

## Appearance of new tetraspanin genes during vertebrate evolution

Antonio Garcia-España<sup>a,\*</sup>, Pei-Jung Chung<sup>b</sup>, Indra Neil Sarkar<sup>c</sup>, Eric Stiner<sup>d</sup>,  
Tung-Tien Sun<sup>b,e,f,g,\*</sup>, Rob DeSalle<sup>d,\*</sup>

<sup>a</sup> *Research Unit, Universitat Rovira i Virgili, Hospital Joan XXIII, Pere Virgili Institute, Tarragona 43007, Spain*

<sup>b</sup> *Department of Cell Biology, New York University School of Medicine, New York, NY 10016, USA*

<sup>c</sup> *Marine Biological Laboratory, Woods Hole, MA 02543, USA*

<sup>d</sup> *American Museum of Natural History, New York, NY 10024, USA*

<sup>e</sup> *Department of Dermatology, New York University School of Medicine, New York, NY 10016, USA*

<sup>f</sup> *Department of Pharmacology, New York University School of Medicine, New York, NY 10016, USA*

<sup>g</sup> *Department of Urology, New York University School of Medicine, New York, NY 10016, USA*

Received 27 March 2007; accepted 17 December 2007

Available online 21 February 2008

### Abstract

A detailed phylogenetic analysis of tetraspanins from 10 fully sequenced metazoan genomes and several fungal and protist genomes gives insight into their evolutionary origins and organization. Our analysis suggests that the superfamily can be divided into four large families. These four families—the CD family, CD63 family, uroplakin family, and RDS family—are further classified as consisting of several ortholog groups. The clustering of several ortholog groups together, such as the CD9/Tsp2/CD81 cluster, suggests functional relatedness of those ortholog groups. The fact that our studies are based on whole genome analysis enabled us to estimate not only the phylogenetic relationships among the tetraspanins, but also the first appearance in the tree of life of certain tetraspanin ortholog groups. Taken together, our data suggest that the tetraspanins are derived from a single (or a few) ancestral gene(s) through sequence divergence, rather than convergence, and that the majority of tetraspanins found in the human genome are vertebrate (21 instances), tetrapod (4 instances), or mammalian (6 instances) inventions.

© 2008 Published by Elsevier Inc.

**Keywords:** Tetraspanins; Evolution; Gene family; Phylogenetics

Tetraspanins (tetraspan or TM4SF) form a large group of integral membrane proteins that we will call a superfamily [1,2]. This superfamily has as many as 33 members in humans. Human tetraspanins are widely distributed in cells and tissues and have homologs conserved through distantly related eukaryotic species. Structurally, tetraspanins are 200- to 300-amino-acid-long proteins with four transmembrane (TM) domains, which delimit one small extracellular loop of 13–30 amino acids, a short intracellular sequence, and a second, large extracellular loop, which is quite variable in sequence and length. Two highly conserved features of

tetraspanin proteins are (i) their second loop harbors a Cys-Cys-Gly sequence (the CCG motif) plus 2 to 6 additional cysteines and (ii) their four TM domains contain some well-conserved residues. These features have been used to distinguish the tetraspanins from other four-transmembranous proteins [1,3–6].

Many tetraspanin proteins had originally been identified as human tumor antigens; in some cases their expression correlates with tumor progression [5,8]. In humans, several forms of retinal degeneration are caused by mutations in the gene encoding peripherin/RDS [9], and mental retardation syndromes have been linked to defects in Tetraspanin 7 (TSPAN7; TM4SF2), [10]. Members of the tetraspanin superfamily can form large integrated signaling complexes or tetraspanin-enriched microdomains by their primary associations with a variety of transmembrane and intracellular signaling/cytoskeletal proteins and secondary associations with themselves [4,11,12]. Tetraspanins participate in a broad

\* Corresponding authors.

*E-mail addresses:* [agarciae.hj23.ics@gencat.net](mailto:agarciae.hj23.ics@gencat.net) (A. Garcia-España), [sunt01@med.nyu.edu](mailto:sunt01@med.nyu.edu) (T.-T. Sun), [desalle@amnh.org](mailto:desalle@amnh.org) (R. DeSalle).

spectrum of membrane-associated cellular activities such as cell adhesion, motility, activation of signaling pathways, facilitation of membrane protein maturation, and cell proliferation. This participation occurs in normal and in pathological conditions such as cancer metastasis or infections by viral, bacterial, or parasitic organisms [7,11,13–23]. Nevertheless, despite the implication of their role in this broad spectrum of important cellular activities, only a relatively small number of the tetraspanins have been studied in detail.

Some specific tetraspanin functions have been described across broad evolutionary divergences. Examples include the PLS1 tetraspanin, which enables the plant pathogenic fungus *Magnaporthe* to invade its rice host's leaves [24]; the LBM tetraspanin, whose mutations cause synaptic defects in *Drosophila*; the CD9 and CD81 tetraspanins, which are involved in mammalian sperm: oocyte fusion [18,22]; CD81, which is involved in immune signaling [25]; peripherin/RDS, which scaffolds vertebrate photoreceptor outer segment structure [26]; and uroplakins, in the maintenance of the urothelial permeability barrier [27–29].

To gain a more complete understanding of tetraspanin biology, we have examined the evolutionary history of members of the tetraspanin superfamily through genomic analysis. There are currently three main approaches to the analysis of large gene families in a phylogenetic context. The first involves searching the database and including every accession with reasonable BLAST or BLAT hit statistics. This is the approach of a recent study on tetraspanins by Huang et al. [30], in which over 200 tetraspanins were included in a distance-based analysis. While these kinds of studies are important in defining the phylogenetic structure of the tetraspanin superfamily, no definitive statements can be made about the *absence* of superfamily members in particular taxa. Information about not only the presence, but also the absence, of gene family members is critical for understanding the phylogenetic classification of gene family members and for understanding the origin of new gene family members. The second approach is to focus on a gene family of a specific taxonomic group such as insect tetraspanins [2]. This approach, while more feasible than more inclusive analyses, does not address broader evolutionary questions about a gene family. A third approach, which we take here, is to analyze a gene family *in a group of fully sequenced and carefully annotated genomes*. This approach has been used successfully to build, e.g., an ortholog identification Web tool for plants [31].

In the present study, as an expansion of our recent analysis of uroplakin tetraspanins [32], we analyzed the complete genomes of nine representative animal species, two plant species, seven fungi, and several other single-celled eukaryotic organisms to define the superfamily composition of tetraspanins and the evolutionary origin of the various superfamily members. Restricting the phylogenetic analysis of the tetraspanin superfamily to only well-annotated complete genomes allowed us to expand the interpretation of the distribution of genes in this superfamily across eukaryotes. This approach also allowed us to test hypotheses about duplications and losses of gene members in the superfamily as well as setting upper limits on divergence times of members of gene families and hence the origination of new tetraspanin genes.

