	Screening with Insv-BEN in the *Drosophila* proteome		Screening with Insv-BEN in the mouse proteome	
	Drosophila Protein	Z-score	Mouse protein	Z-score
1	Elba2	14.9	Bend6	13.2
2	insv	13.6	Bend5	12.7
3	Bsg25A	12.4	Banp	12.6
4	pre-mod(mdg4)-C	8.1	Nacc1	12.4
5	CG17341	7.3	Bend7	12.3
6	CG42854	5.8	Nacc2	12.1
7	PNPase	5.6	Bend4	11.3
8	CG12112	5.1	Gm15262	10.4
9	CG31367	5.1	Bend3	8.0
10	ey	4.9	Tcerg1l	5.2
11	bcd	4.8	Nol4l	5.1
12	stops	4.8	Tcerg1	5.0
13	B-H1	4.8	Dmbx1	4.6
14	Fip1	4.8	Hoxc9	4.5
15	ftz	4.8	Lbx1	4.4
16	tup	4.7	Hoxd1	4.4
17	ro	4.6	Meis3	4.4
18	eve	4.5	Hoxa9	4.4
19	Lmx1a	4.5	Lhx9	4.4
20	Vsx2	4.5	Pou6f2	4.3
21	pdm3	4.4	Scg3	4.3
22	B-H2	4.4	Hoxa3	4.3
23	anon-37Cs	4.3	Gsc	4.3
24	CG11085	4.3	Hoxa7	4.3
25	Dfd	4.3	Hoxc11	4.3
26	gsb	4.2	Evx1	4.3
27	unc-4	4.2	Zhx1	4.3
28	exex	4.2	Six1	4.3
29	unpg	4.2	Lhx1	4.2
30	E5	4.2	Nkx2-3	4.2
31	lbe	4.2	Hlx	4.2
32	toe	4.2	Isl1	4.2
33	en	4.2	Zhx3	4.2
34	GlnRS	4.2	Pou1f1	4.2
35	Tig	4.2	Emx2	4.2
36	CG13141	4.1	Pax6	4.2
37	Vsx1	4.1	Hoxd3	4.2
38	Dr	4.1	Hoxd9	4.1
39	nub	4.1	Pknox1	4.1
40	Lim1	4.0	Hesx1	4.1
41	so	4.0	Gbp5	4.1
42	CG34031	4.0	Hoxa11	4.1
43	oc	4.0	Pitx2	4.1
44	CG4196	4.0	Pde3a	4.0
45	lbl	4.0	Q8K2W9	4.0
46	Rx	4.0	Meis2	4.0
47	eyg	4.0	Vsx1	4.0
48			Lmx1b	4.0
49			Nkx6-3	4.0
50			Six3	4.0
51			Emx1	4.0
52			Pou4f3	4.0

	Screening with BEND3-BEN4 in the *Drosophila* proteome		Screening with BEND3-BEN4 in the mouse proteome	
	Drosophila Protein	Z-score	Mouse protein	Z-score
1	Elba2	7.6	Bend3	17.5
2	pre-mod(mdg4)-C	6.6	Banp	10.3
3	insv	6.4	Bend5	9.5
4	Bsg25A	6.0	Gm15262	9.2
5	PNPase	4.5	Bend6	9.0
6	CG17341	4.4	Nacc1	8.2
7	Fip1	4.0	Bend7	7.9
8	atl	4.0	Nacc2	7.8
9			Bend4	7.7
10			Nat8	4.2
11			Exoc8	4.2
12			Scg3	4.1
13			Nat8f2	4.0
14			Krtcap2	4.0
15			Gbp9	4.0

