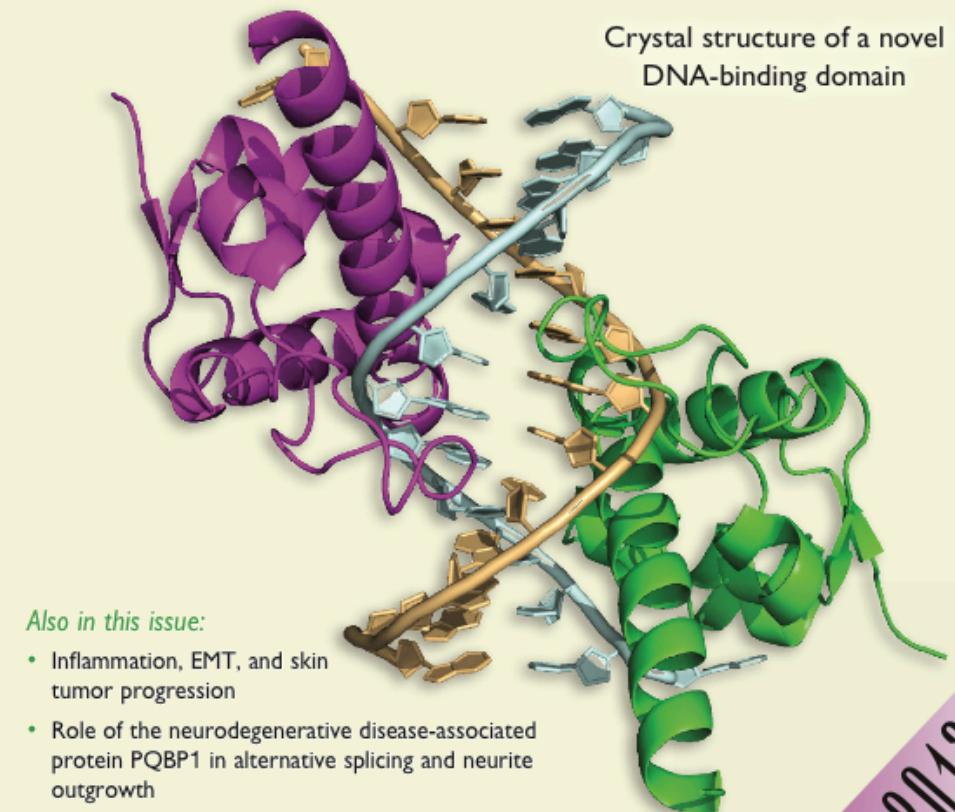


Case study: Finding a new DNA binding domain

Stockholm, November 8 2018

Jakub Orzechowski Westholm
Long-term bioinformatics support
NBIS, SciLifeLab, Stockholm University



Also in this issue:

- Inflammation, EMT, and skin tumor progression
- Role of the neurodegenerative disease-associated protein PQBP1 in alternative splicing and neurite outgrowth



Cold Spring Harbor Laboratory Press

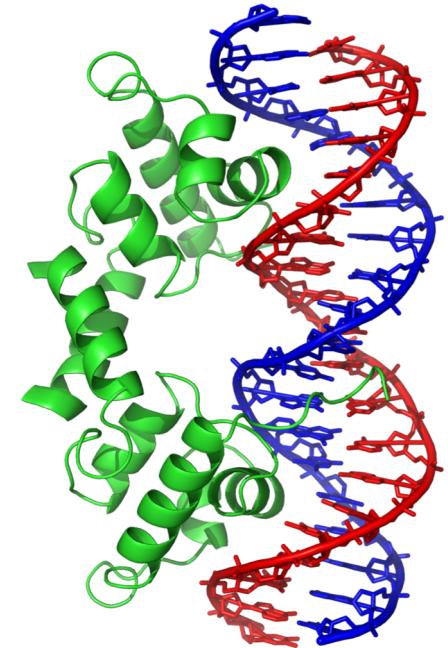
G&D 2013

Transcription factors

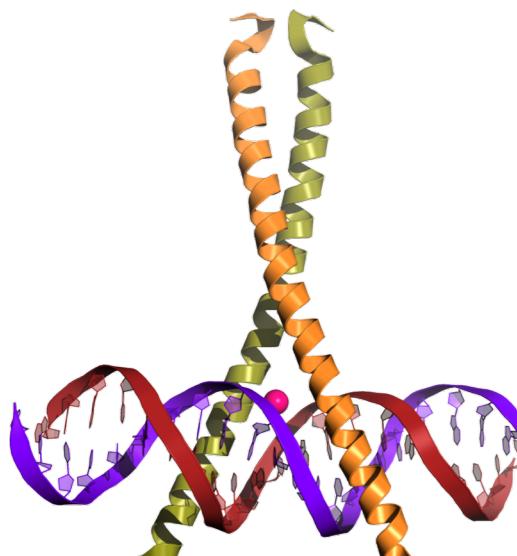
- Transcription factors typically consist of
 - Activation/repression domains
 - A sequence specific DNA binding domain
- The number of such DNA binding domains in eukaryotes is limited:
 - Less than 40 (**Yusuf et al.** *The Transcription Factor Encyclopedia*. Genome Biology 2012)



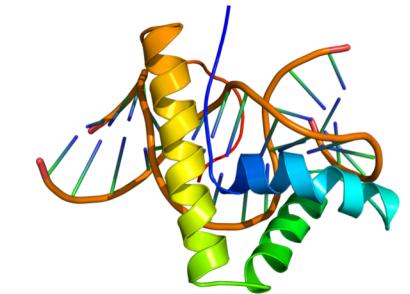
zinc finger



helix-turn-helix



basic leucine zipper



high mobility group box

BEN domains

- Over 100 proteins across animals/metazoans and viruses have BEN domains.

