Research article

Emergence and evolution of the ERM proteins and merlin in metazoans

Shabardina V.[1*], Kashima Y.[2], Suzuki Y.[3], Makalowski W.[1]

[1]Institue of Bioinformatics, University of Muenster, Niels-Stensen-Strasse 14, Muenster, 48149, Germany.

[2]Department of Computational Biology and Medical Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan.

[3]Laboratory of Systems Genomics, Department of Computational Biology and Medical Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8562, Japan.

Author for correspondence: Shabardina V., Institute of Bioinformatics, University of Muenster, Muenster, Germany, +49 251 8353003, shabardi@uni-muenster.de

Abstract

Ezrin, radixin, moesin, and merlin are the cytoskeletal proteins which functions are specific to metazoans. They participate in cell cortex rearrangements, including cell-cell contact formation, and play important role in cancer progression. Here we perform a comprehensive phylogenetic analysis of the proteins spanning 87 species. The results describe a possible mechanism of 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 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 paralogs.

Keywords

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, 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) that binds to the FERM domain and phosphorylation of a conserved threonine

55 in CERMAD (Yonemura et al. 2002; Niggli & Rossy 2008). The important role during this
56 transition belongs to the middle α-helical domain. In the dormant ERMs it forms coiled-coil
57 structure, bringing N- and C-terminal domains together (Schliep *et al.*, 2017).

58 Little is known about individuality of each ERMs in, both, health and disease. The three
59 proteins are paralogs and share high amino acid sequence similarity (~75% in human)
60 (Funayama *et al.*, 1991; Lankes and Furthmayr, 1991). They demonstrate similar cellular
61 localization and are often discussed as functionally redundant. However, data from several
62 studies on knock-out mice revealed different phenotypes targeting different organs, and only
63 ezrin's depletion appeared to be lethal (Kikuchi *et al.*, 2002; Kitajiri *et al.*, 2004; Saotome, Curto
64 and McClatchey, 2004; Liu *et al.*, 2015). Special interest in ERMs' role in cancer stimulated
65 multiple studies. The proteins' dysregulation can lead to disruption of cell-cell contacts,
66 enhanced cell migration and invasion, and higher cancer cells survival (Clucas and Valderrama,
67 2014). Some studies showed that ezrin, radixin and moesin may exploit different cellular
68 mechanisms in tumors. (Pujuguet *et al.*, 2003; Kobayashi *et al.*, 2004; Debnath and Brugge,
69 2005; Estecha *et al.*, 2009; Chen *et al.*, 2012; Valderrama, Thevapala and Ridley, 2012).

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

79 With the more research being done, it is getting clear that ezrin, radixin, and moesin can
80 invoke different physiological effects in different tissue types, especially in cancer (Clucas and
81 Valderrama, 2014). However, their highly conserved sequence and tertiary structure make it
82 not a trivial task to distinguish their functions in vivo. Phylogenetic approach can be an effective
83 tool in resolving the problem, as it enables precise paralogs characterization by tracing
84 evolutionary history of the binding sites and conserved amino acid motifs. So far only few
85 phylogenies of ERMs and merlin have been described in the literature. As a rule, they feature
86 limited taxonomy representation or are included in the studies as an accessory and brief part
87 of the discussions (Turunen *et al.*, 1998; Golovnina *et al.*, 2005; Phang *et al.*, 2016). Thus,

88 although these phylogenies recover some interesting patterns, they do not provide the full
89 understanding of ERMs and merlin evolution. The studies agree on the fact that the proteins are
90 highly conserved within the metazoan clade, especially in vertebrates. Moreover, the
91 appearance of the ERM proteins and merlin in the tree of Life seems to coincide with the origin
92 of multicellularity in animals (Bretscher, Edwards and Fehon, 2002; Omelyanchuk *et al.*, 2009;
93 Nambiar, McConnell and Tyska, 2010; Sebé-Pedrós *et al.*, 2013). This view is supported by the
94 recent discovery of ERM-like proteins in Choanoflagellata and Filasterea, the closets unicellular
95 relatives of metazoans (Fairclough *et al.*, 2013; Suga *et al.*, 2013).
96 The position of merlin relative to the ERM family differs in the literature, and some studies
97 exclude merlin from the discussion of the ERM family altogether. Nevertheless, since these
98 proteins share evolutionary history and structural characteristics, it is reasonable to unite them
99 in one group. In this work we conduct the first comprehensive phylogenetic analysis for the
100 ERM family and merlin, that includes data from all sequenced by the time metazoan orders. The
101 results describe the ERM and merlin sequence conservation and paralog number diversity
102 within the clade of Metazoa. We suggest that the increased organism complexity led to
103 diversification of the protein paralogs in vertebrates. Moreover, we highlight the importance of
104 phylogenetic studies of paralogs, in general, in application to experimental biology, especially
105 in disease-related research.
106
107 Results
108
109 *Full length ERM/merlin-like proteins appeared within Metazoa-Filasterea-Choanoflagellata*
110 *group*
111 Search for the ERM and merlin homologs throughout all eukaryotic clades resulted in the
112 selection of 285 protein sequences spanning 87 species, including metazoan and unicellular
113 organisms. ERM-like proteins are also present in choanoflagellates (*Salpingoeca rosetta* and
114 *Monosiga brevicollis*) and filastereans (*Capsaspora owczarzaki*). A sequence of 298 amino acids
115 (Supplementary file) identified in corallochytrean *Corallochytrium limacisporum* revealed 25%
116 sequence identity to the FERM domain of human ezrin based on the tBLASTn search against the
117 species' whole genome sequence. Although such similarity level does not guarantee structural
118 alikeness, according to (Rost and Sander, 1994), the domain annotation by PFAM indicated that
119 this polypeptide from *C. limacisporum* belongs to the class of FERM domain with the high
120 statistical support (e-value < 10^{-5} for each of the three subdomains of FERM). Sequence-based

121 prediction of biochemical properties revealed that the two FERM domains, human and
122 corallochytrean, exhibit distinct features, including pI (8.75 for human ezrin FERM domain and
123 6.79 for the *C. limacisporum* polypeptide), amino acid content, instability index and
124 hydropathicity. In particular, metazoan FERM domain is predicted to be more hydrophilic than
125 its suggested corallochytrean homolog (grand average of hydropathicity (GRAVY) index is -
126 0.530 and -0.270, respectively) and less stable (instability index estimated to be 43.57, i.e.
127 unstable protein, and 31.80, stable, respectively). The only binding site conserved in the
128 corallochytrean FERM is the site for PIP_2 interaction. No ERM-like proteins or FERM-like
129 domains could be found in the other inspected taxa. The list of all the taxa and the
130 corresponding proteins IDs used for the analysis can be found at the Supplementary (Table S2).
131
132 *ERM+merlin protein family is conserved throughout all the metazoan orders and three unicellular*
133 *species*
134 Sequence comparison of the selected proteins demonstrated that the domain structure and
135 most of the known binding sites are conserved throughout the whole metazoan clade, although
136 there is some length variation. The amino acid motifs conservation analysis (Fig. 1) revealed
137 that the homologs of such early branching animals as *Trixhoplax adhaerens* (Placozoa) and
138 *Amphimedon queenslandica* (Porifera) demonstrate similar pattern to the mammalian proteins
139 with a high statistical significance. The FERM and CERMAD domains are characteristically well
140 preserved, even in the proteins from the unicellular organisms, while α-helical middle domain
141 is the least conserved ERM domain as has been previously described (Phang *et al.*, 2016). The
142 N-terminal part of the FERM shows some length variation throughout different taxa by
143 including short, non-conserved amino acid stretches. More striking length variation can be
144 observed within the region separating the α-helical domain and the CERMAD. It is short for
145 proteins from vertebrate animals, but is increased in all other taxa, the longest is in the protein
146 from *C. owczarzaki*. The sequence of this region is poorly preserved between the taxa. The
147 proteins from the species of flat worms, *Inoshia linei* and *C. owczarzaki* are the most divergent.

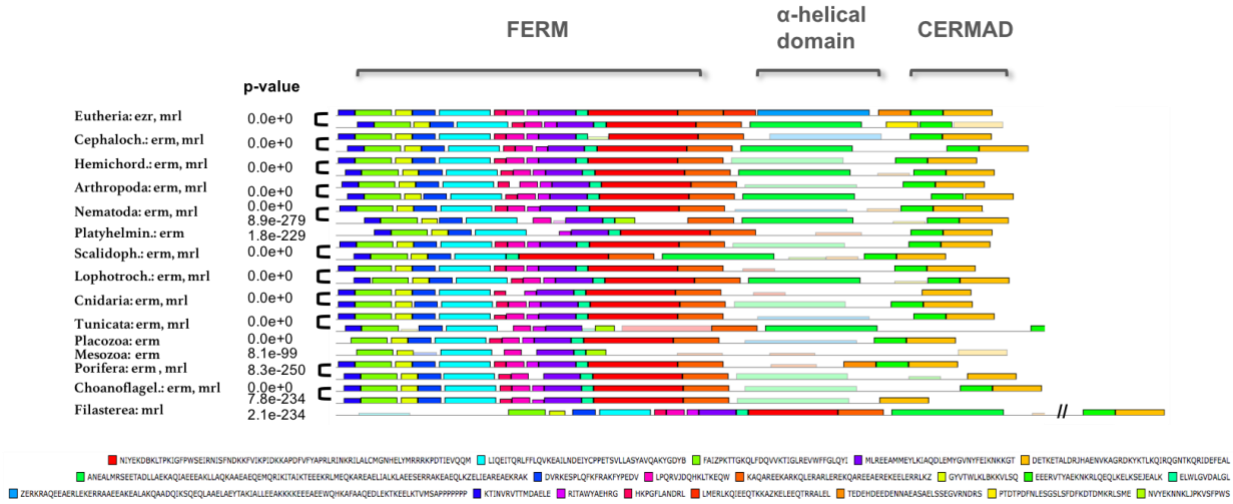


Figure 1. MEME conserved amino acid motif analysis. Motifs in pale color were found by the scanning algorithm based on the de-novo motif identification (bright color). Note the reduced length of the region separating the alpha-helical and CERMAD domains in eutherian proteins.

