Evolutionary stories told by one protein family: ERM phylogeny in metazoans

Shabardina V.[1], Kashima Y.[2], Suzuki Y.[2], Makalowski W.[1]

[1]Institue of Bioinformatics, University of Muenster, Niels-Stensen-Strasse 14, Muenster, 48149, Germany.
[2]Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan.

Abstract

Ezrin, radixin, moesin, and merlin are the cytoskeletal proteins that participate in cell cortex rearrangements and also play role in cancer progression. Here we perform a comprehensive phylogenetic analysis of the protein family in metazoans spanning 87 species. The results describe a possible mechanism of the proteins origin in the root of *Metazoa*, paralogs diversification in vertebrates and acquirement of novel functions, including tumor suppression. In addition, a merlin paralog, present in most of vertebrates, but lost in mammals, has been described. We also highlight the set of amino acid variations within the conserved motifs as the candidates for determining physiological differences between the ERM protein paralogs.

Introduction

Ezrin, radixin and moesin of the ERM protein family, further ERMs, are cytoskeleton proteins that mediate physical connection between intermembrane proteins and actin filaments (Bretscher, Edwards and Fehon, 2002). They also act as signaling molecules, for example, as intermediaries in Rho signaling (Ivetic and

Ridley, 2004). Therefore, ERMs facilitate diverse cellular processes, ranging from cytoskeleton rearrangements to immunity (Bosanquet et al. 2014; McClatchey 2014; Ivetic & Ridley 2004; Marion et al. 2011). Dysregulation of ERMs' activity and expression impairs normal wound healing process and contributes to the progression of different types of tumors (Bosanquet *et al.*, 2014; Clucas and Valderrama, 2014).

The activity of ERMs in the cell is regulated by the conformational switch from the inactive, dormant folding to the active, "stretched" form. The inactive state is established through the auto-inhibitory interaction between the N-terminal FERM and the C-terminal CERMAD domains. That results in masking of the binding sites for membrane proteins in the FERM domain and actin binding cite (ABS) in the CERMAD (Turunen *et al.*, 1998). Upon activation ERMs consequently exposed to their two activating factors: PIP_2 (phosphatidylinositol 4,5-bisphosphate) binding via FERM domain and phosphorylation of a conserved threonine in CERMAD (Yonemura et al. 2002; Niggli & Rossy 2008). The important role during this transition belongs to the middle α-helical domain. In the dormant ERMs it forms coiled-coil structure, bringing N- and C-terminal domains together (Schliep *et al.*, 2017).

Little is known about individuality of each ERMs in, both, sickness and health. The three proteins are paralogs and share high amino acid sequence similarity (~75% in human) (Funayama *et al.*, 1991; Lankes and Furthmayr, 1991). They demonstrate similar cellular localization and are often discussed as functionally redundant. However, data from several studies on knock-out mice revealed different phenotypes targeting different organs; only ezrin's depletion was lethal (Kikuchi *et al.*, 2002; Kitajiri *et al.*, 2004; Saotome, Curto and McClatchey, 2004; Liu *et al.*, 2015). Special interest in ERMs' role in cancer stimulated multiple studies. Their dysregulation can lead to disruption of cell-cell contacts, enhanced cell migration and invasion, and higher cancer cells survival (Clucas and Valderrama, 2014). Some studies showed that ezrin, radixin and moesin may exploit different cellular mechanisms in tumors. (Pujuguet *et al.*, 2003; Kobayashi *et al.*, 2004; Debnath and Brugge, 2005; Estecha *et al.*, 2009; Chen *et al.*, 2012; Valderrama, Thevapala and Ridley, 2012).

One of the well-known tumor suppressor factors is an ERM-like protein merlin. It shares 46% amino acid sequence similarity with the whole-length ezrin and 86% with its FERM domain for human (Turunen *et al.*, 1998). Mutations in merlin results in development of neurofibromatosis type 2 characterized by formation of schwannomas (Stickney *et al.*, 2004; Curto *et al.*, 2007). Tumor suppression activity of merlin is linked to the blue-box region in its FERM domain, conserved serine Ser518 and last 40 residues in the C-terminal (Lallemand, Saint-Amaux and Giovannini, 2009; Cooper and Giancotti, 2014). Two acting binding sites in merlin are mapped to the FERM domain, while typical to ERMs C-terminal ABS is absent (Roy, Martin and Mangeat, 1997; Brault *et al.*, 2001).

With the more research being done, it is getting clear that ezrin, radixin, moesin, and merlin can invoke different physiological effects in different tissue types, especially in cancer (Clucas and Valderrama, 2014). However, their highly conserved sequence and tertiary structure make it not a trivial task to distinguish proteins' functions in vivo. Phylogenetic approach can be an effective tool in resolving the problem, as it enables precise paralogs characterization by tracing evolutionary history of the binding sites and conserved amino acid motifs. So far only few phylogenies of ERMs and merlin have been described in the literature. As a rule, they feature limited taxonomy representation or are included in the studies as an accessory and brief part of discussions (Turunen *et al.*, 1998; Golovnina *et al.*, 2005; Phang *et al.*, 2016). Thus, although these phylogenies recover some interesting patterns, they do not provide with the full understanding of ERMs and merlin evolution. The studies agree on the fact that the proteins are highly conserved within the metazoan clade (multicellular animals), and especially in vertebrates. Moreover, the appearance of the ERM proteins and merlin in the tree of Life seems to coincide with the origin of multicellularity in animals (Bretscher, Edwards and Fehon, 2002; Omelyanchuk *et al.*, 2009; Nambiar, McConnell and Tyska, 2010; Sebé-Pedrós *et al.*, 2013). This view is supported by the recent discovery of ERM-like proteins in *Choanoflagellata* and *Filasterea*, the closets unicellular relatives of metazoans (Fairclough *et al.*, 2013; Suga *et al.*, 2013).

The position of merlin relative to the ERM family is differing in the literature, some researchers excluding merlin from discussions of the ERM family. Nevertheless, regarding that the proteins share evolutionary history and structural characteristics, it is reasonable to unite them in one group. In this work we conduct the first comprehensive phylogenetic analysis for the ERM family and merlin, that includes data from all sequenced by the time metazoan genera. The results describe the ERM and merlin sequence conservation and paralog number diversity within the clade of *Metazoa.* We suggest that the increased organism complexity led to diversification of the protein paralogs in vertebrates. Moreover, we highlight the importance of phylogenetic studies of paralogs, in general, in application to experimental biology, especially in disease-related research.

Materials and methods

Data collection

The amino acid sequences of ezrin, radixin, moesin, and merlin were collected using BLASTp (Altschul *et al.*, 1997) with the human protein sequences as queries (ezrin NP_001104547.1, radixin AAA36541.1, moesin NP_002435.1, merlin NP_000259.1) against non-redundant (nr) protein sequences collection at the NCBI database. The selected sequences were manually monitored to exclude database duplicates, splice variants, truncated sequences and to reconstruct correct protein sequences when needed. If several splice variants were described for an organism, the longest one was chosen for the analysis. Only the sequences that spanned all three ERM domains were selected for the analysis to avoid data contamination, since hits in only FERM domain, the most conserved part of the protein family, can result in pulling members of the 4.1 protein superfamily or other proteins, for example MyTH4-FERM domain type myosins. PFAM (Finn *et al.*, 2006), InterProScan (Jones *et al.*, 2014) and CDD domain search (Marchler-Bauer *et al.*, 2015) were used for domain structure analysis and verification.

The taxa selection was made based on the following requirements: 1) every described order of *Metazoa* should be represented by one species, although some exceptions were made in order to balance taxa representation; 2) whole genome or transcriptome of a representative species should be sequenced and available; 3) high quality of the genome assembly annotation. In the case if no representative genome was available for an order, tBLASTn search was run against all available nucleotide sequences for that taxa. Search for possible homologs of ezrin, radixin, moesin, and merlin was performed among other *Opisthokonta* (*Holozoa*, *Nuclearia*, *Fungi*) and *Amoebozoa*, *Excavata, Archaeplastida* (includes green plants), SAR cluster (*Stramenopiles, Alveolata,* and *Rhizaria*), and other protist groups (refer to the tree of Life scheme (Adl *et al.*, 2012)). *Prokaryota* and *Archaea* nucleotide sequences were scanned for the whole length proteins or only FERM domain using tBLASTn search against nr nucleotide collection at NCBI. The taxonomic structure describing the final dataset can be viewed at the Supplementary (Table S1) and is based on the topologies employed by NCBI Taxonomy database (Federhen, 2012) and the Tree Of Life project (Letunic & Bork). The taxa variety will be further discussed in terms "vertebrate" and "invertebrate", the later including the rest of *Eumetazoa*.

Reconstruction of the ERM+merlin phylogeny

Multiple sequence alignment was generated using MAFFT software (Katoh *et al.*, 2002) with the PAM70 (Dayhoff, 1965) substitution matrix (defined as optimal by ProtTest3 (Darriba *et al.*, 2011)) and manually edited to remove uninformative columns; CLUSTALX (Larkin *et al.*, 2007) and Geneious (Geneious, 2019) were used for visualization. Maximum Likelihood (ML) phylogenetic trees were build using RAxML tool (Stamatakis, 2014) with the parameters estimated by running RAxML parameter test (Stamatakis, 2015). As a result, the PROTGAMMALG model was chosen, where GAMMA model estimates the substitution rate between sites and LG is amino acid substitution matrix (Le and Gascuel, 2008). The statistical support for the tree clustering was calculated by running 1000 bootstrap replicates. The resulting trees were inspected and edited using iTOL online tree viewer (Letunic and Bork, 2019) and FigTree software (Rambaut A., 2019).

