# **Supplementary Material for:**

# Proteome-level assessment of origin, prevalence and function of Leucine-Aspartic Acid (LD) motifs

Tanvir Alam<sup>1,6,#</sup>, Meshari Alazmi<sup>1,8,#</sup>, Rayan Naser<sup>2#</sup>, Franceline Huser<sup>2#</sup>, Afaque A. Momin<sup>2#</sup>, Veronica Astro<sup>4</sup>, SeungBeom Hong<sup>2</sup>, Katarzyna W. Walkiewicz<sup>2,7</sup>, Christian G. Canlas<sup>5</sup>, Raphaël Huser<sup>3</sup>, Amal J. Ali<sup>4</sup>, Jasmeen Merzaban<sup>4</sup>, Antonio Adamo<sup>4</sup>, Mariusz Jaremko<sup>4</sup>, Łukasz Jaremko<sup>4</sup>, Vladimir B. Bajic<sup>1\*</sup>, Xin Gao<sup>1\*</sup>, Stefan T. Arold<sup>2\*</sup>

1. King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Division of Computer, Electrical and Mathematical Sciences & Engineering (CEMSE), Thuwal 23955-6900, Saudi Arabia.

2. King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Division of Biological and Environmental Sciences and Engineering (BESE), Thuwal 23955-6900, Saudi Arabia.

3. King Abdullah University of Science and Technology (KAUST), Division of Computer, Electrical and Mathematical Sciences & Engineering (CEMSE), Thuwal 23955-6900, Saudi Arabia.

4. King Abdullah University of Science and Technology (KAUST), Division of Biological and Environmental Sciences and Engineering (BESE), Thuwal 23955-6900, Saudi Arabia.

5. King Abdullah University of Science and Technology (KAUST), Core Labs, Thuwal, 23955-6900, Saudi Arabia.

\* Correspondence should be sent to STA: stefan.arold@kaust.edu.sa or XG: xin.gao@kaust.edu.sa or VBB: vladimir.bajic@kaust.edu.sa

# Contributed equally

6. Present address: Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Qatar Foundation, Education City, Doha, Qatar

7. Present address: NanoTemper Technologies GmbH, Floessergasse 4, Munich, Germany

8. Present address: College of Computer Science and Engineering, University of Hail, P.O. Box 2440, Hail, 81481, KSA

#### Short Title:

Proteome-wide prediction and function of LD motifs

#### **KEYWORDS**

focal adhesion; protein-ligand interaction; nuclear export signal; machine-learning; evolution

## Content:

Supplementary Material Supplementary Methods Supplementary Figures S1 – S9 Supplementary Tables S1 – S5

# **Supplementary Material**

## Extended description of LDMF-identified proteins

The following characterisation of all LDMF-identified proteins was obtained by combining literature searches with four computational methods, namely PrePPI (a Bayesian framework that combines structural, functional, evolutionary and expression information (Zhang, et al., 2012)), GeneFriends (an RNAseq-based gene co-expression network (van Dam, et al., 2015)), and CoCiter (which evaluates the significance of literature co-citations (Qiao, et al., 2013)). The results from PrePPI, GeneFriends and CoCiter are compiled in **Supplementary Table 3**.

#### Proteins with highly likely LD motif sequences:

Band 4.1-like protein 5 (EPB41L5): The EPB41L5 peptide showed interactions in all methods used, and displayed highest affinities of all LD motifs tested (**Fig. 3A**) towards both FAT and  $\alpha$ -parvin. PrePPI probability scores for the interaction between EPB41L5 and FAK, PYK2 or Talin are >0.9 (**Supplementary Table 3**). EPB41L5 (also called YMO1 and LIMULUS, or yurt in Drosophila) contains an N-terminal FERM domain and a C-terminal flexible region, which harbours the predicted LD motif (residues 634-643) (**Fig. 3B**). EPB41L5/yurt is a critical regulator of the lateral membrane-associated cytoskeleton. It promotes focal adhesion formation by stimulating the interaction between paxillin and integrin. It localizes to focal adhesions where it controls actomyosin contractility and FA maturation (Schell, et al., 2017)

Lipoma-preferred partner (LPP). The LPP peptide showed gualitative and guantitative binding to both FAT and α-parvin. Interactions between LPP and vinculin, α-parvin, PYK2 or GIT1 are suggested by GeneFriends and CoCiter scores (Supplementary Table 3). LPP is a scaffolding protein that plays a structural role in the (dis)assembly of cell adhesions and may be involved in signal transductions from adhesion sites to the nucleus, thus affect activation of gene transcription (Petit, et al., 2000; Petit, et al., 2003). LPP shows many similarities to paxillin family proteins: (i) Its N-terminal half is predicted to be an unstructured region harbouring proline-rich and phospho-tyrosine/threonine sequences in addition to the putative LD motif (residues 123-132); (ii) its C-terminal half contains three LIM domains. (iii) LPP shuffles between the nucleus and cytoplasm and is found at cell adhesions, including focal adhesions; (iv) the putative LPP LD motif sequence overlaps with a functional NES (Gorenne, et al., 2006; Gorenne, et al., 2003; Petit, et al., 2000; Petit, et al., 2003), supported by a PrePPI score of 0.98 for the LPP:XPO1 interaction. Moreover, LPP and FAK appear genetically linked (Gorenne, et al., 2006). LPP and vinculin co-localise at focal adhesions and overexpression of the LPP LIM domains displaces LPP and vinculin from these structures (Gorenne, et al., 2003; Petit, et al., 2000).

Ral GTPase-activating protein subunit alpha-1 (RALGAPA1). The RALGAPA1 peptide showed qualitative and quantitative binding to both FAT and  $\alpha$ -parvin. PrePPI scores suggested RGPA1 interactions with FAK (0.76) and PYK2 (1.0). RALGAPA1 obtained a good coexpression score with the ARF GTPase–activating protein GIT2, and, in additional DA and MST experiments, bound with a  $K_d$  of ~80 µM to the GIT1 FAH domain (GIT1 and GIT2 are close homologues and have an identical LD motif binding site). RALGAPA1 (also called GARNL1 or TULIP1) is the catalytic  $\alpha$ 1 subunit of the heterodimeric RalGAP1 complex. RALGAPA1 functions as an activator of Ras-like small GTPases, including RalA and RalB (Shirakawa, et al., 2009). Activated RalA is involved in cell proliferation, migration, and metastasis. The suggested LD motif resides 100 amino acids upstream of the RapGAP domain, in a region predicted to be flexible (residues 1680-1689).

Serine/threonine-protein phosphatase 2A regulatory subunit B" subunit alpha (PPP2R3A). The PPP2R3A peptide showed qualitative and quantitative binding to both FAT and  $\alpha$ -parvin. The predicted LD motif is located in a flexible region upstream of double EF hand

domains (residues 508-517). PPP2R3A modulates substrate selectivity, catalytic activity and subcellular localisation of protein phosphatase 2A (PP2A) (UniProt, 2015). Indirect evidence suggests that PP2A promotes FAK phosphorylation (Kawada, et al., 1999; Moscardo, et al., 2013), interferes with the DLC1:FAK interaction (Ravi, et al., 2015), and is linked to focal adhesion proteins (Ito, et al., 2000).

Coiled-coil domain-containing protein 158 (CCDC158). The predicted LD motif in CCDC158 showed quantitative and qualitative binding to FAT and interacted specifically with the 1/4 site of FAT in NMR. We only measured significant qualitative binding to  $\alpha$ -parvin. CCDC158 is an 1113-residue protein that contains 3 extended coiled-coil domains. The predicted LD motif region (residues 903-912) is located in a flexible region between the second and third coiled-coil. No literature is available for CCDC158.

*C16orf71 (C16orf71).* C16orf71 is an uncharacterised protein of 520 residues. It is predicted to be mostly disordered, and the predicted LD motif region, which is located in the centre of the protein (residues 267-276), showed qualitative and quantitative binding to both FAT and  $\alpha$ -parvin.

#### Proteins with less likely LD motif sequences:

*Nuclear receptor coactivator 2 [NCOA2*, or steroid receptor coactivator 2 (*SRC-2*)]. NCOA2 is structurally mostly disordered and contains four nuclear receptor box (NR box) LXXLL motifs that mediate hormone-dependent co-activation of several nuclear receptors. A LLXXLXXXL motif in NCOA2 is involved in binding and transcriptional coactivation of CREBBP/CBP (Stashi, et al., 2014). The putative LD motif identified by *LDMF* (residues 805-814) is not part of these motifs, despite the similar consensus. Paxillin family proteins also bind to nuclear receptors, such as the androgen receptor and glucocorticoid receptor (Alam, et al., 2014). The region encompassing the putative NCOA2 LD motif is also predicted to function as an NES, akin to several paxillin LD motifs. NOCA2 obtained a strong CoCiter p-value (0.007) for association with PYK2, and has a large co-expression correlation with GIT2, but failed to show binding to GIT1 FAH in our additional experiments.

*Nuclear receptor coactivator 3 (NCOA3,* or *SRC-3).* Akin to NCOA2, NCOA3 is a scaffolding protein with many known interactors. NCOA3 has three NR box LXXLL motifs to bind to and co-activate several nuclear receptors, and a LLXXLXXXL motif to bind CREBBP/CBP (Stashi, et al., 2014). The predicted LD motif region (residues 799-808) is not part of a known motif.

*Calpastatin (CAST).* CAST is a specific inhibitor of the calcium-dependent cysteine protease calpain. The proposed CAST LD motif (residues 156-165) is a helical protein-protein interaction motif located in an otherwise disordered region; it uses its hydrophobic patch to bind to a helical subdomain of calpain, thus stabilising calpain in its inhibited form (Moldoveanu, et al., 2008). Calpain participates in cell migration and anoikis, and among its substrates are Cas, talin, FAK and PYK2 (Carragher, et al., 2003; Cooray, et al., 1996). Calpain also associates with FAT (Carragher, et al., 2003). Hence a possible competitive interaction of the calpastatin LD motif with FAK and/or PYK2 interaction with CAST could potentially promote cleavage by calpain.

*Cyclic AMP-responsive element-binding protein 3 (CREB3).* CREB3 is a single-pass transmembrane endoplasmic reticulum (ER)-bound transcription factor involved in the unfolded protein response, in cell proliferation and migration, tumor suppression and inflammatory gene expression. The predicted LD motif region (residues 49-58) is located in a flexible acidic transcription activation region downstream of a basic leucine-zipper (bZIP) domain (residues 174-237) and the transmembrane region (residues 255-271).

### Proteins with least likely LD motif sequences:

*Ral GTPase-activating protein subunit alpha-2 (RALGAPA2).* RALGAPA2 is the catalytic  $\alpha$ 2 subunit of the heterodimeric RalGAP2 complex. The putative LD motif (residues 1519-1528) lies in a poorly ordered region. The RALGAPA2:PYK2 interaction has a PrePPI score of 0.93, shows medium-level co-expression with GIT2, but failed to show significant binding to GIT1 FAH.

*C8orf37 (C8orf37).* The 207-amino acid C8orf37 protein is widely expressed, with highest levels in brain and heart, and mutations are associated with ciliopathies and retinal dystrophy (Heon, et al., 2016). The putative LD motif (residues 4-13) is in the disordered N-terminal half of the protein.

# **Supplementary Methods**

# **Computational Methods**

As the number of known LD motifs is small, it becomes an imbalanced dataset problem, which usually causes issues for classification methods. Therefore, we used a two-phase approach for building the prediction model. In the first phase, we considered the known LD motifs as the positive set and the remaining 10-mers extracted from these proteins as the negative set. As expected, these extracted 10-mers can be easily differentiated from the true LD motifs because they do not satisfy sequence patterns, secondary structure patterns or physicochemical patterns of the LD motifs. Therefore, a model trained based on such a trivial negative set may not be practically useful. Yet it provides us a rough predictor by assigning different weights to sequence-, secondary structure- and physiochemical- patterns. In a second phase, we used this predictor to obtain more difficult negative sets. This was done by selecting the 10-mers from the proteins in the Protein Data Bank (PDB) which satisfy some of these patterns according to the first predictor, but not all of them. We then used these new negative sets as well to train the final predictor. This results in an active learning framework to train an LD-motif predictor.

## Features that characterise bona fide LD motifs in silico.

To first determine features that characterise LD motifs *in silico*, we analysed known LD motifcontaining proteins using algorithms to predict protein disorder, secondary and tertiary structures. We found that established LD motifs (paxillin family, DLC1 and RoXan), as well as gelsolin's C-terminal LD-like motif are located within protein regions predicted as disordered (**Supplementary Fig. 1**). Secondary structure prediction assigned a significant  $\alpha$ -helix likelihood to those LD motifs, in agreement with structural studies of paxillin LD motifs 1, 2 and 4, DLC1 and gelsolin (**Fig. 1C**) (Alam, et al., 2014; Hoellerer, et al., 2003; Lorenz, et al., 2008; Nag, et al., 2009; Zacharchenko, et al., 2016) (**Supplementary Fig. 1**). Bona fide LD motifs are therefore computationally characterized as short  $\alpha$ -helical segments within disordered protein regions.

## Initial training data set

Our model uses information from protein sequence content of data-windows of length 10AA. Such windows are denoted as core windows. A core window is shifting one residue ahead. So, if a protein has a length L >= 10 AA residues, then there are L-10+1 possible candidate core window to be considered by scanning the protein sequence as containing a putative LD motif.

By surveying the literature, the known LD motifs were found in Paxillin, Leupaxin, PaxB, Hic-5 (Tumbarello, et al., 2002), RoXaN (Vitour, et al., 2004), and DLC1 (Durkin, et al., 2007) and we selected these LD motifs. This resulted in a set of 18 genuine LD motif windows generated from six proteins. We denote this set as the set of known LD motifs (positive set PS1). All the possible windows of length 10AA from the remaining regions of the above-mentioned six proteins were selected as the core windows of the initial negative set (NS1). This produced a set of 4020 windows from six proteins that formed NS1. To consider the importance of surrounding regions of LD motifs, 20AA residues flanking regions on each side of the scanning window were analysed.

## Feature extraction from protein sequences

From the set of aligned 18 windows with their flanking sequences, position frequency matrix (PFM) was constructed. If the flanking region of scanning window is shorter than 20AA (at N-terminal and C-terminal region) then the positions are filled up by a gap ('-'). PFM was then normalised to produce Position Weight Matrix (PWM) using normalisation technique analogous

to (Bajic, et al., 2003). We only consider twenty IUPAC unambiguous AA codes (<u>http://www.bioinformatics.org/sms/iupac.html</u>) and gap ('-') for building PWM. We built PWM from the scanning core window (PWM<sub>CoreSeq</sub>) which consists of 10 residues, the two flanking regions each with 20 residues produces two other PWMs (PWM<sub>UpSeq</sub>, PWM<sub>DownSeq</sub>) and the whole segment (upstream flanking region + core window + downstream flanking region) of 50 (20 + 10 + 20) AA residues produces the additional PWM. Then, during the scanning of protein sequences, we matched the four PWMs with corresponding window segments to get the respective four matching scores (Bajic, et al., 2003). We also considered the average values of the mapping score from the PWM of core window (PWM<sub>CoreSeq</sub>) and PWM of flanking regions (PWM<sub>UpSeq</sub>, PWM<sub>DownSeq</sub>). Thus, we generated five features for each window. While generating the scores from the core PWM (PWM<sub>CoreSeq</sub>) we used our previous knowledge of the properties of *bona fide* LD motifs (Alam, et al., 2014; Hoellerer, et al., 2003). If there are no acidic residues (Asp or Glu) either at position 0 or 6, we assign the score zero to PWM<sub>CoreSeq</sub>. Proline has a tendency to break the helix. Consequently, if there were two consecutive prolines in core motif we also assigned 0 to PWM<sub>CoreSeq</sub>.

## Feature extraction from secondary structure (SS)

We predicted the secondary structure (SS) of the whole protein using PSIPRED (McGuffin, et al., 2000) against the NR database. Each residue in the 50AA window (core + flanking regions) was tagged as belonging to helix ('H') or coil ('C') or strand ('E'). Gap ('-') was also considered for the windows near N/C-terminal of proteins. From the set of 18 windows that correspond to known LD motifs (with flanking regions), we constructed PFM matrices (analogously as mentioned in the previous section) based on SS annotation of residues. PFM was then normalized to PWM. We built the PWM from the scanning core window (PWM<sub>CoreSS</sub>), the two flanking regions each with 20 residues produces two other PWMs (PWM<sub>UpSS</sub>, PWM<sub>DownSS</sub>) and the whole segment (upstream flanking region + core window + downstream flanking region) of 50 (20 + 10 + 20) AA residues produces the additional PWM. Using PWMs, we were able to generate five features from SS information in the analogous manner as explained in the previous section. In these cases, if the core motif part does not have any helical prediction, we assign zero to the core motif score from PWM<sub>CoreSS</sub>.

#### Feature extraction using AAindex

From Amino Acid Index (AAindex) database (Kawashima, et al., 2008) three physiochemical properties were extracted: hydrophobicity (Backer, et al., 1992), volume, and electric charge (Fauchere, et al., 1988). For each of the 10 residues in a core window, we calculate the AAindex values of the above-mentioned three properties that produced 30 (3\*10) features.

#### Model Development

We generated an initial model based on the initial training data. Since this model is based on data derived from only six proteins and contains a very small number (18) of known LD motifs, we extended the training set by hypothetical LD motifs and additional negative data. For this, we used a procedure (explained below that, among other things, utilizes the initial model) that is likely to generate motifs highly similar to known LD motifs. Once the training set is expanded this way, we retrained the model as we used initially.

#### The Initial Model

We extracted five features using primary sequence information, five features using SS information, and 30 features using AAindex for data-windows as discussed previously. Then we used a support vector machine (SVM) model (Cortes and Vapnik, 1995) with linear kernel (Shawe-Taylor and Cristianini, 2004) to build a predictive model (M1). We used 'svmtrain' function of MATLAB 2012b with default parameter setting to build the model (there was no need

to optimize parameters of the SVM model as the default setting provided an excellent performance).

## LD Motifs from Homologous Proteins

As we have very limited number of known LD motifs, we tried to increase that number using standard protein-protein BLAST (blastp) hits which are similar to motifs (Altschul, et al., 1997). We used the six proteins that contain the known LD motifs for the blastp program and selected the complete sequence of the proteins with the high score of BLAST hits (E-value:1e-7, bit score > 40, against NR database). Then, we applied our M1 to identify the LD motifs from these proteins homologous to the six proteins that contained known LD motifs. In this way, we predicted 40 more LD motifs from these proteins. These additional 40 candidate LD motifs were also considered as correct and used for building our final model.

## Active Learning Dataset from PDB

We downloaded a culling set (Wang and Dunbrack, 2003) of proteins from the Protein Data Bank (PDB) to enhance our negative dataset-. We predicted SS of the full chain using PSIPRED. We built three independent models from the initial dataset based on five sequence features ( $M1_{seq}$ ), five SS features ( $M1_{ss}$ ) and 30 AAindex features ( $M1_{aaindex}$ ). For each of these models, we used an SVM model with linear kernel and default parameter setting.

We applied  $M1_{seq}$  to the culling set to predict windows with LD motifs. These windows formed the set  $S_{seq}$ . Analogously, we generated sets  $S_{ss}$  and  $S_{aaindex}$  using  $M1_{ss}$  and  $M1_{aaindex}$ , respectively. Our hypothesis was that a window that does not belong to the intersection of these three sets is less likely to contain LD motifs. So, we included such windows in the negative set. This has resulted in 2,279 additional negative data-windows used for building the final model.

## The Second Model (M2)

We extracted the features from all (18+40) positive and all (4020+2279) negative data-windows in the same fashion as discussed previously and we used an SVM with the linear kernel to build a predictive model (M2). We used 'svmtrain' function of MATLAB 2012b with default parameters setting to build the final model. This model predicts 13 new LD motif from human proteome. We applied a version of the 18-fold cross-validation (CV) to assess the model accuracy. We divided the negative set randomly into 18 disjoint subsets. At each step of CV, we excluded a different subset from the negative data and the window that corresponds to one of the 18 known LD motifs. Moreover, from the additional 40 positive data (windows) we excluded all windows from proteins homologous to the excluded one to which the known LD motif belongs. This last step is done in order to avoid dependent data in the training set. Then, the model is derived from the remaining data as described in the section above, and it was tested on the excluded data.

#### The Final Model

We experimentally (*in vitro*) verified the 13 new LD motifs and found that four of them show a strong binding affinity ("Highly likely" category) towards their binding partners. So, we integrate these four motifs in the roster of true LD motif and build the final model following the same method described above. This final model predicts eight LD motifs. Three were new LD motifs and five were common to previously predicted 13 LD motifs by M2. Using CV approach, mentioned in the above section, the final model achieved over 88.88% sensitivity and accuracy of 99.97% (**Supplementary Table** 1).

#### Validation of LDMF using Random Sets

To evaluate the robustness of our final model we tested it on random sequences generated by Sequence Manipulation Suite (Stothard, 2000). We generated 1,000 random sequences and applied the model to them. *LDMF* did not predict any LD motif in these sequences.

#### Availability

*LDMF* is available at <u>www.cbrc.kaust.edu.sa/ldmf</u>. For the result mentioned in this manuscript, we used the NR database for PSIPRED predictions (McGuffin, et al., 2000). But for our online *LDMF* server, due to the prohibitive time required to obtain the results from the NR database, we used UNIPROT database for PSIPRED predictions.

