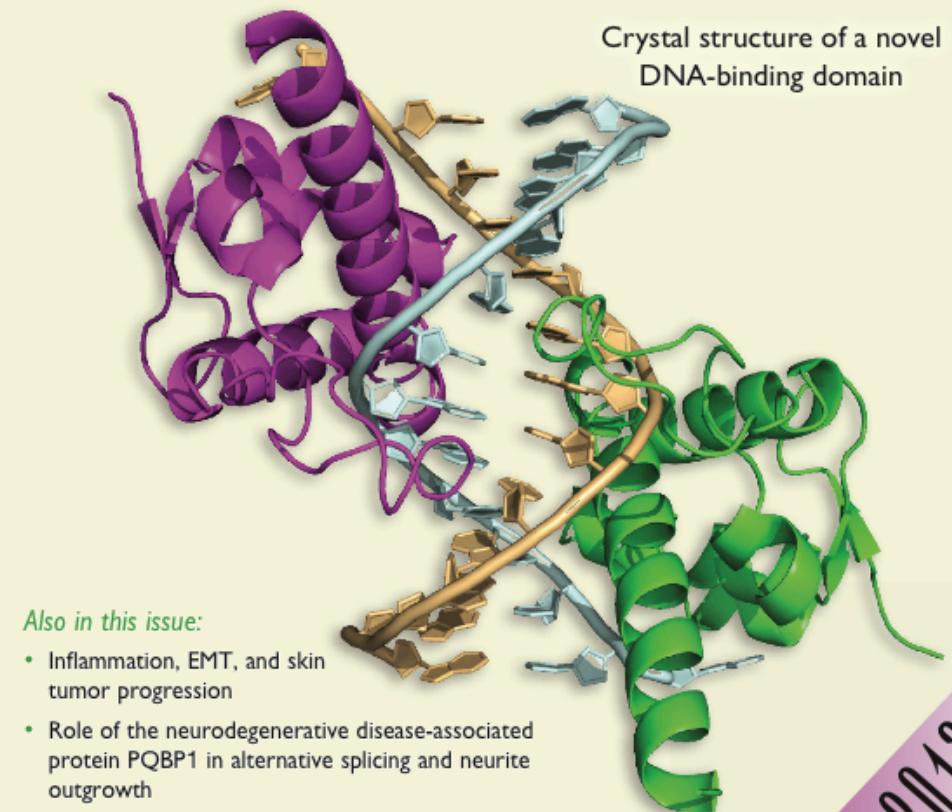


Case study: Finding a new DNA binding domain

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

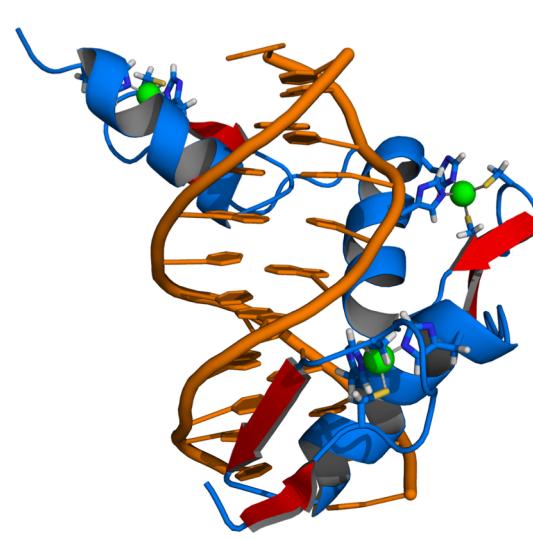


Cold Spring Harbor Laboratory Press

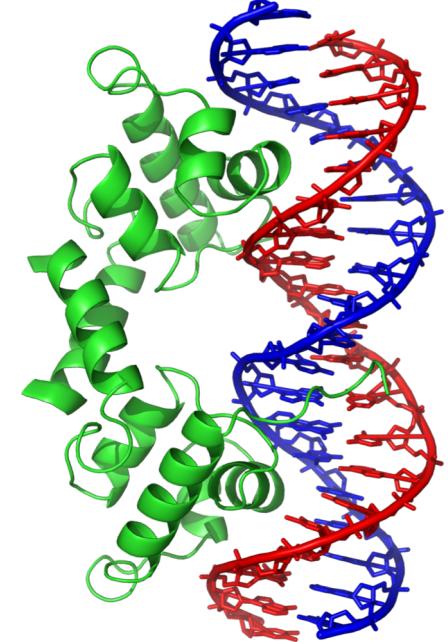
G&D 2013

Transcription factors

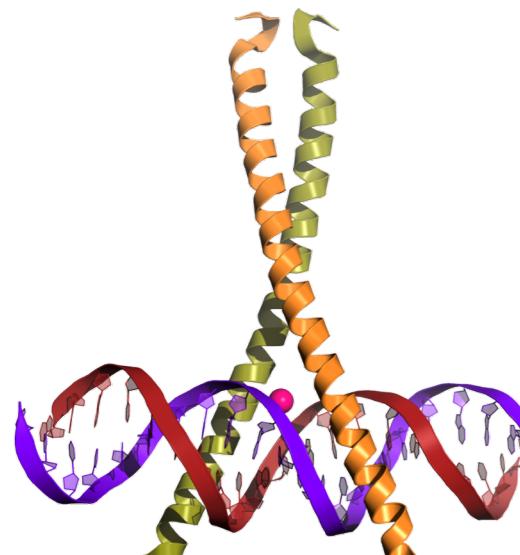
- Transcription factors typically consist of
 - A sequence specific DNA binding domain
 - Activation/repression domains
- The number of such DNA binding domains in eukaryotes is limited:
 - Less than 40 (**Yusuf et al.** *The Transcription Factor Encyclopedia*. Genome Biology 2012)



