Cysteine-rich domains related to Frizzled receptors and Hedgehog-interacting proteins

Jimin Pei[1]* and Nick V. Grishin[1,2]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390
[2]Departments of Biochemistry and Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75390

Received 15 March 2012; Revised 22 May 2012; Accepted 31 May 2012
DOI: 10.1002/pro.2105
Published online 12 June 2012 proteinscience.org

Abstract: Frizzled and Smoothened are homologous seven-transmembrane proteins functioning in the Wnt and Hedgehog signaling pathways, respectively. They harbor an extracellular cysteine-rich domain (FZ-CRD), a mobile evolutionary unit that has been found in a number of other metazoan proteins and Frizzled-like proteins in *Dictyostelium*. Domains distantly related to FZ-CRDs, in Hedgehog-interacting proteins (HHIPs), folate receptors and riboflavin-binding proteins (FRBPs), and Niemann-Pick Type C1 proteins (NPC1s), referred to as HFN-CRDs, exhibit similar structures and disulfide connectivity patterns compared with FZ-CRDs. We used computational analyses to expand the homologous set of FZ-CRDs and HFN-CRDs, providing a better understanding of their evolution and classification. First, FZ-CRD-containing proteins with various domain compositions were identified in several major eukaryotic lineages including plants and Chromalveolata, revealing a wider phylogenetic distribution of FZ-CRDs than previously recognized. Second, two new and distinct groups of highly divergent FZ-CRDs were found by sensitive similarity searches. One of them is present in the calcium channel component Mid1 in fungi and the uncharacterized FAM155 proteins in metazoans. Members of the other new FZ-CRD group occur in the metazoan-specific RECK (reversion-inducing-cysteine-rich protein with Kazal motifs) proteins that are putative tumor suppressors acting as inhibitors of matrix metalloproteases. Finally, sequence and three-dimensional structural comparisons helped us uncover a divergent HFN-CRD in glypicans, which are important morphogen-binding heparan sulfate proteoglycans. Such a finding reinforces the evolutionary ties between the Wnt and Hedgehog signaling pathways and underscores the importance of gene duplications in creating essential signaling components in metazoan evolution.

Keywords: frizzled; Hedgehog-interacting protein; FZ-CRD; HFN-CRD; Mid1; RECK; glypican

Introduction

Wnts and Hedgehogs are secreted morphogenetic proteins that play essential roles in various developmental processes and diseases.[1,2] Wnts are cysteine-rich glycoproteins modified with palmitoyl groups.[3] The major cell surface receptors for Wnts are the seven-transmembrane Frizzled proteins belonging to the superfamily of G-protein-coupled receptors (GPCRs).[4] Wnt interacts with the soluble N-terminal cysteine-rich domain (CRD) of Frizzled, which relays signals to downstream effectors inside the cell including Dishevelled to regulate a number of signaling events, e.g. β-catenin-dependent activation of target genes in the canonical Wnt pathway. Wnts and Frizzleds have been identified in all major groups of metazoans, including Porifera that is considered to be the sister group of all other metazoans.[5]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant number: GM094575 (to N.V.G.); Grant sponsor: Welch Foundation; Grant number: I-1505 (to N.V.G.).

**Correspondence to:* Jimin Pei, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX 75390. E-mail: jpei@chop.swmed.edu

Hedgehogs also undergo lipid modification with added palmitoyl groups. In addition, the functional N-terminal Hedge domain of hedgehog is freed from the Hint domain by proteolytic cleavage and is modified with a cholesterol group attached to its C-terminus. Hedgehogs are present in Cnidaria and most bilaterians, but are apparently absent from Porifera.[6] They are ligands of the transmembrane receptor Patched, which upon Hedgehog binding, releases the seven-transmembrane protein Smoothened to mediate downstream signaling events.[7] The Smoothened proteins are homologs of Frizzled receptors and also possess a related N-terminal cysteine-rich domain.

The CRD in Frizzled and Smoothened is a mobile evolutionary unit that has been found in other metazoan proteins, such as secreted Frizzled-related proteins (SFRPs) and certain receptor tyrosine kinases.[8,9] Recently, several domains distantly related to Frizzled CRDs, revealed by structural comparisons and sensitive sequence similarity searches, have been reported in Hedgehog-interacting proteins, folate receptors and riboflavin-binding proteins, and Niemann-Pick type C1 proteins.[10] Here we employed in-depth sequence and structural analyses to further expand and characterize the diverse repertoire of domains related to Frizzled CRDs, providing new insights into their evolution and classification.

Results and Discussion

Two groups of Frizzled-related cysteine-rich domains—FZ-CRDs and HFN-CRDs

Frizzled-like seven-transmembrane proteins have been identified in various metazoans[5,11] and the amoebozoan *Dictyostelium discoideum*.[12] Domains closely related to Frizzled CRDs are also present in a number of other metazoan proteins with diverse domain compositions, including secreted Frizzled-related proteins (SFRPs),[13] receptor tyrosine kinases Ror[14] and MuSK,[15] carboxypeptidase Z,[16] membrane-associated serine protease Corin,[17] and a long isoform of collagen XVIII.[18] Transitive PSI-BLAST[19] searches (see Materials and Methods) indeed detected Frizzled-related CRDs in all these proteins. The majority of these domains possess 10 conserved cysteines that exhibit a general pattern of “C*C*CX$_8$CX$_6$C*CX$_3$CX$_{6,7}$C*C*C” (C: conserved cysteine; *: a variable number of residues, X$_n$: n residues, and X$_{m,n}$: m to n residues) (Fig. 1 and Supporting Information Fig. S1). The number of residues between the seventh and eighth conserved cysteines is usually six, while receptor-tyrosine kinase-associated CRDs have seven residues between them. Structural studies of three CRDs, in mouse Frizzled,[8] mouse SFRP3,[8] and rat MuSK,[9] revealed a common fold mainly consisting of four core alpha-helices. In all these structures, the disulfide connectivity patterns among the 10 conserved cysteines are C1–C5 (between the first and fifth conserved cysteines), C2–C4, C3–C8, C6–C10, and C7–C9. We collectively refer to these domains as FZ-CRDs.

Previous structural comparisons and profile-profile-based similarity searches also revealed several domains distantly related to FZ-CRDs in several proteins including Hedgehog-interacting proteins (HHIPs), folate receptors and riboflavin-binding proteins (FRBPs), and Niemann-Pick disease Type C1 proteins (NPC1s).[10,20] Transitive PSI-BLAST searches starting from the CRD in human HHIP (gi: 20143973, residues 20–220) found CRDs in HHIPs, FRBPs, and NPC1s with statistically significant scores (e-value inclusion threshold: 1e-4), but did not find known FZ-CRDs. Conversely, transitive PSI-BLAST searches starting from FZ-CRDs did not identify any member in HHIPs, FRBPs, and NPC1s with statistically significant scores. These results suggest that the CRDs in HHIPs, FRBPs, and NPC1s are more closely related to each other than to FZ-CRDs. Therefore, we collectively refer to the CRDs in HHIPs, FRBPs, and NPC1s as HFN-CRDs.

The grouping of HFN-CRDs is also supported by their cysteine patterns. HHIPs and FRBPs share 12 conserved, disulfide-bonded cysteines that adopt a general pattern of “C*C*CC*CX$_8$CX$_{2,3}$C*CX$_3$CX$_6$ C*C*C” (Fig. 1). NPC1s possess eight of these conserved cysteines, while lacking two disulfide bonds formed by C8–C12 and C9–C11 in HHIPs and FRBPs (Fig. 1). The cysteine patterns and disulfide connectivity of HFN-CRDs are similar to those of FZ-CRDs (Fig. 1). Compared with FZ-CRDs with 10 conserved cysteines, the most noticeable difference in HFN-CRDs is the addition of a “CC” motif after the second conserved cysteine. The two cysteines in this “CC” motif (C3 and C4) form disulfide bonds with the first and seventh cysteines (C1 and C7) in HFN-CRDs, respectively (Fig. 1), as revealed by the CRD structures of NPC1[21] and a riboflavin-binding protein.[22] C1 and C7 in the motif of HFN-CRDs correspond to C1 and C5 in the motif of FZ-CRDs (Fig. 1), which form a disulfide bond directly. Another subtle difference of cysteine patterns between HFN-CRDs and FZ-CRDs lies in the number of residues in between C6 and C7 in HFN-CRDs, corresponding to C4 and C5 in FZ-CRDs. While HFN-CRDs have a shorter stretch of residues between C6 and C7 (two residues in HHIPs and three residues in FRBPs and NPC1s), C4 and C5 in FZ-CRDs generally have six residues in between (Fig. 1).