## Results and discussion

### *Definition of the tetraspanin superfamily*

BLAST searches using several prototypic tetraspanins, including CD81, Tsp10 (oculospanin), and other tetraspanin protein sequences, as query sequences yielded many hits (mostly mammalian) with very low *E* values, indicating definite inclusion of these “hits” in the tetraspanin superfamily. We also obtained many other hits with BLAST values larger than *E* -5. The alignment of such proteins with tetraspanins was restricted mainly to regions of highly conserved residues such as the CCG domain of the protein. The tetraspanin-like genes from fungal, protist, plant, *Caenorhabditis elegans*, and *Drosophila melanogaster* showed such low similarity with their mammalian counterparts that we could not state with confidence whether they should be regarded as tetraspanins based on their *E* values alone. To determine whether these nonmammalian sequences should be included in the tetraspanin superfamily, we analyzed the conservation of the intron-exon junctions, as well as their hydrophobicity profiles compared with well-established tetraspanin proteins. The observation in this study and in that of Huang et al. [30], that intron-exon positions appear to be conserved in many tetraspanin genes, validated the inclusion of these nonmammalian proteins from protists, plants, and fungi as divergent invertebrate tetraspanins. Once we determined the validity of inclusion of the highly divergent tetraspanins in the analysis, we aligned all 268 proteins from genomes as described above. Ambiguous regions in the alignment were then trimmed away from the matrix as in Huang et al. [30], resulting in a data matrix with 202 amino acid and gap characters for each protein. The sequences that remained after the trimming process were almost entirely in the four membrane-spanning regions and the second, large extracellular loop region.

### *Tetraspanin superfamily contains four major families*

Phylogenetic methods can aid in defining the membership of many of the tetraspanin ortholog groups. To facilitate a broader understanding of the tetraspanins, we suggest that this large superfamily of transmembrane proteins be classified into four major families—the CD family, the CD63 family, the uroplakin family, and the RDS family [42]. The tree topology obtained after phylogenetic analysis using parsimony with equal weighting is shown in Fig. 1 (see also Table 1).

The three different types of tree building approaches (maximum parsimony or MP, Bayes, and neighbor joining or NJ) yielded trees with many of the same major groupings of tetraspanins as those of Huang et al. [30]. Some of the relationships of major groups to one another are also constant from the Bayes and NJ analysis to the MP analysis. For purposes of clarity, we discuss only the MP trees in this communication. The tree topology obtained after phylogenetic analysis using parsimony with equal weighting is shown in Fig. 1. MP and NJ analysis with the Fitch weighting matrix also yielded similar trees. Jackknife and bootstrap analyses indicated a lack of robustness of all nodes at the base of the MP trees (both equal weighting and Fitch weighting) as well as the NJ and Bayes (see dotted line in Fig. 1). The lack of

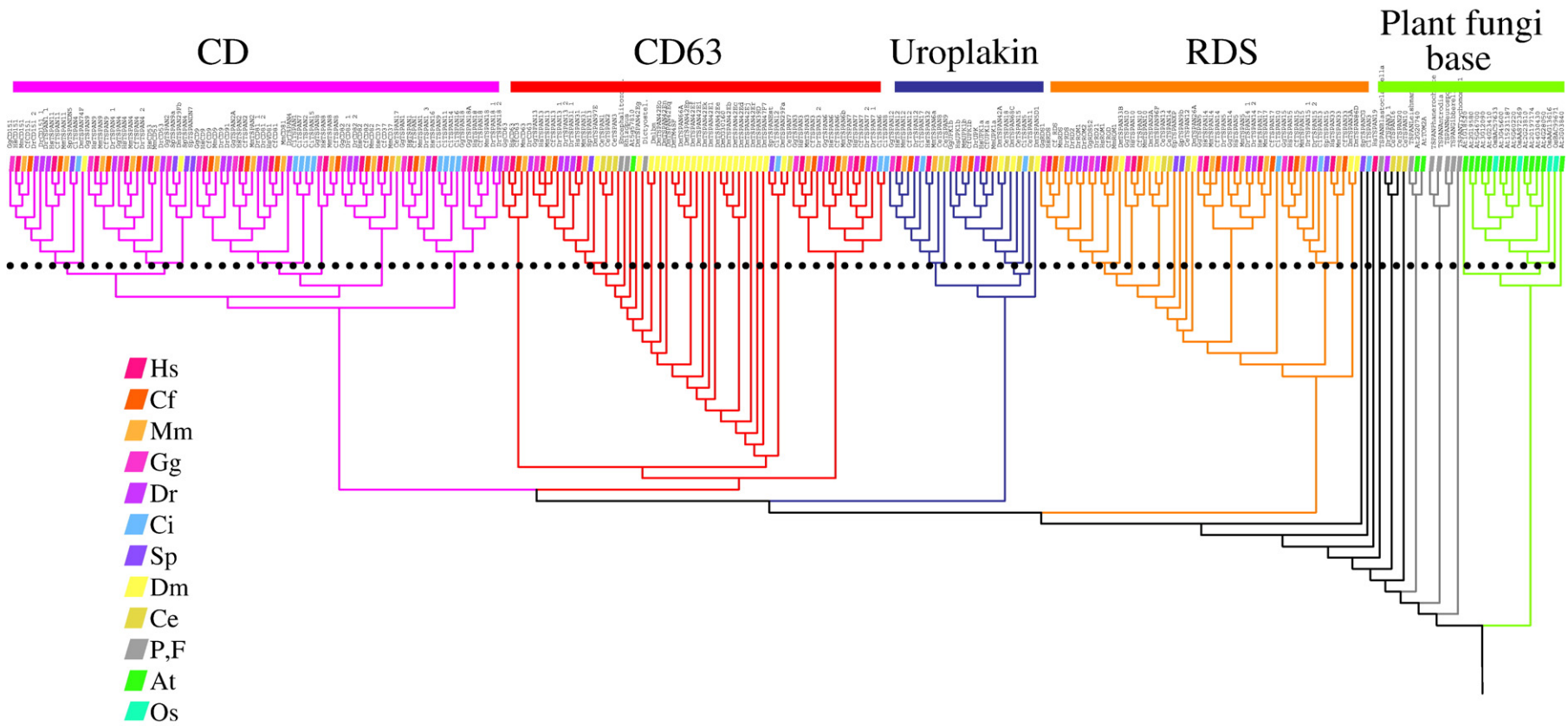


Fig. 1. Results of phylogenetic analysis using the MP approach with the amino acid characters weighted according to the genetic identity approach. Species are designated by colored boxes with a legend for the species designation given (species abbreviations are as in Table 1). More detailed “close-ups” of the four major groups of tetraspanins designated here are available in Supplemental Figs. 1-1 through 1-5. We suggest that the tetraspanin superfamily can be subdivided into four major monophyletic subfamilies (the CD family, the CD63 family, the uroplakin family, and the RDS family) and a group of nonmonophyletic tetraspanins at the base of the tree that comprises fungal, plant, and protist tetraspanins. The black dotted line represents the general area of the tree below which bootstrap and jackknife values drop below 60% and Bayes proportions below 90%.