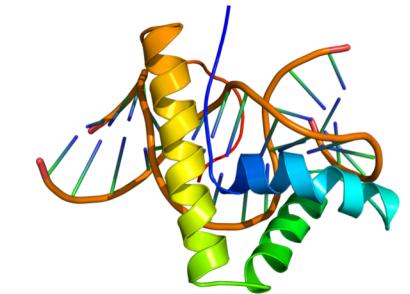
zinc finger



helix-turn-helix



basic leucine zipper



high mobility group box

BEN domains

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

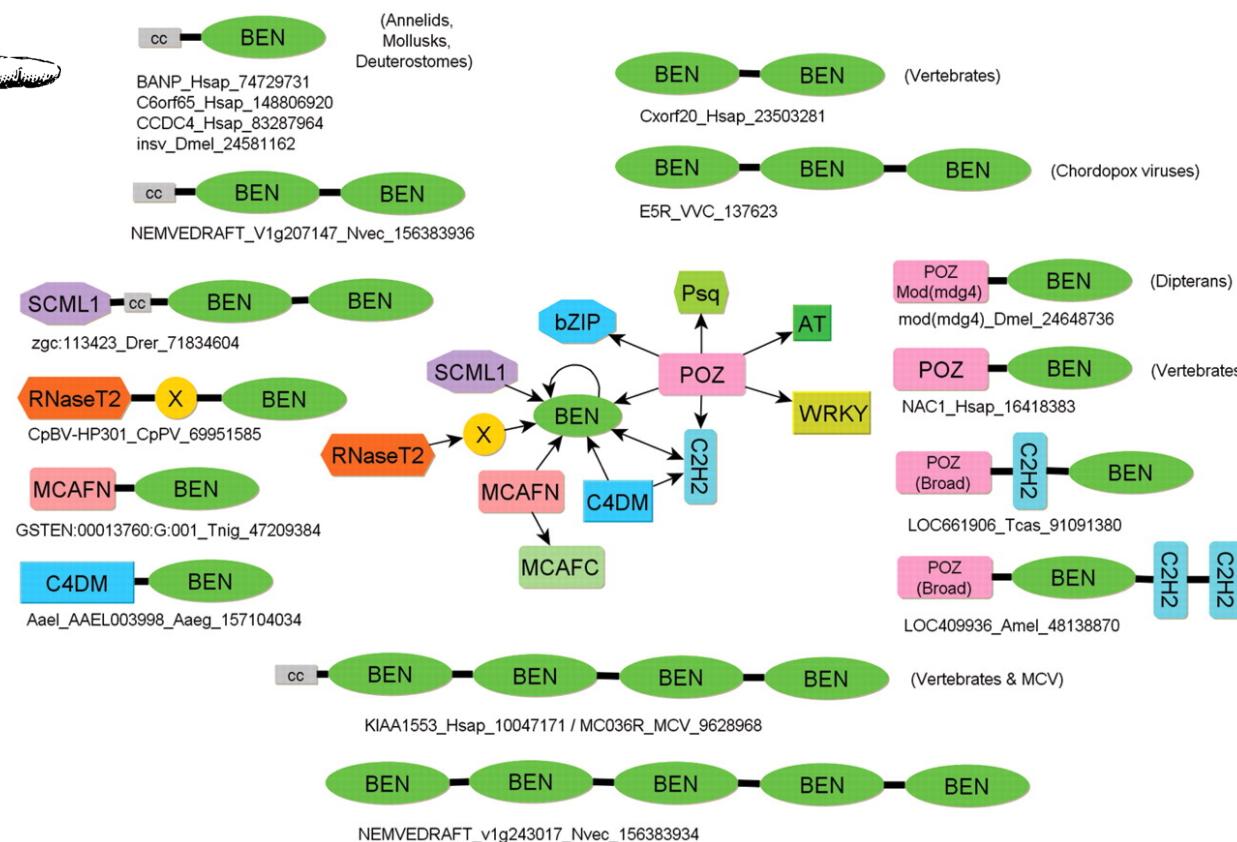
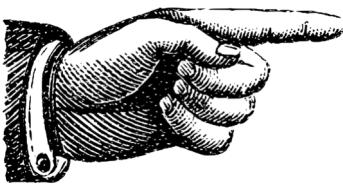
Secondary structure	----- HHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----
insv_Dmel_24581162	PNNTICVPASV FENINWS VC	SLATRK LCLVTFIDRET LATH SMTGKPSP--QDKPLKMQDPGKIQDIX IFAV THKCNASE	-----	EVRNATT KCADENKML 259-356\1
AgaP_ENSANOG0000025789_Agam_118791739	SNNLTVPKRALEAVRWH SY	KFGTRKLQLM TRET LASC SLSRGPCP--DRPVKG AOPVKVAD IVYEV MKKCNVEE	HVRVAT ITNKCADENKML 619-717\	
LLOC24266_Amel_110759165	GEIGIAICEEQ RLAKVWS DY	RKLTRK GAAIL SPTELATEC SVTGQRWS--ERPVKPA D KAQV AI SV T SRPTVTD	SVKVQ LAYKCKENSTAL 27-126\	
bsg25A_Dmel_1930012	PNGTEVSRIS LSAINWD MT	PSITRK LLCEID DRDTLAHH TLSGKPSP--CARPSK QD PLKVA D LVLM TNSLD MTP	EVRTA ITTKCADENKML 102-200\	
Clorf165_Hsap_13375807	GSGIVWDEEK NHQLQVT GD	SKYT KRNLA VH WGT DVLKNR SVTGvatK--DAVVPK P SPRKL SIV RECLYD RAQET	VDETE IAQRLSKVNKYI 133-228\2	
LOC566161_Drer_125823408	GGGIWVDEEK NHQLQRT GD	SKPT KRNLA VW MIWGT ETLKNR SVTGvatK--DALPKP L SPSKL KI IVR CL SRSQET	ADSAE ITQRLSKVNKYI 273-368\	
Aael_AAEI003994_Aaseg_157104034	VRDSLIPQT MDVIDSNS KY	-----	RVQM ARFDDDSIRNRL 364-466\	
mod(mdg4)_Dmel_24648736	GSRVFVS KVALAKAYIP MP	MIYT TCRVM DL WIGKD LR- VR IAQHEETT-----	DKD LQDII TH VCKVF ALRG-- 441-529>3	
NAC1_Hsap_16418383	GTMVYITRAQ LNCNHS HR	KV L RLR RL AS DFD RTLANS C TG IRSS--NDPRPK R PS DRV HL A V K Y CQNPF ANPK	EVQE FDHKLSTL KLMP 441-529>3	
LOC495228_Xlae_148236339	SSGVYVITRAQ LNCNHS SH	KP KLM TR RL DYF ES RE TLARS SAT GQRI A--TMEPKL R P D PKV T TA I K Y V TRACGRG	EMNAIA ADMCTNAR RRVV 374-471\4	
CCDC4_Hsap_83287964	NYPVYITS TSQKDNDEA VNS	KD R RL RL RYI R FV E T DEL KLY C GLG KRKR--ETGP E P DP PKV T CL R E RF IRMH C T SNP	NFNAV INSKCGTS RRAV 384-446\	
GSTM:00029264:G:001_Tnig_47226171	NYPLF ITNKDNDEA VNS	KD R RL RL RYI R FV E T DEL KFS C GLG KRKR--D SGSL ER P LN P PKV V SC L E RF IRMH C A SNP	DWM PSEEQ INKVFS DAVGHA 386-490\5	
LOC560711_Drer_125843107	DYDFV FI PKA QLDS ILLN	RS S LL FR KE CA D DT TL ANS L PNG KRKR--L ND TRK G QD Q N IV GA IK V FT E Y K T C ANG	DWM PSEEQ INKVFS DAVGHA 255-359\	
C10orf30_Hsap_2161768	GFDFV MP KS OLDS ILLN	RS S LL FR KE CA D DT TL ANS L PNG KRKR--L ND TRK G QD Q N IV GA IK V FT E Y K T C ANH	RDW QILQDQIKLARR LRKG 267-373\	
CcBV_3..4_Ccbv_57753424	RTG VV VR KR KE L KRC I RE	ND R T LA RL LL TE V SQ NAL SVC TWTGGAK--NIDIRP G QD EN AR M V ML L FT V EQ QH -G KK	RDW V QILQDQIKLARR LR LRKG 239-345\	
CcBVs6gp3_Ccbv_57753417	QRGVV WV Y GD L KY C QQV	KD K SL AR RL LA V EN R K AL S VC LS ITERA Q--GSNAP E DD HACT V LL N F V LEH-G Q -----	CGWS.ANTSAV M ST IR T KINDI 1112-1213\6	
GIP_L1_00580_GInd_117935419	QSDIYV SY Y GD L KY C QQV	KD K SL AR RL LA V EN R K AL S VC SM SE KA Q A--GSNLP E DE HAK V LL N F V LEH-G Q -----	RGN ND T DIQ IL N TH SK I QE I 1083-1183\	
mbdB_v_Sbgp1_Mdbv_66391199	HTMVY Y NA IKL S NCN K R L	KD K SL AR RL LA V EN R K AL S VC D DT TL ANS L PNG KRKR--L ND TRK G QD Q N IV GA IK V FT E Y K T C ANG	C G WN T DL I PK IL D TH HS K Q I DI 955-1055\	
C6orf65_Hsap_148806920	EKQFQ IE KE WQI ARC N -----	K S .Q K F IND Q QV LY T NE M ATH SLTGAKSS--DKAV KP AMNQN EV Q E I IG V T K Q LF PT ND-----	K G W I L D T QS I Q NS I R N K MF Q E 142-243\	
LOC794392_Drer_125831142	-Y TE FT IP -ELL ERC NT	GT Q LT ND E LR GL Y E RE CLASH S IS GV VV Y-----	SIRR M IG Q KL N-N C T KK P N LS 171-270\7	
BANP_Hsap_74729731	VRC AI IPS-D M LI H ISTN	-----	E IG Y IR Q L NE A KK R LR K RP K 202-300\	
SMAR1_Mmns_10312104	RT E K MA ML LL UD Y D SH RE V Q AV S N LS Q GG KH-----	G K K Q QD PL LT Y IG CH L F Y FK G ITE	SDW Y R IK Q S ID SK C RT AW R R 255-348\8\	
LOC575996_Spur_115728493	VR CA IPS-D M LI H ISTN	RT E K MA ML LL UD Y D SH RE V Q AV S N LS Q GG KH-----	SDW Y R IK Q S ID SK C RT AM R R 237-330\	
Capitella_spi	VR CK IN PT -EM V HIM M N	RT D K LA LK LL DL LL D KE M Q AV S N LS Q GT KH-----	K KK K QD PL LI Y IG CH L V H CH G ITH	EDW Y R IR Q N ID SK C RT A F RR 278-371\
NEMVEDRAFT_vlg243017_Nvec_156383934_4	VR VP IT PS -D L LI H IS <i>N</i>	RT E K MA LS LL DL LL D DT Q AT <i>S</i> N LS Q GM KH-----	G K K Q QD PL MI Y IG CH L Q I Q R FP G ITE	Q DW Y R IK Q N ID SK C RT A F RR 228-321\
PHISDAELQS SL DR ER K	PH I SD AE LS Q SL R DR K	K P EN LA V V LL R LL R TT Q ER E BR G R T VC GF --	GG S Q QD ND V D Q IR Y FY R AL PF DP-----	D KW G QC IS AM S NS Y LR G TR R R 285-375\
XRoxf20_Hsap_23503281_2	WRNIRMP C SV T L AT K	K S .S LS LA Y Q IK Q L FT D LV Q VS N V Y GN LN K H-----	GG S Q QD ND V D Q IR Y FY R AL PF DP-----	SDW Y R IK Q S ID SK C RT AW R R 255-348\8\
LOC100003955_Drer_125851480	LRK W IP Q -CVY KE VK	ET Q KA V AF P V LY S D P IST L SCS AVTG N PEK-----	G I Q Q QD PN K I E AL REF LA M F P Q FD --	SDW Y R IK Q S ID SK C RT AW R R 383-480\
zgc:113423_Drer_718346404_1	ER KV FI SS -F IL Q R AG K	MT S AA V R Y LS RN I TT K EL S QS ST TG NP SR -----	CLL R QD LN T NK V DA I RE A Y V K Y PK FD --	SDW Y R IK Q S ID SK C RT AW R R 239-337\
LOC764357_Spur_115613065	RIQM VM Q DS R WE NT P --	GA R LA AL Y AR Y C D G K IL IR S SV GR -----	S PN K QD FA GL R I K H LL R Q K Y G SR C -----	VIW K TS RE IS Q SQL C KL R R 966-1064\1
NEMVEDRAFT_vlg243017_Nvec_156383934_4	Y QD VL TL DF Q RT Y -----	E I I-S NA Y VA L V R L P D E VL E RA -A AGE -----	G T R S LD DT I LA K Y I AD V LR G FA E K-----	L W D N CLA I CT R IR Q RN P LL GK 604-699\
KIAA1553_Hsap_10047171_1	PPEYQL T AA L K Q IV VD Q-----	LS G DL AC R LV Q Q L PE LS FD S -----	G FA AK R K U ES L H Q LI R Y N EV Y Y P SV-----	AVW Q CL P L N DF F FS R WA QR 85-185\1
KIAA1553_Hsap_10047171_2	ASDHVV D T Q DL T FE FL DE	SS G DA F FA V P Y L H R P E LF D HR-----	G E WL Q LC Q A O R IN DE LE E GL GL DA-----	229-329\
KIAA1553_Hsap_10047171_3	GAD CL LS K SE R Y S I Y ES-----	K G D Q G K Q E Y D PO R Q L I Q Y N TE I Y F PD M Q-----	R W WT E EF V GL K LD E CR R R 392-492\	
KIAA1553_Hsap_10047171_4	PS P Y L LS K DE R Y E IV Q Q-----	-----	E VM H E BC I P S DE R CR R PN R NR-----	558-658\
GSTM:00016974:G:001_Tnig_47220120_1	P Q EY Y LL S RE Q LN I Y E C-----	LS G GN PS R SL V LM I P E LF T AE-----	R V W M E EF V G LD E CR R RR E TE-----	529-629\
NEMVEDRAFT_vlg243810_Nvec_156379688_1	R P Q FA SR S -AV M Q I K C -----	-----	T W W R HC I RA M DE F RP K KK R -----	169-264\
LOC584784_Spur_115651987	K S -A ME LS GM VG W CH Y ER-----	K G E K E Q V L Q I Y E Q EL LS N CS G CT-----	K CC R E BC V Q S I D SH C R Q L FN S Q-----	323-421\
MC036R_MCV_9628968_1	K G -N PA RS V LR K LV D D IL LV K T C SG K R-----	-----	K CC R E BC V Q S I D SH C R Q L FN S Q-----	323-421\
MC036R_MCV_9628968_2	A LE MI P SP A EL CH L H A-----	-----	-----	-----
MC036R_MCV_9628968_3	PA W AG P V T LD I Y E C-----	-----	-----	-----
MC036R_MCV_9628968_4	SC V U AP SL D LR R K M Y G -----	-----	-----	-----
E5R_VVC_137623_2	PA Q Y L IS A K R Y E -----	-----	-----	-----
E5R_VVC_137623_3	-----	-----	-----	-----
GSTM:00013760:G:001_Tnig_472209384	xpat-A_Xlae_14822226	-----	-----	-----
Daphnia_pulex	Branchiostoma_floridae	-----	-----	-----
Consensus/80%	-----	-----	-----	-----