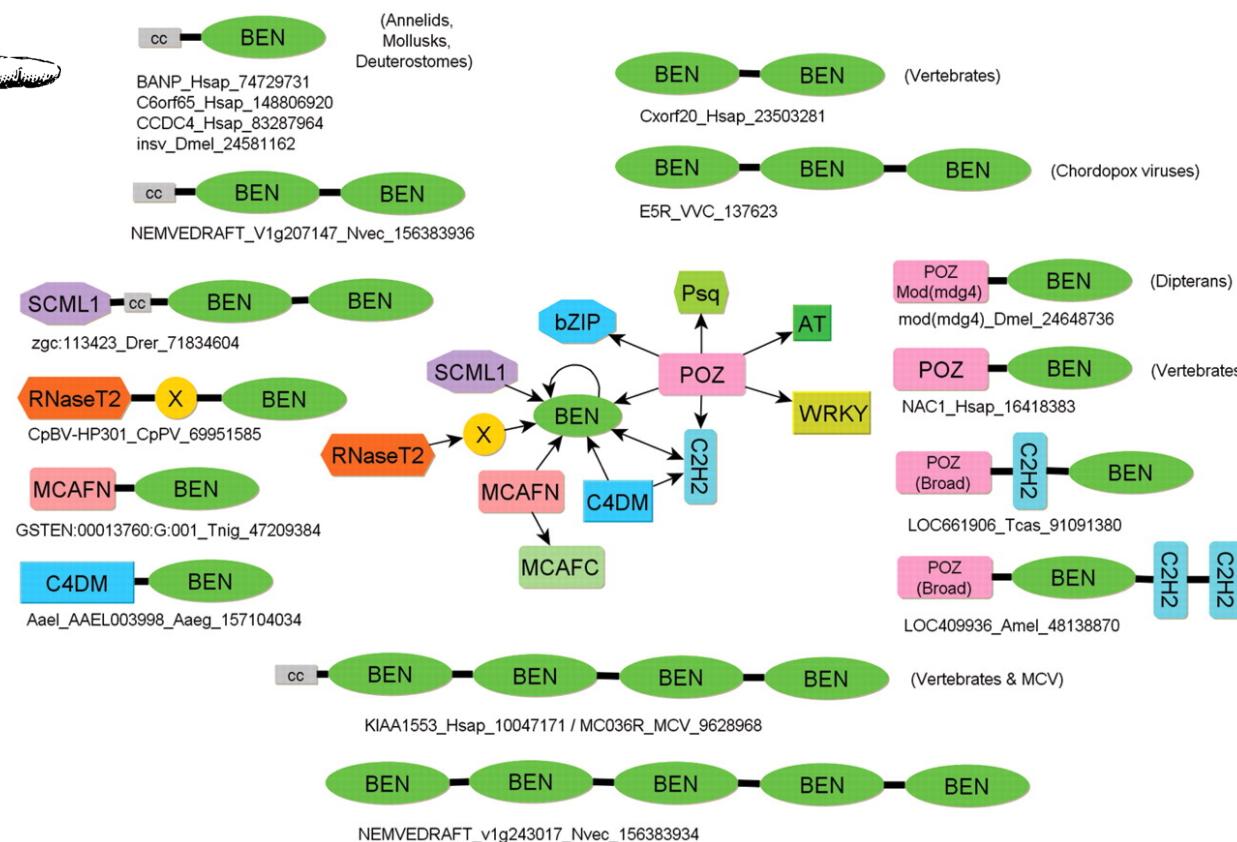
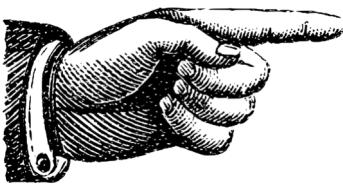
Secondary structure	- HHHHHHHHHH -	- HHHHHHHHHHHH -	- HHHHHHHHHHHHHHHH -	- HHHHHHHHHHHHHHHHHH -	
insv_Dmel_24581162	PNNTCVPAVSFENINNS	VC SLATRKRLVTEIDRETALATH SMTGKPSPLQDKPLKMQDPGKIQDQIFAVTHKCNAE	-- EVRNLATTKCADENKML	259-356\1	
AgaP_ENSANG0000025789_Agam_118791739	SNNLTPKRALEAVRNH	SY KFPTKRLQLMDFLTETLASC SLSRGPCPDRPVKG ALPKVVAIDIVEYVMKKCNVE	-- HVRGVITNPKCADENKML	619-717\1	
LLOC24266_Amel_110759165	GEGIAICEEQLRARVKWS	DY RKLTRKLAAILSPTELATEC SVTGQRWS ERPVPK ALDKAKVAAISYVTSRFPTVD	-- SVKVQVLAYKCKENSTAL	27-126\1	
bsg25A_Dmel_1930012	PNGTEVSRISLSAINWD	MT PSITRKRLCEBIDRDTLAAH TLSGKPSLARPSKQLDPLKVADLVYLMNTSLDMTP	-- EVRTAITTKCADENKML	102-200\1	
Clof165_Hsap_13375807	GSGIVWDEEKWHQLQVT	GD SKYTKRNCAVMINGTDVLKRN SPTVGATK DAVPKP PZSPRKLISIVRECLYDIAQET	-- VDTETAQRLSKVNKYI	133-228\2	
LOC566161_Drer_125823408	GGGIWVDEEKWHQLQRT	GD SKFPTKRNCAVMINGTETLKRN SPTVGATK DALPKP PZSPSKLKIVRECLYDRVSQET	-- ADSAETTORLSKVNKYI	273-368\1	
Aael_AAEI003998_Aaeig_157104034	VRDSLIPYQTMVIDDSV	KY LRPVSKCALALWGHERRLAWS SVTGQRKSNSNSTPSQI QLEPEKFSLIKEVYHRAMQET	-- RVQAMARFDSDRINRLR	364-466\1	
mod(mdg4)_Dmel_24648736	GSRVFVSKVALAKAYIP	MP MIYTCCRMDLWIGKDKL VR IAQHEETT	-- DKDLIQDIIITHCKVFAALRG	441-529>3	
NAC1_Hsap_16418383	GTINVYITRAQLMCHVS	RH KHLVRLRRELASPLDRTNLANS CTTGIRSS NDPRRK PLDSRVLHVAKVYCONFPNPK	-- AVQEFDIDHKLSTLKLMP	374-471\4	
LOC495228_Xlae_148236339	SSGVVITYQQLEDLSHI	KP KLMTRRLDLYD SRETLARS SATGQRIA TMEPKPL RLPDKVTAIAKAVTRACGRGC	-- ENNAIAADMCTNARRVV	374-471\4	
CCDC4_Hsap_83287964	NYPVYIITSQWDEAVNS	KD RRLRLRMYIRFVTTDELKYS CGLGKRRK ETGPER PLDPFKVTLREFIRMHCASNP	-- NFNAVINSKGCTSRRAV	348-446\1	
GSTEN:00029264:G:001_Tnig_47226171	NYPLFITNKQWDEAVNS	KD RRLRLRMYIRFVTTDELKFS CGLGKRRK DSGLER PLNPVVKVSCSLREFIRMHCASNP	-- DWMW PSEEQIINKVPSDAVGH	386-490\5	
LOC560711_Drer_125843107	DYDVFIPKAQOLDSILSN	RS SLLFRNQVCAPEDDTTLANS LPNGKRRK LNDTRK GLDQNIVGAIKVFTEKYCTANG	-- DWMW PSEEQIINKVPSDAVGH	255-359\1	
C10orf30_Hsap_21618768	GFDVFMPSQDLSILSN	RS SLLFRNQVCAPEDDTTLANS LPNGKRRK LNDTRK GLDQNIVGAIKVFTEKYCTANH	-- RDWV QILQDQD1KLAARRRLKRG	267-373\1	
CcBV_3_4_Ccbv_57753424	RTGVVYKRKEKRLCIRE	ND RTLARLRLTEVSQNALSVCTWTGGKAK NDIRP GLDENARMLVLLTFVQHQ-GKK	-- RDWV QILQDQD1KLAARRRLKRG	239-345\1	
CcBVs6gp3_Ccbv_57753417	QRGWVVSYGDLKRYCQQV	KD KSLARRRLLAVENRKALSVCLSIITERAQ GSNARP ELDHHACTVLLNFVLEH-GLQ-	-- CGWSJANTSAVMSTIRTKINDI	1112-1213\1	
GIP_L1_00580_GInd_117935419	QSDIIVYVSYGELEYCQQV	KD KSLARRRLTEVSQNALSVCSMSEKAQA GSNLRP ELDHEKAVLNNFKVIDY-GLQ-	-- RGWN TDIQPILNTHHSKIQBI	1083-1183\1	
MdBV_sBgp1_Mdbv_66391199	HTNIVYINAIKLSNCRKL	KD KSLARRRLTEVSQNALSVCSMSEKAQA GSNLRP ELDHEKAVLNNFKVIDY-GLQ-	-- CGWN TDLKPILDLTHHSKIQDI	995-1055\1	
C6orf65_Hsap_148806920	EKQFQIEKWOIARCN	KD KSLARRRLTEVSQNALSVCSMSEKAQA GSNLRP ELDHEKAVLNNFKVIDY-GLQ-	-- KGWLTDLTSQNSIRNKMQEFS	142-243\1	
LOC794392_Drer_125831142	-YTEFITP-ELLERCRNT	KS QKFINDQMVLYTNEYMATH SLTGAKSS DKAQPK AMNQNQEVEQIIIGVTKQLFPNTD	-- SIRR MIGQKLN-NCTKKPNL	171-270\7	
BANP_Hsap_74729731	GT QKLINDLRLGLEYERELCLASH	SQSGVYVNN RGQPKP ALPTEEVQAILRTWQYFFGKT	-- EIKG YIRQKLQNEAKRLRKKP	202-300\1	
SMAR1_Mmus_10312104	VRCAAIPS-DLHLHISTN	RT EKMALEALDYLQDLYREVQAVS NLSSQGKH	-- GKK QDPPLIYIYICRHLFYKFGITE	255-348\8	
LOC575996_Spur_115728493	VRCAAIPS-DLHLHISTN	RT EKMALEALDYLQDLYREVQAVS NLSSQGKH	-- GKK QDPPLIYIYICRHLFYKFGITE	SDWY RIKQSIDSKCRTAWRRK	237-330\1
Capitella_spi	VRCKINPT-EMVHIMMN	RT DKLALKRLDLYD QKEMQAVS NLSSGTGKH	-- GKK QDPPLIYIYICRHLFYKFGITE	SDWY RIKQSIDSKCRTAWRRK	278-371\1
NEMVEDRAFT_vlg232490_Nvec_156390312	VRVPITPS-DLLHHSN	RT EKMALEALDYLQDLYREVQAVS NLSSGMGKH	-- GKK QDPPLIYIYICRHLFYKFGITE	EDWY RIRQNIDSKCRTAFRRK	228-321\1
CXorf20_Hsap_23503281_2	PHISDAELQSLRDEKRP	KP ENLAVWVRLRLTQREREGR TVCGF	-- GGS GLDNNDVQSDIRRYFYRALPDPP	QDWH RIKQNIIDSKCRTAFRRK	285-375\1
LOC100003955_Drer_125851480	WRNIRMPC-SVLTLAKT	KP ENLAVWVRLRLTQREREGR TVCGF	-- GGS GLDNNDVQSDIRRYFYRALPDPP	DWKW QCISAMNSYLRGTRRKR	255-348\8
zgc:113423_Drer_71834604_1	LS-SLARYLTIQKELTKDVLQVS NVYGNLKNH	-- GLC ALDNPKNASLREFLQENPYICD	-- SDWY RIKQSIDSKCRTAWRRK	235-348\8	
LOC764357_Spur_115613065	LRKVWIPI-CVYKEVFK	ET QKAVAPVLYST PISTLSCS AVTGNPEK	-- GKK QDPPLIYIYICRHLFYKFGITE	SDWY RIKQSIDSKCRTAWRRK	237-330\1
NEMVEDRAFT_vlg243017_Nvec_156383934_4	ERKVIFFS-FIQLRAGK	ET QKAVAPVLYST PISTLSCS AVTGNPEK	-- GKK QDPPLIYIYICRHLFYKFGITE	SDWY RIKQSIDSKCRTAWRRK	278-371\1
KIAA1553_Hsap_10047171_1	MA-SAAVRYLSRNIDTKELESQS	TP MA-SAAVRYLSRNIDTKELESQS	-- GKK QDPPLIYIYICRHLFYKFGITE	EDWY RIRQNIDSKCRTAFRRK	228-321\1
KIAA1553_Hsap_10047171_2	RIQMVMQDSRWEEMPT	GA R-LAIALARYCQGK TGTILIRS SFTGRN	-- GKK QDPPLIYIYICRHLFYKFGITE	QDWH RIKQNIIDSKCRTAFRRK	285-375\1
KIAA1553_Hsap_10047171_3	YQDVTPLDPEFRQITV	EI SNAYAVAVLRLRPEDELEERA-AAGE	-- GKK QDPPLIYIYICRHLFYKFGITE	DWKW QCISAMNSYLRGTRRKR	255-348\8
KIAA1553_Hsap_10047171_4	PPEYQLTAAELKQIVDQ	LS GDLACRLVQL PELFSDV DFSGRCSA	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
GSTEN:00016974:G:001_Tnig_47220120_1	ASDHWVTDQDTEFLDE	SS GDFAVFLHLRL PELFSDH KLGEBQYSC	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
NEMVEDRAFT_vlg243810_Nvec_156379688_1	GADCLLSSKEQIERSIYES	GS GNPASRLVHL PELFTHE-NLRLQKYN	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
LOC584784_Spur_115651987	PSPYLLSDRKEVIVQQ	LS GNPASRLVHL PELFTHE-NLRLQKYN	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
MC036R_MCV_9628968_1	PQEYLLSRSQEQLRNYIYC	LS GNPASRLVHL PELFTHE-NLRLQKYN	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
MC036R_MCV_9628968_2	RPQFASRS-AVQOIKC	LS GNPASRLVHL PELFTHE-NLRLQKYN	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
MC036R_MCV_9628968_3	KS-GNSFVQCLRLYQGEELSNK	LS GNPASRLVHL PELFTHE-NLRLQKYN	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
MC036R_MCV_9628968_4	NGKSGT	-- GFAIDPFLQIKYQTEVYHPI	-- GFAAKR KLESLHLQIIRNYVEVYYPSVK	SDWY RIKQSIDSKCRTAWRRK	237-330\1
E5R_VVC_137623_2	ANEKLSMGVICHLYERA	KG -NPARSVLRKLVDDDLVKS TCGSKGR	-- EQ AIDPDLQYALEBTTYDVYGVEE	KCRA BCVQSIDSHCRQLFNSQ	323-421\1
E5R_VVC_137623_3	CS-ADMARVLLRLYPEV	-- VCGADSE	-- EQ AIDPDLQYALEBTTYDVYGVEE	KCRA BCVQSIDSHCRQLFNSQ	323-421\1
GSTEN:00013760:G:001_Tnig_47209384	ALEMIPSPAEELCHLAH	CS-ADMARVLLRLYPEV	-- EQ AIDPDLQYALEBTTYDVYGVEE	KCRA BCVQSIDSHCRQLFNSQ	323-421\1
xpat-A_Xlae_14822226	PAWAGPVTLIDYECA	SV GELAVLRLHKLQ PQLFDA	-- VCGADSE	YVWQ EGGLPLREFLRLRCLVR	18-107\1
Daphnia_pulex	SCVULPTRAHLRKM	AS YNFPAVRLVLYM PELFTA	-- VCGADSE	YVWQ EGGLPLREFLRLRCLVR	634-734\1
Branchiostoma_floridae	PAQYLIISAKRVEKELAR	AS GHPFAQCTVMDL PELFSS	-- VCGADSE	YVWQ EGGLPLREFLRLRCLVR	807-907\1
Consensus/80%	NQRTYKLFSDISAIGK	AS SKMVKYALLYM PNLPGD	-- VCGADSE	YVWQ EGGLPLREFLRLRCLVR	935-1035\1
IKGKSEED-TLPFIKQMV	IKGKSEED-TLPFIKQMV	VT QELVEKVLKIL RDLFKS	-- GEYKAYRY VENGPFI GLDTLK-LNIVHDVVEPCMPV	PVAK LCKEMVNKYFENPLHII	218-318\1
LRKINSHMEKILFENCK	GV DRYASTYFVRYLFWPYNKYCEW TVKVN	-- GEYKAYRY VENGPFI GLDTLK-LNIVHDVVEPCMPV	-- HWRH EIRDANIEILRVRKRP	221-322\1	
LPLDIILNPLDGGKVLKLSM	GLMGKE ALPNTVRNMRALYIERFRTLS	-- GEYKAYRY VENGPFI GLDTLK-LNIVHDVVEPCMPV	-- HWRH EIRDANIEILRVRKRP	187-287\1	
HMSSEDL-DYCNMMAR	AS GSRGKL GPLPRNL MIDLHQTSKRFV	-- GEYKAYRY VENGPFI GLDTLK-LNIVHDVVEPCMPV	-- KEKR KIKTRNLLLRTRQDRDA	187-287\1	
NN-TMISLMQGKTV	LG	-- RTTKP ALPVVDKVNAAVAKYILKRPDKH	-- EFNQKVTVNYLDRQAAKS	266-354\1	
EQGVVTTYPVILAQAKNK	KT EQFFKIDMGIY	-- RTTKP ALPVVDKVNAAVAKYILKRPDKH	-- KLYRVVNEKGRMRATL	140-229\1	
.....ph..h....	.s...h...Lh...hFsp...b...p...	-- GTHQ ALSPAIISAILTETTKQYGGVQ	-- .Ls...h..lb..h...s...l...h.p....ph..h....	

