Identification and characterization of the human *StARD9* gene in the LGMD2A-region on chromosome 15q15 by *in silico* methods

NIELS HALAMA[1], SILKE A. GRAULING-HALAMA[2] and DIRK JÄGER[1]

[1]National Center for Tumor Diseases, University of Heidelberg, Heidelberg; [2]Institute for Microbiology and Hygiene, Clinical Faculty Mannheim, University of Heidelberg, Mannheim, Germany

Received April 28, 2006; Accepted July 13, 2006

Abstract. StAR (steroidogenic acute regulatory) proteins and proteins with StAR-related lipid transfer (START) domains are involved in lipid transport and metabolism, signal transduction, and transcriptional regulation. In this present study we characterized the *StARD9* gene by bioinformatical methods. The human *StARD9* gene (syn. *KIAA1300*), consisting of 11 exons, was located within human genome sequence CTD-2036P10 (AC090510). The complete coding sequence of human *StARD9* cDNA was determined by DKFZp781J069 cDNA (CR749416), which was identified by using a partial coding sequence of human *StARD9* cDNA (AB037721) as a query. The existing predicted gene model could be refined based on EST-analysis. Comparison of the human protein sequence with chimpanzee, rat, mouse and chicken homologs showed a 98% (chimpanzee), 60% (rat), 60% (mouse) and 45% (chicken) amino-acid identity respectively. A lipid-binding START domain was identified within *StARD9* protein. This is the first report on characterization of the *StARD9* gene.

Introduction

The StAR-related lipid transfer (START) containing proteins that have been analyzed to date have a multitude of functions (1). The START domains are motifs of 200-210 amino acids which occur in a remarkably wide range of proteins involved in diverse cell functions (2). Steroidogenic acute regulatory protein (StAR) is involved in the transfer of cholesterol to the inner mitochondrial membrane of steroidogenic cells. StAR proteins and their homologs are not only found in the adrenal cortex and gonads, but in placenta and brain and as over-expressed proteins in some breast carcinomas. But START domain proteins are not limited to cholesterol transfer. The signal transducing proteins *p122-Rho-GAP*, *Goodpasture antigen binding protein* and others show the variety of possible implications for START containing proteins. Interestingly *StARD9* (syn. *KIAA1300*) lies on the chromosomal region 15q15, which is the target region of genetic analysis of LGMD2A, an autosomal recessive form of limb-girdle muscular dystrophy (3). Seventeen autosomal loci have been identified (4). Besides the known mutations in the *Calpain-3* locus many new genotypes have been found but modifying effects through regulatory genes are still suspected.

We characterize a novel StAR-domain containing gene by bioinformatical methods. A partial coding region of human *StARD9* (AB037721) was found in the HUGO gene database (http://www.gene.ucl.ac.uk/nomenclature/). Using the sequence data from AB037721 we identified DKFZp781J069 cDNA as full-length StARD9. Genomic structure (intron and exon boundaries), detailed human chromosomal localization, expression data, functional relevant SNPs (single nucleotide polymorphisms) and comparative protein homology will be discussed. This is the first report on the identification and characterization of the *StARD9* gene.

Materials and methods

Identification of a novel gene. Automatic annotation of the human genome has produced numerous entries but in many cases these entries are computational predictions and are therefore often incomplete. Human genome sequences, expressed sequence tags (ESTs) and uncharacterized cDNAs were analysed with tBLASTn and BLASTp programs (http://www.ncbi.nlm.nih.gov/BLAST) (5). Amino-acid sequence of human StAR-domain containing proteins was used as a query sequence. HUGO gene database (http://www.gene.ucl.ac.uk/nomenclature/) was used to identify uncharacterized StAR-domain containing proteins.

