DOI: 10.1534/genetics.104.037606

Comparative Genomics and Diversifying Selection of the Clustered Vertebrate Protocadherin Genes

Qiang Wu[1]

Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112

Manuscript received October 14, 2004
Accepted for publication January 14, 2005

ABSTRACT

To explain the mechanism for specifying diverse neuronal connections in the brain, Sperry proposed that individual cells carry chemoaffinity tags on their surfaces. The enormous complexity of these connections requires a tremendous diversity of cell-surface proteins. A large number of neural transmembrane protocadherin (Pcdh) proteins is encoded by three closely linked human and mouse gene clusters (α, β, and γ). To gain insight into *Pcdh* evolution, I performed comprehensive comparative cDNA and genomic DNA analyses for the three clusters in the chimpanzee, rat, and zebrafish genomes. I found that there are species-specific duplications in vertebrate *Pcdh* genes and that additional diversity is generated through alternative splicing within the zebrafish "variable" and "constant" regions. Moreover, different codons (sites) in the mammalian Pcdh ectodomains (ECs) are under diversifying selection, with some under diversity-enhancing positive Darwinian selection and others, including calcium-binding sites, under strong purifying selection. Interestingly, almost all positively selected codon positions are located on the surface of ECs 2 and 3. These diversified residues likely play an important role in combinatorial interactions of Pcdh proteins, which could provide the staggering diversity required for neuronal connections in the brain. These results also suggest that adaptive selection is an additional evolutionary factor for increasing *Pcdh* diversity.

An important mechanism to generate molecular diversity is through alternative splicing. A special form of alternative splicing uses multiple distinct first exons. Mammalian genomes contain a large number of alternatively spliced genes that have multiple "variable" first exons (Zhang *et al.* 2004). The clustered *Pcdh* genes exemplify this type of alternative splicing that utilizes multiple variable first exons. About 60 similar human and mouse *Pcdh* genes are organized into three sequentially linked clusters, designated α, β, and γ (see Figure 1, A and C) (Wu *et al.* 2001). The α and γ clusters have a variable and "constant" genomic organization, similar to that of immunoglobulin (*Ig*) and T-cell receptor (*Tcr*) gene clusters (Wu and Maniatis 1999). Specifically, the variable region of the α cluster contains 15 and 14 highly similar exons in humans and mice, respectively. These variable exons are unusually large (~2.5 kb each) and are organized in a tandem array, which is followed by the constant region of three small exons, in both humans and mice (Figure 1, A and C). Similarly, the variable region of both the human and mouse γ clusters contains a tandem array of 22 large similar exons; while the γ constant region contains three small downstream exons in both species (Figure 1, A and C). In contrast to the α and γ clusters, the human and mouse β clusters contain 16 and 22 variable exons, respectively, but do not contain a constant region. Thus, each member of the human and mouse β clusters is a single-exon gene (Figure 1, A and C).

Each *Pcdh* variable exon is preceded by a distinct promoter (Tasic *et al.* 2002), and these promoters share a highly conserved core motif (Wu *et al.* 2001; Noonan *et al.* 2003, 2004). Specific promoter activation transcribes a high-molecular-weight precursor RNA that extends through all of the downstream variable and constant exons. However, only the 5′-most variable exon is *cis*-spliced to the first constant exon to generate functional mRNAs (Tasic *et al.* 2002; Wang *et al.* 2002a). Pcdh α and γ proteins are generally located at synaptic junctions in the central nervous system (CNS) (Kohmura *et al.* 1998; Wang *et al.* 2002b; Phillips *et al.* 2003), where they may form combinatorial hetero-*cis*-interactions (Murata *et al.* 2004) and specific homophilic *trans*-interactions (Obata *et al.* 1995). Because of their synaptic localization, unusual genomic organization, and characteristic cadherin domains, the Pcdh proteins have been proposed to provide molecular tags for the chemoaffinity hypothesis (Sperry 1963; Shapiro and Colman 1999). This influential hypothesis, posited by Roger Wolcott Sperry more than a half century ago to explain the staggering complexity of neuronal connections in the brain, suggested that the establishment

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY540132–AY540190, AY573971–AY574030, AY576933–AY576986, AY583021–AY583058, and AY583468–AY583498.

[1]*Address for correspondence:* Department of Human Genetics, University of Utah, 15N 2030E, Salt Lake City, UT 84112.
E-mail: qwu@genetics.utah.edu

and maintenance of diverse synaptic junctions were achieved by lock-and-key interactions between molecules specifically expressed by different types of neurons.

An additional notable mechanism for generating molecular diversity is through gene duplication and positive selection. Duplication increases gene numbers while positive selection on duplicated genes can rapidly diversify paralogous protein sequences, resulting in a higher rate of nonsynonymous substitutions relative to synonymous substitutions. Comparative genomics in humans, mice, and rats have revealed that some of the most rapidly evolving genes are those involved in reproduction, adaptive immune response, and chemosensation. For example, positive selection events are known to enhance the diversity of the major histocompatibility complex (*Mhc*) (Hughes and Nei 1988), *Ig* (Tanaka and Nei 1989; Sitnikova and Nei 1998), and *Tcr* (Su and Nei 2001) clusters. Recently, olfactory receptor genes have also been found to be subject to adaptive selection (Emes *et al.* 2004). These examples of positive selection imply that the diversity-enhancing codon substitutions within certain classes of proteins benefit the survival or reproduction of organisms.

To gain insight into *Pcdh* evolution, I performed a comprehensive comparative analysis on the three closely linked neural *Pcdh* clusters in primates, rodents, and fish. I found that the number of *Pcdh* genes is different in each species and that there are additional alternative splice sites within the zebrafish variable and constant regions, generating even more diversity. Moreover, analyzing the pattern of nucleotide substitutions identified codon sites that are likely to have been subject to positive Darwinian selection at the molecular level. Finally, different subfamilies of the *Pcdh* genes in specific mammalian species have distinct sets of sites under positive selection. Interestingly, almost all these nonconserved sites are located in the ECs 2 and 3. These diversified residues may participate in the combinatorial *cis*-interactions between Pcdh molecules expressed on the same neuronal plasma membrane. Thus, both species-specific diversifying selection and the birth-and-death evolution of *Pcdh* genes may have contributed to the molecular coding of the staggering diversity of neuronal connectivity in the mammalian brain.

MATERIALS AND METHODS

Genomic sequence annotation and cDNA cloning and collection: The chimpanzee, rat, and zebrafish *Pcdh* sequences and trace files were identified (AC144823, AC144828, AC144826, AC146480, AL929558, BX119910, BX005294, and BX957322) by iterative BLAST searches of public repositories (GenBank and TraceDB) (Gibbs *et al.* 2004; Noonan *et al.* 2004). The sequences were downloaded from GenBank and the University of California, Santa Cruz (UCSC) genome browser (genome.ucsc.edu) (Kent *et al.* 2002) and by using the FTP from the TraceDB (www.ncbi.nlm.nih.gov/Traces). The sequences were analyzed and annotated as previously described (Wu *et al.* 2001). I manually checked every nucleotide of chimpanzee and rat *Pcdh* exons by using the Sequencher program. Some rat cDNA clones were also sequenced to fill gaps and to confirm splice sites. Gaps in the chimpanzee variable exons were filled by cloning and sequencing PCR products from the chimpanzee genomic DNA (Coriell) with specific primers (supplementary Table S1 at http://www.genetics.org/supplemental/).