Table 1  
Tetraspanins from whole genomes used in this study

Scientific name	Abbreviation	No. of Tsp's	Common name
<i>Homo sapiens</i>	Hs	33	Human
<i>Canis familiaris</i>	Cf	32	Dog
<i>Mus musculus</i>	Mm	34	Mouse
<i>Gallus gallus</i>	Gg	30	Chicken
<i>Danio rerio</i>	Dr	40	Zebrafish
<i>Ciona intestinalis</i>	Ci	17	Sea squirt
<i>Strongylocentrotus purpuratus</i>	Sp	10	Sea urchin
<i>Drosophila melanogaster</i>	Dm	35	Fly
<i>Caenorhabditis elegans</i>	Ce	20	Nematode
<i>Rhizopus oryzae</i>	Rhizopus	1	Fungi
<i>Encephalitozoon cuniculi</i>	Encephalit	1	Fungi
<i>Neurospora crassa</i>	Neurospora	1	Bread mold
<i>Gibberella zeae</i>	Gibberella	1	Fungi
<i>Blastocladiella emersonii</i>	Blastocladiella	1	Fungi
<i>Phanerochaete chrysosporium</i>	Phanerochaete	1	Fungi
<i>Antrodia cinnamomea</i>	Antrodia	1	Fungi
<i>Arabidopsis thaliana</i>	At	16	Thalecress
<i>Oryza sativa</i>	Os	4	Rice

robust inferences using resampling techniques (jackknife and bootstrap) and lack of robustness at the base of the tree comparing the topologies of the MP, Bayes, and NJ analyses are most likely due to the small number of amino acid characters used in the analysis. Despite a lack of robustness at these nodes, the successive weighting procedure resulted in 18 optimal parsimony trees (not shown) with strong consistency. Fig. 1 shows a strict consensus tree of these 18 successively weighted parsimony trees.

Our phylogenetic analyses revealed four major clades (called the CD family, the CD63 family, the uroplakin family, and the RDS family; Fig. 1). In addition, there are several unattached *Drosophila* and *Caenorhabditis* tetraspanins at the base of the tree. One significant result is that the fungal and a few non-fungal single-celled eukaryotes are observed as the most basal nonplant tetraspanins (Fig. 1 and Supplemental Figs. 1 to 5). One might expect an amoeboid species' tetraspanins in our analysis also to be at the base of the tree, but the *Dictyostelium* tetraspanins are found in the CD63 clade. This result is most likely caused by the long branches for these three single-celled eukaryotes that should be found at the base of the tree. The CD63 family also has long branches and is extremely divergent and the placement of *Dictyostelium*, *Rhizopus*, and *Encephalitozoon* near these tetraspanins is most likely the product of long-branch attraction (see above).

The largest cluster of tetraspanins, which we have designated the CD family, comprises proteins previously annotated as vertebrate CD and Tsp proteins with several invertebrate tetraspanins (Supplemental Fig. S1). All of the CD tetraspanins except for CD63 are included in this first large cluster in agreement with [1]. Yet another large cluster is one we have designated the CD63 family. This family contains the CD63 orthologs from several vertebrates as well as the well-known set of genes at chromosome location 42E in the *Drosophila* genome ([42]; Supplemental Fig. S2). This large cluster of tetraspanins is highly divergent and it also contains several vertebrate TSPAN proteins (CD63, TSPAN13, TSPAN31, TSPAN3, TSPAN6, and TSPAN7). This grouping is consistent with the earlier association of CD63

with these other tetraspanins by Maeker et al. [1]. Another major cluster contains the uroplakin proteins (Supplemental Fig. S3). Several *Drosophila* and *Caenorhabditis* and one *Ciona* tetraspanin are included as close relatives of the uroplakins with vertebrate TSPAN12 and TSPAN32 also being included in this family of tetraspanins. The fourth large cluster of tetraspanins with several previously annotated Tsp tetraspanins is designated the RDS family, because it contains the RDS-ROM tetraspanins (Supplemental Fig. S4).

#### Tetraspanin ortholog groups

While the trees we generated have a low robustness at their bases, several common relationships can be seen with the different tree building approaches we used. The different weighting schemes using parsimony (equal weights and Fitch weighting) and the Bayes and NJ approaches have several points of agreement at the level of bootstrap and jackknife robustness. As stated above, both Bayes and NJ analyses agree with MP with respect to the grouping of the major kinds of tetraspanins. For instance, regardless of weighting scheme, CD151's, CD53's, CD9's, CD81's, CD82's, CD37's, CD63's, and uroplakins are all supported strongly as ortholog groups as assessed by Bayesian analysis, bootstrapping, and jackknifing. For the most part, the overall phylogenetic hypothesis is congruent within these well-defined ortholog groups and also between some clusters of ortholog groups.

Table 2  
Annotating *Ciona* (Ci) and *Strongylocentrotus* (Sp) tetraspanins

Name used here	Accession No.	Assigned to ortholog group(s)
<i>Ciona</i>		
CiTSPAN1	ENSCINP00000012953	CD9, Tsp2, CD81
CiTSPAN2	ENSCINP00000012935	CD9, Tsp2, CD81
CiTSPAN4	ESTs; Cin40100141511	CD151, Tsp11
CiTSPAN5	ENSCINP00000010687	Upk
CiTSPAN6	ENSCINP00000015991	Tsp6, Tsp7
CiTSPAN7	ENSCINP00000012956	CD9, Tsp2, CD81
CiTSPAN9	ENSCINP00000003717	Tsp1
CiTSPAN11	ENSCINP00000021588	Tsp1
CiTSPAN12	ENSCINP00000025966	CD63
CiTSPAN13	AABS01000088.1	Tsp15
CiTSPAN14	ENSCINP00000003743	Tsp1
CiTSPAN15	ENSCINP00000026559	CD9, Tsp2, CD81
CiTSPAN16	ENSCINP00000025843	Tsp1
CiTSPAN17	ENSCINP00000015326	Tsp12
<i>Strongylocentrotus</i>		
SpTSPAN15	XP_794304.2	Tsp16
SpTSPANDN5	XP_794023.2	CD151, Tsp11
SpTSPANXP	XP_787025.1	ROM, RDS
SpTSPAN9b	XP_787272.1	ROM, RDS
SpTSPAN13	XP_783097.2	Tsp13, Tsp31
SpTSPANEST	XP_800780.2	CD63

During the sequence download of the Ci and Sp tetraspanins, we noticed that some of the tetraspanins from these species had incomplete annotations that did not coincide with the existing mammalian and other vertebrate annotations. To organize the tetraspanins from these two species we used the original annotated gene names and assigned them to the ortholog groups that the tetraspanins from these species belong to based on the ortholog grouping of mammalian tetraspanins. As well, we include the original accession numbers of the tetraspanins.