# **Bioinformatics**

Prediction of protein disorder: MetaPrDos(Ishida and Kinoshita, 2007) and RaptorX (Kallberg, et al., 2012). Prediction of secondary structure: PSIPRED(McGuffin, et al., 2000) and RaptorX (Kallberg, et al., 2012). Prediction of tertiary structures: SwissModel (Schwede, et al., 2003), RaptorX (Kallberg, et al., 2012). Prediction of transmembrane helices and signal peptides: Phobius (Kall, et al., 2007). Prediction of NES: NetNES1.1 server (Ia Cour, et al., 2004).

# **Biophysical Binding Assays**

#### Overview and Rationale

For initial high-throughput screening, we used three plate assays: 1) differential scanning fluorimetry (DSF) was chosen as a semi-quantitative label-free binding indicator; 2) a direct anisotropy (DA) assay with labelled candidate peptides was chosen to estimate the interaction affinity; and 3) an anisotropy competition assay (ACA) where unlabelled candidate peptides compete against fluorescently labelled known LD motifs, was chosen to assess whether the (unlabelled) candidate motifs bind to the same sites as the known LD motifs. For all candidates, we used microscale thermophoresis (MST) with labelled peptides as an orthogonal quantitative method. ITC was used as an additional label-free method in selected cases to provide an additional binding  $K_d$ , or binding stoichiometry. Nuclear magnetic resonance (NMR) was used in special cases to map binding sites. Peptide sequences included four to eight flanking residues outside the 10-residue core sequence. These additional residues were chosen based on homology modelling, secondary structure and disorder predictions to include helix-capping residues and residues that might additionally contact the LDBDs. Peptides were synthesized with and without a FITC-Ahx N-terminal fluorescent label.

Peptide mimics of paxillin LD4, which were used as positive controls, displayed micromolar  $K_d$  values for FAT and  $\alpha$ -parvin as expected, and competed efficiently against labelled LD4 in ACA (**Fig. 3A, Supplementary Fig. 3**). Although the presence of LD4 resulted in a significant change in melting temperature *Tm* in DSF with FAT, the *Tm* change with  $\alpha$ -parvin was not significant compared to a negative control (a peptide with the scrambled LD4 sequence). This result led us to include an LD2 peptide as a positive control in DSF.

## Protein production

Human  $\alpha$ -parvin-CH<sub>c</sub> (residues 242-372), the FAT domain of human FAK (892-1052), and the rat GIT1 (647-770) were expressed as GST-fusion proteins in *E. coli* BL21 using the expression vectors pGex 6P1, pGexP2, pGex-4T1, respectively. Bacteria were grown in LB medium.  $\alpha$ -parvin-CH<sub>c</sub> and FAT were expressed at 20°C overnight, whereas GIT1 was expressed for 6h at 30°C,  $\alpha$ -parvin-CH<sub>c</sub>,FAT and GIT were purified as described previously (Arold, et al., 2002; Lorenz, et al., 2008; Schmalzigaug, et al., 2007).

## Differential Scanning Fluorimetry

Experiments were performed in 20 mM HEPES pH 7.5, 150 mM NaCl, 2 mM EDTA, 1 mM TCEP. FAT,  $\alpha$ -parvin-CH<sub>c</sub> and GIT1 were used at a concentration of 10  $\mu$ M. Protein stability was assessed for each peptide at 100 and 250  $\mu$ M. SYPRO Orange was used as fluorescent

dye at 1x the protein concentration. The samples were heated from 20°C to 95°C at a rate of 0.03°C/s on a LightCycler 480 II RT-PCR from Roche. To estimate the melting temperature (Tm), a generalized sigmoid was fitted by least squares and the inflection point was computed.

## Direct Anisotropy Assay

Protein was serially diluted in buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 2 mM EDTA, 2 mM DTT, 0.005% Tween-20) and labelled peptides were added at a final concentration of 0.1  $\mu$ M. Fluorescence anisotropy was measured on a PHERAstar FS device (BMG Labtech) using a fluorescence polarization module 485/520/520, at room temperature. Fluorescence anisotropy was determined as: 1000\*(I<sub>1</sub> – I<sub>+</sub>)/(I<sub>1</sub>+2\*I<sub>+</sub>), where I<sub>1</sub> and I<sub>+</sub> are parallel and perpendicular components of fluorescence intensity excited by parallel polarized light. Data were analysed with Origin software using a logistic fit.

## Anisotropy Competition Assay

First, FAT and  $\alpha$ -parvin were titrated, and FITC-Ahx-labelled LD4 was added as described for the direct anisotropy assay. Competition for the LD4 binding site of FAT and  $\alpha$ -parvin was then assessed as follows: the proteins were kept at a concentration corresponding to the  $K_d$  of their interaction with labelled LD4 (10  $\mu$ M for FAT and 25  $\mu$ M for  $\alpha$ -parvin), in the presence of 0.1  $\mu$ M labelled LD4. To that, each non-labelled peptide was added at 100 and 250  $\mu$ M. When competing for the binding site, the unlabelled peptide displaces labelled LD4 resulting in a lower anisotropy. All measurements were performed as for direct anisotropy assay. Values are represented as a ratio to the point estimated to be the  $K_d$  of the protein with LD4 labelled.

## Isothermal Titration Calorimetry

Proteins were dialysed in ITC buffer (20mM HEPES pH 7.5, 150mM NaCl, 1mM EDTA, 1mM TCEP). 1.5 ml of protein solution was placed in the cell at a concentration varying depending on the interaction from 50 to 150  $\mu$ M for FAT and 125  $\mu$ M for GIT1. Peptides were dissolved into the dialysis buffer to a concentration of between 1 to 1.25 mM and placed in the injection syringe. Titrations were performed at 25 °C. As a control, the peptide was titrated into the buffer and the resulting heats subtracted from the protein-binding curve. ITC was performed either on a Nano ITC (TA Instruments), and data were fitted using NanoAnalyze Software, or using a ITC 200 (GE) and data were fitted using Origin Software.

## Microscale Thermophoresis

Serial dilutions of proteins were prepared starting from 630  $\mu$ M (GIT1), 560  $\mu$ M (FAT) or 530  $\mu$ M ( $\alpha$ -parvin) in reaction buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 2 mM EDTA, 2 mM DTT, 0.05% Tween-20). Labelled peptides were added to a final concentration of 0.1  $\mu$ M. The experiment was performed at 20 % LED power and 20, 40 and 60 % MST power in standard capillaries (GIT1) and MST Premium Capillaries (FAT and  $\alpha$ -parvin) on a Monolith NT.115 device at 25 °C (NanoTemper Technologies). Thermophoresis and temperature jump were fitted using the K<sub>D</sub> formula derived from the law of mass action on the provided NT analysis software.

## **Nuclear Magnetic Resonance**

Cells were grown with <sup>15</sup>N-labelled ammonium chloride dissolved in M9 minimal media solution, induced at OD=0.8 with 300uM IPTG and harvested after incubation overnight at 22 °C. Protein samples were purified and NMR samples were prepared by dissolving the <sup>15</sup>N-labelled protein in a 10% D<sub>2</sub>O/90% H<sub>2</sub>O solution with a monomer concentration of 100  $\mu$ L in a total volume of 500  $\mu$ L and pH of 7.5. LD motif-containing peptides were dissolved with FAT gel filtration buffer (20 mM HEPES pH=7.5, 150mM NaCl, 2mM EDTA and 2mM DTT). 2  $\mu$ L of 25 mM 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) sodium salt was added as an internal chemical shift reference for <sup>1</sup>H at 0 ppm. The samples were stable over the course of the NMR experiments.

The <sup>1</sup>H-<sup>15</sup>N HSQC titration experiments were performed at a temperature of 25 °C using a Bruker Avance III 950 MHz NMR spectrometer equipped with a triple resonance inverse TCI CryoProbe. Spectra were acquired with 2048 (<sup>1</sup>H) × 200-256 (<sup>15</sup>N) complex points, a spectral width of 16 ppm for <sup>1</sup>H and 40 ppm for <sup>15</sup>N, and averaged for 36-88 scans depending on sample concentration. Changes in chemical shifts for <sup>1</sup>H and <sup>15</sup>N were measured in ppm ( $\delta H$  and  $\delta N$ ) in comparison to unpublished amino acid backbone NMR assignments of human FAT domain. <sup>15</sup>N shift changes were multipled by a scaling factor  $\alpha$ =0.2, and then the total change in the chemical shift perturbations (CSPi) was calculated following this equation:  $CSPi = \sqrt{\frac{1}{2}} [\delta_H^2 + (\alpha, \delta_N^2)]$  (Williamson, 2013).

# **Data-driven Molecular Docking**

The data-driven HADDOCK 2.1 protocol (van Zundert, et al., 2016) was used to generate the models of complexes for FAT:CCDC158 and FAT:LPP. Crystal structures of FAT (1ow8 and 1ow7) were used for the modelling. Initial models for CCDC158 and LPP were modelled in helical form based on the LD4 peptide. The NMR chemical shift perturbation (CSP) data was used to define the residues, which could be potentially involved in the binding known as active residues. The residues 915, 926, 929, 933, 934, 936, 938, 940, 956, 1031, 1032, 1033, 1035, 1036 and 1038 were marked on FAT helix 1/4 as active residues for FAT:CCDC158. The residues 914, 916, 934, 936, 938, 1022, 1027, 1031, 1032 and 1033 were marked on FAT helix 1/4 and 948, 955, 956, 957, 959, 962, 963, 964, 991 and 1007 were marked on FAT helix 2/3 as active residues for FAT:LPP. The CSP data was only used to define the binding site and not the binding poses. Structures that were listed in the output clusters with best scores were further analysed using PyMol (pymol.org).

## **Cellular Analyses**

## Design and preparation of eGFP-coupled tetra LD motifs

eGFP-LD fusion constructs contained an N-terminal eGFP followed by a HRV3C protease recognition site (LEVLFQGP) and then four times the same LD motif sequence. LD motifs were separated by glycine-serine-threonine linkers of different lengths to enable multivalent associations with LDBDs: LD–GSGST–LD–GSGSTGSGST–LD–GSGSTGSGST–LD–GSGSTGSGST–LD. LD sequences were LD4: TRELDELMASLSD; LPP: EIDSLTSILADLESS; EPB41L5: ATDELDALLASLTENLID; C16orf71: EAWDLDDILQSLQGQ. Constructs were synthesized as gBlock (IDT) fragments separately for the N-terminal eGFP and the C-terminal tetra-LD motif sequences. CPEC cloning (Quan and Tian, 2009) was used to create the construct-including vectors and confirmed by the sequencing.

#### Cell lines, transfection, and antibodies

HeLa cells were cultured in DMEM with 10% FBS and transfected with plasmid DNA using Lipofectamine 3000. For cell spreading and immuno-localization experiments, HeLa cells were plated at low density on fibronectin-coated coverslips, transfected and used for immunofluorescence 24h later, as previously described (Astro, et al., 2011). For live cell imaging, HeLa cells were plated on fibronectin-coated 6-well plates, transfected with GFP-tagged plasmids, manually scratched and recorded 36 h after transfection. The pAb against GFP and Vinculin, and the AlexaFluor 647-conjugated phalloidin were from Thermo Scientific. Fixed cells were observed with the EVOS FL Auto 2 Microscope (Thermo Scientific) using a Plan Apochromat 1.42 NA/60X oil objective (Zeiss).

#### Morphological analysis and functional assays

The measurement of cell area projection, aspect ratio and roundness of transfected HeLa cells spread for 24 hours on fibronectin was evaluated on thresholded images using ImageJ. For wound healing assays, images were captured with a 10x lens at 60-min interval for 30 h using an optical microscope (JuLI<sup>™</sup> Stage Real-Time Cell History Recorder, NanoEntek) equipped with a High-sensitivity monochrome CCD (Sony sensor 2/3") and an automated x-y-z stage, with a 0.3 NA/10X objective (Olympus). During live imaging cells were kept at 37°C and 5% CO2 in a cell incubator (Heracell, 150i, Thermo Scientific). Migration paths were calculated from the nuclear positions of GFP-positive cells obtained from 4 fields per well using two plugins available for ImageJ software (Manual tracking and Chemotaxis tool). The track of each cell was used to measure different parameters of migration: total and Euclidean distances (length of the line segment, calculated between the start and the end point of the cell trajectory), cell velocity and directionality (index of the persistence of the cell movement, given by the ratio between the Euclidean and the total distances. This value may change between 0 and 1, where 1 corresponds to the maximum linearity of the trajectory).

# Supplementary Figure 1. Features that characterise *bona fide* LD motifs *in silico.*

For each known LD motif, we present the secondary structure predictions (SS3: three states, namely H: helix, E: beta strand, C: coil; SS8: eight states, namely H:  $\alpha$  helix, G: 3-helix, I: 5-helix, E: extended  $\beta$  ladder, B:  $\beta$  bridge, T: hydrogen bonded turn, S: bend, L: loop), solvent accessibility (ACC; B: buried; M: medium exposed, E: solvent exposed) and disorder (DISO: order [.] and disorder [\*]) as predicted by the RaptorX server (Kallberg, et al., 2014). Amino acid are numbered starting with 20 positions upstream of the LD motif (unless the LD motif is situated at the N-terminus, which is then taken as number 1).

# 1 Paxillin

LD1							
1	11		21	31			
MDDLDALL	AD LES	STTSHISK	RPVFLSEETP	ΥS			
ССННННН	нн нно	ccccccc	ccccccccc	СС			
LLHHHHHH	нн нні	LLLLLL	LLLLLLLL	$\mathbf{L}\mathbf{L}$			
EEEEEEE	EE EEF	EEEEEE	EEEEEEEEE	ΕE			
	** ***	* * * * * * * *	* * * * * * * * * *	* *			
	1 MDDLDALL CCHHHHHH LLHHHHHH EEEEEEE ******	1 11 MDDLDALLAD LE: CCHHHHHHH HH LLHHHHHHH HH EEEEEEEEE EEI	1 11 MDDLDALLAD LESTTSHISK CCHHHHHHH HHCCCCCCCC LLHHHHHHHH HHLLLLLLL EEEEEEEEE EEEEEEEE ********	11121MDDLDALLADLESTTSHISKRPVFLSEETPCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCLIHHHHHHHHHHLLLLLLLLLLLLLLLLEEEEEEEEEEEEEEEEEEEEEEEEEEE***********************************			

LD2

	1	11	21	31	41
SEQ	QKSAEPSPTV	MSTSLGSNLS	ELDRLLLELN	AVQHNPPGFP	ADEANSSPPL
SS3	ccccccccc	ccccccccc	ннннннн	CCCCCCCCCC	CCCCCCCCCC
SS8	LLLLLLLLL	LLLLLLLL	ннннннн	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEBE	EBMEBBEEBE	EEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

#### LD3

	1	11	21	31	41
SEQ	PLTKEKPKRN	GGRGLEDVRP	SVESLLDELE	SSVPSPVPAI	TVNQGEMSSP
SS3	ccccccccc	ccccccccc	СННННННС	ccccccccc	ECCCCCCCCC
SS8	LLLLLLLL	LLLLLLLL	ГННННННН	LLLLLLLLL	ELLLLLLL
ACC	EEEEEEEEE	EEEEEEEE	EBEEMBEEEE	EEEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

#### LD4

	1	11	21	31	41
SEQ	PQRVTSTQQQ	TRISASSATR	ELDELMASLS	DFKIQGLEQR	ADGERCWAAG
SS3	ccccccccc	ССССНННННН	НННННННН	HHHHCCCCC	ccccccccc
SS8	LLLLLLLL	LLLHHHHHH	ННННННН	HHHHHTLLL	LLLLLHLLL
ACC	EEEEEEEEE	EEEEEEBEE	MBEEBBEEME	EEEEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

#### LD5

	1	11	21	31	41
SEQ	MAQGKTGSSS	PPGGPPKPGS	QLDSMLGSLQ	SDLNKLGVAT	VAKGVCGACK
SS3	ccccccccc	ccccccccc	СННННННН	HHHHCCCCC	ccccccccc
SS8	LLLLLLLL	LLLLLLLL	ГННННННН	HHHHTTLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEE	EEEEEEEE	EEBEEEMEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	*********

# LD1

	1	11	21	31
SEQ	MEELDALLEE	LERSTLQDSD	EYSNPAPLPL	DQ
SS3	ССННННННН	ННННССССС	ccccccccc	CC
SS8	ГГННННННН	HHHHLLLL	LLLLLLLL	$\mathbf{L}\mathbf{L}$
ACC	EEEEEEEE	EEEEEEEEE	EEEEEEEEE	ΕE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* *

# LD2

	1	11	21	31	41
SEQ	YSEAQEPKES	PPPSKTSAAA	QLDELMAHLT	EMQAKVAVRA	DAGKKHLPDK
SS3	ccccccccc	ССССССНННН	ННННННН	ННННННН	ccccccccc
SS8	LLLLLLLL	LLLLHHHHH	НННННННН	НННННННН	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEEE	EEEEBMEEBE	EEEEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD3

	1	11	21	31	41
SEQ	VAVRADAGKK	HLPDKQDHKA	SLDSMLGGLE	QELQDLGIAT	VPKGHCASCQ
SS3	ccccccccc	ccccccccc	СННННННН	HHHHCCCCC	ccccccccc
SS8	LLLLLLLLL	LLLLLLLL	L H H H H H H H H H	HHHHTTLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEE	EEEEBEEME	EEBEEMEMEE	MEEEEBEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# 3 Pax-B

# LD1

	1	11	21	31
SEQ	MATKGLNMDD	LDLLLADLGR	PKSSIKVTAT	VQTTATPSS
SS3	ccccccccc	ННННННСС	ccccccccc	cccccccc
SS8	LLLLLLLL	ННННННТ	LLLLLLLL	LLLLLLL
ACC	EEEEMEMEE	BEEBMEEMEE	EEEEEEEE	MEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * *

# LD2

	1	11	21	31	41
SEQ	VSSQPAPQPP	QQSQQIDGLD	DLDELMESLN	TSISTALKAV	PTTPEEHITH
SS3	CCCCCCCCCC	ccccccccc	НННННННН	ННННННСС	ccccccccc
SS8	LLLLLLLL	LLLLLLHH	ннннннн	ННННННЦ	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEEE	EEEEMEEBE	EEBEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD3

	1	11	21	31	41
SEQ	SQSQPQPYKV	TATNSQPSSD	DLDELLKGLS	PSTTTTTVP	PPVQRDQHQH
SS3	ccccccccc	CCCCCCCCHH	ННННННССС	CCCCCCCCCC	ccccccccc
SS8	LLLLLLLL	LLLLLLHH	HHHHHHTLL	LLLLLLLLL	LLLLLLLL
ACC	EEEEEEMEM	EEEEEEEE	EBMEMBEEME	EEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD4

	1	11	21	31	41
SEQ	NTPNNNNNNN	TNSPKVVHGD	DLDNLLNNLT	SQVKDIDSTG	PTSRGTCGGC
SS3	ccccccccc	CCCCCCCCCH	НННННННН	HHHHCCCCCC	ccccccccc
SS8	LLLLLLLLL	LLLLLLHHH	НННННННН	HHHHLLLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEEE	EBEEBBEEBM	EEBEEMEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD1

	1	11	21	31
SEQ	MEDLDALLSD	LETTTSHMPR	SGAPKERPAE	ΡL
SS3	ссннннннн	HHHCCCCCCC	ccccccccc	СС
SS8	L L H H H H H H H H H	HHHLLLLL	LLLLLLLL	LL
ACC	EEEMEEEEEE	EEEEEEEEE	EEEEEEEEE	ΕE
DISO	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	* *

# LD2

	1	11	21	31	41
SEQ	AAPAAPPFSS	SSGVLGTGLC	ELDRLLQELN	ATQFNITDEI	MSQFPSSKVA
SS3	ccccccccc	ccccccccc	ННННННС	ССССССНННН	HHHCCCCCCC
SS8	LLLLLLLLL	LLLLLLLL	ннннннн	LLLLLHHHH	HHHSLLLLL
ACC	EEEEEEEEE	EEEEEEBM	EBMEBBEEBE	EEEEMMEEB	BEEBEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD3

	1	11	21	31	41
SEQ	SLPSSPSPGL	PKASATSATL	ELDRLMASLS	DFRVQNHLPA	SGPTQPPVVS
SS3	ccccccccc	ССССССНННН	НННННННН	ннннссссс	ccccccccc
SS8	LLLLLLLL	LLLLLHHHHH	НННННННН	HHHHLLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEE	MBEEBBEMBE	EEEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# LD4

	1	11	21	31	41
SEQ	PVVSSTNEGS	PSPPEPTGKG	SLDTMLGLLQ	SDLSRRGVPT	QAKGLCGSCN
SS3	CCCCCCCCCC	CCCCCCCCCC	СННННННН	ннннссссс	ccccccccc
SS8	LLLLLLLLL	LLLLLLLL	LHHHHHHHHH	HHHHTTLLL	LLLLLLLL
ACC	EEEEEEEE	EEEEEEEE	EEEEEEEE	EEBEEEMEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# 5 Roxan

	1	11	21	31	41
SEQ	RTLPSTDSLD	DFSDGDVFGP	ELDTLLDSLS	LVQGGLSGSG	VPSELPQLIP
SS3	ccccccccc	CCCCCCCCH	НННННССС	ccccccccc	ccccccccc
SS8	LLLLLLLLL	LLLLLLLH	HHHHHHTLL	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEEE	EBEEEEEE	EBMEBBEEME	EEEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# 6 DLC1

	1	11	21	31	41
SEQ	SILYSSSGDL	ADLENEDIFP	ELDDILYHVK	GMQRIVNQWS	EKFSDEGDSD
SS3	ccccccccc	CCCCCCCCCH	ннннннн	ННННННН	HHCCCCCCCC
SS8	LLLLLLLL	LLLLLLLH	ннннннн	НННННННН	HHLLLLLL
ACC	EEEEEEEE	EEEEEEMME	EBMEBBEMBM	EMMEEBEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# Supplementary Figure 2. Bioinformatic features of LD motif candidates predicted by other tools.