“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).

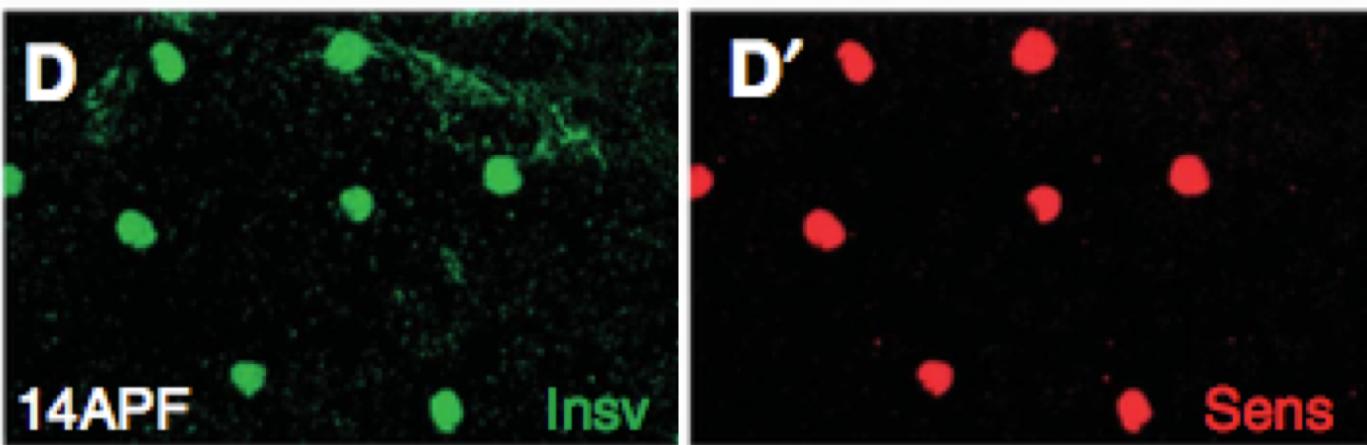


Insensitive protein

- We studied *Insensitive*, a *Drosophila* protein with a 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

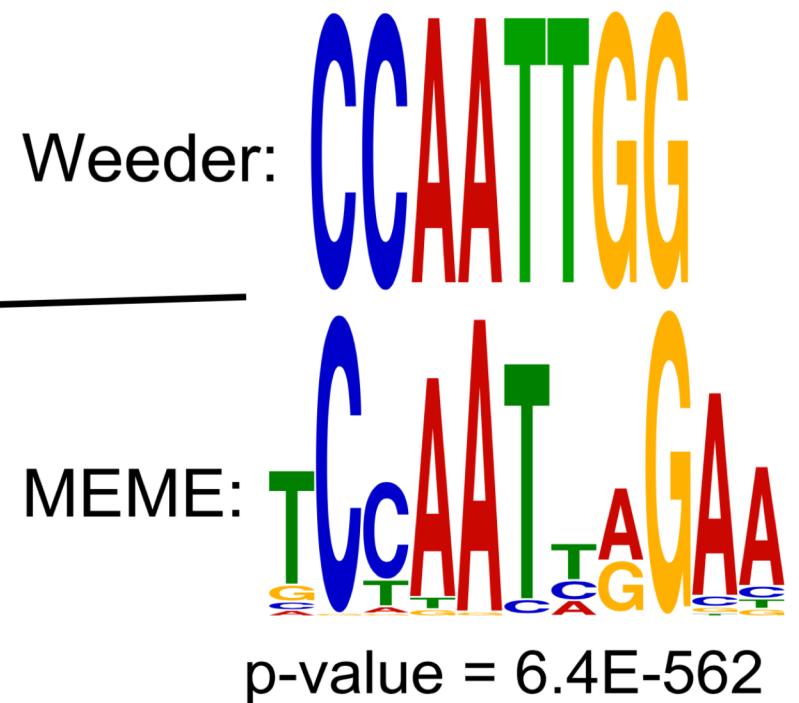
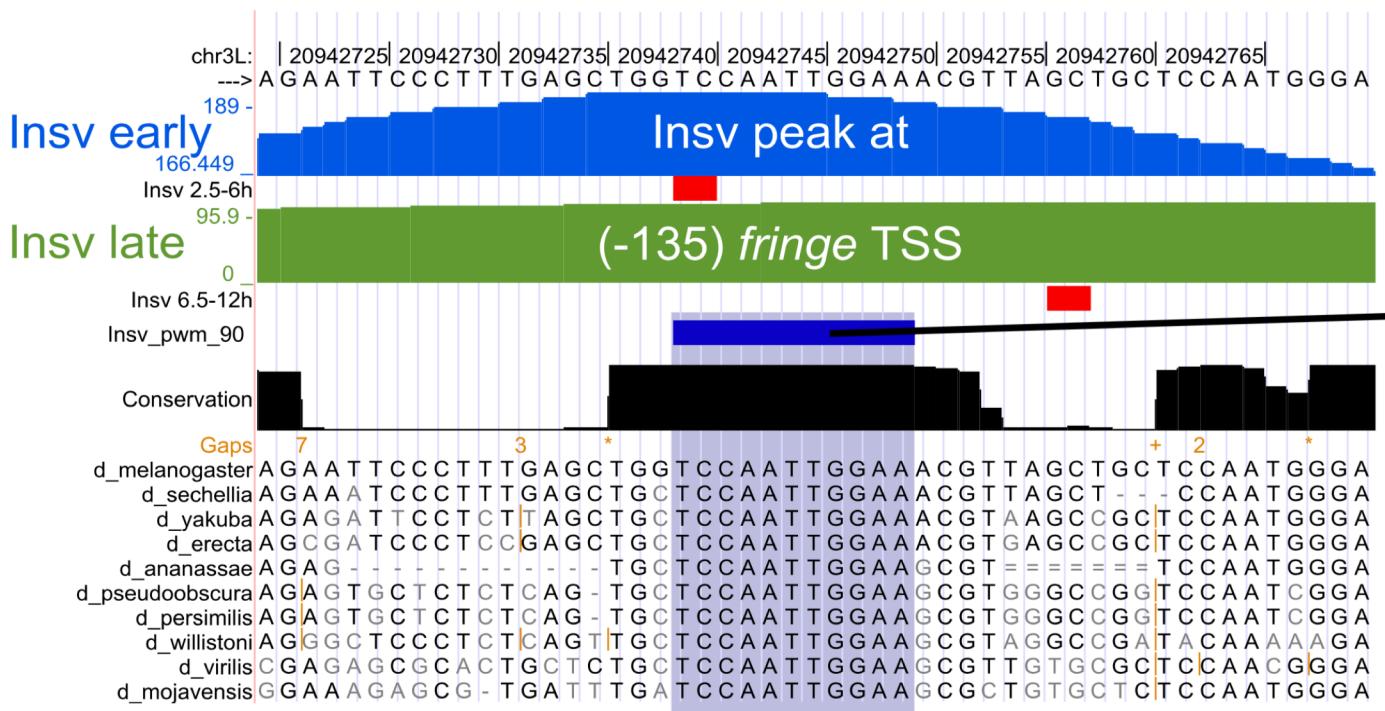
ChIP-seq experiment