To clone members of zebrafish *Pcdh* clusters, I designed specific forward and reverse primers (supplementary Table S1). The adult zebrafish brain tissues were dissected under a dissection microscope and total RNA was isolated by using Trizol (Invitrogen, San Diego) according to the manufacturer's instructions. Extensive RT-PCR and rapid amplification of cDNA ends (RACE) experiments were performed by using the SMART RACE cDNA amplification system (BD Biosciences). The PCR products were cloned and sequenced from both strands with internal primers (supplementary Table S1). The cDNA sequences were compared with the genomic sequences (Noonan *et al.* 2004) to identify the splice sites.

Phylogenetic analysis: The cloned or predicted full-length chimpanzee, rat, and zebrafish variable exon coding sequences were translated, and the resulting polypeptides were aligned by using the PILEUP program of the GCG package. A phylogenetic tree was reconstructed by using the neighbor-joining algorithm in the CLUSTAL W package (Thompson *et al.* 1994). Gaps in the alignment were treated as missing. The robustness of the tree partitions was evaluated by using bootstrap analysis.

Site-specific K_A/K_S analysis: A set of 325 full-length human, chimpanzee, mouse, rat, and zebrafish variable coding sequences was translated. The encoded polypeptide sequences are of about the same length and were aligned by the PILEUP program with very few gaps, especially in ECs 2 and 3. The nucleotide alignment was built by using RevTrans (Wernersson and Pedersen 2003) (www.cbs.dtu.dk/services/RevTrans) according to the protein alignment. The coding regions for ECs 1–3 were extracted and separated into calcium-binding codons according to the structure of the classic C-cadherin (Boggon *et al.* 2002) and the rest as noncalcium binding codons or regions. To compare the positions of codons in different species, the gaps in the alignments were removed.

The standard measure of adaptive molecular evolution at the protein-coding region is to compare the number of nonsynonymous substitutions per nonsynonymous site (K_A) with the number of synonymous substitutions per synonymous site (K_S). K_S reflects the silent mutation rate while K_A reflects the rate of amino acid changes. The substitution rate ratio $\omega = K_A/K_S$ measures the molecular selective pressure. If $\omega = 1$, the amino acid changes are neutral and will be fixed at the same rate as the silent mutations. If $\omega < 1$, the amino acid changes are deleterious and purifying selection will reduce the fixation rate. If $\omega > 1$, the amino acid changes are evolutionarily advantageous and positive selection will increase the fixation rate. As the K_A/K_S value for the calcium-binding codons is 0, I performed only site-specific K_A/K_S analyses on the non-calcium-binding regions. I used the maximum-likelihood codeml program of the PAML package (v3.14beta7) (Yang 1997) (abacus.gene.ucl.ac.uk/software/paml.html) to predict codon sites under Darwinian selection for 22 paralogous *Pcdh* groups (16 mammalian groups, human, chimpanzee, mouse, and rat α, β, γ*a*, and γ*b*, excluding the highly divergent c-type *Pcdh* genes; and 6 zebrafish groups, α1–3 and γ1–3).

The codeml program uses the Markov model of codon substitutions (Yang 1994; Yang and Bielawski 2000; Yang *et al.* 2000). Simple codeml models of null hypothesis with $0 < \omega < 1$ can be compared with more complex models of generalized alternative hypothesis that allow $\omega > 1$. Log-likelihood

values (ℓ) are calculated for each model to enable a likelihood-ratio test (LRT) to be used as a statistical test for significance to accept the complex hypothesis and to reject the null hypothesis. When two models are nested, twice the log-likelihood difference ($2\Delta\ell$) is compared to the chi-square (χ^2) asymptotic distribution (Goldman and Whelan 2000) with the degrees of freedom equal to the difference in the number of parameters between the two models. If the LRT statistic ($2\Delta\ell$) is greater than critical values of the χ^2-distribution and the complex model indicates an estimated $\omega > 1$, Bayesian probabilities are used to infer which codons are most likely to have been subject to positive selection.

For each of the 22 *Pcdh* groups, I first ran model M0 of the codeml program with a nucleotide neighbor-joining tree to obtain a K_S-derived tree (abacus.gene.ucl.ac.uk/software/paml.html; PAML manual). I then used the K_S tree to run three nested pairs of codeml random-sites models: M0 (one ratio) *vs.* M3 (discrete), M1 (neutral) *vs.* M2 (selection), and M7 (β) *vs.* M8 (β + ω). M0 assumes one ω for all sites (Yang 1994) while M3 assumes an unconstrained discrete distribution of ω among sites (Yang *et al.* 2000). M1 assumes a neutral site class with $\omega = 1$ and a conserved site class with $\omega = 0$ while M2 adds an additional site class with ω permitted to be >1 (Nielsen and Yang 1998). M7 assumes a β distribution of ω between 0 and 1 while M8 adds one extra site class with a free ω ratio estimated from the data (Yang *et al.* 2000). Because the iterative estimations of ω values by both M2 and M8 are susceptible to local optima, I ran M2 and M8 with three different initial ω values (0.03, 0.8, and 3.14) and presented only those results with the highest likelihood.

Protein structure analysis: The ECs encoded by members of the clustered *Pcdh* genes were modeled by using SWISS-MODEL (Guex and Peitsch 1997) (swissmodel.expasy.org) with the C-cadherin structure as a template (Boggon *et al.* 2002). Swiss-PDBviewer was used for structural manipulations (Guex and Peitsch 1997). The Pcdh ECs have a Greek-key seven-stranded β sandwich folding topology similar to that of classic cadherins (data not shown). Therefore, I mapped the ω+ sites to the ECs 1–3 tertiary structure of C-cadherin (PDB accession code 1L3W). The ω+ sites were defined as diversified codon positions predicted to be under positive selection with a posterior probability >0.90 by one codeml model (M2, M3, or M8), and >0.50 by at least one other model (Emes *et al.* 2004). Human, chimpanzee, mouse, and rat Pcdh ECs 1–3 sequences were aligned with those of the C-cadherin ECs 1–3 by using hidden Markov models (HMM) (www.cse.ucsc.edu/research/compbio/sam.html) (Krogh *et al.* 1994). The ω+ codon positions were highlighted in the C-cadherin crystal structure by using the PyMOL program (www.pymol.org).

RESULTS

The chimpanzee *Pcdh* clusters: Chimpanzees are the closest evolutionary relative of humans and share nearly 99% DNA sequence identity (Olson and Varki 2003). The two species are thought to have diverged as recently as 4.6 million years ago (Chen and Li 2001). However, chimpanzees and humans have major phenotypic differences in many behavioral, cognitive, and anatomical aspects, such as bipedalism, speech, and brain size. In particular, the human cerebral cortex is dramatically bigger and more complex. The *Pcdh* genes may play important roles in human evolution. For example, *Pcdh X/Y*, which was duplicated and translocated from the X chromosome after the last common ancestor of humans and chimpanzees, has been proposed to be involved in cerebral asymmetry, handedness, lateralization, and the development of language (Crow 2002).