**Supplementary Figure 2.1**. Computational assessment of the structural context of the 18 potential LD motifs proposed by Brown et al. (Brown, et al., 1998). These motifs were suggested based on a pattern search with the sequence pattern (L,V)(D,E)X(L,M)(L,M)XXL used by Brown et al. (Brown, et al., 1998).

**Extended Results**: 16 out of the 18 suggested LD motifs were predicted to be an integral part of a folded protein domain. In 15 out of these 16 cases, the hydrophobic patch of the suggested LD motif is inaccessible to solvent and hence ligands. In the one remaining case (LTK), the suggested LD motif is part of the catalytically important  $\alpha$ C helix of a protein kinase domain. Thus, unless unlikely large unfolding events occur, these 16 putative motifs cannot function as LD motifs despite containing the correct sequence pattern. For the remaining two of the 18 proteins, the suggested LD motif is part of a signalling peptide that is cleaved *in vivo*, and hence an unlikely candidate. Only the remaining LD sequence from chicken tensin was a plausible candidate, being located in an unstructured region and implicated in FAs (Lo, 2004).

## Summary of Previously suggested LD motifs by Brown et al. (Brown, et al., 1998)

	UNIPROT Entry	Motif sequence and location in protein	Sequence identity of 3D templates for suggested LD motif region
1	P09104; γ-Enolase	90-LDNLMLEL-97	100 % identical; *
2	P05937; Calbindin	211-LDALLKDL-218	98 % identical; *
3	P29376; LTK	556-LDFLMEAL-563	77 % identical; *
4	P10911; DBL	662-LDAMLDLL-669	65 % identical; *
5	P22676; Calretinin	220-LDALLKDI-227	59 % identical; *
6	P55039; DRG	276-LDYLLEML-283	55 % identical; *
7	P29461; PTP2	679-LDFLLSIL-686	42 % identical; *
8	P36010; β-Adaptin	409-LDILLELL-416	40 % identity; *
9	P40421; RDG	163-LDDLLVVL-170	40 % identical; *
10	P38570; Integrin $\alpha E$	375-LDGLLSKL-382	38 % identical; *
11	P52306; RAP1 GDS	27-LDCLLQAL-34	24 % identical
12	P53046; Rho1 GEF	713-LDNMLLFL-720	24 % identical; *
13	<mark>P35579; Myosin HC</mark>	1422-LDDLLVDL-1429	17 % identical; coiled-coil
14	P24216; Hap2	443-LDVLMTS-450	13 % / 43 % identical (depending on fragment length); *
15	<mark>P54762; Eph-2</mark>	3-LDYLLLLL-10	Signal peptide; no 3D template
16	P38650; Dynein HC	1361-LDGLLNQL-1368	No template
17	P51592; E3	1453-LDTLLLTL-1460	No template
18	Q04205; tensin	807-LDVLMLDL-814	No template

\*: available in the Protein Model Portal www.proteinmodelportal.org.

*No shading*: proteins where 3D models can be established with good confidence, showing that their LD motifs are implicate in a 3D fold and hence inaccessible for canonical LD motif interactions.

**Yellow** shaded molecules: no high-quality model exists, but either low-identity structural homology or other functionality make an LD-motif function unlikely.

**Green** shading: no 3D model is available, and strong biological assumptions to rule out LD-motif function are lacking. However, known biological function speak against it, and the motif is highly <u>deg</u>enerate.

**Red** shading: this motif is potentially likely to be a bona fide LD motif, because of its structural characteristics and supporting biological evidence.

# **Computational assessment of proposed LD motif–containing proteins** (Brown, et al., 1998)

Homology models are coloured according to their secondary structure (magenta:  $\alpha$ -helix; cyan:  $\beta$ -strands). The putative LD motif is colored in green, with the LD motif positions 0, 3 and 4 (L<sup>0</sup>XXLL) colored in red. MetaPrDos (http://prdos.hgc.jp/ (Ishida and Kinoshita, 2007)) was used for predicting structural order/disorder from the protein sequence. PHOBIUS (http://phobius.sbc.su.se/ {Kall, 2007 #3127}) was used for prediction of transmembrane helices and signal peptides. The structural analysis was carried out using the SWISS-MODEL (https://swissmodel.expasy.org/ (Arnold, et al., 2006)) and RaptorX (http://raptorx.uchicago.edu/ (Kallberg, et al., 2014)) servers.

# 1 P09104; GAMMA ENOLASE; ENO2

Location in protein: 90-LDNLMLEL-97

Structural Information: 100% Sequence Identity with PDB 2akm. The suggested LD motif is part of the catalytic domain.



# 2 **P05937; CALBINDIN**

Location in protein: 211-LDALLKDL-218

Structural Information: 98% Sequence Identity with PDB Template 2f33A. Forms EF-hand helixturn-helix.



# 3 P29376; LTK

Location in protein: 556-LDFLMEAL-563 Structural Information: 77% Sequence Identity with PDB 3ics. The LD motif is situated in the  $\alpha$ C helix of the protein kinase domain.



# 4 P10911; DBL

Location in protein: 662-LDAMLDLL-669 Structural Information: 65 % sequence identity with dbl-homology domain (DH domain); Template PDB 1kz7.



# P22676; CALRETININ; CAB29

Location in protein: 220-LDALLKDI-227 Structural Information: 59 % sequence identify with PDB 2f33; forms EF-hand helix-turn-helix.



# 5 P55039; DRG

Location in protein: 276-LDYLLEML-283 Structural Information: 55% sequence identity with PDB 4a9a.



## Location in protein: 679-LDFLLSIL-686

Structural Information: 23.7% Sequence Identity with PDB 3oc3A and 42% identical with 2cfv in the PTP domain.



# 7 P36000; β-ADAPTIN

Location in protein: 409-LDILLELL-416 Structural Information: 40 % Sequence Identity to 4uqi.



# 8 P40421; RDGC

Location in protein: 163-LDDLLVVL-170

Structural Information: 40 % sequence identity with PBD 5jjtA; LD motif is inaccessible in the catalytic region.



# 9 P38570; Integrin alpha-E; ITGAE

Location in protein: 375-LDGLLSKL-382 Structural Information: 38% sequence identity with PDB 1na4 in VWFA domain.



Location in protein: 27-LDCLLQAL-34

Structural Information: 24.2% sequence identity with PDB 4hxt in ARM repeat.



# 11 P53046; RHO1 GDP-GTP exchange protein 1; ROM1

Location in protein: 713-LDNMLLFL-720 Structural Information: 17.7% identical with PDB 3kz1 (pictured) or 24% identical with PH domain only of 1xcgA.



## 12 P35579; MYOSIN-9; MYH9

Location in protein: 1422-LDDLLVDL-1429 Structural Information: 17% sequence identity to PDB entry 2efr (tropomyosin) in C-terminal coiled-coil region.

## 13 P24216; HAP2

Location in protein: 443-LDVLMTS-450

Structural Information: 43 % identity with PDB 5krw for residues 385-461, but poor model quality. 13% sequence identify with PDB entry 1gk4, similar to human vimentin coil 2b fragment.

# 14 P54762; Ephrin type-B receptor 1; EphB1

Location in protein: 3-LDYLLLL-10

Structural Information: No structure modelling possible for this region. The region is identified as an extracellular signaling peptide (cleaved during maturation) by Phobius (below).



## 15 P38650; CYTOPLASMIC DYNEIN 1 HEAVY CHAIN 1; DYNC1H1

Location in protein: 1361-LDGLLNQL-1368 Structural Information: No homology model possible. The LD motif is found in the coiled-coil STEM region.

#### PREDICTION RESULT: P38650

 Prediction false positive rate:
 5.0% ‡
 Change FP rate

 2-state prediction
 (Red: Disordered residues Black: Ordered residues)

 1
 KTKPPVTGNLR
 PEEALQALTI
 YEGKFGRLKD
 DREKCAKAKE
 ALELTDTGLL
 50

 51
 SGSEERVQVA
 LEELQDLKGV
 WSELSKVWEQ
 IDQMKEQPWV
 SVQPRKLRQN
 100

 101
 LDGLLNQLKN
 FPARLRQYAS
 YEFVQRLLKG
 YMKINMLVIE
 LKSEALKDRH
 150

 151
 WKQLMKRLHV
 NWVSELTLG
 QIWDVDLQKN
 200

Disorder profile plot



## 16 P51592; E3 UBIQUITIN-PROTEIN LIGASE; HYD

Location in protein: 1453-LDTLLLTL-1460 Structural Information: No homologous structure for modelling.

## 17 Q04205; TENSIN; TNS

Location in protein: 807-LDVLMLDL-814

Structural Information: No 3D template is available. This motif is promising, because an interaction between the homologue tensin3 and FAK and Cas has been reported (Cui et al. Mol Cancer Res. 2003). Tensin is also involved in the function of focal adhesions. The LD motif of tensin is located in a disordered region and predicted helical.

Predi	RED		ULT: Q04205	e FP rate						
2-sta (Red:	te pr Diso	ediction	ck: Ordered residues	3)						
	1 51 101 151	LESHSPSLSS MLDLAPSVHK RSFGTSVGTD PEPRSYGSAP	CSPQPSPLQP SQSVPSAATR PLAKPYSPGP ASILPLSASY	MPPHSHSMPE QDKPAAMLSS LVPAARSTAE SPAGSQQLLV	FPRAPSRREI LSAQRLSGHY PDYTVHEYRE SSPPSPTAPA	EQSIEALDVL AQPTPQVVQP TYTPYSYQPV	50 100 150 200			
Diso	der	profile plot				1.0	_		h	
						0.8 Output 0.6 0.5 0.5 0.5	$\sim$	$\sim\sim\sim\sim$	$\mathcal{S}$	
						诺 0.3 0.2 0.1 0.0 <sub>0</sub> 10 2	0 30 40 50 60	70 80 90 10011012013 residue number	0140150160170180190	(FP rate= 5.0%) Prediction (Meta)

**Supplementary Figure 2.2**. Secondary structure predictions (SS3: three states, namely H: helix, E: beta strand, C: coil; SS8: eight states, namely H:  $\alpha$  helix, G: 3-helix, I: 5-helix, E: extended  $\beta$  ladder, B:  $\beta$  bridge, T: hydrogen bonded turn, S: bend, L: loop), solvent accessibility (ACC; B: buried; M: medium exposed, E: solvent exposed) and disorder (DISO: order [.] and disorder [\*]) for the non-paxillin motifs suggested by SlimSearch4 (Krystkowiak and Davey, 2017), which was the only algorithm which predicted a reasonable number of LD motif candidate in the human proteome (see **Supplementary Table 1**). The feature predictions were established by the RaptorX server (Kallberg, et al., 2014). The suggested LD motif (unless the LD motif is situated at the N-terminus, which is then taken as number 1).

According to this analysis, 27/34 of the suggested sequences appear to have secondary structure or order/disorder features unfitting for known LD motifs. Of the remaining ones, 4/7 lack the typical amino acid features, in particular the presence of additional acidic charges (GAPD1, F16B1, TENC1, CK072). Hence, only 3/34 motifs would remain as plausible candidates (MIAP, SRTD1, AZI1).

	1	11	21	31	41
SEQ	PPPPTQQHCI	TDNSLSLKTP	LECLLTPLPP	SADDNLKTPP	ECLLTPLP
SS3	0000000000	ccccccccc	cccccccccc	00000000000	ccccccc
SS8	LLLLLLLL	LLLLLLLL	LGLSLLLLL	LLLTLLLSLL	LLSLLLL
ACC	EEEEEEEM	EEEEMEMEME	MEMMMMEMEE	EEEEMEEME	EMEEEME
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	* * * * * * * *
2.	O43166 SI1L1_H	UMAN			
	1	11	21	31	41
SEQ	MKPYSSSKDS	SPTLASKVDQ	LEGMLKMLRE	DLKKEKEDKA	HLQAEVQH
SS3	ccccccccc	ссснннннн	ннннннн	ннннннн	HHHHHCC
SS8	LLLLLLLLL	LLLHHHHHHH	ннннннн	ннннннн	HHHHHLL
ACC	EEEEEEEE	EEEEEEEE	EEEBEEEEE	EEEEEEEE	EEEEEEE
DISO	*******	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	*******
3.	O60941 DTNB_H	UMAN			
	1	11	21	31	41
SEQ	DELEQRMSAL	QESRRELMVQ	LEELMKLLKE	EEQKQAAQAT	GSPHTSPT
SS3	СННННННН	НННННННН	ннннннн	нннннссс	ccccccc
SS8	LHHHHHHHH	ннннннн	ннннннн	HHHHHHLLL	LLLLLLL
ACC	EEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEE	EEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	* * * * * * * *

#### 1. E5RHQ5|NPB11 HUMAN

#### 4. O75069|TMCC2 HUMAN

	1	11	21	31	41
SEQ	GPGGALGSPK	SNALYGAPGN	LDALLEELRE	IKEGQSHLED	SMEDLKTQ
SS3	cccccccccc	0000000000	ннннннн	***	HHHHHCC
SS8	LLLLLLLL	LLLLLLLL	ннннннн	**	HHHHHLL
ACC	EEEEEEEE	EEEEEEEE	EEEMMEEMEE	EEEEEEEE	EEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * *	* * * * * * * *

# 5. P12270|TPR\_HUMAN

	1	11	21	31	41
SEQ	SRLEEQMNGL	KTSNEHLQKH	VEDLLTKLKE	AKEQQASMEE	KFHNELNA
SS3	СННННННН	ННННННН	нннннннн	НННННННН	HHHHHCC
SS8	LHHHHHHHHH	H H H H H H H H H H	НННННННН	H H H H H H H H H H	HHHHHLL
ACC	EEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEE	EEEEEEE
DISO					* *
6.	P31949 S10AB_H	UMAN			
	1	11	21	31	41
SEQ	LSFMNTELAA	FTKNQKDPGV	LDRMMKKLDT	NSDGQLDFSE	FLNLIGGL
SS3	ссссннннн	ннсссссннн	ннннннсс	ссссссннн	HHHHCCC
SS8	LLLLHHHHHH	HHHSLLLHHH	HHHHHHHLT	TLSSLELHHH	HHHHHTL
ACC	EEMMMEMBME	MBEEEEEM	BEEBBEEMME	EEEEBMBEM	BMEMBEEE
DISO	**				*
7.	Q14602 ID2B HU	JMAN			
	1 1	11 21			
SEO	SIRKNSLLDH H	RLGISOSKTP V	DDLMSLL		
SS3	0 0000000000	2 2222222222	ccccccc		
SS8	LLLLLLLL I	LLLLLLLL L	LLLLLL		
ACC	EEEEEEEE E	EEEEEEEE E	EEEEEE		
DISO	********	* * * * * * * * *	* * * * * *		
8.	Q14C86 GAPD1_	HUMAN			
8.	Q14C86 GAPD1_	HUMAN	21	31	41
8.	Q14C86 GAPD1_	HUMAN 11 GDOLREDRMA	21 LDNLLANLPP	31 AKPGKSSSLE	41 MTPYNTPO
8. SEQ	Q14C86 GAPD1_ 1 LVNFMKSVMS	HUMAN 11 GDQLREDRMA CCCCHHHHHH	21 LDNLLANLPP HHHHHCCCC	31 AKPGKSSSLE CCCCCCCCC	41 MTPYNTPQ cccccccc
8. SEQ SS3	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL	31 AKPGKSSSLE CCCCCCCCC LLLLLLLLL	41 MTPYNTPQ CCCCCCCC LLLLLLL
8. SEQ SS3 SS8 ACC	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMMEM	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEE
8. SEQ SS3 SS8 ACC	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMNEM *****	21 L D N L L A N L P P H H H H H H H C C C C H H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * *	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE *****	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE *******
8. SEQ SS3 SS8 ACC DISC 9.	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LILEEEELLL EEEEEEEEE XXXXXXXXXXXXXXXXXX	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHHH EEEEEEMMEM *********	21 LDNLLANLPP HHHHHCCCC HHHHHHTSLL BEEBMEEMEE *******	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE *******	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******
8. SEQ SS3 SS8 ACC DISC 9.	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE A A A A A A A A A A A A A A A A A A A	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMMEM ***********	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * *	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE *********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE *******
8. SEQ SS3 SS8 ACC DISC 9.	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEEELLL EEEEEEEEE AAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ********************************	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE
8. SEQ SS3 ACC DISC 9. SEQ SEQ	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE AAAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMNEM ************************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M * * * * * * * * * * 21 L D S L L K T L D D H H H H H H H H H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE xxxxxxxxxxxxxxx	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHHH EEEEEEMMEM **************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ****************************	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE AAAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMNEM ************************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ********** 21 LDSLLKTLDD HHHHHHHHH HHHHHHHHH	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEE
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC DISO	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE A A A A A A A A A A A A A A A A A A A	HUMAN 11 GDQLREDRMA CCCCHHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ****************************	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCC LLLLLLL EEEEEEEE ******
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC DISO 10.	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LILEEEELLL EEEEEEEEE 4 Q16760 DGKD_H 1 TESSEESEVM CCCHHHHHHH LLLHHHHHHH EEEEEBEEM 4 5 W0V3 F16B1	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ****************************	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEEE ******
8. SEQ SS3 SS8 ACC 9. SEQ SS3 SS8 ACC DISO DISO 10.	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE A A A A A A A A A A A A A A A A A A A	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * * 21 L D S L L K T L D D H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCC LLLLLL EEEEEEEE *******
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC DISC DISC 10.	Q14C86 GAPD1_ 1 LVNFMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEE 4 Q16760 DGKD_H 1 TESSEESEVM CCCHHHHHHH LLLHHHHHHH EEEEEBEEM 4 4 2 SW0V3 F16B1_ 1	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * * 21 L D S L L K T L D D H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEE **********	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEEE *******
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC DISO 10. SEQ	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE AAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * * 21 L D S L L K T L D D H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEEE ********
8. SEQ SS3 SS8 ACC 9. SEQ SS3 SS8 ACC DISO 10. SEQ SS3 SS8 ACC SS3 SS8	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE AAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 LDNLLANLPP HHHHHHCCCC HHHHHHTSLL BEEBMEEMEE ********************************	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEEE ********
8. SEQ SS3 ACC DISC 9. SEQ SS3 ACC DISO 10. SEQ SS3 SS3 ACC	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE AAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMNEM ************************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * * 21 L D S L L K T L D D H H H H H H H H H H H H H H H H H H B M E B M E E E E E * * * * * * * * 21 L E Q M L D I L V Q H H H H H H H H H H H H H H H H H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLLL EEEEEEEEE ********* 41 PPPTIAEE CCCCCCCC LLLLLLL EEEEEEEE ******** 41 GPCMEYLL CCCHHCCC LLHHHHHL
8. SEQ SS3 SS8 ACC DISC 9. SEQ SS3 SS8 ACC DISO 10. SEQ SS3 SS8 ACC DISO	Q14C86 GAPD1_ 1 LVN FMKSVMS CCCCEECCCC LLLEEELLL EEEEEEEEEE AAAAAAAAAAAAAAA	HUMAN 11 GDQLREDRMA CCCCHHHHH LLLLHHHHHH EEEEEEMMEM **************************	21 L D N L L A N L P P H H H H H H C C C C H H H H H H T S L L B E E B M E E M E E * * * * * * * * * * * 21 L D S L L K T L D D H H H H H H H H H H H H H H H H H H B M E B M E E E E E * * * * * * * 21 L E Q M L D I L V Q H	31 AKPGKSSSLE CCCCCCCCC LLLLLLLL EEEEEEEEEEEEEEE	41 MTPYNTPQ CCCCCCCC LLLLLL EEEEEEEE ******** 41 PPPTIAEE CCCCCCCC LLLLLL EEEEEEEE ********