- Analysis:
 - FastQC
 - Mapping: Bowtie
 - QC: Phantompeaktools
 - 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..

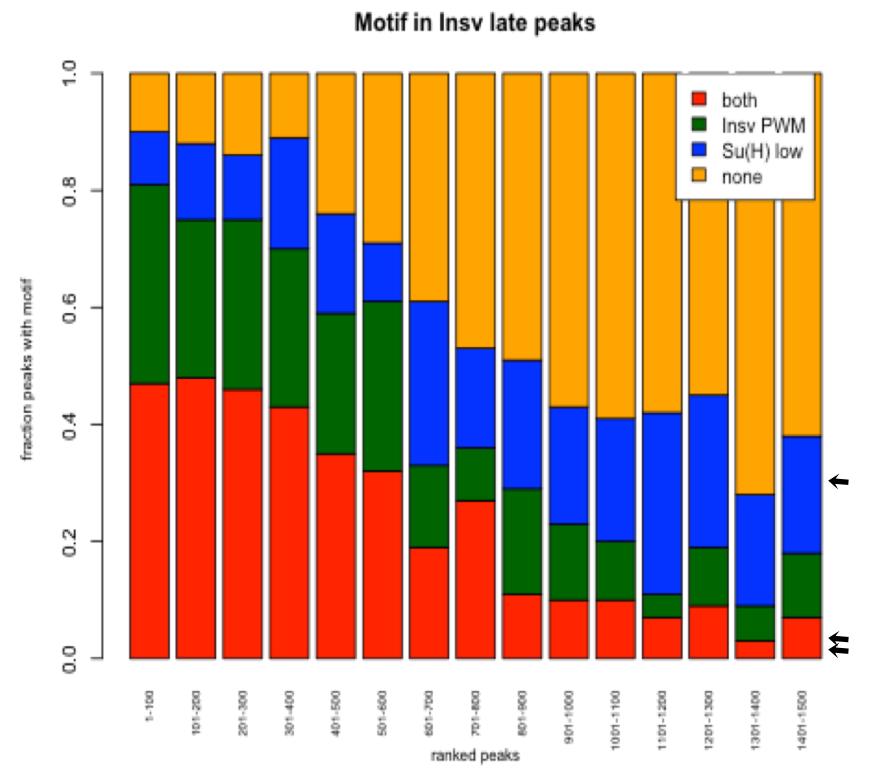
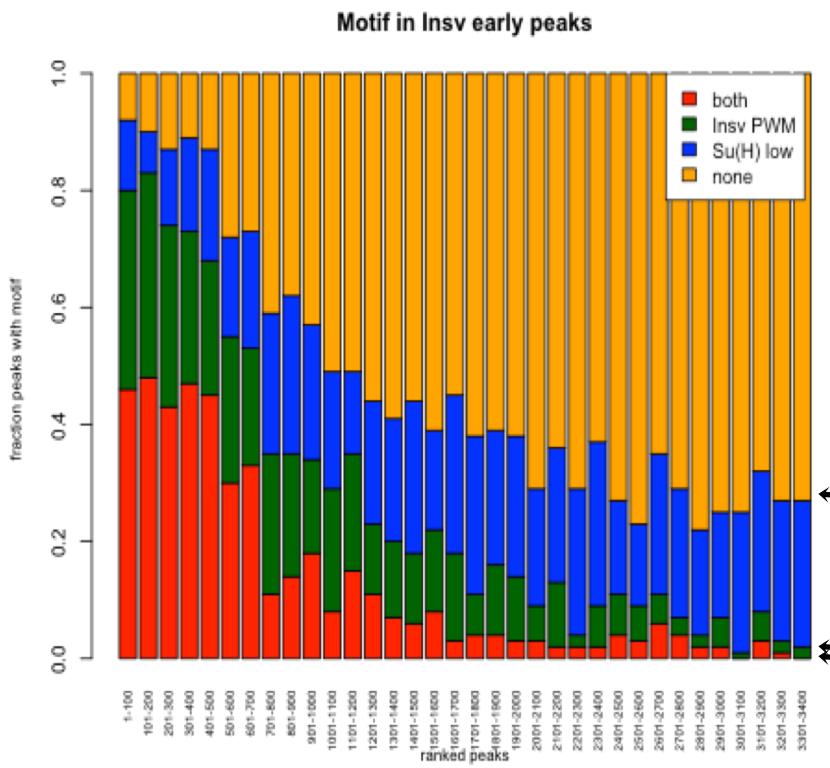
AB	Time	Unique reads mapping	Nr peaks
Insv	2.5-6h	7,473,521 (58%)	5364
Insv	6.5-12h	4,292,248 (61%)	2390

