

Supplementary Material for:

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

Tanvir Alam^{1,2,#}, Meshari Alazmi^{1,2,#}, Rayan Naser^{1,3,#}, Franceline Huser^{1,3,#}, Afaq A. Momin^{1,3,#}, Katarzyna W. Walkiewicz^{1,3,5}, Christian G. Canlas⁴, Raphaël G. Huser², Amal J. Ali³, Jasmeen Merzaban³, Vladimir B. Bajic^{1,2,*}, Xin Gao^{1,2,*}, Stefan T. Arold^{1,3,*}

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

1. Computational Bioscience Research Center (CBRC);
2. Division of Computer, Electrical and Mathematical Sciences & Engineering (CEMSE);
3. Division of Biological and Environmental Sciences and Engineering (BESE);
4. Imaging and Characterization Core Lab
5. Current address: NanoTemper Technologies GmbH, Floessergasse 4, Munich, Germany

* 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

Short Title:

Proteome-wide prediction and function of LD motifs

KEYWORDS

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

Supplementary Figure 1

Disorder predictions for the *bona fide* LD motifs, established by MetaPrDos. The 2-state prediction shows residues predicted to be disordered in red. These red residues correspond to sequence positions with a disorder tendency above 0.5 in the disorder profile plot.

1 Paxillin

PREDICTION RESULT: P49024-pax-chick

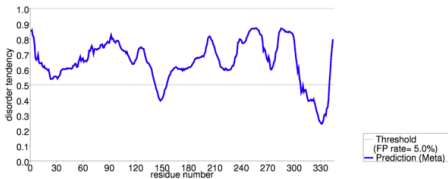
Prediction false positive rate: 5.0%

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	MDDLALLAD	LESTTSHISK	RPVFLTEETP	YSYPTGNHTY	QEIAVPPFPV	90
51	PPPSSEALNG	TVIDPLDQWQ	FSVSRYGHQQ	PQSQSPIYSS	SAKSSSSASVP	100
101	RDGLSSPSR	ASEEEHVYSF	PNKQKSAEPS	FTMTSTSLGS	NLSELDRLLL	150
151	ELNAVQHNP	SGFSADEVSR	SPSLPNVTGP	HYVIPSSSSS	AGGKAAPPTK	200
201	EKPKRNGGRG	IEDVRPVSVE	LLDELESSVP	SPVPAITVSQ	GEVSSPQVRN	250
251	ASQQQTRISA	SSATRELEDEL	MASLSDFKFM	AQKGAGSSSS	PPSTTPKPGS	300
301	QLDTMLGSLQ	SDLNKLGVAT	VARGVCGACK	KPIAGQVVTA	MGKT	350

Disorder profile plot



2 Hic-5 protein

PREDICTION RESULT: hic-5

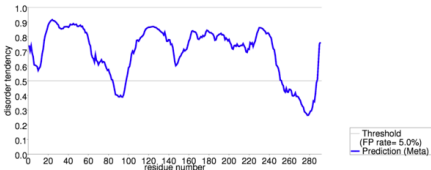
Prediction false positive rate: 5.0%

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	MEDLDALLAD	LQITTPFRCP	VLLTDSPEKP	QPTETRPPPP	PYDEKTAMSN	90
51	KTSDHETFPV	DKDHLYSIVQ	KYPLFSVSPA	LGGGLCELDL	LLNELNATQF	100
101	NITDEIMSQF	PTRDPSEQKA	EAQKEAEKRA	LSASSATLEL	DRIMASLSDF	150
151	HKQNTVSQEV	EAPGAYKGSE	EVSRRPGTDE	LSSPRSTACV	PKDLEDAPTP	200
201	KSPKVVSAAG	HLEVKTQNVN	SDEVTSASRP	DSVSGSKVPE	ATSVPRSDLD	250
251	SMLVKLQSG	KQQGIETYSK	GLCESQRP	AGQVVTA	LGH T	300

Disorder profile plot



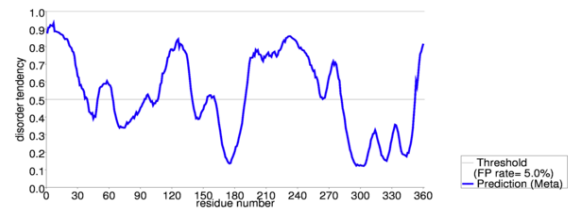
3 DLC1

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	NSTQTSSSSS	QSETSSAVST	PSPVTTRSL	STCNKRVGMY	LEGFDFFSQS	50
51	TLNNVTEQNY	KNRESYPEDT	VFYIPEDHKP	GTFFKALSHG	SFCPSGNSSV	100
101	NWRTGSFHGP	GHLSLRRENS	HDSPKELKRR	NSSSSLSSRL	SIYDNVPGSI	150
151	LYSSSGELAD	LENEDIFPEL	DDILYHVKGM	QRIVNQWSEK	FDEGDSDSA	200
201	LDSVSPCPSS	FKQIHLVDH	DRRTPSDLDS	TGNSLNEFEE	PTDIPERRDS	250
251	GVGASLTRCN	RHRLRWHSFQ	SSHRPSLSNV	SLQINCQSV	QMNLQKQYSL	300
301	LKLTALLEKY	TPSNKHGFSW	AVPKFMKRIK	VPDYKDRSVF	GVPLTVNVQR	350
351	SCQPLPQSIQ					400

Disorder profile plot



4 Roxan

PREDICTION RESULT: roxan

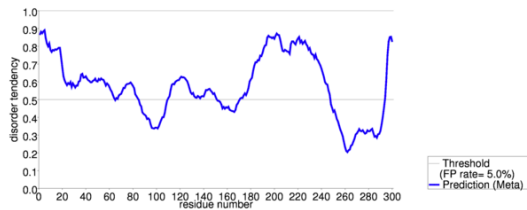
Prediction false positive rate: 5.0%

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	SLLSNGTAAG	VADQGTSNGL	GSIDDIETGN	VPDTREQVEI	GAPRDCYVDP	50
51	RGSPALLEST	PTMPLFFHVL	DLLAPLDSSR	TLPSTDSLDD	FSDGDVFGPE	100
101	LDTLILDSLSL	VQGLSGSGV	PSELPLQIPV	FPGTPLLFP	VVGGSIPVSS	150
151	PLPPASFGLV	MDPSKKLAAS	VLDALDPFPG	TLDPLDLLFY	SETRLDALDS	200
201	FGSTRGSLDK	PDSFMEETNS	QDHRPPSGAQ	KPAPSPEPCM	PNTALLIKNP	250
251	LAATHEFKQA	CQLCYPKTGP	RAGDITYREG	LEHKCKRDIL	LGRLRSSDQ	300

Disorder profile plot



5 E6BP

PREDICTION RESULT: Q05086-E6ap

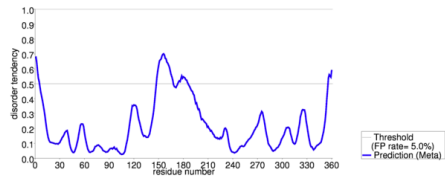
Prediction false positive rate: 5.0% [Change FP rate](#)

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	RVYTRLISNE	KIETAFLNAL	VYLSPNVECD	LTYNHVYSRD	PNYLNLFIIIV	50
51	MENRNLSHPE	YLEMALPLFC	KAMSKLPLAA	QGKLIRLWSK	YNADQIRRMQ	100
101	ETFQQLITYK	VISNEFNRSN	LVNDDAIVA	ASKCLRMVYY	ANVVGGEVDT	150
151	NHNEEDDEEP	IPESSELTLQ	ELLGEERRNK	KGPRVDFLET	ELGVKTLDGR	200
201	KPLIFFEETI	NEPLNEVLEM	DKDYTFKVE	TENKFSFMTG	PFILNAVTKN	250
251	LGLYYDNRIK	MYSERRITVL	YSLVQGQQLN	PYLRLKVRDR	HIIDDALVRL	300
301	EMIAMENPAD	LKKQLYVEFE	GEQGVDEGGV	SKEFFQLVVE	EIFNPDIGMF	350
351	TYDESTKLFW					400

Disorder profile plot



6 E6BP/ERC-55

PREDICTION RESULT: Q14257-Erc55

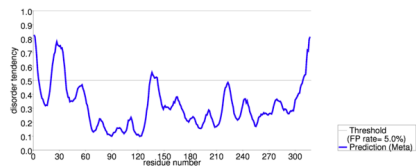
Prediction false positive rate: 5.0% [Change FP rate](#)

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	MRLGPRTAAL	GLLLCAAAA	GAGKAELHY	PLGERRSDDY	REALLGVQED	50
51	VDEYVKLGHE	EQQKRLQAI	KKIDLSDGF	LTESELSSWI	QMSFKHYAMQ	100
101	EAKQQFVEYD	KNSDDTVTWD	EYNIQMYDRV	IDFDENTALD	DAEEESFRKL	150
151	HLKDKKRFEK	ANQDSGPGLS	LEEFIAFEHP	EEVDYMTFV	IQEALAEHDK	200
201	NGDGFVSLEE	FLGDYRNDPT	ANEDPEWILV	EKDRFVNDYD	KDNDGRLDPO	250
251	ELLFWVVPNN	QGIAQEEALH	LIDEMDLNGD	KKLSEEEILE	NPDLFILTSEA	300
301	TDYGRQLHDD	YFYHDEL				350

Disorder profile plot



7 Gelsolin (mouse)

PREDICTION RESULT: P06396-gelsolin

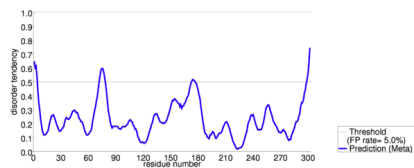
Prediction false positive rate: 5.0%

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	RHGG	RQGQII	YNWGAQSTQ	DEVAASAILT	AQDDEELGGT	PVQSRVVQVK	50
51	EPAHLM	SLFG	GRPMI	IKGG	TSREGGQTAP	ASTRLFQVRA	100
101	VLPKAG	ALNS	NDAFV	LKTPS	AAYLWVG	TGA	150
151	QVAEG	SEPDG	FWALG	GKAA	YRTSP	RLKDK	200
201	IEEV	PGELMQ	EDLAT	DDVML	LDTWD	QVFVM	250
251	IETD	PANRDR	RTPTV	VVKQG	FEPPS	FVGNF	300
301	AA						350

Disorder profile plot



8 CD4

PREDICTION RESULT: P01730-CD4

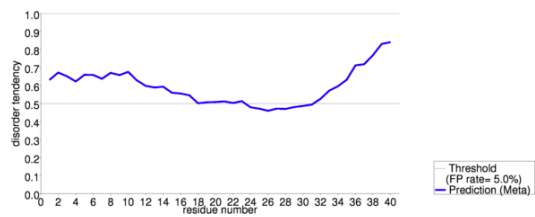
Prediction false positive rate: 5.0%

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	CVRCR	HRRRQ	AERMS	QIKRL	LSEK	KTCQCP	HRFQ	KCSPI	50
---	-------	-------	-------	-------	------	--------	------	-------	----

Disorder profile plot



Supplementary Figure 2

Computational assessment of the structural context of E6BP and of the potential LD motifs proposed by Brown et al. (*Nature Structural Biology*, 1998)

Homology models are coloured according to their secondary structure (magenta: α -helix; cyan: β -strands). The putative LD motif is colored in green, with the LD motif positions 0, 3 and 4 (L⁰XXLL) colored in red.

MetaPrDos (<http://prdos.hgc.jp/>) was used for predicting structural order/disorder from the protein sequence. PHOBIUS (<http://phobius.sbc.su.se/>) was used for prediction of transmembrane helices and signal peptides.

The structural analysis was carried out using the PHYRE2 (<http://www.sbg.bio.ic.ac.uk/phyre2/>), SWISS-MODEL (<http://swissmodel.expasy.org/>) and RaptorX (<http://raptorx.uchicago.edu/>) servers.

Supplementary Figure 2.1. Summary Table

	UNIPROT Entry	Motif sequence and location in protein	Sequence identity of 3D templates for suggested LD motif region
E6BP / ERC-55			
	Q14257; E6BP	208-LEEFLGDYR-216	17-21 % identical
Previously suggested (Brown et al. <i>NSB</i> 1998)			
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-LDAML DLL-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; RDGC	163-LDDLLVVL-170	40 % identical; *
10	P38570; Integrin α 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	P35579; Myosin HC	1422-LDDLLVDL-1429	17 % identical; coiled-coil
14	P24216; Hap2	443-LDVLM TS-450	13 % / 43 % identical (depending on fragment length); *
15	P54762; Eph-2	3-LDYLLLLL-10	Signal peptide; no 3D template
16	P38650; Dynein HC	1361-LDGLLNQL-1368	No template
17	P51592; E3	1453-LDTLLLT L-1460	No template
18	Q04205; tensin	807-LDVLM L DL-814	No template

Legend:

*: 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 degenerate.

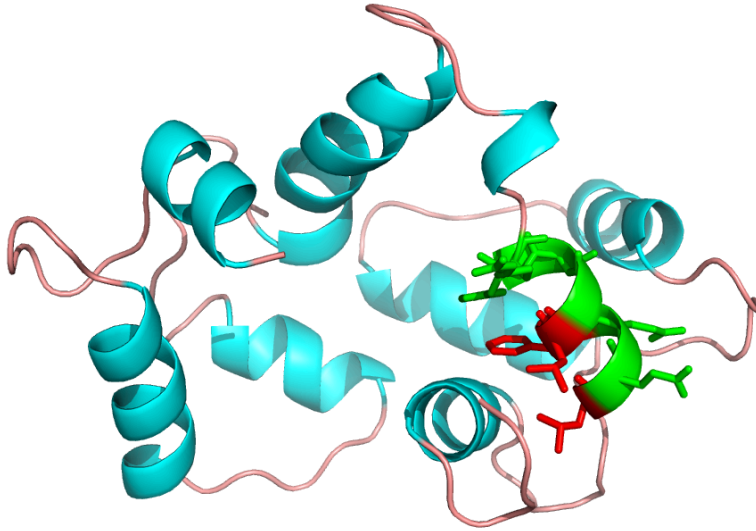
Red shading: this motif is potentially likely to be a bona fide LD motif, because of its structural characteristics and supporting biological evidence.

Supplementary Figure 2.2: Details for E6BP

Q14257; E6BP / ERC-55

Location in protein: 208-LEEFLGDYR-216

Structural Information: 17-21 % sequence identity with EF-hand proteins. Shown is a model built based on PDB 3evv.

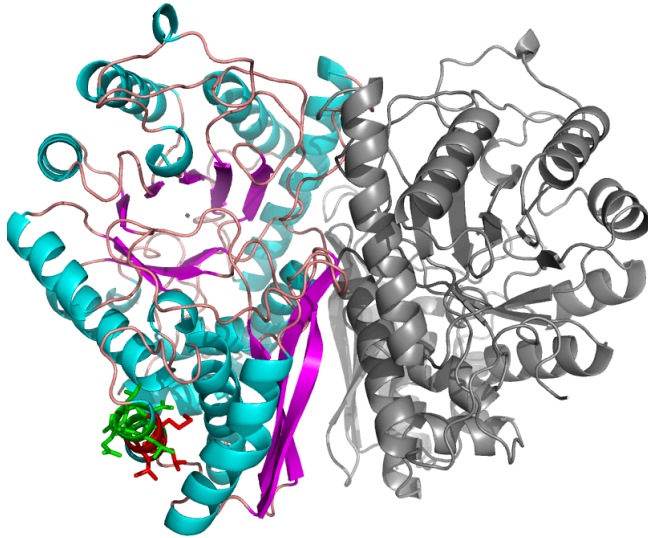


Supplementary Figure 2.3: Details for the 18 LD motifs suggested by Brown et al. (1998)

1 P09104; GAMMA ENOLASE; ENO2

Location in protein: 90-LDNLMLEL-97

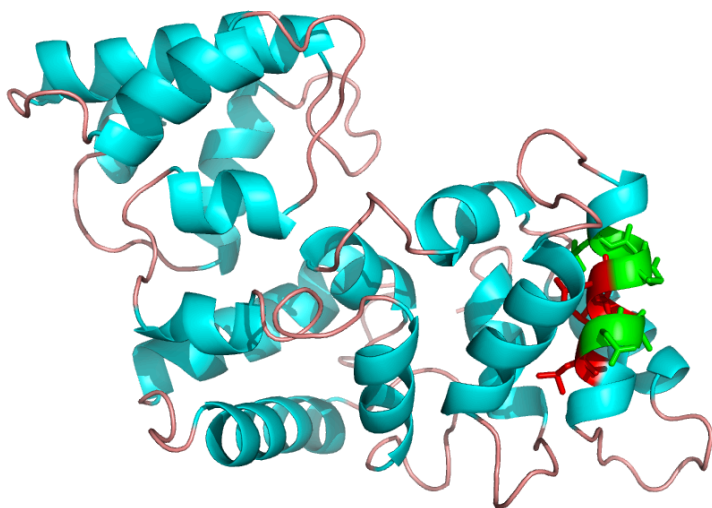
Structural Information: 100% Sequence Identity with PDB 2akm. The suggested LD motif is part of the catalytic domain.



2 P05937; CALBINDIN

Location in protein: 211-LDALLKDL-218

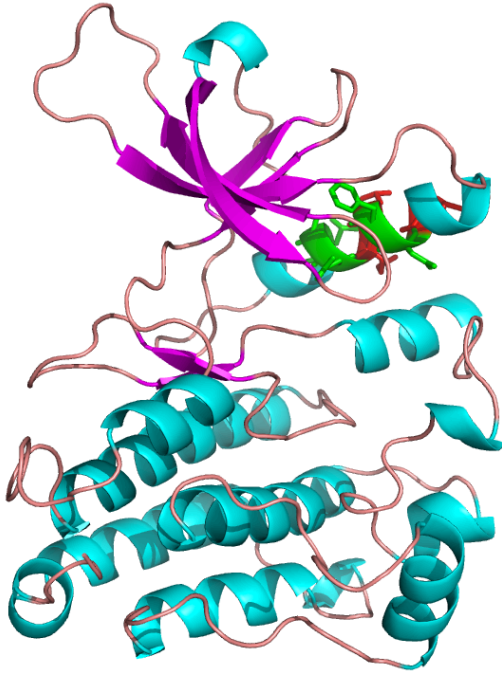
Structural Information: 98% Sequence Identity with PDB Template 2f33A. Forms EF-hand helix-turn-helix



3 P29376; LTK

Location in protein: 556-LDFLMEAL-563

Structural Information: 77% Sequence Identity with PDB 3ics. The LD motif is situated in the α C helix of the protein kinase domain.

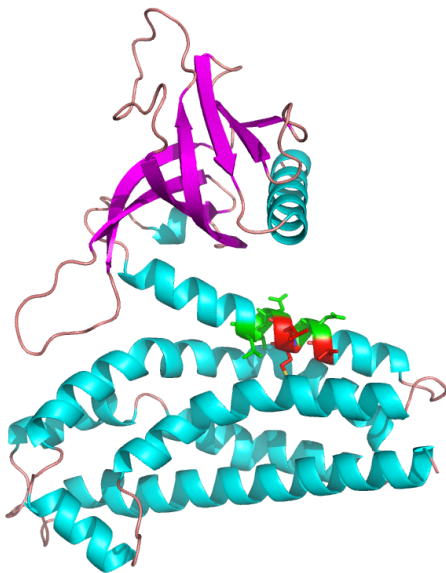


4 P10911; DBL

Location in protein: 662-LDAMLDLL-669

Structural Information: 65 % sequence identity with dbl-homology domain (DH domain);

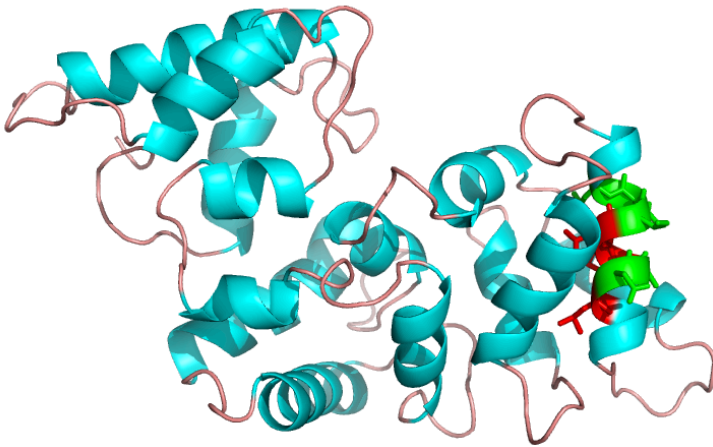
Template PDB 1kz7



5 P22676; CALRETININ; CAB29

Location in protein: 220-LDALLKDI-227

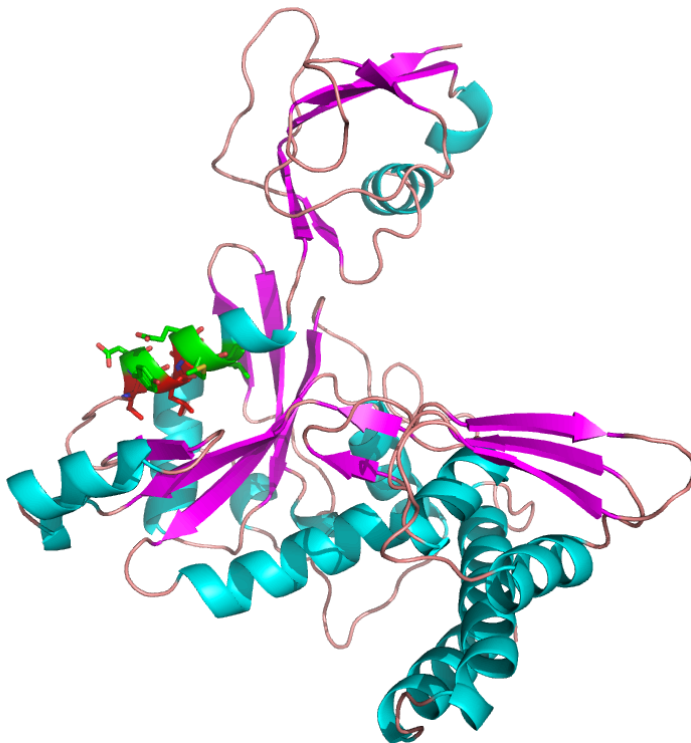
Structural Information: 59 % sequence identity with PDB 2f33; forms EF-hand helix-turn-helix.



