Case study: Finding a new DNA binding domain

Stockholm, November 8 2018

Jakub Orzechowski Westholm

Long-term bioinformatics support

NBIS, SciLifeLab, Stockholm University

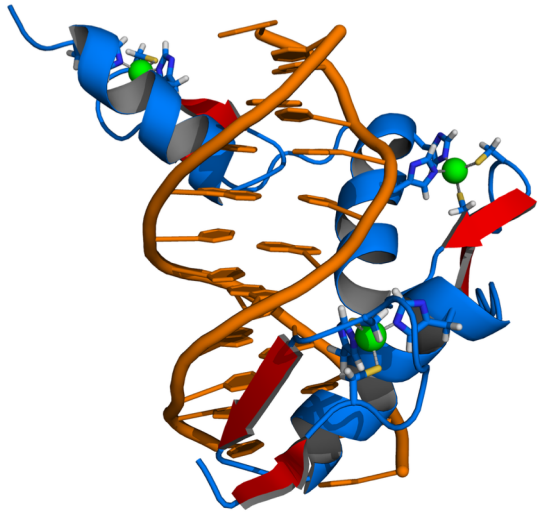
SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

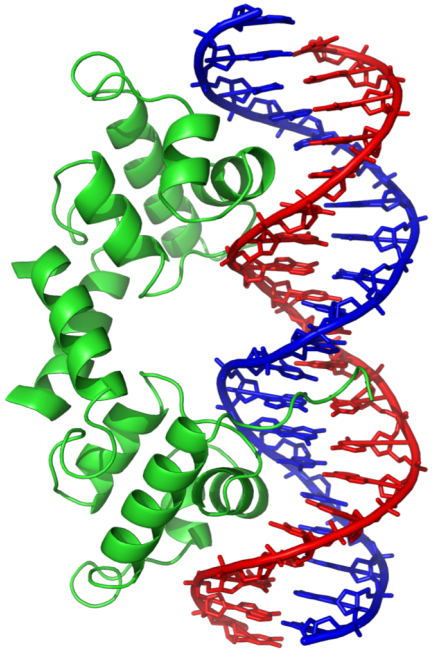
Transcription factors

- Transcription factors typically consist of
  - Activation/repression domains
  - A sequence specific DNA binding domain

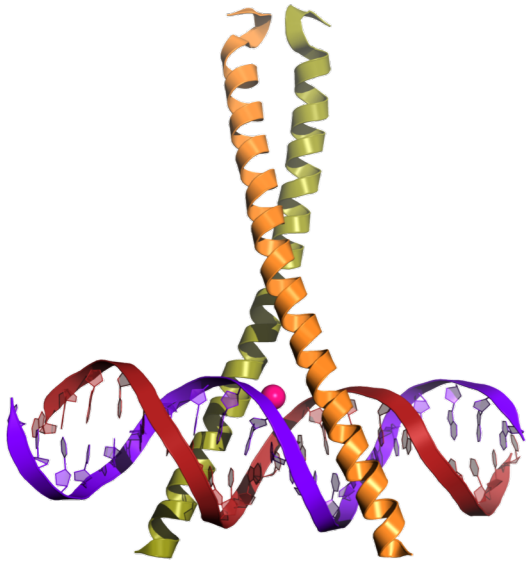
- The number of such DNA binding domains in eukaryotes is limited:
  - Less than 40 (**Yusuf** et al. *The Transcription Factor Encyclopedia*. Genome Biology 2012)



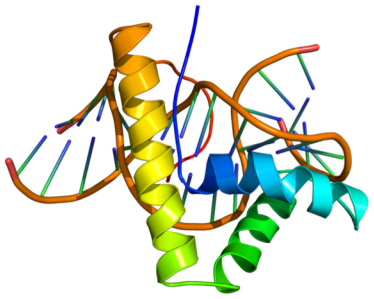
zinc finger



helix-turn-helix



basic leucine zipper



high mobility group box

BEN domains

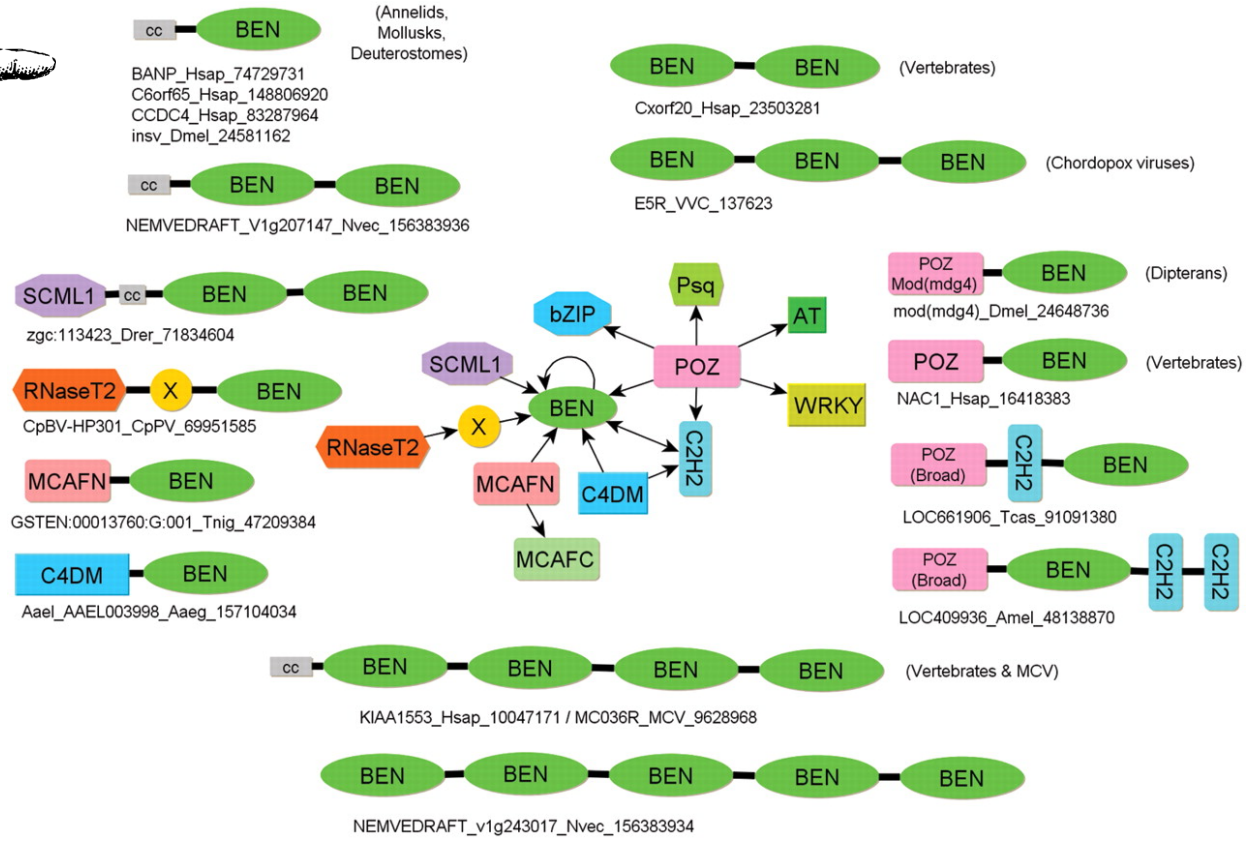
- Over 100 proteins across animals/metazoans and viruses have BEN domains.

"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).

