Protocadherin domain interactions provide a molecular logic of neuronal self-recognition

Rotem Rubinstein[†1,2,4], Chan Aye Thu [†1], Kerry Marie Goodman[†1,2], Holly N Wolcott[†1], Fabiana Bahna[1,2,4], Seetha Mannepalli[1,2], Goran Ahlsen[1,2,4], Maxime Chevee[1], Adnan Halim[5], Henrik Clausen[5], Tom Maniatis[1], Lawrence Shapiro[1,2*], and Barry Honig[1,2,3,4*]

[1]Department of Biochemistry and Molecular Biophysics

[2]Department of Systems Biology

[3]Department of Medicine

[4]Howard Hughes Medical Institute

Columbia University, New York, NY 10032, USA.

[5]Copenhagen Center for Glycomics, Departments of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

[†] Co-first authors.

*Corresponding authors: bh6@columbia.edu; lss8@columbia.edu

Summary

Self-avoidance, a process preventing interactions of axons and dendrites from the same neuron during development, is mediated in vertebrates through the stochastic single-neuron expression of clustered protocadherin protein isoforms. Extracellular cadherin (EC) domains mediate isoform-specific homophilic binding between cells, conferring cell recognition through a poorly understood mechanism. Here, we report crystal structures for the EC1-EC3 domain regions from four protocadherin isoforms representing the α, β and γ subfamilies. All are rod-shaped and monomeric in solution. Biophysical measurements, cell aggregation assays, and computational docking reveal that *trans* binding between cells depends on the EC1-EC4 domains, which interact in an antiparallel orientation. We also show that the EC6 domains are required for the formation of *cis*-dimers. Overall, our results are consistent with a model in which protocadherin *cis*-dimers engage in a head-to-tail interaction between EC1-EC4 domains from apposed cell surfaces, possibly forming a zipper-like protein assembly thus providing a size-dependent self-recognition mechanism.

Introduction

The human brain is comprised of approximately 10 billion neurons, each of which can connect with up to thousands of others. Neuronal self-avoidance is a process in which dendrites and axons originating from the same neuron repel one another, but can freely interact with neurites from other neurons. The combined properties of self-recognition and non-self discrimination require that contacting neurons display diverse cell surface identities that allow for discrimination between self and non-self (Hattori et al., 2009; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010).

In *Drosophila* and other invertebrates self-avoidance is mediated by Dscam1 proteins; immunoglobulin superfamily members produced by alternative splicing of the *DSCAM1* pre-mRNA. This cell-autonomous and stochastic alternative splicing can theoretically produce up to 19,008 Dscam1 isoforms with distinct ectodomains, each of which have highly specific homophilic *trans* binding specificity (Hattori et al., 2008; Miura et al., 2013; Schmucker et al., 2000; Wojtowicz et al., 2007). Distinct cell surface identities are generated in *Drosophila* by the stochastic expression of a small set of Dscam1 isoforms in each neuron (Miura et al., 2013). Homophilic interactions between identical sets of protein isoforms on the surface of neurites from the same neuron result in repulsion and neurite self-avoidance (Hattori et al., 2008). The expression of even a single Dscam1 isoform is sufficient for self-avoidance of neurites from the same neuron (Hughes et al., 2007; Matthews et al., 2007; Soba et al., 2007). However, robust non-self discrimination, which allows processes from different neurons to freely interact, requires thousands of distinct Dscam1 isoforms (Hattori et al., 2009).

Recent studies suggest that in vertebrate nervous systems neuronal self-avoidance functionality is provided, at least in part, by the clustered protocadherins (Pcdhs) (Chen and Maniatis, 2013; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010). Mammalian Pcdhs are encoded in a contiguous genomic locus comprised of three adjacent gene clusters (*Pcdh α, β and γ*), each of which contains close to 60 “variable” exons (58 in mice, Figure 1A) (Wu and Maniatis, 1999). Only a few variable exons are stochastically chosen for expression in each cell by a mechanism involving alternative promoter choice (Ribich et al., 2006; Tasic et al., 2002). Each variable exon encodes an entire Pcdh ectodomain region consisting of six tandem extracellular cadherin (EC) domains, a single transmembrane region, and a short cytoplasmic region.

In the α and γ gene clusters, a “constant” C-terminal cytoplasmic region encoding an intracellular domain (ICD) is joined to the variable ectodomain exon by pre-mRNA splicing. The β cluster does not contain such a constant region and therefore β-Pcdhs are lacking an ICD. The α and γ gene clusters also encode a small set of “C-type” Pcdhs, which are divergent from other members of their respective clusters, and appear to have distinct functions (Figure 1A) (Chen et al., 2012). Deletion of the Pcdhγ gene cluster in mice leads to the disruption of self-avoidance in retinal starburst amacrine cells and Purkinje cells with phenotypes similar to those described for Dscam1 deletion mutants in Drosophila (Lefebvre et al., 2012).

Like invertebrate Dscam proteins, Pcdh isoforms engage in isoform-specific *trans* homophilic interactions (Schreiner and Weiner, 2010; Thu et al., 2014). It is remarkable that Pcdhs, with only 58 isoforms, can mediate neural self-recognition and non-self discrimination similar to Dscams, which have up to tens of thousands of distinct extracellular isoforms. Central to this capability is the observation that a single mismatched Pcdh isoform can interfere with recognition between cells that express an otherwise matching set of Pcdhs (Thu et al., 2014). Understanding the mechanism underlying this “interference” phenomenon is crucial, as it is likely to explain how only 58 Pcdh isoforms can provide sufficient functional diversity to enable self-recognition and non-self discrimination in the nervous system comparable to the much more diverse Drosophila Dscam gene.

Here we report crystal structures of Pcdh extracellular protein fragments comprising the previously mapped Pcdh specificity-determining EC1-EC3 domains for PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 isoforms, thus providing examples from all three Pcdh gene clusters. Guided by these structures we used two orthogonal mutagenesis approaches – surface saturating arginine mutagenesis and bioinformatics-derived predictions – to map the isoform specificity-determining regions at the amino acid level using cell aggregation and biophysical experiments as readouts. The two approaches yielded consistent results, revealing an essential role for EC1 through EC4 in *trans* homophilic interactions and for EC6 in *cis* interactions. On the basis of these findings we propose a model for Pcdh mediated cell-cell recognition that is consistent with the remarkable ability of these cell surface proteins to provide diverse single-cell identities to vertebrate neurons.

Results

Structures of Pcdh EC1-EC3 region fragments from α, β and γ sub-families

We determined crystal structures of proteins composed of the three N-terminal EC domains of mouse PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 to a resolution of 2.4 Å, 3.3 Å, 2.9 Å and 2.9 Å, respectively (Figure 1B, Table S1). We focused on protein fragments containing EC1-EC3, since the results of earlier cell aggregation experiments indicated that Pcdh isoform-specific recognition was mediated via the EC2-EC3 domains and that the EC1 domain is required for *trans* binding (Schreiner and Weiner, 2010).

The four structures show high overall similarity (Figures 1B and S1A). Each structure consists of three EC domains, each with the two-layer β-sheet fold observed in classical cadherins. Successive domains are connected by calcium-binding linkers, each of which coordinate three Ca^{2+} ions utilizing side chains in the same conserved motifs (Figure 1B). These motifs are also conserved within type-I and type-II classical cadherins with the exception of the EE motif (bottom of EC1 domain, Figure 1B), which is present only in type-II cadherins. In contrast with previous conclusions (Schreiner and Weiner, 2010), but consistent with the presence of Ca^{2+} at the inter-domain linkers and in common with classical cadherins, we have found that cell aggregation of Pcdhs is Ca^{2+} dependent (Figure S1B). Despite these similarities to classical cadherins, the Pcdh isoform structures are distinctive in several aspects. Most notably, the overall arrangement of the three EC domains in each structure is much straighter than the curved classical cadherin architecture (Figure 1C). This “straight-rod” architecture arises from an extended zigzagged conformation: an arrangement that is generated primarily by a very different EC2-EC3 angle than classical cadherins (>131° difference, Figure 1D).

In addition, mass spectrometry analyses showed that all four isoforms contain two sites of O-mannosylation at residues 194 and 196 (PcdhγC5 sequence numbering, Figures 1B and S1 panels G and H). These positions are conserved in sequence among most Pcdh isoforms (Fig. S1G) and among classical cadherins (Vester-Christensen et al., 2013), suggesting these O-glycans play important functional roles. O-mannosylation of cadherins and protocadherins were recently discovered (Vester-Christensen et al.,

2013), and it was further shown that O-mannosylation of E-cadherin is essential for preimplantation development of the mouse embryo (Lommel et al., 2013).

The Pcdh structures show local Pcdh-specific embellishments on the EC domain fold. In particular Pcdh EC1 domains show a number of differences from vertebrate cadherin EC1 domains (Figure S1D), as was previously observed in NMR structures of Pcdhα4 and Pcdhβ14 EC1 domains (Morishita et al., 2006). The A-strand is shorter than that of classical cadherins and lacks the conserved Trp-2 residue, which anchors the strand-swap trans-binding interface of classical cadherins (Figures S1C and S1D; Posy et al., 2008). The EC1 EF loop region in each of the Pcdh structures contains a disulfide-constrained loop formed by a Pcdh-specific CX_5C motif. The EC2 and EC3 domains of the Pcdh structures are each most similar to either the EC1 or EC2 domain from the atypical cadherin-23 (RMSD 1.5 and 1.2 Å). However, the D and E strands of Pcdh EC2 domains, and the CD loop region of EC3, are significantly longer than found in cadherin-23 or in classical cadherins (Figure S1E). There are also distinctive differences among the structures of the four Pcdh isoforms. The EC1 BC loop helix, C strand and CD loop regions display distinct conformations in all four structures (Figure S1F). In EC3 the two C-type structures (PcdhαC2 and PcdhγC5) have a longer FG loop than Pcdhβ1 and PcdhγA8, a feature conserved among α and C-type Pcdhs (Figure S1F).

Analysis of the molecular packing of the four Pcdh EC1-EC3 structures revealed different crystallographic contacts for each isoform, with no interfaces in common. Interfaces exhibiting typical protein-protein interface attributes were not identified in any of the crystal forms analyzed.

Analytical ultracentrifugation and cell aggregation assays define the multimeric structure of Pcdhs

We expressed and purified proteins from a C-terminal deletion series comprising EC1-EC6, EC1-EC5, EC1-EC4, and EC1-EC3, and a construct comprising domains EC2-EC6 where EC1 was deleted. Using AUC we assessed the oligomerization state of each of these ectodomain fragments in solution. With the exception of PcdhγA8, all EC1-EC3 Pcdh isoform fragments behaved as monomers (Table 1A). This finding was consistent with our crystal structures in which no apparent binding interfaces were

detected. The PcdhγA8 EC1-EC3 fragment formed a disulfide-linked dimer through cysteine 283 in the EC3 domain (Figure S2A-B); however, this disulfide bond is likely artifactual since it is not detected in the larger PcdhγA8 isoform fragment (EC1-EC4) (Table 1A).

In contrast to monomeric EC1-EC3 fragments, EC1-EC4 or EC1-EC5 Pcdh fragments were observed to self-associate as dimers with dissociation constants (K_D) in the micromolar range (2.9 – 100 μM) that varied significantly between isoforms (Table 1A). The EC1-deleted constructs comprising domains EC2-EC6 also formed homodimers in solution, with K_D values in the low micromolar range (8.9 - 23 μM). Importantly, AUC measurements for complete ectodomains including EC1-EC6 could be fit only to a tetramer (dimer-of-dimers) model, indicating a crucial role for the EC6 domain in Pcdh association (Table 1A).

We expressed similarly truncated Pcdhs in K562 cells and assessed their ability to mediate cell aggregation. K562 cells provide a robust assay for Pcdh cell-cell recognition, as they do not express endogenous Pcdhs and do not spontaneously aggregate in liquid culture (Reiss et al., 2006; Schreiner and Weiner, 2010; Thu et al., 2014). Cells expressing the EC1-EC3 fragment, which was found to be monomeric in solution, failed to produce cell aggregates (Figure 2A). In contrast, with the exception of PcdhgC5 EC1-EC4 which forms a non-natural disulfide between monomers, cells expressing either EC1-EC4, EC1-EC5, or the complete ectodomain (EC1-EC6), showed extensive aggregation for all isoforms tested (Figure 2A). Consistent with previous studies (Schreiner and Weiner, 2010; Thu et al., 2014), cells expressing Pcdh EC2-EC6 fragments, which were shown above to homodimerize in solution, did not aggregate (Figure 2A). Detection of two independent dimers, one of which (generated by EC1-EC4 and EC1-EC5 fragments) correlates with cell-cell aggregation while the other (generated by EC2-EC6 fragments) does not (Figure 2A), strongly suggests that EC1-EC4 and EC1-EC5 fragments mediate *trans* interactions while the EC2-EC6 fragments mediate *cis* interactions involving the most membrane-proximal domain, EC6 (see also below). The observation that full-length ectodomains form apparent tetramers in AUC strongly suggests that this molecular species corresponds to a dimer-of-dimers formed by these two distinct interfaces, one mediating *cis* and the other *trans* interactions.

Structural elements of the *trans*-binding interface

Arginine-scanning mutagenesis - Selected non-basic surface residues of the PcdhγC5 EC1-EC3 domains revealed in the crystal structure were individually mutated to arginine, and the homophilic recognition function of these single-arginine-mutant proteins was assessed using the K562 cell aggregation assay. Selected basic surface residues were mutated to glutamic acid. As expected, the majority of single-point mutant proteins exhibited wild-type cell aggregation phenotypes (Figures S2C). In contrast, cells transfected with the arginine point mutant L87R in the EC1 domain, S116R and T142R in the EC2 domain, and M301R and E302R in the EC3 domain of PcdhγC5 showed no detectable aggregation (Figure 2B-C). Cells transfected with the EC2 S114R mutation showed diminished homophilic binding (Figure S2C). S114 and S116 are located in the AB loop connecting the A and B β-strands in EC2 while M301 and E302 are located in the FG loop of EC3. All are located on one side of the molecule and are very close to one another in space, thus defining a potentially continuous homophilic recognition interface with elements distributed over the EC2 and EC3 domains. Notably, L87 in EC1 faces in the same direction although T142 in EC2 does not.

To determine whether this binding region is unique to PcdhγC5, we produced mutants for isoforms from all three Pcdh gene clusters for residues structurally equivalent to PcdhγC5 positions 87, 116, and 301. Mutations equivalent to 301R abolished homophilic recognition for isoforms from all three gene clusters (Pcdhα7, PcdhαC2, Pcdhβ6, PcdhγA8, and PcdhγB6, Figure 2D). Homophilic recognition was abolished for mutations equivalent to 116R for isoforms from the α and γ gene cluster members (Pcdhα7, PcdhαC2, PcdhγA8), but not for the isoforms we tested from the β and γB cluster (Figure 2D). Finally, mutations equivalent to L87R abolished homophilic recognition for PcdhγA8 and diminished homophilic recognition for Pcdhα7. It is possible that homophilic recognition for the Pcdhβ6 and PcdhγB6 isoforms may not involve residues 87 in EC1 and 116 in EC2 or, alternatively, arginine mutants of these residues might not appropriately test their contribution to binding. Below we show that isoforms from the α and β gene clusters do in fact utilize interface residues in the EC2 AB loop region and others in close structural proximity to EC1 residue 87.

Domain shuffling to identify specificity-determining domains - Within each of the mouse gene clusters there exist pairs of Pcdh isoforms (Pcdhα7-Pcdhα8, Pcdhβ6-Pcdhβ8, and

PcdhγA8-PcdhγA9) with greater than 80% pairwise sequence identity within their EC1-EC4 domain regions. Despite this high identity these pairs display strict homophilic specificities (Thu et al., 2014). In order to help identify the binding interface we produced chimeras in which EC domains were shuffled between the closely related isoforms. These proteins were tagged at the C-terminus with either of the fluorescent proteins mCherry or mVenus, and tested for binding specificity in the K562 cell assay. We confirmed that all three pairs bind strictly homophilically (Figure 3A1-4, 3B1-4, 3C1-4).

The results of cell aggregation experiments using different chimeric constructs are summarized in Figures 3 and S3. These results are presented in such a way that two closely related wild-type “parent” proteins appear at the left of each panel while each figure indicates whether a particular chimera co-aggregates with one or the other parent protein, or prefers to aggregate homophilically. Figure 3D summarizes the data presented in Figures 3A-C. All chimeric constructs containing EC1-EC3 domains from one isoform and EC4-EC6 domains from another co-aggregated with the wild-type “parent” isoform that contained the same EC1-EC3 domains (Figure 3A-C panel 6 and Figure S3B and D panel 13), whereas chimeric constructs with just EC2-EC3 shuffled, preferred to aggregate homophilically (Figure S3A-E panels 11 and 12).