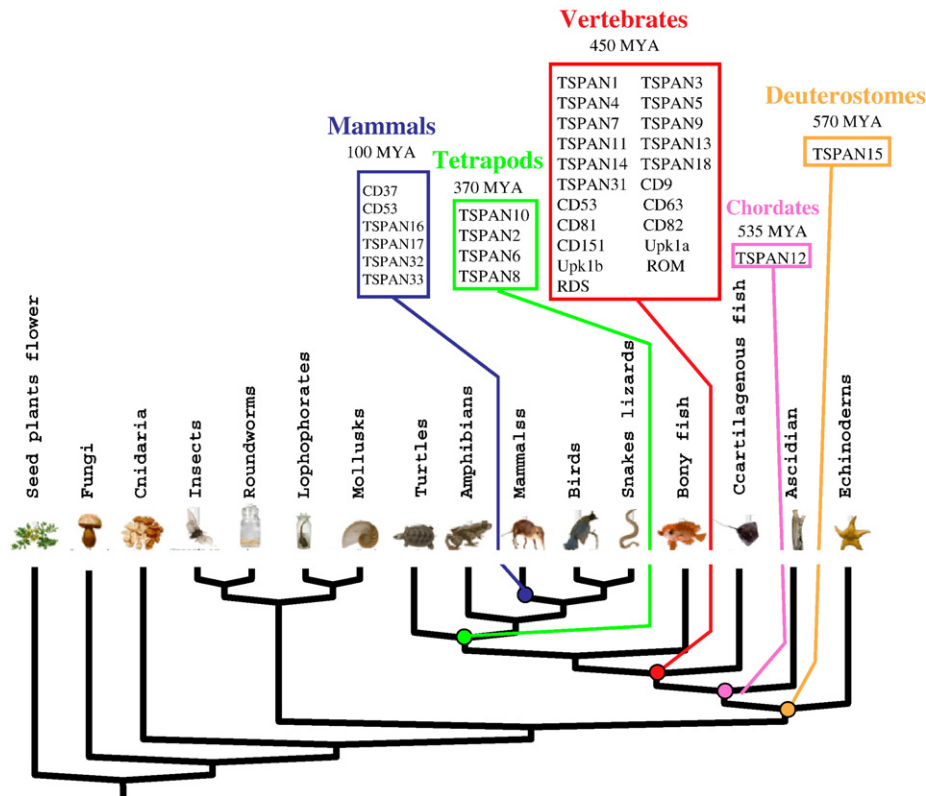


Fig. 2. Ancestral origin of human tetraspanins. The tree shows five ancestral points of origin (mammals, tetrapods, vertebrates, chordates, and deuterostomes) for the human tetraspanins. Different colors indicate different points of origin. The phylogenetic tree is based on our best recent understanding of relationships of major taxonomic groups and the ages of groups are explained in the text.

### Rogue tetraspanins

There are three anomalous groups of “rogue” (not clearly associated with other groups) tetraspanins in the tree (designated by the numbers 1, 2, and 3 in Supplemental Figs. S1, S3, and S4, respectively). Rogue group 1 is at the base of the combination of CD151, CD53, TSPAN11, TSPAN9, and TSPAN4 and indicates an animal origin for this group of four tetraspanins. Rogue 2 is at the base of the combination of TSPAN13, TSPAN31, and the CD63 family, which might result from long-branch attraction (Supplemental Fig. S2). The four tetraspanins in the smaller unattached cluster (one each from *Strongylocentrotus*, *Ciona*, *Caenorhabditis*, and *Drosophila*) might also be grouped as a result of long-branch attraction. Rogue 3 is at the base of a group of tetraspanins including RDS, ROM, TSPAN10, and *Drosophila* nonexpansion tetraspanins DmTs2a and TSPAN96f (Supplemental Fig. S4). This group of rogue tetraspanins probably indicates a bilateral animal origin for the RDS, ROM, and TSPAN10 supergroup of tetraspanins.

### Comparison with Treefam

We also compared this tree structure with the TreeFam (<http://www.treefam.org/>) organization of these genes. TreeFam lists six separate families, TSPAN32, uroplakins, ROM, TSPAN31/TSPAN13, and two mixed families. Mixed family 1 comprises human TSPAN14, TSPAN15, TSPAN5, TSPAN10, and TSPAN17

and orthologs from other organisms and mixed family 2 comprises CD151, TSPAN6, TSPAN18, TSPAN7, CD82, TSPAN3, TSPAN8, CD63, TSPAN2, CD81, TSPAN1, TSPAN12, TSPAN16, TSPAN4, CD53, CD9, and TSPAN9 and orthologs in other organisms. As we show here these families coincide broadly with the families we have designated below except that we sink the TSPAN32 and TSPAN31/TSPAN13 “families” of TreeFam into one of the four major families described above. The separate TSPAN31/TSPAN13 family described in TreeFam is placed as part of the CD63 family in our study, and the TSPAN32 family of TreeFam shows affinity to the uroplakin family in our study. The support measures for nodes in our tree also are in broad agreement with the bootstrap measures in the TreeFam description, in which support values for clusters of orthologs are strong, but those between clusters of orthologs appear weaker.

### CD63 as a particularly ancient tetraspanin

It is interesting that CD63 is associated with what has been called the *Drosophila* expansion tetraspanins [42]. This interesting association (see Fig. 1 and Supplemental Fig. S2) suggests a very ancient origin for the important CD63 tetraspanins found in many vertebrate genomes. This origin may be even more ancient as homologs of CD63 have been reported from sponges [43]. While this association would push the origin of this family back to an ancestor of all animals, it remains to be confirmed because Huang et al. [30] did not observe this association in their

