Emergence and Evolution of ERM Proteins and Merlin in Metazoans

Victoria Shabardina[1],*, Yukie Kashima[2], Yutaka Suzuki[3], and Wojciech Makalowski[1]

[1]Instiue of Bioinformatics, University of Muenster, Germany

[2]Department of Computational Biology and Medical Sciences, The University of Tokyo, Kashiwa, Japan

[3]Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, The University of Tokyo, Kashiwa, Japan

*Corresponding author: E-mail: shabardi@uni-muenster.de.

Abstract

Ezrin, radixin, moesin, and merlin are cytoskeletal proteins, whose functions are specific to metazoans. They participate in cell cortex rearrangement, including cell–cell contact formation, and play an important role in cancer progression. Here, we have performed a comprehensive phylogenetic analysis of the proteins spanning 87 species. The results describe a possible mechanism for the protein family origin in the root of Metazoa, paralogs diversification in vertebrates, and acquisition of novel functions, including tumor suppression. In addition, a merlin paralog, present in most vertebrates but lost in mammals, has been described here for the first time. We have also highlighted a set of amino acid variations within the conserved motifs as the candidates for determining physiological differences between ERM paralogs.

Key words: protein evolution, paralogs fate, ERM phylogeny.

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 et al. 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 (Ivetic and Ridley 2004; Marion et al. 2011; Bosanquet et al. 2014; McClatchey 2014). 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 a conformational switch from the inactive, dormant folding to the active, stretched form. The inactive state is established through the autoinhibitory interaction between the N-terminal FERM and the C-terminal CERMAD domains. This results in the masking of the binding sites for membrane proteins in the FERM domain and actin-binding site (ABS) in the CERMAD (Turunen et al. 1998). Upon activation, ERMs are consequently exposed to their two activating factors: PIP$_2$ (phosphatidylinositol 4,5-bisphosphate) that binds to the FERM domain and

phosphorylation of a conserved threonine in the CERMAD (Yonemura et al. 2002; Niggli and Rossy 2008). An important role during this transition belongs to the middle α-helical domain. In the dormant ERMs, it forms a coiled-coil structure, bringing N- and C-terminal domains together (Hoeflich et al. 2003; Li et al. 2007).

Little is known about individual roles of each of the ERM proteins in both health and disease. The three proteins are paralogs and share high amino acid sequence similarity (~75% in humans) (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, and only ezrin's depletion appeared to be lethal (Kikuchi et al. 2002; Kitajiri et al. 2004; Saotome et al. 2004; Liu et al. 2015). Special interest in ERMs' role in cancer stimulated multiple studies. Their dysregulation can lead to a disruption of cell–cell contacts, enhanced cell migration and invasion, and higher cancer cell survival (Clucas and Valderrama 2014). Importantly, some studies have shown 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 et al. 2012).

One well-known tumor suppressor factor is the ERM-like protein merlin. In humans, it shares 46% amino acid sequence similarity with the whole-length ezrin and 86% similarity when comparing only FERM domains (Turunen et al. 1998). Mutations in merlin result in the 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 the stretch of the last 40 residues in the C-terminus (Lallemand et al. 2009; Cooper and Giancotti 2014). Two ABSs in merlin are located in the FERM domain, whereas the C-terminal ABS, typical to ERMs, is absent (Roy et al. 1997; Brault et al. 2001).

With more research being done, it is getting more clear that ezrin, radixin, and moesin 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 a challenging task to distinguish their functions in vivo. A phylogenetic approach can be an effective tool in resolving this problem, as it enables precise paralogs characterization by tracing the evolutionary history of the binding sites and conserved amino acid motifs. So far, only few phylogenies of ERMs and merlin have been described in literature. As a rule, they feature limited taxonomy representation or are included in the studies as an accessory and brief part of the discussions (Turunen et al. 1998; Golovnina et al. 2005; Phang et al. 2016; Michie et al. 2019). Thus, these phylogenies do not provide a full understanding of ERMs and merlin evolution, although they depict some interesting patterns. These studies agree on the fact that the proteins are highly conserved within the metazoan clade, 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 et al. 2002; Omelyanchuk et al. 2009; Nambiar et al. 2010; Sebé-Pedrós et al. 2013). This view is supported by the recent discovery of ERM-like proteins in Choanoflagellata and Filasterea, the closest unicellular relatives of metazoans (Fairclough et al. 2013; Suga et al. 2013).

The positioning of merlin relative to the ERM family differs in literature, and some studies exclude merlin from the discussion of the ERM family altogether. Nevertheless, because these proteins share evolutionary history and structural characteristics, it is reasonable to unite them in one group. In this work, we conducted the first comprehensive phylogenetic analysis of the ERM family and merlin that includes data from all metazoan orders sequenced by the time. The results describe ERM and merlin sequence conservation and paralog number diversity within the clade of Metazoa. We suggest that increased organism complexity led to a diversification of the protein paralogs in vertebrates. Moreover, we highlight the importance of phylogenetic studies of paralogs in general, and 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, and merlin NP_000259.1) against the nonredundant (nr) protein sequences collection in the NCBI database. The selected sequences were manually monitored to exclude database duplicates, splice variants, and 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. 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.

Taxa selection was done 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) the whole genome or transcriptome of a representative species was sequenced and available, and 3) there was a high quality of the genome assembly annotation. In the case that no representative genome was available for an order, a TBlastN search was run against all available nucleotide sequences for that taxa. A search for possible homologs of ezrin, radixin, moesin, and merlin was performed among other Opisthokonta (Holozoa, Nuclearia, and Fungi) and among Amoebozoa, Excavata, Archaeplastida (includes green plants), SAR (Stramenopiles, Alveolata, and Rhizaria) cluster, and other protist groups (refer to the Tree of Life scheme [Adl et al. 2012]). Prokaryota and Archaea nucleotide sequences were scanned for whole-length proteins or for only the FERM domain using a TBlastN search against the nr nucleotide collection at NCBI. The taxonomic structure describing the final data set can be viewed in the supplementary table S1, Supplementary Material online, and is based on the topologies employed by the NCBI Taxonomy database (Federhen 2012) and the Tree of Life project (Letunic and Bork). The taxa variety will be further discussed as "vertebrate" and "invertebrate," the latter including the remaining Eumetazoa. Sequences were retrieved by July 2018; data for vertebrates were updated in May 2019.

Reconstruction of the ERM + Merlin Phylogeny

Multiple sequence alignment was generated using MAFFT software (Katoh et al. 2002) with the PAM70 substitution matrix (Dayhoff 1965), as defined by ProtTest3 (Darriba et al. 2011),

and manually edited to remove uninformative columns. CLUSTALX (Larkin et al. 2007) and Geneious (Kearse 2012) were used for alignment visualization. Maximum likelihood (ML) phylogenetic trees were build using RAxML tool (Stamatakis 2014) with the parameters estimated by running a RAxML parameter test (Stamatakis 2015). The PROTGAMMALG model was chosen, in which the GAMMA model estimates substitution rate between sites and LG is the amino acid substitution matrix (Le and Gascuel 2008). The statistical support for tree clustering was calculated by running 1,000 bootstrap replicates. The resulting trees were inspected and edited using the iTOL online tree viewer (Letunic and Bork 2007) and FigTree software (Rambaut 2019).

A reduced data set (supplementary table S2, Supplementary Material online) 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 by RAxML with the PROTGAMMALG model. The rooting option and defining marginal ancestral states option were used.

MEGA software (Kumar et al. 2016) was used for an alternative tree reconstruction for Neighbor-Joining and parsimony algorithms with 500 bootstrap replicates and an LG amino acid substitution matrix. A Bayesian inference method was also applied with the use of the MrBayes tool (Ronquist et al. 2012) under the LG matrix. Six chains were run for 3,000,000 generations, every 1,000 generation trees were sampled in two runs. The first 25% of trees were discarded before constructing a consensus tree.

Protein Sequence Analysis

The tertiary structure of polypeptides was predicted by PEP-FOLD3 (Lamiable et al. 2016). Estimation of the 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 within the selected set of protein sequences (supplementary table S2, Supplementary Material online) throughout major metazoan lineages using MEME suite (Bailey et al. 2009).