# 11. Q63HR2|TENC1\_HUMAN

1	11	21	31	
SEQ MKSSGPVERL	LRALGRRDSS	RAASRPRKAE	PHSF	
SS3 CCCCCHHHHH	HHHHCCCCCH	HHHCCCCCCC	CCCC	
SS8 LLLLLHHHHH	HHHHLLLLH	HHHLLLLL	LLLL	
ACC EEEEME BMEB	MEEMEEEEE	EEEEEEEE	EEEE	
DISO * * * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	* * * *	
12. Q7Z3Z2 RD3_HU	JMAN			
1	11	21	31	41
SEQ WLRWNEAPSR	LSTRSPAEMV	LETLMMELTG	QMREAERQQR	ERSNAVRK
SS3 CCCCCCCCCC	CCCCCHHHHH	НННННННН	H H H H H H H H H H	HHHHHCC
SS8 LLLLLLLLL	LLLLHHHHH	ннннннн	H H H H H H H H H H	HHHHHLL
ACC EEEMEEEEEE	EEEEEEEE	MEEEEEEEEE	EEEEEEEE	EEEEEEE
DISO ********	******		* * * * * * * * * *	* * * * * * * *
13. Q8IV76 PASD1	HUMAN			
1	11	21	31	41
SEQ EQLEERTWLL	HDAIQNQQNA	LELMMDHLQK	QPNTLRHVVI	PDLQSSEA
SS3 CHHHHHHHHH	***	ннннннс	0000000000	ccccccc
SS8 LHHHHHHHHH	H H H H H H H H H H	нннннннт	SLLLLLLLL	LLLLLL
ACCEEEEEEEEE	EEEEEEEE	E E E E M E E B E E	EEEEEEEE	EEEEEEE
DISO *			. * * * * * * * * *	* * * * * * * *
14. Q8IWP9 CC28A_	HUMAN			
1	11	21	31	41
1 SEQ LYGELEELPE	11 DKRKTASDSN	21 LDRLLSDLEE	31 LNSSIQKLHL	41 ADAQDVPN
1 SEQ LYGELEELPE SS3 CCCCCCCCC	11 DKRKTASDSN HHHHHHHHH	21 LDRLLSDLEE HHHHHHHHH	31 LNSSIQKLHL HHHHHHHH	41 ADAQDVPN CCCCCCCC
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH	41 ADAQDVPN CCCCCCCC HLLLLLL
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEEE	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH EEEEEEEEE	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEE
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ********	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *****	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH EEEEEEEEE *******	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEEE * * * * * * * *
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ******** 15. Q8N3J3 CQ053_1	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH EEEEEEEEE *****	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEE * * * * * * * *
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ********* 15. Q8N3J3 CQ053_1 1	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEE * * * * * * * * * * *	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEEE *******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ********* 15. Q8N3J3 CQ053_1 1 SEQ TAQNLEAEAS	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H E E E E E E E E E * * * * * * * * * * * 21 L D G L L S E L P E	31 LNSSIQKLHL HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ********** 15. Q8N3J3 CQ053_1 1 SEQ TAQNLEAEAS SS3 CCCCCCCCC	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H E E E E E E E E E E * * * * * * * * * * * 21 L D G L L S E L P E C C C H C C C C C C	31 LNSSIQKLHL HHHHHHHH HHHHHHHH EEEEEEEEE **********	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEE *********	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCC HLLLLLL EEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO *********** 15. Q8N3J3 CQ053_1 1 SEQ TAQNLEAEAS SS3 CCCCCCCCC SS8 LLLLLLLL ACC EEEEEEEEE	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE *********** 31 DFFCGTSS CCCCCCCC LLLLLLL EEEEEEEE	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LLLLLLLH ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEEEEEEEEEE	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCC HLLLLLL EEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS3 LILLLLL ACC EEEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	21 LDRLLSDLEE HHHHHHHH HHHHHHHH EEEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEEE ******
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLLLLH ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH EEEEEEEEE HUMAN 11 PEEELPEADD CCCCCCCCC LLLLLLLLL EEEEEEEEE **********	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCCC HLLLLLL EEEEEEEE ********
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLLILL ACC EEEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEE ******* 41 PRSLLCSI
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLLILL ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH EEEEEEEEEE HHHHHHHHHH EEEEEEE	21 LDRLLSDLEE HHHHHHHHH HHHHHHHHH EEEEEEEEEE ********	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEE ****************	41 ADAQDVPN CCCCCCCC HLLLLL EEEEEEEE ******** 41 PRSLLCSI CCCCCCCC
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLLILL ACC EEEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH EEEEEEEEEE HHHHHHHHHHH EEEEEE	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H E E E E E E E E E E * * * * * * * * * * * 21 L D G L L S E L P E C C C H C C C C C C L L L H H L L L L E E E E E E E M * * * * * * * * * * * 21 21 L L H H L L L L E E E E E E M E E E E E E M * * * * * * * * * *	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEEE ***************	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEEE ******** 41 PRSLLCSI CCCCCCC LLLLLLL
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLLILL ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH EEEEEEEEEE ************ HUMAN 11 PEEELPEADD CCCCCCCCC LLLLLLLLL EEEEEEEEEE **********	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H E E E E E E E E E * * * * * * * * * * * 21 L D G L L S E L P E C C C H C C C C C L L L H H L L L L E E E E E E E M E E * * * * * * * * * * * 21 L E H L L A R L L K H	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEEE ***************	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEEE ******** 41 PRSLLCSI CCCCCCC LLLLLLL EEEEEEEE
1 SEQ LYGELEELPE SS3 CCCCCCCCC SS8 LILLILIH ACC EEEEEEEEE DISO ************************************	11 DKRKTASDSN HHHHHHHHH EEEEEEEEEE *****************	21 L D R L L S D L E E H H H H H H H H H H H H H H H H H H E E E E E E E E E E * * * * * * * * * * * * 21 L D G L L S E L P E C C C H C C C C C C L L L H H L L L L E E E E E E E M E E * * * * * * * * * * * * 21 L E H L L A R L L K H	31 LNSSIQKLHL HHHHHHHHH EEEEEEEEEE ***************	41 ADAQDVPN CCCCCCC HLLLLL EEEEEEEE ******** 41 PRSLLCSI CCCCCCC LLLLLLL EEEEEEEE ******

# 17. Q8NDD1|CA131\_HUMAN

SEQ       FTMSQEQGPG       SSTPPSSPIL       LDALLQNLYD       FGGTEGETEQ       KKIIKKRE         SS       CCCCCCCCC       CCCCCCCCHHH       HHHHHHHC       CCCCCCCCHHH       HHHHHHHC       SCCCCCCCCHHH         SS       LLLLLLL       LLLLLLLH       HHHHHHHC       TTLLLHHH       HHHHHHHC       SCCCCCCCHHH         SS       LLLLLLLH       LLLLLLH       HHHHHHC       SCCCCCCCCC       SCCCCCCCCC       SCCCCCCCCC         ISO       I       11       21       31       41         SS       CCCCCCCCC       CCCCCCCCC       SCCCCCCCC       SCCCCCCCCC         SS       CCCCCCCCC       CCCCCCCCC       SCCCCCCCCC       SCCCCCCCCC         SS       CCCCCCCCCC       CCCCCCCCC       SCCCCCCCCC       SCCCCCCCCC         SS       CCCCCCCCCC       CCCCHHHHH       SEQ       SS       SCCCCCCCCC         SS       CCCCCCCCC       HUMAN       21       31       41         SS       SS       CCCCCCCCC       HUMAN       SCCCCCCCCHHHHH       SS         1       11       21       31       41       SCCCCCCCCCHHHHH         SS       LLLLLHHHH       SCCCCCCCCCHHHHHH       SS       SCCCCCCCCCHHHHHH       SS       SCCCCCCCCCC <td< th=""></td<>
SS3       CCCCCCCCCC CCCCCHH       HHHHHHHCC       CCCCCCCCHHH       HHHHHHHCC         SS3       LILILILLI       LILILILLI       HHHHHHHCC       TTILILHHHH       HHHHHHHCC         SS3       LEEEBEEEE       EEEBEEEE       EEEBEEEE       EEEBEEEE       EEEBEEEE       EEEBEEEE         ISO       1       11       21       31       41         SS3       CCCCCCCCC       CCCCCCCCC       CCCCCCCCC       CCCCCCCCCC         SS3       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC         SS3       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC         SS3       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC         SS3       CCCCCCCCCHHHH       HHHHHHHH       LLILILLL       LIGGGGL         SS3       CCCCCCCCCHHHH       HHHHHHHH       LLILILLL       LIGGGGL         SS3       CCCCCCCCCHHHH       HHHHHHHHH       LLILILLL       LIGGGGL         SS3       CCCCCCCCCHHHH       HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
S8       LILLLILLL       LILLLILLH       HHHHHHH       TILLLHHHH       HHHHHHH         LGG       EEEEEEEEE       EEEEEENEEE       EEEEEEEEE       EEEEEEEE       EEEEEEEE         I8.Q8TC57[MIAP_HUMAN       1       11       21       31       41         SEG       MLPSTFPLLP       EDPHDDSLKN       VESMLDSLEL       EPTYNPLHVQ       SHLYSHLS         S8       CCCCCCCCC       CCCCHHHHH       HHHHHHL       LLLLLLLLL       LLGGGGL         S8       CCCCCCCCC       CCCHHHHH       HHHHHHL       LLLLLLLLL       LGGGGL         S8       CCCCCCCCC       CCCHHHHH       HHHHHHL       LLLLLLLLL       LGGGGL         S9       LLLLLLLL       LLLGGGGL       EEEEENEEE       EEEEENEEE       EEEENEEE         9.Q8WZA0 LZIC_HUMAN       1       21       31       41         S9       CCCCCCCCHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         S9       LLLLLHHHHH       HHHHHHHHHHHHHHHH       HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
1       1       21       31       41         SEQ       MLPSTFPLLP       EDPHDDSLKN       MHEBBEENEE       EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
DISO       18. Q8TC37 M1AP_HUMAN         1       11       21       31       41         SEQ       MLPSTFPLLP       EDPHDDSLKN       VESMLDSLEL       EPTYNPLHVQ       SHLYSHLS         SS3       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC         SS3       LLLLLLLLL       LLILLLLLL       LIGGGGL       SEMBEEN       EENEEBMEE       ENNEEMEE         DISO       ************************************
18. Q8TC57 M1AP_HUMAN         1       11       21       31       41         SEQ       MLPSTFPLLP       EDPHDDSLKN       VESMLDSLEL       EPTYNPLHVQ       SHLYSHLS         SS3       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC       CCCCCCCCCC         SS3       LILILILILI       LILIHHHHHH       HHHHHHIL       LILILILILI       LIGGGGL         SS3       CCCCCCCCHHANHHH       HHHHHHHIL       LILILILILI       LIGGGGL         SS3       CCCCCCCHHAN       HHHHHHHH       HHHHHHHIL       LILILILILI       LIGGGGL         SS3       CCCCCCCHHH       HHHHHHHHH       HHHHHHHHH       HHHHHHHHH       HHHHHHHH         SS3       CCCCCCCHHHH       HHHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         SS3       CCCCCCCCHHHH       HHHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         SS4       LILILLHHH       HHHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         SS4       LILILLHHHH       HHHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         SS6       LILILLHHHH       HHHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         SS6       LILILLHHH       HHHHHHHHHHH       HHHHHHHH       HHHHHHHH<
1       11       21       31       41         SEQ       MLPSTFPLLP       EDPHDDSLKN       VESMLDSL       EL EPTYNPLHVQ       SHLYSHLS         SS3       CCCCCCCCC       CCCCHHHHHH       HHHHHLL       LLLLLLLLL       LLLLLLLLL       LLLLHHHHHH         NCO       EEEEENEEEE       EENEEEMEEE       EENEEEMEEE       EENEEEMEE       HHHHHLL       LLLLLLLLL       LLGGGGL         NO       +************************************
SEQMLPSTFPLLPEDPHDDSLKNVESMLDSL ELEPTYNPLHVQSHLYSHLSSS3CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCSS3LLLLLLLLLLLHHHHHHLLLLLLLLLLGGGGLMC00EEEEEMEEEEEENEEEMEEEEEENEEBMBEMMBMEEMEDISO*********************************
SS3       CCCCCCCCCC       CCCCHHHHHH       HHHHHCCCC       CCCCCCCCC       CCCCCCCCC         SS3       LILILILILI       LILILILILI       LILILILILI       LIGGGGL         SS3       LILILILILI       LILILILILI       LIGGGGL         SS3       LILILILILI       LILILILILI       LIGGGGL         SS3       LILILILILI       LILILILILI       LIGGGGL         SS3       CCCCCCCHHH       HHHHHH       HHHHHHHH         SS3       CCCCCCCHHH       HHHHHHHH       EECREELDTD         SS3       CCCCCCCHHH       HHHHHHHH       HHHHHHHH         SS4       LILILIHHHH       HHHHHHHH       HHHHHHHH         SS4       LILILIHHHH       HHHHHHHH       HHHHHHHH         SS4       CCCCCCCCHHHH       HHHHHHHH       HHHHHHHH         SS4       LILILHHHH       HHHHHHHH       HHHHHHHH         SS4       LILIHHHHH       HHHHHHHHH       HHHHHHHH         SS4       LILIHHHHH       HHHHHHHHH       HHHHHHHH         SS6       CCCCCCCCCC       CCCCHHHHH       HHHHHHHH         SS6       LILEELILI       LHHHHHHHH       HHHHHHHH         SS6       LICCCCCCCCCC       CCCCCCCCCC       HHHHHHHH         SS6       LIC
SS8       LILILILILI       LILILILILI       LILILILILI       LILILILILI       LIGGGGL         NCC       EEEEEMEEEE       EEMEEEMMEE       EEMEEEMMEE       EEEMEEBMEE       EEEMEEBMEE       LILILILILI       LIGGGGGL         19. Q8WZA0 LZIC_HUMAN       1       11       21       31       41         SEQ       MASRGKTETS       KLKQNLEEQL       DRINQQLQDL       EECREELDTD       EYEETKK         SS3       CCCCCCCCHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         ACC       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE         DISO
NCC       EEEEENEEEE       EENEEEMMEE       BEENBEEMEE       EEENEEBMBE       NNBMEEME         19. Q8WZA0 LZIC_HUMAN       1       11       21       31       41         SEQ       MASRGKTETS       KLKQNLEEQL       DRLMQQLQD       EECREELDTD       EYEETKK         SS3       CCCCCCCHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH       HHHHHHHH         ACC       EEEEEEEEEE       EEEEEEEEEE       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE         DS0       ************************************
DISO ************************************
19. Q8WZAO LZIC_HUMAN 1 11 21 31 41 SEQ MASRGKTETS KLKQNLEEQL CCCCCCCHHHH HHHHHHHH HHHHHHHHH HHHHHHHHH HHHHHHHH
1 11 21 31 41 SEQ MASRGKTETS KLKQNLEEQL DRLMQQLQDL EECREELDTD EYEETKK SS3 CCCCCCCHHHH HHHHHHHH HHHHHHHH HHHHHHHH
SEQ MASRGKTETS KLKQNLEEQL DRLMQQLQDL EECREELDTD EYEETKK SS3 CCCCCCHHHH HHHHHHHHHHHHHHHHHHHHHHHHH
SS3 CCCCCCCHHHH HHHHHHHH HHHHHHHH HHHHHHHH
SSB       LIILLIHHHH HHHHHHHH       HHHHHHHHH       HHHHHHHHH       HHHHHHHHH         ACC       EEEEEEEEE       EEEEEEEEEE       EEEEEEEEE       EEEEEEEEE         DISO       ********       ********       ********       ********         1       11       21       31       41         SEQ       LELWMHTDPV       SQKNESVRNQ       VEDLLATLEK       SGAGVPAVIL       RRPNQSQP         SS3       CCCCCCCCC       CCCCHHHHH       HHHHHHHH       HLLLLILLEE       LLLLILLIL         ACC       EEMMBMEEEE       EEEEEBMEM       WEBBENBEE       EEEEEEBMEM       EEEEEEEE         DISO       ************************************
ACO       EEEEEEEEEE       EEEEEEEEEE       EEEEEEEEEE       EEEEEEEEEE         DISO       ************************************
DISO ***** ******************************
20. Q92542 NICA_HUMAN 1 11 21 31 41 SEQ LELWMHTDPV SQKNESVRNQ SS3 CCCCCCCCC CCCHHHHH ACC EEMMBMEEEE EEEEEBMEM DISO ************************************
1       11       21       31       41         SEQ       LELWMHTDPV       SQKNESVRNQ       VEDLLATLEK       SGAGVPAVIL       RRPNQSQP         SS3       CCCCCCCCC       CCCHHHHHH       HHHHHHHH       HCCCCEEE       CCCCCCCCC         SS8       LLLEELLLL       LIHHHHHHHH       HHHHHHHH       HLLLLLEE       LLLLLLLL         ACC       EENNBMEEEE       EEEEEBMEM       BMEBBEMBEE       EEEEEBMM       EEEEEEEE         DISO       ************************************
1       11       21       31       41         SEQ       LELWMHTDPV       SQKNESVRNQ       VEDLLATLEK       SGAGVPAVIL       RRPNQSQP         SS3       CCCCCCCCCC       CCCHHHHH       HHHHHHH       HCCCCCEEE       CCCCCCCCC         SS8       LLEELLLL       LIHHHHHHH       HHHHHHHH       HLLLLEEE       LLLLLLL         ACC       EEMMBMEEEE       EEEEEBMEM       BMEBBEMBEE       EEEEEEBMM       EEEEEEEE         DISO       ************************************
SEQ LELWMRTDPV SQRNESVRNQ VEDLLATLER SGAGVPAVIL RRPNQSQP SS3 CCCCCCCCC CCCHHHHHH ACC EEMNBMEEEE EEEEEBMEM DISO ************************************
SS3       CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
ACC EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
1       11       21       31         SEQ       MLEDSESSYE       PDELTKEMAH       LEGLMKDLNA       ITTA         SS3       CCCCCCCCCCC       HHHHHHHHH       HHHHHHHH       HHCC         SS8       LILILILIL       HHHHHHHHH       HHHHHHHH       HHLL         DISO       ************************************
21. Q92859 NEO1_HUMAN 1 11 21 31 SEQ MLEDSESSYE PDELTKEMAH SS3 CCCCCCCCCC HHHHHHHHH ACCC EEEEEEEEEE EEEEEEEEEEEEEEEEEEEEEEE
1       11       21       31         SEQ       MLEDSESSYE       PDELTKEMAH       LEGLMKDLNA       ITTA         SS3       CCCCCCCCCC       HHHHHHHHH       HHHHHHHH       HHCC         SS8       LLLLLLLL       HHHHHHHHH       HHHHHHHH       HHHHHHHH         ACC       EEEEEEEEE       EEEEEEEEE       EEEEEEEEE         DISO       ************************************
1112131SEQMLEDSESSYEPDELTKEMAHLEGLMKDLNAITTASS3CCCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCSS8LLLLLLLLHHHHHHHHHHHHHHHHHHHHLLACCEEEEEEEEEEEEEEEEEEEEEEEEEEEDISO***
SEQMLEDSESSYEPDELTKEMAHLEGLMKDLNAITTASS3CCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCSS8LLLLLLLLLHHHHHHHHHHHHHHHHHHHHHLLACCEEEEEEEEEEEEEEEEEEEEEEEEEEEDISO*********************************
SS3CCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCSS8LLLLLLLLHHHHHHHHHHHHHHHHHHHHHHHHHHHACCEEEEEEEEEEEEEEEEEEEEEEEEEEEDISO*********************************
SS8       LLLLLLLL HHHHHHHHHH HHHHHHHHHHHHHHHHHH
ACC EEEEEEEEE EEEEEEEEE EEEEEEEEEEEEEEE
DISO ******** ***************************
22. Q96FS4 SIPA1_HUMAN
1 11 21 31 41
SEQ GTPKSDAEPE PGNLSEKVSH LESMLRKLQE DLQKEKADRA ALEEEVRS
SS3 CCCCCCCCC CCCARAAAAA AAAAAAAAA AAAAAAAA
SS3 LILLLLL LLLHHHHHH HHHHHHHHHHHHHHHHHHHH
SS3 LLLLLLLL LLLHHHHHHH HHHHHHHHHHHHHHHH

23. Q96GC5|RM48\_HUMAN

	× · _				
	1	11	21		
SEQ	LSVKEHTEED	FKGRFKARPE	LEELLAKLK		
SS3	сссссссннн	ннсссссснн	нннннсс		
SS8	LLLLLLHHH	HHHLLLLHH	H H H H H H H L L		
ACC	EEEEMEEEM	EEEEEEEEE	BEEMMEEME		
DISO	* * * * * * * * * *	* * * *	* * * * *		
24.	Q9BUN5 CC28B	HUMAN			
	1	11	21	31	41
SEQ	EDGVTEGLPE	EQKKTMADRN	LDQLLSNLED	LSNSIQKLHL	AENAEPEE
SS3	сссссссссн	НННННННН	ннннннн	НННННННН	HHCCCCCC
SS8	LLLLLLLLH	ннннннн	нннннннн	ннннннн	HHLLLLL
ACC	EEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEE	EEEEEE
DISO	* * * * * * * * * *	*****		*	* * * * * * * *
25.	Q9H3T2 SEM6C	HUMAN			
	1	11	21	31	41
SEQ	RVLVRPPPPG	CPGQAVEVIT	LEELLRYLHG	PQPPRKGAEP	PAPLTSRA
SS3	0000000000	CCCCCEEECC	нннннннс	0000000000	ccccccc
SS8	LLLLLLLLL	LLLLEEELL	ннннннн	LLLLLLLL	LLLLLLL
ACC	EEEEEEEE	MEEEEEMEM	BMMBBEMBEE	EEEEEEEE	EEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	*******	* * * * * * *
26.	Q9NVE4 CCD87	HUMAN			
	1	11	21	31	41
SEQ	LPPLLGVVTR	HPAAGHRLEE	LEKMLRNLQE	EEASGQWDPQ	PPKSFPLH
SS3	ссссннннс	ссснннннн	нннннннн	HHHCCCCCCC	ccccccc
SS8	LLLHHHHHHL	LLLHHHHHHH	ннннннн	HHHTLLLLL	
ACC	EEEEEEEEE	EEEEEEMEE	BMEMBEEBEE	EEEEEEEE	EEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * * *	* * * * * * * *
27.	Q9P1Z9 CC180_1	HUMAN			
	1	11	21	31	41
SEQ	KRMEQHRQKH	SLESQVQEAH	LDRLLDQLRQ	QSDKETLAFH	LEKVKDYL
SS3	СННННННН	СННННННН	нннннннс	СССНННННН	Н Н Н Н Н Н С С
SS8	1	H H H H H H H H H H	ннннннн	SLLHHHHHHH	HHHHHLL
ACC	EEEEEEEE	EEEEEEEE	EEEEEEEEEE	EEEEEBEEM	EEEBEEEE
DISO	******				*
28.	Q9UHF0 TKNK	HUMAN			
	1	11	21	31	41
SEQ	SKRDPDLYQL	LQRLFKSHSS	LEGLLKALSQ	ASTDPKESTS	PEKRDMHD
SS3	ссссннннн	ннннссссс	ннннннн	0000000000	CHHHCCCC
SS8	LLLLHHHHHH	HHHHHSLLL	ннннннн	LLLLLLLL	LHHHLLLL
ACC	EEEEEBMEM	BMEBMEEEME	MEEBBEEBEE	EEEEEEEE	EEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * *	* * * * * * * * *	* * * * * * * * *	* * * * * * * *