🟥 Known BEN

🟩 DUF4806

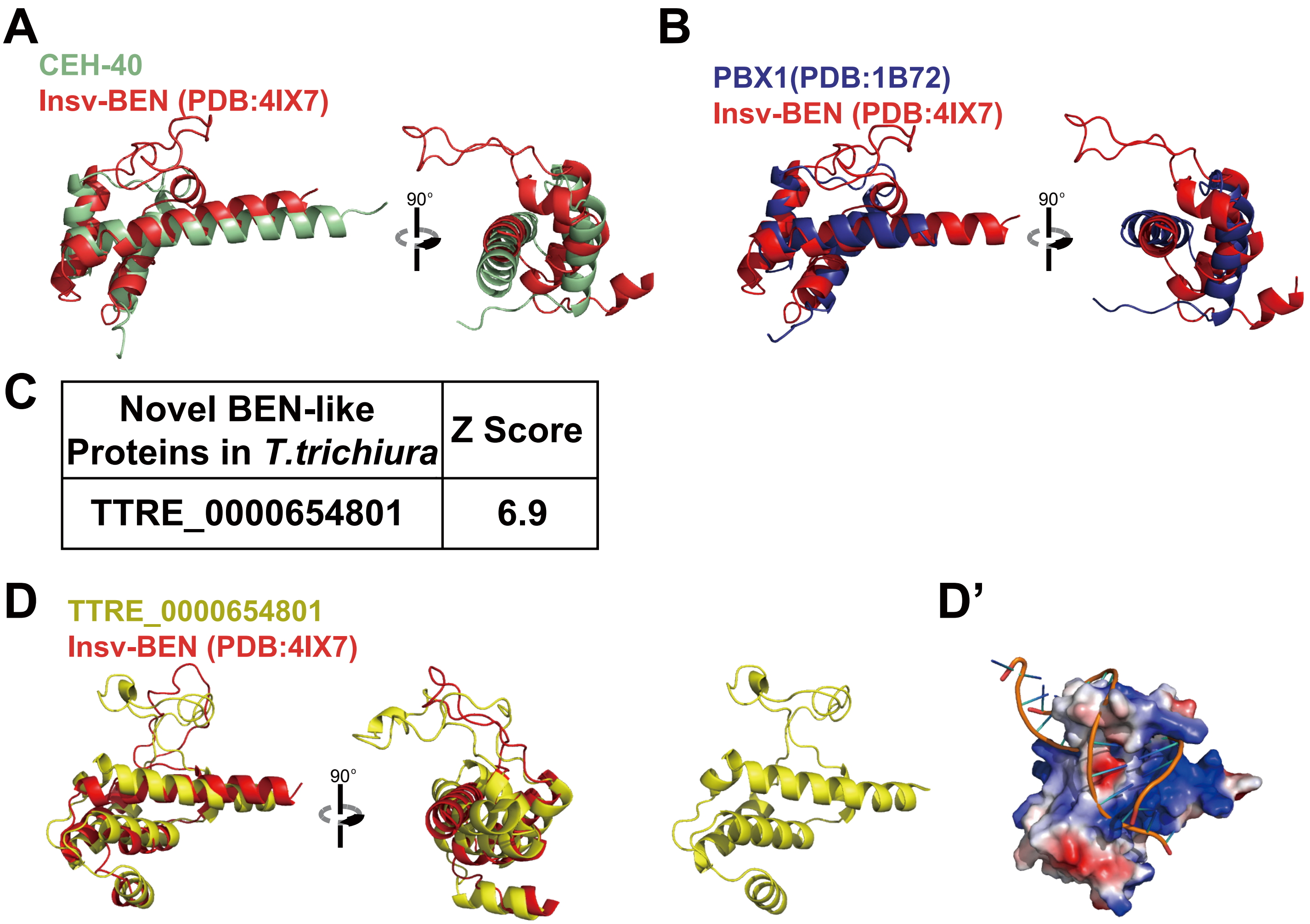
🟪 Homeodomain

🟨 FF

🟦 PH domain

⬜ Unannotated

Figure S1. Structural comparisons reveal BEN domains across species. Related to Figure 1, Data S1A, S1B, S1F and S1G. Structure comparisons are performed between X-ray determined BEN domain structures and AlphaFold predicted models of different species. In particular, the human BEND3-BEN4 structure (PDB: 7W27) was screened in the *Drosophila* proteome, while the *Drosophila* Insv-BEN (PDB: 4IX7) was screened in the mouse proteome. Annotate domains are highlighted with indicated colors.



Novel BEN-like Proteins in *T.trichiura*	Z Score
TTRE_0000654801	6.9

Figure S2. Structure comparisons between the experimentally solved structure of Insv-BEN and AlphaFold predicted models in worms. Related to Figure 2, Data S4A, S4P and S4Q. (A) Structure comparison between the AlphaFold predicted model of CEH-40, a PBX type homeodomain protein, and the solved structure of Insv-BEN (PDB:4IX7). **(B)** Structure comparison between the solved structure of human PBX1 homeodomain (PDB:1B72) and Insv-BEN. **(C)** Structure comparison with Insv-BEN reveals a novel BEN domain the *T. trichiura* proteome. **(D)** Superposition of the structure of Insv-BEN and the predicted model of TTRE_0000654801. The structure prediction also reveals the electrostatic surface potential of this BEN-like domain in complex with the Insv-BEN targeting DNA.