Testing for Positive Selection in Vertebrate Ezrin and Merlin

First, the codeml tool from PAML software (Yang 2007) was run under the branch free-ratio model to test whether the ratio of nonsynonymous to synonymous substitutions varied among tree branches. For statistical assessment, twice the difference between the log likelihood of the alternative and the null (the ratio is the same for all branches) hypotheses were compared with the χ^2 distribution. The branch free-ratio model is useful for overall estimation but not very informative due to its low statistical power, as it uses a big number of parameters for calculations. Therefore, we also performed a branch-site model analysis. In particular, we tested all the

lineages for positive selection under the parameters: model = 2, NSsites = 2, fix_omega = 0, and omega = 1. The null hypothesis in this case was estimated under parameters: model = 2, NSsites = 2, fix_omega = 1, and omega = 1. The χ^2 distribution test for statistical support was used. Each estimation of alternative hypotheses was run twice by codeml with varying parameter "omega" to test for contingency. Parameter "cleandata" was set to 0 allowing for retaining gaps in the alignment.

The PAL2NAL tool (Suyama et al. 2006) was used to generate nucleotide alignments based on amino acid multiple sequence alignments and the corresponding mRNA sequences. The mRNA sequences were extracted from the NCBI data base by using the efetch command from the E-utilities suit (Sayers 2009). The phylogenetic trees for ezrin and merlin protein sets from vertebrate animals were built as described earlier and can be found in supplementary file S1, Supplementary Material online.

Synteny Analysis

A BlastP strategy was applied to analyze syntenic relationships between the genes coding for the ERM and ERM-like proteins in different species. Sequences of the human proteins encoded by the genes surrounding ezrin (seven proteins), radixin (three proteins), and moesin (four proteins) within the region of 1 Mb were used as query in BlastP searches against all proteins in the following species: *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Stylophora pistillata* (cnidarians), *Trichoplax adhaerens*, and *Amphimedon queenslandica*. The expect threshold was set to 1e-10. The analysis was repeated with similarly selected proteins from *Dr. melanogaster* as a query. Intraspecies synteny analysis for *Homo sapiens* was done likewise but with only human proteins as the target for the BlastP search.

For synteny analysis of the merlin gene surroundings, two species with both merlin paralogs were chosen: *D. rerio* and *Gallus gallus*. All proteins, whose genes are located within 1 Mb of merlin1 or merlin2 genes, were taken as queries. BlastP searches with expect threshold 1e-10 were run against all proteins from the corresponding chromosomes: chromosomes 5 (has gene coding for merlin2 protein) and 21 (merlin1) for *D. rerio*; chromosomes 15 (merlin2) and 19 (merlin1) for *G. gallus*. Synteny Database (Catchen et al. 2009) was used for processing the data for *H. sapiens* and *D. rerio*.

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 ERM and merlin homologs throughout all eukaryotic clades resulted in a selection of 285 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*). In addition, a sequence of 298 amino acids (supplementary file S1, Supplementary Material online) from corallochytrean *Corallochytrium limacisporum* revealed a 25% sequence identity with the sequence of the FERM domain of human ezrin, based on a TBlastN search against the species' whole genome sequence. Domain annotation by PFAM indicated that it belongs to a class of FERM domain with the high statistical support (e-value $< 10^{-5}$ for each of the three subdomains of the FERM domain). Furthermore, we predicted biochemical properties of the corallochytrean polypeptide and compared it to the human FERM domain. The analysis revealed that the two FERM domains, human and corallochytrean, exhibit distinct features. For example, different amino acid content, different instability index, hydropathicity, and p*I* (8.75 for the human ezrin FERM domain and 6.79 for the *Co. limacisporum* polypeptide). In particular, the human FERM domain is predicted to be more hydrophilic than its potential corallochytrean homolog: grand average of hydropathicity (GRAVY) index is -0.530 and -0.270, respectively. The human FERM is also less stable: its instability index is 43.57 and higher than the index for the *Co. limacisporum*'s FERM-like protein, that is, 31.80. 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 protein IDs used for the analysis can be found in the supplementary table S2, Supplementary Material online.

ERM + Merlin Protein Family Is Conserved throughout All the Metazoan Orders and Three Unicellular Species

Sequence comparison of the proteins selected for analysis demonstrated that the domain structure and most of the known binding sites are conserved throughout the whole metazoan clade. Even the proteins from such early-branching animals as *T. adhaerens* (Placozoa) and *A. queenslandica* (Porifera) demonstrate high similarity to the mammalian protein pattern of conserved amino acid motifs (fig. 1). The FERM and CERMAD domains are characteristically well preserved, including the proteins from unicellular organisms, whereas the α-helical middle domain is the least conserved ERM domain, in agreement with previous studies (Phang et al. 2016). There is also some length variation of proteins in different lineages. The most variable is the length of the region separating the α-helical domain and the CERMAD. It is short for proteins from vertebrates but is extended in all other taxa, with the longest one to be found in the protein from *C. owczarzaki*. Furthermore, the sequence of this region is poorly preserved between the taxa. Proteins

from species of flat worms, also *Intoshia linei* and *C. owczarzaki* are the most divergent.

The binding sites for the most known binding partners of ERMs, such as ICAM-2 (cell adhesion and immunity [Helander et al. 1996]), EBP50 and NHERF2 (cofactors of sodium/hydrogen exchanger [McClatchey 2014]), lipid PIP$_2$, actin, and the sites for intramolecular interaction can be identified in most of the proteins (supplementary table S2, Supplementary Material online). This signifies that some of the interactions might have been established in early metazoan history. We have also determined variation within some conserved motives between ezrin, radixin, and moesin that can be a clue to understanding the proteins' specificity in cancer (see Discussion). The most conserved regions in the ERM proteins are PIP$_2$ binding sites in F1 and F3 subdomains of the FERM domain and the C-terminal ABS (specifically, KYKTL motif). The most dissimilar proteins come from the early-branching metazoans *Molgula tectiformis* (Tunicata), *A. queenslandica* (Porifera), along with endo- and exo-parasites (*I. linei*, flat worms, and blood-sucking leeches). Interestingly, the PIP$_2$ binding motif within the F1 subdomain is not preserved only in *M. tectiformis*. Small variations in the KYKTL motif can be seen in proteins from Echinodermata and the shark species *Callorhinchus milii*.

A sequence comparison between different animal lineages distinguished three major groups of merlin proteins: 1) nonvertebrate, 2) vertebrate merlin1 (absent in most of Eutheria), and 3) all-vertebrate merlin2 proteins (here, we arbitrarily assigned merlin1 and merlin2 names to the merlin paralogs). Such grouping is also in accordance with our protein phylogeny (fig. 2). Group 2 comprises merlin1 coded by the paralogous gene which has not been described before (supplementary table S2, Supplementary Material online). It is characterized by a unique insertion of 68 amino acids in tetrapods and 15 amino acids motif PPYxPHSNRNSAYMx in bony fishes in the C-terminal domain. It also lacks a tumor-suppression region characterized in human merlin2 between residues 532 and 579, although, merlin1 does have a merlin-specific blue-box region and a conserved Ser518 residue. Interestingly, within the Eutheria clade, only armadillos (super-order Xenarthra) possess merlin1 gene. Invertebrate merlins (group 1) are similar to vertebrate merlin1 but lack the C-terminal domain insertion. Merlin2 (group3) is present in all vertebrate taxa and has an additional ABS in the F1 of the FERM domain and a tumor-suppression amino acid stretch in its CERMAD domain (residues 532–579 in human). The blue-box region can also be identified in two proteins from unicellular species: XP_004364665.2 in *C. owczarzaki* and XP_004991962.1 in *S. rosetta*.