New members of FZ-CRDs in plants and protists

FZ-CRDs have been previously reported only in metazoans and amoebozoans. Our sequence similarity searches expanded the repertoire of FZ-CRDs to include members from several other major eukaryotic lineages, such as green plants (both land plants and green algae), stramenopiles, Alveolata (both Apicomplexa and ciliates), and the Heterolobosea species *Naegleria gruberi* (belonging to the Excavata supergroup). Newly identified FZ-CRD-containing proteins in plants and various protists exhibit diverse domain compositions (Fig. 2).

FZ-CRDs in land plants are highly divergent, as they cannot be linked to metazoan FZ-CRDs by PSI-BLAST searches. Instead, their distant sequence similarities to known FZ-CRDs were revealed by the more sensitive profile-profile-based HHpred method.[23] These land plant proteins co-occur with a leishmanolysin domain[24] with the signature "HEXXH" motif (Peptidase_M8 in Pfam[25]) (Fig. 2). The GPI-anchored leishmanolysin (gp63) is the most abundant surface glycoprotein of *Leishmania* spp.,[26] and leishmanolysin-domain-containing proteins have been identified in various eukaryotic lineages including metazoans, plants, and various protists.[27] These land plant proteins are also predicted to be GPI-anchored.[28,29] Association of an FZ-CRD domain with a leishmanolysin domain appears to be unique in land plants. Interestingly, FZ-CRDs also co-occur with peptidase domains in some metazoan proteins, such as the Type II plasma membrane protein Corin and the extracellular protein carboxypeptidase Z (Fig. 2). FZ-CRDs could facilitate peptidase functions by contributing to substrate binding or interactions with other proteins.

FZ-CRDs associated with the GPCR family of seven-transmembrane proteins have been identified in Metazoa (Frizzled and Smoothened) and Amoebozoa,[12] two major lineages previously classified in the Unikonta supergroup of eukaryotes.[30] On the other hand, we did not find FZ-CRD-containing GPCR proteins from Bikonta lineages (including plants and most protist groups such as stramenopiles, Alveolata, Rhizaria, and Excavata). Therefore, FZ-CRD-containing GPCRs might have originated in the last common ancestor of Opisthokonta and Amoebozoa. We found two divergent FZ-CRD-containing proteins from the amoebozoan *D. discoideum* that are not associated with seven-transmembrane GPCRs, but instead contain other putative extracellular domains (Fig. 2). HHpred[23] search suggests that one of them (gi: 66807523) contains an ependymin-like domain, as it found the Pfam domain Ependymin (PF00811) with a probability score of 99.91%. HHpred probability score (in percentage unit, range: 0–100) provides the estimation of the likelihood that the hit is homologous to the query based on combined profile similarities of amino acids and secondary structures in the framework of a hidden Markov model (scores above 95% are often considered as reliable predictions of homologous relationships). Ependymin is an extracellular protein abundant in fish cerebrospinal fluid[31] and has been previously found only in metazoans.[32] The other FZ-CRD-containing *D. discoideum* protein (gi: 66813234) is predicted to be a GPI-anchored protein and has an uncharacterized N-terminal region with mainly predicted beta-strands. HHpred searches revealed distant similarity between this region and a domain of unknown function DUF3129 (Pfam: DUF3129, probability score: 97.92%) and a chitin-binding domain (Pfam: Chitin_bind_3, probability score: 95.66%).

Lineage-specific expansions and losses were observed for FZ-CRD-containing proteins in nonmetazoans (see Supporting Information Fig. S2 for a list of nonmetazoan FZ-CRDs and their alignment). Most noticeably, several amoebozoans have multiple copies (large than 10) of FZ-CRD-containing proteins (Supporting Information Fig. S2), most of which are associated with GPCRs. Several stramenopiles species, such as *Phytophthora infestans* and *Albugo laibachii*, possess more than one FZ-CRD-containing proteins. Two FZ-CRDs have been identified in the

Figure 1. Multiple sequence alignment of FZ-CRDs and HFN-CRDs. This alignment is divided into four upper sections of FZ-CRDs and four lower sections of HFN-CRDs. Known FZ-CRDs in metazoans are in the subalignment of "FZ-CRD–metazoan," and two newly identified divergent FZ-CRD groups with metazoan members are in the subalignments of "RECK" and "Mid1/FAM155." Each sequence is denoted by its NCBI gene accession number. Available common gene symbols for some proteins are shown after their accession numbers. Domains with solved structures have their PDB codes shown in red letters. Conserved cysteines are shown on black background and substitutions of them are shown on grey background. Disulfide connections are shown and labeled for 10 conserved cysteine positions in FZ-CRDs (labels are: *: C1–C5; #, C2–C4; +, C3–C8; =, C6–C10; and, C7–C9) and 12 conserved cysteine positions in HFN-CRDs (labels are: red *, C1–C3; #, C2–C6; blue *, C4–C7; +, C5–C10; =, C8–C12; and, C9–C11). Starting and ending residues numbers are shown before and after the sequences, respectively. Protein lengths are shown in brackets at the end. Noncharged residues in positions with mainly hydrophobic residues are shaded in yellow. Insertion regions are replaced by the number of inserted residues in parentheses. In glypicans two long inserted regions are highlighted by red numbers. Two-letter species name abbreviations shown after the accession numbers are as follows: Aq, *Amphimedon queenslandica*; At, *Arabidopsis thaliana*; Bn, *Bigelowiella natans*; Dd, *Dictyostelium discoideum*; Dm, *Drosophila melanogaster*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Lm, *Leishmania major*; Ms, *Micromonas* sp.; Mm, *Mus musculus*; Ng, *Naegleria gruberi*; Nv, *Nematostella vectensis*; Pi, *Phytophthora infestans*; Pf, *Plasmodium falciparum*; Rn, *Rattus norvegicus*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; and Tt, *Tetrahymena thermophila*. Their colors are shown as follows: metazoan, black; plants, green; protists, blue; and fungi, orange. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

green alga *Micromonas* sp. RCC299, while they appear to be lacking in the green alga *Chlamydomonas reinhardtii*. Only one FZ-CRD copy seems to be present in the land plant *Arabidopsis thaliana* and the apicomplexan *Plasmodium facalcium*. The Heterolobosea species *N. gruberi*, on the other hand, possesses multiple FZ-CRD-containing proteins (Supporting Information Fig. S2), including a short protein with a single FZ-CRD and a predicted transmembrane segment (TM) (GenBank ID: XP_002670836.1, Fig. 2) and several large proteins with more than 1000 amino acid residues. These large proteins (one of them shown in Fig. 2, GenBank ID: XP_002672407.1) have many EGF-like repeats and a C-terminal REJ module, previously found in cell surface proteins such as PKD1.[33,34]