The binding sites for EBP50, ICAM-2, NHERF2 binding partners of the ERMs, for PIP_2, ABS, and intramolecular interaction sites (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 S2). 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 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 differing proteins come from the early branching metazoans *Molgula tectiformis* (Tunicata), *A. queenslandica* (Porifera*),* and endo- and exo-parasites (*I. linei*, flat worms, and blood sucking leech). Interestingly PIP_2 binding motif within F1 subdomain is not preserved only in *M. tectiformis*. Small variation in KYKTL motif can be seen in the proteins from Echinodermata and shark *Callorhinchus milii*.

Based on the sequence, merlin proteins can be classified into three groups (Fig. 2): 1) non-vertebrate proteins; 2) vertebrate merlin1, absent in most of Eutheria 3) all-vertebrate merlin2 (we arbitrary assign here merlin1 and merlin2 names to the merlin paralogs). The group 2 comprises merlin1 proteins coded by the paralogous gene that has not been described before (the list of the proteins is in the Supplementary, Table S2). It is characterized by a unique insertion of 68 amino acids in tetrapods and 15 amino acids PPYxPHSNRNSAYMx in bonny fishes 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

173 conserved Ser518. Interestingly, within Eutheria clade only armadillos (superorder Xenarthra)
174 possess merlin1 gene. Invertebrate merlins, group 1, are similar to vertebrate merlin1 but lack
175 the C-terminal domain insertion. Merlin2, group 3, is present in all vertebrate taxa and has an
176 additional actin binding site in the F1 of FERM and the tumor-suppression amino acid stretch
177 in its CERMAD domain (residues 532-579 in human). The blue-box region can also be identified
178 in the two proteins from the unicellular species: XP_004364665.2 in *C. owczarzaki* and
179 XP_004991962.1 in *S. rosetta*.
180
181 *Phylogenetic tree of the ERM+merlin family*
182 The reconstructed ML tree resulted in high phylogenetic resolution among vertebrates, while
183 the branching for most of the other taxa have low statistical support (Fig. 2). The alternative
184 methods of phylogenetic reconstruction, including neighbor-joining algorithm, parsimony
185 method, and Bayesian inference, could not improve the resolution (data not shown). Two
186 former trees revealed almost identical branching and statistical confidence. The Bayesian
187 reconstruction could not achieve conversion after 3000000 generations. The run was
188 terminated and the consensus tree was built anyway. It featured clustering of vertebrate
189 proteins supported by high bootstrap values, but unresolved branching for the invertebrate
190 sequences. One of the reasons can be an unequal representation of the taxa due to the lack of
191 sequencing data for the invertebrate clades. Another complication for the analysis could be the
192 high divergence of the amino acid sequences between evolutionary distant lineages. The ML
193 tree is used for the further discussion.
194 The most "eccentric" sequence in the reconstructed phylogeny is the one of the hypothetical
195 protein from *I. linei*, that, although features all three ERM domains, has the highest substitution
196 rate and does not cluster with any other groups. Consequently, it cannot be defined as ERM-like
197 or as merlin-like protein. This is not surprising, as *I. linei*, a representative of orthonectids, is a
198 parasitic animal, and its hermaphrodite nature, fast reproductive cycles and high level of
199 inbreeding can be the reason that makes its genome distant from the genomes of other
200 metazoans (Lu *et al.*, 2017).
201 The protein from *I. linei* was arbitrary chosen to separate the tree into two major clusters:
202 ERM-like and merlin-like groups. Therefore, any "hypothetical" or "unknown" proteins could
203 be annotated whether as ERM-like or as merlin-like, based on their position relative to *I. linei*'s
204 protein. Besides improving annotation of such proteins, some false annotations deposited in
205 the public protein databases were corrected (Supplementary, Table S2).

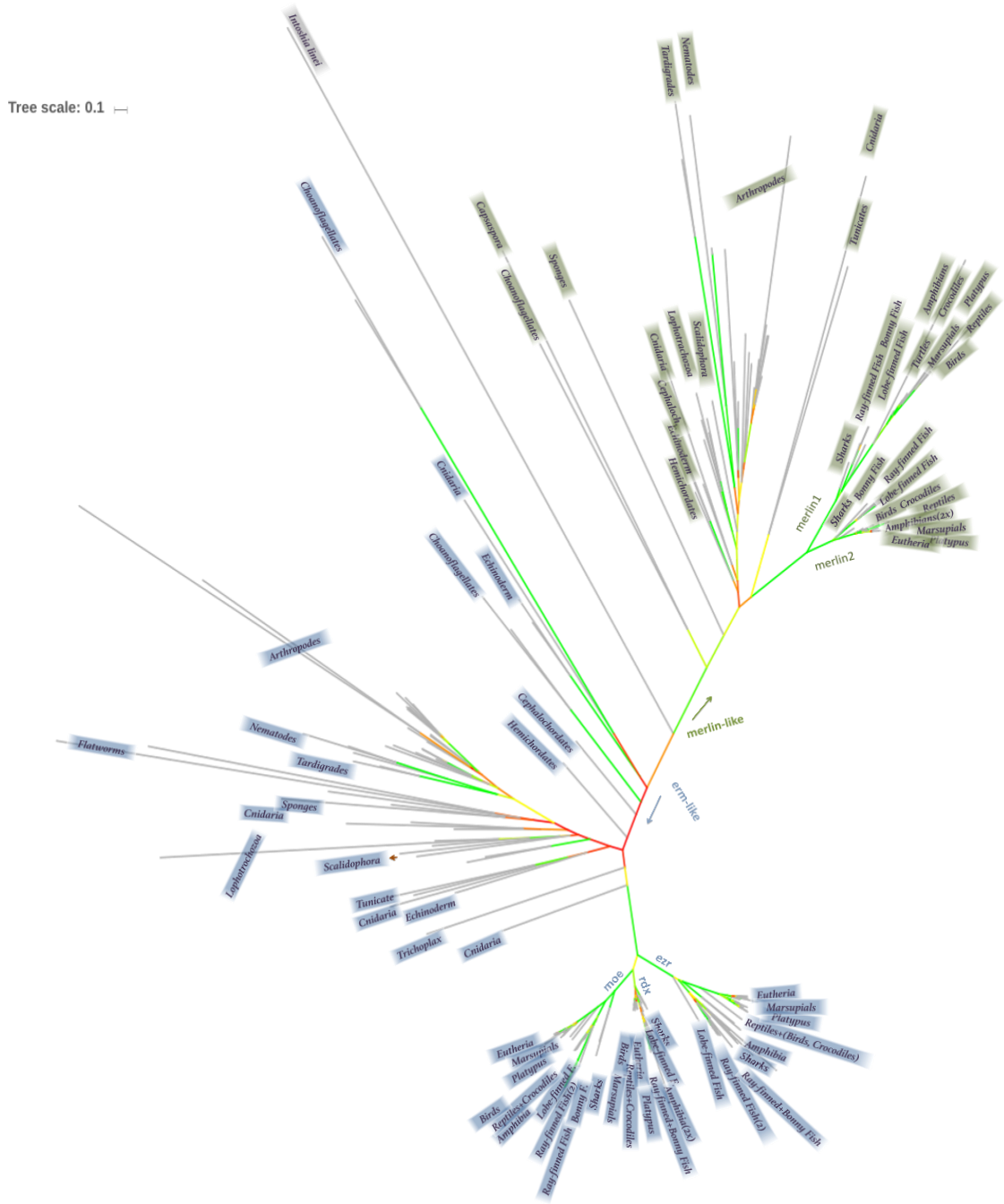


Figure 2. Phylogenetic ML tree: ERM+merlin family. The ERM-like part of the tree is shaded in blue, the merlin-like part is shaded in green. Note the emergence of the ezrin, radixin, moesin paralogs in vertebrates; and divergence of vertebrate merlins on merlin1 and merlin2 paralogs. Some proteins from the unicellular choanoflagellates species and a proteins 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%. Refer to the tree file in Newick format to see branch lengths and bootstrap values (Supplementary file).

215 Interesting observation can be made regarding the proteins from the unicellular species. The
216 *C. owczarzaki*'s protein XP_004364665.2 and S. rosetta's protein XP_004991962.1 cluster as
217 merlins; and XP_004997754.1 from *S. rosetta* and XP_001746613.1 from *M. brevicollis* cluster
218 as ERM-like. However, the bootstrap support is not high enough to confidently assign these
219 annotations. At the same time, the proteins XP_004994097.1 from *S. rosetta* and
220 XP_001743289.1 from *M. brevicollis* rather cluster together with the invertebrate ERM-like
221 group. That provides an insight into the origin of the ERM+merlin family and suggests that
222 merlin and ERMs diverged from the common ancestral protein before emergence of Metazoa.
223 Cnidaria, Placozoa, Echinodermata, Scalidofora, Lophotrochozoa, Porifera, Hemichordata,
224 Cephalochordata, and Mesozoa taxa branching could not be defined by this analysis with
225 statistical confidence. ERM-like proteins from the clade Tunicata are the most related to the
226 vertebrate ERMs (bootstrap support 83%), but the tunicate's merlins do not form clear
227 clustering. The three clearly distinguishable groups, beside vertebrates, are Nematoda (99%)
228 and Tardigrada (100%), and Arthropoda (76%) that includes Insecta (95%). ERM-like proteins
229 from nematodes and tardigrades cluster together with the bootstrap support of 48%. The
230 phylogenetic position of tardigrades is so far unclear and developmental stages and genetics of
231 these animals share features of both, arthropods and round worms (Gabriel *et al.*, 2007;
232 Yoshida *et al.*, 2017).
233 Ezrin divergence from the group radixin-moesin in Vertebrata is supported with 100%
234 bootstrap value; radixin and moesin diverge in the separated clusters with the bootstrap of
235 66% and 100%, respectively. Radixin seems to be the slowest evolving protein in the family
236 judging from the branch lengths. Coelacanthimorpha (*Latimeria chalumnae*), Teleostei (bonny
237 fishes), Holostei (*Lepisosteus oculatus*), Chondrichthyes (cartilaginous fishes) and Amphibia are
238 clearly separated from each other and the closely related group of Prototheria-Metatheria-
239 Theria-Sauria (mammals, reptiles, birds). This is true for each of ezrin, radixin, and moesin
240 clusters. The pattern is similar for the vertebrate merlins, except the merlins from *Xenopus*
241 *laevis* cluster within the mammals-reptiles-birds group. Each ERM protein in the bonny fishes
242 has an additional paralog and the divergence of ezrin, radixin, and moesin into two clusters
243 each can be distinctly seen within the bonny fishes clade. This agrees with the hypothesis of the
244 lineage-specific whole genome duplication (WGD) events (see further).

