

# Chapter 2

## The Evolution of Tetraspanins Through a Phylogenetic Lens

Rob DeSalle, Tung-Tien Sun, Tjard Bergmann, and Antonio Garcia-España

**Abstract** The tetraspanin superfamily of proteins provides an excellent system for examining many important evolutionary phenomena at the level of gene and protein sequences. Because dozens of eukaryotic organisms now have their full genomes sequenced, tetraspanins from these genomes can be compared and placed into a phylogenetic context. The whole genome information allows for researchers to trace with great precision the evolutionary events that have molded the broad array of tetraspanins found in eukaryotic genomes. We first demonstrate that phylogenetic analysis of tetraspanins from the fully sequenced genomes of an exemplar set of eukaryotes can give a fairly complete picture of the relationships of the families and subfamilies of tetraspanins. We can use the phylogenetic analysis of these tetraspanins to classify the various families and subfamilies of tetraspanins and use the

---

R. DeSalle

Sackler Institute for Comparative Genomics, American Museum of Natural History,  
New York, NY 10024, USA

T.T. Sun

Department of Cell Biology, New York University School of Medicine,  
New York, NY 10016, USA

Department of Dermatology, New York University School of Medicine,  
New York, NY 10016, USA

Department of Pharmacology, New York University School of Medicine,  
New York, NY 10016, USA

Department of Urology, New York University School of Medicine,  
New York, NY 10016, USA

T. Bergmann

ITZ, Ecology and Evolution, Tierärztliche Hochschule Hannover, Hannover, Germany

A. Garcia-España (✉)

Research Unit, Universitat Rovira i Virgili, Hospital Joan XXIII,  
Pere Virgili Institute (IISPV), Tarragona 43007, Spain  
e-mail: agarciae.hj23.ics@gencat.cat

sequence information as diagnostics for identifying novel tetraspanins. By using a phylogenetic perspective we also examine several important evolutionary processes in the tetraspanins such as intron evolution and the evolution of small protein motifs. We also describe a website for researchers who are interested in tetraspanin evolution, classification, identification and information called TSPAN4.web (<http://research.amnh.org/users/desalle/data/tspan/>).

## 2.1 Tetraspanins

**An obligatory (but short) introduction:** Tetraspanins are members of a large group of integral membrane proteins (Maecker et al. 1997; Hemler 2001, 2003; Boucheix and Rubinstein 2001). Humans have 33 tetraspanin members scattered throughout the genome (see Table 2.1) whose functions are distributed widely in cells and tissues. The structure of tetraspanins is widely conserved across large phylogenetic with the typical tetraspanin being 200–350-amino-acid-long with four transmembrane (TM) domains. In addition two extracellular loops exist in these proteins, one being small (SEL—about 13–30 amino acids long) and the other large (LEL—up to 150 amino acids long). Many tetraspanin proteins were originally identified as human tumor antigens while others are associated with several forms of retinal degeneration. Still others have been associated with mental retardation syndromes (Zemni et al. 2000). Tetraspanin-enriched microdomains can form through primary associations with a variety of transmembrane and intracellular signaling/cytoskeletal proteins and secondary associations (Levy and Shoham 2005a, b). The conserved structure of tetraspanins over such extreme functional diversity and phylogenetic time makes them an ideal subject for evolutionary analysis. This chapter examines the superfamily of proteins through a phylogenetic “looking glass”, by first explaining the caveats of phylogenetic analysis of tetraspanins. Next, we examine nomenclatural issues that arise as a result of having a phylogenetic framework for this superfamily. We also examine two important aspects of protein evolution using the tetraspanins—intron and short amino acid motif evolution. We conclude by demonstrating how a close up view of the phylogenetics of specific tetraspanins can enhance our understanding of the structure and function of these proteins.

## 2.2 Phylogenomic Methods

**The ins and outs of protein family trees:** Many gene families have been analyzed using phylogenetic approaches. Often the methodology and limitations of such analyses are unclear. While an exhaustive explanation of the phylogenetic approaches is beyond the scope of this chapter, we present here a critical discussion of why we chose our particular approaches to analyze the phylogenetic evolution of tetraspanins. First and foremost to keep in mind when analyzing gene families is the concept