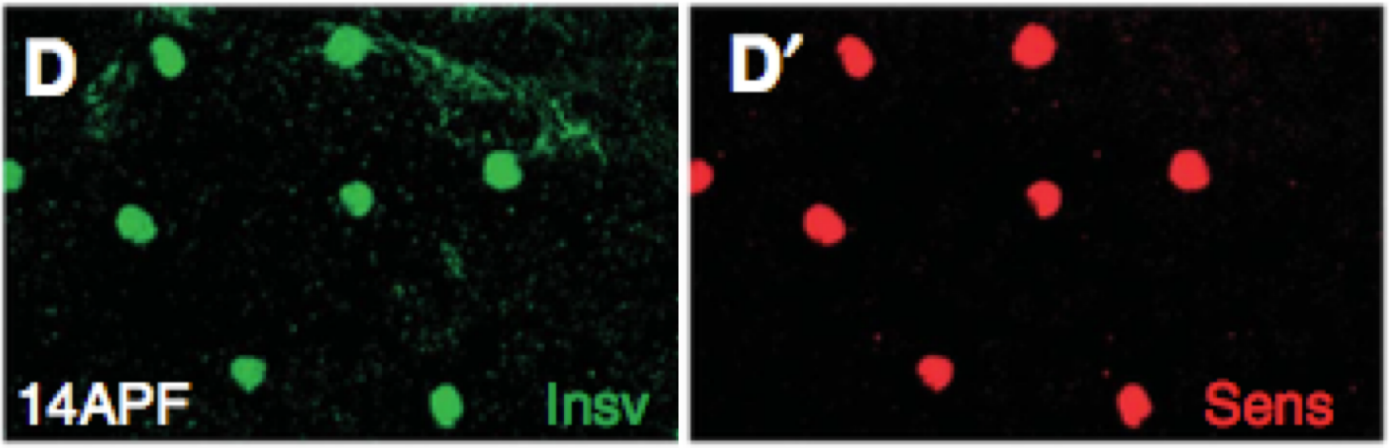


Insensitive protein

- 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

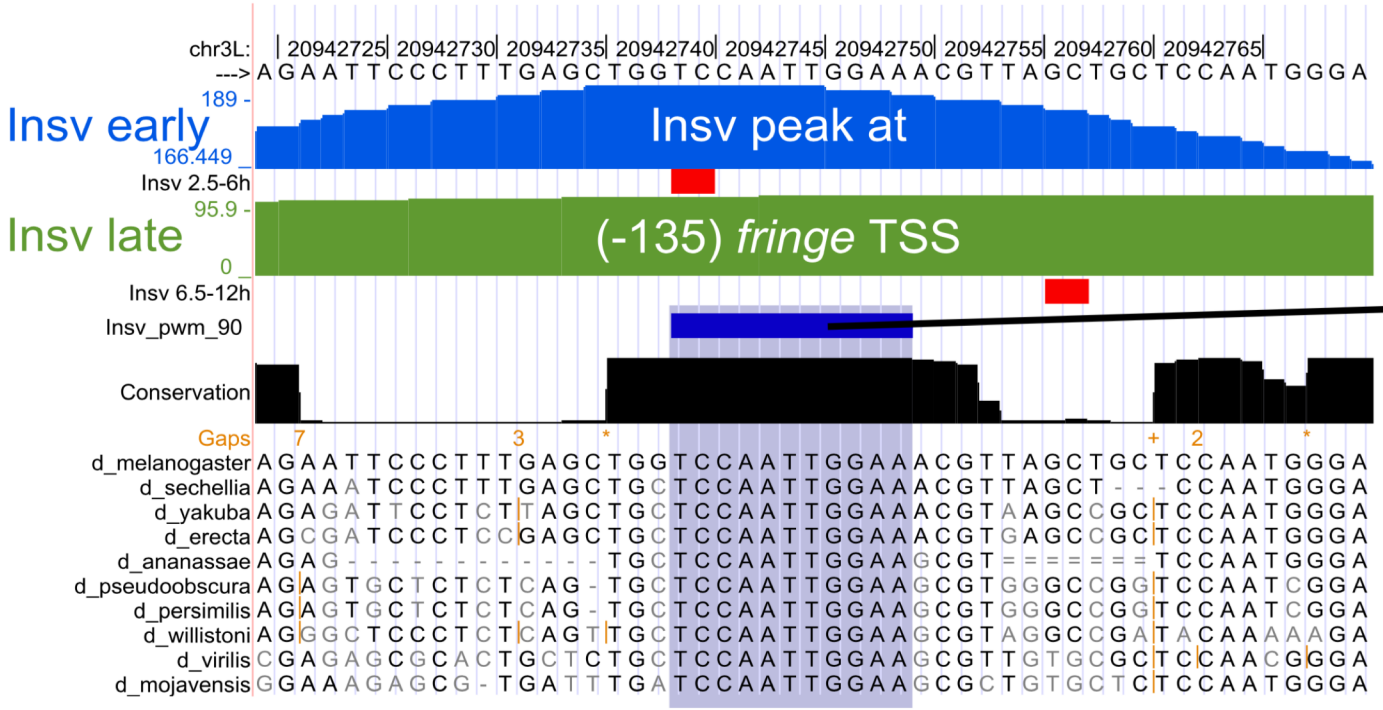
ChIP-seq experiment

- 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..

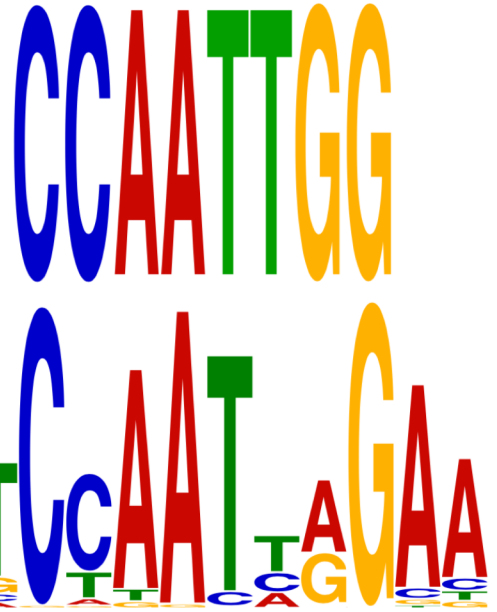
AB	Time	Unique reads mapping	Nr peaks
Insv	2.5-6h	7,473,521 (58%)	5364
Insv	6.5-12h	4,292,248 (61%)	2390

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:

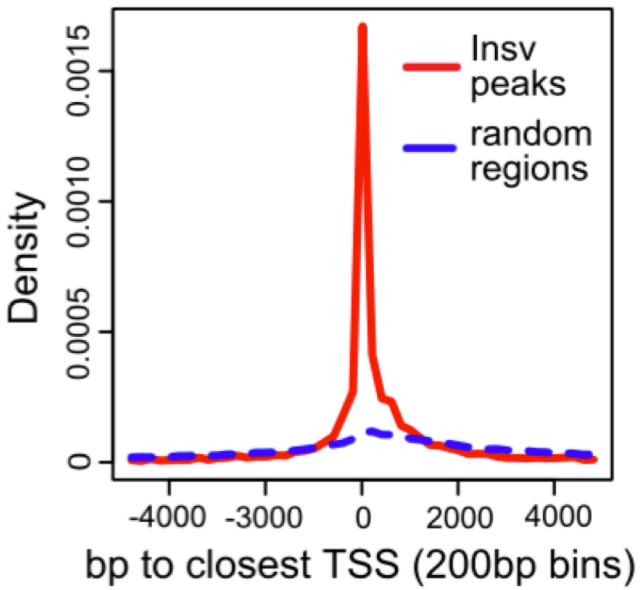


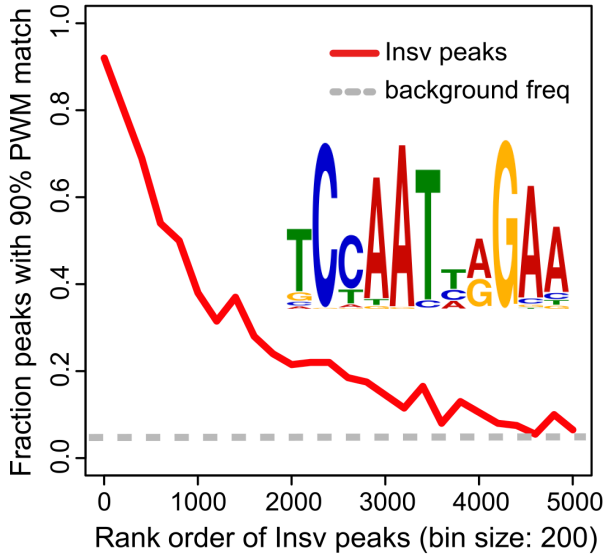
MEME:

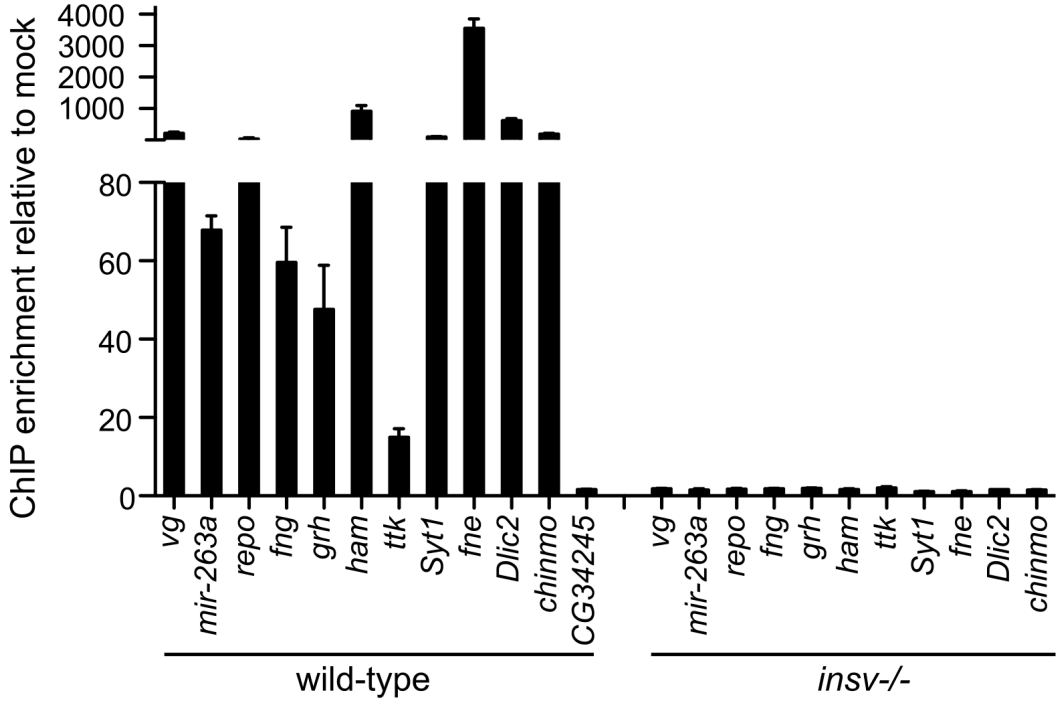
p-value = 6.4E-562

Dai et al. *The BEN domain is a novel sequence-specific DNA-binding domain conserved in neural transcriptional repressors*. Genes & Development, 2013.

Validating peaks



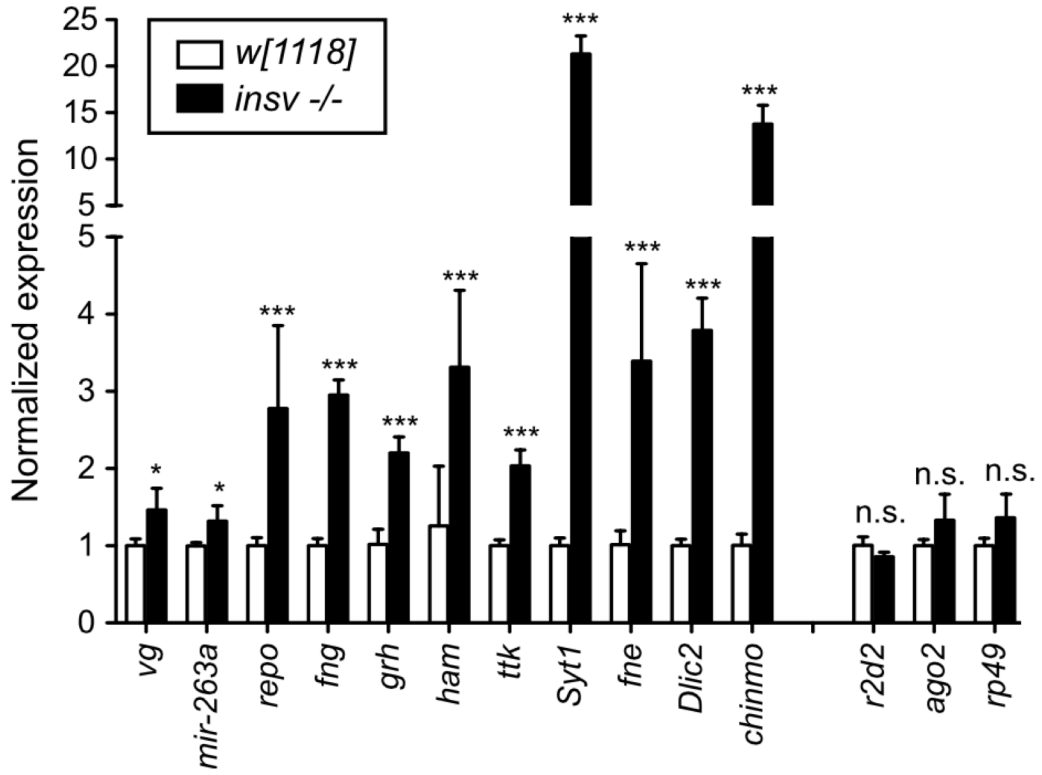




- *Insenstive* peaks are located at promotor regions
- Almost all the top *Insenstive* sites have the motif.
- ChIP-PCR validation of some peaks.