A reduced data set (Supplementary, Table S1) was used to reconstruct a tree for analyzing evolutionary relationships between the proteins from unicellular organisms and metazoans with the same model as described earlier. Ancestral sequence reconstruction was performed for this data set using rooting option and defining marginal ancestral states option by RAxML with the model PROTGAMMALG.

MEGA (Kumar, Stecher and Tamura, 2016) software was used for an alternative tree reconstruction for neighbor-joining and parsimony algorithms with the 500 bootstrap replicates and LG amino acid substitution matrix. Bayesian inference method was also applied on MrBayes tool (Ronquist *et al.*, 2012) under LG matrix. Six chains were run for 3000000 generations, every 1000 generations trees were sampled in two runs. The first 25% of trees were discarded before constructing a consensus tree.

Protein sequence analysis

Tertiary structure of polypeptides was predicted by PEP-FOLD3 (Lamiable *et al.*, 2016). Estimation of proteins' biochemical and biophysical characteristics from their amino acid sequences was done with ExPASy ProtParam (Gasteiger *et al.*, 2005). Conserved amino acid motifs were analyzed using MEME suite (Bailey *et al.*, 2009).

Custom python, perl and bash scripts were used for data processing.

Results

Full length ERM/merlin-like proteins appeared within Metazoa-Filasterea-Choanoflagellata group

Search for the ERM and merlin homologs throughout all eukaryotic clades resulted in the selection of 260 protein sequences spanning 87 species, including metazoan and unicellular organisms. ERM-like proteins are also present in choanoflagellates (*Salpingoeca rosetta* and *Monosiga brevicollis*) and filastereans (*Capsaspora owczarzaki*). A sequence of 298 amino acids (Supplementary File 2) identified in corallochytrean *Corallochytrium limacisporum* revealed 25% sequence identity to the FERM domain of human ezrin based on the tBLASTn search against the species' whole genome sequence. Although such similarity level does not guarantee structural alikeness, as according to (Rost and Sander, 1994) it should be at least 30%, the domain annotation by PFAM indicated that this polypeptide from *C. limacisporum* belongs to the class of FERM domain with high statistical support (e-value < 10^{-5} for each of the three subdomains of FERM). This sequence was not taken for the tree reconstruction. However, sequence-based prediction of biochemical properties revealed that the two FERM domains, human and corallochytrean, exhibit distinct features, including pI (8.75 for human ezrin FERM domain and 6.79 for the *C. limacisporum* polypeptide), amino acid content, instability index and hydropathicity. In particular, metazoan FERM domain is predicted to be more hydrophilic than its suggested corallochytrean homolog (grand average of hydropathicity (GRAVY) index is -0.530 and -0.270, respectively) and less stable (instability index estimated to be 43.57, i.e. unstable protein, and 31.80, stable, respectively). The only binding site conserved in the corallochytrean FERM is the site for PIP_2 interaction. No ERM-like proteins or FERM-like domains could be found in the other inspected taxa. The list of all the taxa and the corresponding proteins IDs used for the analysis can be found at the Supplementary (Table S1).

ERM+merlin protein family is conserved throughout all the metazoan orders and three unicellular species

Sequence comparison of the selected proteins demonstrated that the domain structure and most of the known binding sites are conserved throughout the whole metazoan clade, although there is some length variation. The amino acid motifs conservation analysis (Fig. 1) revealed that the homologs of such early branching animals as *Trixhoplax adhaerens* (*Placozoa*) and *Amphimedon queenslandica* (*Porifera*) demonstrate similar pattern to the mammalian proteins with a high statistical significance. FERM and CERMAD domains are characteristically well preserved, even in the proteins from the unicellular organisms; alpha-helical middle domain is the least conserved ERM domain as has been previously noticed (Phang *et al.*, 2016). The N-terminal part of FERM shows some length variation throughout different taxa by including short, non-conserved amino acid stretch. More striking length variation is within the region separating the alpha-helical domain and CERMAD. It is short for proteins from vertebrate animals, but is increased for all other taxa, the longest is in the protein from *C. owczarzaki* – 335 residues. The sequence of this region is poorly preserved between the taxa, that suggests its specificity to each clade. The proteins from the species of flat worms, *Inoshia linei* and *C. owczarzaki* are the most divergent.

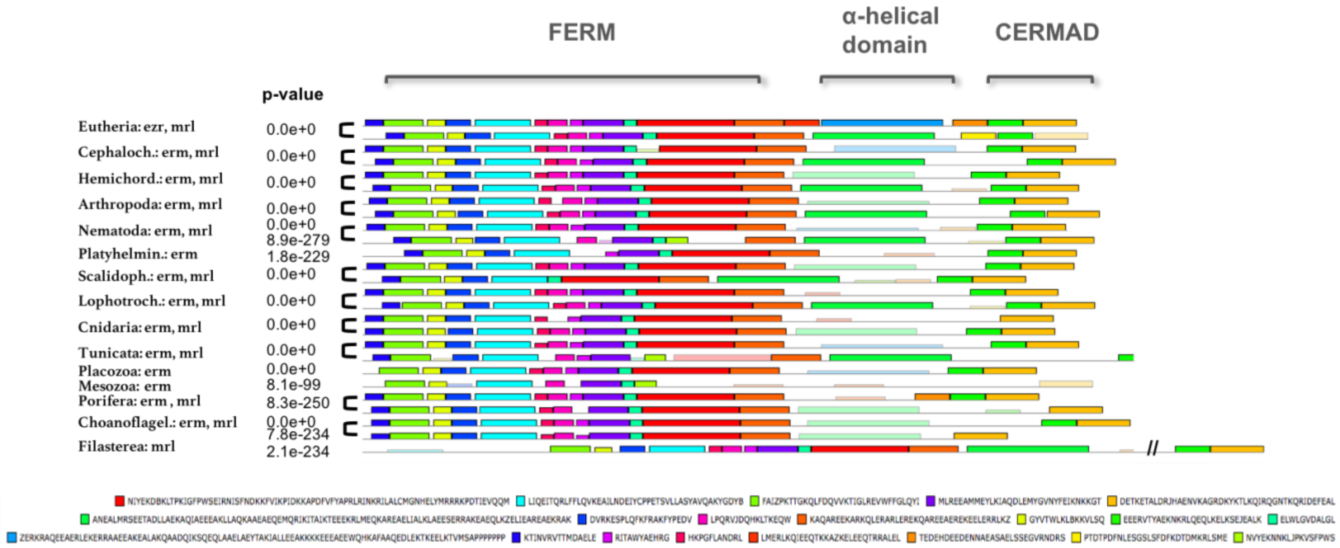
FERM
α-helical domain
CERMAD
p-value
Eutheria: ezr, mrl
Cephaloch.: erm, mrl
Hemichord.: erm, mrl
Arthropoda: erm, mrl
Nematoda: erm, mrl
Platyhelmin.: erm
Scalidoph.: erm, mrl
Lophotroch.: erm, mrl
Cnidaria: erm, mrl
Tunicata: erm, mrl
Placozoa: erm
Mesozoa: erm
Porifera: erm , mrl
Choanoflagel.: erm, mrl
Filasterea: mrl
0.0e+0
0.0e+0
0.0e+0
0.0e+0
0.0e+0
8.9e-279
1.8e-229
0.0e+0
0.0e+0
0.0e+0
0.0e+0
0.0e+0
8.1e-99
8.3e-250
0.0e+0
7.8e-234
2.1e-234
//
NIYEKDBKLTPKIGFPWSEIRNISFNDKKFVIKPIDKKAPDFVFYAPRLRINKRILALCMGNHELYMRRRKPDTIEVQQM
LIQEITQRLFFLQVKEAILNDEIYCPPETSVLLASYAVQAKYGDYB
FAIZPKTTGKQLFDQVVKTIGLREVWFFGLQYI
MLREEAMMEYLKIAQDLEMYGVNYFEIKNKKGT
DETKETALDRJHAENVKAGRDKYKTLKQIRQGNTKQRIDEFEAL
ANEALMRSEETADLLAEKAQIAEEEAKLLAQKAAEAEQEMQRIKITAIKTEEEKRLMEQKAREAELIALKLAEESERRAKEAEQLKZELIEAREAEKRAK
DVRKESPLQFKFRAKFYPEDV
LPQRVJDQHKLTKEQW
KAQAREEKARKQLERARLEREKQAREEAEREKEELERRLKZ
GYVTWLKLBKKVLSQ
EEERVTYAEKNKRLQEQLKELKSEJEALK
ELWLGVDALGL
ZERKRAQEEAERLEKERRAAEEAKEALAKQAADQIKSQEQLAAELAEYTAKIALLEEAKKKKEEEAEEWQHKAFAAQEDLEKTKEELKTVMSAPPPPPPP
KTINVRVTTMDAELE
RITAWYAEHRG
HKPGFLANDRL
LMERLKQIEEQTKKAZKELEEQTRRALEL
TEDEHDEEDENNAEASAELSSEGVRNDRS
PTDTPDFNLESGSLSFDFKDTDMKRLSME
NVYEKNNKLJPKVSFPWS

Figure 1. MEME conserved amino acid motif analysis. Motifs in pale color were found by the scanning algorithm based on the de-novo motif identification (bright color). Note the reduced length of the region separating the alpha-helical and CERMAD domains in eutherian proteins. See Supplementary, Table S1 for the list of the sequences used.