Phylogenetic Tree of the ERM + Merlin Family

The reconstructed ML tree resulted in a high phylogenetic resolution among vertebrates, whereas the branching for

FERM   α-helical domain   CERMAD
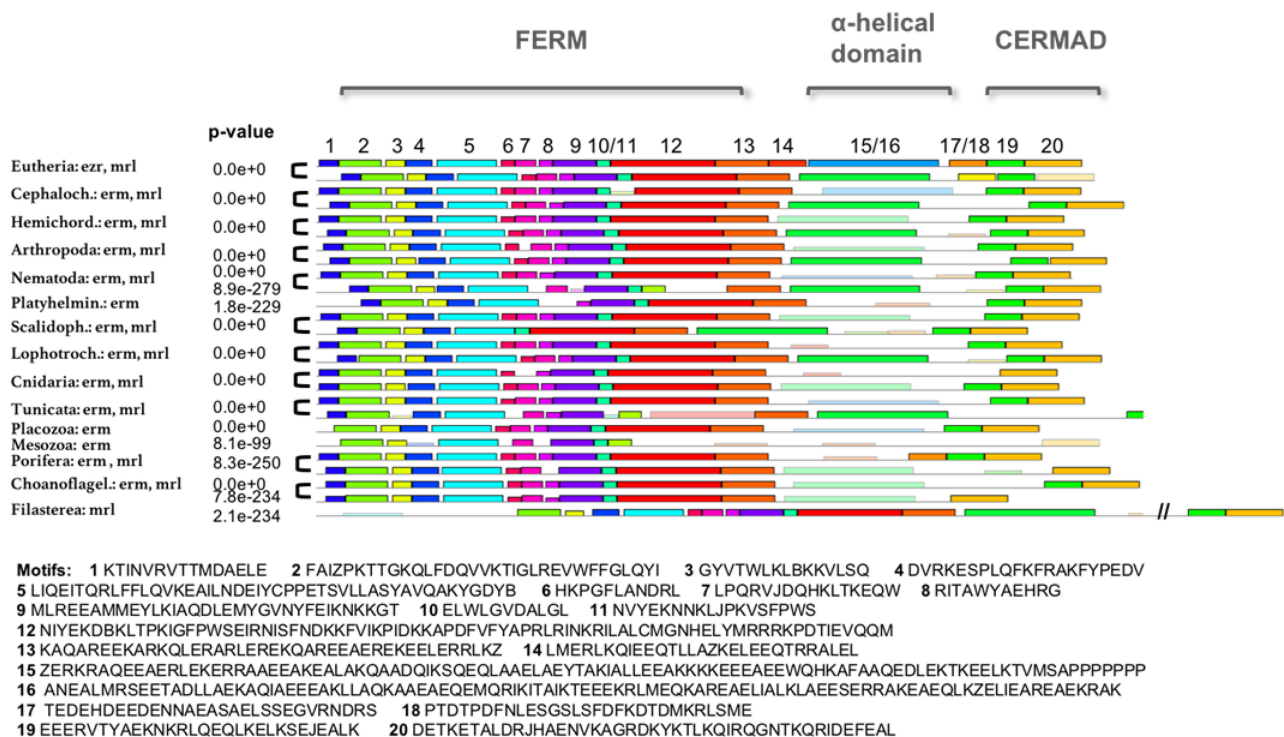
FIG. 1.—MEME conserved amino acid motif analysis. Motifs in pale color were found by the scanning algorithm based on the de novo motif identification (bright color). Note the reduced length of the region separating the α-helical and CERMAD domains in eutherian proteins.

most of the other taxa has a low statistical support (fig. 2). 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 convergence after 3,000,000 generations. The run was terminated and a consensus tree was built anyway. Similarly, to other methods, it featured high bootstrap values for clustering of vertebrate proteins, but unresolved branching for invertebrate sequences. One reason could be an unequal representation of the taxa due to a lack of sequencing data for invertebrates. Another complication could be the high divergence of amino acid sequences between evolutionary distant lineages. The ML tree is used for further discussion.

The most "eccentric" sequence in the reconstructed phylogeny is that of the hypothetical protein from *I. linei*. This protein, although features all three ERM domains, has the highest substitution rate and does not cluster with any other groups. Consequently, it cannot be defined as an ERM-like or as a merlin-like protein. This is not surprising, as *I. linei*, a representative of orthonectids, is a parasitic animal. Its hermaphrodite nature, fast reproductive cycles, and high level of inbreeding can be reasons that make its genome distant from the genomes of other metazoans (Lu et al. 2017).

The protein from *I. linei* was arbitrarily chosen to separate the tree into two major clusters: ERM like and merlin like.

Therefore, any "hypothetical" or "unknown" proteins could be annotated either as ERM like or as merlin like, based on their position relative to *I. linei*'s protein. Besides improving the annotation of such proteins, some false annotations stored in public protein databases were corrected (for details see supplementary table S2, Supplementary Material online).

An interesting observation can be made about the proteins from unicellular species. The *C. owczarzaki*'s protein XP_004364665.2 and *S. rosetta*'s protein XP_004991962.1 cluster together with merlins; and XP_004997754.1 from *S. rosetta* and XP_001746613.1 from *M. brevicollis* cluster with the ERM-like group. However, the bootstrap support is not high enough to confidently assign these annotations. At the same time, the proteins XP_004994097.1 from *S. rosetta* and XP_001743289.1 from *M. brevicollis* cluster together with the invertebrate ERM-like group. This finding provides an insight into the origin of the ERM + merlin family and suggests that merlin and ERMs diverged from the common ancestral protein before the emergence of Metazoa.

Cnidaria, Placozoa, Echinodermata, Scalidofora, Lophotrochozoa, Porifera, Hemichordata, Cephalochordata, and Mesozoa taxa branching could not be defined by this analysis with statistical confidence. ERM-like proteins from the clade Tunicata are the most related to vertebrate ERMs (bootstrap support 83%), but the tunicate's merlins do not form a strictly defined group. Other clearly distinguishable groups, beside vertebrates, are Nematoda (99%),
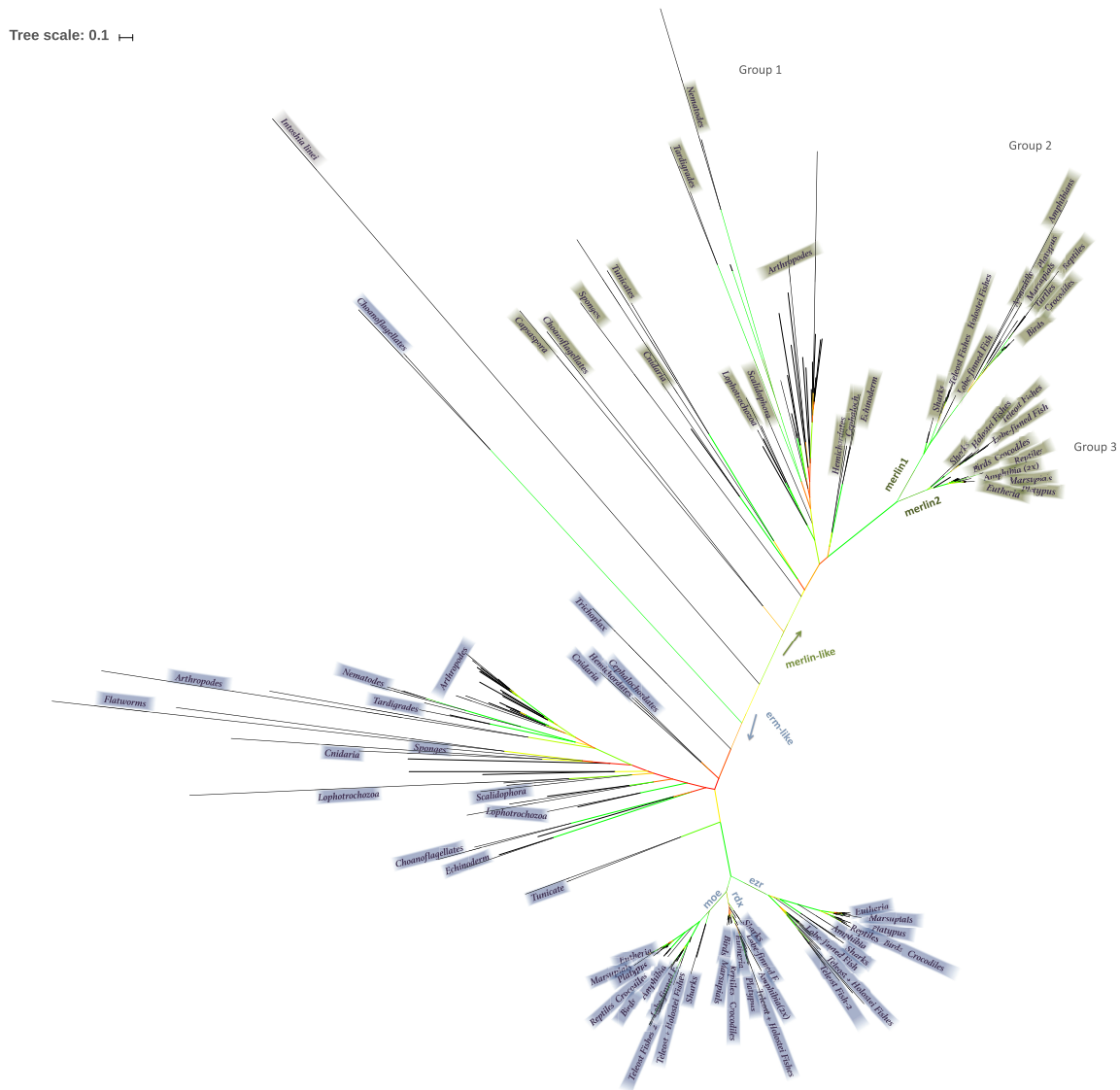
Fig. 2.—Phylogenetic ML tree: ERM + merlin family. Blue color denotes the ERM-like part of the tree and green color is for the merlin-like part. Note the emergence of the ezrin, radixin, and moesin paralogs in vertebrates, and divergence of merlins into nonvertebrate merlins (group 1) and vertebrate merlin1 (group 2) and merlin2 (group 3). Some proteins from the unicellular choanoflagellates species and a protein from *Capsaspora* relate more to merlin-like proteins, some to ERM-like proteins. Color scheme for bootstrap values: green, 70–100%; yellow, 50–70%; red, below 50%. The tree in the Newick format can be found in the supplementary file S1, Supplementary Material online.