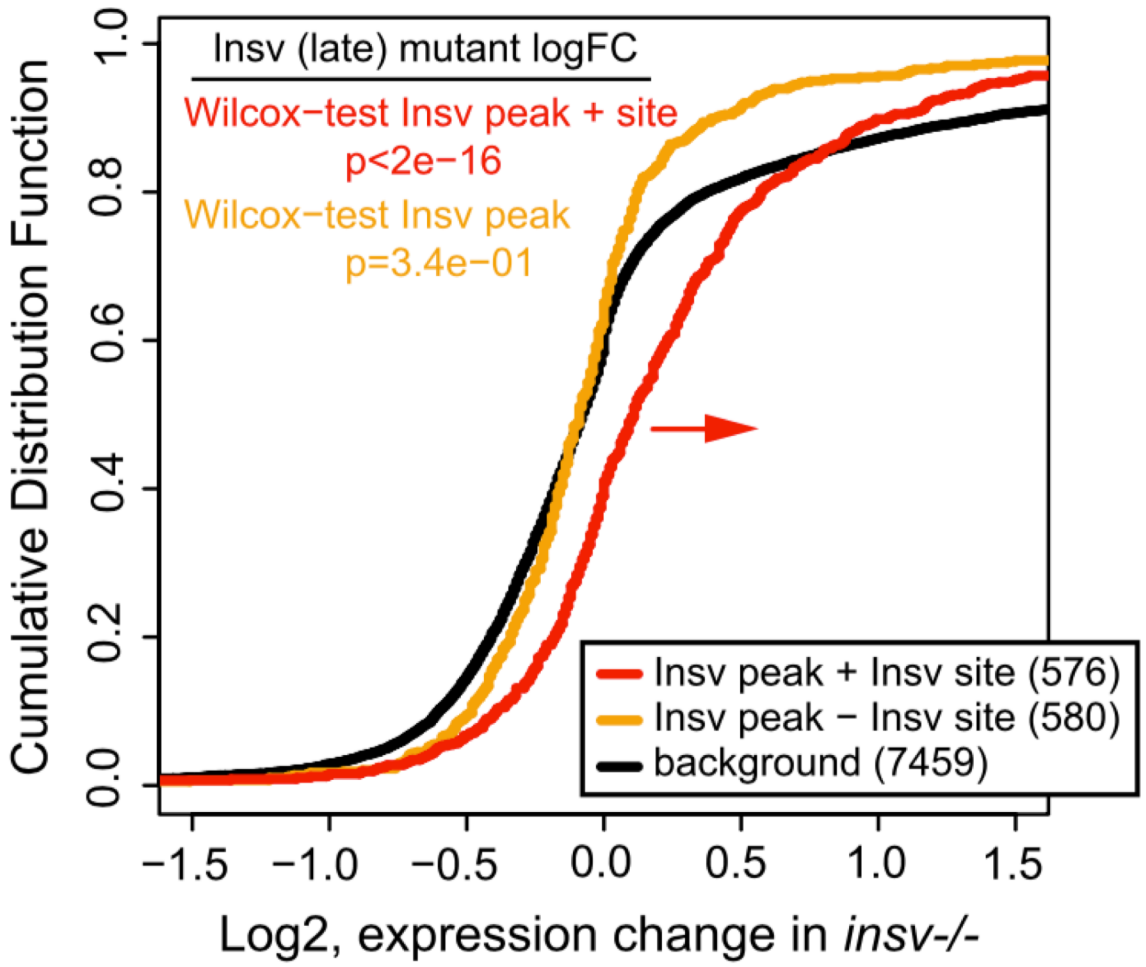
Gene expression

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



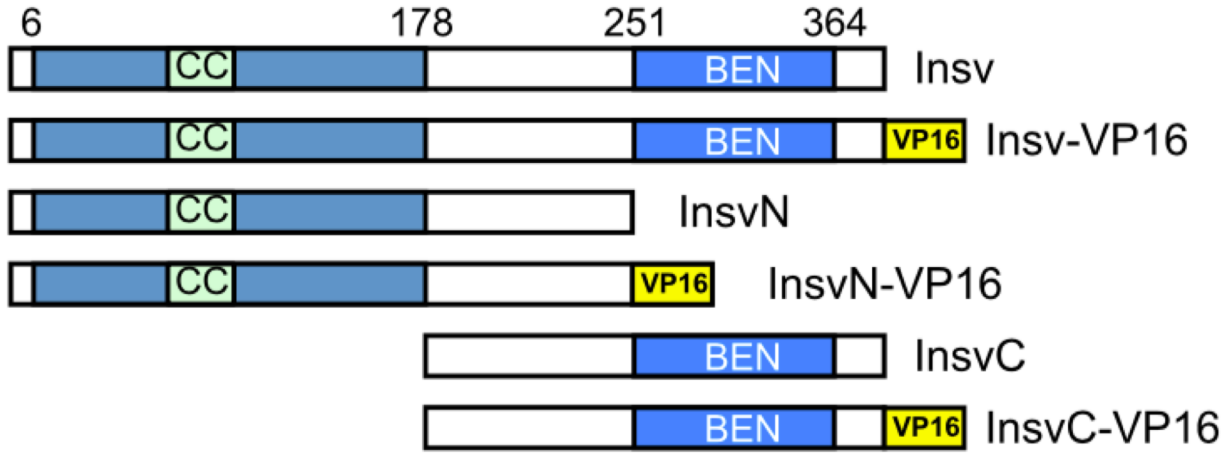
Gene expression, cont.

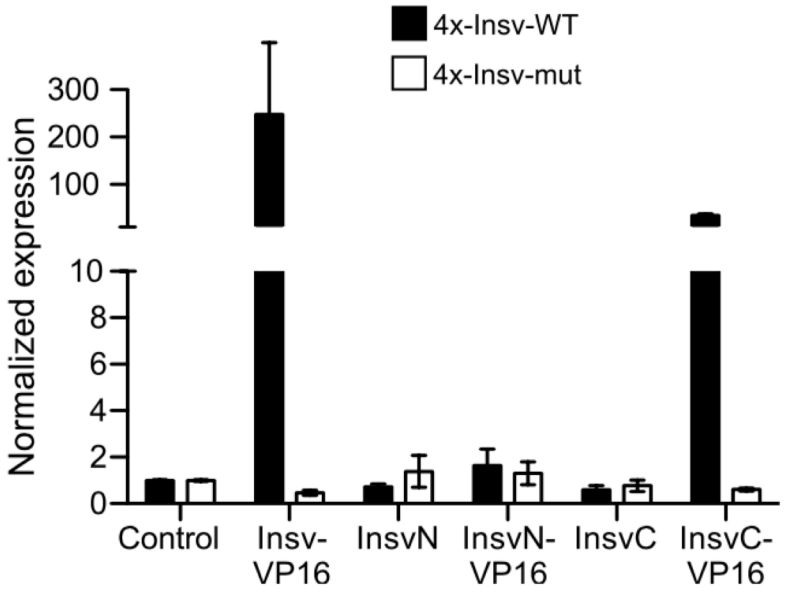
- We also looked a gene expression on a genome-wide scale.
- Genes near Insensitive peaks, that have an Insensitive site, have overall increased expression in an Insensitive mutant.



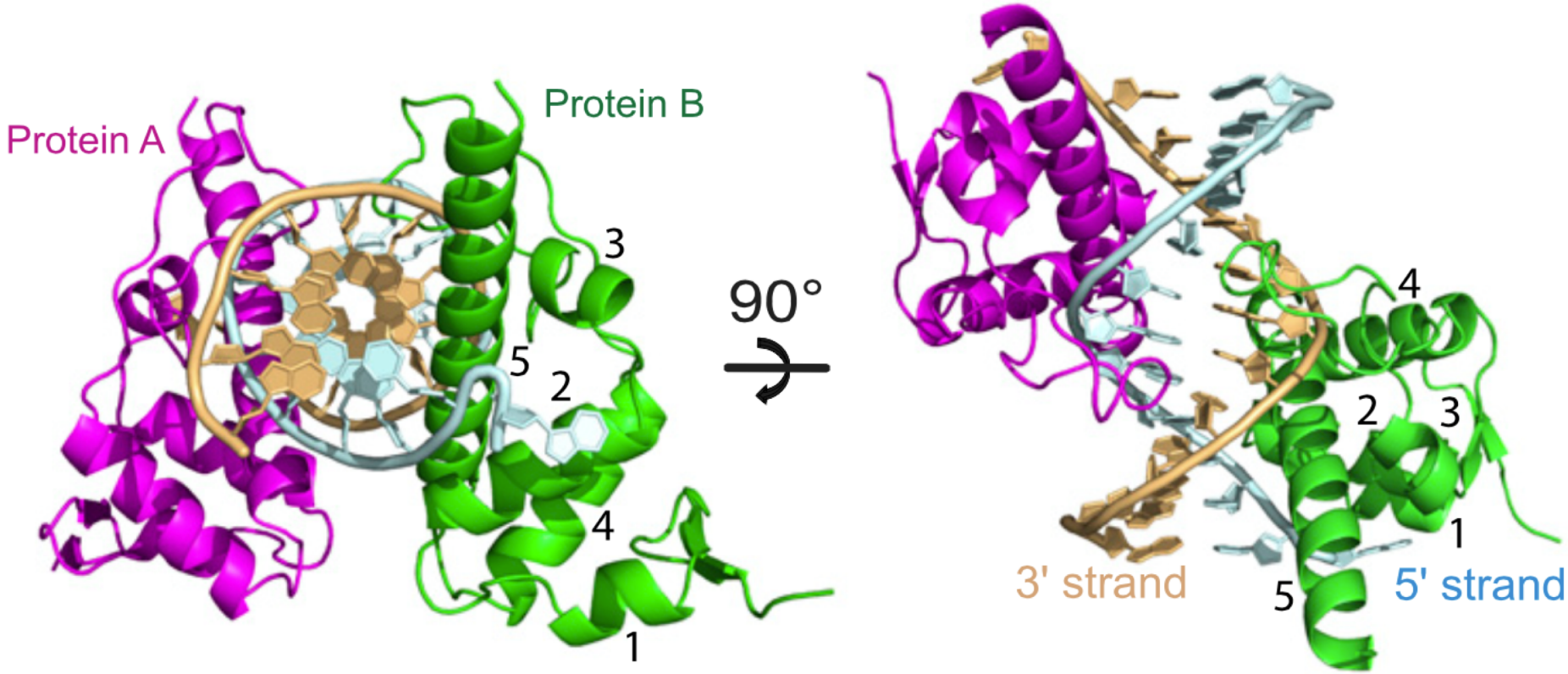
Structure-function experiments

- Actin-luciferase as read-out.
- 4 Insensitive sites in promoter or 4 mutated Insensitive sites
- Different parts of Insensitive, sometimes fused to the V16 activation domain.
- → the (C-terminal) BEN domain is necessary and sufficient for binding to the Insensitive site.



