Unexpected diversity in Shisa-like proteins suggests the importance of their roles as transmembrane adaptors

Jimin Pei [a,*], Nick V. Grishin [a,b]

[a] Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX, 75390, United States
[b] Department of Biochemistry, University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, TX, 75390, United States

ARTICLE INFO

ABSTRACT

The Shisa family of single-transmembrane proteins is characterized by an N-terminal cysteine-rich domain and a proline-rich C-terminal region. Its founding member, *Xenopus* Shisa, promotes head development by antagonizing Wnt and FGF signaling. Recently, a mouse brain-specific Shisa protein CKAMP44 (Shisa9) was shown to play an important role in AMPA receptor desensitization. We used sequence similarity searches against protein, genome and EST databases to study the evolutionary origin and phylogenetic distribution of Shisa homologs. In addition to nine Shisa subfamilies in vertebrates, we detected distantly related Shisa homologs that possess an N-terminal domain with six conserved cysteines. These Shisa-like proteins include FAM159 and KIAA1644 mainly from vertebrates, and members from various bilaterian invertebrates and Porifera, suggesting their presence in the last common ancestor of Metazoa. Shisa-like genes have undergone large expansions in *Branchiostoma floridae* and *Saccoglossus kowalevskii*, and appear to have been lost in certain insects. Pattern-based searches against eukaryotic proteomes also uncovered several other families of predicted single-transmembrane proteins with a similar cysteine-rich domain. We refer to these proteins (Shisa/Shisa-like, WBP1/VOPP1, CX, DUF2650, TMEM92, and CYYR1) as STMC6 proteins (single-transmembrane proteins with conserved 6 cysteines). STMC6 genes are widespread in Metazoa, with the human genome containing 17 members. Frequent occurrences of PY motifs in STMC6 proteins suggest that most of them could interact with WW-domain-containing proteins, such as the NEDD4 family E3 ubiquitin ligases, and could play critical roles in protein degradation and sorting. STMC6 proteins are likely transmembrane adaptors that regulate membrane proteins such as cell surface receptors.

1. Introduction

Adaptor or scaffold proteins are important components with diverse roles in cell signaling networks [1–2]. They are usually non-enzymes that contain modular domains and/or linear peptide motifs to mediate protein–protein interactions and/or protein–membrane associations [3–4]. Modular domains frequently found in adaptor proteins include SH3 [5] and WW domains [6] that recognize proline-rich peptide motifs, SH2 [7] and PTB [8] domains that recognize motifs with phosphorylated residues, PDZ domains [9] that bind specific C-terminal peptides, and PH domains [10] that recognize phosphoinositide lipids. Adaptor proteins regulate diverse aspects of cell surface receptor functions ranging from transducing signals to downstream components, attenuating signals in negative feedback loops, to the trafficking and recycling of receptors. An example of an adaptor protein is β-arrestin [11] that regulates G-protein-coupled receptor (GPCR) signaling. Agonist-bound GPCRs are subject to phosphorylation followed

by β-arrestin recruitment, resulting in the blockage of receptor-G protein coupling and thus receptor desensitization [12]. Furthermore, β-arrestin-bound GPCRs are targeted for clathrin-mediated endocytosis to promote their degradation or recycling. In addition to receptor desensitization, β-arrestins also serve as scaffold molecules that mediate interactions between GPCRs and downstream effectors in G protein-independent signaling events [13]. Like most characterized adaptor proteins, β-arrestins are soluble proteins residing in the cytosol. Some adaptors also have transmembrane domains. One group of transmembrane adaptor proteins (TRAPs), such as LAT, PAG and TRIM, is a key component in immunoreceptor signaling [14].

A recent study showed that overexpression of a brain-specific type I transmembrane protein CKAMP44 (cysteine-knot AMPAR modulating protein, 44 kD) enhanced AMPA receptor desensitization [15], suggesting that it is an adaptor protein. CKAMP44, also named Shisa9, is a member of the Shisa protein family. The founding member of this family is the *Xenopus* Shisa protein (designated later as *Xenopus* Shisa1) that can promote head formation in *Xenopus* development [16]. During amphibian embryonic development, the Spemann organizer produces a number of secreted proteins, such as Noggin, Chordin and Cerberus, that antagonize BMP and Wnt signaling to promote the

* Corresponding author. Tel.: +1 214 645 5951; fax: +1 214 645 5948.
E-mail addresses: jpei@chop.swmed.edu (J. Pei), grishin@chop.swmed.edu (N.V. Grishin).

formation of anterior neuroectoderm (head) and induce dorsoventral differentiation in a non-cell-autonomous fashion [17–18]. On the other hand, *Xenopus* Shisa1 promotes head formation cell-autonomously by retention of Wnt [19–20] and FGF [21] receptors in the endoplasmic reticulum (ER) to inhibit their maturation [16].

Several vertebrate homologs of the original *Xenopus* Shisa1 protein have been cloned and studied, with implicated roles in development, cancer and apoptosis. For example, *Xenopus* Shisa2 was shown to function in segmental patterning during somitogenesis [22]. Expression of Shisa2 in chicken embryos exhibited an anterior pattern, suggesting its role in development [23]. High expression levels of human Shisa2 were found in certain cancer cell lines and its overexpression increased cell growth and invasion [24]. Five mouse homologs of Shisa (Shisa1–5) were shown to have distinct developmental expression patterns, and mouse Shisa2 was also shown to be an antagonist of Wnt and FGF signaling [25]. Mouse Shisa5, also named Scotin, was independently identified as a downstream pro-apoptotic protein of the P53 family proteins [26–28].

Shisa proteins have not been characterized outside vertebrates, and their evolutionary origin remains unclear. Using sensitive sequence similarity searches, we found distant homologs of Shisa proteins (called Shisa-like proteins) within and beyond vertebrates. Noticeably, we found EST evidence of Shisa-like genes in Porifera, suggesting a Metazoa origin for them. Further computational analysis suggests that Shisa and Shisa-like proteins belong to a large and diverse group of predicted single-transmembrane proteins that have undergone large expansions in certain metazoan lineages. These proteins possibly play various roles in regulating the functions of membrane proteins such as cell surface receptors.

2. Materials and methods

2.1. Sequence similarity searches against the non-redundant protein database

PSI-BLAST [29] was used to search for homologs of the Shisa family starting with the mouse Shisa2 protein (NCBI gene identification (gi) number: 21703918; range: 23–137, which excludes the predicted signal peptide and C-terminal low complexity region) against the NCBI non-redundant (nr) protein database (e-value inclusion cutoff: 1e−4). To perform transitive searches, the protein hits found by PSI-BLAST 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. In practice, we noticed that some of the distant Shisa homologs could not be found with e-values less than 1e−4 when composition-based statistics were used. To increase the coverage of Shisa homologs, composition-based statistics was disabled (−t 0) in separate transitive PSI-BLAST searches while the e-value inclusion cutoff was set to be more stringent (1e−6 in the first three rounds of transitive PSI-BLAST and 5e−6 in the fourth round of transitive PSI-BLAST searches, these values were selected based on trials of different cutoff values and manual inspections of profile corruption). The results were manually checked and no false positives were found in the first three rounds of transitive PSI-BLAST searches. A few false positives found in the fourth round of transitive PSI-BLAST searches were removed (false positives were judged by the lack of the cysteine motif C*C*CC*CC in Shisa-like proteins and the fact that PSI-BLAST searches starting from them could not reciprocally find Shisa or Shisa-like proteins). A similar approach was used to search for homologs of the WBP1/VOPP1, CX, DUF2650, TMEM92 and CYYR1 family proteins. The HHpred web server [30] was used for profile-against-profile-based similarity searches using several members found by PSI-BLAST against the Pfam [31] and human proteome databases.

2.2. Sequence similarity searches against EST and genome databases

To increase the coverage of homology detection for Shisa-like proteins and WBP1/VOPP1 homologs, especially in lophotrochozoans where not many protein records were available, we used the NCBI TBLASTN web server to search against the EST database (e-value cutoff: 10) starting with several Shisa and Shisa-like proteins. For each TBLASTN hit, the longest translated amino acid sequence segment covering the hit region and without stop codons was obtained. Sequences without the motif 'C*C*CC*CC' were discarded. The remaining sequences were subject to multiple sequence alignment by MAFFT [32]. The resulting alignment was manually inspected to remove likely false positives (e.g., some cysteine-rich fragments with a lot more cysteines than in Shisa-like proteins). Sequence redundancy was reduced by keeping one sequence for any group of ESTs that appear to have been transcribed from the same gene.

Translated sequences of several plant ESTs and *Plasmodium vivax* ESTs were found to show high similarity to Shisa or Shisa-like proteins. However, these ESTs cannot find high similarity hits in the whole genome sequences of the corresponding plant or *Plasmodium* species. Instead, the transcripts of *Plasmodium* EST sequences (e.g., gi: 260703421) show high similarity (>99%) to a human Shisa5 protein. Two plant ESTs (gi: 76604539 from *Euphorbia esula* and gi: 119017255 from *Manihot esculenta*) show high similarity (99% sequence identity) to an EST from a Platyhelminth species (gi: 84601851 from *Schmidtea mediterranea*). Therefore, the ESTs from plants and *Plasmodium* were considered as contaminations.

To further investigate the species distribution of Shisa subfamilies, TBLASTN was used to search against the NCBI whole genomes, whole-genome shotgun reads and high throughput genomic sequences. BLAT searches were performed in the UCSC genome bioinformatics server [33] against the lamprey genome sequences and TBLASTN searches were performed against the elephant shark genome sequences [34].

2.3. Pattern-based searches in eukaryotic proteomes

To find proteins with a similar domain structure to Shisa-like proteins, we performed pattern-based searches against a number of eukaryotic proteomes sampling various taxa (*Apis mellifera*, *Arabidopsis thaliana*, *Branchiostoma floridae*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Homo sapiens*, *Hydra magnipapillata*, *Monosiga brevicollis*, *Naegleria gruberi*, *Nematostella vectensis*, *Paramecium tetraurelia*, *Saccharomyces cerevisiae*, *Saccoglossus kowalevskii*, *Tetrahymena thermophila*, *Tribolium castaneum* and *Trichoplax adhaerens*). The proteins from these genomes were downloaded from the NCBI genome database. For each proteome, we predicted TMs for its proteins using Phobius [35] and selected those proteins with a single predicted TM. For these predicted single-TM proteins, we identified proteins with a 'C*C*CC*CC' pattern (matching the regular expression of 'C.{2,40}C.{2,40}CC.{2,40}CC') in the sequence segment N-terminal to the predicted TM and excluding any predicted signal peptide. The results were manually inspected to select proteins with a cysteine-rich domain similar to Shisa-like proteins. We found that Shisa and Shisa-like proteins, as well as some other proteins such as WBP1 and TMEM92, usually have a short N-terminal segment before the predict TM (less than 100 residues) with no more than eight cysteines in the N-terminal segment before the predict TM. On the other hand, some proteins identified by pattern match searches have a much longer N-terminal region before the predicted TM with many more cysteines than the six cysteines in the 'C*C*CC*CC' motif (one example is gi: 4557251, ADAM metallopeptidase domain 10). We thus only considered those proteins with no more than eight cysteines in the N-terminal segment as potentially related to Shisa and Shisa-like

SignalP | **Cysteine-rich domain** | **TM** | **C** R/K | **Proline-rich region**

Fig. 1. Domain structure of Shisa and Shisa-like proteins. These proteins usually have a signal peptide (labeled as 'SignalP'), an N-terminal cysteine-rich domain with a distinct cysteine pattern ('C*C*CC*C*CC*C' for Shisa proteins and 'C*C*CC*CC' for Shisa-like proteins), a single predicted transmembrane segment (labeled as 'TM') and a C-terminal proline-rich low complexity region predicted to be largely disordered. Cysteines (labeled as 'C') and positively charged residues (labeled as 'R/K') frequently occur at the C-terminal end of the predicted TM.

proteins. The results of pattern-based searches in selected eukaryotic proteomes are available in Supplementary Figure S14.