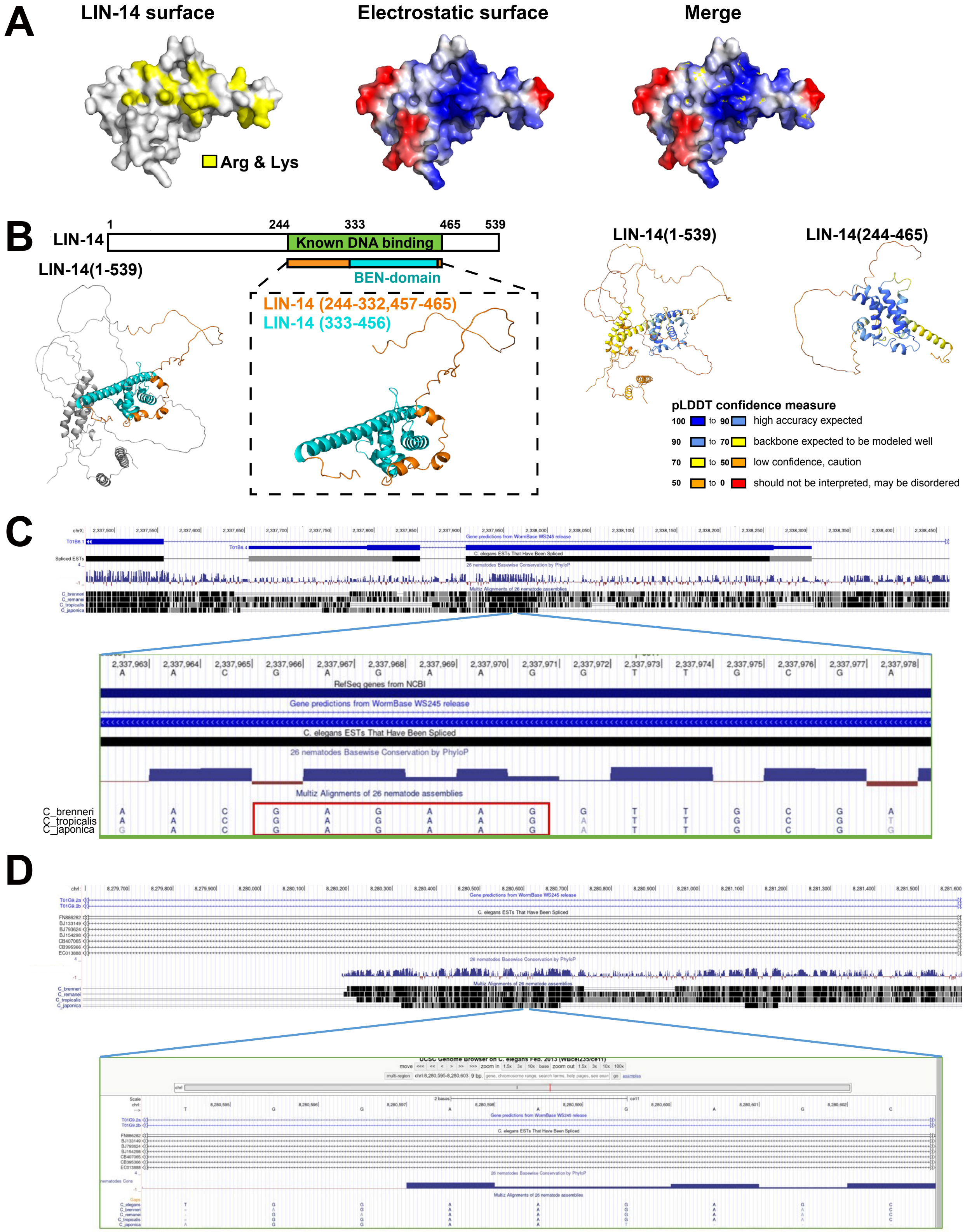
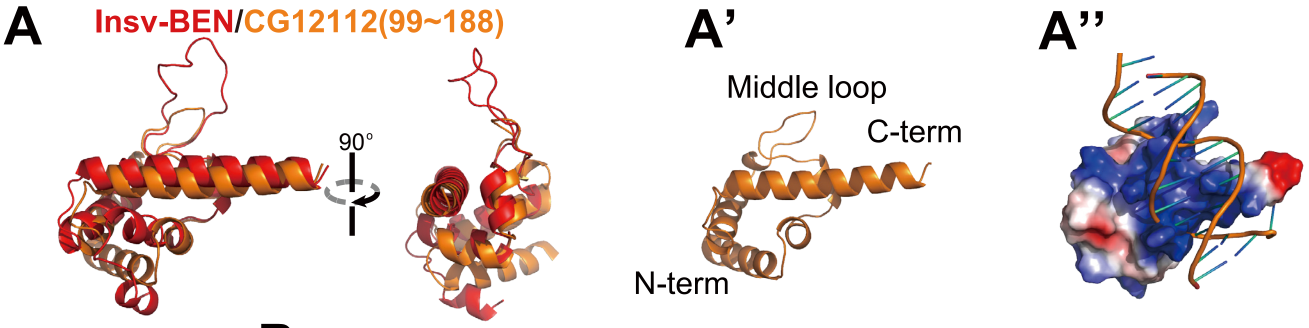