Despite the fact that shuffling EC1-EC3 is sufficient to swap specificity in close pairs, our AUC and cell aggregation assay results (Table 1A and figure 2A) indicate that all four N-terminal domains (EC1-EC4) are required for *trans* homophilic recognition. We therefore generated a chimera of PcdhγA8 in which domains EC2-EC4 were replaced with the corresponding domains of the closely related PcdhγA9 isoform, while domains EC5-EC6 were replaced with the EC5-EC6 domains of the distant PcdhγB6 isoform, which would not be expected to interact in *trans* with PcdhγA8 or PcdhγA9. Cells expressing this chimera adhere to cells expressing PcdhγA9 indicating, consistent with AUC data, that the EC4 domain plays a role in determining homophilic binding specificity (Figure 3C panel 8). This conclusion is also supported by cell aggregation studies using chimeras where EC1 is derived from one parent and EC2-EC6 from another. In all cases, these chimeras co-aggregate with the parent containing the same EC2-EC6 domains (Figure S3A, C, and E panel 1 and S3B and D panel 2). Since domains EC5 and EC6 are not required for *trans* binding these results also implicate EC2-EC4 as sufficient to determine homophilic specificity.

The experiments reported in Figure S3 help define the minimal number of domains within the EC1-EC4 region that determine the binding properties of a chimera. The presence of a single domain is never enough to mediate co-aggregation with a parent isoform containing this domain (Figure S3A, C, and E panels 2, 4 and 6, S3B and S3D panels 1, 3, and 5) but, in some cases, a mismatched single domain is capable of disrupting binding to the parent isoforms (FigureS3C panel 5, S3D panel 6 and S3E panel 3). In a few cases, the presence of just two domains in common is sufficient to mediate co-aggregation with a parent even if the other four domains are different. This can be seen in: a chimera containing EC1 and EC3 from γA9 and EC2 and EC4-EC6 from γA8 which co-aggregates with wild-type γA9 (Figure S3C panel 10), and a chimera containing EC1 and EC2 from β8 and EC3-EC6 from β6 which co-aggregates with wild-type β8 (Figure S3E panel 8). Overall, these results are consistent with all four N-terminal domains, EC1-EC4, contributing to *trans* binding with the relative contributions of each domain to specificity varying from one isoform to another.

Rational design of point mutations to identify specificity-determining residues - Sequence alignment of specificity-determining EC3 domains shows that Pcdhα7 and Pcdhα8 differ in five amino acids whereas PcdhγA8 and PcdhγA9 differ in eight (Figure 4A). Notably, in both cases, three of these residues are located in the same structural element: the FG loop (Figures 4A, 5A, and 5C). In the case of PcdhγA8 and PcdhγA9 the three variable FG loop residues are highly conserved within their respective orthologs (Figure 4B). Together, these data strongly suggest that these three EC3 domain FG loop residues act as specificity determinants for α and γ Pcdh isoforms.

To test this hypothesis experimentally, we swapped the three residues (Figure 5) between the EC3 domains of closely related isoforms and tested their binding specificities with their “parent” native isoforms. We produced chimeras with the three FG-loop residues of one isoform replaced with the corresponding residues of its close-pair isoform. These three-residue-swapped mutants were tested, along with their native “parents”, in the K562 cell aggregation assay. Cells expressing an isoform in which the three FG-loop residues were replaced with those from the close-pair isoform intermixed with cells expressing the wild-type isoform with residues identical to those at the shuffled positions (Figure 5A and 5C). In contrast, these cells segregated from cells expressing the wild-type isoform from which the EC3 domain originated (Figure S4). We conclude

that the three variable residues of the EC3 FG loop are specificity-determining in the closely related α and γ isoforms.

A similar analysis was carried out for EC1 and EC2 domains with comparable results. As with the EC3 domains, we analyzed close isoform pairs (Figure 4A) and identified candidate specificity-determining residues located on the EC1 C strand and EC2 AB region (Figure 5). We validated these assignments by showing that shuffling residues between EC2 domain AB regions resulted in swapped specificities for close-pair isoforms from all three Pcdh gene clusters (Figures 5 & S4). Shuffling residues between EC1 domain C strand regions was sufficient to swap EC1 specificities from Pcdhβ6 to that of Pcdhβ8 or from Pcdhα7 to Pcdhα8. The contribution of this region in the Pcdhγ pair could not be determined since shuffling of residues in this region resulted in a protein that could not mediate cell aggregation (Figures S4D). We note that swapping EC1 specificities from Pcdhβ6 to Pcdhβ8 or EC2 specificities from Pcdhα7 to Pcdhα8 or from PcdhγA9 to PcdhγA8 required the alteration of only a single residue (residue R41N, L114P and S114N for β, α and γ respectively, Figure 5).

Rational and random mutagenesis identify the same functional binding surfaces - Figures 2 and 5 list specificity determining residues identified from arginine scanning and bioinformatics-based mutagenesis. The finding that two different approaches implicate the same structural regions in Pcdh homophilic binding, and that these regions are in common for isoforms from different Pcdh gene clusters, indicates that these regions – the EC1 C, and G strands, the EC2 AB loop and EC3 FG loop (Figure 5D) are likely to contribute to determining the binding specificities for other Pcdh isoforms as well. As shown above, EC4 contributes to the *trans* binding specificity in a similar way to that of EC1. However, we focused on the EC1-EC3 domains because this is the region for which we have atomic-level structures.

AUC experiments on mutant proteins confirm that Pcdh trans-interactions occur via EC1-EC4 domains, whereas cis interactions occur via the EC6 domain - We have provided evidence from both AUC and cell aggregation assays that the EC1-EC4 domains mediate Pcdh *trans* interactions, whereas the EC6 domain mediates an independent Pcdh *cis* interaction. To provide further evidence for these findings we expressed and purified various domain-truncated constructs of PcdhγA8-I116R, PcdhγC5-S116R, and PcdhαC2-S118R. Since an arginine at these positions ablates *trans* binding in cell

aggregation assays these mutant constructs should only affect the Pcdh *trans*-association but not the *cis*-association in AUC experiments. As expected the EC1-EC4 fragment of I116R PcdhγA8 behaved differently from its wild-type counterpart and was monomeric in solution (Table 1B). In contrast, we found that similar to its wild-type counterpart, the EC2-EC6 fragment PcdhγC5-S116R behaved as a dimer with K_D similar to wild-type EC2-EC6. This observation suggests that the EC2-EC6 protein dimerizes in *cis* through a region that is not involved in the *trans* interface (Table 1B). Finally, the complete ectodomain of PcdhαC2 containing an S118R mutation displayed tetramerization affinity, which was an order of magnitude lower than that of the wild-type protein. Similarly, the S116R mutant of PcdhγC5 EC1-EC6 did not form tetramers (as does its wild-type counterpart) but rather, similar to the EC2-EC6 fragment, self-associates as a dimer. Since *trans* binding has been ablated by this mutation, the observed dimer must correspond to association in *cis* (Table 1B).

The *trans* homophilic interface is formed via head-to-tail interactions of EC1-EC4 domains

Computational docking yields antiparallel orientations - We carried out modeling studies in an effort to elucidate the dimerization mode of Pcdhs. We limited our modeling to EC1-EC3 for which we have determined crystal structures and have identified specificity-determining residues. We used the M-zdock program (Pierce et al., 2005) to produce symmetric homodimeric models for the EC1-EC3 domain regions of PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5. We generated thousands of models for each crystal structure, and used the experimentally identified specificity determinant residues to filter the docked models; requiring models to include these residues at the binding interface. A second constraint required docking models to have a buried surface area at the binding interface of more than 1200 $Å^2$ (600 $Å^2$ per protomer). Applying these two conditions reduces the number of docked models from thousands to 149: 23, 40, 40 and 46 for PcdhγA8, Pcdhβ1, PcdhαC2, and PcdhγC5, respectively. We then structurally clustered the filtered docked homodimers with the expectation that there would be more docked structures near the native conformation.

Notably, the majority of the filtered docked homodimeric Pcdhs (62.5%) adopted a head-to-tail orientation of the two molecules in which the EC2 domain of one molecule interacts with the EC3 domain of its partner (Figures 6A and S5Ai-ii). Furthermore, most

structures with this binding mode place the EC1 domain of one molecule adjacent to the expected position of the EC4 domain of its partner (Figure 6A). Only three of the docked and filtered complexes had a head-to-head orientation (two for PcdhγC5 and one for PcdhαC2, Figure 6Aiii) while filtered solutions for Pcdhβ1 and PcdhγA8 resulted solely in solutions with a head-to-tail orientation. We note that it is the application of the two constraints, one of which was experimentally derived, that results in this distribution of binding modes.

Experimental validation of a head-to-tail orientation - The computational evidence for a head-to-tail dimer, taken together with our identification of EC1-EC4 as the specificity-determining region, suggests that EC1 interacts with EC4 and EC2 interacts with EC3. In order to validate this model we carried out cell aggregation assays on chimeras of the γA8 and γA9 Pcdh isoforms, which were designed to determine which domains physically interact. As shown in the schematic, diagrams in Figure 6B panels 1-3, head-to-tail binding would result in a dimer where all EC2/EC3 and EC1/EC4 interactions involve domains from the same wild type protein. In all three cases the chimeras form mixed aggregates thus providing strong evidence for our proposed model of the Pcdh-Pcdh interface. Note that if the monomers bound in a head-to-head orientation, some interacting domains would be derived from different wild type proteins so that mixed aggregates would not be expected to form.

Figure 6B panels 4 and 5 provide direct evidence that EC1 interacts with EC4 and EC2 interacts with EC3. Comparing panel 4 to panel 1, the only difference between the two is that there is a mismatch between EC4 and EC1 in panel 4. The two cell populations in panel 4 form separate aggregates indicating that this single mismatch is sufficient to ablate *trans* dimerization. An identical conclusion regarding EC2 and EC3 is reached by comparison of panel 5 to panel 2. Here again, a single-domain mismatch inhibits co-aggregation even though the remaining three domains are correctly matched.

To further validate the model of head-to-tail binding, we carried out mutagenesis experiments on specificity determining regions. Since, as shown above, for the α and γ close pairs the EC2 AB loop and the EC3 FG loop determine specificities we reasoned that the specificity-determining residues in the EC2 AB loop might interact with corresponding residues in the EC3 FG loop. Notably, the largest cluster of structurally-similar docked and filtered complexes is the only cluster that positions the EC2 AB loop

near the EC3 FG loop and projected to position the EC1 near EC4 (Figures 6A and S5A). To test this model (Figure 6A), we relied on two observations. First, that arginine mutations of residue 301 in the EC3 FG loop region and residue 116 in the EC2 AB loop region (PcdhγC5 numbering) abrogate recognition in isoforms from different gene clusters (Figure 2B-D), and second, that docked models position residue 301 and residue 116 at close distance (less than 6Å, Figure 6A). Hypothesizing that residues 116 and 301 are near each other in the recognition complex, we attempted to rescue single-arginine mutants at residue 303 of PcdhαC2 or 298 of PcdhγA8 and Pcdhβ6 (analogous to PcdhγC5 301) by producing an aspartic acid mutation of PcdhαC2 residue 118, of PcdhγA8 residue 116 or of Pcdhβ6 residue 117 (analogous to PcdhγC5 116). The designed double-mutants could, in principle, form a salt bridge at the interface and thus might rescue recognition.

For all three isoforms (PcdhαC2, Pcdhβ6, and PcdhγA8), cells expressing the double arginine/aspartic-acid mutants tested positive for cell aggregation (Figure 6C), indicating that these two mutated residues (116 and 301), located respectively on domains EC2 and EC3, are in close proximity at the homophilic binding interface. This observation provides strong support for a head-to-tail binding mode where EC2 interacts with EC3 and where EC1 interacts with EC4. Moreover, since PcdhαC2, Pcdhβ6, and PcdhγA8 are not closely related, it is likely that the modeled interface represents the recognition interface for other Pcdhs as well.

Discussion

Counterintuitively, the phenomenon of neuronal self-avoidance is initiated by *trans* homophilic adhesive binding between Pcdhs. Presumably, repulsion is a consequence of the activation of downstream signals via the ICD, which is known to interact with signaling adaptors and kinases (Han et al., 2010; Schalm et al., 2010). This mechanism requires that different neurons express a sufficiently distinct set of Pcdh isoforms so that inappropriate “self”-recognition, and subsequent repulsion, will not occur. In the case of invertebrates, this is accomplished through the stochastic expression of about 10-50 different alternatively spliced Dscam isoforms in each cell (Hattori et al., 2008; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010). With thousands of stochastically generated distinct Dscam isoforms, the probability that two different neurons express the same set of isoforms is extremely low (Miura et al., 2013). Considering the much smaller

number of distinct Pcdh isoforms in vertebrates, isoform diversity alone cannot account for “non-self discrimination”.

As mentioned above, we have shown previously that an interference phenomenon plays a crucial role in Pcdh-based non-self discrimination (Thu et al., 2014). In this paper we present evidence from several independent sources of data that suggest that Pcdh cell-cell recognition is mediated by a mechanism that couples *cis* and *trans* interactions. Specifically, we propose that Pcdh isoforms form promiscuous EC6 dependent *cis*-dimers at the cell surface that associate specifically in *trans* via a stereotyped interface with elements in domains EC1-EC4. Below we summarize our findings and discuss their implications for the molecular mechanisms by which clustered Pcdhs mediate neuronal self-recognition and non-self discrimination.

Pcdh homophilic specificity is determined by a head-to-tail *trans* recognition interface

We found that Pcdh EC1-EC3 fragments do not associate in solution, nor do they mediate homophilic cell-cell recognition in cell aggregation assays. Rather, we showed both in AUC measurements and cell assays that stable *trans* dimerization requires all four of the N-terminal EC1-EC4 domains. Site-directed arginine scanning mutagenesis and rational mutagenesis based on analysis of sequence alignments allowed us to identify key structural elements in a *trans* interface that mediate cell-cell recognition between Pcdhs.

The identification of interfacial regions in EC2 and EC3 through computational modeling and mutagenesis experiments provided strong constraints that made it possible to demonstrate that Pcdh *trans* dimers adopt a head-to-tail orientation where EC2 interacts with EC3. This remarkable anti-parallel trans-interaction is in contrast to the parallel *trans* dimerization of classical cadherins. However, for classical cadherins the parallel binding mode is made possible by a significant intramolecular bend whereby the five EC domains form a highly curved structure so that interacting membrane-distal EC1 domains from apposed cells are parallel to one another. In contrast, since the EC1-EC3 domains in Pcdhs are straight rather than curved, binding in parallel would require a sharp bend between the three N-terminal and three C-terminal domains. Such a bend has been observed only in cadherins lacking inter-domain calcium binding sites (e.g. DN

cadherin (Jin et al., 2012)), and the presence of complete calcium binding sites between all domains renders such significant bending highly unlikely in the case of Pcdhs.

Figure 6A shows the structure of an EC1-EC3 *trans* dimer obtained from our docking studies that satisfy all the constraints established by mutagenesis. The EC4 domain is represented as an ellipse in the diagram since its structure has not yet been determined. In addition to satisfying all the mutagenesis data used as constraints in the docking studies, independent evidence supporting the model include; 1) the set of five cell aggregation studies on γA8 and γA9 chimeras (Figure 6B) that show that EC1 interacts with EC4 and EC2 interacts with EC3; 2) the rescue experiments shown in Figure 6C that reveal that residue 116 in EC2 is in close proximity to residue 301 in EC3, as predicted by the head-to-tail model (Figure 6A).

The head-to-tail model shown in the figure provides a clear explanation of the binding affinity and cell aggregation data. In the model, the free energy of binding is distributed over all four domain-domain interfaces, and all must be present to generate sufficient affinity to produce a stable homodimer. This is evident from the observations that three domain constructs do not dimerize, and that interfacial mutations in only a single domain are sufficient to ablate binding. All EC1-EC3 ectodomain fragments studied here were monomeric and none revealed a likely *trans* interaction. With a head-to-tail orientation, deletion of only one domain in EC1-EC4 effectively removes half the interface, providing a likely explanation for the absence of native dimer interactions.

We note that the structural model itself is unlikely to be accurate in detail and will certainly be superseded once X-ray structures of all four interacting domains are available. The major significance of the model is the demonstration that Pcdhs dimerize in *trans* in a head-to-tail orientation with an extended interface formed from four inter-domain interfaces (two EC2/EC3 and two EC1/EC4). We note that the molecular dimerization logic of Pcdhs where different domains recognize one another through EC1/EC4 and EC2/EC3 *trans* interactions, is fundamentally different from that of Dscam1 where the dimerization interface is formed from three separate self-self interactions, Ig2/Ig2, Ig3/Ig3 and Ig7/Ig7.

Pcdhs form *cis* dimers mediated by EC6

We previously provided evidence for promiscuous Pcdh EC6/EC6 *cis*

interactions. Specifically, any single carrier isoform (β, γ or C-type) can mediate cell-surface delivery of α isoforms, which are otherwise confined within the cell, through interactions involving the EC6 domain (Thu et al., 2014). In addition, the pairwise sequence identity between EC6 domains for all isoforms of Pcdhβ or Pcdhγ clusters averages over 90% (Thu et al., 2014), consistent with the idea of promiscuous interactions.

We show above that the EC6 domain mediates Pcdh *cis* dimerization even in the absence of *trans* interactions. Moreover, as shown in Table 1, the affinity of this interaction is comparable or even stronger than the *trans* interaction involving EC1-EC4. In general, *cis* interactions in the two dimensional environment of the plasma membrane would be significantly enhanced, and the effect is strongest for membrane proximal domains as there would be little entropy loss due to inter-domain flexibility upon binding (Wu et al., 2013; Wu et al., 2011). Indeed, even at low surface densities, molecules with substantial solution (3D) K_Ds, such as that of Pcdhs, will likely form dimers on cell surfaces. The promiscuity of the EC6 carrier function suggests that these dimers can form between essentially any two Pcdh isoforms, which in turn suggests that Pcdhs on cell surfaces exist as *cis* dimers formed by pairs of different isoforms from all three subfamilies as well the C-type isoforms.