To expand the searches against other eukaryotes, we downloaded RefSeq sequences in the following organism categories: fungi, invertebrate, plant, protozoa, vertebrate_mammalian and vertebrate_other (these categories are defined in ftp://ftp.ncbi.nlm.nih.gov/refseq/release). We used the same pattern-based searches on these protein datasets. The results are shown in Supplementary Figure S15.

2.4. Sequence alignment and phylogenetic reconstruction

The multiple sequence alignment for each protein family was made by PROMALS [36]. These alignments were then improved by manual curation. The MOLPHY package [37] was used for phylogenetic

reconstruction based on the alignment of Shisa proteins. The JTT amino acid substitution model [38] was used in MOLPHY. The local estimates of bootstrap percentages were obtained by the RELL method [39] (−R option in the ProtML program of MOLPHY). A phylogenetic reconstruction for WBP1/VOPP1 proteins was conducted in the same way.

3. Results

3.1. The Shisa family proteins in vertebrates

A previous study reported five vertebrate Shisa subfamilies (Shisa1–5) that exhibit varying distributions in mammals, chicken, *Xenopus* and zebrafish [25]. Four additional Shisa subfamilies (Shisa6–9) were found in vertebrates in sequence similarity searches by PSI-BLAST [29] (see Materials and methods). They have a common domain architecture (Fig. 1) characterized by a predicted signal peptide, an N-terminal cysteine-rich domain with eight conserved cysteines [40] exhibiting a distinct pattern ('C*C*CC*C*CC*C', '*' represents a stretch of amino acid residues, Fig. 2), a predicted transmembrane segment (TM) and a proline-rich C-terminal region. For Shisa proteins from subfamilies 1–5, several cysteine residues are present near the C-termini of their predicted TMs (Fig. 2), possibly providing sites for lipid modifications such as palmitoylation to facilitate their localization to specific membrane microdomains [41]. Positively charged residues frequently occur near the C-termini of their

Fig. 2. Multiple sequence alignment of Shisa proteins. The cysteine-rich domains and transmembrane regions of representative Shisa proteins (with a 'C*C*CC*C*CC*C' pattern) from nine vertebrate Shisa subfamilies (Shisa1–9) and one sequence from *Branchiostoma floridae* (Shisa_Bf) are included in this alignment. NCBI accession numbers are shown. A few proteins were derived from genome or EST sequences (accession numbers in italic and underlined). Starting and ending residues numbers for sequences with protein accession numbers are shown before and after the sequences respectively. Protein lengths are shown in brackets. Conserved cysteines are shaded in yellow or red (the additional two cysteine positions compared to Shisa-like proteins). Non-charged residues in positions with mainly hydrophobic residues were shaded in cyan. Prolines and glycines are shown in red letters. Cysteines within and after the predicted transmembrane segments are shaded in gray. Arginines and lysines, occurring frequently C-terminal to the predicted transmembrane segments, are shown in blue letters. Species name abbreviations shown after the accession numbers are as follows: Bf, *Branchiostoma floridae*; Dr, *Danio rerio*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Om, *Oncorhynchus mykiss*; Oa, *Ornithorhynchus anatinus*; Xt: *Xenopus tropicalis*; and Xl, *Xenopus laevis*.

predicted TMs and may help determine the topology of Shisa proteins [42]. They are predicted to be type I transmembrane proteins with their proline-rich, mostly disordered C-terminal regions residing in the cytosol.

Seven of the nine Shisa subfamilies (Shisa2–7 and Shisa9) have orthologous proteins in mammals and zebrafish (Fig. 2), suggesting that they were present in the last common ancestor of vertebrates. The evolutionary origin of the first reported Shisa in *Xenopus laevis* (belonging to the Shisa1 subfamily) [16] has been unclear [25]. In a previous study Shisa1 was not identified in amniotes, and only one putative ortholog was found in zebrafish [25]. We found EST evidence of Shisa1 in another fish *Oncorhynchus mykiss* (gi: 24599333). Moreover, we found genomic sequences harboring putative Shisa1 genes in birds and the elephant shark *Callorhinchus milii* (Fig. 2 and Supplementary Figure S1), suggesting that the Shisa1 subfamily, despite apparent absence from mammals, has a deeper and wider distribution in vertebrates than previously thought [25]. On the other hand, the Shisa8 subfamily is restricted to be in mammals (including one in the basal mammalian species *Ornithorhynchus anatinus*, gi: 149457717). Shisa8 appears to be closely related to Shisa9 (Fig. 3) and could originate from a mammalian-specific gene duplication of Shisa9. Alternatively, the Shisa8 gene could be present in the last common ancestor of vertebrates, but was lost in vertebrate lineages other than mammals. Subfamily-specific gene duplications or gene losses were observed. For example, *Xenopus* appears to lack a Shisa5 ortholog while zebrafish has several Shisa5 proteins (Fig. 2). *Xenopus* also possesses a distantly related Shisa homolog with a pair of cysteines mutated (Supplementary Figure S1, discussed below).

A phylogenetic tree reconstructed by the MOLPHY program [37] supports three major clades among the nine Shisa subfamilies (Shisa1–3, Shisa4–5 and Shisa6–9, Fig. 3). Subfamilies Shisa1–3 form a monophyletic group with a good bootstrap support (Fig. 3). Compared with other Shisa subfamilies, they share several sequence features such as longer insertions between the second and third conserved cysteines in the cysteines-rich domain (Fig. 2) and a 'PxxxP' signature at the beginning of their predicted TMs (Fig. 2). The cysteine patterns at the end of their predicted TMs can serve as subfamily-specific signatures, with 'CCC[KQ]C', 'CCCRC' and 'YCCTC' motifs in Shisa1, Shisa2 and Shisa3 respectively. Subfamilies Shisa6–9 are more closely related to each other than to other Shisa subfamilies (Fig. 3). Unlike other subfamilies, Shisa6–9 lack cysteines near the C-termini of their predicted TMs (Fig. 2), suggesting a loss of lipid modification sites.

We found genomic and/or EST evidence of several Shisa family members in the basal vertebrate lamprey and several Chondrichthyes species (cartilaginous fishes) (Supplementary Figure S1). Among other deuterostomes, only one Shisa protein was found in the amphioxus *B. floridae* (Fig. 2), which belongs to the clade containing subfamilies Shisa4–5 (Fig. 3). The presence of a Shisa protein in *B. floridae* suggests that Shisa proteins probably originate in the ancestor of chordates.

3.2. FAM159 and KIAA1644: two Shisa-like protein families distantly related to Shisa proteins

PSI-BLAST searches detected two protein families FAM159 and KIAA1644 with vertebrate members that are distantly related to the Shisa family proteins (Fig. 4 and Supplementary Figure S2). Proteins in these families share a similar domain architecture to Shisa proteins (Fig. 1), except that FAM159 proteins lack a predicted signal peptide. Functions of these families have not been characterized. Two copies of FAM159 proteins exist in vertebrates (FAM159A and FAM159B, Supplementary Figure S2). While FAM159 proteins were only found in vertebrates, KIAA1644 family members are present in vertebrates and tunicates such as *C. intestinalis*, suggesting an Olfactore origin of KIAA1644 proteins. There are also two KIAA1644 copies in amniotes (Fig. 4 and Supplementary Figure S2).

The cysteine-rich domains in FAM159 and KIAA1644 contain six highly conserved cysteines ('C*C*CC*CC', '*' represents a stretch of amino acid residues) as compared to eight conserved cysteines in the Shisa family proteins (Fig. 4). The extra two cysteines in the Shisa family proteins lie in between the two 'CC's and after the second 'CC', respectively, giving rise to a signature of 'C*C*CC*C*CC*C' (the extra cysteines underlined). Most Shisa homologs outside vertebrates (Fig. 4 and described below) show the six-cysteine pattern similar to KIAA1644 and FAM159 while lacking the extra cysteine pairs in vertebrate Shisa proteins. Therefore, it is likely that Shisa proteins have evolved the extra pair of cysteines that possibly form a new disulfide bond to increase the stability of the cysteine-rich domain. To differentiate the proteins based on their cysteine patterns, we refer to the Shisa homologs with six conserved cysteines and without the specific cysteine pairs in Shisa proteins as 'Shisa-like proteins'. Several positions with mainly hydrophobic residues are also conserved among Shisa and Shisa-like proteins (Fig. 4), supporting the inferred homologous relationships between them. Most noticeably, each of the two 'CC' signatures is often preceded by an aromatic residue (tyrosine being the most frequent) (Fig. 4).

Fig. 3. A phylogenetic tree of Shisa proteins by MOLPHY. Protein names are shown as Shisa subfamily names followed by species name abbreviations defined in Fig. 2. Branch support values above 90 are shown in red numbers.

3.3. Shisa-like proteins in deuterostomes other than vertebrates

In addition to FAM159 and KIAA1644 mainly found in vertebrates, we detected Shisa-like proteins in other deuterostomes by PSI-BLAST.
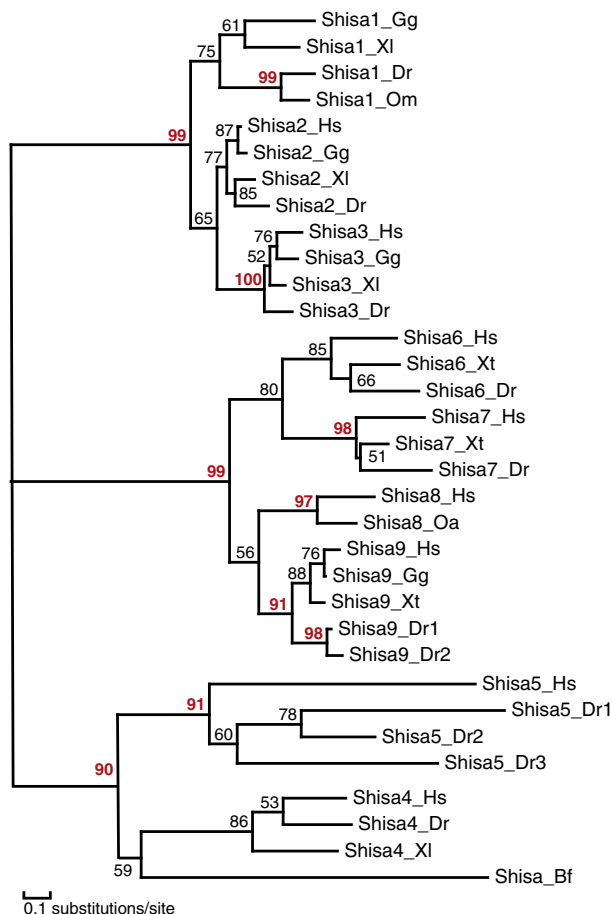
Shisa family proteins:

```
NP_001085264.1  Xl  22  EYCHGWTDSYGIWRPGFQCPERYDPSEATFCCGSCG-----LKYCCSTVESRLDQGLCP   75 [269]
NP_001007539.1  Hs  36  EYCHGWLDAQGVWRIGFQCPERFDGGDATICCGSCA-----LRYCCSSAEARLDQGGCD   89 [295]
NP_001073974.1  Hs  26  EYCHGWVDVQGNYHEGFQCPEDFDTLDATICCGSCA-----LRYCCAAADARLEQGGCT   79 [238]
NP_937792.2     Hs  29  EDCLWYLDRNGSWHPGFNCE------FFTFCCGTCY-----HRYCCRDLTLLI(6)HCL   79 [197]
NP_057563.3     Hs  29  EVCMASR---GLSLFPESCP------DFCCGTCD----DQYCCSDVLKKF(5)RCA   73 [240]
NP_997269.2     Hs  92  ETCWGYYDVSGQYDKEFECNNS--ESGYLYCCGTCY-----YRFCCKKRHEKLDQRQCT  143 [551]
NP_001138648.1  Hs  80  ELCHGYYDVMGQYDATFNCST----GSYRFCCGTCH-----YRFCCEHRHMRLAQASCS  129 [538]
XP_003121062.1  Hs  58  DRCRGYYDVMGQWDPPFNCSS----GAYSFCCGTCG-----YRFCCHDGPRRLDQSRCS  107 [547]
NP_001138676.2  Hs  72  DFCRGYFDVMGQWDPPFNCSS----GDFIFCCGTCG-----FRFCCTFKKRRLNQSTCT  121 [424]
XP_002597311.1  Bf  16  RTCSPFYNPAGIYVPEFSCDD----DVYPYCCGVCD-----NKYCCRDNSSLIS-TTCD   64 [197]
```