245 *Positive selection in fish lineages*

246 The long branch lengths for ezrin and merlin1 proteins in vertebrates suggest that they evolved
247 faster than radixin, moesin and merlin paralogs2. To take a deeper look into the evolution of
248 vertebrate ezrin and merlins we analyzed the ratio of nonsynonymous to synonymous
249 nucleotide substitutions, omega, along ezrin and merlin trees separately from the rest of the
250 data set. The codeml test in PAML for varying omega rates among the tree branches (branch
251 free-ratio model) verified that the varying ratio scenario is more likely than not (p-value 5.7e-
252 109 for ezrin and 1.5e-39 for merlin1-2). The branch free-ratio model is useful for overall
253 estimation but not very informative due to the little statistical power, as it uses a big number of
254 parameters for calculations. Therefore, we performed as well the branch-site model analysis
255 in order to test for positive selection within some lineages, based on the ML tree and the branch-
256 free ratio model estimation.

257 Among ezrin proteins only holostei and teleost fishes are likely to have undergone evolution
258 under positive selection (p-value 0.005). Other lineages seem to evolve under stricter constrain.
259 Such phenomena can be explained by the additional duplication events in fishes and, therefore,
260 weaker selective pressure (Glasauer and Neuhauss, 2014).

261 In the case of merlin, all tested branches of merlin1 showed evidence for positive selection
262 (p-value 4.6e-09). At the same time the sequence of merlin2, the paralog responsible for anti-
263 tumor activity, seems to be highly conserved during evolution in all vertebrates.

264

265 *Paralog number diversity*

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

276 Vertebrates have, as a rule, three ERM proteins (ezrin, radixin, and moesin) and two merlins.
277 Although several very interesting exceptions can be found. As it was already mentioned,

278 Eutheria, except armadillos, keep only merlin2. We indicated a few more cases of lineage
279 specific paralogs lost or gain. For example, teleost fishes have four to six ERM paralogs in
280 different combinations, and Atlantic salmon (*Salmo salar*) has ten ERM genes (two of ezrin, four
281 radixin, and four of moesin). This observation is in accordance with the hypothesis of a lineage
282 specific WGD in salmonids (Glasauer and Neuhauss, 2014). Further, some of Neognathae birds
283 lost moesin gene. *X. laevis* has five ERM proteins (one ezrin, two radixins, two moesins) and
284 three merlin paralogs, likely the result of another lineage specific WGD that took place around
285 40 million years ago (Van de Peer, Maere and Meyer, 2009). Elephant shrew (*Elephantus*
286 *edwardii*) has two ezrin genes and one radixin. Interestingly, one of its ezrin genes is a
287 retrogene. Besides there are at least four pseudogenes descendant from the ERM family in this
288 species. There are few cases of ERM pseudogenes throughout mammals, but the case of
289 elephant shrew is the most prominent.
290 To avoid any errors during the paralogs number estimation, we ran tBLASTn searches
291 against any available RSA sequences and whole genome sequences for the species that show
292 lack of any of the paralogs.

293 Table 1. Number of paralogous genes in different metazoan lineages. For more detailed count
294 refer to the Table S1 in the Supplementary.

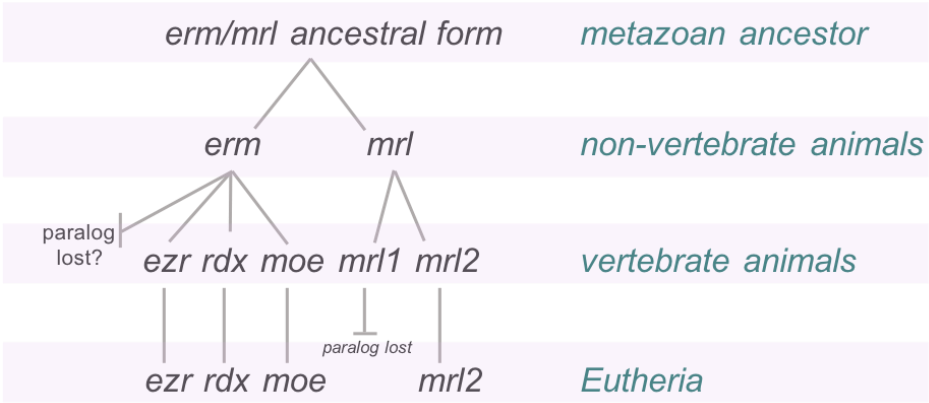
	ERM-like	Merlin-lik	comments
Mesozoa	1	0	*Intoshia linei*
Placozoa	1	0	*Trichoplax adhaerens*
Porifera	1	1	
Cnidaria	1-2	1	Local duplication in *E. pa*
Platyhelmynthes	1	0	
Nematoda	1	0-1	
Arthropoda	0-3	1	
Lophotrochozoa	0-3	0-1	
Hemichordata	1	1	*Saccoglossus kowalevskii*
Cephalochordata	1	1	*Branchiostoma belcheri*
Tunicata	1	1	
Chondrichthyes	3	1-2	cartilage fishes
Coelacanthimorpha	3	2	*Latimeria chalumnae*
Holostei	3	2	*Lepisosteus oculatus*
Teleostei	4-10	2	bonny fishes
Reptilia	3	1-2	
Aves	2-3	1-2	
Crocodylia	3	2	
Amphibia	5	3	*X. laevis*
Eutheria	2-3	1-2	
Metatheria	3	2	marsupials
Prototheria	3	2	platypus

295

296 Emergence of the three ERM paralogous genes in vertebrates could be a result of the two
297 rounds of the WGD that took place in the root of vertebrates and a consequent loss of one copy
298 of the gene. An additional increase of the paralogs number in the teleost fishes is likely to be the
299 result of the lineage specific WGD event. Although the possibility of a duplication event of a local
300 character cannot be excluded.

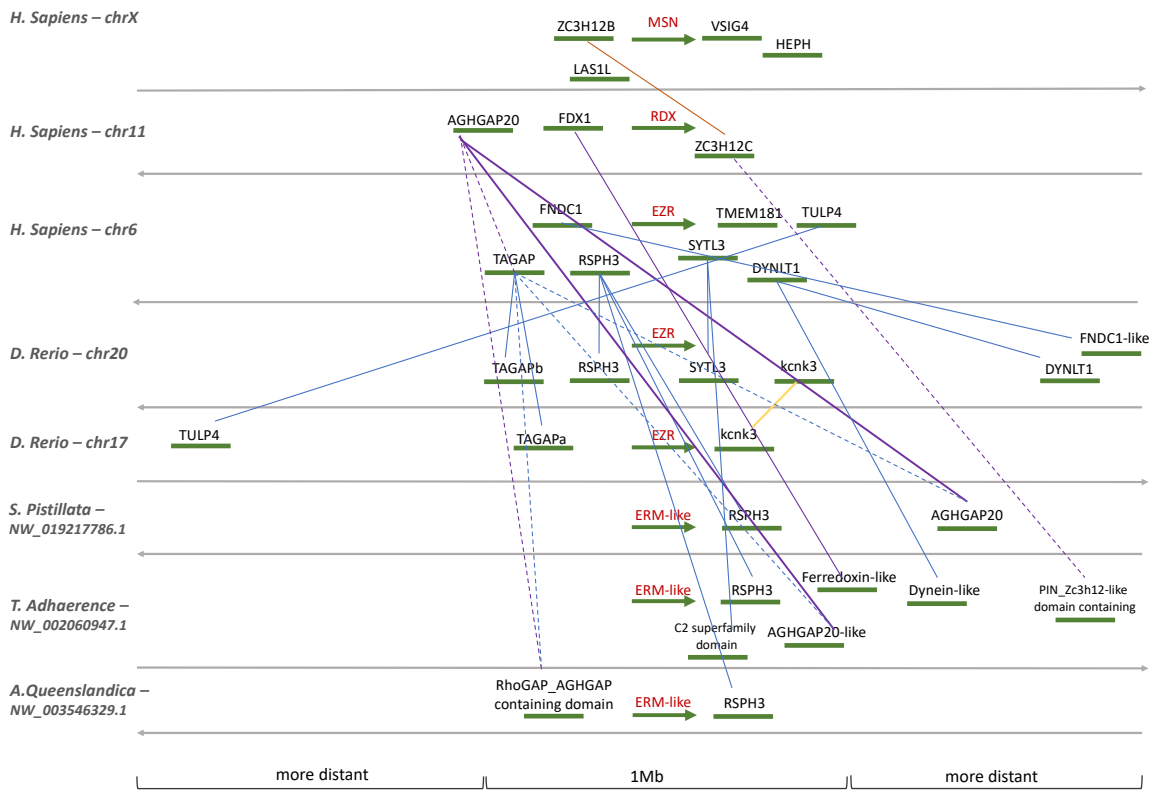
301 The situation with merlin genes is different. Even the lineages that underwent the additional
302 WGD rounds and have increased number of ERM paralogs, still keep strictly two merlin genes.