6 P55039; DRG

Location in protein: 276-LDYLLLEML-283

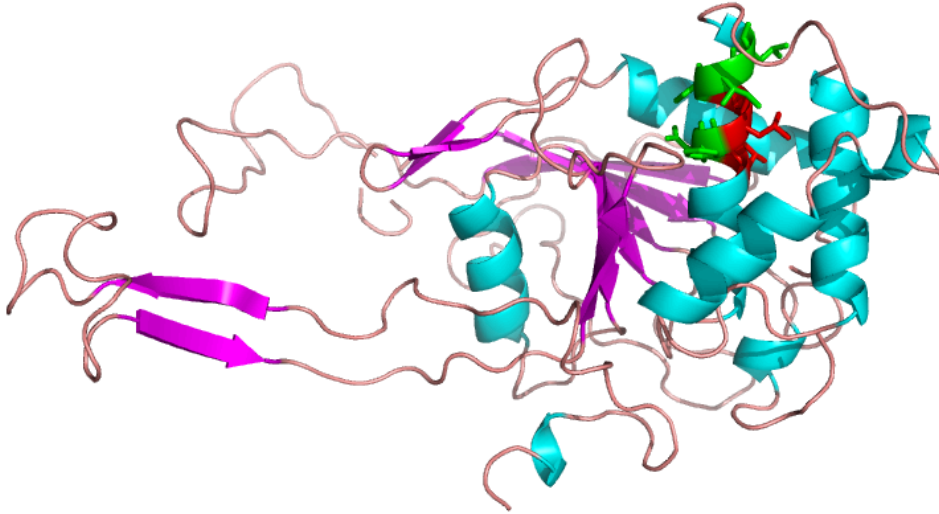
Structural Information: 55% sequence identity with PDB 4a9a



7 P29461; PTP2

Location in protein: 679-LDFLLSIL-686

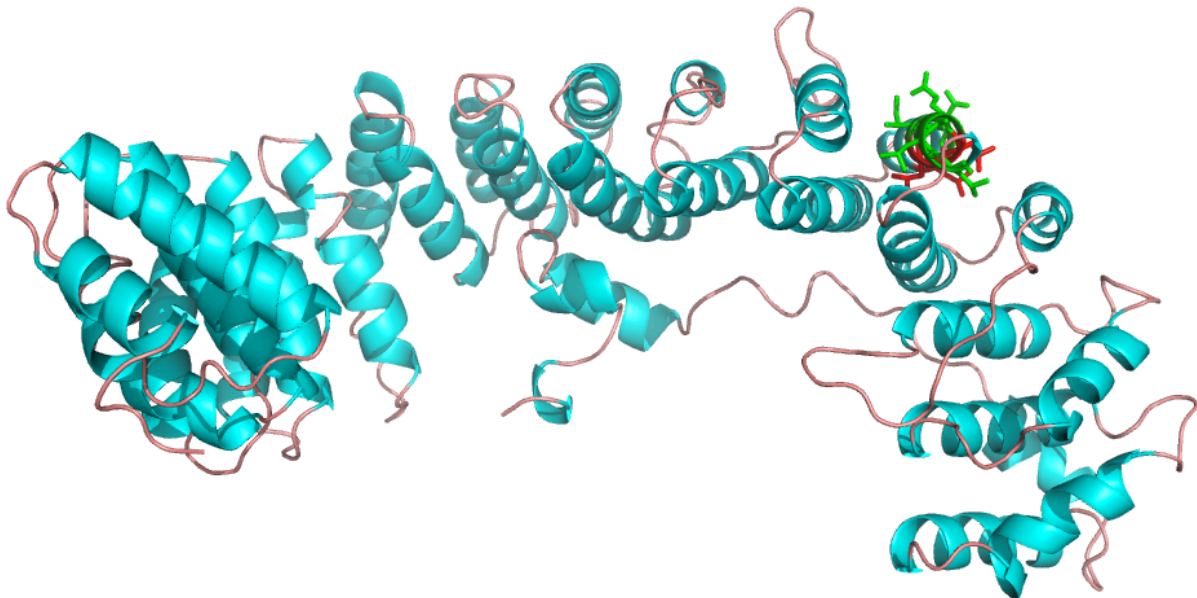
Structural Information: 23.7% Sequence Identity with PDB 3oc3A and 42% identical with 2cfv in the PTP domain.



8 P36000; β -ADAPTIN

Location in protein: 409-LDILLELL-416

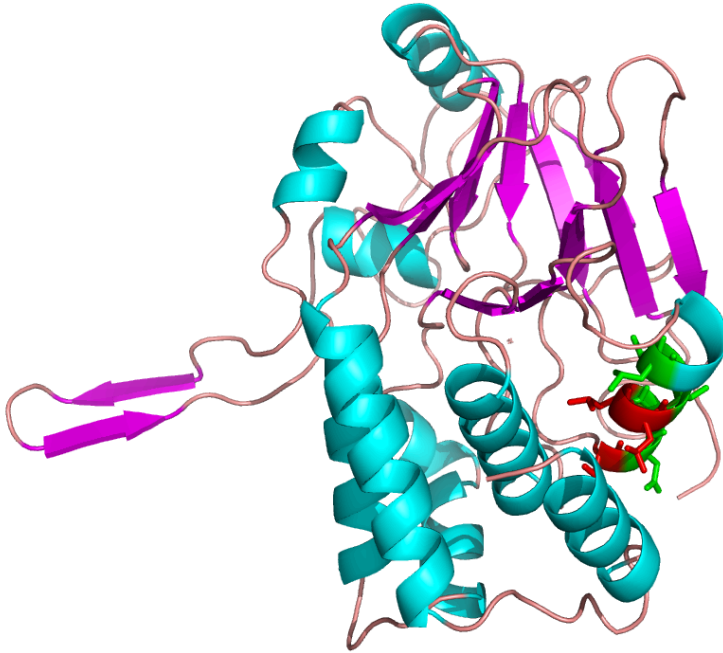
Structural Information: 40 % Sequence Identity to 4uqi.



9 P40421; RDGC

Location in protein: 163-LDDLLVVL-170

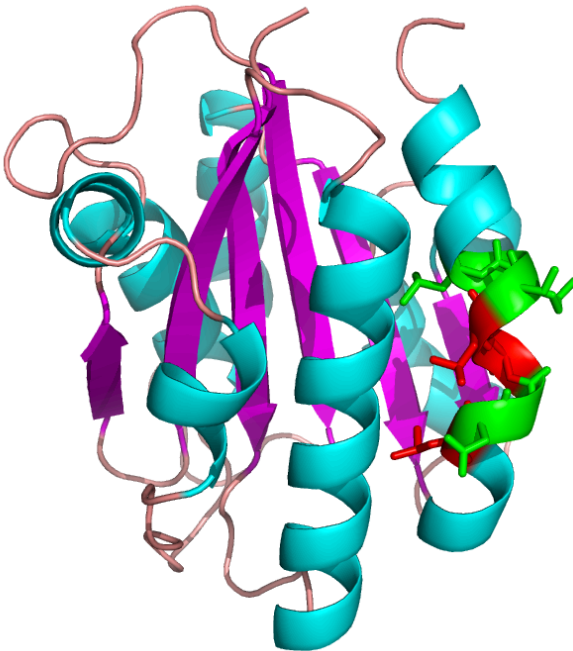
Structural Information: 40 % sequence identity with PDB 5jjaA; LD motif is inaccessible in the catalytic region



10 P38570; Integrin alpha-E; ITGAE

Location in protein: 375-LDGLLSKL-382

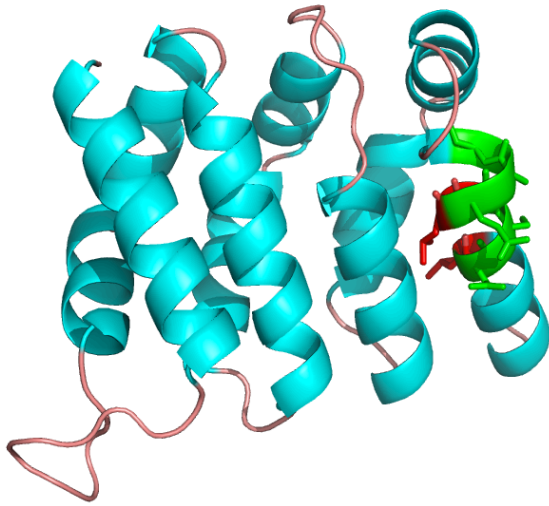
Structural Information: 38% sequence identity with PDB 1na4 in VWFA domain



11 P52306; RAP1 GTPase DISSOCIATION STIMULATOR 1; RAP1GDS

Location in protein: 27-LDCLLQAL-34

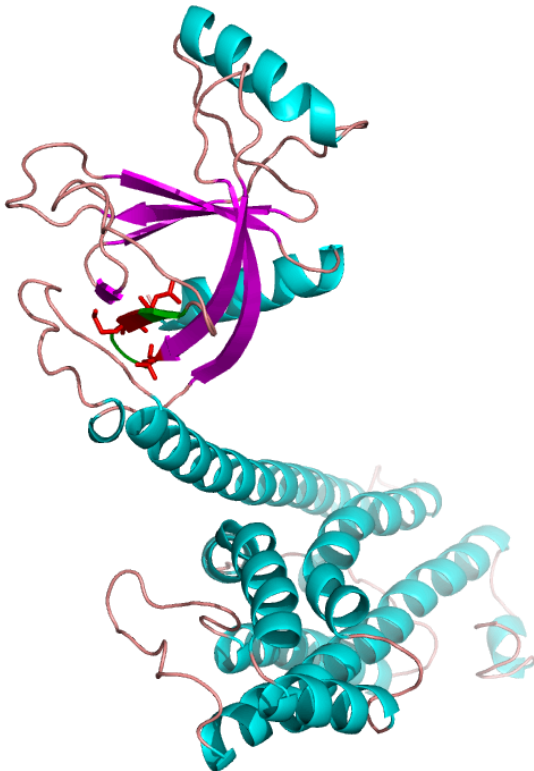
Structural Information: 24.2% sequence identity with PDB 4hxt in ARM repeat.



12 P53046; RHO1 GDP-GTP exchange protein 1; ROM1

Location in protein: 713-LDNMLLFL-720

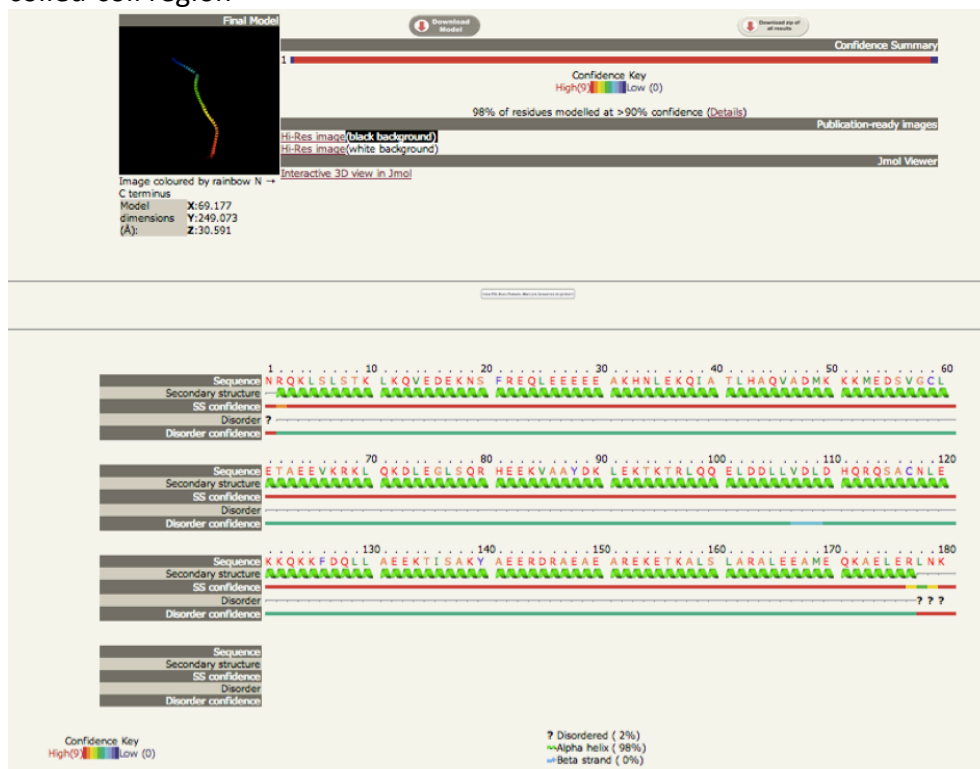
Structural Information: 17.7% identical with PDB 3kz1 (pictured) or 24% identical with PH domain only of 1xcgA;



13 P35579; MYOSIN-9; MYH9

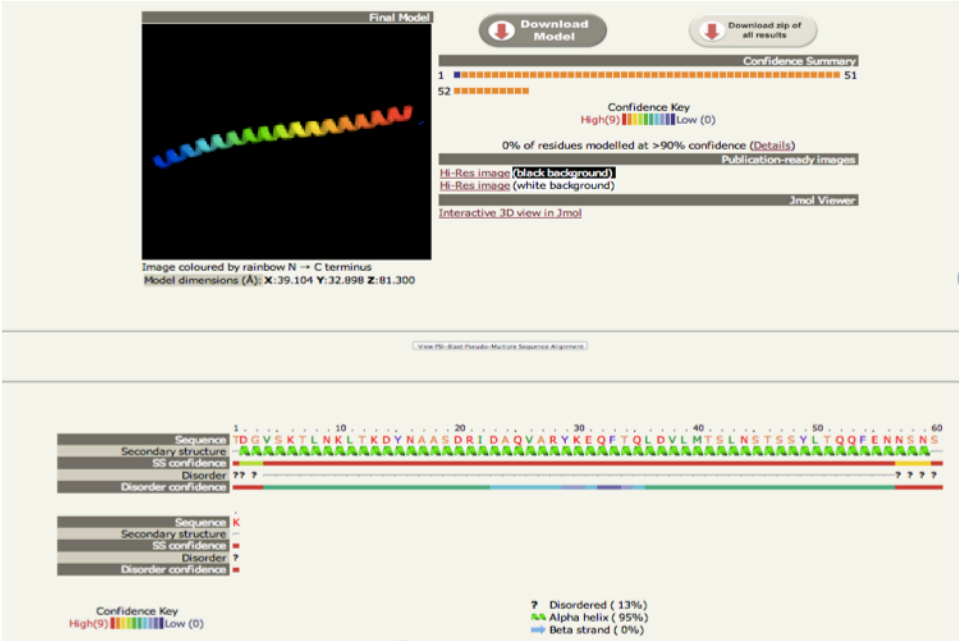
Location in protein: 1422-LDDLLVDL-1429

Structural Information: 17% sequence identity to PDB entry 2efr (tropomyosin) in C-terminal coiled-coil region



14 P24216; HAP2

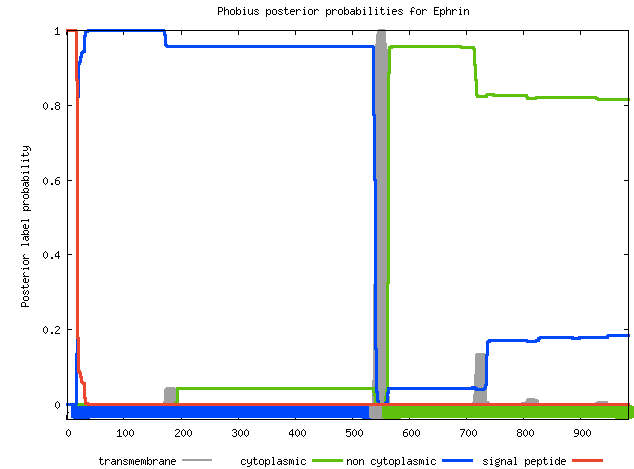
Location in protein: 443-LDVLMTS-450
Structural Information: 43 % identity with PDB 5krw for residues 385-461, but poor model quality. 13% sequence identify with PDB entry 1gk4 (88.8% confidence in phyre2 for the coiled-coil domain), similar to human vimentin coil 2b fragment



15 P54762; Ephrin type-B receptor 1;EphB1

Location in protein: 3-LDYLLLLL-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).

ID	Ephrin		
FT	SIGNAL	1	17
FT	REGION	1	1
FT	REGION	2	12
FT	REGION	13	17
FT	TOPO_DOM	18	540
FT	TRANSMEM	541	563
FT	TOPO_DOM	564	984
FT			
FT			



16 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

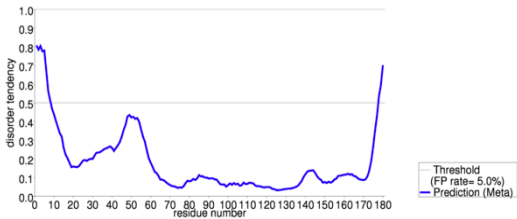
Prediction false positive rate: 5.0% Change FP rate

2-state prediction

(Red: Disordered residues Black: Ordered residues)

1	KTKPVTGNLR	PEEALQALTI	YEGKFGRLLD	DREKCAKAKE	ALELTDTGLL	50
51	SGSEERVQVA	LEELQDLKGV	WSELKVVWEQ	IDQMKEQPWW	SVQPRKLRQN	100
101	LDGLLNQLKN	FPARLRQYAS	YEFVQRLKLG	YMKINMLVIE	LKSEALKDRH	150
151	WKQLMKRLHV	NWVSELTLG	QIWDVLDQKN			200

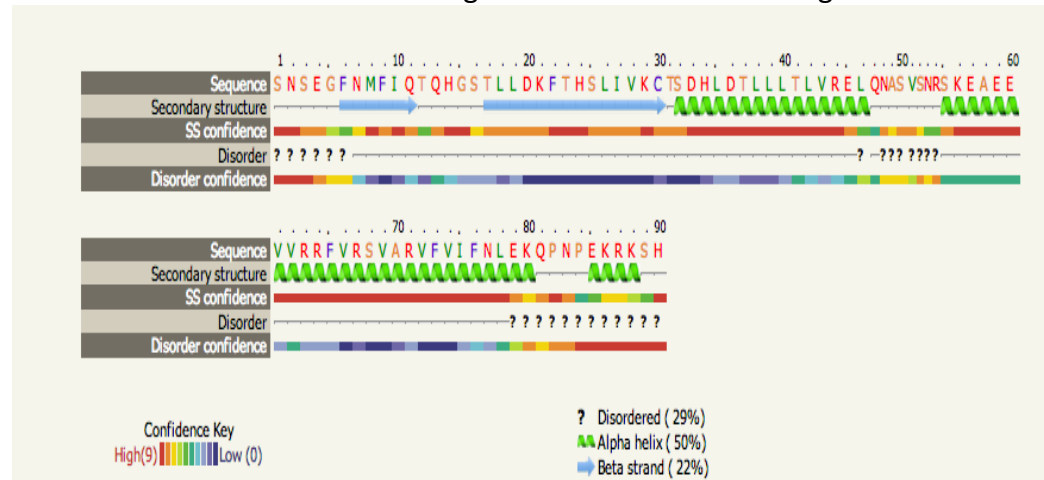
Disorder profile plot



17 P51592; E3 UBIQUITIN-PROTEIN LIGASE; HYD

Location in protein: 1453-LDTLLLT-1460

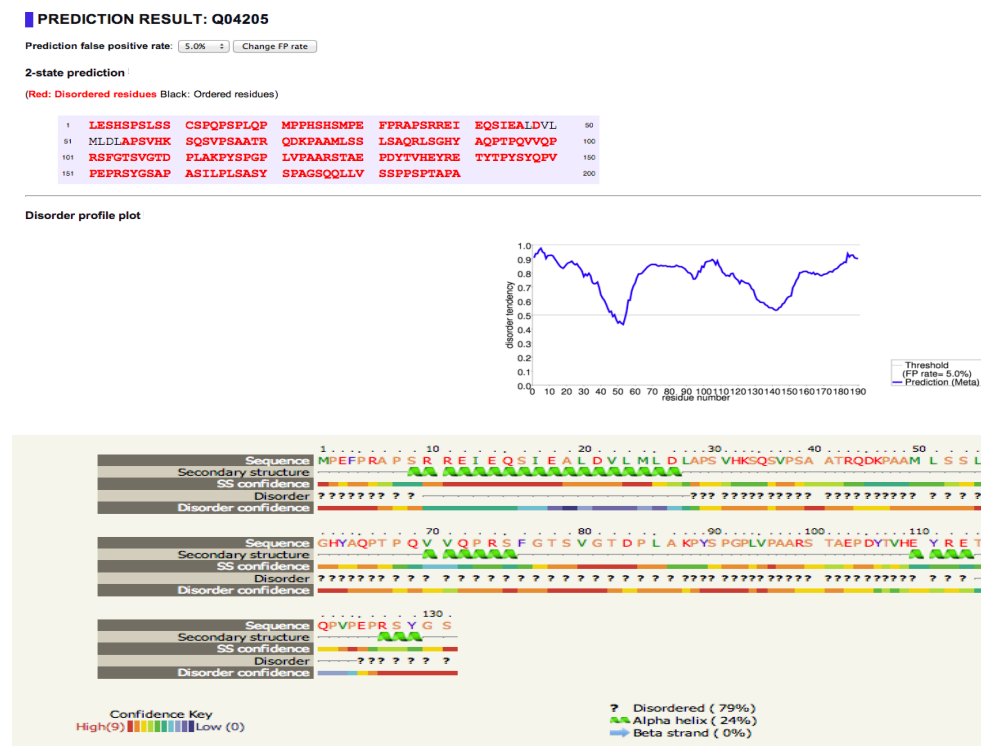
Structural Information: No homologous structure for modelling.