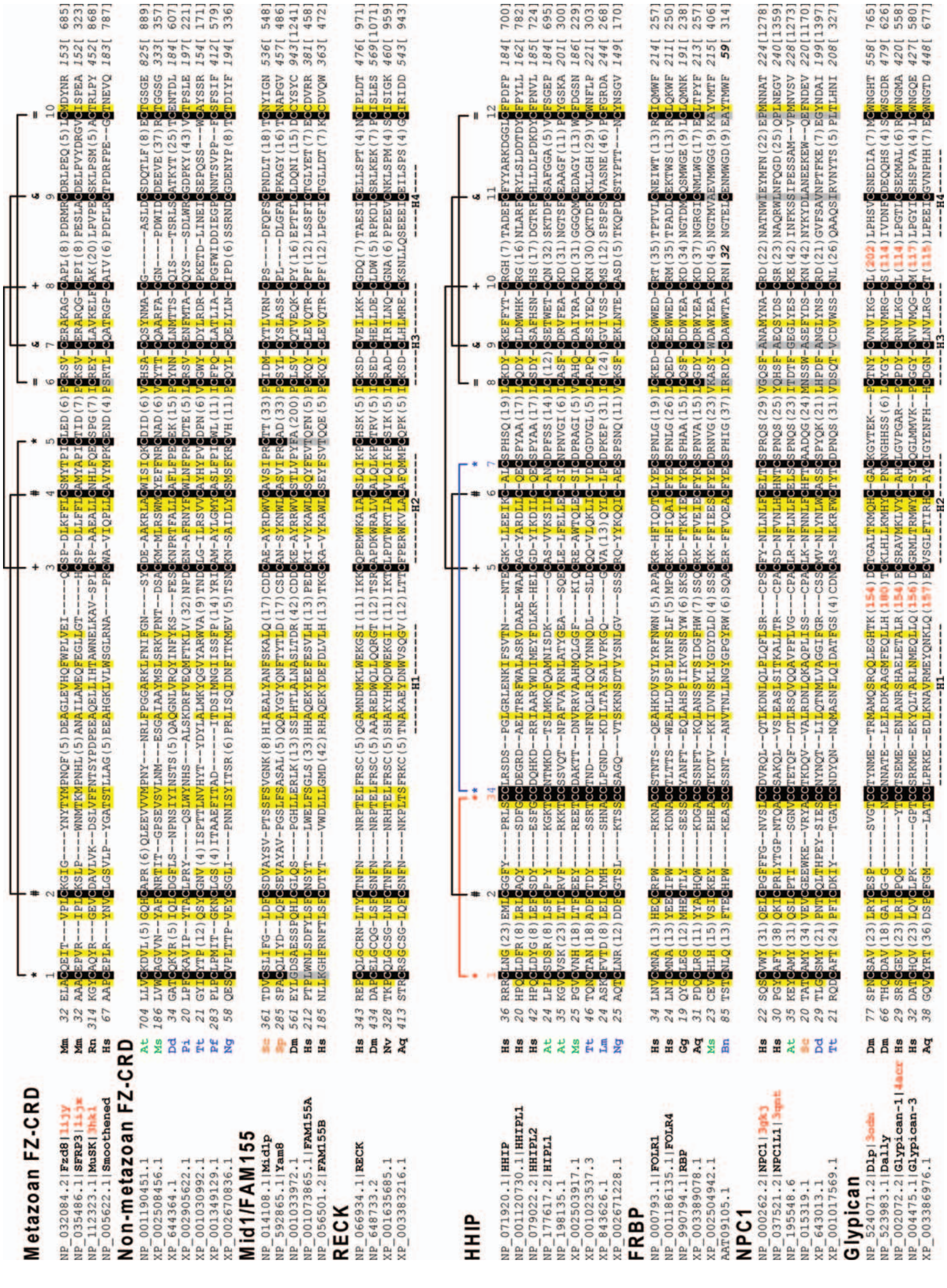

Figure 1

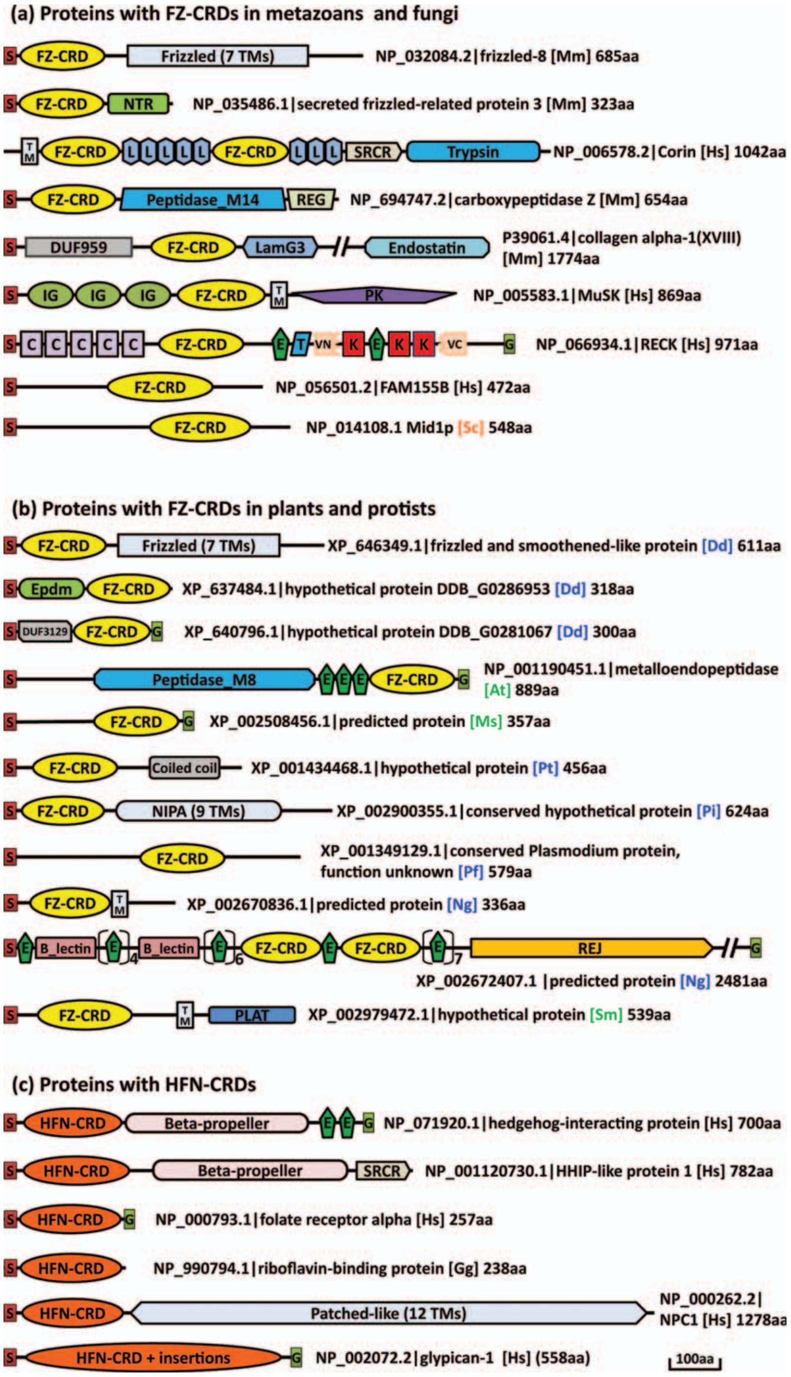
(a) Proteins with FZ-CRDs in metazoans and fungi
S FZ-CRD Frizzled (7 TMs)
NP_032084.2|frizzled-8 [Mm] 685aa
S FZ-CRD NTR
NP_035486.1|secreted frizzled-related protein 3 [Mm] 323aa
TM FZ-CRD L L L L L FZ-CRD L L L SRCR Trypsin
NP_006578.2|Corin [Hs] 1042aa
S FZ-CRD Peptidase_M14 REG
NP_694747.2|carboxypeptidase Z [Mm] 654aa
S DUF959 FZ-CRD LamG3 Endostatin
P39061.4|collagen alpha-1(XVIII) [Mm] 1774aa
S IG IG IG FZ-CRD TM PK
NP_005583.1|MuSK [Hs] 869aa
S C C C C C FZ-CRD E T VN K E K K VC G
NP_066934.1|RECK [Hs] 971aa
S FZ-CRD
NP_056501.2|FAM155B [Hs] 472aa
S FZ-CRD
NP_014108.1 Mid1p [Sc] 548aa
(b) Proteins with FZ-CRDs in plants and protists
S FZ-CRD Frizzled (7 TMs)
XP_646349.1|frizzled and smoothened-like protein [Dd] 611aa
S Epdm FZ-CRD
XP_637484.1|hypothetical protein DDB_G0286953 [Dd] 318aa
S DUF3129 FZ-CRD G
XP_640796.1|hypothetical protein DDB_G0281067 [Dd] 300aa
S Peptidase_M8 E E E FZ-CRD G
NP_001190451.1|metalloendopeptidase [At] 889aa
S FZ-CRD G
XP_002508456.1|predicted protein [Ms] 357aa
S FZ-CRD Coiled coil
XP_001434468.1|hypothetical protein [Pt] 456aa
S FZ-CRD NIPA (9 TMs)
XP_002900355.1|conserved hypothetical protein [Pi] 624aa
S FZ-CRD
XP_001349129.1|conserved Plasmodium protein, function unknown [Pf] 579aa
S FZ-CRD TM
XP_002670836.1|predicted protein [Ng] 336aa
S E B_lectin E 4 B_lectin E 6 FZ-CRD E FZ-CRD E 7 REJ G
XP_002672407.1 |predicted protein [Ng] 2481aa
S FZ-CRD TM PLAT
XP_002979472.1|hypothetical protein [Sm] 539aa
(c) Proteins with HFN-CRDs
S HFN-CRD Beta-propeller E E G
NP_071920.1|hedgehog-interacting protein [Hs] 700aa
S HFN-CRD Beta-propeller SRCR
NP_001120730.1|HHIP-like protein 1 [Hs] 782aa
S HFN-CRD G
NP_000793.1|folate receptor alpha [Hs] 257aa
S HFN-CRD
NP_990794.1|riboflavin-binding protein [Gg] 238aa
S HFN-CRD Patched-like (12 TMs)
NP_000262.2| NPC1 [Hs] 1278aa
S HFN-CRD + insertions G
NP_002072.2|glypican-1 [Hs] (558aa)
100aa