A
LIN-14 surface
Electrostatic surface
Merge
Arg & Lys
B
LIN-14
1
244
333
465
539
Known DNA binding
BEN-domain
LIN-14(1-539)
LIN-14 (244-332,457-465)
LIN-14 (333-456)
LIN-14(1-539)
LIN-14(244-465)
pLDDT confidence measure
100 to 90 high accuracy expected
90 to 70 backbone expected to be modeled well
70 to 50 low confidence, caution
50 to 0 should not be interpreted, may be disordered
C
Gene predictions from WormBase WS245 release
Spliced ESTs
C. elegans ESTs That Have Been Spliced
26 nematodes Basewise Conservation by PhyloP
Multiz Alignments of 26 nematode assemblies
2,337,963
2,337,964
2,337,965
2,337,966
2,337,967
2,337,968
2,337,969
2,337,970
2,337,971
2,337,972
2,337,973
2,337,974
2,337,975
2,337,976
2,337,977
2,337,978
A A C G A G A A G G T T G C G A
RefSeq genes from NCBI
C_brenneri
C_tropicalis
C_japonica
D
Gene predictions from WormBase WS245 release
C. elegans ESTs That Have Been Spliced
26 nematodes Basewise Conservation by PhyloP
Multiz Alignments of 26 nematode assemblies
UCSC Genome Browser on C. elegans Feb. 2013 (WBcel235/ce11)
T G G A A G A G C
C_elegans
C_brenneri
C_remanei
C_tropicalis
C_japonica