“Prediction of the secondary structure using the multiple alignment indicated an all α -fold, with four conserved helices.”

Abhiman et al. *BEN: A novel domain in chromatin factors and DNA viral proteins*. 2008, Bioinformatics

BEN domains, cont.

- The BEN domain sometimes co-occurs with chromatin remodeling domains (e.g for histone deacetylation).

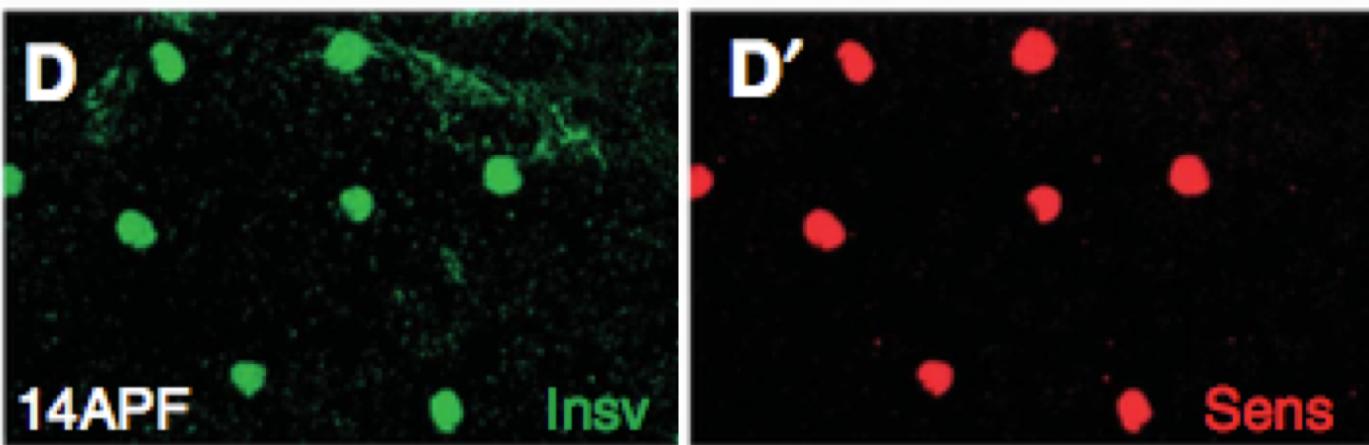


Insensitive protein

- We studied *Insensitive*, a *Drosophila* protein with a single BEN domain.
- *Insensitive* shows nuclear expression in the peripheral nervous system, and is involved in Notch signalling.
- *Insensitive* is expressed ubiquitously in the early embryo and later throughout the developing ectoderm but becomes highly restricted to the developing CNS and PNS. Peak expression at 2-4 hours.

Insensitive protein, cont.

- Previous studies suggested that *Insensitive* was a co-factor of a TF called *Suppressor of hairless*.
- We wanted to see where *Insensitive* bound to DNA, and determine possible targets.
- ChIP-seq from fly embryos, from two time points.
- IgG as control.