The binding sites for EBP50, ICAM-2, NHERF2 binding partners, PIP$_2$, ABS, and intramolecular interaction sites for ERM-like proteins (Bretscher et al. 2002; McClatchey 2014; Ivetic & Ridley 2004; Marion et al. 2011; Li et al. 2015) can be identified in the most of the sequences (Supplementary, Table S1). This signifies that the protein family is responsible for the basic cellular functions and some of the interactions might have been established in the early metazoan history. The most conserved regions in the proteins are PIP$_2$ binding sites in F1 and F3 subdomains of FERM and C-terminal ABS (specifically, KYKTL motif) for ERM-like proteins. The proteins that have deviation come from the early branching metazoans *Molgula tectiformis* (*Tunicata*), *Amphimedon queenslandica* (*Porifera),* and endo- and exo-parasites (*I. linei*, flat worms, and blood sucking leech).

Interestingly F1 PIP_2 binding motif is not preserved only in *M. tectiformis*. Some variation in KYKTL motif can be seen in the proteins from *Echinodermata* and shark *Callorhinchus milii*.

The merlin sequences can be classified into three groups, see also Fig. 2: 1) non-vertebrate proteins; 2) vertebrate merlin1, except *Eutheria* and *Metatheria* 3) all-vertebrate merlin2 (we assign here merlin1 and merlin2 names to the merlin paralogs). The group 2 comprises proteins coded by the paralogous gene that has not been described before (the list of the proteins is in the Supplementary, Table S1). It has a unique insertion of ~27 amino acids (SKHLQEQLNELKTEIEALKLKERET) in C-terminal domain and lacks tumor-suppression region characterized in human merlin2 between residues 532-579. It does have merlin specific blue-box region and the conserved Ser518. Invertebrate merlins, group 1, are similar to vertebrate merlin1 but lack the 27-amino acid insertion. Merlin2, group 3, is present in all vertebrate taxa and has an additional actin binding site in the F1 of FERM and the tumor-suppression amino acid stretch in its CERMAD (residues 532-579 in human). Judging from the short branch lengths of the merlin2 clade, it seems that its evolution went under higher functional constrains than in the case with merlin1. The blue-box region can also be identified in the two proteins from the unicellular species: XP_004364665.2 in *C. owczarzaki* and XP_004991962.1 in *S. rosetta*.

Phylogenetic tree for the ERM+merlin family

The reconstructed ML tree resulted in high phylogenetic resolution among vertebrates, while the branching for most of the other taxa have low statistical support (Fig. 2). The alternative methods of phylogenetic reconstruction, including neighbor-joining algorithm, parsimony method, and Bayesian inference, could not improve the resolution (data not shown). Two former trees revealed almost identical branching and statistical confidence. The Bayesian reconstruction could not achieve conversion after 3000000 generations. The run was terminated and the consensus tree was built anyway. It featured clustering of vertebrate proteins supported by high bootstrap values, but unresolved branching for the invertebrate sequences. One of the reasons can be an unequal representation of the taxa due to the lack of sequencing data or incomplete genomes sequencing of the invertebrate clades (see the taxa missing in the analysis in the Table S1, Supplementary). Another complication for the analysis could be high divergence of the proteins' sequences between evolutionary distant lineages. The ML tree is used for the further discussion.

The most "eccentric" sequence in the reconstructed phylogeny is that of hypothetical protein from *I. linei*, that, although features all three ERM domains, has the highest substitution rate and does not cluster with any other groups, neither it can be defined as ERM-like or as merlin-like protein. This is not surprising, as *I. linei*, a representative of orthonectids, is a parasitic animal, and its hermaphrodite nature, fast reproductive cycles and high level of inbreeding can be the reason that makes its genome distant from the genomes of other metazoans (Lu *et al.*, 2017).

The protein from *I. linei* was arbitrary chosen to separate the tree into two major clusters: ERM-like and merlin-like groups. Therefore, any "hypothetical" or "unknown" proteins can be annotated whether as ERM-like or as merlin-like, based on their position relative to *I. linei*'s sequence. Besides improving annotation of such proteins, some false annotations deposited in the public protein databases were corrected (Supplementary, Table S1).

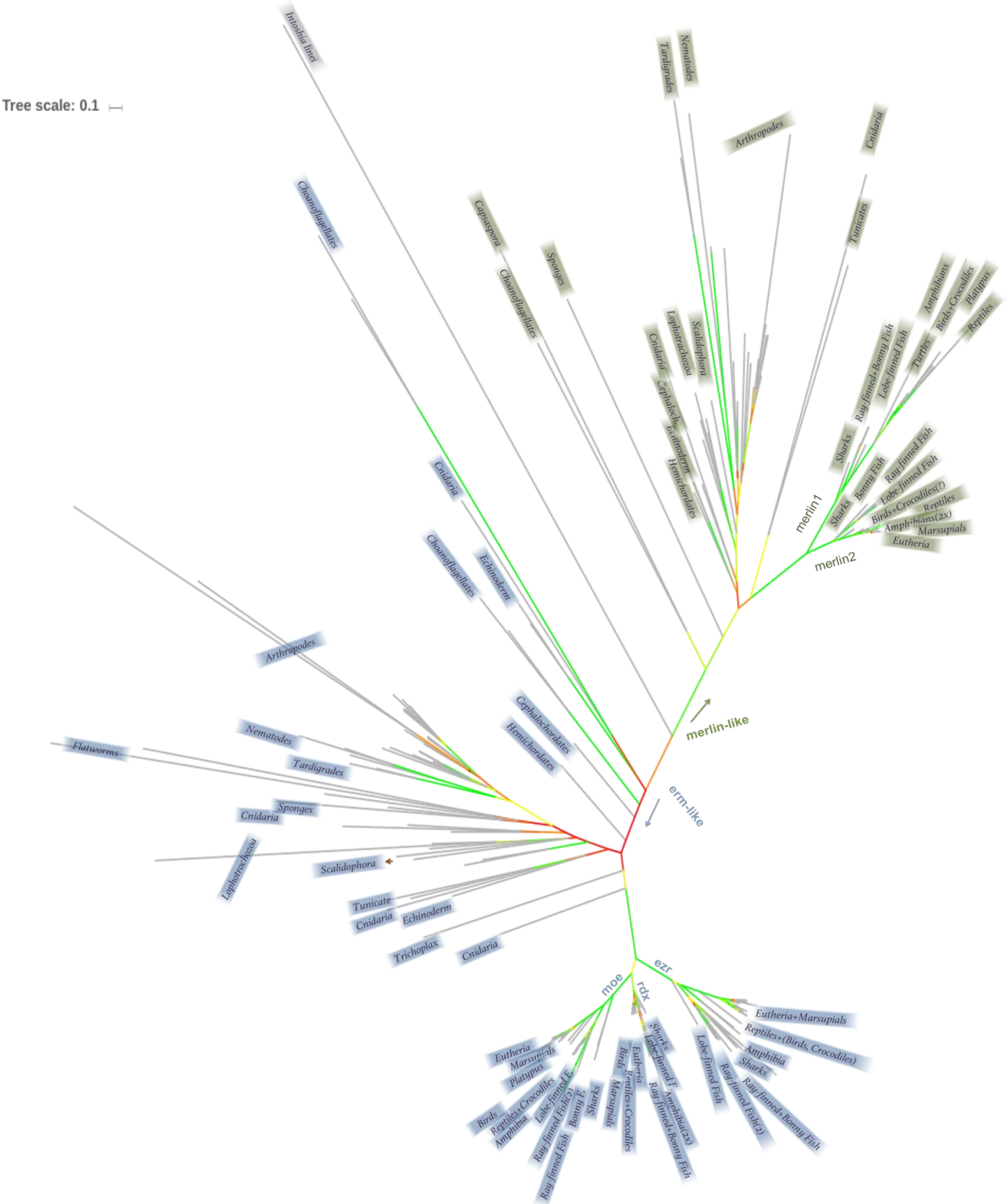

Figure 2. Phylogenetic ML tree: ERM+merlin family. Color scheme for bootstrap values: green – 80-100%, yellow – 50-70%, red – below 50%. Refer to the tree file in Newick format to see branch lengths and bootstrap values (Supplementary File 3).

The most interesting conclusion that can be made from the assumption of ERM-like and merlin-like clustering is characterization of the proteins from the unicellular species: *C. owczarzaki*'s protein XP_004364665.2 and one of the *S. rosetta*'s proteins XP_004997754.1, that can be assigned as merlin-like. The other four proteins, all from choanoflagellate species, are more similar to ERM-like group: XP_001743289.1 and XP_001746613.1 (*M. brevicollis*) and XP_004991962.1, XP_004994097.1 (*S. rosetta*). That provides an insight into the origin of the ERM+merlin family and suggests that merlin and ERMs diverged from the common ancestral protein before emergence of *Metazoa*.