To investigate the role of *Pcdh* genes in human brain evolution, I compared the human and chimpanzee α, β, and γ clusters. Both the human and chimpanzee clusters span a region of ~750 kb of genomic DNA. As expected, almost all members of the human and chimpanzee clusters are conserved (Figure 1, A and B). Surprisingly, I found that three *Pcdh* genes are different. The human β*17* and β*18* genes have two- and one-nucleotide insertions, respectively, while the chimpanzee γ*b3* gene has a one-nucleotide insertion, all of which cause frameshifts (Figure 1, A and B). Interestingly, the γ*b3* gene has further degenerated into a relic in both mice and rats (Figure 1, C and D) (Wu *et al.* 2001). Therefore, it seems that the γ*b3* gene functions only in humans. Relics are sequence fragments with only limited similarity to the corresponding functional genes; while pseudogenes have more extensive sequence similarity but are rendered nonfunctional by mutations.

Loss-of-function mutations are a mechanism for rapid phenotypic evolution between closely related species such as humans and chimpanzees (Olson and Varki 2003). To date, very few genes have been found with mutations in humans but not in chimpanzees. These include the *Tcr* γ*V10* gene, the CMP-Neu5Ac hydroxylase gene, the olfactory receptor *OR912-93* gene, and a type I hair-keratin gene (Olson and Varki 2003). These mutations have contributed to various aspects of human evolution. Given the predominant expression of *Pcdh* genes in the brain, the difference in *Pcdh* gene numbers that resulted from the birth-and-death evolution may have contributed to rapid human brain evolution.

The rat *Pcdh* clusters: The genomic organization of the rat α cluster has been reported (Yanase *et al.* 2004); however, the complete rat *Pcdh* locus has not been fully annotated. I analyzed the three rat *Pcdh* clusters and found that the overall organization is highly conserved among rats, mice, chimpanzees, and humans. Both the rat α and γ clusters are organized into variable and constant regions while the rat β cluster lacks a constant region (Figure 1D). This demonstrates that the organization of the three clusters is highly conserved in mammals.

The genomic organization of the rat *Pcdh* clusters is almost the same as that of the mouse (Wu *et al.* 2001). Members of the three clusters are orthologous between rat and mice. Similar to the mouse β cluster, rats have six more β genes than humans and four more than chimpanzees. The two non-cadherin genes located between the β and γ clusters are also conserved. However, there are some differences between the mouse and rat clusters. First, the mouse α cluster has one variable exon less than the rat because one mouse variable exon has been interrupted by a transposon after the divergence of the two species (Wu *et al.* 2001). Second, the rat

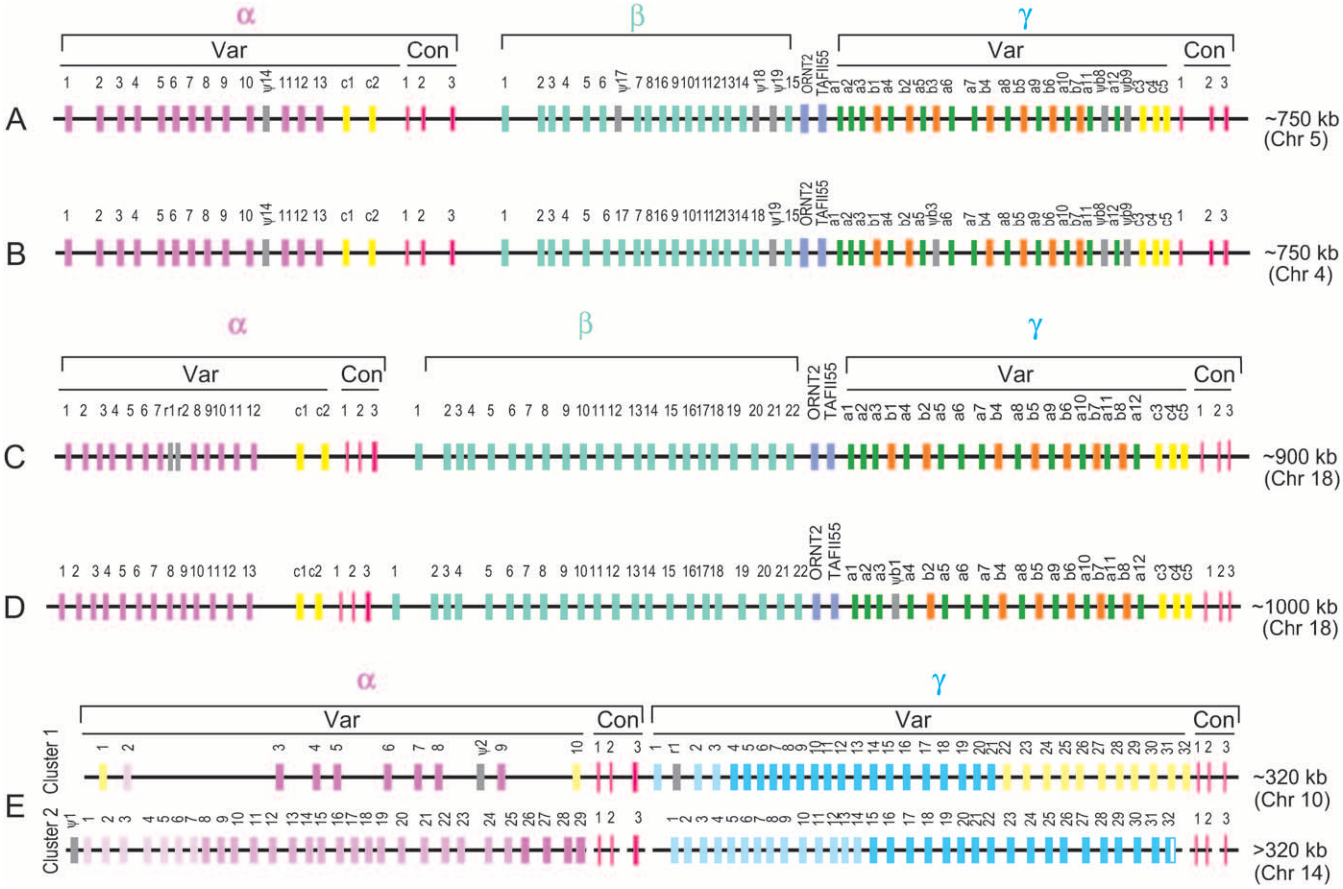

Figure 1.—Comparison of the (A) human, (B) chimpanzee, (C) mouse, (D) rat, and (E) zebrafish *Pcdh* clusters. Each cluster contains multiple, highly similar, tandem variable (Var) exons indicated by vertical color bars: mauve, α variable exons; turquoise, mammalian β genes; green, γ*a* variable exons; orange, γ*b* variable exons; yellow, c-type variable exons; blue, zebrafish γ variable exons; and gray, relic (r) or pseudogene (ψ). Constant (Con) exons are indicated by small red vertical bars following the variable-exon tandem arrays. The two noncadherin genes (*ORNT2* and *TAFII55*) are conserved in mammals but not in fish. The approximate length in each genome is shown on the right. Gaps in sequences of zebrafish *Pcdh* cluster 2 are represented by dashes. The orientation of the zebrafish α cluster 2 and γ cluster 2 is not confirmed, and they may contain additional variable exons in the gaps. Var, variable; Con, constant.

ortholog of the mouse γ*b1* gene has been mutated to a pseudogene (Figure 1, C and D). These observations demonstrated that birth-and-death evolution occurs in the rodent *Pcdh* gene clusters.

