RESEARCH ARTICLE

Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway

Gábor Erdős[1☯], **Tamás Szaniszló**[1☯], **Mátyás Pajkos**[1], **Borbála Hajdu-Soltész**[1], **Bence Kiss**[2], **Gábor Pál**[2], **László Nyitray**[2], **Zsuzsanna Dosztányi**[1]*

1 MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary, **2** Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary

☯ These authors contributed equally to this work.
* dosztanyi@caesar.elte.hu

OPEN ACCESS

Abstract

Protein-protein interactions (PPIs) formed between short linear motifs and globular domains play important roles in many regulatory and signaling processes but are highly underrepresented in current protein-protein interaction databases. These types of interactions are usually characterized by a specific binding motif that captures the key amino acids shared among the interaction partners. However, the computational proteome-level identification of interaction partners based on the known motif is hindered by the huge number of randomly occurring matches from which biologically relevant motif hits need to be extracted. In this work, we established a novel bioinformatic filtering protocol to efficiently explore interaction network of a hub protein. We introduced a novel measure that enabled the optimization of the elements and parameter settings of the pipeline which was built from multiple sequence-based prediction methods. In addition, data collected from PPI databases and evolutionary analyses were also incorporated to further increase the biological relevance of the identified motif hits. The approach was applied to the dynein light chain LC8, a ubiquitous eukaryotic hub protein that has been suggested to be involved in motor-related functions as well as promoting the dimerization of various proteins by recognizing linear motifs in its partners. From the list of putative binding motifs collected by our protocol, several novel peptides were experimentally verified to bind LC8. Altogether 71 potential new motif instances were identified. The expanded list of LC8 binding partners revealed the evolutionary plasticity of binding partners despite the highly conserved binding interface. In addition, it also highlighted a novel, conserved function of LC8 in the upstream regulation of the Hippo signaling pathway. Beyond the LC8 system, our work also provides general guidelines that can be applied to explore the interaction network of other linear motif binding proteins or protein domains.

Author summary

Fine-tuning of many cellular processes relies on weak, transient protein-protein interactions. Such interactions often involve compact functional modules, called short linear

motifs (SLiMs) that can bind to specific globular domains. SLiM-mediated interactions can carry out diverse molecular functions by targeting proteins to specific cellular locations, regulating the activity and binding preferences of proteins, or aiding the assembly of macromolecular complexes. The key to the function of SLiMs is their small size and highly flexible nature. At the same time, these properties make their experimental identification challenging. Consequently, only a small portion of SLiM-mediated interactions is currently known. This underlies the importance of novel computational methods that can reliably identify candidate sites involved in binding to linear motif binding domains. Here we present a novel bioinformatic approach that efficiently predicts new binding partners for SLiM-binding domains. We applied this method to the dynein light chain LC8, a protein that was already known to bind many partners in a wide range of organisms. With this method, we not only significantly expanded the interaction network of LC8, but also identified a novel function of LC8 in a highly important pathway controlling organ size in animals.

Introduction

A large number of protein-protein interactions (PPIs) are mediated by short linear motifs (SLiMs) that are recognized by specific globular domains [1]. SLiM-mediated interactions are involved in a wide range of biological functions and can regulate the formation of transient protein complexes, orchestrate subcellular localization, modulate post-translational modification state, and determine the fate of proteins [1]. Such interactions emerged as key mediators of complex regulatory processes in higher eukaryotic cells and their aberrant functioning can contribute to various diseases as well [2]. The key to the essential nature of SLiMs in biological systems lies in their specific properties. SLiMs correspond to a stretch of approximately 3–10 residues that generally reside within intrinsically disordered regions (IDRs). As a result, they usually form transient, weak interactions with micromolar binding affinity [3]. Due to these specific properties, the identification of linear motif sites is challenging both experimentally and computationally [4]. Currently, the most comprehensive collection, the Eukaryotic Linear Motif database (ELM) holds only 200–300 motif patterns with a few thousands of experimentally verified instances [5]. This number pales in comparison to the expected number of linear motif mediated interactions in the human proteome, estimated to number at least several hundred thousand [6]. Linear binding peptides have been systematically analyzed only for a few specific interaction domains, such as SH2, PTB, 14-3-3, PDZ and SH3 domains [7,8]. However, for most motif binding domains, the interaction network is largely incomplete.

The identification of linear motif mediated interactions is usually divided into two phases. The first phase is the characterization of the common consensus sequence motif that is shared among the diverse set of binding partners of a common domain, i.e. defining the motif class [9,10]. The core motif that mediates interactions with a given domain is usually represented by a sequence pattern or a position specific scoring matrix (PSSM). In the second phase, the core motif is used to identify additional candidate binding sites in the proteome, i.e. finding novel motif instances. However, as the information content of consensus motifs is usually low, predicted motif matches are overwhelmingly dominated by false positive matches that occur purely by chance [3,11]. Therefore, this phase involves additional filtering steps to remove matches that are unlikely to be functional and to prioritize motif hits for further experimental characterization. Various computational tools such as ELM, QuasiMotifFinder, MiniMotifMiner, SLiMSearch, ScanProsite or DOReMi [5,12–16] have been developed to overcome this

problem. These tools scan a defined set of proteins with a single consensus motif and utilize various discriminatory attributes to prioritize motif hits, including structural context, protein disorder, functional ontology, evidence for PPIs and shared cellular localization. Evolutionary conservation can highlight functionally relevant positions in proteins and have been used for globular domains to identify conserved motif-like patterns that are indicative of the function of a protein [12]. However, linear motif sites generally reside within IDRs that are generally less conserved [17]. Within these regions, SLiMs often show a specific pattern of evolutionary conservation that are characterized by a higher relative conservation of the key motif residues compared to their flanking regions [18,19], and this information can be used to highlight true binding motifs. Using various sequence attributes, specific filtering pipelines were constructed to identify novel motif instances for several domains [20–24]. However, the strength and opti-mality of the filtering steps have never been systematically tested. In order to build optimal fil-tering protocols, a more systematic approach is needed that can take into account the specific trade-off between reducing false positive hits while capturing biologically relevant motif matches, which is likely to be specific to individual binding domains.

In this study, we focused on LC8 dynein light chain and its binding partners as a case study, and explored how the interaction network of a specific linear motif binding protein can be expanded in an optimal way. LC8 is a remarkably conserved eukaryotic hub protein [22]. Although LC8 was originally suggested to function as a cargo adaptor for the dynein motor complex, its extensive interaction network suggests a more general role, independent of dynein [22,25]. Recently, the prevailing view has become that LC8 functions as a dimerization or olig-omerization engine for various proteins [25,26]. Known interaction partners link LC8 to pro-cesses such as nuclear transport, tumor suppression, viral replication, DNA damage repair, apoptosis, mitosis and signaling [22,25]. In contrast to their functional heterogeneity, LC8 binding partners generally share a common binding mechanism [22,27]. The known struc-tures of complexes between LC8 and various bound partners show that the binding groove is formed at the dimerization interface of the homodimeric LC8, favoring binding partners that are also dimerized [28,29], often promoted by coiled coil (CC) regions [20,30]. The recognition SLiMs are generally located within intrinsically disordered regions [25] and undergo a disor-der-to-order transition upon complex formation. In their bound form, these segments adopt highly similar conformations that augment the central beta sheet of LC8 on each side of the dimer [27]. Many of the binding segments contain a Thr-Gln-Thr (TQT) motif with additional positions showing larger variations. Apart from the canonical TQT motif, there are also non-canonical binding motif instances, in which the central Gln is replaced by Met or Asn [22]. An even more unusual Thr-Ser-Pro (TSP) binding motif mediates the interaction between Pak1 and LC8. Altogether, more than 50 LC8 binding motif instances were collected from various eukaryotic species [22]. To reveal the optimal binding motif, a phage display study was also carried out, identifying further LC8 binding motif instances [20]. The suggested general role for LC8 raises the possibility of many additional binding partners in the human proteome.

In this work we expanded the interaction network of dynein light chain LC8 using a combi-nation of computational and experimental methods. We introduced a novel measure based on information gain that enabled us to build an optimal bioinformatic pipeline by combining var-ious attributes predicted from the amino acid sequence. We also incorporated known binding partners from PPI databases and exploited the specific evolutionary conservation of binding motifs to increase the likelihood that the motif hit is biologically relevant. The resulting proce-dure enabled us to drastically reduce false positive predictions among putative novel linear motif instances and to expand the interaction network of LC8 with high-confidence predic-tions. We experimentally verified the binding of several novel motif instances to LC8 using surface plasmon resonance (SPR) assay. One of the most interesting outcomes of the extended

interaction network of LC8 revealed a possible new function of the LC8 protein in the Hippo pathway through interaction with WWC and AMOT protein family members. The presented study significantly contributes to the better understanding of the functional and evolutionary properties of the LC8 interactome. Beyond this specific hub protein, it also offers general guidelines for the exploration of additional linear motif-mediated interaction networks.

Results

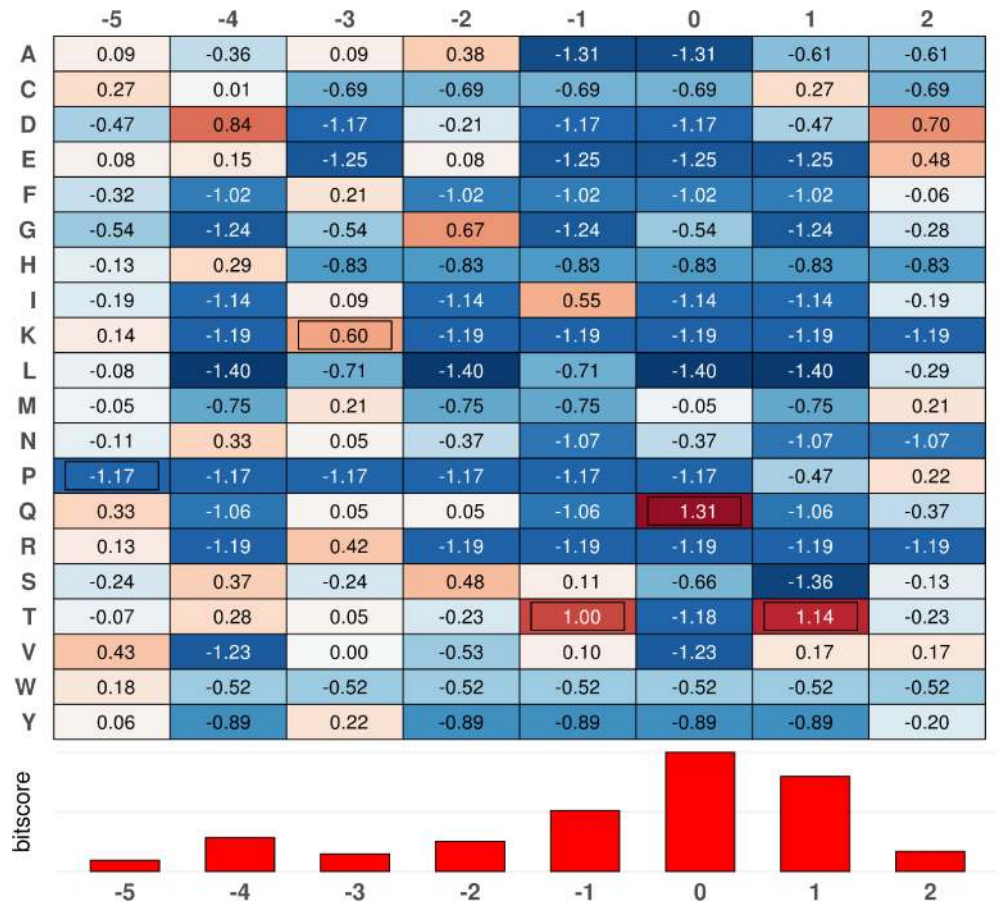### Scoring peptide segments based on known motifs

LC8 recognizes a short linear motif in its partner proteins. In this case, the binding involves both polypeptide chains of the homodimer LC8. However, to emphasize the similar binding mode to many single modular domains that also bind short linear motifs, LC8 is also referred to as a "binding domain" in this article. We assembled a manually curated database of LC8 interaction partners based on literature search, in which LC8 binding was verified at the motif level (see Materials and Methods and S1 Table). The final dataset contained 53 partners with 67 motif instances and covered multiple eukaryotic species and viruses. 40 motifs in 33 proteins belonged to human or could be directly mapped to a human protein based on close homology. The length of the core binding motif was taken as 8 amino acids (S1 Fig; for details, see S1 Text).

