Incorporating Mouse Genome
Mammalian Genome
Genes and Phenotypes

New members of the neurexin superfamily: multiple rodent homologues of the human *CASPR5* gene

Walther Traut,[1] Dieter Weichenhan,[2] Heinz Himmelbauer,[3] Heinz Winking[1]

[1]Institut für Biologie, Zentrum für medizinische Struktur- und Zellbiologie, Universität zu Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany
[2]Innere Medizin III, Universitätsklinikum Heidelberg, Otto-Meyerhof-Zentrum, D-69120 Heidelberg, Germany
[3]Max-Planck-Institut für molekulare Genetik, D-14195 Berlin-Dahlem, Germany

Received: 15 November 2005 / Accepted: 21 February 2006

Abstract

Proteins of the Caspr family are involved in cell contacts and communication in the nervous system. We identified and, by *in silico* reconstruction, compiled three orthologues of the human *CASPR5* gene from the mouse genome, four from the rat genome, and one each from the chimpanzee, dog, opossum, and chicken genomes. Obviously, *Caspr5* gene duplications have taken place during evolution of the rodent lineage. In the rat, the four paralogues are located in one chromosome arm, Chr 13p. In the mouse, however, the three *Caspr5* genes are located in two chromosomes, Chr 1 and Chr 17. RT-PCR shows that all three mouse paralogues are being expressed. Common expression is found in brain tissue but different expression patterns are seen in other organs during fetal development and in the adult stage. Tissue specificity of expression has diverged during evolution of this young rodent gene family.

Introduction

Proteins of the Caspr family play essential roles in the correct development and proper functioning of the peripheral and central nervous system. They are multidomain transmembrane proteins that belong to a subgroup of the neurexin family that is involved in cell adhesion and intercellular communication (for review, see Poliak and Peles 2003). Caspr1 (contactin associated protein = Caspr = CNTNAP1) and Caspr2-4 (contactin-associated protein-like 2–4) are predominantly found in nerve tissue (Peles et al. 1997; Poliak et al. 1999; Spiegel et al. 2002). While Caspr1 is required for axoglial sealing at the paranodes, Caspr2 is found in the adjacent juxtaparanodal regions and is associated with K^+ channels (for review, see Poliak and Peles 2003). Caspr3 and Caspr4 have different distributions in the nervous system (Spiegel et al. 2002) but precise locations along the axons have not yet been worked out. Much less is known of Caspr5.

Full-length cDNAs of the human Caspr5 have been identified (Spiegel et al. 2002), and the gene is annotated in the genome databanks. Molecular mapping of a reciprocal mouse translocation with a lethal effect led us to a disrupted orthologue of the human *CASPR5* (= *CNTNAP5*) gene (D. Weichenhan, W. Traut, H. Himmelbauer, H. Winking, unpublished). It turned out that the mouse genome contained three different genes that are orthologues of the single human gene. Here we present *in silico* reconstruction of the genes, cytogenetic mapping of the three loci by fluorescence *in situ* hybridization (FISH), cDNA sequence data, an RT-PCR study of expression in different organs and developmental stages, and the phylogeny of Caspr5, including *in silico* reconstruction of the homologues from the rat and other species whose whole-genome sequences are available.

Material and methods

Terminology. In this article, for the sake of readability, we use the same gene name for orthologues of all species but give them a two-letter prefix to indicate the respective genus and species, e.g., *Md-Caspr5* for the opossum (*Monodelphis domestica*) *Caspr5*.

Correspondence to: Walther Traut; E-mail: traut@molbio.uni-luebeck.de

Table 1. Primers for PCR and sequencing

Gene	*Name*	*5′–3′ sequence*
MmCaspr5-1	Mm5-1F2	GCTCTCAGGCTTGTGGCATTTA
	Mm5-1B1	CCAGTTGTGCCCTGTGTCACTA
MmCaspr5-2	Mm5-2F2	GCTGTCTGGCTTGTGGCACGTT
	Mm5-2B1a	GTTGTGCCCTGTGTCGCTGAAT
MmCaspr5-3	Mm5-3F2	GCTCTCTGGCTTGTGGCATTTT
	Mm5-3B1	CCAGTTGTGCCCTGTGTCACTG
MmCaspr5-1, -2, -3	Ca5allF1	IRD700-GGATTAACAGCGACAAAYTACA
	Ca5allB1	IRD800-CAGTCAGAGCTYCCATATCTTC
Actin	ActF1	TCCTGACCCTGAAGTACCCC
	ActB1	CGTCAGGCAGCTCATAGCTC

Isolation of total RNA and cDNA preparation. We prepared cDNA from outbred NMRI mice. Tissues were dissected from adult animals and embryos at days 10 and 16 of pregnancy (plug day = day 0) and immediately stored in RNALater (Qiagen, Valencia, CA) at 4°C until further use. Total RNA was isolated by the Trizol™ (Invitrogen, Carlsbad, CA) method according to the instructions of the manufacturer. In brief, tissue samples were submerged in Trizol and homogenized mechanically with a mixer mill, model MM301 (Retsch, Haan, Germany). After extraction with chloroform, RNA was precipitated with 2-propanol and finally dissolved in 20–100 μl DEPC-treated H_2O, depending on the size of the tissue sample. Concentration and integrity of the RNAs were assessed by visual comparison with a known reference sample of total RNA on agarose gels. Depending on the RNA concentration, 3–10 μl of the RNA samples were reverse transcribed at 52°C with Superscript III as recommended by the manufacturer (Invitrogen). After the reaction, mixtures were diluted with 80 μl H_2O and stored at −20°C.