Figure 2. Domain structure diagrams for select proteins containing FZ-CRDs or HFN-CRDs. The diagrams were drawn approximately to scale (lower right corner). Two exceptions are the long mouse collagen alpha-1, in which the middle region with mainly collagen tripeptide repeats is replaced by a double slash, and a long protein from *N. gruberi*, where the repeated EGF domains are shown in parentheses with the number of repeats indicated. NCBI accession number and annotation are shown for each protein, followed by species name abbreviation (in brackets) and protein length. Domains with multiple predicted transmembrane segments (TM) have the number of TMs shown in parentheses. Domain or module name abbreviations are as follows: C, repeated cysteine-rich module with unknown function in RECK proteins; E, EGF-like domain; G, predicted GPI C-terminal signal peptide; K, Kazal domain; L: LDLa domain; S: N-terminal signal peptide; T: TILa domain; Epdm, ependymin domain; IG, immunoglobulin domain; LamG3, laminin_G_3 domain; NIPA, magnesium transporter NIPA domain; NTR, UNC-6/NTR/C345C module; PK, protein kinase domain; PLAT, PLAT (polycystin-1, lipoxygenase, alpha-toxin)/LH2 (lipoxygenase homology) domain; REG, carboxypeptidase regulatory domain; REJ, REJ (receptor for egg jelly) domain; SRCR, scavenger receptor cysteine-rich domain; TM: transmembrane segment; VN, VWC N-terminal subdomain; and VC, VWC C-terminal subdomain. Species name abbreviations are: At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Ms, *Micromonas* sp.; Ng, *Naegleria gruberi*; Pf, *Plasmodium falciparum*; Pi, *Phytophthora infestans*; Pt, *Paramecium tetraurelia*; Sc, *Saccharomyces cerevisiae*; and Sm, *Selaginella moellendorffii*. Color codings for species names are: metazoan, black; plant, green; fungi, orange; and protist, blue. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

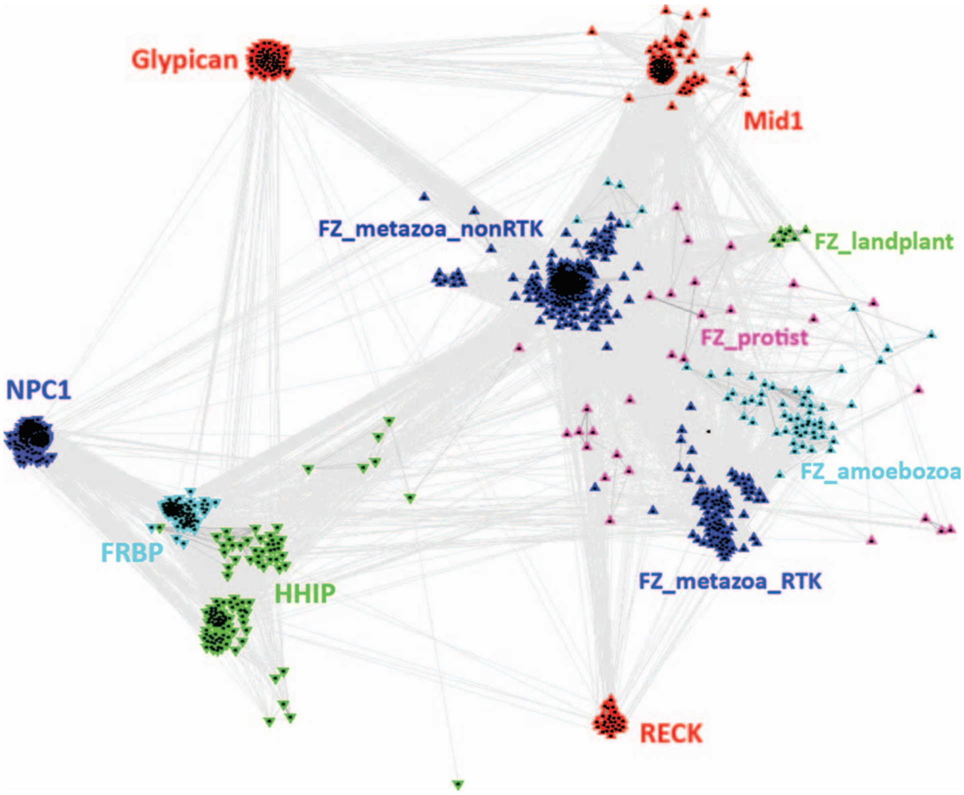

Figure 3. Sequence clustering diagram of FZ-CRDs and HFN-CRDs. This diagram is generated by the CLANS program on a nonredundant set of FZ-CRDs and HFN-CRDs (see Materials and Methods). Protein domains are represented as dots inside triangular shapes, and the connections between them suggest BLAST hits with *P* values less than 1e-4. FZ-CRDs and HFN-CRDs are shown as dots inside upper and lower triangles, respectively. Three newly identified groups (Mid1, RECK, and glypican) are shown in red triangles. FZ-CRDs except for the divergent Mid1 and RECK groups are colored and labeled based on species (blue, metazoans; green, land plants; cyan, amoebozoans; and magenta, other protists). The metazoan FZ-CRDs are mainly in two separate clusters labeled as FZ_metazoa_RTK (FZ-CRDs in receptor tyrosine kinases) and FZ_metazoa_nonRTK (other metazoan FZ-CRDs, including those in Frizzled, Smooothened, SFRPs, and Corin), respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Like metazoan FZ-CRDs, FZ-CRDs in nonmetazoan proteins are most likely located in the extracellular space, based on frequent occurrences of predicted N-terminal signal peptides, predicted TMs, other extracellular domains, and C-terminal hydrophobic segments that are often predicted as GPI anchor-modified sites (Fig. 2). Functions of FZ-CRD-containing proteins in nonmetazoans remain unknown. They could mediate protein-protein interactions, similar to the known metazoan FZ-CRDs in Frizzled receptors and SFRPs that interact with Wnts. Since Wnts are only found in metazoans, the interaction partners of nonmetazoan FZ-CRDs remain to be discovered. Some of these FZ-CRDs may also be capable of binding small molecule ligands such as lipids, based on their distant similarity to HFN-CRDs with ligand-binding capabilities such as FRBPs and NPC1s.[10] The diverse domain compositions of nonmetazoan FZ-CRD-containing proteins indicate that these FZ-CRDs could have evolved differing functions, e.g. through interactions with different molecules.

Two divergent new groups of FZ-CRDs

We identified two new groups of divergent FZ-CRDs present in the fungal protein Mid1 and the metazoan protein RECK, respectively. While transitive PSI-BLAST did not link these two groups to known FZ-CRDs, sensitive profile-profile searches using HHpred supported their homologous relationships with good statistical scores (probability scores above 90%). These inferred homologous relationships are also supported by 10 conserved cysteines (Fig. 1). As revealed by CLANS[35] (Fig. 3), members of these two new FZ-CRDs form two clusters well separated from other FZ-CRD groups, suggesting that they have undergone significant sequence divergence compared with known FZ-CRDs.

A divergent FZ-CRD in Mid1 proteins

A divergent FZ-CRD was found in the fungal protein Mid1 (Pfam: PF12929), a stretch-activated calcium channel component. Mid1 forms a complex with the transmembrane calcium channel Cch1, both shown to be required for calcium influx and mating in fungi.[36–38] Most of the fungal Mid1 sequences have a predicted N-terminal signal peptide. Some of them also possess a hydrophobic segment at the very C-terminal end, suggesting that they are GPI-anchored proteins. Indeed, about half of fungal Mid1 sequences were predicted to be GPI-anchored proteins by the GPI-SOM sever.[28] The cysteine-rich domain is located near the C-terminus of Mid1 and is essential for its function[39] (Fig. 2).

Sequence similarity searches also identified putative metazoan orthologs of Mid1, including those from insects, chordates, and several basal metazoan groups (e.g. Cnidaria, Placozoa, and EST evidence of Porifera, Supporting Information Fig. S3). Human has two copies of putative Mid1 orthologs named FAM155A and FAM155B respectively, both of which do not have known function. Mid1-like sequences have limited distribution outside fungi and metazoans, with members identified from land pants, green algae, stramenopiles, and an EST sequence from the putative Excavata species *Malawimonas jakobiformis* (gi: 110040638) (Supporting Information Fig. S3). Mid1 may have appeared early in the evolution of eukaryotes, but have undergone independent gene losses in many lineages outside Opisthokonta. A lineage-specific gene loss event is evidenced in the group of land plants, as no Mid1-like proteins were found in flowering plants (angiosperms, e.g. *Arabidopsis thaliana*), although they are present in "lower" land plants such as Bryophyta (mosses, e.g. gi: 168019397 from *Physcomitrella patens*), Pinophyta (conifers, e.g. gi: 294460988 from *Picea sitchensis*), and Cycadophyta (cycads, e.g. EST sequence gi: 158959386 from *Cycas sporophyll*). Functions of Mid1-like proteins in nonfungal species remain to be investigated.