FAM159:

```
NP_001157914.1  Hs   6  RLCSGYYSLNQSFVEPFQCPRRGEGAALQYCCGFAD-----LKYCCSEPGSYFPYKHSY   59 [160]
XP_002934242.1  Xt   7  ESCSSYYTANNTFVDTFGCPRKEDGSDLIYCCGFVD-----LKYCCNEPGNYFPYKHGY   60 [159]
NP_001036158.1  Hs   3  GACTSYVSAEQEVVRGFSCPRPGGEAAAVFCCGFRD-----HKYCCDDPHSFFPYEHSY   56 [190]
XP_002931467.1  Xt   3  AECSSYTTAEKVVVGAFSCPKADSDMGAIYCCGFKD-----IKYCCDDPNSFFPYEHSY   56 [168]
XP_422478.1     Gg   3  SECSSYLNADRVLVSGFSCPRVGGDARAVYCCGFQD-----VKYCCDDPHSFFPYEHSY   56 [117]
```

KIAA1644:

```
NP_001092764.1  Hs  29  RVCEPYTDHKGRYHFGFHCPRLSDNKTFILCCHHNNT---VFKYCCNETEFQAVMQANL   84 [199]
NP_001025312.1  Dr  29  RVCEPYSDHKGRYHFGFHCPRLSDNKTYIFCCHHNNT---AFKYCCNETEFQSVMQLNL   84 [207]
BI560596.1      Hs      HLCQTTRDVHGRHYPGFFCPRLSDSPEEAYCCHLQA----AGGSCCTRAEFEALYQVNL
XP_002187587.1  Tg  29  HLCEGYVGPDGRSHPGFYCPRLTDPPGHRYCCQPSLD---ALKSCCSQVALEALTGVNL   84 [163]
XP_002129289.1  Ci  25  EICEDIVNLKGEMLQAFGCPRLHEMKENKYCCGNKS----VSLTCCSNETQSIYYAAKM   79 [126]
```

Shisa-like proteins from deuterostomes:

```
XP_002597315.1  Bf  40  EVCEEYER(6)TLVDSFSCPTNSDQDKDSYCCMAGT-----LKYCCDDCTQSLT-LSCA   95 [246]
XP_002597320.1  Bf  20  EQCRAF-----GGLTTFNCPRPL(4)DDIYCCDN---------DCCDDCWESD(4)ACT   66 [178]
XP_002585700.1  Bf  29  ERCGS----------FSCPRSSDDSDEKYCCGSS--------QCCNDCGRSTV-SSCP   67 [209]
XP_002606575.1  Bf  18  EYCSGYTDAFGDYNSGFSCPRNS(9)DDIYCCGTST-----YKYCCSWRSGG-------   70 [230]
XP_002731303.1  Sk  34  GYCESYYNIYGVEQQGFYCPQEYDLPSETYCCGEG------FRYCCDYDTYQSSSKVNL   86 [218]
XP_002731319.1  Sk  17  YYCPGFLDNAGVWRSGFFCPLEYNSPNYMYCCGQTY-----TRYCCDYDDHQREQDTSN   70 [266]
XP_002731321.1  Sk  27  DYCSGYWDDYNTYQFGFLCD--------SYCCGSDT-----DKYCCSIDPDGIDDLFNS   72 [201]
XP_799962.2     Sp  21  EFCEAYVDESGTYHTSSHCPGPD(6)NNVYCCGTST-----NKYCCSEPTGYFLEH---   74 [163]
AM159798.1      Xb      NYCED----------DLDCTDR---MFETYCCVEEG-------ECCDYDDYLDEWDD--
```

Shisa-like proteins from protostomes:

```
DB744702.1      Am      MECSFGNKN-IISKVIDSCPGILDSSDKSYCCYDFEK---NEMYCCDAVEFASKS----
FE424659.1      Dp      KMCKRKSIF--GVKSYFTCPEPGQTPDMVFCCGDDDD--LGGQFCCRGLTIDNIGS---
EW910016.1      Is      EYCYGYTDKFGKWNNGFPQPQS-SGEPVYCCGTA-----TEPYCCRKKDDDPPMPVHD
CD449869.1      Hd      MECGIA---------GTCPNIVSDSSKTFCCKTET----VTIPYCCDAGDYAKFAFREA
EFV55384.1      Ts  17  RTCPNG----------RDCPGYGDSLDEFYCCDSTVQ--PGDFYCCDYGTKEKTSYRGL   63 [226]
FN882036.1      Ce      FWCPSGY------GYLAFCPQPTDWDNYNWCCTFPYM-GSWKPSCCQFAIP--------
CV738029.1      Sj      EVCEGIPS(9)QTYSHFLCPDNPNQLFEKYCCWDPT---TGMGACCSYDVKMYSTSGRG
FC773981.1      Lg      TICTNNFD(5)TTYGYFICPRAGEPEDHEYCCGDD-----KREHCCEYYNQTGRIV---
EY296348.1      Hr      TICRNEFD(6)THYEWFICPSQNHTSDFHYCCGQP-----DKEYCCRFADNGWRV----
FM925299.1      Bp      NLCVNNEDFNGTKFGSFYCPLPGFNPAAKYCCGEP-----ETQYCCEYFDDNGRK----
```

Shisa-like proteins from Porifera:

```
GW181182.1      Aq      SDCS------------VTCPSSNDASYKTYCCSDSSSMNNNNNNYCCNPYL---------
GH571479.1      Sd      ETCEKF---------SLNCGRS---FSQKYCCFELNG-PYPIHTCCNPYI---------
```

Fig. 4. Multiple sequence alignment of cysteine-rich domains from Shisa proteins and Shisa-like proteins. This alignment is formatted and colored in the same way as in Fig. 2. Shisa proteins and Shisa-like proteins are separated by a horizontal line. Species name abbreviations shown after the accession numbers are as follows: Aq, *Amphimedon queenslandica*; Am, *Apis mellifera*; Bp, *Brachionus plicatilis*; Bf, *Branchiostoma floridae*; Ce, *Caenorhabditis elegans*; Ci, *Ciona intestinalis*; Dr, *Danio rerio*; Dp, *Daphnia pulex*; Gg, *Gallus gallus*; Hr, *Helobdella robusta*; Hs, *Homo sapiens*; Hd, *Hypsibius dujardini*; Is, *Ixodes scapularis*; Lg, *Lottia gigantea*; Sk, *Saccoglossus kowalevskii*; Sj, *Schistosoma japonicum*; Sp, *Strongylocentrotus purpuratus*; Sd, *Suberites domuncula*; Tg, *Taeniopygia guttata*; Ts, *Trichinella spiralis*; Xt, *Xenopus tropicalis*; Xl, *Xenopus laevis* and Xb, *Xenoturbella bocki*.

Noticeably, Shisa-like genes have undergone large expansions in the cephalochordate *B. floridae* [43] and the hemichordate *S. kowalevskii* [44], which possess at least 15 and 14 copies of these genes, respectively (Supplementary Figure S2). Most of the *B. floridae* Shisa-like proteins have an extra pair of cysteines after the second 'CC', giving rise to a pattern of 'C*C*CC*CC*C*C' (the extra cysteines underlined). Such a pattern is not observed in other deuterostome Shisa-like proteins, suggesting they originated and expanded in the amphioxus. In contrast, only one Shisa-like protein was found in the echinoderm *Strongylocentrotus purpuratus* [45]. We also identified EST evidence of several Shisa-like proteins with tandem cysteine-rich domains in *S. kowalevskii*, *Xenoturbella bocki* [46], and two tunicates *Molgula tectiformis* and *Botryllus schlosseri* (Supplementary Figure S3).

3.4. Shisa-like proteins in protostomes

PSI-BLAST searches identified Shisa-like proteins from several protostomes such as insects, nematodes, and flatworm (Platyhelminthes)

species from the *Schistosoma* genus. TBLASTN searches against EST databases using these proteins as queries revealed that Shisa-like proteins are widely distributed in various protostomes, including both ecdysozoans and lophotrochozoans.

Shisa-like proteins derived from ESTs were detected in the Arthropoda phylum (including species from insects, crustaceans and chelicerates) and two other ecdysozoan phyla Tardigrada and Nematoda (Supplementary Figure S4). Various EST hits in Pterygota insects were found, including one species from the Ephemeroptera order (e.g., mayflies) and quite a few species from the Neoptera orders such as Hymenoptera (e.g., bees and ants), Orthoptera (e.g., grasshoppers) and Hemiptera (e.g., true bugs) (Supplementary Figure S4). Among the Endopterygota (Holometabola) orders, Shisa-like proteins are only present in Hymenoptera and are absent from Coleoptera (e.g., beetles), Lepidoptera (e.g., butterflies and moths) and Diptera (e.g., flies) (Supplementary Figure S5). For example, no significant hits of Shisa-like sequences were found in species such as *D. melanogaster* (a dipteran) [47], *T. castaneum* (a coleopteran) [48] and *Bombyx mori*

(a lepidopteran) [49] in BLAST or TBLASTN searches against NCBI RefSeq protein sequences, EST databases and whole genome sequences. Recent phylogenetic studies (e.g., [50]) supported that Hymenoptera is the sister group to other Endopterygota insects. Therefore, a single gene loss event in the last common ancestor of Endopterygota orders other than Hymenoptera could explain the lack of Shisa-like proteins in these orders (Supplementary Figure S5).

EST evidence of Shisa-like proteins was found in several lophotrochozoan phyla such as Annelida, Mollusca, Platyhelminthes and Rotifera (Supplementary Figure S4). Multiple copies of Shisa-like proteins were detected in certain lophotrochozoan species, such as *Schmidtea mediterranea* and *Crassostrea gigas* (Supplementary Figure S4). The cysteine-rich domain in most of the protostomian Shisa-like proteins has six conserved cysteines (C*C*CC*CC) similar to FAM159 and KIAA1644, while lacking the extra pair of cysteines in vertebrate Shisa proteins. A few Shisa-like proteins in Mollusca and Platyhelminthes have an extra pair of cysteines after the second CC signature, giving rise to a similar pattern (C*C*CC*CC*C*C) to some *B. floridae* Shisa-like proteins (Supplementary Figure S4).

