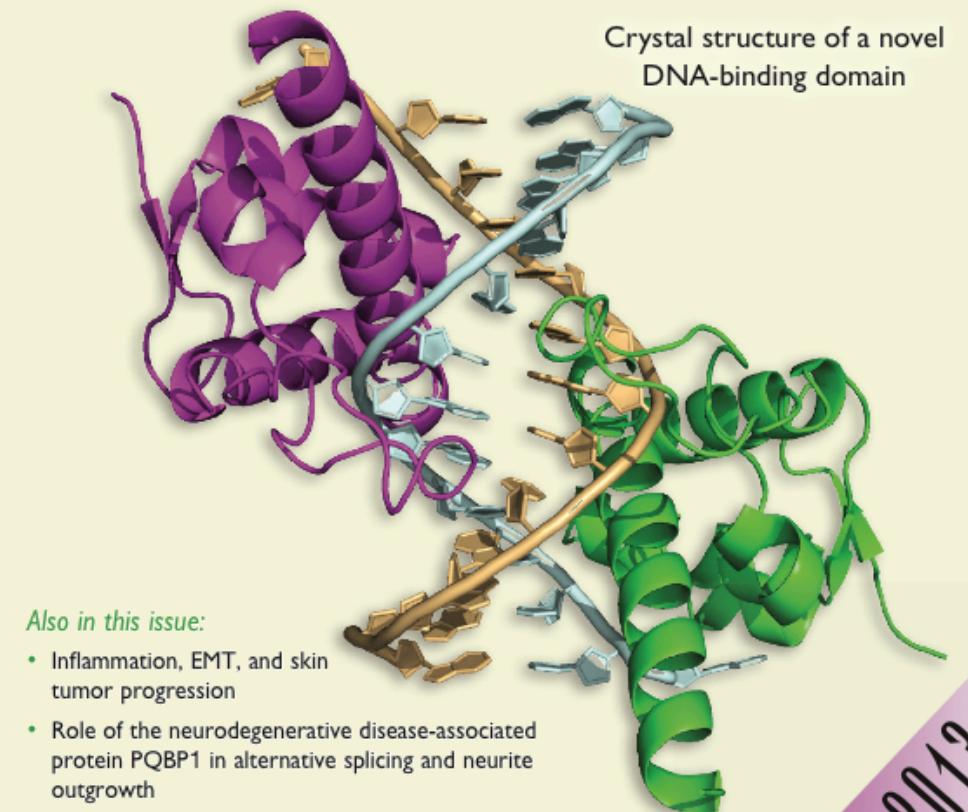
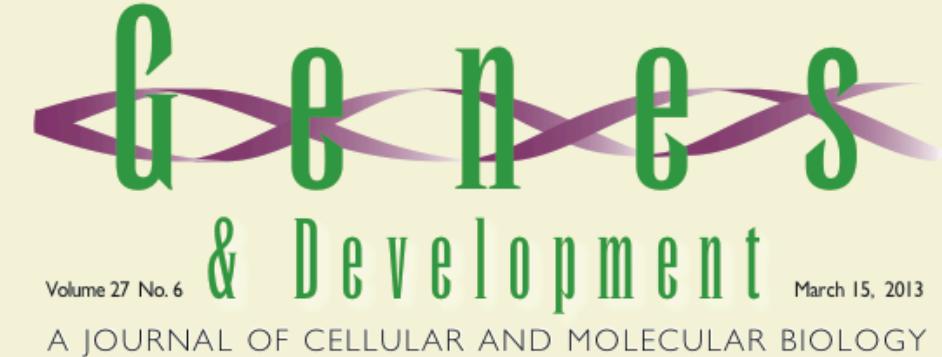


Case study: Finding a new DNA binding domain

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

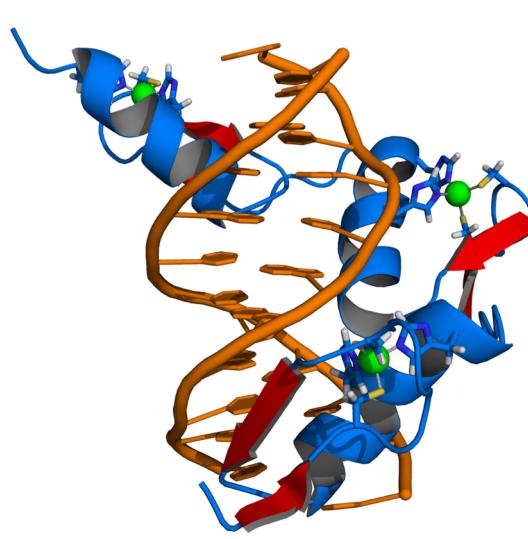


Cold Spring Harbor Laboratory Press

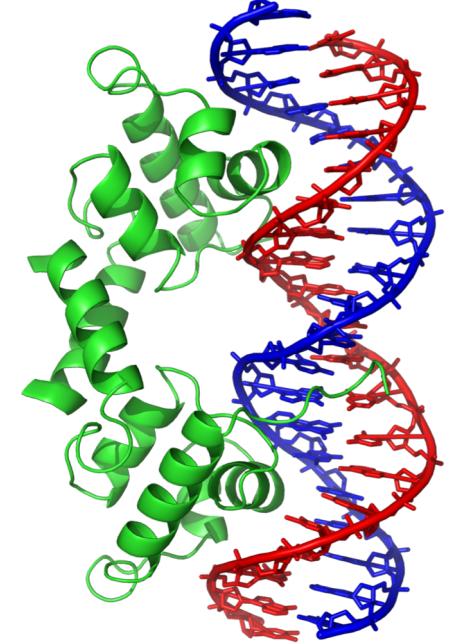
G&D 2013

Transcription factors

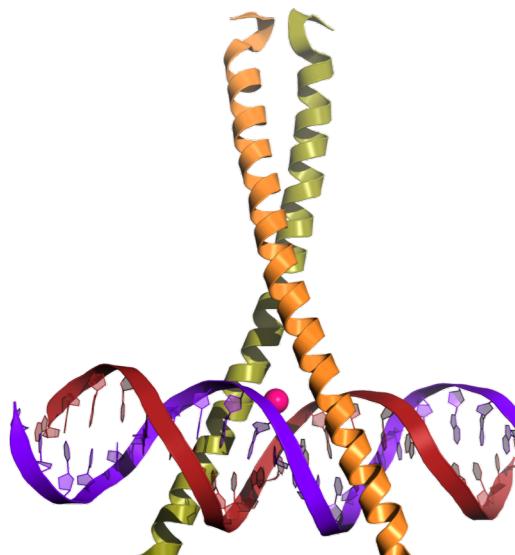
- Transcription factors typically consist of
 - Activation/repression domains
 - A sequence specific DNA binding domain
- The number of such DNA binding domains in eukaryotes is limited:
 - Less than 40 (**Yusuf et al.** *The Transcription Factor Encyclopedia*. Genome Biology 2012)