Structural and chromosomal localization of the novel gene. Exon-intron boundaries were determined by examining the consensus sequence of exon-intron junctions ('gt...ag' rule of intronic sequence) and the codon usage within the coding region. To refine the exon-intron boundaries the existing cDNAs or ESTs (expressed sequence tags) were aligned to the genomic sequence using SPIDEY (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/). This revealed a different genomic structure of *StARD9* as predicted by the automated computational analysis. These findings were then used to identify the human chromosomal localization of the genome clones in the Ensembl database (http://www.ensembl.org/) and

Correspondence to: Dr N. Halama, National Center for Tumor Diseases, University of Heidelberg, INF 350, Heidelberg 69120, Germany

Key words: steroidogenic acute regulatory, chromosome 15q15

subsequently the precise human chromosomal localization of *StARD9*. Existing alternative splicing databases were searched for known alternatively spliced transcripts (EBI alternate splicing database, http://www.ebi.ac.uk/asd/; alternative splicing database, http://hazelton.lbl.gov/~teplitski/alt/; human alternative splicing database http://www.bioinformatics.ucla.edu/~splice/HASDB/) but did not yield any information on alternative transcripts of *StARD9*.

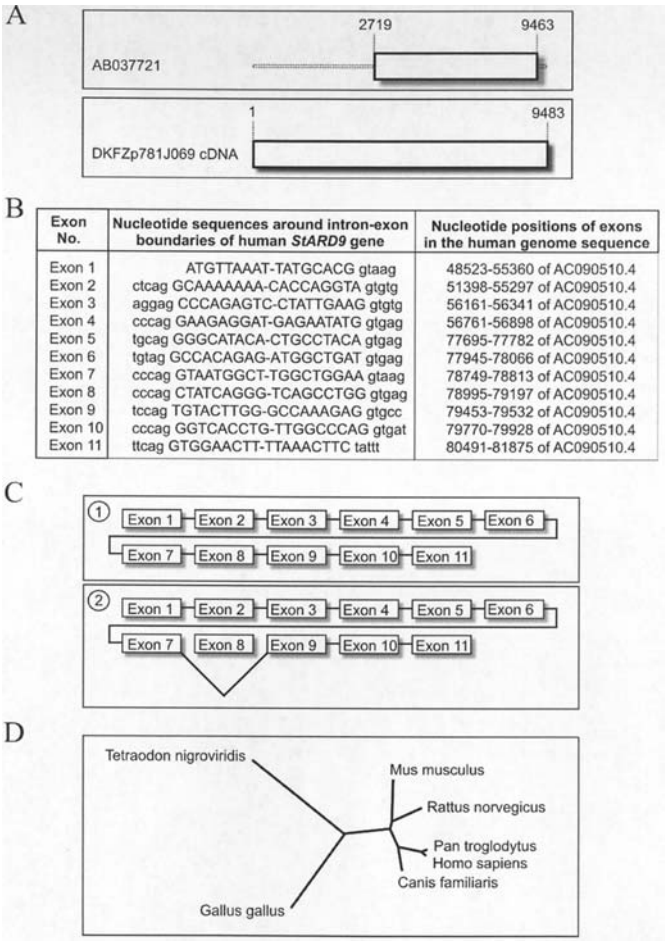
Analysis of deduced amino-acid sequence. Using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and the protein database (http://www.expasy.uniprot.org/) we predicted the coding region. Translation into amino-acid sequence was performed using ORF Finder, analysis of the identified domains was performed using InterProScan (http://www.ebi.ac.uk/InterProScan/) and the Pfam program (http://pfam.wustl.edu/). The comparative analysis of *StARD9* in different species was performed using the ClustalW (http://www.ebi.ac.uk/clustalw/) and Ensembl (http://www.ensembl.org) databases (6,7).

Expression profiling. The expression profiles for normal human tissues were obtained from GeneAnnot (12 normalized tissues were hybridized against Affymetrix GeneChips HG-U95A-E with optimal annotation quality, (http://bioinfo2.weizmann.ac.il/cgi-bin/geneannot/GA_search.pl?keyword_type=gene_symbol&keyword=*STARD9*) and from ArrayExpress (Affymetrix Gene Chip HG-U133A, http://www.ebi.ac.uk/aedw/dwd?uniprot=Q9P2P6). Furthermore electronic Northern analysis of NCBI's UniGene dataset was extracted from GeneCards (http://www.genecards.org).