The ELM database describes the LC8 binding motif using the regular expression "^P.K. TQT". From the 67 known motif instances, only 13 matched this regular expression, and 24 further motifs contained only the canonical "TQT" motif core. Considering human partners only, the corresponding numbers were 7 and 17, respectively. These numbers indicate that the definition in the ELM database is too restrictive to capture the majority of known instances. To better describe the common sequential properties of known LC8 binding sites, we used a position specific scoring matrix.

The calculated PSSM is shown in Fig 1. Positive scores indicate amino acid residues that are favored at a given position. The PSSM clearly captures the frequent occurrence of the canonical "TQT" motif. However, with the exception of Gln in position 0, there are additional favored amino acids in every position. The strong preference for Lys at position -3 from the central Gln is not supported by the current collection of known partners, as neighboring positions have stronger preferences according to the bitscore. Using the obtained PSSM, we scanned the whole human proteome to score every overlapping eight amino acid long peptide segment (S2 Fig). As expected, all known LC8 binding peptides had a positive score. While only 2% of human peptides had a positive score, these still represented more than one hundred thousand cases. This indicates that on the one hand, positive PSSM scores are strongly associated with true binding motifs and can serve as a valid starting point for novel LC8-binding motif discovery. On the other hand, the very large number of initial motif hits underscore the need for that additional filtering steps.

Establishing sequence features for filtering true binding partners

In order to establish additional filtering criteria, we gathered various predicted features of the PSSM-identified peptides. The methods we used included PFAM annotations [31], average disorder prediction scores (using IUPred, PONDR VSL2, Espritz and DISOPRED3) [32–35], average score to be part of disordered binding regions (using ANCHOR, MORF-CHIBI and DISOPRED3-BR) [35–37], and secondary structure prediction scores using PSIPRED [38]. At the protein level, information about predicted cellular localization [20,39] and the presence of coiled coil regions was also collected (See Materials and Methods). The various data for all

	-5	-4	-3	-2	-1	0	1	2
A	0.09	-0.36	0.09	0.38	-1.31	-1.31	-0.61	-0.61
C	0.27	0.01	-0.69	-0.69	-0.69	-0.69	0.27	-0.69
D	-0.47	0.84	-1.17	-0.21	-1.17	-1.17	-0.47	0.70
E	0.08	0.15	-1.25	0.08	-1.25	-1.25	-1.25	0.48
F	-0.32	-1.02	0.21	-1.02	-1.02	-1.02	-1.02	-0.06
G	-0.54	-1.24	-0.54	0.67	-1.24	-0.54	-1.24	-0.28
H	-0.13	0.29	-0.83	-0.83	-0.83	-0.83	-0.83	-0.83
I	-0.19	-1.14	0.09	-1.14	0.55	-1.14	-1.14	-0.19
K	0.14	-1.19	0.60	-1.19	-1.19	-1.19	-1.19	-1.19
L	-0.08	-1.40	-0.71	-1.40	-0.71	-1.40	-1.40	-0.29
M	-0.05	-0.75	0.21	-0.75	-0.75	-0.05	-0.75	0.21
N	-0.11	0.33	0.05	-0.37	-1.07	-0.37	-1.07	-1.07
P	-1.17	-1.17	-1.17	-1.17	-1.17	-1.17	-0.47	0.22
Q	0.33	-1.06	0.05	0.05	-1.06	1.31	-1.06	-0.37
R	0.13	-1.19	0.42	-1.19	-1.19	-1.19	-1.19	-1.19
S	-0.24	0.37	-0.24	0.48	0.11	-0.66	-1.36	-0.13
T	-0.07	0.28	0.05	-0.23	1.00	-1.18	1.14	-0.23
V	0.43	-1.23	0.00	-0.53	0.10	-1.23	0.17	0.17
W	0.18	-0.52	-0.52	-0.52	-0.52	-0.52	-0.52	-0.52
Y	0.06	-0.89	0.22	-0.89	-0.89	-0.89	-0.89	-0.20

Fig 1. Position specific scoring matrix of LC8 binding motifs. PSSM score of each position of the binding motif numbered from the central glutamine of the canonical "TQT" motif core. Amino acids corresponding to the ELM definition are framed. Values are color scaled as a heat map ranging from blue (negative) to red (positive) centered on zero (white). The Shannon entropy of each column is shown below in bitscore.

known true positive motifs from human and other species are available at http://gerdos.web. elte.hu/data/LC8/known_results.html.

These various features can be associated with known motifs to different extents. The main challenge is to find the best tools and parameter settings that enable the prioritization of peptide segments that are the most likely to be biologically relevant motif hits in an optimal way. We introduced a metric based on weighted information gain derived from the Shannon entropy to globally attest the discriminatory power of each criterion. On a dataset containing 40 known human binding partners and 10,000 random human segments from the proteome with a higher than zero PSSM score, the weighted information gain was calculated for each criterion. The suggested measure enabled not only to rank the attributes in terms of their discriminatory power, but also to choose the best parameter settings. According to this protocol, the strongest criterion based on the information gain was the predicted intracellular localization, which was fulfilled by all known motifs, but only 85.53% of random peptides. The next strongest filtering criterion was based on PFAM annotations. Using annotation strictly based on the domain type PFAM families, it was possible to filter out 15% of random motifs, while retaining all known motifs. The third strongest criterion was based on PSSM scores. The highest information gain was achieved with the cutoff value of 3.3. This cutoff value was not met by

only three known human motifs (PAK1, NRF1, MYZAP). At the next level, disorder prediction methods produced the highest information gain. Four methods were tested with different cutoff values corresponding to different false positive rates. The optimal choice was IUPred with the cutoff value of 0.42, which was fulfilled by all but one known motifs.

We also tested three methods (ANCHOR, MoRFchibi, DISOPRED3) [35–37] for predicting disordered binding regions also known as molecular recognition features (MoRFs). The information gain was much lower, compared to previous criteria. The optimal information gain was reached using the ANCHOR method with a cutoff value of 0.57. However, even this criterion would filter out 57.5% of the positive examples. An additional criterion considered was secondary structure prediction. Based on known structures of the complexes, the binding motif is expected to adopt a beta-strand conformation. Surprisingly, predicted beta-strands occurred in only four out of 40 cases, and even in these cases only very short segments were predicted to be in beta conformation. Helical segments were predicted in five cases, including MYO5A, which has been shown to have helical tendencies in the unbound form [40]. However, the lack of residues predicted to be in helix did not perform well as a filtering criterion as the majority of random hits were also predicted to lack regular secondary structural elements, and the overall information gain was below that of predicting disordered binding regions. Similarly, the presence of coiled coil regions in the partner proteins predicted by NCoils [41] did not have a strong discriminatory power either. As these criteria filtered out many known motifs, they were not included in our filtering protocol.

The resulting filtering protocol is shown on Fig 2, indicating the proportion of random hits and known motifs that were eliminated at each step. By the stepwise application of the four filtering steps, a drastic reduction of random hits could be achieved, while still retaining the majority of known motifs: 90% of known motifs were kept while 99.78% of random hits were filtered out (i.e. only 0.22% was kept). By applying the filtering protocol for the complete human proteome, 335 candidate motifs remained. We carried out a 3-fold cross validation to measure the generality of the obtained filtering protocol (see Materials and Methods). On average, we were able to correctly categorize 34 examples from the 40 experimentally validated human binding motifs. This was only marginally worse compared to using the complete database where 36 motifs were categorized correctly. The cut-off values for the PSSM and IUPred scores also did not change largely. This indicates that the protocol is robust and can correctly describe the general attributes of the binding event between LC8 and its known binding partners.

Evolutionary conservation of motif hits

Given the extreme conservation of LC8, it is of special interest how the interaction motifs are conserved in partner proteins. To study this, we generated multiple sequence alignments of orthologous proteins harboring known LC8 binding motifs and categorized them into 5 taxonomic groups: Mammalia, Vertebrata, Metazoa, Fungi and Eukarya. The conservation at the level of protein and motif was tested based on the obtained alignments in each taxonomic group. The definition of motif conservation applied here depends on the PSSM score of the motif (see Materials and Methods). Consequently, this approach cannot be applied to motifs that significantly differ from canonical motifs, i.e. their binding motif had a PSSM value below the threshold. These three examples were excluded from the analysis. Nevertheless, this strict criterion ensured that not only the general sequence similarity is maintained, but also the similarity to known human LC8 motifs is preserved.