# 29. Q9UHV2|SRTD1\_HUMAN

1	11	21	31	41
SEQ KPGPEDGP	GK EEAPELDE	AE LDYLMDVL	VG TQALERPI	PGP GR
SS3 CCCCCCC	cc ccccccc <mark>H</mark>	нн нннннн	HC CCCCCCC	ccc cc
SS8 LLLLLLL	LL LLLLLL <mark>H</mark>	нн нннннн	HH LLLLLL	LLL LL
ACC EEEEEE	EE EEEEEEE	EE EEEEEE	EE EEEEEE	EEE EE
DISO ******	** *******	** ******	** ******	* * * * *
30. Q9UKX3 MYH13	3_HUMAN		_	
1	11	21	31	41
SEQ ETANSKCASL	EKTKQRLQGE	VEDLMRDLER	SHTACATLDK	KQRNFDKV
SS3 CHHHHHHHHH	ннннннн	ннннннн	ннннннн	HHHHHCC
SS8 LHHHHHHHHH	H H H H H H H H H H	ннннннн	***	HHHHHLL
ACC EEEEEEEEE	EEEEEEEE	EEEEEEEEEE	EEEEEEEE	EEEEEEE
DISO *******			*	* * * * * * * *
31. Q9UPN4 AZI1_H	IUMAN			
1	11	21	31	41
SEQ AGDNLEMMAP	SRGSAKSRGP	LEELLHTLQL	LEKEPDVLPR	PRTHHRGR
SS3 CCCCCCCCCC	000000000000000000000000000000000000000	ннннннн	HHHCCCCCCC	ccccccc
SS8 LLLLLLLLL	LLLLLLLLLL	H H H H H H H H H	HHHLLLLLL	LLLLLLL
ACC EEEEEEEEE	EEEEEEEE	BEEBBEMBEM	MEEEEEMEE	EEEEEEE
DISO ********	* * * * * * * * *	**	* * * * * * * * *	* * * * * * * *
32. Q9Y2G9 SBNO2	HUMAN			
1	11	21	31	41
SEQ AQADPAALAH	QGCDINFKEV	LEDMLRSLHA	GPPSEGALGE	GAGAGGAA
SS3 CCCCHHHHHC	сссссснннн	нннннннс	0000000000	CCCCCCCC
SS8 LLLLHHHHHH	TTLLLHHHH	HHHHHHHHL	LLLLLLLLL	LLLLLLL
ACCEEEEEEEEEE	EEMEMEBEEB	BMEBBEEBEE	EEEEEEEE	EEEEEEE
	********	* * * * * * * * * *	* * * * * * * * * *	*******
<u>33. Q9Y4J8 D1NA_I</u>	IUMAN			
1	11	21	31	41
SEQ DELEQRMSAL	QESRRELMVQ	LEGLMKLLKT	QGAGSPRSSP	SHTISRPI
SS3 CHHHHHHHHH	нннннннн	ннннннн	0000000000	ccccccc
SS8 LHHHHHHHHH	ннннннн	нннннннн	LLLLLLLL	LLLLLLL
ACC EEEEEEEEE	EEEEEEEE	EEEEMEEEEE	EEEEEEEE	EEEEEEE
		* * * * * * * * * * *	* * * * * * * * * * *	* * * * * * * *
34. QOLKS2 SKCAP	HUMAN			
1	11	21	31	41
SEQ NDAEAQRREI	ELLRREGELP	LEELLRSLPP	QLLEGPSSPS	QTPSSHDS
SS3 CCHHHHHHHH	нннннсссс	ннннннссс	ccccccccc	ccccccc
SS8 LLHHHHHHH	HHHHHTLSL	HHHHHHSLL	LLLLLLLLL	LLLLLLL
	EEEEEEME	MEEBBEEMEE	EEMEEEEEE	EEEEEEE
0100 ********	*********	* * * * * * * * * * *	*********	*******

# **Supplementary Figure 3. Binding Assays**

Binding assays of known LD motifs and LD motifs proposed by LDMF-proposed to FAT,  $\alpha$ -parvin and GIT1. ACA: anisotropy competition assay; DA: direct fluorescence anisotropy; MST: microscale thermophoresis; DSF: differential scanning fluorimetry.









# Supplementary Figure 3.2: Binding of highly likely LD motifs to FAT





**Supplementary Figure 3.4:** Binding of least likely LD motifs to FAT and motifs discarded in round 1.





# Supplementary Figure 3.5: Binding of LD motif controls to $\alpha$ -parvin





# **Supplementary Figure 3.6:** Binding of highly likely LD motifs to α-parvin



# **Supplementary Figure 3.7:** Binding of less likely LD motifs to α-parvin



α-parvin μM

Supplementary Figure 3.8: Binding of least likely LD motifs to  $\alpha$ -parvin and motifs discarded in round 1



**Supplementary Figure 3.9:** Binding of LD motif candidates to GIT1.





Tm shift in °C for differential scanning fluorimetry for peptides with (a) FAT (b) α-parvin and (c) GIT1. The Uniprot identifiers are given instead of protein gene names. Genes were both identifiers differ are P2R3A: PPP2R3A; CD158:CCDC158; RGPA1/2: RALGAPA1/2; ICAL: CAST; IBP2: IGFBP2; FIP1: FIP1L1; E41L5: EPB41L5; CP071: C16orf71; CP037: C8orf37.

Supplementary Figure 3.11: Anisotropy competition assay plotted as difference in fluorescence anisotropy.



Proteins were kept at a concentration corresponding to the  $K_d$  of their interaction with labeled LD4 (10 µM for FAT and 25 µM for α-parvin), in the presence of 0.1 µM labeled LD4. To that, each non-labeled LD motif candidate peptide was added at 100 or 250 µM. Plotted are the resulting relative changes of the fluorescence anisotropy in presence of the unlabeled candidate peptides. The Uniprot identifiers are given instead of protein gene names. Genes were both identifiers differ are P2R3A: PPP2R3A; CD158:CCDC158; RGPA1/2: RALGAPA1/2; ICAL: CAST; IBP2: IGFBP2; FIP1: FIP1L1; E41L5: EPB41L5; CP071: C16orf71; CP037: C8orf37.



Supplementary Figure 3.13: Titration of GIT1 on to LD2 and EPB41L5



		20 residue	es upstream of pr motif	redicted	20 residues downstream of predicted motif			
LD motif region of protein	LD motif region of protein (Uniprot Name)	Exposed	Moderately exposed	Buried	Exposed	Moderately exposed	Buried	
EPB41L5	E41L5_HUMAN	18	2	0	10	9	1	
LPP	LPP_HUMAN	19	1	0	20	0	0	
RALGAPA1	RGPA1_HUMAN	15	4	1	16	3	1	
PPP2R3A	P2R3A_HUMAN	20	0	0	18	0	2	
CCDC158	CD158_HUMAN	20	0	0	18	1	1	
C16orf71	CP071_HUMAN	13	7	1	17	1	2	
NCOA2	NCOA2_HUMAN	20	0	0	16	3	1	
NCOA3	NCOA3_HUMAN	20	0	0	17	1	2	
CAST	ICAL_HUMAN	20	0	0	20	0	0	
CREB3	CREB3_HUMAN	15	2	3	11	7	2	
RALGAPA2	RGPA2_HUMAN	20	0	0	15	4	1	
C8org37	CH037_HUMAN	3	-	-	20	0	0	

# Supplementary Figure 4: RaptorX results for predicted human proteins

## Summary of RaptorX results for predicted 12 proteins

Secondary structure predictions (SS3: three states, namely H: helix, E: beta strand, C: coil; SS8: eight states, namely H:  $\alpha$  helix, G: 3-helix, I: 5-helix, E: extended  $\beta$  ladder, B:  $\beta$  bridge, T: hydrogen bonded turn, S: bend, L: loop), solvent accessibility (ACC; B: buried; M: medium exposed, E: solvent exposed) and disorder (DISO: order [.] and disorder [\*]) for the non-paxillin motifs suggested by SlimSearch4 (Krystkowiak and Davey, 2017), which was the only algorithm which predicted a reasonable number of LD motif candidate in the human proteome (see **Supplementary Table 1**). The feature predictions were established by the RaptorX server (Kallberg, et al., 2014). The suggested LD motif region is boxed. Amino acid are numbered starting with 20 positions upstream of the LD motif (unless the LD motif is situated at the N-terminus, which is then taken as number 1).

## E41L5\_HUMAN

561	571	581	591	601	611	621	631
SEQ DFKSNILKAQ	VEAVHKVTKE	DSLLSHKNAN	VQDAATNSAV	LNENNVPLPK	ESLETLMLIT	PADSGSVLKE	ATDELDALLA
SS3 ССННННННН	ннннннн	HHCCCCCCCC	CCCHHCCCCC		ссннннсссс		сснннннн
SS8 ТНННННННН	нннннннн	HHHLLLLL	LLLHHHLLLL	LLLLLLLL	LLHHHHLLLL	LLLLLLLLL	LLHHHHHHHHH
ACC EBMBMBBEBB	EMBBEEBEEE	EEEEEEEEEE	EEEEEEEE	${\tt E} {\tt M} {\tt E} {\tt E}$	EEEEEEEE	EEEEEEMEE	EMEEBMEMBE
DISO	* * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
641	651	661	671	681	691	701	711
SEQ SLTENLIDHT	VAPQVSSTSM	ITPRWIVPQS	GAMSNGLAGC	EMLLTGKEGH	GNKDGISLIS	PPAPFLVDAV	TSSGPILAEE
SS3 ННННСССССС		CCCCCECCCC					CCCCCCCHH
SS8 HHHHLTLLL	LLLLLLLLL	LLLL <b>E</b> LLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLLL	LLLLLLHH
ACC EBMEMBMEME	EMEEMEEEMM	MMEEMMBEEE	EEEEEBEEB	EEEEEEEE	EEEEEEMEE	EEMEMMMEEB	EEEEEMEEE

## LPP\_HUMAN

	81	91	101	111	121	131	141	151
SEQ	LPSISGNFPP	PPPLDEEAFK	VQGNPGGKTL	EERRSSLDAE	IDSLTSILAD	LECSSPYKPR	PPQSSTGSTA	SPPVSTPVTG
SS3	ccccccccc			сссссссннн	ссснннннн	HCCCCCCCCC		
SS8	LLLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLHHHH	LLLHHHHHHH	HILLLLLL	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	MEMBEMBBEM	BEEEEEEE	EEEEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	161	171	181	191	201	211	221	231
SEQ	HKRMVIPNQP	PLTATKKSTL	KPQPAPQAGP	IPVAPIGTLK	PQPQPVPASY	TTASTSSRPT	FNVQVKSAQP	SPHYMAAPSS
SS3	ccccccccc							
SS8	LLLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEEE	EEEEEEEEEE	EEEEEEEEEE	EEEEEEEEEE	${\tt E} {\tt M}$	EEEEEEEME	$M \to M \to $	EEEMEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

**RGPA1\_HUMAN** 

	1601	1611	1621	1631	1641	1651	1661	1671
SEQ	DLSGKYSWDS	AILYGPPPVS	GLSEPTSFML	SLSHQEKPEE	PPTSNECLED	ITVKDGLSLQ	FKRFRETVPT	WDTIRDEEDV
SS3	CCCCCEEEEE	EEECCCCCC	0000000000	ccccccccc	ccccccccc	ccccccccc	ccccccccc	CCCCCCCCCH
SS8	LTTSLEEEE	EEEELLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLL <mark>H</mark>
ACC	вммвммввмв	MMBMEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEMEEEBEE	MEEMEEEMM
DISO		* * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	*****	
	1681	1691	1701	1711	1721	1731	1741	1751
SEQ	1681 LDELLQYLG <mark>V</mark>	1691 TSPECLQRTG	1701 ISLNIPAPQP	1711 VCISEKQEND	1721 VINAILKQHT	1731 EEKEFVEKHF	1741 NDLNMKAVEQ	1751 DEPIPQKPQS
SEQ SS3	1681 LDELLQYLGV HHHHHHHH	1691 TSPECLQRTG HCCHHHCCCC	1701 ISLNIPAPQP CCCCCCCCC	1711 VCISEKQEND CCCCHHHHHH	1721 VINAILKQHT HHHHHHHHH	1731 EEKEFVEKHF HHHHHHHH	1741 NDLNMKAVEQ CCCCCCCCC	1751 DEPIPQKPQS CCCCCCCCCC
SEQ SS3 SS8	1681 LDELLQYLGV ННННННННН ННННННННН	1691 TSPECLQRTG HCCHHHCCCC HLHHHHHLLT	1701 ISLNIPAPQP CCCCCCCCCC LLLLLLLLLL	1711 VCISEKQEND CCCCHHHHH LLLLHHHHHH	1721 VINAILKQHT ННННННННН ННННННННН	1731 EEKEFVEKHF HHHHHHHH HHHHHHHH	1741 NDLNMKAVEQ CCCCCCCCC LLLLLLLLL	1751 DEPIPQKPQS CCCCCCCCC LLLLLLLS
SEQ SS3 SS8 ACC	1681 L D E L L Q Y L G V H H H H H H H H H H H H H H H H H H B M E B B E M B M M	1691 TSPECLQRTG HCCHHHCCCC HLHHHHHLLT EMEEBMEEEE	1701 ISLNIPAPQP CCCCCCCCC LLLLLLLLL EEEEEEEEE	1711 VCISEKQEND CCCCHHHHH LLLLHHHHHH EEEEEEBEE	1721 VINAILKQHT HHHHHHHH HHHHHHHH BMEMBMEBME	1731 E E K E F V E K H F H H H H H H H H H H H H H H H H H H	1741 NDLNMKAVEQ CCCCCCCCC LLLLLLLL EEEEEEEEE	1751 DEPIPQKPQS CCCCCCCCC LLLLLLLS EEEEEEMEM
SEQ SS3 SS8 ACC DISO	1681 L D E L L Q Y L G V H H H H H H H H H H H H H H H H H H B M E B B E M B M M 	1691 TSPECLQRTG HCCHHHCCCC HLHHHHHLLT EMEEBMEEEE	1701 ISLNIPAPQP CCCCCCCCC LLLLLLLL EEEEEEEEE	1711 VCISEKQEND CCCCHHHHHH LLLLHHHHHH EEEEEEBEE	1721 VINAILKQHT НННННННН НННННННН ВМЕМВМЕВМЕ	1731 EEKEFVEKHF HHHHHHHH HHHHHHHH EBEEMEEEEE	1741 NDLNMKAVEQ CCCCCCCCC LLLLLLLL EEEEEEEEE	1751 DEPIPQKPQS CCCCCCCCC LLLLLLLS EEEEEEMEM

# P2R3A\_HUMAN

	401	411	421	431	441	451	461	471
SEQ	NPLENVSSDD	LMETLYIEEE	SDGKKALDKG	QKTENGPSHE	LLKVNEHRAE	FPEHATHLKK	CPTPMQNEIG	KIFEKSFVNL
SS3	сссссссснн	HHHHCCCCCC					CCCCCCCCHH	HHCCCCCCCC
SS8	LLLLLLHH	HHHHLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLHH	HHHHLLLLL
ACC	E E E E E E E E E E	EEEBBMBMEE	EEEEEEEE	EEEEEEEE	EEEMEEEEE	MEEMEEEBEE	MEEEEEEM	EMBEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	481	491	501	511	521	531	541	551
SEQ	481 PKEDCKSKVS	491 KFEEGDQRDF	501 TNSSSQE <mark>EID</mark>	511 KLLMDLESFS	521 QKMETSLREP	531 LAKGKNSNFL	541 NSHSQLTGQT	551 LVDLEPKSKV
SEQ SS3	481 PKEDCKSKVS CCCCCCCCC	491 KFEEGDQRDF CCCCCCCCC	501 TNSSSQEEID CCCCCHHHHH	511 KLLMDLESFS HHHHHHHHH	521 QKMETSLREP HHHHCCCCCC	531 LAKGKNSNFL CCCCCCCCC	541 NSHSQLTGQT CCCCCCCCC	551 LVDLEPKSKV CCCCCCCCC
SEQ SS3 SS8	481 PKEDCKSKVS CCCCCCCCCC LLLLLLLLL	491 KFEEGDQRDF CCCCCCCCC LLLLLLLLLL	501 TNSSSQE <mark>EID</mark> CCCCCHHHHH LLLLLHHHHH	511 KLLMDLESFS HHHHHHHHH HHHHHHHH	521 QKMETSLREP HHHHCCCCCC HHHHLLLLLL	531 LAKGKNSNFL CCCCCCCCC LLLLLLLLL	541 NSHSQLTGQT CCCCCCCCC LLLLLLLLL	551 LVDLEPKSKV CCCCCCCCC LLLLLLLLL
SEQ SS3 SS8 ACC	481 PKEDCKSKVS CCCCCCCCC LLLLLLLL EEEEEEEEE	491 KFEEGDQRDF CCCCCCCCC LLLLLLLL EEEEEEEE	501 TNSSSQEEID CCCCCHHHHH LLLLLHHHHH EEEEEEEBE	511 KLLMDLESFS HHHHHHHH HHHHHHHH EBBEEBEEBB	521 QKMETSLREP HHHHCCCCCC HHHHLLLLL EEEEEEEEE	531 LAKGKNSNFL CCCCCCCCC LLLLLLLL EEEEEEEE	541 NSHSQLTGQT CCCCCCCCC LLLLLLLL EEEEEEEE	551 LVDLEPKSKV CCCCCCCCC LLLLLLLL EEEEEEEEE
SEQ SS3 SS8 ACC DISO	481 PKEDCKSKVS CCCCCCCCC LLLLLLLL EEEEEEEEE ******	491 KFEEGDQRDF CCCCCCCCC LLLLLLLL EEEEEEEEE ********	501 TNSSSQEEID CCCCCHHHHH LLLLLHHHHH EEEEEEEBE *****	511 KLLMDLESFS HHHHHHHHH HHHHHHHHH EBBEEBEEBB 	521 QKMETSLREP HHHHCCCCCC HHHHLLLLL EEEEEEEEE *******	531 LAKGKNSNFL CCCCCCCC LLLLLLLL EEEEEEEE ********	541 NSHSQLTGQT CCCCCCCCC LLLLLLLL EEEEEEEE *******	551 LVDLEPKSKV CCCCCCCCC LLLLLLLL EEEEEEEE ********

# CD158\_HUMAN

	881	891	901	911	921	931	941	951
SEQ	STASFLSHHS	TKANTLKEDP	TRDLKQLLQE	LRSVINEEPA	VSLSKTEEDG	RTSLGALEDR	VRDCITESSL	RSDMCHRSNN
SS3	0000000000	ccccccccc	СННННННН	нннннсссс	ccccccccc	СССССННННН	ннннннн	Н Н Н Н Н Н С С С С
SS8	LLLLLLLLL	LLLLLLLL	ннннннн	HHHHHLLL	LLLLLLLL	LLLLHHHHH	ннннннн	HHHHHLLL
ACC	EEEEEEEEE	EEEEEEEEE	EEMBEEBBEE	BMEMBEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	961	971	981	991	1001	1011	1021	1031
SEQ	961 SLRDSTEGSK	971 SSETLSREPV	981 TLHAGDREDP	991 SGCFTFTSAA	1001 SPSVKNSASR	1011 SFNSSPKKSP	1021 VHSLLTSSVE	1031 GSIGSTSQYR
SEQ SS3	961 SLRDSTEGSK CCCCCCCCCC	971 SSETLSREPV CCCCCCCCCC	981 TLHAGDREDP CCCCCCCCC	991 SGCFTFTSAA CCCCCCCCCC	1001 SPSVKNSASR CCCCCCCCCC	1011 SFNSSPKKSP CCCCCCCCCC	1021 VHSLLTSSVE CCCCCCCCCC	1031 GSIGSTSQYR CCCCCCCCCC
SEQ SS3 SS8	961 SLRDSTEGSK CCCCCCCCC LLLLLLLLLL	971 SSETLSREPV CCCCCCCCC LLLLLLLLL	981 TLHAGDREDP CCCCCCCCC LLLLLLLLL	991 SGCFTFTSAA CCCCCCCCC LLLEELLLL	1001 SPSVKNSASR CCCCCCCCC LLLLLLLLL	1011 SFNSSPKKSP CCCCCCCCC LLLLLLLLL	1021 VHSLLTSSVE CCCCCCCCC LLLLLLLLL	1031 GSIGSTSQYR CCCCCCCCC LLLLLLLLL
SEQ SS3 SS8 ACC	961 SLRDSTEGSK CCCCCCCCC LLLLLLLLL EEEEEEEEE	971 SSETLSREPV CCCCCCCCC LLLLLLLLL EEEEEEEEE	981 TLHAGDREDP CCCCCCCCC LLLLLLLL EBMEEEEEEE	991 SGCFTFTSAA CCCCCCCCCC LLLEELLLL EEEMEBEMEE	1001 SPSVKNSASR CCCCCCCCC LLLLLLLLL EEEBEEEEEE	1011 SFNSSPKKSP CCCCCCCCC LLLLLLLLL EEEEEEEEE	1021 VHSLLTSSVE CCCCCCCCCC LLLLLLLLL EEEEEEEEE	1031 GSIGSTSQYR CCCCCCCCC LLLLLLLLL EEEEEEEEE

