Genomic organization and chromosomal mapping of the mouse P-cadherin gene

Masayuki Hatta, Seiji Miyatani, Neal G. Copeland¹, Debra J. Gilbert¹, Nancy A. Jenkins¹ and Masatoshi Takeichi*

Department of Biophysics, Faculty of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto 606, Japan and ¹Mammalian Genetics Laboratory, ABL-Basic Research Program, NCI-Frederick Cancer Research and Development Center, Frederick, MD 21702, USA

Received May 23, 1991; Revised and Accepted July 18, 1991

ABSTRACT

Cadherins are a family of Ca^{2+} -dependent cell adhesion molecules, that includes P-cadherin, E-cadherin, N-cadherin and L-CAM. In this study, the genomic organization of the mouse P-cadherin gene was determined by analyzing overlapping DNA clones obtained from a mouse genomic library. The results showed that this gene spans over 45kb and consists of 15 exons. A marked feature of this gene is that the first intron is 23kbp long accounting for half its length. Comparisons of this structure with that of L-CAM, a chicken cadherin, revealed that the exon-intron boundaries are conserved between the two genes except that the P-cadherin first exon includes the corresponding first and second exons of the L-CAM gene. This gene was also similar to the other in that the second intron, which corresponds to the P-cadherin first intron, is exceptionally longer than other introns. These results suggest that the exon-intron pattern conserved in these genes is of significance for generation of domain structure of cadherin molecules or for their transcriptional regulation. We also determined the chromosomal localization of the P-cadherin gene by interspecific backcross analysis, and found that this gene is located in the central region of mouse chromosome 8 and linked with the E-cadherin locus. This is the first evidence for the linkage of different cadherin genes.

INTRODUCTION

Cadherins are a gene family of Ca^{2+} -dependent cell-cell adhesion receptors (1). Members of the family includes E-cadherin (uvomorulin) (2, 3), P-cadherin (4), N-cadherin (5, 6), R-cadherin (7), L-CAM (8) and EP-cadherin (9). They share a common domain structure, and their amino acid sequences are conserved in a range of 45% to 74%. Recently, novel forms of cadherin have been identified. One is T-cadherin; this molecule is similar to other cadherins at the extracellular domain, but it has no cytoplasmic domain (10). A desmosomal glycoprotein,

desmoglein I, is similar to cadherins, but they have additional sequences at the carboxy terminus as well as some deletions in the extracellular domain (11). Other desmosomal glycoproteins also seem to be modified forms of cadherin (12). Also, *Drosophila fat* has cadherin-like sequences, although its overall structure is different from the original cadherins (P. Mahoney and C. Goodman, personal communication). These findings suggest that cadherins constitute a superfamily through vertebrates to invertebrates.

Each of the members of the cadherin family shows a unique pattern of tissue distribution that is developmentally regulated, and the pattern is generally associated with morphogenetic events such as the separation of cell layers, epithelial-mesenchymal cell transition and condensation or dispersion of cell mass (1). Moreover, *in vitro* studies revealed that these molecules bind cells by their homophilic interactions, and, through this property, they seem to contribute to the selective adhesion of different cell types (1). All these results suggest that the differential expression of cadherins during development play important roles in morphogenesis of embryos.

In order to further pursue the function of cadherins, genetic approaches will be essential. To this aim, it is important to reveal the structure and chromosomal localization of the cadherin genes. The results of such studies should also provide us with significant information for dissecting the cis-acting sequences necessary for tissue specific expression of these genes. So far, the structure of the L-CAM gene has been reported (13), but no information is available for genomic structure of other cadherins. Regarding chromosomal mapping, the human E-cadherin gene has been localized to chromosome 16 in the region of 16p11-16qter (14), and the mouse E-cadherin gene to chromosome 8 (15). Also, the human N-cadherin gene has been mapped to chromosome 18 (16).

The aim of the present study is to reveal the genomic organization of the mouse P-cadherin gene and its chromosomal localization. The results show that the exon-intron boundaries are highly conserved between the mouse P-cadherin and chicken L-CAM genes, and the P-cadherin gene is tightly associated with the E-cadherin gene.