**Table 2.1** Tetraspanin superfamily nomenclature, family designations and divergence times

Protein	Gene	Aliases	Family	DT
TSPAN1	TSPAN1	TSP-1	CD	V
TSPAN2	TSPAN2	TSP-2	CD	T
TSPAN3	TSPAN3	TSP-3	CD63	V
TSPAN4	TSPAN4	TSP-4/NAG2	CD	V
TSPAN5	TSPAN5	TSP-5	RD	V
TSPAN6	TSPAN6	TSP-6	CD63	T
TSPAN7	TSPAN7	CD231/TALLA-1/A15	CD63	V
TSPAN8	TSPAN8	CO-029	CD	T
TSPAN9	TSPAN9	NET-5	CD	V
TSPAN10	TSPAN10	OCULOSPANIN	RD	V
TSPAN11	CD151-like	CD151-like	RD	V
TSPAN12	TSPAN12	NET-2	Uroplakin	C
TSPAN13	TSPAN13	NET-6	CD63	V
TSPAN14	TSPAN14		RD	V
TSPAN15	TSPAN15	NET-7	RD	D
TSPAN16	TSPAN16	TM4-B	CD	M
TSPAN17	TSPAN17		RD	M
TSPAN18	TSPAN18		CD	V
TSPAN19	TSPAN19		CD	V
TSPAN20	UPK1B	UP1b, UPK1B	Uroplakin	V
TSPAN21	UPK1A	UP1a, UPK1A	Uroplakin	V
TSPAN22	RDS	RDS, PRPH2	RD	V
TSPAN23	ROM1	ROM1	RD	V
TSPAN24	CD151	CD151	CD	V
TSPAN25	CD53	CD53	CD	M
TSPAN26	CD37	CD37	CD	M
TSPAN27	CD82	CD82/KAI-1	CD	V
TSPAN28	CD81	CD81	CD	V
TSPAN29	CD9	CD9	CD	V
TSPAN30	CD63	CD63	CD63	V
TSPAN31	TSPAN31	SAS	CD63	V
TSPAN32	TSPAN32	TSSC6	Uroplakin	M
TSPAN33	TSPAN33		CD	M

*M* mammal divergence at 100 MYA, *T* tetrapod divergence at 370 MYA, *V* vertebrate divergence at 450 MYA, *C* chordate divergence at 535 MYA and *D* deuterstome divergence at 570 MYA

of homology of genes and proteins. Genes in a gene family can be **orthologous** or **paralogous** with each other (Thornton and DeSalle 2000). Orthologous genes are those that are in different organisms as a result of common ancestry via speciation. Paralogous genes are those that are in genomes (the same or different genomes) as a result of gene duplication. For example, human UP1A and human UP1B are both considered uroplakins, they are in reality paralogs of each other. By the same token, chimpanzee UP1A and human UP1B share some common ancestry they are also paralogs of each other. On the other hand though, chimpanzee UP1A and human UP1A are considered orthologs of each other. A first approximation of homology is

usually made using similarity via a BLAST score. The determination of orthology can then be made by optimizing some aspect of the similarity scoring or through phylogenetic analysis (Chiu et al. 2006).

A major issue in the analysis of gene families that is also a consideration with tetraspanins, is to decide whether the analysis should be done on protein or DNA sequences. Since gene families that include Bacteria and Archaea and Eukarya will span the entire time that life has existed on this planet—3.5 or so billion years, considerable sequence change has occurred amongst the genes in the gene family. Using DNA sequences at this level is problematic because third positions in the genes will have evolved much more rapidly and the extreme amount of change that has occurred is difficult to compensate for even by modeling nucleotide sequence change. On the other hand, amino acid coding of the sequences evolves at a slower rate making such sequences more amenable to models that have been developed to compensate for such sequence change. Extreme sequence divergence also means that sequence alignment becomes a problem, and the alignment of amino acid sequences is simpler than alignment of DNA sequences at this degree of sequence change (although amino acids can be aligned first and used as a guide for DNA sequence alignment). Alignment and choice of model to compensate for extreme sequence change are two major initial problems to consider when examining gene families. Since tetraspanins appear to be present in all eukaryotic life this means that the common ancestor of the members of this gene family are at least the age of eukaryotes—more than 1.5 billion years. This observation means that amino acids are perhaps an appropriate source of data for phylogenetic studies of this large group of genes.

Another major issue has to do with how to generate phylogenetics, once an alignment of the gene family members has been produced. There are two main approaches to generating phylogenies both with their advantages and detractions. The most commonly used by molecular biologists are what are called distance or phenetic approaches. The linear sequence information in this method are condensed into a distance (or similarity) measure based on a model of sequence change for each pair of genes (or proteins) in the data set. The pairwise distances are then used in an algorithm that generates a phenogram that represents the distance information in the condensed matrix. The advantage of this kind of approach is its computational ease and rapidity. A second category of approaches leaves the sequence information intact as unitary characters and utilizes a character by character methods to generate a phylogenetic hypothesis. In this approach, the DNA sequence positions in the gene or the amino acid positions in the protein are assessed with optimality criteria for their fit onto a phylogenetic hypothesis. What this means is that the data for the genes or proteins in the analysis needs to be assessed for optimality with respect to each tree that can be generated for the genes or proteins in the analysis. For instance, for three proteins, three trees need to be assessed for optimality [if the three proteins are A, B and C, then the three trees are ((A,B)C), ((A,C),B) and ((B,C)A)]. When the number of proteins or genes in an analysis is over 15 or so, the ability of computers to compute exact solutions is prohibitive (the NP complete problem) and heuristic approaches to get a best estimate

of optimality are used (Felsenstein 2004). These character-based approaches can use parsimony or likelihood methods for assessing optimality of a tree topology, given the assumptions of the approaches. The choice one makes as to which method to use is often based on accessibility and speed.