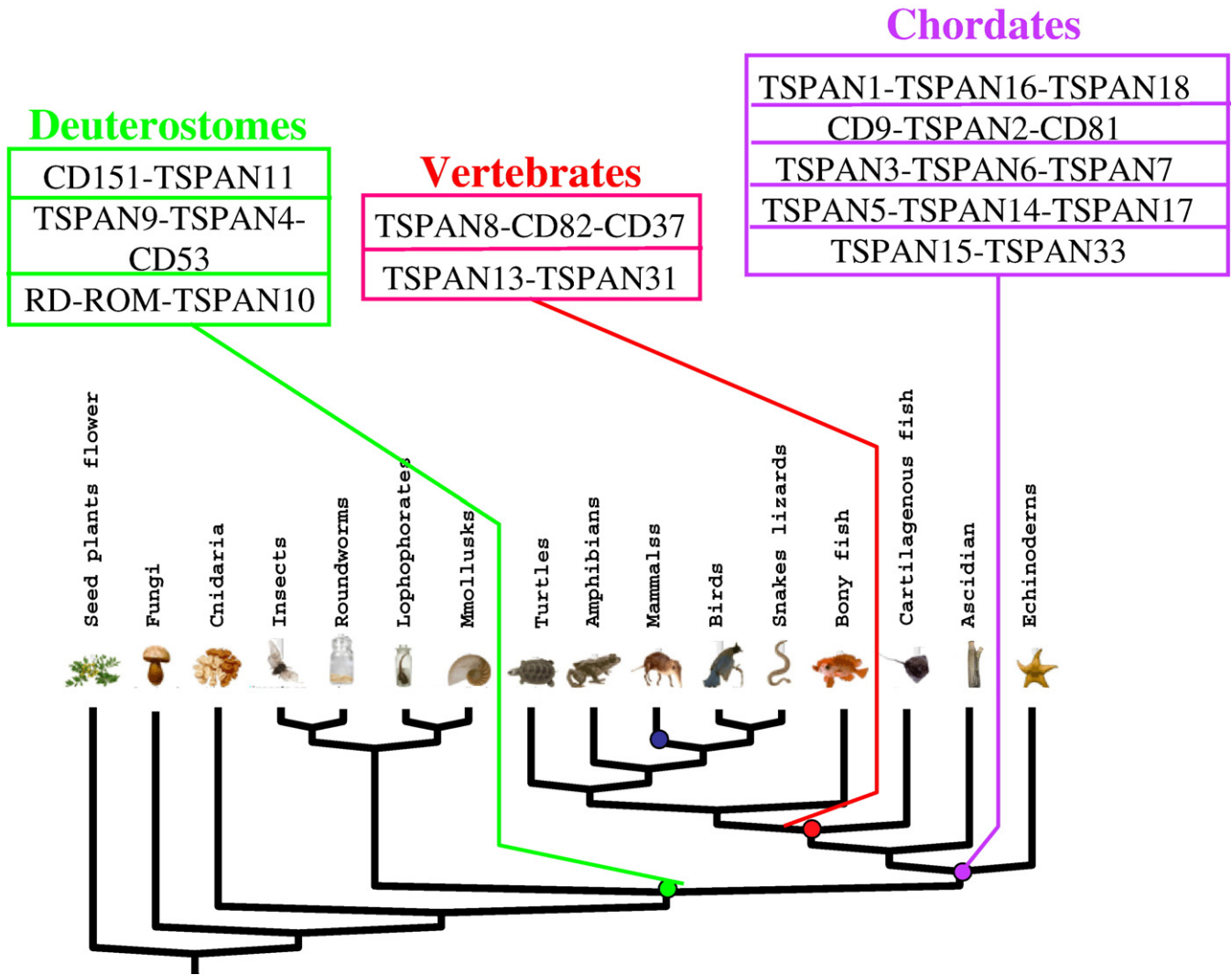


Fig. 3. A phylogenetic tree showing the point of ancestral origin for clusters of tetraspanin groups. Three points of origin are shown—vertebrate, chordate, and deuterostome.

analysis of tetraspanins. Nevertheless, the closer association of CD63 with an ancient duplication event in the ancestor of bilateral animals (Fig. 3) suggests a radically different origin for CD63 from most of the other CD genes. The same can be said for the several other vertebrate TSPAN's (TSPAN3, TSPAN6, TSPAN7, TSPAN13, and TSPAN31) found in the CD63 family in comparison with other vertebrate TSPAN genes.

#### *Ciona* and *strongylocentrotus* tetraspanins

Using the approach of Sarkar et al. ([41]; also see <http://research.amnh.org/users/desalle/data/tspan>), we assigned *Ciona* and *Strongylocentrotus* tetraspanins to known ortholog groups based on their phylogenetic affinity to the well-annotated tetraspanins from vertebrate and invertebrate genomes. Table 2 lists the annotation of several of these tetraspanins from *Ciona* and *Strongylocentrotus*. We note that some of the *Ciona* and *Strongylocentrotus* genes cannot be annotated to specific vertebrate tetraspanin groups. This result is not problematic, however, be-

cause both *Strongylocentrotus* and *Ciona* are basal nonvertebrate deuterostomes and the lack of ability to determine the exact ortholog group of these *Ciona* and *Strongylocentrotus* tetraspanins simply indicates that they might be basal to multiple ortholog groups as a result of duplication events in the vertebrates.

#### *Earliest common ancestor and age of origin of tetraspanins and their ortholog groups*

Examination of all 33 vertebrate tetraspanin ortholog groups allowed us to assign whether an ortholog group originated in the ancestor of mammals (100 Mya [44]), tetrapods (370 Mya [44]), vertebrates (450-510 Mya [44]), chordates (525 Mya [45]), protostome-deuterostome (570 Mya [46]), or bilateral animals (1200 Mya [47,48]). Fig. 2 shows the evolutionary origins for the 33 tetraspanin paralogs of humans. Most of the tetraspanins have a vertebrate origin (21), with 4 tetraspanins originating in the ancestor of tetrapods and 6 being strictly mammalian inventions.

One tetraspanin (TSPAN12) originated in the ancestor of chordates, and 1 (TSPAN15) originated after the protostome-deuterostome split. Fig. 2 also shows the ages of these single tetraspanin groups based on fossil evidence of the ancestors as listed above.

In addition to determining the earliest common ancestor of single tetraspanins, we also determined the origin of larger groups of tetraspanins (Fig. 3). For instance, CD9, TSPAN2, and CD81 form a triad of tetraspanin groups. Based on the ancestral tetraspanins related to this triad (four *Ciona* tetraspanins), we infer that this triad had an origin in the chordate ancestor. On the other hand, CD151 and TSPAN11 have a *Drosophila* tetraspanin–DmTsp74F—in an ancestral position to this pair in the phylogeny, suggesting a bilateral animal origin for this pair. This Figure therefore shows the earliest common ancestral position for strongly supported pairs and triads of tetraspanin groups. The majority of the pairs and triads originate in the deuterostome ancestor. Only three of the pairs and triads examined here have origins in the animal ancestor (CD151/TSPAN11; TSPAN9/TSPAN4/CD53; RDS/ROM/TSPAN10), indicating that the large number of *Drosophila* and *Caenorhabditis* tetraspanins have few orthologs in vertebrates.