303 However, there is a large exception within the Eutheria clade: majority of the lineages there lost
304 merlin1 gene (Fig. 3). This fact was previously unknown or unappreciated, likely because most
305 of the merlin studies were done on the representatives of Eutheria clade (human, mouse), thus
306 describing only merlin2 paralog. Therefore, it is not surprising that merlin1 went unnoticed.
307



308
309 Figure 3. Scheme of gene duplication and paralogs lost in vertebrates. Erm – ERM-like, mrl – merlin-
310 like, ezr – ezrin, rdx – radixin, moe – moesin.
311
312 *Synteny relationships*
313 We analyzed synteny conservation for ERM proteins at two levels: intra-synteny for ezrin,
314 radixin, and moesin within *Homo sapiens* species, and synteny for ERMs and ERM-like proteins
315 between metazoan lineages. Interestingly, ezrin, radixin, and moesin genes very poorly
316 preserve synteny, neither in their closest neighborhood (~500 Gb up- and downstream), nor
317 on the whole chromosomes (Fig. 4). Such observation suggests that the appearance of the three
318 paralogs can be of a local character, rather than a result of the WGD event. However, in such
319 case the persistence of the three paralogs in all the diverse vertebrate lineages would be also
320 difficult to explain.
321 On the other hand, the similarity can be found in genes location between mammals (*H.*
322 *sapiens, Mus musculus*), fish (*Danio rerio*), cnidarians (*Stylophora pistillata*), *T. adhaerens*, and
323 *A. queenslandica*. The remarkable conservation of the syntenic block can be seen between
324 human and mouse, for example, where all seven protein coding genes surrounding ezrin in
325 human match those in mouse (data not shown). Insects (*Drosophila melanogaster*), however,
326 show no trace of synteny with any of the lineages, that is surprising regarding the high sequence
327 similarity of the ERM-like protein in fruit fly and human ERMs. Such divergence could reflect
328 the specificity of insects' - or fruit fly, in particular - genome architecture and evolution.

329 Perhaps, analysis of conservation of non-coding sequences, including repetitive elements, could
330 be useful in understanding the syntenic divergence in *D. melanogaster* and character of
331 duplication events in vertebrate lineage.
332



333
334 Figure 4. Synteny relationships for ezrin, radixin, and moesin genes in the genome of *H. sapiens* and
335 of ezrin and ERM-like proteins in different animals. Color scheme is used for the better visualization.
336 Solid lines connect syntenic genes, dash lines are used in case if only parts of genes (proteins domains)
337 are syntenic.
338
339 To assess the evolutionary background of the two merlin paralogs, we compared gene
340 content around merlin1 and merlin2 genes in chicken and zebrafish. The revealed synteny
341 suggests the higher genomic structure conservation between regions containing merlin2 genes.
342 In particular, more syntenic genes located within 1 Mb interval and shared higher sequence
343 identity. Synteny between the two different merlin paralogs (merlin1 and merlin2) is the least
344 conserved, with only one syntenic gene pare present in the close proximity to the merlin genes
345 (Tab. 2). This pattern is more striking if comparing human and zebrafish. The data present in
346 the Synteny Database show 19 synteny pares around merlin2 gene in human and zebrafish and

347 no similarity when comparing human merlin2 and zebrafish merlin1 (see Supplementary,
348 TableS2).
349
350 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 pares	14	16	34
Synteny within 1 Mb interval	1	2	9
Syntheny: same chromosome	13	14	25
Sequence identi estimator[a]	51%	66%	67%

351 [a] Sequence identity estimator shows how much of sequence identity share 50% of all
352 syntenic pares. The detailed comparison table can be found at the Supplementary table S2.
353
354 *Unicellular ancestry of ERM+merlin family*
355 With the assumption that the closest unicellular relatives of animals are choanoflagellates and
356 filastereans, we assumed that their ERM-like and merlin-like proteins are the best candidates
357 for speculating about the proteins ancestral form. We built a small phylogeny tree, including
358 only few sequences from an eutherian representative (*H. sapiens*) and *S. rosetta, M. brevicollis,*
359 *and C. owczarzaki*. Eliminating the rest of the taxa decreases the reliability of the reconstruction,
360 but, regarding the high sequence diversity and scarcity of the data for invertebrates, this
361 approach is the most straightforward. The tree highlights three groups: 1) merlins (Eutheria,
362 Choanoflagellata, Filasterea), although this cluster can be separated into two subgroups:
363 filasteran and eutarian+choanoflagellate; 2) highly specialized (or the result of genome mis-
364 assembly) proteins of the two choanolagellate short proteins lacking most of the middle
365 domain; 3) ERM group comprising eutherian and choanoflagellate homologs (Fig. 5).
366

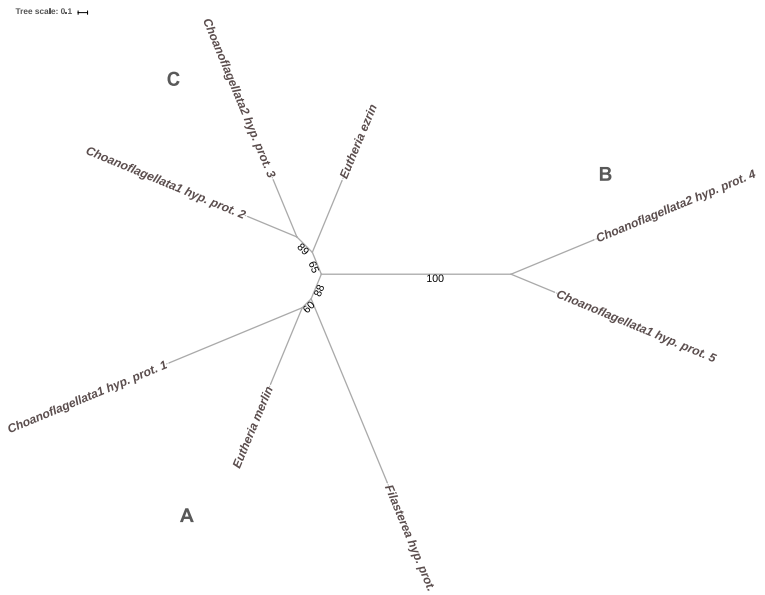


Figure 5. ML tree indicating three clusters of ERM/merlin-like proteins in the unicellular organisms: two species form choanoflagellates and one filasterean species. Group A unites merlin and merlin-like proteins, clade B is represented by the two proteins with a rudimental middle domain, clade C – ERM-like proteins. The tree in Newick format is in Supplementary file.

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). 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 α-helix (Fig. 6).

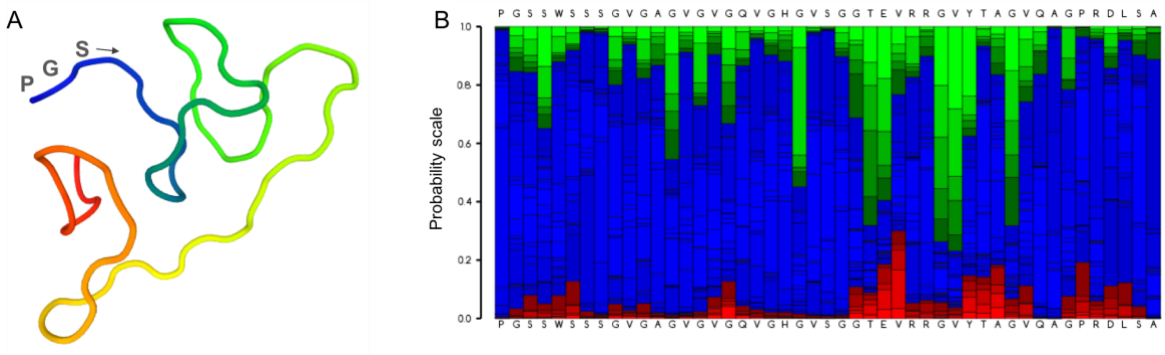


Figure 6. Structural modelling. A: 3D peptide structure prediction for the first 50 residues of the insertion in the reconstructed ancestral protein. Coloring is used for a better visualization. B: Probability plot for the first 47 amino acids of the insertion, each amino acid is assigned a probability to be included in a particular structure: red – 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.

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 moesin.

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 *Hydra 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 in the studies with vertebrate ERMs. For example, there are experimental data discussed for chicken moesin (Winckler *et al.*, 1994; Li and Crouch,

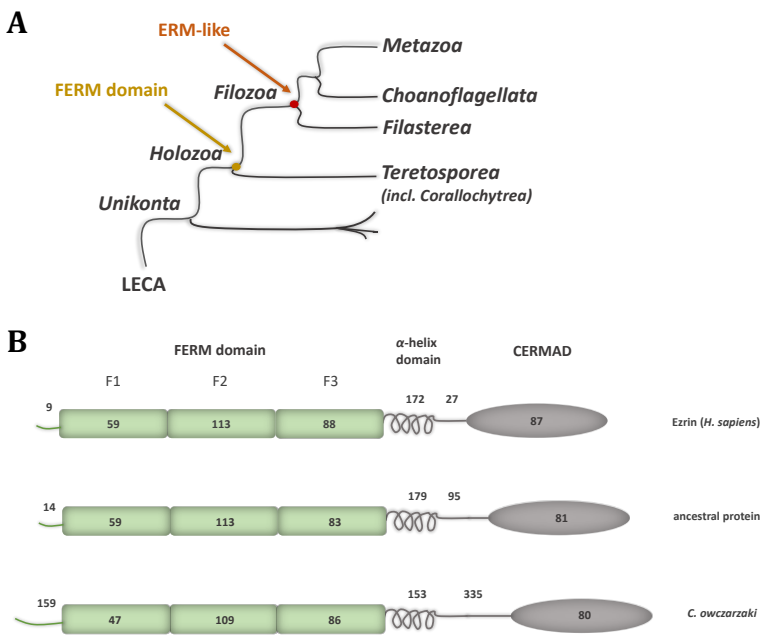
409 2000), even though the authors did not find enough evidence to confidently claim moesin's
410 expression in the samples. Indeed, such experiments are questionable, as our phylogenetic
411 analysis indicates that chicken (*Gallus gallus*) lost moesin gene and has only ezrin and radixin.
412 Such an example shows the importance of incorporating bioinformatics milieu in the wet-lab
413 studies, especially for the proteins from the less studied, not model organisms, to avoid
414 confusion and data discrepancy.