Cnidaria, Placozoa, Tunicata, Echinodermata, Scalidofora, Lophotrochozoa, Porifera, Hemichordata, Cephalochordata, and *Mesozoa* taxa branching could not be defined by this analysis with statistical confidence. Although, several interesting conclusions can be drawn. Proteins representing three cnidarian species from different orders are not monophyletic. The three animals exhibit very different life styles: *Hydra vulgaris* lives in fresh waters, *Exaiptasia pallida* is anemone and dwells in the waters of the Atlantic Ocean, and *Stylophora pistillata* is a coral and common in the Indo-Pacific region. Therefore, it is possible that these different life styles caused the high diversification of the ERM+merlin protein family.

The three clearly distinguishable groups, beside vertebrates, are *Insecta* (99% bootstrap value), *Nematoda* (99%) and *Tardigrada* (100%). ERM-like proteins from nematodes and tardigrades cluster together with the bootstrap support of 60%. The phylogenetic position of tardigrades is so far unclear and developmental stages and genetics of these animals share features of both, arthropods and round worms (Gabriel *et al.*, 2007; Yoshida *et al.*, 2017). Interesting to note that the proteins from the parasitic organisms are characterized by higher substitution rate (longer branches). For example, the ERM-like protein in *Pediculus humanus corporis* (body louse), one of the three ERM-like proteins in *Helobdella robusta* (leech) and ERM-like protein in all three representatives of *Platyhelminthes*.

Ezrin divergence as a first ERM paralog in *Vertebrata* is supported with 100% bootstrap value, with radixin and moesin separating later from a preceding radixin/moesin form. Radixin seems to be the slowest evolving protein in the family. *Coelacanthimorpha* (*Latimeria chalumnae*), *Teleostei* (bonny fishes), *Holostei* (*Lepisosteus oculatus*, spotted gar), *Chondrichthyes* (cartilaginous fishes) and *Amphibia* are clearly separated from the closely related group of *Prototheria-Metatheria-Theria-Sauria* (reptiles, turtles, crocodiles, birds). This is true for each of ezrin, radixin, and moesin clusters. Interesting to note that the phylogeny of the birds' proteins parallel their habitat geography: *Calypte anna* (natural areal is in New Zealand) cluster together with *Gallus gallus* that originated in the area of Indochina, Indonesia, and Phillipines, while proteins from birds of America form separate group.

It is likely that the three ERM paralogous genes appeared as a result of the two rounds of the whole genome duplication (WGD) that took place in the root of vertebrates and a consequent loss of one copy of the gene. An additional increase of the paralogs number in teleost fishes is likely to be the result of the lineage specific WGD event. Although the possibility of a duplication event of a local character cannot be excluded. Intriguing observation is the existence of the two merlin paralogs, merlin1 and merlin2, in vertebrates, as already mentioned in the above section. Merlin1 was lost in *Eutheria* and *Metatheria* (Fig. 3). This was previously unknown or unappreciated, that can be explained by the fact that most of the merlin studies were done on the representatives of *Eutheria* clade (human, mouse), thus describing only merlin2 paralog. Therefore, it is not surprising that merlin1 went unnoticed.

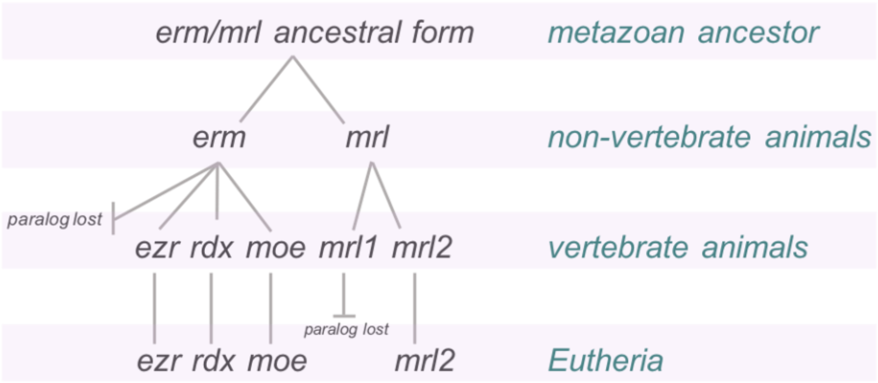

Figure 3. Scheme of gene duplication and paralogs lost in vertebrates. Erm – ERM-like, mrl – merlin-like, ezr – ezrin, rdx – radixin, moe – moesin.

Paralog number diversity

All invertebrate taxa included in this analysis are characterized by the presence of one or two ERM-like paralogs and zero or one merlin-like paralogs (Table 1). In particular, *I. linei*, *T. adhaerens* and all three species of *Platyleminthes* have only one gene coding for an ERM-like protein and no merlins. Among four *Nematoda* species (from four orders) only *Trichinella pseudospiralis* possesses both, ERM-like and merlin-like, proteins. These observations highlight the trend of simplification in parasites. All other invertebrate taxa have at least one ERM-like and one merlin-like proteins.

Vertebrates have, as a rule, three ERM proteins (ezrin, radixin, and moesin) and two merlins. Although several very interesting exceptions can be found. As it was already mentioned, *Metatheria* (marsupials) and *Eutheria*, or *Theria,* keep only merlin2. Opposite to that, *Prototheria* (represented by platypus) lost merlin2, but settled with merlin1 only. We indicated a few more cases of lineage specific paralogs lost or gain. Thus, teleost fishes have four to six ERM paralogs in different combinations; some teleosts have both merlins, some lost merlin1, but no pattern can be identifiable. Atlantic salmon (*Salmo salar*) has seven ERM genes (two of ezrin and of moesin, and three of radixin) that is in accordance with the hypothesis of a lineage specific WGD in salmonids (Glasauer and Neuhauss, 2014). Further, *Neognathae* birds possess only two of the three ERM proteins, ezrin always being present; while *Paleoghathae* birds (namely, *Tinamus guttatus*) have all three ERM proteins and one merlin. *Xenopus laevis* has four ERM proteins and three merlin paralogs, likely the result of another lineage specific WGD that took place around 40 million years ago (Van de Peer, Maere and Meyer, 2009). The rarest case comes from the exotic taxa: *Sarcophilus harrisii* (Tasmanian devil, *Metatheria*) has one ezrin and one merlin only; further, *Ornithorhynchus anatinus* (platypus, *Prototheria*) has one moesin and one merlin only. And, finally, *Phascolarctos cinereus*, a koala, although belongs to *Metatheria*, has both merlins, plus all three ERMs. Also, elephant shrew (*Elephantus edwardii, Eutheria*) has two ezrin genes, no moesin or radixin, and one merlin gene. To avoid any errors during the paralogs number estimation, we conducted tBLASTn searches against any available RSA sequences and whole genome sequences for the species that show lack of any of the paralogs.

Table 1. Number of paralogous genes in different metazoan lineages. Lineage specific paralog number is highlighted in red. For more detailed count refer to the Table S1 in the Supplementary.

	ERM-like	Merlin-like	comments
Mesozoa	1	0	*Intoshia linei*
Placozoa	1	0	*Trichoplax adhaerens*
Porifera	1	1	
Cnidaria	1	1	
Platyhelmynthes	1	0	
Nematoda	1	1	
Arthropoda	1-2	0-1	
Lophotrochozoa	1-2	1	
Hemichordata	1	1	*Saccoglossus kowalevskii*
Cephalochordata	1	1	
Tunicata	1-2	1	
Chondrichthyes	3	1-2	cartilage fishes
Coelacanthimorpha	3	2	*Latimeria chalumnae*
Holostei	3	2	*Lepisosteus oculatus*
Teleostei	4-6	1-2	bonny fishes
Reptilia	3	1-2	
Aves	2-3	1-2	
Crododylia	3	2	
Amphibia	3-4	2-3	
Eutheria	3	1	
Metatheria	1-3	1-2	marsupials
Prototheria	1	1	platypus

Unicellular ancestry of ERM+merlin family

With the assumption that the closest unicellular relatives of animals are choanoflagellates and filastereans, we assumed that their ERM-like and merlin-like proteins are the best candidates for speculating about the proteins ancestral form. We built a small phylogeny tree, including only few sequences from a eutherian representative (*H. sapiens*) and *S. rosetta, M. brevicollis, and C. owczarzaki*. Eliminating the rest of the taxa decreases the reliability of the reconstruction, but, regarding the high sequence diversity and scarcity of the data for invertebrates, this approach is the most straightforward. The tree highlights three groups: 1) merlins (*Eutheria, Choanoflagellata, Filasterea*), although this cluster can be separated into two subgroups: filasteran and eutarian+choanoflagellate; 2) highly specialized, or the result of genome mis-assembly, proteins of the two choanolagellate short proteins lacking most of the middle domain; 3) ERM group comprising eutherian and choanoflagellate homologs (Fig. 4).

