Comparative DNA Sequence Analysis of Mouse and Human Protocadherin Gene Clusters

Qiang Wu, Theresa Zhang, Jan-Fang Cheng, et al.

Genome Res. 2001 11: 389-404
Access the most recent version at doi:10.1101/gr.167301

Letter

Comparative DNA Sequence Analysis of Mouse and Human Protocadherin Gene Clusters

Qiang Wu,[1] Theresa Zhang,[2] Jan-Fang Cheng,[3] Youngwook Kim,[1] Jane Grimwood,[4] Jeremy Schmutz,[4] Mark Dickson,[4] James P. Noonan,[4] Michael Q. Zhang,[2] Richard M. Myers,[4] and Tom Maniatis[1,5]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA; [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; [3]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [4]Department of Genetics and The Stanford Human Genome Center, Stanford University School of Medicine, Stanford, California 94305, USA

The genomic organization of the human protocadherin α, β, and γ gene clusters (designated *Pcdh*α [gene symbol PCDHA], *Pcdh*β [PCDHB], and *Pcdh*γ [PCDHG]) is remarkably similar to that of immunoglobulin and T-cell receptor genes. The extracellular and transmembrane domains of each protocadherin protein are encoded by an unusually large "variable" region exon, while the intracellular domains are encoded by three small "constant" region exons located downstream from a tandem array of variable region exons. Here we report the results of a comparative DNA sequence analysis of the orthologous human (750 kb) and mouse (900 kb) protocadherin gene clusters. The organization of *Pcdh*α and *Pcdh*γ gene clusters in the two species is virtually identical, whereas the mouse *Pcdh*β gene cluster is larger and contains more genes than the human *Pcdh*β gene cluster. We identified conserved DNA sequences upstream of the variable region exons, and found that these sequences are more conserved between orthologs than between paralogs. Within this region, there is a highly conserved DNA sequence motif located at about the same position upstream of the translation start codon of each variable region exon. In addition, the variable region of each gene cluster contains a rich array of CpG islands, whose location corresponds to the position of each variable region exon. These observations are consistent with the proposal that the expression of each variable region exon is regulated by a distinct promoter, which is highly conserved between orthologous variable region exons in mouse and human.

[The sequence data described in this paper have been submitted to the GenBank/EMBL/DDB] data library under accession nos. AY013756–AY013813, AY013873–AY013878, AF332005, and AF332006.]

Cadherin superfamily proteins are calcium-dependent cell-adhesion molecules that have been implicated in tissue morphogenesis during embryonic development and in the maintenance of selective neuronal connections in the adult brain (Dreyer and Roman-Dreyer 1999; Shapiro and Colman 1999; Steinberg and McNutt 1999; Bruses 2000; Gumbiner 2000; Yagi and Takeichi 2000). Classic cadherins and protocadherins are two subfamilies within the cadherin superfamily (Suzuki 1996; Nollet et al. 2000; Wu and Maniatis 2000). Classic cadherins have five ectodomain repeats, a transmembrane segment, and a conserved cytoplasmic domain that interacts with β-catenin. In contrast, protocadherins have six or more ectodomain repeats, which are encoded by unusually large exons, and have other sequence features that distinguish them from the classic cadherins, including distinct intracellular domains (Suzuki 1996; Wu and Maniatis 2000).

A new set of mouse protocadherin cDNA clones, designated *CNR*, was previously isolated in a yeast two-hybrid screen that used the Fyn tyrosine kinase as bait. The CNR proteins are expressed at synaptic junctions in different regions of the adult brain, and individual neurons appear to express a distinct subset of *CNR* mRNAs (Kohmura et al. 1998). A remarkable feature of these protocadherin cDNAs is that the sequence of the 5′ region of each cDNA, which encodes the extracellular and transmembrane domains, differs from each other, whereas the 3′ region of each cDNA, which encodes the intracellular Fyn-interaction domain, is identical.

To investigate the mechanism of cell-specific protocadherin gene expression, we determined the genomic organization of the human protocadherin genes (Wu and Maniatis 1999; also see human genes in Fig. 1). Three closely linked human protocadherin gene clusters, designated *Pcdh*α (which are the orthologs of the mouse *CNR* genes), *Pcdh*β, and *Pcdh*γ, were identified in the 5q31 region of human chromosome 5. Remarkably, the variable 5′ region of each human protocadherin cDNA was found to be encoded by a differ-

[5]Corresponding author.
E-MAIL maniatis@biohp.harvard.edu; FAX (617) 495 3537.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.167301.

ent large exon, and the variable region exons are organized in a tandem array in each gene cluster. Sequence comparisons of genomic DNA and cDNAs identified three small exons downstream from *Pcdh*α and *Pcdh*γ variable region exons that encode the common 3′ region of protocadherin cDNAs and were therefore designated *Pcdh*α and *Pcdh*γ constant region exons. Surprisingly, the *Pcdh*β gene cluster does not contain constant region sequences. Thus, all *Pcdh*β genes consist of a single exon that encodes the extracellular, transmembrane and short cytoplasmic domains of the protein. Further studies revealed that each of the variable region exons of *Pcdh*α and *Pcdh*γ gene clusters is independently spliced to the respective three constant region exons. Therefore, all of the protocadherin proteins encoded in the *Pcdh*α and *Pcdh*γ gene clusters have similar but non-identical N-terminal extracellular and transmembrane domains, whereas the identical C-terminal cytoplasmic domains within each cluster are encoded by the constant region exons unique to each cluster. This variable and constant region organization of Pcdhα and Pcdhγ proteins suggests that diverse extracellular signals could converge on a single cytoplasmic signal transduction pathway. We noted that the organization of the *Pcdh*α and *Pcdh*γ gene clusters is strikingly similar to that of both the immunoglobulin (*Ig*) and T-cell receptor (*TCR*) gene clusters (Wu and Maniatis 1999). Comparison of genomic and cDNA sequences of *Pcdh*α and *Pcdh*γ genes suggests that the patterns of cell-specific expression of individual protocadherin protein are established by a novel mechanism. Subsequently, an almost identical organization was reported for the mouse *CNR* (mouse *Pcdh*α) gene cluster (Sugino et al. 2000).

A puzzling feature of the human *Pcdh*α and *Pcdh*γ gene clusters is the presence of variable region exons near the end of the two gene clusters that are more similar to each other than to the other variable region sequences within each cluster. These exons were designated *Pcdh*α-*C1* and -*C2* in the *Pcdh*α gene cluster, and *Pcdh*γ-*C3*, -*C4,* and -*C5* in the *Pcdh*γ gene cluster. In contrast, the *Pcdh*β gene cluster does not have a C-type protocadherin variable region sequence. All members of the *Pcdh*β gene cluster are very similar to each other and have features distinct from members of the *Pcdh*α and *Pcdh*γ gene clusters (Wu and Maniatis 1999).

Protocadherin genes are expressed in specific regions of the brain (Kohmura et al. 1998; Hirano et al. 1999a,b; Yamagata et al. 1999; Redies 2000), and they have been proposed to be a part of the molecular code for establishing and maintaining specific neuronal connections in the brain (Hagler and Goda 1998; Dreyer and Roman-Dreyer 1999; Serafini 1999; Shapiro and Colman 1999; Wu and Maniatis 1999). An understanding of the mechanism of cell-specific protocadherin gene expression may therefore provide insights into the specificity of neuronal cell–cell connections during development and in response to cognitive and sensory inputs. On the basis of the unusual genomic organization of protocadherin gene clusters, we proposed four models for the cell-specific expression of protocadherins, which included a cell-specific DNA rearrangement, and *cis*- or *trans*- alternative splicing mechanisms (Wu and Maniatis 1999).

Here we report the complete DNA sequence of the mouse protocadherin gene clusters on chromosome 18 and present a comparative analysis of the mouse and human protocadherin gene clusters. This sequence comparison provides insights into the mechanism of protocadherin gene expression, and the mouse sequence will provide information necessary for studies in the more experimentally tractable mouse model. We have identified ~60 mouse protocadherin genes in this region, and find that the overall organization of the mouse and human *Pcdh*α and *Pcdh*γ gene clusters is essentially identical (Fig. 1). However, the mouse *Pcdh*β gene cluster has six more genes than the corresponding human *Pcdh*β cluster. Comparative analysis of intergenic regions revealed sequences upstream of each variable region exon that are highly conserved between human and mouse, but less conserved between genes within each gene cluster in either human or mouse. In addition, the pattern of CpG island distribution corresponds with that of variable region exons. These observations suggest that each variable region exon is transcribed from its own promoter.

RESULTS

Genomic Organization of the Mouse Protocadherin Gene Clusters on the 18c Region of Chromosome 18

Based on the organization of human protocadherin gene clusters in the 5q31 region of chromosome 5, and available mouse cDNA and EST sequence information from GenBank, we designed 19 pairs of PCR primers to amplify genomic DNA containing the homologous mouse protocadherin genes. We used these primers to screen a mouse BAC genomic DNA library (RPCI-23), and isolated 21 BAC clones containing sequences of the mouse protocadherin gene clusters. From the restriction maps of these BAC clones, seven minimally overlapping clones were selected for DNA sequencing (RPCI-23_193o23, 6p18, 72c14, 92d17, 161o8, 56b11, and 19k11) (Fig. 1A). The total extent of genomic DNA included in the seven BACs (excluding the overlapping regions) was estimated by pulse-field gel electrophoresis to be ~1MB. All seven clones were mapped by fluorescence in situ hybridization (FISH) to the 18c region of mouse chromosome 18, which is homologous to the 5q31 region of human chromosome 5.