RT-PCR and sequencing. Gene-specific primers spanning exons 1-3 were designed by making use of the sequence differences among the three *Caspr5* mouse genes. These primers and, as a control, primers for actin were used in RT-PCR (listed in Table 1 and synthesized by MWG-Biotech, Ebersberg, Germany). The reaction mix of 10 μl contained 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, 2.5 mM of each dNTP, 0.3 μM of each primer, 0.1 U *Taq* polymerase, and 0.5 μl cDNA sample. PCR cycling conditions were an initial 2 min at 94°C denaturation, 40 cycles of 15 sec at 94°C, 15 sec at 60°C (for *Mm-Caspr5-1* and *-3*) or 64°C (for *Mm-Caspr5-2*), 2 min at 68°C, and a final extension of 5 min at 68°C. The PCR products were separated by gel electrophoresis in 3% agarose gels.

For sequencing, the respective bands were cut out from the gel and the DNA fragments recovered from the gel by centrifugation through blotting paper (Weichenhan 1991). Sequencing was done on the LI-COR DNA Sequencer 4200 (Bad Homburg, Germany) according to the multiplex sequencing protocol using the SequiTherm EXCEL sequencing kit LC (Epicenter Technologies, Madison, WI) with infrared-labeled primers Ca5allF1 and Ca5allB1 designed to match all three *Caspr5* mouse genes (Table 1).

BAC-FISH. As probes for cytogenetic mapping by FISH, we selected two bacterial artificial chromosomes (BACs) each for the three mouse genes from the Ensembl v31 (EBI and Sanger Institute) database using BLASTN searches, one with exon 1, the other with exon 2. The BAC probes chosen were RP23-66a20 and RP23-406i13 for *Mm-Caspr5-1*, RP23-202a19 and RP23-472i23 for *Mm-Caspr5-2*, and RP23-211h11 and RP23-198m15 for *Mm-Caspr5-3*. BACs were labeled by nick translation with biotin-dATP (BioNick Labeling System, Invitrogen, Karlsruhe, Germany) and used for *in situ* hybridization on mouse NMRI chromosomes. Biotin label was detected with Cy3-conjugated streptavidin (Jackson ImmunoResearch Laboratories, Soham, UK). Counterstaining of chromosomes was done with DAPI (4, 6-diamidino-2-phenylindole). Fluorescence images in black and white were taken with a cooled CCD camera through the Zeiss Pinkel filter set (to avoid offset), pseudocolored, and merged with Adobe Photoshop.

Alignments and* in silico *reconstructions. For optimal alignment of two sequences, we used BioEdit v7.0.1 (Hall 1999). Multiple alignments were done with ClustalW v1.82 at EBI. Phylogenetic reconstruction was conducted using MEGA2 (Kumar et al. 2001). Exon–intron structure of the human gene was inferred from comparison of the cDNA (GenBank accession Nos. AB077881 and AK056528) with the respective region of the genomic contig using ASSEMBLE v3.11, a program written in Java by

one of the authors (WT). Identical stretches in cDNA and genomic sequence identified exons. Exon–intron borders were defined (1) by the ends of identical sequence and (2) by the presence of canonical 5′- or 3′-splice signals adjacent to or within a few nucleotides of the identical sequence. A test for correctness was the precise continuation of the cDNA sequence across introns. All reconstructions passed this test.

The excellent sequence conservation of Caspr5 allowed *in silico* reconstruction of the open reading frames (ORFs) of unknown orthologous cDNAs from whole-genome sequences. TBLASTN searches with the amino acid sequence of the model, human Caspr5 isoform 1 (accession No. AB077881), exposed exon stretches. In some cases, this had to be complemented by FASTA searches (FASTA and TFASTX v3.4; Pearson and Lipman 1988; Pearson et al. 1997) in the respective downloaded genomic segments. Plausibility required that the stretches with highest similarity to parts of the query be located in the genome in the correct order and have the same reading polarity. These stretches were retrieved with some 50 nucleotides more at both ends. They were conceptually translated, and both nucleotide and amino acid sequence aligned to the model nucleotide and amino acid sequence using ASSEMBLE. Exon–intron borders were identified as described in the previous paragraph.

Domains and substitution rates. For signal peptide detection, we used the SignalP v3.0 server (http://www.cbs.dtu.dk; Bendtsen et al. 2004). The transmembrane domain and the inside–outside prediction was done with TMHMM (http://www.cbs.dtu.dk; Sonnhammer et al. 1998). For general domain prediction, we used ELM (http://elm.eu.org, Puntervoll et al. 2003).

