277B16 (AC129919). These sequences had no annotation for *IRSp53* to date.

Comparing the cDNA (accession numbers NM_017450, NM_017451 and NM_006340 for the S-, L- and T-forms) and genomic sequences, the human *IRSp53* gene spanned about 82 kb, and consisted of 17 exons (Fig. 1). Except for the transcriptional termini (exons 14, 16 and 17), all the exon-intron boundaries were accorded for the consensus GT-AG rule (Fig. 2). The common part of the four isoforms was encoded by exon 1 through 13 (Fig. 3). The S-form went through to exon 14 and ended with a polyadenylation (polyA) signal. The nucleotide sequence we previously determined (AB017120) had 2,033 bp followed by the polyA tail, while the NM_017450 sequence had additional 135 bp.

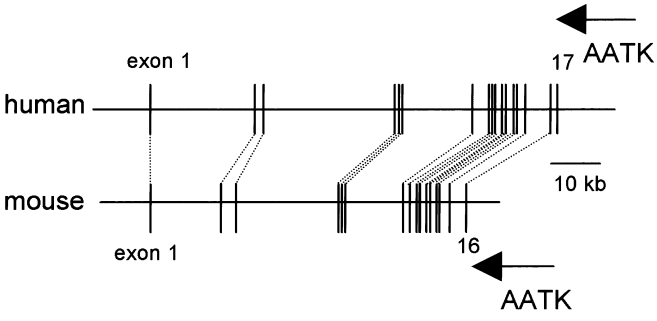

Fig. 1 Genomic organization of human and mouse *IRSp53* genes. Schematic illustrations showing that the genomic organization is generally well conserved between human and mouse, except for exon 17. Note that although we have tried to illustrate it as faithfully as possible, the exons and several introns are too small to draw in scale

human

intron		exon		intron
---tggttcgggt	CCGCT--	1	--ATAAG	gtgagcgccc---
22544		1	147	22692
---tttcttccag	ACCAT--	2	--GGCAG	gtggaactgc---
41041		148	223	41118
---tctctctcag	GTGTG--	3	--ACTCG	gtgagacccc---
45254		224	310	45342
---tgcttcccag	GAGAC--	4	--AAATG	gtgagtccac---
72205		311	372	72268
---ccatgtccag	CTGAA--	5	--TGAGT	gtaagtgcac---
73027		373	444	73100
---tgaatctcag	GCTGC--	6	--TGCAG	gtaggcccgc---
73816		445	582	73955
---ccctccacag	TACAT--	7	--CCAAG	gtgaggcggc---
87301		583	735	87455
---tccacttcag	GGCAA--	8	--TGGGG	gtgagtctgt---
90875		736	957	91098
---ttccctgcag	CGGAT--	9	--CACCA	gtaagggctc---
91280		958	1159	91483
---gtcccagcag	CCGAG--	10	--AAGAT	gtgagtgttt---
91887		1160	1361	92090
---tctcttccag	GCGGG--	11	--ATGAG	gtgagctctg---
93451		1362	1430	93521
---gtccctacag	CCTGC--	12	--CCCAG	gtcagtgggc---
94118		1431	1593	94282
---ctgcccccag	GGCCT--	13	--AGCAG	gtaaggggac---
95848		1594	1628	95884
---tgtttcacag	TGGCA--	14	--ATTGC	acgagttggg---
96341	S 1629		S 3168	97887
---ttccttgcag	CGCCG--	15	--AGTTA	gtaagttgcc---
98287	T 1629		T 1674	98334
---tctctttcag	GAATC--	16	--CCTGC	accagGTGTG---
103143	T 1675		T 2129	103599
---CCTGCaccag	GTGTG--	17	--ACAAT	aacttaaaat---
103603	L 1677		L 2877	104807