Analysis of the mouse genomic DNA sequences revealed 14 *Pcdh*α genes that are highly similar to the

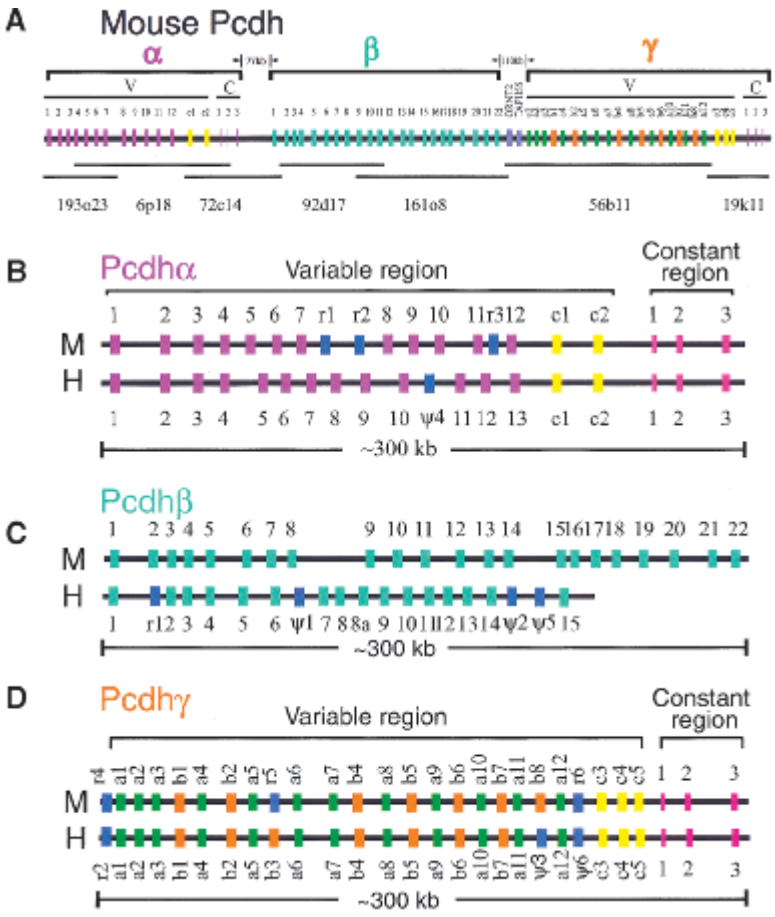


Figure 1 Comparison of the organization of mouse and human protocadherin gene clusters. Shown are the genomic organization of three closely linked mouse protocadherin gene clusters (*A*) and comparisons of the genomic organization of mouse and human *Pcdh*α/*CNR* (*B*), *Pcdh*β (*C*), and *Pcdh*γ (*D*) gene clusters. The BAC clones used in the sequence analysis are shown below (*A*). The length of sequences between clusters is also shown in (*A*). Each gene family contains multiple tandem variable region exons indicated by a vertical color bar: (mauve) Pcdhα variable region exons; (turquoise) Pcdhβ genes; (orange) *Pcdh*γ-b variable region exons; (green) *Pcdh*γ-a variable region exons; (yellow) C-type *Pcdh* variable region exons (present in both the *Pcdh*α and *Pcdh*γ gene clusters); (blue) relic or pseudogene variable region sequences (present in all three gene clusters); (pink) constant region exons. Abbreviations: Pcdh, protocadherin; V, variable region; C, constant region; M, mouse; H, human; r, relic; Ψ, pseudogene.

human *Pcdh*α genes. The variable region exons of these mouse *Pcdh*α genes are organized in a tandem array spanning a region of 250 kb mouse genomic DNA. Like the human protocadherin gene clusters, the constant region of the mouse *Pcdh*α gene cluster is organized into three small exons located downstream from the variable region tandem array (Fig. 1B). Following the *Pcdh*α gene cluster there is a second cluster of mouse *Pcdh*β genes, which is followed in turn by a third cluster of *Pcdh*γ genes. Like the human *Pcdh*β gene cluster, no constant region exons were found for the mouse *Pcdh*β gene cluster (Fig. 1C). However, three small constant region exons are located downstream of the mouse *Pcdh*γ variable region exons (Fig. 1D). Thus, the overall genomic organization of the three protocadherin gene clusters is highly conserved between mouse and human. In total, we identified ~60 protocadherin genes in this region. The upstream and downstream limits of the gene clusters were defined by the presence of a histidyl-tRNA synthetase homologous gene (O'Hanlon et al. 1995) upstream of the variable region exon of *Pcdh*α*1*, and a nonsyndromic deafness (*diaphanous*) gene (Lynch et al. 1997) downstream from the *Pcdh*γ constant region exon 3. These noncadherin genes are also conserved between human and mouse.

Comparison of the Organization of the Mouse and Human *Pcdh*α Gene Clusters

Sequence analysis of the genomic DNA containing the mouse *Pcdh*α genes revealed 14 large variable region exons encoding the protocadherin extracellular and transmembrane domains highly similar to those of the human Pcdhα proteins (Fig. 1B). Sequencing of the cDNA fragments of all mouse *Pcdh*α genes confirmed the consensus splice sites at the ends of all 14 variable region exons (Fig. 2A). The first 12 mouse *Pcdh*α genes are highly similar to each other, and eight of them are identical to the previously cloned mouse protocadherin genes (Kohmura et al. 1998). The last two mouse *Pcdh*α genes (*Pcdh*α*-C1* and *-C2*) are highly similar to the last two human *Pcdh*α genes. Like the corresponding human genes, mouse *Pcdh*α*-C1* and *-C2* genes are more similar to each other than to the 12 upstream *Pcdh*α genes. Similar to the organization of human *Pcdh*α constant region, the three small mouse *Pcdh*α constant region exons are located ~10 kb downstream from the last variable region exon. The constant region exons of *Pcdh*α are highly conserved between mouse and human. Specifically, the nucleotide sequences of constant region exons 1, 2, and 3 are 92%, 99%, and 89% identical between mouse and human, respectively. Moreover, both human and mouse *Pcdh*α constant regions have two alternatively spliced forms (Sugino et al. 2000).

Although there is one less variable region exon in the mouse *Pcdh*α gene cluster, as compared to human, the gene order is essentially conserved between mouse and human (Fig. 1B). However, the distance between some orthologous genes in mouse is very different from that in human. For example, the distance be-

A

	5′ splice site
mPcdhα1	GCAGAAGTAAATTCAGATCTTTCTGGTAAT **GTAAGT**CCAA
mPcdhα2	CCACCCTTAGAGGCCGAATACTTAGGAAAG **GTAGGT**CTTT
mPcdhα3	CTGGTTGGAGACATTGATTTCTCCATCAAA **GTGAGT**AAAT
mPcdhα4	TTGCAGTCTGCAGAGGATTCCTCTGGAAAG **GTTAGT**TTTA
mPcdhα5	CTACCTCAGGGACCCAGCTCTACAGAGAAT **GTGAGT**TTTG
mPcdhα6	CAGGATTTGAACGACGATCATTGCTCAAGA **GTAAGT**TCAA
mPcdhα7	CTACCTCAGGGTCCCAGCTCTACAGATAAC **GTGAGT**CTTA
mPcdhα8	CCATCTGTTTCTTTGGACTCCTCCGAGAAG **GTGAGT**TGTT
mPcdhα9	CATTCTGTTGGAGGAGACGTGCCCGGAAAG **GTGGGC**TATT
mPcdhα10	AGTGGAGTCCAGCAAGAGATCTTTGAGAAT **GTAAGT**ACAA
mPcdhα11	CAGGAGTCAGAGTCAAGTCACCCTGGACAG **GTGAGT**TTTC
mPcdhα12	AGAGAAAGGCAGGTAGAACATTTGAAAGAG **GTAAGT**TCAT
mPcdhα-c1	AAAGGGGATCATTCAAATGTGGAAGCCGTG **GTAAGC**ACAT
mPcdhα-c2	CTAAAAAATGATGCTGGCTCTCAAAATGAG **GTGAGA**TGGT
Consensus	AG **GTAAGT**

B

	5′ splice site
mPcdhγ-a1	CTTGAAGATAAAGAGGAAATATTTTCTCAG **GTAAAT**TTGT
mPcdhγ-a2	CTTGATGATAAAAGAGAAGAAACCCCTCAG **GTAAAA**AAAA
mPcdhγ-a3	GAAACGAAAGAAGACCCCACGCTGCCTCAG **GTAAGT**CTAT
mPcdhγ-b1	CCCCAAGTCTCGGACGACTCTGCTTTCCAG **GTAAGT**TTCT
mPcdhγ-a4	CTTGAAATAAAAGGAGACTCCAGTCTGCAG **GTGAGT**GACT
mPcdhγ-b2	GTTCCGTTTGTCTCGGATTCTATCTCAAAG **GTGAGC**TTCA
mPcdhγ-a5	ACACACAAAGAAGAGCCCGGAGATGCTCAG **GTCAGT**TTTC
mPcdhγ-a6	TTTTGCAAAGAGGAAGACTCTCTTGATCAG **GTAAGG**TTAT
mPcdhγ-a7	GAATGTAAGGGTGAAGCCCCAAGTTCCCAG **GTGAGT**TAAT
mPcdhγ-b4	TGTAATTCCAATGAGTCAACCTCCCATCAG **GTAAGT**TTCC
mPcdhγ-a8	GAGAATAAGGATGAAGATGCTTGCGCTCCG **GTAAGT**TCCA
mPcdhγ-b5	CCGACTTCCCACCCCGAACCTCTAACACCG **GTGAGT**TCCA
mPcdhγ-a9	TTTCCTATTGATGACACTCCTTTGGTTCCT **GTGAGT**TCTG
mPcdhγ-b6	TTGACAACACATCCTGAGACTCTAACACCG **GTAAGT**TTCA
mPcdhγ-a10	TGTCCTGTAGAAGACGCTCCTTTGGTGCCA **GTGAGT**TCTG
mPcdhγ-b7	GTTGAAGCAGACGAGAAGACCTTTAACCAG **GTATTT**AATG
mPcdhγ-a11	TTAGGCAAATGTGAACCGACAGATATTCAG **GTGAGA**GATA
mPcdhγ-a12	TTTTCTAAAGACAATCATGCATTAAATCAG **GTGAGT**CTAT
mPcdhγ-c3	TTGGGTGCAGAGAGCGCCCCCCCTGGACAG **GTAAGT**GTTA
mPcdhγ-c4	CCACCCTCTGATCTTCTCTATGGGCTAGAG **GTGAGA**CCTT
mPcdhγ-c5	CGCTCTAGTACGCTGCGGGAGCGGAGCCAG **GTGAGG**GCTG
Consensus	AG **GTAAGT**

Figure 2 Alignments of variable region 5′ splice sites of mouse *Pcdh*α (*A*) and *Pcdh*γ (*B*) gene clusters. The 5′ splice site sequences are shown in bold, with the consensus below each panel.

tween mouse *Pcdhα4* and *Pcdhα5* genes is only 5 kb in contrast to the large 12 kb intergenic region between the corresponding human genes. Three "relic" sequences were identified in the mouse *Pcdh*α gene cluster, and only one pseudogene was identified in the corresponding human cluster. Relics are defined as sequence fragments with only limited similarity to the corresponding functional genes (Rowen et al. 1996). In contrast, pseudogenes show more extensive sequence similarity but are rendered nonfunctional by mutations.