Duan et al. *Insensitive is a corepressor for Suppressor of Hairless and regulates Notch signalling during neural development*. 2011, EMBO J

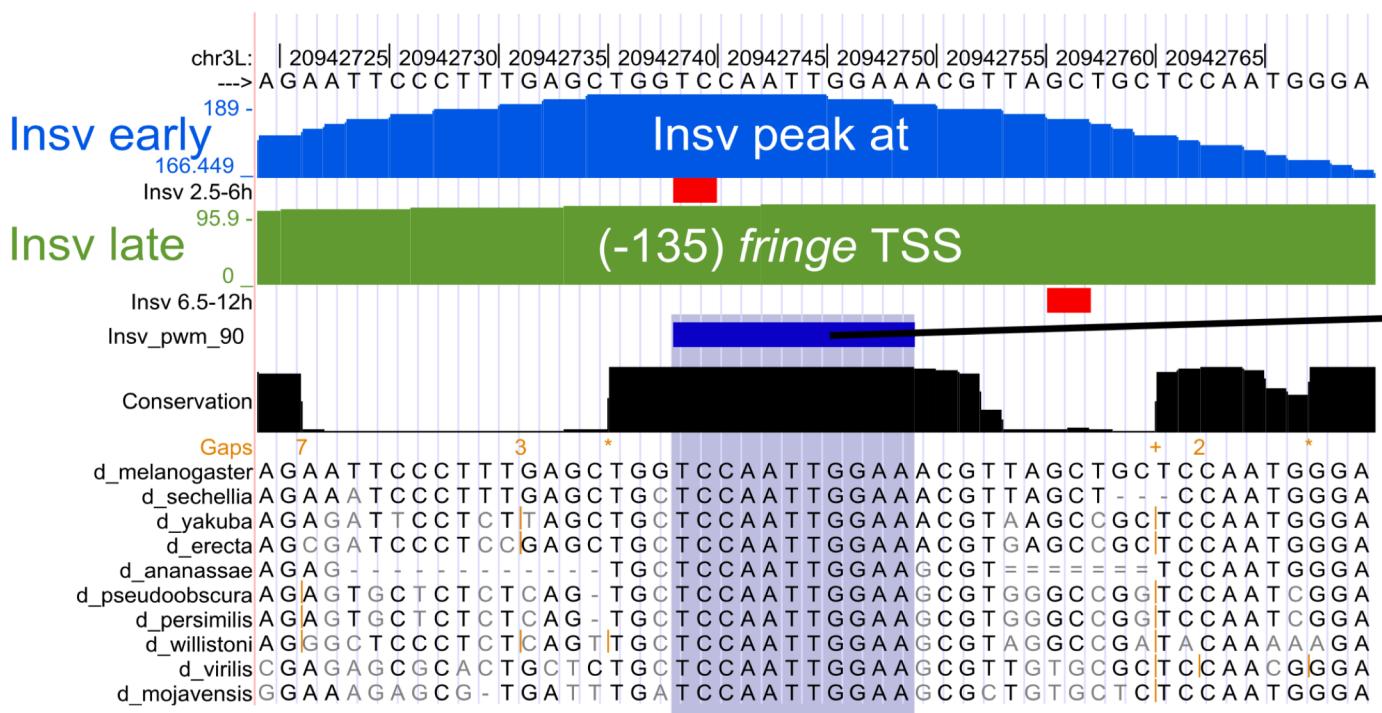
ChIP-seq experiment

- Analysis:
 - FastQC
 - Mapping: Bowtie
 - QC: Phantompeakqualtools
 - Peak calling: Quest (Valouev et al. *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data*. Nature methods, 2008)
 - Peak annotation: chippeakanno
 - Motif finding: MEME, Weeder
 - Custom scripts..

AB	Time	Unique reads mapping	Nr peaks
Insv	2.5-6h	7,473,521 (58%)	5364
Insv	6.5-12h	4,292,248 (61%)	2390

Insenstive seems to bind to a new motif

We were expecting to find the *Suppressor of Hairless* motif, but instead found a new site.

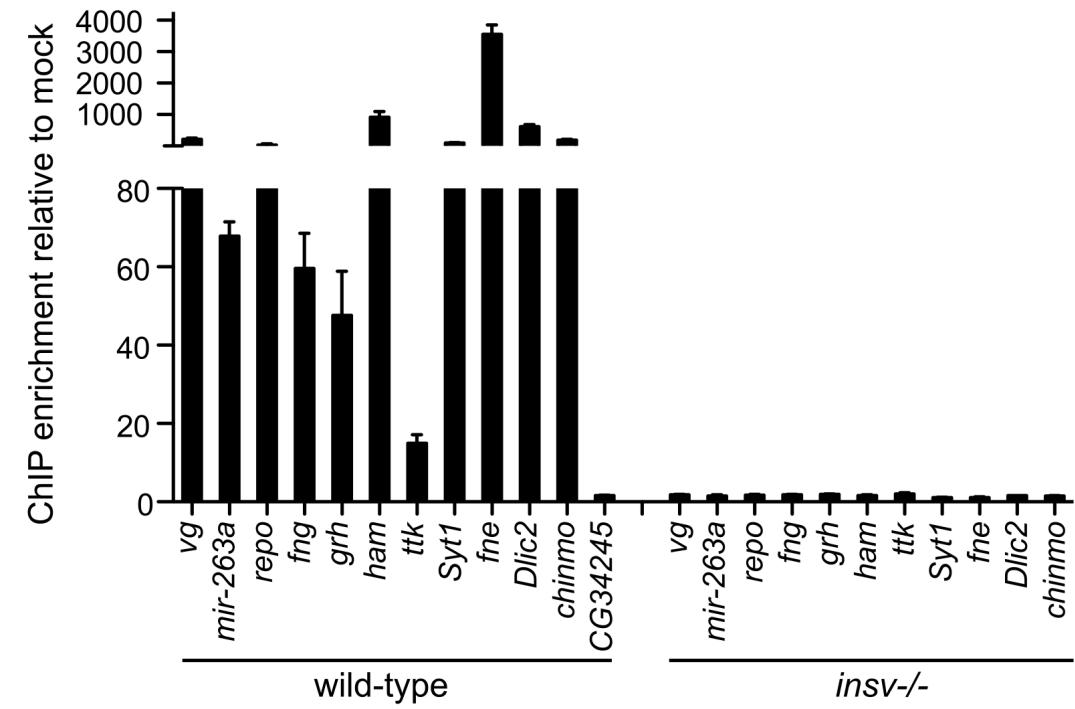
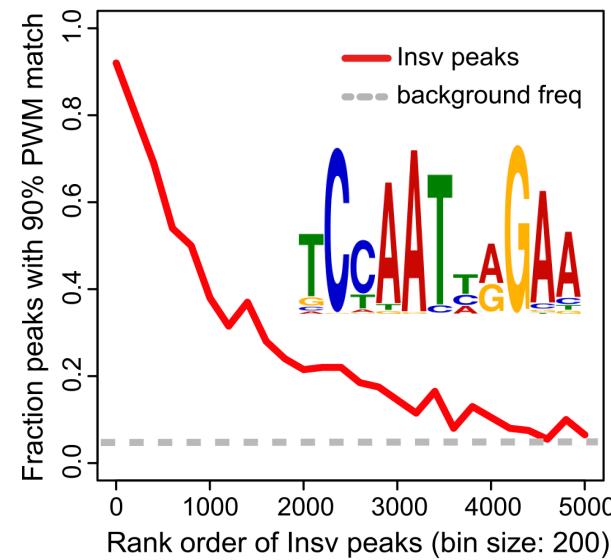
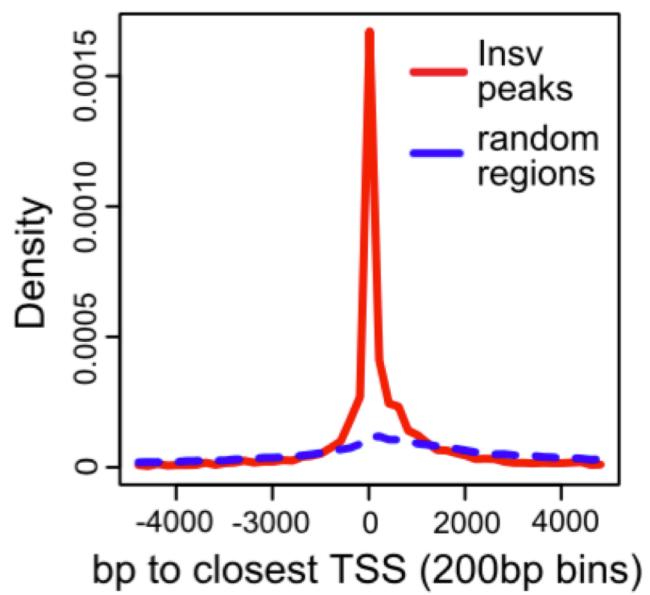


Weeder:

MEME:

CCAAATTGG
CCTCAATGAA
p-value = 6.4E-562

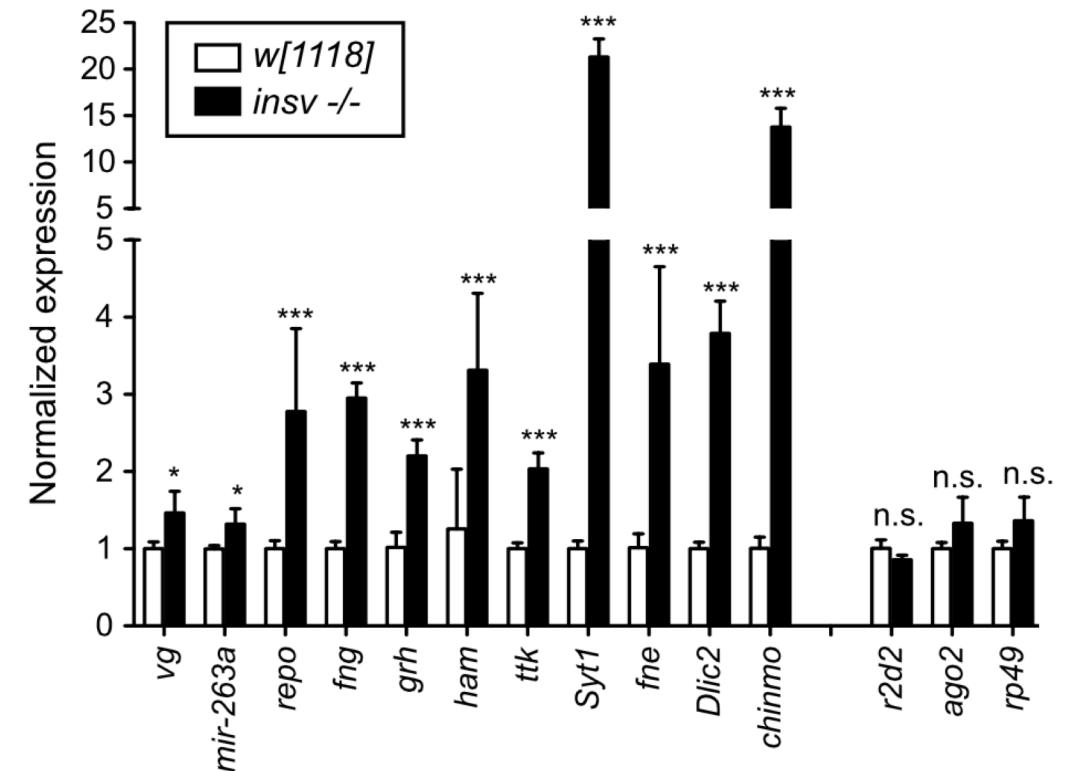
Validating peaks



- *Insenstive* peaks are located at promotor regions
- Almost all the top *Insenstive* sites have the motif.
- ChIP-PCR validation of some peaks.

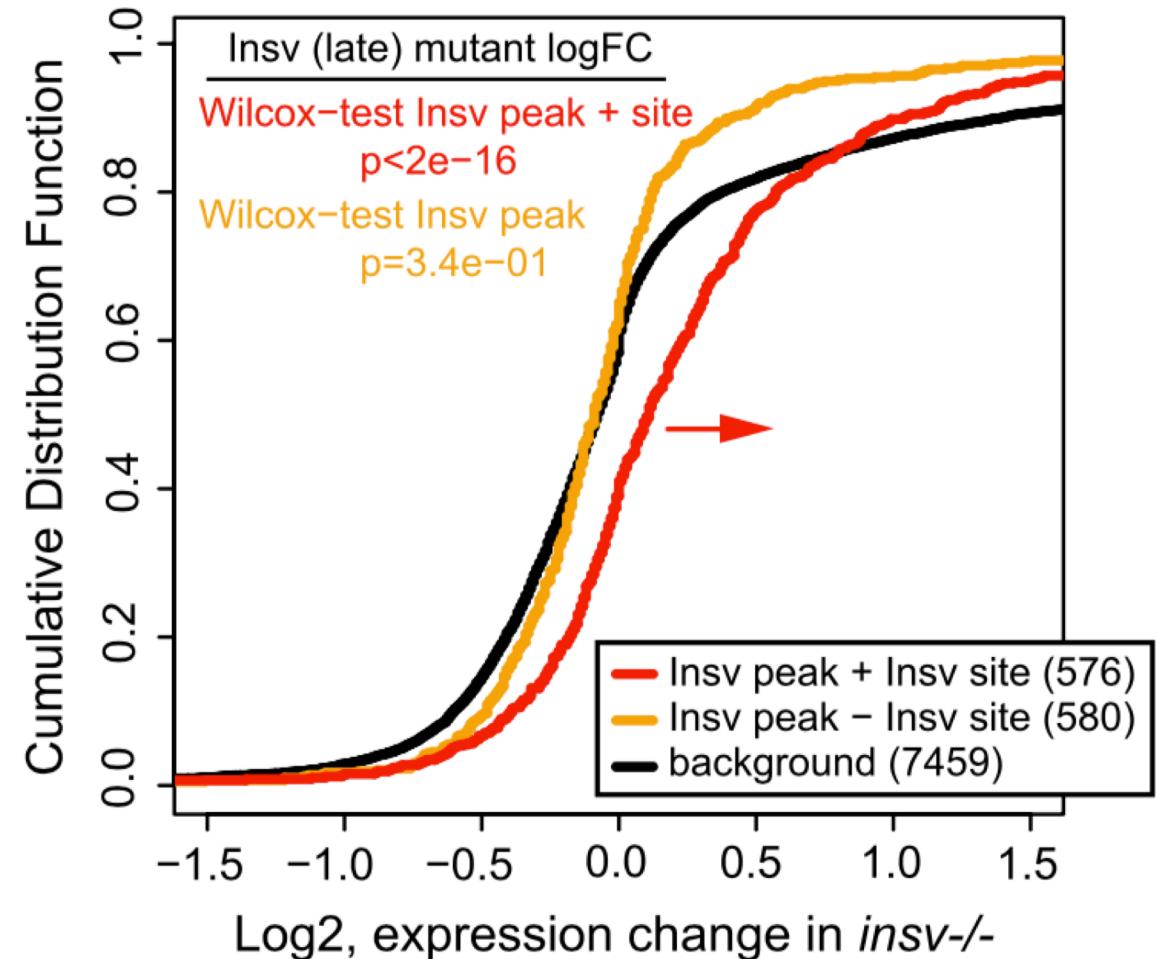
Gene expression

- rt-qPCR on selected genes → genes near Insensitive peaks have increased expression in an Insensitive mutant.



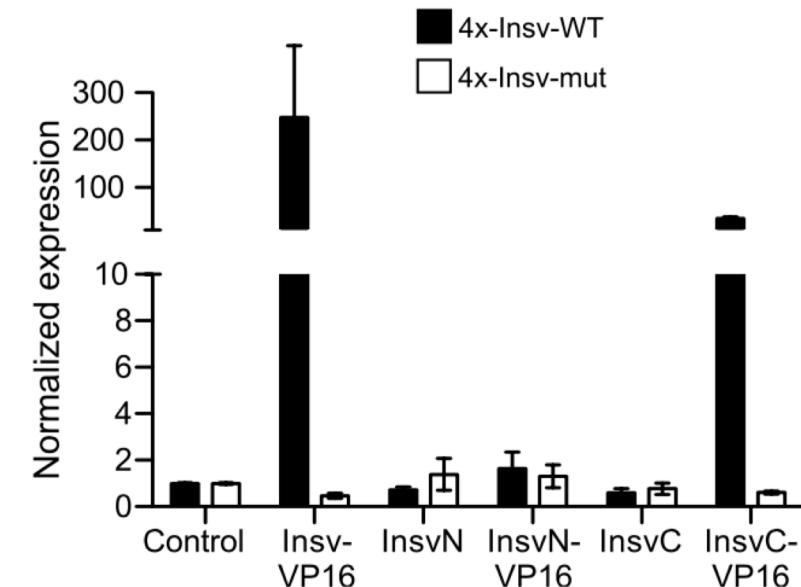
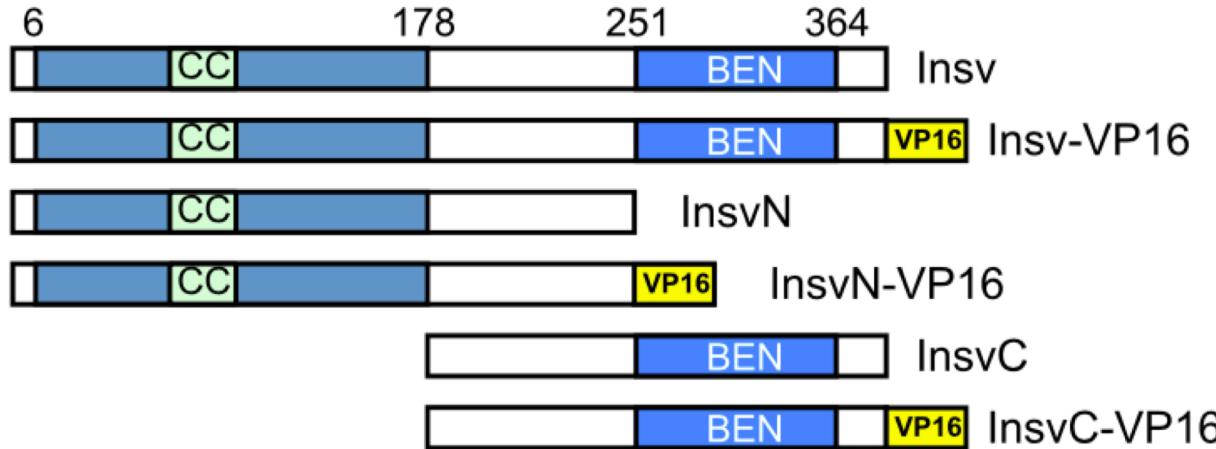
Gene expression, cont.

- We also looked at gene expression on a genome-wide scale.
- Genes near Insensitive peaks, that have an Insensitive site, have overall increased expression in an Insensitive mutant.