By analyzing tetraspanins from only taxa with complete genomes, we can determine the earliest common ancestor for each of the different tetraspanins (Fig. 2) and their ortholog groups (Fig. 3). By far the majority of tetraspanins found in the human genome are either vertebrate (21 instances) or mammalian (6 instances) inventions. The tetraspanins in nondeuterostomes also show large sequence divergence as evidenced by the diversity of the genes in the *Drosophila* tetraspanins that have expanded in chromosomal region 42E and the large number of *C. elegans* tetraspanins with unique intron structure. When the different tetraspanin subgroups are clustered together in accordance with their phylogenetic patterns, the majority of the secondary groups are chordate or deuterostome inventions, suggesting that the large number of tetraspanins in mammals like mice and humans is a result of duplication events in the ancestor of vertebrates and the ancestor of mammals.

#### *Evolution of the structure and function of tetraspanins*

Our data suggest that the large superfamily of tetraspanin proteins be classified into four major families—the CD family, the CD63 family, the uroplakin family, and the RDS family. Because of weak support for relationships at the base of the tetraspanin tree, we believe that further clustering of these four groups is not possible. Within these four families, our tree structure provides strong support for most of the specific ortholog groups for tetraspanins. In addition, some associations of specific ortholog groups with each other are also well supported—such as the association of CD151 and TSPAN11; the association of TSPAN9, TSPAN4, and CD53; and the association of CD9, CD81, and TSPAN2.

Detailed analyses of members within each of the major tetraspanin families can yield insights into how the structure and function of the member genes may have evolved. For example, we recently studied the available genomic and cDNA sequences of uroplakins Ia and Ib. Although these two tetraspanins were thought to be produced as major differentiation products only by mammalian bladder urothelia, uroplakin-related genes were

recently found to exist in lower vertebrates, including chicken, frog, and fish [32,49]. UPIa and Ib bind specifically with two associated proteins, uroplakins II and IIIa, respectively, forming heterodimers before they can exit from the endoplasmic reticulum [50,51]. The UPIa/II and UPIb/IIIa heterodimers assemble into 16-nm particles that are packed hexagonally, forming two-dimensional crystals, called urothelial plaques—which cover almost the entire apical surface of mammalian urothelia and contribute to the remarkable transcellular permeability barrier function of the bladder [50,52]. Our analyses revealed that the UPIa and UPIb genes, and their associated UPII and UPIII genes, evolved by gene duplication with the appearance of vertebrates; that various combinations of uroplakin genes can be discarded during vertebrate evolution depending on the form of the nitrogenous waste (i.e., urea, uric acid, or ammonium) that is produced by the organism; and that UPIa and UPIb genes co-evolved with their partner UPII and UPIIIa genes, respectively [32]. Analyses of other tetraspanin families using completely sequenced genomes, in a manner similar to that of Garcia-España et al. [32], may lead to a better understanding of the structural and functional relationships among various tetraspanin gene members. The present study is a step toward this end in that we have better defined tetraspanin subfamilies, identified novel tetraspanin members, and assigned orthologs of many mammalian tetraspanins that are now more amenable to genetic and functional analyses.

#### **Announcement: T4NET Web site for tetraspanin nomenclature and research**

We also announce the launch of a Web site (<http://research.amnh.org/users/desalle/data/tspan>) for tetraspanin researchers. The Web site details the phylogenetic hypothesis described in this paper and serves as a Web identification tool for putative tetraspanins in genome research.

#### **Materials and methods**

##### *Tetraspanin matrix and alignment*

Tetraspanin sequences were collected from diverse sources: the pFAM site at the Sanger Institute (<http://www.sanger.ac.uk>) and the Locus Link site at the National Center for Biotechnology Information (NIH, Bethesda; <http://www.ncbi.nlm.nih.gov>) or by searching the various genome sequencing projects using the Blast-T program with multiple starting queries and *E* value threshold of *E* -5. This threshold is very relaxed and several hits were obtained that had minimal sequence similarity. We restricted our searches to 11 well-characterized and fully sequenced metazoan genomes—*Homo*, *Canis*, *Drosophila*, *Mus*, *Ciona*, *Danio*, *Gallus*, *Strongylocentrotus*, *Caenorhabditis*, *Arabidopsis*, *Oryza*, and several fungi (*Rhizopus*, *Encephalitozoon*, *Neurospora*, *Gibberella*, *Blastocladiella*, *Phanerochaete*, *Antrodia*) and the fully sequenced single-celled eukaryotic genomes *Leishmania*, *Trypanomonas*, and *Dictyostelium*. We performed very aggressive searches for tetraspanins in yeast genomes by lowering the *E* value thresholds and found no remnants of TSPAN's in the genomes of those organisms. While a few other mammalian species have complete and well-annotated genomes, such as *Rattus rattus*, *Pan troglodytes*, and *Macaca mulatta*, these species are very closely related to either human or mouse. A preliminary examination of the TSPAN's in these organisms suggests that their TSPAN's have clear orthologs to either human or mouse TSPAN's. Furthermore, inclusion of these TSPAN sequences from rat, chimp, and rhesus macaque in the analysis does not reveal new orthologs.

Multiple alignments of cDNA sequences were performed using the Dialign software from the GenomatixSuite ([www.genomatix.de/cgi-bin/dialign/dialign.pl](http://www.genomatix.de/cgi-bin/dialign/dialign.pl)). Alignments of protein sequences were performed using the default parameters of the ClustalW program and manually adjusted with the program MacClade4 PPC [33]. Amino-terminus sequences up to the first transmembrane domain were trimmed away as well as the carboxyl-terminus sequences after the fourth transmembrane domain.

### Phylogenetic analysis

#### Establishing paralog grouping and alignment

The first step in any phylogenetic or genealogical analysis is establishment of membership in a phylogenetically defined group and the second step is the construction of the phylogenetic tree. Systematic theory suggests that phylogenetic analysis is really a two-step procedure [34,35]. The first step is establishment of topological similarity and this step is usually accomplished based on some non-phylogenetic criteria such as sequence similarity when examining DNA sequences of proteins and protein domains or topological locations of anatomical features used in morphological systematics.

Since BLAST scores at the level of  $e^{-5}$  and  $-6$  are borderline with respect to showing good similarity, we used the presence or absence and positions of introns in the genes in this superfamily as an indicator of membership in the superfamily. In particular, a gene sequence was included in the analysis if it had a Blast score of  $e^{-5}$  or better AND at least one conserved intron-exon junction with other genes in the superfamily. To establish intron-exon positions in tetraspanins we used the better annotated genomes of *Drosophila*, *Homo*, *Canis*, *Caenorhabditis* and *Mus*. We used the annotation information in the Ensembl search engine for tetraspanins. Searches in Ensembl clearly indicate intron-exon junctions for all of the tetraspanins we include in this study. These genomes have fully annotated intron-exon junction information for almost all of the tetraspanins in them. Tetraspanins from other genomes—*Ciona*, *Danio*, *Gallus*, *Strongylocentrotus*, *Leishmania*, *Encephalitozoa*, *Basidomycota*, *Trychomonas*, *Dictyostelium*, *Neurospora*, *Gibberella*, *Blastocladiella*, and *Trypanosoma*—were obtained and intron-exon junctions inferred from whole genome sequences. Alignment of amino acid sequences was performed using the default setting in ClustalW [36]. We followed the approach of Huang et al. [30], by which ambiguously aligned [37] regions of the various superfamily members were trimmed away from the unambiguous aligned regions of the four transmembrane regions and the large internal loop leaving a matrix with sequence from only these unambiguously aligned regions.