Comparison of the Organization of Human and Mouse *Pcdh*β Gene Clusters

Sequence analysis of the genomic DNA downstream from the mouse *Pcdh*α gene cluster revealed a large exon located ~77 kb downstream from the last Pcdhα constant region exon (Fig. 1A).

This single large exon encodes an 818aa protein containing a signal peptide, six typical protocadherin ectodomains, a transmembrane segment, and a short cytoplasmic domain. The encoded protein is highly similar to the human *Pcdhβ1* protein: 88% identity and 92% similarity with no gaps over the entire length. Thus, we designated this gene mouse *Pcdhβ1*. Following the mouse *Pcdhβ1* gene, there are 21 additional *Pcdh*β genes that are more similar to the human *Pcdh*β genes than to the human *Pcdh*α and *Pcdh*γ genes. We have therefore designated these genes mouse *cdhβ2*–*Pcdhβ22* (Fig. 1C). We previously identified 15 *Pcdh*β genes in the human *Pcdh*β locus. We have now isolated a clone (CTD-2130B15) that covers the gap between the human *Pcdhβ8* and *Pcdhβ9* genes, and found that the gap sequence contains only one additional *Pcdh*β gene (therefore designated *Pcdhβ8a*). Thus, mouse has six more *Pcdh*β genes than human does, and the *Pcdh*β locus is expanded in mouse compared to that in human (Fig. 1C).

The predicted amino acid sequences of the mouse Pcdhβ proteins are more similar to each other than to the mouse Pcdhα or Pcdhγ proteins. The Pcdhβ proteins have highly conserved extracellular and transmembrane domains. The nucleotide and amino acid sequences in the region around the transmembrane domains of Pcdhβ proteins are almost identical, and these proteins have a very short cytoplasmic domain. In contrast to the *Pcdh*α and *Pcdh*γ gene clusters, neither mouse nor human *Pcdh*β gene clusters contain constant region exons. Moreover, all of the *Pcdh*β EST and cDNA clones currently in the GenBank database correspond to unspliced mRNAs. Therefore, Pcdhβ proteins do not appear to contain a common C-terminal intracellular domain. However, we noted that a highly conserved 5′ splice site is located at the end of most *Pcdh*β variable region exons (Wu and Maniatis 1999), and this splice site is conserved between mouse and human (data not shown). Thus, it seems likely that the conserved *Pcdh*β 5′ splice sites do function. However, neither the cell type in which this splicing occurs nor the target 3′ splice site has been identified.

Identification of Two Noncadherin Genes between the *Pcdh*β and *Pcdh*γ Gene Clusters

Both the mouse and human protocadherin gene clusters are interrupted by two noncadherin-like genes located between the *Pcdh*β and *Pcdh*γ gene clusters. The first gene is an ornithine transporter gene (*ORNT2*), and the second gene encodes a component (*TAFII55*) of the human TFIID complex (Fig. 1A). The coding regions of both genes are located on the opposite strand that encodes the protocadherins. The mitochondrial ornithine transporter 1 (*ORNT1*) gene, which is defec-

tive in hyperornithinemia–hyperammonemia-homocitrullinuria syndrome, had been previously mapped to human chromosome 13q14 (Camacho et al. 1999). The human *ORNT2* gene is a paralog of the human *ORNT1* gene and has a full-length coding region. However, the corresponding mouse *ORNT2* gene has a single nucleotide deletion near the 5′ end of the coding region. This single nucleotide deletion is not a consequence of sequencing error, because three genome-sequencing centers independently determined the same sequence. Thus, the mouse *Ornt2* gene may be a pseudogene as a consequence of a very recent mutation. Alternatively, a second methionine codon located 107 nucleotides downstream from the first one may actually be the translational start codon. If so, the single nucleotide deletion in the mouse sequence would not inactivate the gene. Both human and mouse genes are transcribed because they have numerous EST matches in the database. The *TAFII55* gene, which encodes a subunit of TFIID complex (Chiang and Roeder 1995), consists of a single exon located between the *Pcdh*β and *Pcdh*γ gene clusters in both mouse and human.

Comparison of the Organization of Human and Mouse *Pcdh*γ Gene Clusters

DNA sequence analysis identified 22 mouse *Pcdh*γ variable region exons and three small constant region exons in the region downstream from the *Pcdh*β gene cluster (Fig. 1A,D). One of the mouse *Pcdh*γ genes is identical to previously cloned protocadherin 2C gene (Hirano et al. 1999a). Sequencing of cDNAs spanning the splice sites between variable and constant regions confirmed that cDNA fragments of all mouse *Pcdh*γ genes share an identical constant region sequence. Thus, each variable region exon is independently spliced to the first constant region exon. Comparison of the sequences of cDNAs with those of the genomic DNA identified a consensus splice site downstream from each variable region exon (Fig. 2B).

The organization of mouse *Pcdh*γ gene cluster is essentially the same as that of human *Pcdh*γ gene cluster (Fig. 1D). Both have >20 variable region exons and both have three downstream constant region exons. The constant region exon sequences are highly conserved between mouse and human. Specifically, constant region exons 1, 2, and 3 have 95%, 90%, and 80% identity, respectively, between mouse and human at the nucleotide level. In addition, we found that each of the mouse *Pcdh*γ genes has the corresponding orthologous human gene except the mouse *Pcdh*γ-*b8* gene, whose orthologous gene is the human *Pcdh*ψ*3* gene. Moreover, the mouse has a relic sequence at the location corresponding to the human *Pcdh*γ-*b3* gene. Similar to the *Pcdh*α gene cluster, the last three *Pcdh*γ genes (*C3*, *C4*, and *C5*) are conserved between mouse and human (Fig. 1D). All five mouse C-type protocadherin genes, *C1* and *C2* in the *Pcdh*α cluster and *C3*, *C4*, and *C5* in the *Pcdh*γ cluster, are similar to each other and are distinct from other members in the clusters.

Evolutionary Relationships among Members of the Human and Mouse *Pcdh*α, *Pcdh*β, and *Pcdh*γ Genes

The proteins encoded by the protocadherin loci in human and mouse are highly similar. The evolutionary relationships between human and mouse *Pcdh*α genes are displayed in Figure 3A. The phylogenetic tree shows that most individual *Pcdh*α genes are orthologous between human and mouse. Thus, it is likely that each Pcdhα protein has a distinct, highly conserved function. However, the human *Pcdh*α*7* and *Pcdh*α*9*, and the mouse *Pcdh*α*7* and *Pcdh*α*8* genes are paralogous, and the four genes are within a small branch in the tree. Therefore, the human *Pcdh*α*7* and *Pcdh*α*9*, and the mouse *Pcdh*α*7* and *Pcdh*α*8* genes are probably the consequence of duplications of their respective ancestors after divergence of primates and rodents. Moreover, human *Pcdh*α*6* and *Pcdh*α*8* are paralogous, and there is a single orthologous mouse *Pcdh*α*6* gene. This observation suggests that human *Pcdh*α*6* and *Pcdh*α*7*, and *Pcdh*α*8* and *Pcdh*α*9* are duplicated from a single ancestral gene pair. The *Pcdh*α-*c1* and *Pcdh*α-*c2* variable regions are distinct from other *Pcdh*α proteins, and their high conservation between human and mouse strongly suggests that they have specific functions distinct from those of the other *Pcdh*α genes.

The evolutionary relationships between human and mouse *Pcdh*β genes are displayed in Figure 3B. The human and mouse *Pcdh*β genes display both orthologous and paralogous relationships. For example, the human *Pcdh*β*1*, *2*, *3*, *6*, *7*, *13*, *14*, and *15* genes appear to be the orthologs of the mouse *Pcdh*β*1*, *2*, *3*, *13*, *15*, *8*, *20*, and *22*, respectively. However, three mouse *Pcdh*β genes (*5*, *7*, and *9*) are paralogous and in a small branch with the human *Pcdh*β*4* gene, and six mouse *Pcdh*β genes (*4*, *6*, *8*, *10*, *11*, and *12*) are paralogous and in a small branch with a single human *Pcdh*β*5* gene. This observation suggests that the mouse *Pcdh*β gene cluster expanded after the divergence of mouse and human.

In contrast to both *Pcdh*α and *Pcdh*β genes, members of *Pcdh*γ genes are strictly conserved between mouse and human. As shown in Figure 3C, each mouse gene and its human ortholog form a small branch in the phylogenetic tree. Therefore, members of *Pcdh*γ gene cluster are orthologous between mouse and human. However, the mouse ortholog of human *Pcdh*γ-*b3* gene has degenerated into a relic sequence, and the human ortholog of mouse *Pcdh*γ-*b8* has become a pseudogene (Fig. 1D).