The results of the conservation analysis for the known human motifs are presented in Fig 3. Among the known partners, the LC8 binding sites located within dynein intermediate chains
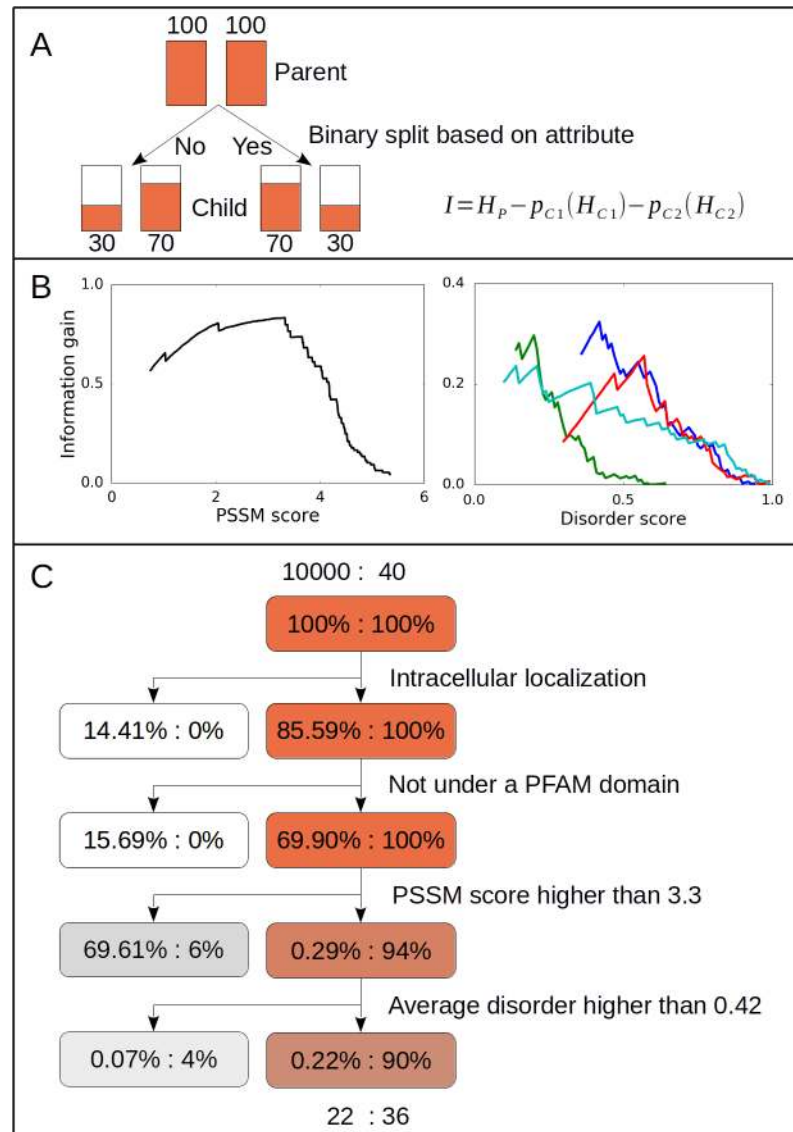
Fig 2. Filtering protocol to find true binding partners. (A) Schematic diagram of the binary filtering protocol we created utilizing information gain. A given attribute provides a binary split of the Parent group into the *C1* and *C2* child groups. The information gain (*I*) is then calculated as the difference of the Shannon entropy (*H*) of the Parent group minus the Shannon entropy of the Children groups weighted by their relative probabilities (*p*). These values were calculated over the dataset containing 40 known human binding partners and 10,000 random human segments from the proteome with a higher than zero PSSM score. (B) The information gain of the PSSM score (left panel) and four disorder prediction methods as a function of different cut-off values (right panel). The disorder prediction method used here were: IUPred (blue), Espritz Disprot (green) and VSL2 (red line), DISOPRED3 (cyan). Optimal cut-off values were obtained from the cut-off value corresponding to the maximum of the information gain, yielding 3.3 for the PSSM score, and 0.42 for IUPred disorder prediction score. (C) The outline of the final filtering protocol indicating the number of elements and percentage of cases in each Child group with the applied binary split.

https://doi.org/10.1371/journal.pcbi.1005885.g002

(DYNC1|1, DYNC1|2) exhibited the most pronounced conservation, spreading across a wide range of eukaryotic species (e.g. starlet sea anemone, *C. elegans*, slime mold). In contrast, other motifs identified in human or other mammalian species were generally not conserved beyond Vertebrata (e.g., BMF, DLGAP1, MTCL1). In the majority of cases, not only the LC8 binding motif but the complete protein was lost beyond this evolutionary distance (e.g. BMF, BSN,

Fig 3

Protein	Position	Mammalia	Vertebrata	Metazoa	Fungi	Eukaryota
DYNC1I1	166	100%	86%	83%	50%	67%
DYNC1I2	156	100%	83%	83%	50%	67%
AMBRA1	1100	100%	57%	0%		0%
ATMIN	668	100%	80%	33%		
BCAS1	566	100%				
BCL2L11	110	100%	50%			
BMF	66	100%	75%			
BSN	1426	100%	20%			
BSN	1502	86%	40%			
BSN	1530	100%	60%			
DAZL	248	100%	67%	0%		
DLGAP1	650	100%	60%	0%		
DLGAP1	675	100%	80%	0%		
EML3	81	83%	25%	0%	0%	
FAM83D	387	100%	25%			
FAM83D	405	60%	25%			
FAM83D	438	100%	75%			
GLCCI1	344	100%	67%			
GPHN	206	100%	60%	0%	0%	0%
MORC3	636	100%	40%	0%		0%

Protein	Position	Mammalia	Vertebrata	Metazoa	Fungi	Eukaryota
MTCL1	1634	100%	80%	0%		
MYO5A	1284	83%	40%	0%	0%	0%
NEK9	943	100%	17%	0%		
NOS1	234	86%	29%	0%		0%
RASGRP3	610	100%	57%	0%		
SNPH	310	100%	67%			
SPAG5	467	100%				
TP53BP1	1150	86%	40%	0%	0%	
TP53BP1	1167	100%	20%	0%	0%	
TRPS1	1205	100%	67%			0%
UHRF1BP1L	1395	100%	33%	0%		0%
WWC1	278	100%	86%	75%		
WWC1	887	100%	71%	75%		
ZMYM3	851	100%	60%	80%		
ZMYND11	451	100%	67%	0%		
ZNF354A	93	100%				
MAP4	800	50%	0%	0%		
MYZAP	422	0%	0%			
NRF1	3	0%	17%	0%		
PAK1	215	0%	0%	0%	0%	0%

Fig 3. Summarized evolutionary conservation results of known human LC8 binding partners. Protein names and motif start positions are indicated in the first and second columns. The colored boxes represent the presence of orthologues of the known partner at different evolutionary levels. The percentages and colour scheme of the boxes show the PSSM based motif conservation across all species. Conservation values increase from white (low motif conservation) to red (high motif conservation).

SNPH). Similarly, limited conservation of LC8 motifs was observed for binding regions experimentally verified in other species (S3 Fig). For example, the EGL protein from *Drosophila melanogaster* had orthologues at each level, but its binding motif showed conservation only in metazoan species. While the centriole duplication functionality is conserved across metazoan species, the LC8 binding protein ANA2 involved in this process is specific to *Drosophila* species [42]. Lowering the PSSM cutoff value used to define the conservation of the motifs perturbed the results only in three cases: NRF1, FAM83D and MYO5A, but had no major impact on the overall trends (S4A and S4B Fig). We can conclude from this analysis of motif conservation that while the LC8 binding interface is highly conserved, known partners and especially the binding motifs located within them are significantly less conserved.

We tested the specific conservation pattern of motif binding sites using the SLiMPrints algorithm [19]. This method identifies short stretches of residues within disordered regions that show high relative conservation compared to their flanking region as calculated from multiple sequence alignments of orthologous sequences. In our dataset, 13 out of the 40 known human motifs exhibited high relative conservation, which is only 32.5% of all cases. This suggests that the approach based on the island-like conservation identified by SLiMPrints has a limited ability to highlight putative functional binding motifs and overall it is not a good filtering criterion, as it misses the majority of functional motifs.

Here, we suggest an alternative filtering criterion based on the observation that known motifs are generally conserved within their own taxonomic group even if they lack motif level conservation over wider evolutionary distances. For the majority of the cases, known motifs in human sequences were conserved in at least 80% of mammalian species. The exceptions included only the three cases that had a PSSM value below our cutoff, and two additional examples (the MAP4 motif and the second motif of FAM83D at position 405). In contrast, randomly chosen motif hits located within disordered regions generally did not possess this property. Applying this discriminatory technique (see Materials and Methods), we could filter out 163 of the possible candidate peptides, reducing the number of candidate motifs to 172. Therefore, for human sequences the conservation within mammalian species can be used as an efficient filter to further reduce likely false positives.

Integrating protein-protein interaction data

The predicted motif hits were also analyzed in relation to experimental data in PPI databases. Known interaction partners of LC8 were collected using the integrative PSICQUIC approach [43], which enabled us to search multiple databases simultaneously and to group the experiments based on detection methods and association types (see Materials and Methods). Altogether, 381 LC8 interaction partners were collected (S2 Table). Several of these interaction partners were supported by multiple lines of evidence, with a total of 782 independent experiments. Known motif partners were well represented in current PPI databases, with only three missing out of these 33 partners. The comparison of interaction partners collected from PPI databases with those that contained known motifs indicated that there is no single criterion that can identify biologically relevant interactions from candidates in PPI databases. However, PPIs from low-throughput and direct experiments are more likely to be biologically relevant, especially when they are supported by multiple independent measurements (see S5 Fig).

Overall, there were a large number of potential partners recorded in PPI databases (323) that contained neither known nor predicted LC8 binding motif instances. By looking at the distribution of partners over the number of supporting experiments, we can see that partners with known and predicted motifs were largely evenly distributed, similarly to the partners detected by direct methods (Fig 4A). Known and predicted motifs dominated in partners with at least 7 supporting measurements, with a single exception corresponding to KANK2, in which no motif-like segments could be identified. In contrast, most of the partners without a likely motif candidate were detected by only a single, indirect method. Although it cannot be ruled out that some of these interactions represent an alternative binding mechanism to LC8, the large number of PPIs that are not compatible with the existing binding mode underlines that data from PPI databases should be treated with caution as they may contain several false positives.

Among the 172 predicted motifs in 152 proteins that satisfied our filtering criteria, there were 43 novel motif instances predicted by our pipeline that had corresponding PPI data. While most of these interactions were based on indirect, high-throughput experiments, many of them were supported by multiple measurements. 8 of these motifs represented additional instances in proteins that already contained an experimentally verified binding motif. In 35 cases, the likely binding region for LC8 could be identified for proteins whose interaction were studied only at the level of the protein (see S2 Table). Furthermore, the presence of PPIs lends support to the biological relevance of the predicted motifs in these cases. By considering motif hits located in proteins with corresponding PPI data, we created a dataset that contained a list of high confidence motif instances for LC8. In addition to PPIs, the presence of island-like conservation was also taken as an indication of likely motif hits. Altogether, we collected 71
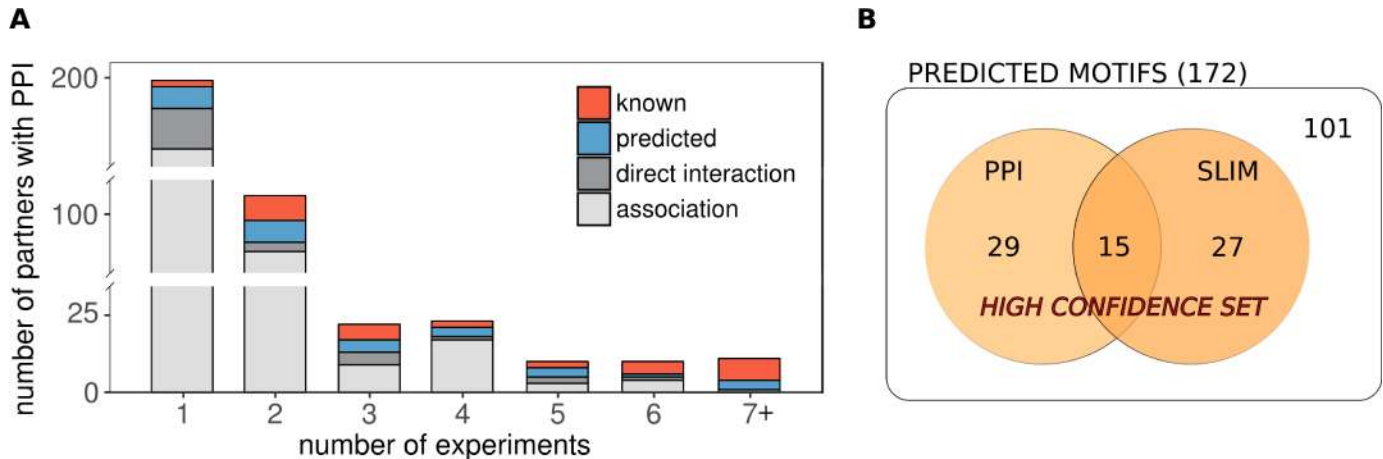
Fig 4. Distribution of LC8 partners and predicted motifs. (A) Distribution of LC8 interaction partners from PPI databases over the number of experiments. Partners that contain neither known nor predicted motifs are shaded grey according to the interaction type. (B) The composition of the high confidence prediction set shown as a Venn diagram. PPI: motifs in partners that appear in PPI databases. SLIM: motifs that show island-like conservation with the SLiMPrints method.

https://doi.org/10.1371/journal.pcbi.1005885.g004

high confidence motif instances, supported by PPI data (29 cases), the presence of island-like conservation (27 cases), or both (15 cases) (Fig 4B). Using these complementary information, the number of interaction partners of LC8 could be practically doubled (S3 Table). The high confidence motifs are available at http://gerdos.web.elte.hu/data/LC8/HUMAN/high_confidence_hits.html.