Synonymous and nonsynonymous substitution rates (K_s, K_n) were calculated in pairwise comparisons with CODONML in the PAML v3.15 collection (http://abacus.gene.ucl.ac.uk/software/paml.html; Yang 1997). For site-specific evolutionary analysis, we used the sitewise likelihood-ratio (SLR) test of Massingham and Goldman (2005) and the CODONML program of the PAML package.

Sequence data

Homo sapiens. The present knowledge of *Hs-Caspr5* (= *CASPR5* = *CNTNAP5*, see Terminology subsection) is based on two cDNAs (accession Nos. AB077881 and AK056528) and seven expressed sequence tags (ESTs) (accession Nos. BX492462, BM668869, BM693542, AW896193, AA069426, AI199572, AL707802). The gene is contained in the Chr 2 supercontig NT_022135.14 and maps cytogenetically to 2q14.3 according to Entrez Gene. There are three obvious splice variants. In variant 1 (as in AB077881), the ORF is distributed among 24 exons, all with canonical splice sites, and stretches over a genomic region of 888,638 bp (from start to stop codon). The conceptual translation product is called Caspr5 isoform 1. This variant is used here as a reference and model for *in silico* reconstruction of orthologues. Variant 2 (as in AK056528) may not be a bona fide cDNA. It consists of 19 exons. Exons 1–17 were almost as in the reference sequence but with a consensus splice site three nucleotides further downstream in exon 7, plus parts of exons 18 and 24 without canonical splice recognition sites. Variant 3 (represented in BM693542 and BM668869) consists of three exons. Exons 1–2 are as in the reference cDNA, exon 3 is not spliced at its downstream donor splice recognition site and runs into the following intron with a polyadenylation site shortly in front of a poly(A)-tail.

Canis familiaris (dog). *Cf-Caspr5* was found in contig NW_139886.1 of Chr 19. The TBLASTN search detected 22 of the 24 exons. The missing exons 2 and 23 were found by TFASTX searches offline in the downloaded genomic segment. The 24 exons containing the ORF extend over a genomic stretch of 787,828 bp from start to stop codon (inclusive).

Gallus gallus (chicken). *Gg-Caspr5* was found in the Chr 7 contig NW_060410.1. The cDNA ORF is divided into 24 exons, as in the human sequence, and extends over a genomic stretch of 256,003 bp from start to stop codon (inclusive).

Monodelphis domestica (opossum). The *Md-Caspr5* orthologue was found with Ensembl. It is contained in scaffold AB_13358.1. The scaffold was downloaded and FASTA34 searches with single human exons were performed. Exons 15 and 17 have not been recognized, probably because of incomplete sequence data. The reconstructed cDNA ORF covers a genomic stretch of 891,918 bp.

Mus musculus (house mouse). In the genomic contigs, three complete genes, *Mm-Caspr5-1, –2, –3*, were found with equally strong similarity to the human *Hs-Caspr5*. *Mm-Caspr5-1, Mm-Caspr5-2*, and two incomplete copies were contained in contig NT_078297.3. *Mm-Caspr5-1* covers a genomic stretch of 896,815 bp (positions 22001911–22898725), which maps to position 1E2.1 at 115.7–116.6 Mb according to GenBank Map Viewer. We sequenced a

partial cDNA that runs from exon 1 to exon 3 (accession No. AM076973). *Mm-Caspr5-2* covers 715,690 bp (positions 5851423–6566247, but note that 865 bp are missing in this section of NT_078297.3), which maps to position 1E1.1 at 99.6–100.3 Mb according to GenBank. Exon 21 is not represented in NT_078297.3. It was found instead in the correct context in a partial cDNA (accession No. AK133394) and in the genomic Celera contig GA_x6K02T2R7CC and included in the reconstructed sequence. Our partial *Mm-Caspr5-2* cDNA runs from exon 1 to exon 3 (accession No. AM076974). The two incomplete versions in contig NT_078297.3 were at positions from ≈8240555 to ≈8658407 and from ≈8904324 to ≈9590825. Only 9 and 10 of the complete set of 24 exons were recovered, 5 and 3 of them, respectively, displayed frameshifts and/or in-frame stops. According to NCBI Map Viewer, the incomplete versions map to 1E1.1 at 102.0–102.4 Mb and 1E1.2 at 102.7-103.3 Mb. *MmCaspr5-3* is contained in contig NT_039657.3 and covers 640,286 bp (positions 1238586–1878871), which maps to Chr 17D at 55.5–56.1 Mb according to GenBank MapViewer. Our partial cDNA runs from exon 1 to exon 3 (accession No. AM076975).

Pan troglodytes (chimpanzee). *Pt-Caspr5* was reconstructed from contig NW_104188.1 of Chr 2B. Only exon 17 was missing in the genomic sequence. The gene covers 875,021 bp of genomic DNA (positions 9676720-10551740).