The overall organization of the protocadherin

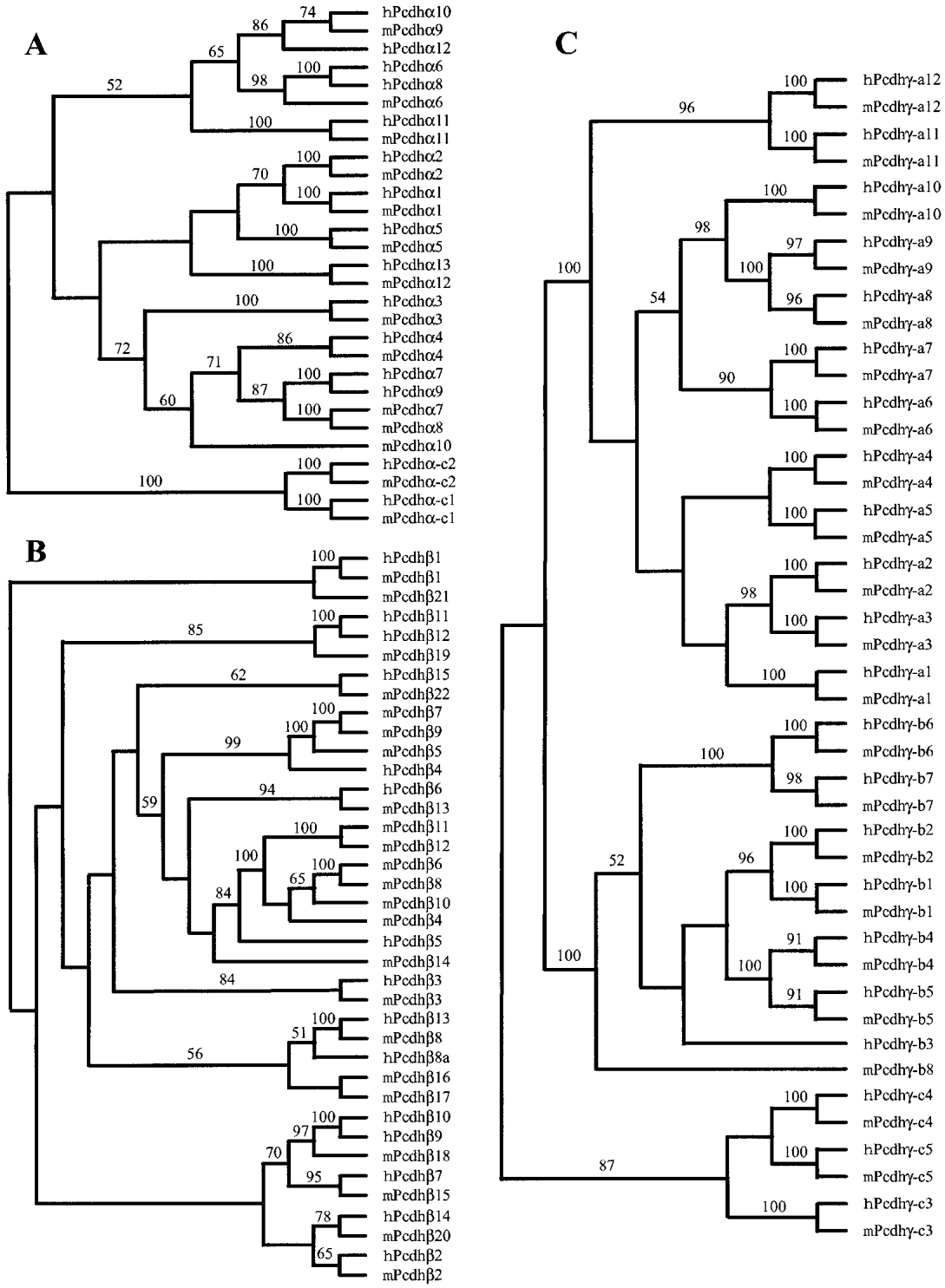


Figure 3 Phylogenetic trees of human and mouse *Pcdhα* (*A*), *Pcdhβ* (*B*), and *Pcdhγ* (*C*) gene clusters. The trees were reconstructed using the neighbor-joining method of the PAUP program. The tree branches are labeled with the percentage support for that partition based on 1000 bootstrap replicates. Only bootstrap values of >50% are shown. The unrooted trees are rooted by midpoint prior to output.

gene clusters in mouse and human is essentially the same. First, both mouse and human have three protocadherin gene clusters, in the same order and orientation (Fig.1). Second, the C-type protocadherin genes, the last two *Pcdhα* genes and the last three *Pcdhγ* genes, are more similar to each other, and are separated from

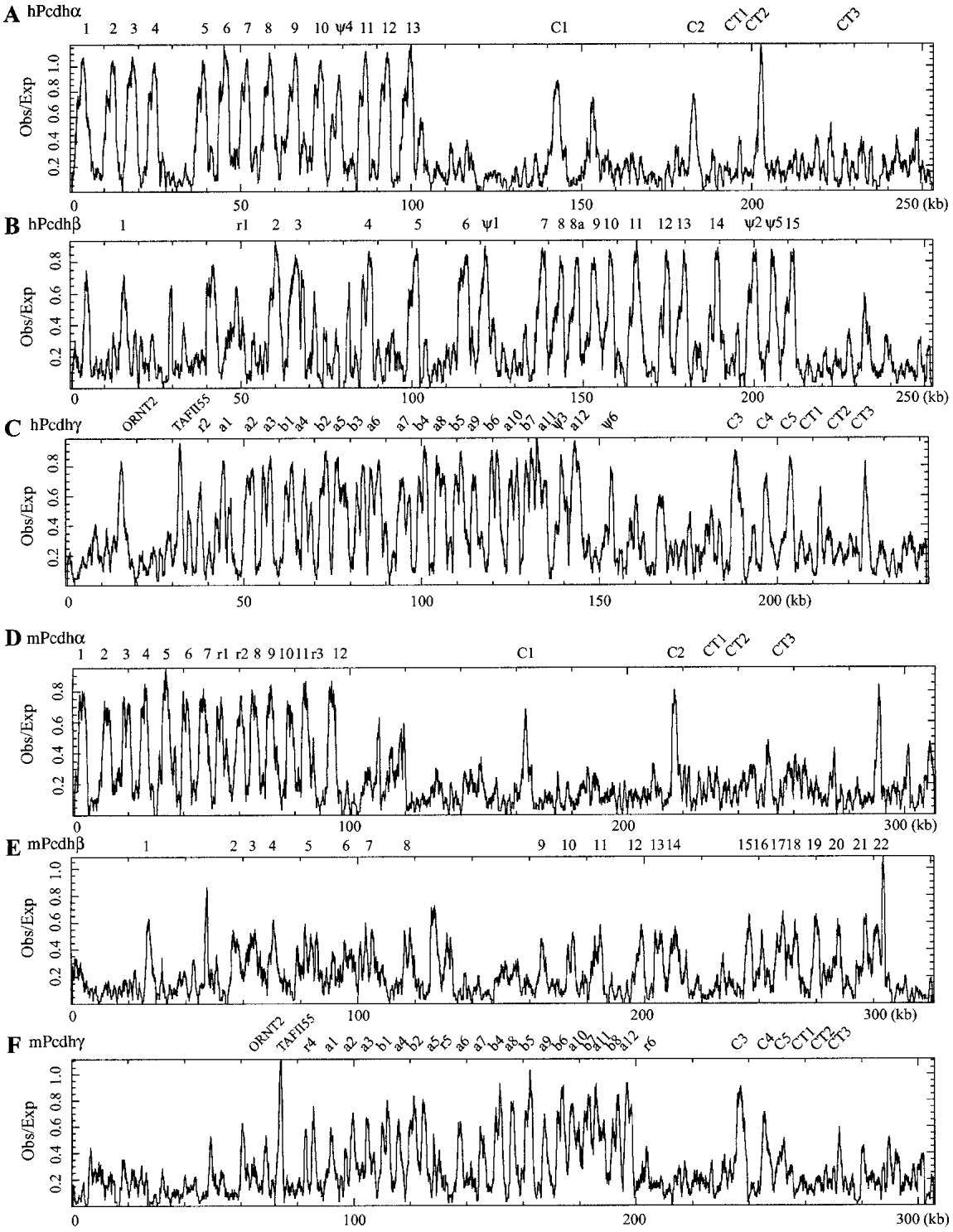
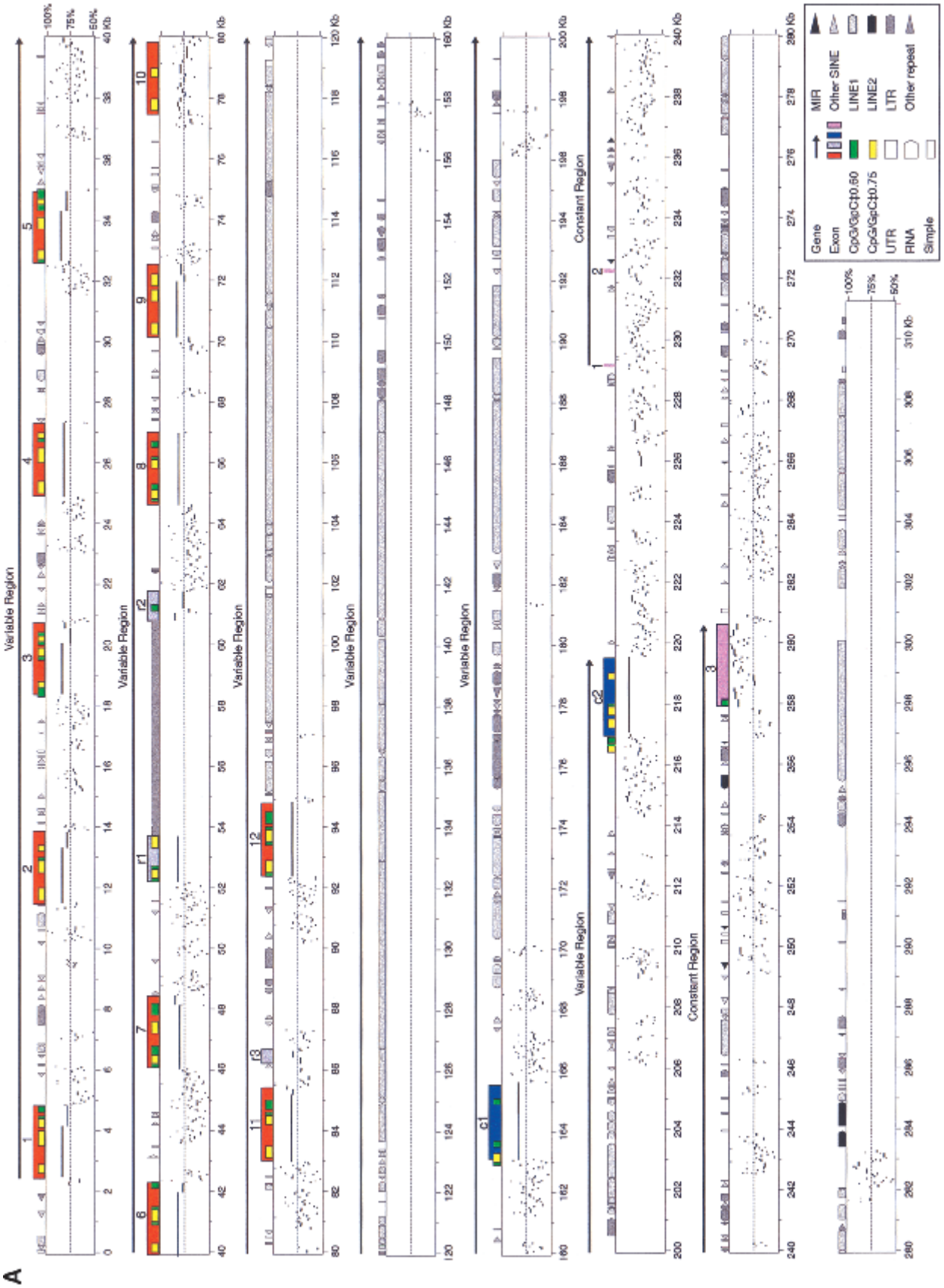
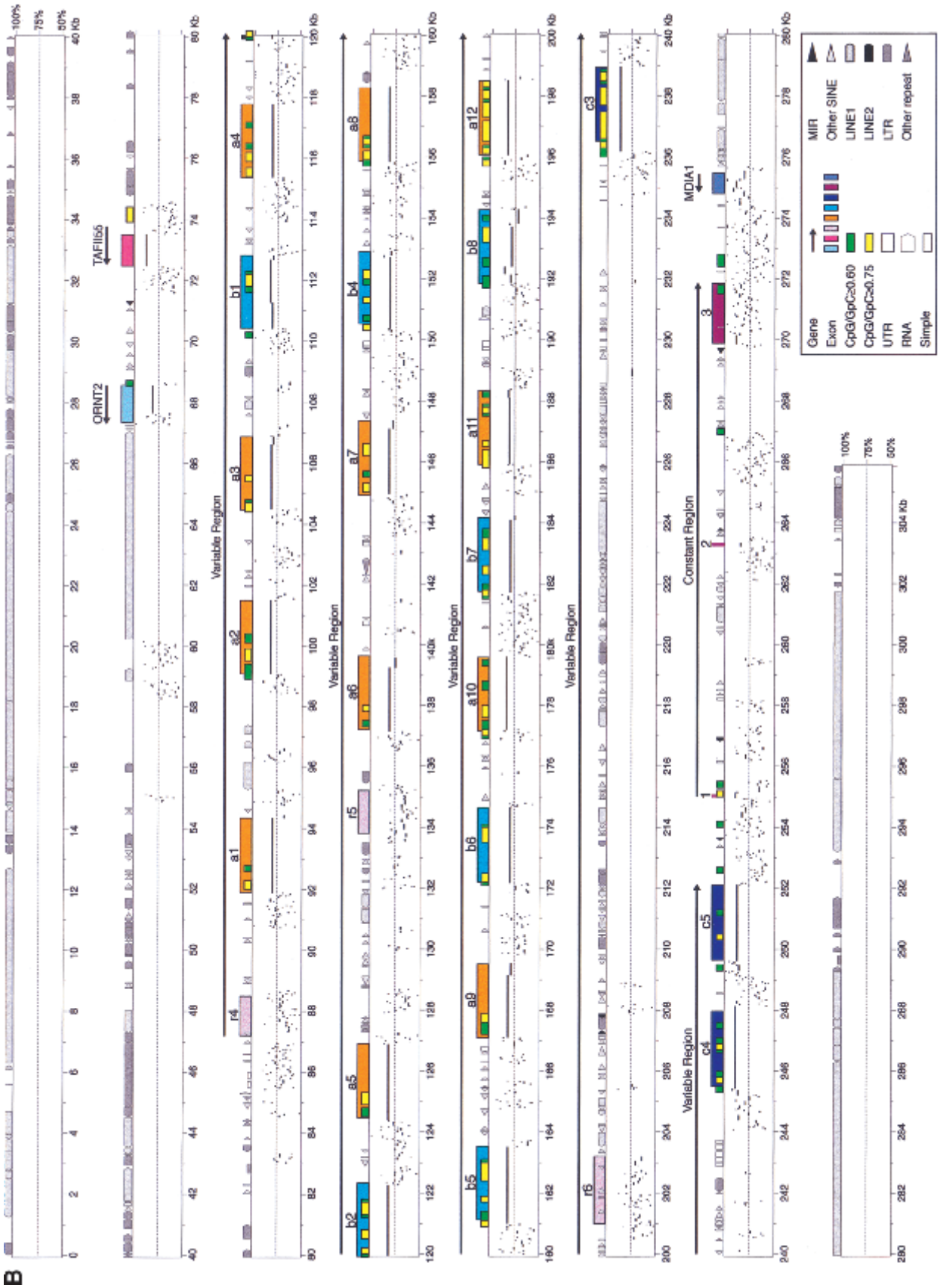