The zebrafish *Pcdh* clusters: Recent genomic sequence analyses identified three zebrafish *Pcdh* clusters (Noonan *et al.* 2004; Tada *et al.* 2004). Here I annotated the complete zebrafish *Pcdh* repertoire (Figure 1E). By sequencing large numbers of cDNAs and comparing them with genomic DNA sequenced by the Stanford Human Genome Center (Noonan *et al.* 2004) and the Sanger Institute, I identified all splice sites in the zebrafish *Pcdh* variable and constant regions. In total, I found that zebrafish have 102 *Pcdh* variable exons organized into four clusters, considerably more than that in mammals (Figure 1E). In contrast to mammals, zebrafish have two α and two γ clusters, but lack the β cluster. Each of the four clusters contains both variable and constant regions (Figure 1E). Similar to the mammalian *Pcdh* genes, the zebrafish *Pcdh* genes are also expressed in the CNS. For example, I detected the expression of the *Pcdh 1*γ and *2*γ clusters from zebrafish brain total RNA preparations by using RT-PCR (supplementary Figure S1 at http://www.genetics.org/supplemental/).

The zebrafish constant sequences are similar between the two α clusters and also between the two γ clusters (supplementary Figure S2 at http://www.genetics.org/supplemental/). Specifically, the two α constant polypeptide sequences share 84% similarity and 80% identity (supplementary Figure S2A), while the two γ constant polypeptides share 82% similarity and 79% identity (supplementary Figure S2B). Thus, the α and γ clusters were duplicated in the zebrafish genome. The *Pcdh* clusters appear to be quite divergent between teleosts and mammals. For example, the lengths of the constant region exon 2 are identical in mammals but different in teleosts; and the sequence conservation is too low for these exons to be identified by sequence comparisons only.

Additional diversity generated by alternative splicing within the *Pcdh* variable and constant regions: In addi-

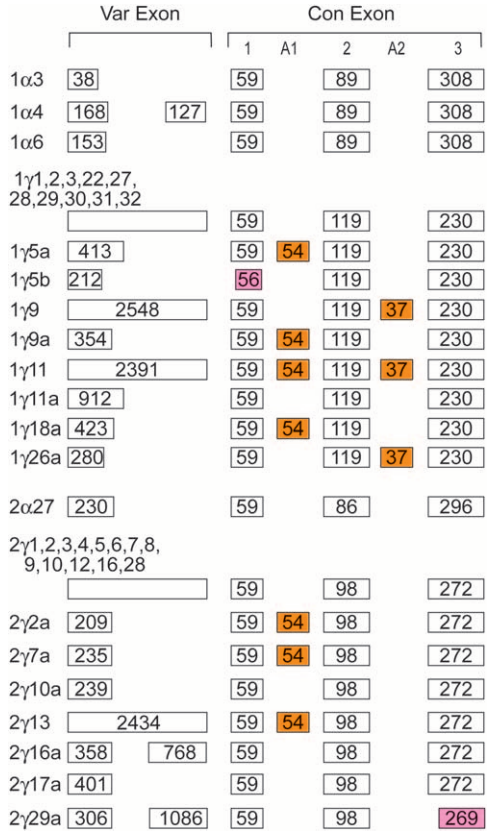

FIGURE 2.—Diverse mRNA repertoire generated by alternative splicing of zebrafish *Pcdh* clusters. The corresponding cDNAs were cloned from zebrafish total RNA preparations and sequenced from both strands. Boxes with nucleotide length indicate exons. Orange boxes indicate alternative cassette constant exons. Pink boxes indicate constant exons that use an alternative 3′ splice site. Var, variable; Con, constant.

tion to the full-length forms, there are internal alternative 5′ splice sites within the mammalian α and γ variable exons. These internal 5′ splice sites are spliced to the 3′ splice site of the first constant exon to generate shorter mRNAs. The encoded polypeptides have a signal peptide but lack a transmembrane segment. Therefore, the encoded proteins may be secreted *Pcdh* isoforms. In addition, there is an alternatively spliced intron within the constant exon 3 of the mammalian α cluster, which can be retained or excluded to generate two sets of α mRNAs (SUGINO *et al.* 2000; WU *et al.* 2001). In contrast, no alternative splicing event has been found in the constant region of the two zebrafish α clusters.

I identified extensive alternative splicing in the constant region of the two zebrafish γ clusters. For example, the zebrafish *1*γ cluster has two novel cassette exons (A1 and A2), located between the three constant exons, which have not been observed in the constant region of the mammalian γ cluster. I cloned cDNAs that contain all four combinations of these two cassette exons: exclusion of both and inclusion of A1, A2, or both (Figure 2). Therefore, the zebrafish *1*γ cluster can potentially encode four sets of proteins, each consisting of 32 full-length Pcdhs. I also cloned a cassette exon (A1) between zebrafish *2*γ cluster constant exons 1 and 2 (Figure 2). The length of this cassette exon is the same as that of the corresponding exon in the *1*γ cluster. In addition, they display 48% nucleotide identity, while the encoded polypeptides have a 41% sequence similarity. These observations suggest that the first cassette exon is conserved between the two zebrafish γ clusters, and its existence precedes the duplication of the γ clusters. Given the existence of the A2 cassette exon in the *1*γ cluster, I reasoned that a similar cassette exon may also exist in the *2*γ cluster (Figure 2).

Similar to the mammalian clusters, I observed additional alternative splice sites within variable exons in all four zebrafish clusters (Figure 2). Most splice sites in the zebrafish clusters conform to the canonical GT-AG consensus. Interestingly, there are several introns with noncanonical splice sites: *1α1* and *1γ11* have a GC-AG intron; *1γ5* and *2γ29* have an AT-AA intron; and *2γ16* has an AT-AC intron. On the basis of the sequence context of their splice sites, splicing of all these introns seems to use the major U2-dependent splicing pathway (WU and KRAINER 1999). The alternatively spliced mRNAs may encode short-form Pcdh proteins that lack a transmembrane segment and may be secreted. Moreover, there are additional alternative 3′ splice sites within constant exons 1 and 3 that generate even more diversity (Figure 2). Interestingly, in both cases the alternative 3′ splice site is only three nucleotides downstream from the normal 3′ splice site (Figure 2). Therefore, the encoded constant polypeptides lack one conserved amino acid residue (glutamine or glutamic acid).

Phylogenetic relationships of the chimpanzee, rat, and zebrafish *Pcdh* genes: The variable regions of all chimpanzee, rat, and zebrafish proteins are similar and of almost the same length. They also have the same domain structure. Each variable polypeptide consists of a signal peptide, followed by six tandem EC repeats, a transmembrane segment, and a very short cytoplasmic fragment. The evolutionary relationships between these genes are shown as an unrooted phylogenetic tree (Figure 3). This phylogenetic tree demonstrated that a mixture of divergent groups of *Pcdh* genes exists in specific vertebrate lineages. This analysis suggests that the birth-and-death evolution occurs in this multigene family of the vertebrate nervous system.