415 To get an insight on the evolution of the ERM+merlin protein family, we collected protein
416 sequences from 84 species of Metazoa and 3 unicellular species from the clades
417 Choanoflagellata and Filasterea. We could also identify a FERM-like domain with the conserved
418 PIP_2 binding site in a species from Corallochytrea clade, likely the first lineage with FERM
419 domain in the tree of Life (Fig. 7A). This finding is in agreement with the study of the domain
420 gain and lost in different taxa, that estimated that the FERM's origin took place in Holozoa
421 (Grau-Bové *et al.*, 2017). This is also consistent with the fact that FERM domain is the most
422 conserved part of the proteins in the family. It is possible, that this corallochytrean FERM
423 domain is a full-length protein, as computational prediction by ExPASy ProtParam algorithm
424 estimated that it likely folds into a stable structure. We suggest that this FERM-like protein is a
425 membrane binding protein involved in communication of the cell and extracellular
426 environment. It can be similar to a predecessor of the ERM+merlin family. Domain shuffling
427 and/or shifts within the open reading frame of the predecessor gene could lead to the origin of
428 a longer protein with an acting binding capability of its newly acquired C-terminal part, i.e. with
429 a scaffolding function similar to that of the modern ERMs. Such mechanisms, for example, were
430 discussed in the studies about emergence of animal multicellularity (Shalchian-Tabrizi *et al.*,
431 2008; Richter and King, 2013).

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

441 An ancestor of filasterean and choanoflagellates was likely able to form transient cell-cell or
442 cell-surface contacts and could exploit its ERM/merlin-like protein for this purpose, as *S. rosetta*
443 and *C. owczarzaki* probably do, as suggested by sequence analysis of their homologs. This trend
444 could be expanded in primitive metazoans to the scaffolding function within cell-cell contacts,
445 for example, adherens junctions in trichoplax. ERM-like protein of this early branching
446 metazoan already possesses the key features of the ERM family: ABS and binding sites for PIP_2,
447 EBP50, ICAM-2, NHERF2, and intramolecular binding sites. Similarly, one study suggested that
448 ERM proteins were involved in the development of the filopodia in metazoans (Sebé-Pedrós *et*
449 *al.*, 2013). However, more sequencing data from other unicellular taxa and in vivo experiments
450 are required to support or reject this hypothesis.
451 Function of the protein family in the unicellular organisms was probably limited and
452 restricted to scaffolding, partly because of inability to activity regulation. Indeed, the auto-
453 inhibitory interaction in the unicellular proteins is questionable, due to the presence of an extra
454 amino acid stretch between the middle and CERMAD domains (Fig. 7B). Prediction of the
455 tertiary structure of this insertion indicates that it is unlikely an α-helix, therefore, such
456 inclusion could drastically change protein folding.

457



458

459

460 Figure 7. Early metazoan history of ERM+merlin protein family. A: Schematic illustration of the early
461 ERM+merlin phylogenetic history. The arrows show first appearance of the protein structures. B:
462 Domain structure comparison. The predicted ERM/merlin ancestral protein and the protein from *C.*
463 *owczarzaki* demonstrate longer insertions between the α-helix and CERMAD. Numbers indicate number
464 of amino acids.

465

466 Further in the metazoan evolution, decreasing distance between the middle and the C-
467 terminal domain could be one of the evolutionary modifications that facilitated ERMs'
468 characteristic auto-inhibitory, intramolecular binding. Therefore, our hypothesis is in
469 accordance with the rheostat-like model of ezrin activation that ascribes the major role in this
470 process to the alpha-helical domain (Li *et al.*, 2007). The rheostat-like manner of activation
471 allows intermediate protein states between its inactive and active form. This multilevel manner
472 of conformational regulation granted biochemical flexibility to ERM and merlin proteins.
473 Consequently, they evolved more regulatory and binding sites and, therefore, acquired more
474 functions in the cell, and eventually, became involved in signaling pathways. ERM's intricate
475 activity regulation mechanism became beneficial in vertebrate animals with increasing
476 complexity of their cellular physiology and number of cell types. As a result, three ERM and two
477 merlin paralogs diverged, acquiring some tissue specific functions in a process that can be
478 described by the birth-and-death model (Nei and Rooney, 2005).

479 In agreement with the high sequence conservation among different lineages, ERM proteins
480 also show high synteny preservation, especially within mammalian taxa, but also in such a
481 distantly related animals as trichoplax and sponges. Surprisingly, very poor synteny
482 conservation can be observed between ERM proteins within species. In the case of human
483 ERMs, one syntenic genes pare for regions around moesin and radixin genes, and one – for
484 regions around ezrin and radixin genes. Keeping in mind the mentioned conserved features of
485 the proteins and their coding genes it is unlikely that the ERM paralogs appeared within a local
486 duplication event. Rather the genomic regions including the ERM genes underwent significant
487 diversification within species.
488 Spatial expression of paralogs is often shown to be different from the ancestral gene, that can
489 be a sign of sub-functionalization (Glasauer and Neuhauss, 2014). Indeed, the existing RNAseq
490 data of ezrin, radixin, and moesin demonstrate different expression patterns in different human
491 tissues (Fig. 8). This can explain the inconsistency of some data about ezrin, moesin, radixin,
492 and merlin roles in cancer, since the experimental results can be influenced by the cell/tissue
493 type, momentary availability of interacting partners or ERMs' binding sites exposure.
494 Therefore, it is important to investigate in details the ERMs and merlin expression levels, for
495 example, with the use of long read sequencing, and including experiments in different
496 developmental stages. Tissue distribution of different splice isoforms is another interesting
497 topic in ERM+merlin research: for example, in humans there are two ezrin splice variants, six –
498 for radixin, five – for moesin, and eleven - for merlin. Unfortunately, not much is known about
499 functions of different isoforms.

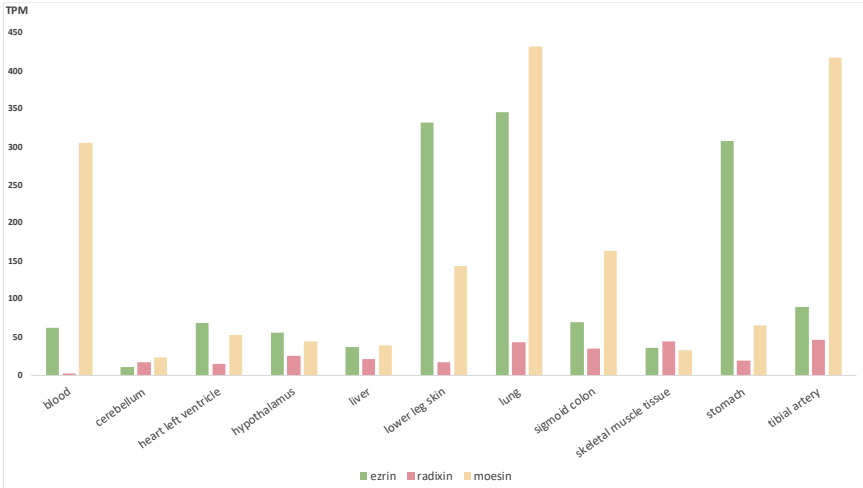


Figure 8. Expression levels for human ezrin, radixin, and moesin in different tissues. TPM – transcripts per kilobase million. Based on the data from Genotype tissue expression portal (GTExPortal, 2019).

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 α-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 α-helix region that is known to be essential for binding regulatory Rll 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 $H_{/Q}$DENxA in radixin and moesin and xxEx$S_{/x}$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

525 regions of cell-cell contacts. Ezrin is considered to be the major, indispensable paralog. Indeed,
526 its knock out in mice causes early death of the animals. However, genomes of Tasmanian devil
527 lack ezrin gene, at least based on the sequencing material that is available for this species at the
528 moment. Some birds and mammals (elephant shrew) lost either radixin or moesin genes. This
529 suggest the genetic plasticity of ERM family between different vertebrate lineages that is likely
530 due to the conformational plasticity of the proteins.

531 In this work, we for the first time, to our knowledge, stress the existence of the two merlin
532 paralogs in the vertebrate genomes: merlin1 and merlin2. Merlin1 was, apparently, lost in the
533 Eutheria lineage, while merlin2 is present in all vertebrates. Merlin1 contains an additional
534 amino acid stretch within its CERMAD that is absent from merlin2. Also, merlin1 has a weak
535 sequence similarity to merlin2 in the last 30 amino acids, that is responsible for anti-tumor
536 activity in human merlin2. Strikingly, merlin1 lacks one of the two N-terminal actin binding
537 sites (in the F1 subdomain). These observations together with the fact that actin binding is
538 important for merlin anti-proliferative activity (Cooper & Giancotti 2014), suggests that
539 merlin1 is unlikely to exhibit tumor suppressive effect. It can, therefore, perform a novel,
540 unknown function or specifically participate in cytoskeleton scaffolding, similar to merlin2 in
541 its unfolded conformation (Lallemand, Saint-Amaux and Giovannini, 2009).