18 Q04205; TENSIN; TNS

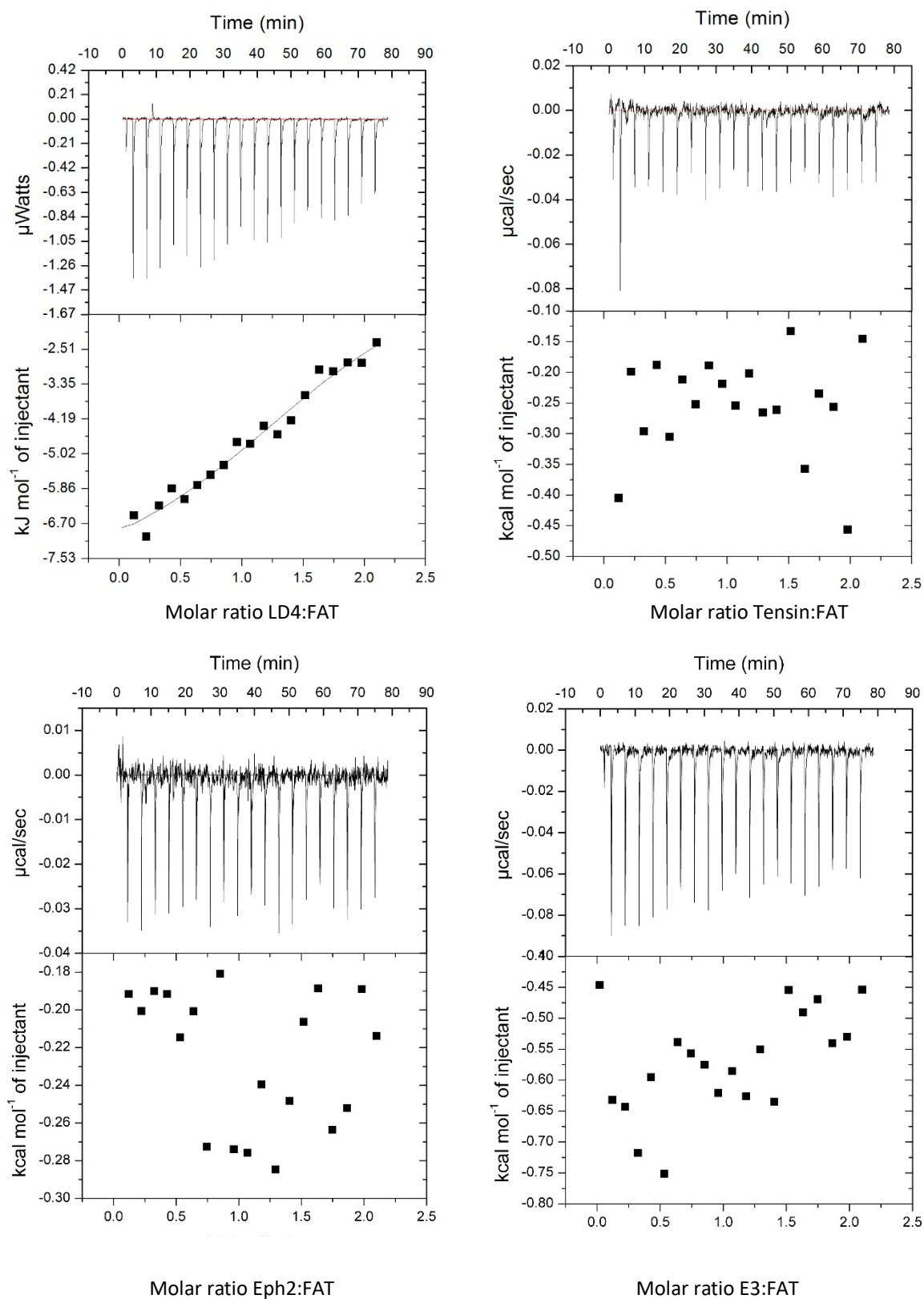
Location in protein: 807-LDVLMIDL-814

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



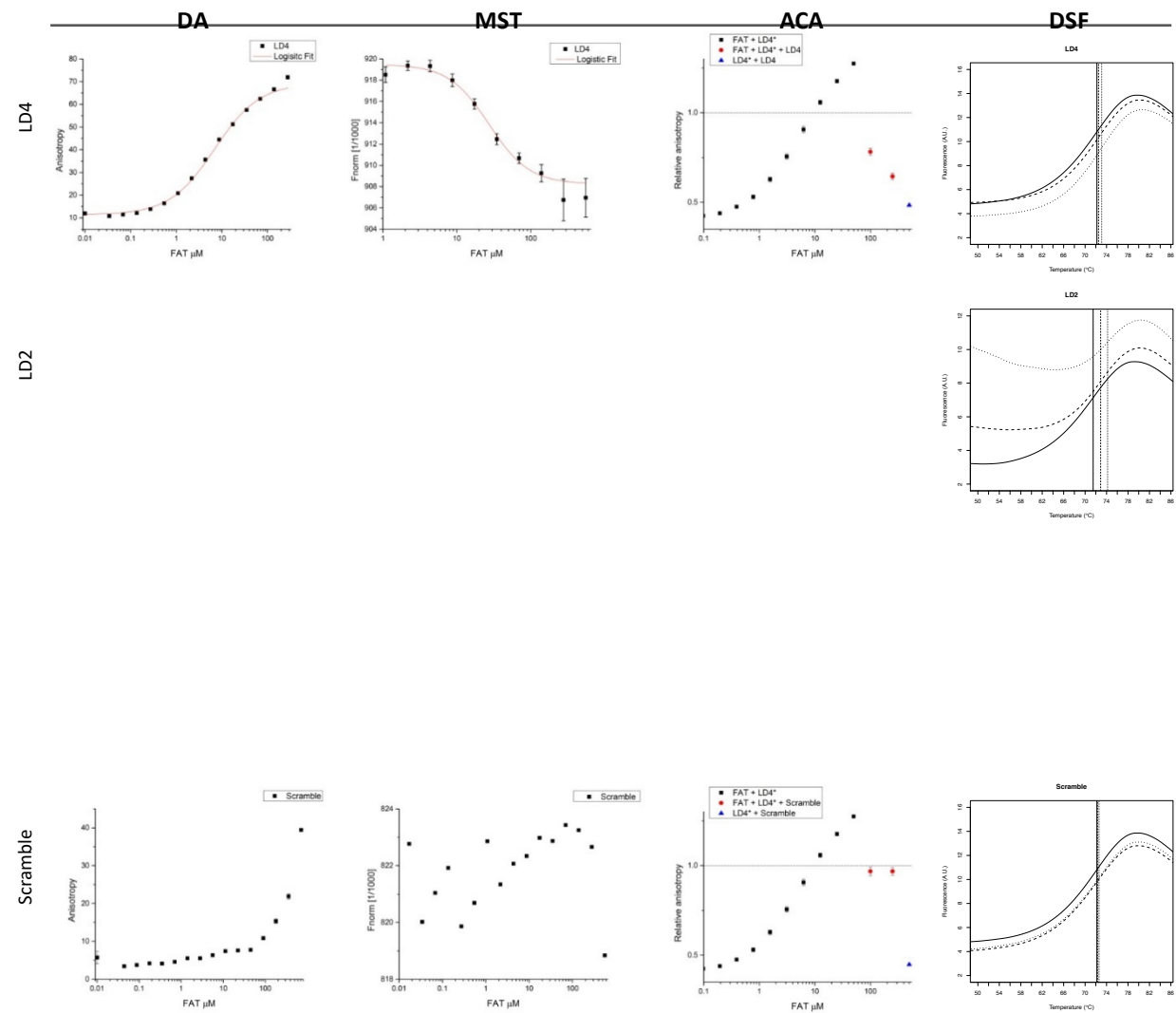
Supplementary Figure 3:

Isothermal calorimetric titration of FAT with LD4 and previously proposed LD motifs

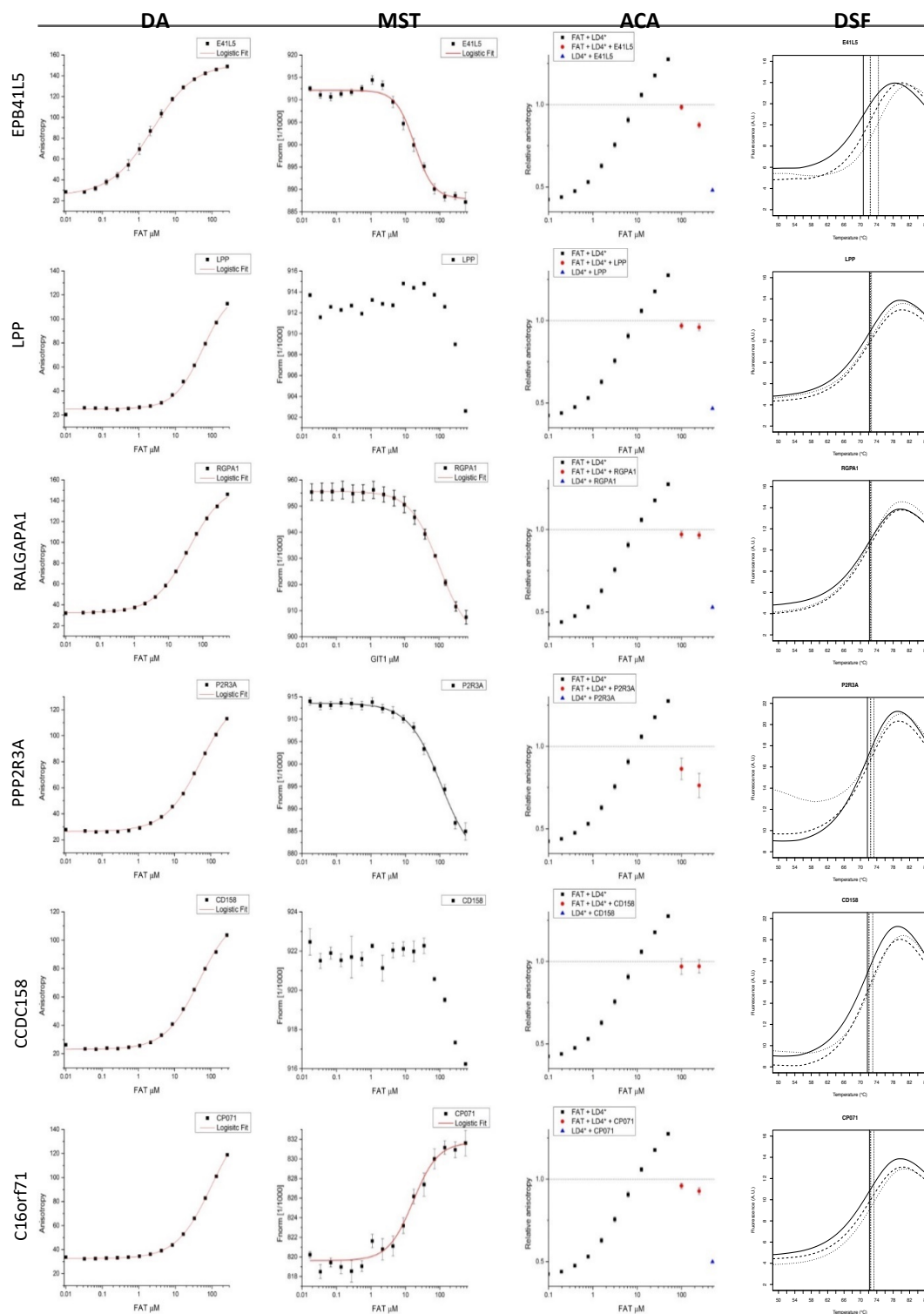


Supplementary Figure 4

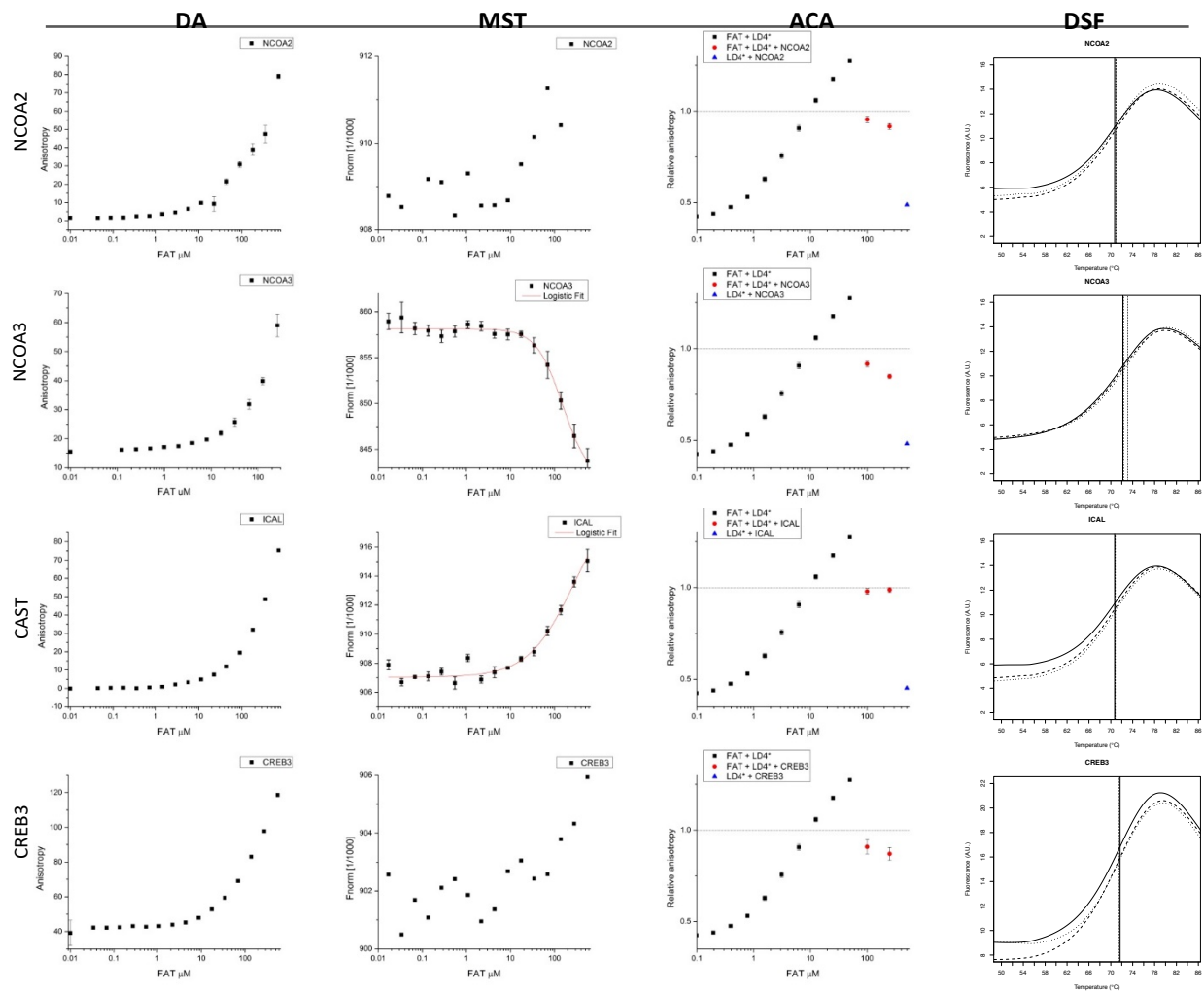
Binding assays of known LD motifs and LD motifs proposed by LDMF-proposed to FAT, α -parvin and GIT1. ACA: anisotropy competition assay; DA: direct fluorescence anisotropy; MST: microscale thermophoresis; DSF: differential scanning fluorimetry.



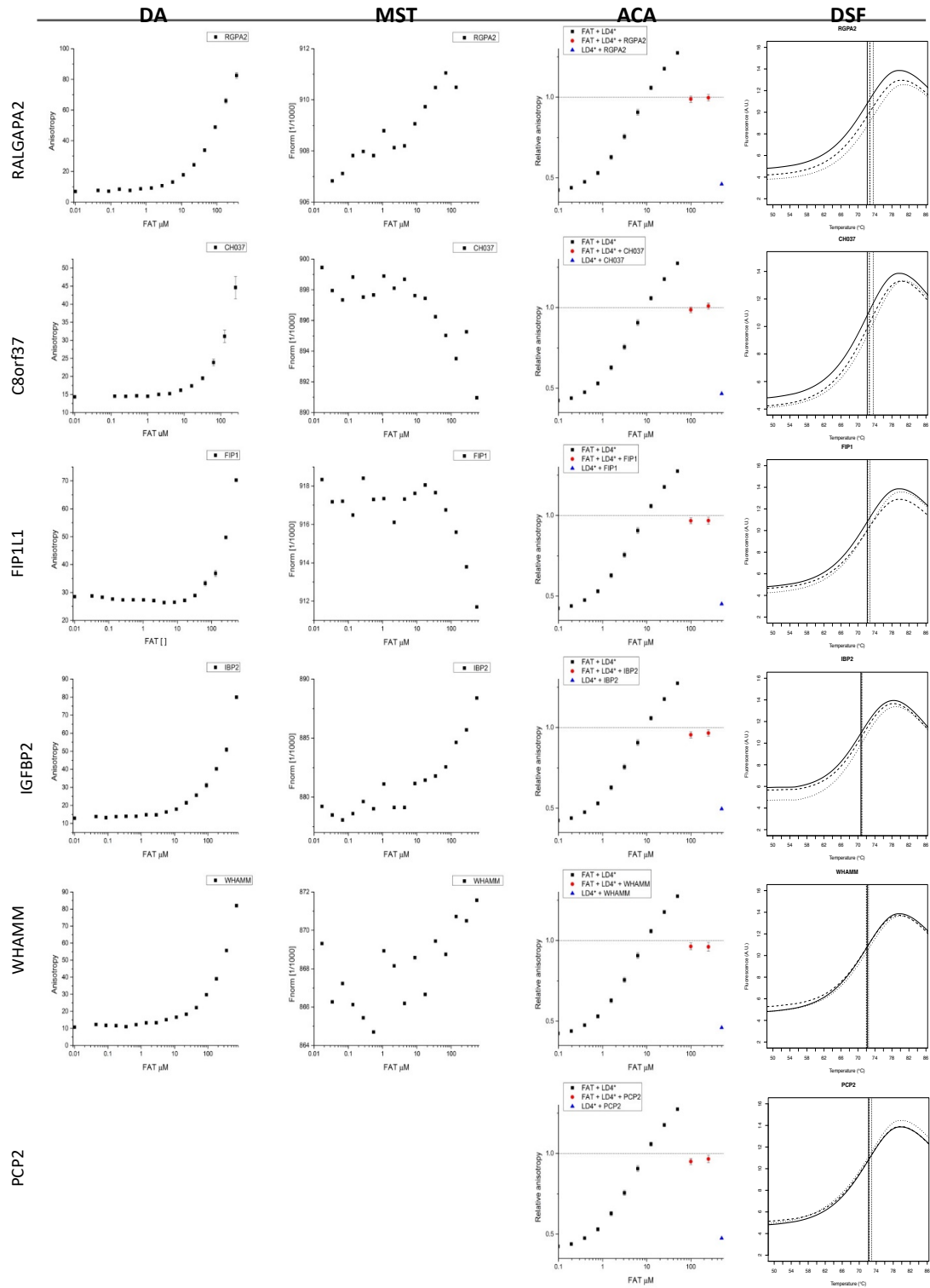
Supplementary Figure 4.1: Binding of LD motif controls to FAT