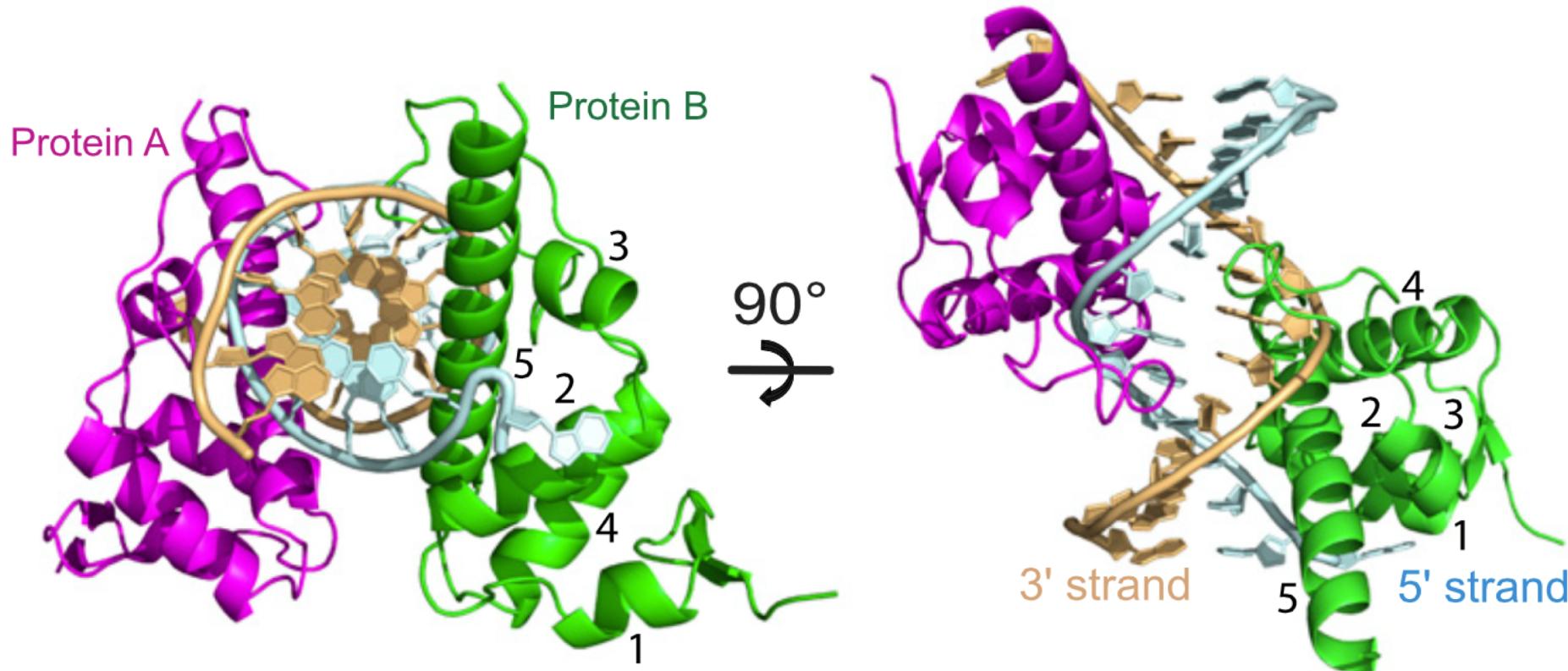


Structure-function experiments

- Actin-luciferase as read-out.
- 4 Insensitive sites in promoter or 4 mutated Insensitive sites
- Different parts of Insensitive, sometimes fused to the VP16 activation domain.
- → the (C-terminal) BEN domain is necessary and sufficient for binding to the Insensitive site.



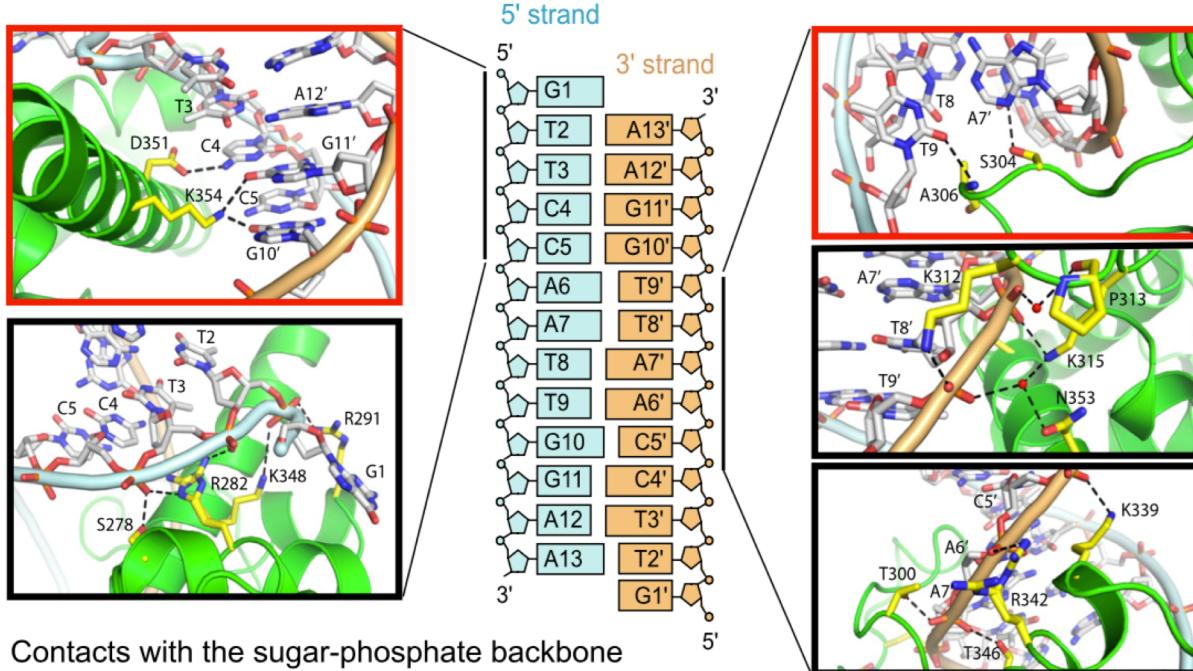
Crystal structure of BEN domain bound to DNA



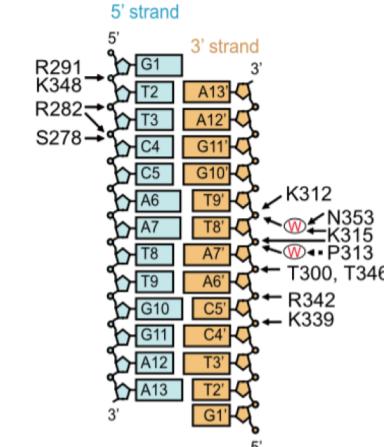
Validating the structure

- From the structure, we can see which amino acids make contact with which nucleotides.
- We can make predictions about how amino acid and DNA mutations will affect binding, and test these predictions.

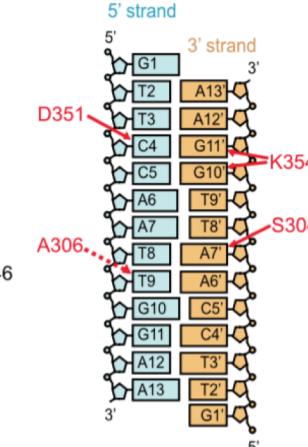
Base-specific hydrogen-bonding contacts



A BEN contacts with sugar-phosphate backbone



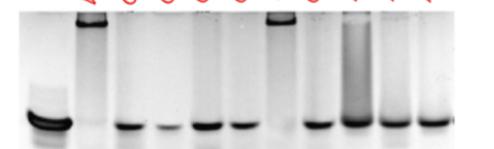
B Base-specific contacts of BEN domain



C Insv-BEN variants tested on wt probe

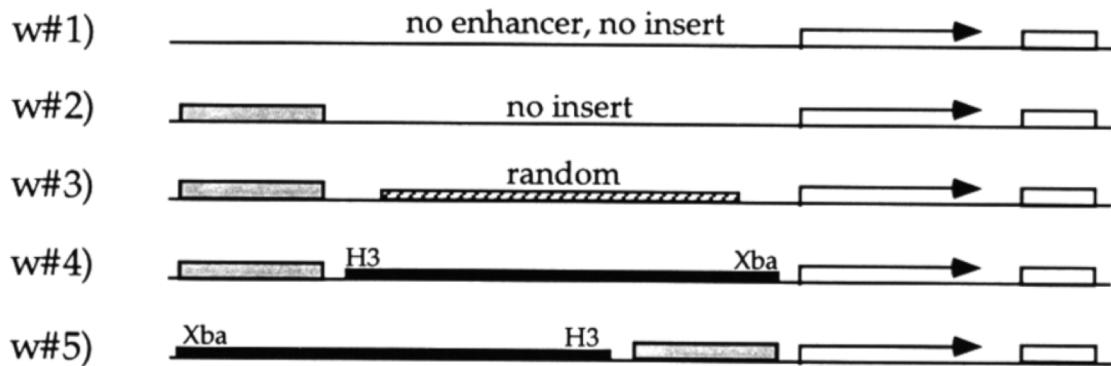


D WT Insv-BEN on variant probes



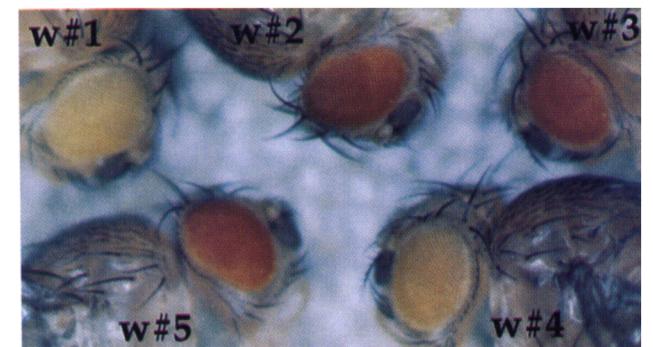
Insulator elements

- Insulator elements were first described as DNA elements that can restrict e.g. interactions between enhancers and target genes or the spread of heterochromatin.