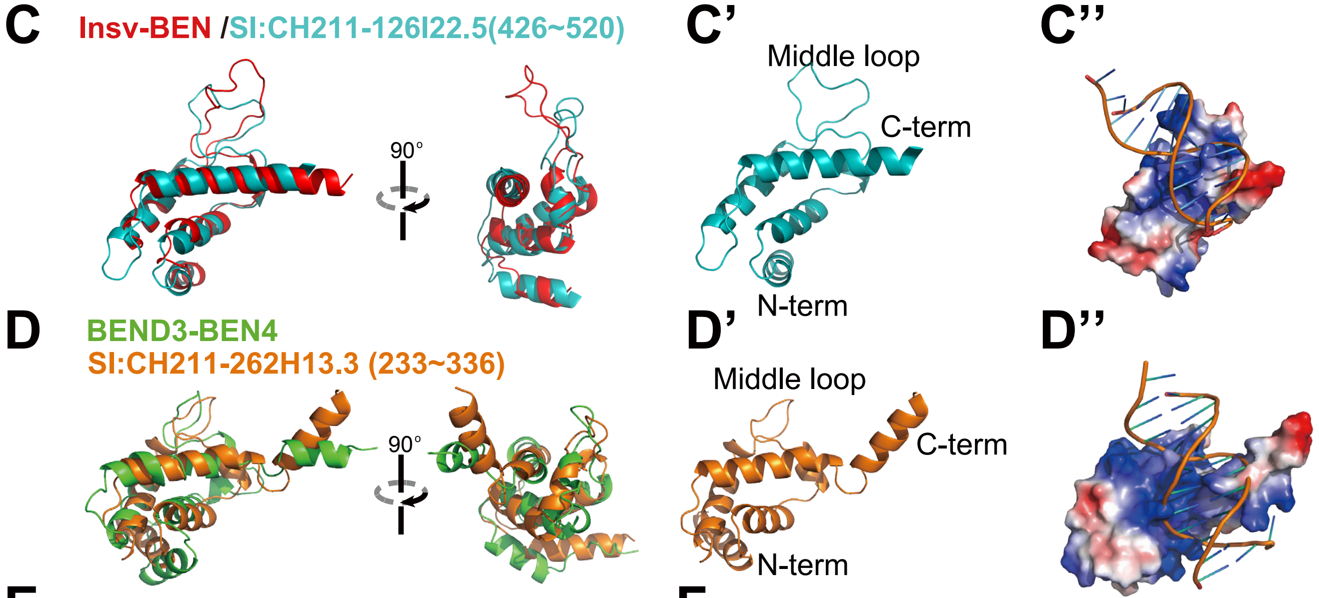
Figure S3. LIN-14 is a novel BEN domain. Related to Figure 2. (A) The distribution of Lys and Arg residues (yellow) in the α5 helix of LIN-14 BEN domain with electrostatic surface view. **(B)** The region comprising amino acids 244-332 of LIN-14 is predicted to be disordered, while amino acids 244-465 has been previously identified as a DNA-binding region. Structure comparisons reveal that the region between amino acids 333 and 465 contains a BEN domain (also shown in Figure 2G-I'). **(C-D)** LIN-14 binding sites on *nlp-45* and *dma-1* genes contain conserved GA-rich LIN-14 targeting DNA motif.



B

Novel BEN Domains in the Zebrafish

Gene Symbol	Uniprot ID	Z Score	Annotated Domain	BEN-Type
si:dkey-266f7.4	A0A0R4IGE0	9.2	DUF4806	Type I
si:dkey-266f7.5	I3IRX2	9.0	DUF4806	Type I
si:ch211-126i22.5	E9QFL0	8.6	N.A.	Type I
si:ch211-67e16.4	E7FEY9	8.5	DUF4806	Type I
si:ch211-262h13.3	Q1LUV1	8.5	DUF4806	Type II
si:ch211-244k5.1	X1WEX0	8.2	DUF4806	Type II
si:dkeyp-107f9.2	A0A0R4IKR7	8.0	DUF4806	Type I
si:dkey-65l23.2	A5WUV4	7.9	DUF4806	Type I
si:ch211-73m21.1	K7DY41	7.0	DUF4806	Type I



E

Screening for BEN domain proteins with CG42854 (67-160) in the mouse proteome

Rank	Gene Symbol	Uniprot ID	Z Score	r.m.s.d.	lali	nres	%id
1	Bend6	Q6PFX2	6.8	3.4	86	281	10
2	Nacc1	Q7TSZ8	5.6	3.3	82	514	12
4	Banp	Q8VBU8	5.3	3.8	80	548	11
5	Bend5	Q8C6D4	5.1	3.0	83	421	11
6	Nacc2	Q9DCM7	5.0	3.4	78	586	13
8	Bend3	Q6PAL0	4.8	3.1	83	825	13
9	Gm15262/Bend2	A0A140LIQ5	4.6	3.3	83	728	17
11	Bend4	P86174	4.0	3.5	71	541	20
12	Bend7	Q8BSV3	4.0	3.6	73	434	11

F

Screening for BEN domain proteins with CG12112 (99-188) in the mouse proteome

Rank	Gene Symbol	Uniprot ID	Z Score	r.m.s.d.	lali	nres	%id
1	Bend6	Q6PFX2	6.3	3.0	81	281	10
2	Nacc1	Q7TSZ8	5.1	3.1	76	514	13
4	Bend5	Q8C6D4	4.9	2.9	78	421	13
6	Banp	Q8VBU8	4.6	3.7	80	548	10
7	Nacc2	Q9DCM7	4.6	3.2	75	586	15
8	Gm15262/Bend2	A0A140LIQ5	4.1	3.3	83	728	13
26	Bend7	Q8BSV3	3.6	3.0	72	434	13
106	Bend4	P86174	3.3	3.5	70	541	17