Rattus norvegicus (rat). Four *Caspr5* genes, *Rn-Caspr5-1, –2, –3,* and –4, have been reconstructed and two more incomplete versions were found in their neighborhood. From the two incomplete versions, 6 and 10 exons were discovered, respectively, with 5 and 4 of them with frameshifts and/or in-frame stops. *Rn-Caspr5-1* and *Rn-Caspr5-4* are contained in contig NW_047391.2 at positions 9428368–10449076 and 3574928–4632210, repectively. These positions map to chromosome band 13p11 and 13p11-12 according to NCBI Map Viewer. *Rn-Caspr5-2* and *Rn-Caspr5-3* are contained in contig NW_047390.2 at positions 1192186–2293551 and 7714567–8649647, respectively. According to Map Viewer, these sequences map to chromosome band 13p13.

Accession numbers. The accession numbers of partial cDNAs were AM076973 for mouse *Caspr5-1*, AM076974 for *Caspr5-2*, and AM076975 for *Caspr5-3*. The data of the *in silico* reconstructed genes have been submitted as third-party annotations (*Cf-Caspr5*: BN000917; *Gg-Caspr5*: BN000918; *Md-Caspr5*: BN000919, BN000920, BN000921; *Mm-Caspr5-1*: BN000865; *Mm-Caspr5-2*: BN000866; *Mm-Caspr5-3*: BN000867; *Pt-Caspr5*: BN000915, BN000916; *Rn-Caspr5-1*: BN000868; *Rn-Caspr5-2*: BN000869; *Rn-Caspr5-3*: BN000870; *Rn-Caspr5-4*: BN000871).

Results

Three mouse* Caspr5 *genes. In a TBLASTN search with the human Caspr5 isoform1 sequence against the mouse reference genome, annotated and non-annotated exons from three different genomic regions were returned as best hits. The sequences were well conserved between human and mouse. *In silico* reconstruction on the basis of the human cDNA sequence yielded the complete coding regions of three *Caspr5* genes and two incomplete, obviously defective copies with only a few recoverable exons left. The same three complete genes and two defective copies were found in the public reference genome derived from mouse strain C57BL/6J and in the Celera assembly, which was derived fom various strains. The three complete genes are termed *Mm-Caspr5-1*, *Mm-Caspr5-2*, and *Mm-Caspr5-3* here.

They are rather large genes. The coding region of each of these genes is distributed among 24 exons and extends over more than 640 kb of genomic length. Some of the introns are longer than 100 kb. The three genes share 83%–84% identity with the human sequence on the nucleotide level and 82%–85% on the amino acid level. Among each other, they have 89%–90% identity on the nucleotide level and 86%–88% on the amino acid level. A specific feature of *Mm-Caspr5-2* is the truncated exon 14; 39 nucleotides are deleted at the 3′ end of the exon, leaving the reading frame and the splice site intact.

Specific primer sets for each of the three genes were designed from differential regions of exons 1 and 3. RT-PCR with these primers on cDNA from adult brain yielded fragments of the expected size (307, 310, and 307 bp for *Caspr5-1*, –2, and –3, repectively) from all three genes (Fig. 1). The fragments were recovered and sequenced. The partial cDNA sequences (deposited under accession Nos. AM076973, AM076974, and AM076975) were identical with those expected from the reconstructed genes. This proves that all three genes are being transcribed in the mouse.

For cytogenetic mapping of the mouse genes, two BACs were selected for each of the three genes according to databank data, one including exon 1, the other one exon 2. Using FISH with the BACs as probes (BAC-FISH), we mapped *Caspr5-1* to Chr

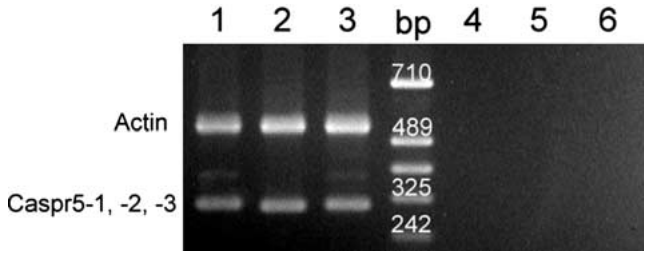

Fig. 1. RT-PCR with primers specific for *MmCaspr5-1* (lanes 1, 4), *MmCaspr5-2* (lanes 2, 5), *MmCaspr5-3* (lanes 3, 6), and *Actin*. Lanes 1–3: adult brain cDNA, lanes 4–6: blanks.

1E2.3, *Caspr5-2* to Chr 1E2.1, and *Caspr5-3* to Chr 17E1.1 (Fig. 2). This confirms the origin of the probes from Chrs 1 and 17. The specific band locations, however, appear shifted relative to the positions predicted by EBI Ensembl Cytoview or NCBI GenBank Map Viewer. Presumably, the chromosomal anchor points for the large contigs are far away from the *Caspr5* sites in the data sets used to construct the specific regions in Cytoview or Map Viewer.