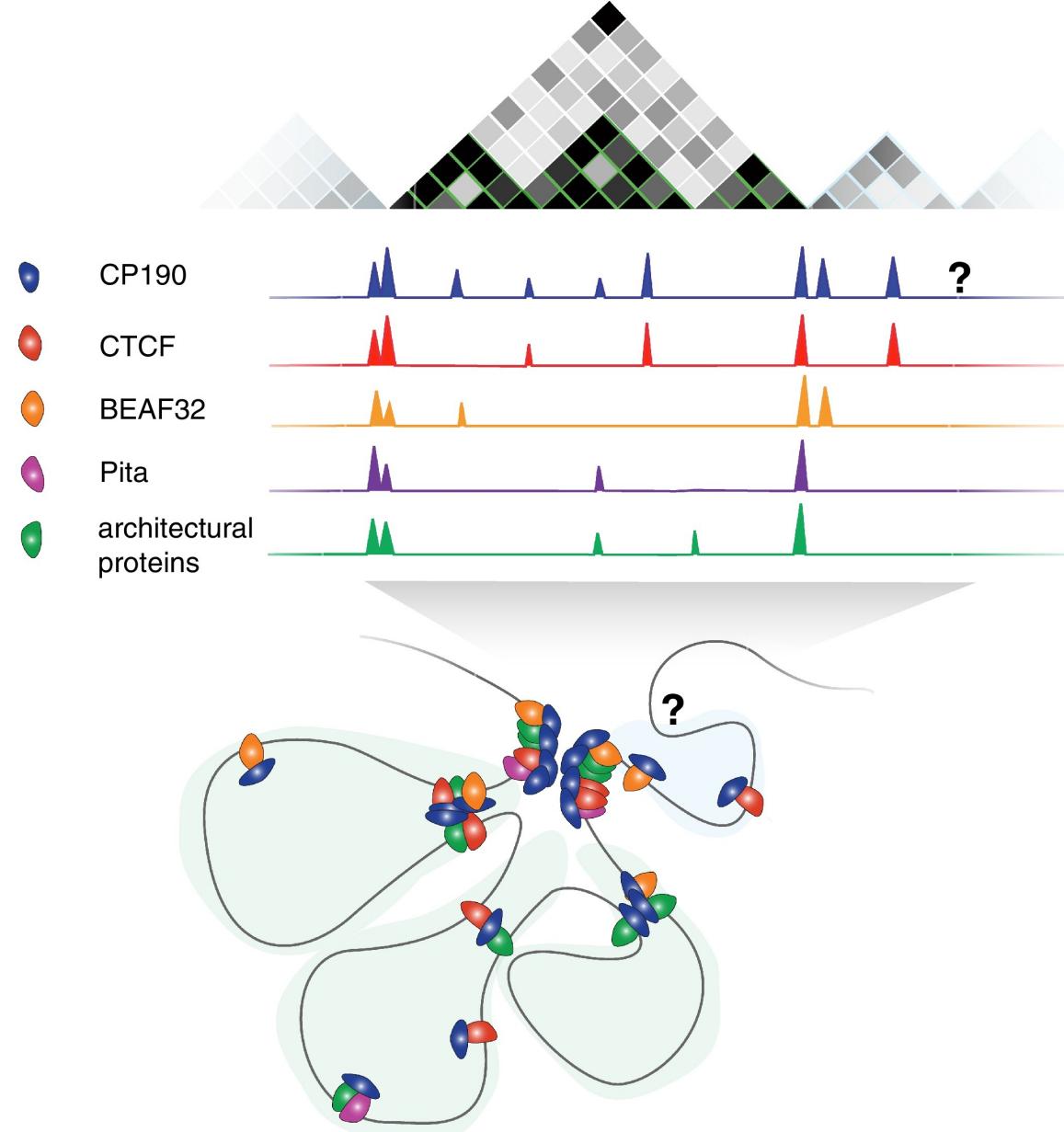
Enhancer blocking
(#light eyes/total)

—	(5/5)
—	(0/6)
—	(0/5)
+	(12/24)
—	(1/11)



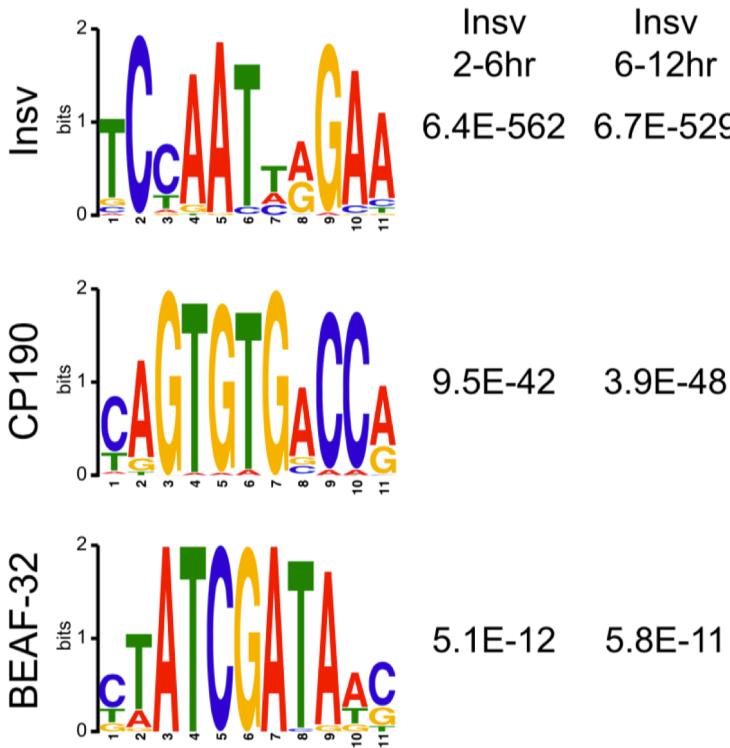
Insulator elements, cont.

- Insulator elements control DNA looping.
- Enhancers and target genes can end up in different loop domains (\approx topologically associated domains, TADs)



Ali et al. *Insulators and domains of gene expression*.
Current Opinion in Genetics & Development, 2016.