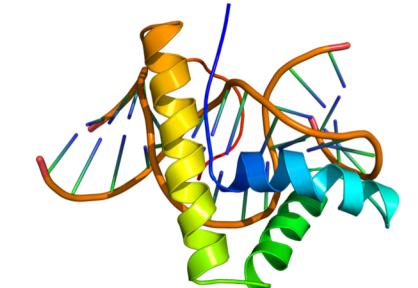
zinc finger



helix-turn-helix



basic leucine zipper



high mobility group box

BEN domains

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

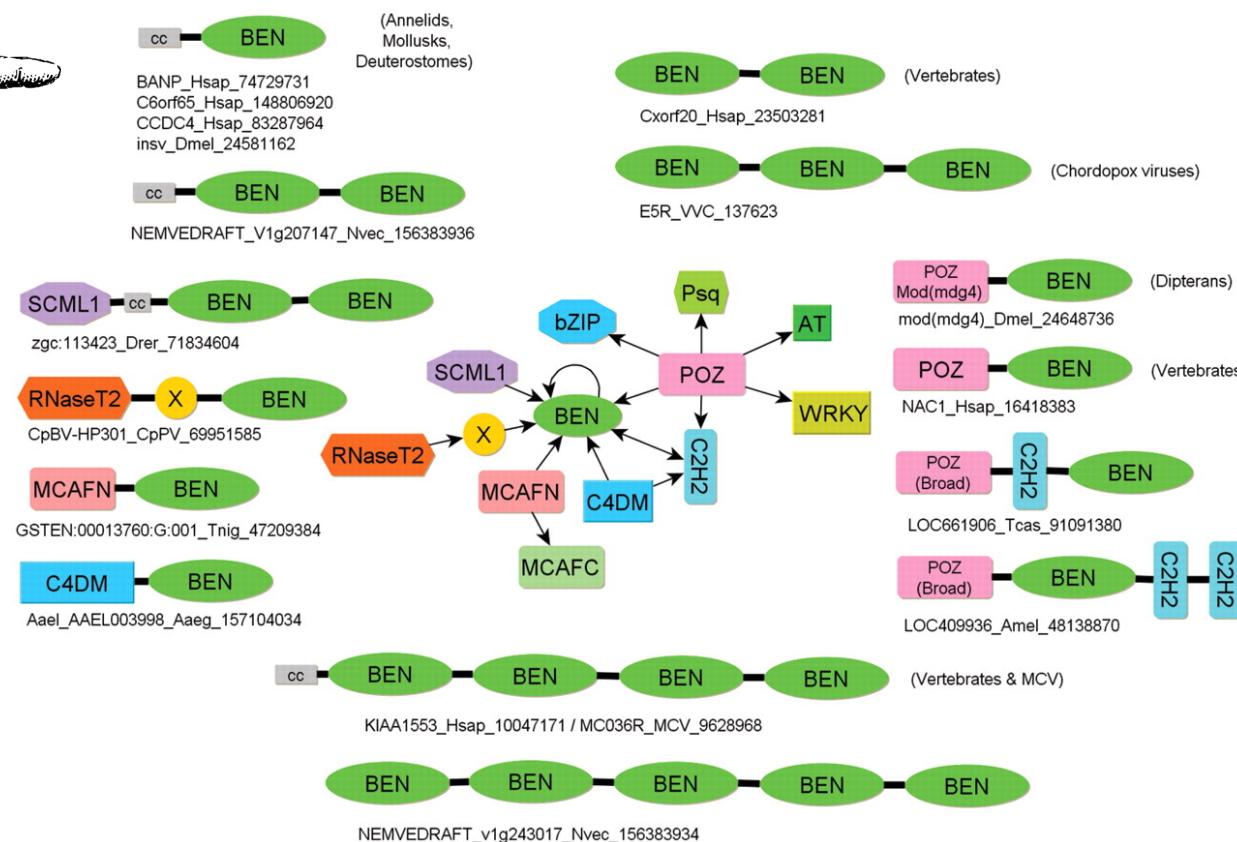
Secondary structure	----- HHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----	----- HHHHHHHHHHHHHH -----
insv_Dmel_24581162	PNNTICVPASV FENINWS VC	SLATRK LCLVTFIDRET LATH SMTGKPSP--QDKPLKMQDPGKIQDIX IFAV THKCNASE	-----	EVRNATT KCADENKML 259-356\1
AgaP_ENSANOG0000025789_Agam_118791739	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 SRPTVDS	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\	
Clof165_Hsap_13375807	GSGIVWDEEK NHQLQVT GD	SKYT KRNLAHV WIGTDVLKNR SVTGVATK--DAVVPK P SPRKL SIV RECLYD RIAQET	VDETE IAQRLSKVNKYI 133-228\2	
LOC566161_Drer_125823408	GGGIWVDEEK NHQLQRT GD	SKPT KRNLAHV WIGTDVLKNR SVTGVATK--DALPKP L SPSKL KI IVR CL SRSQET	ADSAE ITQRLSKVNKYI 273-368\	
Aael_AAEI003994_Aaseg_157104034	VRDSLIPQT MDVIDSNS KY	LPFS SKRNLA LW GHMER LAWS SVTGKRSN--NSTPSI Q ZEPE FLSP IKEK VYHRAMQET	RVQMARFDDDSIRNLRR 364-466\	
mod(mdg4)_Dmel_24648736	GSRVFVS KVALAKAYIP MP	MIYT TCRVM D IVIGKD KL-VR IAQHEETT--DKDL IQDII TH VCKVF ALRG	AVQEFD DHKLSTL KLMP 441-529>3	
NAC1_Hsap_16418383	GTIVVYITRAQ LNCNHS RV	KV HLRRL LA SDFT LNLS CCGT GI RS--NDPRPK P DSRVL HA V KV Y CQNPFAPNFK	EMNAIA ADMCTN ARRVV 374-471\4	
LOC495228_Xlae_148236339	SSGVYVITRAQ LNCNHS SH	KP KLM TRRL LD YF SE RTL ARS SAT Q RIA--TM E PKL R PD PKV T TA I K Y TR AC GRGC	NFNAV INSKCGTS RRAV 348-446\	
CCDC4_Hsap_83287964	NYPVYITS TSQKDN DEAVNS	KD R RLRLRY I RFV T TTD ELKYS C GLG KRKR--ETGP ER P PD PKV T CL REF IRM HCTSNP	DWM PSEEQ INKVFS DAVGHA 386-490\5	
GSTEN:00029264:G:001_Tnig_47226171	NYPLF ITNKDN DEAVNS	KD R RLRLRY I RFV T TTD ELKFS C GLG KRKR--D SGSL ER P PN PKV V SC L REF IRMHCASNP	DWM PSEEQ INKVFS DAVGHA 255-359\	
LOC560711_Drer_125843107	DYDFV FI PKA QLDS ILLN	RS S LLFRK W CAFE D DTT LNLS LPNGKRKR--L NDTRK G DQ NI V GA I K V TE KY C TANG	RDW Y Q ILQDQ IK LARR LRK G 267-373\	
C10orf30_Hsap_2161768	GFDFV MPK QS ODS ILS IL	RS S LLFRK W CAFE D DTT LNLS LPNGKRKR--L NDNRK G DQ NI V GA I K V TE KY C TANH	RDW Y Q ILQDQ IK LARR LRK G 239-345\	
CcBV_3..4_Ccbv_57753424	RTG VV VR KRKEL KRC IRE	ND R TLAR LL TEV ES Q NAL SVC TW TW GG KAK --N IDIRP G DEN AR M V LLT F V EQ QH -G KK	CGWS.ANTSAV M ST IRT KINDI 1112-1213\6	
CcBVs6gp3_Ccbv_57753417	QRGVV WVSYGDL KY CQQV	KD K SLAR LL LA V ENR K AL SVC LS ITERA Q--GS NARP E DDHACT V LLN F V LEH-G Q	RGN W TD Q IL NTL H SK I QE I 1083-1183\	
GIP_L1_00580_GInd_117935419	QSDIIVY S Y GELK Y CQQV	KD K SLAR LL TEV EN K K AL SVC.S ME SK EA KA--GS NLR E DEHAKS V LLN F V IDY-G Q	C GWN TDL PI LD T H SK I QD I 955-1055\	
mbdB_v_sBgp1_Mdb_v_66391199	HTMVY Y NA IKL S NCN K R L	KD K SLAR LL LE V E T PK SALTICL TG S RARA --GAT IRP G DET ART V LT TY EE Y GRE	K GW I LD T Q S I NR K MF E 142-243\	
C6orf65_Hsap_148806920	EKQFQ PI KE WQI ARC N ---	KS Q KF IND Q W V LY T NE M ATH SL TG AK SS --DK AV KP AM N Q N EV Q E II G V T K Q LP PT ND	S IRR M IG Q KL N-N C T KK P N LS 171-270\7	
LOC794392_Drer_125831142	-Y TE FitP-ELL ERC NT	GT Q LT ND DL R LG Y E RE CLASH S IS G V V Y N--R GQ KP A P T EE V Q A IL R T V Q Y F PP G K	E IKG Y IR Q L Q NE AK R L R K K P 202-300\	
BANP_Hsap_74729731	VRC AI IPS-D ML H ISTN	RT E KM AL LL D Y D Y RE V Q A WS N LS G Q G KH --G KK Q DP PL T Y IG CH L F Y PK G ITE	SDW Y R IK Q S IDS K CR T AW R R 255-348\8\	
SMAR1_Mmns_10312104	VRC AI IPS-D ML H ISTN	RT E KM AL LL D Y D Y RE V Q A WS N LS G Q G KH --G KK Q DP PL T Y IG CH L F Y PK G ITE	SDW Y R IK Q S IDS K CR T AN R R 237-330\	
LOC575996_Spur_115728493	VR CKIN PT-EM VHIMNM	RT D KL AL LL D Y D Y RE V Q A WS N LS G Q G KH --K KK Q DP PL T Y IG Y CH L V K H G I TH	EDW Y R IK Q N IDS K CR T A FRR K 278-371\	
Capitella_spi	VR VP ITPS-D ML H ISN	RT E KM AL LL D Y D Y RE V Q A WS N LS G Q G KH --G KK Q DP ML Y IG IR CH L Q IR PK G ITE	Q DW YR IK Q N IDS K CR T AP FR R 228-321\	
NEMVEDRAFT_vlg232490_Nvec_156390312	PHISDAELQS SLR DE DKR	K P EN LA V LV RR LT RE ER BR G TV CG --G GS Q DN D V Q D Y IR Y F Y R AL P DF P	DKW G QC IS AM S NS L R G TR R R 285-375\	
CXorf20_Hsap_23503281_2	WRNIRMP-C SV T L AT K S	K S LS LA Y Y Q L E T D BL W V Q N Y GN LN KH --G LC A D P N K IS AL R E Q EN Y IC D	SDW Y R IK Q S IDS K CR T AW R R 667-765\9	
LOC100003955_Drer_125851480	LRK W IPQ-C VY KE VF K	ET Q KAV A F V Y LS I D P ST L SCS AV TG N PEK --G I Q Q DP N K I E AL R FL E AM F P Q FD--VAWA Q C LG V IN-S IT KNL K T	383-480\	
zgc:113423_Drer_718346404_1	ERK VI FFS-F IL Q R AG K	MT S AAV R Y LS R N I E T K EL S Q S ST TG N PSR --C LL R LD T N K V DA I RE A W V K Y R K PF D --K DW K V CL A V IN ST ARY Y R F MD	239-337\	
LOC764357_Spur_115613065	RIQM VMQ DS R WE ENT --	GA R LA Y Y AL Y RC I D G T K IL IR S SV GR N--S KN P D FA G LR K I KH L Q F K Y G SR C --VIWK TS RE IS Q LC K RL R Y K 966-1064\1	604-699\	
NEMVEDRAFT_vlg243017_Nvec_156383934_4	YQD VT DL P DF Q RT Y --	E I S NAV A Y AV R L P DE VL ER--A AGE --G TR S LD T I KA K AD V LR G FA E AK--L IM D N CLA Y I Q R IN P LL GK	85-185\	
KIAA1553_Hsap_10047171_1	PPEYQLT AA L K Q IV D Q	LS G DL AC R L V Q L P E LF SD V D F SR G CSA G FAA K R K LS HL Q L IR Y EV Y Y P SV K --AVW Q E CL P L N D FF S RF W Q R	85-185\	
KIAA1553_Hsap_10047171_2	ASD H V D T Q D LT E FL D E	SS G DA F FA V Y L H R L P E LF D H R -K LG Q E Y S C--G GD G KQ E Y D PO R L Q I Y R Y T E I Y F DM Q --E AWL Y Q C AO R IN DE E LG L D A	229-329\	
KIAA1553_Hsap_10047171_3	GADC LL SK S RS Y I RS Y E	S S G GN P AS R Y L V H L P E LF T HE--N LR K Y Q N C G --G S LG KK Q DP PS R I KL IR Y Q L Y P R A K--R VW T E EF V GK L DER C R R D TE EQ	392-492\	
KIAA1553_Hsap_10047171_4	PSP Y LL S DK E Y RE V I Q Q	LS G GN P AS R Y L V H L P E LF T AE--N LR L Y Q N Y H NC AN CK R --G S LG KK Q DP VR V N L IR Y Q L V P Q A Q Q --R VW M E EF V GK L DER C R R RE TE Q	558-658\	
GSTEN:00016974:G:001_Tnig_47220120_1	PQEYLL S RE Q LN I Y E C	LS G GN P AS R Y L V H L P E LF T QE--N ARR Y NC --G S LG KK Q DP VR V N L IR Y Q L V P Q A Q Q --R VW M E EF V GK L DER C R R RE TE Q	529-629\	
NEMVEDRAFT_vlg243810_Nvec_156379688_1	RQ P FA S RS-R AV Q I CK S --	K S G N F S Q V Y LL R I Y E Q GE LS N CS G TR--G K E I O P V K Q Y I E K Q V Y H Y I PT--T WW R H C I R AM D E F P R K K ER 169-264\	169-264\	
LOC584784_Spur_115651987	AWEK LS MG V CH Y LE Y RA	K G -N PA RS V L R K U D D IL V K S T C SG K R R --E QA I A D P V K Q Y Y A LT YY D Y Y G VE--K CR R E BC V Q S I D S HC R Q L F N S Q 323-421\	323-421\	
MC036R_MCV_9628968_1	ALEM I P S PA E LL CH LA	CS A D MR AV LL R LE Y PE--V CG AD SE --A E -P AI Y F D A V R C V SE Y Y P L V C--Y V W Q E LG FL R E FL V R C L R V 18-107\	18-107\	
MC036R_MCV_9628968_2	PAWAG P V T LD I Y E C AS	S V -G E LA V LL R HK V Q EL FD A --Q LR R C Y S C--G GD RT H C O P A R L Q I LR H C V AL C FP S MS--G GM GR K D PL P R L LL R LR H Y V Q L Y Q LL H AA R --R WV I K F L A LC D ER R RR R CA T 807-907\	807-907\	
MC036R_MCV_9628968_3	SCV U PL D LR R LR K Y M G	A S -Y N F A VR M Y V Y P L E FT A --N JL H T PN C --G GM GR K D PL P R L LL R LR H Y V Q L Y Q LL H AA R --R WV I K F L A LC D ER R RR R CA T 935-1035\	935-1035\	
MC036R_MCV_9628968_4	PAQY L IS A K R Y K EL A R	R S -G HF A Q A O T V M L P E LF S SS--T ER Q F SC C --G SD EH L R D PL P V R V R L I R Y H V R A VL C PG A --N Q R T Y K LF S D O IS A IG K --A S -SK Y W M Y Y LL R Y P FG D --H RF I Y R M -S IK HH K I F SP F K L N L IR I L Y EV F YN N --N N K W I I GT Q Y D K M I A E SD K Y 102-204\	102-204\	
E5R_VVC_137623_2	IKG K SEED-T L PI K Q M V	Y T .Q EL V E K V L K I L R D LF K S--G EY K A Y R .--V EN G PI G D OT L K -L N I V H D I V E PC PM V--L R K I N S H M E K I L F E N C Y N Q Y 218-318\	218-318\	
E5R_VVC_137623_3	L R K I N S H M E K I L F E N C Y N Q Y	G W V Y A S Y Y Y P Y V N K Y C E W T K V N Y--G GM KE B OP T N V R AM R Y I E R PT L S--D W H R E IRD A NE I E LR V K R KE P 221-322\	221-322\	
GSTEN:00013760:G:001_Tnig_472209384	xpat-A_Xlae_14822226	I P D L I U N P D G K L V K Y S Y N --I N -H R FA E LL F Q Y HP V MS L Q F W A N K V N Y--G SR KL G L M DI H Q T S K R V F E --K E K R I K T R LL N LL R TR Q D R A T 187-287\	187-287\	
Daphnia_pulex	HMSSED-L-DYC N MMAR	NN T KM IS L M M G K V T VE E LL T TK S LT G K--R TT K P A P V D K V NA A Y K I L R K R D K H --E FN Q K V T N Y LR D Q A AK S --K L Y R V V NE K GR M R AT L 266-354\	266-354\	
Branchiostoma_floridae	EQGVVTTYPYI LA Q AK NK T	KT E Q F K I M Y G Y I E T E E LL N LG H GG--G G TH Q A L S P A I S A I S LT E T K Q V GG V --K L Y R V V NE K GR M R AT L 140-229\	140-229\	
Consensus/80%ph....s...h...Lh...h.Fsp...b...p...s...h...Lh...h.Fsp...b...p...Ls...h...Lb...h...s...l...h.p...		

“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).