Insenstive seems to bind to a new motif

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

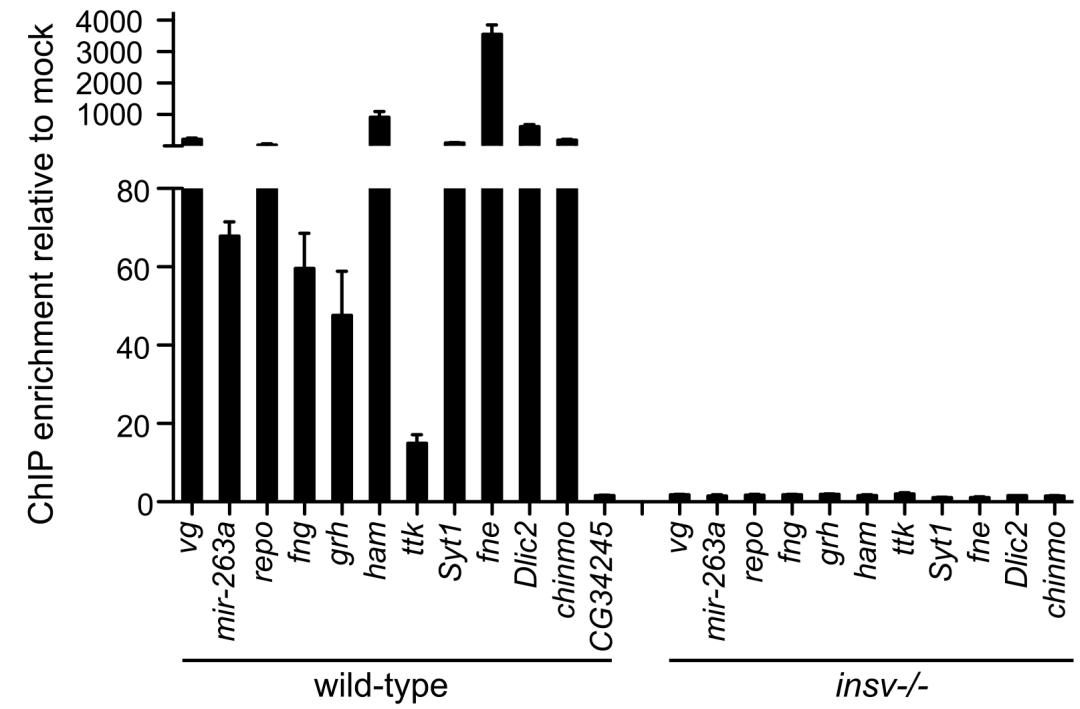
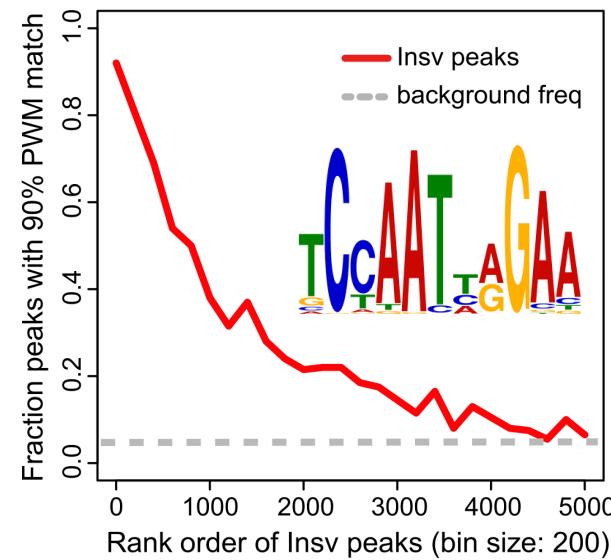
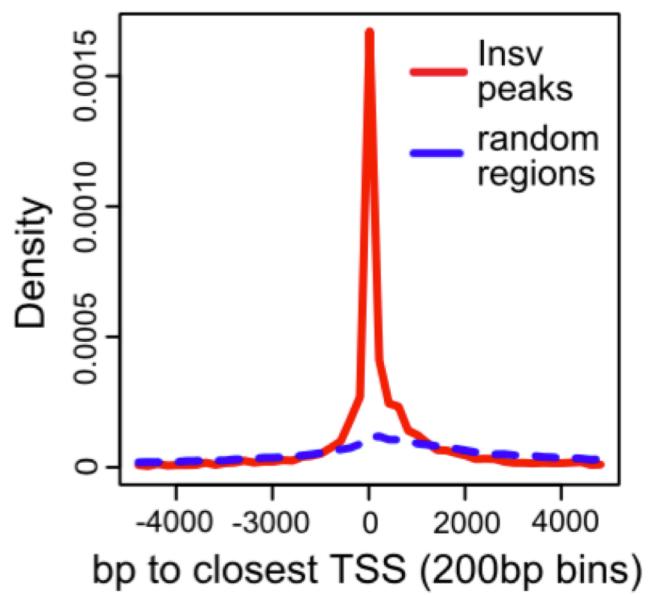