Figure 4 Distribution of CpG islands in the genomic sequences of human and mouse protocadherin gene clusters. Shown are ratios of observed to expected CpG dinucleotide frequency of a 1000 bp sliding window in the region of human *Pcdh*α (*A*), *Pcdh*β (*B*), and *Pcdh*γ (*C*) and mouse *Pcdh*α (*D*), *Pcdh*β (*E*), and *Pcdh*γ (*F*) gene clusters. The peak of ratios correlates with the position of protocadherin variable region exons but not constant region exons. The position of each variable and constant region exon is indicated at the top of each panel. (CT), constant region exon.





(See following page for legend.)

Figure 5 Percent identity plot (PIP) of the *Pcdh*α (*A*) and *Pcdh*γ (*B*) genomic sequences between mouse and human by using the PipMaker program with the chaining option. The mouse genomic sequences are shown on the *x*-axis, and the percentage sequence identities (50%–100%) are shown on the *y*-axis. Annotation of the mouse sequences is illustrated at the top of the sequences by solid color boxes. The repeats of mouse sequence are depicted as follows: (black pointed boxes) LINE2s; (light gray pointed boxes) LINE1s; (dark gray pointed boxes) LTRs; (black triangles) MIRs; (light gray triangles) SINEs other than MIRs; (dark gray triangles) other repeats; (white boxes) simple repeats. Short yellow boxes are CpG islands where the ratio of CpG/GpC exceeds 0.75, and short green boxes are CpG islands where the ratio of CpG/GpC is between 0.60 and 0.75. (*MDIA1*) the last exon of mouse *diaphanous* gene *1*.

corresponding upstream genes by a very large intergenic region (>40 kb) in both mouse and human (Fig. 1B and 1D). Third, the members of the *Pcdh*α and *Pcdh*γ gene clusters are strikingly conserved in both gene order (Fig. 1B,D) and gene sequences (Fig. 3A,C). Finally, the *Pcdh*α and *Pcdh*γ gene clusters have highly conserved constant region exons between mouse and human whereas the *Pcdh*β gene cluster does not have constant region exons in both mouse and human (Fig. 1).

The Distribution of CpG Islands Corresponds to the Locations of the Variable Region Exons

At present, it is not known whether each protocadherin gene cluster is transcribed from a single promoter, or whether each variable region exon has its own promoter. Insights into this problem could be provided by examining the sequences immediately surrounding each variable region in the mouse and human protocadherin gene clusters. One characteristic shared by ~50% of mammalian promoters is the occurrence of CpG islands located near the 5′ ends of genes (Antequera and Bird 1993). Close examination of the sequences around the translation start sites of mouse and human protocadherin variable region exons revealed a high density of CpG dinucleotides, suggesting that they are CpG islands. Indeed, the sequences near the human *Pcdh*α*2*, *Pcdh*β*1*, *Pcdh*γ*-a10*, and *Pcdh*γ*-b3* translation start codons match four previously isolated CpG islands (Cross et al. 1994) (GenBank accession nos. Z65300, Z59266, Z60764, and Z58035, respectively).

We therefore searched the entire human and mouse gene clusters for CpG islands using the CpGplot program (Larsen et al. 1992). As shown in Figure 4, the ratio of observed to expected CpG dinucleotide frequency peaks at the locations of each variable region exon in both mouse and human. It is known that mouse genome lost some CpG dinucleotides after the divergence of mouse and human (Antequera and Bird 1993). Consistent with this, we note that the ratio is slightly lower in mouse than in human (comparing Fig. 4A,B,C to 4D,E,F, respectively). Nevertheless, this distribution supports the proposal that each variable region exon has its own promoter and a transcriptional start site is located upstream from each variable region exon.

Noncoding Sequence Conservation Within the Variable Region of Mouse and Human Protocadherin Gene Clusters

We used the PipMaker program (Schwartz et al. 2000) to compare sequences of the entire mouse and human *Pcdh*α and *Pcdh*γ gene clusters (Fig. 5). Interestingly, the first two relics (r1 and r2) in the mouse *Pcdh*α gene cluster appear to result from interruption of an archaic protocadherin gene by repetitive elements (Fig. 5A). Although there are many conserved intergenic sequences in the protocadherin variable region, the most striking features are the occurrence of highly conserved sequences upstream of each variable region exon (Fig. 5A,B). For example, in the *Pcdh*γ variable region, almost all conserved segments above 70% identity and longer than 100 base pairs (bp) are immediately upstream of variable coding regions.

A systematic analysis of these sequences revealed that the 5′ flanking sequences of orthologous variable region exons have a significantly higher percentage identity than the corresponding paralogous sequences within *Pcdh*α and *Pcdh*γ gene clusters in both mouse and human (Fig. 6A,B). In both the *Pcdh*α and *Pcdh*γ gene clusters, there is a peak of sequence identity at the region ~200 bp upstream from the translation start codon. In contrast, a lower level of sequence identity, which is only slightly above the baseline for random sequences, is observed in the upstream sequences between the paralogous genes within either *Pcdh*α or *Pcdh*γ gene cluster in both human and mouse (broken lines in Fig. 6A,B). We also observed that some variable region exons have a conserved element further upstream of the coding region. These results are consistent with the notion that there is a distinct promoter upstream of each variable region exon. The high level of sequence conservation upstream of variable region exons is in contrast to the sequences downstream from the variable region 5′ splice site, in which there is no conservation of sequences between the two species.