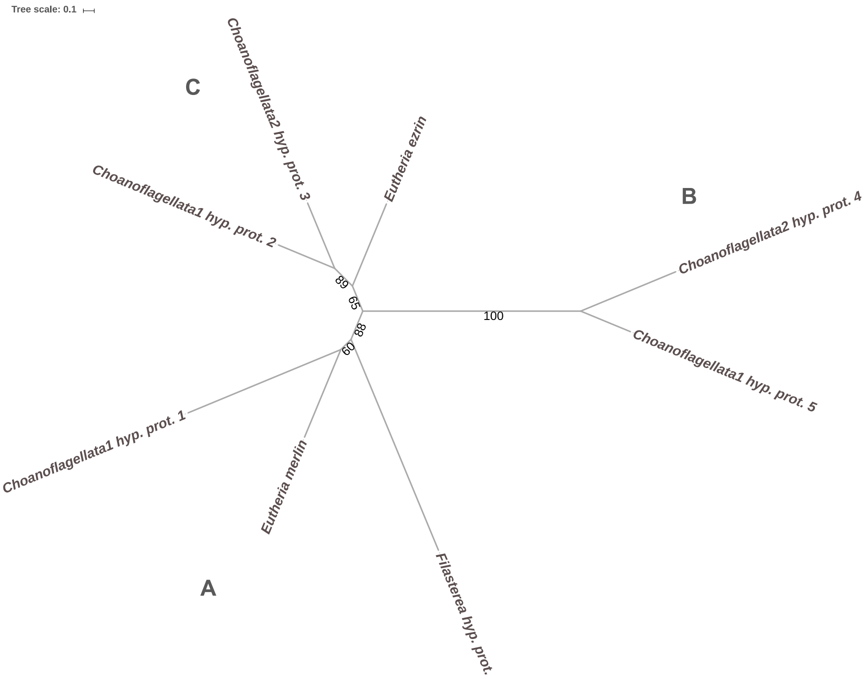
Tree scale: 0.1
C
Choanoflagellata2 hyp. prot. 3
Eutheria ezrin
Choanoflagellata1 hyp. prot. 2
B
Choanoflagellata2 hyp. prot. 4
89
65
100
88
60
Choanoflagellata1 hyp. prot. 5
Choanoflagellata1 hyp. prot. 1
Eutheria merlin
A
Filasterea hyp. prot.

Figure 4. ML tree indicating three clusters of ERM/merlin-like proteins in the unicellular organisms: two species form choanoflagellates and one filasterean species. Group A unites merlin and merlin-like proteins, clade B is represented by the two proteins with a rudimental middle domain, clade C – ERM-like proteins. The tree in Newick format is in Supplementary File 4.

Although the tree is more a scheme than an illustration of the phylogenetic relationships, it is useful for speculating about the origin of the modern protein family. We, therefore, modelled an ancestral sequence for ERM+merlin family based on this tree (Supplementary File 1). It suggests the conserved domain structure and presence of the most characteristic binding sites (for PIP_2, intramolecular interaction, ABS), but includes an additional 63 amino acid region separating alpha-helical domain and CERMAD. Computational prediction of the tertiary structure of this insertion characterized the corresponding polypeptide as an extended structure with a low probability to form alpha-helix (Fig. 5).

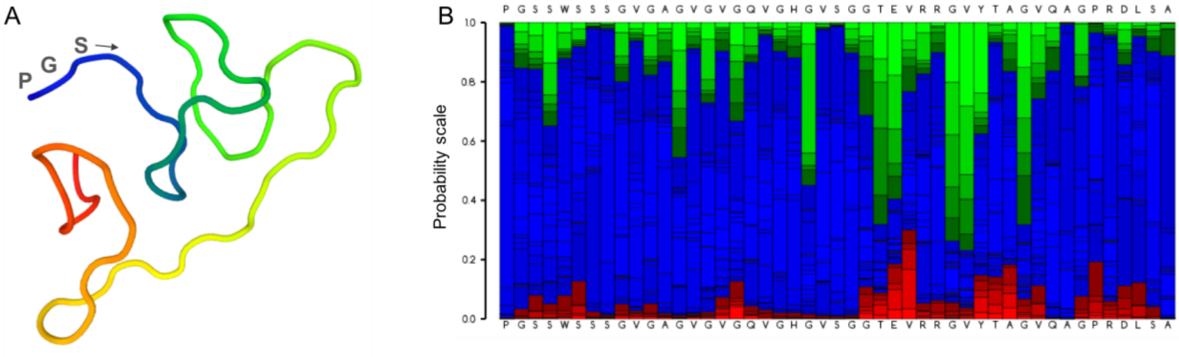

Figure 5. Structural modelling. A: 3D peptide structure prediction for the first 50 residues of the insertion in the reconstructed ancestral protein. Coloring is used for a better visualization. B: Probability plot for the first 47 amino acids of the insertion, each amino acid is assigned a probability to be included in a particular structure: red – alpha-helix, blue – random coil, green – extended structure. Higher values mean higher probability. The plot suggests that the analyzed polypeptide folds into an extended structure. Reconstruction was done in PEP-FOLD3.

Discussion

Phylogenetic analysis is a valuable approach in protein annotation and characterization and can significantly improve genome annotations. Unfortunately, it requires more time and efforts than an automated genome annotation pipelines.

However, it can and should be routinely used for proteins that are actively studied in vivo and for medical applications, if not for evolutionary studies. The presented here phylogenetic tree allows distinguishing between ERM-like and merlin-like subgroups of the protein family and between different ERM paralogs in the approach proposed more than 20 years ago by Jonathan Eisen (Eisen, 1998). As a result, we were able to improve annotations of these proteins in different species, as well correcting errors in several cases when, for example, an ezrin was erroneously called a radixin or a merlin.

Furthermore, the tree clearly demonstrates that assignment of the invertebrate proteins to “ezrin”, “radixin” or “moesin” is inconsistent, since the divergence of the three ERM paralogs happened in the root of *Vertebrata*. A good example are the two incorrect assignments at the NCBI protein database: XP_002160112.1 protein annotated as radixin in *H. vulgaris* and NP_727290.1 protein annotated as moesin in *D. melanogaster*. We suggest to restrict the names ezrin, radixin, moesin only to the vertebrate proteins, while referring to others as ERM-like or merlin-like.

Some inconsistencies also happen for the studies with vertebrate ERMs. For example, there are expression data published for chicken moesin (Li and Crouch, 2000) and experiments done in chicken erythrocytes with endogenous moesin (Winckler *et al.*, 1994), that is questionable, as our phylogenetic analysis indicates that chicken (*Gallus gallus*) lost moesin gene and has only ezrin and radixin. Such an example shows the importance of incorporating bioinformatics milieu in the wet-lab studies, especially for the proteins from the less studied, not model organisms, to avoid confusion and data discrepancy.

To get an insight on the evolution of the ERM+merlin protein family, we collected protein sequences that span 84 species of *Metazoa* and 3 unicellular species from the clades *Choanoflagellata* and *Filasterea.* We could also identify a FERM-like domain with the conserved PIP_2 binding site in a species from *Corallochytrea* clade, likely the first lineage with FERM domain in the tree of Life

(Fig. 6A). This finding is in agreement with the study of the domain gain and lost in different taxa, that estimated that the FERM's origin took place in *Holozoa* (Grau-Bové *et al.*, 2017). This is also consistent with the fact that FERM domain is the most conserved part of the proteins from the family. It is possible, that this corallochytrean FERM domain is a full-length protein, as computational prediction by ExPASy ProtParam algorithm estimated that it likely folds into a stable structure. We suggest that this FERM-like protein is a membrane binding protein involved in communication of the cell and extracellular environment. It can be similar to a predecessor of the ERM+merlin family. Domain shuffling and/or shifts within the open reading frame of the predecessor gene could lead to the origin of a longer protein with an acting binding capability of its newly acquired C-terminal part, i.e. with a scaffolding function similar to that of the modern ERMs. Such mechanisms, for example, were described in some works discussing origin of novel proteins within emergence of animal multicellularity (Shalchian-Tabrizi *et al.*, 2008; Richter and King, 2013).

We found first full-length ERM+merlin-like proteins in the closest unicellular relatives of metazoan – choanoflagellates (*S. rosetta* and *M. brevicollis*), and filasterean (*C. owczarzaki*). These proteins combine some characteristics of ERM-like (for example, C-terminal ABS) and merlin-like (multiple dispersed prolines at the C-terminal end of the alpha-helical domain, absence of the ERM-specific $R_{/K}EK_{/R}EEL$ repeat within the alpha-helix) groups. It points out that modern merlin-like and ERM-like proteins likely emerged from the same ancestral form in the root of *Metazoa*. Phylogenetic reconstruction of a sequence of this ancestral form suggests that it could bind actin filaments and PIP_2 lipids, therefore, could perform the function of mechanical linkage of the cell membrane and underlying actin filaments.

An ancestor of filasterean and choanoflagellates was likely able to form transient cell-cell or cell-surface contacts and could exploit its ERM/merlin-like protein for this purpose, as *S. rosetta* and *C. owczarzaki* probably do, as suggested by sequence analysis of their homologs. This trend could be expanded in primitive metazoans to the scaffolding function within cell-cell contacts, for example, in *Trichoplax*. ERM-like protein of this early branching metazoan already possesses the key features of the ERM family: ABS and binding sites for PIP_2, EBP50, ICAM-2, NHERF2, and intramolecular binding sites. It can likely participate in the arrangement of the only type of cell-cell contact the animal exploits - adherens junctions. Similarly, one study suggested that ERM proteins were involved in the development of the filopodia in metazoans (Sebé-Pedrós *et al.*, 2013). However, more sequencing data from other unicellular taxa and in vivo experiments are required to support or reject this hypothesis.