Motif combinations



Background ↗

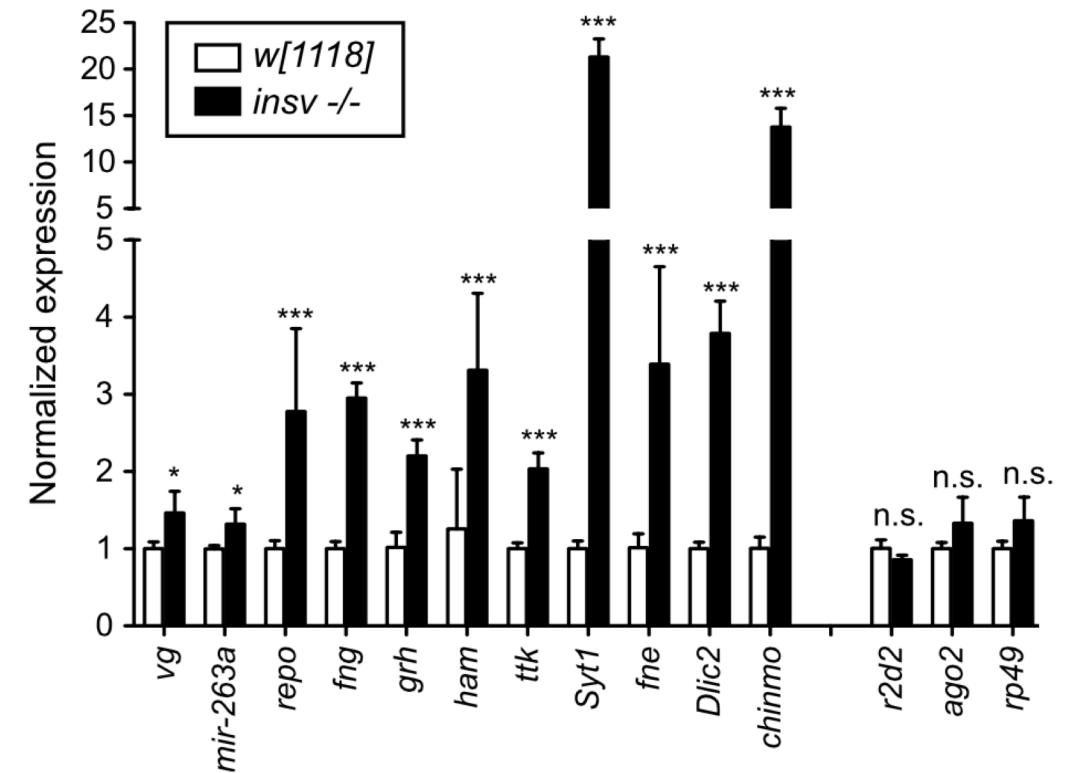
Validating peaks



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

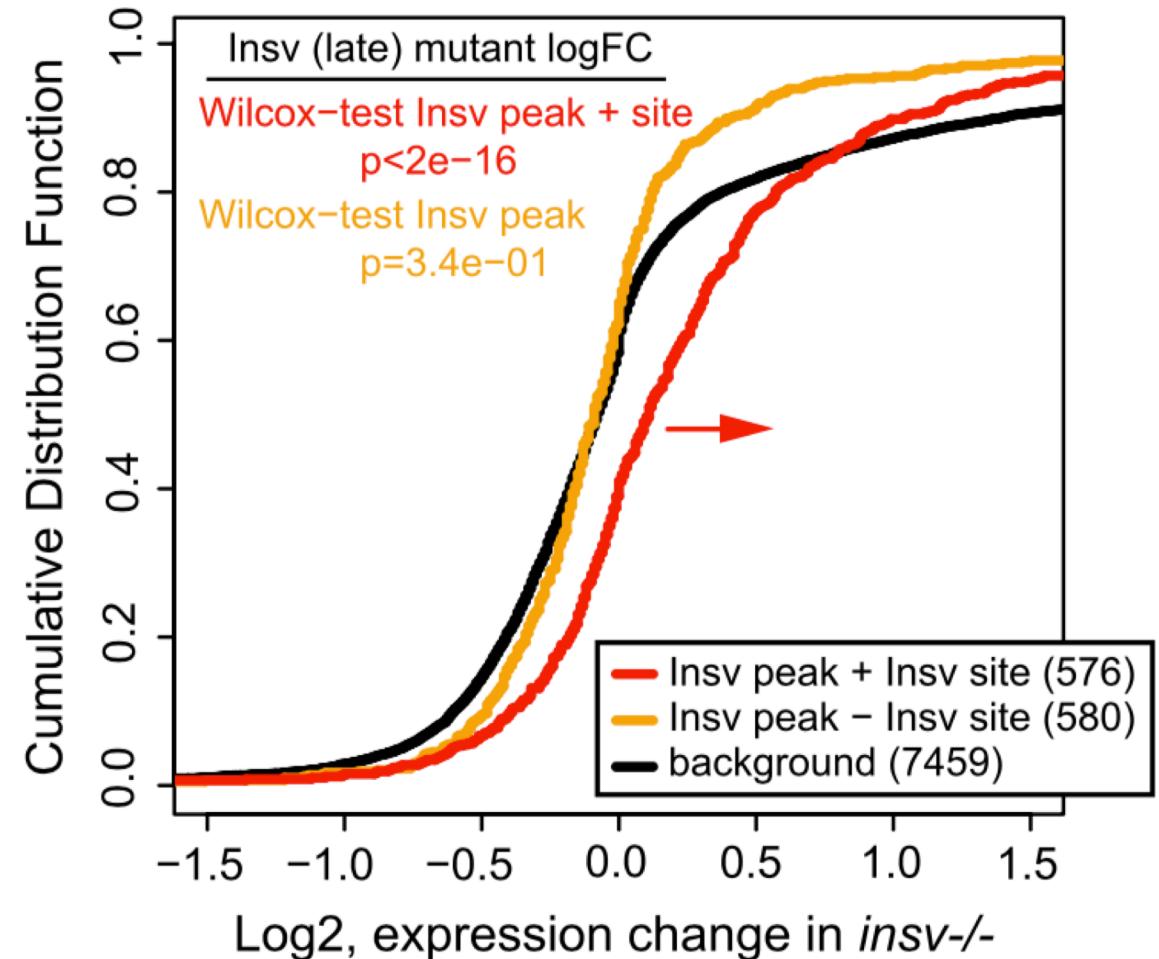
Gene expression

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



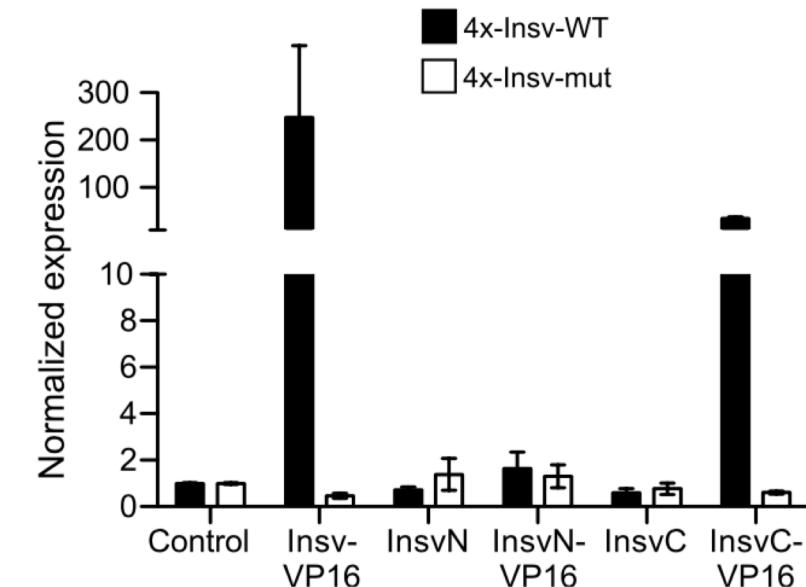
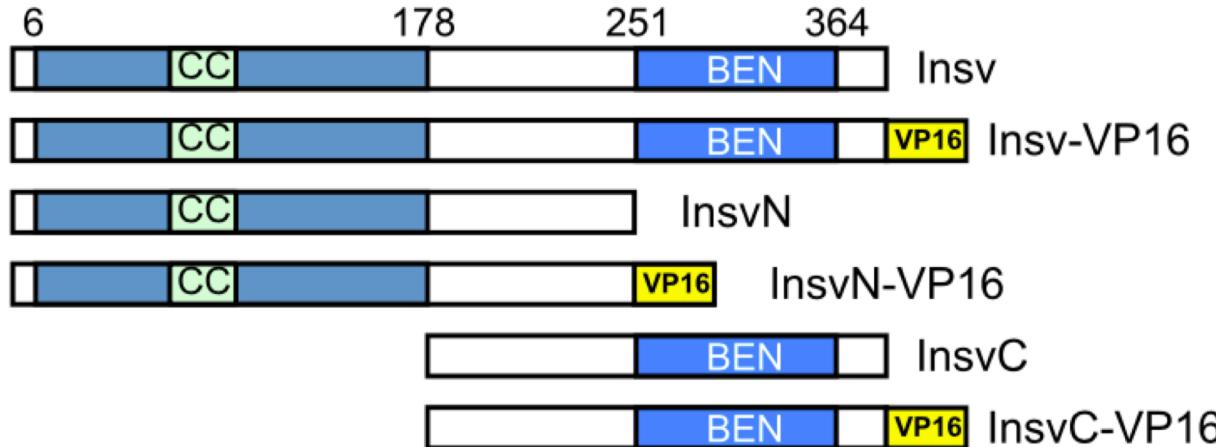
Gene expression, cont.

- 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.

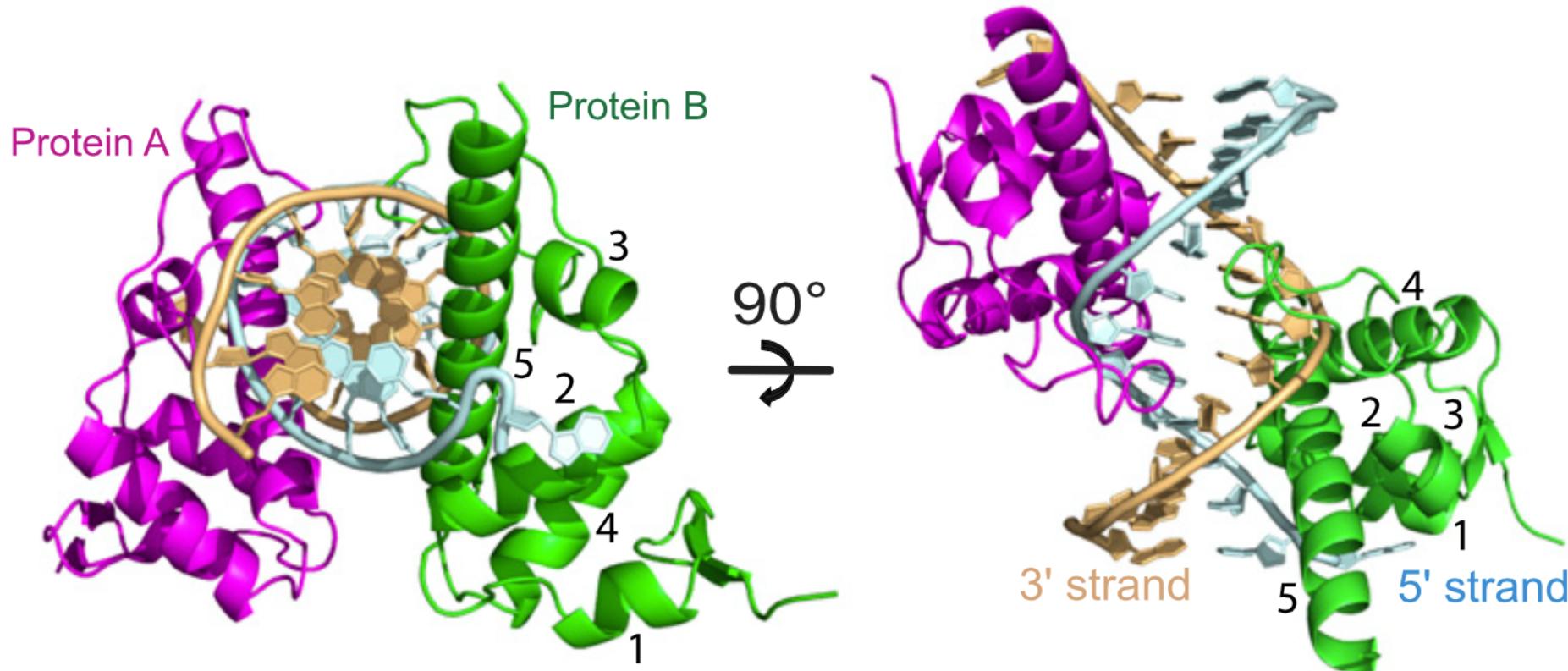


Structure-function experiments

- 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.



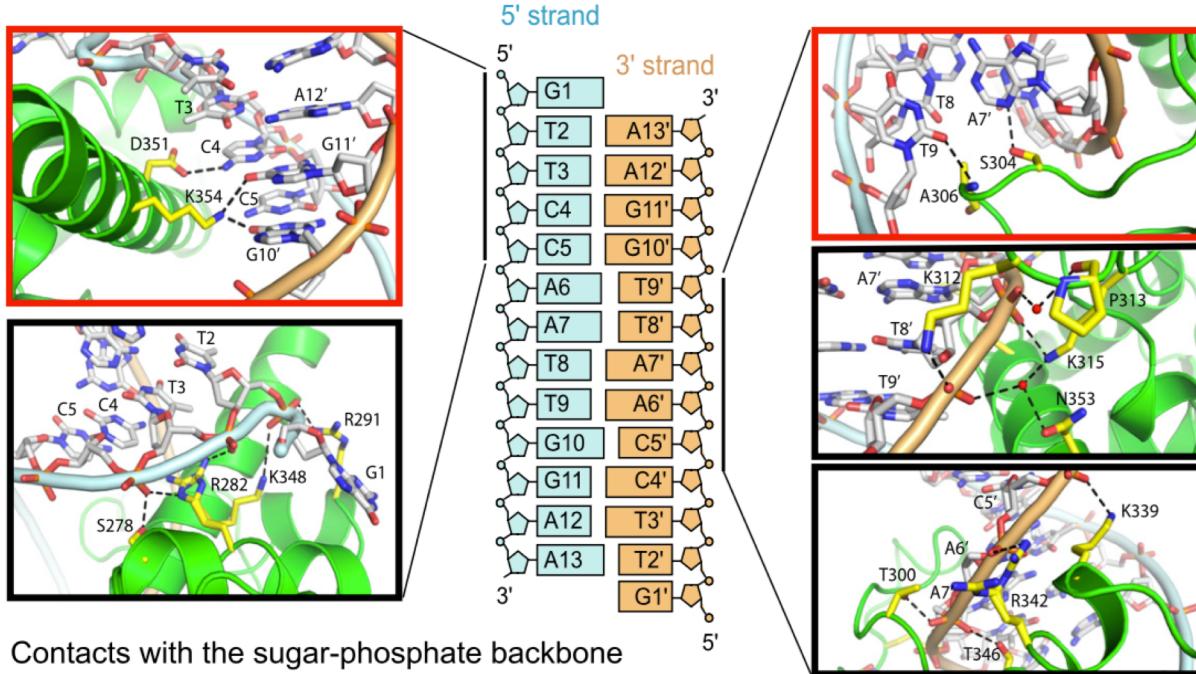
Crystal structure of BEN domain bound to DNA