* To whom correspondence should be addressed

Figure 1. Genomic organization of the mouse P-cadherin gene. The 15 exons encoding the mouse P-cadherin gene are shown as vertical bars for coding region and open boxes for noncoding regions. The genomic clones are aligned with respect to the genomic DNA at the middle, and restriction maps are shown at the bottom of the figure. The sites for NotI and SalI were not found. RI, EcoRI; Ba, BamHI; Xh, XhoI; RV, EcoRV; Sp, SphI; Sm, SmaI; Sc, SacI; Hd, HindIII.

MATERIALS AND METHODS

Construction of a mouse genomic library

Genomic DNA purified from E14 embryos of C57/BL6 mice was digested partially with Sau3AI, and approximately 15kbp fragments of the DNA were isolated by sucrose density gradient centrifuge. These DNA fragments were ligated with EMBL4 phage arms that were purified after digestion with BamHI and SalI, and packaged *in vitro* (Gigapack Gold, Stratagene, U.S.A.). The titer of the obtained library was 6×10^6 p.f.u.

Isolation and analysis of genomic clones

Gene cloning, restriction mapping and sequence analysis were done by standard methods (17) with some modifications. The mouse genomic library was screened with the full length mouse P-cadherin cDNA (4) labelled with [α^{32} P]dCTP (Amersham, UK) using a nick translation or a random primer labeling kit (Takara, Japan). Positive clones were digested with EcoRI or BamHI, and each fragment was cloned into pUC vector. These subclones were used as templates for restriction mapping and sequence analysis. DNA sequences were determined by the method of ExoIII-Mung bean nuclease deletion, followed by 7-deaza dideoxy chain terminating reaction (Takara, Japan) using double strand templates.

Interspecific backcross mapping

Interspecific backcross progeny were generated by mating (C57BL/6J $\times$ *M. spretus*)F₁ females and C57BL/6J males as previously described (18). A total of 205 N₂ progeny were obtained; a random subset of these N₂ mice were used to map the P-cadherin locus (see text for details). DNA isolation, restriction enzyme digestion, agarose gel electrophoresis, Southern blot transfer, and hybridization were performed essentially as described (19). All blots were prepared with Zetabind nylon membrane (AMF-Cuno). The P-cadherin probe, a 3.2 kb full length mouse cDNA (4), was labeled with [α^{32} P]dCTP using a nick translation labeling kit (Boehringer Mannheim). Washing was done to a final stringency of $0.5 \times$ SSCP, 0.1% SDS, 65°C. Fragments of 6.0, 4.6, 2.6, 2.3, 1.6, 1.1 and 0.5 kb were detected in PvuII digested C57BL/6J DNA; fragments of 4.6, 3.1, 2.6, 2.4, 1.6 1.1 and 0.5 kb were

No.	EXON (bp)	INTRON (bp)	DONOR	ACCEPTOR
1	1k	23k	AAA G gtaagacccc.....ttcctcacag TA GCC 55 Lys 56 Val 57 Ala	
2	74	1077	CAG gtaaatgtgt.....ttgcttcag GGA 81 Gln 82 Gly	
3	126	128	CAG gtatgaccag.....ctgtttccag CTC 122 Gln 123 Leu	
4	156	100	GAG gtaaatagcc.....gtgccacag CTT 174 Glu 175 Leu	
5	145	1769	CCT G gtaagaaggg.....accccgagc GC ACT 222 Pro 223 Gly 224 Thr	
6	176	608	GAG gtgagtagct.....tctcccctag AAA 281 Glu 282 Lys	
7	129	1366	AAG gtaacgccc.....ctccccacag TAT 324 Lys 325 Tyr	
8	186	2.3k	AAG gtaagtctgt.....tctctcctag GGT 386 Lys 387 Gly	
9	241	223	ATC AG gtacctggg.....tcctttccag C TAC 466 Ile 467 Ser 468 Tyr	
10	146	740	AGT G gtgagagctt.....ttgtaaacag GA AAC 515 Ser 516 Gly 517 Asn	
11	225	2389	AAA G gtaactctg.....gtctctccag GA GAC 590 Lys 591 Gly 592 Asp	
12	210	4559	CTG A gtgagtacca.....ctcatttcag CC CTT 660 Leu 661 Thr 662 Leu	
13	131	346	CAG gtgagtcaca.....tctcataaag GAC 704 Gln 705 Asp	
14	147	2224	GAG gtaaggcatg.....gcttcctag AAC 753 Glu 754 Asn	
15	883+poly(A)			