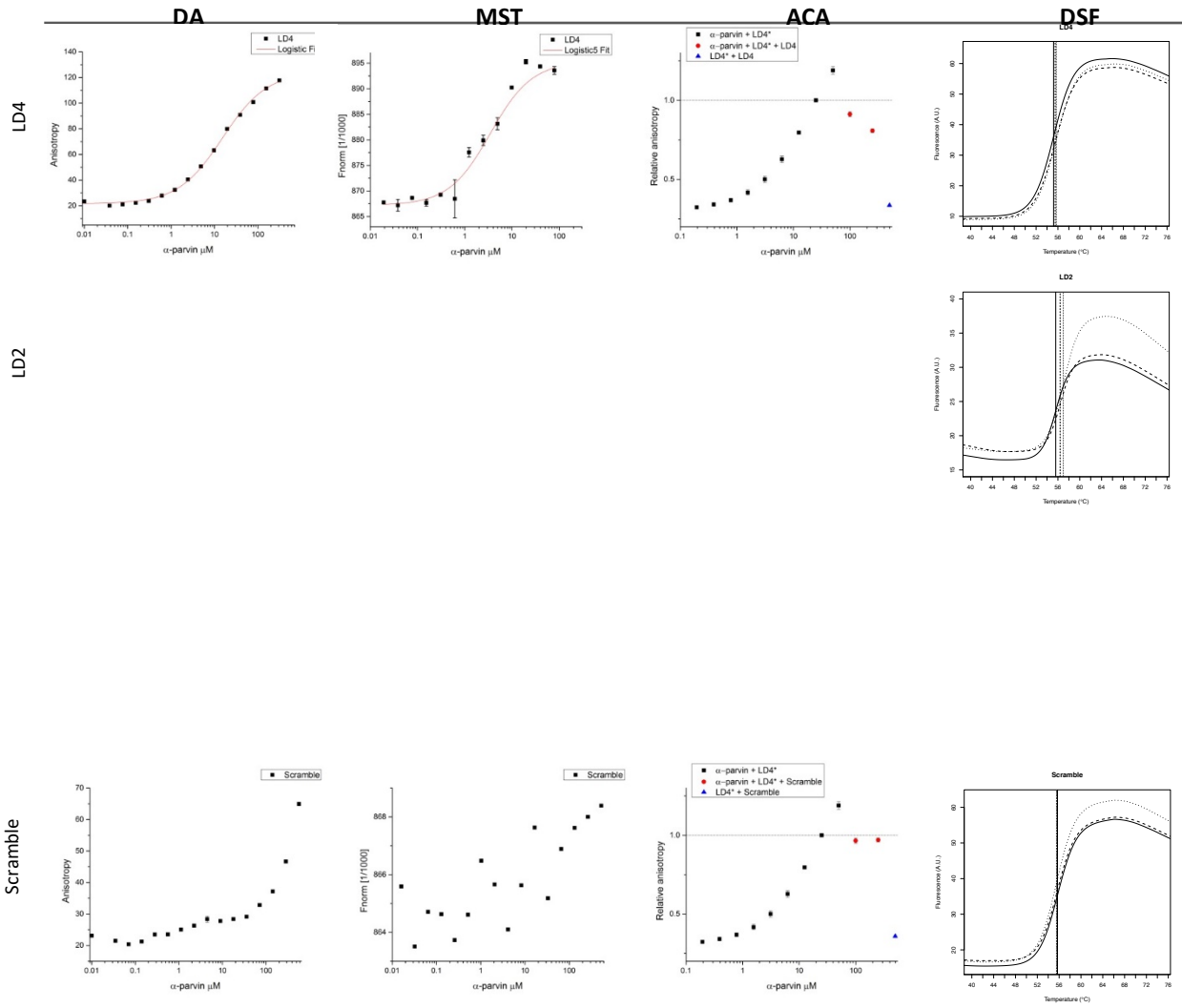
Supplementary Figure 4.2: Binding of highly likely LD motifs to FAT



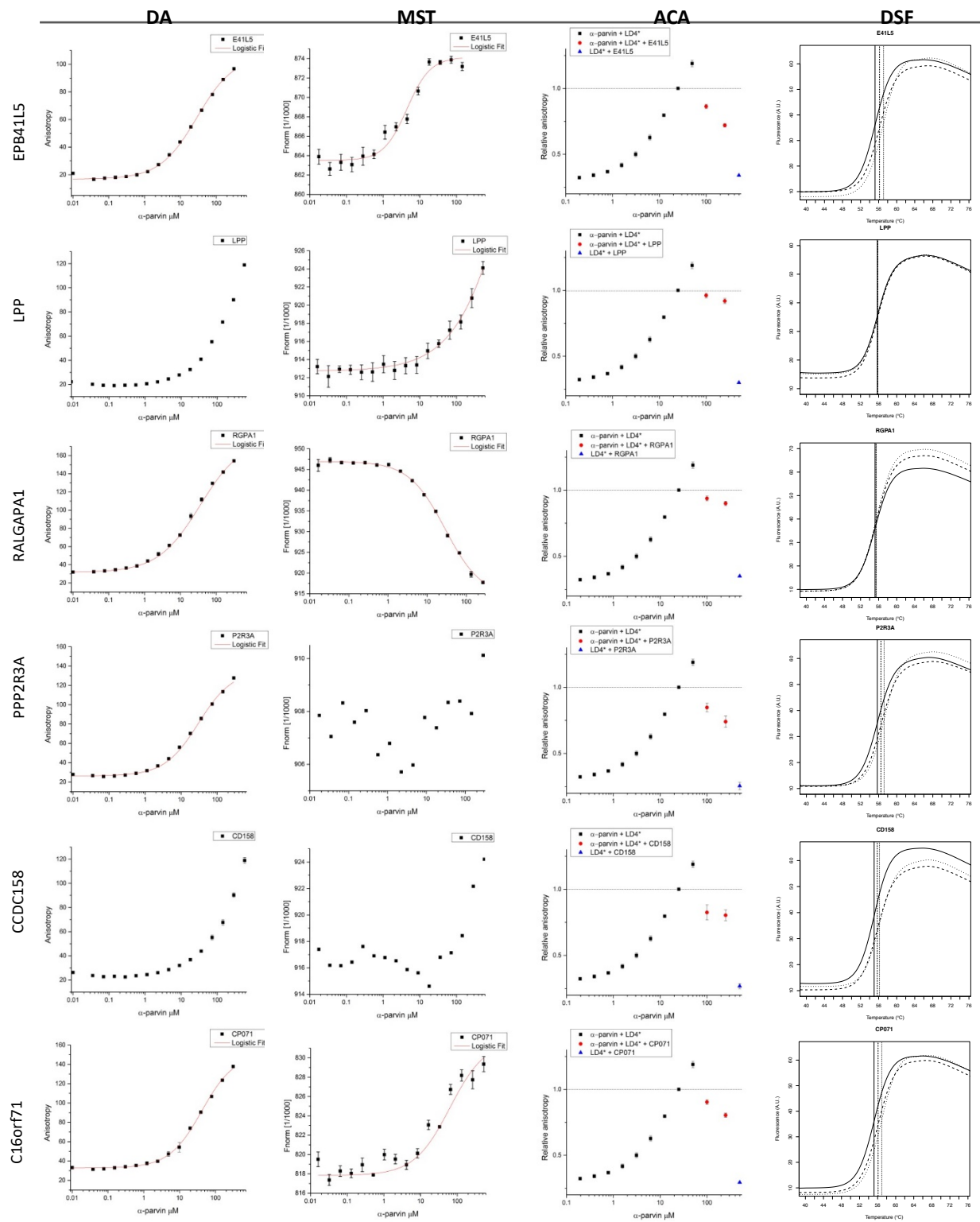
Supplementary Figure 4.3: Binding of less likely LD motifs to FAT



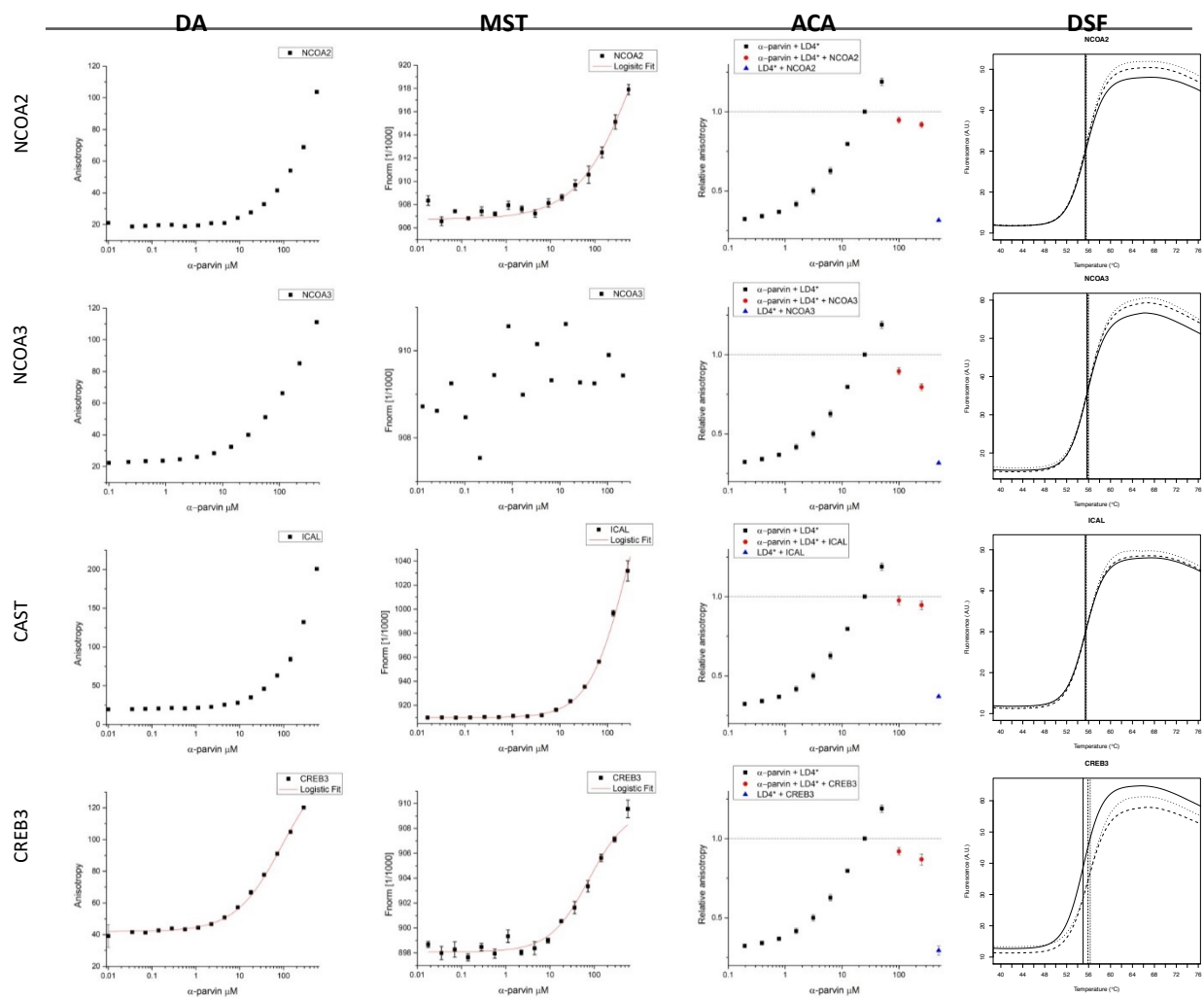
Supplementary Figure 4.4: Binding of least likely LD motifs to FAT and motifs discarded in round 1.



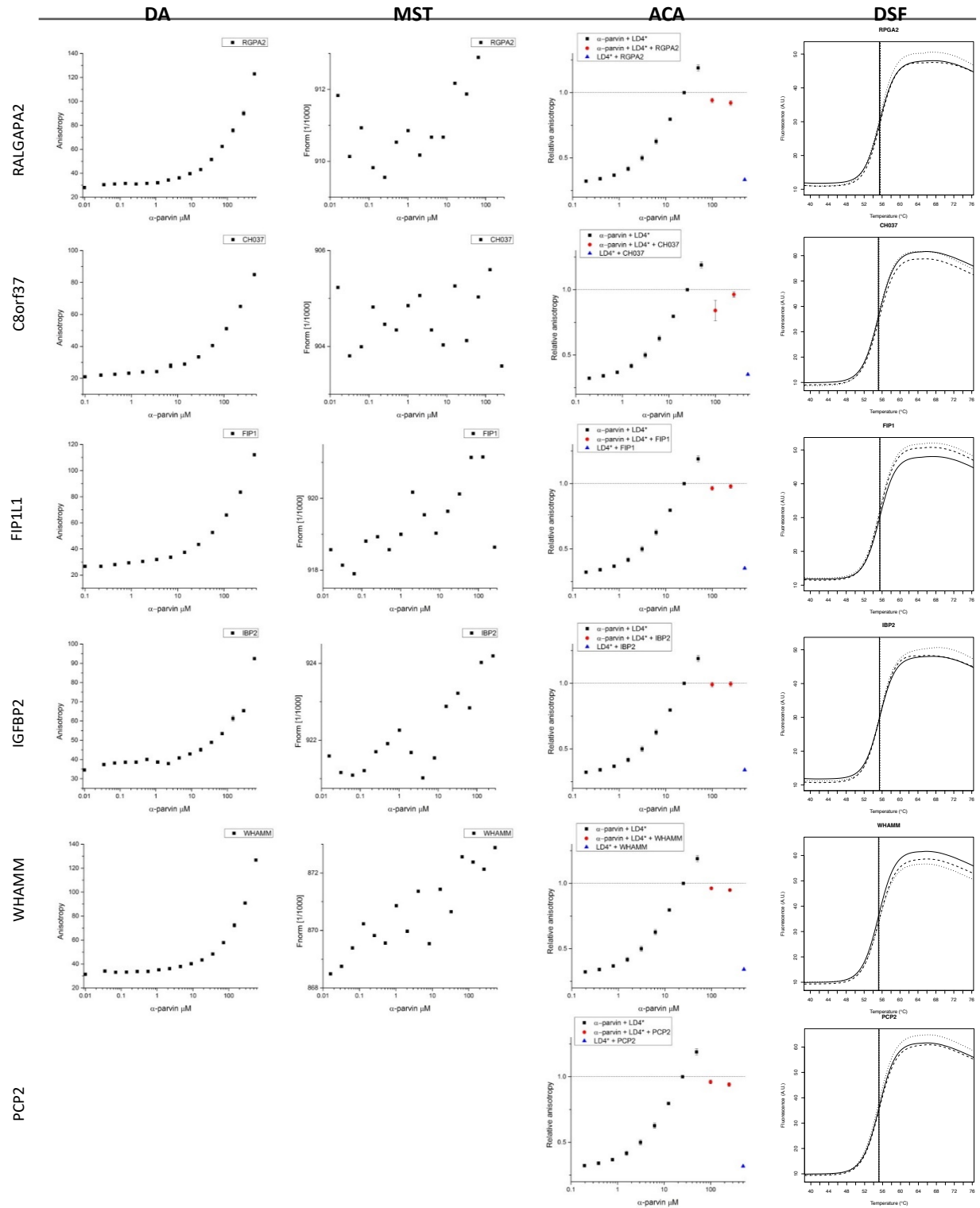
Supplementary Figure 4.5: Binding of LD motif controls to α -parvin



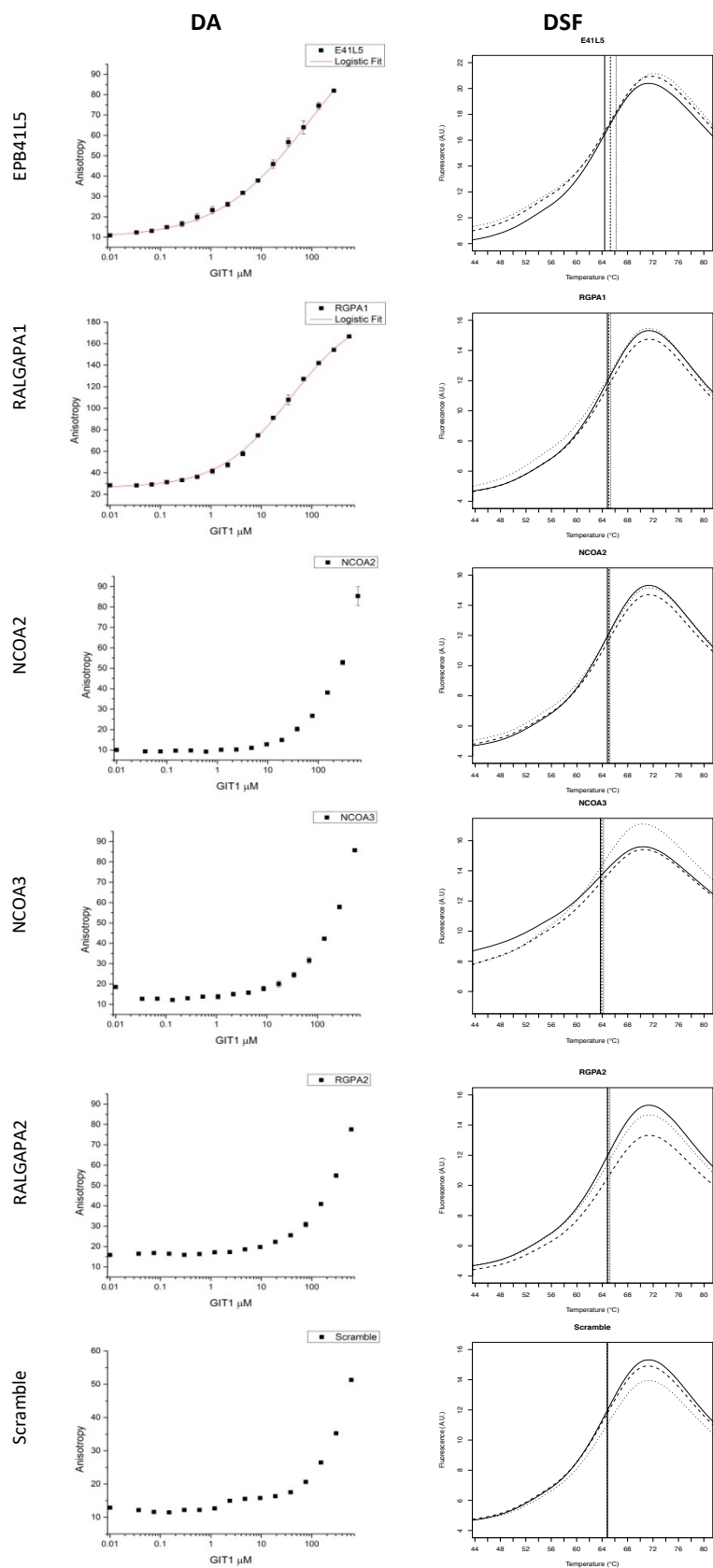
Supplementary Figure 4.6: Binding of highly likely LD motifs to α -parvin



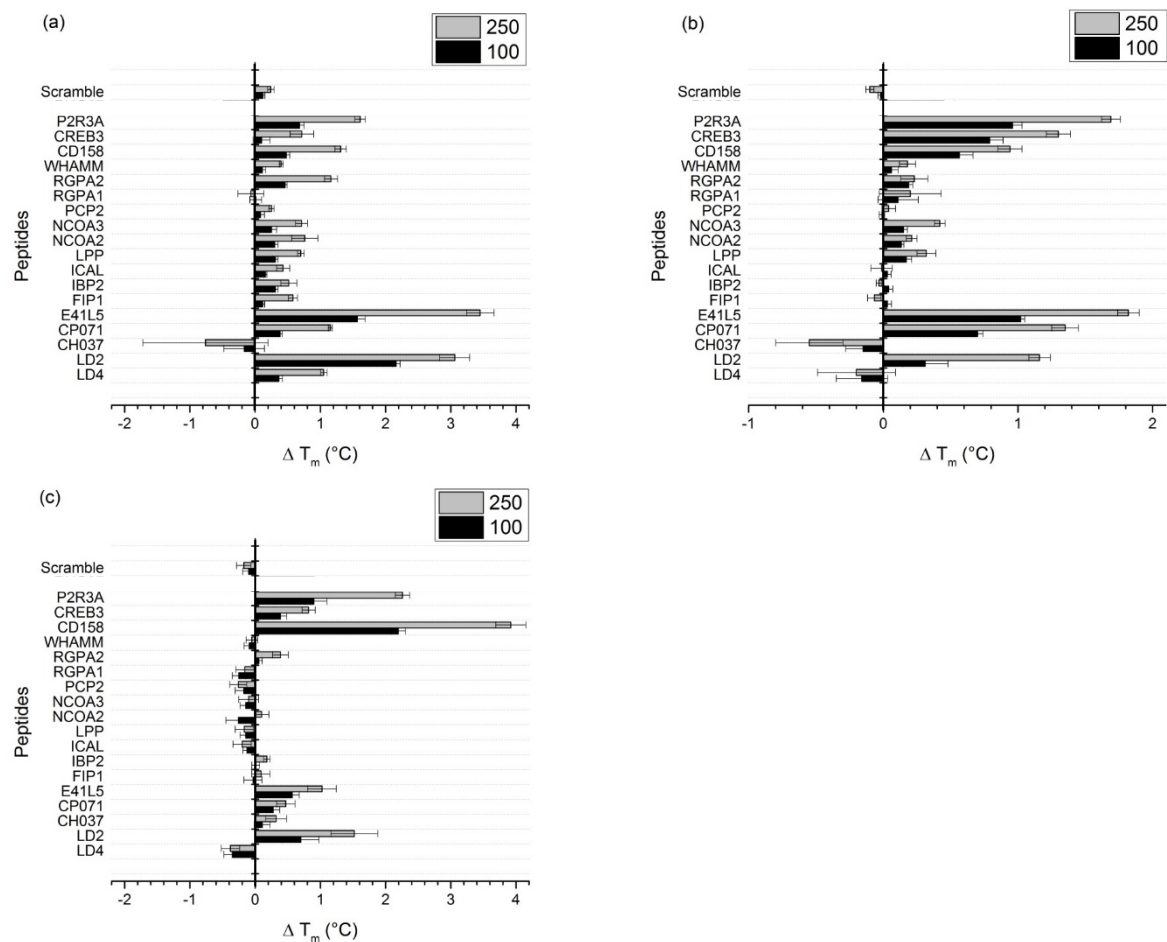
Supplementary Figure 4.7: Binding of less likely LD motifs to α -parvin