Domain composition of mouse Caspr5 proteins. The conceptually translated ORFs of *Mm-Caspr5-1*, –2, and –3 present precursor proteins with an N-terminal signal peptide and a domain structure that is identical among the three mouse proteins and the human Caspr5 protein (Fig. 3) and similar to the human Caspr1–4 proteins (Spiegel et al. 2002). A transmembrane domain divides the proteins into long N-terminal extracellular and short C-terminal cytoplasmic regions. The outer moiety contains a discoidin (FA58C) domain which is thought to mediate intercellular contacts, four laminin G (LAMG) domains, and two epidermal growth factor-like (EGF) domains. The fibrinogen-related domain that has been considered present in Caspr1–4 (Spiegel et al. 2002) may also be present in Caspr5 (marked FBG? in Fig. 3). ELM predicts it for human and mouse Caspr1 and Caspr2 but not for the remaining members of the family, although there is good sequence conservation in the respective region in all members of the Caspr family (not shown). In Mm-Caspr5-2, the partial deletion of exon 14 abolishes a part (13 amino acids) of this domain. On the cytoplasmic side, a juxtamembrane domain that binds protein 4.1 homologues had been found in Caspr1 and 2 but not in Caspr3 and 4 (Spiegel et al. 2002). It is also absent in the three Mm-Caspr5 proteins.

Expression pattern. For an overview on the activity of the three genes, we checked various organs and body parts of adults and fetal mice by RT-PCR (Table 2). Like other members of the Caspr family, *Mm-Caspr5-1*, *Mm-Caspr5-2*, and *Mm-Caspr5-3* are expressed in adult brains. They are also active in fetal heads at day 10 and brains at day 16. In some other tissues, however, the transcriptional patterns differ. *Mm-Caspr5-1* and *Mm-Caspr5-2* activity was found in the rumps of day-10 fetuses but *Mm-Caspr5-3* activity was not. Lungs of day-16 fetuses showed activity of *Mm-Caspr5-1* and *Mm-Caspr5-2* but not of *Mm-Caspr5-3*, while the reverse was true for lungs of adults. Thus, there is some differential activity of the three genes. Inconsistent results of repeat experiments in some tissues (indicated by ± in Table 2) may have been caused by very low levels of activity or the presence of only a few cells with activity in the respective tissue or organ.

Caspr5 *genes of other mammals and chicken.* To investigate the phylogenetic relationship of the three mouse genes, we reconstructed *Caspr5* orthologues from some more species. Like with the mouse, we used human Caspr5 isoform 1 for TBLASTN searches in publicly available whole genomes and as a

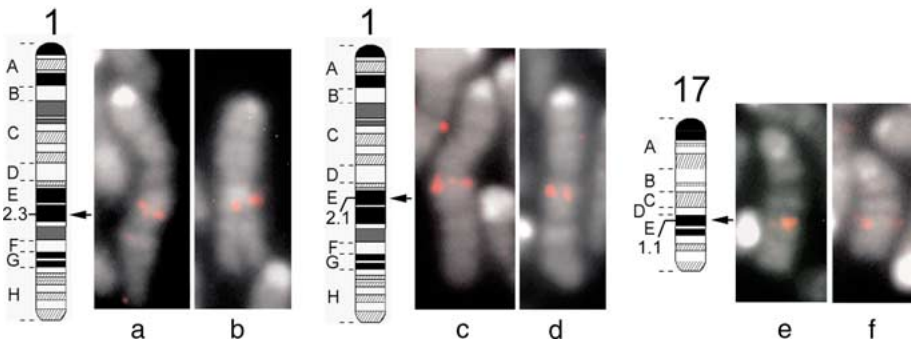

Fig. 2. BAC-FISH on mouse chromosomes with *Mm-Caspr5-1* (**a, b**), *Mm-Caspr5-2* (**c, d**) and *Mm-Caspr5-3* (**e, f**) probes. Partial karyotypes compared with standard Chromosomes 1 and 17. The biotinylated probes were detected with Cy3-conjugated streptavidin (red signals), chromosomes were stained with DAPI. The probes were RP23-66A20 (**a**), RP23-406I13 (**b**), RP23-202A19 (**c**), RP23-472I23 (**d**), RP23-211H11 (**e**), RP23-198M15 (**f**).

Fig. 3. Domain composition of Mm-Caspr5-1, –2, and –3 precursor proteins. FA58C = coagulation factor 5/8 C-terminal domain, discoidin domain, LAMG = laminin G domain, EGF = epidermal growth factor-like domain, FBG = fibrinogen-related domain; TM = transmembrane helix.

model for *in silico* reconstruction. In this way we retrieved one orthologue each from the chicken (*Gg-Caspr5*), the opossum (*Md-Caspr5*), the dog (*Cf-Caspr5*), and the chimpanzee (*Pt-Caspr5*) but four from the rat (*Rn-Caspr5-1, –2, –3, –4*), not including two defective copies in the rat.

Sequence conservation is rather high in Caspr5, 76% identity between human and chicken on the nucleotide level and 79% on the amino acid level. Even more striking is the conservation of the nucleotide sequence in specific exons, e.g., exon 10 with 98% identity between opossum and human. One of the four rat genes, *Rn-Caspr5-2*, had a truncated exon 14, precisely like that of *Mm-Caspr5-2* from the mouse.