Assembly termination by mismatched isoforms distinguishes self from non-self

We have shown above that full-length Pcdh ectodomains in solution form tetramers (a *cis/trans* dimer of dimers) mediated by head-to-tail *trans* interactions involving EC1-EC4, and a *cis* interaction involving EC6. A schematic of this molecular arrangement is shown in the left panel of Figure 6D. If Pcdhs on cell surfaces interacted in this manner, cellular recognition would be based on dimeric recognition units. However, as we have discussed in a previous study, dimeric recognition units are unlikely to provide sufficient diversity for neuronal non-self discrimination, and indeed all models based on multimeric recognition units encounter difficulties in accounting for both self-recognition and non-self discrimination (Thu et al., 2014). For this reason, we previously proposed an alternative recognition mechanism based on “junction-like” molecular assemblies at least partially reminiscent of those formed by classical cadherins.

As discussed above, each Pcdh molecule forms strong independent *trans* and *cis* interactions. This is in contrast to classical cadherins, in which each molecule forms relatively strong *trans* interactions and two weak asymmetrical *cis* interactions that become stronger on cell surfaces only once the *trans* interactions have been formed (Wu et al., 2011). In the case of classical cadherins, the combination of *cis* and *trans* interactions generates a two-dimensional lattice that corresponds to the extracellular structure of adherens junctions (Harrison et al., 2011). In contrast, the interactions defined here for Pcdhs suggest the formation of a one-dimensional zipper-like structure involving symmetrical *cis* and *trans* interactions. This structure is depicted in the right panel of Figure 6D, which shows how each bivalent Pcdh *cis* dimer could recognize two other dimers via independent *trans* interactions so as to form a connected ribbon of molecules that emanate from two apposed cell surfaces. We note that still undiscovered extracellular, trans-membrane or cytoplasmic interactions may ultimately reveal a more complex network of interactions than the one depicted in the figure. For example, the receptor tyrosine kinase Ret has been shown to associate with, and directly or indirectly phosphorylate Pcdhα and γ tyrosine residues in their ICD's (Schalm et al., 2010). In any case, the existence of even a one-dimensional network would provide a mechanism for interference that does not encounter the problems based on models of isolated multimeric recognition units.

Figure 6E illustrates that cells with the same isoform composition would be able to form a large assembly upon contact. In contrast, cells with different isoform compositions would incorporate mismatches, preventing further growth of the lattice (Figure 6F). If downstream signaling leading to neurite repulsion depends on the size of the assembly, which in turn depends on isoform composition, the model offers a natural mechanism for Pcdh interference. Indeed, there is a striking dependency of the size of Pcdh assemblies on the number of mismatched Pcdh isoforms. Figure 6G plots the average size of such linear assemblies as a function of the number of mismatched isoforms between two contacting neurons. Assembly size is obtained from Monte-Carlo calculations based on a model that assumes that each cell contains a stable set of *cis* dimers formed from the random association of monomers present in each cell. When all isoforms are identical assembly size is limited solely by the number of copies of each isoform. Remarkably, the presence of even a single mismatched isoform is sufficient to reduce the average size of an assembly by at least two orders of magnitude. The results presented in Figure 6G thus suggest that a mechanism based on mismatched-isoform

chain termination of a linear Pcdh-assembly could provide a binary definition of self and non-self.

While we recognize that this isoform mismatch chain-termination model is speculative, it is consistent with the presence of strong independent *cis* and *trans* interactions. Such signaling systems have been observed previously, including the one-dimensional network of CTLA-4/B7 immune receptors (Schwartz et al., 2001) where signaling has also been proposed to be based on large cell surface assemblies. Most importantly, the model provides a mechanism whereby 58 Pcdhs can generate the high level of diversity sufficient to allow for neuronal self-avoidance without encountering the problems for self-recognition, which is implicit in previous models that depend on discrete combinatorial multimeric recognition units.

Experimental procedures

Protein production and Crystallography: Proteins for crystallization or biophysical analysis were expressed in suspension-adapted HEK293 Freestyle cells (Invitrogen) and purified by nickel affinity and size exclusion chromatography. Pcdh crystals were grown by vapor diffusion in 1-2μl hanging drops, except the Pcdhβ1 EC1-3 crystals, which were grown in 0.2μl sitting drops. The PcdhγC5 EC1-3 $P4_32_12$ crystal structure was solved using the MIRAS technique while all the other Pcdh crystal structures were solved by molecular replacement. See Extended Experimental Procedure for details.

Cell aggregation assays: Pcdh expression constructs were transfected into K562 cells by electroporation. The transfected cells were grown in culture for 24 hours. Cells were then allowed to aggregate for one to three hours on a rocker inside an incubator at 37ºC. The cells were then fixed in 4% PFA for 10 minutes, washed in PBS, and cleared with 50% glycerol for imaging. See Extended Experimental Procedure for details.

Sedimentation equilibrium analytical ultracentrifugation: Proteins were diluted to an absorbance at 10mm path length and 280 nm of 0.65, 0.43 and 0.23 absorbance units. All samples were run at four speeds: 11000, 14000, 17000 and 20000 rpm (all EC 1-EC3 constructs) or 9000, 11000, 13000 and 15000 rpm (all EC1-EC4, EC1-EC5 and EC1-EC6 constructs), respectively. Measurements were carried out at 25ºC, and detection was by UV at 280 nm.

Monte-Carlo simulations – A stochastic algorithm was used to estimate the average size of Pcdh-assemblies (number of linked *cis* dimers) formed between a pair of neurons each expressing 15 distinct isoforms with 0–15 common isoforms. It was assumed that a neuron expresses an equal number of copies of each of the 15 Pcdh isoforms, with either 1000 or 100 copies per isoform (i.e., 15,000 or 1,500 total Pcdh monomers respectively). 10^6 simulations were performed and in each simulation stable *cis* dimers were randomly and independently generated for the contacting neurons. Note that the distribution of *cis* dimers on both neurons will not in general be identical even for neurons with an identical set of monomers. A linear network was initiated by randomly choosing a dimer on one of the cells. In the next step, a *cis* dimer is chosen on the second cell where one of its monomer constituents matches one of the monomers in the dimer chosen on the first cell. This matching process is then repeated with the search for matching dimers alternating between the contacting neurons moving from one cell to the other as the chain extends in two directions. This extension process was repeated until there remained no matching dimers either due to a mismatch or to a depletion of dimers.

Author Contributions

R.R., C.A.T., K.M.G., T.M., L.S., and B.H. designed research, analyzed data, and assembled and wrote the paper. R.R. carried out the computational analysis, C.A.T. carried out cell aggregation assays and analysis, F.B., S.M., H.N.W. and K.M.G. prepared and crystallized all proteins. H.N.W. and K.M.G. determined the crystal structures. C.A.T., S.M., H.N.W., and M.C. prepared Pcdh mutants. G.A. performed and analyzed the AUC experiments. A.H. and H.C. performed the glycosylation analysis.

Acknowledgments

We thank Dr. David Hirsh for valuable comments on the manuscript. We thank Igor Kourinov, Surajit Banerjee, Narayanasami Sukumar at Advanced Photon Source (APS) for support with synchrotron data collection. This work was supported by the National Science Foundation grant to B.H. (MCB- 1412472), National Institutes of Health grant to L.S. (R01GM062270), joint NIH grant to T.M. and L.S. (R01GM107571), NIH Training Programs to H.W. (T32GM008281), and Danish National Research Foundation (DNRF107) to A.H. and H.C..

References

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

Chen, W.V., Alvarez, F.J., Lefebvre, J.L., Friedman, B., Nwakeze, C., Geiman, E., Smith, C., Thu, C.A., Tapia, J.C., Tasic, B.*, et al.* (2012). Functional significance of isoform diversification in the protocadherin gamma gene cluster. Neuron *75*, 402-409.

Chen, W.V., and Maniatis, T. (2013). Clustered protocadherins. Development *140*, 3297-3302.

Han, M.H., Lin, C., Meng, S., and Wang, X. (2010). Proteomics analysis reveals overlapping functions of clustered protocadherins. Mol Cell Proteomics *9*, 71-83.

Harrison, O.J., Jin, X., Hong, S., Bahna, F., Ahlsen, G., Brasch, J., Wu, Y., Vendome, J., Felsovalyi, K., Hampton, C.M.*, et al.* (2011). The extracellular architecture of adherens junctions revealed by crystal structures of type I cadherins. Structure *19*, 244-256.

Hattori, D., Chen, Y., Matthews, B.J., Salwinski, L., Sabatti, C., Grueber, W.B., and Zipursky, S.L. (2009). Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms. Nature *461*, 644-648.

Hattori, D., Millard, S.S., Wojtowicz, W.M., and Zipursky, S.L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. Annual review of cell and developmental biology *24*, 597-620.

Hughes, M.E., Bortnick, R., Tsubouchi, A., Baumer, P., Kondo, M., Uemura, T., and Schmucker, D. (2007). Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron *54*, 417-427.

Jin, X., Walker, M.A., Felsovalyi, K., Vendome, J., Bahna, F., Mannepalli, S., Cosmanescu, F., Ahlsen, G., Honig, B., and Shapiro, L. (2012). Crystal structures of Drosophila N-cadherin ectodomain regions reveal a widely used class of Ca(2)+-free interdomain linkers. Proc Natl Acad Sci U S A *109*, E127-134.

Lefebvre, J.L., Kostadinov, D., Chen, W.V., Maniatis, T., and Sanes, J.R. (2012). Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature *488*, 517-521.

Lommel, M., Winterhalter, P.R., Willer, T., Dahlhoff, M., Schneider, M.R., Bartels, M.F., Renner-Muller, I., Ruppert, T., Wolf, E., and Strahl, S. (2013). Protein O-mannosylation is crucial for E-cadherin-mediated cell adhesion. Proceedings of the National Academy of Sciences of the United States of America *110*, 21024-21029.

Matthews, B.J., Kim, M.E., Flanagan, J.J., Hattori, D., Clemens, J.C., Zipursky, S.L., and Grueber, W.B. (2007). Dendrite self-avoidance is controlled by Dscam. Cell *129*, 593-604.

Miura, S.K., Martins, A., Zhang, K.X., Graveley, B.R., and Zipursky, S.L. (2013). Probabilistic splicing of Dscam1 establishes identity at the level of single neurons. Cell *155*, 1166-1177.

Morishita, H., Umitsu, M., Murata, Y., Shibata, N., Udaka, K., Higuchi, Y., Akutsu, H., Yamaguchi, T., Yagi, T., and Ikegami, T. (2006). Structure of the cadherin-related neuronal receptor/protocadherin-alpha first extracellular cadherin domain reveals diversity across cadherin families. The Journal of biological chemistry *281*, 33650-33663.

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics *21*, 1472-1478.

Posy, S., Shapiro, L., and Honig, B. (2008). Sequence and structural determinants of strand swapping in cadherin domains: do all cadherins bind through the same adhesive interface? Journal of molecular biology *378*, 954-968.

Reiss, K., Maretzky, T., Haas, I.G., Schulte, M., Ludwig, A., Frank, M., and Saftig, P. (2006). Regulated ADAM10-dependent ectodomain shedding of gamma-protocadherin C3 modulates cell-cell adhesion. The Journal of biological chemistry *281*, 21735-21744.

Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in the protocadherin-alpha gene cluster. Proc Natl Acad Sci U S A *103*, 19719-19724.

Schalm, S.S., Ballif, B.A., Buchanan, S.M., Phillips, G.R., and Maniatis, T. (2010). Phosphorylation of protocadherin proteins by the receptor tyrosine kinase Ret. Proceedings of the National Academy of Sciences of the United States of America *107*, 13894-13899.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671-684.

Schreiner, D., and Weiner, J.A. (2010). Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proceedings of the National Academy of Sciences of the United States of America *107*, 14893-14898.

Schwartz, J.C., Zhang, X., Fedorov, A.A., Nathenson, S.G., and Almo, S.C. (2001). Structural basis for co-stimulation by the human CTLA-4/B7-2 complex. Nature *410*, 604-608.

Soba, P., Zhu, S., Emoto, K., Younger, S., Yang, S.J., Yu, H.H., Lee, T., Jan, L.Y., and Jan, Y.N. (2007). Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. Neuron *54*, 403-416.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol Cell *10*, 21-33.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between alpha, beta, and gamma protocadherins. Cell *158*, 1045-1059.

Vester-Christensen, M.B., Halim, A., Joshi, H.J., Steentoft, C., Bennett, E.P., Levery, S.B., Vakhrushev, S.Y., and Clausen, H. (2013). Mining the O-mannose glycoproteome reveals cadherins as major O-mannosylated glycoproteins. Proceedings of the National Academy of Sciences of the United States of America *110*, 21018-21023.

Wojtowicz, W.M., Wu, W., Andre, I., Qian, B., Baker, D., and Zipursky, S.L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell *130*, 1134-1145.

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.

Wu, Y., Honig, B., and Ben-Shaul, A. (2013). Theory and simulations of adhesion receptor dimerization on membrane surfaces. Biophysical journal *104*, 1221-1229.

Wu, Y., Vendome, J., Shapiro, L., Ben-Shaul, A., and Honig, B. (2011). Transforming binding affinities from three dimensions to two with application to cadherin clustering. Nature *475*, 510-513.

Zipursky, S.L., and Grueber, W.B. (2013). The molecular basis of self-avoidance. Annual review of neuroscience *36*, 547-568.

Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell *143*, 343-353.

Figure legends

Figure 1. **Crystal structures of four Pcdh EC1-EC3 isoforms.**

A) The Pcdh genomic locus contains three adjacent clusters of variable exons. Each exon encodes an entire ectodomain comprising six EC domains, a transmembrane (TM) domain, and a short cytoplasmic region. Alpha and gamma clusters also contain three constant exons that encode a cluster-specific intracellular domain (ICD) which are joined by pre-mRNA splicing for alpha and gamma clusters. C-type Pcdh exons are shown in pink and light blue for the alpha and gamma clusters, respectively.

B) Crystal structures of EC1-EC3 regions from PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 shown in ribbon representation. Ca^{2+} ions are drawn as green spheres. N-glycans and conserved O-mannose residues are drawn as sticks. The inter-domain calcium binding sites are arranged similarly to those observed in classical cadherins (expanded view). See also Figure S1 and Table S1.

C) Comparison of the PcdhγC5 and type I classical C-cadherin structures. The overall architecture of classical cadherin ectodomains have a curved shape with an approximate 90° angle between EC1 and EC5 (Boggon et al., 2002). In contrast, the architecture of Pcdh EC1-EC3 domain regions is characterized by an extended zigzagged conformation.

D) EC2-EC3 angles distinct from classical cadherins account for the extended zigzagged conformation of the Pcdh structures. EC1-EC3 domains are drawn as blue (PcdhγC5) and yellow (C-cadherin) ovals. Angles shown are between principal axes of inertia for adjacent domains.

Figure 2. Elements of Pcdh *cis* and *trans* binding.

A) Correlating multimerization states of truncated Pcdh proteins with their cell-cell recognition properties. Cells transfected with Pcdh deletion series plasmid constructs were tested for aggregation. With the exception of EC2-EC6 Pcdh fragments and PcdhγC5 EC1-EC4, all deletion proteins that formed oligomers in solution also mediated

cell aggregation. Full-length Pcdhα4 include the EC6 domain from PcdhγC3 so it could be delivered to cell surface.

B) Probing homophilic interaction interface by arginine-scanning mutagenesis. Residues mutated to arginine are drawn in space filling representation. In blue are mutations that did not disrupt recognition, in orange are mutations that weakened recognition and in red are mutations that abolished cell-cell recognition. Excluding residue 142, all the effective arginine mutants are located along one side of the molecule.

C) Cell aggregation experiments showing the mutations in part (B) that weakened or abolished interactions. See also Figure S2C.

D) In other Pcdh isoforms, residues analogous to the effective PcdhγC5 arginine mutants had similar effects on the cell-cell recognition in the majority of cases.

Figure 3. Pcdh *trans* binding depends on the four N-terminal domains EC1-EC4.

A-C) Domain-shuffled chimeras of closely related isoforms and their wild-type counterparts were assayed for binding specificity. Swapped specificity was noted for chimeras in which either the EC1-EC3 or EC2-EC4 domains were replaced with the corresponding domains of closely related isoforms. See also Figure S3.

D) Schematic representation of the domain-shuffled isoforms and their observed binding specificities to their wild-type isoform counterparts.

Figure 4. Candidate specificity determining residues.

A) Multiple sequence alignment of the three closely related Pcdh isoform pairs, along with PcdhγC5. Highlighted in gray are positions conserved in all Pcdh sequences. Sequence positions that differ between the closely related isoforms are shown in red; a subset of these residues determines binding specificity. Residues swapped between isoforms and assayed for binding properties are boxed. Secondary structure from PcdhγC5 is shown at the top of the alignment.

B) Multiple sequence alignment of the FG-loop region for PcdhγA8 and PcdhγA9 orthologs. Three of the residues that differ between mouse PcdhγA8 and PcdhγA9 are highly conserved in orthologs (highlighted in red), suggesting their functional importance.