mouse

---tgcagttgtc	GCTTT--	1	--ACAAG	gtgagtttcc---
31401786		1	115	31401902
---tcttttccag	ACCAT--	2	--GGCAG	gtatagctgg---
31415754		116	191	31415831
---ctcccctcag	GTGTC--	3	--ACTTG	gtaagaccct---
31418809		192	278	31418897
---tttcctccag	GGGAC--	4	--AGACG	gtgagtttgg---
31439446		279	340	31439509
---ctgtgtccag	CTGAA---	5	--TAAGT	gtaagtacag---
31440050		341	412	31440123
---gggtctccag	GCTGC--	6	--TGCAG	gtaggtctgc---
31440646		413	550	31440785
---ctgtccccag	TACAT--	7	--CCAAG	gtgagctagg---
31452354		551	703	31452508
---tccactccag	GGCAA--	8	--TGGGG	gtgagtcctg---
31455181		704	928	31455407
---cttcctccag	CGGAT--	9	--CACCA	gtaagggcct---
31455547		929	1130	31455750
---gtctcggcag	CTGAG--	10	--AAGAT	gtgagcaccc---
31456147		1131	1332	31456350
---tttcttctag	GCGGG--	11	--ATGAG	gtaagcatac---
31457292		1333	1401	31457362
---cgccccgcag	CCTGC--	12	--CTCAG	gtgaggcctg---
31457793		1402	1564	31457957
---ttgttcccag	GGTCT--	13	--AGCAG	gtaagaggtt---
31459210		1565	1599	31459246
---tgttccacag	TGGCA--	14	--GCTCT	ctgcgcccct---
31459732		1600	1798	31459932
---ttccttacag	CGCGG--	15	--AGTTA	gtaagttgcc---
31461696				31461745
---ctttccccag	GAATC--	16	--TCACC	atgtgtagtg---
31465018				31465438

Fig. 2 Exon-intron boundaries of the human and mouse *IRSp53* genes. The boundaries were defined by alignment of the cDNA and genomic sequences. It should be noted that the downstream boundaries for exons 14, 16 and 17 are the end of the cDNA sequences indicated, and does not necessarily mean the position of the polyA tail. *Upper and lower* cases indicate exon and intron sequences, respectively, and the position of the boundary nucleotide in the given sequences are indicated. The accession numbers of referenced sequences are AC115099 for human genome, NM_017450 for human S-form cDNA, NM_017451 for human L-form cDNA, NM_006340 for human T-form cDNA, NT_039521 for mouse genome and AF390178 for mouse cDNA

There was a typical polyA signal in the genomic sequence near the end of AB017120, but also a continuous A sequence as well in the genome (at 96747 in AC115099). Thus, the exact termini of transcripts are

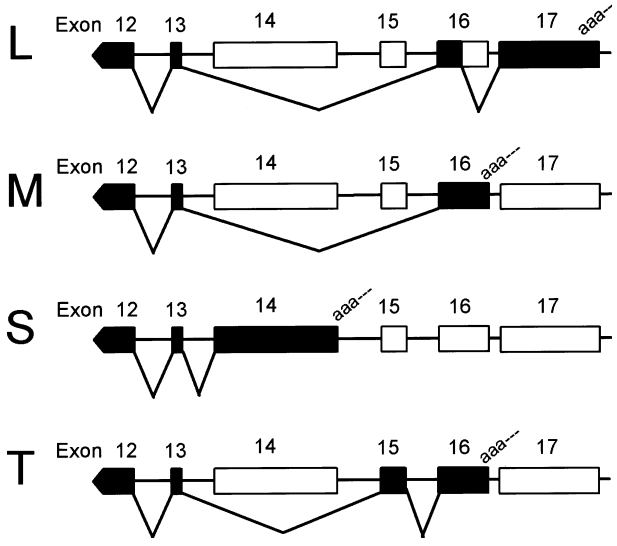

Fig. 3 Pattern of alternative splicing of the human *IRSp53* gene to generate four isoforms. The *black shaded* regions are used by respective isoforms; *aaa* polyA tails

somewhat uncertain, and we use a longer transcript in comparison in Fig. 2. The T-form skipped exon 14 and used exons15 and 16. Both the L and M forms skipped exons14 and 15 and reached to exon 16. The M-form ended with the polyA signal near the downstream boundary of exon 16. In contrast, the L form left exon 16 halfway, just ahead of the stop codon in frame, and resumed at exon 17, resulting in further extension of amino acid coding. The splice donor site in exon 16 for the L-form was also accorded for the consensus GT-AG rule, but slightly irregular as intronic +3 and 4 was CT (see below). It should be noted that there were only 5 bp between exons 16 and 17, but the downstream boundary of exon 16 did not provide a splice donor site.