542 The paralog number of merlin proteins is extremely conserved among all metazoan: while
543 ERMs' gene number can vary as a result of lineage specific duplications, there is always only
544 one merlin gene in invertebrates and mammals (except armadillo, who keeps, both merlin1 and
545 merlin2) and two merlin paralogs (merlin1 and merlin2) in other vertebrates. Such
546 conservation may signify the sensitivity to gene dosage effect for both merlin paralogs. It is
547 unclear, though, why merlin1 was lost in the most of mammalian lineages and what protein
548 took over its function. One could also speculate that this paralog was lost together with the
549 organ/tissue-specific function that are present in Amphibia, Sauria, and fishes, but not in
550 Mammalia.

551 At the same time merlin2 seems to be evolved specifically to counteract cellular
552 dysregulations leading to cancer. It is likely that within evolution merlin2 paralog acquired
553 additional to scaffolding function to contribute to evolving of the intricate anti-cancer
554 protective system in vertebrates. Its increased in comparison to merlin1 importance is
555 mirrored in higher synteny conservation and lower rate of nonsynonymous substitutions
556 appearance. One can say that merin2 protein is one of the features that were evolved by the
557 animal cell as adaptation to complex multicellularity.

558 Emergence of new proteins and new protein functions is an important question in
559 evolutionary biology, and fate of paralogous forms is probably one of the least understood
560 aspects of the process. Based on sequence comparison and phylogenetic reconstruction we
561 hypothesized the way the ERM+merlin protein family could have gone from the first
562 appearance of the FERM domain in holozoans to the functionally multifaceted group of five
563 homologs with tissue specificity. We propose that the three ERM paralogs retained in the
564 vertebrates due to their conformational plasticity that appeared to be beneficial in the
565 conditions of the vertebrate evolution: increased complexity of organisms' physiology and
566 biochemistry of the cells. Merlin1 paralog is for the first time discussed here, and suggested to
567 perform a yet unknown function specific to non-mammalian vertebrates. Merlin2 is the most
568 interesting example of this protein family evolution, as it seems to be specifically adapted in
569 vertebrates to anti-cancer protection.
570
571 Materials and methods
572
573 *Data collection*
574 The amino acid sequences of ezrin, radixin, moesin, and merlin were collected using BLASTp
575 (Altschul *et al.*, 1997) with the human protein sequences as queries (ezrin NP_001104547.1,
576 radixin AAA36541.1, moesin NP_002435.1, merlin NP_000259.1) against non-redundant (nr)
577 protein sequences collection at the NCBI database. The selected sequences were manually
578 monitored to exclude database duplicates, splice variants, truncated sequences and to
579 reconstruct correct protein sequences when needed. If several splice variants were described
580 for an organism, the longest one was chosen for the analysis. Only the sequences that spanned
581 all three ERM domains were selected for the analysis. PFAM (Finn *et al.*, 2006), InterProScan
582 (Jones *et al.*, 2014) and CDD domain search (Marchler-Bauer *et al.*, 2015) were used for domain
583 structure analysis and verification.
584 Taxa selection was done based on the following requirements: 1) every described order of
585 Metazoa should be represented by one species, although some exceptions were made in order
586 to balance taxa representation; 2) whole genome or transcriptome of a representative species
587 should have been sequenced and available; 3) high quality of the genome assembly annotation.
588 In the case if no representative genome was available for an order, tBLASTn search was run
589 against all available nucleotide sequences for that taxa. Search for possible homologs of ezrin,
590 radixin, moesin, and merlin was performed among other Opisthokonta (Holozoa, Nuclearia,

591 Fungi) and Amoebozoa, Excavata, Archaeplastida (includes green plants), SAR cluster
592 (Stramenopiles, Alveolata, and Rhizaria), and other protist groups (refer to the tree of Life
593 scheme (Adl *et al.*, 2012)). Prokaryota and Archaea nucleotide sequences were scanned for the
594 whole length proteins or only FERM domain using tBLASTn search against nr nucleotide
595 collection at NCBI. The taxonomic structure describing the final dataset can be viewed at the
596 Supplementary (Table S1) and is based on the topologies employed by NCBI Taxonomy
597 database (Federhen, 2012) and the Tree Of Life project (Letunic & Bork). The taxa variety will
598 be further discussed in terms "vertebrate" and "invertebrate", the later including the rest of
599 Eumetazoa. The sequences were retrieved by July 2018; data for vertebrates were updated in
600 May 2019.
601
602 *Reconstruction of the ERM+merlin phylogeny*
603 Multiple sequence alignment was generated using MAFFT software (Katoh *et al.*, 2002) with
604 the PAM70 (Dayhoff, 1965) substitution matrix, as defined by ProtTest3 (Darriba *et al.*, 2011),
605 and manually edited to remove uninformative columns. CLUSTALX (Larkin *et al.*, 2007) and
606 Geneious (Geneious, 2019) were used for the alignment visualization. Maximum Likelihood
607 (ML) phylogenetic trees were build using RAxML tool (Stamatakis, 2014) with the parameters
608 estimated by running RAxML parameter test (Stamatakis, 2015). As a result, the
609 PROTGAMMALG model was chosen, where GAMMA model estimates the substitution rate
610 between sites and LG is amino acid substitution matrix (Le and Gascuel, 2008). The statistical
611 support for the tree clustering was calculated by running 1000 bootstrap replicates. The
612 resulting trees were inspected and edited using iTOL online tree viewer (Letunic and Bork,
613 2019) and FigTree software (Rambaut A., 2019).
614 A reduced data set (Supplementary, Table S2) was used to reconstruct a tree for analyzing
615 evolutionary relationships between the proteins from the unicellular organisms and metazoans
616 with the same model as described earlier. Ancestral sequence reconstruction was performed
617 for this data set using rooting option and defining marginal ancestral states option by RAxML
618 with the model PROTGAMMALG.
619 MEGA (Kumar, Stecher and Tamura, 2016) software was used for an alternative tree
620 reconstruction for neighbor-joining and parsimony algorithms with the 500 bootstrap
621 replicates and LG amino acid substitution matrix. Bayesian inference method was also applied
622 on MrBayes tool (Ronquist *et al.*, 2012) under LG matrix. Six chains were run for 3000000

623 generations, every 1000 generations trees were sampled in two runs. The first 25% of trees
624 were discarded before constructing a consensus tree.
625
626 *Protein sequence analysis*
627 Tertiary structure of polypeptides was predicted by PEP-FOLD3 (Lamiable *et al.*, 2016).
628 Estimation of proteins' biochemical and biophysical characteristics from their amino acid
629 sequences was done with ExPASy ProtParam (Gasteiger *et al.*, 2005). Conserved amino acid
630 motifs were analyzed throughout major metazoan lineages using MEME suite (Bailey *et al.*,
631 2009) on the selected set of protein sequences (see Supplementary, Table S2).
632
633 *Testing for positive selection in vertebrate ezrin and merlin*
634 Codeml tool from PAML software (Yang, 2007) was run under model = 1 to test whether ratio
635 of nonsynonymous-synonymous substitutions varies among tree branches. Twice difference
636 between log likelihood of the alternative and the null (the ratio is the same for all branches)
637 hypotheses were compared to χ^2 distribution. Several lineages then were tested for positive
638 selection: model = 2, NSsites = 2, fix_omega = 0, omega = 1. The null hypothesis in this case was
639 estimated under parameters model = 2, NSsites = 2, fix_omega = 1, omega = 1. As well, χ^2
640 distribution test for statistical support was used. Each codeml estimation of alternative
641 hypothesis was run twice with varying parameter 'omega' to test for contingency. Parameter
642 'cleandata' was set to 0 allowing for retaining gaps in alignment.
643 PAL2NAL tool (Suyama, Torrents and Bork, 2006) was used to generated nucleotide
644 alignments based on amino acid multiple sequence alignments and the corresponding mRNA
645 sequences. The mRNA sequences were extracted from NCBI data base by using efetch command
646 from E-utilities suit (Sayers, 2018). The phylogenetic trees for ezrin and merlin protein set from
647 vertebrate animals were built as described earlier and can be found in Supplementary file.
648
649 *Synteny analysis*
650 A BLASTp strategy was applied to analyze syntenic relationships between genes coding for the
651 ERM and ERM-like proteins in different species. Sequences of the proteins encoded by the genes
652 surrounding ezrin (seven proteins), radixin (three proteins), and moesin (four proteins) from
653 *H. sapiens* within the region of 1 Mb were used as query in BLASTp search against all proteins
654 in the following species: *M. musculus*, *D. rerio*, *D. melanogaster*, *S. pistillata* (cnidarians), *T.*
655 *adhaerens*, and *A. queeslandica* with expect threshold of 1e-10. The analysis was repeated with

656 the similarly selected proteins from *D. melanogaster* as a query. Intraspecies synteny analysis
657 for *H. sapiens* was done in a similar manner, but with only human proteins in the target, with
658 expected threshold of 1e-10.
659 For synteny analysis of the merlin gene surrounding, two species with both merlin paralogs
660 were chosen: *D. rerio* and *G. gallus.* All proteins, which genes are located within 1Mb of merlin1
661 or merlin2 genes, were taken as queries. BLASTp searches with expected threshold 1e-10 were
662 run against all proteins from the corresponding chromosomes: chromosomes 5 (has gene
663 coding for merlin2 protein) and 21 (merlin1) for D. rerio, and chromosomes 15 (merlin2) and
664 19 (merlin1). Synteny Database (Catchen, Conery and Postlethwait, 2009) was used for
665 processing the synteny data for *H. sapiens* and *D. rerio*.
666
667 Custom python, perl and bash scripts were used for data processing.
668
669 <u>Acknowledgments</u>
670
671 This work was supported by the Institute of Bioinformatics, University of Muenster. The open
672 access publishing is supported by the University of Muenster.
673
674 <u>Competing interests</u>
675
676 The authors declare no competing interests.
677
678 All authors have read and approved the final version of the manuscript.
679
680
681 References:
682 Adl, S. M. *et al.* (2012) 'The Revised Classification of Eukaryotes Eukaryotic Microbiology', *J.*
683 *Eukaryot. Microbiol*, 59(5), pp. 429–493. doi: 10.1111/j.1550-7408.2012.00644.x.
684 Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database
685 search programs.', *Nucleic acids research*, 25(17), pp. 3389–402. Available at:
686 http://www.ncbi.nlm.nih.gov/pubmed/9254694 (Accessed: 23 July 2017).
687 Bailey, T. L. *et al.* (2009) 'MEME SUITE: tools for motif discovery and searching.', *Nucleic acids*
688 *research*. Oxford University Press, 37(Web Server issue), pp. W202-8. doi: 10.1093/nar/gkp335.
689 Bosanquet, D. C. *et al.* (2014) 'FERM family proteins and their importance in cellular movements
690 and wound healing (Review)', *International Journal of Molecular Medicine*, 34(1), pp. 3–12. doi:

691 10.3892/ijmm.2014.1775.
692 Brault, E. *et al.* (2001) 'Normal membrane localization and actin association of the NF2 tumor
693 suppressor protein are dependent on folding of its N-terminal domain.', *Journal of cell science*, 114(Pt
694 10), pp. 1901–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11329377 (Accessed: 22 April
695 2019).
696 Bretscher, A., Edwards, K. and Fehon, R. G. (2002) 'ERM proteins and merlin: integrators at the cell
697 cortex.', *Nature reviews. Molecular cell biology*, 3(8), pp. 586–99. doi: 10.1038/nrm882.
698 Catchen, J. M., Conery, J. S. and Postlethwait, J. H. (2009) 'Automated identification of conserved
699 synteny after whole-genome duplication', *Genome Research*, 19(8), pp. 1497–1505. doi:
700 10.1101/gr.090480.108.
701 Chen, S.-D. *et al.* (2012) 'Knockdown of radixin by RNA interference suppresses the growth of
702 human pancreatic cancer cells in vitro and in vivo.', *Asian Pacific journal of cancer prevention : APJCP*,
703 13(3), pp. 753–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22631643 (Accessed: 26
704 January 2019).
705 Clucas, J. and Valderrama, F. (2014) 'ERM proteins in cancer progression', *Journal of Cell Science*,
706 127(2), pp. 267–275. doi: 10.1242/jcs.133108.
707 Cooper, J. and Giancotti, F. G. (2014) 'Molecular insights into NF2/Merlin tumor suppressor
708 function.', *FEBS letters*. NIH Public Access, 588(16), pp. 2743–52. doi: 10.1016/j.febslet.2014.04.001.
709 Curto, M. *et al.* (2007) 'Contact-dependent inhibition of EGFR signaling by Nf2/Merlin', *The Journal*
710 *of Cell Biology*, 177(5), pp. 893–903. doi: 10.1083/jcb.200703010.
711 Darriba, D. *et al.* (2011) 'ProtTest 3: fast selection of best-fit models of protein evolution',
712 *Bioinformatics*, 27(8), pp. 1164–1165. doi: 10.1093/bioinformatics/btr088.
713 Dayhoff, M. O. (1965) *No Title*. Silver Spring: National Biomedical research Foundation.
714 Debnath, J. and Brugge, J. S. (2005) 'Modelling glandular epithelial cancers in three-dimensional
715 cultures', *Nature Reviews Cancer*, 5(9), pp. 675–688. doi: 10.1038/nrc1695.
716 Dransfield, D. T. *et al.* (1997) 'Ezrin is a cyclic AMP-dependent protein kinase anchoring protein.',
717 *The EMBO journal*. European Molecular Biology Organization, 16(1), pp. 35–43. doi:
718 10.1093/emboj/16.1.35.
719 Eisen, J. A. (1998) 'Phylogenomics: improving functional predictions for uncharacterized genes by
720 evolutionary analysis.', *Genome research*, 8(3), pp. 163–7. Available at:
721 http://www.ncbi.nlm.nih.gov/pubmed/9521918 (Accessed: 30 April 2019).
722 Estecha, A. *et al.* (2009) 'Moesin orchestrates cortical polarity of melanoma tumour cells to initiate
723 3D invasion.', *Journal of cell science*, 122(Pt 19), pp. 3492–501. doi: 10.1242/jcs.053157.
724 Fairclough, S. R. *et al.* (2013) 'Premetazoan genome evolution and the regulation of cell
725 differentiation in the choanoflagellate Salpingoeca rosetta', *Genome Biology*. BioMed Central, 14(2), p.
726 R15. doi: 10.1186/gb-2013-14-2-r15.
727 Federhen, S. (2012) 'The NCBI Taxonomy database.', *Nucleic acids research*. Oxford University Press,
728 40(Database issue), pp. D136-43. doi: 10.1093/nar/gkr1178.
729 Fievet, B. T. *et al.* (2004) 'Phosphoinositide binding and phosphorylation act sequentially in the
730 activation mechanism of ezrin.', *The Journal of cell biology*, 164(5), pp. 653–9. doi:

731 10.1083/jcb.200307032.
732 Finn, R. D. *et al.* (2006) 'Pfam: clans, web tools and services', *Nucleic Acids Research*. Oxford
733 University Press, 34(90001), pp. D247–D251. doi: 10.1093/nar/gkj149.
734 Funayama, N. *et al.* (1991) 'Radixin is a novel member of the band 4.1 family.', *The Journal of cell*
735 *biology*, 115(4), pp. 1039–48. Available at:
736 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2289953&tool=pmcentrez&rendertype=
737 abstract (Accessed: 15 August 2015).
738 Gabriel, W. N. *et al.* (2007) 'The tardigrade Hypsibius dujardini, a new model for studying the
739 evolution of development', *Developmental Biology*. Academic Press, 312(2), pp. 545–559. doi:
740 10.1016/J.YDBIO.2007.09.055.
741 Gasteiger, E. *et al.* (2005) 'Protein Identification and Analysis Tools on the ExPASy Server', in *The*
742 *Proteomics Protocols Handbook*. Totowa, NJ: Humana Press, pp. 571–607. doi: 10.1385/1-59259-890-
743 0:571.
744 Geneious (2019) *Geneious Prime | Sequence Analysis Software | Academic & Government*, *2019*.
745 Available at: https://www.geneious.com/academic/ (Accessed: 22 April 2019).
746 Glasauer, S. M. K. and Neuhauss, S. C. F. (2014) 'Whole-genome duplication in teleost fishes and its
747 evolutionary consequences', *Molecular Genetics and Genomics*, 289(6), pp. 1045–1060. doi:
748 10.1007/s00438-014-0889-2.
749 Golovnina, K. *et al.* (2005) 'Evolution and origin of merlin, the product of the Neurofibromatosis
750 type 2 (NF2) tumor-suppressor gene.', *BMC Evolutionary Biology*, 5(1), p. 69. doi: 10.1186/1471-2148-
751 5-69.
752 Grau-Bové, X. *et al.* (2017) 'Dynamics of genomic innovation in the unicellular ancestry of animals',
753 *eLife*, 6. doi: 10.7554/eLife.26036.
754 GTExPortal (2019) *GTEx Portal*. Available at: https://gtexportal.org/home/ (Accessed: 4 December
755 2018).
756 Ivetic, A. and Ridley, A. J. (2004) 'Ezrin/radixin/moesin proteins and Rho GTPase signalling in
757 leucocytes.', *Immunology*, 112(2), pp. 165–76. doi: 10.1111/j.1365-2567.2004.01882.x.
758 Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*,
759 30(9), pp. 1236–1240. doi: 10.1093/bioinformatics/btu031.
760 Katoh, K. *et al.* (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast
761 Fourier transform', *Nucleic Acids Research*. Oxford University Press, 30(14), pp. 3059–3066. doi:
762 10.1093/nar/gkf436.
763 Kikuchi, S. *et al.* (2002) 'Radixin deficiency causes conjugated hyperbilirubinemia with loss of Mrp2
764 from bile canalicular membranes.', *Nature genetics*, 31(3), pp. 320–5. doi: 10.1038/ng905.
765 Kitajiri, S. *et al.* (2004) 'Radixin deficiency causes deafness associated with progressive
766 degeneration of cochlear stereocilia.', *The Journal of cell biology*, 166(4), pp. 559–70. doi:
767 10.1083/jcb.200402007.
768 Kobayashi, H. *et al.* (2004) 'Clinical significance of cellular distribution of moesin in patients with
769 oral squamous cell carcinoma.', *Clinical cancer research : an official journal of the American Association*
770 *for Cancer Research*, 10(2), pp. 572–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14760079