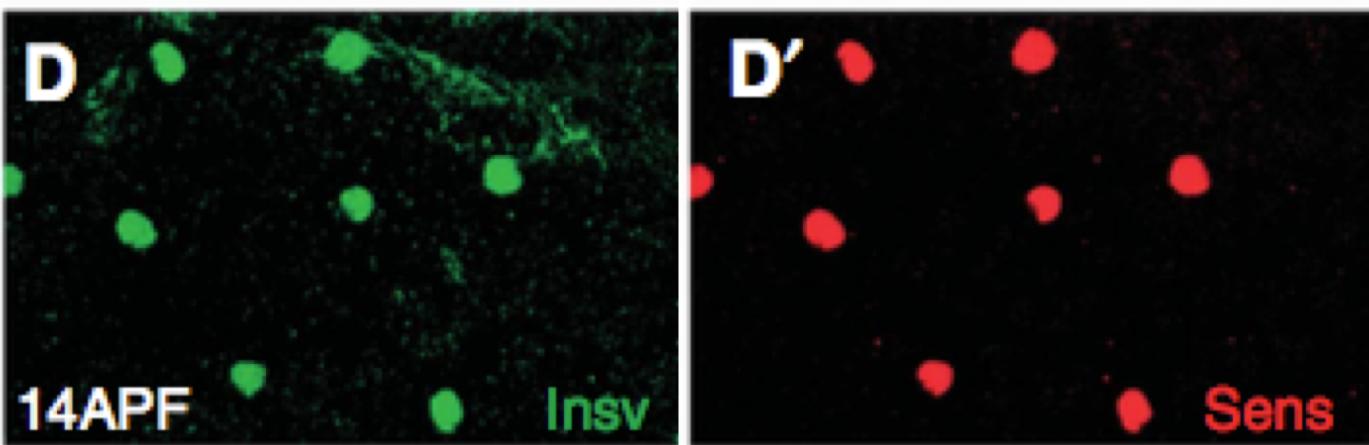


Insensitive protein

- 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

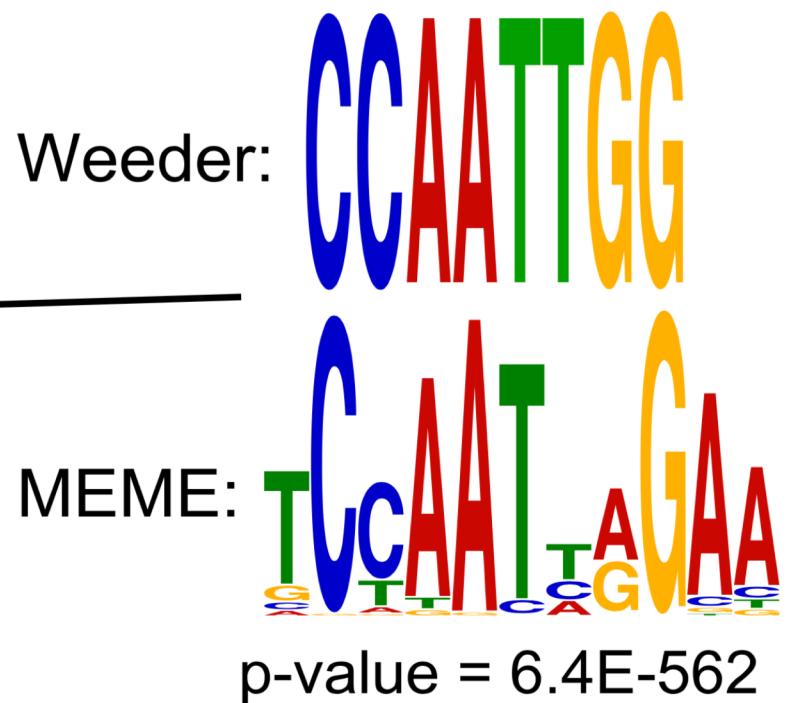
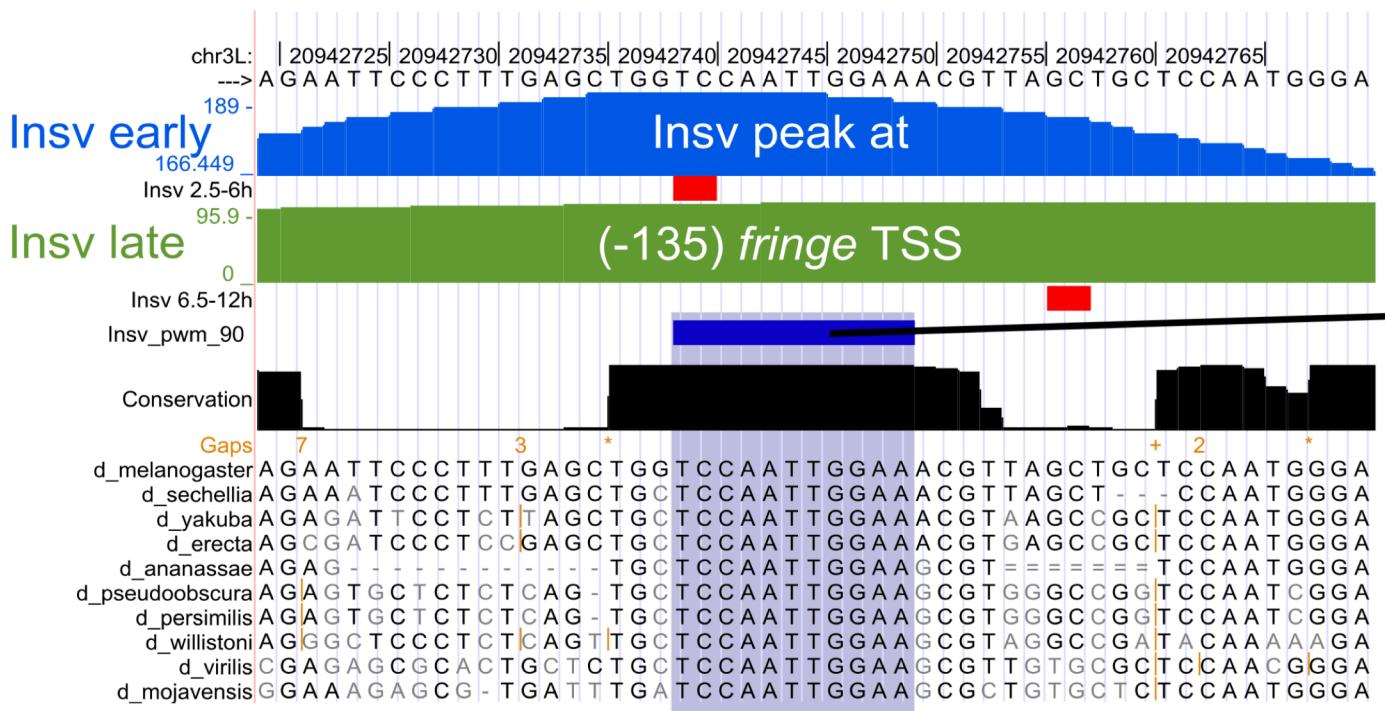
ChIP-seq experiment

- 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..

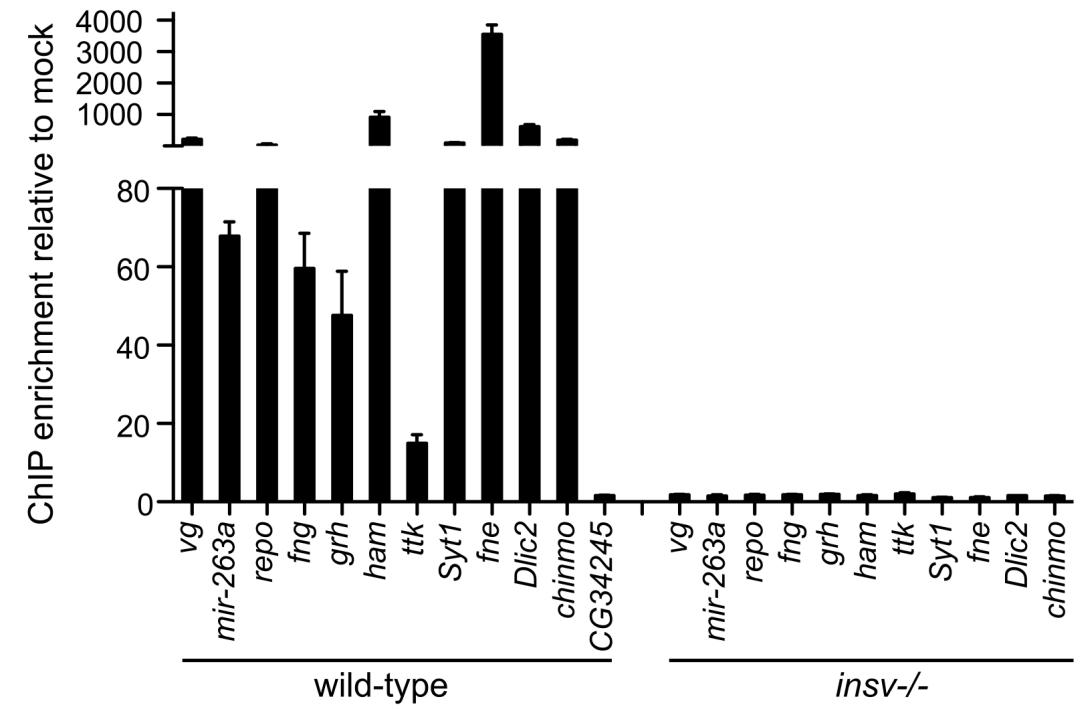
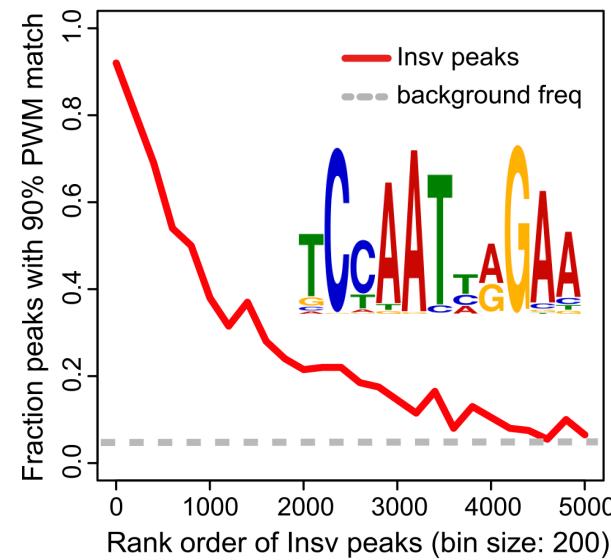
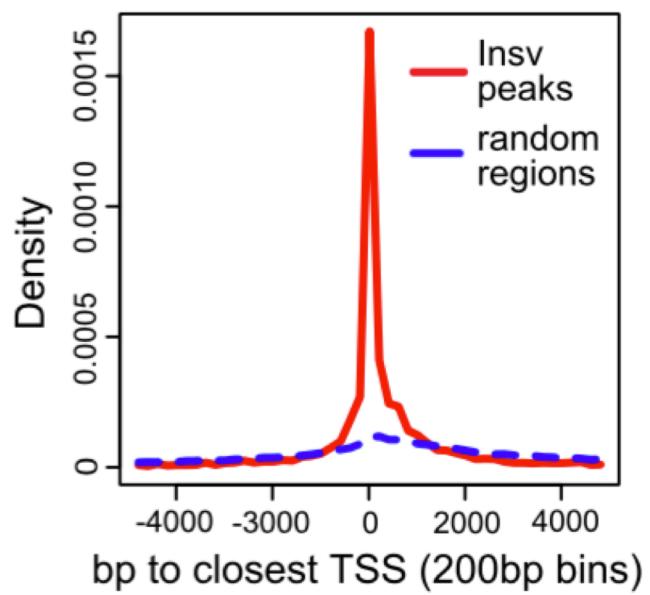
AB	Time	Unique reads mapping	Nr peaks
Insv	2.5-6h	7,473,521 (58%)	5364
Insv	6.5-12h	4,292,248 (61%)	2390