Genomic organization in rodents

Mouse and human *IRSp53* were well conserved despite of long evolutional history. When compared with the S-form cDNA, they were 96% identical over the 522 amino acids and 87% identical at the nucleotide level over the entire coding region. Although three isoforms (S, T and M) were identified in rodents (including mouse, rat and hamster), it has been argued whether the L-form existed in rodents (Alvarez et al. 2002). We determined mouse and rat genomic sequences covering the region downstream of exon 16 (AB105194 and AB105195, respectively). Rat has been frequently used in studies on IRSp53 as its brain, one of the main expression sites, is larger than mouse. Recently, a mouse genomic sequence covering the entire coding region of *IRSp53* was deposited in the public database (NT_039521). Comparing the cDNA (AF390178) and genomic sequences, the mouse *IRSp53* gene spanned about 64 kb, and consisted of 16 exons (Fig. 1). The

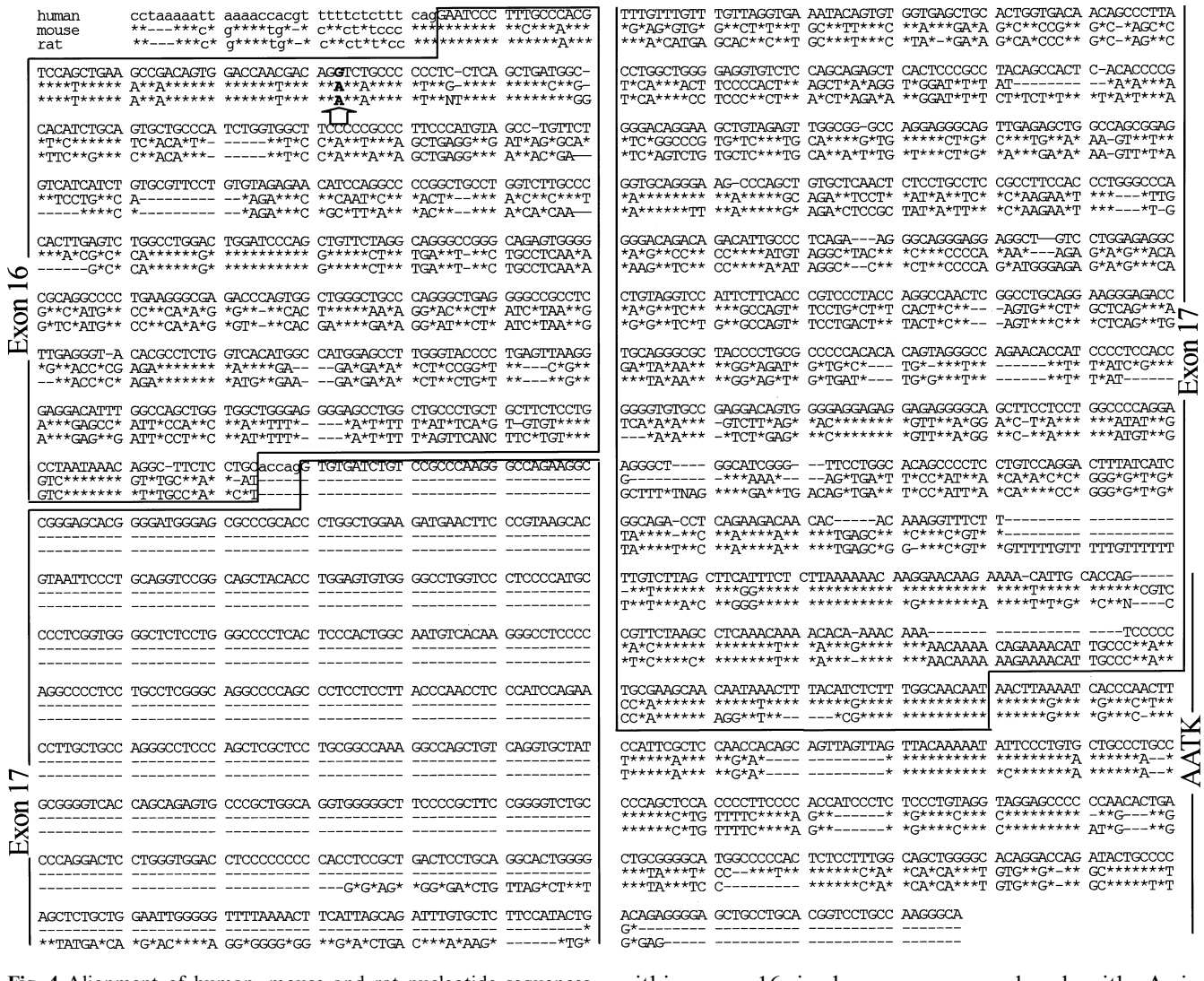

Fig. 4 Alignment of human, mouse and rat nucleotide sequences covering exons 16 and 17 of *IRSp53*. *Upper and lower cases* indicate exon and intron sequences, respectively. The identical nucleotides and gaps are indicate with an *asterisk* and a *dash*, respectively. The nucleotide change affecting the generation of the L-form is indicated by an *arrow*, and is also shown in Fig. 5. The accession numbers of the sequences are AC115099 for human, NT_039521 for mouse and AC105195 for rat. The regions for exon 16 and 17 for *IRSp53* are *boxed*. As the 3′ terminus of the *AATK* gene is unknown, it is not boxed but just indicated