Functional relevant SNP evaluation. To identify possible functional relevant SNPs (single nucleotide polymorphisms) that could disrupt ESE/ESS (exonic splicing enhancer/exonic splicing silencer) motifs we extracted SNP data from Ensembl (http://www.ensembl.org) and NCBI's SNPdb (http://www.ncbi.nlm.nih.gov). Using RESCUE-ESE web server software (http://genes.mit.edu/cgi-bin/rescue-ese_new.pl) we analyzed 10 bp in either direction around each SNP.

Results

Human StARD9 gene. Among the characterized StAR-domain containing proteins found in the HUGO gene database (http://www.gene.ucl.ac.uk/nomenclature/), *StARD9* cDNA (AB037721) was a partial cDNA. To find the full-length *StARD9* cDNA, AB037721 was used as a query sequence for the BLAST program. As a result human DKFZp781J069 cDNA was found to encode full-length human *StARD9* (Fig. 1). Human *StARD9* gene fragments were identified within human genome sequence CTD-2036P10 (AC090510) by using human full-length *StARD9* cDNA (1-9481) as a query sequence for the BLAST program. Boundaries of exon-intron junctions were determined through alignment of ESTs to the genomic sequence and through observation of consensus sequences at exon-intron junctions. *StARD9* consists of 11 exons. *StARD9* was localized to the genomic region of chromosome 15q15 due to the mapping of the human genome clone AC090510 to this region. The complete coding sequence of human *StARD9* cDNA was deduced



Exon No.	Nucleotide sequences around intron-exon boundaries of human *StARD9* gene	Nucleotide positions of exons in the human genome sequence
Exon 1	ATGTTAAAT-TATGCACG gtaag	48523-55360 of AC090510.4
Exon 2	ctcag GCAAAAAAA-CACCAGGTA gtgtg	51398-55297 of AC090510.4
Exon 3	aggag CCCAGAGTC-CTATTGAAG gtgtg	56161-56341 of AC090510.4
Exon 4	cccag GAAGAGGAT-GAGAATATG gtgag	56761-56898 of AC090510.4
Exon 5	tgcag GGGCATACA-CTGCCTACA gtgag	77695-77782 of AC090510.4
Exon 6	tgtag GCCACAGAG-ATGGCTGAT gtgag	77945-78066 of AC090510.4
Exon 7	cccag GTAATGGCT-TGGCTGGAA gtaag	78749-78813 of AC090510.4
Exon 8	cccag CTATCAGGG-TCAGCCTGG gtgag	78995-79197 of AC090510.4
Exon 9	tccag TGTACTTGG-GCCAAAGAG gtgcc	79453-79532 of AC090510.4
Exon 10	cccag GGTCACCTG-TTGGCCCAG gtgat	79770-79928 of AC090510.4
Exon 11	ttcag GTGGAACTT-TTAAACTTC tattt	80491-81875 of AC090510.4

Figure 1. Structure and phylogeny of StARD9. (A) Partial cDNA and complete cDNA. (B) Exon-Intron boundaries and nucleotide positions. (C1) Full-length transcript. (C2) Alternate transcript with skipping of exon 8. (D) Phylogenic tree of StARD9 and its orthologs in different species.

using full-length *StARD9* cDNA (Fig. 2). Human *StARD9* cDNA consists of a 9483-bp coding region, a 152-bp 5'-UTR and it encodes an 1820-aa KIAA1300 protein (Fig. 2). Beside the full-length transcript exists an alternatively spliced transcript (*StARD9* isoform 2), which does not have exon 8. There was however no full-length cDNA derived from the alternatively spliced *StARD9* form in the public database. This indicates that full-length human *StARD9* is the major transcript and the alternatively spliced transcript is only a minor transcript, which might be prone to nonsense mediated mRNA decay.

Expression profiles of human StARD9. The investigation of available microarray experiments and the results of the 'virtual Northern blot' showed a predominant expression of *StARD9* in the central nervous system, muscle cells (heart and skeletal muscle), pancreatic and prostate tissue and in cells of the lung.