For the sequences upstream of the C-type protocadherin variable region exons, not only does each orthologous gene pair have a higher sequence identity than paralogous gene pairs, but also the conserved regions are much larger than those of other *Pcdh*α and *Pcdh*γ genes (Fig. 6C). Although there is no conserved segment above 70% identity and longer than 100 bp at the 5′ segment flanking the *C1* protocadherin gene, there are five, three, three, and two such highly conserved segments upstream of the *C2*, *C3*, *C4*, and *C5* genes, respectively. This observation suggests that the

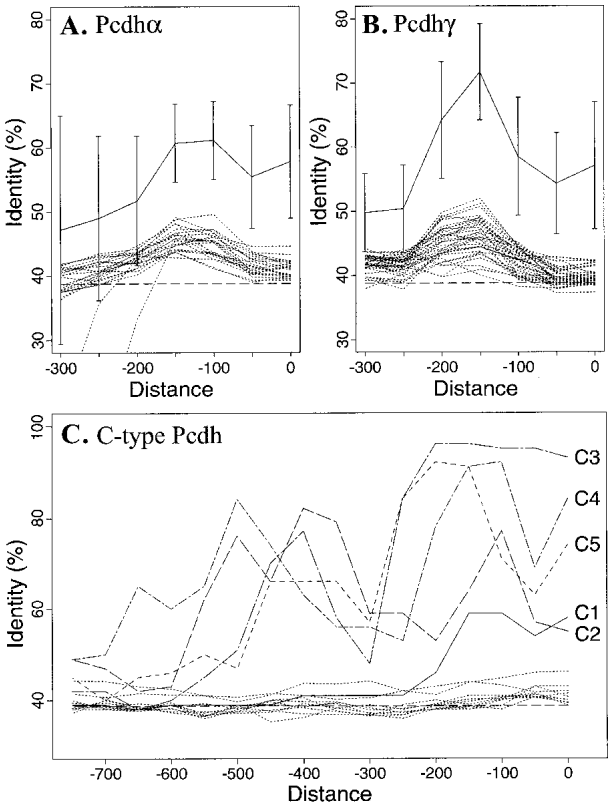


Figure 6 Upstream sequences of orthologous genes are more conserved than paralogous genes. The maximal sequence identities of all 100-bp segments within a 150-bp sliding window were computed for each gene pair. The *x*-axis represents the end position of the sliding window relative to the translation-start codon. The *y*-axis represents the percentage sequence identities. Shown are the average of 100-bp-segment maximal identities of all orthologous (solid lines with standard deviation) gene pairs in *Pcdh*α (*A*) and *Pcdh*γ (*B*) gene clusters. Also shown are the maximal identities between each gene and all the other paralogous members (excluding C-type protocadherin genes) of the same gene cluster (broken lines without standard deviation). The maximal identities for each orthologous gene pair in C-type protocadherin genes are shown individually in *C*. Note that the conserved region upstream of C-type protocadherin genes is larger than that of other protocadherin genes.

regulation of C-type protocadherins is different from that of other protocadherins.

Comparison of Protocadherin Constant Region Sequences

We noted previously that human *Pcdh*α constant region exons 1 and 2 are the same length as and similar to the corresponding *Pcdh*γ constant region exons (Wu and Maniatis 1999). The mouse *Pcdh*α constant region exons 1 and 2 are also the same length as the corresponding *Pcdh*γ constant region exons. The nucleotide sequences of mouse *Pcdh*α constant region exon 1 are 63% identical to that of the corresponding *Pcdh*γ constant region exon. The constant region exon sequences are also highly conserved between human and mouse. Specifically, the *Pcdh*α and *Pcdh*γ constant coding regions have 96% and 91% nucleotide identities between human and mouse, respectively, while the amino acid sequences are 99% identical for both Pcdhα and Pcdhγ constant regions. Therefore, the intracellular signal transduction pathway must be conserved between human and mouse.

There are many conserved noncoding segments in the constant region of both *Pcdh*α and *Pcdh*γ gene clusters, as shown by PIP plot (Fig. 5A,B). The most prominent one is a conserved sequence segment upstream of the constant region exon 1 in both *Pcdh*α and *Pcdh*γ gene clusters (Fig. 7A,B). Specifically, there is an 83% sequence identity in a 200 bp intronic region and 83% sequence identity in a 300 bp intronic region upstream of *Pcdh*α and *Pcdh*γ constant region exon 1, respectively. These regions contain ~50 continuous identical nucleotides between mouse and human (Fig. 7A,B). The functional significance of these highly conserved sequences remains to be established.

Identification of a DNA Sequence Motif Upstream of Protocadherin Variable Region Exons

Because the members of each protocadherin gene cluster are very similar to each other and upstream sequences are conserved between orthologous gene pairs in *Pcdh*α and *Pcdh*γ gene clusters, we used a version of the Gibbs sampler program called `GibbsDNA` (Z. Ioschikz and M.Q. Zhang, unpubl.) to determine whether the upstream sequences share any motif. Strikingly, there is a highly conserved sequence motif upstream of all variable region exons in each protocadherin gene cluster in both mouse and human (Fig. 8). The motif cannot be found in transcription factor binding site databases. Moreover, this motif is located at about the same distance from the translation start codon of each variable region exon (Fig. 8). In addition, we noted that there are several more nucleotides immediately upstream of this motif that appear to be conserved. We also noted that the distribution of motifs for C-type protocadherin genes is different from others, in which only the first C-type genes in both clusters (*C1* and *C3* in *Pcdh*α and *Pcdh*γ gene clusters, respectively) have the motifs. Although human *Pcdh*γ*-C4* has a weak motif, the orthologous mouse *Pcdh*γ*-C4* does not have the motif. Interestingly, both human and mouse *Pcdh*β*1* genes do not have the motif.

A careful examination of the motif from all three gene clusters revealed a common core sequence, "CGCT" (Fig. 8). Moreover, this core sequence is surrounded by additional conserved sequences that are specific for each gene cluster (Fig. 8). For example, in both human and mouse, a CC dinucleotide is found at fixed distances upstream and downstream from the core sequence in the *Pcdh*α gene cluster (Fig. 8A,D). Similarly, other cluster-specific sequences are found in

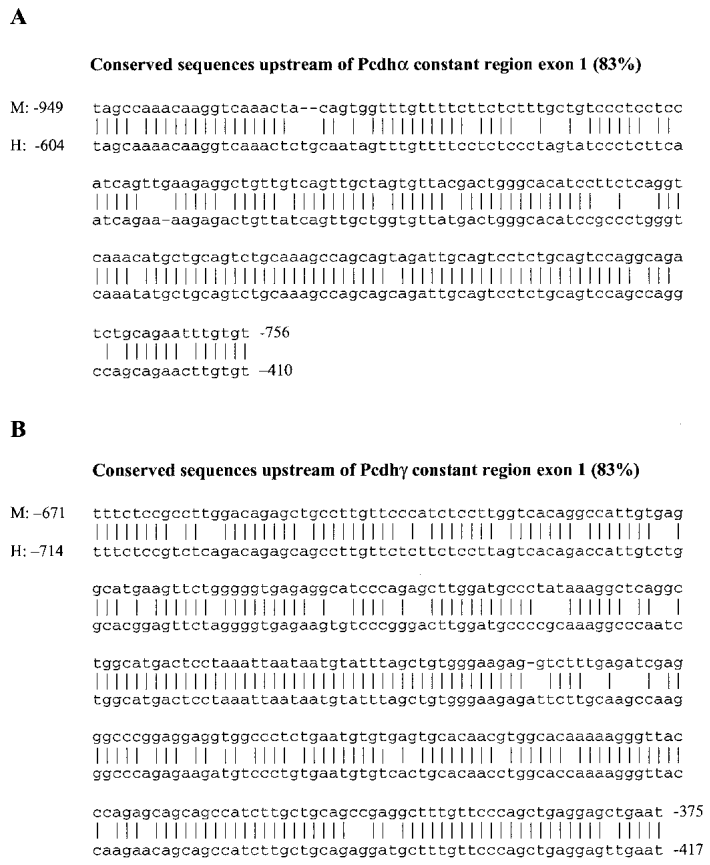


Figure 7 Conserved sequences upstream from constant region exon 1 of *Pcdh*α (*A*) and *Pcdh*γ (*B*) gene clusters. The identical nucleotides are shown by short vertical lines. The relative positions to the start nucleotide of constant region exon 1 are shown at the beginning and end of each sequence.

the *Pcdh*β and *Pcdh*γ gene clusters. This remarkable similarity of sequence motifs among genes within a cluster, and between the same clusters in human and mouse, and their striking locations in the loci strongly suggest that they are important for the regulation of protocadherin gene expression.

DISCUSSION

Protocadherins are members of the cadherin superfamily of cell-adhesion proteins (Kohmura et al. 1998; Hirano et al. 1999b; Yoshida and Sugano 1999; Kim et al. 2000; Nollet et al. 2000; Wu and Maniatis 2000). A subset of these proteins, originally designated *CNR* (*Pcdh*α), has been shown to be expressed at synaptic junctions and to display distinct patterns of cell-specific expression in different regions of the brain (Kohmura et al. 1998). The human counterparts of the CNR proteins were recently shown to be part of a larger family of proteins encoded by a cluster of genes, designated *Pcdh*α. This cluster was found to be located upstream of two additional protocadherin gene clusters, designated *Pcdh*β and *Pcdh*γ (Wu and Maniatis 1999). The striking immunoglobulin-like organization of these gene clusters suggested that novel mechanisms may be involved in the regulation of their cell-specific expression in the brain (Chun 1999; Shapiro and Colman 1999; Wu and Maniatis 1999; Yagi and Takeichi 2000). To gain insight into these mechanisms, we determined the complete DNA sequence of the corresponding mouse protocadherin gene clusters. We then performed a comparative sequence analysis to identify potential regulatory sequences involved in determining the cell-specific expression of individual variable region exons.

Interspecies comparative sequence analysis is a powerful tool for obtaining information on gene organization and regulation. To date, comparative sequencing studies have been achieved for relatively few chromosomal loci, and the conservation of noncoding sequences varies widely between different loci (Ansari-Lari et al. 1998; Jang et al. 1999; Endrizzi et al. 2000). For example, there is relatively little sequence conservation in the intergenic regions of mammalian globin gene clusters, or in the excision repair cross-complementing repair group 2 (*ERCC2*) regions between human and mouse (Lamerdin et al. 1996; Hardison et al. 1997). In contrast, there is a very high level of noncoding sequence identity (~71%) within a 100-kb region of the human and mouse T-cell receptor gene clusters (Koop and Hood 1994). We have found that the DNA sequences immediately upstream of each variable region exon are highly conserved between mouse and human orthologs (Figs. 5 and 6). A striking example of this is the 90% sequence identity within 338 bp upstream of the mouse and human *Pcdh*γ*-C3* variable region exons. Other highly conserved intergenic sequences were identified in the region between the last variable region exon and the first constant region exon. For example, one of the most conserved sequences is located approximately 500 bp upstream of the first constant region exon in both the *Pcdh*α and *Pcdh*γ gene clusters (Fig. 7).