Figure 2. Exon/intron boundaries of the mouse P-cadherin gene. Exon sequences are in capital letters and intron sequences are in lowercase letters. The amino acids with respect to the boundaries are indicated with the numbers in P-cadherin cDNA (4).

detected in PvuII digested *M. spretus* DNA. The P-cadherin probe also detected fragments of 10.0, 6.6 and 3.3 kb in BglI digested C57BL/6J DNA and fragments of 10.0, 8.2, 6.6, 3.3 and 2.9 kb in BglI digested *M. spretus* DNA. A description of the probes and RFLPs for the loci linked to P-cadherin including metallothionein-1 (*Mt-1*) haptoglobin (*Hp*) and adenosine

Figure 3. Comparisons of the exon/intron pattern of the mouse P-cadherin and the chicken L-CAM genes. The amino acid sequences, represented by single letter codes, of the mouse P-cadherin (P) and the chicken L-CAM (L) are aligned according to their optimal intergenic and intragenic similarity. Some of typical repeated sequences are boxed, and exon boundaries are indicated by arrowheads. Exon numbers for P-cadherin (P) and L-CAM (L) are shown. SIG, signal peptide; PRE, precursor sequence; EC, extracellular domain; TM, transmembrane domain; CP, cytoplasmic domain; Ca, a calcium binding site; CAT, the region containing catenin binding sites.

phosphoribosyltransferase (*Aprt*) have been reported previously (20). One locus included in this study has not previously been reported for our interspecific mouse backcross. The probe for the E-cadherin (uvomorulin) locus (*F5H3*) was a 900 bp mouse cDNA (15). A 3.2 kb fragment was detected in PstI digested C57BL/6J DNA; a 5.6 kb fragment was detected PstI digested M. spretus DNA. The E-cadherin locus has previously been assigned to the central region of mouse chromosome 8 (15). The P-cadherin probe does not cross hybridize to the E-cadherin locus.

Recombination distances were calculated as described (21) using the computer program SPRETUS MADNESS developed by D. Dave (Data Management Services, Inc., Frederick, Maryland) and A.M. Buchberg (NCI-FCRDC, ABL-BRP, Frederick, Maryland). Gene order was determined by minimizing the number of recombination events required to explain the allele distribution patterns.

RESULTS AND DISCUSSION

Genomic organization of the mouse P-cadherin gene

We screened a mouse genomic library and isolated 12 independent clones. All EcoRI fragments obtained from the phage clones were subcloned, and the nucleotide sequences at each end of the fragments were determined. This sequence analysis allowed to confirm the linkage of the phage clones. Seven of 12 clones are aligned in Figure 1.

We also found that the P-cadherin gene is a single copy in the mouse genome by Southern blot analysis (data not shown).

Restriction mapping and sequence analysis showed that the mouse P-cadherin gene spreads over 45 kbp and consists of 15

exons. The whole structure of the gene and restriction sites are shown in Figure 1, and the DNA sequences of exon-intron boundaries are shown in Figure 2. All the boundaries fit with the GT-AG rule. The length of the introns varies from 128bp to 23kbp. A marked feature of this gene is that the first intron spreads over 23kbp and occupies more than half the length of the gene. The relationship between the exon-intron boundaries and the primary structure of the protein is described in Figure 3.