The high confidence set together with the already known motif hits was used to carry out a GO enrichment analysis against the human proteome with the DAVID server [44]. The analysis showed an enrichment of multiple functionalities previously associated with LC8, such as cytoskeletal organization, microtubule binding or cell morphogenesis (Fig 5). Besides the already known activities, the enrichment analysis revealed a highly over-represented, yet previously undiscovered function of LC8 that links the protein to the regulation of cell and tissue growth, and in particular, to the Hippo pathway. This result suggests that LC8 might play a critical role in the regulation of the Hippo pathway.

High confidence motif hits establish a role of LC8 in the Hippo pathway

In order to understand how LC8 is connected to the Hippo pathway, we took a closer look at the relevant partners. The motif hits highlighted two families involved in the upstream regulation of the Hippo pathway, the WWC (WWC1, WWC2 and WWC3) and angiomotin (AMOT, AMOTL1, AMOTL2) families [45,46,47]. The WWC family member WWC1, also known as KIBRA, was shown previously to interact with LC8 [48], and two binding motifs were also identified, located at positions 278 and 887 (highlighted in Table 1) [49]. However, no binding sites have been previously identified in the other two members of the family, WWC2 and WWC3. Interactions between LC8 and the AMOT family members were previously reported in high-throughput studies to be LC8 interactors [50], but the binding motif has not been located in any of these proteins as of yet. Here, we identified the likely LC8 binding regions in all WWC and AMOT family members and analyzed their evolutionary conservation. The binding of these peptides was verified and quantitatively characterized by SPR (Table 1).

The list of putative LC8 binding peptides identified by our protocol are shown in Table 1, together with the measured binding constants, which confirmed the binding of the selected
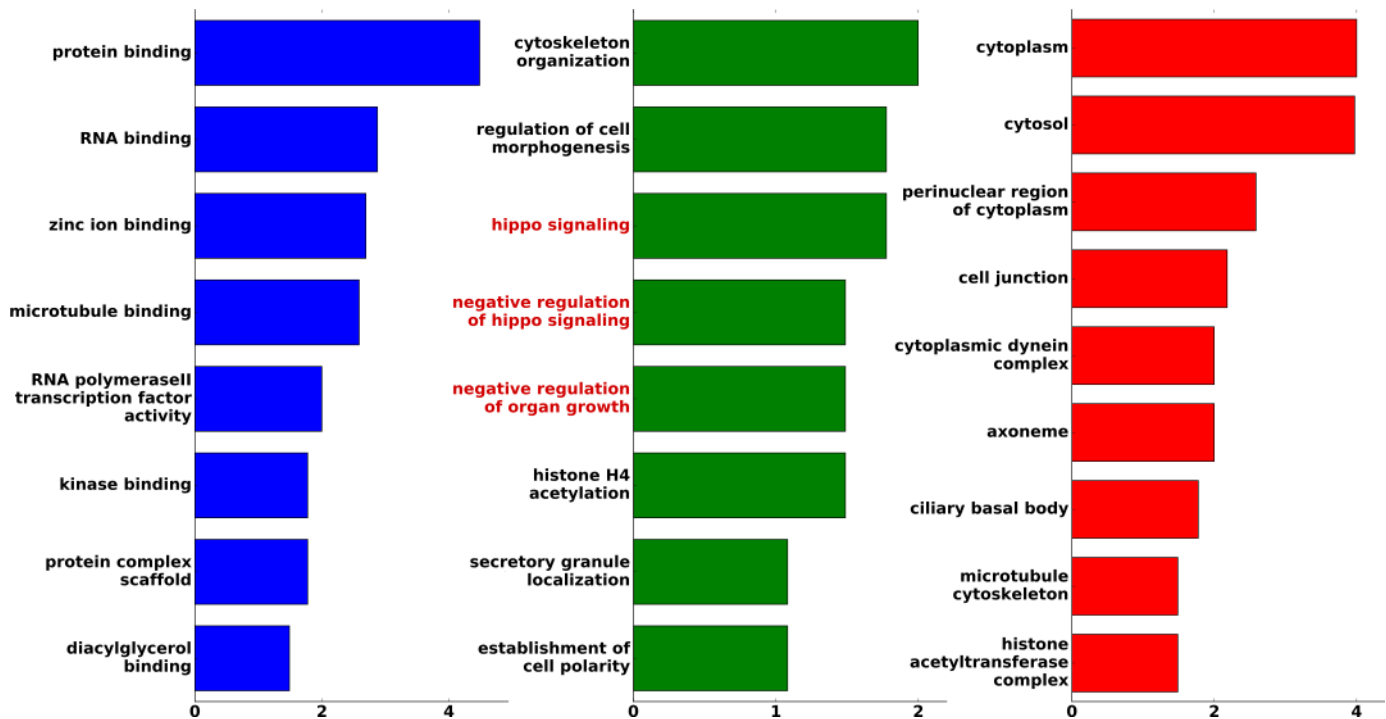
Fig 5. GO enrichment in the high confidence set of LC8 binding partners. Molecular Function (blue), Biological Process (green) and Cellular Component (red) categories that are enriched in the high-confident LC8 binding partners compared to human background. The x-axis represents the log-odds ratio of each enriched GO category. Process names related to the Hippo pathway are colored in red.

https://doi.org/10.1371/journal.pcbi.1005885.g005

peptides to LC8. As an example, results are shown for the motif in AMOTL2 in Fig 6. Although the apparent K_d values indicate weak interactions, these binding affinities are similar to other LC8 interactions [25,20], and in biological settings they can increase in strength due to avidity caused by dimerization [30]. Besides the canonical TQT motifs these peptides also contain SQT and TNT motifs. Additional peptides that further expand the repertoire of compatible amino acids at various positions were also tested for their binding to LC8 (S4 Table).

In order to properly assess the conservation of LC8 binding motifs in WWC and AMOT family members, we traced the evolutionary history of these proteins and their identified binding regions and additional functional modules. Almost all vertebrate species have three paralogs of both the WWC and AMOT family members (Figs 7 and 8). For both families, the

Table 1. Experimentally validated LC8 binding motifs in the Hippo pathway.

Gene name	UniProt Accession	Start	Binding motif sequence	$K_d(\mu M)\pm SD$
WWC1*	Q8IX03	278	LDVSSQTD	0.45±0.03
WWC1*	Q8IX03	887	VDKETNTE	4.00±0.50
WWC2	Q6AWC2	282	LDAGSQTS	32.7±2.97
WWC2	Q6AWC2	962	VDKETNTD	17.6±0.54
WWC3	Q9ULE0	208	CDAGSQTD	5.34±0.18
WWC3	Q9ULE0	867	VDKETNTE	3.50±0.15
AMOT	Q4VCS5	860	RDCSTQTE	1.20±0.04
AMOTL1	Q8IY63	873	KDSSTQTD	1.55±0.04
AMOTL2	Q9Y2J4	724	RDGSTQTE	1.98±0.19

*The two WWC1 motifs were validated in a previous study [49]

https://doi.org/10.1371/journal.pcbi.1005885.t001
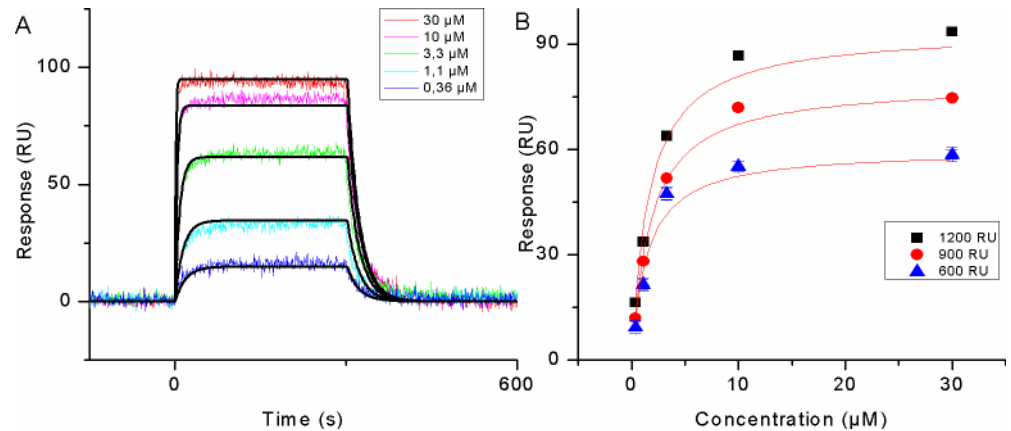
Fig 6. SPR analysis of the LC8 –AMOTL2 interaction. (A) AMOTL2 peptide was injected to His$_6$-LC8 immobilized on Tris-NTA sensor chip as described in Materials and Methods. Sensorgrams derived from five different analyte injections are indicated with colored lines. Black lines show the global fit of the experimental data to a 1: 1 Langmuir model. (B) Equilibrium analysis of the LC8 –AMOTL2 interaction at different LC8 immobilization levels.

multiple paralogs observed at the level of vertebrates could be traced back to a single common ancestor gene, as shown in the case of Florida lancelet. KIBRA orthologs were detected beyond the level of chordates, not only in arthropods, but also in cnidaria (Fig 7). AMOT was also present in arthropods, but it was missing in the fruit fly (Fig 8). The most likely explanation for this is that the angiomotin was originally present in all bilaterian animals, but was lost relatively recently in the dipteran lineage that includes Drosophila [51]. Altogether, the WWC family members seem to have a more ancient evolutionary origin compared to angiomotins.

Both WWC and AMOT family members showed a highly conserved domain and linear motif organization across a wide range of species (Figs 7 and 8). While most LC8 binding motifs verified in human sequences are not conserved beyond vertebrates, KIBRA and the other WWC family members represent an important exception in this regard. The two LC8 binding motifs of this protein family are well conserved in all three paralogs in vertebrates, lancelet and fruit fly. Furthermore, one of the motifs is also present in a sea anemone, matching the strong evolutionary conservation of the WW and C2 domains, the defining functional modules of this family (Fig 7). The AMOT family members contained a single LC8 binding motif, which seems to have emerged more recently. While orthologous sequences could be detected in non-vertebrate species based on the conserved coiled coil region together with the angiomotin domain, these sequences lacked the LC8 binding motif (Fig 8). The LC8 motif, similarly to the PDZ and WW binding motifs, has become conserved within vertebrate species. In both of these families, the redundancy and conservation underline the functional importance of LC8 binding motifs in the upstream regulation of the Hippo pathway.