The maximum parsimony tree (Fig. 4) constructed from the sequence alignment shows the sequences to be arranged according to the phylogeny of the species with the exception of the rodent sequences. Mouse and rat sequences are not on different branches but form a single cluster. Each of the three mouse sequences has one or two rat partners as closest relation. Mm-Caspr5-1 is most closely related to Rn-Caspr5-1 and Rn-Capr5-4, Mm-Caspr5-2 is on a separate branch with Rn-Caspr5-2, while Mm-Caspr5-3 is closest to Rn-Caspr5-3. Critical nodes at the bases of the rodent branch, the Mm-Caspr5-2 / Rn-Caspr5-2 and RnCaspr5-1 / Rn-Caspr5-4 branches are supported by high bootstrap values. The same tree topology was found in neighbor-joining (NJ) trees and when cDNA sequences were used (not shown). When the amino acid sequences were truncated to exons 1–14 to avoid biases and errors introduced by the truncated exon 14 in Mm-Caspr5-2 and Rn-Caspr5-2, and by the missing exon 17 in the chimpanzee and exons 15 and 17 in the opossum, Rn-Caspr5-4 shifted closer to the Mm-Caspr5-2 plus Rn-Caspr5-2 branch but otherwise the topology remained the same (not shown). The trees reveal that *Caspr5* genes have undergone duplications in the ancestral rodent genome before separation of the mouse and rat lineages; the one creating *Rn-Caspr5-1* and *Rn-Caspr5-4* has happened in the rat lineage after separation.

An analysis of the nonsynonymous/synonymous substitution rate ratio (K_n/K_s) shows all *Caspr5* sequences under purifying selection (K_n/K_s from 0.10 to 0.36 in pairwise comparisons). Selection appears to have acted stringently on *Caspr5* in species with

Table 2. Expression of Mm-Caspr5-1, -2, and -3 in various tissues detected by RT-PCR

		Mm-Caspr5-1	*Mm-Caspr5-2*	*Mm-Caspr5-3*
Day-10 fetus	head	+	+	+
	rump	+	+	–
	Yolk sac	–	–	±
	placenta	±	–	–
Day-6 fetus	brain	++	++	++
	Legs	+	+	±
	heart	–	–	–
	Liver	–	–	–
	kidney	+	+	+
	Lung	+	±	–
	yolk sac	–	–	–
	placenta	–	–	–
Adult	brain	++	++	++
	sk. muscle	–	–	–
	heart	–	–	–
	liver	–	–	–
	kidney	–	–	–
	lung	–	–	+
	spleen	±	–	–
	uterus	–	–	±
	ovary	±	–	–
	testis	–	±	–

Each RT-PCR was repeated at least twice on cDNAs prepared separately from two adults, four day-16 embryos, and two samples of ten day-10 embryos each. ++ strong signal, + weak signal, ± signal present in some samples, – no signal.

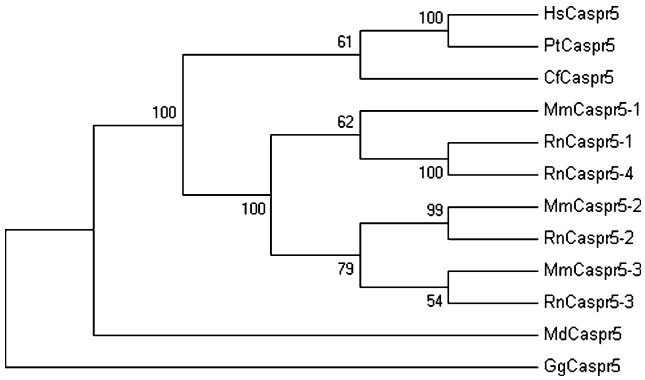

Fig. 4. Phylogeny of Caspr5. The maximum parsimony tree was constructed from the human (Hs–), chimpanzee (Pt–), dog (Cf–), rat (Rn–), mouse (Mm–), opossum (Md–), and chicken (Gg–) amino acid sequences, aligned by Clustal W. The chicken sequence was used as an outgroup. Bootstrap values are listed at nodes.

Table 3. Genomic length (coding start to coding stop) and cytogenetic position of Caspr5 genes from mammals and chicken

Species	*Gene*	*Genomic length (bp)*	*Cytogenetic location*
Mouse	*Mm-Caspr5-1*	896,815	1E2.3[a]
Mouse	*Mm-Caspr5-2*	715,690	1E2.1[a]
Mouse	*Mm-Caspr5-3*	640,286	17E1.1[a]
Rat	*Rn-Caspr5-1*	1,020,709	13p11[b]
Rat	*Rn-Caspr5-2*	1,101,366	13p13[b]
Rat	*Rn-Caspr5-3*	935,081	13p13[b]
Rat	*Rn-Caspr5-4*	1,057,283	13p11[b]
Human	*Hs-Caspr5*	888,638	2q14.3[b]
Chimpanzee	*Pt-Caspr5*	>875,021	2B[b]
Dog	*Cf-Caspr5*	787,828	19[b]
Opossum	*Md-Caspr5*	>891,918	n.d.
Chicken	*Gg-Caspr5*	256,003	7[b]

[a]FISH data, this article.
[b]Data from NCBI MapViewer.

only one gene (K_n/K_s from 0.10 to 0.11 in human/dog, human/opossum, dog/opossum) and in a more relaxed fashion on the multiple *Caspr5* genes in rodents (K_n/K_s from 0.14 to 0.19 in rodent versus human, dog, opossum comparisons; K_n/K_s from 0.25 to 0.36 in mouse/mouse and rat/rat comparisons). Since the K_n/K_s shift may also have been caused by some sites being under positive selection, we tested the alignment for such sites. We detected many sites under purifying selection but only one candidate for positive selection (position 256) above the 95% probability threshold using the SLR method of Massingham and Goldman (2005) and none with statistical significance using the CODONML program of the PAML package (Yang 1997).