Human StARD9 (syn. KIAA1300) protein. The identified full-length mRNA codes for a 1820-aa long protein. It contains a StART (StAR-related lipid-transfer) lipid-binding domain at the amino-acid positions 1721-1813. Further analysis with InterProScan (http://www.ebi.ac.uk/InterProScan/) did not

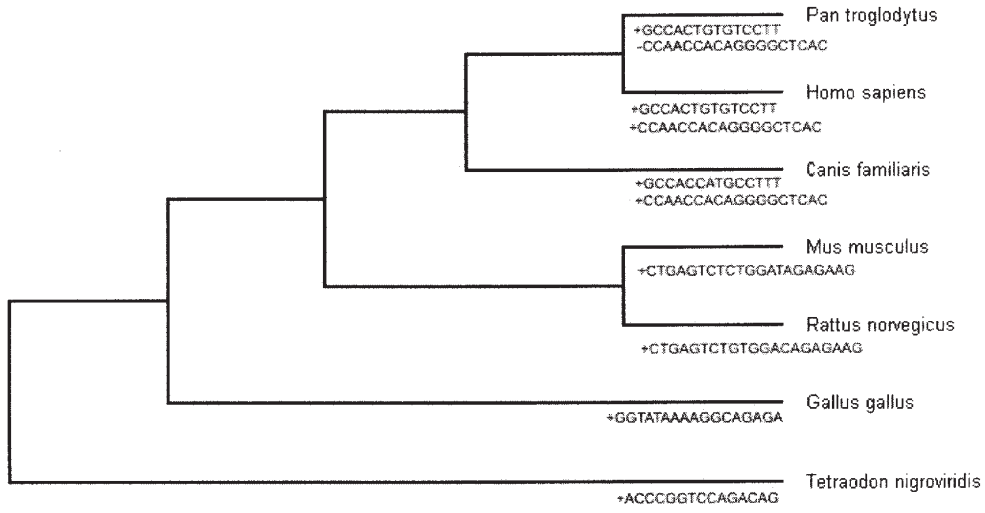


Figure 2. Identogram showing examples of unique additional (+) or unique absent (-, sequence is present in all other species) sequences, dendrogram only for structural purposes.

Figure 3. Single nucleotide polymorphisms of *StARD9* with functional effects. Lower sequence, disrupted ESE; upper sequence, ESE sequence; italic and bold, ESE-hexamer(s). (A) rs17346082 (no allele frequencies were obtainable). (B) rs9269. (C) rs1995939.

reveal any more known domains. The full protein sequence of the full-length transcript and of the alternatively spliced transcript can be found at http://www.halama.org/protein.htm.

Comparative genomics on StARD9 orthologs. Data mining revealed putative orthologs of human *StARD9* in chimpanzee, mouse, rat and chicken. Human *StARD9* showed 98% amino-acid identity with the chimpanzee (XP_510339.1), 60% with the mouse protein (AAH32885), 60% with the rat protein (XP_230493) and 60% with the chicken protein (XP_421162). Interestingly the more distant the relation of the orthologs are the more different are the numbers of exons and the numbers of existing alternatively spliced transcripts. Fig. 1D shows the phylogenetic tree of the (predicted) *StARD9* homologs. Analysis of the protein structure in the different species revealed a high conservation of the StART domain including the region between amino-acids 1390 and 1813. The full multiple sequence alignment of different species can be found at http://www.halama.org/seqalignment.htm. Among the species analyzed different unique sequence features were found. Sample sequence features are shown in Fig. 2.

SNP analysis. From the data of 15 available SNPs (rs8031218, rs11857283, rs3742995, rs3742994, rs17346082, rs3742993, rs3742992, rs16957061, rs16957063, rs938046, rs1058846, rs3199486, rs9269, rs1995939 and rs8028863) in the genomic region of *StARD9* we identified three as functionally relevant, i.e. one of the available alleles disrupted an existing exonic splicing enhancer. Sequences and available allele frequencies are shown in Fig. 3.