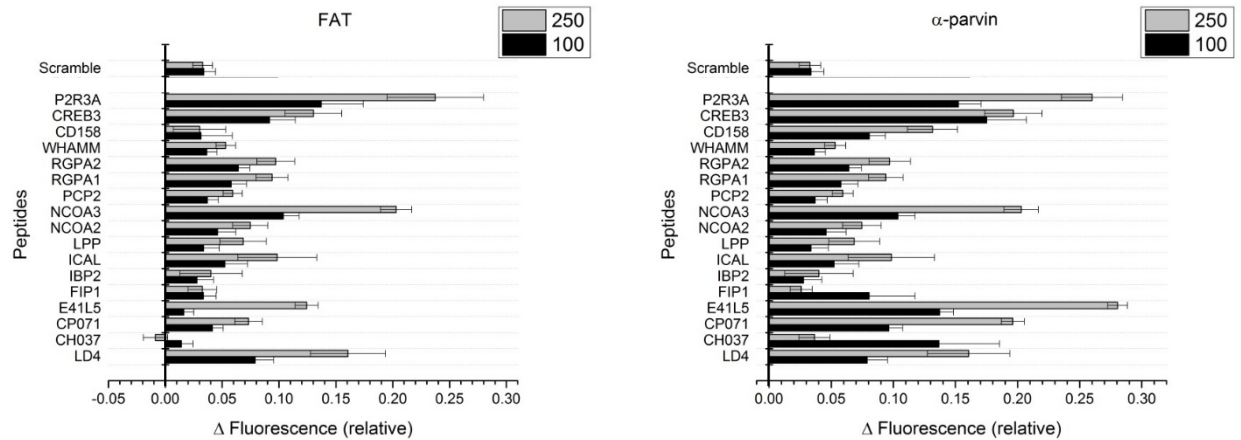
Supplementary Figure 4.8: Binding of least likely LD motifs to α -parvin and motifs discarded in round 1



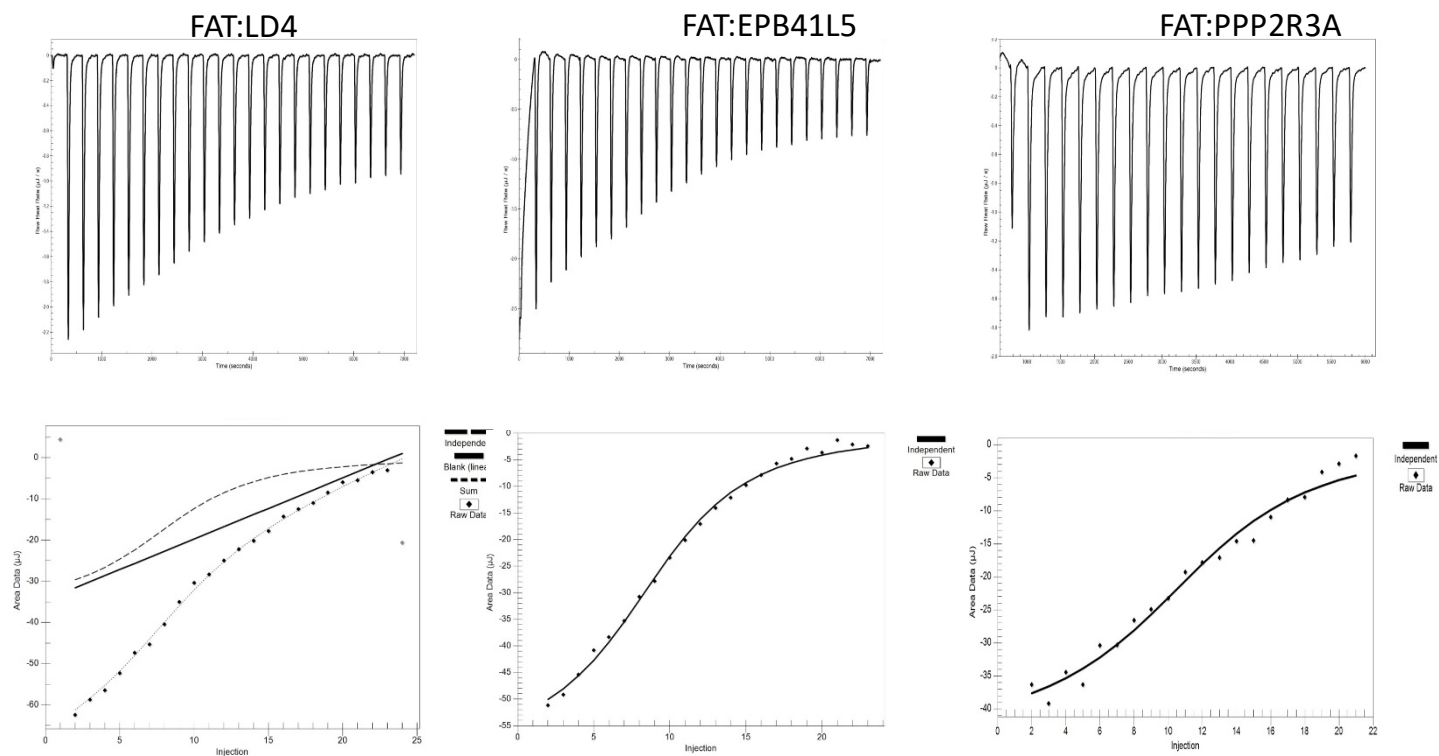
Supplementary Figure 4.9: Binding of LD motif candidates to GIT1.



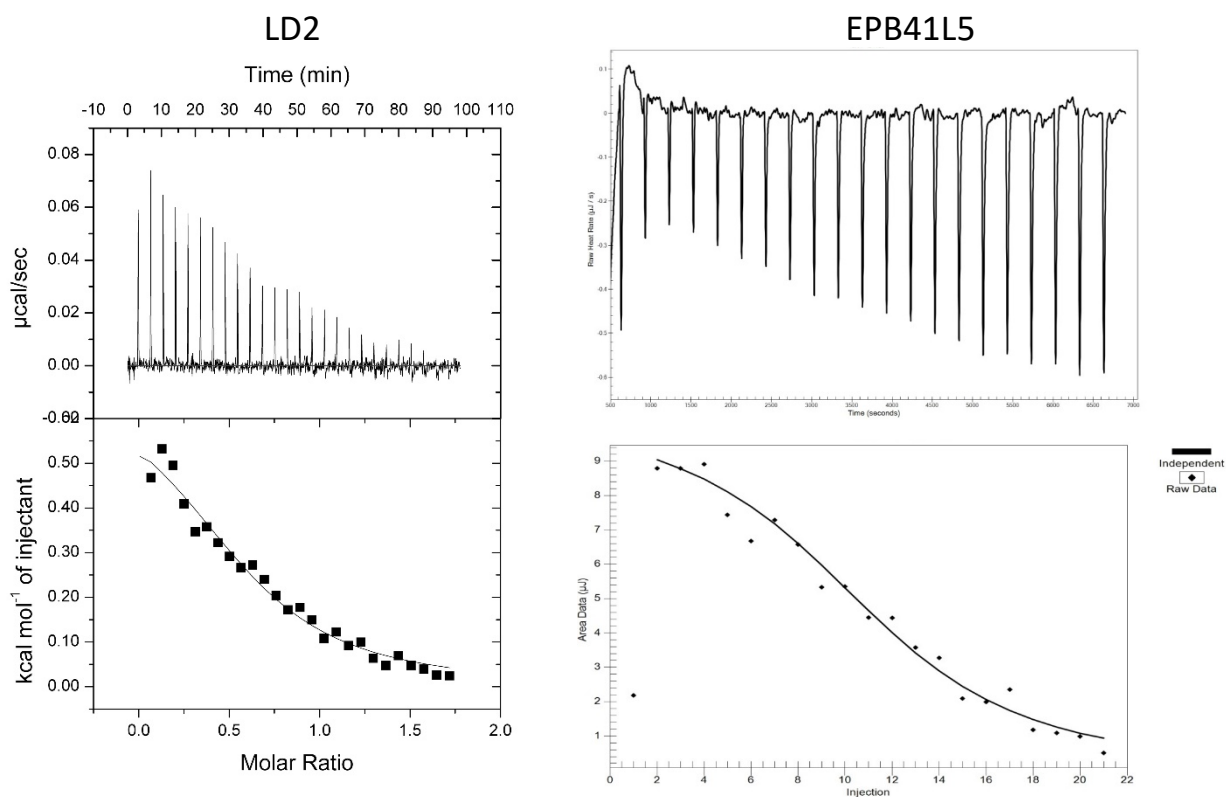
Supplementary Figure 4.10: T_m shift in °C for differential scanning fluorimetry for peptides with (a) FAT (b) α -parvin and (c) GIT1. The Uniprot identifiers are given instead of protein gene names. Genes were both identifiers differ are P2R3A: PPP2R3A; CD158:CCDC158; RGPA1/2: RALGAPA1/2; ICAL: CAST; IBP2: IGFBP2; FIP1: FIP1L1; E41L5: EPB41L5; CP071: C16orf71; CP037: C8orf37.



Supplementary Figure 4.11: 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 μ M for FAT and 25 μ M for α -parvin), in the presence of 0.1 μ M labeled LD4. To that, each non-labeled LD motif candidate peptide was added at 100 or 250 μ M. 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; RGA1/2: RALGAP1/2; ICAL: CAST; IBP2: IGFBP2; FIP1: FIP1L1; E41L5: EPB41L5; CP071: C16orf71; CP037: C8orf37.



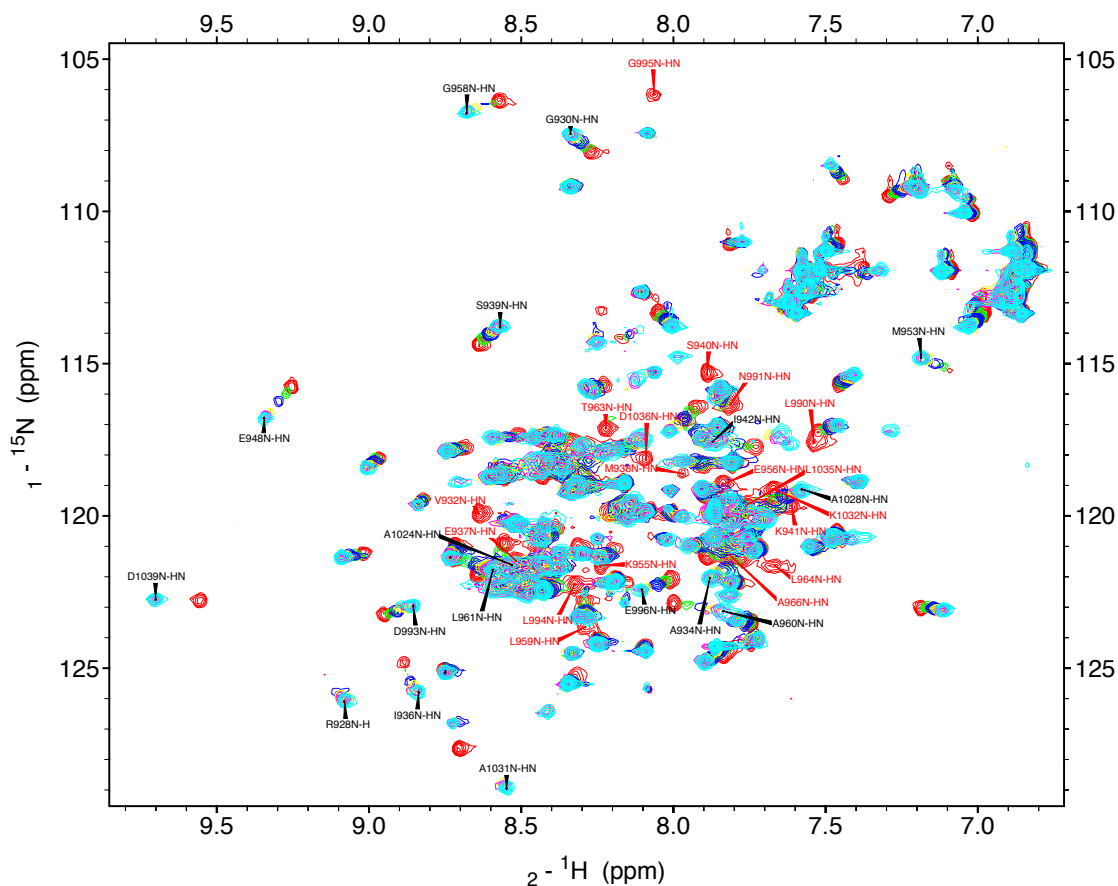
Supplementary Figure 4.12: Titration of FAT on to LD motifs



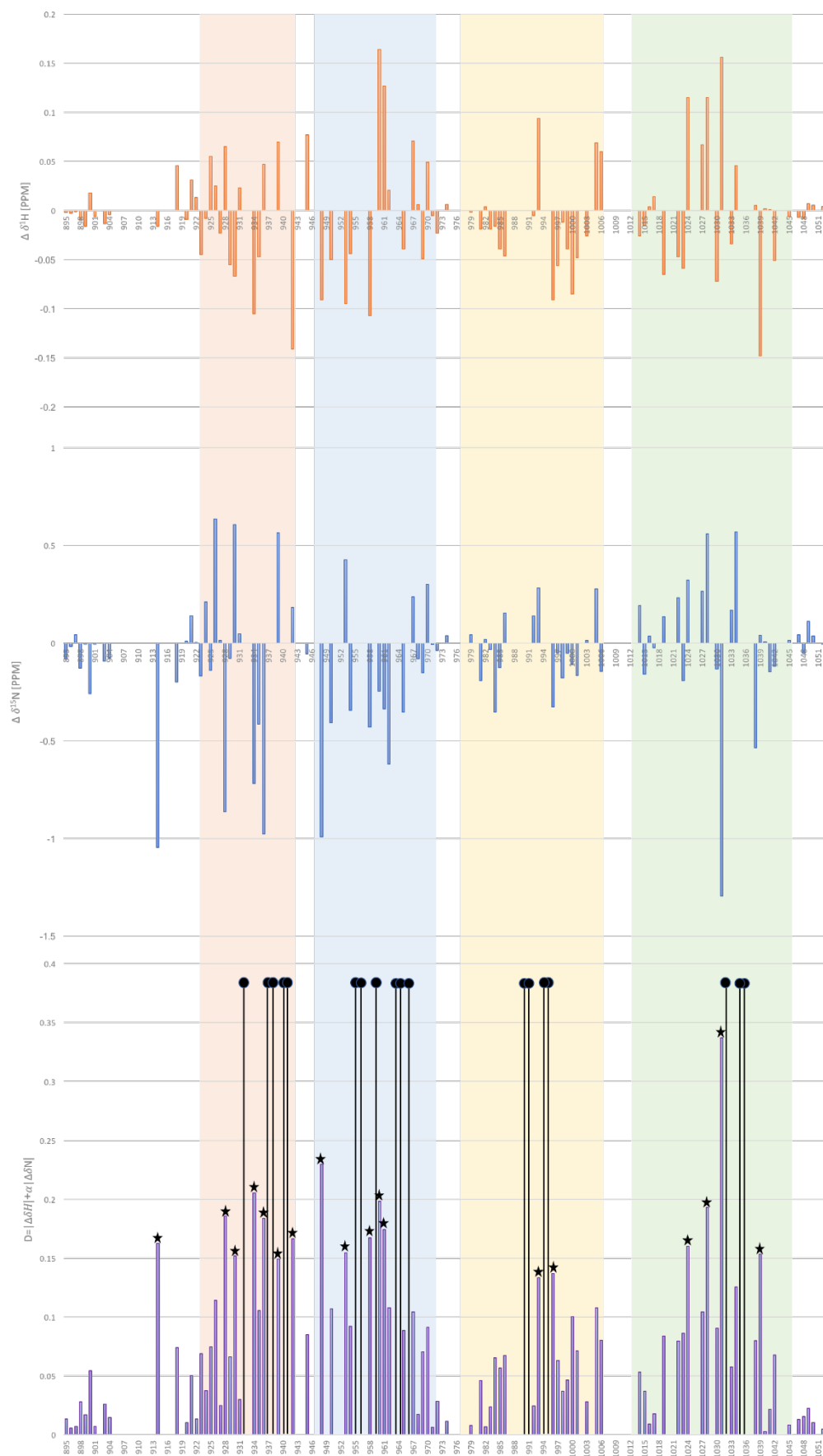
Supplementary Figure 4.13: Titration of GIT1 on to LD2 and EPB41L5

Supplementary Figure 5:

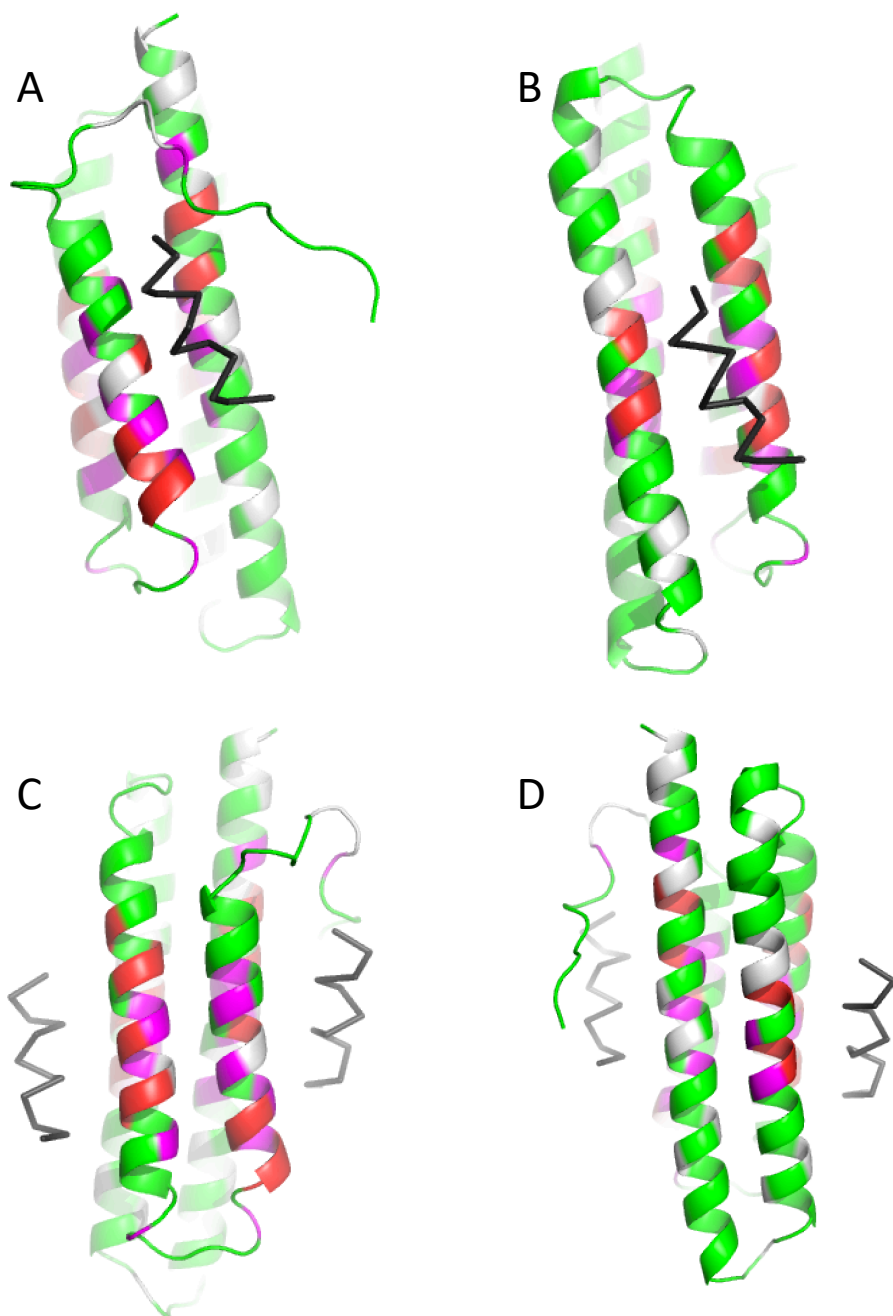
^1H - ^{15}N HSQC titration experiments. Shown are the NMR chemical shifts of ^{15}N -labelled FAT domain titrated with LD4, LD2, DLC1, LPP, and CD158.



