CMLS, Cell. Mol. Life Sci. 61 (2004) 1157–1166
1420-682X/04/101157-10
DOI 10.1007/s00018-004-3470-z

CMLS Cellular and Molecular Life Sciences

Research Article

Phylogenetic origin of LI-cadherin revealed by protein and gene structure analysis

R. Jung[a,‡], M. W. Wendeler[a], M. Danevad[a], H. Himmelbauer[b] and R. Geßner[a,*]

[a] Institute of Laboratory Medicine and Biochemistry, Virchow-Hospital of Charité Medical School,
Humboldt University of Berlin, Augustenburger Platz 1, 13353 Berlin (Germany), Fax: +49 30 450 569907,
e-mail: gessner@charite.de
[b] Max-Planck-Institute of Molecular Genetics, Ihnestr. 73, 14195 Berlin (Germany)

Received 22 December 2003; received after revision 9 February 2004; accepted 27 February 2004

Abstract. The intestine specific LI-cadherin differs in its overall structure from classical and desmosomal cadherins by the presence of seven instead of five cadherin repeats and a short cytoplasmic domain. Despite the low sequence similarity, a comparative protein structure analysis revealed that LI-cadherin may have originated from a five-repeat predecessor cadherin by a duplication of the first two aminoterminal repeats. To test this hypothesis, we cloned the murine LI-cadherin gene and compared its structure to that of other cadherins. The intron-exon organization, including the intron positions and phases, is perfectly conserved between repeats 3–7 of LI-cadherin and 1–5 of classical cadherins. Moreover, the genomic structure of the repeats 1–2 and 3–4 is identical for LI-cadherin and highly similar to that of the repeats 1–2 of classical cadherins. These findings strengthen our assumption that LI-cadherin originated from an ancestral cadherin with five domains by a partial gene duplication event.

Key words. LI-cadherin; cadherin evolution; partial gene duplication; cadherin repeat; phylogeny; gene structure.

Cadherins are a heterogenic superfamily of transmembrane glycoproteins which mediate Ca^{2+}-dependent cell-cell adhesion [1, 2]. They play an important role during tissue development and are critical for the maintenance of junctional complexes between epithelial cells [3, 4]. The common feature of all cadherins is a variable number of 4 to 34 cadherin repeats comprising their ectodomain [5]. The cadherin repeat consists of about 110 amino acids arranged in seven β-strands and two short α-helices forming a β-barrel structure that resembles the topology of Ig domains [6, 7].

The number of cadherins identified in various organisms increases steadily. So far, 18 genes containing cadherin repeats have been found in the *Caenorhabditis elegans* genome and more than 70 genes have been identified in the human genome [8, 9]. Classical-type cadherins are the most extensively studied members of this protein family. They contain five extracellular cadherin repeats, a single transmembrane domain and a cytoplasmic domain of about 160 amino acids [10]. This intracellular domain is highly conserved among classical cadherins [11] and associates with a group of cytoplasmic proteins, termed catenins [12]. Catenins link classical cadherins to the actin cytoskeleton and regulate their adhesive properties [13]. Another well-defined and widely expressed cadherin subfamily, the desmosomal cadherins, share with the classical cadherins a highly similar ectodomain also comprising five cadherin repeats, but contain a distinct cytoplasmic domain that indirectly links them to the intermediate filament network [14].

[*] Corresponding author.
[‡] Present address: Schering AG, Müllerstr. 178, 13342 Berlin (Germany)

LI-cadherin represents a novel type of cadherin with a distinct structure compared to other members of the cadherin superfamily [15]. In contrast to classical cadherins, LI-cadherin consists of seven extracellular cadherin repeats and a rather short cytoplasmic domain comprising only about 25 amino acids. This cytoplasmic domain shares no homologies to that of classical cadherins and does not bind to *β*-catenin [16]. Nevertheless, LI-cadherin is able to mediate Ca^{2+}-dependent cell-cell adhesion independent of cytoplasmic interactions [16]. Interestingly, LI-cadherin is coexpressed basolaterally with E-cadherin, the prototype of classical cadherins, in all epithelial cells of the intestine, but both proteins are localized in different membrane regions [17]. While E-cadherin is concentrated in adherens junctions [18], LI-cadherin is evenly distributed along the lateral contact areas but it is excluded from adherens junctions and desmosomes [15]. LI-cadherin is not found in healthy gastric epithelia, but is highly expressed in intestinal metaplasia and adenocarcinomas of the stomach [19]. Due to its specific expression in certain types of cancer, LI-cadherin has been proposed as a differentiation marker for human gastric, pancreatic and hepatocellular carcinomas [20–22].

To reveal the origin of the two additional cadherin repeats of LI-cadherin compared to classical and desmosomal cadherins, we cloned the murine LI-cadherin gene, analyzed its structure and compared it with that of various classical cadherins. Combining these data with the results of our protein sequence analyses led us to conclude that LI- and E-cadherin originated from a common ancestor molecule with five cadherin repeats by a partial gene duplication event.

Materials and methods

Cloning of the murine LI-cadherin gene

A 129 SVJ mouse genomic lambda FIX II phage library (Stratagene, La Jolla, Calif.), prepared from liver genomic DNA, was screened with cDNA probes covering the total protein coding sequence of the rat LI-cadherin [15]. The probes were labeled with digoxigenin-11-dUTP (Roche, Mannheim, Germany) using random primers [23]. Hybond-N filters (Amersham Pharmacia, Freiburg, Germany) were hybridized at 38 °C for 14 h in 5 × standard saline citrate (SSC), 50% deionized formamide, 2% blocking reagent (Roche), 0.02% sodium dodecylsulfate (SDS) and 0.1% N-lauroylsarcosine and subsequently washed with 0.5 × SSC at 46 °C. Digoxigenin-labeled probes were detected with anti-digoxigenin Fab fragments conjugated to alkaline phosphatase and visualized with the chemiluminescent substrate CSPD (Roche). Clones of interest were plaque-purified by three cycles of rescreening. For phage DNA purifications, plate lysates were collected and subjected to Qiagen column chromatography (Qiagen, Hilden, Germany). Restriction fragments of the phages were characterized by Southern blot analysis and subcloned in the vector $pBlueScriptSK^+$ (Stratagene).