A third major concern regarding phylogenetic analysis of gene and protein families has to do with the robustness of the inferences made when using small numbers of characters. Methods such as bootstrapping (Felsenstein 1985) and jackknifing (Farris et al. 1996) can be used to assess the robustness of inference at nodes in the tree. These methods are resampling techniques that can be applied to both distance and character-based analyses. For most phylogenetic comparisons, the robustness of inference at nodes is roughly correlated with the amount of sequence information for each taxon, so inferences made with single genes or proteins for each taxon are not necessarily robust. As a general rule of thumb, any bootstrap or jackknife value above 65% is credible (Hillis and Huelsenbeck 1992), but bootstrap and jackknife values of gene and protein phylogenies should probably be viewed differently from the same measures for organismal trees. This is because these measures, when used in organismal studies, can tell the systematist where further work is needed and where more sequence information needs to be collected from the genomes of the organisms being analyzed, to converge on a robust inference. In gene and protein family trees no new data can be added. Other approaches for assessing robustness that are based on character-by-character analysis exist that place phylogenetic analysis in more of a statistical context such as Bayesian Phylogenetic analysis that estimates a posterior probability for each node in a phylogenetic tree (Huelsenbeck and Ronquist 2001). Since the Bayesian posterior is a probability, researchers have used the classical *p* value cutoff of 0.05 as an indicator of significance for these statistics.

A fourth aspect of gene and protein family phylogenetics concerns sampling. Some phylogenetic studies take the approach of including just the genes from a single group, like the uroplakins, to obtain as many representatives of the genes from as many organisms as possible regardless of whether a whole genome for the organism exists. Other researchers have limited their analyses to those organisms with fully sequenced genomes to examine all of the genes in a gene family from existing full genomes. We have argued elsewhere (Garcia-España et al. 2008) that analysis of organisms with fully sequenced genomes is the most efficient and informative approach, because in this case the *absence* of a gene in a subgroup of species can infer special significance. If genes from an organism without a fully sequenced genome are used then no inference about its absence can be made, thus imposing severe limitations on the interpretations of the data.

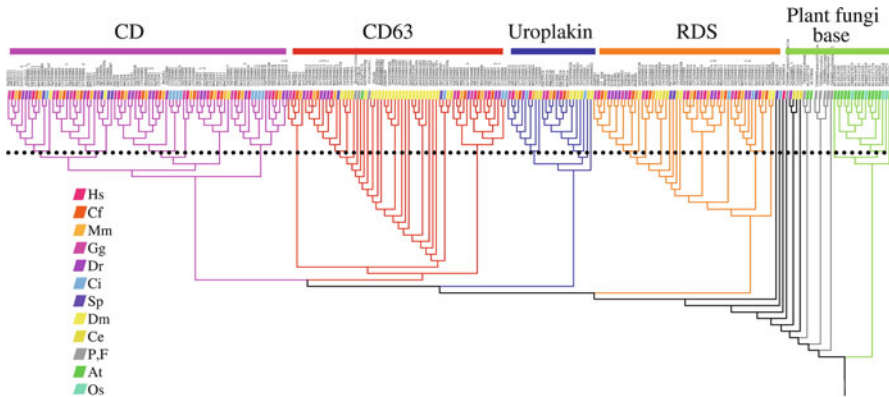
Finally, it is critical when thinking about protein and gene family phylogenetics to define the role of rooting or choice of outgroups. Of course, results of gene and protein family phylogenies can be presented as unrooted networks, and these can be quite informative, but being able to root the network renders polarity to the changes in the tree that can be inferred from the topology of the tree. Choice of outgroups in gene and protein family analysis can come from two sources. Firstly, if one is clear that a group of genes is orthologous, a closely related gene family

that is not part of that group can be used as the outgroup. Secondly, if one has a clearly defined group of orthologous genes in a gene family, then the gene or protein from the most primitive organism in the analysis can be used as an outgroup. While it is easy to feed a lot of tetraspanin data through a phylogenetic analysis program, the nuances discussed above concerning choice of characters (i.e. DNA sequences or amino acid sequences), orthology, choice of algorithm or optimality criteria (i.e. distance analysis, parsimony or likelihood), robustness of inference (bootstrap, jackknife or Bayesian posteriors) and outgroup choice can all have a huge impact on interpreting results.

### 2.3 Classification System for Tetraspanins

**A tree-based nomenclature for tetraspanins:** We begin this discussion with a note on nomenclature. This large group of genes (proteins) includes 33 members in the human genome (Table 2.1). The members of this group of proteins are also sometimes called the transmembrane 4 superfamily (TM4SF) proteins. The nomenclature of the genes and proteins within this large group of proteins is most clearly articulated by the HUGO (Human Genome Organization) Gene Nomenclature Committee (<http://www.genenames.org/index.html>). According to the HUGO nomenclature system, there are 33 genes that exist in the human genome that encode the tetraspanin proteins and some of their “aliases” (Table 2.1). Note that some of the tetraspanin proteins produced by these 33 genes have been annotated as TSPAN followed by a number. Still others in the large group of genes are named uroplakins (UPK), Retinal degeneration slow (RDS) and the well known CD proteins (followed by a number) because of their specific cell expression pattern and function. These proteins are found in a wide range of living species and present in plants, animals, fungi and protists. Because of the breadth of organismal range and functionality of these proteins, we have adopted the convention of calling the entire group of tetraspanins a superfamily as in the TM4SF tradition. We then divide this tetraspanin superfamily into families, which are then divided into groups based on the existing annotations of genes and proteins in this superfamily.