Figure S4. Identification of novel BEN-like structures in *Drosophila* and zebrafish. Related to Figure3, Data S3B, S3C, S4A, S4H, S4R and S4H. (A-A’) Superposition of the solved structure of Insv-BEN (PDB:4IX7, red) and the predicted model of CG12112. **(A’’)** shows the electrostatic surface potential of the CG12112 BEN-like structure in complex with the Insv targeting DNA. **(B)** Zebrafish proteins containing BEN-like structures revealed by structure comparisons. Except for E9QFL0, all the rest BEN-like regions overlap with presumed DUF4806 motifs. **(C-C’)** Superposition of Insv-BEN (PDB: 4IX7, red) and the predicted model of zebrafish SI:CH211-126I22.5 (426~520). **(C’’)** shows the electrostatic surface potential of the SI: CH211-126I22.5 (426~520) BEN-like region in complex with the Insv-BEN targeting DNA. **(D-D’)** Superposition of the structure of BEND3-BEN4 (PDB:7W27, green) and the predicted model of zebrafish SI: CH211-262H13.3 (233~336). **(D’’)** shows the electrostatic surface potential of the SI:Ch211-262H13.3 BEN-like region in complex with the BEND3--BEN4 targeting DNA. **(E-F)** Structure screening reveal proteins with DUF4806-like structures in the mouse proteome. Predicted models of DUF4806 (with the downstream α5 helices) of *Drosophila* CG42854 and CG12112 were used to screen for proteins with 3D similarities. Both queries revealed known BEN-containing factors.