Additionally, a 129 mouse cosmid library (Resource Center of the German Human Genome Project, Berlin, Germany), cloned in Lawrist 7, was screened with a probe covering the nucleotides +50 to +800 of the murine LI-cadherin cDNA [17]. The probe was amplified and digoxigenin-11-dUTP-labeled with the forward primer: 5′-GTG GAT ATG GCG AAG AAG GGA AGT TCA GCG-3′ and the reverse primer: 5′-GGT CGA TCG AGA ATG GGA AC-3′ using the PCR DIG Probe Synthesis Kit (Roche). Filters were hybridized at 42.5 °C for 14 h in a buffer containing 5× SSC, 50% deionized formamide, 2% blocking reagent (Roche), 0.02% SDS and 0.1% N-lauroylsarcosine and washed subsequently in 0.5× SSC at 58 °C. Two cosmid clones were identified as described above and obtained from the Resource Center of the German Human Genome Project (clone MPMGc121I10150Q2 and clone MPMGc121M1327Q3). Cosmid DNA was purified on Qiagen columns and used directly for restriction typing, PCR analysis and DNA sequencing.

DNA sequencing

DNA sequences were determined by the dideoxy chain termination method [24] using fluorescent dye/Big-Dye terminators on 373A and 377 automated sequencers (Applied Biosystems, Darmstadt, Germany). Sequence alignment was performed using the MacMolly Tetra software package (version 3.7; Soft Gene, Berlin, Germany). The intron-exon boundaries were identified by sequence comparison with murine LI-cadherin cDNA [17] assuming conserved consensus sites.

Northern blot

Total RNA from murine tissues was isolated using TRIZOL reagent according to the manufacturer’s protocol (Gibco BRL, Grand Island, N. Y.). Approximately 20 µg of total RNA from each mouse tissue was separated on a 1.2% agarose gel, transferred to a nylon membrane (Hybond-N; Amersham-Pharmacia) and UV cross-linked (UV Stratalinker 1800; Stratagene). A digoxigenin-labeled antisense RNA in vitro transcript carrying 1400 bp of the 3′ end of rat LI-cadherin cDNA [15] was used as a probe at a concentration of 5 ng/ml. Hybridization was performed overnight at 50 °C in hybridization buffer containing 7% SDS, 50 mM sodium phosphate, 50% deionized formamide, 5% standard saline phosphate EDTA (SSPE), 0.1% N-laurylsarcosine and 2% blocking reagent (Roche). After hybridization, the membrane was washed at a final stringency of 0.1 × SSC/0.1% SDS at 68 °C. Detection of the digoxigenin-labeled nucleic acids was performed as described above.

Primer extension

Primer extension analysis was performed according to standard protocols [25]. The synthetic oligonucleotides ‘primer 1’ (5′-TGT CGT CCA TTC AGC CGT GGA GAC-3′) and ‘primer 2’ (5′-CAG TAA GTA AGA AAT GCT GC-3′) were end-labeled with [γ-^{32}P]ATP using a T4 polynucleotide kinase (MBI Fermentas, St. Leon-Rot, Germany) and purified by gel filtration chromatography on Centri-Sep columns (Princeton Separations, Adelphia, NJ, USA). Radiolabeled oligonucleotides (10^5 cpm) were hybridized to 100 µg of total RNA in reaction buffer (Gibco BRL, Karlsruhe, Germany). The extension reaction was performed with 200 U of SuperScriptII reverse transcriptase (Gibco BRL) for 30 min at 42 °C. The reaction products were analyzed on a 6% denaturing acrylamide gel.

Protein sequence analysis

Protein sequences were analyzed with the MacMolly Tetra software package (version 3.7; Soft Gene). The module Align was used to calculate the amino acid identity of cadherin repeats and full-length proteins with the following parameters: pairwise local alignment, PAM 250; gap penalties: opening (–5), extending (0). Matrix Plots were performed with the module Complign using the default gap and mismatch penalties. The minimal match length was set to 21 and the number of mismatches tolerated was set to 14.

Results

Protein sequence analysis of LI and classical cadherins

The protein structure of LI-cadherin reveals some striking differences when compared to E-cadherin, the most extensively studied classical cadherin (fig. 1 A). LI-cadherin exhibits two additional extracellular cadherin repeats but lacks the prosequence typical for classical cadherins. Interestingly, parts of the conserved Ca^{2+}-binding motifs at the junctions of the cadherin repeat are missing between cadherin repeat EC2 and EC3 of LI-cadherin. The cytoplasmic domain of LI-cadherin comprises only about 25 amino acids and is thus much shorter than that of E-cadherin with a length of 160 amino acids.

In figure 1 B, the protein sequences of rat LI-, as well as murine E-, P- and N-cadherin derived from published cDNA sequences (accession numbers X06115, X06340 and M31131, respectively) are compared to that of murine LI-cadherin [17] using the unbiased matrix plot analysis [26]. The x-axis corresponds to the amino acid sequence of murine LI-cadherin and the y-axis to the indicated sequences of the other cadherins. In this plot, identical amino acids on both coordinates produces a small dot. However, dots were only printed when 14 out of 21 successive amino acids were identical. Thus, stretches of homology produce short diagonal lines. In this plot, two identical or highly similar proteins like murine and rat LI-cadherin yield a long central diagonal marked in red (fig. 1 B). The parallel red lines in this plot at a distance of about 220 amino acids indicate that the cadherin repeats EC1–2 of rat and murine LI-cadherin are not only similar to each other but also closely related to the repeats EC3–4. The same parallel lines at 220 amino acids distance are also seen when comparing murine LI-cadherin with various classical cadherins. These findings indicate that the cadherin repeats EC1–2 and EC3–4 of murine LI-cadherin are both homologous to the repeats EC1–2 of classical cadherins. The long central diagonal in those plots indicates the overall homology of the repeats EC3–7 of LI-cadherin to the repeats EC1–5 of classical cadherins.