Tardigrada (100%), and Arthropoda that includes Insecta (95%). ERM-like proteins from nematodes and tardigrades cluster together with the bootstrap support of 48%. Indeed, so far the phylogenetic position of tardigrades is an unanswered question, as these animals share features of their developmental stages and genetics with both arthropods and round worms (Gabriel et al. 2007; Yoshida et al. 2017).

Ezrin divergence from the group radixin–moesin in Vertebrata is supported with 100% bootstrap value; radixin and moesin diverge into the separated clusters with bootstraps of 66% and 100%, respectively. Judging from the branch lengths, radixin seems to be the slowest evolving

protein in the family. Coelacanthimorpha (*Latimeria chalumnae*), Teleostei (bony fishes), Holostei (*Lepisosteus oculatus*), Chondrichthyes (cartilaginous fishes), and Amphibia are clearly separated from each other and the closely related group of Prototheria–Metatheria–Theria–Sauria (mammals, reptiles, and birds). This is true for each of ezrin, radixin, and moesin clusters. The pattern is similar for the vertebrate merlins, except, the merlins from *Xenopus laevis* cluster within the mammals–reptiles–birds group. Each ERM protein in the bony fishes has an additional gene copy and, respectively, two clusters for each paralog. This agrees with the hypothesis of lineage-specific whole genome duplication (WGD) events.

Positive Selection in Fish Lineages

The long branch lengths for the vertebrate ezrin and merlin1 proteins suggest that they evolved faster than radixin, moesin, and merlin2 paralogs. To take a deeper look into the evolution of the vertebrate ezrin and the merlins, we analyzed the ratio of nonsynonymous to synonymous nucleotide substitutions, omega. For this, we built ezrin and merlin trees separately and performed the test on positive selection using PAML. First, we confirmed that omega is varying among the tree branches (P value 5.7e-109 for ezrin and 1.5e-39 for merlin1-2). Further, we tested for positive selection within the lineages. The results show that only ezrin in holostei and teleost fishes likely undergone evolution under positive selection (P value 0.005). None of the other lineages showed evidence for positive selection. Such phenomena can be explained by the additional duplication events in fishes and, therefore, weaker selective pressure (Glasauer and Neuhauss 2014). Interestingly, analysis of merlin1 evolution showed evidence of positive selection in all lineages (P value 4.6e-09). At the same time, the sequence of merlin2, the paralog responsible for antitumor activity, seems to be highly conserved during evolution in all vertebrates.

Paralog Number Diversity

All invertebrate taxa included in this analysis are characterized by the presence of zero to two ERM-like paralogs and zero to one merlin-like paralogs (table 1 and supplementary table S1, Supplementary Material online). In particular, *I. linei*, *T. adhaerens*, and all three species of Platyhelminthes have only one gene coding for an ERM-like protein and no merlins. Among Nematoda species, *Loa loa* has only an ERM-like protein, no merlin. These observations highlight the trend of simplification in parasites. All other invertebrate taxa have at least one ERM-like and one merlin-like protein. However, some paralogs can be missing from our data due to incomplete genome assemblies. It is interesting to note that some cnidarians and lophotrochozoans underwent a local duplication of ERM-like genes that resulted in paralogous genes located within few thousand nucleotides from each other (*Lingula anatina* and *Exaiptasia pallida*).

As a rule, vertebrates have three ERM proteins (ezrin, radixin, and moesin) and two merlins, although several very interesting exceptions can be found. As already mentioned, Eutheria, except armadillos, keep only merlin2. We indicated a few more cases of lineage-specific paralogs loss or gain. For example, teleost fishes have four to six ERM paralogs in different combinations, and Atlantic salmon (*Salmo salar*) has ten ERM genes (two of ezrin, four radixin, and four of moesin). This observation is in accordance with the hypothesis of a lineage-specific WGD in salmonids (Glasauer and Neuhauss 2014). Furthermore, some Neognathae birds have lost the moesin gene. *Xenopus laevis* have five ERM proteins (one ezrin, two radixins, and two moesins) and three merlin

paralogs, likely the result of another lineage-specific WGD that took place around 40 Ma (Van de Peer et al. 2009). Elephant shrews (*Elephantus edwardii*) have two ezrin genes and one radixin. Interestingly, one of its ezrin genes is a retrogene, which is not typical for mammalian ERMs. Besides, there are at least four pseudogenes descendants from the ERM family in this species. There are a few more cases of ERM pseudogenes throughout mammals, but the case of the elephant shrew is the most prominent.

To avoid any errors during the paralogs' number estimation, we ran TBlastN searches against any available Sequence Read Archive (SRA) sequences and whole genome sequences for the species which lack any of the paralogs.

Emergence of the three ERM paralogous genes in vertebrates could be a result of the two rounds of the 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 a lineage-specific WGD event, although the possibility of a duplication event of a local character cannot be excluded.

The situation with merlin genes is different. Even the lineages that underwent the additional WGD rounds and have an increased number of ERM paralogs still keep strictly two merlin genes. However, there is a large exception within the Eutheria clade with the majority of these lineages having lost a merlin1 gene (fig. 3). This fact was previously unknown or unappreciated, likely because most of the merlin studies were done on the representatives of the Eutheria clade (human and mouse), thus describing only the merlin2 paralog. Therefore, it is not surprising that merlin1 went unnoticed.

Synteny Relationships

We analyzed synteny conservation for ERM proteins at two levels: intrasynteny for ezrin, radixin, and moesin within *H. sapiens* species, and synteny for ERMs and ERM-like proteins between metazoan lineages. Interestingly, ezrin, radixin, and moesin genes preserve synteny very poorly, neither in their closest neighborhood (~500 Gb up- and down-stream) nor on the whole chromosomes (fig. 4). This observation suggests that the appearance of the three paralogs can be of a local character, rather than a result of a WGD event. However, in such a case, the persistence of the three paralogs in all the diverse vertebrate lineages would be difficult to explain.