genomic organization was also well conserved between human and mouse, although the size of several intron sequences varied. The exon-intron boundaries were similar (Fig. 2), and the generation mechanism of the M, S and T-forms was identical to human. However, there was a notable difference between human and rodent genomes, which affected the generation of the L-form. Mouse and rat sequences were shorter by about 400 bp in the region corresponding to human exon 17. In addition, there were many discordant nucleotides in the distal half of exon 16 and most parts of exon 17 when these were aligned (Fig. 4). The G nucleotide situated at the position corresponding to the splice donor site within exon 16 in human, was replaced with A in rodents (Fig. 4 *arrow*). Although several computer programs to predict splice sites (Splice view, http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html and Splice Site Prediction by Neural Network in Berkley *Drosophila* Genome Project http://www.fruitfly.org/seq_tools/splice.html) poorly recognized the splice donor site in human, the nucleotide substitution in rodents further decreased the possibility as the substitution abolished the GT consensus sequence at the boundary. Together with the finding of lack of the coding sequence specific to the L-form, we conclude that rodents do not generate the L-form.

We previously detected each protein isoform with specific antibodies recognizing the unique amino acid sequences at the C-terminus, where the L-form-specific antibody recognized some protein species in rat brain tissues (Okamura-Oho et al. 2001). Based on the study described here, the detected protein species were not derived from the IRSp53 L-form, although other results are still valid. The L-form-specific amino acid sequence may be coded for by another gene in rodents, as suggested (Alvarez 2002). The fifth isoform was recently

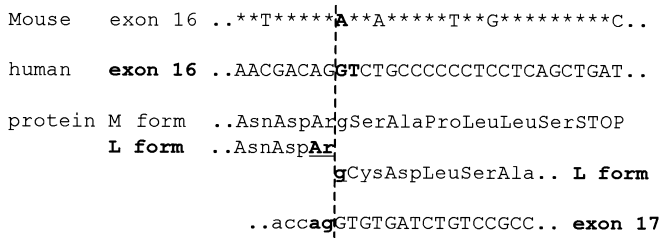

Fig. 5 Nucleotide change affecting in the generation of the L-form of *IRSp53*. *Upper and lower cases* indicate exon and intron sequences, respectively. The identical nucleotides between human and mouse are indicated with an *asterisk*. The amino acid sequences in the M-form, which reads through the indicated point, and the L-form, which is generated by splicing, are indicated *below* the nucleotide sequences

reported in rodents, which lacks 40 amino acids encoded by exon 9 (Alvarez et al. 2002). This is generated by the use of an additional splice acceptor site within exon 9 instead of the start position of exon 9 described in this report. A similar mechanism may be possible in human; however, no such transcripts detectable in human by RT-PCR were reported. The isoforms with and without the 40 amino acids may reflect the size difference in protein between 53 and 58 kDa. However, there is still confusion regarding the identity of protein species, some of which may due to different conditions for SDS-PAGE and as yet unknown posttranscriptional modifications. Therefore further studies will be required to identify protein species.

As regards evolution, it is interesting whether humans (or ancestor species) have gained the additional L-form or rodents have lost one of the isoforms. The former is plausible as each isoform of IRSp53 is involved in fine-tuning of its function, which may contribute to advancement of the central nerve system. This should be clarified by examination of other mammalian genomes, as well as by functional analyses of each isoform.

Finally, the downstream sequence (about 130 bp) of *IRSp53* was overlapped with the *AATK* gene, which encoded apoptosis-associated tyrosine kinase (Gaozza et al. 1997) (Fig. 4). The orientation of the transcripts was opposite (thus, encoded by the complementary strands) and the overlap occurred in their 3′-non-coding regions. This is highlighted by homology between human and rodent sequences in the vicinity of the end of exon 17, although the region was not used for *IRSp53* in rodents. The overlap is one of the examples for enriched gene distribution in a particular region of genome.