A divergent FZ-CRD domain in the metazoan RECK proteins

A second group of previously unidentified FZ-CRDs occurs in RECK proteins (reversion-inducing cysteine-rich proteins with Kazal motifs). The human glycoprotein RECK was identified in a screen of putative suppressors of tumor invasion.[40] Presence of an N-terminal signal peptide and a C-terminal hydrophobic segment suggests that RECK is GPI-anchored to the plasma membrane[40] (both GPI-SOM and FragAnchor severs gave a positive GPI anchor prediction for human RECK, gi: 11863156). RECK plays an essential role in development, and it can suppress tumor invasion and metastasis by acting as inhibitors of several matrix metalloproteases that break down extracellular matrix.[41,42]

Several cysteine-rich domains or motifs have been reported in RECK, including five N-terminal repeats of a module with six cysteines (proposed to form cysteine knots), three Kazal domains, and two putative EGF-like domains[40] (Fig. 2). Kazal domain was originally identified as a serine protease inhibitor. The second and third Kazal domains in RECK proteins were shown to be able to inhibit the matrix metalloprotease MMP9.[43] Frequently occurring in extracellular proteins, some Kazal domains may possess other functions. For example, Kazal domains in follistatin are involved in binding and inhibiting the TGF-β family proteins such as activin.[44] They are also present in complement proteins C6 and C7 and play roles in interactions with other complement proteins.[45]

HHpred results also suggested a VWC (von Willebrand factor type C domain) N-terminal subdomain and a VWC C-terminal subdomain sandwiching the three Kazal domains in RECK (Fig. 2). The VWC modules frequently occur in extracellular proteins, such as the Chordin family proteins, and are involved in binding bone marrow proteins (BMPs, belonging to the TGF-β family).[46] Regions in RECK with distant sequence similarities to two EGF-like domains (as proposed before)[40] and a TILa domain were also suggested by HHpred.

The newly identified FZ-CRD domain in RECK lies in the middle region after the five N-terminal repeats and before the Kazal domains and the other domains (Fig. 2). A subtle cysteine pattern variation in RECK FZ-CRDs lies in the number of residues between the third and fourth conserved cysteines: RECK FZ-CRDs have nine residues while most of other FZ-CRDs and HFN-CRDs have eight residues (Fig. 1). RECK proteins were only identified in metazoans, including basal groups such as Porifera, Cnidaria, and Placozoa. RECK thus could be a metazoan invention that coincides with the occurrence of extracellular matrix and multicellularity. The presence of multiple domains in RECK suggests that it could interact with different extracellular proteins and have a diverse array of functions.

Divergent evolution of known HFN-CRD families

As described above, HFN-CRDs, occurring in HHIPs, FRBPs, and NPC1s, are distantly related to FZ-CRDs and have a similar, yet unique cysteine pattern compared with that of FZ-CRDs. Separation of HFN-CRDs from FZ-CRDs is also reflected in the sequence clustering result generated by CLANS (Fig. 3). We investigate phylogenetic relationships of proteins in each of the three known HFN-CRD families.

HFN-CRDs in HHIPs and HHIP-like proteins. Human HHIP is a GPI-anchored protein with an N-terminal HFN-CRD, a beta-propeller domain, and two C-terminal EGF domains.[20,47] Two HHIP-like proteins (HHIPL1 and HHIPL2) also exist in the human genome (Fig. 1). HHIPL1 and HHIPL2 possess the HFN-CRD domain and the beta-propeller domain, but lack the EGF domains. HHIP-like proteins with a beta-propeller domain were also found in land plants and some stramenopiles species. In addition, our similarity searches also revealed other HHIP-like proteins that do not possess beta-propeller domains. HFN-CRDs in most of these HHIPs and HHIP-like proteins exhibit a cysteine pattern of "$C*C*CC*CX_8CX_2C*CX_3CX_6C*C*C$" (Fig. 1 and Supporting Information Fig. S5). A phylogenetic tree reconstruction using the MOLPHY program[48] revealed that HHIPs and HHIP-like proteins with beta-propeller domains form a well-supported clade (Group 1 in Supporting Information Fig. S9). The phylogenetic tree also suggests that HHIPs could originate from a chordate-specific duplication of HHIP-like proteins, an event likely coupled with the addition of two C-terminal EGF-like domains in HHIPs (Supporting Information Fig. S9). Group 1 HHIP-like proteins with beta-propeller domains appear to be absent in insects and fungi, suggesting that they have been lost in these lineages.

HHIP-like proteins without beta-propeller domains form four major groups (Groups 2–5 in Supporting Information Fig. S9). Group 2 proteins, containing one HFN-CRD and sometimes a C-terminal predicted transmembrane segment, are from species of green plants, stramenopiles, and Alveolata. Group 3 consists of proteins from choanoflagellates, basal metazoan groups Cnidaria and Placozoa, and some deuterostomes such as Hemichordata and the amphioxus *Branchiostoma floridae*. Group 3 proteins are apparently lost in the vertebrate lineage. They are mostly single-domain proteins, and some members also possess a predicted transmembrane segment. Group 4 proteins are all from *N. gruberi*, likely resulted from gene expansion events. Group 5 proteins have a patchy phylogenetic distribution, with species from diverse sources such as *Leishmania*, choanoflagellates, and stramenopiles.

HFN-CRDs in FRBPs. Most of HFN-CRDs in FRBPs (Supporting Information Fig. S6) follow the cysteine pattern of "$C*C*CC*CX_8CX_3C*CX_3CX_6C*C*C$", with three residues in between the sixth and seventh conserved cysteines, instead of two residues in HHIPs and HHIP-like proteins. HFN-CRDs in FRBPs are mainly found in metazoans. Phylogenetic tree reconstruction suggested three major groups (Groups 1–3 in Supporting Information Fig. S10) of FRBPs. Group 1 consists of metazoan proteins that include the experimentally studied mammalian folate receptors. Noticeably, gene duplications have resulted in multiple folate receptors in mammals, including four copies in the human genome. Gene duplication events of Group 1 FRBPs also occurred in several other deuterostomes such as *Ciona intestinalis* and *B. floridae* (Supporting Information Fig. S10). On the other hand, Group 1 FRBP proteins have a limited distribution in ecdysozoans, as they appear to be absent from insects. Group 2 FRBPs consist of metazoan proteins that include the avian riboflavin-binding proteins[22,49] and the highly divergent mammalian retibindin proteins (long branches in Supporting Information Fig. S10) that are expressed in retina and with unknown function.[50] Group 3 FRBPs contain a couple of divergent proteins from lower organisms such as green algae, stramenopiles, and Rhizaria (based on the EST sequence gi: 48374354), as well as three sequences from *B. floridae*. Interestingly, these proteins, except the one from green algae, have circularly permuted HFN-CRDs (Supporting Information Fig. S6).

HFN-CRDs in NPC1s. The membrane-bound protein NPC1 is a key player in regulating subcellular cholesterol transportation from the lysosome to the plasma membrane and the endoplasmic reticulum (ER). Mutations of the human NPC1 gene have been associated with Niemann-Pick Type C disease. NPC1 has an N-terminal HFN-CRD domain responsible for binding cholesterol, and a Patched-like 12-transmembrane domain that serves to transport cholesterol molecules across the lysosomal membrane. Although four cysteines in the general HFN-CRD pattern ("$C*C*CC*CX_8CX_{2,3}C*\underline{C}X_3\underline{C}X_6C*\underline{C}*\underline{C}$") are substituted (underlined in the previous pattern) by noncysteine residues in NPC1 CRDs, they have evolved several extra disulfide bonds compared with other HFN-CRDs. NPC1 CRDs have a wider distribution in eukaryotes than the CRDs in HHIPs and FRBPs, with proteins and/or ESTs detected in metazoans, fungi, plants, and protists from various lineages such as Amoebozoa, Alveolata, stramenopiles, Heterolobosea, and Jakobida (Supporting Information Fig. S7). Such a wide distribution suggests that NPC1 CRD plays an important role in cholesterol regulation in most eukaryotes. We observed that some species, such as those of the *Plasmodium* genus, seem to be lacking of NPC1 CRDs, but still possess proteins with Patched-like transmembrane domains.