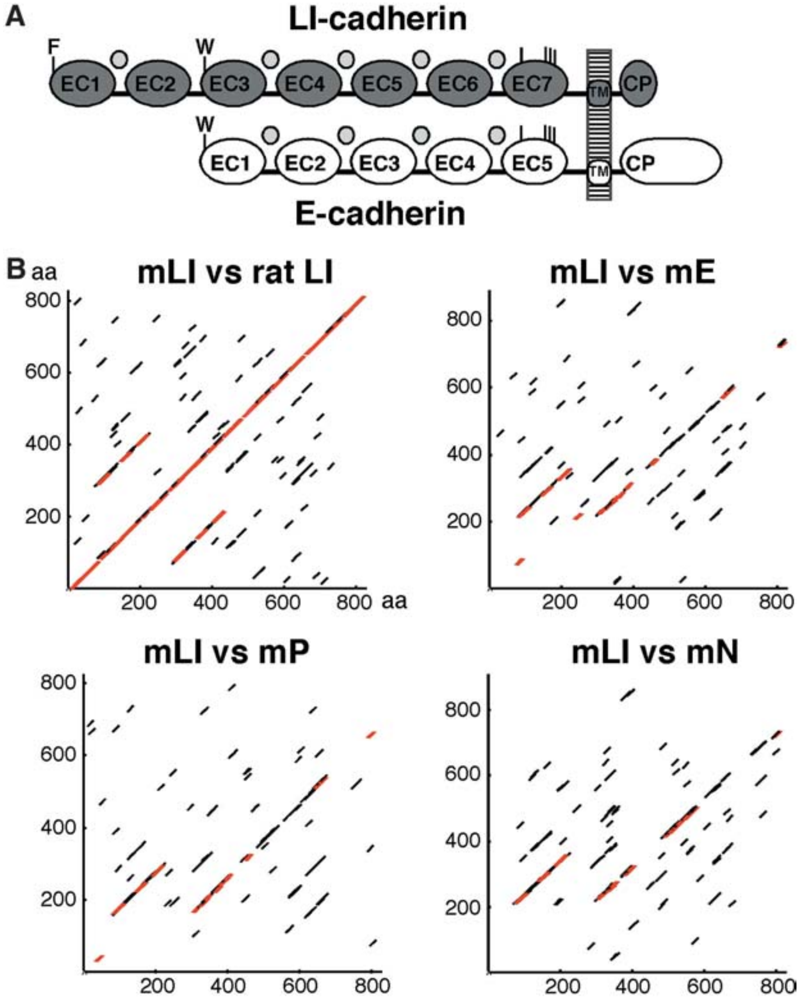
Due to these results, we compared in detail the amino acid sequences of all LI-cadherin repeats to those of E-cadherin (fig. 1 C). The first two cadherin repeats of LI- and E-cadherin show a higher similarity to each other than to any other repeat. As already anticipated from the previous analysis, the isolated LI-cadherin repeats EC3 to EC7 all exhibit the highest similarity to repeats EC1 to EC5 of E-cadherin. Although the overall amino acid identity of LI- and E-cadherin is only about 25%, these results suggest that the LI-cadherin repeats EC1–2 originated from a duplication of the first two repeats of an ancestral cadherin molecule with five repeats. Since the best evidence for this mechanism should be found on the genomic level, we cloned and analyzed the murine LI-cadherin gene and compared it to the published gene structures of classical cadherins.

Cloning of the murine LI-cadherin gene

Using a rat cDNA probe covering the entire LI-cadherin cDNA sequence [15], a murine genomic library was screened and nine λ phage clones were isolated (fig. 2). Hybridization experiments and sequencing of the genomic inserts with phage primers revealed that these clones contained about 70% of the corresponding LI-cadherin cDNA but lacked large parts of the 5′ region (fig. 2). In order to isolate the remaining regions of the LI-cadherin gene, we screened in addition a murine cosmid library. Two cosmid clones, designated cLI1 and cLI2, were identified with a probe encoding the first 300 amino acids of murine LI-cadherin [17]. The sequenced cosmid DNA contained all of the remaining genomic regions. In total, the murine LI-cadherin gene spans about 59 kb and contains 18 exons (fig. 2). As already observed for other cadherin genes, the intron positions do not correspond to the domain structure of the encoded protein [27].

Determination of the transcriptional start site of the murine LI-cadherin gene

To determine the transcriptional start site of the LI-cadherin gene, we performed a primer extension assay. Total



C

E \ LI	EC1	EC2	EC3	EC4	EC5	EC6	EC7
EC1	36	19	28	22	19	25	17
EC2	24	35	19	26	19	29	15
EC3	20	26	18	21	32	26	16
EC4	26	18	24	18	22	34	21
EC5	20	17	17	18	15	17	23

Figure 1. Comparative protein structure analysis of LI- and E-cadherin. (*A*) Schematic representation of the LI- and E-cadherin protein structures. Extracellular cadherin repeats are labeled as EC1 to EC7, the transmembrane domain as TM and the cytoplasmic domain as CP. The small circles in between cadherin repeats symbolize Ca^{2+}-binding pockets and the small vertical lines in EC7 and EC5 represent conserved cysteines. Conserved aromatic amino acids at the second position of the cadherin repeats EC1 and EC3 are indicated by F (phenylalanine) and W (tryptophan). (*B*) Matrix plot analysis of murine LI-cadherin compared to rat LI-, as well as murine E-, P- and N-cadherin (mE, mP and mN). Major regions of homology are shown in red. The long diagonal lines indicate the overall homology between the compared protein sequences. Parallel lines at a distance of 110 amino acids from the central diagonal (or at multiples thereof) are related to the homology of successive extracellular cadherin repeats to each other. The minimal match length was set to 21 amino acids (aa) and the number of mismatches tolerated was set to 14 aa. (*C*) Shotgun comparison of individual cadherin repeats of murine LI- and E-cadherin. The best matches (in percent aa identity) for every LI-cadherin repeat (top) are shaded in gray.

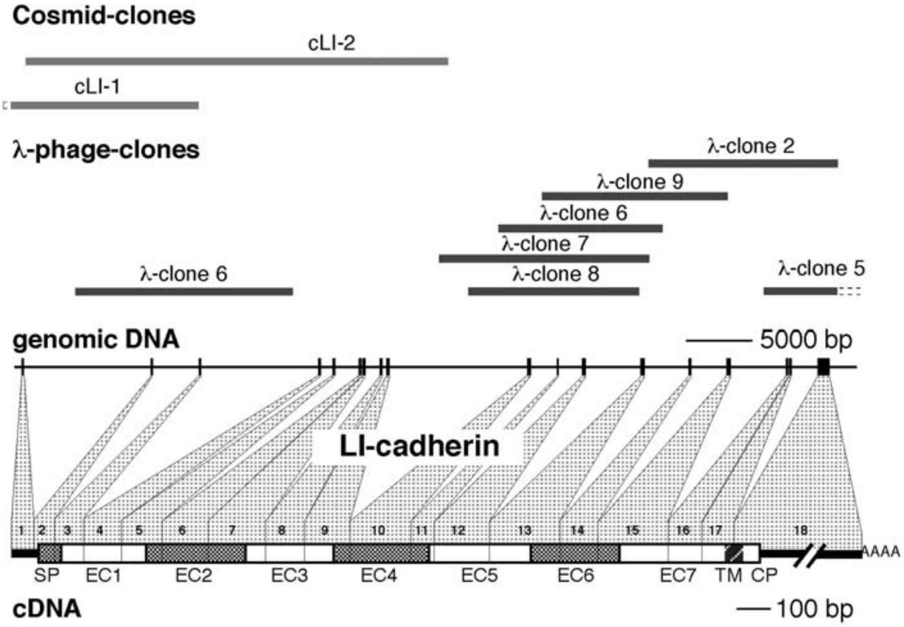


Figure 2. Cloning of the murine LI-cadherin gene. Initially, seven different λ clones were identified in a mouse genomic library using a 2.8-kb rat LI-cadherin cDNA probe. To retrieve the missing 5′ region of the gene, two cosmid clones were isolated by screening a spotted library with murine LI-cadherin cDNA probes. The initial alignment was achieved by Southern blot analysis and subsequently verified by DNA sequencing. The complete LI-cadherin gene covers more than 59 kb of genomic DNA and consists of 18 exons that are marked in the schematic representation linking the gene and cDNA sequence. SP, signal peptide; EC1–EC7, extracellular cadherin repeats 1 to 7; TM, transmembrane region; CP, cytoplasmic domain.