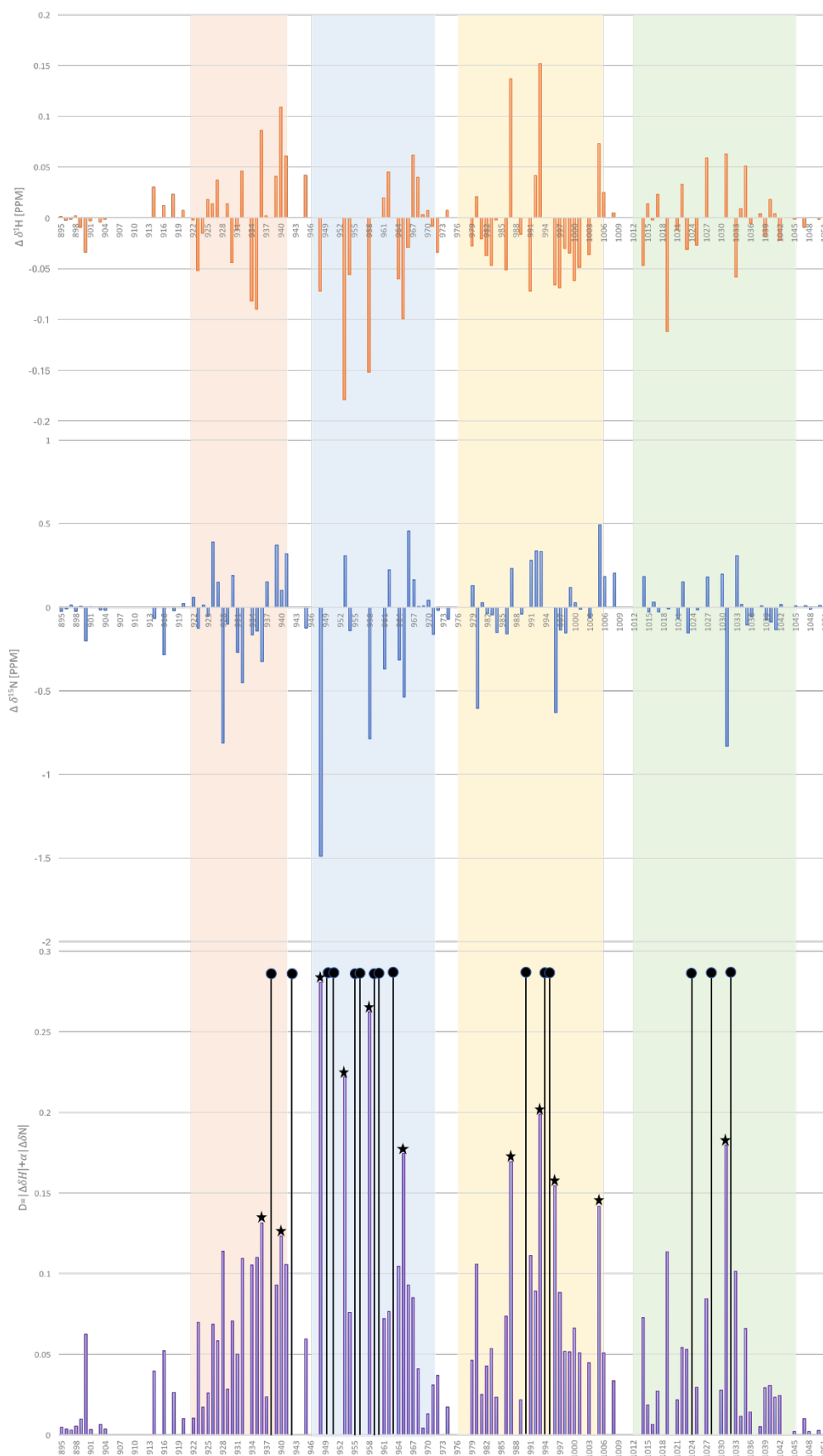
Supplementary Figure 5.1: FAT/LD2 Titration. Overlay of ^1H - ^{15}N -HSQC spectra of 100 μM ^{15}N -FAT in the absence (red) and presence of 0.5 (green), 1 (blue), 2 (yellow), 3 (magenta) and 4 (cyan) times molar excess of LD2 peptide. Resonances that disappeared upon LD2 addition are labelled in red. Resonances that significantly shifted $>2\sigma = 0.13$ are labelled in black. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



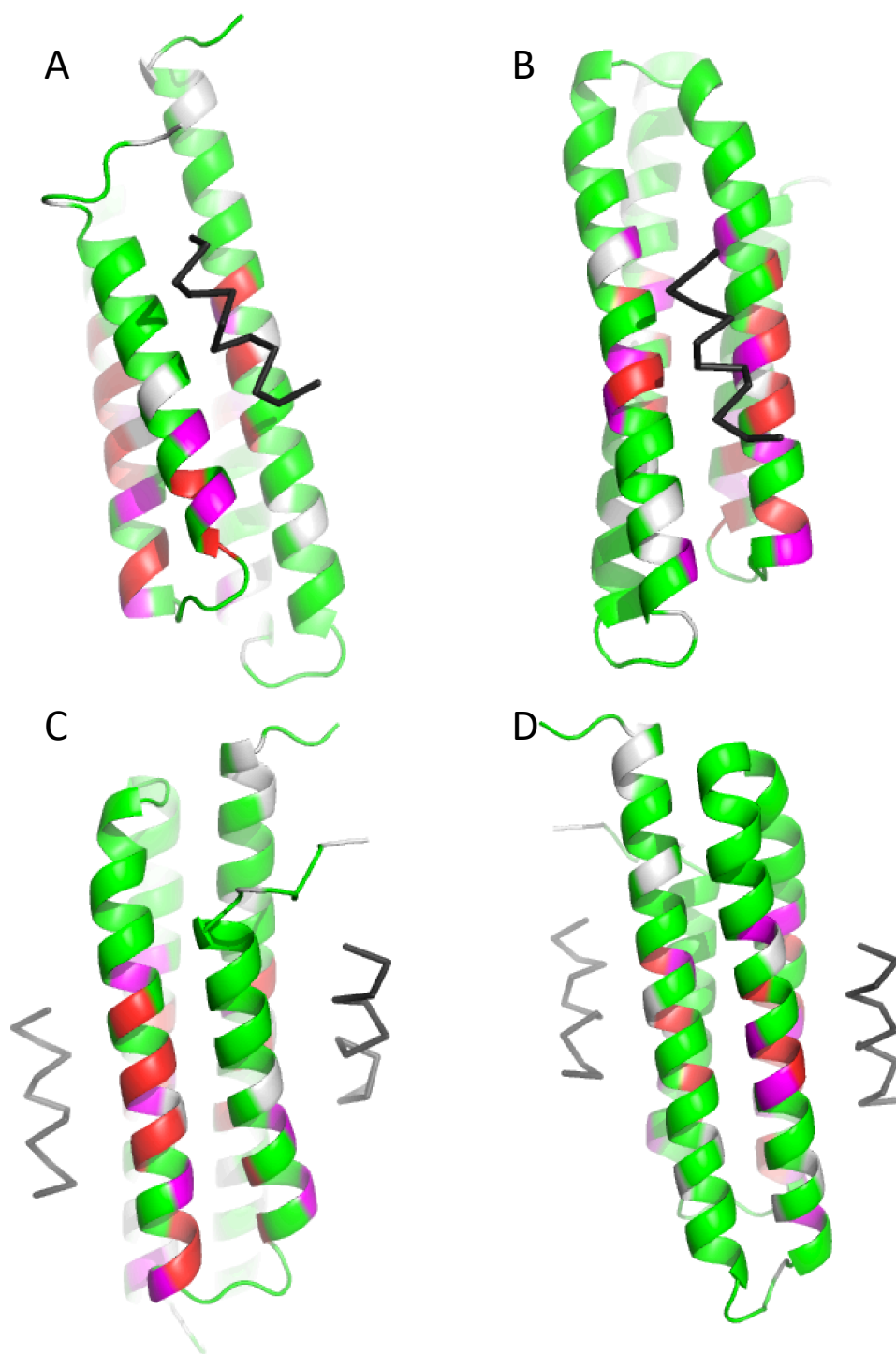
Supplementary Figure 5.2: Chemical shift changes in FAT induced by the LD2 peptide. Chemical shift differences in ppm were calculated for ^1H (top panel), ^{15}N (middle panel) and the weighted combined $^1\text{H},^{15}\text{N}$ (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of LD2 peptide. Resonances showing significant change (greater than $2\sigma = 0.13$) are marked by stars. 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).



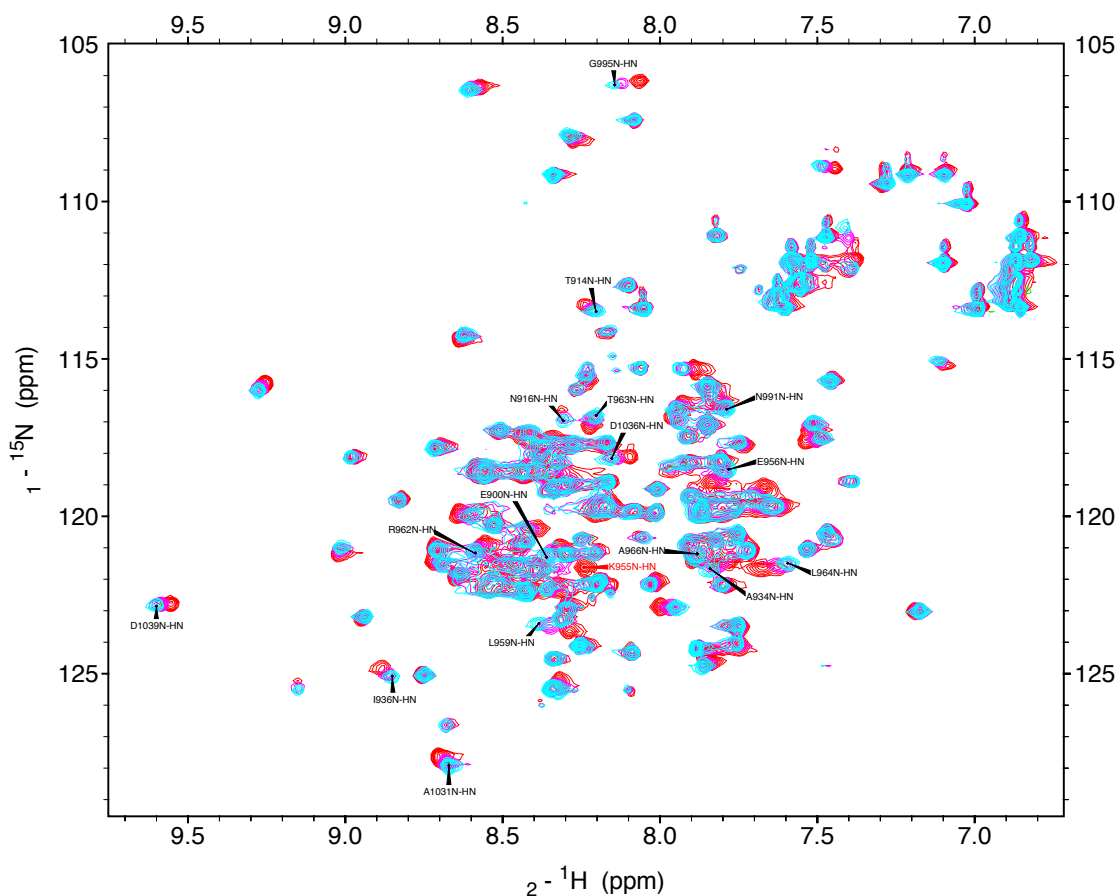
Supplementary Figure 5.3: Mapping changes of NMR resonances on FAT structure upon titration with LD2 where A is the face of FAT helices 1 and 4, B is 2-3 helices, C is 1-2 helices and D is 3-4 helices. Resonances showing significant changes are labeled in magenta. LD2 is represented as a ribbon. Others that disappeared are labeled in red. Gray residues represent amino acids whose peaks are not assigned. This figure was prepared using the PDB file 1OW8 (Paxillin LD2 motif bound to FAT of FAK)



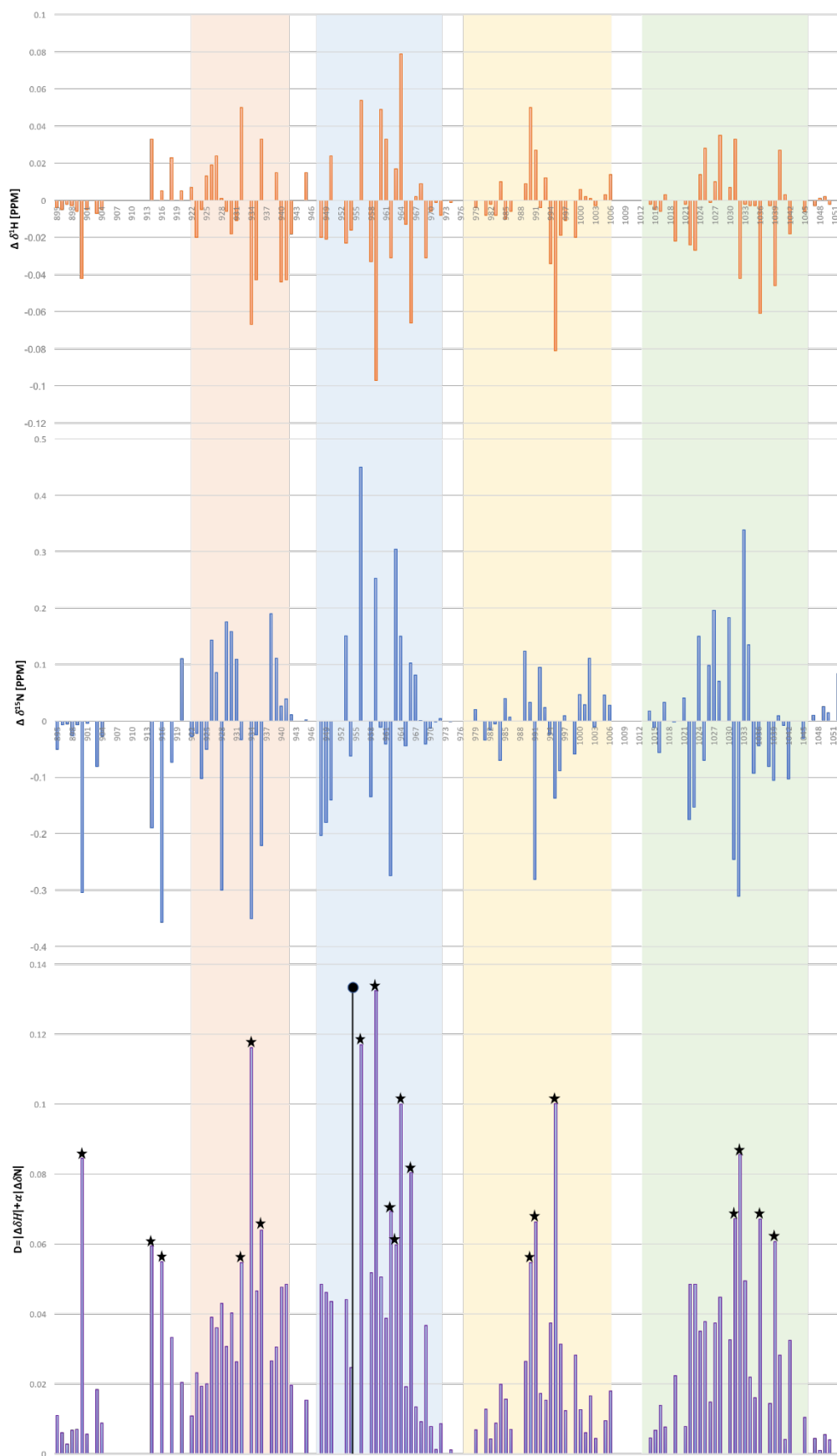
Supplementary Figure 5.5: Chemical shift changes in FAT induced by LD4 peptide. Chemical shift differences in ppm where calculated for ^1H (top panel), ^{15}N (middle panel) and the weighted combined $^1\text{H},^{15}\text{N}$ (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of LD4 peptide. Resonances showing significant change (greater than $2\sigma=0.12$) are marked by stars. 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).



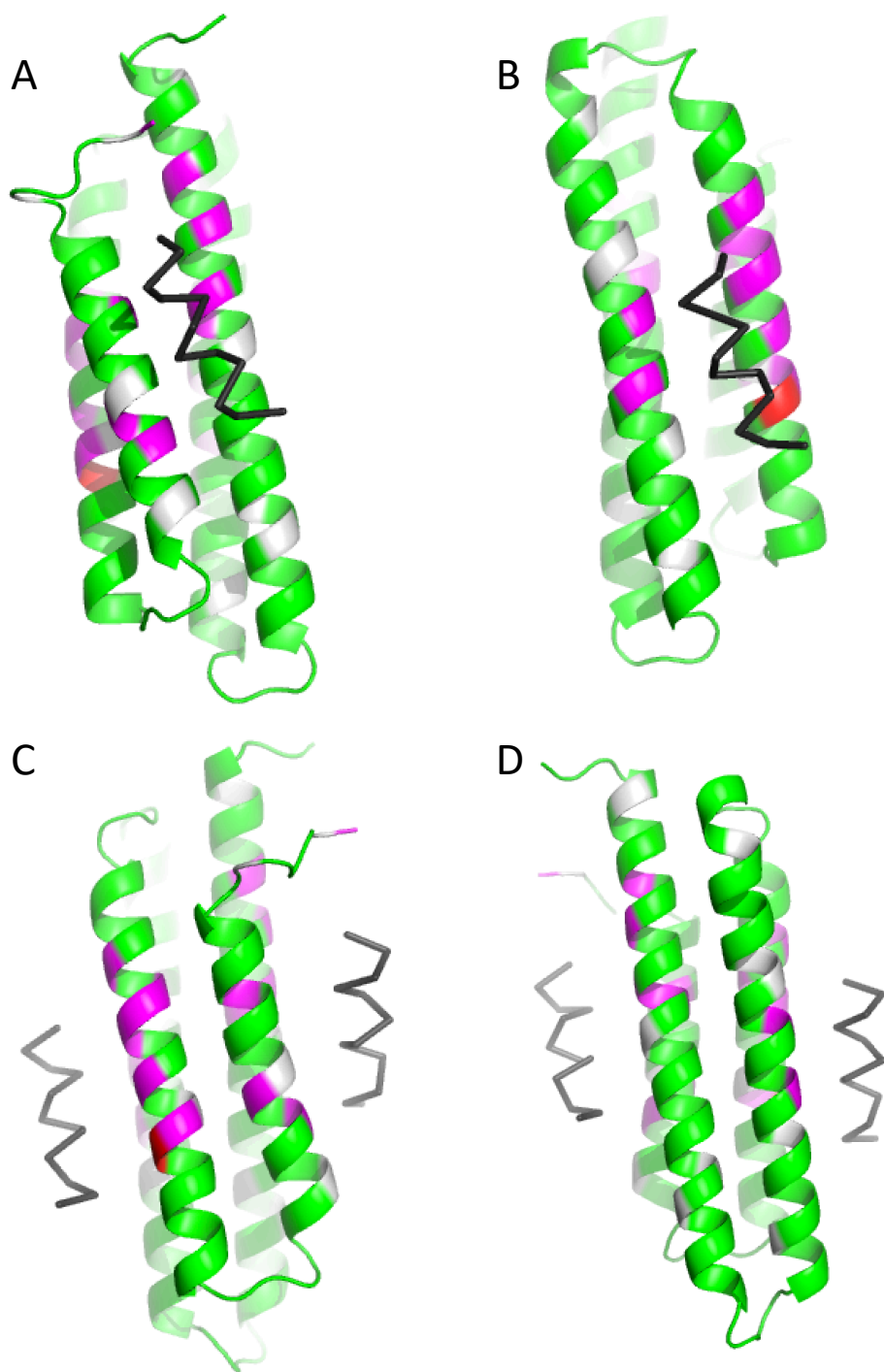
Supplementary Figure 5.6: Mapping changes of NMR resonances on FAT structure upon titration with LD4 where A is the face of FAT helices 1 and 4, B is 2-3 helices, C is 1-2 helices and D is 3-4 helices. Resonances showing significant changes are labeled in magenta. LD4 is represented as a ribbon. Others that disappeared are labeled in red. Gray residues represent amino acids whose peaks are not assigned. This figure was prepared using the PDB file 1OW7 (Paxillin LD4 motif bound to FAT of FAK).