EST searches retrieved five Shisa homologs in the Rotifera species *Brachionus plicatilis*. One of them has six cysteines in the N-terminal domain, characteristic of Shisa-like proteins (C*C*CC*CC), while the other four sequences match the motif of vertebrate Shisa proteins with eight cysteines (C*C*CC*C*CC*C) (Supplementary Figures S1 and S4). These four *B. plicatilis* sequences are quite interesting as no Shisa-specific cysteine patterns (C*C*CC*C*CC*C) were detected in other protostomian species. *B. plicatilis* Shisa proteins have predicted TMs quite different from those observed in any subfamily of Shisa proteins from chordates (Supplementary Figure S1). Besides, two of the four *B. plicatilis* Shisa proteins possess two tandem cysteine-rich domains. Such tandem repeats of cysteine-rich domains were not observed in any Shisa proteins from chordates. Thus, they are unlikely to be contaminations of chordate cDNAs in the *B. plicatilis* EST library. In one possible evolutionary scenario, Shisa proteins (with the C*C*CC*C*CC*C pattern) were present in the common ancestor of Bilateria, but were subsequently lost in all lineages but chordates and Rotifera. A more parsimonious scenario is that *B. plicatilis* acquired Shisa genes through horizontal transfer from certain chordate species, and these genes have since undergone divergent evolution such as domain duplications.

3.5. Shisa-like proteins in Porifera

A TBLASTN search using the *A. mellifera* Shisa-like protein (gi: 110760589) as the query against EST database found a weak hit to an EST (gi: 282462982) from the Porifera (sponge) species *Amphimedon queenslandica* (e-value: 0.68, with five out of the six conserved cysteines in the cysteine-rich domain and the predicted TM aligned). The translated sequence of this EST sequence has the same domain architecture as Shisa-like proteins. TBLASTN using this translated sequence as query found an EST from a different Porifera species (gi: 300475707 from *Suberites domuncula*, Fig. 4) with an e-value of 0.002. These results suggest that Shisa-like proteins were present in the last common ancestor of metazoans.

Aside from apparent contaminations (see Materials and methods), no protein or EST evidence of Shisa-like proteins was found outside metazoans. We also did not detect Shisa-like genes in Cnidaria or Placozoa, suggesting that they were lost in these metazoan lineages.

3.6. Prediction of disulfide bond connectivity in Shisa-like proteins

Conserved cysteines are often involved in disulfide bond formation. A newly added conserved pair of cysteines could form a disulfide bond, as in the case of Shisa proteins with eight cysteines compared to Shisa-like proteins with six cysteines. Similarly, concerted mutations in an otherwise conserved pair of cysteine positions suggest disulfide bond connectivity between them [51]. Several divergent homologs of Shisa or Shisa-like proteins have one pair of cysteines mutated, suggesting possible connectivity between the mutated positions (Fig. 5). A few Mollusca Shisa-like proteins have the first and third cysteines in C*C*CC*CC mutated, suggesting that they form a disulfide bond. In a divergent Shisa homolog from *Xenopus tropicalis* (gi: 113205932) and one cysteine-rich domain of a *B. plicatilis* Shisa protein (gi: 225249116) (Supplementary Figure S1), the fourth and sixth cysteines in the general C*C*CC*CC pattern were mutated, suggesting a disulfide bond between them. Thus the inferred cysteine connectivity for the six conserved cysteines would be 1–3, 2–5 and 4–6 (Fig. 5).

Manual inspection of the results of a regular expression search of current protein structure database revealed two structural folds with the C*C*CC*CC pattern (a few other hits also encompassing this pattern were not considered since they have a lot more conserved cysteines). One structural fold contains penaeidins, which are antimicrobial peptides found in shrimps [52] (protein databank (pdb) id: 1ueo). They have six conserved cysteines following the C*C*CC*CC pattern and exhibit a 1–3, 2–5, and 4–6 disulfide bond connectivity that coincides with the inferred connectivity pattern for Shisa and Shisa-like proteins (Fig. 5). The other structural fold consists of granulin domains [53] (pdb id: 2jye), which have a palindromic cysteine pattern of 'C*C*CC*CC*CC*CC*C*C', with two subdomains exhibiting a mirror symmetry of structures and cysteine patterns (Fig. 5). The first subdomain with a C*C*CC*CC pattern also has the 1–3 and 2–5 cysteine connectivity. The fourth and sixth cysteines of the first subdomain make a disulfide bond with the 6′ and 4′ cysteines of the second subdomain respectively, and vice versa (Fig. 5). The first subdomain of granulin can also form a folding unit as its structure alone is available [54] (pdb id: 1g26). In this case, the fourth and sixth cysteines in this subdomain are spatially close and can potentially form a disulfide bond. It should be noticed that penaeidin and the first subdomain of granulin have different structural folds despite sharing the same cysteine pattern and similar connectivity, suggesting that sequences with a C*C*CC*CC pattern, such as Shisa-like proteins and the WBP1/VOPP1 proteins (described below) could also adopt different 3-dimensional structural folds. No statistically significant sequence similarities were found between Shisa-like proteins and proteins with known structures. Structural determination of the cysteine-rich domains in Shisa and Shisa-like proteins could reveal possible structural similarities and homologous relationships to other proteins.

3.7. Additional protein families with Shisa-like domain architectures — defining the STMC6 proteins

We searched a number of eukaryotic proteomes for predicted single-transmembrane proteins that possess an N-terminal cysteine-rich domain with a C*C*CC*CC pattern similar to the one in Shisa-like proteins (see Materials and methods). Several additional families were uncovered, including WBP1/VOPP1, CX, DUF2650, TMEM92, and CYYR1 (Fig. 6 and Supplementary Figures S6–S11). Together with Shisa and Shisa-like proteins, we refer to them as STMC6 proteins (single-transmembrane proteins with conserved 6 cysteines) based on the shared cysteine pattern and the common domain structure.

Sequence similarities among the WBP1/VOPP1, TMEM92, CX and DUF2650 families were found by PSI-BLAST searches with statistical supports, suggesting that members of these four families are homologous. For example, a CX family member (gi: 268555046) found a WBP1/VOPP1 family member (gi: 169234886) with an e-value of 7e−5 in the seventh PSI-BLAST iteration (e-value inclusion threshold: 1e−4, no composition-based statistics). Another CX family member (gi: 253683499) found a TMEM92 protein (gi: 73966363) with an e-value
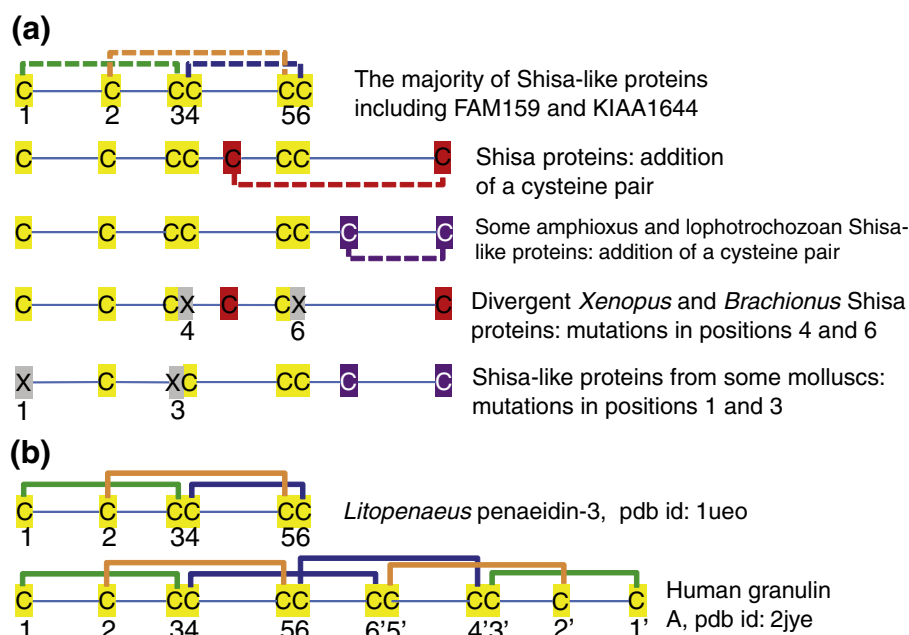
Fig. 5. Cysteine patterns and disulfide bond connectivity predictions for Shisa and Shisa-like proteins.(a) The six conserved cysteines in Shisa-like proteins are labeled from 1 to 6. Cysteine pairs with inferred disulfide bond connectivity are connected by dashed lines. Mutations in the cysteine positions are marked by 'X's. (b) Disulfide bond connections are shown for two protein domains with known structures and containing the cysteine pattern 'C*C*CC*CC'.

of 1e−5 in the third PSI-BLAST iteration (e-value inclusion threshold: 1e−4, no composition-based statistics). A DUF2650 member (gi: 115532536) found a TMEM92 member (gi: 269954664) with an e-value of 8e−04 in the seventh PSI-BLAST iteration (e-value inclusion threshold: 1e−4, no composition-based statistics). Distant homologous relationships among WBP1/VOPP1, CX, DUF2650 and TMEM92 were also supported by the sensitive profile-profile-based HHpred searches [30]. For example, an HHpred search using a CX protein (gi: 17557266; range: 132–191, covering the cysteine-rich domain and the predicted TM) as the query against the Pfam and human proteome databases found the Pfam domain DUF2650, the human VOPP1 protein and the human TMEM92 protein with probability scores of 96.38, 91.98 and 86.89, respectively. An HHpred search using a TMEM92 protein (gi: 253683499; range: 28–85, covering the cysteine-rich domain and the predicted TM) as the query against the Pfam and human proteome databases found Pfam domains WBP-1 and DUF2650 with probability scores of 96.75 and 93.62, respectively.

The WBP1/VOPP1 family proteins include vertebrate proteins WBP1 (WW domain-binding protein 1) [55], its close homolog OPAL1 (outcome predictor in acute leukemia 1) [56], and a less similar protein VOPP1 (Vesicular Over-expressed in cancer Prosurvival Protein 1) [57–58] (Fig. 6). These proteins have a sequence motif of four consecutive positions with mainly aromatic residues in their predicted TMs (Fig. 6, Supplementary Figures S6 and S7). The WBP1/VOPP1 family proteins in *Caenorhabditis remanei* and *Caenorhabditis briggsae* have been expanded (Supplementary Figure S6). The second and fifth cysteines of the C*C*CC*CC signature in these nematode proteins are mutated (C*X*CC*XC), suggesting a disulfide bond connection between them (Supplementary Figure S6). Coupled with concerted mutations in the fourth and sixth cysteines in some of the CX family members (C*C*CX*CX, Supplementary Figure S8), these proteins were inferred to have 1–3, 2–5 and 4–6 disulfide bond connectivity, identical to the predicted cysteine connectivity pattern of Shisa and Shisa-like proteins (Fig. 5). Some WBP1/VOPP1 proteins (e.g., gi: 268567772 from *C. briggsae* and gi: 291234472 from *S. kowalevskii*, Supplementary Figure S6) have more than one cysteine-rich domains before their predicted TMs.

Similar to Shisa/Shisa-like proteins, most members in WBP1/VOPP1, CX, DUF2650, TMEM92 and CYYR1 families have cysteines

and positively charged residues near the C-termini of their predicted TMs, suggesting that they are type I transmembrane proteins that can potentially undergo lipid modifications. WBP1, OPAL1 and VOPP1 proteins have CCCAφ (φ, an aromatic residue), CCCVC and CCGAG signatures near the C-termini of their predicted TMs, respectively (Supplementary Figure S6). A phylogenetic reconstruction for WBP1/VOPP1 proteins suggests that WBP1 and OPAL1 proteins form a major group (Supplementary Figure S12, the bootstrap support is low at 61). OPAL1 proteins, being closely related to WBP1, probably resulted from a gene duplication of WBP1 in vertebrates (Supplementary Figure S12). Vertebrate VOPP1 proteins together with a number of invertebrate proteins form a second major group (Supplementary Figure S12, the bootstrap support is 84), although the relative positions of many invertebrate proteins do not have strong bootstrap supports. Insect and nematode sequences form long branches in the tree (Supplementary Figure S12), reflecting elevated evolutionary rates in these lineages.