Discussion

Impediments for automatic annotation. Large size is an inherent problem for automated gene prediction (Wang et al. 2003). It is caused by the high number of exons and the presence of excessively large introns. In the case of the rodent *Caspr5* genes (Table 3), besides the presence of 24 exons and several excessively large introns (>100 kb), there were additional problems. No full-sized cDNA was available except from the human. Thus, reconstruction had to rely on homology. This was particularly hampering as several orthologues of the human *CASPR5* gene were present in the mouse and rat genome. Finally, sequencing of the mouse reference genome was not complete, and gaps in the assembled sequence are more likely to hit large genes like *Caspr5* than small genes. In the public genome databases, one of the 24 exons, #21, of *Mm-Caspr5-2* was not represented, while exons 5 and 6 were missing in the Celera mouse genomic sequence, although otherwise the same three complete genes and two incomplete genes were found in both assemblies. Thus, ''manual'' compilation has been a necessity for these genes.

Evolution of paralogues. We show here that mouse and rat genomes contain three and four paralogues, respectively, that are orthologues of the single *Caspr5* gene present in nonrodent mammals and chicken. Besides complete genes, two more yet very incomplete copies have been detected in the rat and mouse genomes. At least three rounds of duplication of the ancestral gene must have taken place in the rodent lineage to generate the actually present complete and incomplete copies of the gene. The association of each of the three complete mouse homologues with one or two rat counterparts in the maximum parsimony tree (Fig. 4) suggests that they originated from duplication events in common rat–mouse ancestors. Convincing additional evidence comes from exon 14, which is truncated at identical positions in *Mm-Caspr5-2* and *Rn-Caspr5-2* while all other copies have the ancestral form of exon 14. Hence, an amplification of the Caspr5 family has taken place after separation of Primates from Rodentia but before separation of the mouse and rat lineages. One duplication, the *Rn-Caspr5-1* / *Rn-Caspr5-4* duplication, must have happened later, in the rat lineage after separation from the mouse lineage.

The primate–rodent split is estimated to have occurred between 85 and 87 MYA (Springer et al. 2003), while the mouse–rat split is dated between 14 and 24 MYA by molecular and fossil evidence (Adkins et al. 2001; Jacobs and Pilbeam 1980; Springer

et al. 2003). Thus, the rodent *Caspr5* family is a rather young gene family.

In the rat, the complete and incomplete *Caspr5* versions are contained within a 29-Mb region on the small arm of Chr 13 (Table 3). Hence, the amplification process has probably taken place *in situ*. In the mouse, all but one of the *Caspr5* genes are harbored in a 17-Mb cluster on Chr 1. The region is part of the conserved synteny block between Chr 1 of the mouse and Chr 13 of the rat, as shown by reciprocal chromosome painting (Stanyon et al. 1999) and displayed in the mouse-rat orthology map at MGI (http://www.informatics.jax.org). One of the mouse paralogues (*Mm-Caspr5-3*), however, has left the cluster and moved to Chr 17. The jump of *Mm-Caspr5-3* must have happened after separation of the mouse and rat branches. No other synteny relationship is evident between this chromosome and rat Chr 13. The mechanism of the jump is not clear. It was not a retrotransposition event because the complete gene with introns, a segment of more than 640 kb, has moved to mouse Chr 17. But it was not a singular event. Several other cases of single genes moving into a new neighborhood are documented in the orthology map.

The rodent *Caspr5* family is not the only rodent-specific gene family. The Mouse Genome Sequencing Consortium (2002) lists 25 families that have expanded in the rodent lineage and are, in most cases, represented by only one orthologue in humans. An even younger gene family is the *Sp100-rs* cluster in mouse Chr 1. It contains 50–2000 gene copies, depending on the source of the chromosome, and is confined to *Mus musculus* and its closest relations (Traut et al. 2001; Weichenhan et al. 2001).

Diverging functions. Three main alternative routes are generally considered as evolutionary outcomes of gene duplications: (1) nonfunctionalization, the silencing of one copy by deleterious mutations; (2) neofunctionalization, the acquisition of a novel function by one copy while the other serves the former function; and (3) subfunctionalization, the distribution of subsets of the former functions to both copies (Lynch and Conery 2000). In the rodent *Capr5* family, we see products of several rounds of duplication. Among them are degenerating copies with frameshifts and in-frame stops and only remnants of the former coding sequence left over. Other copies are active and under purifying, though somewhat relaxed, selection. We detected a shift from stringent selection ($K_n/K_s \approx 0.1$) in single-copy *Caspr5* genes of nonrodents to somewhat relaxed selection pressure ($K_n/K_s \approx 0.3$) in the multiple copies of rodents. This is in line with observations on large though unrelated samples of orthologue and paralogue comparisons in bacteria and mammals (Kondrashov et al. 2002).