Insensitive binds at insulator elements

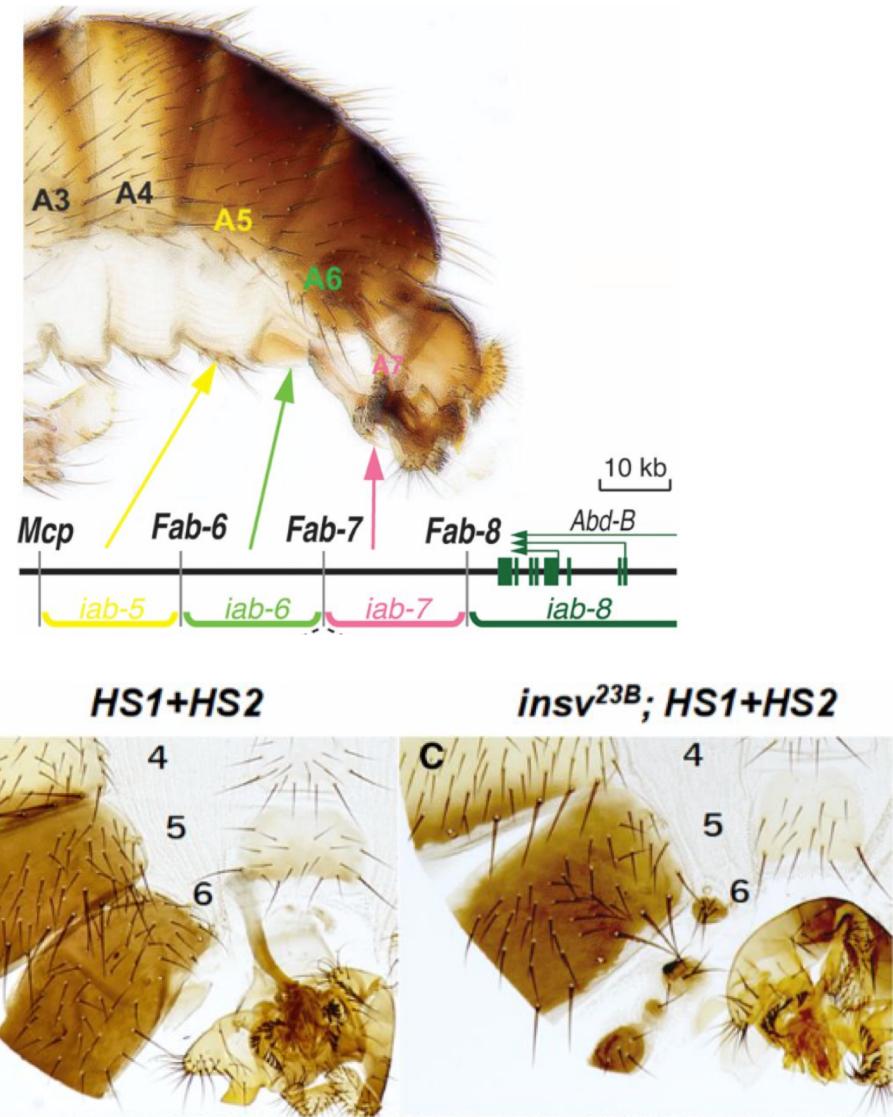
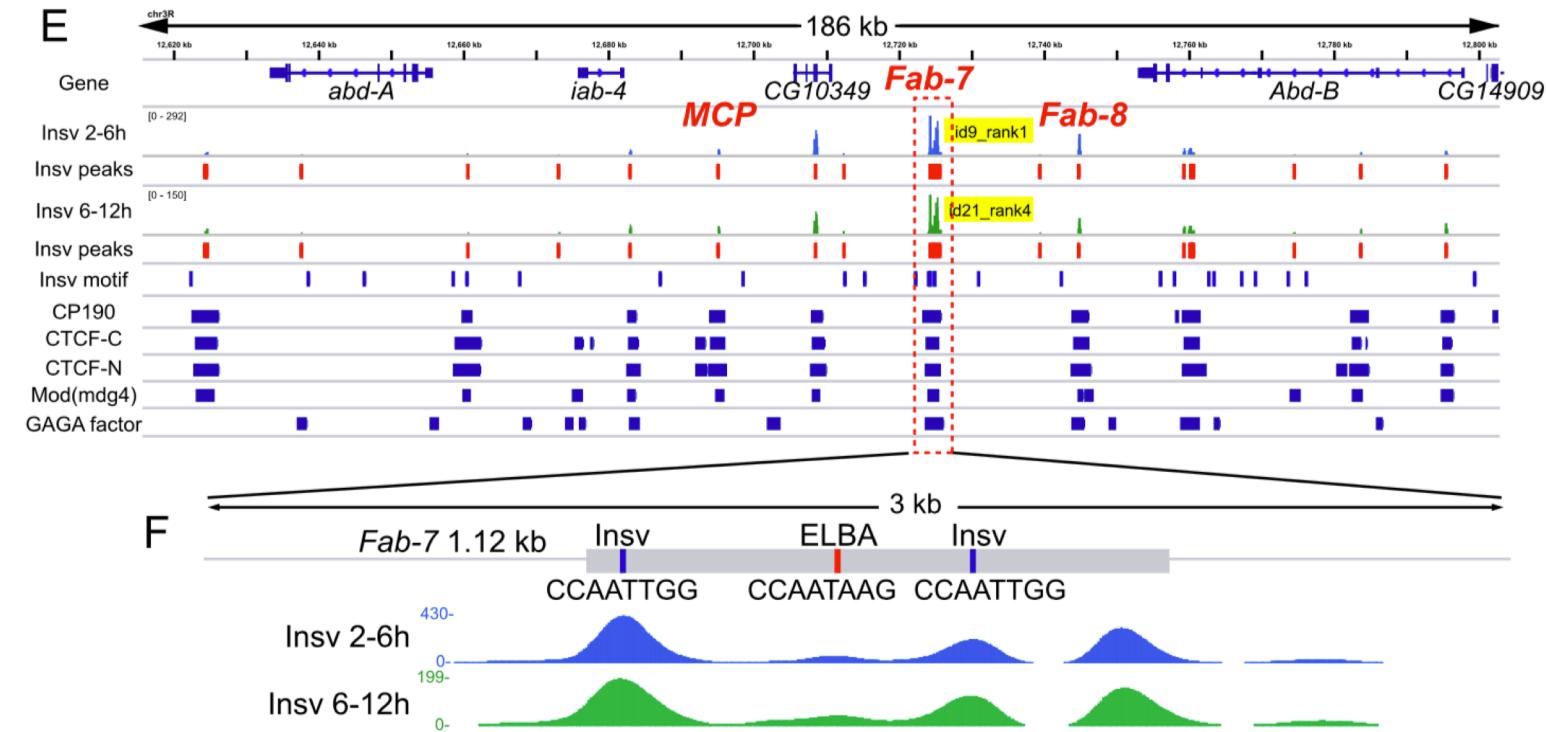


		# fraction of peaks in row data (# peaks in genome)										Insulators		Control TFs	
		Insv 2-6hr (4197)	Insv 6-12hr (1870)	BEAF-32 (4709)	CP190 (6652)	CTCF_C (3154)	CTCF_N (2532)	GAF (3904)	Mod(mdg4) (3060)	Su(hw)_1 (3420)	Su(hw)_2 (3630)	CtBP (4946)	Gro (1337)		
Insv	2-6hr (4197)	1	0.41	0.6	0.72	0.38	0.34	0.36	0.36	0.11	0.11	0.08	0.02		
Insv	6-12hr (1870)	0.95	1	0.6	0.74	0.43	0.4	0.4	0.43	0.11	0.12	0.06	0.02		
BEAF-32	(4709)	0.4	0.19	1	0.81	0.38	0.29	0.28	0.29	0.08	0.08	0.19	0.02		
CP190	(6652)	0.35	0.17	0.6	1	0.38	0.3	0.24	0.4	0.29	0.3	0.15	0.02		
CTCF_C	(3154)	0.42	0.22	0.59	0.82	1	0.79	0.28	0.46	0.17	0.16	0.12	0.02		
CTCF_N	(2532)	0.45	0.25	0.55	0.8	0.98	1	0.29	0.5	0.18	0.18	0.14	0.02		
GAF	(3904)	0.28	0.15	0.34	0.4	0.23	0.19	1	0.26	0.12	0.14	0.14	0.06		
Mod(mdg4)	(3060)	0.4	0.22	0.45	0.89	0.48	0.42	0.33	1	0.42	0.43	0.12	0.02		
Su(hw)_1	(3420)	0.11	0.05	0.11	0.56	0.15	0.13	0.14	0.37	1	0.95	0.09	0.02		
Su(hw)_2	(3630)	0.1	0.05	0.1	0.54	0.14	0.12	0.14	0.35	0.88	1	0.09	0.02		
CtBP	(4946)	0.06	0.02	0.19	0.21	0.08	0.08	0.11	0.08	0.06	0.07	1	0.03		
Gro	(1337)	0.06	0.02	0.08	0.09	0.04	0.04	0.2	0.05	0.04	0.06	0.11	1		

- Insensitive peaks are enriched for C190 and BEAF-32 motifs
- Insensitive peaks overlap C190, BEAF-32 and CTCF peaks

Dai et al. *Common and distinct DNA-binding and regulatory activities of the BEN-solo transcription factor family*. Genes & Development, 2015.

Insensitive binding at the Fab-7 insulator



BEN domain protein function

- Insulators:
 - Elba1, Elba2, Elba3 (Aoki et al. *Elba, a novel developmentally regulated chromatin boundary factor is a hetero-tripartite DNA binding complex*. eLife, 2012)
- TFs:
 - BEND5 (Dai et al. *The BEN domain is a novel sequence-specific DNA-binding domain conserved in neural transcriptional repressors*. Genes Dev. 2013)
 - BEND6 (Dai. et al. *BEND6 is a nuclear antagonist of Notch signaling during self-renewal of neural stem cells*. Development, 2013)
- Chromatin remodelers:
 - BEND3 involved in heterochromatin formation (Saksouk et al. *Redundant Mechanisms to Form Silent Chromatin at Pericentromeric Regions Rely on BEND3 and DNA Methylation*. Mol Cell, 2014)
- Chromatin component?
 - Elba2 (Xu et al. *BEN domain protein Elba2 can functionally substitute for linker histone H1 in Drosophila in vivo*. Scientific Reports, 2016)

Some conclusions

- The BEN domain is a new DNA binding domain.
 - Gene annotation: clues about the function of over 100 genes with the BEN domain:
 - Transcription factors
 - Chromatin remodelers
 - insulator proteins etc.
- Insensitive is a transcriptional repressor
- Insensitive (and other BEN-proteins) have insulator activity.
- ChIP-seq was one (but important) method in this story

Acknowledgements

Eric Lai (Sloan-Kettering)

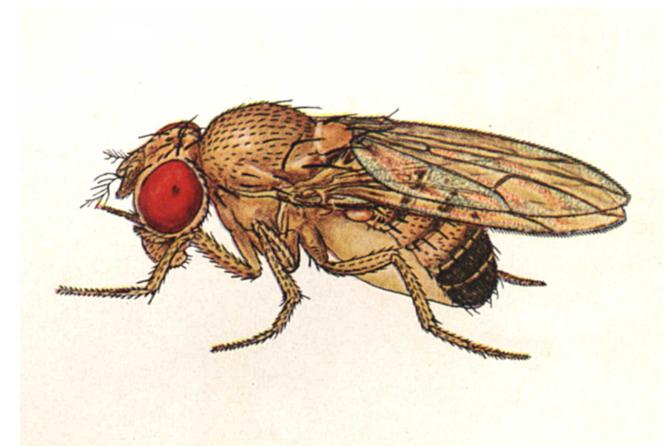
Qi Dai

Hong Duan

Dinshaw Patel (Sloan-Kettering)

Aiming Ren

Artem Serganov



The End