The zebrafish *Pcdh* genes do not display orthologous relationships with the mammalian genes and all zebrafish *Pcdh* genes display paralogous relationships (Figure 3). Members of the two zebrafish α clusters can be divided into three groups: Group 1 includes *1α2* and *2α1–2α7*, group 2 includes *2α8–2α25*, and group 3 includes *1α3–1α9* and *2α26–2α29*. Members of the α group 3 are distantly related to the mammalian α genes (Figure 3). Similarly, members of the two zebrafish γ clusters can also be divided into three groups: Group

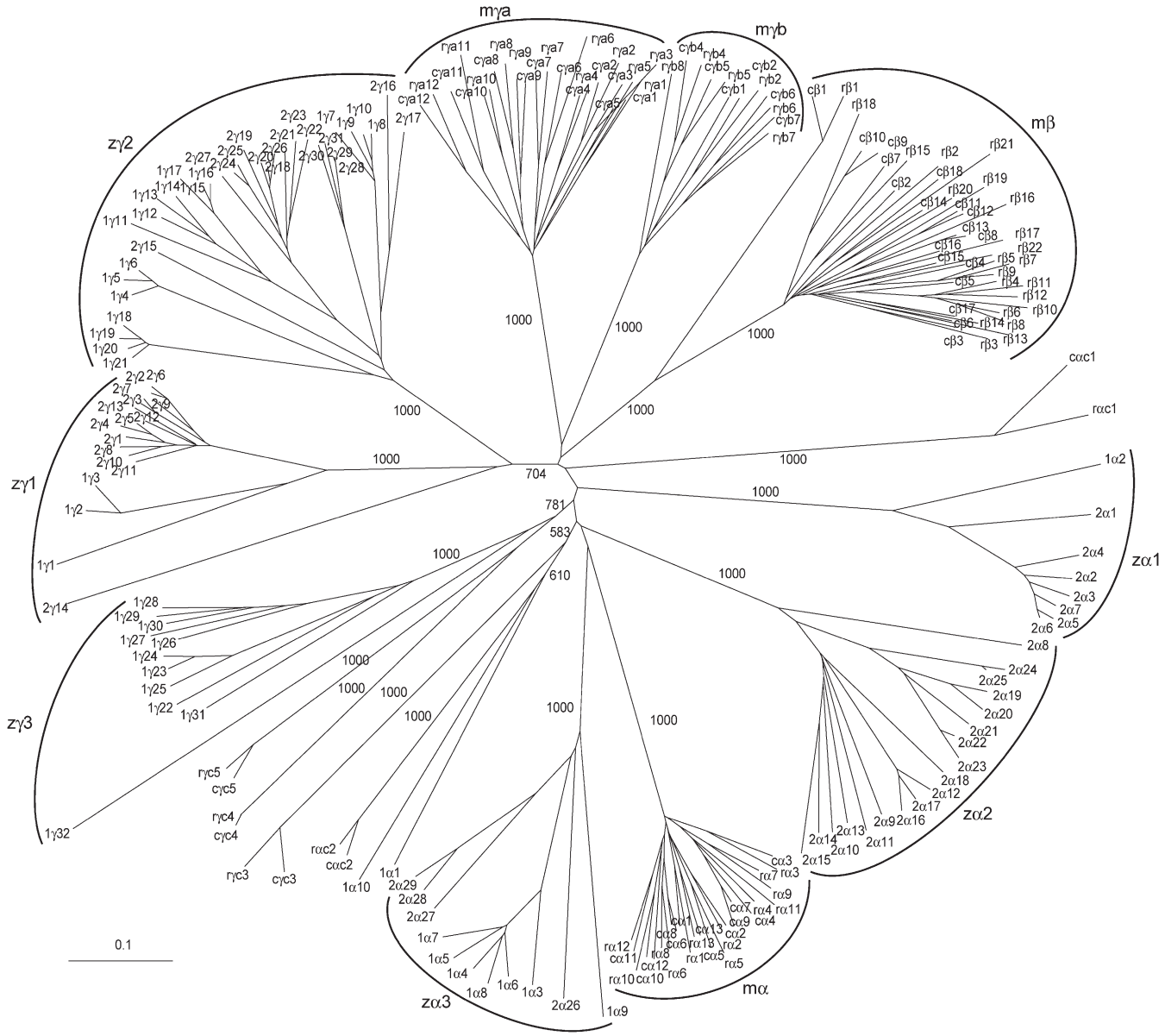

Figure 3.—Phylogenetic tree of chimpanzee (c), rat (r), and zebrafish (z) (*1*α, *1*γ, *2*α, and *2*γ) *Pcdh* clusters. The tree branches are labeled with the percentage support for that partition on the basis of 1000 bootstrap replicates. Only bootstrap values of >50% on major branches are shown. The scale bar equals a distance of 0.1. mα, mβ, mγa, and mγb, mammalian α, β, γ*a*, and γ*b* groups; zα1–3 and γ1–3, zebrafish α and γ groups 1–3.

1 includes *1*γ*1–1*γ*3* and *2*γ*1–2*γ*14*, group 2 includes *1*γ*4–1*γ*21* and *2*γ*15–2*γ*31*, and group 3 includes *1*γ*22–1*γ*32*. Members of the γ group 3 are remotely related to the mammalian *c4* and *c5* genes (Figure 3).

The mammalian *Pcdh* genes can be divided into four groups: α, β, γ*a*, and γ*b*. These four groups and the six zebrafish groups each have a long major branch while members within each group have relatively short secondary branches in the phylogenetic tree (Figure 3). In addition, members within each group share conserved promoter motifs that are related but have diversified considerably among all groups (supplementary text and Figure S3 at http://www.genetics.org/supplemental/). These observations suggest that an ancestral variable exon with its preceding promoter existed for a long time for each group during early vertebrate evolution. Extensive duplications of each ancestral exon occurred separately during zebrafish and mammalian lineages and resulted in the expansion seen within each *Pcdh* group. Interestingly, members of the mammalian β cluster appear to be evolutionarily closer to the mammalian γ*a* and γ*b* genes and are relatively more similar to the zebrafish γ groups 1 and 2 genes. Moreover, the mammalian α cluster is closer to the zebrafish α group 3 genes. These data suggest that the mammalian β, γ*a*, and γ*b* genes are derived from a common ancestor while the mammalian α cluster has a distinct ancestor. Because variable exons of any of the zebrafish groups are

located physically close to each other on the genome (Figure 1), they are likely the results of tandem duplications from an ancestral variable exon of each group. Interestingly, mammalian c-type and zebrafish *1α1* and *1α10* genes seem to be unduplicated singletons (Figure 3). These observations suggest that the mammalian c-type *Pcdh* genes are ancient.