# CP071\_HUMAN

	241	251	261	271	281	291	301	311
SEO	APRSKMPLVE	PPEGPPVLSL		TLOSLAGOED	NOGNBAPGTV	WWAADHROVO	DRMVPSAHNR	LMEOLALLCT
CC2		ccccccccc	синссссини	нининососс	cccccccc	cccccccc	сссссенни	нининини
555 558	LILLLLLLLL	LLLLLLLL	нннннциннн	HHHHHHTLLL	LLLLLLLLL	HGGGLLLLL	LLLLLHHHHH	
ACC	FFFFFFFFFF	MEEMEEMMMM	EEBEEMMBME	BBEEBEEEEE	FFFFFFFFFF	BMEEEEEEE	EEEEEBMEM	BBEMBMEBBE
DISO	* * * * * * * * * *	* * * *		***	******	*	**	*****
	321	331	341	351	361	371	381	391
SEQ	TQSKASACAR	KVPADTPQDT	KEADSGSRCA	SRKQGSQAGP	GPQLAQGMRL	NAESPTIFID	LRQMELPDHL	SPESSSHSSS
SS3	нссссссссс		ннннссссс			CCCCCEEEEE		
SS8	HHLLLLLLL	LLLLLLLLL	HHHHHLLLL	LLLLLLLL	LLLLLLLLL	LLLLEEEEE	LLLLLLLLL	LLLLLLLL
ACC	EEEEEEEE	EMEEEEEEE	EEEEEEEE	EEEEEEEE	EEEEEEMEM	EEEMMMBMBM	BMEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	******	* * * * * * * * * *	* * * * * * * * * *

	721	731	741	751	761	771	781	791
SEQ	STAPGSEVTI	KQEPVSPKKK	ENALLRYLLD	KDDTKDIGLP	EITPKLERLD	SKTDPASNTK	LIAMKTEKEE	MSFEPGDQPG
SS3	0000000000	CCCCCCCCHH	ННННННС					
SS8	LLLLLLLL	LLLLLHHHH	ННННННЦ	TTLLLLTTLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEE	EEEEEMEEEE	MMBBBMBBBM	EEMEEEEBE	EMEEEEEEE	EEEEEEEEEE	EEEEEEEEE	EEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	*	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	801	811	821	831	841	851	861	871
SEQ	801 SELD <mark>NLEEIL</mark>	811 DDLQ <mark>NSQLPQ</mark>	821 LFPDTRPGAP	831 AGSVDKQAII	841 NDLMQLTAEN	851 S P V T P V G A Q K	861 TALRISQSTF	871 NNPRPGQLGR
SEQ SS3	801 SELD <mark>NLEEIL</mark> CCCC <mark>CHHHH</mark>	811 DDLQNSQLPQ HHHHCCCCCC	821 LFPDTRPGAP CCCCCCCCCC	831 AGSVDKQAII CCCCCHHHHH	841 NDLMQLTAEN HHHHHCCCC	851 SPVTPVGAQK CCCCCCC <mark>HHH</mark>	861 TALRISQSTF HHHCCCCCCC	871 NNPRPGQLGR CCCCCCCCC
SEQ SS3 SS8	801 SELDNLEEIL CCCCCHHHHH LLLLLHHHHH	811 DDLQNSQLPQ HHHHCCCCCC HHHHLLLLL	821 LFPDTRPGAP CCCCCCCCC LLLLLLLLL	831 AGSVDKQAII CCCCCHHHH LLLLLHHHHH	841 NDLMQLTAEN HHHHHHCCCC HHHHHHLLLL	851 SPVTPVGAQK CCCCCCCHHH LLLLLLHHHH	861 TALRISQSTF HHHCCCCCCC HHHLLLLLL	871 NNPRPGQLGR CCCCCCCCC LLLLLLLLL
SEQ SS3 SS8 ACC	801 SELDNLEEIL CCCCCHHHHH LLLLLHHHHH EEEEBBEBB	811 DDLQNSQLPQ HHHHCCCCCC HHHHLLLLL EMBEEEEMEE	821 LFPDTRPGAP CCCCCCCCC LLLLLLLLL MMEEEEEEE	831 AGSVDKQAII CCCCCHHHHH LLLLLHHHHH EEEBEMEEBB	841 NDLMQLTAEN HHHHHHCCCC HHHHHHLLLL EEBMEMEEEE	851 SPVTPVGAQK CCCCCCCHHH LLLLLHHHH EEEEEEEEE	861 TALRISQSTF HHHCCCCCCC HHHLLLLLL EEEEBEEEEE	871 NNPRPGQLGR CCCCCCCCC LLLLLLLL EEEEEEEEE
SEQ SS3 SS8 ACC DISO	801 SELDNLEEIL CCCCCHHHHH LLLLLHHHHH EEEEBBEBB * * * * * * * * * * *	811 DDLQNSQLPQ HHHHCCCCCC HHHHLLLLL EMBEEEEMEE ****	821 LFPDTRPGAP CCCCCCCCC LLLLLLLLL MMEEEEEEE *******	831 AGSVDKQAII CCCCCHHHH LLLLLHHHHH EEEBEMEEBB	841 NDLMQLTAEN HHHHHHCCCC HHHHHHLLLL EEBMEMEEEE *******	851 SPVTPVGAQK CCCCCCCHHH LLLLLHHHH EEEEEEEEE * * * * * * * * * * *	861 TALRISQSTF HHHCCCCCCC HHHLLLLLL EEEEBEEEEE ******	871 NNPRPGQLGR CCCCCCCCC LLLLLLLL EEEEEEEE * * * * * * * * * * *

# NCOA3\_HUMAN

	721	731	741	751	761	771	781	791
SEQ	VVKQEQLSPK	KKENNALLRY	LLDRDDPSDA	LSKELQPQVE	GVDNKMSQCT	SSTIPSSSQE	KDPKIKTETS	EEGSGDLDNL
SS3	ccccccccc	НННССННННН	HHCCCCCCCC	ccccccccc	ccccccccc	ccccccccc	ccccccccc	CCCCCCCCC
SS8	LLLLLLLLL	НННННННН	HHLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLL <mark>L</mark> H
ACC	EEEEEEEEE	EEEMMBBBMB	BBMEEMEEEE	EBEEMEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEB
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	801	811	821	831	841	851	861	871
SEQ	DAILGDLTSS	DFYNNSISSN	GSHLGTKQQV	FQGTNSLGLK	SSQSVQSIRP	PYNRAVSLDS	PVSVGSSPPV	KNISAFPMLP
SS3	нннннсссс	ccccccccc	CCCCCCCCCH	ccccccccc	ccccccccc	ccccccccc	ccccccccc	ccccccccc
SS8	ннннннгг	LLLLLLLL	LLLLLLLH	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL
ACC	EEBBEEBEEE	$\mathbb{E} \xrightarrow{B} \mathbb{M} \xrightarrow{E} $	EEEEBEEEB	BEEBEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *

# ICAL\_HUMAN

	81	91	101	111	121	131	141	151
SEQ	KKAVSRSAEQ	QPSEKSTEPK	TKPQDMISAG	GESVAGITAI	SGKPGDKKKE	KKSLTPAVPV	ESKPDKPSGK	SGMDAALDDL
SS3	0000000000	0000000000	ccccccccc	ccccccccc	ccccccccc	ccccccccc	ccccccccc	ссссннннн
SS8	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLHHHHHH
ACC	EEEEEEEEE	EEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEEEEE	EEEEEBEEB
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	161	171	181	191	201	211	221	231
SEQ	161 IDTLGGPEET	171 EEENTTYTGP	181 EVSDPMSSTY	191 IEELGKREVT	201 IPPKYRELLA	211 KKEGITGPPA	221 DSSKPIGPDD	231 AIDALSSDFT
SEQ SS3	161 IDTLGGPEET HHHHCCCCCC	171 EEENTTYTGP CCCCCCCCCC	181 EVSDPMSSTY CCCCCCC <mark>HHH</mark>	191 IEELGKREVT HHHHCCCCCC	201 I P P K Y R E L L A C C H H H H H H H H	211 KKEGITGPPA HCCCCCCCCC	221 DSSKPIGPDD CCCCCCC <mark>HHH</mark>	231 AIDALSSDFT HHHHHCCCC
SEQ SS3 SS8	161 IDTLGGPEET HHHHCCCCCC HHHHLLLLLL	171 EEENTTYTGP CCCCCCCCC LLLLLLLLLL	181 EVSDPMSSTY CCCCCCCHHH LLLLLLHHH	191 IEELGKREVT HHHHCCCCCC HHHHTLLLLL	201 I P P K Y R E L L A C C H H H H H H H H L L H H H H H H H H H	211 KKEGITGPPA HCCCCCCCC HLTLLLLLL	221 DSSKPIGPDD CCCCCCCHHH LLLLLLHHH	231 AIDALSSDFT HHHHHHCCCC HHHHHHLLLL
SEQ SS3 SS8 ACC	161 I DTLGGPEET HHHHCCCCCC HHHHLLLLL MEEEEEEEEE	171 EEENTTYTGP CCCCCCCCCC LLLLLLLLL EEEEEMEEE	181 EVSDPMSSTY CCCCCCCHHH LLLLLLHHH EMEEEEEEM	191 IEELGKREVT HHHHCCCCCC HHHHTLLLL MEEBEEEEEE	201 I P P K Y R E L L A C C H H H H H H H H L L H H H H H H H H B M E E B M E M B E	211 KKEGITGPPA HCCCCCCCC HLTLLLLLL EEEEMEEEEE	221 DSSKPIGPDD CCCCCCCHHH LLLLLLHHH EEEEEMEEEE	231 AIDALSSDFT HHHHHHCCCC HHHHHHLLLL BMEEBMEEBE
SEQ SS3 SS8 ACC DISO	161 I D T L G G PE E T H H H H C C C C C C H H H H L L L L L MEE E E E E E E E E * * * * * * * * * * *	171 EEENTTYTGP CCCCCCCCC LLLLLLLL EEEEEMEEE *******	181 EVSDPMSSTY CCCCCCCHHH LLLLLLHHH EMEEEEEEM	191 IEELGKREVT HHHHCCCCCC HHHHTLLLLL MEEBEEEEEE *********	201 IPPKYRELLA CCHHHHHHHH LLHHHHHHHH BMEEBMEMBE *	211 KKEGITGPPA HCCCCCCCC HLTLLLLLL EEEEMEEEEE *******	221 DSSKPIGPDD CCCCCCHHH LLLLLLHHH EEEEMEEEE	231 AIDALSSDFT HHHHHHCCCC HHHHHHLLLL BMEEBMEEBE *******

# **CREB3\_HUMAN**

	1	11	21	31	41	51	61	71
SEQ	MELELDAGDQ	DLLAFLLEES	GDLGTAPDEA	VRAPLDWALP	LSEVPSDWEV	DDLLCSLLSP	PASLNILSSS	NPCLVHHDHT
SS3	0000000000	СННННННСС	CCCCCCCCHH	HHCCCCCCCC	CCCCCCCCC	ннннннсс		
SS8	LLLLLLLLL	L H H H H H H H H <b>T</b>	LLLLLLHH	HHLLLLLL	LLLLLHL	нннннны	LLLLLLLL	LLLLLLLL
ACC	EEEEEEEME	EBBEMBMEEE	EEEEEEEEEE	EEBEEEBEME	EEEMBEEMEB	MEBBMEBBME	EEBMEEMEEE	EEMMBMMMMM
DISO	* * * * * * * * * * .						* * * * * * * * * *	* * * * * * * * * *
	81	91	101	111	121	131	141	151
SEQ	YSLPRETVSM	DLGECEISLT	GRTGFMGLAI	HTFPFAESES	CRKEGTQMTP	QHMEELAEQE	IARLVLTDEE	KSLLEKEGLI
SS3	0000000000					ссссснинсс	сссссснин	ннинисссс
						00000		
SS8	LLLLLLLLL	LLLLLLLL	LLLLLLLLL	LLLLLLLLL	LLLLLLLLL	LLLLHHHLL	LLLLLLHHH	HHHHHTTLL
SS8 ACC	LLLLLLLLL MMBMEEEBBB	LLLLLLLL MEEEMEEEEE	LLLLLLLLL	LLLLLLLLL	LLLLLLLLL	LLLLHHHLL EEEEEEEE	LLLLLLHHH MEEMEMMEEM	H H H H H H T T L L MEMBEEEME

# RGPA2\_HUMAN

	1441	1451	1461	1471	1481	1491	1501	1511
SEQ	QTPTEGPVGG	SPVGSLSDVR	VIVRDISGKY	SWDGKVLYGP	LEGCLAPNGR	NPSFLISSWH	RDTFGPQKDS	SQVEEGDDVL
SS3	ECCCCCCCCC	CCCCCCCEEE	EEECCCCCE	EEEEEECC	ccccccccc	ccccccccc	ccccccccc	сссссссенн
SS8	EELLLLLLL	LLLLLLEEE	EEELTTSLE	EEEEEELL	LLLLLLLLL	LLLLLLLLL	LLLLLLLLL	LLLLLLIHH
ACC	BMMEEEEEE	EEEEEMMBM	ввввимвим	BBMBEBMMME	MEEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEB
DISO	*****	* * * * • • • • • • •			* * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	*****
	1521	1531	1541	1551	1561	1571	1581	1591
SEQ	DKLLENIGHT	SPECLLPSQL	NLNEPSLTPC	GMNYDQEKEI	IEVILRQNAQ	EDEYIQSHNF	DSAMKVTSQG	QPSPVEPRGP
SS3	ннннннссс	ccccccccc	ccccccccc	СССНННННН	нннннннн	ннннннсс	ccccccccc	ccccccccc
SS8	нннннннгг	LLLLLLLLL	LLLLLLLLL	LLLHHHHHHH	нннннннн	$\tt H  H  H  H  H  H  H  H  L  L$	LLLLLLLLL	LLLLLLSH
ACC	МЕВВЕМВЕМЕ	MEEBEMEEEM	EEEEEEEEE	EEEEMBEEB	MEMBMEBMEE	BEEEEEEE	EEEEEEEEE	EEEEEEME
DISO						* *	********	

### CH037\_HUMAN

	1	11	21	31	41	51	61	71
SEQ	MAEDLDELLD	EVESKFCTPD	LLRRGMVEQP	KGCGGGTHSS	DRNQAKAKET	LRSTETFKKE	DDLDSLINEI	LEEPNLDKKP
SS3	ссссннннн	нннннсссс	CCCCCCCCCC	ccccccccc	ccccccccc	ccccccccc	СННННННН	HCCCCCCCCC
SS8	LLLLHHHHHH	HHHHHLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	LLLLLLLL	ГННННННН	HTLLLLLL
ACC	EEEMEEBME	EBMEEEEEE	EEEEEEEE	EEEEEEEE	EEEEEEEEE	EEEEEEEE	EEBMEBBEMB	${\tt M} {\tt E} {\tt E}$
DISO	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *	* * * * * * * * * *
	81	91	101	111	121	131	141	151
SEQ	81 SKLKSKSSGN	91 TSVRASIEGL	101 GKSCSPVYLG	111 GSSIPCGIGT	121 NISWRACDHL	131 RCIACDFLVV	141 SYDDYMWDKS	151 C D Y L F F R N N M
SEQ SS3	81 SKLKSKSSGN CCCCCCCCC	91 TSVRASIEGL CCCCCCCCC	101 GKSCSPVYLG CCCCCCCEC	111 GSSIPCGIGT CCCCCCCCCC	121 NISWRACDHL CCCCCCCCC	131 RCIACDFLVV CCCCCCCEEE	141 SYDDYMWDKS EEECCEECCC	151 CDYLFFRNNM CCEEEECCCC
SEQ SS3 SS8	81 SKLKSKSSGN CCCCCCCCC LLLLLLLLL	91 TSVRASIEGL CCCCCCCCC LLLLLLLLL	101 GKSCSPVYLG CCCCCCCEEC LLLLLLEEL	111 GSSIPCGIGT CCCCCCCCC LLLSLTLLTT	121 NISWRACDHL CCCCCCCCC TLLTSLLTTL	131 RCIACDFLVV CCCCCCCEEE LLLTTLEEEE	141 SYDDYMWDKS EEECCEECCC EEETEEELTT	151 CDYLFFRNNM CCEEEECCCC LSEEEEETTL
SEQ SS3 SS8 ACC	81 SKLKSKSSGN CCCCCCCCC LLLLLLLLL EEEEEEEEE	91 TSVRASIEGL CCCCCCCCC LLLLLLLL EEEEEEEEE	101 GKSCSPVYLG CCCCCCCEEC LLLLLLEEL EEEBMBBBBB	111 GSSIPCGIGT CCCCCCCCC LLLSLTLLTT EMEMEEEEEE	121 NISWRACDHL CCCCCCCCCC TLLTSLLTTL EEEMMBBEMB	131 RCIACDFLVV CCCCCCCEEE LLLTTLEEEE MBMEBMBMBB	141 SYDDYMWDKS EEECCEECCC EEETEEELTT MBEEMEMMEE	151 CDYLFFRNNM CCEEEECCCC LSEEEETTL BMBBBBMMMB

# **Supplementary Figure 5:**



GO ANALYSIS: **A)** Distribution of Semantic Similarity between LDMF-predicted proteins and known LD motif proteins (cyan) and between LDMF-predicted proteins and all proteins, except the known LD motif proteins (red). The p-value of Mann-Whitney U test for the distributions is 6.32e-10.

**B)** Distribution of Semantic Similarity between LDMF-predicted proteins and known LD motif proteins (cyan) and between LDMF-predicted proteins and same number of random proteins from human (red). The p-value of Mann-Whitney U test for the distributions is 30648 e-14.

# Supplementary Figure 6: <sup>1</sup>H-<sup>15</sup>N HSQC titration experiments.

Shown are the NMR chemical shifts of <sup>15</sup>N-labelled FAT domain titrated with LD4, LD2, LPP, and CD158.



Supplementary Figure 6.1: FAT/LD2 Titration.

Overlay of <sup>1</sup>H-<sup>15</sup>N-HSQC spectra of 100  $\mu$ M <sup>15</sup>N-FAT in the absence (red) and presence of 0.5 (green), 1 (blue), 2 (yellow), 3 (magenta) and 4 (cyan) times molar excess of LD2 peptide. Resonances that disappeared upon LD2 addition are labelled in red. Resonances that significantly shifted >2 $\sigma$ = 0.16 are labelled in blue. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



Supplementary Figure 6.2: Chemical shift changes in FAT induced by the LD2 peptide. Chemical shift differences in ppm where calculated for <sup>1</sup>H (top panel), <sup>15</sup>N (middle panel) and the weighted combined <sup>1</sup>H, <sup>15</sup>N (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of LD2 peptide. Red dashed line indicates the upper threshold of  $2\sigma$ = 0.16 and the blue double-dashed line indicates the lower threshold of  $\sigma$ = 0.08. Others that disappeared upon LD2 addition are marked by full black circles. The shaded areas represent the helices (orange for helix1, blue for helix2, yellow for helix3, green for helix4).



**Supplementary Figure 6.3: FAT/LD4 Titration.** Overlay of <sup>1</sup>H-<sup>15</sup>N-HSQC spectra of 100  $\mu$ M <sup>15</sup>N-FAT in the absence (red) and presence of 0.5 (green), 1 (blue), 2 (yellow), 3 (magenta) and 4 (cyan) times molar excess of LD4 peptide. Resonances that disappeared upon LD4 addition are labelled in red. Resonances that significantly shifted >2 $\sigma$ = 0.12 are labelled in blue. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



Supplementary Figure 6.4: Chemical shift changes in FAT induced by LD4 peptide. Chemical shift differences in ppm where calculated for <sup>1</sup>H (top panel), <sup>15</sup>N (middle panel) and the weighted combined <sup>1</sup>H, <sup>15</sup>N (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of LD4 peptide. Red dashed line indicates the upper threshold of  $2\sigma$ = 0.12 and the blue double-dashed line indicates the lower threshold of  $\sigma$ = 0.06. Others that disappeared upon LD4 addition are marked by full black circles. The shaded areas represent the helices (orange for helix1, blue for helix2, yellow for helix3, green for helix4).