WBP1/VOPP1 proteins are widespread in Metazoa (Supplementary Figure S13). Outside metazoans, they are only found in the ciliate *T. thermophila* based on both protein and EST evidence (Fig. 6 and Supplementary Figure S7). CX and DUF2640 are nematode-specific families, while TMEM92 appears to be mammalian-specific. The other family CYYR1 (cysteine and tyrosine rich protein 1) is vertebrate-specific [59]. Weak sequence similarities were found between CYYR1 and Shisa/Shisa-like proteins by PSI-BLAST, suggesting that they are evolutionarily related. For example, A PSI-BLAST search (e-value inclusion threshold 1e−4, with no composition-based statistics) starting with a zebrafish Shisa protein (gi: 192453554) identified a chicken CYYR1 protein (gi: 118083813) with an e-value of 1e−4 in the eighth iteration.

4. Discussion

4.1. Homology among the STMC6 proteins?

Homology inference for STMC6 proteins was complicated by the biased amino acid composition in the cysteine-rich domains, the predicted TM regions and the C-terminal proline-rich low complexity regions. To reduce the chance of finding false hits, we removed the

```
Vertebrate WBP1
NP_036609.1      Hs  26 ELCPGVNNQPYLCES-GHCCGE-TGCCT----YYYELWWFWLLWTVLILFSCCCAFRHRRAKLR   83[269]
NP_001015807.1   Xt  21 EHCPGVNNRPYVCET-GHCCGE-TGCCT----YYYELWWFWLLWSVLILLSCCCAYRHRRAKLR   78[311]
NP_991197.1      Dr  24 VFCSGVNNEQYRCEM-GYCCGE-TECCT----YYYELWWFWLVWTLIILMSCCCAYRHRRVKMR   81[227]
Vertebrate OPAL1
NP_001077382.1   Hs  33 EACVGTNNQSYICDT-GHCCGQ-SQCCN----YYYELWWFWLVWTIIIILSCCCVCHHRRAKHR   90[363]
NP_001090795.1   Xt  12 ETCIGINNQSYICET-GHCCGQ-SQCCN----YYYELWWFWLVWTIIIILSCCCVCHHRRAKHR   69[337]
NP_001018467.1   Dr  23 LHCEGVNNKSYVCES-GHCCGE-SQCCS----YYYELWWFWLVWAIIFILSCCCVCHHRRTKHR   80[334]
Vertebrate VOPP1
NP_110423.3      Hs  24 KHCWYF(6)YYICRSYEDCCGS--RCCVR-ALSIQRLWYFWFLLMMGVLFCCGAGFFIRRRMYP   87[172]
NP_001072426.1   Xt  23 KHCWYF(6)YYICRAYEDCCGS--RCCVR-ALSIQRLWYFWFLLMMGVLFCCGAGFFIRRRMYP   86[167]
CAQ15188.1       Dr  24 KYCWYF(6)YFICKAYEDCCGT--ECCVR-ALAVQRIWYFWLLLIIAILCCCSTGYFIRRRAHP   87[162]
Other deuterostomes
XP_002120990.1   Ci  41 ETCNGSDGESYVCNV-GHCCED-DNCCT----YYYELWWFWLIWGLIILMSCCCAYHHRQKRRL   98[378]
XP_002121928.1   Ci  17 EYCGD----GYYCRD(7)CCGK(7)CCSY--TAFWGMWYWMGMSLVFVIVIVVLKECYKMMKK   82[191]
XP_002131028.1   Ci  24 EVCATDPDHALICPD(7)CCNY(7)CCTY--TAVWGLWYFWFCLVSFFFIVGGVVALCKYCATK   93[168]
XP_002127288.1   Ci   2 SFCG-----NNICND(7)CCGF(4)CCYK(4)EWYSMWYIWVGIVSFLILLSLCFRCWRQQQQI   66[135]
XP_002588917.1   Bf  24 EYCTG---YGHYCDL(4)CCGY--TCCRY(7)SVWNLWYFWFVVIFVMMACCGGCGYYRRQHL   87[246]
XP_002603751.1   Bf  29 EFCKKQDKEGYWCET-GHCCGD-GGCCT----YYYELWYAILKLPSYQEAITQQVDTSPPPRYT   86[259]
XP_782362.1      Sp  28 EICPS----GIECDYNEHCCED--GCCA(11)GFYNLWYFWFVIIMVMMTCCGACSYVKKRQYY   92[258]
XP_800378.1      Sp  22 ISCSYY(8)YYRCLSSQYCCGT-SSCCYR(5)SVYSLWYFWFCIILFFTLCSGASGAYYRKRQR   91[186]
XP_791639.1      Sp  27 EFCQV---EGYWCET-GHCCGS-GECCT----YYYELWMFWLVWFLIVFIIGFCVWQRKRFHPW   81[312]
XP_002737172.1   Sk  83 HNCD-----DIECRDDQYCCER-NVCCNF--SEIYHNWWFWFSILFIAAIVLGSIVGMACAGRR  138[161]
XP_002741429.1   Sk  22 EECAG---YEHSCSPPEYCCNY--TCCRA(4)SFWSMWYFWCIVVFILMACFGGCGYYRRRHVL   81[185]
Ecdysozoans
XP_393257.1      Am  42 RLCNG----GHVCSPPKECCSF--GCCY(14)LIWTYWYLWAAVLVGLAIAAACGCWLWKRHHS  109[422]
EFZ09690.1       Si  25 KYCASD(6)YYECSRTEYCCAF--GCCVSPGFHFHHLWYWILVIIMFLVCSGGGWWYRYWLQG   89[158]
EFN66503.1       Cf  35 KLCDG----GHVCSPPKNCCQH--GCCY(16)LIWTYWYLWVAVLAALAIAAICGFWLWKRRRA  104[412]
XP_001664191.1   Aa   4 RFCEG----GHRCLPPKECCAQ--GCCY(21)FLFNHWYFWCFVLAIVLALLCACSLWKRRQL   78[445]
NP_572909.1      Dm  34 RFCEG-------CTQ--GCCTQ--GCCP(20)FFATHWYFWCVVVAIILAILCAYSLWKKRRTL  102[518]
ADV40322.1       Lh  29 YYCN-----HDLCREDQYCCGD-NLCCD----YVYYPFYFWAGVVFMLIALSACGGLFRYCYDN   82[145]
NP_493831.1      Ce  11 QNCKQ(16)MRTNSMYLECCVK-HRYCR----WYGNTGFFILMLFVVIICIIVSVAASVVCCCT   81[117]
XP_003091617.1   Cr  50 KVCKAY(9)IKNTHGYGSCCEV-GGICY(14)FYKQTWFFVACGGVLLLLIIVGVVVGYFICRR  128[158]
XP_002647920.1   Cb  61 VECDKY(9)TKGTYGQVICCDE-LKFCSG--PFYTQIWFFAVCGGVALLLIGIIGVVVFLLCCR  127[151]
Cnidarians
XP_001633593.1   Nv  25 DYCGRKGKKEFYCDK-GHCCGN-FECCVY--YKYYEMWWFWLIWCVITILGCSCAYLKRRQIGL   84[191]
XP_001630062.1   Nv 128 MTCD-----EDYCAS--YCCYD(6)CCSE--TDITDMWWFWVSFATGLLFLSLCCGLCCRRHLR  185[244]
Placozoans
XP_002118935.1   Ta  22 TFHTY----YSQCLMQEYCCGG-TSCCF----AFYKVWYFWASIGFLILAFAAMITWLRYRRMQ   76[114]
XP_002114788.1   Ta  30 LRCV-------GCMS--KCCCN-GKCCRC-RHLIYEMWWFWLLWCLVVLLSCCCAYIKRRHRRF   82[205]
Porifera
GH561511.1       Sd     SYCSDF(3)GWYCTF--GCCYSFTGCCY----GAGSYWYAWTIVSLFFLVCFCACAGCGGYGYR
GO092717.1       Lc     NQCTTYIRDYYTCASYQYCCGQ-CSCCS----SVYALWWFWAVTTPIFIFIVTCCTFWICRRRR
Tetrahymena
XP_001023744.1   Tt  20 DICYQK(7)TAYCDSGYKCCFYQDSCCR----DFGYYWWIFVIAAIFVIILLFSCCLYKFGCCK   83[133]
XP_001020578.1   Tt  19 TYCGSSYSYNYYCPSGYYCCST-TECCQ----NVYSSWWFITIIVVSVILIIVAIVLGVIKRRR   77[156]
                                                                 =========TM=========
```

Fig. 6. Multiple sequence alignment of WBP1/VOPP1 family and its homologous proteins. This alignment includes aligned regions of the cysteine-rich domains and predicted trans-membrane segments (labeled TM at the bottom). The sequences are denoted by NCBI accession numbers. EST-derived proteins have italic and underlined accession numbers. Starting residue numbers for sequences with protein accession numbers are shown before the sequences. Conserved cysteines are shaded in yellow. Mutations in conserved cysteines are shaded in black. Aromatic residues in the four positions with mainly aromatic residues are in red letters. Cysteines near the C-terminal ends of predicted transmembrane segments are shaded in gray. Arginines and lysines near the C-terminal ends of the predicted TMs are shown as blue letters. Species abbreviations shown after the accession numbers are as follows: Aa, *Aedes aegypti*; Am, *Apis mellifera*; Bf, *Branchiostoma floridae*; Cb, *Caenorhabditis briggsae*; Ce, *Caenorhabditis elegans*; Cr, *Caenorhabditis remanei*; Cf, *Camponotus floridanus*; Ci, *Ciona intestinalis*; Dr, *Danio rerio*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Lh, *Latrodectus hesperus*; Lc, *Leucetta chagosensis*; Nv, *Nematostella vectensis*; Sk, *Saccoglossus kowalevskii*; Si, *Solenopsis invicta*; Sp, *Strongylocentrotus purpuratus*; Sd, *Suberites domuncula*; Tt, *Tetrahymena thermophila*; Ta, *Trichoplax adhaerens* and Xt, *Xenopus tropicalis*.

predicted signal peptides and the C-terminal low complexity regions. The remaining regions covering the cysteine-rich domains and predicted TMs were subject to PSI-BLAST searches under more strict e-value cutoffs (1e−4 or lower) than default (2e−3 for the local blastpgp program or 5e−3 in online server). Shisa and Shisa-like proteins form a clear homologous group, as they can be readily linked by PSI-BLAST under stringent settings. On the other hand, CYYR1 can only be linked to Shisa proteins under more relaxed PSI-BLAST settings (disabling of composition-based statistics). WBP1 and VOPP1 belong to another clear homologous group, while CX, DUF2650 and TMEM92 can be linked to WBP1/VOPP1 proteins under less stringent PSI-BLAST settings (disabling of low complexity statistics and with higher e-value inclusion cutoffs). HHpred search results also support the distant homologous relationships among WBP1/VOPP1, CX, DUF2650 and TMEM92.

Sequence similarity search methods such as PSI-BLAST and HHpred did not link the group of Shisa/Shisa-like/CYYR1 to the group of WBP1/VOPP1/CX/DUF2650/TMEM92 with statistically significant scores. However, these proteins share common sequence features including a similar domain architecture (Fig. 1), a cysteine-rich domain with the general C*C*CC*CC pattern, and frequently a C-terminal low complexity region (Supplementary Figures S6–S11) with PY motifs (discussed below). These features collectively tend to argue that all these families of STMC6 proteins are evolutionarily related. Future structural determination of the cysteine-rich domains of these families can provide insights into the possible homologous

relationships among STMC6 proteins. For example, a shared structural fold and the same disulfide bond connectivity would favor the proposed common evolutionary origin of STMC6 proteins.