Diversifying selection of the clustered *Pcdh* genes: In the adaptive immune system, antigen presentation, binding, and elimination require unlimited diversity. Positive molecular selection is an important factor for enhancing the IG, MHC, and TCR diversity (Hughes and Nei 1988, 1989; Tanaka and Nei 1989; Sitnikova and Nei 1998; Su and Nei 2001). The complexity of neuronal connections in the CNS also requires staggering diversity. Pcdh proteins have been proposed to provide the specificity required for these diverse neuronal connections. Sequence analyses have demonstrated that the first half of variable exons are divergent and those of the second half are highly conserved (Sugino *et al.* 2000; Wu *et al.* 2001; Noonan *et al.* 2004). Given that vertebrate-specific *Pcdh*, *Ig*, and *Tcr* clusters have similar genomic organizations and that they may provide enormous diversity for the CNS and adaptive immune system, respectively, I hypothesized that some *Pcdh* variable coding regions may also have been subject to diversity-enhancing positive Darwinian selection if the diversity leads to better fitness for the organisms. Recent studies demonstrated that gene conversion occurs in the 3′ extracellular and cytoplasmic coding sequences (Noonan *et al.* 2004). Gene conversion or recombination is known to interfere with the detection of selection sites (Anisimova *et al.* 2003; Shriner *et al.* 2003). Thus, I estimated the ω values of individual sites only on the EC1–3 coding region for various groups of Pcdh proteins in different species.

In the cases of the *Ig*, *Tcr*, and *Mhc* clusters, only the complementarity-determining region (CDR) and the antigen recognition site (ARS) were found to be under positive selection (Hughes and Nei 1988, 1989; Tanaka and Nei 1989; Sitnikova and Nei 1998; Su and Nei 2001). The Pcdh protein sequences are generally conserved among paralogs and between orthologs, suggesting that they are under purifying selection. However, some *Pcdh* coding regions or sites may still be under positive selection. Without knowing which regions or sites are important in protein-protein interactions *a priori*, I separated the coding region into calcium-binding sites and non-calcium-binding regions. Because the calcium-binding sites are absolutely conserved among all members of Pcdh proteins, these sites are under strong purifying selection with $\omega = 0$. I analyzed the non-calcium-binding sites of the first three ECs by using the codeml program (Yang 1997) to estimate the nonsynonymous and synonymous rate ratio.

I ran three pairs of nested codeml models on the 16 data sets of human, chimpanzee, mouse, and rat α, β, γ*a*, and γ*b* paralogous groups and on the 6 data sets of the zebrafish α1–3 and γ1–3 groups to infer positively selected sites (see materials and methods for details). Members of each of these 22 groups are closely related paralogs. The parameter estimates for the 22 paralogous groups are shown in supplementary Table S2 (http://www.genetics.org/supplemental/). The positively selected ω+ sites (Emes *et al.* 2004) in each group are shown in supplementary Table S3 (http://www.genetics.org/supplemental/). The human, chimpanzee, mouse, and rat have overlapping but distinct ω+ site profiles. Even between human and chimpanzee, the ω+ sites are not identical. Although zebrafish has a large number of clustered *Pcdh* genes, no ω+ sites are predicted to be under positive selection (supplementary Table S3). This result suggests that the zebrafish *Pcdh* genes may have been duplicated recently.

I aligned the EC1–3 sequences of the classic C-cadherin to those of the human (supplementary Figure S4A at http://www.genetics.org/supplemental/), chimpanzee (supplementary Figure S4B), mouse (supplementary Figure S4C), and rat (supplementary Figure S4D) Pcdhs by the profile HMM. I then mapped the Pcdh ω+ sites onto the X-ray crystal structure of the first three ECs of C-cadherin (Boggon *et al.* 2002) on the basis of the alignments. Almost all positively selected sites are located on the surface of ECs 2 and 3 (Figure 4). Interestingly, some of the positively selected sites are mapped on the *cis*-interaction interface of C-cadherin (Boggon *et al.* 2002). These diversified sites may participate in combinatorial *cis*-interactions between Pcdh proteins expressed on the same plasma membrane. These results suggest that positive Darwinian selection may be an additional evolutionary factor for increasing *Pcdh* diversity.

DISCUSSION

Lineage-specific duplication and birth-and-death evolution of vertebrate clustered *Pcdh* genes: I show that the *Pcdh* gene clusters are vertebrate specific and are conserved throughout vertebrate evolution (see supplementary text at http://www.genetics.org/supplemental/). Specifically, both zebrafish and mammalian α and γ clusters contain variable and constant regions; and the constant sequences are highly conserved among the major branches of vertebrates (supplementary Figure S2). The clustered variable exons were duplicated in tandem in vertebrate genomes. This large repertoire of similar variable exons is the major source of *Pcdh* diversity. However, the duplications are very different between fish and mammals. For example, mammals have distinct subtypes of the γ*a* and γ*b* genes that were duplicated as pairs (Wu and Maniatis 1999) while zebrafish lacks these subtypes (Noonan *et al.* 2004). Zebrafish also appears to lack the β cluster (Figure 1E). In addition, members of the mammalian β cluster appear to be duplicated in

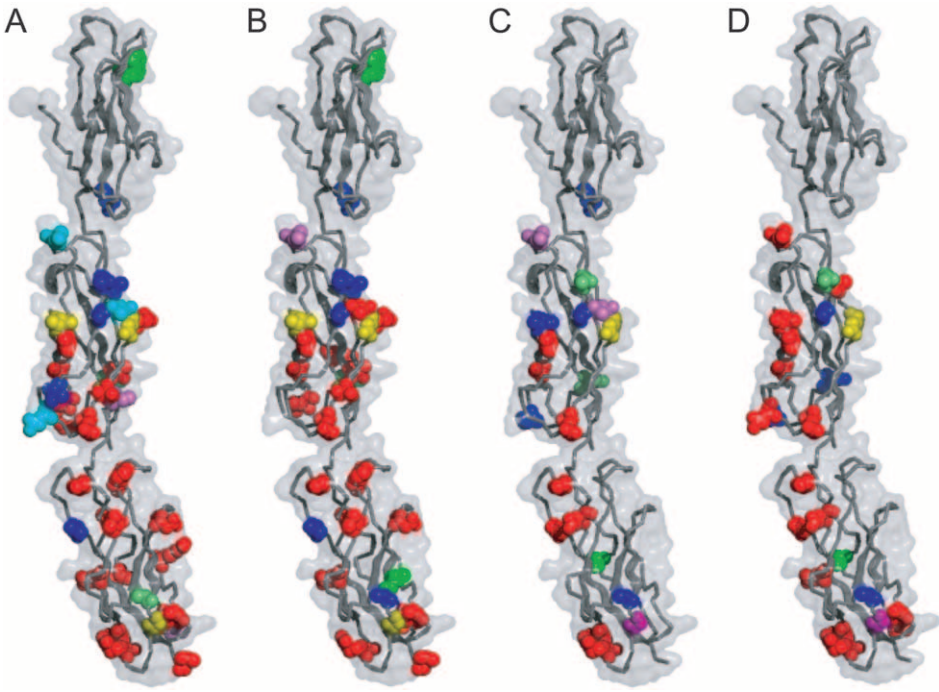

FIGURE 4.—Site-specific K_A/K_S analysis of (A) human, (B) chimpanzee, (C) mouse, and (D) rat *Pcdh* clusters. Shown are the positively selected ω+ sites mapped to a ribbon diagram of the ECs 1–3 crystal structure (from top to bottom) of C-cadherin (BOGGON *et al.* 2002). The residues mapped are comparably numbered according to the sequence of C-cadherin and are equivalent between different Pcdh groups and among the four species. The molecular surface is shown in transparency. The positively selected ω+ sites for the α group only are highlighted in red spheres; for the β group only, in green; for the γ*a* group only, in blue; for the γ*b* group only, in violet; for both the α and γ*a* groups, in yellow; for both the α and γ*b* groups, in cyan; for both the γ*a* and γ*b* groups, in lime; and for the α, β, and γ*a* groups, in magenta. Note that almost all positively selected nonconserved ω+ sites are mapped to the surface of ECs 2 and 3.