Insenstive seems to bind to a new motif

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



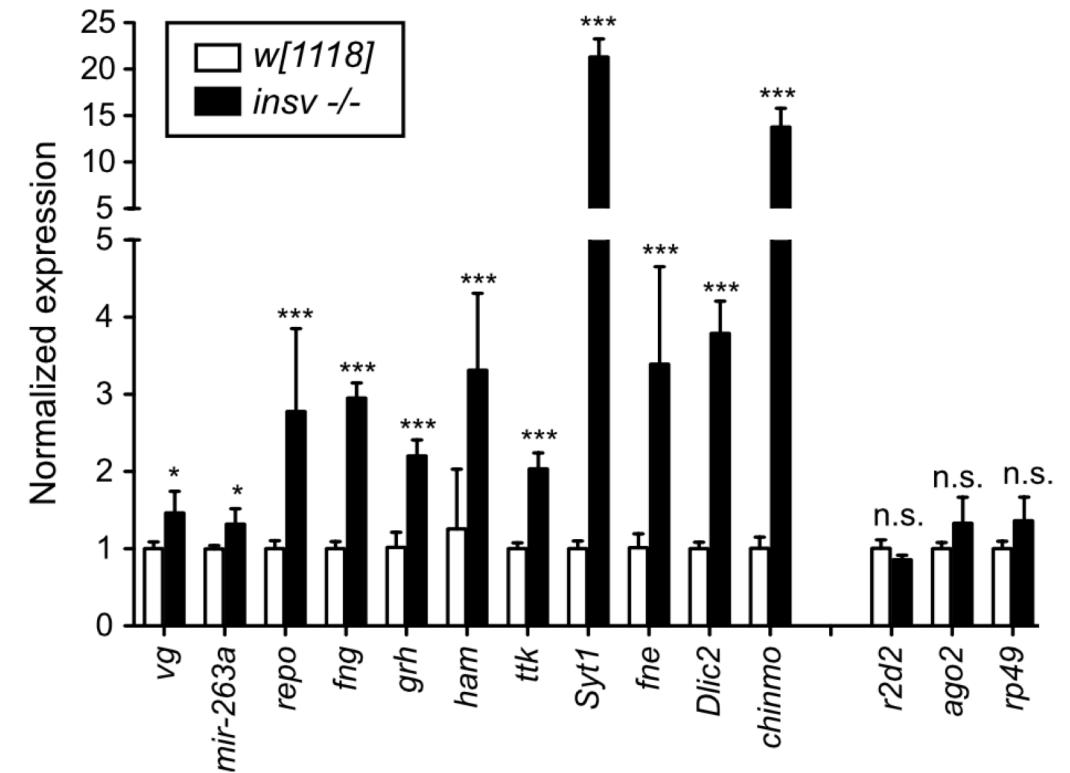
Validating peaks



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

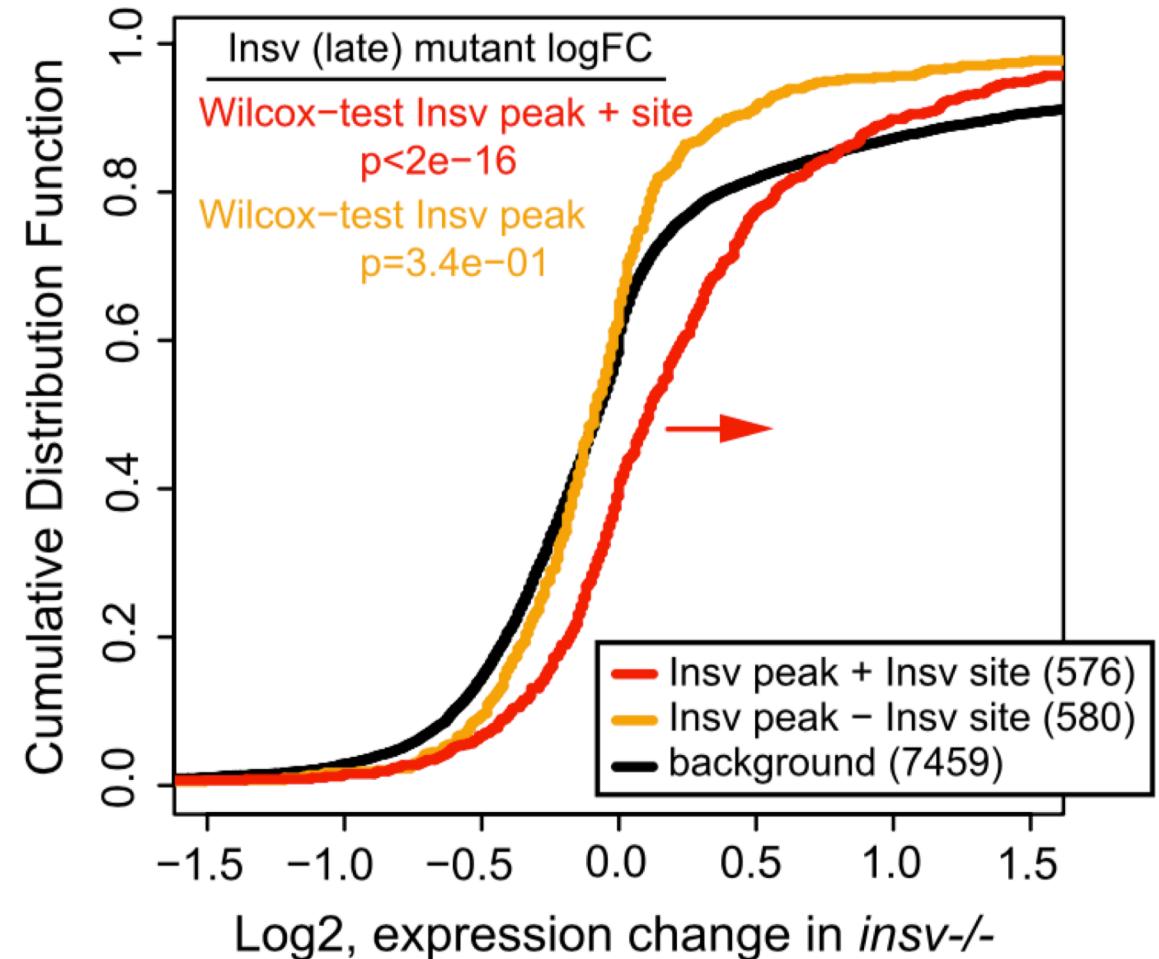
Gene expression

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



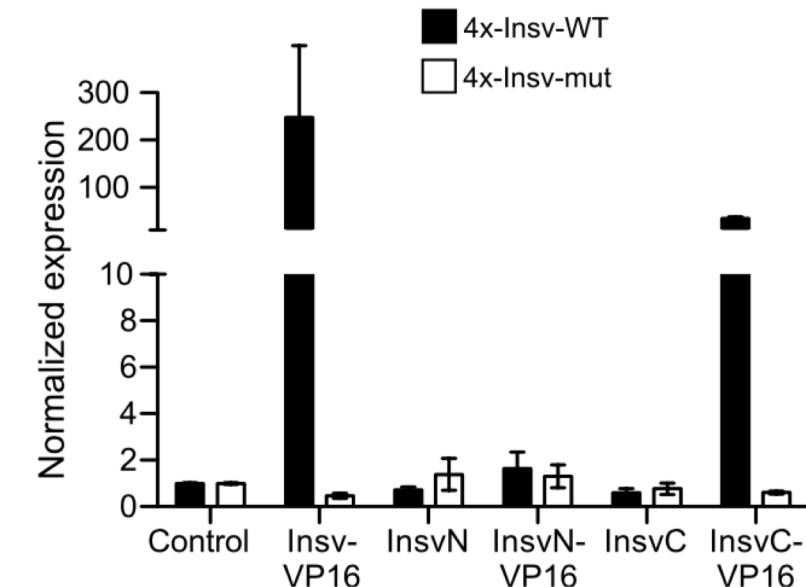
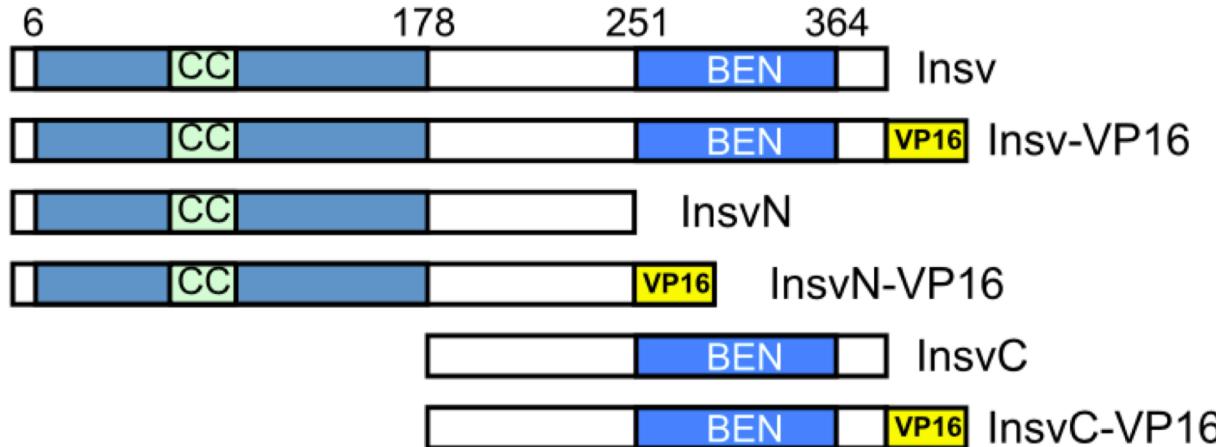
Gene expression, cont.

- 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.

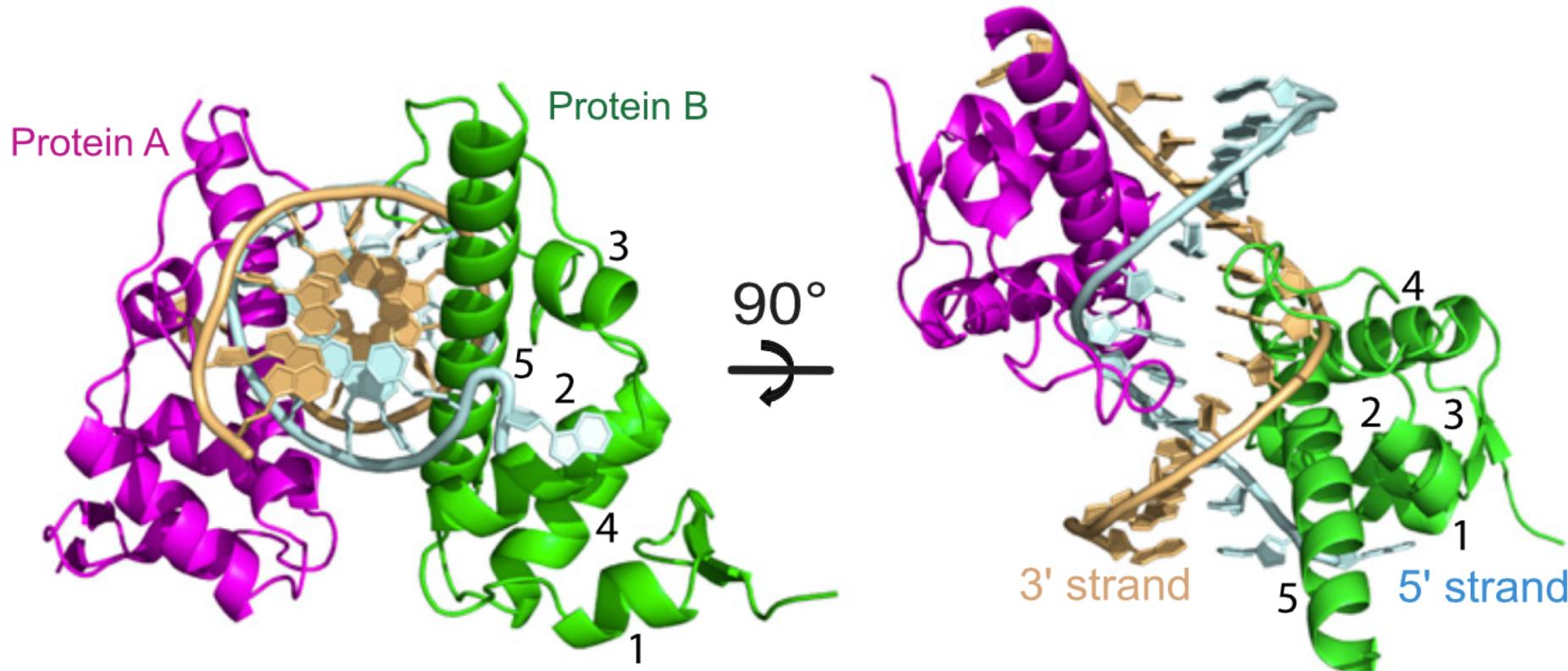


Structure-function experiments

- 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.



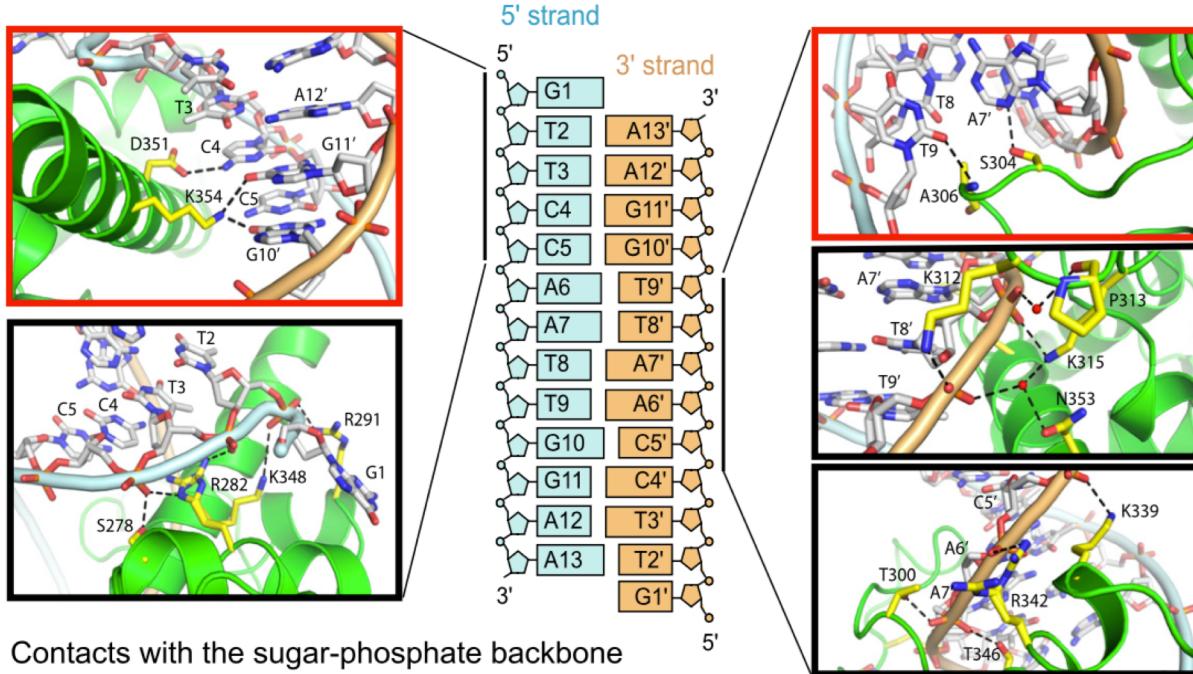
Crystal structure of BEN domain bound to DNA



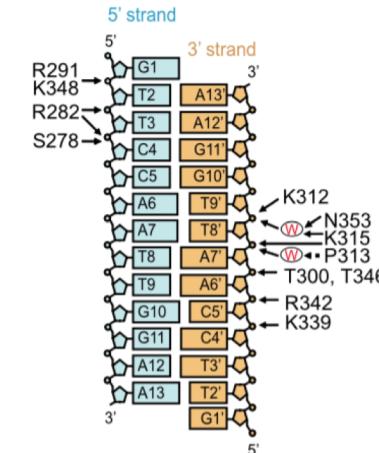
Validating the structure

- 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.