Discussion

In this work, we established a systematic filtering protocol that can be used to expand the interaction network of a given linear motif binding domain with high confidence motif hits and reduce the number of random motif matches. While similar bioinformatic pipelines have been used before [14], the optimality of the applied filtering steps could not be guaranteed or even assessed due to the lack of appropriate measures. Here, we offered a solution to this problem by introducing a decision tree-like filtering procedure together with the weighted information gain that enabled the construction of an optimal bioinformatic pipeline. The presented
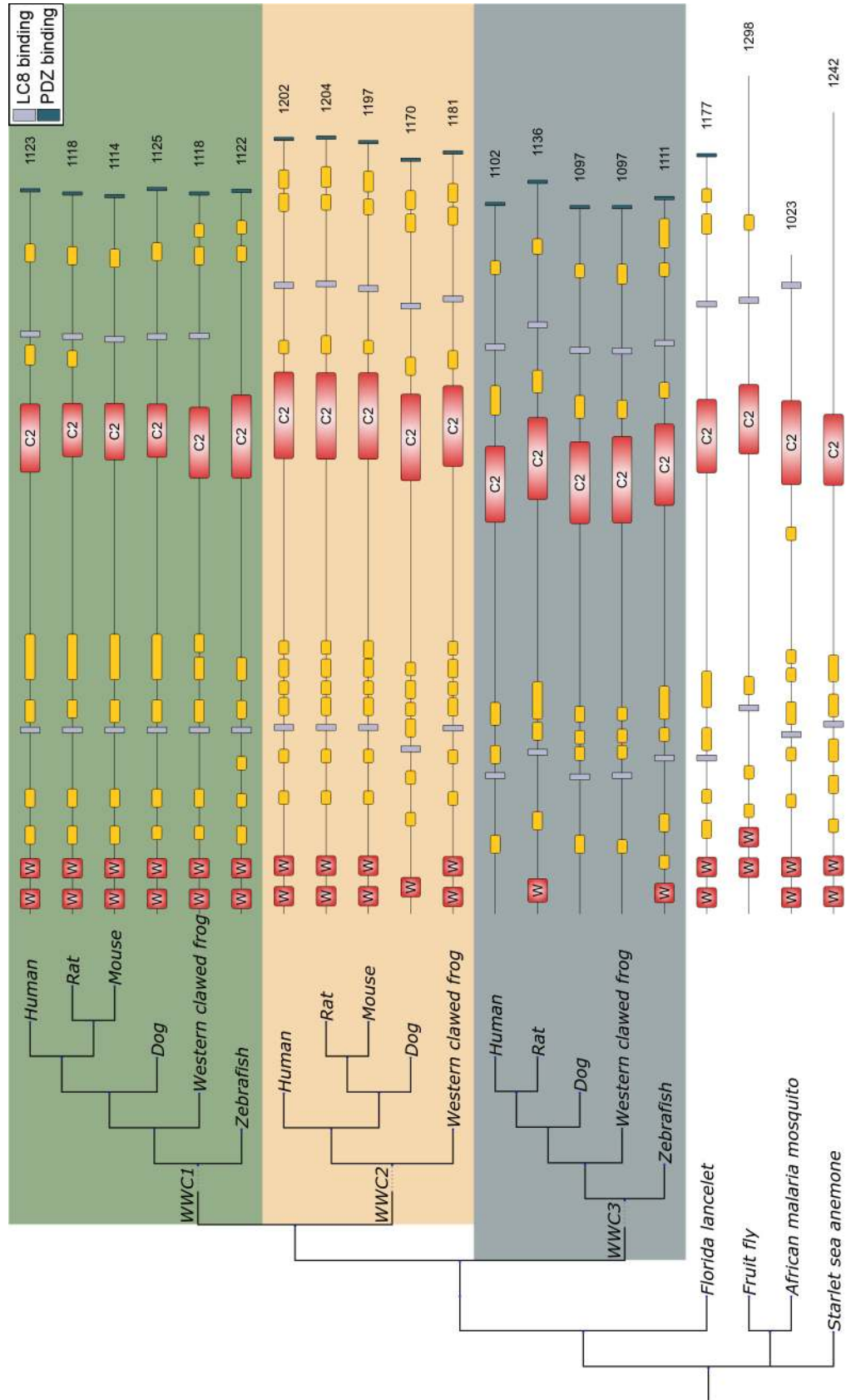
Fig 7. Structural organization and phylogenetic tree of the WWC family members. The tree branch lengths do not represent real values. The subtrees of the three vertebrate WWC paralogs are highlighted by colored rectangles. Red and yellow boxes indicate the domains and coiled coil regions, respectively. Small purple and dark green boxes mark the location of the LC8 and PDZ binding motifs, respectively.

approach was applied to expand the interaction network of LC8, a highly conserved eukaryotic hub protein that binds its partners via a specific linear motif [22]. By combining our motif filtering protocol with data collected from protein-protein interaction databases and the information on the presence of island-like conservation, we created a dataset of 71 novel high confidence motif instances (S3 Table). These novel binding sites significantly enriched the interaction network of LC8 with novel partners and highlighted a previously unknown, important function of LC8 in the Hippo pathway.

A list of predicted LC8 binding motifs was also created in an earlier work [20]. In this case, the PSSM was calculated based on a library of binding motifs evolved through phage selection instead of using the collection of naturally occurring motif instances. Although some of the filtering criteria, like IUPred based disorder prediction or intracellular localization were used in both studies, the details of implementation differed significantly, including the optimal cutoff for the PSSM score. The comparison of the two sets showed that among the top 83 hits identified based on the in vitro evolution and absent from the list of known naturally occurring motifs, 58 (70%) were also uncovered by our presented motif discovery pipeline. However, our evolutionary filtering criteria reduced this overlap to 38 (46%). While to some extent, the limited overlap could be due to the technical differences in the implementation, a more appealing explanation is that binding motifs selected by phage display are optimized for binding strength, while the sequence of naturally occurring motifs were shaped by various evolutionary requirements in the cell, from which strength is just one, and not necessarily the dominant constraint. Despite these differences, in vitro evolution by phage selection [20,52] represent a complementary approach to predict potential motif instances for specific binding domains and to explore the binding preferences of such domains.

While the filtering protocol implemented here is specific to LC8, the presented work also has important implications beyond this system. We suggested here a general framework to find the optimal selection of methods for the filtering steps. The importance of this optimization can be best demonstrated through the example of disorder prediction methods. Several methods, such as DISOPRED3, PONDR VSL2 or ESpritz Disprot [33–35], that were tested in this work, perform better on specific datasets of ordered and disordered proteins [53–55]. Nevertheless, IUPred achieved the best results for this specific problem, supporting the choice of this approach in motif-centric application of protein disorder [4,14,19]. The pipeline applied here also enabled us to identify features that had limited discriminatory power and therefore were not incorporated into the current pipeline. For example, all known LC8 binding motifs adopt a β-strand conformation upon binding, but the current tools are not capable of capturing this property efficiently. Additional features, such as predicted disordered binding regions/ MoRFs, the presence of coiled coils or island-like conservation also had limited discriminatory power. As these properties are clearly associated with at least a subset of LC8 binding partners, they highlight potential areas where more efficient methods are needed.

The analysis of evolutionary information was also highly valuable to increase the biological relevance of potential functional sites, but also provided some unexpected results. LC8 is ubiquitous in all eukaryotes, presenting a highly conserved binding interface for its interaction partners, which is practically identical from Drosophila to human. One of the most well-characterized interaction partner of LC8, the dynein intermediate chain is also conserved, at least in the animal kingdom. Therefore, it was surprising to find that most of the other known
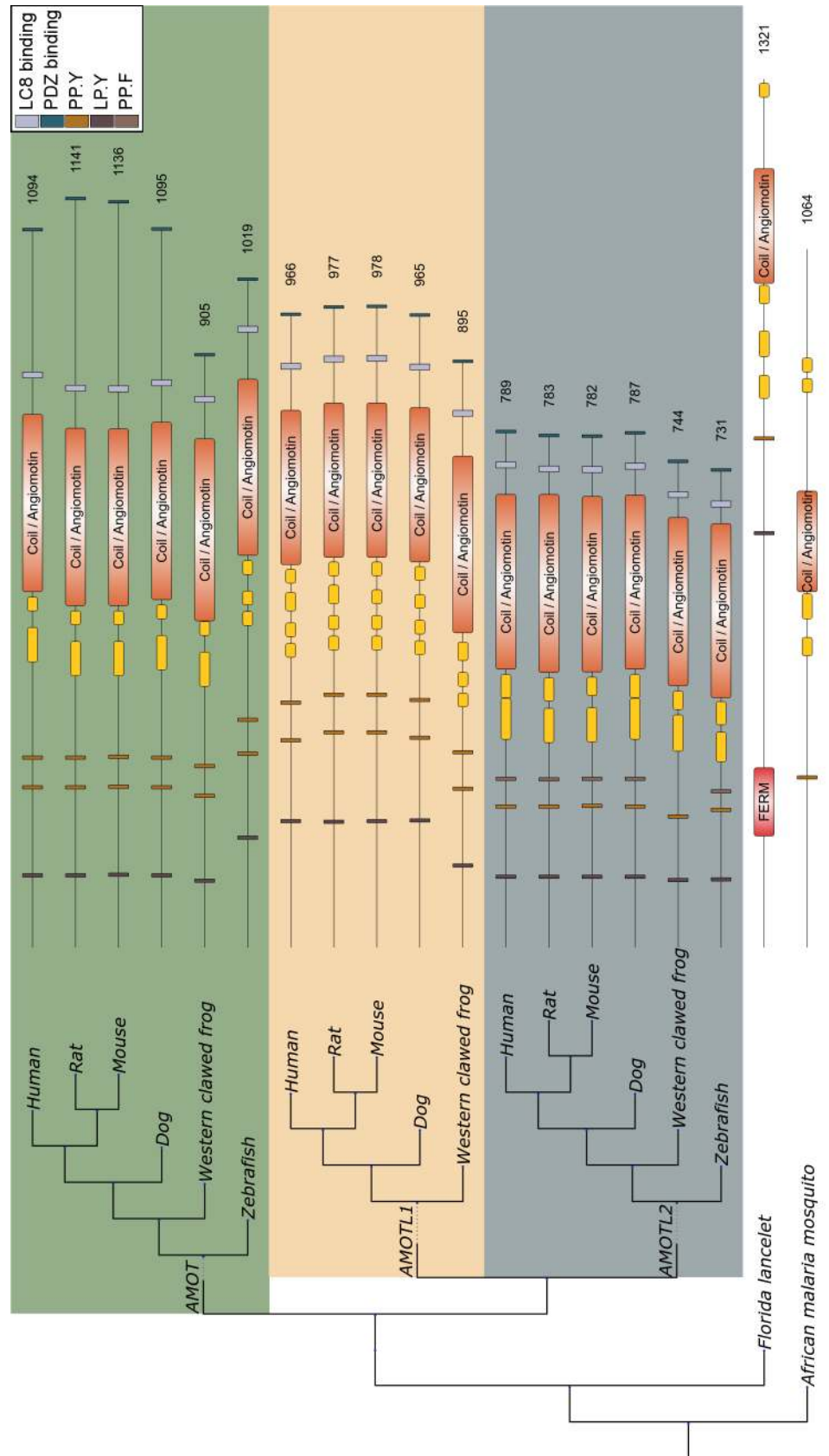
Fig 8. Structural organization and phylogenetic tree of the AMOT family members. The tree branch lengths do not represent real values. The subtrees of the three vertebrate AMOT paralogs are highlighted by colored rectangles. The angiomotin PFAM family was merged with overlapping coiled coil regions and indicated as angiomotin/Coil domain in orange. Other coiled coil regions and domains are colored yellow and red, respectively. Small boxes mark the location of linear motifs: brown (dark, light, medium)—WW binding, purple—LC8 binding, and dark green—PDZ binding motifs.