**Supplementary Figure 6.5: FAT/LPP titration.** Overlay of <sup>1</sup>H-<sup>15</sup>N-HSQC spectra of 100  $\mu$ M <sup>15</sup>N-FAT in the absence (red) and presence of 1 (green), 2 (blue), 3 (yellow), 4 (magenta) and 5 (cyan) times molar excess of LPP peptide. Resonances that disappeared upon LPP addition are labelled in red. Resonances that significantly shifted >2 $\sigma$ = 0.06 are labelled in blue. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



Supplementary Figure 6.6: Chemical shift changes in FAT induced by LPP peptide. Chemical shift differences in ppm where calculated for <sup>1</sup>H (top panel), <sup>15</sup>N (middle panel) and the weighted combined <sup>1</sup>H, <sup>15</sup>N (lower panel) chemical shift perturbation of FAT in the presence of a five times molar excess of LPP peptide. Red dashed line indicates the upper threshold of  $2\sigma$ = 0.06 and the blue double-dashed line indicates the lower threshold of  $\sigma$ = 0.03. The shaded areas represent the helices (orange for helix1, blue for helix2, yellow for helix3, green for helix4).



**Supplementary Figure 6.7: FAT/CCDC158 Titration.** Overlay of 1H15N-HSQC spectra of 100  $\mu$ M 15N-FAT in the absence (red) and presence of 0.5 (green), 1 (blue), 2 (yellow), 3 (magenta) and 4 (cyan) times molar excess of CCDC158 peptide. Resonances that disappeared upon CCDC158 addition are labelled in red. Resonances that significantly shifted >2 $\sigma$ = 0.114 are labelled in blue. All spectra were recorded at 25°C at a proton frequency of 900 MHz.



Supplementary Figure 6.8: Chemical shift changes in FAT induced by CCDC158 peptide. Chemical shift differences in ppm where calculated for <sup>1</sup>H (top panel), <sup>15</sup>N (middle panel) and the weighted combined <sup>1</sup>H, <sup>15</sup>N (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of CCDC158 peptide. Red dashed line indicates the upper threshold of  $2\sigma$ = 0.114 and the blue double-dashed line indicates the lower threshold of  $\sigma$ = 0.057. Others that disappeared upon CCDC158 addition are marked by full black circles. The shaded areas represent the helices (orange for helix1, blue for helix2, yellow for helix3, green for helix4).

# Supplementary Figure 7: Analysis of homology among unicellular LD motifs candidates.



Pairwise distance for stem eukaryotes proteins containing LD motifs

The conservation of LD motif-containing proteins in unicellular eukaryotes. Heat map shows pairwise identity matrix (in percentage) where E-value<1e-10. Proteins with annotated domains from PFAM are clustered on the sequence labels (on Y-axis) as follow: (1) LIM domain, (2) Protein kinase domain, (3) Formin Homology 2, (4) Retinal Maintenance, (5) Ubiquitin-activating enzyme active site (Thif family), (6) Mitochondrial carrier protein, (7) Ankyrin repeat, (8) RasGEF domain (RhoGEF), and (9) Ras association (RalGDS/AF-6) domain. The Y-axis shows the gene names. Each gene starts with the abbreviation of the species coming from. Abbreviations are as follows (1) Homo sapiens: *Homo sapiens*, (2) C. owczarzaki: *Capsaspora owczarzaki*, (3) M. brevicollis: *Monosiga brevicollis*, (4) M. verticillata: *Mortierella verticillata*, (5) D. discoideum: *Dictyostelium discoideum*, (6) S. arctica: *Sphaeroforma arctica*, (7) S. rosetta: *Salpingoeca rosetta*, (8) S. punctatus: *Spizellomyces punctatus*, (9) F. alba: *Fonticula alba*, (10) T. trahens: *Thecamonas trahens*, and (11) A. macrogynus: *Allomyces macrogynus*.

# Supplementary Figure 8: Conservation of human non-paxillin LD motif

**proteins across species**. In a first instance, we searched homologues of the newly identified LD motif-containing proteins in the most commonly used model species (i.e. fruit fly, zebrafish, chicken and mouse). When a homologue was not found in the model species, we considered other close species that have been sequenced (such as sharks or ostrich). The region encompassing the identified 10-residue LD motifs in humans are boxed.

#### EBP41L

E41L5\_human EDL39813.1\_mouse XP\_007894558.1\_shark XP\_015145518.1\_chicken

#### LPP

LPP\_human AAH85321.1\_mouse AAI71497.1\_zebrafish NP\_001026738.1\_chicken

#### RALGAPA1

RGPA1\_human AAI45120.1\_mouse BAH47605.1\_zebrafish XP\_023034350.1\_fruit\_fly

#### PPP2R3A

P2R3A\_human XP\_021062079.1\_mouse XP\_007885555.1\_shark XP\_422556.4\_chicken

#### CCDC158

CD158\_human NP\_796204.1\_mouse XP\_020373213.1\_shark xp\_009685592.1\_ostrich

#### C16orf71

CP071\_human PNI44266.1\_chimpanzee NP\_001258515.1\_mouse

#### NCOA2

NCOA2\_human BAF69036.1\_mouse AAI63724.1\_zebrafish

#### NCOA3

NCOA3\_human XP\_021041194.1\_mouse XP\_692938.5\_zebrafish ETLMLITPADSGSVLKEATDELDALLASLTENLIDHTV-APQVSSTSMITP LTPVHGTAADSASVLKDATDELDALLLSLTENLMDHTV-TPQVSSPSMITP SPPASIPPPAGSPLTQNAAQDLDRLLASLTENLIDFTDPTPQSAVMANGV-LISANVAPAEEAAVSKSAPDDRDVSLTSLTENLIDFTEATPRVSSQPTIT-. : : \* : \* \* \*\*\*\*\*\*:\*.\* : \*: : .

GNPGGKTLEERRSSLDAEIDSLTSILADLECSSPYKPRPPQSSTGSTASP GNPGGKTLEERRSSLDAEIDSLTSILADLECSSPYKPRPPQGSASSIASP IKGPEKTLEERRSSLDAEIDSLTSILADLESSSPYKPRTQQNSASPAATG VNPGGKTLEERRSSLDAEIDSLTSILADLESSSPYKPRIQQGSGTSSSAA

QFKRFRETVPTWDTIRDEEDVLDELLQYLGVTSPECLQRTGISLNIPAPQ QLRRFRETVPTWSTIREEEDVLDELLQYLGTTSPECLQRTGISLNVPAPQ AKRRCREVIPNWDSLQDGEDSLDEMLQYLGVSSPECLQRTGAPLNIPAPP LRHRPAGVLPLAKDMAPDLDQLDDMLAYIGHTSPECVPPTVTQLNAPSSS :\* .:\* .:\* \* \* \* \* \* \* \*

ASFLSHHSTKANTLKEDPTFDLKQLLQELRSVINEEPA--VSLSKTEEDGRT-TSFLSHHSIKTNTPKEDPTFDLKQLLQELRTVINEEPA--MALSKTEEDGRT-SGYLSSAVSKTSCLQEDLNFDLKNLLMDLKSTACDGCNRSNSINYTECEGERT ASCVPHMTSKPGILKEEPLFDLKRILQELSSDS-EIPS--VVLSKNDSDGGRT :. : \* . :\*: \*\*: :\* : : : : : : : :\*

----QEKDPKIKTETSEEGSGDLDNLDAILGDLTSSDFYN-NSISSNGSHLGTKQ------QEKDPKIKTETSEEVSGDLDNLDAILGDLTSSDFYN-NPT--NGSHPGAKQQM TSSSSNQESKVKL----EQPDELESLESILGGLRNPGPGMFVDSGSGGSEVGNK---.::: \*:\* \* .:\*:.\*:\*\*\*\*\*

#### CAST

ICAL\_human ICAL\_mouse XP\_007891289.1\_shark ABP68381.1 chicken

#### CREB3

CREB3\_human CREB3 mouse XP\_009678890.1\_ostrich

#### RALGAPA2

RGPA2\_human BAA92774.1 mouse XP\_005158962.1\_zebrafish XP\_023168385.1\_fruit\_fly

#### C8orf37

CH037 human NP\_080281.3\_mouse XP\_007886093.1\_shark

#### RoXaN

Z3H7B\_human NP 001074485.1 mouse XP\_007907761.1\_shark KFO61317.1\_chicken

#### DLC1

RHG07\_human XP\_021074615.1\_mouse XP 007889463.1 shark PAVPVESKPDKPSGKSGM-DAALDDLIDTLGGPEETEEENTTYTGPEVSDP-PASPVQSTPSKPSDKSGM-DAALDDLIDTLGGHEDTNRDDPPYTGPVVLDP-AAFSVSASQPAPKGKTGD-MGALDALSDML-DPEAPVHSGPKYTGPEVKEKA -VASMVAAADKPNSEPAMDESALDSLIDTLGSSEEDVATRPVYTGPEITEN-\*\*\* \* \* \* \* \*\*\*\*

. . . \*... .

EAVRAPLDWALPL---SEVPSDWEVDDLLCSLLSPPASLNILSSSNPCLVHHD --VKASLDLELSPSENSVOELSDWEVEDLLSSLLSPSVSRDVLGSSSSSILHD-EESSLLEDWGLSD----AQLLNNKEMDDFISSLLSPFVDEPGTLQGYSPTDSDS \* \* \* \* \* \* \* \* \* \* \* \* . .:

-WHRDTFGPQKDSSQVEEGDDVLDKLLENIGHTSPECLLPSQLNLNEPSLT -QLRRFRETVPTWSTIQEEEdVLDELLQYLGTTSPECLQRTGISLNVPAPQ -RSSVRFSKCSSELDVEDGVDVLDQLLEDLGHSSPECLPEPQLRLTQPPSP LRHR-PPGVLPLAKDVAPDLDQLDDMLAYIGHTSPECVPPTISELNAPSLS \*\* \* \* \*\*\*\* \* \*

MAEDLDELLDEVESKFCTPDLLRRG---MVEOPKGC MAKDLDELLDEVETKFCRLDPLRLD---LGERPKGD MADDLDQLLDEVEBRFCGQSAPRQRNKGAGERV---\*\* \*\*\* \*\*\*\*\*

RTLPSTDSLDDFSDGDVFGPELDTLLDSLS\_VQGGLSGSGVPSELPQLIP----RTLPGTEGLDDFSDGDVFGPELDTLLDSLS\_VQGGLPGSGMPSELPQLIP----VSLSVPESTEEFTDGEIIGEDIDTLLDSIQD----YPMSSAPGPIPTNVPANVS VPLPVPENVEDFTDGDIIGELDSLLDSLAEGS-PYPLGTIPTNLPTEMPO---\* \* \*

GSILYSSSGDLADLENEDIFPELDDILYHVKGMQRIVNQWSEKFSDEGDS GSILYSSSGELADLENEDIFPELDDILYHVKGMQRIVNQWSEKFSDEGDS GSHLYASTGDLLDLEKEDIFPHLDDILOHVNGLOOIVDHWSKNVLPELOD 

# **Supplementary Figure 9: Spreading Assay**



A) HeLa cells transfected with GFP alone (control cells) and GFP-tagged C16orf71 were plated on fibronectin, fixed and stained for the indicated antibodies. B) A large meshwork of actin stress fibres is observed in <sup>eGFP</sup>C16orf71, but not in control cells. The percentage of cells with high stress fibre density is shown. The quantification was performed on 23-42 cells. Scale bar=  $50 \ \mu m$ .

# Supplementary Table 1: Prediction results from other existing tools

## Table 1.1. Predictions for known LD motifs

We used [LV] [DE] X [LM] [LM] XXL as a regular expression for generating output from existing tools. SlimSearch4 returned 37 proteins (44 motifs). PSSM search returned 881 proteins (1000 motifs). FIMO search returned 1432 proteins (1614 motifs). The table shows the amino acid positions of known human LD motifs. Rank refers to a motif's rank based on the conservation score for SlimSearch4, based on the PWM p-value for PSSMSearch, and based on the p-value for FIMO.

Index	Protein name	Start position	End position	Uniprot ID	SlimSearch4 rank	PSSMSearch rank	FIMO rank
1	PXN	3	12	P49023	12	6	447
2	PXN	144	153	P49023	1	167	460
3	PXN	216	225	P49023	14	22	386
4	PXN	265	274	P49023	Not found	Not found	216
5	PXN	333	342	P49023	8	11	235
6	LPXN	3	12	O60711	3	35	453
7	LPXN	92	101	O60711	2	85	43
8	LPXN	127	136	O60711	5	114	223
9	Hic-5	3	12	O43294	4	2	448
10	Hic-5	92	101	O43294	Not found	Not found	305
11	Hic-5	157	166	O43294	6	4	196
12	Hic-5	203	212	O43294	5	152	233
13	RoXaN	280	289	Q9UGR2	Not found	Not found	420
14	DLC1	905	914	Q96QB1	Not found	Not found	Not found

# Supplementary Table 1.2: Predictions for additional LD motifs in the human proteome

We used [LV] [DE] X [LM] [LM] XXL as a regular expression for generating output from existing tools. SlimSearch4 returned 37 proteins (44 hits). PSSM search returned 881 proteins (1000 hits). PSSM search rank is out of 1000 hits. FIMO search returned 1432 proteins (1614 hits). FIMO search rank is out of 1614 hits. The table reports the rank of the LDMF predicted 12 proteins, based on the conservation score for SlimSearch4; based on the PWM p-value for PSSMSearch; based on p-value for FIMO (Grant, et al., 2011; Krystkowiak and Davey, 2017; Krystkowiak, et al., 2018).

Index	Protein name	Start position	End position	Uniprot ID	SlimSearch 4 rank	PSSMSearch rank	FIMO rank
1	EPB41L5	634	643	Q9HCM4	Not found	Not found	518
2	LPP	123	132	Q93052	Not found	515	Not found
3	RALGAPA1	1680	1689	Q6GYQ0	Not found	Not found	236
4	PPP2R3A	508	517	Q06190	Not found	10	Not found
5	CCDC158	903	912	Q5M9N0	Not found	40	512
6	C16orf071	267	276	Q8IYS4	Not found	101	Not found
7	NCOA2	805	814	Q15596	Not found	42	Not found
8	NCOA3	799	808	Q9Y6Q9	Not found	63	Not found
9	CAST	156	165	P20810	Not found	33	Not found
10	CREB3	49	58	O43889	Not found	Not found	122
11	RALGAPA2	1519	1528	Q2PPJ7	Not found	70	1432
12	C8orf37	4	13	Q96NL8	Not found	163	Not found

# Supplementary Table 2: Results of predictions from final model

Features	Number of Features	Sensitivity (%)	Specificity (%)	Accuracy (%)
All	40	88.889	100.00	99.968
Sequence	5	83.333	99.968	99.921
Secondary Structure	5	94.444	80.251	80.292
AAindex	30	66.667	97.143	97.056

Prediction results of the final LDMF model using different combination of features

The sensitivity, specificity, accuracy stated are based on the performance of the machinelearning model on the test set. We used the known LD motifs to build the machine learning model. We then tested the performance of the computational model using a leave-one-out cross validation approach. Given the imbalanced nature of our training data, 'sensitivity' appears as the most appropriate evaluation metric.

# Supplementary Table 3: Round1-round2\_predictions

The LD motif sequences used in LDMF are given, according to: *bona fide* LD motifs used in the initial training of LDMF, and LD motif candidates predicted in the second round of LDMF. **Supplementary Table 3.A**: Information for the *bona fide* LD motifs.

Index	Protein name	Start position	End position
1.	Paxillin   PXN	3	12
	Primary and secondary s	sequence	
	MDDLDALLADLESTTSF	IISKRPVFLSEETPYS	
2.	Paxillin   PXN	144	153
3	Pavillin   PXN	216	225
0.	Primary and secondary s	sequence	220
	PLTKEKPKRNGGRGLEDVRPSVESLLDELES	SSVPSPVPAITVNQGEMSSP	
	ССССССССССССССССССННННССССС	222222222222222222222222222222222222222	
4.	Paxillin   PXN	265	274
	Primary and secondary s	sequence	
	PQRVTSTQQQTRISASSATRELDELMASLSD	FKIQGLEQRADGERCWAAG	
			0.10
э.	Paxilin   PAN Primary and secondary	333	342
	MAQGKTGSSSPPGGPPKPGSQLDSMLGSLQ	SDI NKI GVATVAKGVCGACK	
	ССССССССССССССССССССССННННН	НННННСССССССССССССССССССССССССССССССССС	
6.	Leupaxin   LPXN	3	12
	Primary and secondary s	sequence	
	MEELDALLEELERSTLQ	DSDEYSNPAPLPLDQ	
	ССННННННННННКСССС	222222222222222222222222222222222222222	
7.	Leupaxin   LPXN	92	101
8		127	136
0.	Primary and secondary s	sequence	150
	VAVRADAGKKHLPDKQDHKASLDSMLGGLE	QELQDLGIATVPKGHCASCQ	
	НССССССССССССССССССССННСССН	НННННСССССССССССССССССССССССССССССССССС	
9.	Paxillin-B   paxB	10	19
	Primary and secondary s	sequence	
	MATKGLNMDDLDLLLADLGRPK	SSIKVTATVQTTATPSS	
10.	Paxillin-B   paxB	108	117
		ITSISTAI KAVPTTPEEHITH	
	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	
11.	Paxillin-B   paxB	231	240
	Primary and secondary s	sequence	
	SQSQPQPYKVTATNSQPSSDDLDELLKGLSF	PSTTTTTTVPPPVQRDQHQH	
	СССССССССССССССССННННННСССС		
12.	Paxillin-B   paxB	310	319
	Primary and secondary s		
	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	HHHCCCCCCCCCCCCCCCCC	
13	Transforming growth factor beta-1-induced transcript 1	3	12
	protein   TGFB111	, , , , , , , , , , , , , , , , , , ,	
	Primary and secondary s	sequence	
	MEDLDALLSDLETTTSH	MPRSGAPKERPAEPL	
14.	Transforming growth factor beta-1-induced transcript 1	92	101
	Protein   TGFBTT	equence	
	AAPAAPPFSSSSGVLGTGLCELDRLLQELN	ATQFNITDEIMSQFPSSKVA	
	СССССССССССССССНННННННННН	222222222222222222222222222222222222222	
15.	Transforming growth factor beta-1-induced transcript 1	157	166
	protein   TGFB1I1		
	Primary and secondary s		
		FRVQNHLPASGPTQPPVVS	
16	Transforming growth factor bate 1 induced transprint (		212
10.	mansionning grown factor beta- i-induced transcript	203	212
	Primarv and secondarv s	sequence	
	PVVSSTNEGSPSPPEPTGKGSLDTMLGLLQS	DLSRRGVPTQAKGLCGSCN	
	СССССССССССССССССССССННННННН	ннсссссссссссссссссс	
17.	Zinc finger CCCH domain-containing protein 7B   ZC3H	7B 280	289
	Primary and secondary s		
	RTLPSTDSLDDFSDGDVFGPELDTLLDSLSL	VQGGLSGSGVPSELPQLIP	
10			014
10.	Primary and secondary	sequence	314
	SILYSSSGEI ADI ENEDIEPEI DDII VHV/KGM	IORIVNOWSEKFSDEGDSD	
	ССССССССССССССССССНННННННН	ннннннсссссссссс	

# **Supplementary Table 3.B**: Information of the 13 predict LD motifs from round1 predicted by LDMF

	Index	Protein name	Start position	End position
1.		Band 4.1-like protein 5   EPB41L5	634	643
		Primary and secondary s	sequence	
		ETLMLITPADSGSVLKEATDELDALLASLTE	ENLIDHTVAPQVSSTSMITP	
0			AHHUUUUUUUUUUUUUUUUUUU	
2.		Insulin-like growth factor-binding protein 2	230	239
		Primary and secondary	sequence	
		LGLEEPKKLRPPPARTPCQQELDQVLERIS	[MRLPDERGPLEHLYSLHIP	
		ССССССССССССССННННННННННН	000000000000000000000000000000000000000	C
3.		Protein C8orf37   C8orf37	4	13
		Primary and secondary s	sequence	
		MAEDLDELLDEVESKFCT	PDLLRRGMVEQPKGC	
4		CHHHHHHHHHHHHHHHH		1690
4.			1680	1689
		Primary and secondary s	sequence	
		QFKRFRETVPTWDTIRDEEDVLDELLQYLG	VTSPECLQRTGISLNIPAPQ	
		СССССССССССССССННННННННННС	000000000000000000000000000000000000000	C
5.		Uncharacterized protein C16orf71   C16orf71	267	276
		Primary and secondary s	sequence	
		PLVEPPEGPPVLSLQQLEAWDLDDILQSLAG	QEDNQGNRAPGTVWWAADH	
0.		Primary and secondary		132
		GNPGGKTLEERRSSLDAEIDSLTSILADLEC	SSPYKPRPPQSSTGSTASP	
		СССССССССССССССНННННННННСС	000000000000000000000000000000000000000	C
7.		Pre-mRNA 3'-end-processing factor FIP1	5	14
		FIP1L1		
		CCHHHHHHHHHHHHHCCCCC		
8.		Calpastatin   CAST	156	165
0.		Primary and secondary s	sequence	
		PAVPVESKPDKPSGKSGMDAALDDLIDTLG	<b>GPEETEEENTTYTGPEVSDP</b>	
		ССССССССССССССССННННННННСС	000000000000000000000000000000000000000	C
9.		Nuclear receptor coactivator 2   NCOA2	805	814
			CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	20
10		Nuclear receptor coactivator 3   NCOA3	799	808
		Primary and secondary s	sequence	
		QEKDPKIKTETSEEGSGDLDNLDAILGDLTS	SDFYNNSISSNGSHLGTKQ	
		ССССССССССССССССССНННННСССС	000000000000000000000000000000000000000	C
11.		WASP homolog-associated protein with actin,	22	31
		membranes and microtubules   WHAMM		
		VCESPAERPRDSI ESESCPGSMDEVI ASI RI	HGRAPI RKVEVPAVRPPHAS	
		СССССССССССССССССНННННННН	000000000000000000000000000000000000000	c
12.		Ral GTPase-activating protein subunit alpha-2	1519	1528
		RALGAPA2		
		Primary and secondary		
				°C
13		Purkinie cell protein 2 62	71	
		homoloa   PCP2	71	
		Primary and secondary s	sequence	
		RCSLQAGPGQTTKSQSDPTPEMDSLMDMLA	STQGRRMDDQRVTVSSLPGI	=
		СССССССССССССССССНННННННН	НННСССССССССССССССССССССССССССССССССССС	C