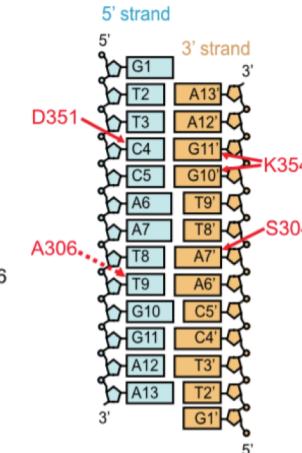
Base-specific hydrogen-bonding contacts



A BEN contacts with sugar-phosphate backbone



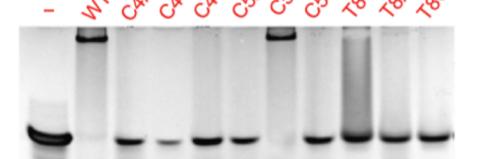
B Base-specific contacts of BEN domain



C Insv-BEN variants tested on wt probe

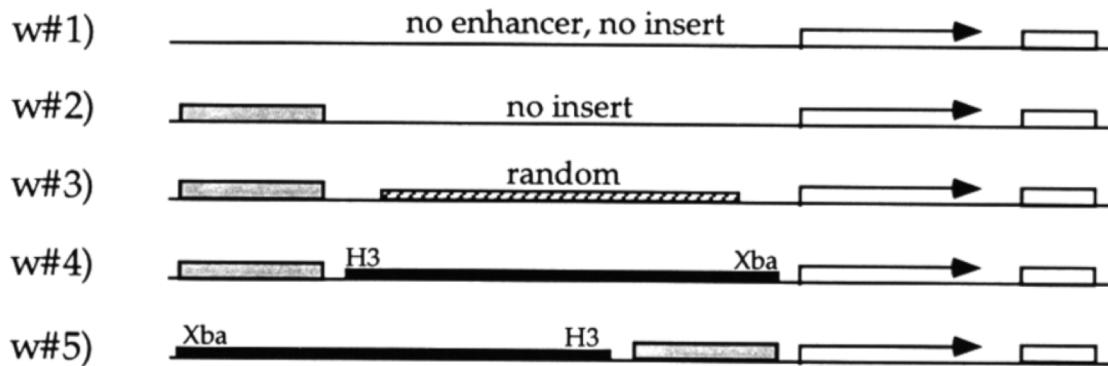


D WT Insv-BEN on variant probes



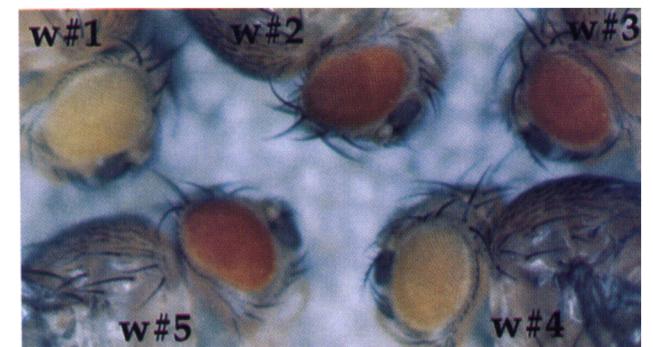
Insulator elements

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



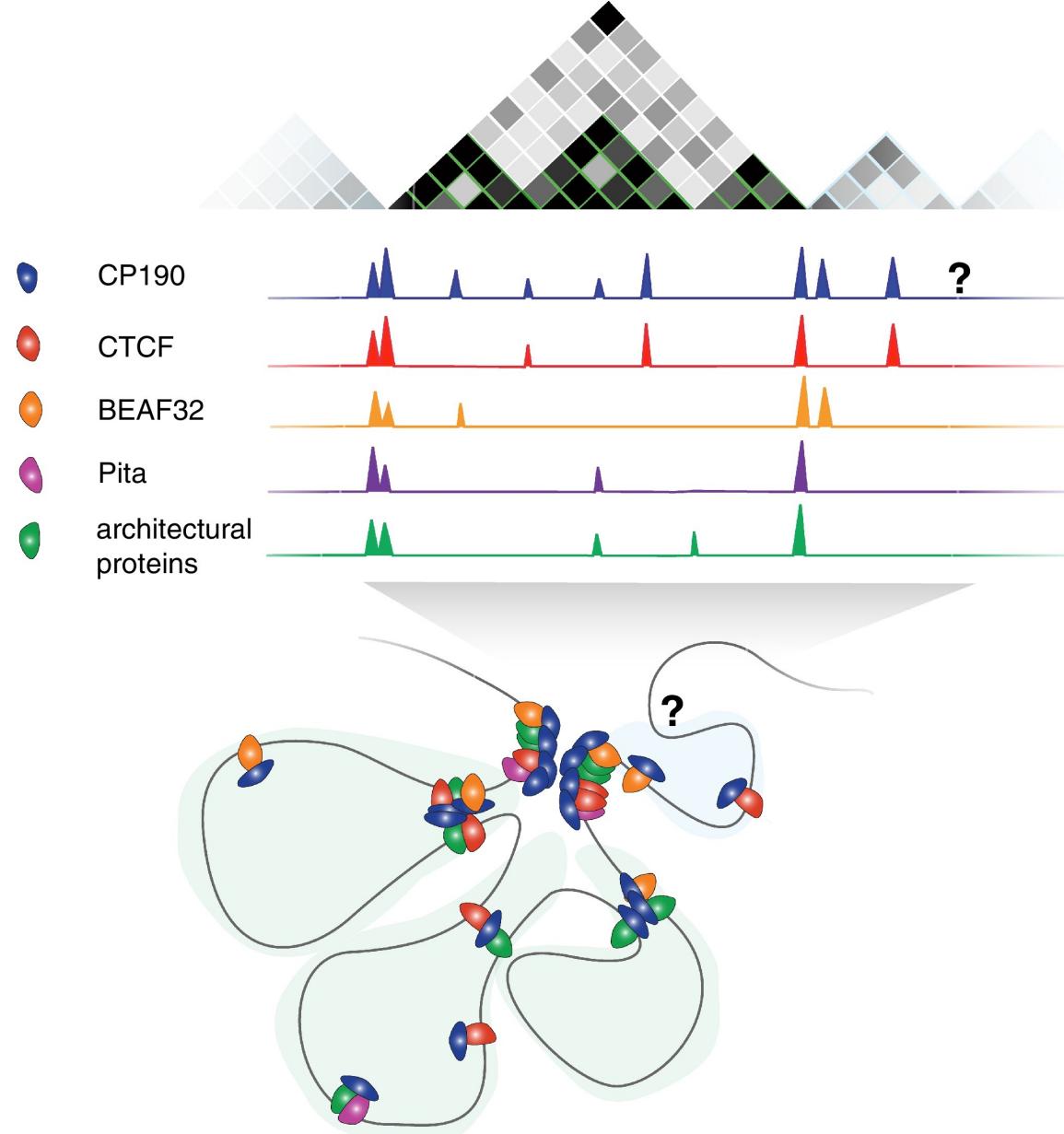
Enhancer blocking
(#light eyes/total)

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



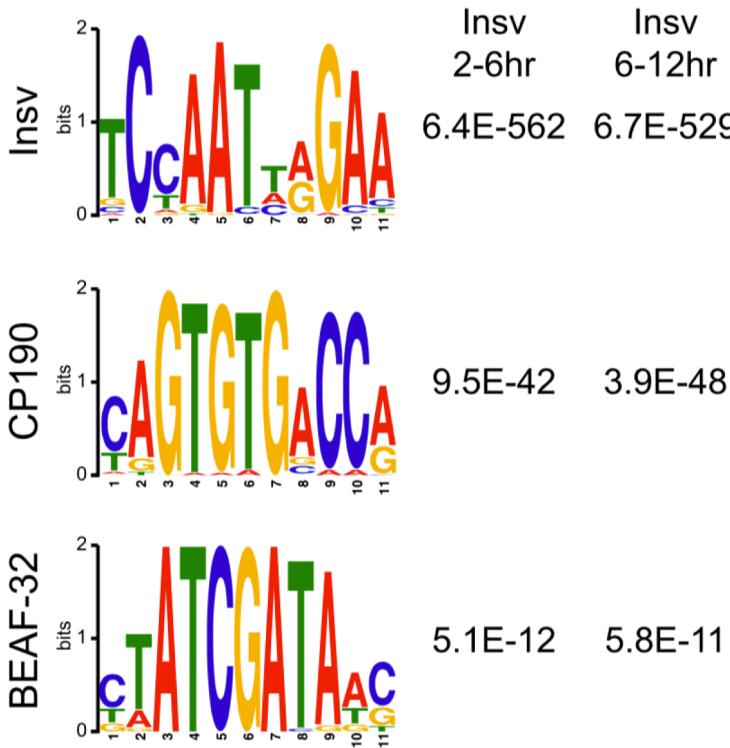
Insulator elements, cont.

- 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

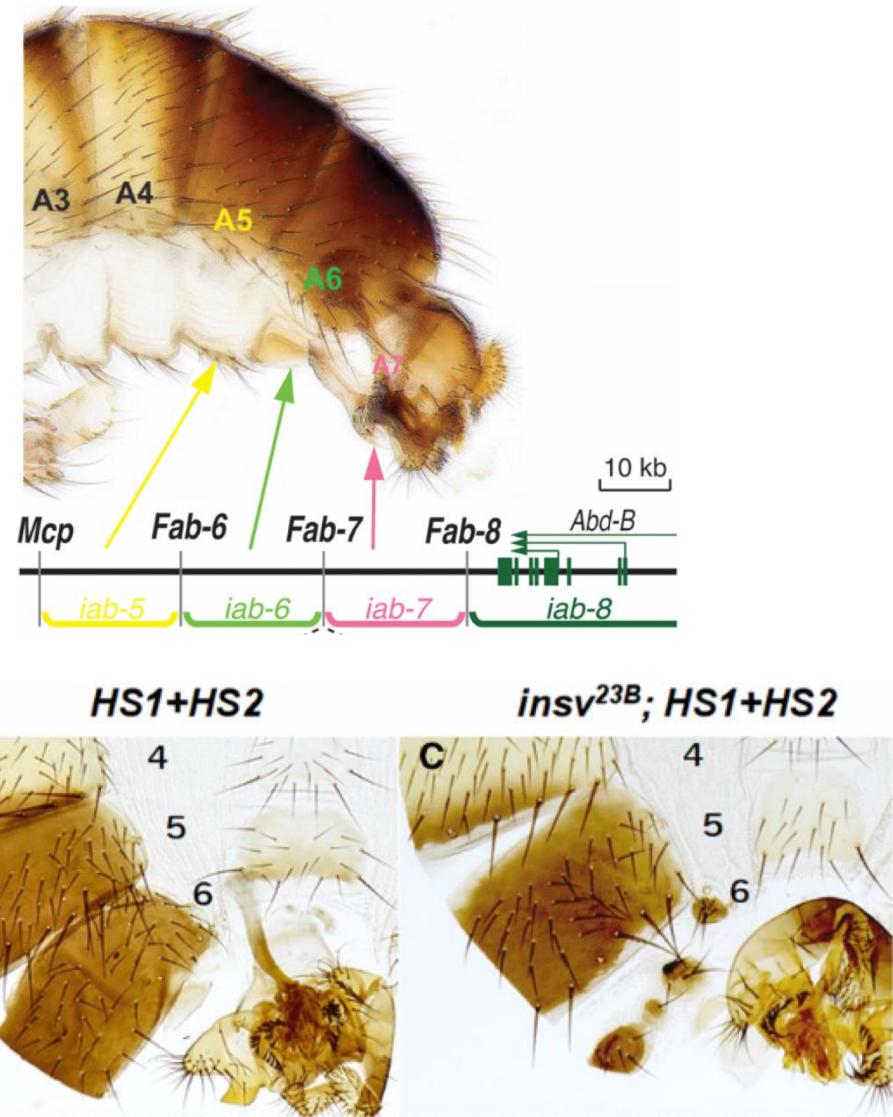
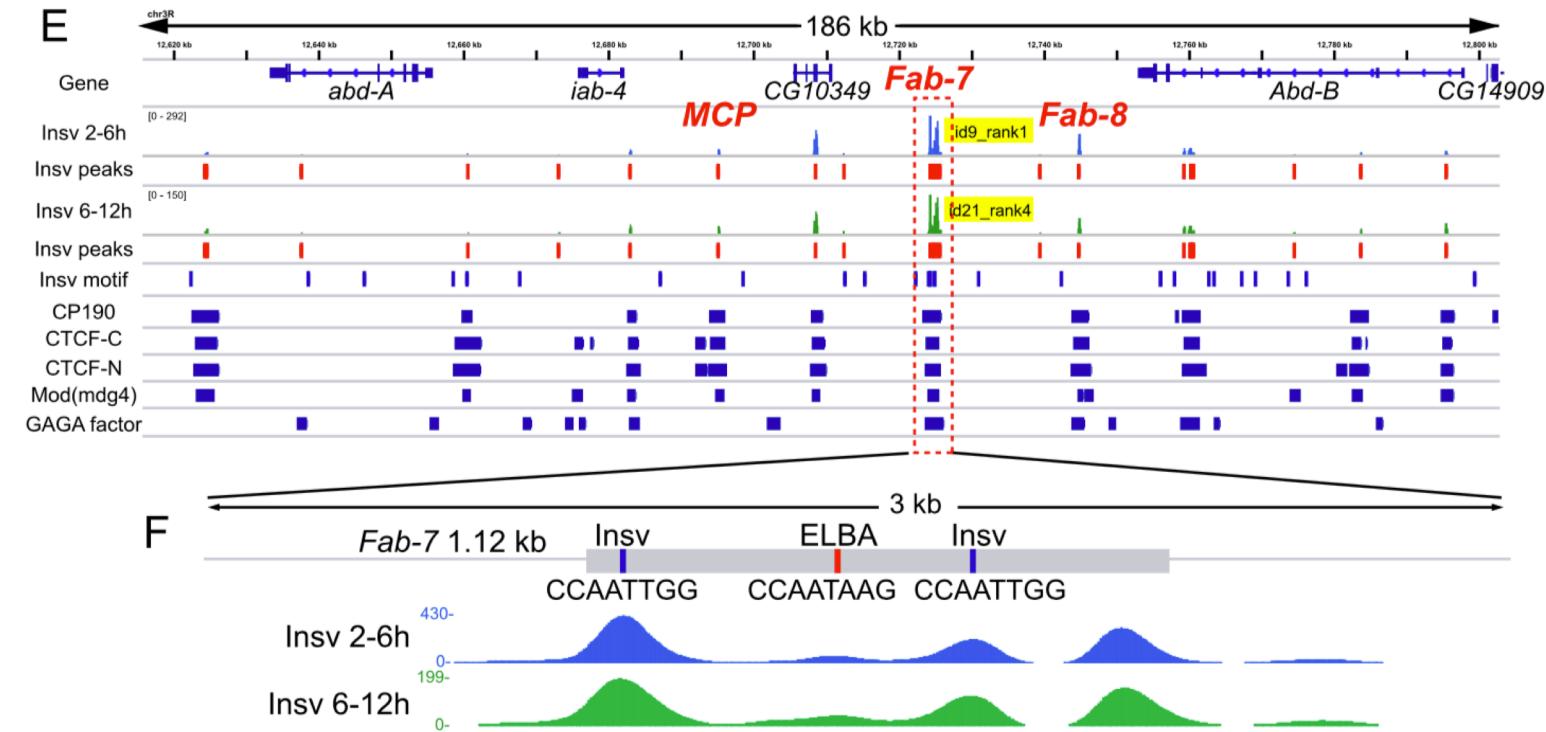