interaction partners of LC8 showed limited evolutionary conservation at the motif level. The overwhelming majority of both known and predicted motifs in the human proteome could not be traced beyond vertebrates. These results are in agreement with the evolutionary plasticity of linear motifs, as SLiMs can be introduced or eliminated by a few point mutations, or created ex-nihilo [56]. However, the lack of motif conservation was unexpected in the light of the strong evolutionary constraints acting upon LC8. Nevertheless, signs of purifying selection could be detected for the binding motifs over smaller evolutionary distances. In a subset of cases, the SLiMPrints method [19] was able to detect that motif residues showed higher relative conservation compared to their flanking regions. However, this approach failed to highlight many validated binding motifs, either because of problems with alignments, a strong overall conservation or the lack of the sensitivity of the method. In our experience, a more efficient filtering criterion was the conservation of motif residues within mammalian species. This criterion was true for nearly all verified human binding motifs, but was not met by most of the random motif hits that were located within the evolutionary variable disordered regions. Therefore, it is a highly efficient filtering approach. Nevertheless, we understand that by filtering based on mammalian-level conservation comes with a risk of losing some instances that are conserved only in a smaller taxonomic unit within mammalia, e.g. in primates.

The expanded interaction network of LC8 highlighted novel functions and drew attention to the potential role of LC8 in the Hippo pathway which controls organ size and cell growth by regulating cellular proliferation and apoptosis [57]. In our high confident hits, two putative LC8 binding partner families, the WWC and the angiomotin family were identified. Both families are tightly connected to the Hippo pathway and are known to be involved in its upstream regulation pathway in multiple ways [58]. WWC and AMOT family members can act either by activating the core kinase modules or by forming complexes with YAP/TAZ sequestering them at cell-cell junctions and preventing their nuclear access [59–64]. Previously, the interaction between WWC family members and LC8 has been established at the protein level [65], but the binding motifs were identified only in WWC1/KIBRA [49]. In addition, a study of the protein interaction network of the mammalian Hippo pathway also revealed interaction between AMOT proteins and LC8 [66], confirmed by additional large-scale studies [67,68]. A recent work established a connection between LC8 and the Hippo pathway in Drosophila [69]. In the present work, the LC8 binding sites were experimentally verified both in WWC and the AMOT protein families and were shown to be evolutionarily conserved.

The verification of binding sites established a direct connection between LC8 and the Hippo pathway through the WWCs and AMOT family proteins, nevertheless, the exact role of these interactions in the regulation of Hippo pathway is still unclear. Recently, LC8 was suggested to promote dimerization of partially disordered proteins by stabilizing their coiled coil regions [25]. This model was based on detailed characterization of the complex formation of several scaffolding proteins, such as Swallow, dynein intermediate chain DYNC1|1 and GKAP [70]. However, there are still open questions regarding the regulation of binding of LC8 to its partners. In agreement with the general model for LC8 function, both families contain coiled coil (CC) regions and the dimerization of both AMOT [71,64] and WWC family members [72] might be necessary for their scaffolding activity. While the functional role of the CC region

in the WWC family has not been explored, the CC regions of AMOT family members were shown to form interactions with various components of the Hippo pathway and influence their activity and localization [59,63,73,64]. Therefore, the potential stabilization of coiled coil regions driven by LC8 could also play important roles in regulation of the Hippo pathway. However, the LC8 function related to intracellular trafficking cannot be ruled out entirely either, as both KIBRA and AMOT proteins alternate between cytoplasmic and nuclear locations depending on various signals [39,74]. While LC8 might provide an important additional layer to the regulation of the Hippo pathway, further investigations are needed to understand how the interaction of LC8 affects WWC and AMOT proteins in the Hippo pathway and in other processes.

In conclusion, the novel binding partners presented in this work provide exciting new opportunities to study how the binding of LC8 influences the stability and binding properties of various proteins. In addition, a general framework was also proposed here for the optimization of the linear motif filtering. While the protocol was applied specifically to LC8 and takes advantage of its relatively well-characterized interaction network with over 60 known linear motif instances, it can be applied in a similar way for other linear motif binding domains for which similar number of partners have been identified. In a more general sense, the novel motif filtering protocol further underlines that computational approaches can complement large-scale experimental studies of linear motif binding systems, and advance our understanding of how these relatively weak, transient interactions contribute to highly dynamic, complex regulatory processes in the cell.

Materials and methods

Computational filtering procedure

We collected 76 LC8 binding motifs from the literature, however, upon further investigation we rejected 9 of them due to the lack of motif level evidence. The resulting list is given in S1 Table (for details, see S1 Text). The two vertebrate isoforms of LC8, DYNLL1 and DYNLL2 were not discriminated in this study. From the collected 67 binding motifs we generated a non-redundant set using the CD-HIT suite tool [75]. The remaining 62 motifs were aligned and used to construct a PSSM. The values in the PSSM matrix capture the preferences of each amino acid for over-representation or under-representation at each position in the binding motif. The elements of the PSSM ($P_{i, j}$) were expressed as the log-odds score of amino acid frequency in each position in the known binding partners divided by the background frequency:

$$P_{i,j} = \log\left(\frac{A_{i,j}}{D_i}\right)$$

where $A_{i, j}$ is the frequency of amino acid i at position j in the alignment of known binding partners, and D_i is the background frequency of amino acid i. The background frequency was derived from the frequency of amino acids in the UniProt Eukaryotic proteome. As not every amino acid was present in each position in the known partners, we incorporated a pseudocount correction to account for these zero occurrences.

$$P_{i,j} = \log\left(\frac{\frac{A_{i,j}+\frac{B}{20}}{m+B}}{D_i}\right)$$

Where B is the pseudocount with a value of 5, as suggested in previous works [76], m is the number of sequences, and 20 is the number of amino acids.

Additional sequential features were calculated for each known binding partner to identify the attributes of true binding events. The binding partners were analyzed using more than 20

different tools, to predict disorder content (IUPred, PONDR VSL2 and DISOPRED3) [32,33,35], the tendency to be involved in disordered binding regions (ANCHOR, MoRFchibi and DISOPRED3-BR) [35–37], secondary structure and subcellular localization predictions. For each potential binding motif, position-based numerical values were averaged. The subcellular localization of a motif was predicted by a combination of methods, and only putative motifs with predicted intracellular localization were considered to be accessible for interaction with LC8. A motif was considered to be intracellular if it contained no predicted signal sequence according to the SignalP predictor [77]. Proteins with a predicted signal sequence were further analysed using PHOBIUS [78] for transmembrane topology and the motif was categorized according to its result to either be in intracellular (INT), transmembrane (TM) or extracellular space (EXT). PFAM identifies conserved sequence families, which can be of different types, including domains, families, motifs, repeats and also intrinsically disordered domains [31]. Each motif was categorized as Domain or Family if it overlapped in at least one position with a respective PFAM annotation [31]. Secondary structure predictions were carried out using PSIPRED [38]. Proteins were checked with NCOILS to predict coiled-coil regions [41]. The calculated features can be found at http://gerdos.web.elte.hu/data/LC8/known_results.html.

A protocol was generated to find the optimal filtering criteria based on these sequence features. Due to the limited number of positive examples, the full tree space was not explored, instead, optimal filtering methods and the best parameter settings were selected globally. For this, we defined two sets, one from the known human motifs, and one from 10,000 randomly selected peptides from the human proteome with PSSM scores above 0. Various attributes were calculated for each motif in both sets. We sought to find attributes that can keep all or most of the known motifs while discarding the largest number of the randomly selected peptides. The optimization was guided by using the information gain theory derived from the Shannon entropy.

$$I = H_p - p_{c1}(H_{c1}) - p_{c2}(H_{c2})$$

Where I is the information gain, H_p is the Shannon entropy of the parent group, c_1 and c_2 are the child groups on a specific split, and p is the probability of an attribute in the given group. During the entropy calculations, the number of occurrences with a specific attribute were normalized with the sum of the examples to weight the imbalanced sets. The information gain reaches its theoretical maximum in the ideal case, which results in a clean set, e.g. when a criterion is true for all entries in the known motif set. In these cases, the best attribute was defined as the one that filtered out the largest number of random peptides.

We validated the filtering protocol using a 3-fold cross-validation. The set of known human interaction partners and their respective attributes were split into three equal, non-overlapping subsets. Two sets were used to calculate the new maximum points of the information gain for each attribute, and the elements of the remaining third set were used as a test set to evaluate the resulting filters. The number of correctly categorized examples were calculated by summing over all three possible test set.

Evolutionary analysis

In order to compile evolutionary data on each protein having any putative motif according to our pipeline, we generated a dataset of orthologous sequences. These hits were obtained by running the GOPHER prediction algorithm with default settings against the QFO database [79,80]. Then, we constructed the multiple sequence alignments of orthologs for each protein using the MAFFT algorithm (default parameters) [81]. The ortholog sequences were classified

into the most specific term using the five main evolutionary levels according to the UniProt taxonomic lineage: Mammalia, Vertebrata, Metazoa, Fungi and Eukarya.

According to the multiple sequence alignments of orthologs, each aligned instance of the candidate LC8 binding motifs was analyzed in PSSM based conservation terms. For the evolutionary analysis at least two predicted orthologs were required at each level. An aligned motif was considered to be conserved if its PSSM score exceeded the cutoff value (3.3). In the evolutionary filtering step a candidate motif was kept if it was conserved in at least 80% of the mammalian species.

Motif conservation was also analyzed using the SLiMPrints method [19]. This method searches for regions with island-like conservation, i.e. regions that have high conservations relative to their flanking regions. A motif was considered to have an island-like conservation if it had at least 3 overlapping positions with a significant region (p-value<0.05) detected by SLiM-Prints. The method was applied to known human binding motifs of LC8 as well as predicted motifs.

In order to study the evolutionary history of the structural organization of AMOT and WWC families, the domain annotation and coiled coil region predictions were retrieved from the InterPro resource (version 63) [82]. The inter-species motif mapping of the known LC8 binding motifs of WWC and AMOT family members was carried out by using the canonical ELM definition and literary data. The definitions of the C-terminal class 3 PDZ-binding motif and the WW-binding motif of group I were obtained from the ELM database [5]. The first WW-binding motif of the AMOT family members was defined as LPxY [74]. In addition, within the orthologues of AMOTL2 the WW-binding motif of group I was defined as PPxF. The identified ortholog sequences were used to generate the multiple sequence alignment and phylogenetic tree of the families by applying the PhyML algorithm (default settings) [83].