tandem from an ancestral variable exon remotely related to the γ variable exon and are unique in that they have become single-exon genes. The variable exons of the β cluster appear to have lost the ability to splice to the constant regions. However, they still have remnant 5′ splice sites conserved among paralogs (WU *et al.* 2001) and can be spliced to the γ constant exons at very low levels (TASIC *et al.* 2002). Moreover, extensive tandem duplications of the zebrafish variable exons occurred after the cluster-wide duplications, since each of the six zebrafish groups has a large number of variable exons and all members within a group are located close to each other (Figures 1 and 3). These observations suggest that *Pcdh* genes are duplicated in a lineage-specific manner. Finally, the *Pcdh* duplication appears to include the variable exon and its promoter. Diversified promoter sequences (supplementary text and Figure S3; WU *et al.* 2001; TASIC *et al.* 2002) and the balancing selection of polymorphic sites in the promoter regions may provide the diversity for gene regulation (NOONAN *et al.* 2003).

Even closely related species have distinct numbers of functional *Pcdh* genes. For example, two members of the β cluster appear to be functional in chimpanzees but have been mutated to pseudogenes in humans. Similarly, the γ*b3* gene appears to be functional in humans but has been mutated to a pseudogene in chimpanzees. A comparison between mouse and rat *Pcdh* clusters reveals that one member of the α cluster has been rendered nonfunctional in mice by a transposon insertion while γ*b1* has been mutated to a pseudogene in rats. These observations indicate that members of the *Pcdh* clusters are subject to evolution by a birth-and-death process (Figure 1) in addition to concerted evolution (NOONAN *et al.* 2004). Therefore, the diversification of *Pcdh* genes required for complex neuronal connectivity in the mammalian brain is achieved through the birth-and-death evolution of species-specific duplication and independent variable-exon mutation, in conjunction with alternative splicing and diversifying selection.

Positively selected residues of Pcdh proteins may participate in combinatorial interactions in the mammalian brain: Cadherin superfamily proteins function in cell plasma membrane adhesion through direct *cis*- and *trans*-interactions of their extracellular ECs (PATEL *et al.* 2003). Members of the mammalian *Pcdh* clusters are expressed mainly in the CNS, where they display distinct cell-specific expression patterns (KOHMURA *et al.* 1998; WANG *et al.* 2002b). The encoded proteins have weak cell adhesion activities (OBATA *et al.* 1995; SAGO *et al.* 1995) and are proposed to provide the vast diversity for specific cell-cell connections in the brain through combinatorial interactions (SHAPIRO and COLMAN 1999).

It is usually assumed that the conserved residues of orthologous proteins are of functional importance because they play the same roles in different species. However, in cases that require great diversity, such as the CDR of IG and the ARS of MHC, the nonconserved residues are of functional importance because they provide the enormous diversity required for adaptive immune defense (HUGHES and NEI 1988; TANAKA and NEI 1989). The diversity-enhancing positive selection operating on these clustered genes provides a mecha-

nism for generating the genetic diversity required for combating pathogens.

I reasoned that adaptive selection in the variable coding regions may be an important source of *Pcdh* diversity. To estimate positively selected sites, I used the maximum-likelihood codeml program because maximum-parsimony and *ad hoc* methods did not account for major features of molecular evolution, such as unequal nucleotide frequencies, transition/transversion rate bias, and codon usage bias (Bielawski *et al.* 2000). I found that positive Darwinian selection operates on a set of sites within specific mammalian EC-coding regions. Given that these exons were duplicated prior to the divergence of rodents and primates, and that some members may not be under positive selection, it is remarkable that positive selection can still be detected within an entire group. It is also striking that almost all positively selected sites are located on the surface of ECs 2 and 3. The enhanced rate of nonsynonymous substitutions at specific sites in ECs 2 and 3 may allow very large numbers of combinatorial *cis*-interactions among paralogous Pcdh proteins expressed on the same synaptic surface. The fact that the nonsynonymous substitution rate is higher than the synonymous substitution rate at these sites suggests that diversity-enhancing selection actively creates differences among *Pcdh* paralogs in mammalian species. These diversified residues in ECs 2 and 3 may be functionally important regions of the Pcdh proteins.

Classic cadherins have five repetitive ECs, each of which has a Greek-key β sandwich folding topology with four strands facing one side of the molecule and three strands facing the opposite side (Patel *et al.* 2003). Recent structural studies clearly show that EC1 functions in *trans*-interaction between cadherins expressed on the plasma membranes of the neighboring cells (Boggon *et al.* 2002), consistent with numerous biochemical and cell biological studies. However, molecular force measurements and cell-based assays suggest that additional ECs play a role in cell adhesion (Sivasankar *et al.* 1999; Chappuis-Flament *et al.* 2001). The crystal structure of the entire extracellular domain of C-cadherin reconciles the apparent discrepancy by demonstrating that EC2 may participate in *cis*-interactions between cadherins expressed on the same cell surface (Boggon *et al.* 2002).

I have modeled the first three ECs of the Pcdh proteins. Each EC displays a Greek-key β sandwich folding structure (data not shown). I propose that ECs 2 and 3 function in *cis*-interactions between Pcdhs expressed on the same cell membrane. Because a single neuron expresses multiple *Pcdh* genes (Kohmura *et al.* 1998; Tasic *et al.* 2002) and Pcdh proteins of different groups may interact in *cis* (Murata *et al.* 2004), the *cis*-interactions between the positively selected sites in ECs 2 and 3 potentially generate a large spectrum of combinations. Specific *trans*-interaction between the EC1s of these vast numbers of *cis*-combinations could provide enormous diversity for neuronal connectivity in the CNS. Consistent with this idea, some positively selected sites are mapped to the residues in C-cadherin EC2 that are located in the *cis*-interaction interface (Boggon *et al.* 2002). Different species have distinct ω+ site profiles (supplementary Table S3). This supports that each species has a lineage-specific evolutionary process for the neuronal connections in the brain. Interestingly, primates have more positively selected sites than rodents (supplementary Table S4 at http://www.genetics.org/supplemental/), consistent with the increased brain complexity in primates. In addition, zebrafish has no positively selected sites although it has more *Pcdh* genes than mammals. Thus, the zebrafish Pcdh proteins may provide much less combinatorial diversity. If my conclusion that diversity-enhancing positive selection occurs at the *Pcdh* variable coding region is correct, it will be the first evidence that adaptive evolution actively selects diversity for *Pcdh* gene clusters. The combinatorial interactions between Pcdh proteins have not been proved by experiments. Nevertheless, this analysis suggests a direction for future structural and mutagenesis studies on the importance of positively selected residues.