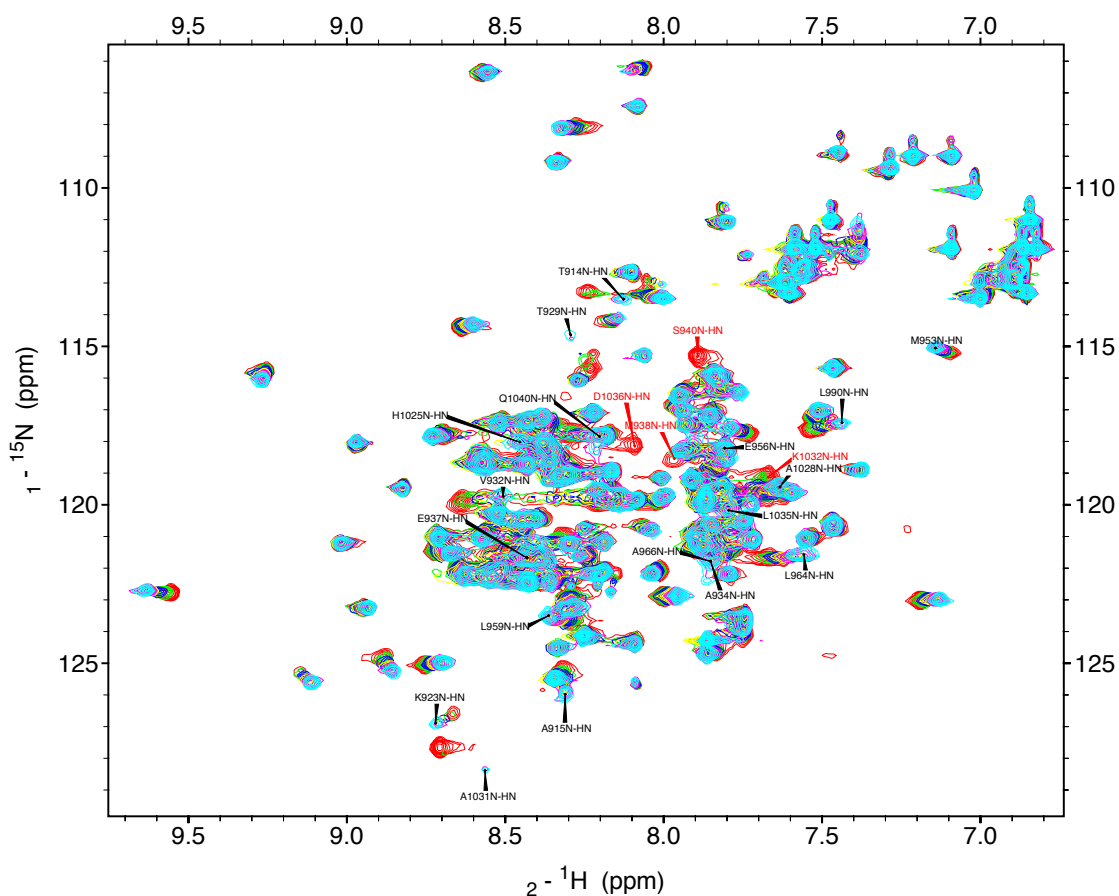
Supplementary Figure 5.7: FAT/LPP titration. Overlay of ^1H - ^{15}N -HSQC spectra of 100 μM ^{15}N -FAT in the absence (red) and presence of 3 (magenta) and 5 (cyan) times molar excess of LPP peptide. Resonances that disappeared upon LPP addition are labelled in red. Resonances that significantly shifted $>2\sigma = 0.055$ are labelled in black. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



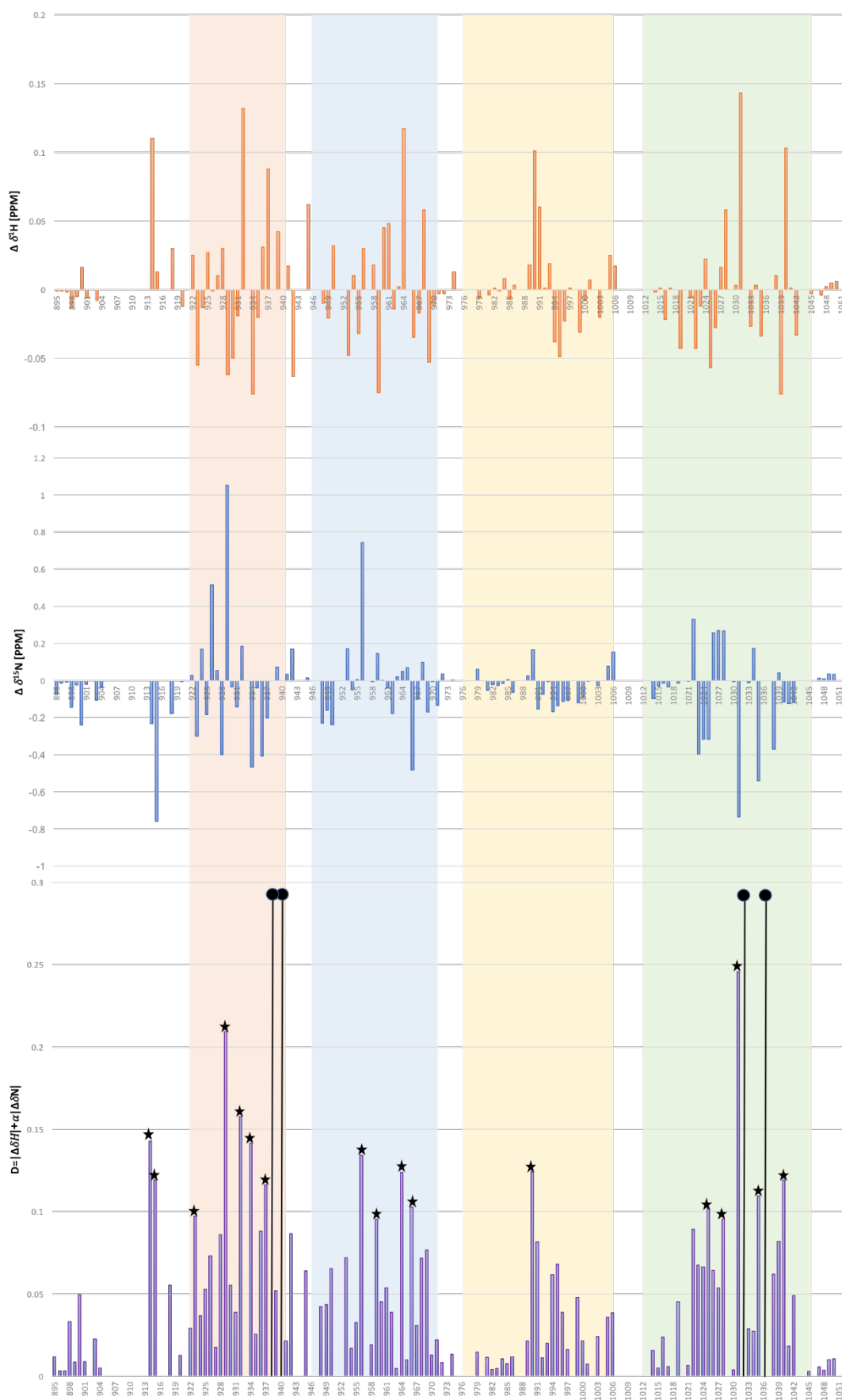
Supplementary Figure 5.8: Chemical shift changes in FAT induced by LPP peptide. Chemical shift differences in ppm were calculated for ^1H (top panel), ^{15}N (middle panel) and the weighted combined $^1\text{H},^{15}\text{N}$ (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of LPP peptide. Resonances showing significant change (greater than $2\sigma=0.055$) are marked by stars. Others that disappeared upon ID2 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).



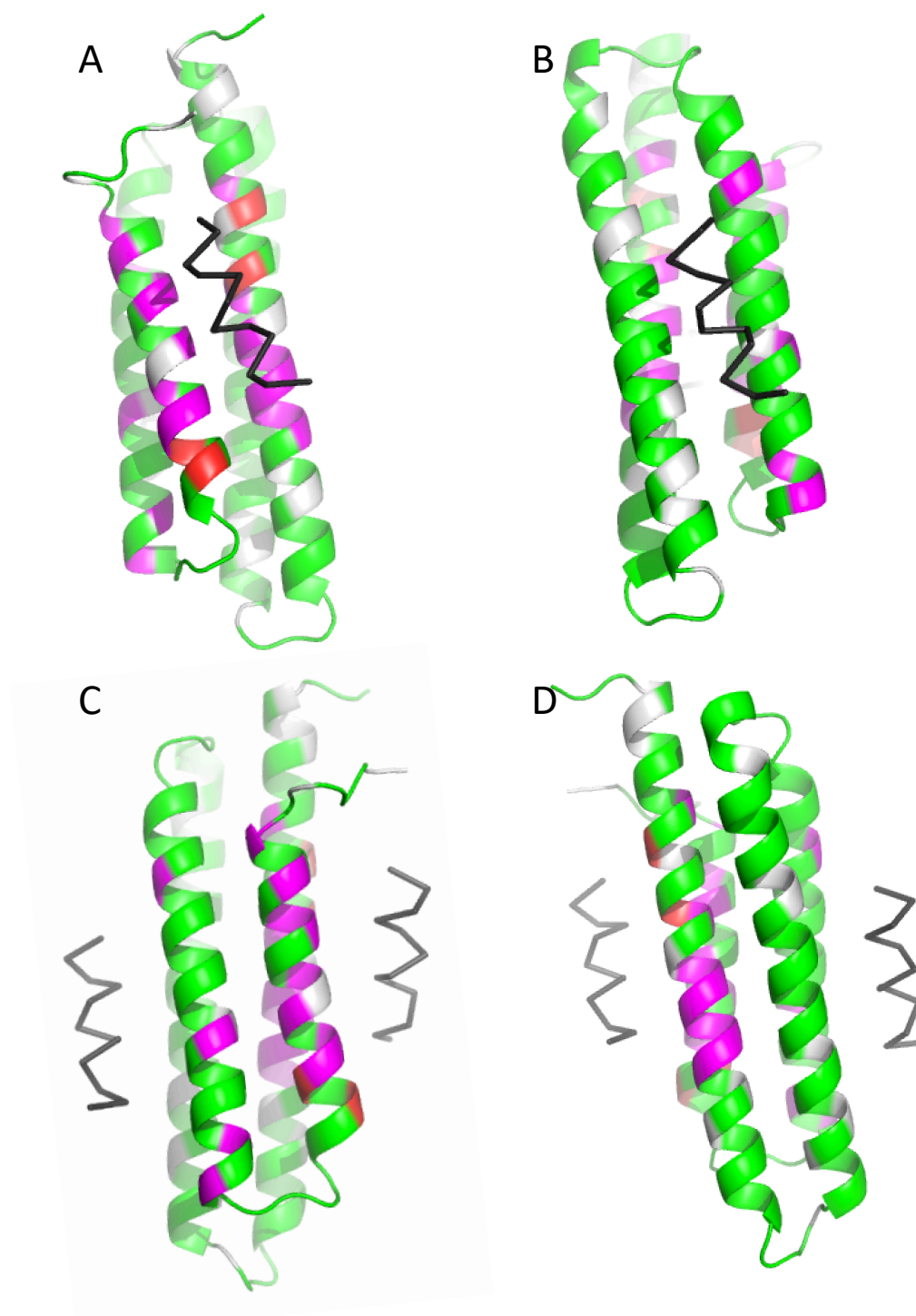
Supplementary Figure 5.9: Mapping changes of NMR resonances on FAT structure upon titration with LPP where A is the face of FAT helices 1 and 4, B is 2-3 helices, C is 1-2 helices and D is 3-4 helices. Resonances showing significant changes are labeled in black. LPP is represented as a ribbon. Others that disappeared are labeled in red. Gray residues represent amino acids whose peaks are not assigned. This figure was prepared using the PDB file 1OW7 and modified (Paxillin LD2 motif bound to 1-4 binding site and LD4 motif bound to 2-3 binding site on FAT of FAK).



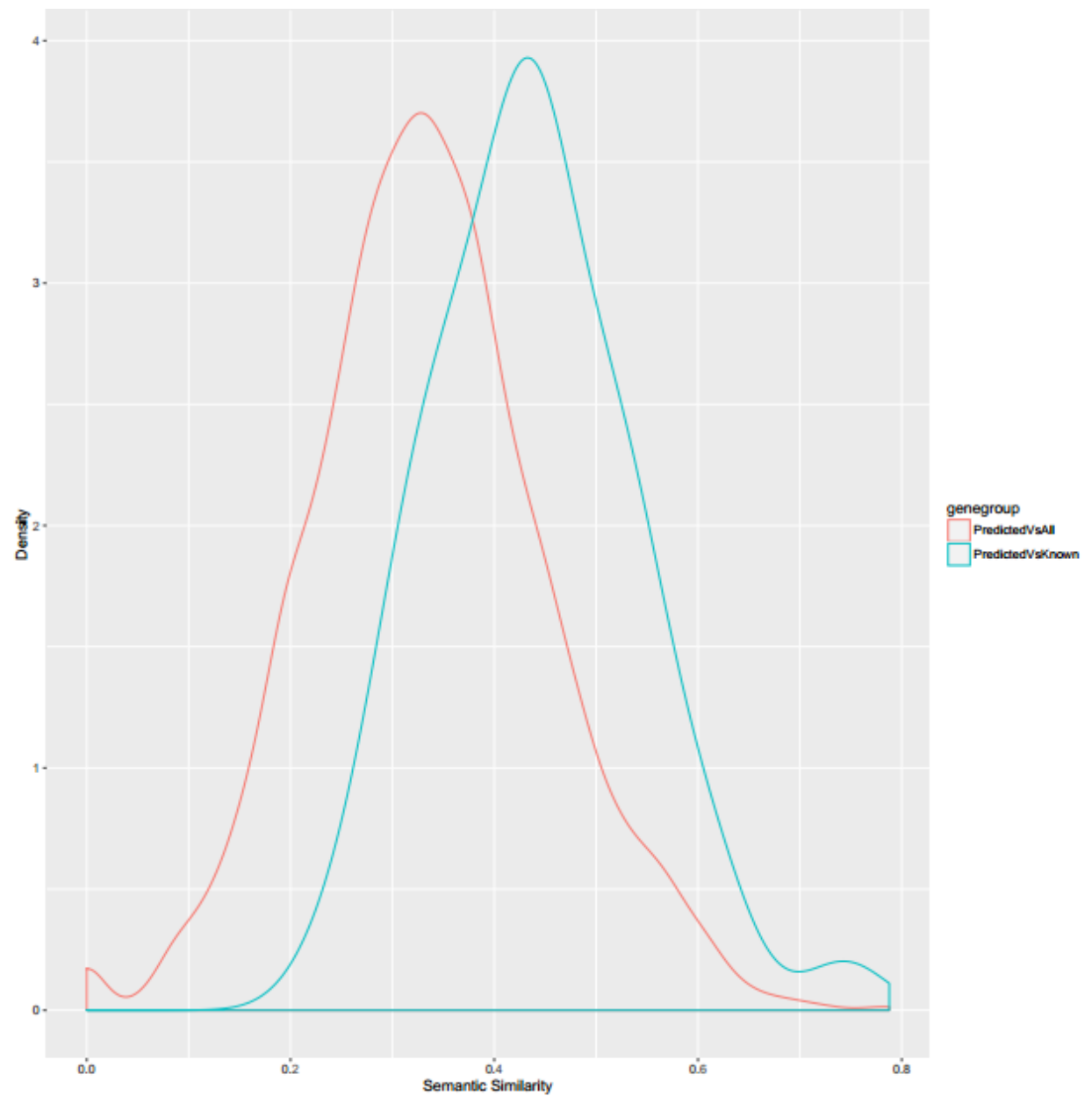
Supplementary Figure 5.10: FAT/CD158 Titration. Overlay of ^1H - ^{15}N -HSQC spectra of 100 μM ^{15}N -FAT in the absence (red) and presence of 0.5 (green), 1 (blue), 2 (yellow), 3 (magenta) and 4 (cyan) times molar excess of CD158 peptide. Resonances that disappeared upon CD158 addition are labelled in red. Resonances that significantly shifted $>2\sigma = 0.092$ are labelled in black. All spectra were recorded at 25°C at a proton frequency of 950 MHz.



Supplementary Figure 5.11: Chemical shift changes in FAT induced by CD158 peptide. Chemical shift differences in ppm were calculated for ^1H (top panel), ^{15}N (middle panel) and the weighted combined ^1H , ^{15}N (lower panel) chemical shift perturbation of FAT in the presence of a four times molar excess of CD158 peptide. Resonances showing significant change (greater than $2\sigma = 0.092$) are marked by stars. 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 5.12: Mapping changes of NMR resonances on FAT structure upon titration with CD158 where A is the face of FAT helices 1 and 4, B is 2-3 helices, C is 1-2 helices and D is 3-4 helices. Resonances showing significant changes are labeled in magenta. CD158 is represented as a ribbon. Others that disappeared are labeled in red. Gray residues represent amino acids whose peaks are not assigned. This figure was prepared using the PDB file 1OW7 and modified (Paxillin LD2 motif bound to 1-4 binding site and LD4 motif bound to 2-3 binding site on FAT of FAK).



Supplementary Figure 6:

GO ANALYSIS: 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 $6.32e-10$.

Conservation of human non-paxillin LD motif proteins across species. The region encompassing the identified 10-residue LD motifs in humans are boxed.

SPPASIPPPAGSPLTQNAAQOLDRL LASLTENLIDFTDPTPQSAMVANGV-
ETLMLITPADSGSVLK EATDELDALLASLTENLIDHTV-APQVSSTSMITP
ETLMLITPADSGSVLK EATDELDALLASLTENLIDHTV-APQVSSTSMITP
LTPVHGTAA DSASVLKDATDEL DALLLSLTENLM DHTV-TPQVSSPSMITP
LI SAN VAPAEEAAVSK SAPDORDVSL TSLTENLIDFTEATPRVSSQPTIT-

. : . * : * * ** ** * : * : :

IKGPEKLTLEERRSSLDAEIDSLTSLADLESSSPYKPRQTQNSASPAATG
VNPGGKLTLEERRSSLDAEIDSLTSLADLESSSPYKPRIQQSGTSSAA
GNPGGKLTLEERRSSLDAEIDSLTSLADLECSSPYKPRPPQGSASSIASP
GNPGGKLTLEERRSSLDAEIDSLTSLADLECSSPYKPRPPQSSTGSTASP
GNPRGKLTLEERRSSLDAEIDSLTSLADLECSSPYKPRPPQSSTGSTASP

L R H R P A G V L P L A K D M A P D L D Q L D D M L A Y I G H T S P E C V P P T V T Q L N A P S S S
 A K R R C R E V I P N W D S L Q D G E D S L D E M L Q Y L G Y S S P E C L Q R T G A P L N I P A P P
 Q F K R F R E T V P T W D T I R D E E D V L D E L L Q Y L G V T S P E C L Q R T G I S L N I P A P Q
 Q F K R F R E T V P T W D T I R D E E D V L D E L L Q Y L G V T S P E C L Q R T G I S L N I P A P Q
 Q L R R F R E T V P T W S T I R E E E D V L D E L L Q Y L G T T S P E C L Q R T G I S L N V P A P Q
 . * . * . : * * * . * * * . * * * * * . * * * :

----YNFTDPSKVIIPKNDKDIQQEDIKLLMELESFSQTIEDGQKYNSSRGNDV
 LCGSQAGVSK----GEQMCKIDKDEIDKLLLDLEHFSQKMESTFRESPKKESEF
 KVS^KFEEDQ----RDFTNSSSQEEDKLLMDLESFSQKMETSLREPLAKGKNS
 KVS^KFEEDQ----RDFTNSSSQEEDKLLMDLESFSQKMETSLREPLAKGKNS
 EVPTTLELRDQ----RHFMPNPSQEEIDKLLMDLESFSQKMETSLGEPLARGKSI

SGYLSFPVCRKSCFQEDLNSDPNNLQLDLKOTFCDEHSMKMSNKNDTNECE-
 ASCVPHMTSKPGILKEEPLRLKRIQLQLSSDS-EIPSVVLSKNDSDGGRV
 TSFLSHHSIKTNTPKEDPTRDLKQLLQELRTVINNEEPAMALSKTEEDGRT-
 ASFLSHHSIKANTLKEDPTRDLKQLLQELRSVINNEEPAVSLSKTEEDGRT-
 ASFLSHHSIKANTLKEDPTRDLKQLLQELRSMINNEEPAVSLSKTEEDGRT-
 : : : : * : * : : : * : : : * : :

MADDLDQLLDEVESRF CGQSAPRQ R NKGAGERV ---
MAEDLDELLDEVESK FCTPDLLRRG --- MVEQPKGC
MAEDLDELLDEVESK FCTPDLLRRG --- MVEQPKGC
MAKDLELLDEVET KFCRLDPLRLD --- LGERPKGD