Several research groups have used tetraspanins as the subject of gene family analysis (Huang et al. 2005; Todres et al. 2000; Garcia-España et al. 2008). In general, their results are congruent with respect to the monophyly of members in the major groups of tetraspanins. Some differences occur between the phylogenies when deeper nodes are examined, and hence relationships of families of genes within the superfamily may be different between the two studies. However, these differences in interpretation are due to the lack of robustness at nodes at the base of the trees in all of the studies accomplished so far. In this chapter, we discuss the classification system of Garcia-Espana et al. 2008 as a framework for tetraspanin evolution. A detailed phylogeny of the tetraspanins from this study can be found at the TSPAN4 website (see last section of this chapter for a full discussion of the website as a research tool). Since both studies that use large



**Fig. 2.1** Phylogenetic tree from Garcia-Espana et al. (2008). Species are designated by *colored boxes* with a legend for the species designation given (species abbreviations are given in Garcia-Espana and at <http://research.amnh.org/users/desalle/data/tspan/>). More detailed “close-ups” of the four major groups of tetraspanins designated here are available on the TSPAN4 website <http://research.amnh.org/users/desalle/data/tspan/>. The tetraspanin superfamily can be subdivided into four major monophyletic families (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%.

sampling of tetraspanin genes (Huang et al. 2005; Garcia-España et al. 2008) tend to group genes similarly, the groupings can serve as a basis for the classification of tetraspanins.

The phylogenetic analyses summarized in the tree in Fig. 2.1 (all trees discussed in this chapter can also be found at <http://research.amnh.org/users/desalle/data/tspan/>) shows four major clades that we have given the rank of family (called the **CD family**, the **CD63 family**, the **uroplakin family**, and the **RDS family**; Table 2.1). The largest cluster of tetraspanins, i.e., the CD family, comprises proteins annotated in existing genome databases as vertebrate CD and Tsp proteins with several invertebrate tetraspanins. This group includes all of the previously annotated tetraspanins with the designation CD in their name (151, 53, 9, 81, 82, 37) except for CD63. This latter CD tetraspanin is placed into its own family, the second largest of the four families with respect to members. This family contains the CD63 orthologs from several vertebrates and cluster of genes at chromosome location 42E in the *Drosophila* genome. The CD63 family of tetraspanins is highly divergent with several previously annotated vertebrate TSPAN proteins (TSPAN13, TSPAN31, TSPAN3, TSPAN6, and TSPAN7). The uroplakin family is made up of the classically named vertebrate uroplakin (UP) genes and several invertebrate tetraspanins (represented by the well-characterized *Drosophila* tetraspanin expansion group) as well as TSPAN32 and TSPAN12. The final large family of animal tetraspanins is called the RDS family, because it includes the RDS-ROM tetraspanins and this family also includes the orthologs of Human TSPAN10, TSPAN14, TSPAN5, TSPAN17, TSPAN15, TSPAN33.

## 2.4 The Origin of the Tetraspanins

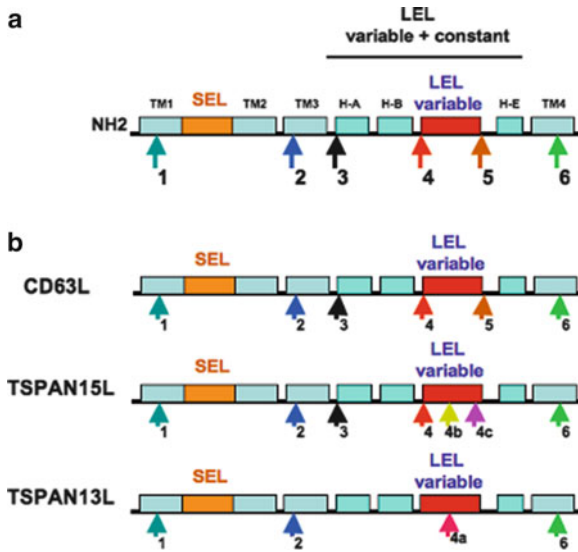
**Superfamilies, families and groups:** By examining the clustering of tetraspanin orthologs and assaying the taxonomic representation within the ortholog groups, we can estimate the times of origin and divergence of the various groups. For instance, each of the four major families, i.e., CD, CD63, Uroplakin and RDS, have both vertebrate and invertebrate representatives of the Bilateria, but no fungal, plant or protist members. In addition, recent analysis of Cnidaria, Placozoa and Poriferan tetraspanins indicates that these phyla also have representatives of all four families of tetraspanins and that, while the Choanoflagellate, *Monosiga* also has a tetraspanin (data not shown), it is not orthologous to any of the tetraspanins in the CD, CD63, Uroplakin or RDS families of tetraspanins. The distribution of tetraspanins in the genomes of all of these animals and the choanoflagellate indicates that the expansion of this superfamily into the four large families we describe above was an “invention” in the genome of the ancestor of the Metazoa. These observations suggest that the origin of the expanded tetraspanin superfamily into four families corresponds to a divergence of more than about 540–650 million years in the ancestor of all metazoans that most likely existed prior to the Vendian period (Hedges and Kumar 2002; Doolittle et al. 1996).