Function of the protein family in the unicellular organisms was probably limited and restricted to scaffolding, partly because of inability to activity regulation. Indeed, the auto-inhibitory interaction in the unicellular proteins is questionable, due to the presence of an extra amino acid stretch between the middle and CERMAD domains (Fig. 6B). Prediction of the tertiary structure of this insertion indicates that it is unlikely an alpha helix, therefore, such inclusion could drastically change protein folding. Further in the metazoan evolution, decreasing of the distance between the middle and the C-terminal domain could be one of the evolutionary modifications that facilitated ERMs' characteristic auto-inhibitory, intramolecular binding. Therefore, our hypothesis is in accordance with the rheostat-like model of ezrin activation that ascribes the major role in this process to the alpha-helical domain (Li *et al.*, 2007). The rheostat-like manner of activation allows intermediate protein states between its inactive and active form. This multilevel manner of conformational regulation granted biochemical flexibility to ERM and merlin proteins. Consequently, they evolve more regulatory and binding sites and, therefore, acquired more functions in the cell, and eventually, became involved in signaling pathways. ERM's intricate activity regulation mechanism became beneficial in vertebrate animals with increasing complexity of their cellular physiology and number of cell types. As a result, three ERM and two merlin paralogs diverged, acquiring some tissue specific functions in a process that can be described by the birth-and-death model (Nei and Rooney, 2005). In addition, spatial expression of paralogs is often shown to be different from the ancestral gene, that can be a sign of sub-functionalization (Glasauer and Neuhauss, 2014). This can explain the inconsistency of some data about ezrin, moesin, radixin, and merlin roles in cancer, since the experimental results can be influenced by the cell/tissue type, momentary availability of interacting partners or ERMs' binding sites exposure.

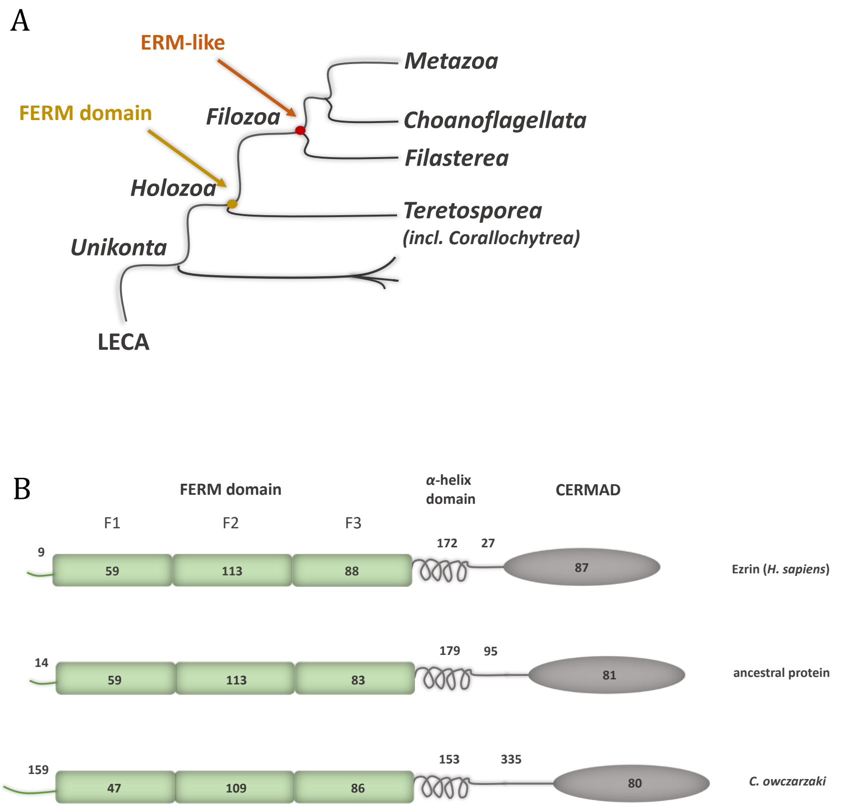

Figure 6. A: Schematic illustration of the early ERM+merlin phylogenetic history. The arrows show first appearance of the protein structures. B: Domain structure comparison. The predicted ERM+merlin ancestral protein and the protein from *C. owczarzaki* demonstrate longer insertions between the alpha-helix and CERMAD. Numbers indicate number of amino acids.

Indeed, the existing RNAseq data of ezrin, radixin, and moesin demonstrate different expression patterns in different human tissues (Fig. 7), although, it is worth to notice that the available data for these proteins is often controversial. The present time RNAseq techniques cannot always achieve good resolution for paralogs and splice isoforms. Therefore, it is important to investigate the ERMs and merlin expression levels in a more precise manner, for example, with the use of long read sequencing, and including experiments in different developmental stages. The study of splice isoforms tissue distribution is another interesting topic in ERM+merlin research: for example, in humans there are two ezrin splice variants, six – for radixin, five – for moesin, and eleven - for merlin. Nothing yet is known about functions of different isoforms.

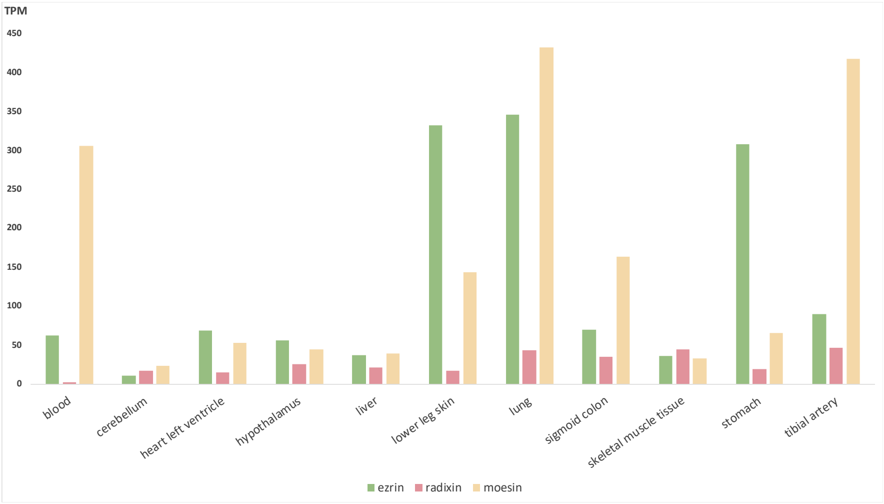

Figure 7. Expression levels for human ezrin, radixin, and moesin in different tissues. TPM – transcripts per kilobase million. Based on the data from Genotype tissue expression portal (GTExPortal, 2019).

Biochemical and physiological studies of the paralog specific amino acid variations are important for understanding each protein's role in healthy and cancer cells. Based on such variations' conservation within vertebrate lineages, we highlighted several motifs that can be candidates for such studies. For example, there is not much known about the role of the polyproline stretch in ezrin and radixin. Moesin lacks this stretch although has a structural analog (Li *et al.*, 2007), and merlin possesses multiple discontinues prolines in the homologous site. The polyproline stretch can possibly bind SH3 domain as another way of the proteins' activity regulation (Li *et al.*, 2007) that would not be possible in moesin and merlin. Another motif of interest is located at the beginning of the alpha-helix and is specific to each of the proteins: EREKEQ in ezrin, EKEKEK in radixin, EKEKER in moesin, and ERTR/EKEK/EREK in merlin. This motif was earlier shown to be important for supporting the coiled-coil folding (Phang *et al.*, 2016). Next to it there is the REKEEL motif that is specific to ERMs and is absent from merlins. Another candidate is the amphipathic stretch of 14 amino acids within the alpha-helix region that is known to be essential for binding regulatory RII subunit of protein kinase A (Dransfield *et al.*, 1997). This region is highly conserved in ezrin and radixin but in moesin the conservation level is only 70%. The six amino acids motif in the N-terminal end of CERMAD is $\text{H}_{/\text{Q}}$DENxA in radixin and moesin and $\text{xxExS}_{/\text{x}}$x in ezrin (where x is any amino acid, S/x meaning that S is conserved in half of the cases).

At the same time, ezrin, radixin, and moesin retained the set of overlapping, redundant functions, most essential for cell survival, such as organizing molecular complexes in the regions of cell-cell contacts. Ezrin is considered to be the major, indispensable paralog. Indeed, its knock out in mice causes early death of the animals. However, genomes of Tasmanian devil and platypus lack ezrin gene, at least based on the sequencing material that is available for these species at the moment. Some birds and mammals (elephant shrew) lost either radixin or moesin genes. This suggest the genetic plasticity of ERM family between different vertebrate lineages that is likely due to the conformational plasticity of the proteins.

In this work, we for the first time, to our knowledge, stress the existence of the two merlin paralogs in the vertebrate genomes: merlin1 and merlin2. Merlin1 was, apparently, lost in the two lineages, *Eutheria* and *Metatheria* (except for koala that has both genes), while merlin2 is present in all vertebrates. Merlin1 contains an additional amino acid stretch within its CERMAD that is absent from merlin2. Also, merlin1 has a weak sequence similarity to merlin2 in the last 30 amino acids, that is responsible for anti-tumor activity in human merlin2. Strikingly, merlin1 lacks one of the two N-terminal actin binding sites (in the F1 subdomain). These observations together with the fact that actin binding is important for merlin anti-proliferative activity (Cooper & Giancotti 2014), suggests that merlin1 is unlikely to exhibit tumor suppressive effect. It, therefore, should perform a specific, unknown function. At the same time merlin2 seems to be evolved specifically to counteract cellular dysregulations leading to cancer. It is unclear, though, why merlin1 was lost in the major mammalian lineages and what protein took over its function. One could also speculate that this paralog was lost together with the organ/tissue-specific function that are present in *Amphibia, Sauria,* and fishes, but not in *Mammalia*. As it was shown for the aryl hydrocarbon receptor paralogs in vertebrates, they diversified and specialized along with the development of complex xenobiotics metabolizing and adaptive immunity (Hahn, Karchner and Merson, 2017). Similarly, merlin2 paralog could acquire additional to scaffolding function to contribute to evolving of the intricate anti-cancer protective system in vertebrates.