* * * * * * . * * * * *

NCOA2

AAI63724.1_danio_rerio
sp|Q15596|NCOA2_HUMAN
PNI56205.1_chimpanzees
BAF69036.1_mouse

```
TGIKTEKTDGGYD---RVEPSSELDLDDLQNSQ-PGLFTDSRPVSLPSAVDK
----TEKEEMSFEPGDQPGSELNLEEILDDLQNSQLPQLFPDTRPGAPAGSVD-
----TEKEEMSFEPGDQPGSELNLEEILDDLQNSQLPQLFPDTRPGAPAGSVD-
----TVKEEVSFEPDQPGSELNLEEILDDLQNSQLPQLFPDTRPGAPTGSVD-
* * : : : * * : : : * * : : *
```

NCOA3

XP_692938.5_danio_rerio
sp|Q9Y6Q9|NCOA3_HUMAN
PNI59688.1_chimpanzees
XP_021041194.1_mouse

```
TSSSSNQESKVKL---EQPDELESLESLGGLRNP GPMFVDSGSGGSEVGNK---
----QEKDPKIKTETSEEGSGDLNLDAILGDLTSSDFYN-NSISSNGSHLGTQK--
----QEKDPKIKTETSEEGSGDLNLDAILGDLTSSDFYN-NSISSNGSHLGTQK--
----QEKDPKIKTETSEEVSGDLNLDAILGDLTSSDFYN-NPT--NGSHPGAKQQM
. . . : * * * : : * * : : * * : : *
```

CAST

XP_007891289.1_shark
sp|P20810|ICAL_HUMAN
tr|K7DU47|K7DU47_PANTR_chimpanzees
sp|P51125|ICAL_MOUSE
ABP68381.1_chicken

```
AAFSVSA SQPAPKGTGD-MGALDALSDML-QPEAPVHSGPKYTGPVEVKEKA
PAVPVESKPDKPSGKSGM-DAALDDLIDTLGGPEETEENTTYTGPEVSDP-
PAVPVESKPDKPSGKSGM-DAALDDLIDTLGGPEETEENTTYTGPEVSDP-
PASPVQSTPSKPSDKSGM-DAALDDLIDTLGHEDTNRDDPPYTGPVVLDP-
-VASMAAADKPNSEPMDESALDSLIDTLGSEEDVATRPVYTGPTEITEN-
. : : * : : * * * * * * * * * * *
```

CREB3

XP_009678890.1_ostrich
sp|Q61817.2|CREB3_MOUSE
sp|O43889|CREB3_HUMAN
XP_003829889.1_chimpanzees

```
EESLLLEDWGLSD---AQLLNKEMDDFISLLSPFVDEPGTLQGYSPDSDS
--VKASLDLELSPSENSVQELSDWEVDLLSLLSPSVSDVLGSSSSSILHD-
EAVRAPLDWALPL----SEVPSDWEVDLLCSLLSPASLNILSSSNPCLVHHD
EAVRAPLDWELPL----SEVPSDWEVDLLCSLLSPASLNILSSSNPCLVHHD
* * : : : * * : : * * * * * * * * *
```

RALGAP2

BAA92774.1_mouse
XP_023168385.1_drosophila
sp|Q2PPJ7|RGPA2_HUMAN
PNI39515.1_chimpanzees
XP_005158962.1_danio_rerio

```
-QLRRFRETVPWTSTIQEEDVLDELLQYLGTTSPECLQRTGISLNVPAQ
LRHR-PPGVLP LAKDVAPDLQ LDDMLAYIGHTSPECVPPTISELNAPSLS
-WHRDTFGPQKDSSQVEEGDVLDKLLENIGHTSPECLLPSQLNLNEPSLT
-WHRDTFGPQKDSSQVEEGDVLDKLLENIGHTSPECLLPSQLNLNEPSLP
-RSSVRFSEKCSSELDVEDGVLDQLLEDLGSSPECLPEPQLRLTQPPS-
: * * * * * * * * * * * * * * *
```

C16orf71

NP_001258515.1_mouse
C16orf71_human
PNI44266.1_chimpanzees

```
-----SLKQLESWLDYILQSLPGRQDSQGNSASRSAWWLADR
PLVEPPEGPPVLSLQQLQSLAGQEDNQGNRAPGTVWVAADH
-----SLQQLQSLAGQEDNQGNRAPGTVWVAAD-
* * * * * * * * * * * * * * *
```

RoXaN

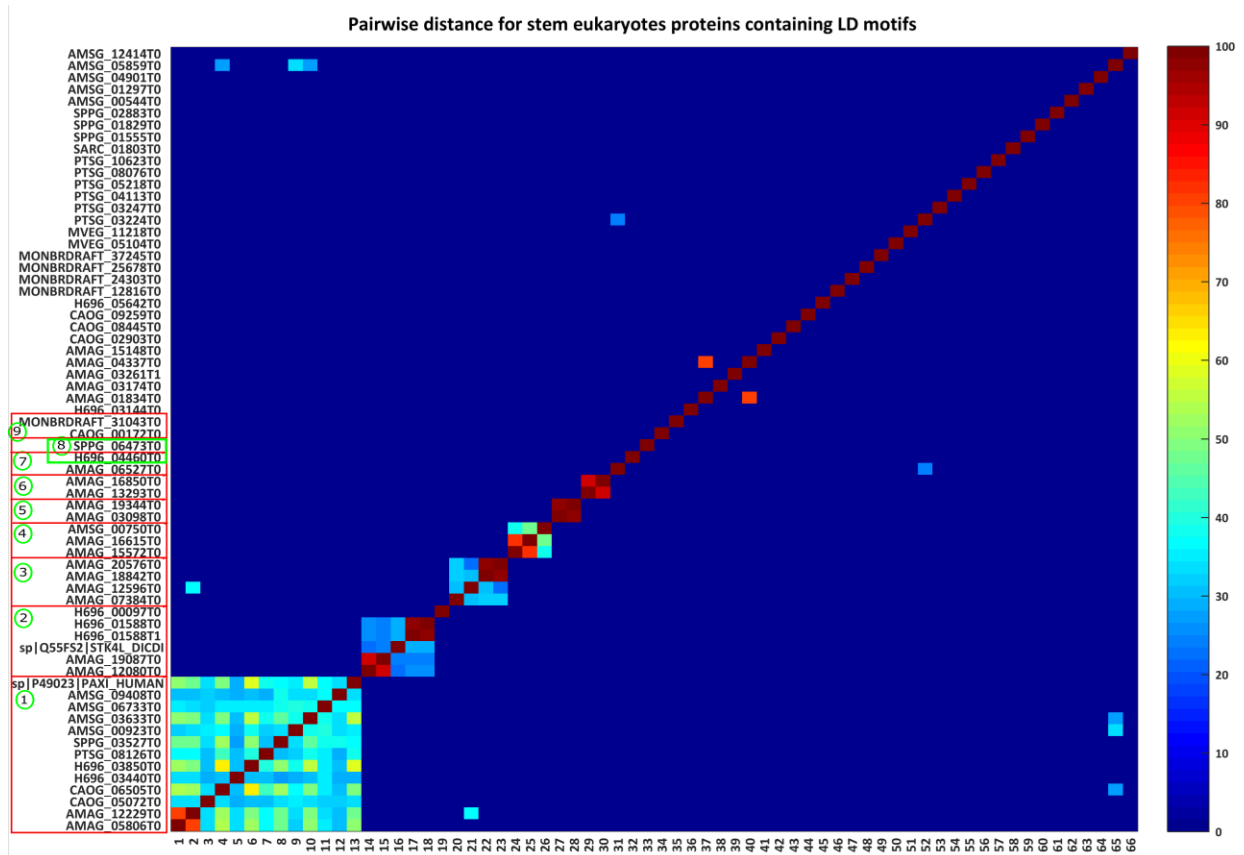
sp|Q9UGR2|Z3H7B_HUMAN
XP_009436757.1_chimpanzees
NP_001074485.1_mouse
KF061317.1_chicken
XP_007907761.1_shark

```
RTLPSDLSDDFSDGDFVGP ELDTL LDSL SVQGGLSGSGVPS ELPQLIP----
RTLPSDLSDDFSDGDFVGP ELDTL LDSL SVQGGLSGSGVPS ELPQLIP----
RTLPGTEGLDDFSDGDFVGP ELDTL LDSL SVQGGLP GSGMPSEL PQLIP----
VPLPVPENVEDFTDGIIEGELDSLDSLAE GS-PYPLGTIPTNLPT EMPQ---
VLSVPESTEEFTDGEIIEGIDTL LDSIQD-----YPMSSAPGIPTNVPANVS
* : : * * * * * * * * * * * * * * *
```

DLC1

XP_007889463.1_shark
sp|Q96QB1|RHG07_HUMAN
XP_016814601.1_chimpanzees
XP_021074615.1_mouse

```
GSHLYASTGDL LDKEDIFPHLDDILQHVNGLQQIVDHWKKNVLPQLQD
GSILYSSSGDLADLENE DIFPELDDILYHVKGMRIVNQWSEKFSDEGDS
GSILYSSSGDLADLENE DIFPELDDILYHVKGMRIVNQWSEKFSDEGDS
GSILYSSSGELADLENE DIFPELDDILYHVKGMRIVNQWSEKFSDEGDS
* * * * * * * * * * * * * * *
```



Supplementary Figure 8:

Unicellular Homology: The conservation of LD motif-containing proteins in unicellular eukaryotes. Heat map shows pairwise identity matrix (in percentage) where $E\text{-value} < 1e-10$. 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. Abbreviaions are as follows (1) P49023: *Homo sapiens*, (2) CAOG: *Capsaspora owczarzakii*, (3) MONBRDRAFT: *Monosiga brevicollis*, (4) MVEG: *Mortierella verticillata*, (5) Q55FS2: *Dictyostelium discoideum*, (6) SARC: *Sphaeroforma arctica*, (7) PTSG: *Salpingoeca rosetta*, (8) SPPG: *Spizellomyces punctatus*, (9) H696: *Fonticula alba*, (10) AMSG: *Thecamonas trahens*, and (11) AMAG: *Allomyces macrogynus*.

Supplementary Table 1: Results of predictions from final model

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

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

Supplementary Table 2: 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.

Index	Protein name	Start position	End position
1.	Paxillin PXN	3	12
Primary and secondary sequence			
-----MDDLADLESTTSHISKRPVFLSEETPYS			
-----CCHHHHHHHHHHCCCCCCCCCCCCCCCC			
2.	Paxillin PXN	144	153
Primary and secondary sequence			
OKSAEPSPTVMSTSLGSNLSELDRLLELNAVQHNPPGFPADEANSSPPL			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
3.	Paxillin PXN	216	225
Primary and secondary sequence			
PLTKEKPKRNGRGLEDVRPSVESLLDELESSVPSPVPAITVNOGEMSSP			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
4.	Paxillin PXN	265	274
Primary and secondary sequence			
PQRVTSTQQTRISASSATRELDLMASLSDFKIQGLEQRADGERCWAAG			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
5.	Paxillin PXN	333	342
Primary and secondary sequence			
MAQGKTGSSSPGPGPKPSQLDSMLGSLQSDLNKLG VATVAKGVC GACK			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
6.	Leupaxin LPXN	3	12
Primary and secondary sequence			
-----MEELDALLELERSTLQDSDEYSNPAPLPLDQ			
-----CCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
7.	Leupaxin LPXN	92	101
Primary and secondary sequence			
YSEAQEPKESPPPSKTAAAQLDELMALHTEMQAKVAVRADAGKKHLPDK			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCC			
8.	Leupaxin LPXN	127	136
Primary and secondary sequence			
VAVRADAGKKHLPDKQDHKASLDSMLGGLEQELQDLGIATVPKGHCASCQ			
HCCCCCCCCCCCCCCCCCHHCCCCCHHCCCCCHHHHHHHHHCCCCCCCCCCCC			
9.	Paxillin-B paxB	10	19
Primary and secondary sequence			
-----MATKGLNMDLDDLADLGRPKSSIKVTATVQTATPSS			
-----CCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCC			
10.	Paxillin-B paxB	108	117
Primary and secondary sequence			
VSSQPAPQPQQSQIDGLDDLDELMESLNTSISTALKAVPTTPEEHITH			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
11.	Paxillin-B paxB	231	240
Primary and secondary sequence			
SQSQPQPYKVTATNSQSSDDLDELLKGLSPSTTTTTTVPPPVRDQHQH			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
12.	Paxillin-B paxB	310	319
Primary and secondary sequence			
NTPNNNNNNNTNSPKVVHGGDDLNNLNNLTSQVKDIDSTGPTSRGTGGC			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCC			
13.	Transforming growth factor beta-1-induced transcript 1 protein TGFB1I1	3	12
Primary and secondary sequence			
-----MEDLDALLSDLETTTSHMPRSGAPKERPAEPL			
-----CCHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
14.	Transforming growth factor beta-1-induced transcript 1 protein TGFB1I1	92	101
Primary and secondary sequence			
AAPAAPPFSSSGVLGTGLCELDRLQLNATQFNITDEIMSQFPSSKVA			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
15.	Transforming growth factor beta-1-induced transcript 1 protein TGFB1I1	157	166
Primary and secondary sequence			
SLPSSSPGLPKASATSATLELDRLMASLSDFRVQNHLPASGPTQPPVVS			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCHHHCCCCCCCCCCCC			
16.	Transforming growth factor beta-1-induced transcript 1 protein TGFB1I1	203	212
Primary and secondary sequence			
PVVSSTNEGSPSPPEPTGKGLDMLGLLQSDLSRRGVPTQAKGLCGSCN			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
17.	Zinc finger CCCH domain-containing protein 7B ZC3H7B	280	289
Primary and secondary sequence			
RTLPSSTDLSLDFSDGDVFGPELDTLLDSLVLVQGGLSGSGVPSELPLQIP			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
18.	Rho GTPase-activating protein 7 Dlc1	905	914
Primary and secondary sequence			
SILYSSSGELADLENEDIPELDDILYHVKGMRIVNQWSEKFSDEGDSD			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCC			

Supplementary Table 2.a: Information for the *bona fide* LD motifs.

Index	Protein name	Start position	End position
1.	Band 4.1-like protein 5 EPB41L5	634	643
Primary and secondary sequence			
ETLMITPADSGSVLKEATDELDALLASLTENLIDHTVAPQVSSTSMITP			
HHHHCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
2.	Insulin-like growth factor-binding protein 2 IGFBP2	230	239
Primary and secondary sequence			
LGLEPKKLRPPPARTPCQELDQVLERISTMRLPDERGPLEHLYSLHIP			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
3.	Protein C8orf37 C8orf37	4	13
Primary and secondary sequence			
-----MAEDLDELLDEVESKFCTPDLLRRGMVEQPKGC			
-----CHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
4.	RaI GTPase-activating protein subunit alpha-1 RALGAP1	1680	1689
Primary and secondary sequence			
QFKRFRETVPTWDTIRDEEDVLDLQYLGVTSPLELQRTGISLNIPAPQ			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
5.	Uncharacterized protein C16orf71 C16orf71	267	276
Primary and secondary sequence			
PLVEPPEGPPVLSLQQLA WDLDDILQSLAQEDNQGNRAPGTVWAAADH			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
6.	Lipoma-preferred partner LPP	123	132
Primary and secondary sequence			
GNPGGKTLERRSSLD A EIDSLTSLADLECSPYKPRPPQSSTGSTASP			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
7.	Pre-mRNA 3'-end-processing factor FIP1 FIP1L1	5	14
Primary and secondary sequence			
-----MSAGEVERLVSELGGTGGDEEEEWLYGGPWVDVH			
-----CCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
8.	Calpastatin CAST	156	165
Primary and secondary sequence			
PAVPVESKPKPSGKSGMDAALDDLIDTLGGPEETEEENTTYTGPEVSDP			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
9.	Nuclear receptor coactivator 2 NCOA2	805	814
Primary and secondary sequence			
KTEKEEMSFEPGDQPGSELNLEEILDDLQNSQLPQLFPDTRPGAPAGSV			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
10.	Nuclear receptor coactivator 3 NCOA3	799	808
Primary and secondary sequence			
QEKDPKIKTETSEEGSGDLNLDAILGDLTSSDFYNNSISSNGSHLGTKQ			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
11.	WASP homolog-associated protein with actin, membranes and microtubules WHAMM	22	31
Primary and secondary sequence			
VCESPAERPRDSLESFSCPGSMDEVLASLRHGRAPLRKVEVPAVRPPHAS			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
12.	RaI GTPase-activating protein subunit alpha-2 RALGAP2	1519	1528
Primary and secondary sequence			
WHRDTFGPQKDSSQVEEGDDVLDKLENNIGHTSPECLLPQLNLNEPSLT			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
13.	Purkinje cell protein 2 homolog PCP2	62	71
Primary and secondary sequence			
RCSLQAGPGQTTKSQSDPTPEMDSLMDMLASTQGRRMDDQRTVSSLPGF			
CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			

Supplementary Table 2.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 sequence			
ETLMLITPADSGSVLKEATDELDALLASLTENLIDHTVAPQVSSTSMITP HHHHCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
2.	Serine/threonine-protein phosphatase 2A regulatory subunit B" subunit alpha PPP2R3A	508	517
Primary and secondary sequence			
KVSKEEGDQORDFTNSSSQEEIDKLLMDLESFSQKMETSLREPLAKGKNS CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCC			
3.	Coiled-coil domain-containing protein 158 CCDC158	903	912
Primary and secondary sequence			
ASFLSHHSTKANTLKEDPTRDLKQLQELRSVINEEPAVLSKTEEDGRT HHHHCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC			
4.	Ral GTPase-activating protein subunit alpha-1 RALGAP1	1680	1689
Primary and secondary sequence			
QFKRFRETVPWTDITRDEEDVLDELLQYLGVTSPCLQRTGISLNIPAPQ CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
5.	Uncharacterized protein C16orf71 C16orf71	267	276
Primary and secondary sequence			
PLVEPPEGPPVLSLQQLA WDLDDILQSLAQEDNQGNRAPGTVWVAADH CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
6.	Lipoma-preferred partner LPP	123	132
Primary and secondary sequence			
GNPGGKTLEERRSLDAEIDSLTSLADLECSSPYKPRPPQSSTGSTASP CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
7.	Cyclic AMP-responsive element-binding protein 3 CREB3	49	58
Primary and secondary sequence			
EAVRAPLDWALPLSEVPSDWEVDDLCSLLSPPASLNILSSSNPCLVHHH HHHHCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
8.	Calpastatin CAST	156	165
Primary and secondary sequence			
PAVPVESKPDKPSGKSGMDAALDDLIDTLGGPEETEEENTTYTGPEVSDP CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
9.	Nuclear receptor coactivator 2 NCOA2	805	814
Primary and secondary sequence			
KTEKEEMSFEPGDQPGSELDNLEEILDDLQNSQLPQLFPDTRPGAPAGSV CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
10.	Nuclear receptor coactivator 3 NCOA3	799	808
Primary and secondary sequence			
QEKDPKIKTETSEEGSGDLNLDAILGDLTSSDFYNNSSISNGSHLGTKQ CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			
11.	Protein C8orf37 C8orf37	4	13
Primary and secondary sequence			
-----MAEDLDELLEVESKFCTPDLLRRGMVEQPKGC -----CHHHCCCCCCCCCCCCCCCCCCCC			
12.	Ral GTPase-activating protein subunit alpha-2 RALGAP2	1519	1528
Primary and secondary sequence			
WHRDTFGPQKDSSQVEEGDDVLDKLENNIGHTSPECLLPQLNLNEPSLT CCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCC			

Supplementary Table 2.c: Information of the 12 new LD motifs finally suggested by LDMF.

Supplementary Table 3: Computational Validation

Summary of the bioinformatic search for evidence supportive of interactions between known LD-motif binding proteins and the LDMF-predicted LD motif-containing proteins from the human proteome.

To allow straightforward reproducibility, gene names are given in the table. The corresponding protein names for the LDBD-containing proteins are XPO1: exportin; PABPC1: polyadenylate-binding protein 1 (PABP-1); VCL: vinculin; TLN1: talin; PARVA: α -parvin; PARVB: β -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 2 β (PYK2); GIT1: Arf GTPase-activating protein/GRK-interacting protein 1 (GIT1); GIT2: Arf GTPase-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 α -1 (RGPA1); PPP2R3a: Serine/threonine-protein phosphatase 2A regulatory subunit B' subunit α (P2R3A); CCDC158: coiled-coil domain-containing 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 α -2 (RGPA2); C8orf37: uncharacterized protein C8orf37 (CP037).

Highly likely													
gene name	XPO1	PABPC1	VCL	TLN1	PARVA	PARVB	PARVG	PDCD10	PTK2	PTK2B	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	Small, 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													
gene name	XPO1	PABPC1	VCL	TLN1	PARVA	PARVB	PARVG	PDCD10	PTK2	PTK2B	GIT1	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													
gene name	XPO1	PABPC1	VCL	TLN1	PARVA	PARVB	PARVG	PDCD10	PTK2	PTK2B	GIT1	GIT2	BCL2
RALGAPA2	-	Small	-	Small	Small*	-	Small	-	-	0.93, Small	-	Medium	Small
C8orf37	Small	-	-	Small*	-	-	-	Small	Small	-	-	-	-

Blue is probability score > 0.5 from PrePP1 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.