Figure 5. **Structural elements of the canonical Pcdh *trans* binding interface.**

A-C) Assessing specificity-determining residues. Binding properties of wild-type isoforms (left side of each panel) or constructs with shuffled residues (top of each panel) were tested separately for each EC domain. Cases in which shuffled residues swapped specificities are indicated by an orange outline. Residues shuffled between closely related isoforms are shown in magenta on surface representations of the Pcdhα7, Pcdhβ6, and PcdhγA8 structures. Sequence alignments of shuffled regions are shown. See also Figure S4.

D) Correspondence between *trans* interface residues identified by arginine scanning and close-isoform pair analysis. Single arginine mutant residues that abolish or diminish homophilic binding, highlighted in red and orange respectively, are found in the same structural regions as the shuffled residues (see also Figure 2). Residues that swap binding specificity between closely related isoforms are shown in magenta on surface representations of the Pcdh-γC5 crystal structure.

Figure 6. Molecular logic of Pcdh-mediated cell-cell recognition.

A) Shown in ribbon representation is the only orientation observed for docking of the four EC1-EC3 domains structures which position the EC2 AB loop in close proximity to the EC3 FG loop. EC2 AB loop residue 116 and FG loop residue 301 are drawn as space filling and colored red and blue respectively. The vast majority of the docked complexes were observed to interact in this mode. See also Figure S5A.
B) Cell aggregation assays on chimeric proteins that show EC1 interacts with EC4 and EC2 interacts with EC3. Schematic representation of the head-to-tail interaction between the domain-shuffled chimeras is shown above each panel. Mixed aggregates were formed where all interactions involve “matching” domains (panels 1-3). Separate aggregates were formed when there is a mismatch between EC1/EC4 (panel 4) or between EC2/EC3 (panel 5).
C) The EC2 domain AB region recognizes the EC3 domain FG loop. Cells expressing isoforms with single arginine mutants in the EC3 FG loop region, or with double mutations (aspartate at the AB region and arginine at the FG loop), were assayed for

aggregation. The double-mutation rescued the non-adhesive phenotype, supporting the head-to-tail binding orientation shown in part (A).

D) Two possible models of Pcdh interaction. A discrete tetramer composed of a dimer of dimers is observed in analytical ultracentrifugation, but we suggest that a connected ribbon of molecules can form between cells via the *trans* and *cis* interactions.

E & F) A model for Pcdh mediated cell-cell recognition based on formation of a superstructure defined by promiscuous *cis* and specific *trans* interactions. Growth of the chain of molecules requires matching of all isoforms; a single mismatch can terminate chain extension. Dendrites of the same neuron will have the same isoform repertoire while dendrites of different neurons will differ. In this model, repulsion signaling is triggered, or achieves a sufficient level for response, only through the formation of an extended chain of Pcdhs.

G) For the case of 15 distinct Pcdh isoforms expressed per cell, Monte-Carlo simulations were used to estimate the average size of one-dimensional Pcdh assemblies between contacting cells. The average number of *cis* dimers that comprise such assemblies is shown on a logarithmic scale as a function of the number of mismatched isoforms. Two cases are shown: one for 15000 total Pcdh monomers (1000 per isoform, red), and one for 1500 total copies (100 per isoform). The model assumes that each cell contains a stable set of *cis* dimers formed from the random association of monomers present in each cell. See also Figure S5B.

Figure 1

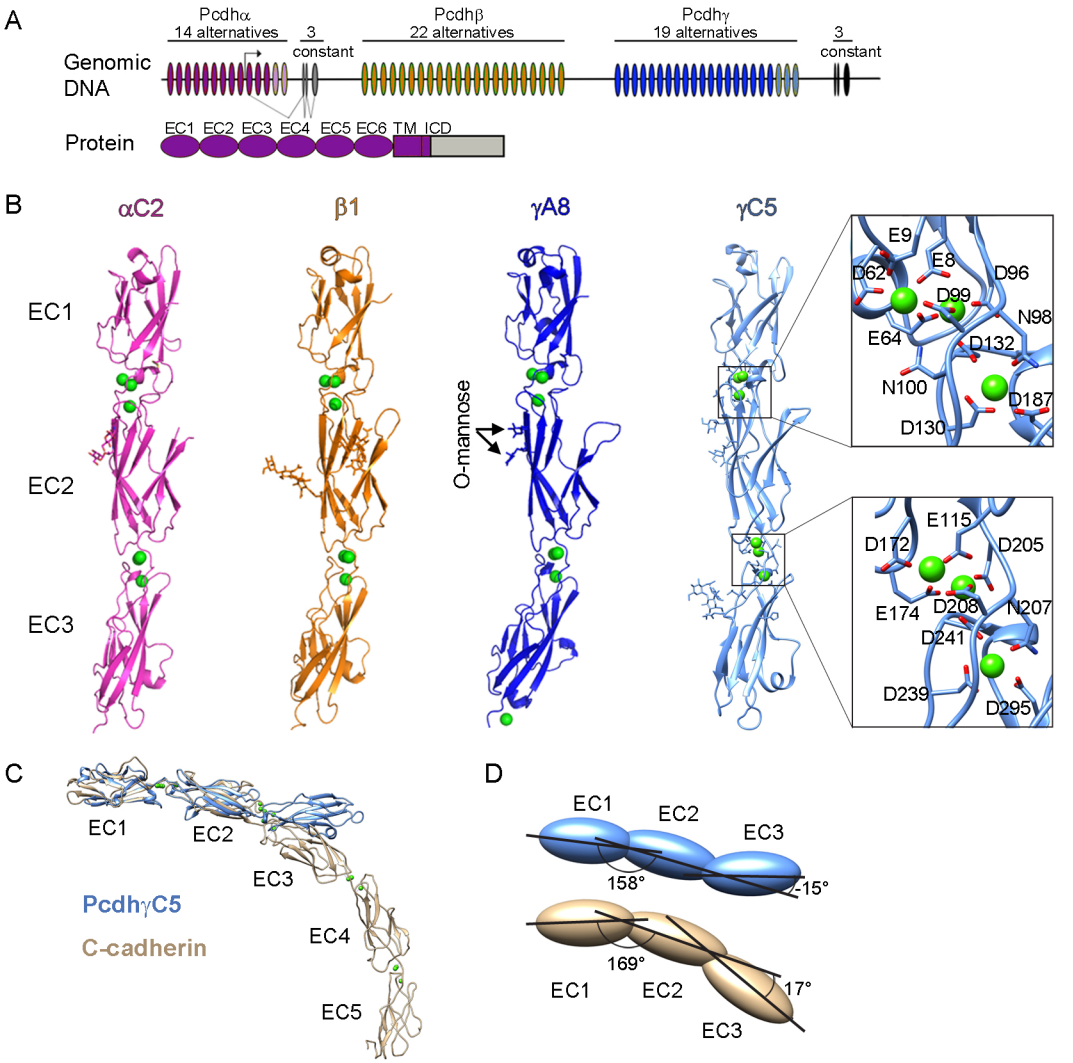

A
Pcdhα
14 alternatives
3
constant
Pcdhβ
22 alternatives
Pcdhγ
19 alternatives
3
constant
Genomic
DNA
Protein
EC1 EC2 EC3 EC4 EC5 EC6 TM ICD
B
αC2
β1
γA8
γC5
EC1
EC2
EC3
O-mannose
E9
D62
E8
D96
D99
N98
E64
D132
N100
D187
D130
E115
D172
D205
E174
D208
N207
D241
D239
D295
C
EC1
EC2
EC3
EC4
EC5
PcdhγC5
C-cadherin
D
EC1
EC2
EC3
158°
-15°
169°
EC1
EC2
17°
EC3

Figure 2

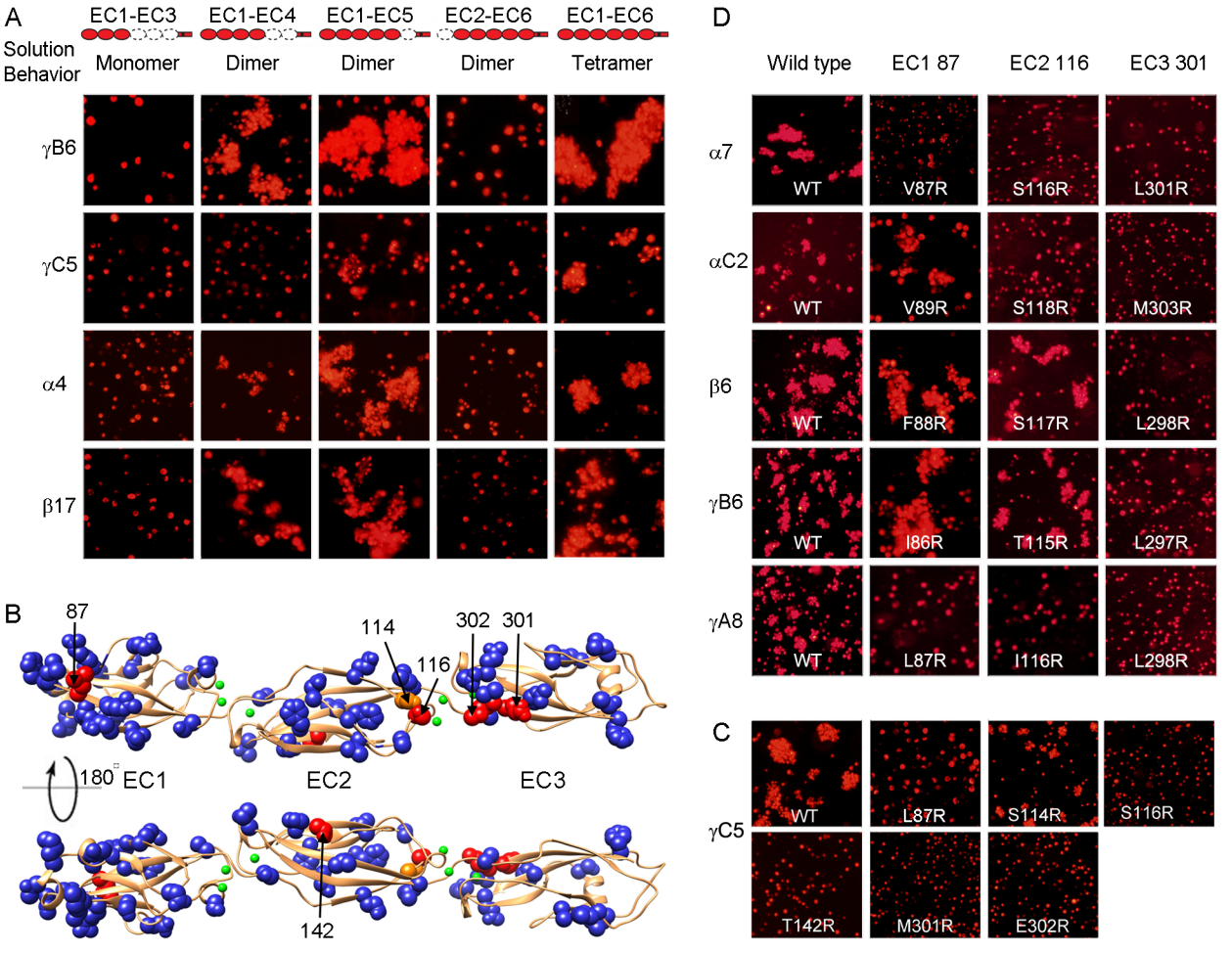

A
EC1-EC3
EC1-EC4
EC1-EC5
EC2-EC6
EC1-EC6
Solution Behavior
Monomer
Dimer
Dimer
Dimer
Tetramer
γB6
γC5
α4
β17
B
87
114
116
302
301
180°
EC1
EC2
EC3
142
D
Wild type
EC1 87
EC2 116
EC3 301
α7
WT
V87R
S116R
L301R
αC2
WT
V89R
S118R
M303R
β6
WT
F88R
S117R
L298R
γB6
WT
I86R
T115R
L297R
γA8
WT
L87R
I116R
L298R
C
γC5
WT
L87R
S114R
S116R
T142R
M301R
E302R

Figure 3

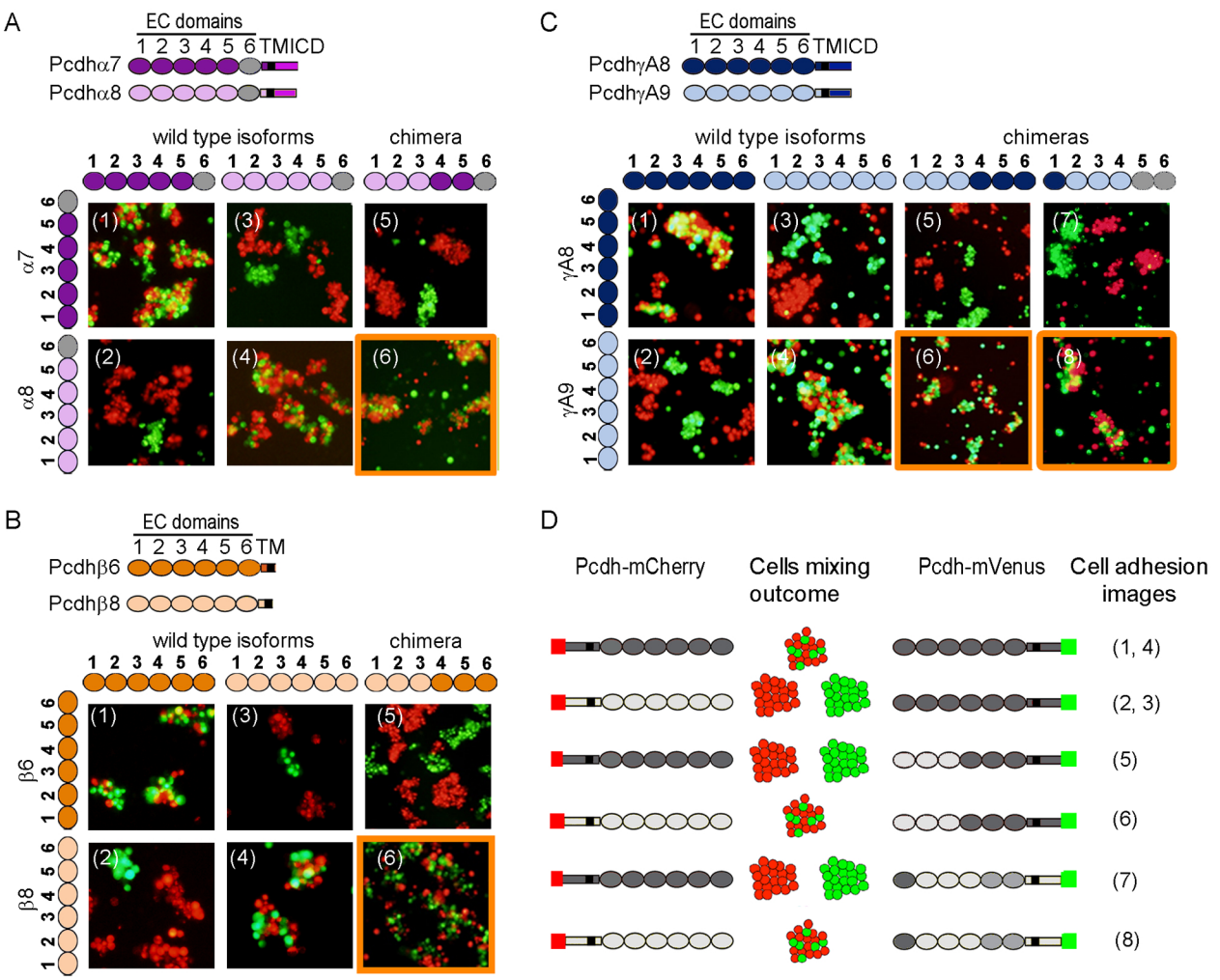

A
EC domains
1 2 3 4 5 6 TMICD
Pcdhα7
Pcdhα8
wild type isoforms
chimera
1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6
α7
α8
(1)
(2)
(3)
(4)
(5)
(6)
B
EC domains
1 2 3 4 5 6 TM
Pcdhβ6
Pcdhβ8
wild type isoforms
chimera
1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6
β6
β8
(1)
(2)
(3)
(4)
(5)
(6)
C
EC domains
1 2 3 4 5 6 TMICD
PcdhγA8
PcdhγA9
wild type isoforms
chimeras
1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6
γA8
γA9
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
D
Pcdh-mCherry
Cells mixing outcome
Pcdh-mVenus
Cell adhesion images
(1, 4)
(2, 3)
(5)
(6)
(7)
(8)