Acknowledgements We thank A. Asaka and Y. Ohtsuka for technical assistance, and K. Saito for preparing the manuscript. This study was supported in part by Grants for Human Genome, for Brain Science and for Pediatric Research from the Ministry of Health, Labor and Welfare, a Grant from the Organization for Pharmaceutical Safety and Research (Millennium Genome Project-5) and a Grant for Organized Research Combination System from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Abbott MA, Wells DG, Fallon JR (1999) The insulin receptor tyrosine kinase substrate p58/53 and the insulin receptor are components of CNS synapses. J Neurosci 19:7300–7308

Alvarez CE, Sutcliffe JG, Thomas EA (2002) Novel isoform of insulin receptor substrate p53/p58 is generated by alternative splicing in the CRIB/SH3-binding region. J Biol Chem 227:24728–24734

Gaozza E, Baker SJ, Vora RK, Reddy EP (1997) AATYK: a novel tyrosine kinase induced during growth arrest and apoptosis of myeloid cells. Oncogene 15:3127–3135

Govind S, Kozma R, Monfries C, Lim L, Ahmed S (2001) Cdc42Hs facilitates cytoskeletal reorganization and neurite outgrowth by localizing the 58-kD insulin receptor substrate to filamentous actin. J Cell Biol 152:579–594

Hubbard SR, Till JH (2000) Protein tyrosine kinase structure and function. Annu Rev Biochem 69:373–398

Koide R, Ikeuchi T, Onodera O, Tanaka H, Igarashi S, Endo K, Takahashi H, Kondo R, Ishikawa A, Hayashi T, Saito M, Tomoda A, Miike T, Naito H, Ikuta F, Tsuji S (1994) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nat Genet 6:9–13

Krugmann S, Jordens I, Gevaert K, Driessens M, Vandekerckhove J, Hall A (2001) Cdc42 induces filopodia by promoting the formation of an IRSp53:Mena complex. Curr Biol 11:1645–1655

Miki H, Yamaguchi H, Suetsugu S, Takenawa T (2000) IRSp53 is an essential intermediate between Rac and WAVE in the regulation of membrane ruffling. Nature 408:732–735

Miki H, Takenawa T (2002) WAVE2 serves a functional partner of IRSp53 by regulating its interaction with Rac. Biochem Biophys Res Commun 293:93–99

Modrek B, Lee C (2002) A genomic view of alternative splicing. Nat Genet 30:13–19

Nagafuchi S, Yanagisawa H, Sato K, Shirayama T, Ohsaki E, Bundo M, Takeda T, Tadokoro K, Kondo I, Murayama N, Tanaka Y, Kikushima H, Umino K, Kurosawa H, Furukawa T, Nihei K, Inoue T, Sano A, Komure O, Takahashi M, Yoshizawa T, Kanazawa I, Yamada M (1994a) Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nat Genet 6:14–18

Nagafuchi S, Yanagisawa H, Ohsaki E, Shirayama T, Tadokoro K, Inoue T, Yamada M (1994b) Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). Nat Genet 8:177–182

Naito H, Oyanagi S (1982) Familial myoclonus epilepsy and choreoathetosis: Hereditary dentatorubral-pallidoluysian atrophy. Neurology 32:798–807

Oda K, Shiratsuchi T, Nishimori H, Inazawa J, Yoshikawa H, Taketani Y, Nakamura Y, Tokino T (1999) Identification of BAIAP2 (BAI-associated protein 2), a novel human homologue of hamster IRSp53, whose SH3 domain interacts with the cytoplasmic domain of BAI1. Cytogenet Cell Genet 84:75–82

Okamura-Oho Y, Miyashita T, Ohmi K, Yamada M (1999) Dentatorubral-pallidoluysian atrophy protein interacts through a proline-rich region near polyglutamine with the SH3 domain of an insulin receptor tyrosine kinase substrate. Hum Mol Genet 8:947–957

Okamura-Oho Y, Miyashita T, Yamada M (2001) Distinctive tissue distribution and phosphorylation of IRSp53 isoforms. Biochem Biophys Res Commun 289:957–960

Soltau M, Richter D, Kreienkamp HJ (2002) The insulin receptor substrate IRSp53 links postsynaptic shank1 to the small G-protein cdc42. Mol Cell Neurosci 21:575–583

Tadokoro K, Fujii H, Ohshima A, Kakizawa Y, Shimizu K, Sakai A, Sumiyoshi K, Inoue T, Hayashi Y, Yamada M (1992) Intragenic homozygous deletion of the *WT1* gene in Wilms' tumor. Oncogene 7:1215–1221

Yeh TC, Ogawa W, Danielsen AG, Roth RA (1996) Characterization and cloning of a 58/53-kDa substrate of the insulin receptor tyrosine kinase. J Biol Chem 271:2921–2928