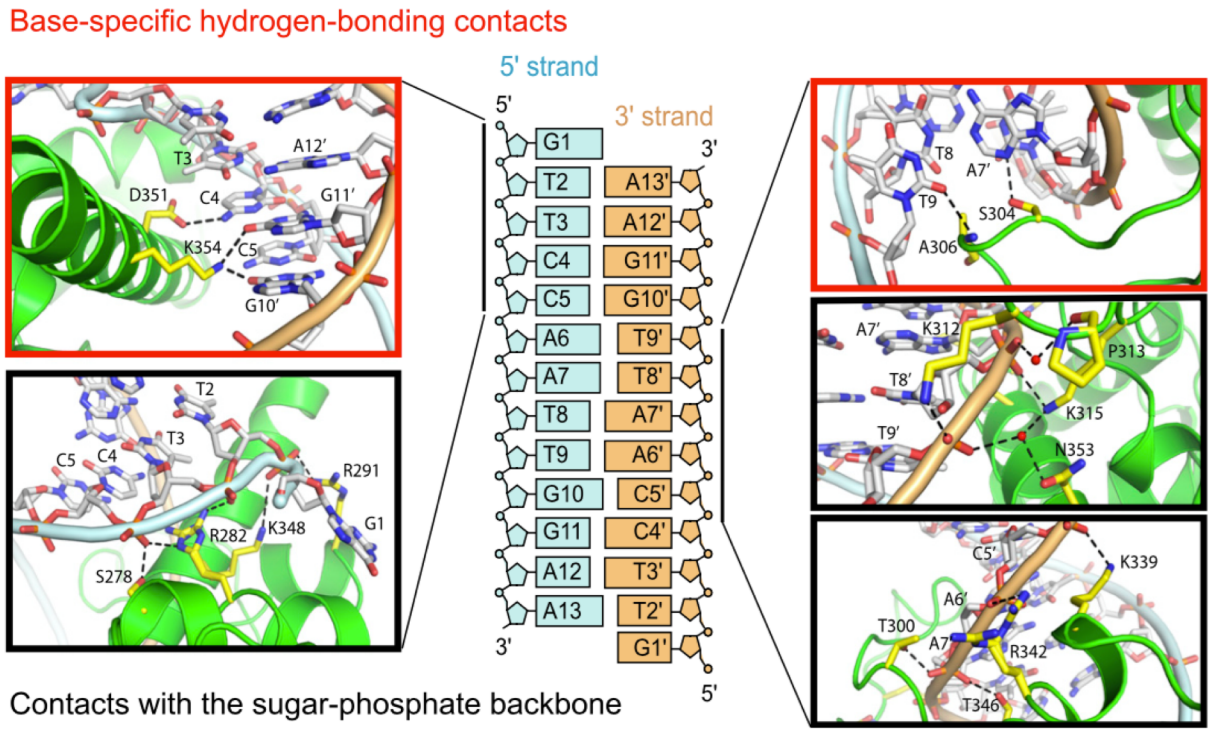


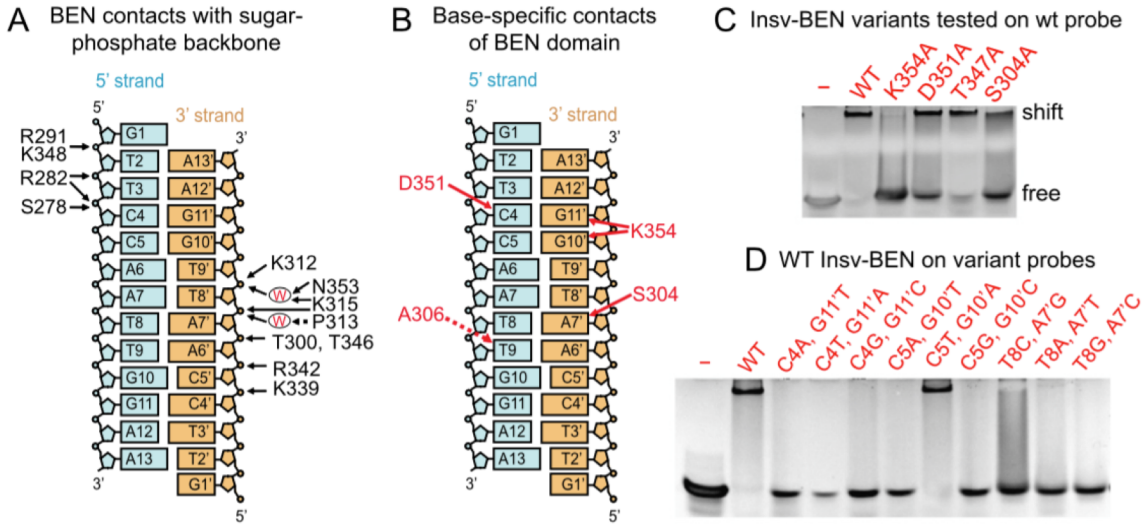
Crystal structure of BEN domain bound to DNA



Validating the structure

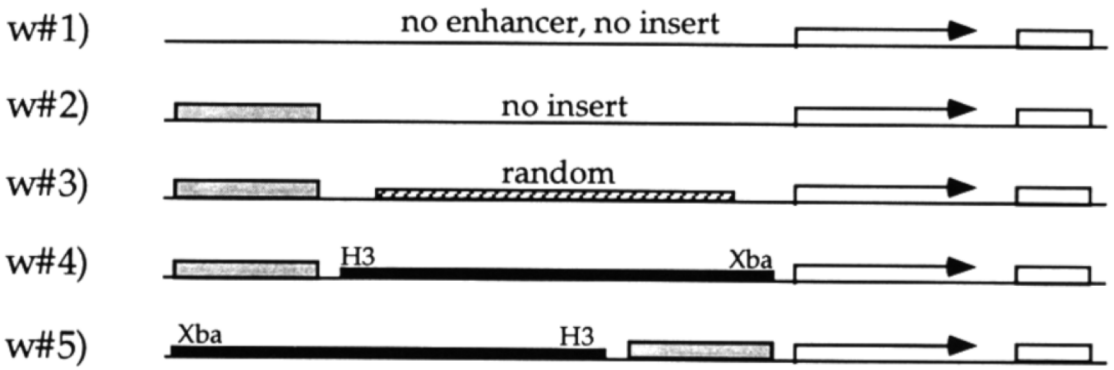
- From the structure, we can see with amino acids make contact with which nucleotides.
- We can make predictions about how amino acid and DNA mutations will affect binding, and test these predictions.