#### Tree building

Once sequences from protein domains were determined to be valid members of the tetraspanin superfamily and aligned as discussed above, three major kinds of phylogenetic analysis—neighbor joining, parsimony, and Bayes analysis—were performed. For neighbor-joining and parsimony analysis two weighting schemes were used: (1) equal weights for all characters and (2) a genetic identity cost matrix (Fitch matrix). In all similarity analyses gaps were scored as missing. While there is no relevant published information to guide us as to which proteins to use as outgroups, we chose the plant tetraspanins as outgroups to root the animal tetraspanins. All phylogenetic analyses (parsimony and neighbor joining) were performed using PAUP\* [38]. Parsimony searches were performed using the parsimony ratchet implemented in PAUPRAT [39] with 5000 ratchet replicates and a search on all shortest trees from the ratchet by a heuristic method using the ratchet trees as starting trees with tree bisection-reconnection branch swapping and the retention of all shortest trees. To enhance the resolution of the parsimony search we used the successive weighting procedure implemented in PAUP (the reweighting using rescaled consistency index option in PAUP). This method allows for the choice of parsimony trees that are more consistent with the data. In this study we obtained over 5000 parsimony trees with our searches. The successive weighting procedure was able to choose 18 of these 5000 equally parsimonious trees as being more consistent with the character information. Bootstrap and jackknife trees were also generated using PAUP\* [38]. Bayesian analysis of the sequence data was conducted using MrBayes [40] with Parsmodel active and 1,000,000 MCMC replicates with default burning parameters.

#### Identification of orthologs

We used the approach of [41] to determine ortholog relationships of tetraspanins from the *Ciona* and *Strongylocentrotus*, in which a “guide tree” [31,41] was produced using the *Canis*, *Homo*, *Drosophila*, *Caenorhabditis*, and *Mus*

tetraspanins. The ortholog relationships of query tetraspanins from the sea urchin and the ascidian were determined from their phylogenetic affinity to tetraspanin groups from the guide tree.

#### Web site

To facilitate further genome level research on tetraspanins we have developed a Web site called T4NET (<http://research.amnh.org/users/desalle/data/tspan>). This Web site has three major inaugural functions. First, the Web site shows the phylogenetic hypothesis for all of the tetraspanins described in this paper. This phylogenetic hypothesis is discussed in detail above and is presented in “pop-up” format to show in detail aspects of the relationships of tetraspanins discussed in this paper. Second, the Web site serves as a resource server for papers, other Web sites, and reference material on tetraspanin biology. Finally, the Web site can be used as a Web identification tool for putative tetraspanins. This tool uses the phylogenetic tree on the Web site as a guide tree and the methods described in Sarkar et al. [41] to allow for researchers to enter their putative tetraspanin sequence. The Web site then rapidly determines the best ortholog group to which the putative tetraspanin belongs.

### Acknowledgments

We thank Fedor Berditchevski (University of Birmingham), Andrew F.X. Goldberg (Oakland University), Shoshana Levy (Stanford University), E. Rubinstein (INSERM U602, France), Derek Wilson (MRC, Cambridge, UK), and Mark D. Wright (Austin Research Institute, Australia) for critical reading of the manuscript and helpful suggestions. These studies were funded by NIH Grants DK39753 and DK52206 (T.T.S.), FIS 02/3003 (A.G.-E.), the Arnold and Arlene Family Foundation (T.T.S.), and the Sackler Institute of Comparative Genomics and the Dorothy and Louis Cullman Program in Molecular Systematics at the American Museum of Natural History (R.D.).

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.12.005.

### References

- [1] H.T. Maecker, S.C. Todd, S. Levy, The tetraspanin superfamily: molecular facilitators, *FASEB J.* 11 (1997) 428–442.
- [2] E. Todres, J.B. Nardi, H.M. Robertson, The tetraspanin superfamily in insects, *Insect Mol. Biol.* 9 (2000) 581–590.
- [3] M.E. Hemler, Specific tetraspanin functions, *J. Cell. Biol.* 155 (2001) 1103–1107.
- [4] M.E. Hemler, Tetraspanin proteins mediate cellular penetration, invasion, and fusion events and define a novel type of membrane microdomain, *Annu. Rev. Cell Dev. Biol.* 19 (2003) 397–422.
- [5] C. Boucheix, E. Rubinstein, Tetraspanins, *Cell Mol. Life Sci.* 58 (2001) 1189–1205.
- [6] C.S. Stipp, T.V. Kolesnikova, M.E. Hemler, Functional domains in tetraspanin proteins, *Trends Biochem. Sci.* 28 (2003) 106–112.
- [7] C. Boucheix, G.H. Duc, C. Jasmin, E. Rubinstein, Tetraspanins and malignancy, *Expert Rev. Mol. Med.* (2001) 1–17.
- [8] F. Martin, et al., Tetraspanins in viral infections: a fundamental role in viral biology? *J. Virol.* 79 (2005) 10839–10851.
- [9] S. Kohla, et al., The role of the peripherin/RDS gene in retinal dystrophies, *Acta Anat. (Basel)* 162 (1998) 75–84.
- [10] R. Zemni, et al., A new gene involved in X-linked mental retardation identified by analysis of an X;2 balanced translocation, *Nat. Genet.* 24 (2000) 167–170.