Conclusion

Emergence of new proteins and new protein functions is an important question in evolutionary biology, and fate of paralogous forms is probably one of the least understood aspects of the process. Based on sequence comparison and phylogenetic reconstruction we hypothesized the way the ERM+merlin protein family could have gone from the first appearance of the FERM domain in holozoans to the functionally multifaceted group of five homologs with tissue specificity. We propose that the three ERM paralogs retained in the vertebrates due to their conformational plasticity that appeared to be beneficial in the conditions of the vertebrate evolution: increased complexity of organisms' physiology and biochemistry of the cells. Merlin1 paralog is for the first time discussed here, and suggested to perform a yet unknown function specific to non-mammalian vertebrates. Merlin2 is the most interesting example of this protein family evolution, as it seems to be specifically adapted in vertebrates to anti-cancer protection.

References:

Adl, S. M. *et al.* (2012) 'The Revised Classification of Eukaryotes Eukaryotic Microbiology', *J. Eukaryot. Microbiol*, 59(5), pp. 429–493. doi: 10.1111/j.1550-7408.2012.00644.x.

Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.', *Nucleic acids research*, 25(17), pp. 3389–402. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9254694 (Accessed: 23 July 2017).

Bailey, T. L. *et al.* (2009) 'MEME SUITE: tools for motif discovery and searching.', *Nucleic acids research*. Oxford University Press, 37(Web Server issue), pp. W202-8. doi: 10.1093/nar/gkp335.

Bosanquet, D. C. *et al.* (2014) 'FERM family proteins and their importance in cellular movements and wound healing (Review)', *International Journal of Molecular Medicine*, 34(1), pp. 3–12. doi: 10.3892/ijmm.2014.1775.

Brault, E. *et al.* (2001) 'Normal membrane localization and actin association of the NF2 tumor suppressor protein are dependent on folding of its N-terminal domain.', *Journal of cell science*, 114(Pt 10), pp. 1901–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11329377 (Accessed: 22 April 2019).

Bretscher, A., Edwards, K. and Fehon, R. G. (2002) 'ERM proteins and merlin: integrators at the cell cortex.', *Nature reviews. Molecular cell biology*, 3(8), pp. 586–99. doi: 10.1038/nrm882.

Chen, S.-D. *et al.* (2012) 'Knockdown of radixin by RNA interference suppresses the growth of human pancreatic cancer cells in vitro and in vivo.', *Asian Pacific journal of cancer prevention : APJCP*, 13(3), pp. 753–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22631643 (Accessed: 26 January 2019).

Clucas, J. and Valderrama, F. (2014) 'ERM proteins in cancer progression', *Journal of Cell Science*, 127(2), pp. 267–275. doi: 10.1242/jcs.133108.

Cooper, J. and Giancotti, F. G. (2014) 'Molecular insights into NF2/Merlin tumor suppressor function.', *FEBS letters*. NIH Public Access, 588(16), pp. 2743–52. doi: 10.1016/j.febslet.2014.04.001.

Curto, M. *et al.* (2007) 'Contact-dependent inhibition of EGFR signaling by Nf2/Merlin', *The Journal of Cell Biology*, 177(5), pp. 893–903. doi: 10.1083/jcb.200703010.

Darriba, D. *et al.* (2011) 'ProtTest 3: fast selection of best-fit models of protein evolution', *Bioinformatics*, 27(8), pp. 1164–1165. doi: 10.1093/bioinformatics/btr088.

Dayhoff, M. O. (1965) *No Title*. Silver Spring: National Biomedical research Foundation.

Debnath, J. and Brugge, J. S. (2005) 'Modelling glandular epithelial cancers in three-dimensional cultures', *Nature Reviews Cancer*, 5(9), pp. 675–688. doi: 10.1038/nrc1695.

Dransfield, D. T. *et al.* (1997) 'Ezrin is a cyclic AMP-dependent protein kinase anchoring protein.', *The EMBO journal*. European Molecular Biology Organization, 16(1), pp. 35–43. doi: 10.1093/emboj/16.1.35.

Eisen, J. A. (1998) 'Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.', *Genome research*, 8(3), pp. 163–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9521918 (Accessed: 30 April 2019).

Estecha, A. *et al.* (2009) 'Moesin orchestrates cortical polarity of melanoma tumour cells to initiate 3D invasion.', *Journal of cell science*, 122(Pt 19), pp. 3492–501. doi: 10.1242/jcs.053157.

Fairclough, S. R. *et al.* (2013) 'Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta', *Genome Biology*. BioMed Central, 14(2), p. R15. doi: 10.1186/gb-2013-14-2-r15.

Federhen, S. (2012) 'The NCBI Taxonomy database.', *Nucleic acids research*. Oxford University Press, 40(Database issue), pp. D136-43. doi: 10.1093/nar/gkr1178.

Fievet, B. T. *et al.* (2004) 'Phosphoinositide binding and phosphorylation act sequentially in the activation mechanism of ezrin.', *The Journal of cell biology*, 164(5), pp. 653–9. doi: 10.1083/jcb.200307032.

Finn, R. D. *et al.* (2006) 'Pfam: clans, web tools and services', *Nucleic Acids Research*. Oxford University Press, 34(90001), pp. D247–D251. doi: 10.1093/nar/gkj149.

Funayama, N. *et al.* (1991) 'Radixin is a novel member of the band 4.1 family.', *The Journal of cell biology*, 115(4), pp. 1039–48. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2289953&tool=pmcentrez&rendertype=abstract (Accessed: 15 August 2015).

Gabriel, W. N. *et al.* (2007) 'The tardigrade Hypsibius dujardini, a new model for studying the evolution of development', *Developmental Biology*. Academic Press, 312(2), pp. 545–559. doi: 10.1016/J.YDBIO.2007.09.055.

Gasteiger, E. *et al.* (2005) 'Protein Identification and Analysis Tools on the ExPASy Server', in *The Proteomics Protocols Handbook*. Totowa, NJ: Humana Press, pp. 571–607. doi: 10.1385/1-59259-890-0:571.

Geneious (2019) *Geneious Prime | Sequence Analysis Software | Academic &amp; Government*, *2019*. Available at: https://www.geneious.com/academic/ (Accessed: 22 April 2019).

Glasauer, S. M. K. and Neuhauss, S. C. F. (2014) 'Whole-genome duplication in teleost fishes and its evolutionary consequences', *Molecular Genetics and Genomics*, 289(6), pp. 1045–1060. doi: 10.1007/s00438-014-0889-2.

Golovnina, K. *et al.* (2005) 'Evolution and origin of merlin, the product of the Neurofibromatosis type 2 (NF2) tumor-suppressor gene.', *BMC Evolutionary Biology*, 5(1), p. 69. doi: 10.1186/1471-2148-5-69.

Grau-Bové, X. *et al.* (2017) 'Dynamics of genomic innovation in the unicellular ancestry of animals', *eLife*, 6. doi: 10.7554/eLife.26036.

GTExPortal (2019) *GTEx Portal*. Available at: https://gtexportal.org/home/ (Accessed: 4 December 2018).

Hahn, M. E., Karchner, S. I. and Merson, R. R. (2017) 'Diversity as Opportunity: Insights from 600 Million Years of AHR Evolution.', *Current opinion in toxicology*. NIH Public Access, 2, pp. 58–71. doi: 10.1016/j.cotox.2017.02.003.

Ivetic, A. and Ridley, A. J. (2004) 'Ezrin/radixin/moesin proteins and Rho GTPase signalling in leucocytes.', *Immunology*, 112(2), pp. 165–76. doi: 10.1111/j.1365-2567.2004.01882.x.

Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30(9), pp. 1236–1240. doi: 10.1093/bioinformatics/btu031.

Katoh, K. *et al.* (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*. Oxford University Press, 30(14), pp. 3059–3066. doi: 10.1093/nar/gkf436.

Kikuchi, S. *et al.* (2002) 'Radixin deficiency causes conjugated hyperbilirubinemia with loss of Mrp2 from bile canalicular membranes.', *Nature genetics*, 31(3), pp. 320–5. doi: 10.1038/ng905.

Kitajiri, S. *et al.* (2004) 'Radixin deficiency causes deafness associated with progressive degeneration of cochlear stereocilia.', *The Journal of cell biology*, 166(4), pp. 559–70. doi: 10.1083/jcb.200402007.

Kobayashi, H. *et al.* (2004) 'Clinical significance of cellular distribution of moesin in patients with oral squamous cell carcinoma.', *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(2), pp. 572–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14760079 (Accessed: 26 January 2019).