		# fraction of peaks in row data (# peaks in genome)										Insulators		Control TFs	
		Insv 2-6hr (4197)	Insv 6-12hr (1870)	BEAF-32 (4709)	CP190 (6652)	CTCF_C (3154)	CTCF_N (2532)	GAF (3904)	Mod(mdg4) (3060)	Su(hw)_1 (3420)	Su(hw)_2 (3630)	CtBP (4946)	Gro (1337)		
Insv	2-6hr (4197)	1	0.41	0.6	0.72	0.38	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)

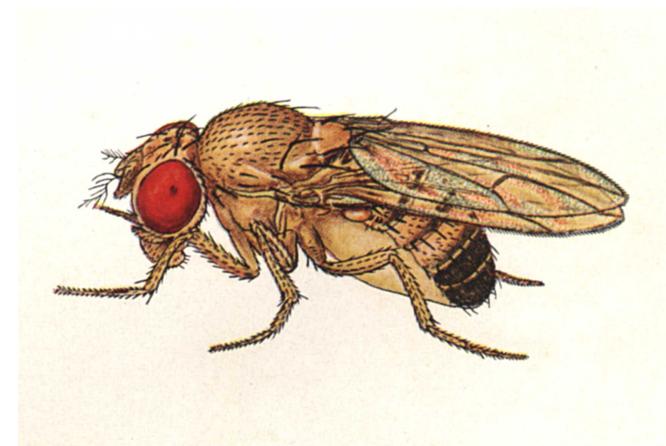
Qi Dai

Hong Duan

Dinshaw Patel (Sloan-Kettering)

Aiming Ren

Artem Serganov



Extensions of ChIP-seq

Stockholm, November 8 2018

Jakub Orzechowski Westholm

Long-term bioinformatics support

NBIS, SciLifeLab, Stockholm University

SciLifeLab



So far..

.. you have seen how to use ChIP-seq for

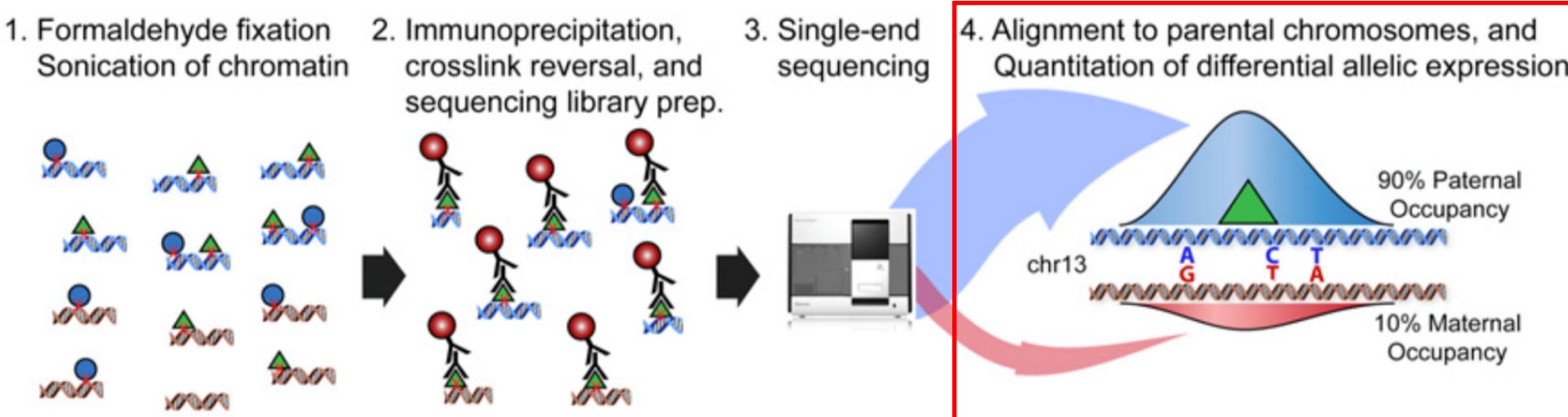
- analyzing which regions of the DNA a protein interacts with
- using a lot of material (millions of cells)

This lecture

- Allele-specific binding of transcription factors
- ChIP-seq from small numbers of cells
- Single cell ChIP-seq

Allele-specific binding

- Using ChIP-seq data it's possible to find variants that affect protein binding.
- If there are heterozygous sites, it's possible to see differences in binding to the two alleles.



Reddy et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Research 2012.

Why is this interesting?

- GWAS studies have found many mutations involved in disease and other traits in non-coding regions.
- It's harder to figure out the effect of such mutations, compared to mutations in coding regions.
- But many non-coding mutations might influence DNA binding of transcription factors or other proteins.
- It's possible to use ChIP-seq data to see which transcription factors are affected, giving an mechanism to the mutations.

Early example:

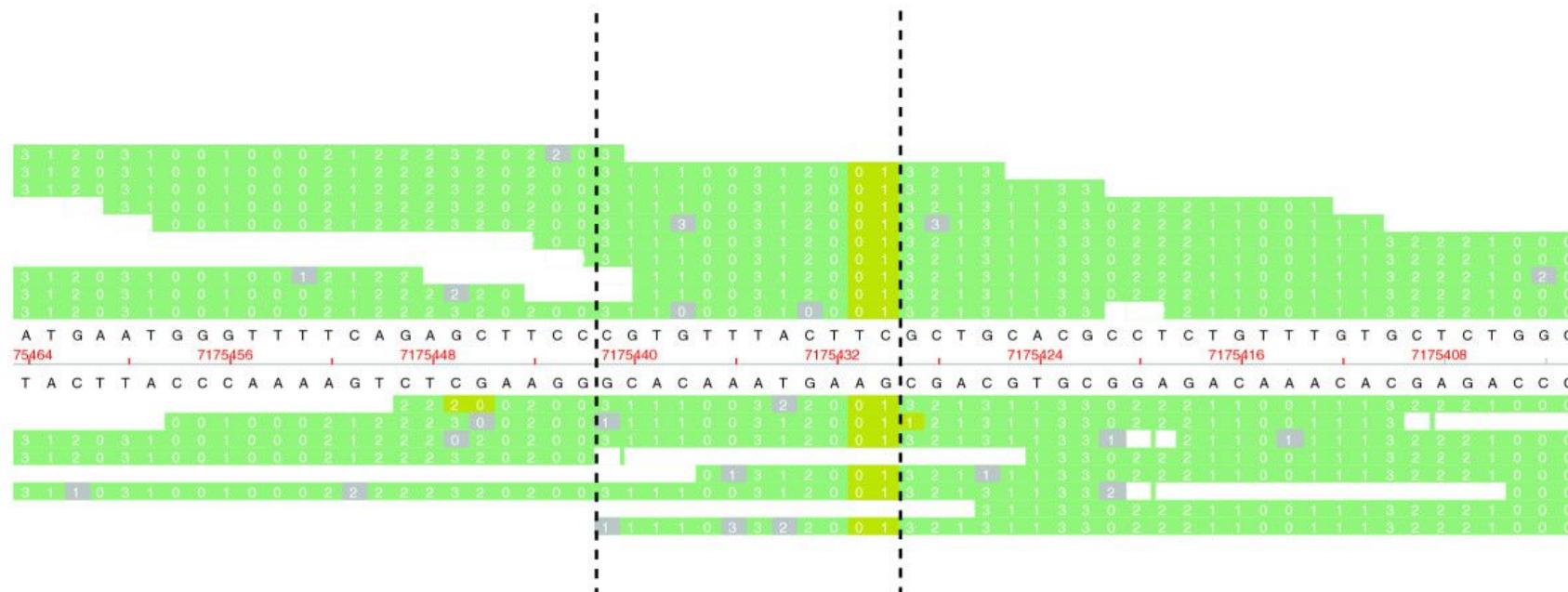
FOXA1 top motif:



HepG2 genomic DNA: CGTGTTTACTT[T/C]

Reference genome: CGTGTTTACTTC

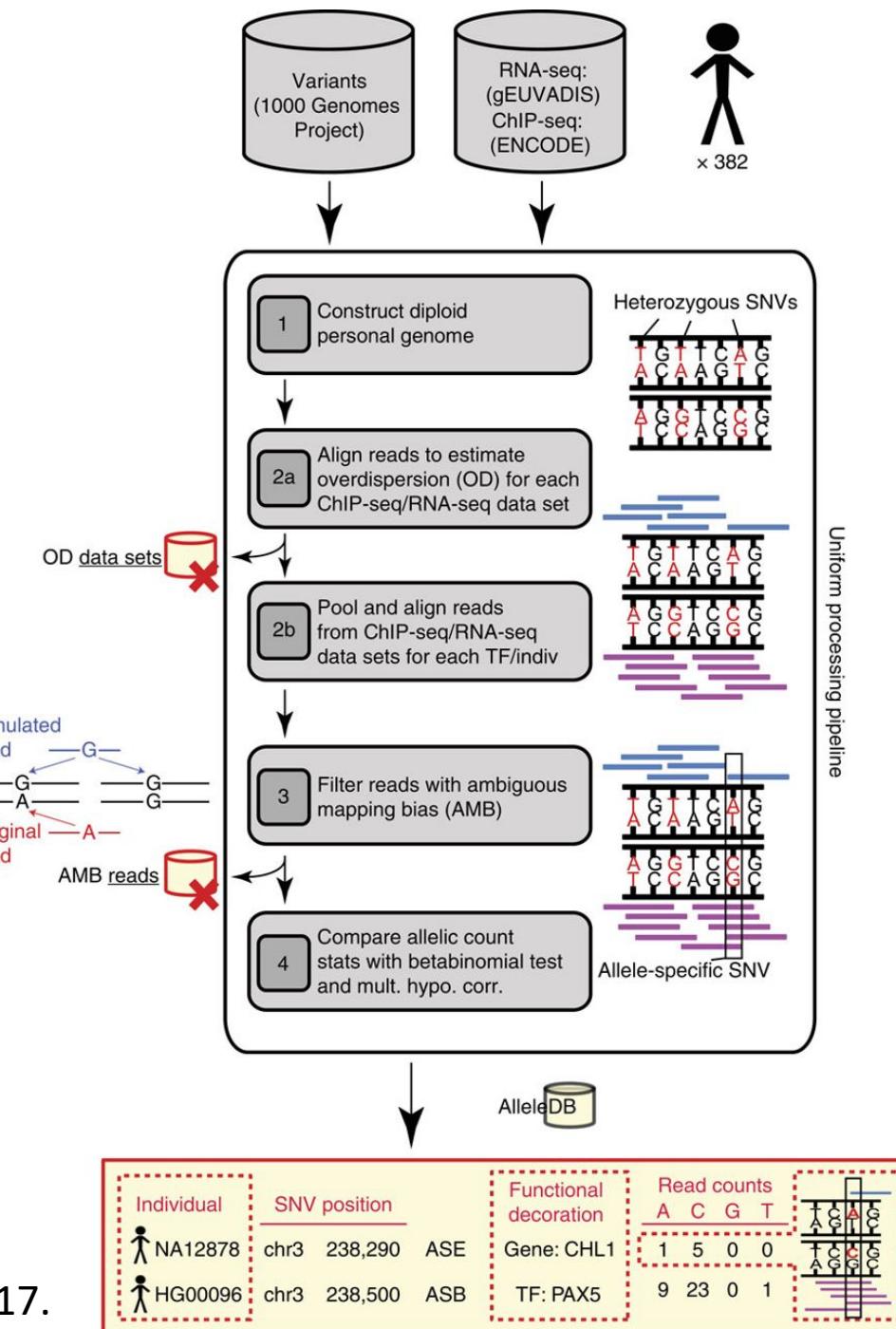
FOXA1 ChIP DNA: CGTGTTTACTTT



Motallebipour et al. Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. Genome Biology 2009.