RNA was isolated from mouse small intestine and checked for integrity as well as the presence of LI-cadherin transcripts by Northern blot analysis (fig. 3A). A single band was only detected in the RNA fraction from intestine, whereas kidney and testis RNA stained negative for LI-cadherin transcripts. The size of the LI-cadherin transcript was estimated to be 3.6 kb. Splice variants could not be detected.

The primer extension analysis was performed with two antisense primers differing 31 bp in position at the 5′ end of the LI-cadherin cDNA. Maximum length products of 67 bp and 98 bp were obtained for primer 1 and primer 2, respectively (fig. 3B). Both fragments indicate independently a transcription initiation site at an adenine 134 bp upstream from the ATG (fig. 3C). The presence of several major primer extension products of shorter length might reflect multiple start sites as has been described for other TATA-less promotors [28]. No products were obtained with the same primers when yeast tRNA was used instead of intestinal RNA. On the genomic DNA, a cytosine is located in front of the first transcribed adenine and the second transcribed nucleotide is a guanine, which resembles perfectly well the consensus transcriptional initiation sequence.

At the 3′ end of the LI-cadherin gene, a single poly-A signal was identified. Assuming a poly-A tail length of 200–250 bp [29], the predicted size of the LI-cadherin transcript, 3417 bp from the start point to the poly-A site, is in good agreement with the detected 3.6-kb mRNA band revealed in the Northern blot.

Gene structure analysis of the murine LI-cadherin gene

The locations of the exon-intron boundaries were derived by comparing the genomic sequence of LI-cadherin to the corresponding cDNA sequence (AF177669) [16]. When in doubt, the general rules for splice sites were used to define their exact position [30]. The resulting sequences of the exon/intron boundaries are consistent with the reported consensus sequence for splice donor (GT) and acceptor (AG) sites [31] as shown in table 1. The exon sizes of the murine LI-cadherin gene range between 70 and 889 bp and the intron sizes vary between 95 and 10,218 bp (table 1).

Gene structure comparison of various cadherin genes

As already described above, the intron positions within the cDNA sequence do not correspond to the repeat structure of the derived protein sequence. We therefore compared the intron positions within the LI-cadherin cDNA to those of murine E- [27], P- [32] and N-cadherin [33], as shown in figure 4A. This analysis revealed that the intron pattern within the LI-cadherin extracellular repeats EC3-7 is identical to that of murine E-, P- and N-cadherin repeats EC1-5 with the exception of an additional intron close to the transmembrane domain. Remarkably, the po-

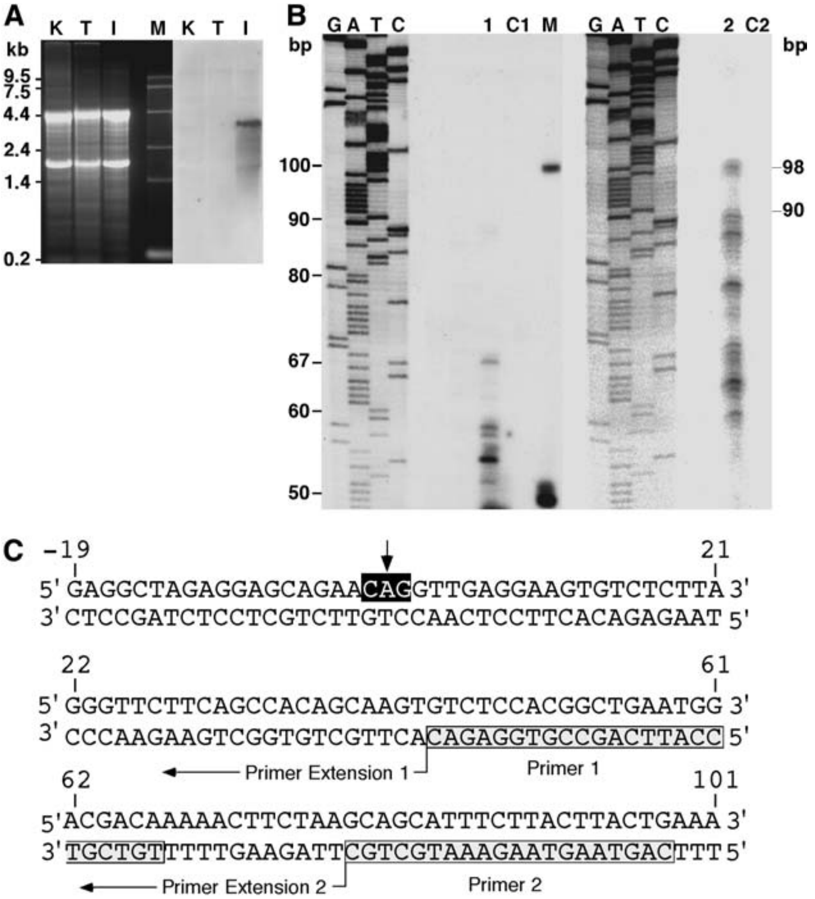

A
K T I M K T I
kb
9.5
7.5
4.4
2.4
1.4
0.2
B
G A T C
bp
100
90
80
67
60
50
1 C1 M
G A T C
2 C2
bp
98
90
C
-19
21
5' GAGGCTAGAGGAGCAGAACAGGTTGAGGAAGTGTCTCTTA 3'
3' CTCCGATCTCCTCGTCTTGTCCAACTCCTTCACAGAGAAT 5'
22
61
5' GGGTTCTTCAGCCACAGCAAGTGTCTCCACGGCTGAATGG 3'
3' CCCAAGAAGTCGGTGTCGTTCACAGAGGTGCCGACTTACC 5'
Primer Extension 1
Primer 1
62
101
5'ACGACAAAAACTTCTAAGCAGCATTTCTTACTTACTGAAA 3'
3' TGCTGTTTTTGAAGATTCGTCGTAAAGAATGAATGACTTT 5'
Primer Extension 2
Primer 2