Phylogenetic analysis revealed two major groups of proteins with an NPC1 CRD (Group 1 and Group 2 in Supporting Information Fig. S11). Group 1 proteins, usually with more than 1000 amino acid residues, contain an NPC1 CRD fused with a Patched-like transmembrane domain. The majority of Group 1 proteins are from metazoans, fungi, and green plants. Lineage-specific gene duplications were observed in vertebrates, insects, and nematodes (Supporting Information Fig. S11). Human has an NPC1 paralog called NPC1L1 that plays a special role in intestinal uptake of cholesterol.[51,52] Group 2 proteins, containing only an NPC1 CRD and without a Patched-like domain, are from protists such as

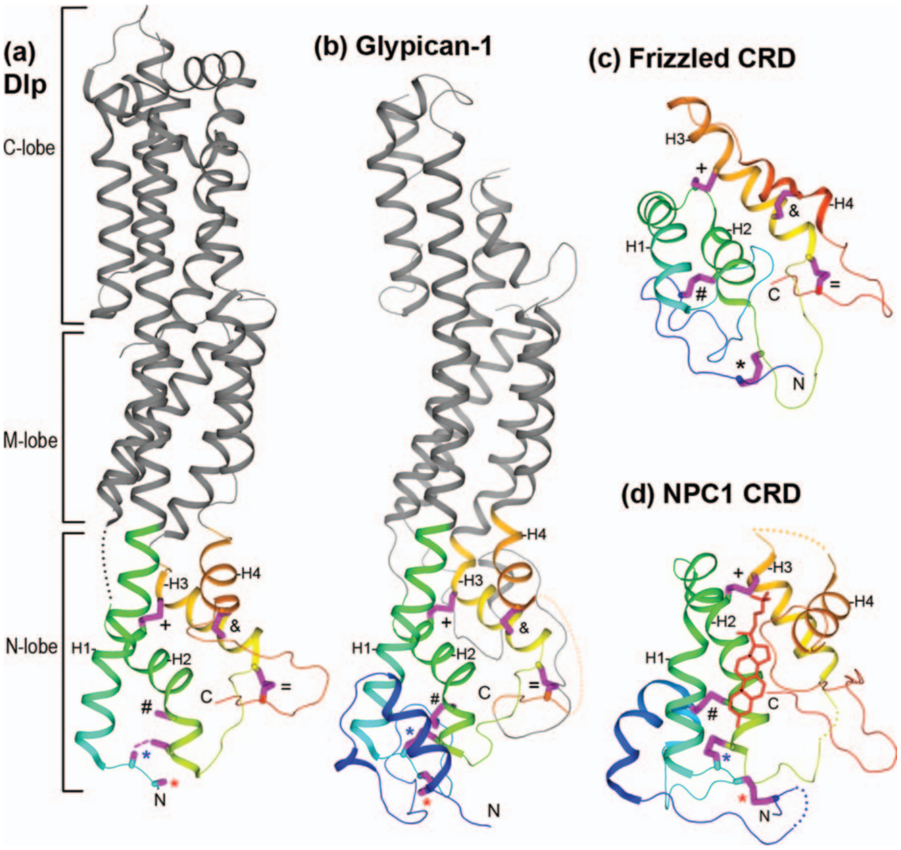

Figure 4. Structural diagrams of HFN-CRD and FZ-CRD domains. N- and C-termini are labeled for the CRDs. Four core alpha helices in these CRDs are labeled H1, H2, H3, and H4, respectively. Sidechains of conserved cysteines, most of which form disulfide bonds in the structures, are shown as sticks and colored magenta. They are labeled the same way as in Figure 1. (a) The glypican core region of *Drosophila* Dlp (PDB: 3odn). Three previously defined sub-domains of glypican are labeled as N-, M-, and C-lobes, respectively. The N-lobe has an HFN-CRD fold and is colored in rainbow from N- to C-terminus, while the M- and C-lobes, formed mainly by insertions from the N-lobe, are colored in grey. The N-terminal sequence segment before the ‘CC’ motif in the ‘C*C*CC*CX$_9$CX$_2$C*CX$_3$CX$_6$C*C*C’ pattern is disordered. Sidechains of the two cysteines in the ‘CC’ motif (C3 and C4) are also partially disordered. Therefore, the inferred disulfide bond between C4 and C7 (labeled by a blue *) are shown as a dashed line. (b) The glypican core region of human Glypican-1 (PDB: 4acr). One disordered region in the N-lobe (between the last two conserved cysteines) is displayed by a dotted connection. (c) The spatial structure of FZ-CRD (PDB: 1ijy) from the mouse protein Frizzled8. (d) The spatial structure of an HFN-CRD (PDB: 3gkj) from human NPC1. To aid the view of conserved disulfide bonds, some insertions to the NPC1 CRD are replaced by dotted connections. The bound cholesterol molecule is shown in red line. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Heterolobosea, stramenopiles, and Alveolata. Patched-like proteins without NPC1 CRDs are also present in these organisms. It is thus likely that two separate genes, encoding an NPC1 CRD and a Patched-like domain respectively, existed in the ancestor of eukaryotes, and their fusion has occurred in some lineages such as Opisthokonta and green plants. It should be noted that some stramenopiles species contain both Group 1 and Group 2 proteins (Supporting Information Fig. S11). They could have acquired Group 1 proteins (fusion of an NPC1 CRD and the Patched-like domain) via ancient endosymbiotic event(s).[53]

A novel HFN-CRD in glypicans

Glypicans are a family of heparan sulfate proteoglycans (HSPGs), proteins modified with heparan sulfate glycosaminoglycan (HS GAG) chains. They are attached to the outer leaflet of the plasma membrane by a C-terminal GPI anchor and play important roles in regulating the distribution and signaling of various morphogens including Hedgehog, Wnt, FGF, and BMP.[54,55] *Drosophila* and human have two (Dally and Dally-like) and six (glypican 1–6) copies of glypicans, respectively. Human glypican-3 is expressed in hepatocellular carcinomas and promotes their growth by stimulating Wnt signaling.[56] Different glypicans in *Drosophila* and mammals have been found to either inhibit or stimulate Hedgehog signaling.[57,58] A glypican molecule consists of an N-terminal globular core region and a C-terminal tail that is modified with HS GAGs. While HS GAGs may be responsible for binding of certain morphogens like FGF,[55] the core region of *Drosophila* glypican Dally-like (Dlp) has been shown to be essential for Hedgehog signaling.[58]

Recently the structure of the core region of *Drosophila* Dlp was reported.[59] It consists of mainly alpha helices with an overall cylinder-like shape and can be roughly divided into three portions (N-, M-, and C-lobes)[59] [Fig. 4(a)]. A more recent structure of the core region of human glypican-1 was obtained that shares a similar fold[60] [Fig. 4(b)]. However, the evolutionary origin of glypicans remains unclear despite the elucidation of glypican core structures. Structural and sequence comparisons by computational tools such as DaliLite and HHpred revealed that the N-lobe of Dlp bears similarities to HFN-CRDs and FZ-CRDs, suggesting that glypicans are evolutionarily related to them. A DaliLite structural comparison revealed distant structural similarity of Dlp (PDB: 3odn) to an NPC1 HFN-CRD structure (PDB: 3gkj) (*Z*-score: 3.2 and RMSD: 3.7 over 88 aligned positions) and an FZ-CRD structure (PDB: 1ijy) (*Z*-score: 3.2 and RMSD: 4.1 over 70 aligned positions). Such Dali *Z*-scores fall in the "grey area" of homology (between 2 and 8).[61] These structures exhibit a core of four alpha helices (labeled H1–H4 in Fig. 4) arranged in the same topology (left-handed connection between the first three core alpha helices and right-handed connection between the last three alpha helices). Noticeably, several disulfide bond-forming cysteine pairs are shared in these structures (Fig. 4), consistent with the sequence alignment (Fig. 1). The sequence similarity is also manifested in the HHpred search result using the human glypican-1 as the query, which identified the FZ-CRD domain (Pfam: PF01392) as a marginal hit (probability score: 71.6%) with several conserved cysteines aligned consistently with the DaliLite alignment.

Glypicans possess 14 conserved cysteines. Two of them form a disulfide bond in the C-lobe, while the remaining 12 cysteines all reside in the N-lobe and exhibit a pattern ("C*C*CC*CX$_9$CX$_2$C*CX$_3$CX$_6$ C*C*C") highly similar to that of HFN-CRDs from HHIPs (C*C*CC*CX$_8$CX$_2$C*CX$_3$CX$_6$C*C*C"). Like known HFN-CRDs, glypicans contain an extra "CC" motif compared with FZ-CRDs with 10 conserved cysteines. In the structure of Dlp [Fig. 4(a)], the N-terminal segment before the "CC" motif is disordered, and the sidechains of the two cysteines in the "CC" motif are partially disordered too. Still, the second cysteine in the "CC" motif is spatially close to, and likely forms a disulfide bond [marked by a blue star in Fig. 4(a)] with the seventh conserved cysteine in the general pattern of HFN-CRDs, the same way as seen in the NPC1 CRD structure. The more recent structure of human glypican-1[60] [Fig. 4(b)] indeed revealed that the cysteines in the substructure missing in Dlp exhibit the same disulfide connectivity as NPC1 CRD, following the general pattern of HFN-CRDs. Therefore, we classify the glypican N-lobe as an HFN-CRD.