The function of Caspr5 is not known yet nor are the sites of expression in species with only one gene, e.g., human. Thus, we cannot distinguish between neofunctionalization and subfunctionalization. We found expression of all three mouse genes in brain tissue, a common expression site also for other members of the Caspr family. Nevertheless, we found differences of expression in fetal and adult organs other than brain among the three genes. The divergent expression patterns indicate that at least some functional divergence has taken place during evolution. Direct support for functional differentiation of the Caspr5 family comes from a mouse chromosome translocation that disrupts *MmCaspr5-2*. Homozygotes of the translocation are embryonic lethals. The missing gene product of *Mm-Caspr5-2* is obviously not substituted sufficiently by those of *Mm-Caspr5-1* and *Mm-Caspr5-3* or else their products are not available in the right tissue or at the right stage of development (D. Weichenhan, W. Traut, H. Himmelbauer, H. Winking, unpublished).

The problems in identifying *Caspr5* in rodents have presumably inhibited research on this gene and its proteins in the commonly used model species mouse and rat. Identification of the multiple genes now provides the means to investigate them with presently available tools, e.g., knockouts and RNAi, and to precision map the proteins relative to each other and to other Caspr proteins in the nervous system. It will be interesting to learn how the relatively young family of rodent Caspr5 proteins have acquired new functions or partitioned among them the function or functions served by a single protein in other mammals.

Acknowledgments

The authors thank Enno Hartmann (Lübeck, Germany) for helpful discussions and two anonymous reviewers for useful comments. Tim Massingham (Hinxton, UK) kindly provided the SLR program. The technical assistance of Constanze Reuter, Ella Manthey, and Heidemarie Riechers (Lübeck, Germany) is highly appreciated.

References

1. Adkins R, Gelke E, Rowe D, Honeycutt R (2001) Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. Mol Biol Evol 18, 777–791

2. Bendtsen J, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: Signal P 3.0. J Mol Biol 340, 783–795
3. Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nuclei Acids Symp Ser 41, 95–98
4. Jacobs L, Pilbeam D (1980) Of mice and men: fossil-based divergence dates and molecular 'clocks'. J Hum Evol 9, 551–555
5. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:research 0008.1–0008.9
6. Kumar S, Tamura K, Jakobsen I, Nei M (2001) MEGA: Molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245
7. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155
8. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169, 1753–1762
9. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562
10. Pearson W, Lipman D (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85, 2444–2448
11. Pearson W, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. Genomics 46, 24–36
12. Peles E, Nativ M, Lustig M, Grumet M, Schilling J, et al. (1997) Identification of a novel contactin-associated transmembrane receptor with multiple domains implicated in protein-protein interactions. EMBO J 16, 978–988
13. Poliak S, Peles E (2003) The local differentiation of the myelinated axons at nodes of Ranvier. Nat Rev Neurosci 4, 968–980
14. Poliak S, Gollan L, Martinez R, Custer A, Einheber S, et al. (1999) Caspr2, a new member of the neurexin superfamily, is localized at the juxtaparanodes of myelinated axons and associates with K+ channels. Neuron 24, 1037–1047
15. Puntervoll P, Linding R, Gemünd C, Mattingsdale M, Costantini A, et al. (2003) ELM server, a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31, 3625–3630
16. Sonnhammer E, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, J. Glasgow et al. eds. (Menlo Park, CA: AAAI Press), pp 175–182
17. Spiegel I, Salomon D, Erne B, Schaeren-Wilmers N, Peles E (2002) Caspr3 and Caspr4, two novel members of the Caspr family are expressed in the nervous system and interact with PDZ domains. Mol Cell Neurosci 20, 283–297
18. Springer M, Murphy W, Eizirik E, O'Brien S (2003) Placental mammal diversification and the Cretaceous–Tertiary boundary. Proc Natl Acad Sci USA 100, 1056–1061
19. Stanyon R, Yang F, Cavagna P, O'Brien P, Bagga M, et al. (1999) Reciprocal chromosome painting shows that genomic arrangements between rat and mouse proceeds ten times faster than between humans and cat. Cytogenet Cell Genet 84, 150–155
20. Traut W, Rahn IM, Winking H, Kunze B, Weichenhan D (2001) Evolution of a 6–200 Mb long-range repeat cluster in the genus Mus. Chromosoma 110, 247–252
21. Wang J, Li S, Zhang Y, Zheng H, Xu Z, et al. (2003) Vertebrate gene predictions and the problem of large genes. Nat Rev Genet 4, 741–749
22. Weichenhan D (1991) Fast recovery of DNA from agarose gels by centrifugation through blotting paper. Trends Genet 7, 109
23. Weichenhan D, Kunze B, Winking H, van Geel M, Osoegawa K, et al. (2001) Source and component genes of a 6–200 Mb gene cluster in the house mouse. Mamm Genome 12, 590–594
24. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13, 555–556