Figure 3. Identification of the transcriptional start site of the LI-cadherin gene. (*A*) Equal amounts of total RNA from mouse kidney (K), testis (T) and intestine (I) were subjected to Northern blot analysis. Integrity of the 28S and 18S rRNA was verified by ethidium bromide staining. Hybridization with a rat LI-cadherin cDNA probe revealed a single band of about 3600 bp only in RNA obtained from the intestine. (*B*) Mapping of the LI-cadherin transcriptional start site was done by primer extension using two murine LI-cadherin-specific reverse DNA primers, differing by 31 bp in their target position. The maximum length of the resulting products was 67 bp (primer 1, left side) and 98 bp (primer 2, right side). Yeast tRNA served as a negative control (lanes C1 and C2). Dideoxy sequencing products (lanes G, A, T and C) were obtained with a third primer and served in conjunction with a [^{32}P]-labeled 50-bp DNA ladder (lane M) as a reference. (*C*) Genomic region surrounding the transcriptional start site. The first transcribed nucleotide within the typical CAG motif is marked by an arrow. The position of the two reverse primers used for the primer extension analysis are shown in shaded boxes.

sition and the phasing of the first three introns within the LI-cadherin repeats EC1–2 (LI-cadherin introns 3–5) match perfectly with those of LI-cadherin repeats EC3–4 (LI-cadherin introns 7–9) as well as with those of the E-, P- and N-cadherin repeats EC1–2 (E-cadherin introns 4–6). However, the fourth introns in both EC1–2 and EC2–4 of LI-cadherin (introns 6 and 10), have a different phasing (phase 1) compared to the corresponding introns of classical cadherins (phase 0). Since any duplication of cadherin repeats must have taken place at the genomic level, we also analyzed the same part of the LI- and E-cadherin sequences with respect to their exon structures (fig. 4B). The protein sequences encoded by the LI-cadherin exons 4–6 match without major gaps with those of exons 8–10 within the molecule and with those of exons 5–7 of E-cadherin. In contrast, the derived protein sequences of LI-cadherin exons 7 and 11 as well as those of the corresponding E-cadherin exon 8 align only partially. This observation as well as the phase shift between LI-cadherin introns 6 and 10 on the one hand and E-cadherin intron 7 on the other suggest that the duplication may have involved either exons 4–7 or 5–8 (E-cadherin numbering) of a five-domain precurser molecule of LI-cadherin.

Discussion

All classical and desmosomal cadherins as well as the glycosyl phosphatidyl inositol-anchored T-cadherin [34] are characterized by an extracellular region built from five structurally defined homology domains called cadherin repeats [17]. LI-cadherin is an intestine-specific member of the cadherin superfamily with distinct structural features including seven cadherin repeats [15]. In this study, we ex-

Table 1. LI-cadherin gene structure.

No.	Exon size (bp)	Intron size (bp)	5′ splice donor site	Intron phase	3’-splice acceptor site
1	112	9452	aag **gt**aagg...	–	ctggttcactcacag g**ag**
2	22 + 48	3464	TTG **gt**aagc...	0	tctctttattaac**ag** ACC
3	99	8726	CAG **gt**aaag...	0	ttttctttttctc**ag** TTT
4	135	924	CAG **gt**gagc...	0	tttcttcaacaac**ag** CTT
5	139	1792	CCA G **gt**agaa...	1	atgctttgtcacc**ag** GA AAG
6	159	95	GAA G **gt**aagt...	1	ccccaacttctac**ag** GA TCC
7	200	1094	CAG **gt**agtg...	0	gctttggccgtgc**ag** GTG
8	132	364	TCA **gt**gagt...	0	accttcacattgc**ag** CAT
9	152	10218	TTG G **gt**aaga...	1	ctttcttctccaa**ag** GT AAC
10	216	1994	GTA G **gt**aagc...	1	gactttggttttc**ag** AT TTC
11	77	1791	AAT **gt**gagt...	0	ttgtttcttcctc**ag** TAT
12	192	4126	AAG **gt**agat...	0	tctctccctccac**ag** CCT
13	245	3310	GTG AG **gt**acag...	2	tctcccctcatga**ag** T TAT
14	131	2675	GTA G **gt**gagc...	1	ccttcaatttctt**ag** GT GGG
15	240	4140	AAT G **gt**gagt...	1	tccctgttctgac**ag** GT ACA
16	117	129	CCA G **gt**aggt...	1	tgttctgcttttt**ag** TT ACT
17	114	2015	ATT G **gt**aagt...	1	gttttctcttttc**ag** GT ATA
18	89 + 799		(AGAAATC-poly-A)		
Consensus sequence			AG **gt**aagt...		(t/c)$_{10}$ nc**ag** G
mRNA		3409			
Coding region		2484			
Genomic region		59726			

amined the phylogenetic origin of its seven cadherin repeats by analyzing the protein and gene structure of LI-cadherin and comparing it to that of classical cadherins.

The protein structures of LI- and E-cadherin [35], the prototype of classical cadherins, exhibit some striking differences, suggesting a poor relationship of both proteins within the cadherin family. Compared to E-cadherin, LI-cadherin exhibits two additional cadherin repeats and a short cytoplasmic domain. Furthermore, the overall amino acid identity of LI- and E-cadherin is less than 25%. In contrast, a closer investigation of the protein structures of LI- and various classical cadherins using the unbiased matrix plot analysis revealed a homology of EC1–5 of classical cadherins to EC3–7 of LI-cadherin. Moreover, not only EC3–4, but also EC1–2 of LI-cadherin are homologous to EC1–2 of classical cadherins. A domain-specific protein sequence analysis comparing every LI-cadherin repeat to every E-cadherin repeat supports this result. Based on the comparison of the protein structures of LI- and E-cadherin, we assumed a phylogenetic relationship between the classical five-domain cadherins and the seven-domain LI-cadherin. In particular, we hypothesized that the two additional extracellular repeats of LI-cadherin originated by a partial gene duplication event from a common ancestor with five cadherin repeats.