Figure 4

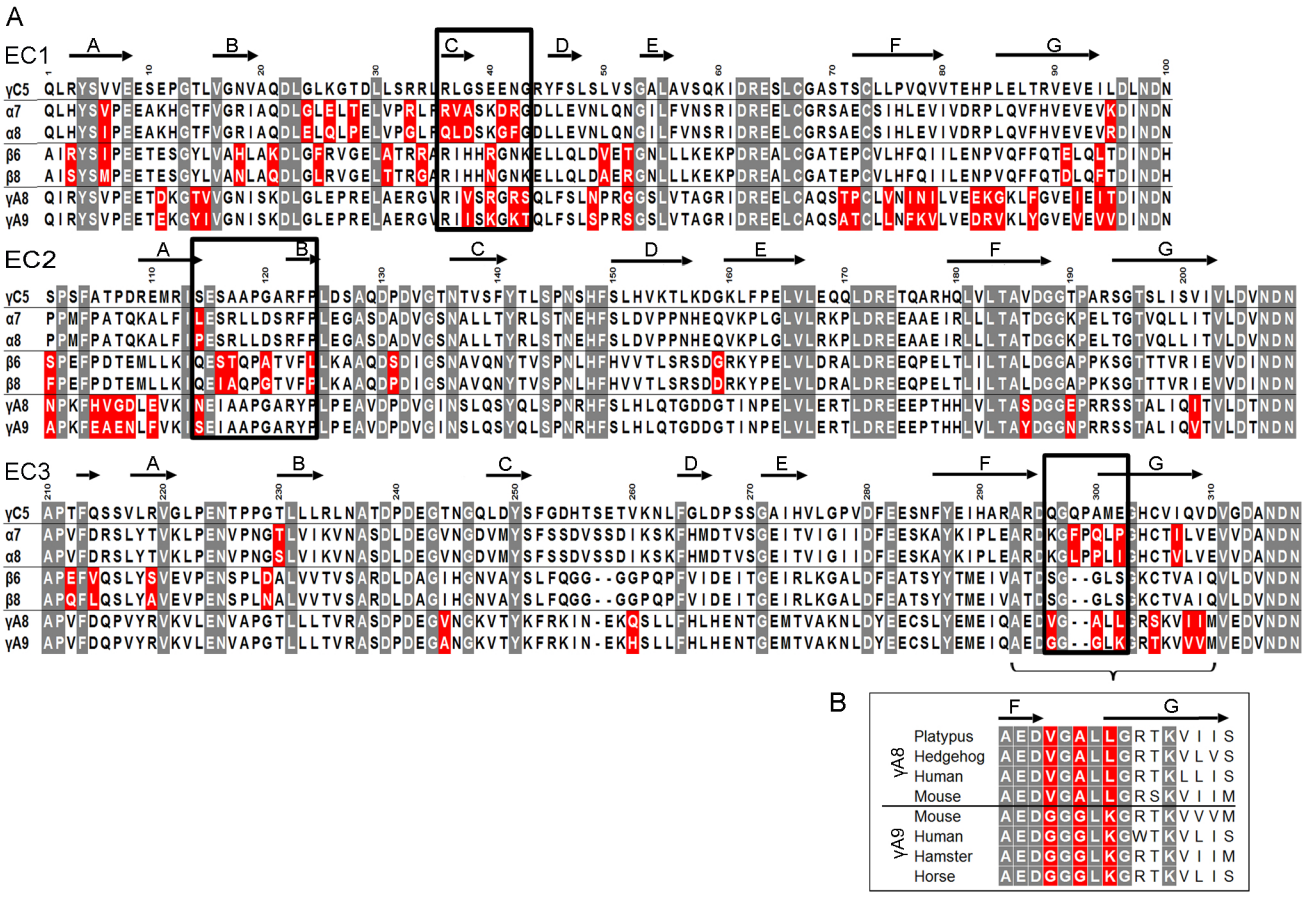

A
EC1
EC2
EC3
γC5
α7
α8
β6
β8
γA8
γA9
B
Platypus
Hedgehog
Human
Mouse
Hamster
Horse

Figure 5

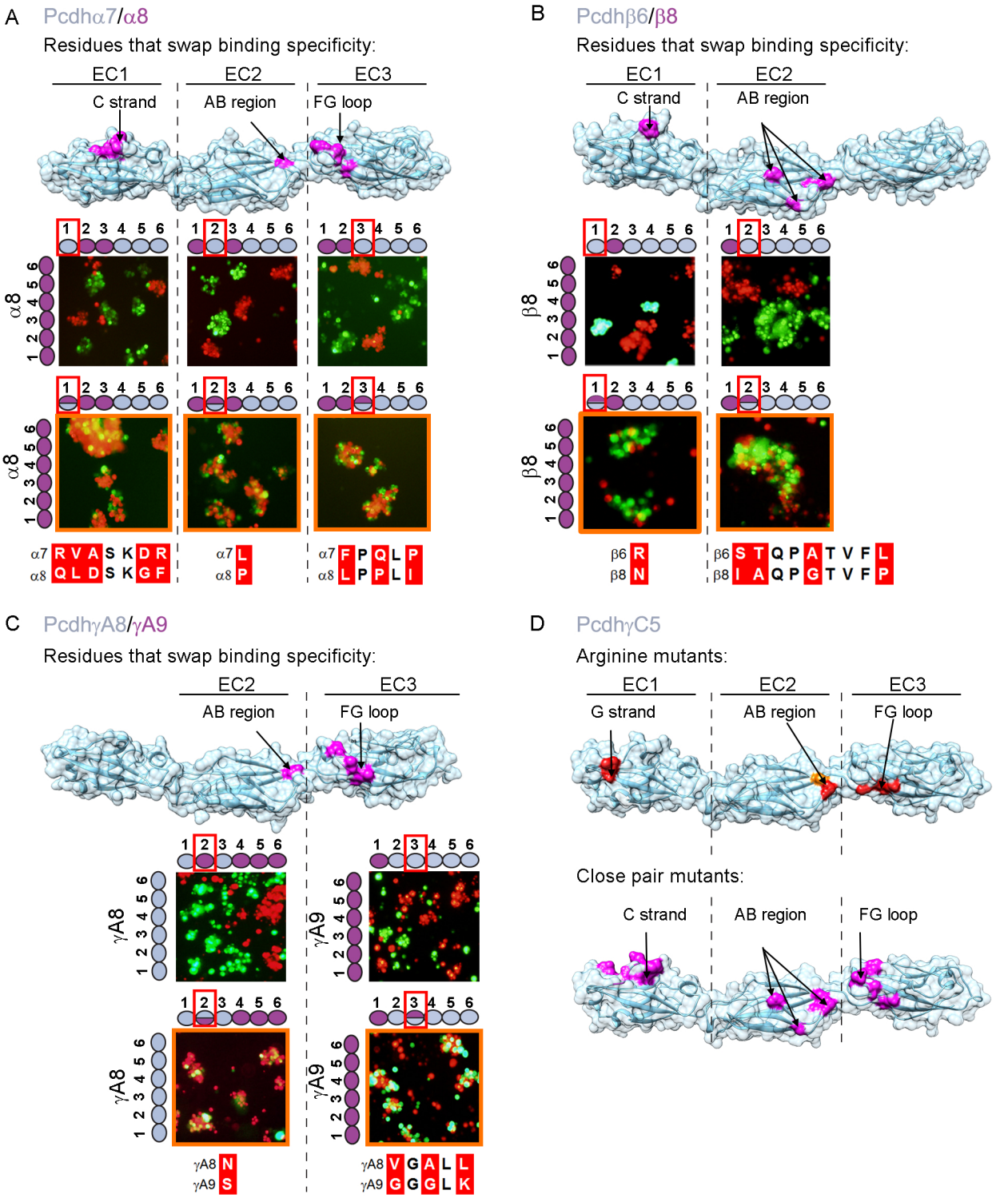
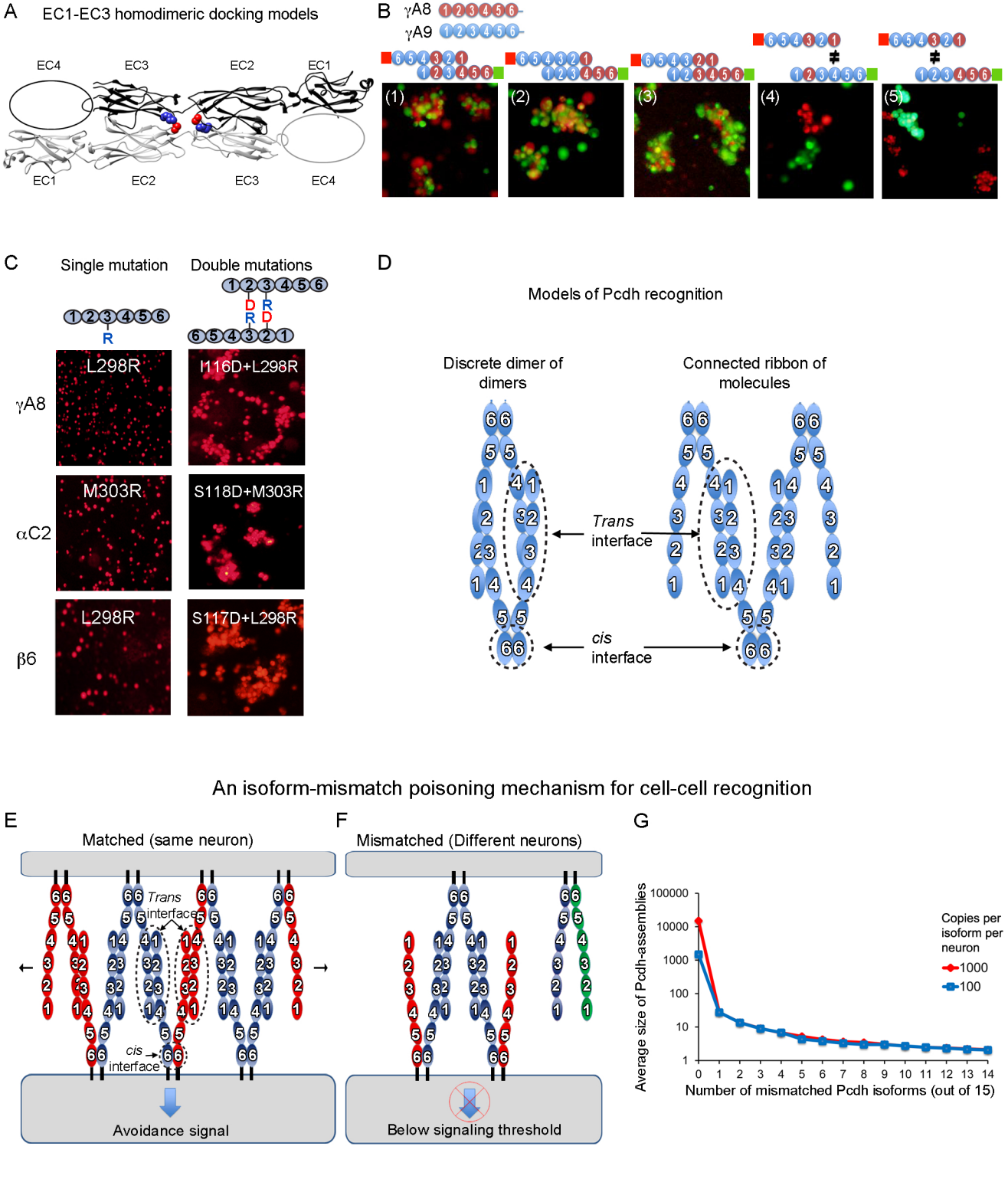

A
Pcdhα7/α8
Residues that swap binding specificity:
EC1
EC2
EC3
C strand
AB region
FG loop
α8
α7 R V A S K D R
α8 Q L D S K G F
α7 L
α8 P
α7 F P Q L P
α8 L P P L I
B
Pcdhβ6/β8
Residues that swap binding specificity:
EC1
EC2
C strand
AB region
β8
β6 R
β8 N
β6 S T Q P A T V F L
β8 I A Q P G T V F P
C
PcdhγA8/γA9
Residues that swap binding specificity:
EC2
EC3
AB region
FG loop
γA8
γA9
γA8 N
γA9 S
γA8 V G A L L
γA9 G G G L K
D
PcdhγC5
Arginine mutants:
EC1
EC2
EC3
G strand
AB region
FG loop
Close pair mutants:
C strand
AB region
FG loop

Figure 6


A
EC1-EC3 homodimeric docking models
EC4
EC3
EC2
EC1
EC1
EC2
EC3
EC4
B
γA8
γA9
(1)
(2)
(3)
(4)
(5)
C
Single mutation
Double mutations
R
D
R
D
L298R
I116D+L298R
γA8
M303R
S118D+M303R
αC2
L298R
S117D+L298R
β6
D
Models of Pcdh recognition
Discrete dimer of dimers
Connected ribbon of molecules
Trans interface
cis interface
An isoform-mismatch poisoning mechanism for cell-cell recognition
E
Matched (same neuron)
Trans interface
cis interface
Avoidance signal
F
Mismatched (Different neurons)
Below signaling threshold
G
Average size of Pcdh-assemblies
100000
10000
1000
100
10
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of mismatched Pcdh isoforms (out of 15)
Copies per isoform per neuron
1000
100

Supplemental Experimental Procedures

Protein production

Pcdh cDNA for Pcdhα7 EC1-3 (Q1–D316) and EC1-5 (Q1–D530); PcdhαC2 EC1-3 (Q1–D318), EC1-4 (Q1–D422), EC1-5 (Q1–D532), EC1-6 (Q1–K644) and EC2-6 (S103–K644); Pcdhβ1 EC1-3 (A1–D314); PcdhγA8 EC1-3 (Q1–D313), EC1-4 (Q1–D418) and EC2-6 (N101–E641); PcdhγB6 EC1-4 (G1–D418), EC1-5 (G1–D528) and EC1-6 (G1–E641); PcdhγC5 EC1-3 (Q1–D316), EC1-3 with extended N-terminus (including GWSSG before Q1), EC1-4 (Q1–D420), EC1-5 (Q1–D530), EC1-6 (Q1–E643) and EC2-6 (S101–E643), excluding the predicted signal sequence, were cloned into a modified pαSHP-H mammalian expression vector (a kind gift from Daniel J. Leahy, John Hopkins University) with a BiP signal sequence and a C-terminal octahistidine tag. The cDNA for PcdhγC5 EC1-3 was codon optimized to improve protein yields. The γA8 I116R, αC2 S118R and γC5 S116R mutations were introduced by the Quikchange method (Stratagene). These constructs were transfected using polyethyleneimine (Polysciences Inc.) into suspension-adapted HEK293 Freestyle cells (Invitrogen) in serum free media (Invitrogen). The media was harvested 6 days after transfection and the secreted proteins were purified by nickel affinity chromatography followed by size exclusion chromatography. Purified proteins were concentrated to 2-23mg/ml for AUC or crystallography experiments.

Crystallography

Pcdh crystals were grown by vapor diffusion in 1-2μl hanging drops, except the Pcdhβ1 EC1-3 crystals, which were grown in 0.2μl sitting drops. The crystallization conditions were: 28% PEG MME 500, 100mM sodium acetate, pH 4 for PcdhαC2 EC1-3; 24% (w/v) PEG1500, 20% (v/v) glycerol, 3% (w/v) glucose for Pcdhβ1 EC1-3; 28% PEG400, 100mM Tris-Cl, pH 8.7 for

PcdhγA8 EC1-3; 40% (v/v) MPD, 5% (w/v) PEG 8000, 100mM sodium cacodylate, pH 6 for the PcdhγC5 EC1-3 $P4_32_12$ crystal form; 16% (w/v) PEG 6000, 200mM calcium acetate, 100mM imidazole, pH 8.0 (30% (v/v) PEG 400 added cryoprotectant) for the PcdhγC5 EC1-3 C2 crystal form; 25.5% (w/v) PEG 4000, 15% (v/v) glycerol, 3mM calcium chloride, 85mM Tris-Cl, pH 8.5 (20% (v/v) ethylene glycol added cryoprotectant) for the PcdhγC5 EC1-3 $P2_1$ crystal form; 8% (w/v) PEG3350, 200mM potassium nitrate, 3% (v/v) glycerol (30% PEG400 added cryoprotectant) for PcdhγC5 EC1-3 with extended N-terminus. Heavy atom derivatives of the PcdhγC5 EC1-3 $P4_32_12$ crystals were obtained by soaking the crystals in the crystallization condition supplemented with 1mM ethyl mercuric phosphate (EMP) or K_2HgI_4 for 2-16h.

Complete native and derivative datasets were collected from single crystals at 100K on either the Northeastern Collaborative Access Team beamline 24-ID-E at the Advanced Photon Source, Argonne National Laboratory or beamline X4C at National Synchrotron Light Source, Brookhaven National Laboratory. The PcdhγC5 EC1-3 crystal data was indexed using DENZO and scaled and merged with SCALEPACK (Otwinowski and Minor, 1997, 2001). All other data was indexed with MOSFLM (Battye et al., 2011) and scaled and merged with Scala (PcdhαC2 EC1-3 and PcdhγC5 EC1-3 extended N-terminus) or Aimless (PcdhγA8 EC1-3 and Pcdhβ1 EC1-3) (Evans, 2007).

The PcdhγC5 EC1-3 $P4_32_12$ crystal structure was solved using the MIRAS technique. Initial heavy atom sites were located using SOLVE/RESOLVE (Terwilliger and Berendzen, 1999). Optimization of heavy atom sites and solvent flattening was then carried out in SHARP (delaFortelle and Bricogne, 1997) to generate the initial electron density map. Initial model building into the map was carried out in Coot (Emsley et al., 2010) and iterative refinement and model building were carried out using Phenix (Adams et al., 2010) and Coot.

All other Pcdh crystal structures were solved by molecular replacement with Phaser (McCoy et al., 2007) using the PcdhγC5 EC1-3 $P4_32_12$ crystal structure as a search model, except Pcdhβ1 EC1-3, for which PcdhαC2 EC1-3 was used as the search model. Iterative refinement and model building were then conducted using Phenix and Coot.