Table 2.1 lists the human tetraspanin genes and their approximate time of origin using this approach. Using this approach we can designate certain tetraspanins as “inventions” of particular ancestors in the history of animals. For instance, there appears to have been a burst of tetraspanin “invention” in the ancestor of vertebrates. This burst also corresponds with well-known genome duplications in the ancestor of particular lineages of vertebrates. Another significant “burst” of tetraspanin origin also occurred in the ancestor of mammals, where CD37, CD53, TSPAN16, TSPAN17, TSPAN32 and TSPAN33 arose (Garcia-España et al. 2008). Future work using this approach should incorporate the newly emerging mammalian genomes to determine whether any of the tetraspanins are specific to orders of mammals such as the primates.

## 2.5 Introns and Cysteines

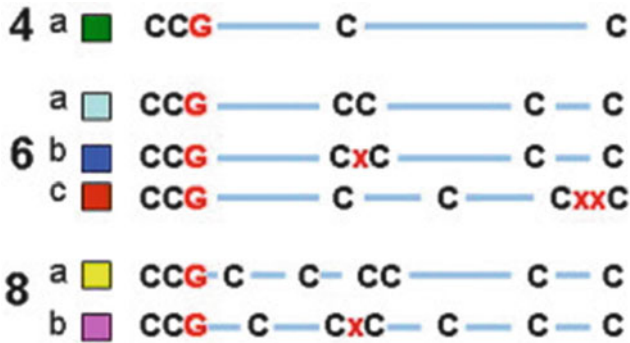
**Evolution of intron junctions and protein motifs:** The tetraspanin superfamily offers an excellent system for examining specific aspects of genome and protein evolution. In this section, we examine two evolutionary phenomena specific to tetraspanins. The first concerns the evolution of introns (Garcia-España et al. 2009; Garcia-España and DeSalle 2009) and the second concerns the evolution of repeated motifs in proteins (DeSalle et al. 2010). Using the robust phylogeny of the tetraspanins these interesting aspects of the gene family can be examined in precise detail. While several elegant studies of intron evolution using whole genome approaches have been useful in detecting genome-wide intron evolutionary trends, taking a





**Fig. 2.2** Cartoons of intron positions in the tetraspanin genes. (a) The small (SEL; orange) and large (LEL; red) extracellular loops are indicated. Light blue represents the four transmembrane domains while no color, represents the intracellular regions. Ancestral intron positions 1–6 are indicated on the protein by colored arrows of the same color that will be used through all of the figures. Abbreviations are *TM-1–4* transmembrane domains; *H-A*, *H-B* and *H-E* constant helices in the LEL. (b) Animal CD63L (top), TSPAN15L (middle) and TSPAN13L (bottom) tetraspanins' consensus intron structure. Purple red (intron 4a), yellow (intron 4b) and (intron 4c) indicate four new intron junctions discussed in the text

gene family-specific approach can also be useful. The analysis of gene family, again, in fully sequenced genomes, can reveal patterns of intron gain and loss more precisely. In addition, if a gene family is used where specific function of the gene products is known, more precise interpretation of the gain and loss patterns can be made. The intron structure of tetraspanins is also interesting because of the relatively large number of introns in the genes in this superfamily. While an examination of the range of tetraspanin genes for intron position reveals that there are at least 105 unique intron positions in the tetraspanins of fungi, plants, protists and animals, the most common intron structure of tetraspanins in animals is a six intron scheme (Fig. 2.2a). More precisely, there are three major intron patterns (Fig. 2.2b) from which all other animal intron patterns are derived. As with the appearance of new tetraspanins in the genomes of animals, we can use the phylogenomic approach to give dates to the gain and loss of new introns in tetraspanins. These data show that there is a strong correlation of the appearance of tetraspanins with novel functions with the insertion of introns in new positions in the overall tetraspanin gene structure (Garcia-España et al. 2009). For instance, as we discussed above, there was a burst of appearance of novel tetraspanins in the ancestor of vertebrates, and this burst of new tetraspanins is accompanied by the appearance of six new introns in