Comparison with L-CAM gene structure

The structure of the mouse P-cadherin gene was compared with that of chicken L-CAM, a member of the cadherin gene family, reported by Sorkin et al. (13). The size of these genes was quite different. While the P-cadherin gene is 45kb long, the L-CAM gene is only 9kbp. However, it was found that, in spite of such a difference, all the exon-intron boundaries were conserved between the two cadherin genes, except that the P-cadherin gene has 15 exons whereas L-CAM has 16 exons (Figure 3). The first P-cadherin exon covers the regions corresponding to the first and the second exons of L-CAM. To confirm the number of exons at the 5' end of the P-cadherin gene, we isolated more cDNA clones by screening a cDNA library with a 5'-probe. The longest cDNA clone isolated had a 791bp 5'-untranslated sequence. We found that this sequence is included in the first exon, and thus could not find any other introns in more upstream regions.

The conservation of the exon-intron boundaries between the P-cadherin and L-CAM genes suggests that this pattern may be associated with the domain structure of cadherin molecules. It appears that introns do not split amino acid sequences of the putative functional domains so far identified. The amino acid

sequence DADD at position 280 to 283 in E-cadherin (uvomorulin) was recently identified as a Ca^{2+} binding domain (22), and the DXDD sequence is preserved in all the cadherins identified, in which X represents A or E. The corresponding sequence is located within the 6nd exon of the P-cadherin gene (Figure 3). The cleavage site for generating the mature form of cadherins at their amino terminus (23) is located in the center of the P-cadherin 3rd exon. Cadherins are associated with cytoplasmic components, called catenins, and the deletion of 64 amino acids from a region of the cytoplasmic domain of E-cadherin abolishes this activity (24–27). The P-cadherin 15nd exon roughly corresponds to this catenin-binding region (Figure 3). The genomic organization of the cadherin genes therefore might reflect the conserved motifs of these proteins, as discussed for other proteins (28).

We recently found that the mouse N-cadherin gene also has the same exon-intron boundaries (S. Miyatani and M. Takeichi, in preparation). This finding further supports the significance of the conserved exon-intron pattern in cadherin gene organization.

Sequence analysis of cadherins indicates that the extracellular domain contains four units of repeated sequences (Figure 3), of which the first and second repeats most resembles each other (5). However, the pattern of exon-intron boundaries does not correspond to such repeated sequences. For considering the significance of the exon-intron pattern in more detail, it would be necessary to obtain more information about the function of each domain of cadherin molecules.

It should also be emphasized that the overall pattern of introns of the P-cadherin gene is similar to that of L-CAM. The first intron of the mouse P-cadherin gene spreads over half the gene, spanning 23kb, as described above, and the corresponding intron of chicken L-CAM is 3.5kbp long and occupies more than one third of the 9kbp L-CAM gene. Our recent results indicate that the first intron of the P-cadherin gene contains enhancer activities (M. Hatta and M. Takeichi, in preparation). Therefore, the long size of this intron might have to do with transcriptional regulation of this gene.

Chromosomal localization of the P-cadherin gene

The mouse chromosomal location of the P-cadherin locus (*Pcad*) was determined by interspecific backcross analysis using progeny derived from matings of [(C57BL/6J $\times$ *Mus spretus*)F₁ $\times$ C57BL/6J] mice. This interspecific backcross mapping pane 1 has been typed for over 660 loci that are well distributed among all the autosomes as well as the X chromosome. C57BL/6J and *M. spretus* DNAs were digested with several enzymes and analyzed by Southern blot hybridization for informative restriction fragment length polymorphisms (RFLPs) first using the P-cadherin probe. A 3.1 kb *M. spretus*-specific PvuII fragment and the 8.2 and 2.9 kb *M. spretus*-specific BglI fragments were followed (see Materials and Methods). Analysis of the backcross progeny indicated that *M. spretus* specific PvuII and BglI RFLPs were not fixed in the *M. spretus* mice used to establish the F₁ hybrid animals. Consequently, only the 50 animals inheriting the *M. spretus* specific RFLPs were scored in this analysis. The two BglI RFLPs cosegregated and all the mapping results indicated that *Pcad* is located in the central region of mouse chromosome 8 linked to the *Hp* locus and the E-cadherin locus (*Ecad*). The additional loci included in this analyses are *Mt-1* and *Aprt*. The segregation analysis for the 110 mice typed in common for the loci linked to *Pcad* is shown in Figure 4. Up to 122 mice were typed for some markers. Each locus was analyzed in pairwise