The missing Ca^{2+}-binding pocket between EC2 and EC3 of LI-cadherin also supports our hypothesis. The acidic binding motif DXNDN in EC2 and DXD in EC3 are absent in LI-cadherin [15, 36]. A duplication of the first two repeats would lead to an incomplete Ca^{2+}-binding pocket, because the N-terminal cadherin repeat (EC1) lacks the DXD motif in general. The lack of Ca^{2+}-binding motifs between the LI-cadherin repeats EC2 and EC3 is likely to influence Ca^{2+}-binding and thus has functional implications. Moreover, the conserved tryptophan at position 2 in EC1 of classical cadherins [7, 37] is found at the same position within EC3 of LI-cadherin, whereas it is replaced by a phenlyalanine in EC1 of LI-cadherin. The tryptophan within EC1 of classical cadherins has been described as a key residue for the cadherin adhesion mechanism [37, 38]. One might thus assume that the function of the tryptophan within EC1 of classical cadherins is preserved in EC3 and not in EC1 of LI-cadherin, although the aromatic phenylalanine in EC1 of LI-cadherin might also adopt a similar function as tryptophan.

To test our hypothesis of a common phylogenetic origin of LI- and classical cadherins at the genomic level, we cloned the murine LI-cadherin gene and compared its structure to that of various classical cadherins. The LI-cadherin gene spans 59 kb and contains 18 exons. Like other characterized cadherin genes in mouse and human, the LI-cadherin gene harbors large introns, especially in the 5′ region [27, 32, 33]. The identified major transcriptional start site fits with the initiator consensus sequence established by a statistical analysis of 502 eukaryotic initiation sites of RNA polymerase II promoters [39]. The proximal 5′-flanking region lacks TATA boxes but contains a GC-box element at –36, which is known to determine the start point in TATA-less promoters [39, 40].

The intron positions within the cDNA sequence do not correspond to the cadherin repeat structure of the derived protein. Within the classical cadherins, every repeat has its