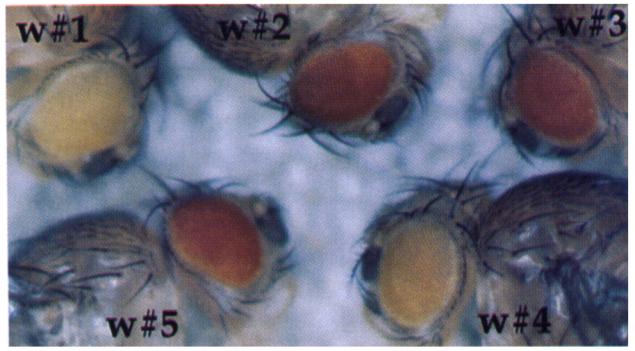


Insulator elements

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



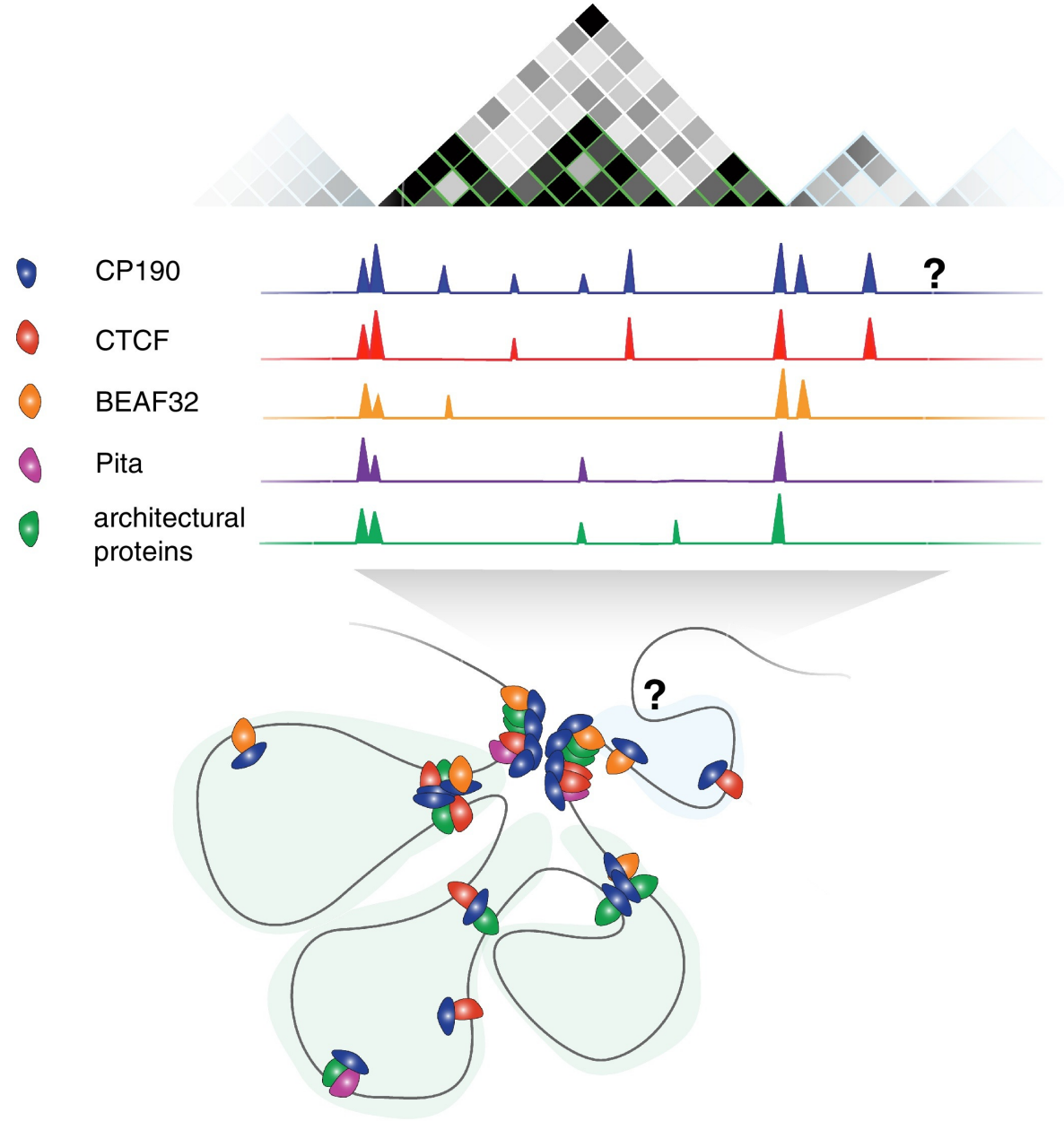
Enhancer blocking (#light eyes/total)	
—	(5/5)
—	(0/6)
—	(0/5)
+	(12/24)
—	(1/11)



Hagstrom et al. *Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex*. Genes & Development 1996.

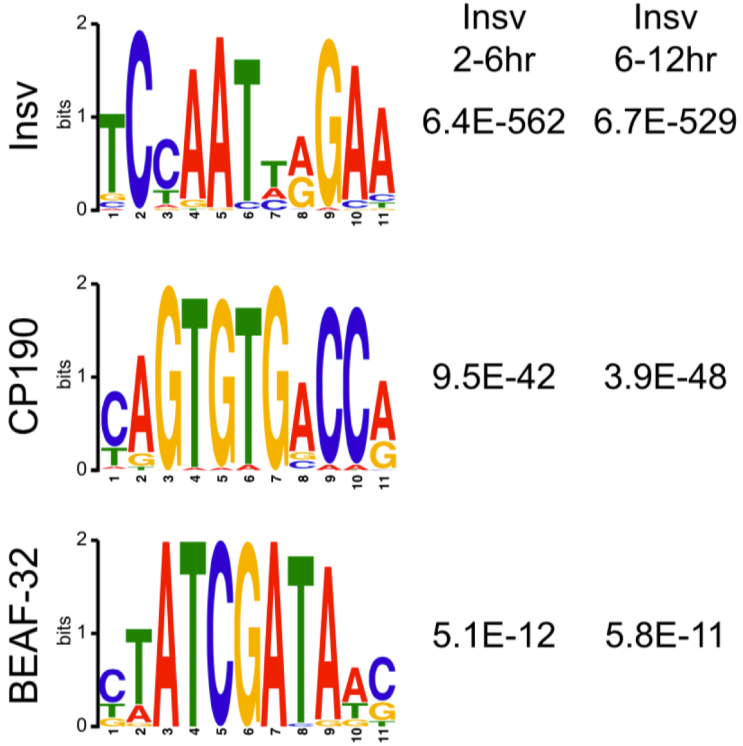
Insulator elements, cont.

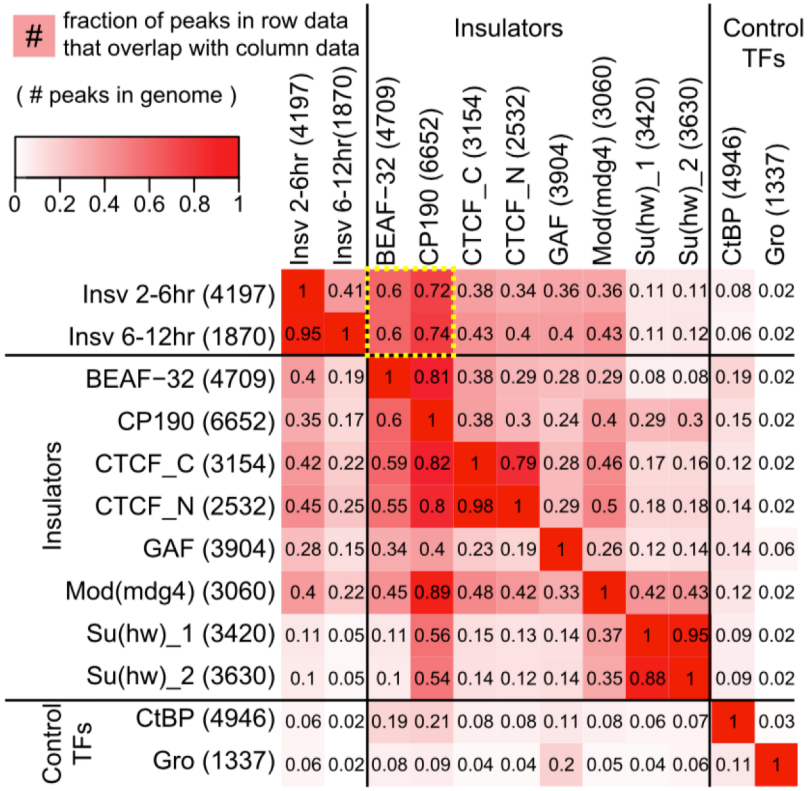
- Insulator elements control DNA looping.
- Enhancers and target genes can end up in different loop domains (≈ topologically associated domains, TADs)