Although interspersed repeats are considered "junk" DNA sequences, recent studies have shown that some of them may be active in modifying the genome (Moran et al. 2000). The interspersed repeats occupy 41% and 36% of the genomic sequences in the protocadherin loci in mouse and human, respectively. This is much higher than that (30%) in the human β T-cell receptor locus (Rowen et al. 1996). The number of short interspersed nucleotide elements (SINEs) is much higher than that of long interspersed nucleotide elements (LINEs), in contrast to almost equal numbers of SINEs and LINEs in the human β T-cell receptor locus (Rowen et al. 1996) and the *Bpa/Str* region (Mal-

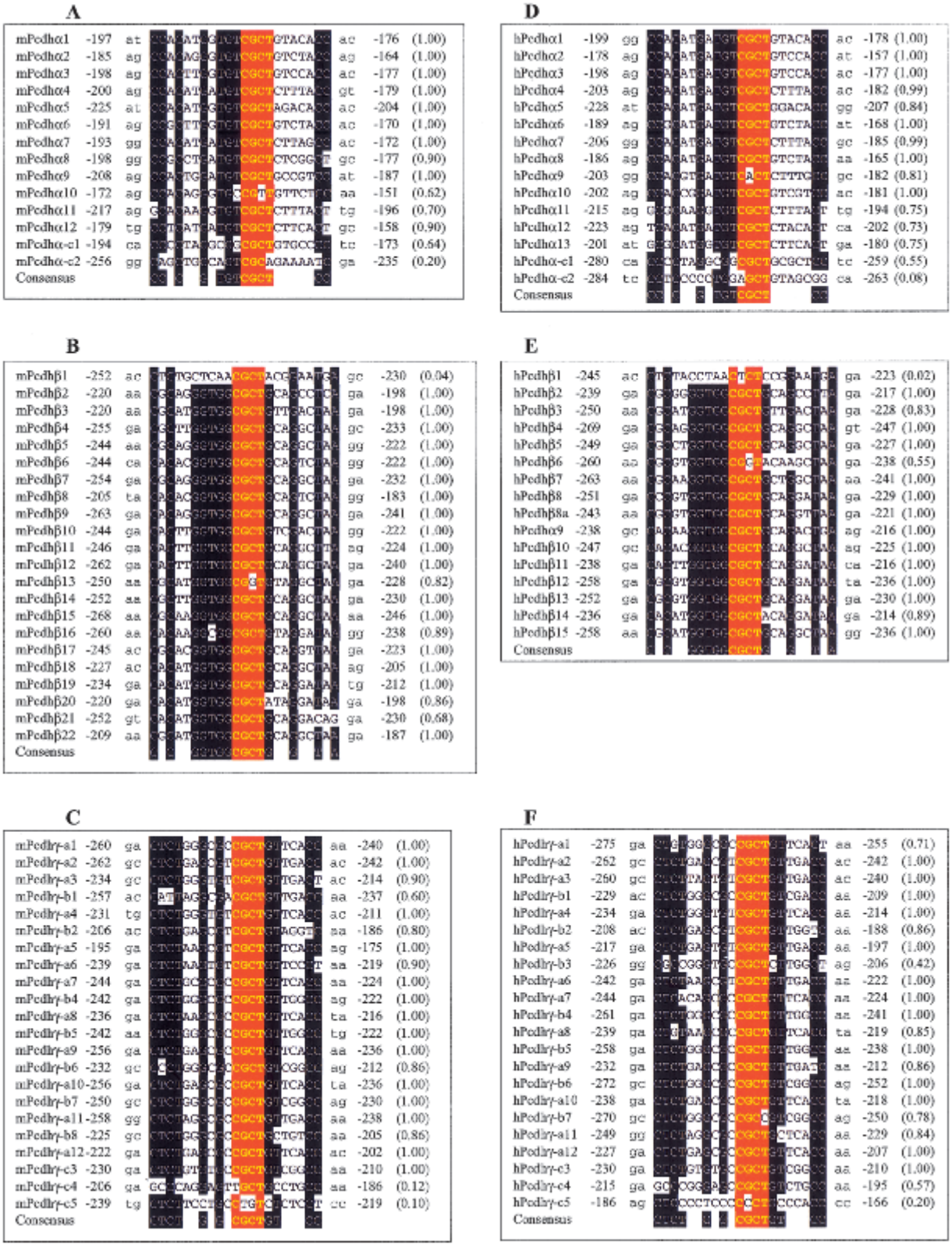


Figure 8 Alignment of conserved sequence motif upstream of protocadherin coding region. Shown are the conserved sequences and their relative positions to the translation start codon in mouse *Pcdh*α (*A*), *Pcdh*β (*B*), and *Pcdh*γ (*C*) and human *Pcdh*α (*D*), *Pcdh*β (*E*), and *Pcdh*γ (*F*) gene clusters. The probability of finding the motif within -290 to -150 nucleotides upstream of the translation start codon is shown within parentheses at *right*. The consensus sequences are shown below each panel. The conserved nucleotides are shown with white letters on a black background. The core sequences are highlighted with yellow bold letters on a red background.

lon et al. 2000). Interestingly, most of the variable region 5′ splice sites are immediately followed by repeat sequences.

Remarkably, the large regions between the first C-type protocadherin gene (*C1* and *C3* in *Pcdh*α and *Pcdh*γ gene clusters, respectively) and the other upstream variable region exons are almost entirely occupied by repeats in both mouse and human (Fig. 5). In contrast, the regions between the last C-type protocadherin gene (*C2* and *C5* in the *Pcdh*α and *Pcdh*γ gene clusters, respectively) and the first constant region exon contain relatively few repeats in both mouse and human. Instead, this region has a relatively high sequence conservation between mouse and human in both the *Pcdh*α and *Pcdh*γ gene clusters (Fig. 5A,B). The most conserved segments are shown in Figure 7, where long stretches of exact sequence identity are observed between mouse and human.

The bulk of the mammalian genome has a GC content of 40% and is poor in CpG dinucleotides, with only 25% of the expected CpG dinucleotide frequency based on the GC content. However, there are regions of genomic DNA that contain the CpG dinucleotides at about its expected frequency, which are known as CpG islands (Antequera and Bird 1993). Both mouse and human genomic DNA of protocadherin gene clusters have ~41% GC content. However, the distribution of CpG dinucleotides is highly specific, as the ratio of observed to expected CpG dinucleotide frequency peaks at the location of each variable region exon (Fig. 4). It is usually assumed that each island identifies a gene, because the number of CpG islands that are not associated with genes is likely to be small (Antequera and Bird 1993).

In summary, we annotated the mouse protocadherin genomic DNA sequence and found that the overall genomic organization of the three protocadherin gene clusters is highly conserved between mouse and human. Moreover, we identified the orthologous mouse and human gene pairs in the *Pcdh*α and *Pcdh*γ gene clusters, and found that the number and order of *Pcdh*α and *Pcdh*γ genes are essentially conserved between mouse and human. We also found, however, that the mouse and human *Pcdh*β genes display both orthologous and paralogous relationships, and the mouse *Pcdh*β locus is larger and has six more genes than the human locus. Finally, we showed that the upstream sequences of each variable coding region are more conserved between orthologous than between paralogous genes. Within these upstream sequences, there is a conserved motif shared by almost all members of the three closely linked gene clusters. In addition, the distribution of CpG islands correlates with the locations of variable region exons. Taken together, these results strongly suggest that each protocadherin variable region exon has a distinct promoter.

METHODS

Mouse BAC Isolation and Sequencing

Nineteen PCR primer pairs were designed to screen a mouse BAC library (RPCI-23). The primer sequences are: ATCCCAA AATGGTGATGAAACTG and CGCTGGCAGAGGCCAAGAT CA (length of product: 89 bp); CTCTGTGCACCTGGAGGAG GC and CTGGTGTTGCACTGGATACTGTT (89 bp); GAAGTG GCCAGGAATCCCAGC and CTCAGGGATGGAGTAGTGGA TC (95 bp); CCACTGAAGGCCGACTGGGAAC and CTCTGG GACGGAGTAATGAAGC (101 bp); CTTCGGATGCAGACATC GGAAC and TCTTTAACACTAGTTGGAGTGG (120 bp); CGT CAGATGCAGATGTCGGTTC and AGCCCAAGAGGTTTCAC CTGC (110 bp); ATCCGATGCAGATATCGGAGTC and CTT TAACACAAGGGATAACGAAG (120 bp); ATCTGATTTGG ATATAGGAGCC and GAGCAACAAACGATGCTCTTGG (165 bp); CGGACATAGGAGAGAACGCTG and CCTTCTTTAATA TAAGTGACGGTC (120 bp); CTAGAAGGCGCCTCTGATGC AG and AGTTTTCGAAGAACAAGCACTGG (140 bp); AAGA GACGGTTCCGGAAGACAG and AACGAGTACTGACAGCT TCTGC (110 bp); CAGAGTGGATCGAGTGCCCTTG and GGTCACCATCTACTGTGGCTAC (140 bp); CTGGCTGTCAT TCCAACTTCTC and GTAGCCACAGTAGATGGTGACC (140 bp); CCAAGTCTCCTACACCATGCTC and GTGATGTGGGC ATTGGAGCCTG (100 bp); CTGCATGGATGTGCAATCTGAG and CTCTCTGTTTCTTCCTCTATGG (200 bp); GCAGGCTAT TAACTGACAGGTC and GAGAAAGATCAACAGAACTTGCC (120 bp); GTCCCAGAACTACCAATATGAG and AGGGTCA TGGAGCTGAAGACTG (100 bp); AAATGTGCTGTGGTTG TAGAGG and ACAGCAACAACTGTCTCTTGTG (110 bp); GAAGGTATTTGAGCGTGATCTAG and CTTCTTCTAGTCAG TTTCAATCCAC (120 bp). A total of 21 mouse BAC clones were isolated, and their sizes were estimated by pulse-field gel electrophoresis. The clones were digested with *Bst*Z17I. The restriction map was assembled from the resulting fragments. Seven minimally overlapping mouse clones were selected for shotgun sequencing. The chromosomal locations of the selected clones were mapped by FISH. Draft sequences of these BAC clones were produced by the DOE Joint Genome Institute. The sequences of all the mouse BACs and four human gap-closing clones were finished by the sequencing group at the Stanford Human Genome Center. All of the other human clones were finished by the DOE Joint Genome Institute. The finished sequences for the mouse and human clones contain no gaps and are estimated to contain less than one error per 200,000 bp. The GenBank accession nos. for the mouse clones are AC020967, AC020968, AC020969, AC020971, AC020972, AC020973, and AC020974. The GenBank accession nos. for the human clones are AC005366, AC004776, AC005609, AC005618, AC005752, AC005754, AC008468, AC010223, AC025436, and AC074130. In addition, the sequences and quality scores for each base position can be found at http://www-shgc.stanford.edu/Seq/Status/doe.html.