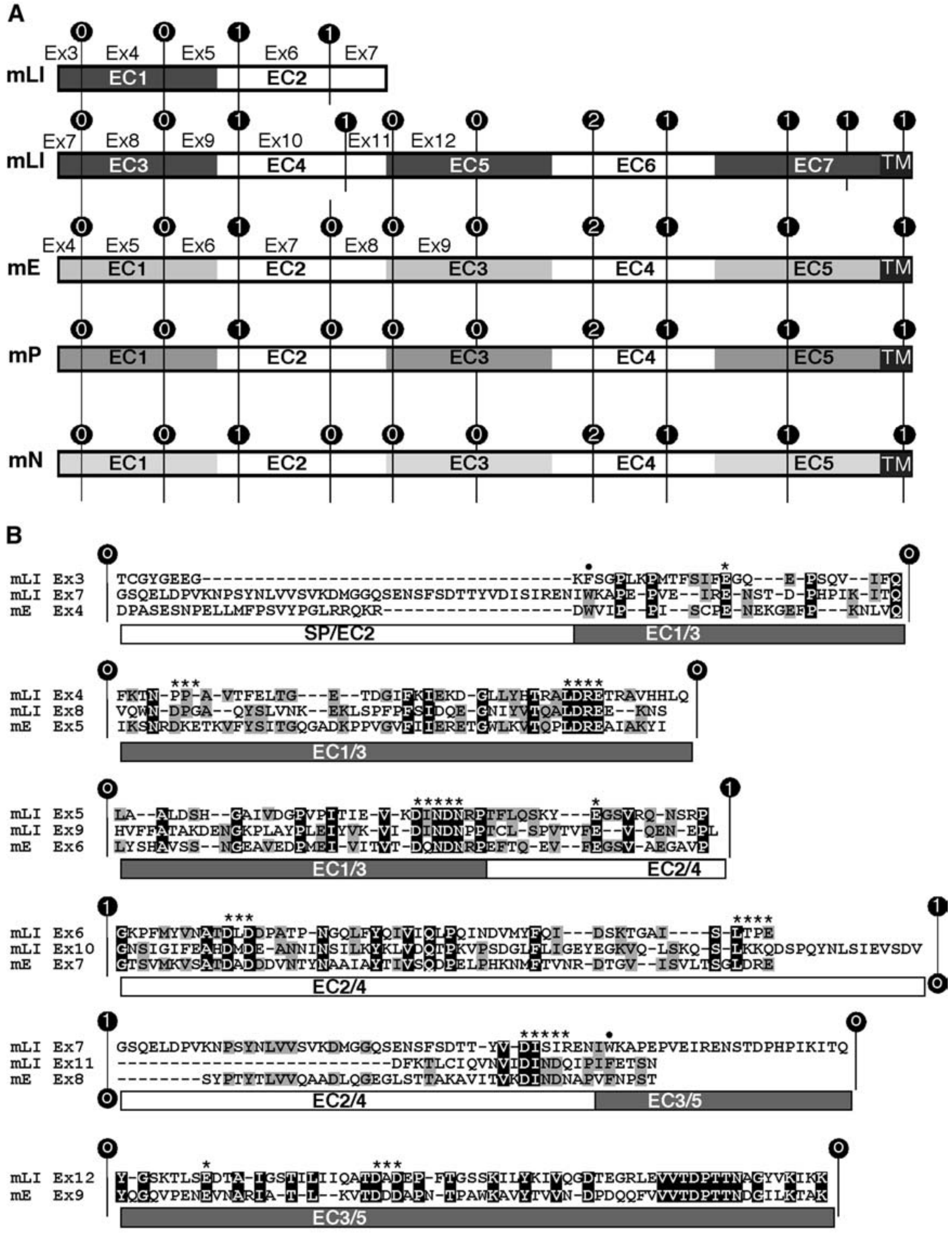

A
mLI
Ex3 Ex4 Ex5 Ex6 Ex7
EC1 EC2
0 0 1 1
mLI
Ex7 Ex8 Ex9 Ex10 Ex11 Ex12
EC3 EC4 EC5 EC6 EC7 TM
0 0 1 1 0 0 2 1 1 1 1
mE
Ex4 Ex5 Ex6 Ex7 Ex8 Ex9
EC1 EC2 EC3 EC4 EC5 TM
0 0 1 0 0 0 2 1 1 1
mP
EC1 EC2 EC3 EC4 EC5 TM
0 0 1 0 0 0 2 1 1 1
mN
EC1 EC2 EC3 EC4 EC5 TM
0 0 1 0 0 0 2 1 1 1
B
mLI Ex3 TCGYGEEG------------------------------------KFSGPLKPMTFSIFEGQ---E-PSQV--IFQ
mLI Ex7 GSQELDPVKNPSYNLVVSVKDMGGQSENSFSDTTYVDISIRENIWKAPE-PVE--IRE-NST-D-PHPIK-ITQ
mE Ex4 DPASESNPELLMFPSVYPGLRRQKR-------------------DWVIP--PI--SCPE-NEKGEFP---KNLVQ
SP/EC2
EC1/3
mLI Ex4 FKTN-PP-A-VTFELTG---E--TDGIFKIEKD-GLLYHTRALDRETRAVHHLQ
mLI Ex8 VQWN-DPGA--QYSLVNK--EKLSPFPFSIDQE-GNIYVTQALDREE--KNS
mE Ex5 IKSNRDKETKVFYSITGQGADKPPVGVFIIERETGWLKVTQPLDREAIAKYI
EC1/3
mLI Ex5 LA--ALDSH--GAIVDGPVPITIE-V-KDINDNRPTFLQSKY---EGSVRQ-NSRP
mLI Ex9 HVFFATAKDENGKPLAYPLEIYVK-VI-DINDNPPTCL-SPVTVFE--V-QEN-EPL
mE Ex6 LYSHAVSS--NGEAVEDPMEI-VITVT-DQNDNRPEFTQ-EV--FEGSV-AEGAVP
EC1/3
EC2/4
mLI Ex6 GKPFMYVNATDLDDPATP-NGQLFYQIVIQLPQINDVMYFQI---DSKTGAI----S-LTPE
mLI Ex10 GNSIGIFEAHDMDE-ANNINSILKYKLVDQTPKVPSDGLFLIGEYEGKVQ-LSKQ-S-LKKQDSPQYNLSIEVSDV
mE Ex7 GTSVMKVSATDADDDVNTYNAAIAYTIVSQDPELPHKNMFTVNR-DTGV--ISVLTSGLDRE
EC2/4
mLI Ex7 GSQELDPVKNPSYNLVVSVKDMGGQSENSFSDTT-YV-DISIRENIWKAPEPVEIRENSTDPHPIKITQ
mLI Ex11 --------------------------DFKTLCIQVNVIDINDQIPIFETSN
mE Ex8 --------SYPTYTLVVQAADLQGEGLSTTAKAVITVKDINDNAPVFNPST
EC2/4
EC3/5
mLI Ex12 Y-GSKTLSEDTA-IGSTILIIQATDADEP-FTGSSKILYKIVQGDTEGRLEVVTDPTTNAGYVKIKK
mE Ex9 YQGQVPENEVNARIA-T-L--KVTDDDAPN-TPAWKAVYTVVN-DPDQQFVVVTDPTTNDGILKTAK
EC3/5

Figure 4. Comparison of the exon-intron structure of LI-cadherin and various classical cadherins. (*A*) Alignment with respect to the cadherin repeat structure of murine LI-, E-, P- and N-cadherin (mLI, mE, mP and mN) indicating schematically the positions of the splice sites and the intron phasing. Exons are labeled Ex1, Ex2, etc. and intron phases are indicated by white numbers on black bullets. Whereas the intron-exon structure of EC3–7 of LI-cadherin almost perfectly matches that of EC1–5 of classical cadherins, the LI-cadherin intron-exon structure of EC1–2 is homologous to both EC3–4 of LI-cadherin and EC1–2 of classical cadherins. (*B*) Alignment with respect to the exons encoding the LI-cadherin repeats EC1–5 (exons 3–12) and the E-cadherin repeats EC1–3 (exons 4–9). Similar to *A*, the positions of cadherin repeats are represented by white and gray bars, the intron positions by vertical lines and their phases by white numbers on black bullets. The conserved tryptophan at position 2 within E-cadherin EC1 and LI-cadherin EC3 is marked with a dot and the positions of the conserved cadherin Ca^{2+}-binding motifs (E, DXD, LDRE, DXNDN) are indicated with asterisks. Note that LI-cadherin exons 4, 5 and 6 can be fully aligned with exons 8, 9 and 10 of LI-cadherin and 5, 6 and 7 of E-cadherin, respectively, whereas only partial alignments are possible for LI-cadherin exons 3 and 7 with exons 7 and 11 of LI-cadherin and exons 4 and 8 of E-cadherin, respectively.

particular intron position that is well preserved between different members of the cadherin subfamily, like E-, P-, and N-cadherin [27, 32, 33, 41]. Thus, the intron pattern can be used in addition to the sequence analysis as an independent tool to find homologies between two genes.

A superposition of the intron positions of LI-cadherin onto those of classical cadherins revealed that the intron pattern of the cDNA region encoding the LI-cadherin repeats EC3–7 perfectly matches that of the repeats EC1–5 of E-, P- and N-cadherin. The only difference we found is an additional intron in the LI-cadherin cDNA encoding the membrane-proximal last cadherin repeat.

A third line of evidence can be drawn from our observation that the phasing of the introns is also almost completely preserved between the respective gene fragments of LI-cadherin (EC3–7) and the compared classical cadherins (EC1–5), implying that both proteins originated from a common ancestral cadherin molecule with five extracellular cadherin repeats. The only difference in phasing is found in the second intron of LI-cadherin EC4 compared to that of EC2 of classical cadherins. We thus assume that this phase shift took place after the separation of the five-domain precurser proteins of LI- and classical cadherins.

Subsequently, a partial gene duplication event involving either exons 4–7 or 5–8 (E-cadherin exon numbering scheme) may have led to the generation of the seven-domain LI-cadherin. This model would explain the sequence similarity, the conserved intron position pattern and the intron phasing between the cadherin repeats EC1–2 and EC3–4 of LI-cadherin. It would also resolve the different phasing of the second intron of both EC2 and EC4 of LI-cadherin with respect to that of EC2 of classical cadherins.

In conclusion, our data strongly support the hypothesis that LI- and classical cadherins originated from a precurser cadherin with five cadherin repeats and that LI-cadherin separated from the other members of this family by a partial gene duplication event.

Acknowledgements. We thank P. Krienke for excellent technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 366, Teilprojekt C2 and Graduiertenkolleg 276).

1 Takeichi M. (1991) Cadherin cell adhesion receptors as a morphogenetic regulator. Science **251:** 1451–1455

2 Geiger B. and Ayalon O. (1992) Cadherins. Annu. Rev. Cell Biol. **8:** 307–332

3 Nelson W. J. (1992) Regulation of cell surface polarity from bacteria to mammals. Science **258:** 948–955

4 Tepass U., Truong K., Godt D., Ikura M. and Peifer M. (2000) Cadherins in embryonic and neural morphogenesis. Nat. Rev. Mol. Cell. Biol. **1:** 91–100

5 Nollet F., Kools P. and Roy F. van (2000) Phylogenetic analysis of the cadherin superfamily allows identification of six major subfamilies besides several solitary members. J. Mol. Biol. **299:** 551–572

6 Overduin M., Harvey T. S., Bagby S., Tong K. I., Yau P., Takeichi M. et al. (1995) Solution structure of the epithelial cadherin domain responsible for selective cell adhesion. Science **267:** 386–389