The structural core of the HFN-CRD in glypicans (N-lobe) likely evolved two long insertions (each more than 100 amino acids, Figs. 1 and 4) between the first and second core alpha helices (H1 and H2), and between the third and fourth core alpha helices (H3 and H4), respectively. These long insertions mainly consist of alpha helices that together form the M- and C-lobes. Such long insertions to the glypican HFN-CRD domain probably hindered the detection of its sequence and structural similarities to other HFN-CRDs and FZ-CRDs. A site important for mediating Hedgehog signaling has been mapped to the C-lobe by mutagenesis studies,[59] suggesting that the insertion events are important to develop new functional sites. Wnt interaction site in glypicans has not been mapped. One hypothesis generated based on homology is that the N-lobe of glypican, like some other HFN-CRDs such as FRBPs and NPC1s, retains the ability to bind small molecules, and thus interacts with Wnts by binding to their lipid moieties.

Glypicans are only found in metazoans (including the basal Porifera group). Therefore, they could have evolved from a gene duplication of an HFN-CRD in the last common ancestor of metazoans. The inferred homology between glypicans and other HFN-CRDs illustrates the importance of gene duplications in creating molecular components in the transition from unicellular eukaryotes to multicellular animals. Other notable components likely resulted from gene duplication events in the Hedgehog signaling pathway included Smoothened (duplication of Frizzled in the Wnt pathway) and Patched and Dispatched (duplications of NPC1 in the cholesterol transport process). Glypicans constitute a major class of HPSGs and function in many cellular processes including modulation of various extracellular signals, such as Wnt and Hedgehog.[54,62] The discovery of an HFN-CRD in glypicans suggests diversified functions of the FZ-CRD/HFN-CRD module in the Wnt and Hedgehog pathways, which harbor several components with this module such as Frizzled, Smoothened, and HHIP. While the FZ-CRD in Frizzled has been found to be involved in the interactions with Wnt ligands, the functions of the FZ-CRD/HFN-CRD module in Smoothened, HHIP, and glypican molecules remain to be elucidated.

Materials and Methods

Sequence similarity searches

PSI-BLAST[19] iterations were conducted to search for homologs of Frizzled CRDs starting from three domains with known structures (protein databank (PDB) ids: 1ijy (mouse Frizzled 8), 1ijx (mouse secreted Frizzled-related protein 3), and 3hkl (rat MuSK)) against the NCBI nonredundant (nr) protein database (e-value inclusion cutoff: 1e-4). To perform transitive searches, PSI-BLAST hits were grouped by BLASTCLUST (with the score coverage threshold ($-S$, defined as the bit score divided by alignment length) set to 1, length coverage threshold ($-L$) set to 0.5, and no requirement of length coverage on both sequences ($-b$ F)), and a representative sequence from each group was used to initiate new PSI-BLAST searches. Such an iterative procedure was repeated until convergence. This procedure was also used for transitive PSI-BLAST searches for the Mid1 family (starting query—gi: 6324038 (Mid1 from *Saccharomyces cerevisiae*)), Frizzled-like CRDs in RECK proteins (starting query—gi: 11863156 from human RECK, range: 353–475), the glypican family (starting query—gi: 167001141 (human glypican-1)), and CRDs in Hedgehog-interacting proteins (staring query—gi: 20143973 (human HHIP), range: 20–220). HHpred[23] was used for profile-profile-based similarity searches to identify distant homologous relationships of FZ-CRDs and HFN-CRDs (profile databases used: Pfam,[25] SCOP,[63] and the eukaryotic proteome databases). To further investigate the distribution of FZ-CRDs and HFN-CRDs in lower organisms where protein records might be lacking, TBLASTN was used to search against nonhuman, nonmouse EST database in NCBI (est_others).

Domain architecture analysis

HMMER3[25] and HHpred were used to detect known Pfam domains (Pfam version: 25.0) in FZ-CRD and HFN-CRD-containing proteins with default parameter settings. Phobius[64] was used to predict transmembrane segments and membrane topology. N-terminal signal peptides were predicted by Phobius and SignalP 3.0.[65] GPI-SOM[28] and FragAnchor[29] were used for GPI anchor signal prediction.

Sequence alignment, clustering, and phylogenetic reconstruction

The multiple sequence alignment for select members of FZ-CRDs and HFN-CRDs was made by PROMALS3D[66] and improved by manual adjustment. Multiple sequence alignments for eight FZ-CRD and HFN-CRD groups were made by MAFFT[67] with some manual adjustment and shown in Supporting Information Figures S1–S8. For sequence clustering analysis, highly similar domains were filtered by CD-HIT[68] at the sequence identity cutoff of 90%. CLANS[35] was used to cluster and visualize the remaining nonredundant domains based on pairwise BLAST P values. CLANS is a tool for graph visualization of all-against-all similarities for a set of sequences. Sequences are represented as vertices in a CLANS graph. The length of an edge between two sequences is correlated with their similarity, which can be set as the logarithm of the BLAST hit P value between them. The sequences are first randomly placed in a two-dimensional or three-dimensional space. The vertices (sequences) are then moved iteratively, based on the attractive forces designed to be proportional to the sequence similarities and a small repulsive force that serves to prevent collapse of sequences into one point.[35] Such a process can be run to equilibrium where movements of the vertices are negligible. For clustering and visualization of FZ-CRDs and HFN-CRDs, the movement process was run to equilibrium in a two-dimensional representation. The BLAST P value cutoff was set to 1e-4 to show the connections between the sequences in the final graph.

The MOLPHY package[48] was used for phylogenetic reconstruction for three known HFN-CRD groups (HHIPs, FRBPs, and NPC1s). The JTT amino acid substitution model[69] was used in MOLPHY. The local estimates of bootstrap percentages were obtained by the RELL method[70] (-R option in the ProtML program of MOLPHY).

Acknowledgment

The authors thank Lisa Kinch for critical reading of the manuscript.

References

1. Logan CY, Nusse R (2004) The Wnt signaling pathway in development and disease. Annu Rev Cell Dev Biol 20:781–810.
2. Jiang J, Hui CC (2008) Hedgehog signaling in development and cancer. Dev Cell 15:801–812.
3. Mikels AJ, Nusse R (2006) Wnts as ligands: processing, secretion and reception. Oncogene 25:7461–7468.
4. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63:1256–1272.
5. Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, Richards GS, Zwafink C, Degnan BM (2010) Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. Evol Dev 12:494–518.
6. Adamska M, Matus DQ, Adamski M, Green K, Rokhsar DS, Martindale MQ, Degnan BM (2007) The evolutionary origin of Hedgehog proteins. Curr Biol 17:R836–R837.
7. Ingham PW, Nakano Y, Seger C (2011) Mechanisms and functions of Hedgehog signalling across the metazoa. Nat Rev Genet 12:393–406.
8. Dann CE, Hsieh JC, Rattner A, Sharma D, Nathans J, Leahy DJ (2001) Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. Nature 412:86–90.
9. Stiegler AL, Burden SJ, Hubbard SR (2009) Crystal structure of the Frizzled-like cysteine-rich domain of the receptor tyrosine kinase MuSK. J Mol Biol 393:1–9.
10. Bazan JF, de Sauvage FJ (2009) Structural ties between cholesterol transport and morphogen signaling. Cell 138:1055–1056.
11. Huang HC, Klein PS (2004) The Frizzled family: receptors for multiple signal transduction pathways. Genome Biol 5:234.
12. Prabhu Y, Eichinger L (2006) The Dictyostelium repertoire of seven transmembrane domain receptors. Eur J Cell Biol 85:937–946.