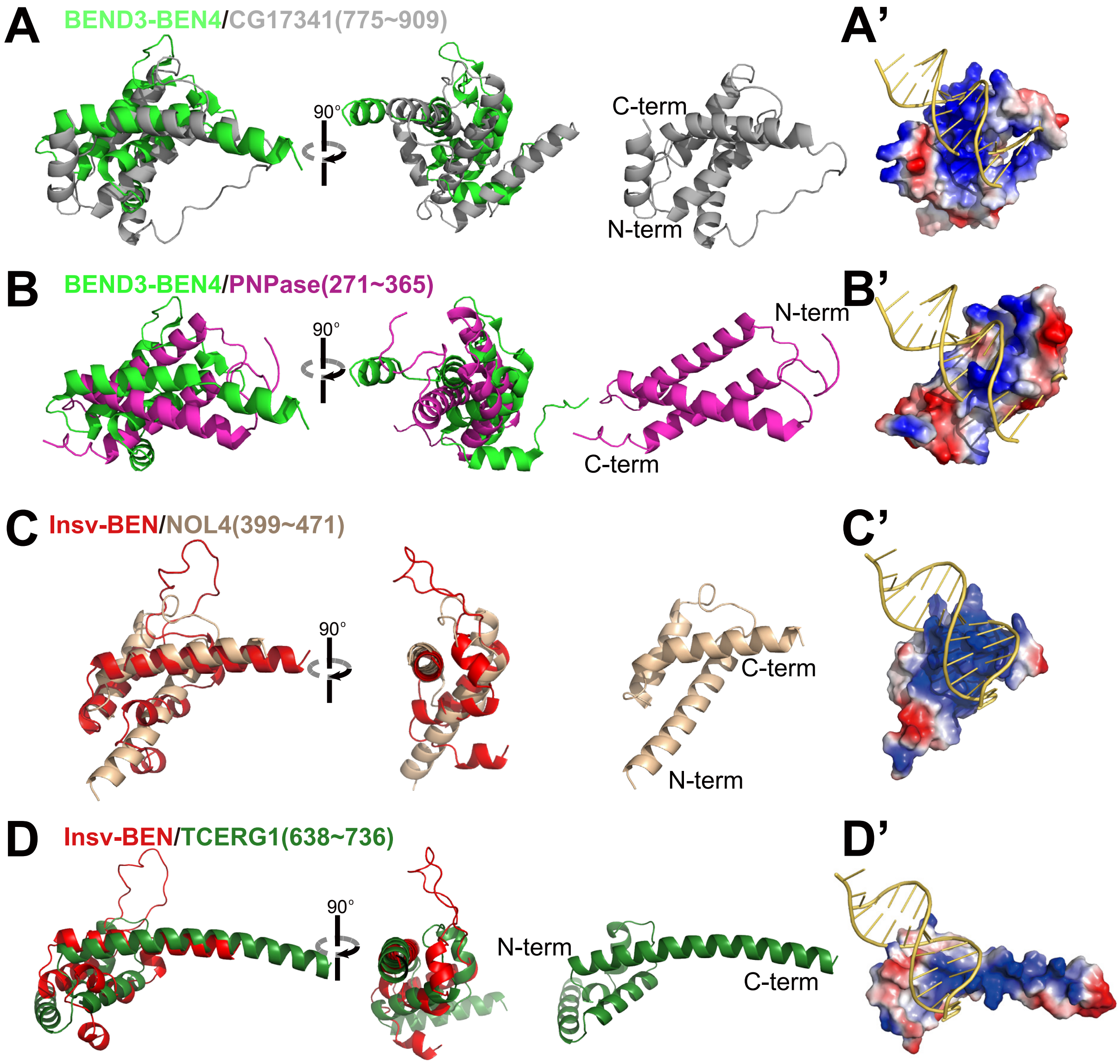


Figure S5. Structure comparisons reveal the similarity between BEN domain and HTH motif. Related to Figure 5, Data S4A, S4T-S4W. (A) Superposition of the structure of BEND3-BEN4 (PDB: 7W27, green) and the predicted model of *Drosophila* CG17341 (775~909). While the amino acids 775-909 in CG17341 is highly structured and have an HTH core, it has not been previously annotated with a known domain. **(B)** Superposition of the structure of BEND3-BEN4 (PDB: 7W27, green) and the predicted model of *Drosophila* PNPase (271~365). The amino acids 272-365 in PNPase have been annotated as a PH domain (also Polyribonucleotide nucleotidyltransferase, RNA-binding domain). **(C)** Superposition of the structure of Insv BEN domain (PDB: 4IX7, red) and the predicted model of mouse NOL4 (388~471). While the amino acids 388-471 in NOL4 is structured and have an HTH core, it has not been previously annotated as a known. **(D)** Superposition of the structure of Insv-BEN (PDB: 4IX7, red) and the predicted model of mouse TCERG1. The amino acids (638~736) in TCERG1 have been annotated as a FF1 domain.