Kumar, S., Stecher, G. and Tamura, K. (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets', *Molecular Biology and Evolution*, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Lallemand, D., Saint-Amaux, A. L. and Giovannini, M. (2009) 'Tumor-suppression functions of merlin are independent of its role as an organizer of the actin cytoskeleton in Schwann cells', *Journal of Cell Science*, 122(22), pp. 4141–4149. doi: 10.1242/jcs.045914.

Lamiable, A. *et al.* (2016) 'PEP-FOLD3: faster *de novo* structure prediction for linear peptides in solution and in complex', *Nucleic Acids Research*, 44(W1), pp. W449–W454. doi: 10.1093/nar/gkw329.

Lankes, W. T. and Furthmayr, H. (1991) 'Moesin: a member of the protein 4.1-talin-ezrin family of proteins.', *Proceedings of the National Academy of Sciences of the United States of America*, 88(19), pp. 8297–301. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=52495&tool=pmcentrez&rendertype=abstract (Accessed: 15 August 2015).

Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Le, S. Q. and Gascuel, O. (2008) 'An Improved General Amino Acid Replacement Matrix', *Molecular Biology and Evolution*. Oxford University Press, 25(7), pp. 1307–1320. doi: 10.1093/molbev/msn067.

Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation'. Available at: http://itol.embl.de/help/17050570.pdf (Accessed: 11 May 2017).

Li, Q. *et al.* (2007) 'Self-masking in an Intact ERM-merlin Protein: An Active Role for the Central α-Helical Domain', *Journal of Molecular Biology*, 365(5), pp. 1446–1459. doi: 10.1016/j.jmb.2006.10.075.

Li, Q. *et al.* (2015) 'Ezrin/Exocyst Complex Regulates Mucin 5AC Secretion Induced by Neutrophil Elastase in Human Airway Epithelial Cells', *Cellular Physiology and Biochemistry*, 35(1), pp. 326–338. doi: 10.1159/000369699.

Li, W. and Crouch, D. H. (2000) 'Cloning and expression profile of chicken radixin.', *Biochimica et biophysica acta*, 1491(1–3), pp. 327–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10760599 (Accessed: 16 December 2018).

Liu, X. *et al.* (2015) 'Moesin and myosin phosphatase confine neutrophil orientation in a chemotactic gradient.', *The Journal of experimental medicine*. Rockefeller University Press, 212(2), pp. 267–80. doi: 10.1084/jem.20140508.

Lu, T.-M. *et al.* (2017) 'The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution', *Zoological Letters*, 3(1), p. 6. doi: 10.1186/s40851-017-0068-5.

Marchler-Bauer, A. *et al.* (2015) 'CDD: NCBI's conserved domain database', *Nucleic Acids Research*, 43(D1), pp. D222–D226. doi: 10.1093/nar/gku1221.

Marion, S. *et al.* (2011) 'Ezrin promotes actin assembly at the phagosome membrane and regulates phago-lysosomal fusion.', *Traffic (Copenhagen, Denmark)*, 12(4), pp. 421–37. doi: 10.1111/j.1600-0854.2011.01158.x.

McClatchey, a. I. (2014) 'ERM proteins at a glance', *Journal of Cell Science*, (June), pp. 1–6. doi: 10.1242/jcs.098343.

Nambiar, R., McConnell, R. E. and Tyska, M. J. (2010) 'Myosin motor function: the ins and outs of actin-based membrane protrusions', *Cellular and Molecular Life Sciences*, 67(8), pp. 1239–1254. doi: 10.1007/s00018-009-0254-5.

Nei, M. and Rooney, A. P. (2005) 'Concerted and birth-and-death evolution of multigene families.', *Annual review of genetics*. NIH Public Access, 39, pp. 121–52. doi: 10.1146/annurev.genet.39.073003.112240.

Niggli, V. and Rossy, J. (2008) 'Ezrin/radixin/moesin: Versatile controllers of signaling molecules and of the cortical cytoskeleton', *International Journal of Biochemistry and Cell Biology*, 40(3), pp. 344–349. doi: 10.1016/j.biocel.2007.02.012.

Omelyanchuk, L. V *et al.* (2009) 'Evolution and origin of HRS, a protein interacting with Merlin, the Neurofibromatosis 2 gene product.', *Gene regulation and systems biology*, 3, pp. 143–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20054405 (Accessed: 20 November 2018).

Pataky, F., Pironkova, R. and Hudspeth, A. J. (2004) 'Radixin is a constituent of stereocilia in hair cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 101(8), pp. 2601–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14983055 (Accessed: 16 December 2018).

Van de Peer, Y., Maere, S. and Meyer, A. (2009) 'The evolutionary significance of ancient genome duplications.', *Nature reviews. Genetics*, 10(10), pp. 725–32. doi: 10.1038/nrg2600.

Phang, J. M. *et al.* (2016) 'Structural characterization suggests models for monomeric and dimeric forms of full-length ezrin.', *The Biochemical journal*. Portland Press Limited, 473(18), pp. 2763–82. doi: 10.1042/BCJ20160541.

Pujuguet, P. *et al.* (2003) 'Ezrin Regulates E-Cadherin-dependent Adherens Junction Assembly through Rac1 Activation', *Molecular Biology of the Cell*, 14(5), pp. 2181–2191. doi: 10.1091/mbc.e02-07-0410.

Rambaut A. (2019) *FigTree, 2009*. Available at: http://tree.bio.ed.ac.uk/software/figtree/ (Accessed: 18 April 2019).

Richter, D. J. and King, N. (2013) 'The Genomic and Cellular Foundations of Animal Origins', *Annual Review of Genetics*. Annual Reviews , 47(1), pp. 509–537. doi: 10.1146/annurev-genet-111212-133456.

Ronquist, F. *et al.* (2012) 'MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space', *Systematic Biology*, 61(3), pp. 539–542. doi: 10.1093/sysbio/sys029.

Rost, B. and Sander, C. (1994) 'Combining evolutionary information and neural networks to predict protein secondary structure', *Proteins: Structure, Function, and Genetics*. John Wiley & Sons, Ltd, 19(1), pp. 55–72. doi: 10.1002/prot.340190108.

Roy, C., Martin, M. and Mangeat, P. (1997) 'A dual involvement of the amino-terminal domain of ezrin in F- and G-actin binding.', *The Journal of biological chemistry*, 272(32), pp. 20088–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9242682 (Accessed: 22 September 2015).

Saotome, I., Curto, M. and McClatchey, A. I. (2004) 'Ezrin is essential for epithelial organization and villus morphogenesis in the developing intestine.', *Developmental cell*, 6(6), pp. 855–64. doi: 10.1016/j.devcel.2004.05.007.

Schliep, K. *et al.* (2017) 'Intertwining phylogenetic trees and networks', *Methods in Ecology and Evolution*. Edited by R. Fitzjohn. doi: 10.1111/2041-210X.12760.

Sebé-Pedrós, A. *et al.* (2013) 'Insights into the Origin of Metazoan Filopodia and Microvilli', *Molecular Biology and Evolution*. Oxford University Press, 30(9), pp. 2013–2023. doi: 10.1093/molbev/mst110.

Shalchian-Tabrizi, K. *et al.* (2008) 'Multigene Phylogeny of Choanozoa and the Origin of Animals', *PLoS ONE*. Edited by R. Aramayo, 3(5), p. e2098. doi: 10.1371/journal.pone.0002098.

Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.', *Bioinformatics (Oxford, England)*, 30(9), pp. 1312–3. doi: 10.1093/bioinformatics/btu033.

Stamatakis, A. (2015) 'Using RAxML to Infer Phylogenies', in *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 6.14.1-6.14.14. doi: 10.1002/0471250953.bi0614s51.

Stickney, J. T. *et al.* (2004) 'Activation of the tumor suppressor merlin modulates its interaction with lipid rafts.', *Cancer research*. American Association for Cancer Research, 64(8), pp. 2717–24. doi: 10.1158/0008-5472.CAN-03-3798.

Suga, H. *et al.* (2013) 'The Capsaspora genome reveals a complex unicellular prehistory of animals', *Nature Communications*. Nature Publishing Group, 4(1), p. 2325. doi: 10.1038/ncomms3325.

Turunen, O. *et al.* (1998) 'Structure-function relationships in the ezrin family and the effect of tumor-associated point mutations in neurofibromatosis 2 protein.', *Biochimica et biophysica acta*, 1387(1–2), pp. 1–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9748471 (Accessed: 8 October 2017).

Valderrama, F., Thevapala, S. and Ridley, A. J. (2012) 'Radixin regulates cell migration and cell-cell adhesion through Rac1', *Journal of Cell Science*, 125(14), pp. 3310–3319. doi: 10.1242/jcs.094383.

Winckler, B. *et al.* (1994) 'Analysis of a cortical cytoskeletal structure: a role for ezrin-radixin-moesin (ERM proteins) in the marginal band of chicken erythrocytes.', *Journal of cell science*, 107 (Pt 9), pp. 2523–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7531201 (Accessed: 5 May 2019).

Yonemura, S. *et al.* (2002) 'Rho-dependent and -independent activation mechanisms of ezrin/radixin/moesin proteins: an essential role for polyphosphoinositides in vivo.', *Journal of cell science*, 115(Pt 12), pp. 2569–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12045227 (Accessed: 13 July 2015).

Yoshida, Y. *et al.* (2017) 'Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus', *PLOS Biology*. Edited by C. Tyler-Smith. Public Library of Science, 15(7), p. e2002266. doi: 10.1371/journal.pbio.2002266.