**Fig. 2.3** The LEL patterns of cysteines according to their number and relative position to each other. *Blue lines* indicate any number of residues between adjacent cysteines. Each small *x* (*x*) indicates a single residue. Capital *red G* (*G*) indicates the glycine residue of the CCG motif. Each of the six cyteine patterns is represented by a distinct *colored box*

these genes. Furthermore, the position of these new introns in tetraspanins is non-random, as nearly 50% of new introns appear in the small extracellular loop (SEL), which accounts for only 10% of the entire length of most tetraspanin proteins. When the large extracellular loop (LEL), which on average makes up only 25% of tetraspanin proteins, is examined for novel intron occurrence Garcia-Espana and colleagues (2009) observed that another 25% of novel introns accrue in this region. Clearly the two extracellular loops (totaling ~35% of the entire amino acid sequence of most tetraspanins) are accruing the vast majority (70–75%) of novel introns. Using the patterns of intron gain and loss, it was also estimated that there are 105 intron gain events (42 alone in *C. elegans*) and only four intron loss events (Garcia-Espana et al. 2009). Finally, using the phylogenomic approach, we were able to determine that indels (i.e. insertions or deletions) at the ends of DNA exonic sequences could have caused the appearance of two discordant intron positions between orthologous tetraspanins (Garcia-España and DeSalle 2009). These data suggest that an intron sliding mechanism (Tarrío et al. 2008) can be used to explain these observations. This intron-sliding mechanism could have been important in generating functional diversity in this superfamily of tetraspanin proteins.

Cysteine residues have been used to characterize tetraspanins in the past because these cysteine reside in distinct motifs and because these cysteine residues may play important roles in the secondary and tertiary structure of tetraspanin proteins. Most of these cysteine motifs are found in the large extracellular loop (LEL) and so the phylogenomic analysis used to examine them focused on this region of the tetraspanin structure. In general, the number of cysteines is even (four, six or eight) suggesting that they interact in pairs in disulfide bonding. The exceptions to this even number of cysteines are the RDS/ROM and plant tetraspanins. There are six easily recognized cysteine motifs that can be examined in a phylogenomic context (Fig. 2.3). The results of this analysis suggest that the cysteine motifs are correlated closely with phylogeny

and hence novel cysteine motifs are correlated with the bursts of appearance of new tetraspanins. In addition, DeSalle et al. (2010) suggest that the four cysteine motif (see Fig. 2.3) is highly derived occurring at the tips of the tetraspanin tree. What this means is that the four cysteine motif is a new evolutionary innovation. This suggests that the reduction in number of cysteines in the LEL is a recurring and more recent event in the evolution of the animal tetraspanins. One of the more important results of this study concerns the examination of the highly conserved CCG motif in the LEL. A detailed phylogenetic analysis of this motif reveals that it originated in the common ancestor of Unikonts—animals, Fungi and Amoebozoa. The CCG motif does not appear to be in all tetraspanins of plants, Stramenophiles, Alveolata, Discicristata, or Excavata, suggesting that it did not exist in the common ancestor of Bikonts.

## 2.6 Up Close and Personal

**The uroplakins:** An examination of closely related tetraspanins within specific families and groups can shed light on the evolutionary steps leading to the structure and function of proteins in the superfamily. As an example, we discuss the analysis of uroplakins by a detailed phylogenetic analysis in Garcia-Espana et al. (2006). This family of proteins are the integral membrane subunits of urothelial plaques (also known as the Asymmetric Unit Membrane [AUM]) that line the specialized apical surface of the mammalian urinary bladder epithelium. While there are four major uroplakins (UP1a, UP1b, UPK2 and UPK3) only two of them are tetraspanins (UP1a and UP1b), while the other two (UP2, UP3a) span the membrane only once. UP1a and UP1b interact selectively with UP2 and UP3a, respectively, to form Ia/II and Ib/IIIa complexes that further assemble to make up the urothelial plaques. A detailed phylogenetic analysis of all four of these four major uroplakin proteins individually, revealed a general correlation of protein evolution with organismal evolution. By tracing the presence and absence of the genes for these proteins in the genomes of vertebrates, Garcia-Espana et al. (2006) demonstrated two major phenomena relevant to the evolution of these proteins: (1) the UPIa and UPIb genes co-evolved by gene duplication in the common ancestor of vertebrates, as did UPII and UPIIIa; and (2) uroplakins can be lost in different combinations in vertebrate lineages generating a great deal of variability in the functionality of the proteins. Specifically, duplication of an ancestral UPI gene into UPIa and UPIb occurred in the common ancestor of cartilaginous fish and other vertebrates. Concomitantly, the duplication of a UPII/UPIIIa gene occurred in the same ancestor to produce the UPII and UPIIIa proteins. In addition, using a coevolutionary approach (where the proteins were compared in the following pairs UPIa/UPII, UPIa/UPIII, UPIb/UPII and UPIb/UPIIIa), these authors also showed that only the UPIa/UPII and UPIb/UPIIIa coevolutionary pairs showed statistical correlation suggesting there is a strong co-evolutionary relationship between UPIa and UPIb and their partners UPII and UPIIIa/IIIb, respectively (Garcia-Espana et al. 2006). These results further support

the biochemical analyses regarding the stoichiometric interactions (Tu et al. 2002; Hu et al. 2005) and strengthen the hypothesis that UPIa and UPII interact with each other and that UPIb and UPIIIa interact with each other in the formation of urothelial plaques.