Note added in proof: While this manuscript was under review, J. P. Noonan, J. Grimwood, J. Danke, J. Schmutz, M. Dickson *et al.* (2004, Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. Genome Res. **14:** 2397–2405) reported the coelacanth *Pcdh* gene clusters. By comparison with the *Pcdh* gene clusters of other vertebrates, they concluded that the coelacanth *Pcdh* clusters are likely very similar to those of the tetrapod ancestor. Therefore, valuable knowledge on vertebrate evolution would be gained by obtaining a complete coelacanth genome sequence.

I am indebted to M. Capecchi and T. Maniatis for encouragement. I am grateful to L. Jorde, J. Metherall, R. Myers, J. Seger, W. Sundquist, and D. Witherspoon for critical reading of the manuscript and to C.B. Chien and D. Grunwald for providing zebrafish. I thank P. Haws for technical assistance and H. Peng, F. Whitby, D. Witherspoon, S. Wu, and G. Ying for many useful suggestions. Q.W. is a March of Dimes Basil O'Connor Scholar and an American Cancer Society Research Scholar.

LITERATURE CITED

Anisimova, M., R. Nielsen and Z. Yang, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164:** 1229–1236.

Bielawski, J. P., K. A. Dunn and Z. Yang, 2000 Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics **156:** 1299–1308.

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

Chappuis-Flament, S., E. Wong, L. D. Hicks, C. M. Kay and B. M. Gumbiner, 2001 Multiple cadherin extracellular repeats mediate homophilic binding and adhesion. J. Cell Biol. **154:** 231–243.

Chen, F. C., and W. H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

Crow, T. J., 2002 Handedness, language lateralisation and anatomical asymmetry: relevance of protocadherin XY to hominid speciation and the aetiology of psychosis. Br. J. Psychiatry **181:** 295–297.

Emes, R. D., S. A. Beatson, C. P. Ponting and L. Goodstadt, 2004

Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. Genome Res. **14:** 591–602.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren *et al.*, 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428:** 493–521.

Goldman, N., and S. Whelan, 2000 Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. **17:** 975–978.

Guex, N., and M. C. Peitsch, 1997 SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis **18:** 2714–2723.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

Hughes, A. L., and M. Nei, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86:** 958–962.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res. **12:** 996–1006.

Kohmura, N., K. Senzaki, S. Hamada, N. Kai, R. Yasuda *et al.*, 1998 Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. Neuron **20:** 1137–1151.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler, 1994 Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. **235:** 1501–1531.

Murata, Y., S. Hamada, H. Morishita, T. Mutoh and T. Yagi, 2004 Interaction with protocadherin-gamma regulates the cell-surface expression of Protocadherin-alpha. J. Biol. Chem. **279:** 49508–49516.

Nielsen, R., and Z. Yang, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929–936.

Noonan, J. P., J. Li, L. Nguyen, C. Caoile, M. Dickson *et al.*, 2003 Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin alpha cluster. Am. J. Hum. Genet. **72:** 621–635.

Noonan, J. P., J. Grimwood, J. Schmutz, M. Dickson and R. M. Myers, 2004 Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. **14:** 354–366.

Obata, S., H. Sago, N. Mori, J. M. Rochelle, M. F. Seldin *et al.*, 1995 Protocadherin Pcdh2 shows properties similar to, but distinct from, those of classical cadherins. J. Cell Sci. **108:** 3765–3773.

Olson, M. V., and A. Varki, 2003 Sequencing the chimpanzee genome: insights into human evolution and disease. Nat. Rev. Genet. **4:** 20–28.

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

Phillips, G. R., H. Tanaka, M. Frank, A. Elste, L. Fidler *et al.*, 2003 Gamma-protocadherins are targeted to subsets of synapses and intracellular organelles in neurons. J. Neurosci. **23:** 5096–5104.

Sago, H., M. Kitagawa, S. Obata, N. Mori, S. Taketani *et al.*, 1995 Cloning, expression, and chromosomal localization of a novel cadherin-related protein, protocadherin-3. Genomics **29:** 631–640.

Shapiro, L., and D. R. Colman, 1999 The diversity of cadherins and implications for a synaptic adhesive code in the CNS. Neuron **23:** 427–430.

Shriner, D., D. C. Nickle, M. A. Jensen and J. I. Mullins, 2003 Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. **81:** 115–121.

Sitnikova, T., and M. Nei, 1998 Evolution of immunoglobulin kappa chain variable region genes in vertebrates. Mol. Biol. Evol. **15:** 50–60.

Sivasankar, S., W. Brieher, N. Lavrik, B. Gumbiner and D. Leckband, 1999 Direct molecular force measurements of multiple adhesive interactions between cadherin ectodomains. Proc. Natl. Acad. Sci. USA **96:** 11820–11824.

Sperry, R. W., 1963 Chemoaffinity in the orderly growth of nerve fiber patterns and connections. Proc. Natl. Acad. Sci. USA **50:** 703–710.

Su, C., and M. Nei, 2001 Evolutionary dynamics of the T-cell receptor VB gene family as inferred from the human and mouse genomic sequences. Mol. Biol. Evol. **18:** 503–513.

Sugino, H., S. Hamada, R. Yasuda, A. Tuji, Y. Matsuda *et al.*, 2000 Genomic organization of the family of CNR cadherin genes in mice and humans. Genomics **63:** 75–87.

Tada, M. N., K. Senzaki, Y. Tai, H. Morishita, Y. Z. Tanaka *et al.*, 2004 Genomic organization and transcripts of the zebrafish Protocadherin genes. Gene **340:** 197–211.

Tanaka, T., and M. Nei, 1989 Positive Darwinian selection observed at the variable-region genes of immunoglobulins. Mol. Biol. Evol. **6:** 447–459.

Tasic, B., C. E. Nabholz, K. K. Baldwin, Y. Kim, E. H. Rueckert *et al.*, 2002 Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol. Cell **10:** 21–33.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Wang, X., H. Su and A. Bradley, 2002a Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. Genes Dev. **16:** 1890–1905.

Wang, X., J. A. Weiner, S. Levi, A. M. Craig, A. Bradley *et al.*, 2002b Gamma protocadherins are required for survival of spinal interneurons. Neuron **36:** 843–854.

Wernersson, R., and A. G. Pedersen, 2003 RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. **31:** 3537–3539.

Wu, Q., and A. R. Krainer, 1999 AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol. Cell. Biol. **19:** 3225–3236.

Wu, Q., and T. Maniatis, 1999 A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell **97:** 779–790.

Wu, Q., T. Zhang, J. F. Cheng, Y. Kim, J. Grimwood *et al.*, 2001 Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res. **11:** 389–404.

Yanase, H., H. Sugino and T. Yagi, 2004 Genomic sequence and organization of the family of CNR/Pcdhalpha genes in rat. Genomics **83:** 717–726.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Yang, Z., and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:** 496–503.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Zhang, T., P. Haws and Q. Wu, 2004 Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. Genome Res. **14:** 79–89.

Communicating editor: Z. Yang