Procedure

- Need reference genome. Otherwise heterozygous regions where the TF only binds to one allele are missed.
- Need good way to call variants and avoid biases when mapping reads
 - SNVs are easy
 - Small indels also quite easy
 - Large variations harder
- Binomial test for differential binding.



Chen et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nature Communications 2017.

Overall results:

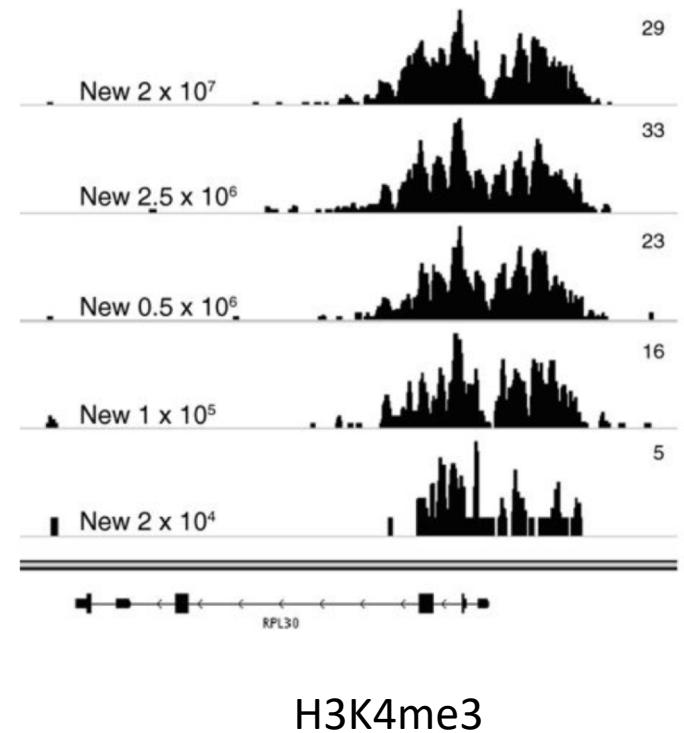
- 1-11% of sites have been reported to have allele specific binding (MacDaniell 2010, Rozowski 2011, Reddy 2012)
- Resolution: enrichment for mutations within 50bp of highest point of peak (Reddy 2012)
- TF binding is strongly heritable, more than gene expression (MacDaniell 2010, Reddy 2012, Chen 2017)
- Sites with allele specific binding were significantly enriched for variants associated with disease. (Reddy 2012)
- Some mutations hit the transcription factor motif, but most do not. (Reddy, 2012)
 - other mechanisms for transcription factor recruitment. Co-factors?

Low input ChIP-seq

- Usually ChIP-seq requires a lot of starting material: around 1-10 million cells
- This is a problem when we want to study rare cell types/populations
 - Nervous system
 - Cancer
 - ..

Methods for low input ChIP-seq

- Native ChIP - no cross-linking
- Micrococcal nuclease
- Gilfillan et al. Limitations and possibilities of low cell number ChIP-seq, BMC Genomics 2012
 - Down to 100,000 cells with good quality
 - down to 20,000 cells with ok quality
- Brind'Amour et al. Ultra-low-input native ChIP-seq for rare cell populations. Protocol Exchange, 2015
 - Down to 1000 cells



Application with low cell numbers

- Rare neural cell populations:
 - Midbrain dopamine-producing neurons
 - 20,000–30,000 cells per mouse, yield when sorting cells is around 5000 cells
- If we need 1 millions cells per ChIP, it would take over 200 mice
- Now one mouse gives enough cells for 3 ChIPs + input + RNA-seq



Corrected: Author correction

ARTICLE

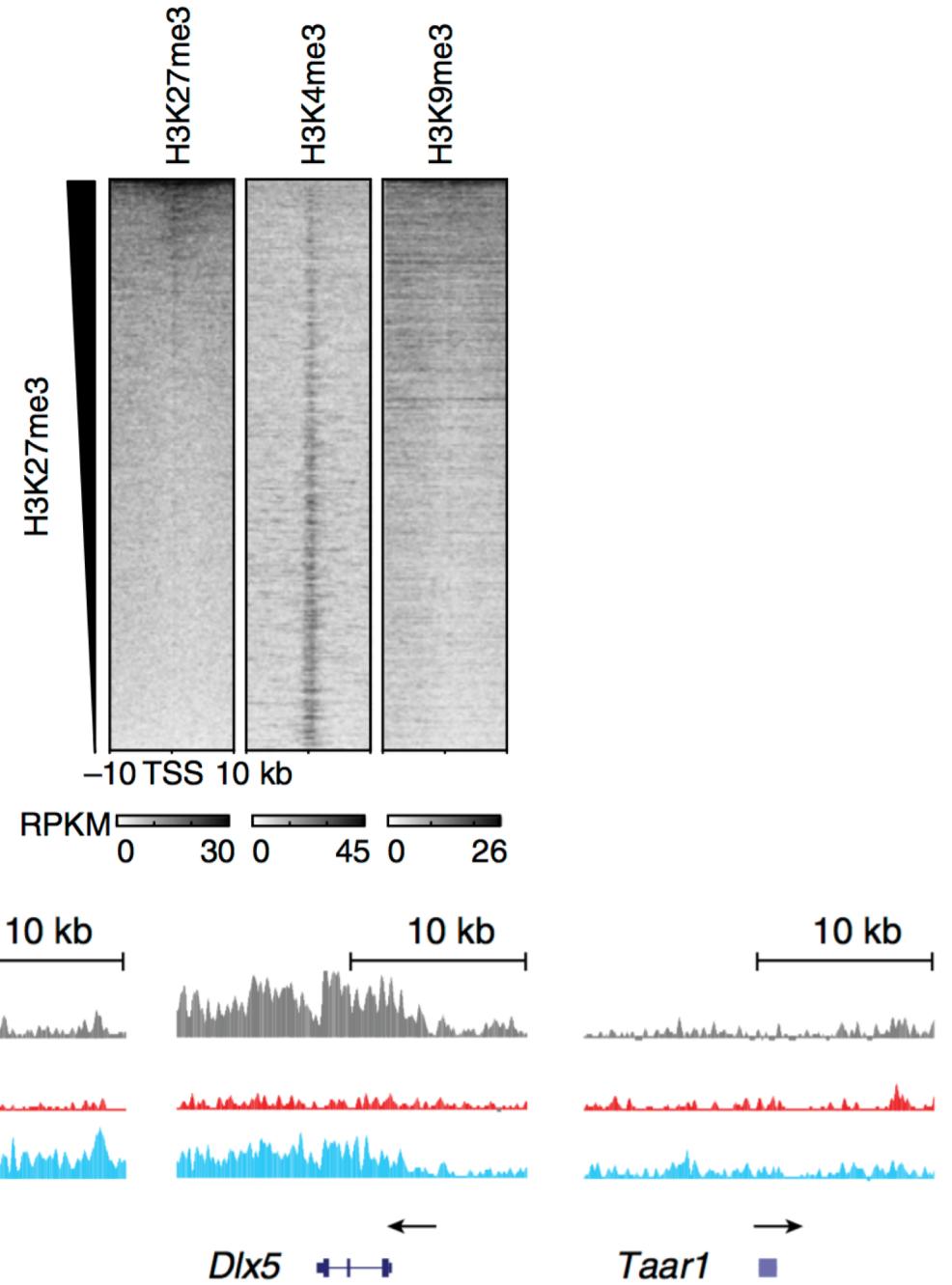
DOI: 10.1038/s41467-018-03538-9

OPEN

A comprehensive map coupling histone modifications with gene regulation in adult dopaminergic and serotonergic neurons

Erik Södersten¹, Konstantinos Toskas¹, Vilma Rraklli¹, Katarina Tiklova¹, Åsa K. Björklund^{1,2}, Markus Ringnér^{1,3}, Thomas Perlmann¹ & Johan Holmberg¹

- They were able to get useful data for 3 histone marks.
- Also comparison with RNA-seq data.
- No big changes to analysis
 - Some quality measures might not look as good, e.g. duplication rates
 - QC even more important!



Single cell ChIP-seq

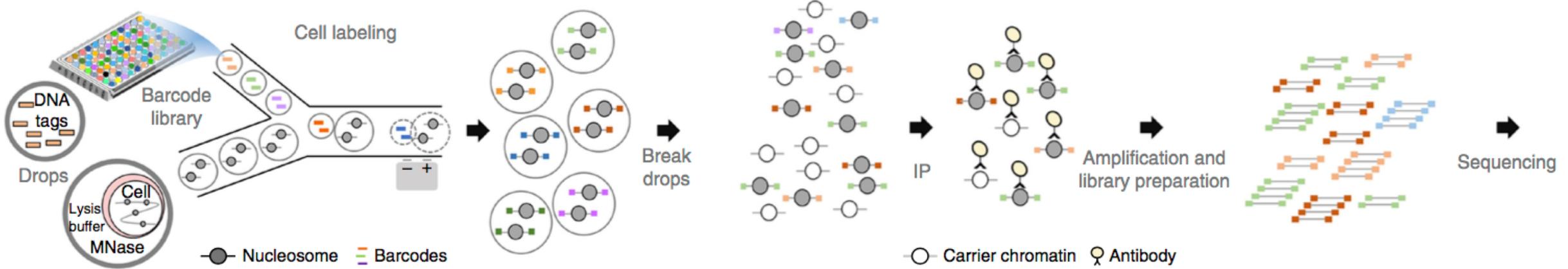
- The signal we get from normal ChIP-seq is an average over all cells in the sample
- This misses heterogeneity
 - Cell types
 - Primed vs unprimed cells
 - Response to stimuli
- With single cell ChIP-seq, we get data for each cell separately
- This is similar to single cell RNA-seq, but much harder (since we only have two chromosome copies, compared to many RNA molecules).

nature
biotechnology

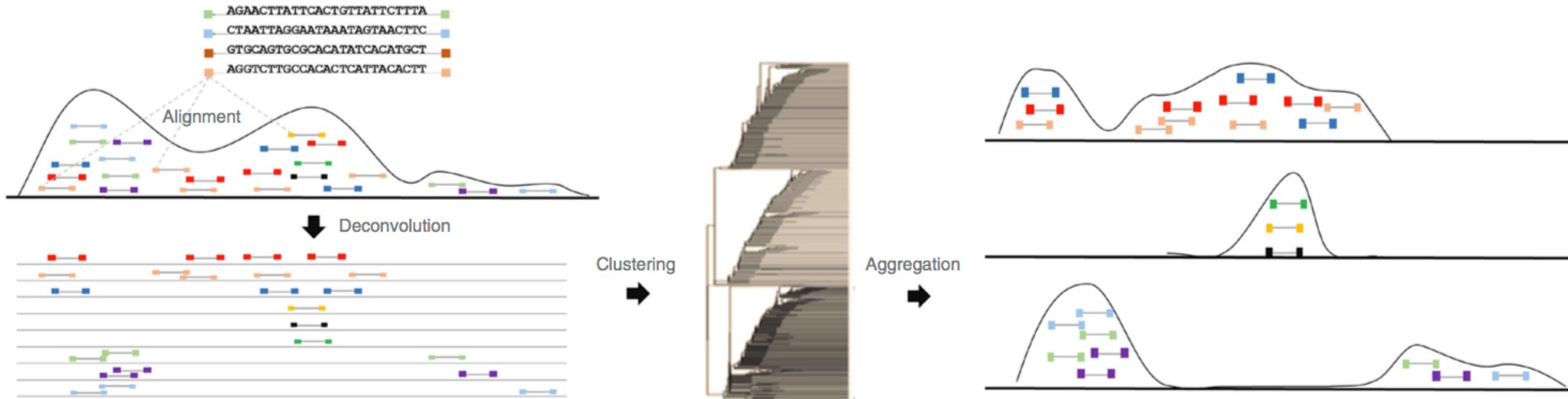
Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state

Assaf Rotem^{1,2,7}, Oren Ram^{2–4,7}, Noam Shores^{2,7}, Ralph A Sperling^{1,6}, Alon Goren⁵, David A Weitz¹ & Bradley E Bernstein^{2–4}