4.2. Divergent evolution of STMC6 proteins

Several types of recurring evolutionary events were observed for STMC6 proteins. First, gene expansion by duplications is a frequent scheme in the evolution of different STMC6 families, resulting in quite a number of STMC6 genes in some metazoan genomes. For example, the human genome has 17 STMC6 genes (eight copies of Shisa; two copies of FAM159, two copies of KIAA1644, three copies of WBP1/VOPP1, one TMEM92, and one CYYR1). Large gene expansions of Shisa-like genes were observed in deuterostomes such as *B. floridae* and *S. kowalevskii* (Supplementary Figure S2). High gene copy numbers of WBP1/VOPP1, CX and DUF2650 families represent a prominent feature in some nematode genomes (Supplementary Figures S6 and S8–S11). Duplicated STMC6 gene products could evolve new functions and potentially regulate the functions of different membrane proteins in a variety of ways (discussed below).

Second, domain-level duplications resulting in tandem cysteine-rich domains occurred occasionally and independently in several STMC6 families, including Shisa-like proteins from a few deuterostomes (Supplementary Figure S3), two *B. plicatilis* Shisa proteins (Supplementary Figure S4), and some WBP1/VOPP1 proteins from *C. briggsae* and *S. kowalevskii* (Supplementary Figure S6). Such domain-level duplication events support that the cysteine-rich domain is an independent evolutionary and folding unit in STMC6 proteins. Duplicated cysteine-rich domains could contribute to extra binding sites for their interaction partners.

Third, at the amino acid residue level, concerted mutations of the conserved cysteines were found independently in a few Shisa and Shisa-like proteins (Fig. 5), some of the WBP1/VOPP1 proteins in nematodes (Supplementary Figure S7), and some CX family proteins (Supplementary Figure S8). These mutations suggest the 1–3, 2–5 and 4–6 disulfide bond connectivity in the general C*C*CC*CC pattern of the cysteine-rich domains in STMC6 proteins. On the other hand, the extra cysteine pairs in Shisa proteins and some *B. floridae* and lophotrochozoan Shisa-like proteins could form an additional disulfide bond. Structural studies of the cysteine-rich domains in STMC6 proteins would be required to validate these disulfide bond connectivity predictions based on co-varying positions.

4.3. Functions of STMC6 proteins

The STMC6 proteins, especially WBP1/VOPP1 and Shisa-like proteins, are widespread in Metazoa (Supplementary Figure S13), yet not much functional information is available for most of them except for *Xenopus* Shisa1 [16] and mouse Shisa9/CKAMP44 [15]. WBP1 is one of the first WW domain-interacting proteins identified more than a decade ago [55, 60]. The C-terminal region of WBP1 has several PY motifs ([LP]PxY) that are supposed to interact with WW domains. However, no experimental work has been performed to study the cellular functions of WBP1 ever since. High gene expression of OPAL1, a close homolog of WBP1, was originally indicated as an outcome predictor of acute lymphoblastic leukemia (ALL) [61–62], although a later multivariate analysis has shown that its expression may not be an independent prognostic feature for ALL [56]. VOPP1 (previously named ECOP for EGFR-Coamplified and Over-expressed Protein), a distantly related WBP1 homolog, was implicated in several types of human cancers as a pro-survival protein [57, 63–64]. VOPP1 was found to localize in cytoplasmic vesicles [58]. Its knockdown induces apoptosis that is likely related to oxidative cellular injury [65]. Despite these sporadic studies, the specific functional roles of members in the WBP1/VOPP1 family are not understood. Members from the

CX, DUF2650 and TMEM92 families also lack experimental studies and detailed information about their cellular functions.

Like WBP1, the majority of STMC6 proteins contain PY motifs ([LP] PXY) in their C-terminal proline-rich, mainly disordered regions (Supplementary Figures S1, S2, S4 and S6–S11). PY motifs are the most common WW-domain-interacting motifs [6]. WW domains often occur in tandem in a protein. Consistently, two or more PY motifs have been observed in quite a number of STMC6 proteins, suggesting they could interact with multiple WW domains from a single protein. One important class of proteins with tandem WW domains is the NEDD4 family HECT-domain-containing E3 ubiquitin ligases [66]. The NEDD4 family proteins, such as Rsp5 from yeast and the mammalian proteins NEDD4, NEDD4-2, SMURF1, SMURF2 and ITCH, target a variety of proteins for ubiquitination to facilitate their degradation by the proteasome and the lysosome or play roles in their sorting to specific subcellular locations [66]. The presence of an N-terminal C2 domain facilitates membrane localization of the NEDD4 family proteins, and thus many of their targets are transmembrane proteins including ion channels such as epithelial sodium channel (EnaC) [67] and cell surface receptors such as TGF-β [68]. The interactions between the NEDD4 family proteins and their targets are often mediated by the WW domains in the NEDD4 family proteins and the PY motifs in the targets [66].

Some transmembrane proteins do not possess the PY motifs themselves, but are still targeted by the NEDD4 family proteins through interactions with PY-motif-containing adaptor proteins. One example is TGF-β, which degradation can be mediated by SMAD7 [68]. Several transmembrane adaptors with PY motifs have also been identified, such as LAPTM5 [69–70] and NDFIP proteins [71–72], that can promote the degradation of transmembrane proteins. The presence of PY motifs in STMC6 proteins suggests that they could play similar roles. In one simplified model, STMC6 proteins could interact with target membrane proteins and recruit the NEDD4 family E3 ligases to ubiquitinate the targets and promote their degradation (Fig. 7a). In this way, the STMC6 proteins act as down-regulators of cell surface expression of the target proteins. Alternatively, competitive binding of STMC6 proteins to NEDD4 members can cause up-regulation of cell surface expression of transmembrane proteins that are otherwise targeted by NEDD4 members, as proposed by some researchers [73–74] (Fig. 7b).

PY motifs are present in STMC6 members from the basal metazoan branches, such as Shisa-like proteins from Porifera (Supplementary Figure S4) and WBP1/VOPP1 members from Porifera, Placozoa and Cnidaria (Supplementary Figure S7), suggesting that their interactions with WW-domain containing proteins existed in the ancestor of metazoans.

Some STMC6 proteins do not contain PY motifs in their C-terminal disordered regions, such as Shisa-like protein families FAM159 and KIAA1644 (Supplementary Figure S2) and many nematode members of the WBP1/VOPP1, CX and DUF2650 families (Supplementary Figures S6, S8 and S9). STMC6 proteins without PY motifs may have evolved new functions that do not depend on the interactions with WW-domain-containing proteins. Functional divergence after gene duplications is evidenced for the Shisa family proteins in vertebrates, which have nine subfamilies with divergent C-terminal regions among them (Supplementary Figure S1). Most members of Shisa subfamilies 3–8 contain PY motifs in their C-terminal regions (Supplementary Figure S1), suggesting that they could potentially serve as adaptor proteins to regulate the ubiquitination of other target proteins through the interactions with the NEDD4 family proteins. On the other hand, most members of Shisa subfamilies 1, 2 and 9 lack PY motifs (Supplementary Figure S1), suggesting that they do not target other proteins for ubiquitination and degradation. The mouse Shisa9 protein (CKAMP44) functions to desensitize AMPA receptors in the plasma membrane [15] (Fig. 7c). The *Xenopus* Shisa1 protein plays a role in embryonic development via the retention of Wnt and FGF receptors in the endoplasmic
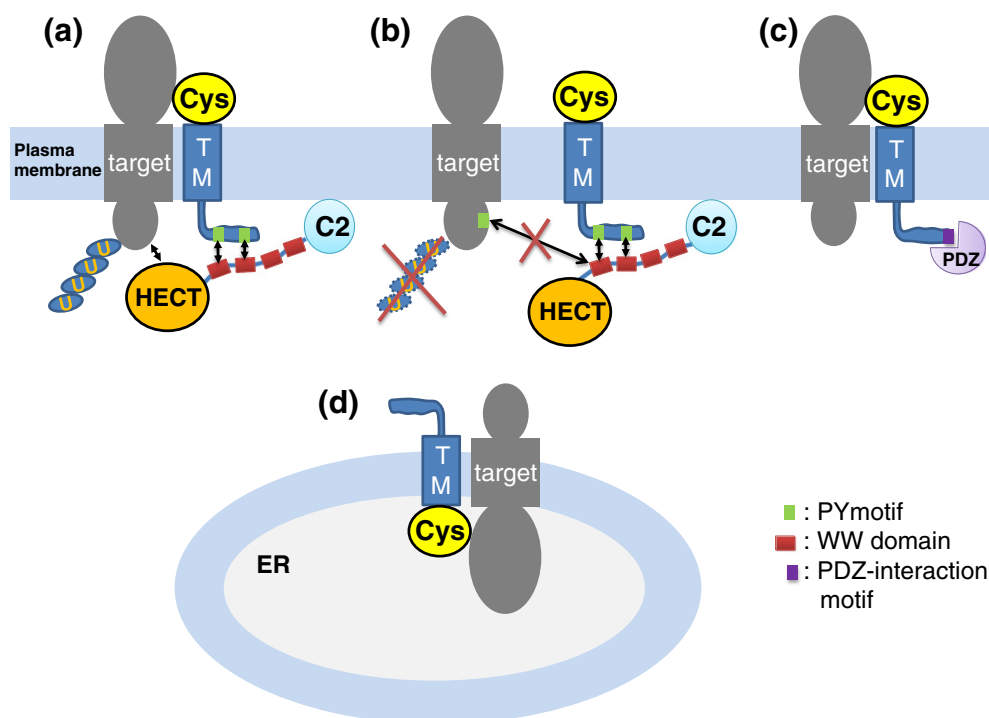
Fig. 7. Putative mechanisms of action of STMC6 proteins. This diagram depicts four putative mechanisms of STMC6 proteins in regulating other transmembrane proteins. (a). The STMC6 protein can decrease cell surface expression of a target transmembrane protein by recruiting a NEDD4 family E3 ubiquitin ligase to ubiquitinate the target, possibly resulting in its degradation or relocalization. (b). The STMC6 protein can enhance cell surface expression of a target transmembrane protein by competitively interacting with the WW domains of a NEDD4 family E3 ubiquitin ligase and thus blocking the interaction between the E3 ligase and the target. (c). The STMC6 protein Shisa9 (CKAMP44) desensitizes the AMPA receptor by interacting with it. Shisa9 can also potentially interact with PDZ-domain-containing proteins via a C-terminal motif. (d). The STMC6 protein Shisa1 from *Xenopus* retains receptors in the ER and inhibits their cell surface expression.

reticulum [16] (Fig. 7d). Therefore, existing experimental results suggest that Shisa1 and Shisa9 are both adaptor proteins that regulate the localization or function of receptors. However, their targets, subcellular localizations and cellular functions are different. New functions for the Shisa paralogs could be acquired by evolving new linear sequence motifs in the C-terminal low complexity region. For example, it has been observed that mouse Shisa9 contains a type 2 PDZ-domain interacting signature ([ST]x[VLI]) at the C-terminus [15]. Examination of other Shisa members revealed that such a PDZ-interacting motif is a conserved feature in subfamilies Shisa2 and Shisa6–9, while it is lacking in subfamilies Shisa3–5 (Supplementary Figure S1).

4.4. Comparison of STMC6 proteins with a class of immunoreceptor-related transmembrane adaptor proteins

Several transmembrane adaptor proteins (TRAPs), such as LAT, NTAL, PAG and SIT, play important roles in mediating the functions of immunoreceptors [14]. These adaptor proteins do not interact with receptors, but function as downstream targets of immunoreceptor-activated tyrosine kinases. Phosphorylated TRAPs then serve as scaffolds to recruit other proteins to regulate downstream signaling. These immunoreceptor-related TRAPs and STMC6 proteins are all single-transmembrane proteins with cytosolic C-terminal regions that contain linear peptide motifs. However, all characterized immunoreceptor-related TRAPs have very short N-terminal regions before their TMs [14], while STMC6 proteins possess an N-terminal cysteine-rich domain that could interact with the ectodomains of receptors or other transmembrane proteins. In this sense, STMC6 proteins are similar to some receptor-associated proteins that also possess an ectodomain for mediating protein–protein interactions, such as T-cell-receptor-associated CD3 components [75] and B-cell-receptor-associated CD79 components [76]. STMC6 proteins could also interact with receptors or other transmembrane proteins via their predicted TMs. For WBP1/VOPP1 proteins (Fig. 6), the conserved motif with several aromatic residues in their predicted TMs could serve as interaction sites for other transmembrane proteins. Future experimental investigation is required to identify the interaction partners of STMC6 proteins, which could help reveal their cellular functions.