- [11] S. Levy, T. Shoham, Protein-protein interactions in the tetraspanin web, *Physiology (Bethesda)* 20 (2005) 218–224.
- [12] M. Sala-Valdés, et al., EWI-2 and EWI-F link the tetraspanin web to the actin cytoskeleton through their direct association with ezrin-radixin-moesin proteins, *J. Biol. Chem.* 281 (2006) 19665–19675.
- [13] O. Barreiro, et al., Endothelial tetraspanin microdomains regulate leukocyte firm adhesion during extravasation, *Blood* 105 (2005) 2852–2861.
- [14] M.E. Hemler, Tetraspanin functions and associated microdomains, *Nat. Rev., Mol. Cell Biol.* 6 (2005) 801–811.
- [15] S. Levy, S.C. Todd, H.T. Maeker, CD81 (TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system, *Annu. Rev. Immunol.* 16 (1998) 89–109.
- [16] S. Levy, T. Shoham, The tetraspanin web modulates immune-signalling complexes, *Nat. Rev. Immunol.* 5 (2005) 136–148.
- [17] G. Zhou, et al., Uroplakin Ia is the urothelial receptor for uropathogenic *Escherichia coli*: evidence from in vitro FimH binding, *J. Cell Sci.* 114 (2001) 4095–4103.
- [18] F. Le Naour, et al., Severely reduced female fertility in CD9-deficient mice, *Science* 287 (2000) 319–321.
- [19] M. Gordón-Alonso, et al., Tetraspanins CD9 and CD81 modulate HIV-1-induced membrane fusion, *J. Immunol.* 177 (2006) 5129–5137.
- [20] M.H. Tran, et al., Tetraspanins on the surface of *Schistosoma mansoni* are protective antigens against schistosomiasis, *Nat. Med.* 12 (2006) 835–840.
- [21] O. Silvie, et al., Cholesterol contributes to the organization of tetraspanin-enriched microdomains and to CD81-dependent infection by malaria sporozoites, *J. Cell Sci.* 119 (2006) 1992–2002.
- [22] E. Rubinstein, et al., Reduced fertility of female mice lacking CD81, *Dev. Biol.* 290 (2006) 351–358.
- [23] S.H. Ho, et al., Recombinant extracellular domains of tetraspanin proteins are potent inhibitors of the infection of macrophages by human immunodeficiency virus type 1, *J. Virol.* 80 (2006) 6487–6496.
- [24] P.H. Clergeot, et al., PLS1, a gene encoding a tetraspanin-like protein, is required for penetration of rice leaf by the fungal pathogen *Magnaporthe grisea*, *Proc. Natl. Acad. Sci. U. S. A.* 98 (2001) 6963–6968.
- [25] T. Shoham, R. Rajapaksa, C.C. Kuo, J. Haimovich, S. Levy, Building of the tetraspanin web: distinct structural domains of CD81 function in different cellular compartments, *Mol. Cell Biol.* 26 (2006) 1373–1385.
- [26] A.F.X. Goldberg, Role of peripherin/RDS in vertebrate photoreceptor architecture and inherited retinal degenerations, *Int. Rev. Cytol.* 253 (2006) 131–175.
- [27] X.T. Kong, et al., Roles of uroplakins in plaque formation, umbrella cell enlargement, and urinary tract diseases, *J. Cell Biol.* 167 (2004) 1195–1204.
- [28] P. Hu, et al., Ablation of uroplakin III gene results in small urothelial plaques, urothelial leakage, and vesicoureteral reflux, *J. Cell Biol.* 151 (2000) 961–972.
- [29] P. Hu, et al., Role of membrane proteins in permeability barrier function: uroplakin ablation elevates urothelial permeability, *Am. J. Physiol., Renal Physiol.* 283 (2002) F1200–F1207.
- [30] S. Huang, et al., The phylogenetic analysis of tetraspanins projects the evolution of cell-cell interactions from unicellular to multicellular organisms, *Genomics* 86 (2005) 674–684.
- [31] J.C. Chiu, et al., OrthologID: automation of genome-scale ortholog identification within a parsimony framework, *Bioinformatics* 22 (2006) 699–707.
- [32] A. Garcia-España, et al., Origin of the tetraspanin uroplakins and their co-evolution with associated proteins: implications for uroplakin structure and function, *Mol. Phylogenet. Evol.* 41 (2006) 355–367.
- [33] D.R. Maddison, W.P. Maddison, MacClade4, version 4.05, Sinauer, Sunderland, MA, 2000.
- [34] M.C.C. DePinna, Concepts and tests of homology in the cladistic paradigm, *Cladistics* 7 (1991) 367–394.
- [35] A.V.Z. Brower, V. Schawaroch, Three steps of homology assessment, *Cladistics* 12 (1996) 265–272.
- [36] R. Chenna, et al., Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res.* 31 (2003) 3497–3500.
- [37] J. Gatesy, W. Wheeler, R. DeSalle, Alignment of ambiguous nucleotide sites and the exclusion of data, *Mol. Phylogenet. Evol.* 2 (1993) 152–157.
- [38] D.L. Swofford, et al., PAUP\*: Phylogenetic Analysis Using Parsimony (and Other Methods), Sinauer, Sunderland, MA, 2000.
- [39] D.S. Sikes, P.O. Lewis, Beta software, version 1. PAUPRat: PAUP implementation of the parsimony ratchet. Distributed by the authors. Department of Ecology and Evolutionary Biology, Univ. of Connecticut, Storrs (2001).
- [40] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogeny, *Bioinformatics* 17 (2001) 754–755.
- [41] I.N. Sarkar, et al., An automated phylogenetic key for classifying homeoboxes, *Mol. Phylogenet. Evol.* 24 (2002) 388–399.
- [42] L.G. Fradkin, J.T. Kamphorst, A. DiAntonio, C.S. Goodman, J.N. Noordermeer, Genomewide analysis of the *Drosophila* tetraspanins reveals a subset with similar function in the formation of the embryonic synapse, *Proc. Natl. Acad. Sci. U. S. A.* 99 (2002) 13663–13668.
- [43] W.E. Muller, et al., Initiation of an aquaculture of sponges for the sustainable production of bioactive metabolites in open systems: example, *Geodia cydonium*, *Mar. Biotechnol.* 1 (1999) 569–579.
- [44] S.B. Hedges, S. Kumar, Genomics: vertebrate genomes compared, *Science* 297 (2002) 1283–1285.
- [45] D. Shu, et al., A new species of yunnanozoan with implications for deuterostome evolution, *Science* 299 (2003) 1380–1384.
- [46] D. Erwin, The origin of body plans, *Am. Zool.* 39 (1999) 617–629.
- [47] R.F. Doolittle, et al., Determining divergence times of the major kingdoms of living organisms with a protein clock, *Science* 271 (1996) 470–477.
- [48] M. Nei, et al., Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms, *Proc. Natl. Acad. Sci. U. S. A.* 98 (2002) 2497–2502.
- [49] K. Sakakibara, et al., Molecular identification and characterization of *Xenopus* egg uroplakin III, an egg raft-associated transmembrane protein that is tyrosine phosphorylated upon fertilization, *J. Biol. Chem.* 280 (2005) 15029–15037.
- [50] L. Tu, T.T. Sun, G. Kreibich, Specific heterodimer formation is a prerequisite for uroplakins to exit from the endoplasmic reticulum, *Mol. Biol. Cell* 13 (2002) 4221–4230.
- [51] C. Hu, et al., Assembly of urothelial plaques: tetraspanin function in membrane protein trafficking, *Mol. Biol. Cell* 16 (2005) 3937–3950.
- [52] T.T. Sun, Altered phenotype of cultured urothelial and other stratified epithelial cells: implications for wound healing, *Am. J. Physiol. (Renal Physiol.)* 291 (2006) F9–F21.