7 Shapiro L., Fannon A. M., Kwong P. D., Thompson A., Lehmann M. S., Grubel G. et al. (1995) Structural basis of cell-cell adhesion by cadherins. Nature **374:** 327–337

8 Angst B. D., Marcozzi C. and Magee A. I. (2001) The cadherin superfamily. J. Cell Sci. **114:** 625–626

9 Hynes R. O. (1999) Cell adhesion: old and new questions. Trends Cell Biol. **9:** M33–M37

10 Kemler R. (1992) Classical cadherins. Semin. Cell Biol. **3:** 149–155

11 Stappert J. and Kemler R. (1994) A short core region of E-cadherin is essential for catenin binding and is highly phosphorylated. Cell Adhes. Commun. **2:** 319–327

12 Kemler R. (1993) From cadherins to catenins: cytoplasmic protein interactions and regulation of cell adhesion. Trends Genet. **9:** 317–321

13 Gumbiner B. M. (2000) Regulation of cadherin adhesive activity. J. Cell Biol. **148:** 399–404

14 Troyanovsky S. M., Troyanovsky R. B., Eshkind L. G., Leube R. E. and Franke W. W. (1994) Identification of amino acid sequence motifs in desmocollin, a desmosomal glycoprotein, that are required for plakoglobin binding and plaque formation. Proc. Natl. Acad. Sci. USA **91:** 10790–10794

15 Berndorff D., Geßner R., Kreft B., Schnoy N., Lajous-Petter A. M., Loch N. et al. (1994) Liver-intestine cadherin: molecular cloning and characterization of a novel Ca(2+)-dependent cell adhesion molecule expressed in liver and intestine. J. Cell Biol. **125:** 1353–1369

16 Kreft B., Berndorff D., Bottinger A., Finnemann S., Wedlich D., Hortsch M. et al. (1997) LI-cadherin-mediated cell-cell adhesion does not require cytoplasmic interactions. J. Cell Biol. **136:** 1109–1121

17 Angres B., Kim L., Jung R., Geßner R. and Tauber R. (2001) LI-cadherin gene expression during mouse intestinal development. Dev. Dyn. **221:** 182–193

18 Boller K., Vestweber D. and Kemler R. (1985) Cell-adhesion molecule uvomorulin is localized in the intermediate junctions of adult intestinal epithelial cells. J. Cell Biol. **100:** 327–332

19 Grötzinger C., Kneifel J., Patschan D., Schnoy N., Anagnostopoulos I., Faiss S. et al. (2001) LI-cadherin: a marker of gastric metaplasia and neoplasia. Gut **49:** 73–81

20 Hippo Y., Taniguchi H., Tsutsumi S., Machida N., Chong J. M., Fukayama M. et al. (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res. **62:** 233–240

21 Takamura M., Sakamoto M., Ino Y., Shimamura T., Ichida T., Asakura H. et al. (2003) Expression of liver-intestine cadherin and its possible interaction with galectin-3 in ductal adenocarcinoma of the pancreas. Cancer Sci. **94:** 425–430

22 Wong B. W., Luk J. M., Ng I. O., Hu M. Y., Liu K. D. and Fan S. T. (2003) Identification of liver-intestine cadherin in hepatocellular carcinoma – a potential disease marker. Biochem. Biophys. Res. Commun. **311:** 618–624

23 Feinberg A. P. and Vogelstein B. (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:** 6–13

24 Sanger F., Nicklen S. and Coulson A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467

25 Sambrook J., Fritsch E. F. and Maniatis T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.

26 Vingron M. and Argos P. (1991) Motif recognition and alignment for many sequences by comparison of dot-matrices. J. Mol. Biol. **218:** 33–43

27 Ringwald M., Baribault H., Schmidt C. and Kemler R. (1991) The structure of the gene coding for the mouse cell adhesion molecule uvomorulin. Nucleic. Acids Res. **19:** 6533–6539

28 Clark M. P., Chow C. W., Rinaldo J. E. and Chalkley R. (1998) Correct usage of multiple transcription initiation sites and C/EBP-dependent transcription activation of the rat XDH/XO TATA-less promoter requires downstream elements located in the coding region of the gene. Nucleic. Acids Res. **26:** 1801–1806

29 Wahle E. (1995) Poly(A) tail length control is caused by termination of processive synthesis. J. Biol. Chem. **270:** 2800–2808

30 Breathnach R. and Chambon P. (1981) Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. **50:** 349–383

31 Mount S. M. (1982) A catalogue of splice junction sequences. Nucleic Acids Res. **10:** 459–472

32 Hatta M., Miyatani S., Copeland N. G., Gilbert D. J., Jenkins N. A. and Takeichi M. (1991) Genomic organization and chromosomal mapping of the mouse P-cadherin gene. Nucleic Acids Res. **19:** 4437–4441

33 Miyatani S., Copeland N. G., Gilbert D. J., Jenkins N. A. and Takeichi M. (1992) Genomic structure and chromosomal mapping of the mouse N-cadherin gene. Proc. Natl. Acad. Sci. USA **89:** 8443–8447

34 Ranscht B. and Dours-Zimmermann M. T. (1991) T-cadherin, a novel cadherin cell adhesion molecule in the nervous system lacks the conserved cytoplasmic region. Neuron **7:** 391–402

35 Schuh R., Vestweber D., Riede I., Ringwald M., Rosenberg U. B., Jackle H. et al. (1986) Molecular cloning of the mouse cell adhesion molecule uvomorulin: cDNA contains a B1-related sequence. Proc. Natl. Acad. Sci. USA **83:** 1364–1368

36 Geßner R. and Tauber R. (2000) Intestinal cell adhesion molecules: liver-intestine cadherin. Ann. N. Y. Acad. Sci. **915:** 136–143

37 Boggon T. J., Murray J., Chappuis-Flament S., Wong E., Gumbiner B. M. and Shapiro L. (2002) C-cadherin ectodomain structure and implications for cell adhesion mechanisms. Science **296:** 1308–1313

38 Patel S. D., Chen C. P., Bahna F., Honig B. and Shapiro L. (2003) Cadherin-mediated cell-cell adhesion: sticking together as a family. Curr. Opin. Struct. Biol. **13:** 690–698

39 Bucher P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. **212:** 563–578

40 Kageyama R., Merlino G. T. and Pastan I. (1989) Nuclear factor ETF specifically stimulates transcription from promoters without a TATA box. J. Biol. Chem. **264:** 15508–15514

41 Gallin W. J. (1998) Evolution of the 'classical' cadherin family of cell adhesion molecules in vertebrates. Mol. Biol. Evol. **15:** 1099–1107