Protein-protein interaction data

To collect PPIs for LC8 from public databases we applied the PSICQUIC approach [43]. Known non-linear motif binding partners (DYNLL1, DYNLL2 and UBC) were omitted. To remove redundant hits, the following four filtering steps were applied: 1. For each partner and study only the direct interactions were kept, except when the given study had only non-direct interactions for that protein. 2. The less informative 'other affinity chromatography technology' annotations were filtered out when the study had 'TAP', 'CO-IP' or 'pulldown' annotation as well. 3. Annotations with unknown association type were filtered out if another database had the same interaction from the same study with full annotation. 4. Cases where the PubMed identifier could not be retrieved were filtered out if the given partner was annotated with the same detection method in an identifiable study. The process resulted in a collection of 559 interaction partners for LC8. The partners came from 7 eukaryotic species and 23 virus strains. Finally, for subsequent analysis, the subset of human, mouse and rat data were merged, and this way 381 partners remained. The resulting interaction data is given in S2 Table.

PSICQUIC gives the PSI-MI identifier for detection methods and association types. In the case of the detection methods, based on their meaning and position in the PSI-MI hierarchy, we (i) assigned the labels "biochemical or biophysical", "PCA" and "other" to them; (ii) merged the detection method descriptions into 8 groups (see the legend on S5A Fig). Similarly, we labelled the association types as "direct" or "non-direct" interaction. PSICQUIC doesn't provide direct information about the scale of the study, therefore we queried each PubMed identifier in our set. Since the output was, again, redundant, we sorted the number of annotations by source, and took only the highest one. If this number was higher than 50, we considered it as "high throughput", otherwise as "low throughput".

Experimental procedure

The full-length His$_6$-tagged LC8 (DYNLL2; UniProt accession number Q96FJ2) was cloned into a pET21-derived pBH4 vector and expressed as described previously [84]. The 11-residue-long fragments of the predicted binding partners were synthesized using solid-phase technique by GenoSphere Ltd.

SPR measurements were performed on a ProteOn XPR36 (Bio-Rad) instrument equipped with HTG sensor chip (Bio-Rad). The sensor chip was activated for 300 s with 150 μl 10 mM NiCl$_2$ solution followed by a washing step for 300 s with the running buffer containing 20 mM Hepes, 150 mM NaCl, 0.05% Tween-20, 0.1 mM TCEP, 50 μM EDTA, pH 7.5 buffer. The immobilization of His$_6$-tagged LC8 was performed on the activated Ni^{2+}-NTA surface at three different densities (1200 RU, 900 RU, 600 RU). The putative binding peptides were injected onto the chip at five different concentrations simultaneously at a flow rate of 60 μl · min^{-1} for 400 s, while the dissociation of the peptides was recorded for 600 s. In the sixth analyte channel, running buffer was injected for double referencing. The double referenced data were global fitted to the 1: 1 Langmuir model using the ProteOn software. The presented K$_d$ values were obtained from the mean of the kinetic and the equilibrium analysis delivered K$_d$ values, and the standard deviations were calculated from three individual K$_d$ values.

Supporting information

S1 Text. Supporting information.
(DOCX)

S1 Fig. Example structures of the complexes formed between LC8 and selected binding peptides.
(DOCX)

S2 Fig. Distribution of PSSM scores across the human proteome.
(DOCX)

S3 Fig. Summarized evolutionary conservation results of known non-human LC8 binding partners.
(DOCX)

S4 Fig. Summarized evolutionary conservation results of known human LC8 binding motif applying lower PSSM cutoffs.
(DOCX)

S5 Fig. Statistics of PSICQUIC results.
(DOCX)

S1 Table. Collection of experimentally verified LC8 binding motifs from the literature.
(XLSX)

S2 Table. LC8 binding partners from PPI databases.
(XLSX)

S3 Table. High confidence prediction set for LC8 binding motifs.
(XLSX)

S4 Table. Experimentally validated LC8 binding motifs.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Zsuzsanna Dosztányi.

Data curation: Gábor Erdős, Mátyás Pajkos, Borbála Hajdu-Soltész.

Formal analysis: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész.

Funding acquisition: Gábor Pál, Zsuzsanna Dosztányi.

Investigation: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész, Zsuzsanna Dosztányi.

Methodology: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész, Zsuzsanna Dosztányi.

Project administration: Zsuzsanna Dosztányi.

Resources: Bence Kiss, Gábor Pál, László Nyitray.

Software: Gábor Erdős, Mátyás Pajkos, Borbála Hajdu-Soltész.

Supervision: Zsuzsanna Dosztányi.

Validation: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész, Zsuzsanna Dosztányi.

Visualization: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész.

Writing – original draft: Gábor Erdős, Tamás Szaniszló, Mátyás Pajkos, Borbála Hajdu-Soltész, Gábor Pál, László Nyitray, Zsuzsanna Dosztányi.

References

1. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev. 2014; 114: 6733–6778. https://doi.org/10.1021/cr400585q PMID: 24926813

2. Uyar B, Weatheritt RJ, Dinkel H, Davey NE, Gibson TJ. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? Mol Biosyst. 2014; 10: 2626–2642. https://doi.org/10.1039/c4mb00290c PMID: 25057855

3. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol Biosyst. 2012; 8: 268–281. https://doi.org/10.1039/c1mb05231d PMID: 21909575

4. Gibson TJ, Dinkel H, Van Roey K, Diella F. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. Cell Commun Signal. 2015; 13: 42. https://doi.org/10.1186/s12964-015-0121-y PMID: 26581338

5. Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, et al. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 2016; 44: D294–300. https://doi.org/10.1093/nar/gkv1291 PMID: 26615199

6. Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. Mol Cell. 2014; 55: 161–169. https://doi.org/10.1016/j.molcel.2014.05.032 PMID: 25038412

7. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol. 2001; 19: 348–353. https://doi.org/10.1038/86737 PMID: 11283593

8. Ivarsson Y, Arnold R, McLaughlin M, Nim S, Joshi R, Ray D, et al. Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. Proc Natl Acad Sci U S A. 2014; 111: 2542–2547. https://doi.org/10.1073/pnas.1312296111 PMID: 24550280

9. Weatheritt RJ, Luck K, Petsalaki E, Davey NE, Gibson TJ. The identification of short linear motif-mediated interfaces within the human interactome. Bioinformatics. 2012; 28: 976–982. https://doi.org/10.1093/bioinformatics/bts072 PMID: 22328783

10. Edwards RJ, Palopoli N. Computational prediction of short linear motifs from protein sequences. Methods Mol Biol. 2015; 1268: 89–141. https://doi.org/10.1007/978-1-4939-2285-7_6 PMID: 25555723

11. Mészáros B, Dosztányi Z, Simon I. Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition. PLoS One. 2012; 7: e46829. https://doi.org/10.1371/journal.pone.0046829 PMID: 23056474

12. Gutman R, Berezin C, Wollman R, Rosenberg Y, Ben-Tal N. QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. Nucleic Acids Res. 2005; 33: W255–61. https://doi.org/10.1093/nar/gki496 PMID: 15980465

13. Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, Brooks AW, et al. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. Nucleic Acids Res. 2012; 40: D252–60. https://doi.org/10.1093/nar/gkr1189 PMID: 22146221

14. Krystkowiak I, Davey NE. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx238 PMID: 28387819

15. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006; 34: W362–5. https://doi.org/10.1093/nar/gkl124 PMID: 16845026

16. Horn H, Haslam N, Jensen LJ. DoReMi: context-based prioritization of linear motif matches. PeerJ. 2014; 2: e315. https://doi.org/10.7717/peerj.315 PMID: 24711967

17. Brown CJ, Johnson AK, Daughdrill GW. Comparing models of evolution for ordered and disordered proteins. Mol Biol Evol. 2010; 27: 609–621. https://doi.org/10.1093/molbev/msp277 PMID: 19923193

18. Davey NE, Shields DC, Edwards RJ. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics. 2009; 25: 443–450. https://doi.org/10.1093/bioinformatics/btn664 PMID: 19136552

19. Davey NE, Cowan JL, Shields DC, Gibson TJ, Coldwell MJ, Edwards RJ. SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res. 2012; 40: 10628–10641. https://doi.org/10.1093/nar/gks854 PMID: 22977176

20. Rapali P, Radnai L, Süveges D, Harmat V, Tölgyesi F, Wahlgren WY, et al. Directed evolution reveals the binding motif preference of the LC8/DYNLL hub protein and predicts large numbers of novel binders in the human proteome. PLoS One. 2011; 6: e18818. https://doi.org/10.1371/journal.pone.0018818 PMID: 21533121

21. Zeke A, Bastys T, Alexa A, Garai Á, Mészáros B, Kirsch K, et al. Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases. Mol Syst Biol. 2015; 11: 837. https://doi.org/10.15252/msb.20156269 PMID: 26538579

22. Rapali P, Szenes Á, Radnai L, Bakos A, Pál G, Nyitray L. DYNLL/LC8: a light chain subunit of the dynein motor complex and beyond. FEBS J. 2011; 278: 2980–2996. https://doi.org/10.1111/j.1742-4658.2011.08254.x PMID: 21777386

23. Davey NE, Morgan DO. Building a Regulatory Network with Short Linear Sequence Motifs: Lessons from the Degrons of the Anaphase-Promoting Complex. Mol Cell. 2016; 64: 12–23. https://doi.org/10.1016/j.molcel.2016.09.006 PMID: 27716480

24. Michael S, Travé G, Ramu C, Chica C, Gibson TJ. Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. Bioinformatics. 2008; 24: 453–457. https://doi.org/10.1093/bioinformatics/btm624 PMID: 18184688

25. Barbar E. Dynein light chain LC8 is a dimerization hub essential in diverse protein networks. Biochemistry. 2008; 47: 503–508. https://doi.org/10.1021/bi701995m PMID: 18092820

26. Clark SA, Jespersen N, Woodward C, Barbar E. Multivalent IDP assemblies: Unique properties of LC8-associated, IDP duplex scaffolds. FEBS Lett. 2015; 589: 2543–2551. https://doi.org/10.1016/j.febslet.2015.07.032 PMID: 26226419

27. Fan J, Zhang Q, Tochio H, Li M, Zhang M. Structural basis of diverse sequence-dependent target recognition by the 8 kDa dynein light chain. J Mol Biol. 2001; 306: 97–108. https://doi.org/10.1006/jmbi.2000.4374 PMID: 11178896

28. Hall J, Karplus PA, Barbar E. Multivalency in the assembly of intrinsically disordered Dynein intermediate chain. J Biol Chem. 2009; 284: 33115–33121. https://doi.org/10.1074/jbc.M109.048587 PMID: 19759397

29. Williams JC, Roulhac PL, Roy AG, Vallee RB, Fitzgerald MC, Hendrickson WA. Structural and thermo-dynamic characterization of a cytoplasmic dynein light chain-intermediate chain complex. Proc Natl Acad Sci U S A. 2007; 104: 10028–10033. https://doi.org/10.1073/pnas.0703614104 PMID: 17551010

30. Radnai L, Rapali P, Hódi Z, Süveges D, Molnár T, Kiss B, et al. Affinity, avidity, and kinetics of target sequence binding to LC8 dynein light chain isoforms. J Biol Chem. 2010; 285: 38649–38657. https://doi.org/10.1074/jbc.M110.165894 PMID: 20889982

31. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families data-base: towards a more sustainable future. Nucleic Acids Res. 2016; 44: D279–85. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716

32. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005; 347: 827–839. https://doi.org/10.1016/j.jmb.2005.01.071 PMID: 15769473

33. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrin-sic disorder. BMC Bioinformatics. 2006; 7: 208. https://doi.org/10.1186/1471-2105-7-208 PMID: 16618368

34. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein dis-order. Bioinformatics. 2012; 28: 503–509. https://doi.org/10.1093/bioinformatics/btr682 PMID: 22190692

35. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-bind-ing activity. Bioinformatics. 2015; 31: 857–863. https://doi.org/10.1093/bioinformatics/btu744 PMID: 25391399

36. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009; 5: e1000376. https://doi.org/10.1371/journal.pcbi.1000376 PMID: 19412530

37. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res. 2016; 44: W488–93. https://doi.org/10.1093/nar/gkw409 PMID: 27174932

38. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000; 16: 404–405. PMID: 10869041

39. Rayala SK, den Hollander P, Balasenthil S, Yang Z, Broaddus RR, Kumar R. Functional regulation of oestrogen receptor pathway by the dynein light chain 1. EMBO Rep. 2005; 6: 538–544. https://doi.org/10.1038/sj.embor.7400417 PMID: 15891768

40. Bodor A, Radnai L, Hetényi C, Rapali P, Láng A, Kövér KE, et al. DYNLL2 dynein light chain binds to an extended linear motif of myosin 5a tail that has structural plasticity. Biochemistry. 2014; 53: 7107–7122. https://doi.org/10.1021/bi500574z PMID: 25312846

41. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science. 1991; 252: 1162–1164. https://doi.org/10.1126/science.252.5009.1162 PMID: 2031185

42. Slevin LK, Romes EM, Dandulakis MG, Slep KC. The mechanism of dynein light chain LC8-mediated oligomerization of the Ana2 centriole duplication factor. J Biol Chem. 2014; 289: 20727–20739. https://doi.org/10.1074/jbc.M114.576041 PMID: 24920673

43. del-Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, et al. A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. 2013; 41: W601–6. https://doi.org/10.1093/nar/gkt392 PMID: 23671334

44. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annota-tion, Visualization, and Integrated Discovery. Genome Biol. 2003; 4: P3. PMID: 12734009

45. Yu F-X, Guan K-L. The Hippo pathway: regulators and regulations. Genes Dev. 2013; 27: 355–371. https://doi.org/10.1101/gad.210773.112 PMID: 23431053

46. Mana-Capelli S, Paramasivam M, Dutta S, McCollum D. Angiomotins link F-actin architecture to Hippo pathway signaling. Mol Biol Cell. 2014; 25: 1676–1685. https://doi.org/10.1091/mbc.E13-11-0701 PMID: 24648494

47. Genevet A, Wehr MC, Brain R, Thompson BJ, Tapon N. Kibra is a regulator of the Salvador/Warts/Hippo signaling network. Dev Cell. 2010; 18: 300–308. https://doi.org/10.1016/j.devcel.2009.12.011 PMID: 20159599

48. Traer CJ, Rutherford AC, Palmer KJ, Wassmer T, Oakley J, Attar N, et al. SNX4 coordinates endosomal sorting of TfnR with dynein-mediated transport into the endocytic recycling compartment. Nat Cell Biol. 2007; 9: 1370–1380. https://doi.org/10.1038/ncb1656 PMID: 17994011

49. Nyarko A, Hall J, Hall A, Hare M, Kremerskothen J, Barbar E. Conformational dynamics promote bind-ing diversity of dynein light chain LC8. Biophys Chem. 2011; 159: 41–47. https://doi.org/10.1016/j.bpc.2011.05.001 PMID: 21621319

50. Wang W, Li X, Huang J, Feng L, Dolinta KG, Chen J. Defining the protein-protein interaction network of the human hippo pathway. Mol Cell Proteomics. 2014; 13: 119–131. https://doi.org/10.1074/mcp.M113.030049 PMID: 24126142

51. Bossuyt W, Chen C-L, Chen Q, Sudol M, McNeill H, Pan D, et al. An evolutionary shift in the regulation of the Hippo pathway between mice and flies. Oncogene. 2014; 33: 1218–1228. https://doi.org/10.1038/onc.2013.82 PMID: 23563179

52. Blikstad C, Ivarsson Y. High-throughput methods for identification of protein-protein interactions involving short linear motifs. Cell Commun Signal. 2015; 13: 38. https://doi.org/10.1186/s12964-015-0116-8 PMID: 26297553

53. Necci M, Piovesan D, Dosztanyi Z, Tompa P, Tosatto SCE. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics. 2017; https://doi.org/10.1093/bioinformatics/btx590 PMID: 28968848

54. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SCE. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics. 2015; 31: 201–208. https://doi.org/10.1093/bioinformatics/btu625 PMID: 25246432

55. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. Proteins. 2014; 82 Suppl 2: 127–137.

56. Davey NE, Cyert MS, Moses AM. Short linear motifs—ex nihilo evolution of protein regulation. Cell Commun Signal. 2015; 13: 43. https://doi.org/10.1186/s12964-015-0120-z PMID: 26589632

57. Zhao B, Li L, Lei Q, Guan K-L. The Hippo-YAP pathway in organ size control and tumorigenesis: an updated version. Genes Dev. 2010; 24: 862–874. https://doi.org/10.1101/gad.1909210 PMID: 20439427

58. Hansen CG, Moroishi T, Guan K-L. YAP and TAZ: a nexus for Hippo signaling and beyond. Trends Cell Biol. 2015; 25: 499–513. https://doi.org/10.1016/j.tcb.2015.05.002 PMID: 26045258

59. Li Y, Zhou H, Li F, Chan SW, Lin Z, Wei Z, et al. Angiomotin binding-induced activation of Merlin/NF2 in the Hippo pathway. Cell Res. 2015; 25: 801–817. https://doi.org/10.1038/cr.2015.69 PMID: 26045165

60. Baumgartner R, Poernbacher I, Buser N, Hafen E, Stocker H. The WW domain protein Kibra acts upstream of Hippo in Drosophila. Dev Cell. 2010; 18: 309–316. https://doi.org/10.1016/j.devcel.2009.12.013 PMID: 20159600

61. Zhang N, Bai H, David KK, Dong J, Zheng Y, Cai J, et al. The Merlin/NF2 tumor suppressor functions through the YAP oncoprotein to regulate tissue homeostasis in mammals. Dev Cell. 2010; 19: 27–38. https://doi.org/10.1016/j.devcel.2010.06.015 PMID: 20643348

62. Yu J, Zheng Y, Dong J, Klusza S, Deng W-M, Pan D. Kibra functions as a tumor suppressor protein that regulates Hippo signaling in conjunction with Merlin and Expanded. Dev Cell. 2010; 18: 288–299. https://doi.org/10.1016/j.devcel.2009.12.012 PMID: 20159598

63. Zhao B, Li L, Lu Q, Wang LH, Liu C-Y, Lei Q, et al. Angiomotin is a novel Hippo pathway component that inhibits YAP oncoprotein. Genes Dev. 2011; 25: 51–63. https://doi.org/10.1101/gad.2000111 PMID: 21205866

64. Wells CD, Fawcett JP, Traweger A, Yamanaka Y, Goudreault M, Elder K, et al. A Rich1/Amot complex regulates the Cdc42 GTPase and apical-polarity proteins in epithelial cells. Cell. 2006; 125: 535–548. https://doi.org/10.1016/j.cell.2006.02.045 PMID: 16678097

65. Wennmann DO, Schmitz J, Wehr MC, Krahn MP, Koschmal N, Gromnitza S, et al. Evolutionary and molecular facts link the WWC protein family to Hippo signaling. Mol Biol Evol. 2014; 31: 1710–1723. https://doi.org/10.1093/molbev/msu115 PMID: 24682284

66. Couzens AL, Knight JDR, Kean MJ, Teo G, Weiss A, Dunham WH, et al. Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. Sci Signal. 2013; 6: rs15. https://doi.org/10.1126/scisignal.2004712 PMID: 24255178

67. Hauri S, Wepf A, van Drogen A, Varjosalo M, Tapon N, Aebersold R, et al. Interaction proteome of human Hippo signaling: modular control of the co-activator YAP1. Mol Syst Biol. 2013; 9: 713. https://doi.org/10.1002/msb.201304750 PMID: 24366813

68. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009; 138: 389–403. https://doi.org/10.1016/j.cell.2009.04.042 PMID: 19615732

69. Barron DA, Moberg K. Inverse regulation of two classic Hippo pathway target genes in Drosophila by the dimerization hub protein Ctp. Sci Rep. 2016; 6: 22726. https://doi.org/10.1038/srep22726 PMID: 26972460

70. Kidane AI, Song Y, Nyarko A, Hall J, Hare M, Löhr F, et al. Structural features of LC8-induced self-association of swallow. Biochemistry. 2013; 52: 6011–6020. https://doi.org/10.1021/bi400642u PMID: 23914803

71. Ernkvist M, Birot O, Sinha I, Veitonmaki N, Nyström S, Aase K, et al. Differential roles of p80- and p130-angiomotin in the switch between migration and stabilization of endothelial cells. Biochim Biophys Acta. 2008; 1783: 429–437. https://doi.org/10.1016/j.bbamcr.2007.11.018 PMID: 18164266

72. Zhang L, Yang S, Wennmann DO, Chen Y, Kremerskothen J, Dong J. KIBRA: In the brain and beyond. Cell Signal. 2014; 26: 1392–1399. https://doi.org/10.1016/j.cellsig.2014.02.023 PMID: 24642126

73. Kim M, Kim M, Park S-J, Lee C, Lim D-S. Role of Angiomotin-like 2 mono-ubiquitination on YAP inhibition. EMBO Rep. 2016; 17: 64–78. https://doi.org/10.15252/embr.201540809 PMID: 26598551

74. Moleirinho S, Guerrant W, Kissil JL. The Angiomotins—from discovery to function. FEBS Lett. 2014; 588: 2693–2703. https://doi.org/10.1016/j.febslet.2014.02.006 PMID: 24548561

75. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010; 26: 680–682. https://doi.org/10.1093/bioinformatics/btq003 PMID: 20053844

76. Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 2009; 37: 939–944. https://doi.org/10.1093/nar/gkn1019 PMID: 19106141

77. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011; 8: 785–786. https://doi.org/10.1038/nmeth.1701 PMID: 21959131

78. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004; 338: 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016 PMID: 15111065

79. Davey NE, Edwards RJ, Shields DC. The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Res. 2007; 35: W455–9. https://doi.org/10.1093/nar/gkm400 PMID: 17576682

80. Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, et al. Standardized benchmarking in the quest for orthologs. Nat Methods. 2016; 13: 425–430. https://doi.org/10.1038/nmeth.3830 PMID: 27043882

81. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059–3066. PMID: 12136088

82. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017; 45: D190–D199. https://doi.org/10.1093/nar/gkw1107 PMID: 27899635

83. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59: 307–321. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638

84. Hódi Z, Németh AL, Radnai L, Hetényi C, Schlett K, Bodor A, et al. Alternatively spliced exon B of myosin Va is essential for binding the tail-associated light chain shared by dynein. Biochemistry. 2006; 45: 12582–12595. https://doi.org/10.1021/bi060991e PMID: 17029413