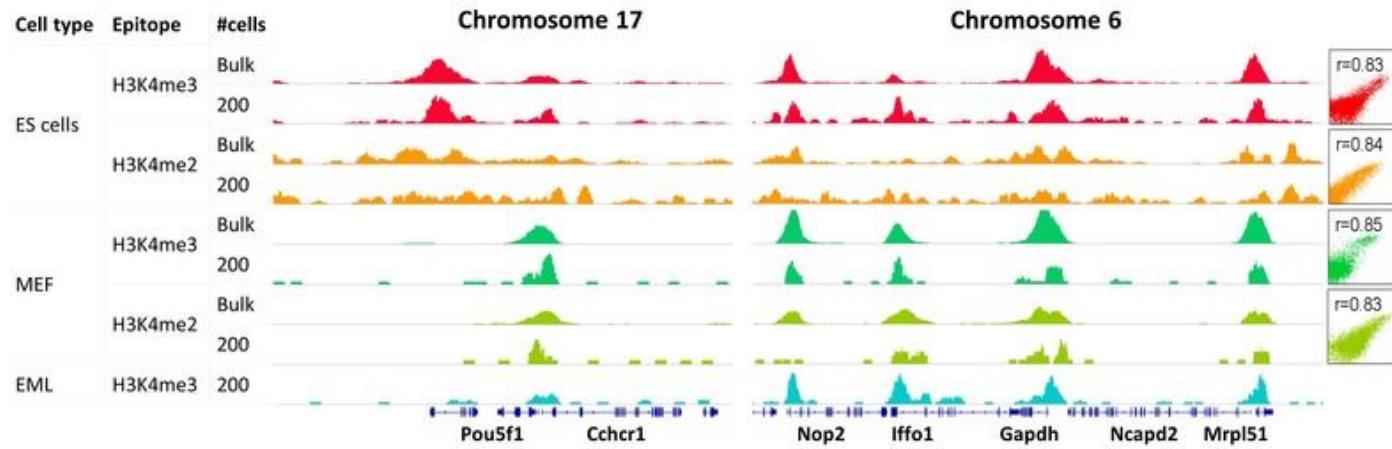
Experiment overview



Analysis overview



Aggregated single cell vs bulk data



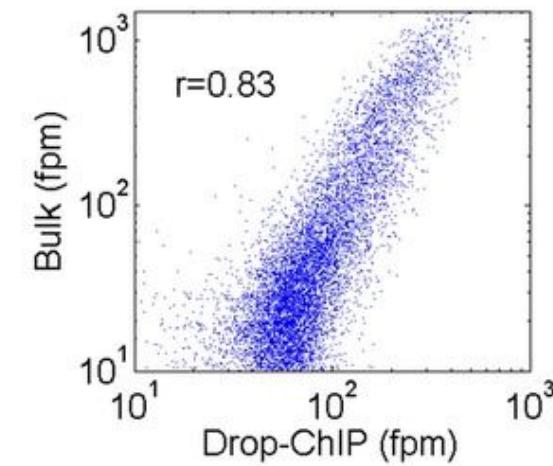
Drop-ChIP
13,900
promoters

Bulk
13,300
promoters

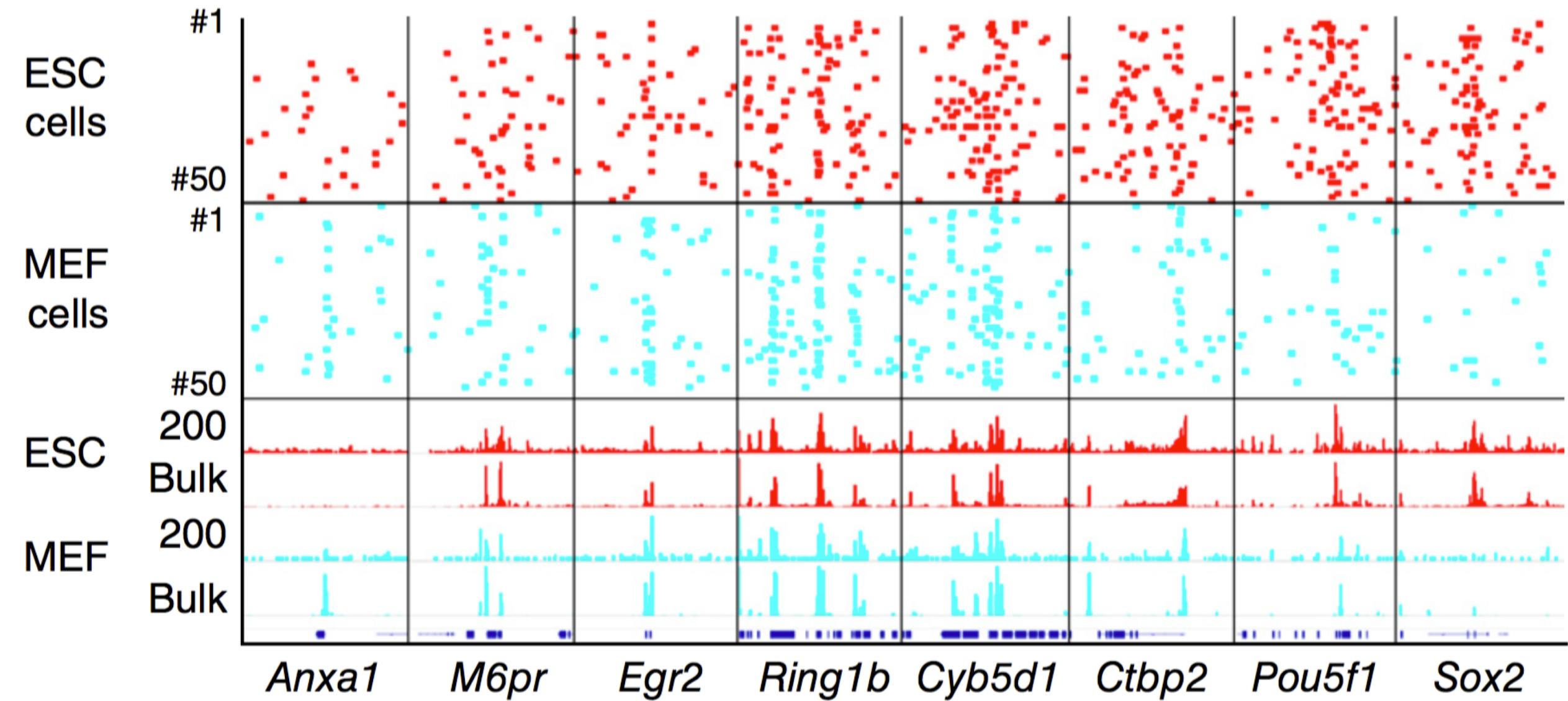


Drop-ChIP
11,100
enhancers

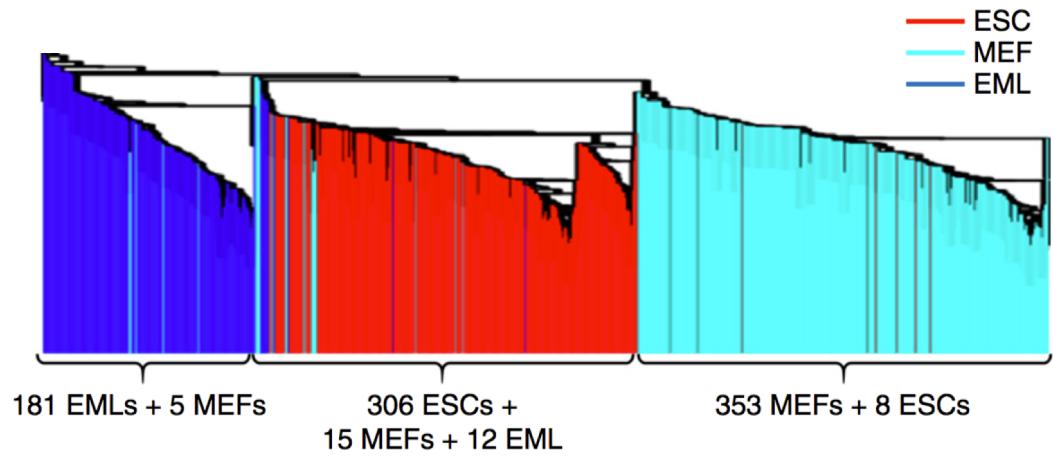
Bulk
8,400
enhancers



Data from individual cells

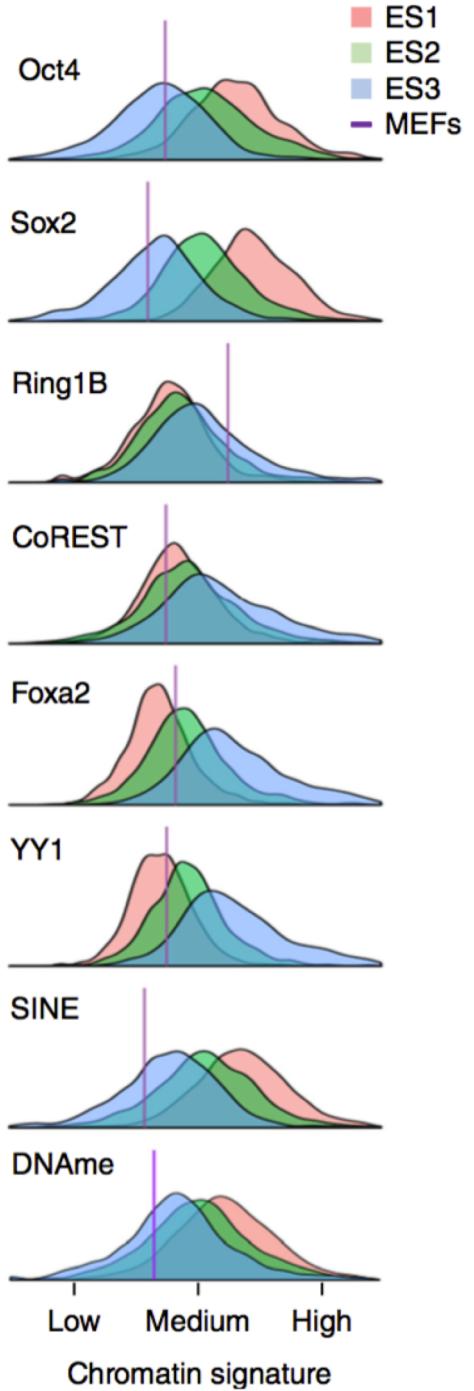
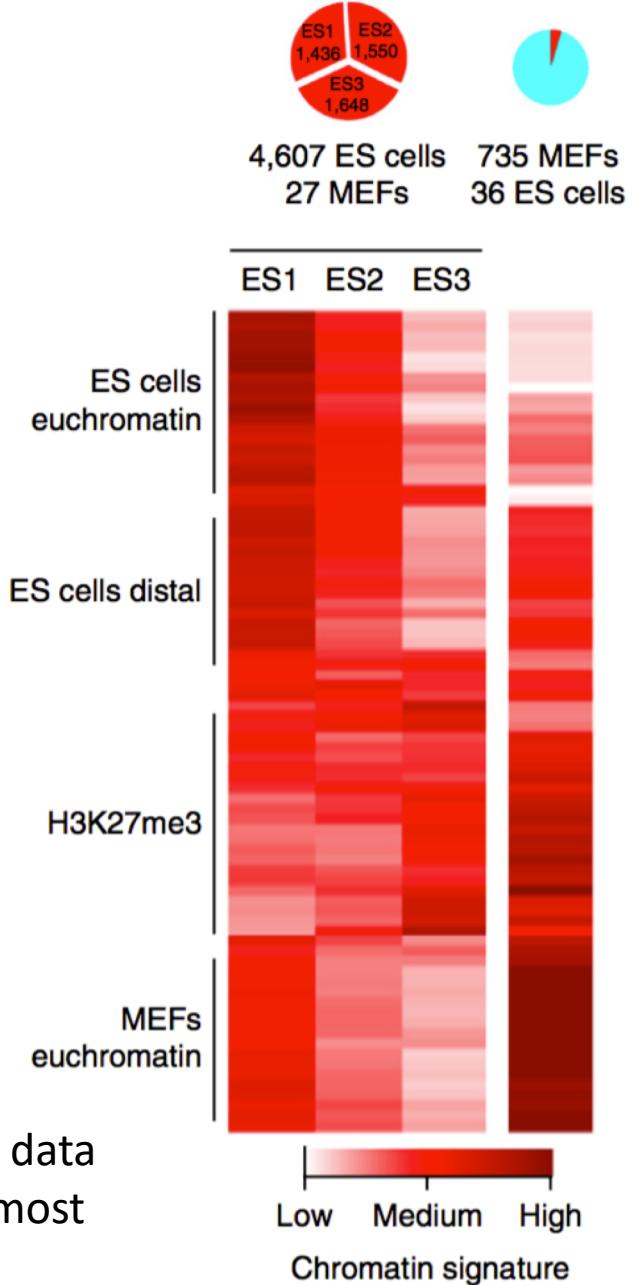


Clustering of single cells



Using promoters and enhancers
→ Possible to separate cell types

Using “chromatin signatures” derived from other data
→ Also possible to separate subpopulations (E1 most pluripotent, then E2m then E3)



Conclusions

- Works
 - Aggregate data look good
 - It's possible (but not easy!) to cluster cells, and find new cell types
- Data from each cell is very sparse
 - This is still enough to cluster cells
 - But this may not be good enough for studying rare cell types
- (Other single cell methods are getting more popular
 - ATAC-seq
 - Bisulphite seq, for DNA methylation).

The End