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

Insensitive binds at insulator elements

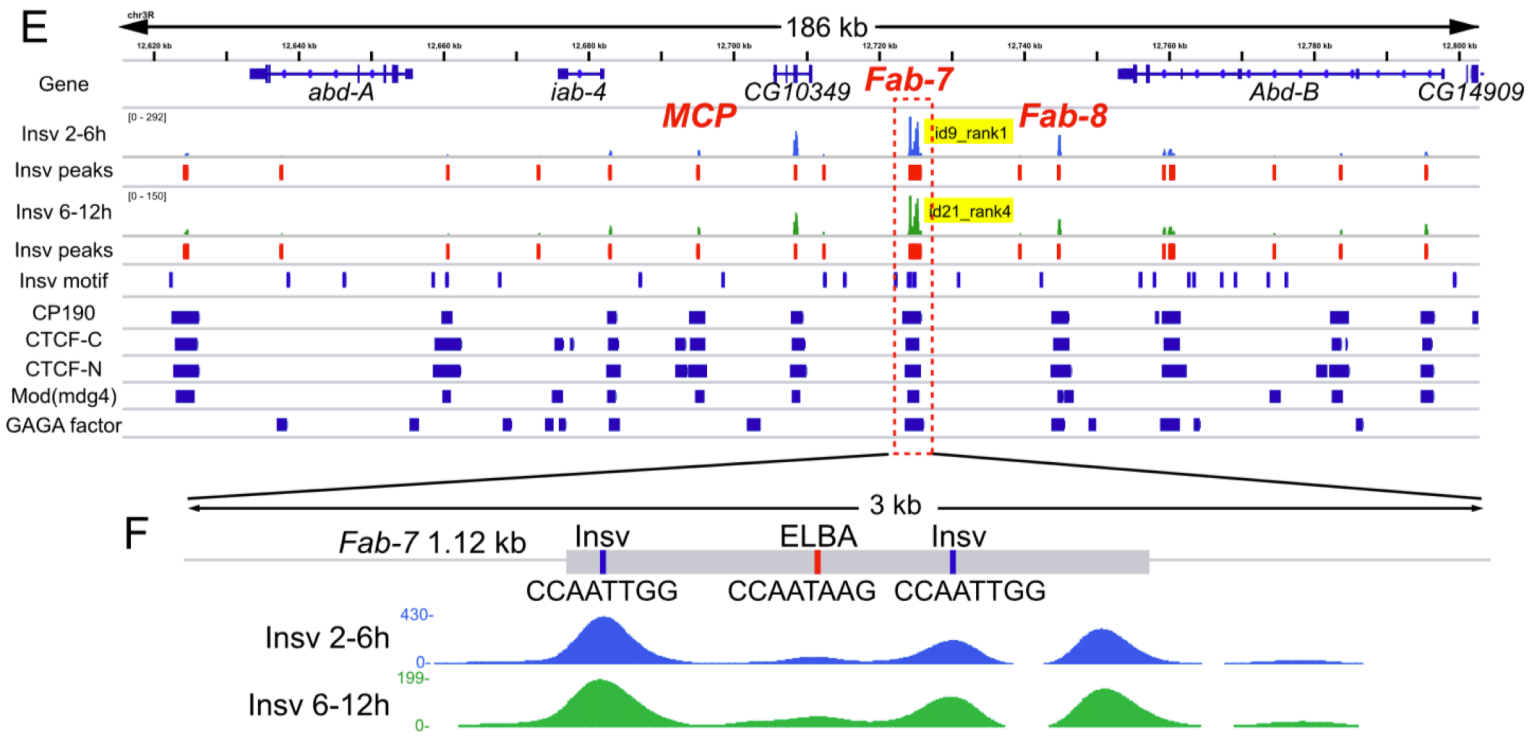


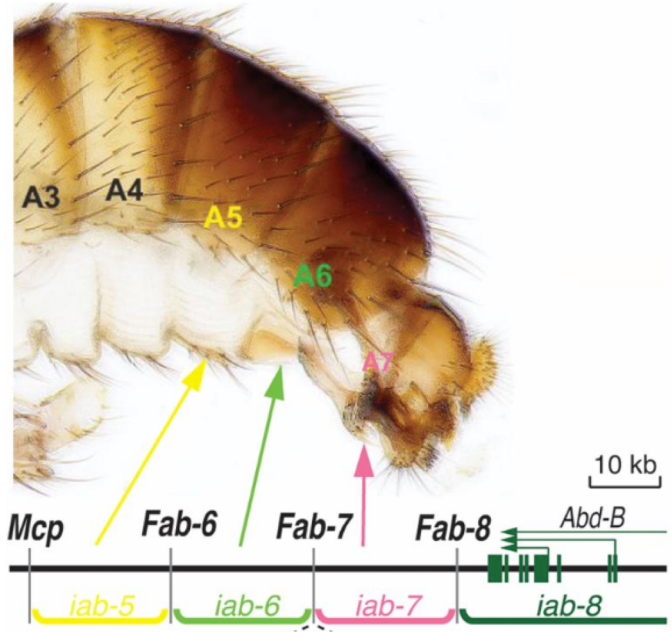


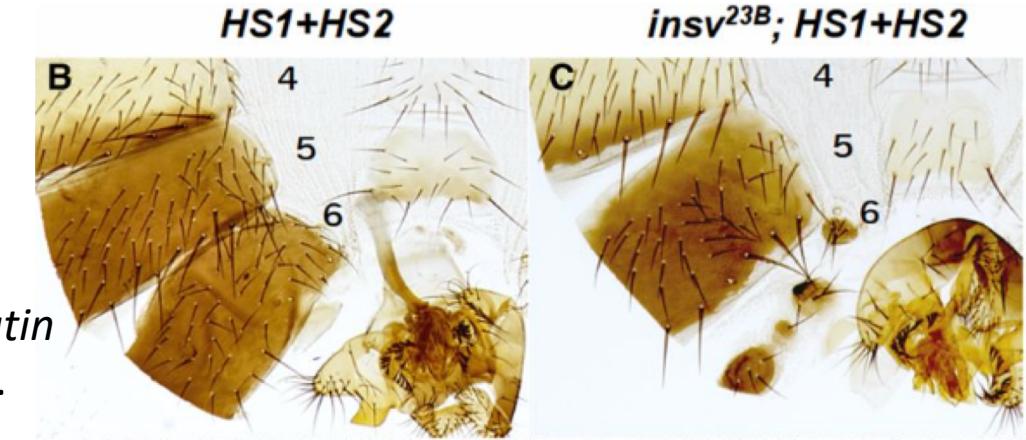
- *Insenstive* peaks are enriched for C190 and BEAF-32 motifs
- *Insenstive* peaks overlap C190, BEAF-32 and CTCF peaks

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







Fedotova et al. *The BEN Domain Protein Insensitive Binds to the Fab-7 Chromatin Boundary To Establish Proper Segmental Identity in Drosophila*. Genetics 2018.

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 remodellers
    - 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)

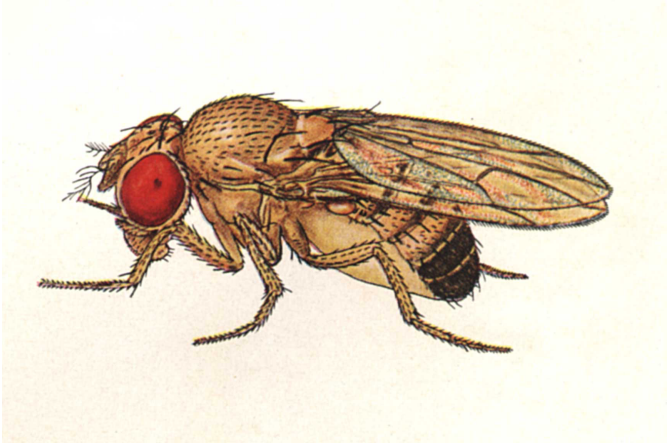
Qi Dai

Hong Duan

Dinshaw Palel (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

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

So far..

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

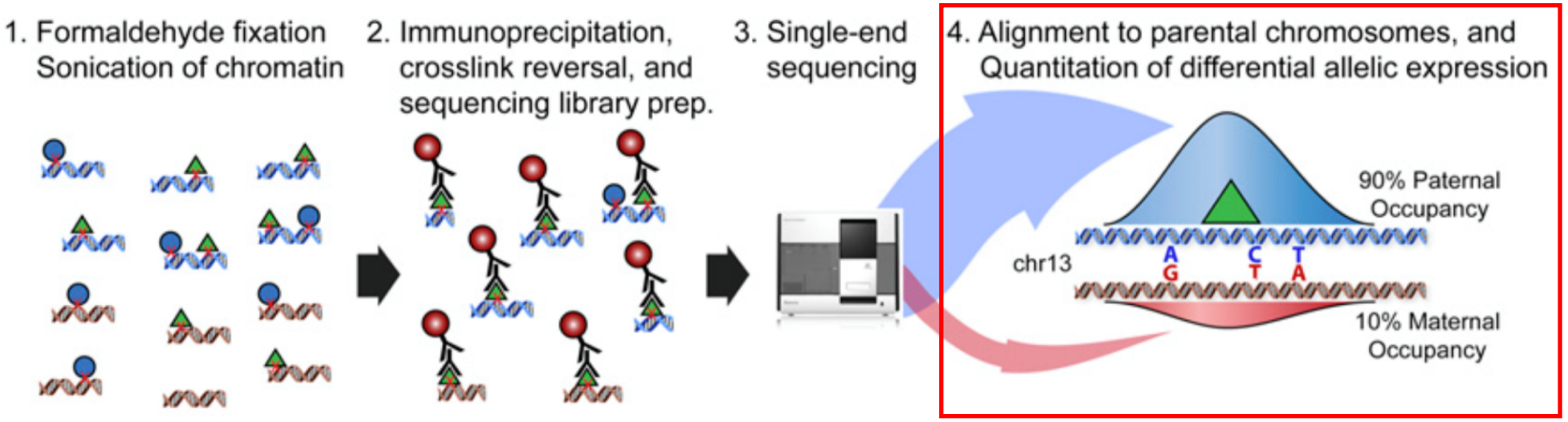
- analyzing which regions of the DNA a protein interacts with
- using a lot of material (millions of cells)

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.