**Supplementary Table 3.C**: Information of the 12 new LD motifs finally suggested by LDMF.

	Index	Protein name	Start position	End position
1.	macx	Band 4.1-like protein 5   EPB41L5	634	643
		Primary and secondary se	quence	
		ETLMLITPADSGSVLKEATDELDALLASLTEN HHHHCCCCCCCCCCCCCCHHHHHHHHHH	ILIDHTVAPQVSSTSMITP HHCCCCCCCCCCCCCCCC	C
2.		Serine/threonine-protein phosphatase 2A	508	517
		regulatory subunit B'' subunit alpha   PPP2R3A		
		Primary and secondary se	quence	
		KVSKFEEGDQRDFTNSSSQEEIDKLLMDLESF	SQKMETSLREPLAKGKNS	
		СССССССССССССССНННННННННН	нннннннссссссссс	C
3.		Colled-coll domain-containing protein 158	903	912
		Primary and secondary se	quence	
		ASFLSHHSTKANTLKEDPTRDLKQLLQELRSV	/INEEPAVSLSKTEEDGRT	
		НННСССССССССССССННННННННН	нннсссссссссссссссс	C
4.		Ral G I Pase-activating protein subunit alpha-1 I RALGAPA1	1680	1689
		Primary and secondary se	quence	
		QFKRFRETVPTWDTIRDEEDVLDELLQYLGV	TSPECLQRTGISLNIPAPQ	•
				C 070
5.		Uncharacterized protein C160ff71   C160ff71 Primary and secondary se	207 quence	276
		PLVEPPEGPPVLSLQQLEAWDLDDILQSLAGQE	EDNQGNRAPGTVWWAADH	
		СССССССССССССССССННННННННС	000000000000000000000000000000000000000	C
6.		Lipoma-preferred partner   LPP	123	132
		Primary and secondary se		
		CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC		C
7.		Cyclic AMP-responsive element-binding	49	58
		protein 3   CREB3		
		Primary and secondary se	QUENCE	
		ННННСССССССССССССССНННННННССС		C
8.		Calpastatin   CAST	156	165
			COCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	°C
9.		Nuclear receptor coactivator 2   NCOA2	805	814
		Primary and secondary se	quence	
		KTEKEEMSFEPGDQPGSELDNLEEILDDLQNS	QLPQLFPDTRPGAPAGSV	
10			700	
10.		Primary and secondary se	auence	000
		QEKDPKIKTETSEEGSGDLDNLDAILGDLTSS	DFYNNSISSNGSHLGTKQ	
		ССССССССССССССССССНННННССССС	000000000000000000000000000000000000000	C
11.		Protein C8orf37   C8orf37	4	13
		MAEDLDELLDEVESKFCTPI	DLLRRGMVEQPKGC	
		СННННННННННННННССССС	222222222222222222222222222222222222222	
12.		Ral GTPase-activating protein subunit alpha-2	1519	1528
		RALGAPA2	guonoo	
		Primary and secondary se WHRDTEGPOKDSSOVEEGDDVI DKI I ENIGH	TSPECI I PSQI NI NEPSI T	
		ссссссссссссссссссснннннннсс		C

# **Supplementary Table 4: Computational Validation**

Summary of the bioinformatic search for evidence supportive of interactions between known LDmotif binding proteins and the LDMF-predicted LD motif–containing proteins from the human proteome. We assessed the *LDMF* predictions using four computational methods, namely PrePPI (a Bayesian framework that combines structural, functional, evolutionary and expression information (Zhang, et al., 2012)), GeneFriends (an RNAseq-based gene co-expression network (van Dam, et al., 2015)), CoCiter (which evaluates the significance of literature co-citations (Qiao, et al., 2013)).

To allow straightforward reproducibility, gene names are given in the table. The corresponding protein names for the LDBD-containing proteins are XPO1: exportin; PABPC1: polyadenulatebinding protein 1 (PABP-1); VLC: vinculin; TLN1: talin; PARVA:  $\alpha$ -parvin; PARVB:  $\beta$ -parvin; PARVG: γ-parvin; PDCD10: programmed cell death protein 10/cerebral cavernous malformations 3 protein (CCM3); PTK2: focal adhesion kinase (FAK); PTK2B: Protein-tyrosine kinase 2β (PYK2); GIT1: Arf GPTase-activating protein/GRK-interacting protein 1 (GIT1); GIT2: Arf GPTase-activating protein/GRK-interacting protein 2 (GIT2); BCL2: Apoptosis regulator Bcl-2. The corresponding protein names for the predicted LD motif-containing proteins are EPB41L5: Band 4.1-like protein 5 (E41L5); LPP: lipoma-preferred partner (LPP); RALGAPA1: Ral GTPase-activating protein subunit  $\alpha$ -1 (RGPA1); PPP2R3a: Serine/threonine-protein phosphatase 2A regulatory subunit B" subunit  $\alpha$  (P2R3A); CCDC158: coiled-coil domaincontaining protein 158 (CD158); C16orf71: uncharacterized protein C16orf71 (CP071); NCOA2: nuclear receptor coactivator 2 (NCOA2); NCOA3: nuclear receptor coactivator 3 (NCOA3); CAST: calpastatin (CAST); CREB3: cyclic AMP-responsive element-binding protein 3 (CREB3); RALGAPA2: Ral GTPase-activating protein subunit  $\alpha$ -2 (RGPA2); C8orf37: uncharacterized protein C8orf37 (CP037).

	Highly likely													
	gene name	XP01	PABPC1	VCL	TLN1	PARVA	PARVB	PARVG	PDCD10	РТК2	РТК2В	GIT1	GIT2	BCL2
	EPB41L5	0.51	-	0.78	0.99, Small*	0.55	Small*	0.6, Small*	0.67	0.99, Small	1, Small*	-	Small	0.9
	LPP	0.98	0.019	Medium, 0.011	<mark>Small,</mark> 0.015	0.51, Large, 0.01	Small*	Small*	-	Small	Small*, 0.002	0.001	Small	0.047
	RALGAPA1	Small	Small*	-	-	-	Small*	-	-	0.76, Small	1	-	Medium	Small
	PPP2R3A	-	Small*	Small	-	Small	-	Small*	-	Small	Small*	-	-	-
	CCDC158	Small*	-	-	-	-	-	-	-	-	-	Small	-	-
	C16orf71	-	-	-	-	-	-	-	-	-	-	Small	-	-
	less likely													
Г														

gene name	XPO1	PABPC1	VCL	TLN1	PARVA	PARVB	PARVG	PDCD10	РТК2	РТК2В	GΠ1	GIT2	BCL2
NCOA2	Small	-	0.041	Small	Small*	-	Small	-	-	Small, 0.007	0.046	Large	Small, 0.015
NCOA3	0.98, Small	Small	-	-	Small*	-	-	-	-	Small, 0.016	-	Small	0.77, Small, 0.026
CAST	-	-	Medium, 0.033	Medium	Small	-	-	-	-	Small*, 0.007	0.036	Small	Small, 0.047
CREB3	-	-	Small	Small	Medium	-	Small*	-	0.77	0.87, Small*	Small	Small*	-
least likely													

PABPC1 XPO1 VCL TLN1 PARVA PARVB PARVG PDCD10 РТК2 РТК2В GIT1 GIT2 BCI 2 gene name Small Small\* 0.93, Small Mediu Small RALGAPA2 Small --Small -\_ Small\* C8orf37 Small ----Small Small -

Blue is probability score>0.5 from PrePPI tool.

Red is the pearson correlation score from GeneFriends tool between 2 associated genes based on the idea of co-expression. Small is in a range [0.1,0.3). Medium is in a range [0.3,0.5). Large is in a range [0.5,1]. Star(\*) means negative correlation.

Green is the p-value<0.05 from CoCiter tool.

# Supplementary Table 5: List of protein accession IDs containing an LD motif

Index	Protein name	Uni-prot	Organism
4		accession ID	
1		P49023	Homo sapiens
2		060711	Homo sapiens
3			Dictyostelium discoldeum (Slime mold)
4	TGFI1_HUMAN	043294	Homo sapiens
5	Z3H7B_HUMAN	Q9UGR2	Homo sapiens
6	RHG07_HUMAN	Q96QB1	Homo sapiens
7	E41L5_HUMAN	Q9HCM4	Homo sapiens
8	P2R3A_HUMAN	Q06190	Homo sapiens
9	CD158_HUMAN	Q5M9N0	Homo sapiens
10	RGPA1_HUMAN	Q6GYQ0	Homo sapiens
11	CP071_HUMAN	Q8IYS4	Homo sapiens
12	LPP_HUMAN	Q93052	Homo sapiens
13	CREB3_HUMAN	O43889	Homo sapiens
14	ICAL_HUMAN	P20810	Homo sapiens
15	NCOA2_HUMAN	Q15596	Homo sapiens
16	NCOA3_HUMAN	Q9Y6Q9	Homo sapiens
17	CH037_HUMAN	Q96NL8	Homo sapiens
18	RGPA2_HUMAN	Q2PPJ7	Homo sapiens
19	CAOG 06505	A0A0D2WU78	Capsaspora owczarzaki
20	H696 03850	A0A058Z589	Fonticula alba
21	AMSG 03633	A0A0L0D4P1	Thecamonas trahens
22	AMAG 12229	A0A0L0SXV7	Allomyces macrogynus
23	AMAG 05806	A0A0L0SDD3	Allomyces macrogynus
24	PTSG 08126	F2UI26	Salpingoeca rosetta
25	SPPG 03527	A0A0L0HLP9	Spizellomyces punctatus
26	AMSG 09408		Thecamonas trahens
27	AMSG 06733		Thecamonas trahens
28	AMAG 15572		Allomyces macrogynus
29	AMAG 16615		Allomyces macrogynus
30	AMSG 00750		Thecamonas trahens
31	CAOG 05072		Cansaspora owczarzaki
32	H696 03440	A0A05876W/5	Fonticula alba
33	CAOG 08445		Cansaspora owczarzaki
34	CAOG 09259		Capsaspora owczarzaki
35	H696 03144	Δ0Δ0587Δ31	Fonticula alba
36	H696_05642	A0A050ZA51 A0A05870V1	Fonticula alba
37	MONBRODAET 12816		Monosiga brovicollis
38	MONBRODAET 24303		Monosiga brovicollis
30	MONBRODAET 25678	A907700	Monosiga brovicollis
40	MONBRODAET 37245		Monosiga brovicollis
40	AMAC 10344		
10			
42	ANAC 16050		
43	SPPC 06472		
44	SFFG_004/3		
40	MONPODALT 24042		
40			
4/	AIVIAG_U 1834		
40			
49			

50	AMAG_04337	A0A0L0S8Q4	Allomyces macrogynus
51	MVEG_05104	KFH68286	Mortierella verticillata
52	MVEG_11218	A0A086TMK6	Mortierella verticillata
53	AMSG_00923	A0A0L0DIS1	Thecamonas trahens
54	H696_01588T0	A0A058ZE07	Fonticula alba
55	H696_01588T1	A0A058ZFD2	Fonticula alba
56	STK4L_DICDI	Q55FS2	Dictyostelium discoideum
57	AMSG_12414	A0A0L0DSV4	Thecamonas trahens
58	H696_00097	A0A058ZGA4	Fonticula alba
59	AMAG_19087	A0A0L0SN30	Allomyces macrogynus
60	AMAG_12080	A0A0L0SYP4	Allomyces macrogynus
61	AMAG_07384	A0A0L0SI62	Allomyces macrogynus
62	AMAG_20576	A0A0L0TDJ3	Allomyces macrogynus
63	AMAG_12596	A0A0L0SZN5	Allomyces macrogynus
64	AMAG_18842	A0A0L0SIJ4	Allomyces macrogynus
65	AMAG_03098	A0A0L0S4M3	Allomyces macrogynus
66	AMAG_15148	A0A0L0T5Z2	Allomyces macrogynus
67	CAOG_02903	A0A0D2U9S0	Capsaspora owczarzaki
68	PTSG_03224	F2U4K6	Salpingoeca rosetta
69	PTSG_03247	F2U4M7	Salpingoeca rosetta
70	PTSG_04113	F2U6M3	Salpingoeca rosetta
71	PTSG_05218	F2UAU8	Salpingoeca rosetta
72	PTSG_08076	F2UHX6	Salpingoeca rosetta
73	PTSG_10623	F2URW3	Salpingoeca rosetta
74	SARC_01803	A0A0L0GAM7	Sphaeroforma arctica
75	SPPG_01555	A0A0L0HRY4	Spizellomyces punctatus
76	SPPG_01829	A0A0L0HMU1	Spizellomyces punctatus
77	SPPG_02883	A0A0L0HMU5	Spizellomyces punctatus
78	AMSG_00544	A0A0L0D8R0	Thecamonas trahens
79	AMSG_01297	A0A0L0DMR4	Thecamonas trahens
80	AMSG_04901	A0A0L0D889	Thecamonas trahens
81	AMSG_05859	A0A0L0DD16	Thecamonas trahens
82	H696_04460	A0A058Z470	Fonticula alba
82	AMAG_06527	A0A0L0SH78	Allomyces macrogynus

# References

Alam, T., *et al.* How to find a leucine in a haystack? Structure, ligand recognition and regulation of leucine-aspartic acid (LD) motifs. *The Biochemical journal* 2014;460(3):317-329.

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997;25(17):3389-3402.

Arnold, K., et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006;22(2):195-201.

Arold, S.T., Hoellerer, M.K. and Noble, M.E. The structural basis of localization and signaling by the focal adhesion targeting domain. *Structure* 2002;10(3):319-327.

Astro, V., *et al.* Liprin-alpha1 regulates breast cancer cell invasion by affecting cell motility, invadopodia and extracellular matrix degradation. *Oncogene* 2011;30(15):1841-1849.

Backer, J.M., *et al.* Phosphatidylinositol 3'-kinase is activated by association with IRS-1 during insulin stimulation. *The EMBO journal* 1992;11(9):3469-3479.

Bajic, V.B., *et al.* Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. *Journal of molecular graphics & modelling* 2003;21(5):323-332.

Brown, M.C., Curtis, M.S. and Turner, C.E. Paxillin LD motifs may define a new family of protein recognition domains. *Nat Struct Biol* 1998;5(8):677-678.

Carragher, N.O., *et al.* A novel role for FAK as a protease-targeting adaptor protein: regulation by p42 ERK and Src. *Curr Biol* 2003;13(16):1442-1450.

Cooray, P., et al. Focal adhesion kinase (pp125FAK) cleavage and regulation by calpain. *The Biochemical journal* 1996;318 (Pt 1):41-47.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning* 1995;20(3):273-297. Durkin, M.E., *et al.* DLC-1:a Rho GTPase-activating protein and tumour suppressor. *Journal of cellular and molecular medicine* 2007;11(5):1185-1207.

Fauchere, J.L., *et al.* Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research* 1988;32(4):269-278.

Gorenne, I., *et al.* LPP expression during in vitro smooth muscle differentiation and stentinduced vascular injury. *Circ Res* 2006;98(3):378-385.

Gorenne, I., et al. LPP, a LIM protein highly expressed in smooth muscle. American journal of physiology. Cell physiology 2003;285(3):C674-685.

Grant, C.E., Bailey, T.L. and Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 2011;27(7):1017-1018.

Heon, E., *et al.* Mutations in C8ORF37 cause Bardet Biedl syndrome (BBS21). *Hum Mol Genet* 2016;25(11):2283-2294.

Hoellerer, M.K., *et al.* Molecular recognition of paxillin LD motifs by the focal adhesion targeting domain. *Structure* 2003;11(10):1207-1217.

Ishida, T. and Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic acids research* 2007;35(Web Server issue):W460-464.

Ito, A., *et al.* A truncated isoform of the PP2A B56 subunit promotes cell motility through paxillin phosphorylation. *The EMBO journal* 2000;19(4):562-571.

Kall, L., Krogh, A. and Sonnhammer, E.L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic acids research* 2007;35(Web Server issue):W429-432.

Kallberg, M., *et al.* RaptorX server: a resource for template-based protein structure modeling. *Methods in molecular biology* 2014;1137:17-27.

Kallberg, M., *et al.* Template-based protein structure modeling using the RaptorX web server. *Nature protocols* 2012;7(8):1511-1522.

Kawada, M., *et al.* Cytostatin, an inhibitor of cell adhesion to extracellular matrix, selectively inhibits protein phosphatase 2A. *Biochimica et biophysica acta* 1999;1452(2):209-217. Kawashima, S., *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;36(Database issue):D202-205.

Krystkowiak, I. and Davey, N.E. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic acids research* 2017;45(W1):W464-W469.

Krystkowiak, I., Manguy, J. and Davey, N.E. PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic acids research* 2018;46(W1):W235-W241.

la Cour, T., *et al.* Ànalysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel* 2004;17(6):527-536.

Lo, S.H. Tensin. *The international journal of biochemistry & cell biology* 2004;36(1):31-34. Lorenz, S., *et al.* Structural analysis of the interactions between paxillin LD motifs and alpha-parvin. *Structure* 2008;16(10):1521-1531.

McGuffin, L.J., Bryson, K. and Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics (Oxford, England)* 2000;16(4):404-405.

Moldoveanu, T., Gehring, K. and Green, D.R. Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains. *Nature* 2008;456(7220):404-408.

Moscardo, A., *et al.* Serine/threonine phosphatases regulate platelet alphallbbeta3 integrin receptor outside-in signaling mechanisms and clot retraction. *Life Sci* 2013;93(20):707-713.

Nag, S., et al. Ca2+ binding by domain 2 plays a critical role in the activation and stabilization of gelsolin. *Proc Natl Acad Sci U S A* 2009;106(33):13713-13718.

Petit, M.M., *et al.* LPP, an actin cytoskeleton protein related to zyxin, harbors a nuclear export signal and transcriptional activation capacity. *Mol Biol Cell* 2000;11(1):117-129.

Petit, M.M., Meulemans, S.M. and Van de Ven, W.J. The focal adhesion and nuclear targeting capacity of the LIM-containing lipoma-preferred partner (LPP) protein. *The Journal of biological chemistry* 2003;278(4):2157-2168.

Qiao, N., *et al.* CoCiter: an efficient tool to infer gene function by assessing the significance of literature co-citation. *PloS one* 2013;8(9):e74074.

Quan, J. and Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PloS one* 2009;4(7):e6441.

Ravi, A., *et al.* Epidermal growth factor activates the Rho GTPase-activating protein (GAP) Deleted in Liver Cancer 1 via focal adhesion kinase and protein phosphatase 2A. *The Journal of biological chemistry* 2015;290(7):4149-4162.

Schell, C., *et al.* The FERM protein EPB41L5 regulates actomyosin contractility and focal adhesion formation to maintain the kidney filtration barrier. *Proceedings of the National Academy of Sciences of the United States of America* 2017;114(23):E4621-E4630.

Schmalzigaug, R., *et al.* GIT1 utilizes a focal adhesion targeting-homology domain to bind paxillin. *Cellular signalling* 2007;19(8):1733-1744.

Schwede, T., *et al.* SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003;31(13):3381-3385.

Shawe-Taylor, J. and Cristianini, N. Kernel methods for pattern analysis. Cambridge university press; 2004.

Shirakawa, R., *et al.* Tuberous Sclerosis Tumor Suppressor Complex-like Complexes Act as GTPase-activating Proteins for Ral GTPases. *Journal of Biological Chemistry* 2009;284(32):21580-21588.

Stashi, E., York, B. and O'Malley, B.W. Steroid receptor coactivators: servants and masters for control of systems metabolism. *Trends Endocrinol Metab* 2014;25(7):337-347.

Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 2000;28(6):1102, 1104.

Tumbarello, D.A., Brown, M.C. and Turner, C.E. The paxillin LD motifs. *FEBS letters* 2002;513(1):114-118.

UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* 2015;43(Database issue):D204-212.

van Dam, S., Craig, T. and de Magalhaes, J.P. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic acids research* 2015;43(Database issue):D1124-1132.

van Zundert, G.C.P., *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of molecular biology* 2016;428(4):720-725.

Vitour, D., *et al.* RoXaN, a novel cellular protein containing TPR, LD, and zinc finger motifs, forms a ternary complex with eukaryotic initiation factor 4G and rotavirus NSP3. *J Virol* 2004;78(8):3851-3862.

Wang, G. and Dunbrack, R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* (*Oxford, England*) 2003;19(12):1589-1591.

Williamson, M.P. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Mag Res Sp* 2013;73:1-16.

Zacharchenko, T., *et al.* LD Motif Recognition by Talin: Structure of the Talin-DLC1 Complex. *Structure* 2016;24(7):1130-1141.

Zhang, Q.C., *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 2012;490(7421):556-560.