Figure 4. Position of the *Pcad* locus on mouse chromosome 8. *Pcad* was placed on mouse chromosome 8 by interspecific backcross analysis. The segregation patterns of the flanking genes in 110 backcross animals that were typed in common is shown at the top of the figure. For individual pairs of loci, more than 110 animals were typed (see text). Each column represents the chromosome identified in the backcross progeny that was inherited from the (C57BL/6J $\times$ *M. spretus*)F₁ parent. The shaded boxes represent the presence of a C57BL/6J allele and white boxes represent the presence of a *M. spretus* allele. The number of offspring inheriting each type of chromosome is listed at the bottom of each column. The RFLPs associated with the *Pcad* locus were not fixed in our *M. spretus* mice at the time the backcross progeny were derived. Therefore, only those mice inheriting the RFLPs were scored. No crossovers between *Pcad*, *Ecad* or *Hp* were observed (see text for more details). A partial chromosome 8 linkage map showing the location of *Pcad* in relation to linked genes is shown at the bottom of the figure. Recombination distances between loci in centimorgans are shown to the left of the chromosome and the positions of all loci except *Pcad* in human chromosomes are shown to the right.

combinations for recombination frequencies using the additional data. The ratios of the total number of mice exhibiting recombinant chromosomes to the total number of mice analyzed for each pair of loci and the most likely gene order are: centromere-*Mt-1* -9/110-*Ecad*-0/110-*Hp*-13/122-*Aprt*. The recombination frequencies [expressed as genetic distances in centimorgans (cM) $\pm$ the standard error] are *Mt-1*-8.2 $\pm$ 2.6-[*Ecad*, *Hp*]-10.7 $\pm$ 2.8-*Aprt*. No crossovers were detected between *Ecad* and *Hp*, giving an upper 95% confidence limit of 2.7 cM between the two loci. Of the 50 mice typed for *Pcad*, 38 had been typed for *Ecad* and 32 had been typed for *Hp*. No recombination was observed between *Pcad* and *Ecad* or *Hp*, giving an upper 95% confidence limit of 7.6 cM between *Pcad* and *Ecad* and 8.9 cM between *Pcad* and *Hp*. The placement of these loci relative to other chromosome 8 markers and a comparison of the interspecific backcross map with the composite intraspecific backcross map has been reported previously (20).

The linkage of *Ecad* and *Pcad* on mouse chromosome 8 is of

particular interest since the two gene products are structurally related and the loci may have arisen by gene duplication. In addition, the distal region of mouse chromosome 8 shares a region of homology with human chromosome 16q (summarized in Ceci et al., (20). In particular, *Ecad* and *Hp* have been assigned to 16q22.1 (14, 29). The placement of *Pcad* in this region suggests that the human homolog of P-cadherin may reside on the long arm of chromosome 16, perhaps at band 22.1.

The number of members of the cadherin gene family is increasing, and they seem to be divided into subfamilies on the basis of sequence similarities (30). It should be important to determine the genomic organization and chromosomal localization of all these members for understanding of their functional and evolutionary relationships in future studies.

ACKNOWLEDGMENTS

We thank Martin Ringwald and Rolf Kemler for the E-cadherin (uvomorulin) probe, Debbie Swing, Denise Angle and Brian Cho for excellent technical assistance, and Glenn Radice for critical comments on the manuscript. This work was supported by research grants (no. 02258102 and 02262216) from the Ministry of Education, Science and Culture of Japan to M.T., and, in part, by the National Cancer Institute, DHHS, under contract NO1-CO-74101 with ABL.