On the other hand, a similarity can be found in gene locations between mammals (*H. sapiens* and *Mus musculus*), fish (*D. rerio*), cnidarians (*Stylophora pistillata*), *T. adhaerens*, and *A. queenslandica*. The remarkable conservation of the syntenic block can be seen between humans and mice, for example, where all seven protein coding genes surrounding ezrin in humans match those in mice (data not shown). Insects (*Dr. melanogaster*), however, show no trace of synteny with any of the lineages, which is surprising regarding the high sequence similarity of the ERM-like protein in fruit fly

Table 1
Number of Paralogous Genes in Different Metazoan Lineages

	ERM Like	Merlin Like	Comments
Mesozoa	1	0	*Intoshia linei*
Placozoa	1	0	*Trichoplax adhaerens*
Porifera	1	1	
Cnidaria	1–2	1	Local duplication in *Exaiptasia pallida*
Platyhelminthes	1	0	
Nematoda	1	0–1	
Arthropoda	0–3	1	
Lophotrochozoa	0–3	0–1	
Hemichordata	1	1	*Saccoglossus kowalevskii*
Cephalochordata	1	1	*Branchiostoma belcheri*
Tunicata	1	1	
Chondrichthyes	3	1–2	Cartilage fishes
Coelacanthimorpha	3	2	*Latimeria chalumnae*
Holostei	3	2	*Lepisosteus oculatus*
Teleostei	4–10	2	Bony fishes
Reptilia	3	1–2	
Aves	2–3	1–2	
Crocodylia	3	2	
Amphibia	5	3	*Xenopus laevis*
Eutheria	2–3	1–2	
Metatheria	3	2	Marsupials
Prototheria	3	2	Platypus

Note.—For more detailed counts, refer to the supplementary table S1, Supplementary Material online.
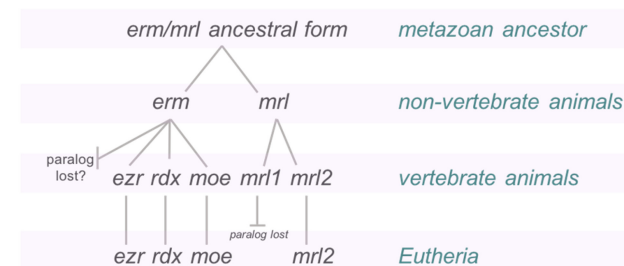
Fig. 3.—Scheme of gene duplication and paralogs loss in vertebrates. Erm, ERM like; mrl, merlin like; ezr, ezrin; rdx, radixin; moe, moesin.

and human ERMs. Such divergence could reflect the particular specificity of genome architecture and evolution in insects or in fruit flies. Perhaps, analysis of conservation of noncoding sequences, including repetitive elements, could be useful in understanding the syntenic divergence in *Dr. melanogaster* and the character of duplication events in the vertebrate lineage.

To assess the evolutionary background of the two merlin paralogs, we compared the gene content around merlin1 and merlin2 genes in chickens and zebrafish (Table 2). The revealed synteny suggests a higher genomic structure conservation between regions containing merlin2 genes. In

particular, more syntenic genes are located within 1-Mb intervals and share higher sequence identity. Synteny between the two different merlin paralogs (merlin1 and merlin2) is the least conserved, with only one syntenic gene pair present in close proximity to the merlin genes (Tab. 2). This pattern is more striking if comparing humans and zebrafish. The data present in the Synteny Database show 19 synteny pairs around the merlin2 gene in humans and zebrafish and no similarity when comparing human merlin2 and zebrafish merlin1 (see supplementary table S2, Supplementary Material online).

## Unicellular Ancestry of ERM + Merlin Family

Within the hypothesis that the closest unicellular relatives of animals are choanoflagellates and filastereans, we also 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 a 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, and Filasterea), although this cluster can be separated into two subgroups: filasteran and eutarian + choanoflagellate, 2) highly specialized proteins of the two choanoflagellate short proteins lacking a part of the middle domain, and 3) ERM group comprising eutherian and choanoflagellate homologs (fig. 5).

We modeled an ancestral sequence for the ERM + merlin family based on this tree (supplementary file S1, Supplementary Material online). It suggests a highly 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 the α-helical domain and the CERMAD. Computational prediction of the tertiary structure of this insertion characterized it as an extended structure with a low probability to form an α-helix (fig. 6).

Discussion

Phylogenetic analysis is a valuable approach in protein annotation and characterization and can significantly improve genome annotations. Unfortunately, it requires more time and effort than 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 phylogenetic tree here allows distinction 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 Eisen (1998). As a result, we were able to improve
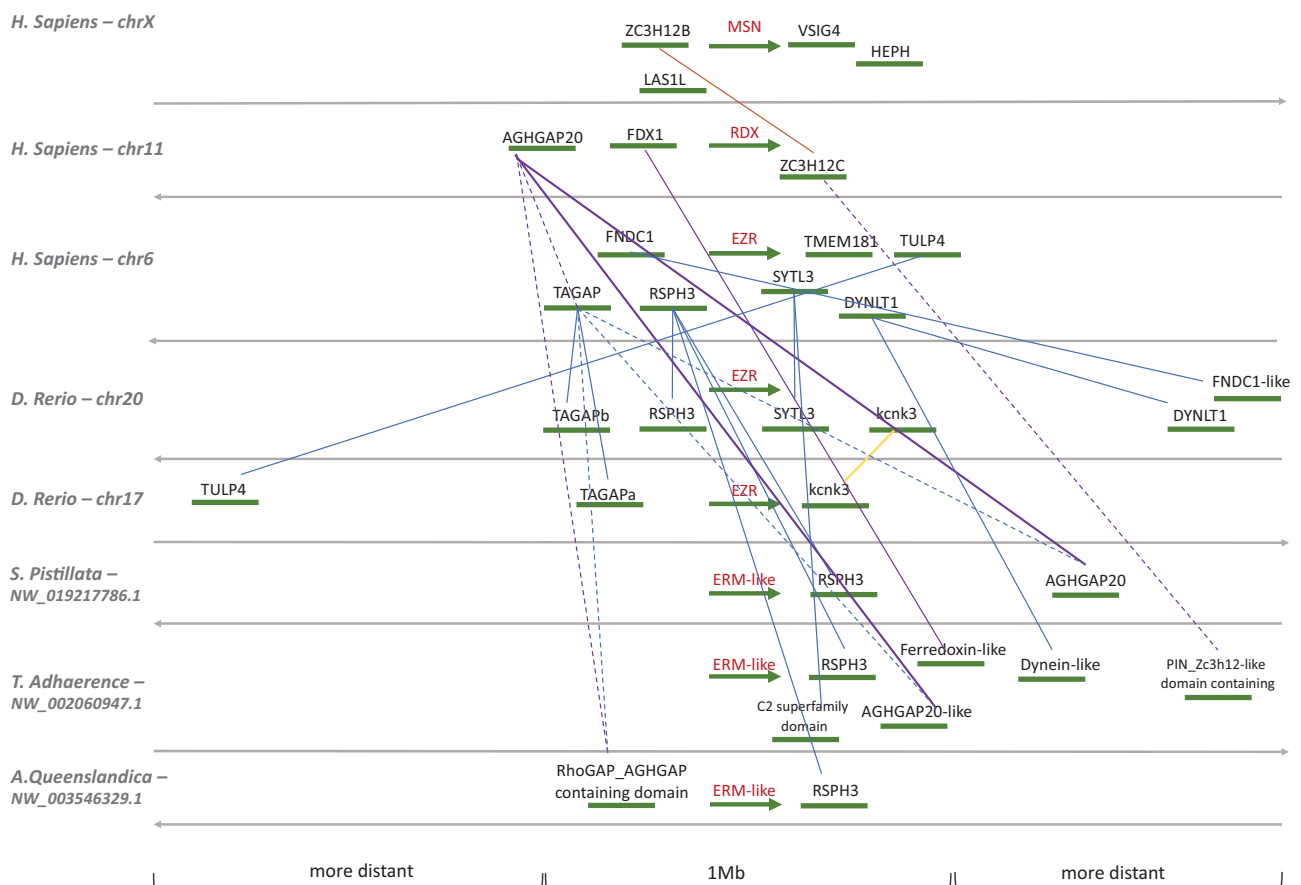
FIG. 4.—Synteny relationships for ezrin, radixin, and moesin genes in the genome of *Homo sapiens* and for ezrin and ERM-like proteins in different animals. A color scheme is used for better visualization. Solid lines connect syntenic genes and dash lines are used in cases where only parts of genes (proteins domains) are syntenic.