## 2.7 TSPAN 4

**A website for studying the evolution of tetraspanins:** To facilitate broader use of the phylogenomic approach to the analysis of tetraspanins, we have developed a website for the use of tetraspanin researchers. The website can be accessed at <http://research.amnh.org/users/desalle/data/tspan/>. Figure 2.4 shows the homepage of the website that we have named T4NET. There are five functions that the website performs. First, the website holds the most recent tetraspanin tree (\* in Fig. 2.4) using accessions from whole genomes. This tree has a pull down function that allows the user to click on one of the four major families of proteins to get an upclose view of the structure and organization of protein groups. Second, the website allows users to input a new putative tetraspanin sequence and identify it using the “gene identifier” (# in Fig. 2.4) function on the website. Figure 2.5 shows the webpage where gene identification can be accomplished. Users simply cut and paste their putative tetraspanin gene sequence into the query box (large arrow) and use the Submit button (small arrow) for the webpage to determine the identity of their putative tetraspanin. Input sequences have to be in amino acid FASTA format. The third function of the website allows the user to view the uroplakin 1a, 1b phylogenetic

Rob DeSalle -- T4NET

© AMNH

\* # @ & \$

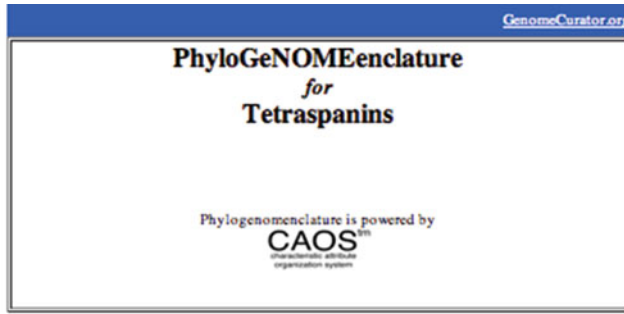
tetraspanin tree gene identifier uroplakin tree figures sequences

**Online material for the article:**  
**"Evolutionary History of the Tetraspanin Protein Superfamily"**  
**Garcia-España A, Chung PJ, Sarkar IN, Stiner E, Sun TT, DeSalle R (2007) *Genomics***

**Site contents**

- Tetraspanin protein superfamily phylogenetic tree
- Phylogenomenclature for tetraspanins
- Uroplakin tree with accession numbers and links to GenBank
- Figures
- Sequences

**Fig. 2.4** Homepage for the TSPAN4 website. There are five functions that the website performs. A tetraspanin tree (\*) function, a “gene identifier” function a uroplakin tree function (@), access the figures in tetraspanin paper (&) and accession numbers for tetraspanin genes that are in fully sequenced genomes (\$). The website can be accessed at <http://research.amnh.org/users/desalle/data/tspan/>



Indra Neil Sarkar\*, Marine Biological Laboratory  
Sergios-Orestis Kolokotronis, American Museum of Natural History  
Rob DeSalle, American Museum of Natural History

The entered sequence is used to search a database of Tetraspanin genes organized into a phylogenetic tree. Based on the closest match, the query sequence is then checked for characteristic attributes which indicate specific classification.

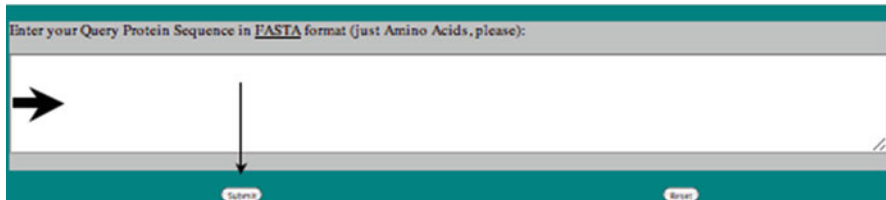
Limitations and important points:

1. This technique is not a replacement for phylogenetic analysis as it cannot describe the relationships between individual genes; it *only* places sequences in a groups based on the sequences used to create the phylogeny..
2. As new sequences become available some characteristic attributes may have to be altered or excluded altogether. Therefore, we will be constantly updating this database. Diagnostic positions with few possible character states that represent many taxa in the tree are the most reliable.

*Try this sequence:*

```
>MmRDS  
AQGLWLMNWL SVLAGIVL FSLGLFLKHEL RH FVPNSLIGVGVLS CVFNALDP AKYAKWKPWL  
KPYLAVCFHFNVILFLVALCCFLLRGSL ESTLAYGLKNGMKYYRDTDIPGRCKTIDMLQHIF  
KCCGNGFRDWFIEQVDGVPFSCCGCRAALLNYYSMNSMGVVTLVWLFVSVITAGLRYLHFA  
LESVSNPE
```

*Just "copy and paste" the full FASTA Amino Acid Sequence(s) into Query Box:*



**Fig. 2.5** Webpage showing the “gene identifier” on the TSPAN 4 website. The *large arrow* indicates the query box where a novel tetraspanin sequence can be pasted and the *small arrow* indicates the “Submit” button

tree (@ in Fig. 2.4) and to rapidly access the sequences that were used to generate the tree. The fourth function on the website allows users to access the figures (& in Fig. 2.4) produced for the various tetraspanin studies that were reviewed in this chapter. This page will also have the full references and link outs or pdfs for all of the papers relevant to tetraspanin evolution. The final function of the website is a list of all of the tetraspanin genes that are in fully sequenced genomes (\$ in Fig. 2.4).