5. Conclusions

We catalogued a large and diverse group of STMC6 proteins that include the Shisa/Shisa-like, WBP1/VOPP1, CX, DUF2650, TMEM92, and CYYR1 family proteins. STMC6 members are widespread in Metazoa, yet with scarce experimental data. They share a domain structure characterized by an N-terminal cysteine-rich domain with a distinct cysteine pattern, a single predicted transmembrane segment, and a proline-rich C-terminal region. Most STMC6 proteins possess PY motifs in their C-terminal regions, suggesting that they could interact with WW-domain-containing proteins such as the NEDD4 family E3 ubiquitin ligases and play important roles in protein sorting and degradation. These proteins are likely transmembrane adaptors that regulate the functions of other transmembrane proteins such as cell surface receptors, as exemplified by *Xenopus* Shisa1 and mouse Shisa9/CKAMP44. Future experimental studies on STMC6 proteins could reveal new functional aspects of receptor regulation.

Supplementary materials related to this article can be found online at doi:10.1016/j.cellsig.2011.11.011.

Competing interests

None declared.

Acknowledgments

References

[1] T. Pawson, J.D. Scott, Science 278 (1997) 2075–2080.
[2] A. Zeke, M. Lukacs, W.A. Lim, A. Remenyi, Trends in Cell Biology 19 (2009) 364–374.
[3] T. Pawson, Nature 373 (1995) 573–580.
[4] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N.P. Brown, G. Trave, T.J. Gibson, Frontiers in Bioscience 13 (2008) 6580–6603.
[5] T. Kaneko, L. Li, S.S. Li, Frontiers in Bioscience 13 (2008) 4938–4952.
[6] M.J. Macias, S. Wiesner, M. Sudol, FEBS Letters 513 (2002) 30–37.
[7] T. Pawson, G.D. Gish, P. Nash, Trends in Cell Biology 11 (2001) 504–511.
[8] P. van der Geer, T. Pawson, Trends in Biochemical Sciences 20 (1995) 277–280.
[9] J. Saras, C.H. Heldin, Trends in Biochemical Sciences 21 (1996) 455–458.
[10] M.A. Lemmon, Biochemical Society Symposium (2007) 81–93.
[11] S.M. DeWire, S. Ahn, R.J. Lefkowitz, S.K. Shenoy, Annual Review of Physiology 69 (2007) 483–510.
[12] S.S. Ferguson, Pharmacological Reviews 53 (2001) 1–24.
[13] L.M. Luttrell, D. Gesty-Palmer, Pharmacological Reviews 62 (2010) 305–330.
[14] V. Horejsi, W. Zhang, B. Schraven, Nature Reviews Immunology 4 (2004) 603–616.
[15] J. von Engelhardt, V. Mack, R. Sprengel, N. Kavenstock, K.W. Li, Y. Stern-Bach, A.B. Smit, P.H. Seeburg, H. Monyer, Science 327 (2010) 1518–1522.
[16] A. Yamamoto, T. Nagano, S. Takehara, M. Hibi, S. Aizawa, Cell 120 (2005) 223–235.
[17] E.M. De Robertis, Mechanisms of Development 126 (2009) 925–941.
[18] C. Niehrs, Trends in Genetics 15 (1999) 314–319.
[19] C.Y. Logan, R. Nusse, Annual Review of Cell and Developmental Biology 20 (2004) 781–810.
[20] B.T. MacDonald, K. Tamai, X. He, Developmental Cell 17 (2009) 9–26.
[21] K. Dorey, E. Amaya, Development 137 (2010) 3731–3742.
[22] T. Nagano, S. Takehara, M. Takahashi, S. Aizawa, A. Yamamoto, Development 133 (2006) 4643–4654.
[23] T.A. Hedge, I. Mason, The International Journal of Developmental Biology 52 (2008) 81–85.
[24] Y. Zhu, A. Tsuchida, A. Yamamoto, K. Furukawa, O. Tajima, N. Tokuda, S. Aizawa, T. Urano, K. Kadomatsu, J. Nagoya, Medical Science 70 (2008) 73–82.
[25] K. Furushima, A. Yamamoto, T. Nagano, M. Shibata, H. Miyachi, T. Abe, N. Ohshima, H. Kiyonari, S. Aizawa, Developmental Biology 306 (2007) 480–492.
[26] J.C. Bourdon, J. Renzing, P.L. Robertson, K.N. Fernandes, D.P. Lane, The Journal of Cell Biology 158 (2002) 235–246.
[27] A. Terrinoni, M. Ranalli, B. Cadot, A. Leta, G. Bagetta, K.H. Vousden, G. Melino, Oncogene 23 (2004) 3721–3725.
[28] L. Zocchi, J.C. Bourdon, A. Codispoti, R. Knight, D.P. Lane, G. Melino, A. Terrinoni, Biochemical and Biophysical Research Communications 367 (2008) 271–276.
[29] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Research 25 (1997) 3389–3402.
[30] J. Soding, A. Biegert, A.N. Lupas, Nucleic Acids Research 33 (2005) W244–248.
[31] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, Nucleic Acids Research 38 (2010) D211–222.
[32] K. Katoh, K. Kuma, H. Toh, T. Miyata, Nucleic Acids Research 33 (2005) 511–518.
[33] B. Rhead, D. Karolchik, R.M. Kuhn, A.S. Hinrichs, A.S. Zweig, P.A. Fujita, M. Diekhans, K.E. Smith, K.R. Rosenbloom, B.J. Raney, A. Pohl, M. Pheasant, L.R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R.A. Harte, B. Giardine, T.R. Dreszer, H. Clawson, G.P. Barber, D. Haussler, W.J. Kent, Nucleic Acids Research 38 (2010) D613–619.
[34] B. Venkatesh, E.F. Kirkness, Y.H. Loh, A.L. Halpern, A.P. Lee, J. Johnson, N. Dandona, L.D. Viswanathan, A. Tay, J.C. Venter, R.L. Strausberg, S. Brenner, PLoS Biology 5 (2007) e101.
[35] L. Kall, A. Krogh, E.L. Sonnhammer, Nucleic Acids Research 35 (2007) W429–432.
[36] J. Pei, N.V. Grishin, Bioinformatics 23 (2007) 802–808.
[37] J. Adachi, M. Hasegawa, Computer Science Monographs, 28, The Institute of Statistical Mathematics, 1996, pp. 1–150.
[38] D.T. Jones, W.R. Taylor, J.M. Thornton, Computer Applications in the Biosciences 8 (1992) 275–282.
[39] M. Hasegawa, H. Kishino, N. Saitou, Journal of Molecular Evolution 32 (1991) 443–445.
[40] Y. Katoh, M. Katoh, International Journal of Molecular Medicine 16 (2005) 181–185.
[41] M.E. Linder, R.J. Deschenes, Nature Reviews Molecular Cell Biology 8 (2007) 74–84.
[42] G. von Heijne, Nature 341 (1989) 456–458.
[43] N.H. Putnam, T. Butts, D.E. Ferrier, R.F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J.K. Yu, E.L. Benito-Gutierrez, I. Dubchak, J. Garcia-Fernandez, J.J. Gibson-Brown, I.V. Grigoriev, A.C. Horton, P.J. de Jong, J. Jurka, V.V. Kapitonov, Y. Kohara, Y. Kuroki, E. Lindquist, S. Lucas, K. Osoegawa, L.A. Pennacchio,

[44] A.A. Salamov, Y. Satou, T. Sauka-Spengler, J. Schmutz, I.T. Shin, A. Toyoda, M. Bronner-Fraser, A. Fujiyama, L.Z. Holland, P.W. Holland, N. Satoh, D.S. Rokhsar, Nature 453 (2008) 1064–1071.
[44] C.J. Lowe, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 363 (2008) 1569–1578.
[45] E. Sodergren, G.M. Weinstock, E.H. Davidson, R.A. Cameron, R.A. Gibbs, R.C. Angerer, L.M. Angerer, M.I. Arnone, D.R. Burgess, R.D. Burke, J.A. Coffman, M. Dean, M.R. Elphick, C.A. Ettensohn, K.R. Foltz, A. Hamdoun, R.O. Hynes, W.H. Klein, W. Marzluff, D.R. McClay, R.L. Morris, A. Mushegian, J.P. Rast, L.C. Smith, M.C. Thorndyke, V.D. Vacquier, G.M. Wessel, G. Wray, L. Zhang, C.G. Elsik, O. Ermolaeva, W. Hlavina, G. Hofmann, P. Kitts, M.J. Landrum, A.J. Mackey, D. Maglott, G. Panopoulou, A.J. Poustka, K. Pruitt, V. Sapojnikov, X. Song, A. Souvorov, V. Solovyev, Z. Wei, C.A. Whittaker, K. Worley, K.J. Durbin, Y. Shen, O. Fedrigo, D. Garfield, R. Haygood, A. Primus, R. Satija, T. Severson, M.L. Gonzalez-Garay, A.R. Jackson, A. Milosavljevic, M. Tong, C.E. Killian, B.T. Livingston, F.H. Wilt, N. Adams, R. Belle, S. Carbonneau, R. Cheung, P. Cormier, B. Cosson, J. Croce, A. Fernandez-Guerra, A.M. Geneviere, M. Goel, H. Kelkar, J. Morales, O. Mulner-Lorillon, A.J. Robertson, J.V. Goldstone, B. Cole, D. Epel, B. Gold, M.E. Hahn, M. Howard-Ashby, M. Scally, J.J. Stegeman, E.L. Allgood, J. Cool, K.M. Judkins, S.S. McCafferty, A.M. Musante, R.A. Obar, A.P. Rawson, B.J. Rossetti, I.R. Gibbons, M.P. Hoffman, A. Leone, S. Istrail, S.C. Materna, M.P. Samanta, V. Stolc, W. Tongprasit, Q. Tu, K.F. Bergeron, B.P. Brandhorst, J. Whittle, K. Berney, D.J. Bottjer, C. Calestani, K. Peterson, E. Chow, Q.A. Yuan, E. Elhaik, D. Graur, J.T. Reese, I. Bosdet, S. Heesun, M.A. Marra, J. Schein, M.K. Anderson, V. Brockton, K.M. Buckley, A.H. Cohen, S.D. Fugmann, T. Hibino, M. Loza-Coll, A.J. Majeske, C. Messier, S.V. Nair, Z. Pancer, D.P. Terwilliger, C. Agca, E. Arboleda, N. Chen, A.M. Churcher, F. Hallbook, G.W. Humphrey, M.M. Idris, T. Kiyama, S. Liang, D. Mellott, X. Mu, G. Murray, R.P. Olinski, F. Raible, M. Rowe, J.S. Taylor, K. Tessmar-Raible, D. Wang, K.H. Wilson, S. Yaguchi, J. Gaasterland, B.E. Galindo, H.J. Gunaratne, C. Juliano, M. Kinukawa, G.W. Moy, A.T. Neill, M. Nomura, M. Raisch, A. Reade, M.M. Roux, J.L. Song, Y.H. Su, I.K. Townley, E. Voronina, J.L. Wong, G. Amore, M. Branno, E.R. Brown, V. Cavalieri, V. Duboc, L. Duloquin, C. Flytzanis, C. Gache, F. Lapraz, T. Lepage, A. Locascio, P. Martinez, G. Matassi, V. Matranga, R. Range, F. Rizzo, E. Rottinger, W. Beane, C. Bradham, C. Byrum, T. Glenn, S. Hussain, G. Manning, E. Miranda, R. Thomason, K. Walton, A. Wikramanayke, S.Y. Wu, R. Xu, C.T. Brown, L. Chen, R.F. Gray, P.Y. Lee, J. Nam, P. Oliveri, J. Smith, D. Muzny, S. Bell, J. Chacko, A. Cree, S. Curry, C. Davis, H. Dinh, S. Dugan-Rocha, J. Fowler, R. Gill, C. Hamilton, J. Hernandez, S. Hines, J. Hume, L. Jackson, A. Jolivet, C. Kovar, S. Lee, L. Lewis, G. Miner, M. Morgan, L.V. Nazareth, G. Okwuonu, D. Parker, L.L. Pu, R. Thorn, R. Wright, Science 314 (2006) 941–952.
[46] M.J. Telford, Genesis 46 (2008) 580–586.
[47] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H. Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor, J.F. Abril, A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A.A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon, D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, R. Remington, R.D. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Siden-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H. Wang, X. Wang, Z.Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, T. Woodage, K.C. Worley, D. Wu, S. Yang, Q.A. Yao, J. Ye, R.F. Yeh, J.S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X.H. Zheng, F.N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin, J.C. Venter, Science 287 (2000) 2185–2195.
[48] S. Richards, R.A. Gibbs, G.M. Weinstock, S.J. Brown, R. Denell, R.W. Beeman, R. Gibbs, G. Bucher, M. Friedrich, C.J. Grimmelikhuijzen, M. Klingler, M. Lorenzen, S. Roth, R. Schroder, D. Tautz, E.M. Zdobnov, D. Muzny, T. Attaway, S. Bell, C.J. Buhay, M.N. Chandrabose, D. Chavez, K.P. Clerk-Blankenburg, A. Cree, M. Dao, C. Davis, J. Chacko, H. Dinh, S. Dugan-Rocha, G. Fowler, T.T. Garner, J. Garnes, A. Gnirke, A. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S.N. Jhangiani, V. Joshi, Z.M. Khan, L. Jackson, C. Kovar, A. Kowis, S. Lee, L.R. Lewis, J. Margolis, M. Morgan, L.V. Nazareth, N. Nguyen, G. Okwuonu, D. Parker, S.J. Ruiz, J. Santibanez, J. Savard, S.E. Scherer, B. Schneider, E. Sodergren, S. Vattahil, D. Villasana, C.S. White, R. Wright, Y. Park, J. Lord, B. Oppert, S. Brown, L. Wang, G. Weinstock, Y. Liu, K. Worley, C.G. Elsik, J.T. Reese, E. Elhaik, G. Landan, D. Graur, P. Arensburger, P. Atkinson, J. Beidler, J.P. Demuth, D.W. Drury, Y.Z. Du, H. Fujiwara, V. Maselli, M. Osanai, H.M. Robertson, Z. Tu, J.J. Wang, S. Wang, H. Song, L. Zhang, D. Werner, M. Stanke, B. Morgenstern, V. Solovyev, P. Kosarev, G. Brown, H.C. Chen, O. Ermolaeva, W. Hlavina, Y. Kapustin, B. Kiryutin, P. Kitts, D. Maglott, K. Pruitt, V. Sapojnikov, A. Souvorov, A.J. Mackey, R.M. Waterhouse, S. Wyder, E.V. Kriventseva, T. Kadowaki, P. Bork, M. Aranda, R. Bao, A. Beermann, N. Berns, R. Bolognesi, F. Bonneton, D. Bopp,