Table 2

Comparison of Synteny between merlin1 and merlin2 Genes

	Da.re mrl1–Da.re mrl2	Da.re mrl1–Ga.gl mrl1	Ga.gl mrl2–Da.re mrl2
Number of syntenic pairs	14	16	34
Synteny within 1 Mb interval	1	2	9
Synteny: same chromosome	13	14	25
Sequence identity estimator[a] (%)	51	66	67

[a]Sequence identity estimator shows how much of sequence identity is shared by 50% of all syntenic pairs. The detailed comparison table can be found at the supplementary table S2, Supplementary Material online.

annotations of these proteins in different species, as well as correct mistakes in several cases when, for example, an ezrin was erroneously assigned as a radixin or a moesin.

Furthermore, the tree clearly demonstrates that assignment of invertebrate proteins to "ezrin," "radixin," or "moesin" is inconsistent, because the divergence of the three ERM paralogs happened at the root of Vertebrata. A good

example is the two incorrect assignments at the NCBI protein database: XP_002160112.1 protein annotated as radixin in *Hydra vulgaris* and NP_727290.1 protein annotated as moesin in *Dr. melanogaster*. We suggest to restrict the names ezrin, radixin, and moesin only to vertebrate proteins while referring to others as ERM like or merlin like.

Some inconsistencies also happened in the studies of vertebrate ERMs. For example, there are experimental data discussed about chicken moesin (Winckler et al. 1994; Li and Crouch 2000), even though the authors did not find enough evidence to confidently claim moesin expression in the samples. Indeed, such experiments are questionable, as our phylogenetic analysis indicates that chickens (*G. gallus*) lost the moesin gene and have only ezrin and radixin. This exemplary case shows the importance of incorporating bioinformatics milieu in wet-lab studies, especially for less studied, nonmodel organisms, to avoid confusion and data discrepancy.

To get an insight into the evolution of the ERM + merlin protein family, we collected protein sequences from 84 species of Metazoa and 3 unicellular species. 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 a FERM domain in the Tree of Life (fig. 7A). Other FERM-containing proteins from amoebas Dictyostelia are related to talin and include a talin-specific N-terminal domain, therefore they are not discussed here. Our findings are in agreement with the study of the domain gain and loss in different taxa, which estimated that the FERM domain's origin took place in Holozoa (Grau-Bové et al. 2017). This is also consistent with the fact that the FERM domain is the proteins' most conserved part in this family. It is possible, that this corallochytrean FERM domain is a full-length protein, as computational predictions estimate that it folds into a stable structure. We suggest that this FERM-like protein is a membrane binding protein as it has a PIP_2 binding site. 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 actin binding capability of its newly acquired C-terminal part. This resulted in the acquisition of the scaffolding function similar to that of modern ERMs. Such mechanisms, for example, were discussed in the studies about the emergence of animal multicellularity (Shalchian-Tabrizi et al. 2008; Richter and King 2013).

Indeed, the first full-length ERM + merlin-like proteins can be found in the closest unicellular relatives of metazoans—choanoflagellates *S. rosetta* and *M. brevicollis* and filasterean *C. owczarzaki*. These proteins combine some characteristics of the ERM-like group (e.g., C-terminal ABS) and the merlin-like group (multiple dispersed prolines at the C-terminal end of the α-helical domain, absence of the ERM-specific R$_{/K}$EK$_{/R}$EEL motif within the α-helix). This finding shows that modern merlin-like and ERM-like proteins presumably emerged from the same ancestral form at the root of Metazoa. Phylogenetic reconstruction of a sequence of this ancestral form suggests that it could bind actin filaments and PIP_2 lipid, and therefore, perform the function of mechanical linkage of the cell membrane and underlying actin filaments.

Therefore, we can hypothesize that the ancestor of filasterean and choanoflagellates was able to form transient cell–cell or cell–surface contacts and could utilize its ERM + merlin-like protein for this purpose. This trend could be expanded in primitive metazoans to the scaffolding function within cell–cell contacts, for example, adherens junctions in Trichoplax. The ERM-like protein of Trichoplax already possesses the key features of the ERM family: ABS and binding sites for PIP_2, EBP50, ICAM-2, NHERF2, and intramolecular binding sites. 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.

Fig. 5.—ML tree indicating three clusters of ERM + merlin-like proteins in unicellular organisms: two species form choanoflagellates and one filasterean species. Group A unites merlin and merlin-like proteins, group B is represented by the two proteins with a rudimental middle domain, and group C represents ERM-like proteins. The protein length is shown in amino acids (aa). The tree in Newick format is in the supplementary file S1, Supplementary Material online.

Fig. 6.—Structural modeling. (A) 3D peptide structure prediction for the first 50 residues of the insertion in the reconstructed ancestral protein. Coloring is used for better visualization. (B) Probability plot for the first 47 amino acids of the insertion. Each amino acid is assigned to the probability to be included in a particular structure: red, α-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.

Fig. 7.—Early metazoan history of ERM + merlin protein family. (*A*) Schematic illustration of the early ERM + merlin phylogenetic history. The arrows indicate the first appearance of the protein structures. (*B*) Domain structure comparison. The predicted ERM + merlin ancestral protein and the protein from *Capsaspora owczarzaki* demonstrate longer insertions between the α-helix and the CERMAD. Numbers indicate the number of amino acids.

We propose that the function of the protein family in unicellular organisms was probably limited and restricted to scaffolding, partly because of the limitations in conformational regulation. Indeed, the autoinhibitory interaction in unicellular proteins can be altered, due to the presence of an extra amino acid stretch between the middle and CERMAD domains (fig. 7*B*). Prediction of the tertiary structure of this insertion indicates that it is unlikely an α-helix and such inclusion could drastically change protein folding, according to earlier studies (Hoeflich et al. 2003).

Further in the metazoan evolution, a decreasing distance between the middle and the C-terminal domain could be one of the evolutionary modifications that facilitated ERMs' autoinhibitory intramolecular binding. Therefore, our hypothesis is in accordance with the rheostat-like model of ezrin activation that ascribes a major role in this process to the α-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 and the ability to interact with diverse proteins in the cell. Consequently, they evolved more regulatory and binding sites, and eventually, acquired more functions in the cell, including signaling functions. ERM's intricate activity regulation mechanism became beneficial in vertebrate animals with an 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 accordance with the high level of sequence conservation among different lineages, ERM proteins also show high synteny preservation. It is more pronounced within mammalian taxa but also can be seen in such distantly related animals as Trichoplax and sponges. Surprisingly, very poor synteny conservation could be observed between ERM proteins within species. In the case of human ERMs, only one syntenic gene pair for moesin and radixin could be found, and one for ezrin and radixin. Because the proteins demonstrate a high level of sequence conservation and similarity in functions, it is unlikely that the ERM paralogs appeared within a local duplication event. It is rather the case that the genomic regions around the ERM genes underwent significant diversification within species.