## 2.8 Conclusion

A phylogenetic context for tetraspanins allows for a precise and logical classification system for this large superfamily of genes. Placing the tetraspanins in a phylogenetic context and using fossil dates for divergence of the model organism fully sequenced genomes allows us to establish the dates of origin for the major tetraspanin groups. In addition, the phylogenetic approach allows for a detailed examination of two phenomena involved in the structure and function of tetraspanins. First we can use the phylogenetic approach to examine the evolution of intron position in the tetraspanin families. Second we can use the tree-based approach to examine the evolution of the structurally and functionally important cysteine residues in the proteins in this superfamily. By examining very closely the evolution of members of a well known group of tetraspanins, the uroplakins, we demonstrate how the phylogenetic approach can be useful in the reconstruction of evolutionary events for a functionally important protein family.

## References

- Boucheix C, Rubinstein E (2001) Tetraspanins. *Cell Mol Life Sci* 58:1189–1205
- Chiu JC et al (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22:699–707
- DeSalle R, Mares R, Garcia-España A (2010) Evolution of cysteine patterns in the large extracellular loop of tetraspanins from animals, fungi, plants and single-celled eukaryotes. *Mol Phylogenet Evol* 56:486–491
- Doolittle RF et al (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271:470–477
- Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics* 12:99–124
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791
- Felsenstein J (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland
- Garcia-España A, DeSalle R (2009) Intron sliding in tetraspanins. *Commun Integr Biol* 2:1–3
- Garcia-España A, Chung PJ, Zhao X, Lee A, Pellicer A, Yu J, Sun TT, DeSalle R (2006) Origin of the tetraspanin uroplakins and their co-evolution with associated proteins: implications for uroplakin structure and function. *Mol Phylogenet Evol* 41:355–367
- Garcia-España A, Chung PJ, Sarkar IN, Stiner E, Sun TT, DeSalle R (2008) Appearance of new tetraspanin genes during vertebrate evolution. *Genomics* 91:326–334
- Garcia-España A, Mares R, Sun TT, DeSalle R (2009) Intron evolution: testing hypotheses of intron evolution using the phylogenomics of tetraspanins. *PLoS One* 4:e4680
- Hedges SB, Kumar S (2002) Genomics: vertebrate genomes compared. *Science* 297:1283–1285
- Hemler ME (2001) Specific tetraspanin functions. *J Cell Biol* 155:1103–1107
- Hemler ME (2003) Tetraspanin proteins mediate cellular penetration, invasion, and fusion events and define a novel type of membrane microdomain. *Annu Rev Cell Dev Biol* 19:397–422
- Hillis DM, Huelsenbeck JP (1992) Signal, noise and reliability in molecular phylogenetic analyses. *J Hered* 83:189–195
- Hu C et al (2005) Assembly of urothelial plaques: tetraspanin function in membrane protein trafficking. *Mol Biol Cell* 16:3937–3950

- Huang S et al (2005) The phylogenetic analysis of tetraspanins projects the evolution of cell-cell interactions from unicellular to multicellular organisms. *Genomics* 86:674–684
- Huelsenbeck JP, Ronquist F (2001) Bayesian, is a reference to Mr. Bayes ([http://en.wikipedia.org/wiki/Bayesian\\_inference](http://en.wikipedia.org/wiki/Bayesian_inference)). *Bioinformatics* 17:754–755
- Levy S, Shoham T (2005a) Protein-protein interactions in the tetraspanin web. *Physiology (Bethesda)* 20:218–224
- Levy S, Shoham T (2005b) The tetraspanin web modulates immune-signalling complexes. *Nat Rev Immunol* 5:136–148
- Maecker HT, Todd SC, Levy S (1997) The tetraspanin superfamily: molecular facilitators. *FASEB J* 11:428–442
- Tarrío R, Ayala FJ, Rodríguez-Trelles F (2008) Alternative splicing: a missing piece in the puzzle of intron gain. *Proc Natl Acad Sci USA* 105:7223–7228
- Thornton JW, DeSalle R (2000) Phylogenetics meets genomics: homology and evolution in gene families. *Annu Rev Genomics Hum Genet* 1:43–72
- Todres E, Nardi JB, Robertson HM (2000) The tetraspanin superfamily in insects. *Insect Mol Biol* 9:581–590
- Tu L, Sun TT, Kreibich G (2002) Specific heterodimer formation is a prerequisite for uroplakins to exit from the endoplasmic reticulum. *Mol Biol Cell* 13:4221–4230
- Zemni R et al (2000) A new gene involved in X-linked mental retardation identified by analysis of an X;2 balanced translocation. *Nat Genet* 24:167–170