T. Butts, A. Chaumot, R.E. Denell, D.E. Ferrier, C.M. Gordon, M. Jindra, Q. Lan, H.M. Lattorff, V. Laudet, C. von Levetsow, Z. Liu, R. Lutz, J.A. Lynch, R.N. da Fonseca, N. Posnien, R. Reuter, J.B. Schinko, C. Schmitt, M. Schoppmeier, T.D. Shippy, F. Simonnet, H. Marques-Souza, Y. Tomoyasu, J. Trauner, M. Van der Zee, M. Vervoort, N. Wittkopp, E.A. Wimmer, X. Yang, A.K. Jones, D.B. Sattelle, P.R. Ebert, D. Nelson, J.G. Scott, S. Muthukrishnan, K.J. Kramer, Y. Arakane, Q. Zhu, D. Hogenkamp, R. Dixit, H. Jiang, Z. Zou, J. Marshall, E. Elpidina, K. Vinokurov, C. Oppert, J. Evans, Z. Lu, P. Zhao, N. Sumathipala, B. Altincicek, A. Vilcinskas, M. Williams, D. Hultmark, C. Hetru, F. Hauser, G. Cazzamali, M. Williamson, B. Li, Y. Tanaka, R. Predel, S. Neupert, J. Schachtner, P. Verleyen, F. Raible, K.K. Walden, S. Angeli, S. Foret, S. Schuetz, R. Maleszka, S.C. Miller, D. Grossmann, Nature 452 (2008) 949–955.

[49] Q. Xia, Z. Zhou, C. Lu, D. Cheng, F. Dai, B. Li, P. Zhao, X. Zha, T. Cheng, C. Chai, G. Pan, J. Xu, C. Liu, Y. Lin, J. Qian, Y. Hou, Z. Wu, G. Li, M. Pan, C. Li, Y. Shen, X. Lan, L. Yuan, T. Li, H. Xu, G. Yang, Y. Wan, Y. Zhu, M. Yu, W. Shen, D. Wu, Z. Xiang, J. Yu, J. Wang, R. Li, J. Shi, H. Li, J. Su, X. Wang, Z. Zhang, Q. Wu, J. Li, Q. Zhang, N. Wei, H. Sun, L. Dong, D. Liu, S. Zhao, X. Zhao, Q. Meng, F. Lan, X. Huang, Y. Li, L. Fang, D. Li, Y. Sun, Z. Yang, Y. Huang, Y. Xi, Q. Qi, D. He, H. Huang, X. Zhang, Z. Wang, W. Li, Y. Cao, Y. Yu, H. Yu, J. Ye, H. Chen, Y. Zhou, B. Liu, H. Ji, S. Li, P. Ni, J. Zhang, Y. Zhang, H. Zheng, B. Mao, W. Wang, C. Ye, G.K. Wong, H. Yang, Science 306 (2004) 1937–1940.

[50] K. Meusemann, B.M. von Reumont, S. Simon, F. Roeding, S. Strauss, P. Kuck, I. Ebersberger, M. Walzl, G. Pass, S. Breuers, V. Achter, A. von Haeseler, T. Burmester, H. Hadrys, J.W. Wagele, B. Misof, Molecular Biology and Evolution 27 (2010) 2451–2464.

[51] R. Rubinstein, A. Fiser, Bioinformatics 24 (2008) 498–504.

[52] Y. Yang, J. Poncet, J. Garnier, C. Zatylny, E. Bachere, A. Aumelas, The Journal of Biological Chemistry 278 (2003) 36859–36867.

[53] D. Tolkatchev, S. Malik, A. Vinogradova, P. Wang, Z. Chen, P. Xu, H.P. Bennett, A. Bateman, F. Ni, Protein Science 17 (2008) 711–724.

[54] D. Tolkatchev, A. Ng, W. Vranken, F. Ni, Biochemistry 39 (2000) 2878–2886.

[55] H.I. Chen, A. Einbond, S.J. Kwak, H. Linn, E. Koepf, S. Peterson, J.W. Kelly, M. Sudol, The Journal of Biological Chemistry 272 (1997) 17070–17077.

[56] A. Holleman, M.L. den Boer, M.H. Cheok, K.M. Kazemier, D. Pei, J.R. Downing, G.E. Janka-Schaub, U. Gobel, U.B. Graubner, C.H. Pui, W.E. Evans, R. Pieters, Blood 108 (2006) 1984–1990.

[57] G.D. Eley, J.L. Reiter, A. Pandita, S. Park, R.B. Jenkins, N.J. Maihle, C.D. James, Neuro-Oncology 4 (2002) 86–94.

[58] A. Baras, C.A. Moskaluk, Journal of Molecular Histology 41 (2010) 153–164.

[59] L. Vitale, R. Casadei, S. Canaider, L. Lenzi, P. Strippoli, P. D'Addabbo, S. Giannone, P. Carinci, M. Zannotti, Gene 290 (2002) 141–151.

[60] H.I. Chen, M. Sudol, Proceedings of the National Academy of Sciences of the United States of America 92 (1995) 7819–7823.

[61] M. Mosquera-Caro, P. Helman, R. Veroff, J. Shuster, S. Martin, G. Davidson, J. Potter, R. Harvey, R. Hromas, E. Andries, S. Atlas, C. Wilson, K. Ar, Y.X. Xu, I.M. Chen, A. Carroll, B. Camitta, C. Willman, Blood 102 (2003) 4A–A.

[62] W.L. Carroll, D. Bhojwani, D.J. Min, E. Raetz, M. Relling, S. Davies, J.R. Downing, C.L. Willman, J.C. Reed, Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program (2003) 102–131.

[63] A. Baras, Y. Yu, M. Filtz, B. Kim, C.A. Moskaluk, Oncogene 28 (2009) 2919–2924.

[64] S. Park, C.D. James, Oncogene 24 (2005) 2495–2502.

[65] A.S. Baras, A. Solomon, R. Davidson, R. Moskaluk, Lab Invest (2011).

[66] R.J. Ingham, G. Gish, T. Pawson, Oncogene 23 (2004) 1972–1984.

[67] D.C. Eaton, B. Malik, H.F. Bao, L. Yu, L. Jain, Proceedings of the American Thoracic Society 7 (2010) 54–64.

[68] P. Kavsak, R.K. Rasmussen, C.G. Causing, S. Bonni, H. Zhu, G.H. Thomsen, J.L. Wrana, Molecular Cell 6 (2000) 1365–1375.

[69] R. Ouchida, T. Kurosaki, J.Y. Wang, The Journal of Immunology 185 (2010) 294–301.

[70] R. Ouchida, S. Yamasaki, M. Hikida, K. Masuda, K. Kawamura, A. Wada, S. Mochizuki, M. Tagawa, A. Sakamoto, M. Hatano, T. Tokuhisa, H. Koseki, T. Saito, T. Kurosaki, J.Y. Wang, Immunity 29 (2008) 33–43.

[71] T. Mund, H.R. Pelham, EMBO Reports 10 (2009) 501–507.

[72] J. Howitt, U. Putz, J. Lackovic, A. Doan, L. Dorstyn, H. Cheng, B. Yang, T. Chan-Ling, J. Silke, S. Kumar, S.S. Tan, Proceedings of the National Academy of Sciences of the United States of America 106 (2009) 15489–15494.

[73] A.A. Konstas, L.M. Shearwin-Whyatt, A.B. Fotia, B. Degger, D. Riccardi, D.I. Cook, C. Korbmacher, S. Kumar, The Journal of Biological Chemistry 277 (2002) 29406–29416.

[74] C. Ohzono, S. Etoh, M. Matsumoto, K.I. Nakayama, Y. Hirota, Y. Tanaka, H. Fujita, Biological and Pharmaceutical Bulletin 33 (2010) 951–957.

[75] K.W. Wucherpfennig, E. Gagnon, M.J. Call, E.S. Huseby, Call ME, Cold Spring Harbor Perspectives in Biology 2 (2010) a005140.

[76] M.S. Neuberger, K.J. Patel, P. Dariavach, K. Nelms, C.J. Peaker, G.T. Williams, Immunological Reviews 132 (1993) 147–161.