Spatial expression of paralogs is often shown to be different from the ancestral gene, which can be a sign of subfunctionalization (Lynch and Conery 2000). Indeed, the existing RNAseq data of ezrin, radixin, and moesin demonstrate different expression patterns in different human tissues (fig. 8). This can explain the inconsistency in some of the data about ezrin, moesin, radixin, and merlin roles in cancer. It is often disregarded that experimental results can be influenced by the cell/tissue type, current availability of interacting partners, or ERMs' binding sites exposure. However, different databases have shown discrepancy in expression data for the ERM proteins (data not shown). Therefore, it is important to
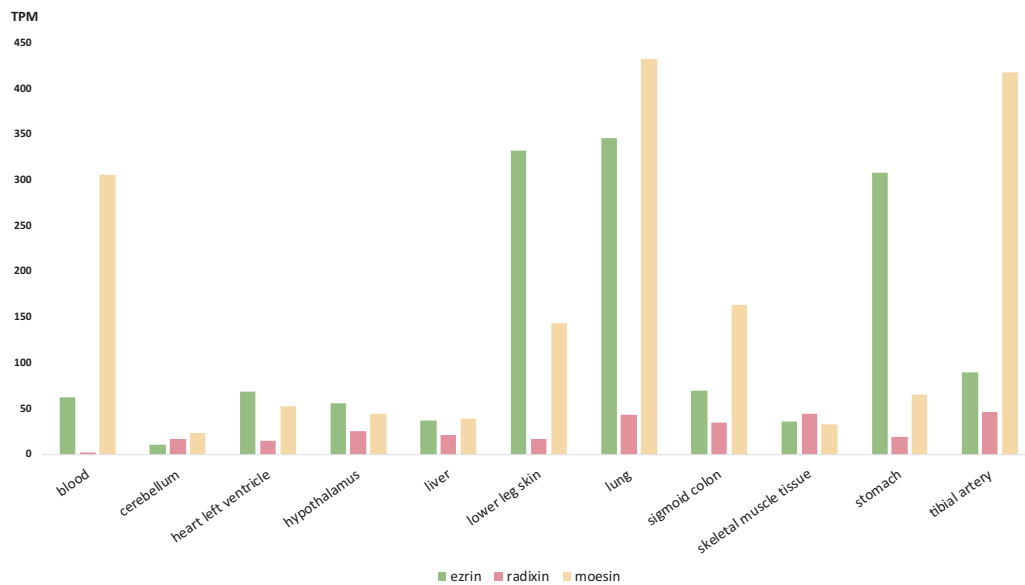
Fig. 8.—Expression levels for human ezrin, radixin, and moesin in different tissues. TPM stands for transcripts per kilobase million. The graph is based on data from the Genotype tissue expression portal (GTExPortal 2019).

investigate in details ERMs and merlin expression levels, for example, with the use of long-read sequencing, and including experiments in different developmental stages and different species. Tissue distribution of different splice isoforms is another interesting topic in ERM + merlin research: for example, in humans, there are 2 ezrin splice variants, 6 fradixin, 5 moesin, and 11 merlin splice variants. Unfortunately, not much is known about the functions of these different isoforms.

Studies of the paralog-specific sequence variation are important for the understanding of the role of each ERM protein, especially in cancer cells. Based on the variation analysis within the vertebrate lineages, we highlighted several motifs as candidates for experimental studies of the ERM proteins (supplementary table S2, Supplementary Material online). For example, there is not much known about the role of the polyproline stretch in ezrin and radixin. Moesin lacks this stretch although it has a structural analog (Li et al. 2007), and merlin possesses multiple discontinued prolines in the homologous site. The polyproline stretch can presumably bind an SH3 domain as another way of the proteins' activity regulation (Li et al. 2007), which would not be possible in moesin and merlin. Another motif of interest is located at the beginning of the α-helix and is specific in each of the proteins: EREKEQ in ezrin, EKEKEK in radixin, EKEKER in moesin, and ERTR/EKEK/EREK in merlin. These motif analogs were earlier shown to be important for supporting coiled-coil folding (Phang et al. 2016). The nearby REKEEL motif is specific to ERMs and is absent in merlins. However, its role in the folding of the proteins is unknown. Another candidate is the amphipathic stretch of 14 amino acids within the α-helix region that is known to be essential for binding the regulatory RII subunit of the protein kinase A (Dransfield et al. 1997). This region is

highly conserved in ezrin and radixin but less so in moesin that could reflect a functional difference between the proteins. Another difference between the three ERMs is the motif of six amino acids in the N-terminal end of CERMAD: $H_{/Q}$DENxA in radixin and moesin and xxExS$_{/x}$x in ezrin (where x is any amino acid).

At the same time, ezrin, radixin, and moesin retained a set of redundant functions, such as scaffolding molecular complexes in the regions of cell–cell contacts, which are most essential for cell survival. Ezrin is considered to be the major, indispensable paralog. Indeed, its knock-out in mice causes early death of the animals. However, the present assembly of the Tasmanian devil's genome lacks the ezrin gene. Some birds and mammals (elephant shrew) lost either radixin or moesin genes. This suggests that the ERM family can be characterized by genetic plasticity that is likely due to the conformational plasticity of the proteins as described by the rheostat-like model.

Our work, for the first time to our knowledge, stresses the existence of the two merlin paralogs in the vertebrate genomes: merlin1 and merlin2. Merlin1 was, apparently, lost in the Eutheria lineage, whereas 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 stretch of the last 30 amino acids, which is responsible for antitumor activity in human merlin2. Strikingly, merlin1 lacks one of the two N-terminal ABSs (in the F1 subdomain). These observations, together with the fact that actin binding is important for merlin antiproliferative activity (Cooper and Giancotti 2014), suggest that merlin1 is unlikely to exhibit tumor-suppressive effects. It can, therefore, perform a novel and unknown function or

specifically participate in cytoskeleton scaffolding, similar to merlin2 in its unfolded conformation (Lallemand et al. 2009).

The paralog number of merlin proteins is extremely conserved among all metazoan. Although ERMs' gene number can vary as a result of lineage-specific duplications, there is always only one merlin gene in invertebrates and mammals (except armadillos, in which both merlin1 and merlin2 are kept) and two merlin paralogs (merlin1 and merlin2) in other vertebrates. Such conservation may signify the sensitivity to gene dosage effect for both merlin paralogs. It is unclear though, why merlin1 was lost in most of mammalian lineages and whether another protein took over its function. One could also speculate that this paralog was lost together with organ/tissue-specific functions that are present in Amphibia, Sauria, and fishes, but not in Mammalia.

At the same time, merlin2 could have evolved specifically to counteract cellular dysregulations leading to cancer. It is not unlikely that during animal evolution the merlin2 paralog acquired a function additional to scaffolding that contributed to the evolution of the intricate anticancer protective system in vertebrates. Its increased importance, in comparison to merlin1, is mirrored in higher synteny conservation and small omega, the ratio of nonsynonymous to synonymous substitution rate.

The emergence of new proteins and new protein functions is an important question in evolutionary biology, whereas the 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 evolved from the first appearance of the FERM domain in holozoans to the functionally multifaceted group of the five homologs with tissue specificity. We propose that the three ERM paralogs retained in the vertebrates due to their conformational plasticity. This plasticity appeared to be beneficial in the conditions of the vertebrate evolution, such as increased complexity of organisms' physiology and biochemistry of cells. Here, the merlin1 paralog is for the first time discussed and is suggested to perform a yet unknown function specific to nonmammalian vertebrates. Merlin2 is the most interesting example of this protein family evolution, as it seems to be specifically adapted in vertebrates as anticancer protection.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Literature Cited

Adl SM, et al. 2012. The revised classification of eukaryotes eukaryotic microbiology. J Eukaryot Microbiol. 59(5):429–493.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server):W202–W208.

Bosanquet DC, et al. 2014. FERM family proteins and their importance in cellular movements and wound healing (Review). Int J Mol Med. 34(1):3–12.

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. J Cell Sci. 114(Pt 10):1901–1912.

Bretscher A, Edwards K, Fehon RG. 2002. ERM proteins and merlin: integrators at the cell cortex. Nat Rev Mol Cell Biol. 3(8):586–599.

Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. Genome Res. 19(8):1497–1505.

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 Pac J Cancer Prev. 13(3):753–759.

Clucas J, Valderrama F. 2014. ERM proteins in cancer progression. J Cell Sci. 127(2):267–275.

Cooper J, Giancotti FG. 2014. Molecular insights into NF2/Merlin tumor suppressor function. FEBS Lett. 588(16):2743–2752.