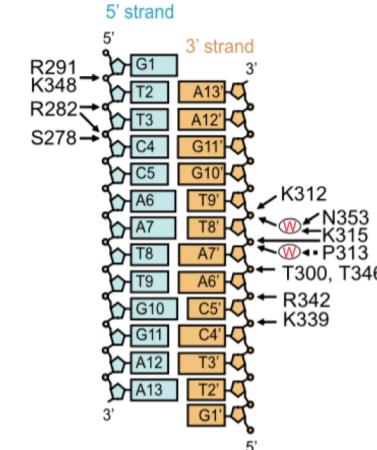
Validating the structure

- 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.

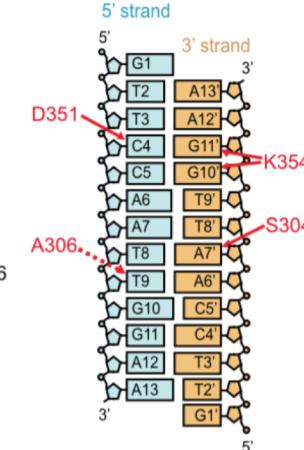
Base-specific hydrogen-bonding contacts



A BEN contacts with sugar-phosphate backbone



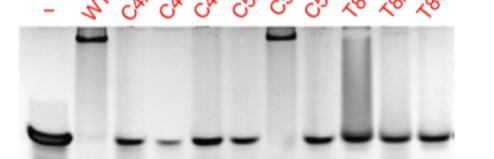
B Base-specific contacts of BEN domain



C Insv-BEN variants tested on wt probe

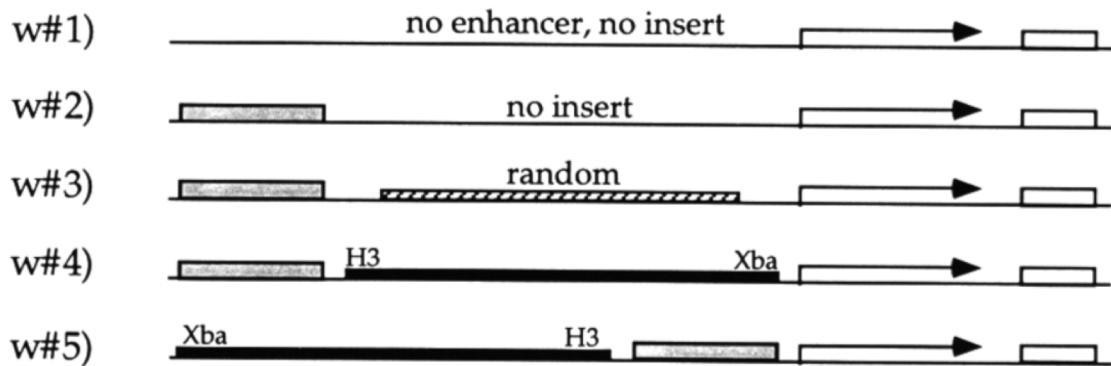


D WT Insv-BEN on variant probes



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.



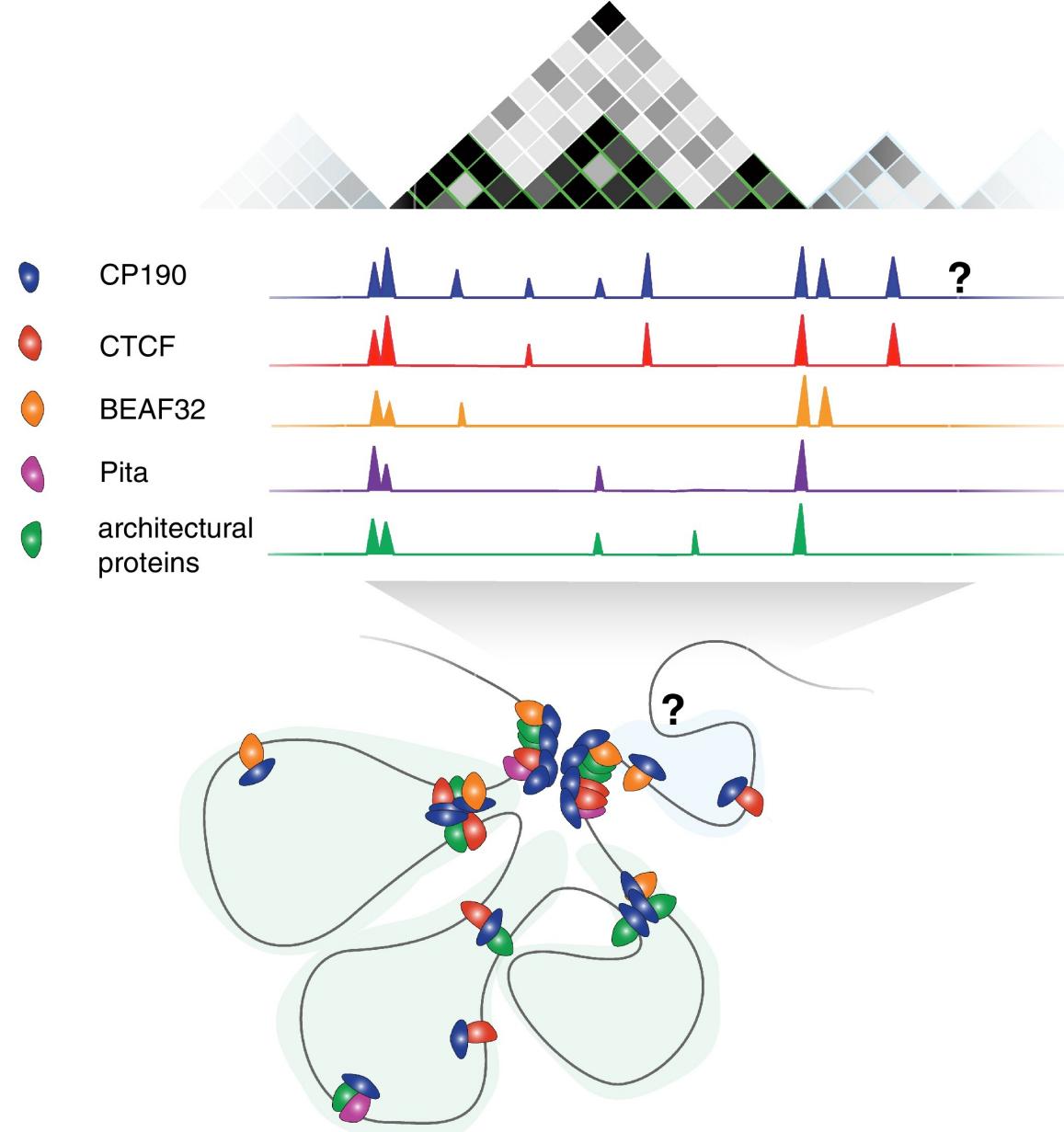
Enhancer blocking
(#light eyes/total)

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



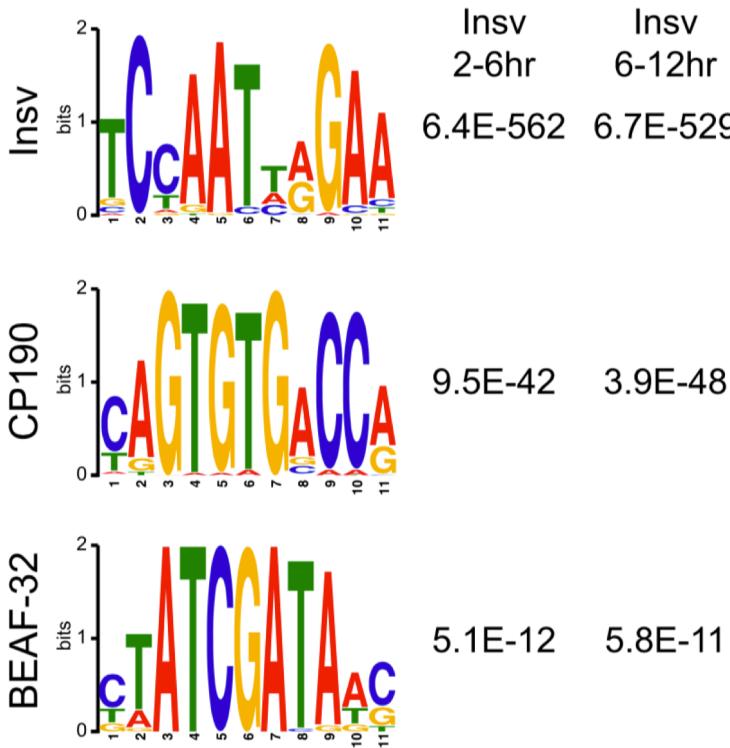
Insulator elements, cont.