Phylogenetic Analysis

The variable region coding sequences were translated, and the resulting polypeptides were aligned using the `Pileup` program of the GCG sequence analysis package (Genetics Computer Group 1999) with default parameters. A phylogenetic tree was reconstructed by using `PAUP` (Phylogenetic Analysis Using Parsimony), version 4.0.0 (Swofford et al. 1996), with distance as an optimality criterion. Gaps in the alignment were treated as missing. The robustness of the tree partitions

was evaluated by using the bootstrap analysis with a neighbor-joining search.

Sequence Analysis

Annotation

The mouse protocadherin coding regions were annotated by using the BLAST program (Altschul et al. 1997) and the GCG (Genetics Computer Group 1999) sequence analysis package. The potential coding sequences were aligned by using the multiple sequence alignment program Pileup. The 5′ splice sites were identified manually by inspecting the alignment of mouse sequences to the corresponding human and mouse cDNA sequences. All of the variable region 5′ splice sites conform to the splice site consensus sequences and were verified by sequencing the cDNA fragments spanning splice junctions between variable and constant region exons. The putative translation start codon was determined by inspecting the translated signal peptide sequences.

Comparison

Human and mouse *Pcdh*α, *Pcdh*β, and *Pcdh*γ genomic sequences were assembled from the finished sequences of the respective BAC clones using the Seqed program of the GCG package. The CpG island distribution was plotted by using the CpGplot program (Larsen et al. 1992). The repeats were masked with the use of the RepeatMasker program (A.F.A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html). The masked mouse genomic sequences of *Pcdh*α and *Pcdh*γ gene clusters were compared with the corresponding human genomic sequences by using PipMaker (Schwartz et al. 2000) on the Web server http://bio.cse.psu.edu/pipmaker/. We used the chaining option of PipMaker for the *Pcdh*α and *Pcdh*γ clusters because their gene orders are conserved.

We assigned human and mouse orthologous protocadherin gene pairs based on the phylogenetic trees (Fig. 3). In cases of paralogous relationships in the phylogenetic tree, we assigned orthologs based on highest sequence identity. To systematically compare the upstream sequences of orthologous and paralogous genes, the upstream sequences were extracted according to our annotation from RepeatMasked genomic sequences (without masking low-complexity DNA). For each orthologous and paralogous gene pair, the maximal sequence identity among all 100 bp segments within a 150 bp sliding window was calculated (any masked sequences were counted as mismatches in the calculation). In the case of human *Pcdhα2* and *Pcdhα-C1*, the sliding window size was 250 bp. The maximal 100 bp-segment identity within each window was plotted against its end position relative to the translation start codon.

We used a version of the Gibbs sampler program to identify the conserved sequence motifs upstream of all variable region exons within each protocadherin gene cluster. The program also calculates the probability (ranging from 0 to 1) of finding the motif within -290 to -150 nucleotides upstream of each variable region start codon.

ACKNOWLEDGMENTS

We thank E. Branscomb and T. Hawkins for supporting the DNA sequence determination of the mouse protocadherin gene clusters at the Joint Genome Institute. We also thank the Sequencing Group at the Stanford Human Genome Center for finishing the clones. We are grateful to W. Miller for advice on PIP analysis, S. Ribich, B. Tasic, P. Cramer, and C. Nabholz for discussion and critical comments on the manuscript. This work was supported by grants from the NIH to T.M. (GM42231), from the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation to Q.W. (DRG-1559), from the NIH to M.Q.Z. (HG01696), from the DOE to J.-F.C (DE-AC03–76SF00098) and to R.M.M. (DE-FC03–99ER62873).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Ansari-Lari, M.A., Oeltjen, J.C., Schwartz, S., Zhang, Z., Muzny, D.M., Lu, J., Gorrell, J.H., Chinault, A.C., Belmont, J.W., Miller, W., et al. 1998. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* **8:** 29–40.

Antequera, F. and Bird, A. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci.* **90:** 11995–11999.

Bruses, J.L. 2000. Cadherin-mediated adhesion at the interneuronal synapse. *Curr. Opin. Cell Biol.* **12:** 593–597.

Camacho, J.A., Obie, C., Biery, B., Goodman, B.K., Hu, C.A., Almashanu, S., Steel, G., Casey, R., Lambert, M., Mitchell, G.A., et al. 1999. Hyperornithinaemia-hyperammonaemia-homocitrullinuria syndrome is caused by mutations in a gene encoding a mitochondrial ornithine transporter. *Nat. Genet.* **22:** 151–158.

Chiang, C.M. and Roeder, R.G. 1995. Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science* **267:** 531–536.

Chun, J. 1999. Developmental neurobiology: A genetic Cheshire cat? *Curr. Biol.* **9:** R651–R654.

Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. 1994. Purification of CpG islands using a methylated DNA binding column. *Nat. Genet.* **6:** 236–244.

Dreyer, W.J. and Roman-Dreyer, J. 1999. Cell-surface area codes: Mobile-element related gene switches generate precise and heritable cell-surface displays of address molecules that are used for constructing embryos. *Genetica* **107:** 249–259.

Endrizzi, M.G., Hadinoto, V., Growney, J.D., Miller, W., and Dietrich, W.F. 2000. Genomic sequence analysis of the mouse naip gene array. *Genome Res.* **10:** 1095–1102.

Genetics Computer Group. 1999. Program Manual for the Wisconsin Package Version 10.0. Genetics Computer Group (GCG), Madison, WI.

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

Hagler, D.J., Jr. and Goda, Y. 1998. Synaptic adhesion: The building blocks of memory? *Neuron* **20:** 1059–1062.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Hirano, S., Ono, T., Yan, Q., Wang, X., Sonta, S., and Suzuki, S.T. 1999a. Protocadherin 2C: A new member of the protocadherin 2 subfamily expressed in a redundant manner with OL-protocadherin in the developing brain. *Biochem. Biophys. Res. Commun.* **260:** 641–645.

Hirano, S., Yan, Q., and Suzuki, S.T. 1999b. Expression of a novel protocadherin, OL-protocadherin, in a subset of functional systems of the developing mouse brain. *J. Neurosci.* **19:** 995–1005.

Jang, W., Hua, A., Spilson, S.V., Miller, W., Roe, B.A., and Meisler, M.H. 1999. Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res.*

9: 53–61.
Kim, S.H., Jen, W.C., De Robertis, E.M., and Kintner, C. 2000. The protocadherin PAPC establishes segmental boundaries during somitogenesis in *Xenopus* embryos. *Curr. Biol.* **10:** 821–830.
Kohmura, N., Senzaki, K., Hamada, S., Kai, N., Yasuda, R., Watanabe, M., Ishii, H., Yasuda, M., Mishina, M., and Yagi, T. 1998. Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron* **20:** 1137–1151.
Koop, B.F. and Hood, L. 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* **7:** 48–53.
Lamerdin, J.E., Stilwagen, S.A., Ramirez, M.H., Stubbs, L., and Carrano, A.V. 1996. Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* **34:** 399–409.
Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. 1992. CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107.
Lynch, E.D., Lee, M.K., Morrow, J.E., Welcsh, P.L., Leon, P.E., and King, M.C. 1997. Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the *Drosophila* gene diaphanous. *Science* **278:** 1315–1318.
Mallon, A.M., Platzer, M., Bate, R., Gloeckner, G., Botcherby, M.R., Nordsiek, G., Strivens, M.A., Kioschis, P., Dangel, A., Cunningham, D., et al. 2000. Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res.* **10:** 758–775.
Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. 2000. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530–1534.
Nollet, F., Kools, P., and van Roy, F. 2000. Phylogenetic analysis of the cadherin superfamily allows identification of six major subfamilies besides several solitary members. *J. Mol. Biol.* **299:** 551–572.
O'Hanlon, T.P., Raben, N., and Miller, F.W. 1995. A novel gene oriented in a head-to-head configuration with the human histidyl-tRNA synthetase (HRS) gene encodes an mRNA that predicts a polypeptide homologous to HRS. *Biochem. Biophys. Res. Commun.* **210:** 556–566.
Redies, C. 2000. Cadherins in the central nervous system. *Prog. Neurobiol.* **61:** 611–648.
Rowen, L., Koop, B.F., and Hood, L. 1996. The complete 685-kilobase DNA sequence of the human β T-cell receptor locus. *Science* **272:** 1755–1762.
Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker–A web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.
Serafini, T. 1999. Finding a partner in a crowd: Neuronal diversity and synaptogenesis. *Cell* **98:** 133–136.
Shapiro, L. and Colman, D.R. 1999. The diversity of cadherins and implications for a synaptic adhesive code in the CNS. *Neuron* **23:** 427–430.
Steinberg, M.S. and McNutt, P.M. 1999. Cadherins and their connections: Adhesion junctions have broader functions. *Curr. Opin. Cell Biol.* **11:** 554–560.
Sugino, H., Hamada, S., Yasuda, R., Tuji, A., Matsuda, Y., Fujita, M., and Yagi, T. 2000. Genomic organization of the family of CNR cadherin genes in mice and humans. *Genomics* **63:** 75–87.
Suzuki, S.T. 1996. Protocadherins and diversity of the cadherin superfamily. *J. Cell Sci.* **109:** 2609–2611.
Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. 1996. Phylogenetic inference. In *Molecular systematics*, (eds. D.M. Hillis et al.), pp. 407–514. Sinauer Associates, Sunderland, Massachusetts.
Wu, Q. and Maniatis, T. 1999. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97:** 779–790.
———. 2000. Large exons encoding multiple ectodomains are a characteristic feature of protocadherin genes. *Proc. Natl. Acad. Sci.* **97:** 3124–3129.
Yagi, T. and Takeichi, M. 2000. Cadherin superfamily genes: Functions, genomic organization, and neurologic diversity. *Genes & Dev.* **14:** 1169–1180.
Yamagata, K., Andreasson, K.I., Sugiura, H., Maru, E., Dominique, M., Irie, Y., Miki, N., Hayashi, Y., Yoshioka, M., Kaneko, K., et al. 1999. Arcadlin is a neural activity-regulated cadherin involved in long term potentiation. *J. Biol. Chem.* **274:** 19473–19479.
Yoshida, K. and Sugano, S. 1999. Identification of a novel protocadherin gene (PCDH11) on the human XY homology region in Xq21.3. *Genomics* **62:** 540–543.

Received October 16, 2000; accepted in revised form January 9, 2001.