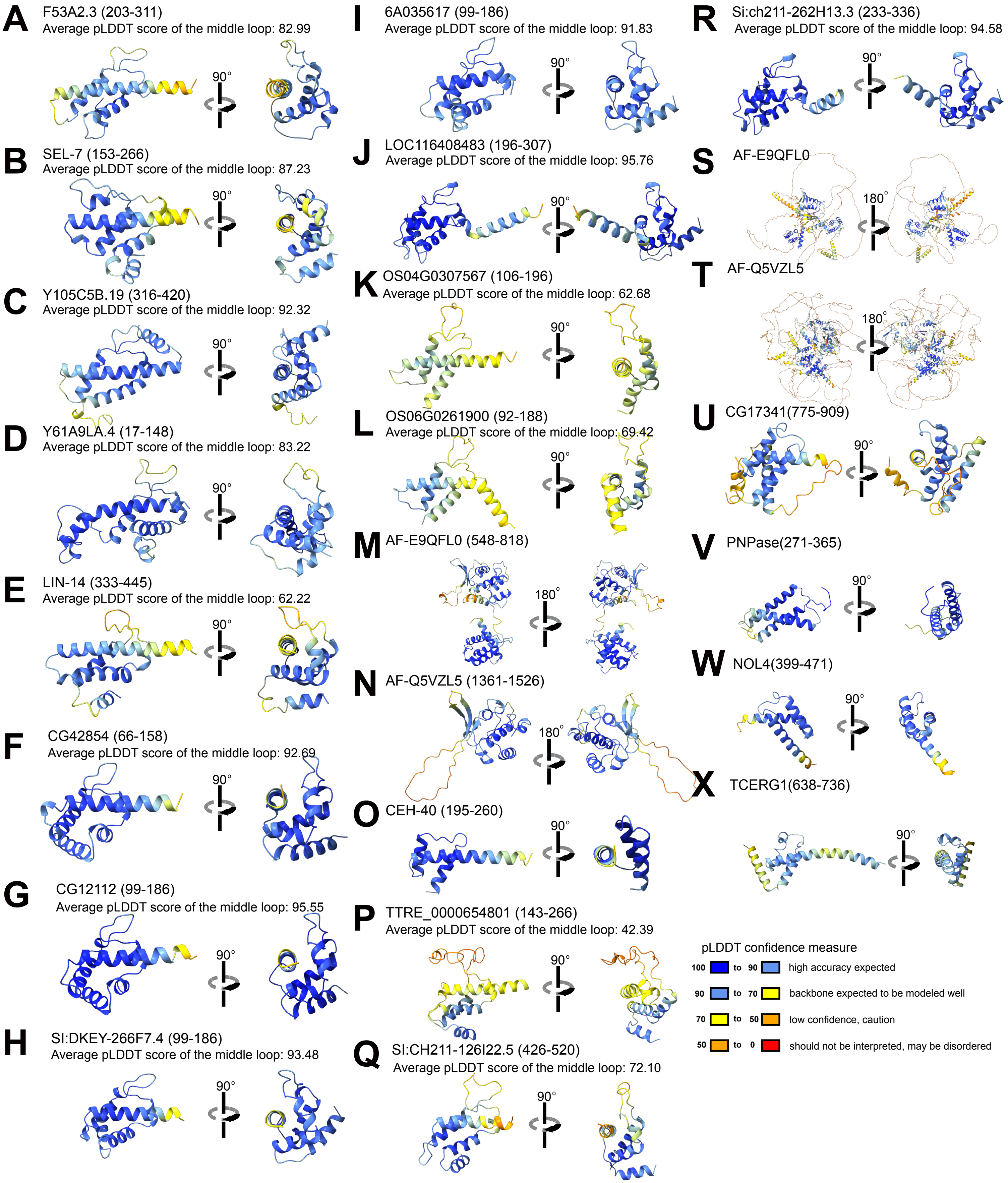

A
F53A2.3 (203-311)
Average pLDDT score of the middle loop: 82.99
90°
B
SEL-7 (153-266)
Average pLDDT score of the middle loop: 87.23
90°
C
Y105C5B.19 (316-420)
Average pLDDT score of the middle loop: 92.32
90°
D
Y61A9LA.4 (17-148)
Average pLDDT score of the middle loop: 83.22
90°
E
LIN-14 (333-445)
Average pLDDT score of the middle loop: 62.22
90°
F
CG42854 (66-158)
Average pLDDT score of the middle loop: 92.69
90°
G
CG12112 (99-186)
Average pLDDT score of the middle loop: 95.55
90°
H
SI:DKEY-266F7.4 (99-186)
Average pLDDT score of the middle loop: 93.48
90°
I
6A035617 (99-186)
Average pLDDT score of the middle loop: 91.83
90°
J
LOC116408483 (196-307)
Average pLDDT score of the middle loop: 95.76
90°
K
OS04G0307567 (106-196)
Average pLDDT score of the middle loop: 62.68
90°
L
OS06G0261900 (92-188)
Average pLDDT score of the middle loop: 69.42
90°
M
AF-E9QFL0 (548-818)
180°
N
AF-Q5VZL5 (1361-1526)
180°
O
CEH-40 (195-260)
90°
P
TTRE_0000654801 (143-266)
Average pLDDT score of the middle loop: 42.39
90°
Q
SI:CH211-126I22.5 (426-520)
Average pLDDT score of the middle loop: 72.10
90°
R
Si:ch211-262H13.3 (233-336)
Average pLDDT score of the middle loop: 94.58
90°
S
AF-E9QFL0
180°
T
AF-Q5VZL5
180°
U
CG17341(775-909)
90°
V
PNPase(271-365)
90°
W
NOL4(399-471)
90°
X
TCERG1(638-736)
90°
pLDDT confidence measure
100 to 90 high accuracy expected
90 to 70 backbone expected to be modeled well
70 to 50 low confidence, caution
50 to 0 should not be interpreted, may be disordered

Figure S6. AlphaFold predicts BEN/BEN-like structures with high confidence. Related to Figure 2B-2E, 2H, 3B, 3E-3J, 5F, 5G, 6C, 6F, S2A, S2D, S4A, S4C, S4D, S5A-D and Data S4B-S4W. AlphaFold predicted models presented in this paper are colored with pLDDT scores. The average pLDDT scores of middle loops are calculated accordingly.