Cell aggregation assay

Plasmids. DNA fragments encoding fluorescent fusion full length Pcdh isoforms were generated as previously described (Thu et al., 2014). The domain deletion and domain swapping between different Pcdh isoforms were made by performing overlapped PCR. The arginine mutations, the double mutations, and the mutations between close pairs were generated by the Quikchange method (Stratagene).The PCR products were then sub-cloned into gateway entry vectors and corresponding expression vectors. EC domains were assigned as previously described (Thu et al., 2014). Transmembrane domains (TM) were predicted by using TMHIM web (http://www.cbs.dtu.dk/services/TMHMM-2.0/). Signal peptides (SP) were predicted by SignalP prediction tools from SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP/). Primer sequences used for PCR amplifications in domain deletion/swapping studies and site directed mutagenesis will be provided upon request.

Cell aggregation assays. Aggregation assays were performed as previously described (Thu et al., 2014). Expression constructs generated by gateway cloning system were transfected into K562 cells (human leukemia cell line, ATCC CCL243) by electroporation method using an Amaxa 4D-Nucleofactor (Lonza). The transfected cells were grown in culture for 24 hours. Cells were then allowed to aggregate for one to three hours on a rocker inside an incubator at 37ºC.

The cells were then fixed in 4% PFA for 10 minutes, washed in PBS, and cleared with 50% glycerol for imaging. The images were taken using an Olympus fluorescent microscope.

Co-aggregation assays. Differentially tagged wild-type or modified Pcdh expression constructs were transfected into K562 cells as described above. K562 cells expressing mCherry- or mVenus-tagged Pcdhs were mixed after 24 hours by shaking for one to three hours. Images of red and green cell aggregates were taken using an Olympus fluorescent microscope.

Effect of the Ca^{2+} chelators EDTA or EGTA on cell aggregation. To assess the requirement for calcium ions in cell-cell adhesion mediated by over-expressed N-cadherin and four Pcdh isoforms, K562 cells transfected with constructs encoding these proteins were treated with EDTA (10mM) or EGTA (5mM) prior to performing the cell aggregation assay. The assay was performed as described above and the fluorescent images were taken with an Olympus inverted microscope.

Sedimentation equilibrium Analytical Ultracentrifugation

Experiments were performed in a Beckman XL-A/I analytical ultracentrifuge (Beckman-Coulter, Palo Alto CA, USA), utilizing six-cell centerpieces with straight walls, 12 mm path length, and sapphire windows. Samples were dialyzed over-night and then diluted in 10 mM Tris, 150 mM NaCl, 3 mM $CaCl_2$, pH 8.0, with varying concentrations of imidazole, as follows: 100 mM (Pcdhα7 EC1-3 and EC1-5; Pcdhβ1 EC1-3; PcdhγC5 EC1-3, EC1-3 extended N-term, and EC1-5; PcdhγA8 EC1-3 and EC2-6; PcdhαC2 EC1-5; PcdhγC5 EC1-4 and EC1-5) 200 mM (PcdhαC2 EC1-3 EC1-4, EC1-6, EC2-6, and EC1-6 S118R; PcdhγA8 EC1-4 and EC1-4 I116R; PcdhγC5 EC2-6 and EC2-6 S116R) or 250 mM (PcdhγC5 EC1-6 and EC1-6 S118R; PcdhγB6 EC1-6). Proteins were diluted to an absorbance at 10 mm and 280 nm of 0.65, 0.43 and 0.23 in

channels A, B and C, respectively. Dilution buffer was used as blank. All samples were run at four speeds, the lowest speed held for 20h then four scans with 1h interval, the second lowest held for 10h then four scans with 1h interval, the third lowest and the highest speed as the second lowest. The speeds were 11000, 14000, 17000 and 20000 rpm (all EC 1-3 constructs) or 9000, 11000, 13000 and 15000 rpm (all EC 1-4, EC1-5 and EC 1-6 constructs), respectively. Measurements were done at 25^{o}C, and detection was by UV at 280 nm. Solvent density and protein v-bar were determined using the program SednTerp. (Alliance Protein Laboratories, Corte Cancion, Thousand Oaks, CA, USA) For calculation of dimeric K_D and apparent molecular weight, all useful data were used in a global fit, using the program HeteroAnalysis, obtained from University of Connecticut. (www.biotech.uconn.edu/auf) Calculation of tetrameric K_D was carried out with the program Sedphat (http://www.analyticalultracentrifugation.com/sedphat/index.htm).

O-mannosylation

Mass spectrometric analyses were performed essentially as previously described (Halim et al., 2015). Briefly, 5 µg of each protein was reduced (5 mM dithiothreitol, 60 °C, 30 min) and alkylated (10 mM iodoacetamide, RT, 30 min) before a 16h, 37°C incubation with 1 µg trypsin (Roche). Tryptic digests were analyzed on a setup composed of an EASY-nLC 1000 UHPLC (Thermo Scientific) interfaced via a nanoSpray Flex ion source to an LTQ-Orbitrap Velos Pro hybrid mass spectrometer. The analytical column (PicoFrit Emitters, New Objectives, 75 µm inner diameter) was packed in-house with Reprosil-Pure-AQ C18 phase (Dr. Maisch GmbH, 1.9 µm particle size). Tryptic digests were separated using a 60 min LC gradient operated at 200 nL/min. MS1 precursor scan (m/z 350–1500) acquisition was performed in the orbitrap using a nominal resolution of 30,000, followed by HCD-MS2 and ETD-MS2 fragmentation of the five

most abundant multiply charged precursor ions. Data were processed using the Sequest HT node of the Proteome Discoverer 1.4 software (Thermo Fisher Scientific). Spectra matched to glycosylated peptides were inspected manually to verify the accuracy of the assignments.

Computational docking analysis

Using the crystal structures of the EC1-EC3 regions from PcdhαC2, Pcdhβ1, PcdhγA8, and PcdhγC5 determined here, we produced docking models of Pcdh *trans* homodimers. As the results from AUC showed that the EC1-EC4 domain region self-associates to form dimers, we used the M-zdock program to generate Pcdh homodimers. For each of the crystal structures we generated 1,500 docking models resulting in a total of 6,000 models. Docked models were filtered by requiring them to include the experimentally identified specificity determinant residues at the binding interface. For docked models of PcdhαC2, PcdhgA8, and PcdhγC5 isoforms, the specificity determinant residues used were: 114, 116, 301, and 302 (PcdhγC5 numbering), whereas the specificity determinant residues 117, 118 and 121 were used to filter Pcdhβ1 isoform docking models. We did not assume an interaction between these specificity determinant residues and did not apply any distance constraints. Applying these filter conditions reduced the number of docked models from thousands to 287. The second constraint required all filtered docking models to have a buried surface area of more than 1200 $Å^2$ (600 $Å^2$ per protomer) at the binding interface, which further reduced the number of docked models to 149. To identify near-native docked homodimers we implemented the structural clustering algorithm described in (Lorenzen and Zhang, 2007). Briefly, we clustered all filtered docked models by generating an all-against-all RMSD matrix that was calculated by comparing the coordinates of protomers from different homodimer models after superposing the Cα atoms of their homophilic binding partners. Clusters were then defined by selecting a representative homodimeric model

with the most near-structural neighbors as defined by RMSD below an empirically selected threshold of 8-12 Å. Once the selected cluster representative and all its near-structural neighbors were removed from the docked-model pool, the homodimeric representative model for the next cluster was defined similarly. This procedure was repeated iteratively.

Sequence and structural alignment and Homology modelling

We modeled the structures for EC1-EC3 regions of Pcdhα7, Pcdhα8, Pcdhβ6, and Pcdhβ8 using Modeller (Eswar et al., 2006) with the crystal structures of PcdhαC2 and Pcdhβ1 as structural templates. Regions containing insertions relative to the templates were built using the LOOPY program (Soto et al., 2008). Structural alignments were calculated using DALI (Holm and Park, 2000) and SKA (Yang and Honig, 2000). Clustal Omega (Sievers et al., 2011) was used to calculate multiple sequence alignments.

Table S1: Data-collection and processing statistics for X-ray structures, related to Figure 1

	γC5 EC1-3 (Native)	γC5 EC1-3 (K_2I_4Hg)	γC5 EC1-3 (EMP)	γC5 EC1-3 (Native)	γC5 EC1-3 (Native)	γC5 EC1-3 Ext. N-terminus	αC2 EC1-3	β1 EC1-3	γA8 EC1-3
Data collection									
Date	8/15/2013	8/15/2013	8/15/2013	7/19/2013	7/19/2013	6/14/2014	9/26/2014	10/23/2014	10/23/2014
Beamline	APS 21ID-E	APS 21ID-E	APS 21ID-E	APS 21ID-E	APS 21ID-E	APS 21ID-E	BNL X4C	APS 21ID-E	APS 21ID-E
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$	$C2$	$P2_1$	$P2_1$	$P2_1$	$I2_12_12_1$	$I2_12_12_1$
Cell dimensions									
a, b, c (Å)	108.637, 108.637, 96.614	108.637, 108.637, 96.614	108.637, 108.637, 96.614	190.806, 104.916, 80.066	67.188, 84.563, 109.144	51.027, 108.874, 86.612	24.990, 97.130, 147.680	74.990, 106.520, 149.350	64.060, 78.120, 167.730
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 97.03, 90	90, 106.43, 90	90, 101.61, 90	90, 94.18, 90	90, 90, 90	90, 90, 90
Resolution (Å)	30.00−2.90 (3.08-2.90)	30.00-2.90 (3.00-2.90)	30.00-3.50 (3.62-3.50)	40.00-3.10 (3.21-3.10)	40.00-3.00 (3.11-3.00)	108.87-3.30 (3.48-3.30)	31.62-2.40 (2.53-2.40)	43.36-3.30 (3.56-3.30)	42.65-2.90 (3.08-2.90)
R_{merge}	0.10 (0.36)	0.15 (0.43)	0.17 (0.48)	0.11 (0.39)	0.07 (0.39)	0.12 (0.55)	0.08 (0.46)	0.15 (0.49)	0.08 (0.40)
$I/\sigma I$	19.8 (5.7)	13.5 (3.6)	16.0 (5.7)	10.0 (2.5)	16.8 (2.6)	9.6 (3.0)	9.1 (1.6)	4.6 (1.6)	10.2 (3.2)
Completeness (%)	99.9 (100.0)	98.5 (89.4)	99.8 (100.0)	94.4 (76.2)	99.1 (93.5)	99.0 (95.0)	95.8 (93.4)	97.9 (97.5)	98.9 (96.4)
Redundancy	14.2 (14.7)	11.7 (8.3)	14.0 (14.5)	3.5 (2.9)	3.7 (2.8)	3.8 (3.8)	3.0 (2.7)	2.6 (2.5)	3.1 (2.9)
Refinement									
Resolution (Å)	30.0−2.9			30.0-3.1	30.0-3.0	66.9-3.3	31.6-2.4	41.7-3.3	42.7-2.9
Number of reflections	13086			26871	23383	13880	26357	9005	9489
R_{work} / R_{free}	21.2 / 24.8			22.2 / 26.6	22.5 / 26.1	20.9 / 26.0	20.8 / 25.2	22.5 / 27.6	26.4 / 28.5
Number of residues									
Protein	311			938	642	640	634	316	313
Carbohydrate	5			14	8	9	5	12	2
Ion	6			18	12	15	12	6	7
Water	19			23	52	0	40	4	20
R.m.s. deviations									
Bond lengths (Å)	0.006			0.004	0.012	0.004	0.003	0.003	0.003
Bond angles (°)	1.229			0.780	0.871	0.753	0.730	0.878	0.654
Ramachandran									
Favored (%)	94.5			95.6	95.2	95.4	99.2	93.3	94.9
Allowed (%)	5.5			4.4	4.8	4.6	0.8	6.4	4.5
Outliers (%)	0.0			0.0	0.0	0.0	0.0	0.3	0.6
Wilson B	57.1			56.8	74.1	73.0	43.3	70.6	44.5
Average B	77.7			87.8	84.8	88.0	56.8	86.3	64.9
PDB ID	4ZPO			4ZPQ	4ZPP	4ZPN	4ZPM	4ZPL	4ZPS

Supplemental References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica Section D, Biological crystallography *66*, 213-221.

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R., and Leslie, A.G.W. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr D *67*, 271-281.

delaFortelle, E., and Bricogne, G. (1997). Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Method Enzymol *276*, 472-494.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta crystallographica Section D, Biological crystallography *66*, 486-501.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 5*, Unit 5 6.

Evans, P.R. (2007). An introduction to stereochemical restraints. Acta crystallographica Section D, Biological crystallography *63*, 58-61.

Halim, A., Carlsson, M.C., Madsen, C.B., Brand, S., Moller, S.R., Olsen, C.E., Vakhrushev, S.Y., Brimnes, J., Wurtzen, P.A., Ipsen, H.*, et al.* (2015). Glycoproteomic analysis of seven major allergenic proteins reveals novel post-translational modifications. Molecular & cellular proteomics : MCP *14*, 191-204.

Holm, L., and Park, J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics *16*, 566-567.

Lorenzen, S., and Zhang, Y. (2007). Identification of near-native structures by clustering protein docking conformations. Proteins *68*, 187-194.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. Journal of applied crystallography *40*, 658-674.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Method Enzymol *276*, 307-326.

Otwinowski, Z., and Minor, W. (2001). Precision of data in diffraction experiments. Nato Sci Ser I Life *325*, 55-60.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Soto, C.S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. Proteins *70*, 834-843.

Terwilliger, T.C., and Berendzen, J. (1999). Automated MAD and MIR structure solution. Acta crystallographica Section D, Biological crystallography *55*, 849-861.

Thu, C.A., Chen, W.V., Rubinstein, R., Chevee, M., Wolcott, H.N., Felsovalyi, K.O., Tapia, J.C., Shapiro, L., Honig, B., and Maniatis, T. (2014). Single-cell identity generated by combinatorial homophilic interactions between alpha, beta, and gamma protocadherins. Cell *158*, 1045-1059.

Yang, A.S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. Journal of molecular biology *301*, 665-678.

Legends for Supplemental Figures

Figure S1. Structural comparison of Pcdh EC1-EC3 domains, related to Figure 1

A) Comparison of the overall structures of Pcdhα2, Pcdhβ1, PcdhγA8, and PcdhγC5 EC1-EC3 domains (see also panel F).

B) Effect of the Ca^{2+} chelators EDTA or EGTA on cell aggregation mediated by N-cadherin or four Pcdh isoforms.

C) Assessing the binding properties of PcdhγC5 with the N-terminus used for crystallization and biophysical studies (left). Crystal structure of the EC1 domain of PcdhγC5 showing the extended structure of five N-terminal amino acids as cartoon (right).

D) Comparison of the crystal structures of the EC1 domain of PcdhαC2 and E-cadherin (PDB: 1L3W).

E) Comparisons of the structures most similar to the PcdhαC2 EC1, EC2 or EC3 domains.

F) Structural comparisons between Pcdhα2, Pcdhβ1, PcdhγA8, and PcdhγC5 EC1, 2 and 3 domains. The structural regions that differ between isoforms are noted.

G) Schematic representation of extracellular Pcdh domains EC1-EC3. White circles show sequence-based predicted O-Man glycosylation sites. Green-white circles show ambiguously identified glycosylation sites. Experimentally identified O-man glycosylation sites are shown in green circles. Evolutionary conservation of predicted sites of O-mannosylation is evident for Pcdhs in Human, Mouse, Frog, and Zebra fish.

H) Electrospray ionization (ESI) and Orbitrap-MS2 fragmentation of Pcdh glycopeptides. Higher-energy collisional dissociation (HCD-MS2) of Pcdh alpha-C2 glycopeptide. Shown is the spectrum of Ser196-Arg204 fragment only with mannose residues attached to Ser196 and Thr198. ETD-MS2 was performed for all glycopeptides and used for glycosylation site assignments (not shown). HCD-MS2 induced loss of mannose residues (green circles) from precursor ions or fragment ions is indicated in the spectra.

Figure S2. Elements of trans binding, related to Figure 2

A) Crystal structure of the PcdhγA8 EC1-EC3 is shown in cartoon. The disulfide bond formed between Cys 283 of each protomer is drawn as sticks and circled.

B) Reducing and non-reducing SDS gel of the chromatography fraction corresponding to the two picks in the elution profile consistent with dimeric or monomeric species.

C) The outcome of arginine scanning mutations on Pcdh binding in the cell aggregation assay.

Figure S3. EC domain shuffling to identify specificity-determining domains, related to Figure 3

A-E) Cells expressing chimeric proteins in which EC domains were shuffled between closely related isoforms were differentially tagged and tested for binding specificity with cells expressing their “parent” isoforms. Co-aggregation was noted for chimeras in which either the EC1-EC3 or EC2-EC4 domains matched with the corresponding domains of closely related isoforms. In addition we report cases for chimeras where three non-consecutive domains EC1, EC2 and EC4 or EC1, EC3 and EC4 are sufficient to mediate co-aggregation with the wild-type protein containing the same three domains (Figure S3A panels 3 and 5, S3B panels 4 and 6, S3D panel 4, and S3E panel 5).

Figure S4. Assessing specificity determining residues, related to Figure 5

A-C) Wild type isoforms appear at the left of each panel and the chimera with the shuffled residues appear on top of each panel. Each cell aggregation assay indicates whether a particular chimera recognizes the parent protein. Shown are only wild-type isoforms from which the EC domain originate. In all cases shown the chimera and wild type isoforms prefer to bind homophilically. Sequence alignments of shuffled regions are shown.