13. Kawano Y, Kypta R (2003) Secreted antagonists of the Wnt signalling pathway. J Cell Sci 116:2627–2634.
14. Green JL, Kuntz SG, Sternberg PW (2008) Ror receptor tyrosine kinases: orphans no more. Trends Cell Biol 18:536–544.
15. Ghazanfari N, Fernandez KJ, Murata Y, Morsch M, Ngo ST, Reddel SW, Noakes PG, Phillips WD (2011) Muscle specific kinase: organiser of synaptic membrane domains. Int J Biochem Cell Biol 43:295–298.
16. Song L, Fricker LD (1997) Cloning and expression of human carboxypeptidase Z, a novel metallocarboxypeptidase. J Biol Chem 272:10543–10550.
17. Yan W, Sheng N, Seto M, Morser J, Wu Q (1999) Corin, a mosaic transmembrane serine protease encoded by a novel cDNA from human heart. J Biol Chem 274:14926–14935.
18. Seppinen L, Pihlajaniemi T (2011) The multiple functions of collagen XVIII in development and disease. Matrix Biol 30:83–92.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.
20. Bosanac I, Maun HR, Scales SJ, Wen X, Lingel A, Bazan JF, de Sauvage FJ, Hymowitz SG, Lazarus RA (2009) The structure of SHH in complex with HHIP reveals a recognition role for the Shh pseudo active site in signaling. Nat Struct Mol Biol 16:691–697.
21. Kwon HJ, Abi-Mosleh L, Wang ML, Deisenhofer J, Goldstein JL, Brown MS, Infante RE (2009) Structure of N-terminal domain of NPC1 reveals distinct subdomains for binding and transfer of cholesterol. Cell 137: 1213–1224.
22. Monaco HL (1997) Crystal structure of chicken riboflavin-binding protein. EMBO J 16:1475–1483.
23. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.
24. Santos AL, Branquinha MH, D'Avila-Levy CM (2006) The ubiquitous gp63-like metalloprotease from lower trypanosomatids: in the search for a function. An Acad Bras Cienc 78:687–714.
25. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38:D211–D222.
26. Etges R, Bouvier J, Bordier C (1986) The major surface protein of Leishmania promastigotes is a protease. J Biol Chem 261:9098–9101.
27. Cobbe N, Marshall KM, Gururaja Rao S, Chang CW, Di Cara F, Duca E, Vass S, Kassan A, Heck MM (2009) The conserved metalloprotease invadolysin localizes to the surface of lipid droplets. J Cell Sci 122:3414–3423.
28. Fankhauser N, Maser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 21:1846–1852.
29. Poisson G, Chauve C, Chen X, Bergeron A (2007) FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. Genomics Proteomics Bioinformatics 5:121–130.
30. Cavalier-Smith T (2009) Megaphylogeny, cell body plans, adaptive zones: causes and timing of eukaryote basal radiations. J Eukaryot Microbiol 56:26–33.
31. Shashoua VE (1991) Ependymin, a brain extracellular glycoprotein, and CNS plasticity. Ann NY Acad Sci 627: 94–114.
32. Suarez-Castillo EC, Garcia-Arraras JE (2007) Molecular evolution of the ependymin protein family: a necessary update. BMC Evol Biol 7:23.
33. Moy GW, Mendoza LM, Schulz JR, Swanson WJ, Glabe CG, Vacquier VD (1996) The sea urchin sperm receptor for egg jelly is a modular protein with extensive homology to the human polycystic kidney disease protein, PKD1. J Cell Biol 133:809–817.
34. Schroder S, Fraternali F, Quan X, Scott D, Qian F, Pfuhl M (2011) When a module is not a domain: the case of the REJ module and the redefinition of the architecture of polycystin-1. Biochem J 435:651–660.
35. Frickey T, Lupas A (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20:3702–3704.
36. Fischer M, Schnell N, Chattaway J, Davies P, Dixon G, Sanders D (1997) The Saccharomyces cerevisiae CCH1 gene is involved in calcium influx and mating. FEBS Lett 419:259–262.
37. Iida H, Nakamura H, Ono T, Okumura MS, Anraku Y (1994) MID1, a novel *Saccharomyces cerevisiae* gene encoding a plasma membrane protein, is required for Ca2+ influx and mating. Mol Cell Biol 14:8259–8271.
38. Kanzaki M, Nagasawa M, Kojima I, Sato C, Naruse K, Sokabe M, Iida H (1999) Molecular identification of a eukaryotic, stretch-activated nonselective cation channel. Science 285:882–886.
39. Maruoka T, Nagasoe Y, Inoue S, Mori Y, Goto J, Ikeda M, Iida H (2002) Essential hydrophilic carboxyl-terminal regions including cysteine residues of the yeast stretch-activated calcium-permeable channel Mid1. J Biol Chem 277:11645–11652.
40. Takahashi C, Sheng Z, Horan TP, Kitayama H, Maki M, Hitomi K, Kitaura Y, Takai S, Sasahara RM, Horimoto A, Ikawa Y, Ratzkin BJ, Arakawa T, Noda M (1998) Regulation of matrix metalloproteinase-9 and inhibition of tumor invasion by the membrane-anchored glycoprotein RECK. Proc Natl Acad Sci USA 95: 13221–13226.
41. Oh J, Takahashi R, Kondo S, Mizoguchi A, Adachi E, Sasahara RM, Nishimura S, Imamura Y, Kitayama H, Alexander DB, Ide C, Horan TP, Arakawa T, Yoshida H, Nishikawa S, Itoh Y, Seiki M, Itohara S, Takahashi C, Noda M (2001) The membrane-anchored MMP inhibitor RECK is a key regulator of extracellular matrix integrity and angiogenesis. Cell 107:789–800.
42. Clark JC, Thomas DM, Choong PF, Dass CR (2007) RECK--a newly discovered inhibitor of metastasis with prognostic significance in multiple forms of cancer. Cancer Metastasis Rev 26:675–683.
43. Chang CK, Hung WC, Chang HC (2008) The Kazal motifs of RECK protein inhibit MMP-9 secretion and activity and reduce metastasis of lung cancer cells in vitro and in vivo. J Cell Mol Med 12:2781–2789.
44. Harrington AE, Morris-Triggs SA, Ruotolo BT, Robinson CV, Ohnuma S, Hyvonen M (2006) Structural basis for the inhibition of activin signalling by follistatin. EMBO J 25:1035–1045.
45. Phelan MM, Thai CT, Soares DC, Ogata RT, Barlow PN, Bramham J (2009) Solution structure of factor I-like modules from complement C7 reveals a pair of follistatin domains in compact pseudosymmetric arrangement. J Biol Chem 284:19637–19649.
46. Zhang JL, Huang Y, Qiu LY, Nickel J, Sebald W (2007) von Willebrand factor type C domain-containing proteins regulate bone morphogenetic protein signaling through different recognition mechanisms. J Biol Chem 282:20002–20014.
47. Bishop B, Aricescu AR, Harlos K, O'Callaghan CA, Jones EY, Siebold C (2009) Structural insights into Hedgehog ligand sequestration by the human

Hedgehog-interacting protein HHIP. Nat Struct Mol Biol 16:698–703.

48. Adachi J, Hasegawa M (1996) MOLPHY version 2.3, programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr 28:1–150.
49. White HB, 3rd, Merrill AH, Jr (1988) Riboflavin-binding proteins. Annu Rev Nutr 8:279–299.
50. Wistow G, Bernstein SL, Wyatt MK, Ray S, Behal A, Touchman JW, Bouffard G, Smith D, Peterson K (2002) Expressed sequence tag analysis of human retina for the NEIBank Project: retbindin, an abundant, novel retinal cDNA and alternative splicing of other retina-preferred gene transcripts. Mol Vis 8:196–204.
51. Altmann SW, Davis HR, Jr, Zhu LJ, Yao X, Hoos LM, Tetzloff G, Iyer SP, Maguire M, Golovko A, Zeng M, Wang L, Murgolo N, Graziano MP (2004) Niemann-Pick C1 Like 1 protein is critical for intestinal cholesterol absorption. Science 303:1201–1204.
52. Kwon HJ, Palnitkar M, Deisenhofer J (2011) The structure of the NPC1L1 N-terminal domain in a closed conformation. PLoS One 6:e18722.
53. Dorrell RG, Smith AG (2011) Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. Eukaryot Cell 10:856–868.
54. Filmus J, Capurro M, Rast J (2008) Glypicans. Genome Biol 9:224.
55. Lin X (2004) Functions of heparan sulfate proteoglycans in cell signaling during development. Development 131:6009–6021.
56. Capurro MI, Xiang YY, Lobe C, Filmus J (2005) Glypican-3 promotes the growth of hepatocellular carcinoma by stimulating canonical Wnt signaling. Cancer Res 65: 6245–6254.
57. Capurro MI, Xu P, Shi W, Li F, Jia A, Filmus J (2008) Glypican-3 inhibits Hedgehog signaling during development by competing with patched for Hedgehog binding. Dev Cell 14:700–711.
58. Williams EH, Pappano WN, Saunders AM, Kim MS, Leahy DJ, Beachy PA (2010) Dally-like core protein and its mammalian homologues mediate stimulatory and inhibitory effects on Hedgehog signal response. Proc Natl Acad Sci USA 107:5869–5874.
59. Kim MS, Saunders AM, Hamaoka BY, Beachy PA, Leahy DJ (2011) Structure of the protein core of the glypican Dally-like and localization of a region important for Hedgehog signaling. Proc Natl Acad Sci USA 108:13112–13117.
60. Svensson G, Awad W, Hakansson M, Mani K, Logan DT (2012) Crystal structure of *N*-glycosylated human glypican-1 core protein: structure of two loops evolutionarily conserved in vertebrate glypican-1. J Biol Chem 287:14040–14051
61. Holm L, Kaariainen S, Wilton C, Plewczynski D (2006) Using Dali for structural comparison of proteins. Curr Protoc Bioinformatics Chapter 5:Unit 5.5.
62. Fico A, Maina F, Dono R (2011) Fine-tuning of cell signaling by glypicans. Cell Mol Life Sci 68:923–929.
63. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
64. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35:W429–W432.
65. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971.
66. Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295–2300.
67. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.
68. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.
69. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282.
70. Hasegawa M, Kishino H, Saitou N (1991) On the maximum likelihood method in molecular phylogenetics. J Mol Evol 32:443–445.