- 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

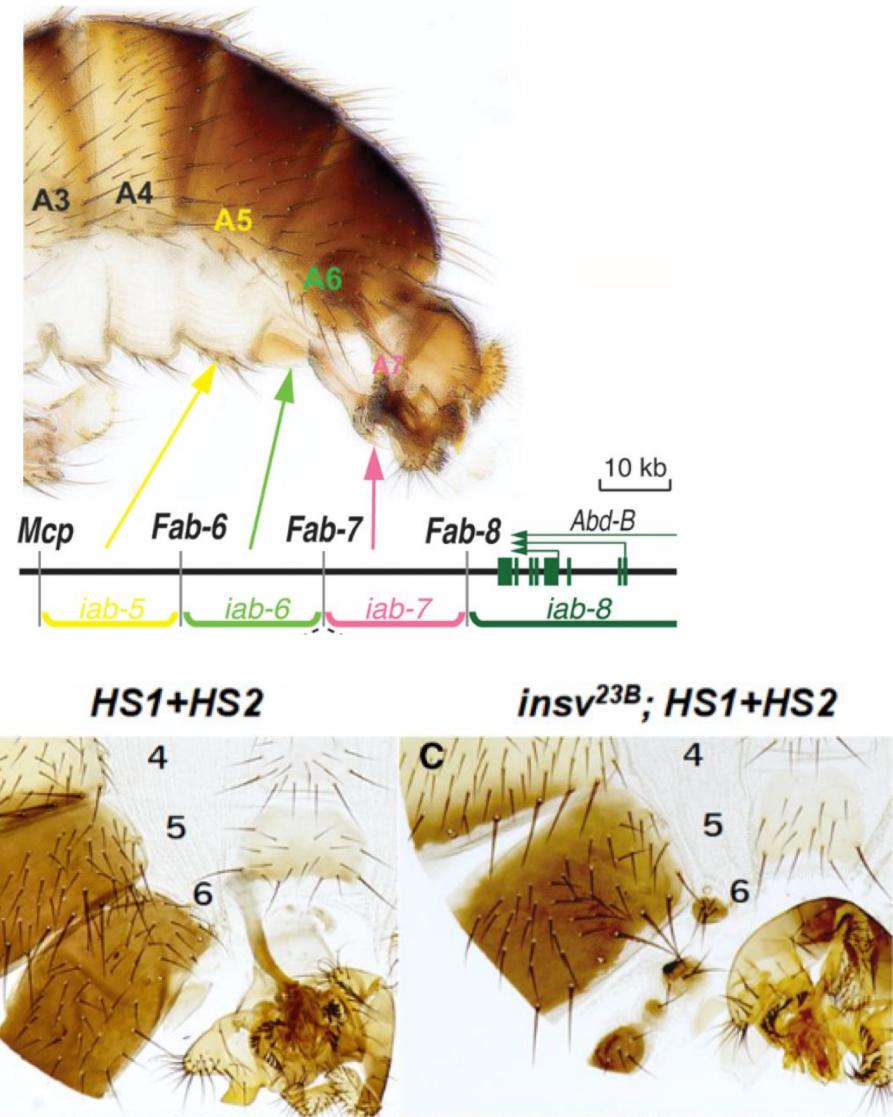
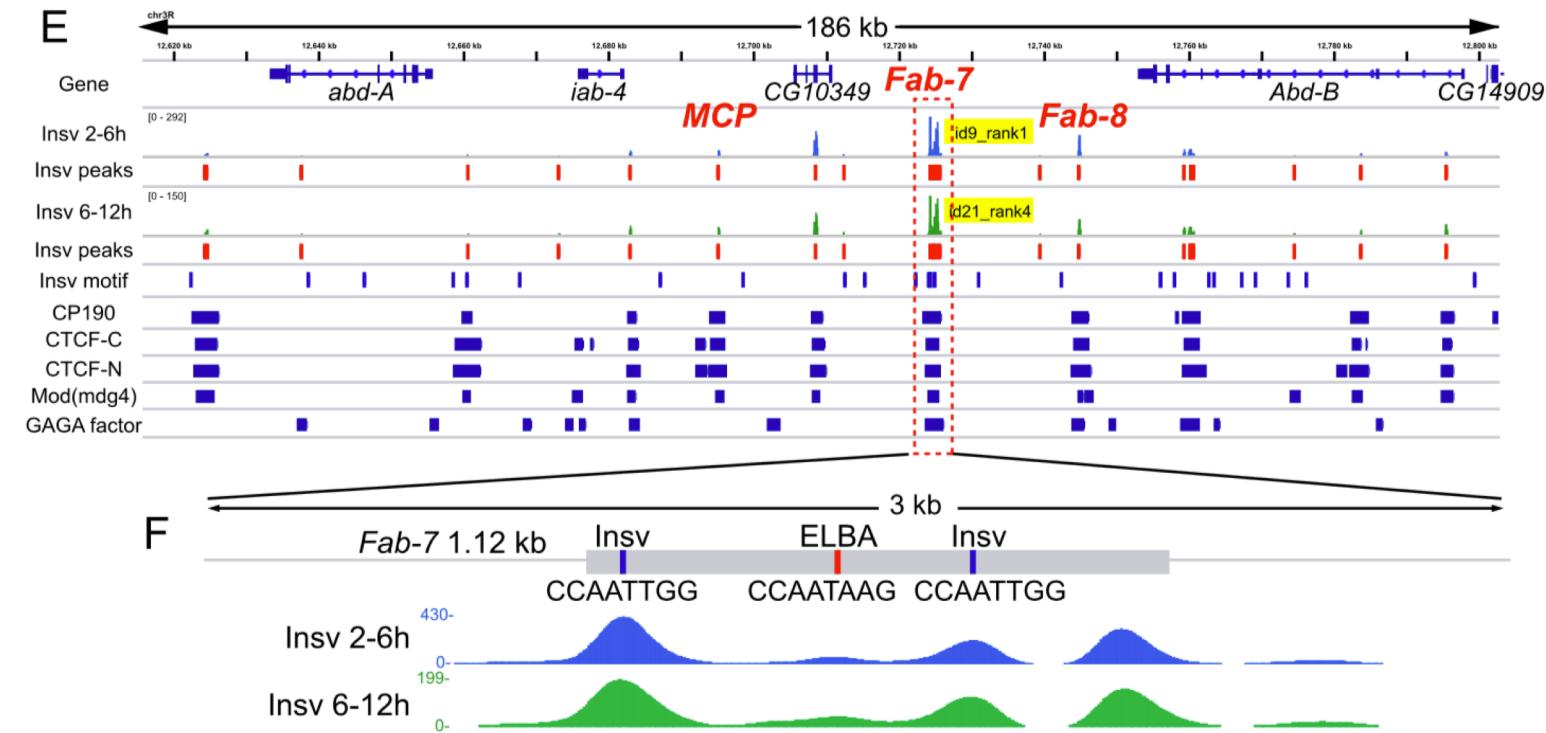


		# fraction of peaks in row data (# peaks in genome)										Insulators		Control TFs	
		Insv 2-6hr (4197)	Insv 6-12hr (1870)	BEAF-32 (4709)	CP190 (6652)	CTCF_C (3154)	CTCF_N (2532)	GAF (3904)	Mod(mdg4) (3060)	Su(hw)_1 (3420)	Su(hw)_2 (3630)	CtBP (4946)	Gro (1337)		
Insv	2-6hr (4197)	1	0.41	0.6	0.72	0.38	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: TFs, insulator proteins etc.
- Insensitive is a transcriptional repressor
- Insensitive (and other BEN-proteins) have insulator activity.

Acknowledgements

Eric Lai (Sloan-Kettering)

Qi Dai

Hong Duan

Dinshaw Patel (Sloan-Kettering)

Aiming Ren

Artem Serganov