REFERENCES

1. Takeichi, M. (1991) *Science*, **251**, 1451–1455.
2. Nagafuchi, A., Shirayoshi, Y., Okazaki, K., Yasuda, K. and Takeichi, M. (1987) *Nature*, **329**, 341–343.
3. Ringwald, M., Schuh, R., Vestweber, D., Eistetter, H., Lottspeich, F., Engel, J., Dolz, R., Jahhnig, F., Epplen, J., Mayer, S., Muller, C., and Kemler, R. (1987) *EMBO J.*, **6**, 3647–3653.
4. Nose, A., Nagafuchi, A. and Takeichi, M. (1987) *EMBO J.*, **6**, 3655–3661.
5. Hatta, K., Nose, A., Nagafuchi, A. and Takeichi, M. (1988) *J. Cell Biol.*, **106**, 873–881.
6. Miyatani, S., Shimamura, K., Hatta, M., Nagafuchi, A., Nose, A., Matsunaga, M., Hatta, K. and Takeichi, M. (1989) *Science*, **245**, 631–635.
7. Inuzuka, H., Miyatani, S. and Takeichi, M. (1991) *Neuron* (in press).
8. Gallin, W., Sorkin, B.C., Edelman, G.M. and Cunningham, B.A. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 2808–2812.
9. Ginsberg, D., DeSimone, D. and Geiger, B. (1991) *Development*, **111**, 315–325.
10. Ranscht, B. and Bronner-Fraser, M. (1991) *Development*, **111**, 15–22.
11. Koch, P.J., Walsh, M.J., Schmelz, M., Goldschmidt, M.D., Zimbelmann, R. and Franke, W.W. (1990) *European J. Cell Biol.*, **53**, 1–12.
12. Holton, J.L., Kenny, T.P., Legan, P.K., Collins, J.E., Keen, J.N., Sharma, R. and Garrod, D.R. (1990) *J. Cell Science*, **97**, 239–246.
13. Sorkin, B.C., Hemperly, J.J., Edelman, G.M. and Cunningham, B.A. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7617–7621.
14. Mansouri, A., Spurr, N., Goodfellow, P.N. and Kemler, R. (1988) *Differentiation*, **38**, 67–71.
15. Eistetter, H.R., Adolph, S., Ringwald, M., Simon-Chazottes, D., Schuh, R., Guenet, J.-L. and Kemler, R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 3489–3493.
16. Walsh, F.S., Barton, C.H., Putt, W., Moore, S.E., Kessel, D., Spurr, N. and Goodfellow, P.N. (1990) *J. Neurochem.*, **55**, 805–812.
17. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor.
18. Buchberg, A.M., Bedigian, H.G., Taylor, B.A., Brownell, E., Ihle, J.N., Nagata, S., Jenlins, N.A. and Copeland, N.G. (1988) *Oncogene Res.*, **2**, 149–165.
19. Jenkins, N.A., Copeland, N.G., Taylor, B.A. and Lee, B.K. (1982) *J. Virol.*, **43**, 26–36.
20. Ceci, J.D., Justice, M.J., Lock, L.F., Jenkins, N.A. and Copeland, N.G. (1990) *Genomics*, **6**, 72–79.
21. Green, E.L. (1981) *Genetics and Probability in Animal Breeding Experiments*. Macmillan, New York, pp. 77–113.
22. Ozawa, M., Engel, J. and Kemler, R. (1990) *Cell*, **63**, 1033–1038.
23. Ozawa, M. and Kemler, R. (1990) *J. Cell Biol.*, **111**, 1645–1650.
24. Nagafuchi, A. and Takeichi, M. (1989) *Cell Regulation*, **1**, 37–44.
25. Nagafuchi, A. and Takeichi, M. (1988) *EMBO J.*, **7**, 3679–3684.
26. Ozawa, M., Ringwald, M. and Kemler, R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4246–4250.
27. Ozawa, M., Baribault, H. and Kemler, R. (1989) *EMBO J.*, **6**, 1711–1717.
28. Traut, T.W. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2944–2948.
29. Westphal, E.M., Burmeister, M., Wienker, T.F., Lehrach, H., Bender, K. and Scherer, G. (1987) *Genomics*, **1**, 313–319.
30. Suzuki, S., Sano, K. and Tanihara, H. (1991) *Cell Regul.*, **2**, 261–270.