Curto M, et al. 2007. Contact-dependent inhibition of EGFR signaling by Nf2/Merlin. J Cell Biol. 177(5):893–903.

Darriba D, et al. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27(8):1164–1165.

Dayhoff MO. 1965. Atlas of protein sequence and structure. Silver Spring (MD): National Biomedical Research Foundation.

Debnath J, Brugge JS. 2005. Modelling glandular epithelial cancers in three-dimensional cultures. Nat Rev Cancer 5(9):675–688.

Dransfield DT, et al. 1997. Ezrin is a cyclic AMP-dependent protein kinase anchoring protein. EMBO J. 16(1):35–43.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8(3):163–167.

Estecha A, et al. 2009. Moesin orchestrates cortical polarity of melanoma tumour cells to initiate 3D invasion. J Cell Sci. 122(19):3492–3501.

Fairclough SR, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 14(2):R15.

Federhen S. 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40(D1):D136–D143.

Finn RD, et al. 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34(90001):D247–D251.

Funayama N, et al. 1991. Radixin is a novel member of the band 4.1 family. J Cell Biol. 115(4):1039–1048.

Gabriel WN, et al. 2007. The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development. Dev Biol. 312(2):545–559.

Gasteiger E, et al. 2005. Protein identification and analysis tools on the ExPASy server. In: The proteomics protocols handbook. Totowa (NJ): Humana Press. p. 571–607.

Glasauer SMK, Neuhauss SCF. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 289(6):1045–1060.

Golovnina K, et al. 2005. Evolution and origin of merlin, the product of the Neurofibromatosis type 2 (NF2) tumor-suppressor gene. BMC Evol Biol. 5(1):69.

Grau-Bové X, et al. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. eLife 6:pii: e26036.

GTExPortal 2019. *GTEx Portal*. Available from: https://gtexportal.org/home/ (accessed December 4, 2018).

Helander TS, et al. 1996. ICAM-2 redistributed by ezrin as a target for killer cells. Nature 382(6588):265–268.

Hoeflich KP, et al. 2003. Insights into a single rod-like helix in activated radixin required for membrane-cytoskeletal cross-linking. Biochemistry 42(40):11634–11641.

Ivetic A, Ridley AJ. 2004. Ezrin/radixin/moesin proteins and Rho GTPase signalling in leucocytes. Immunology 112(2):165–176.

Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240.

Katoh K, et al. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Kearse M, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649.

Kikuchi S, et al. 2002. Radixin deficiency causes conjugated hyperbilirubinemia with loss of Mrp2 from bile canalicular membranes. Nat Genet. 31(3):320–325.

Kitajiri S, et al. 2004. Radixin deficiency causes deafness associated with progressive degeneration of cochlear stereocilia. J Cell Biol. 166(4):559–570.

Kobayashi H, et al. 2004. Clinical significance of cellular distribution of moesin in patients with oral squamous cell carcinoma. Clin Cancer Res. 10(2):572–580.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.

Lallemand D, Saint-Amaux AL, Giovannini M. 2009. Tumor-suppression functions of merlin are independent of its role as an organizer of the actin cytoskeleton in Schwann cells. J Cell Sci. 122(22): 4141–4149.

Lamiable A, et al. 2016. PEP-FOLD3: faster *de novo* structure prediction for linear peptides in solution and in complex. Nucleic Acids Res. 44(W1):W449–W454.

Lankes WT, Furthmayr H. 1991. Moesin: a member of the protein 4.1-talin-ezrin family of proteins. Proc Natl Acad Sci U S A. 88(19):8297–8301.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25(7):1307–1320.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1): 127–128. doi: 10.1093/bioinformatics/btl529.

Li Q, et al. 2007. Self-masking in an intact ERM-merlin protein: an active role for the central α-helical domain. J Mol Biol. 365(5):1446–1459.

Li W, Crouch DH. 2000. Cloning and expression profile of chicken radixin. Biochim Biophys Acta 1491(1–3):327–332.

Liu X, et al. 2015. Moesin and myosin phosphatase confine neutrophil orientation in a chemotactic gradient. J Exp Med. 212(2):267–280.

Lu T-M, et al. 2017. The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution. Zool Lett. 3(1): 6.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155.

Marchler-Bauer A, et al. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43(D1):D222–D226.

Marion S, et al. 2011. Ezrin promotes actin assembly at the phagosome membrane and regulates phago-lysosomal fusion. Traffic 12(4):421–437.

McClatchey AI. 2014. ERM proteins at a glance. J Cell Sci. 127(Pt 15):3199–3204.

Michie KA, et al. 2019. Two sides of the coin: ezrin/radixin/moesin and merlin control membrane structure and contact inhibition. Int J Mol Sci. 20(8):1996.

Nambiar R, McConnell RE, Tyska MJ. 2010. Myosin motor function: the ins and outs of actin-based membrane protrusions. Cell Mol Life Sci. 67(8):1239–1254.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39(1):121–152.

Niggli V, Rossy J. 2008. Ezrin/radixin/moesin: versatile controllers of signaling molecules and of the cortical cytoskeleton. Int J Biochem Cell Biol. 40(3):344–349.

Omelyanchuk LV, et al. 2009. Evolution and origin of HRS, a protein interacting with Merlin, the Neurofibromatosis 2 gene product. Gene Regul Syst Biol. 3:143–157.

Phang JM, et al. 2016. Structural characterization suggests models for monomeric and dimeric forms of full-length ezrin. Biochem J. 473(18):2763–2782.

Pujuguet P, et al. 2003. Ezrin regulates E-cadherin-dependent adherens junction assembly through Rac1 activation. Mol Biol Cell 14(5):2181–2191.

Rambaut A. 2019. *FigTree*, *2009*. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. Available from: http://tree.bio.ed.ac.uk/software/figtree/ (accessed April 18, 2019).

Richter DJ, King N. 2013. The genomic and cellular foundations of animal origins. Annu Rev Genet. 47(1):509–537.

Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3):539–542.

Roy C, Martin M, Mangeat P. 1997. A dual involvement of the amino-terminal domain of ezrin in F- and G-actin binding. J Biol Chem. 272(32):20088–20095.

Saotome I, Curto M, McClatchey AI. 2004. Ezrin is essential for epithelial organization and villus morphogenesis in the developing intestine. Dev Cell 6(6):855–864.

Sayers E. 2009. The E-utilities In-Depth: Parameters, Syntax and More. Bethesda (MD): National Center for Biotechnology Information (US); Available from: https://www.ncbi.nlm.nih.gov/books/NBK25499/ last assesed June 30, 2019.

Sebé-Pedrós A, et al. 2013. Insights into the origin of metazoan filopodia and microvilli. Mol Biol Evol. 30(9):2013–2023.

Shalchian-Tabrizi K, et al. 2008. Multigene phylogeny of choanozoa and the origin of animals. PLoS One 3(5):e2098.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Stamatakis A. 2015. Using RAxML to infer phylogenies. In: Current protocols in bioinformatics. Hoboken (NJ): John Wiley & Sons, Inc. p. 6.14.1–6.14.14.

Stickney JT, et al. 2004. Activation of the tumor suppressor merlin modulates its interaction with lipid rafts. Cancer Res. 64(8):2717–2724.

Suga H, et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun. 4(1):2325.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server):W609–W612.

Turunen O, et al. 1998. Structure–function relationships in the ezrin family and the effect of tumor-associated point mutations in neurofibromatosis 2 protein. Biochim Biophys Acta 1387(1–2):1–16.

Valderrama F, Thevapala S, Ridley AJ. 2012. Radixin regulates cell migration and cell–cell adhesion through Rac1. J Cell Sci. 125(14):3310–3319.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725–732.

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. J Cell Sci. 107 (Pt 9):2523–2534.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 24(8):1586–1591.

Yonemura S, et al. 2002. Rho-dependent and -independent activation mechanisms of ezrin/radixin/moesin proteins: an essential role for polyphosphoinositides in vivo. J Cell Sci. 115(Pt 12):2569–2580.

Yoshida Y, et al. 2017. Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. PLoS Biol. 15(7):e2002266.

Associate editor: Sabyasachi Das