D) Shuffling of PcdhγA8 C-strand residues to PcdhγA9 residues resulted in a protein that could not mediate cell aggregation.

Figure S5. Domain shuffling to identify specificity-determining domains, related to Figure 6

A) Shown in cartoons are six symmetric homodimeric arrangements generated by docking of the four EC1-EC3 domains structures. The EC2 AB loop residue 116 and FG loop residue 301 are drawn as space filling and colored red and blue, respectively. The distribution of docked models is indicated as percentage for each arrangement.

B) A schematic representation of the Monte Carlo simulation used to estimate the average Pcdh assembly size for the simple case of three distinct isoforms expressed per cell. This model assumes that each cell contains a stable set of *cis* dimers formed from the random association of monomers present in each cell. The random incorporation of dimers with mismatched isoforms results in Pcdh chain termination (as indicated by an asterisk).

C) The average size of Pcdh assemblies, shown as the average number of *cis* dimers that comprise such assemblies, is depicted as a function of the number of mismatched isoforms between two contacting cells and the number of copies of each isoform.

Figure S1

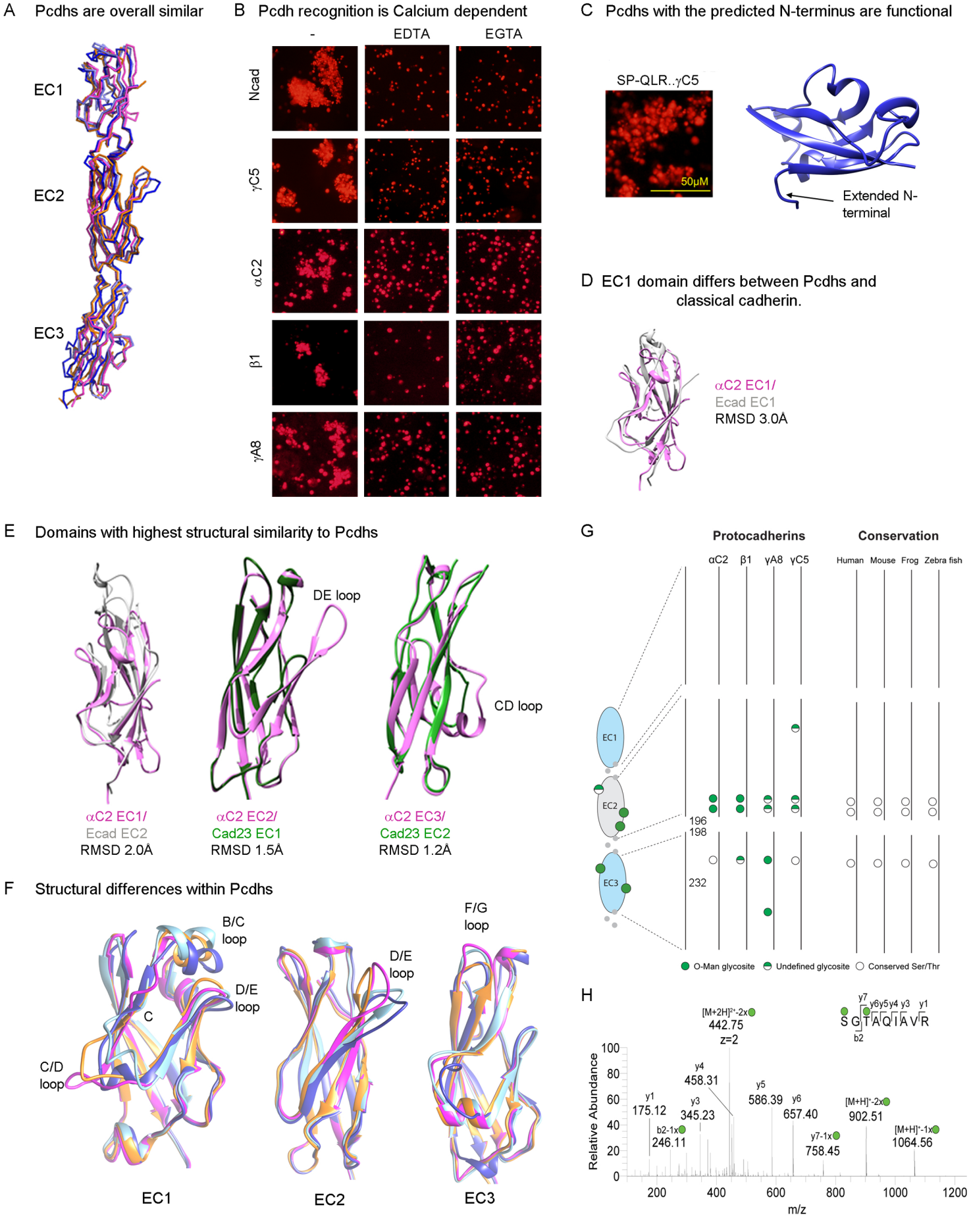
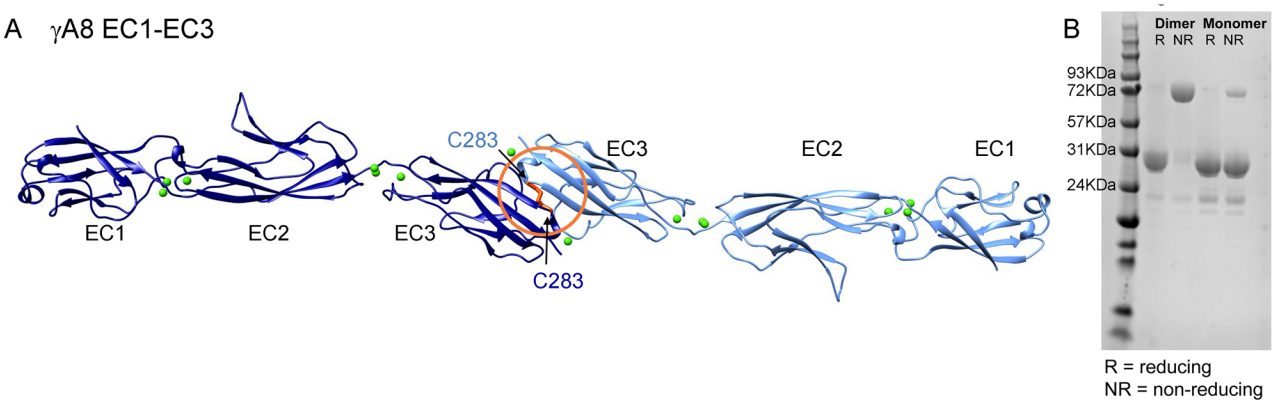

A Pcdhs are overall similar
EC1
EC2
EC3
B Pcdh recognition is Calcium dependent
-
EDTA
EGTA
Ncad
γC5
αC2
β1
γA8
C Pcdhs with the predicted N-terminus are functional
SP-QLR..γC5
50μM
Extended N-terminal
D EC1 domain differs between Pcdhs and classical cadherin.
αC2 EC1/
Ecad EC1
RMSD 3.0Å
E Domains with highest structural similarity to Pcdhs
DE loop
CD loop
αC2 EC1/
Ecad EC2
RMSD 2.0Å
αC2 EC2/
Cad23 EC1
RMSD 1.5Å
αC2 EC3/
Cad23 EC2
RMSD 1.2Å
F Structural differences within Pcdhs
B/C loop
D/E loop
C
C/D loop
D/E loop
F/G loop
EC1
EC2
EC3
G
Protocadherins
Conservation
αC2
β1
γA8
γC5
Human
Mouse
Frog
Zebra fish
EC1
EC2
EC3
196
198
232
O-Man glycosite
Undefined glycosite
Conserved Ser/Thr
H
[M+2H]2+-2x
442.75
z=2
S G T A Q I A V R
b2
y7
y6
y5
y4
y3
y1
Relative Abundance
m/z
y1
175.12
y4
458.31
y3
345.23
b2-1x
246.11
y5
586.39
y6
657.40
y7-1x
758.45
[M+H]+-2x
902.51
[M+H]+-1x
1064.56

Figure S2


A
γA8 EC1-EC3
C283
EC3
EC2
EC1
EC1
EC2
EC3
C283
B
Dimer
Monomer
R NR R NR
93KDa
72KDa
57KDa
31KDa
24KDa
R = reducing
NR = non-reducing

C

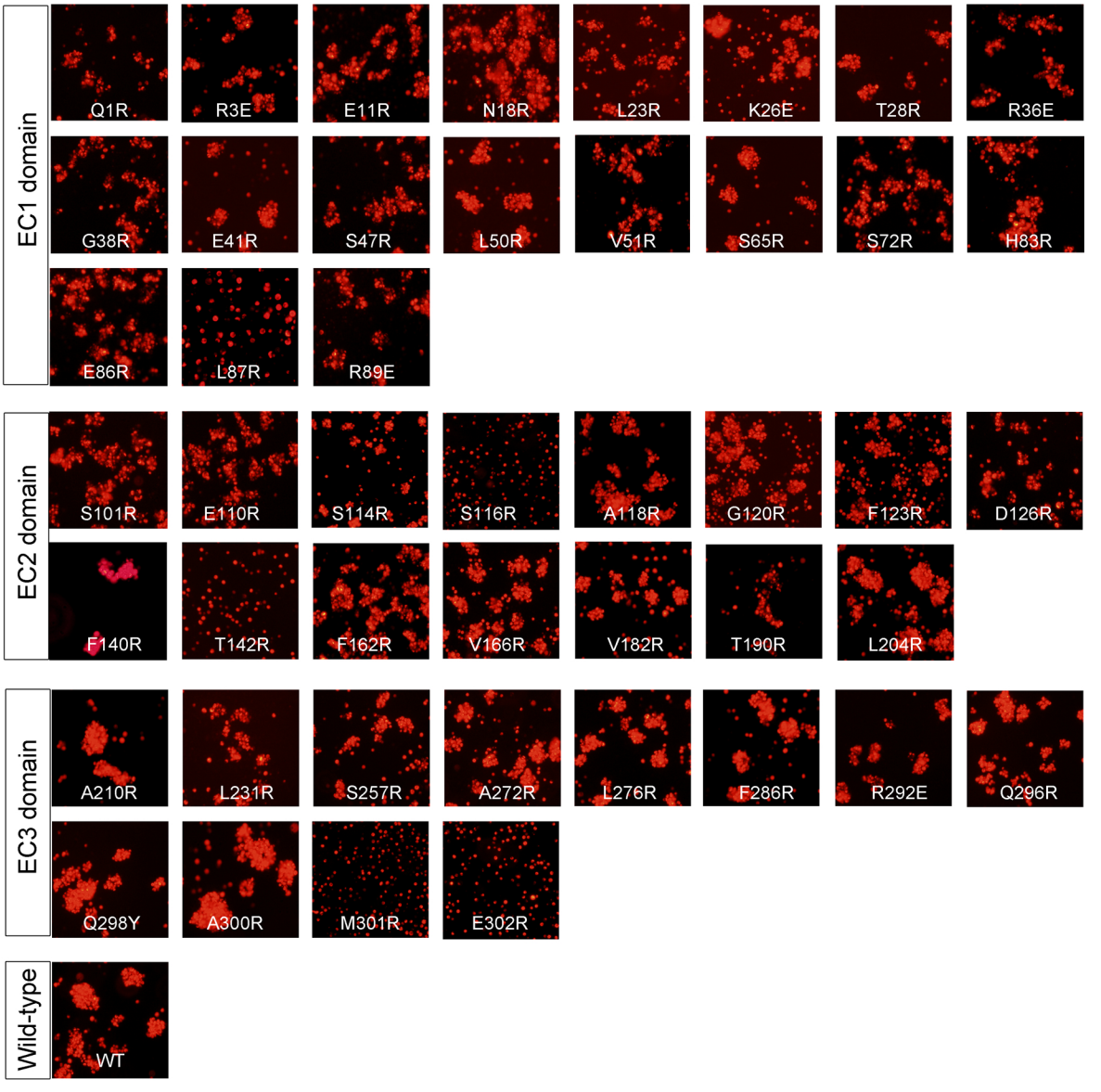
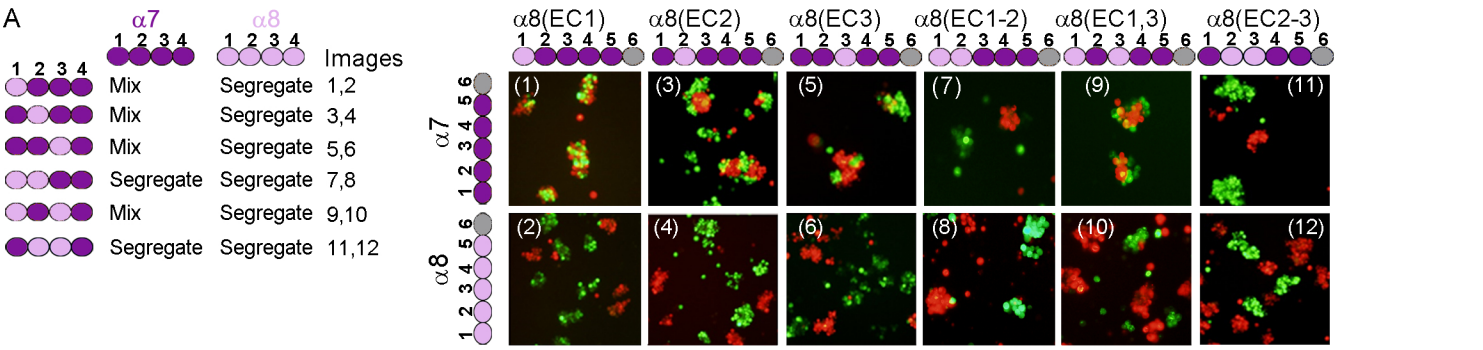
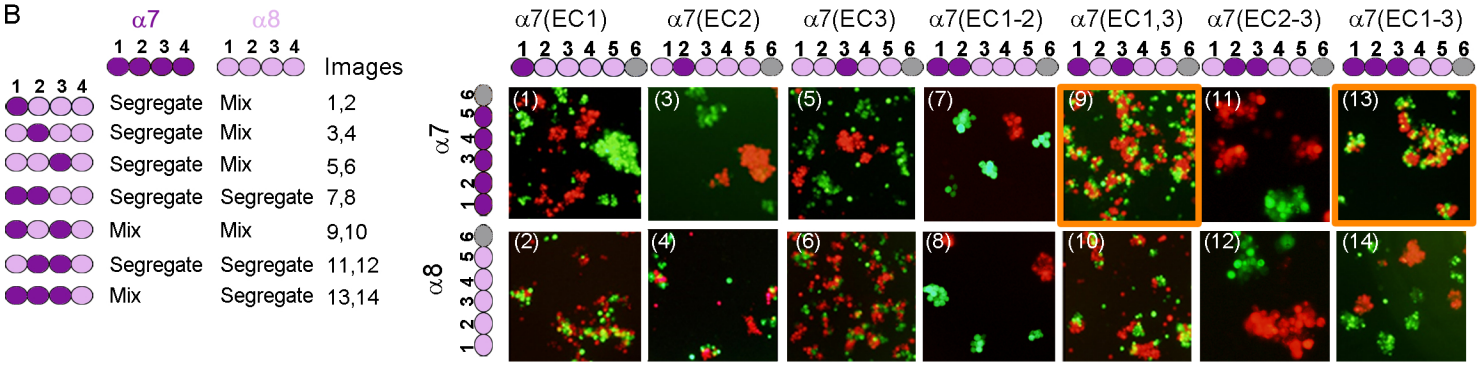
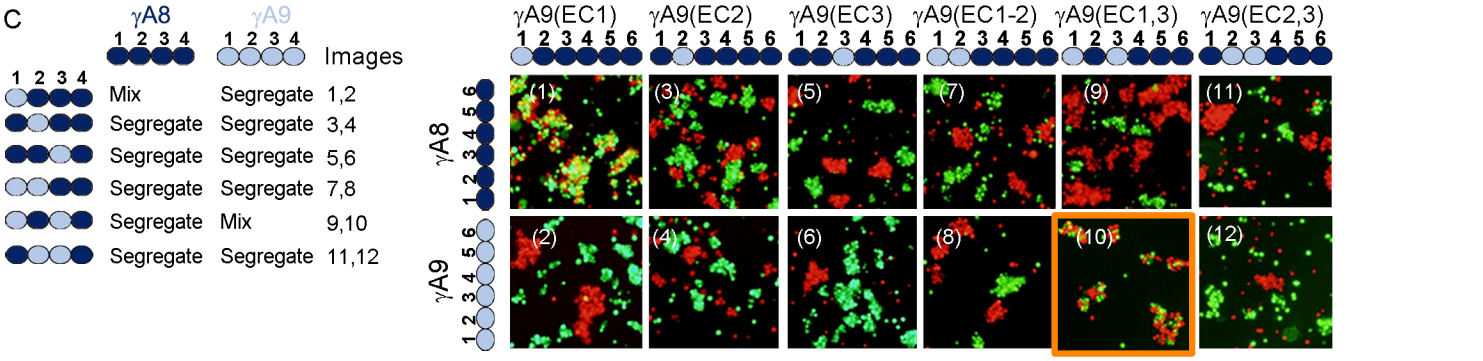
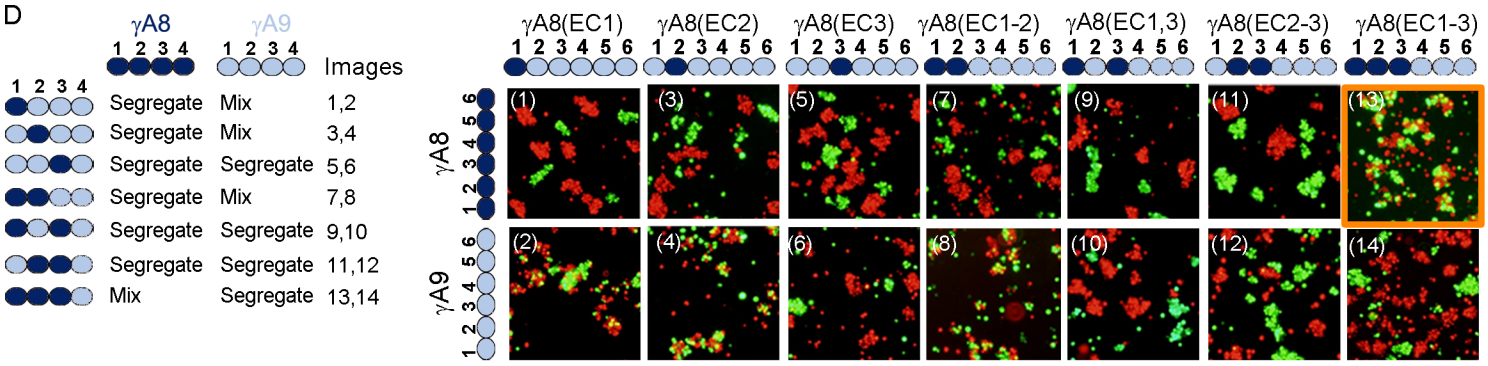