771 (Accessed: 26 January 2019).
772 Kumar, S., Stecher, G. and Tamura, K. (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis
773 Version 7.0 for Bigger Datasets', *Molecular Biology and Evolution*, 33(7), pp. 1870–1874. doi:
774 10.1093/molbev/msw054.
775 Lallemand, D., Saint-Amaux, A. L. and Giovannini, M. (2009) 'Tumor-suppression functions of merlin
776 are independent of its role as an organizer of the actin cytoskeleton in Schwann cells', *Journal of Cell*
777 *Science*, 122(22), pp. 4141–4149. doi: 10.1242/jcs.045914.
778 Lamiable, A. *et al.* (2016) 'PEP-FOLD3: faster *de novo* structure prediction for linear peptides in
779 solution and in complex', *Nucleic Acids Research*, 44(W1), pp. W449–W454. doi: 10.1093/nar/gkw329.
780 Lankes, W. T. and Furthmayr, H. (1991) 'Moesin: a member of the protein 4.1-talin-ezrin family of
781 proteins.', *Proceedings of the National Academy of Sciences of the United States of America*, 88(19), pp.
782 8297–301. Available at:
783 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=52495&tool=pmcentrez&rendertype=ab
784 stract (Accessed: 15 August 2015).
785 Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, 23(21), pp. 2947–
786 2948. doi: 10.1093/bioinformatics/btm404.
787 Le, S. Q. and Gascuel, O. (2008) 'An Improved General Amino Acid Replacement Matrix', *Molecular*
788 *Biology and Evolution*. Oxford University Press, 25(7), pp. 1307–1320. doi: 10.1093/molbev/msn067.
789 Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree
790 display and annotation'. Available at: http://itol.embl.de/help/17050570.pdf (Accessed: 11 May
791 2017).
792 Li, Q. *et al.* (2007) 'Self-masking in an Intact ERM-merlin Protein: An Active Role for the Central α-
793 Helical Domain', *Journal of Molecular Biology*, 365(5), pp. 1446–1459. doi: 10.1016/j.jmb.2006.10.075.
794 Li, Q. *et al.* (2015) 'Ezrin/Exocyst Complex Regulates Mucin 5AC Secretion Induced by Neutrophil
795 Elastase in Human Airway Epithelial Cells', *Cellular Physiology and Biochemistry*, 35(1), pp. 326–338.
796 doi: 10.1159/000369699.
797 Li, W. and Crouch, D. H. (2000) 'Cloning and expression profile of chicken radixin.', *Biochimica et*
798 *biophysica acta*, 1491(1–3), pp. 327–32. Available at:
799 http://www.ncbi.nlm.nih.gov/pubmed/10760599 (Accessed: 16 December 2018).
800 Liu, X. *et al.* (2015) 'Moesin and myosin phosphatase confine neutrophil orientation in a
801 chemotactic gradient.', *The Journal of experimental medicine*. Rockefeller University Press, 212(2), pp.
802 267–80. doi: 10.1084/jem.20140508.
803 Lu, T.-M. *et al.* (2017) 'The phylogenetic position of dicyemid mesozoans offers insights into
804 spiralian evolution', *Zoological Letters*, 3(1), p. 6. doi: 10.1186/s40851-017-0068-5.
805 Marchler-Bauer, A. *et al.* (2015) 'CDD: NCBI's conserved domain database', *Nucleic Acids Research*,
806 43(D1), pp. D222–D226. doi: 10.1093/nar/gku1221.
807 Marion, S. *et al.* (2011) 'Ezrin promotes actin assembly at the phagosome membrane and regulates
808 phago-lysosomal fusion.', *Traffic (Copenhagen, Denmark)*, 12(4), pp. 421–37. doi: 10.1111/j.1600-
809 0854.2011.01158.x.
810 McClatchey, a. I. (2014) 'ERM proteins at a glance', *Journal of Cell Science*, (June), pp. 1–6. doi:

811 10.1242/jcs.098343.
812 Nambiar, R., McConnell, R. E. and Tyska, M. J. (2010) 'Myosin motor function: the ins and outs of
813 actin-based membrane protrusions', *Cellular and Molecular Life Sciences*, 67(8), pp. 1239–1254. doi:
814 10.1007/s00018-009-0254-5.
815 Nei, M. and Rooney, A. P. (2005) 'Concerted and birth-and-death evolution of multigene families.',
816 *Annual review of genetics*. NIH Public Access, 39, pp. 121–52. doi:
817 10.1146/annurev.genet.39.073003.112240.
818 Niggli, V. and Rossy, J. (2008) 'Ezrin/radixin/moesin: Versatile controllers of signaling molecules
819 and of the cortical cytoskeleton', *International Journal of Biochemistry and Cell Biology*, 40(3), pp. 344–
820 349. doi: 10.1016/j.biocel.2007.02.012.
821 Omelyanchuk, L. V *et al.* (2009) 'Evolution and origin of HRS, a protein interacting with Merlin, the
822 Neurofibromatosis 2 gene product.', *Gene regulation and systems biology*, 3, pp. 143–57. Available at:
823 http://www.ncbi.nlm.nih.gov/pubmed/20054405 (Accessed: 20 November 2018).
824 Pataky, F., Pironkova, R. and Hudspeth, A. J. (2004) 'Radixin is a constituent of stereocilia in hair
825 cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 101(8), pp.
826 2601–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14983055 (Accessed: 16 December
827 2018).
828 Van de Peer, Y., Maere, S. and Meyer, A. (2009) 'The evolutionary significance of ancient genome
829 duplications.', *Nature reviews. Genetics*, 10(10), pp. 725–32. doi: 10.1038/nrg2600.
830 Phang, J. M. *et al.* (2016) 'Structural characterization suggests models for monomeric and dimeric
831 forms of full-length ezrin.', *The Biochemical journal*. Portland Press Limited, 473(18), pp. 2763–82. doi:
832 10.1042/BCJ20160541.
833 Pujuguet, P. *et al.* (2003) 'Ezrin Regulates E-Cadherin-dependent Adherens Junction Assembly
834 through Rac1 Activation', *Molecular Biology of the Cell*, 14(5), pp. 2181–2191. doi: 10.1091/mbc.e02-
835 07-0410.
836 Rambaut A. (2019) *FigTree*, *2009*. Available at: http://tree.bio.ed.ac.uk/software/figtree/
837 (Accessed: 18 April 2019).
838 Richter, D. J. and King, N. (2013) 'The Genomic and Cellular Foundations of Animal Origins', *Annual*
839 *Review of Genetics*. Annual Reviews , 47(1), pp. 509–537. doi: 10.1146/annurev-genet-111212-
840 133456.
841 Ronquist, F. *et al.* (2012) 'MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice
842 Across a Large Model Space', *Systematic Biology*, 61(3), pp. 539–542. doi: 10.1093/sysbio/sys029.
843 Rost, B. and Sander, C. (1994) 'Combining evolutionary information and neural networks to predict
844 protein secondary structure', *Proteins: Structure, Function, and Genetics*. John Wiley & Sons, Ltd, 19(1),
845 pp. 55–72. doi: 10.1002/prot.340190108.
846 Roy, C., Martin, M. and Mangeat, P. (1997) 'A dual involvement of the amino-terminal domain of
847 ezrin in F- and G-actin binding.', *The Journal of biological chemistry*, 272(32), pp. 20088–95. Available
848 at: http://www.ncbi.nlm.nih.gov/pubmed/9242682 (Accessed: 22 September 2015).
849 Saotome, I., Curto, M. and McClatchey, A. I. (2004) 'Ezrin is essential for epithelial organization and
850 villus morphogenesis in the developing intestine.', *Developmental cell*, 6(6), pp. 855–64. doi:

851 10.1016/j.devcel.2004.05.007.
852 Sayers, E. (2018) 'The E-utilities In-Depth: Parameters, Syntax and More'. National Center for
853 Biotechnology Information (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK25499/
854 (Accessed: 30 June 2019).
855 Schliep, K. *et al.* (2017) 'Intertwining phylogenetic trees and networks', *Methods in Ecology and*
856 *Evolution*. Edited by R. Fitzjohn. doi: 10.1111/2041-210X.12760.
857 Sebé-Pedrós, A. *et al.* (2013) 'Insights into the Origin of Metazoan Filopodia and Microvilli',
858 *Molecular Biology and Evolution*. Oxford University Press, 30(9), pp. 2013–2023. doi:
859 10.1093/molbev/mst110.
860 Shalchian-Tabrizi, K. *et al.* (2008) 'Multigene Phylogeny of Choanozoa and the Origin of Animals',
861 *PLoS ONE*. Edited by R. Aramayo, 3(5), p. e2098. doi: 10.1371/journal.pone.0002098.
862 Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
863 phylogenies.', *Bioinformatics (Oxford, England)*, 30(9), pp. 1312–3. doi:
864 10.1093/bioinformatics/btu033.
865 Stamatakis, A. (2015) 'Using RAxML to Infer Phylogenies', in *Current Protocols in Bioinformatics*.
866 Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 6.14.1-6.14.14. doi: 10.1002/0471250953.bi0614s51.
867 Stickney, J. T. *et al.* (2004) 'Activation of the tumor suppressor merlin modulates its interaction with
868 lipid rafts.', *Cancer research*. American Association for Cancer Research, 64(8), pp. 2717–24. doi:
869 10.1158/0008-5472.CAN-03-3798.
870 Suga, H. *et al.* (2013) 'The Capsaspora genome reveals a complex unicellular prehistory of animals',
871 *Nature Communications*. Nature Publishing Group, 4(1), p. 2325. doi: 10.1038/ncomms3325.
872 Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: robust conversion of protein sequence
873 alignments into the corresponding codon alignments.', *Nucleic acids research*. Oxford University Press,
874 34(Web Server issue), pp. W609-12. doi: 10.1093/nar/gkl315.
875 Turunen, O. *et al.* (1998) 'Structure-function relationships in the ezrin family and the effect of
876 tumor-associated point mutations in neurofibromatosis 2 protein.', *Biochimica et biophysica acta*,
877 1387(1–2), pp. 1–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9748471 (Accessed: 8
878 October 2017).
879 Valderrama, F., Thevapala, S. and Ridley, A. J. (2012) 'Radixin regulates cell migration and cell-cell
880 adhesion through Rac1', *Journal of Cell Science*, 125(14), pp. 3310–3319. doi: 10.1242/jcs.094383.
881 Winckler, B. *et al.* (1994) 'Analysis of a cortical cytoskeletal structure: a role for ezrin-radixin-
882 moesin (ERM proteins) in the marginal band of chicken erythrocytes.', *Journal of cell science*, 107 (Pt
883 9), pp. 2523–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7531201 (Accessed: 5 May
884 2019).
885 Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and*
886 *Evolution*. Narnia, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.
887 Yonemura, S. *et al.* (2002) 'Rho-dependent and -independent activation mechanisms of
888 ezrin/radixin/moesin proteins: an essential role for polyphosphoinositides in vivo.', *Journal of cell*
889 *science*, 115(Pt 12), pp. 2569–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12045227
890 (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.