Discussion

The human *StARD9* gene, consisting of 11 exons is located at human chromosome 15q15. We characterized the human *StARD9* gene in this study. To reveal the genomic structure of human *StARD9* we identified the complete coding sequence. We observed a START lipid-binding domain near the 3'-end. This domain belongs to the START domain family, which has a broad spectrum of functions. Among these are signaling functions and the acute regulation of steroidogenesis (with or without hormonal regulation). The reference protein for this domain family is MLN64. The crystal structure of the START domain of human MLN64 shows an α/ß fold built around a U-shaped incomplete ß-barrel. Most importantly, the interior of the protein encompasses a hydrophobic tunnel that binds a single cholesterol molecule. The START domain structure shows a similarity to the birch pollen allergen Bet v1 and to bacterial polyketide cyclases/aromatases (8).

The possibility of hormonal regulative functions is interesting in the context of adjacent structures to *StARD9*: the locus for LGMD2A, an autosomal recessive form of limb-girdle muscular dystrophy. The course of this disease is more favorable for females than males and a wide variety of underlying genotypes is documented. One hypothesis could be a

possible regulatory role of *StARD9* due to different hormonal situations in both sexes. The effects of lutenizing hormone (LH) on StART-domain containing genes are already known (8,9). The rest of the protein has an unknown structure. Comparison with orthologs in different species revealed a high conservation of the StART-domain in all species. Comparing the genomic sequence we found 10-20 bp inserts only in human, chimp and dog genomes, not in other genomes. On the other hand we found large stretches of different genomic sequences in distantly related genomes (*Tetraodon negroviridis* and *Gallus gallus*), which are not found in mammals. These facts indicate that multiple rearrangements occurred during evolution, resulting in a (partially) highly conserved and specialized gene. In the human *StARD9* gene we observed two splicing variants, the major transcript includes all exons and the minor transcript lacks exon 8 (Fig. 1C). Additionally we provided information on three functionally relevant SNPs. Expression profiling showed an increased expression of *StARD9* in muscle tissue and tissue of the nervous system. This makes *StARD9* an interesting candidate for future investigations of possible modifying genes for LGMD2A. This is the first report on the identification and characterization of the *StARD9* gene and it provides useful data for further experimental studies.

References

1. Ponting CP and Aravind L: START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. Trends Biochem Sci 24: 130-132, 1999.
2. Tsujishita Y and Hurley JH: Structure and lipid transport mechanism of a StAR-related domain. Nat Struct Biol 7: 408-414, 2000.
3. Chiannilkulchai N, Pasturaud P, Richard I, Auffray C and Beckmann JS: A primary expression map of the chromosome 15q15 region containing the recessive form of limb-girdle muscular dystrophy (LGMD2A) gene. Hum Mol Genet 4: 717-725, 1995.
4. Piluso G, Politano L, Aurino S, Fanin M, Ricci E, Ventriglia VM, Belsito A, Totaro A, Saccone V, Topaloglu H, Nascimbeni AC, Fulizio L, Broccolini A, Canki-Klain N, Comi LI, Nigro G, Angelini C and Nigro V: Extensive scanning of the calpain-3 gene broadens the spectrum of LGMD2A phenotypes. J Med Genet 42: 686-693, 2005.
5. Katoh M: Paradigm shift in gene-finding method: from benchtop approach to desk-top approach. Int J Mol Med 10: 677-682, 2002.
6. Huang J and Lee V: Identification and characterization of a novel human nephronectin gene *in silico*. Int J Mol Med 15: 719-724, 2005.
7. Yoshida K: Identification and characterization of human ZNF18 gene *in silico*. Int J Mol Med 15: 545-548, 2005.
8. Iyer LM, Koonin EV and Aravind L: Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. Proteins 43: 134-144, 2001.
9. Clark BJ, Wells J, King SR and Stocco DM: The purification, cloning, and expression of a novel luteinizing hormone-induced mitochondrial protein in MA-10 mouse Leydig tumor cells. Characterization of the steroidogenic acute regulatory protein (StAR). J Biol Chem 269: 28314-28322, 1994.