EC1 domain
Q1R
R3E
E11R
N18R
L23R
K26E
T28R
R36E
G38R
E41R
S47R
L50R
V51R
S65R
S72R
H83R
E86R
L87R
R89E
EC2 domain
S101R
E110R
S114R
S116R
A118R
G120R
F123R
D126R
F140R
T142R
F162R
V166R
V182R
T190R
L204R
EC3 domain
A210R
L231R
S257R
A272R
L276R
F286R
R292E
Q296R
Q298Y
A300R
M301R
E302R
Wild-type
WT

Figure S3


A
α7
α8
Images
Mix Segregate 1,2
Mix Segregate 3,4
Mix Segregate 5,6
Segregate Segregate 7,8
Mix Segregate 9,10
Segregate Segregate 11,12
α8(EC1)
α8(EC2)
α8(EC3)
α8(EC1-2)
α8(EC1,3)
α8(EC2-3)
α7
α8


B
α7
α8
Images
Segregate Mix 1,2
Segregate Mix 3,4
Segregate Mix 5,6
Segregate Segregate 7,8
Mix Mix 9,10
Segregate Segregate 11,12
Mix Segregate 13,14
α7(EC1)
α7(EC2)
α7(EC3)
α7(EC1-2)
α7(EC1,3)
α7(EC2-3)
α7(EC1-3)
α7
α8


C
γA8
γA9
Images
Mix Segregate 1,2
Segregate Segregate 3,4
Segregate Segregate 5,6
Segregate Segregate 7,8
Segregate Mix 9,10
Segregate Segregate 11,12
γA9(EC1)
γA9(EC2)
γA9(EC3)
γA9(EC1-2)
γA9(EC1,3)
γA9(EC2,3)
γA8
γA9


D
γA8
γA9
Images
Segregate Mix 1,2
Segregate Mix 3,4
Segregate Segregate 5,6
Segregate Mix 7,8
Segregate Segregate 9,10
Segregate Segregate 11,12
Mix Segregate 13,14
γA8(EC1)
γA8(EC2)
γA8(EC3)
γA8(EC1-2)
γA8(EC1,3)
γA8(EC2-3)
γA8(EC1-3)
γA8
γA9

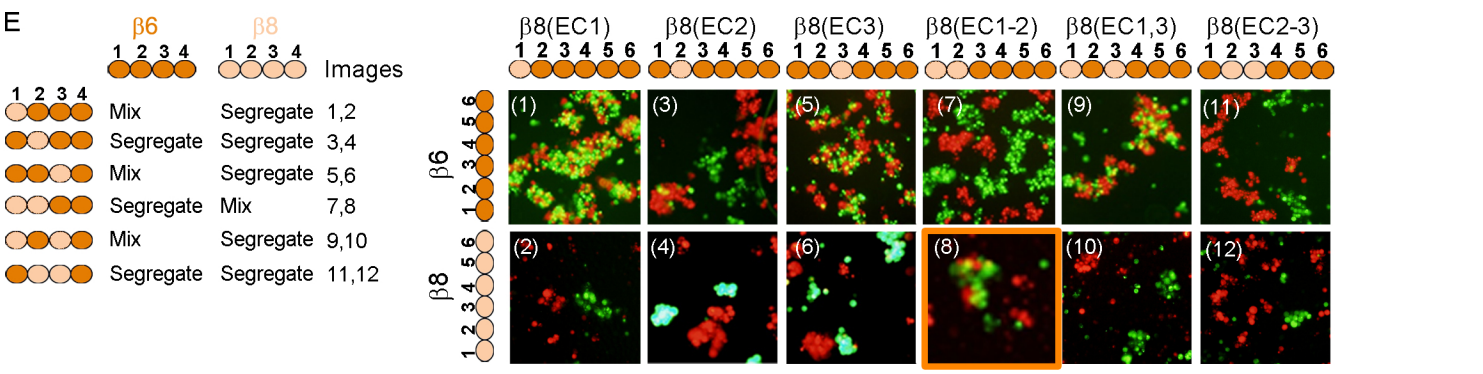

E
β6
β8
Images
Mix Segregate 1,2
Segregate Segregate 3,4
Mix Segregate 5,6
Segregate Mix 7,8
Mix Segregate 9,10
Segregate Segregate 11,12
β8(EC1)
β8(EC2)
β8(EC3)
β8(EC1-2)
β8(EC1,3)
β8(EC2-3)
β6
β8

Figure S4

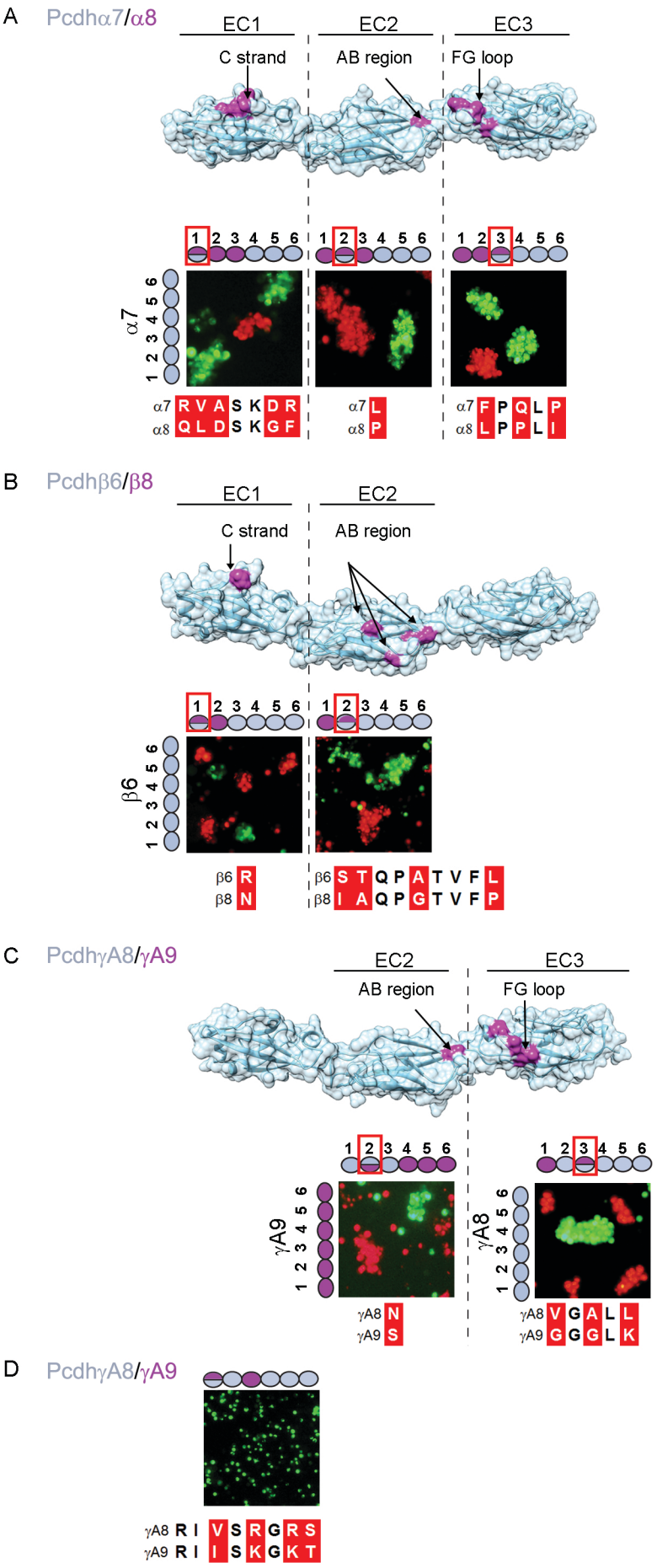
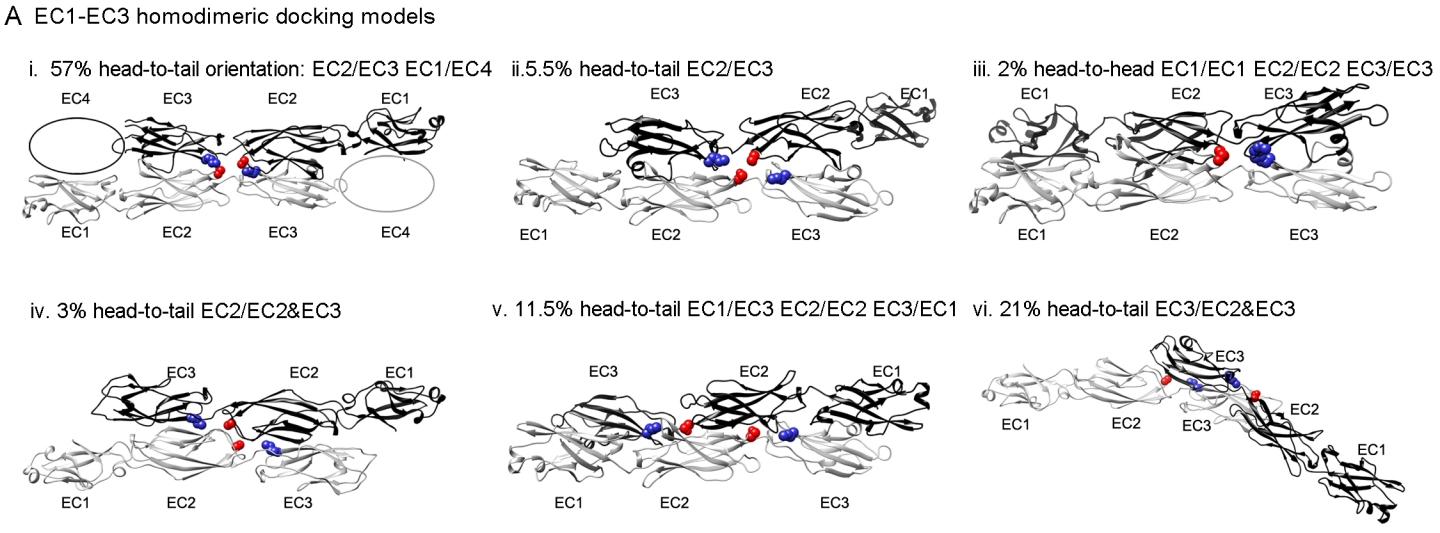
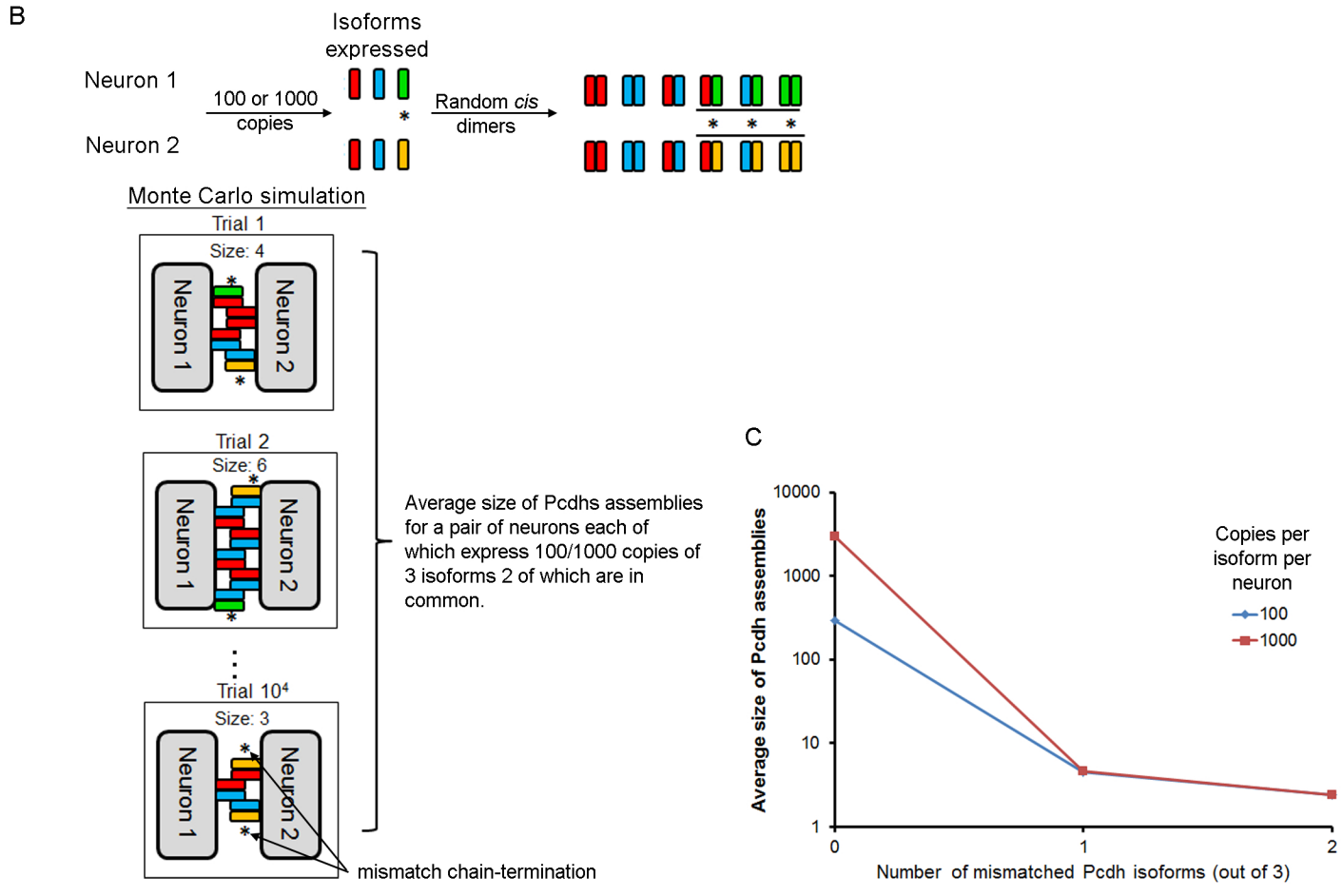

A
Pcdhα7/α8
EC1
EC2
EC3
C strand
AB region
FG loop
α7
α8
B
Pcdhβ6/β8
β6
β8
C
PcdhγA8/γA9
γA9
γA8
D
PcdhγA8/γA9
γA8 R I V S R G R S
γA9 R I I S K G K T

Figure S5


A EC1-EC3 homodimeric docking models
i. 57% head-to-tail orientation: EC2/EC3 EC1/EC4
EC4 EC3 EC2 EC1
EC1 EC2 EC3 EC4
ii.5.5% head-to-tail EC2/EC3
EC3 EC2 EC1
EC1 EC2 EC3
iii. 2% head-to-head EC1/EC1 EC2/EC2 EC3/EC3
EC1 EC2 EC3
EC1 EC2 EC3
iv. 3% head-to-tail EC2/EC2&EC3
EC3 EC2 EC1
EC1 EC2 EC3
v. 11.5% head-to-tail EC1/EC3 EC2/EC2 EC3/EC1
EC3 EC2 EC1
EC1 EC2 EC3
vi. 21% head-to-tail EC3/EC2&EC3
EC3
EC2
EC1
EC1 EC2 EC3
EC1


B
Isoforms expressed
Neuron 1
Neuron 2
100 or 1000 copies
Random cis dimers
Monte Carlo simulation
Trial 1
Size: 4
Neuron 1
Neuron 2
Trial 2
Size: 6
Neuron 1
Neuron 2
Trial 10⁴
Size: 3
Neuron 1
Neuron 2
mismatch chain-termination
Average size of Pcdhs assemblies for a pair of neurons each of which express 100/1000 copies of 3 isoforms 2 of which are in common.
C
Average size of Pcdh assemblies
10000
1000
100
10
1
0
1
2
Number of mismatched Pcdh isoforms (out of 3)
Copies per isoform per neuron
100
1000