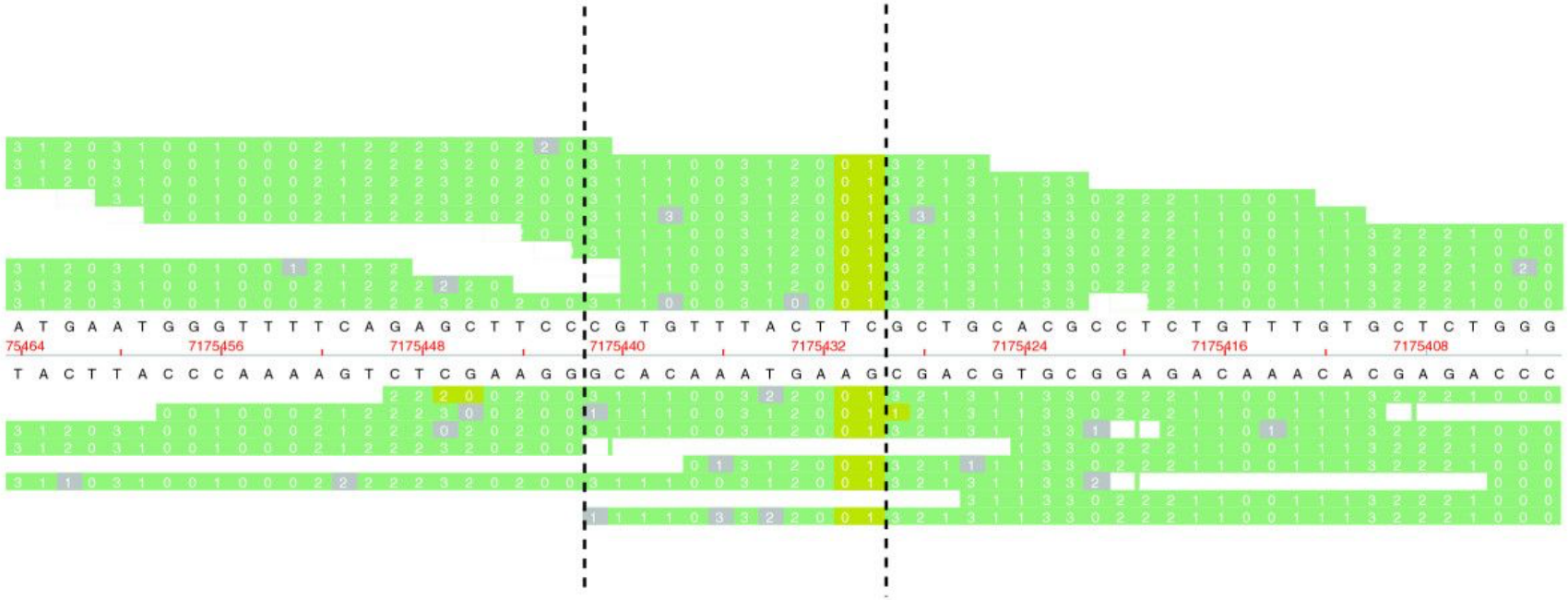
Early example:

FOXA1 top motif:

HepG2 genomic DNA: CGTGTTTACTT[T/C]

Reference genome: CGTGTTTACTTC

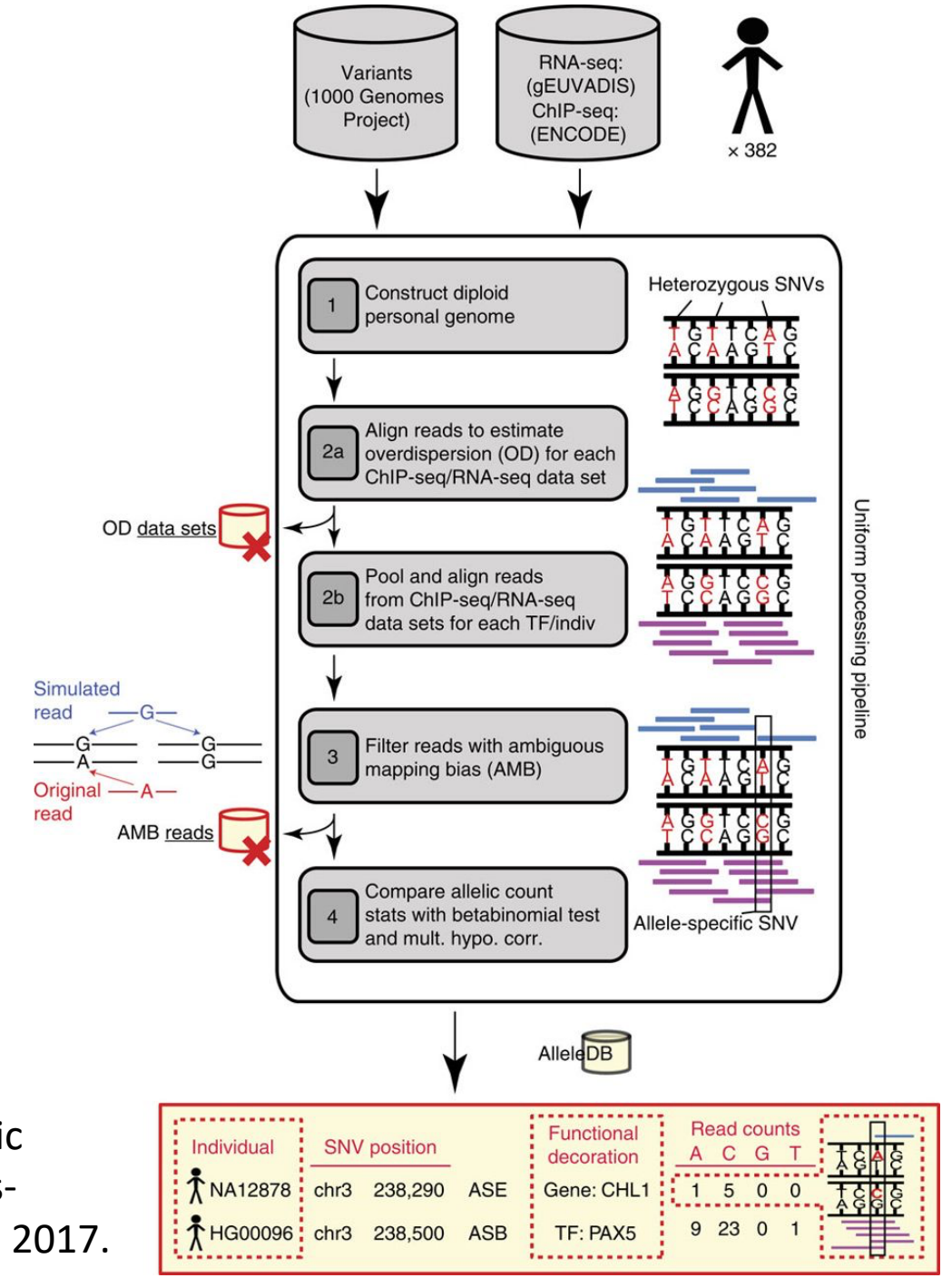
FOXA1 ChIP DNA: CGTGTTTACTTT



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)

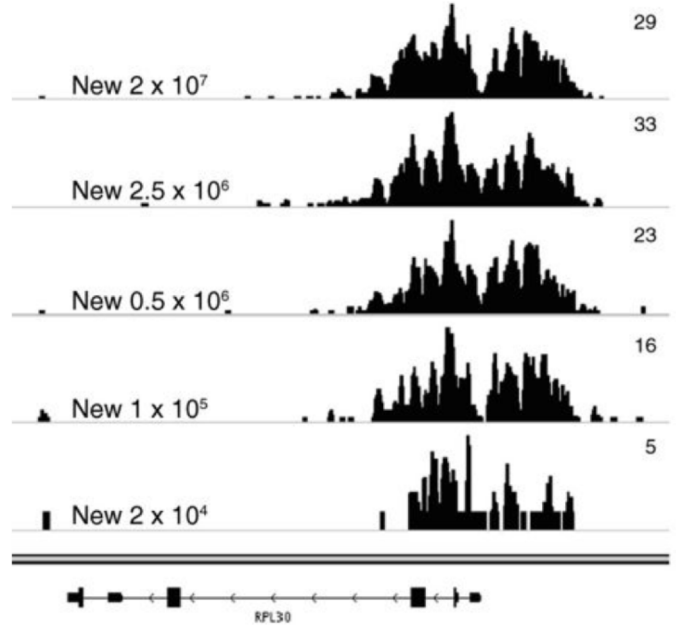
  → other mechanisms for transcription factor recruitment. Co-factors?

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



H3K4me3

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

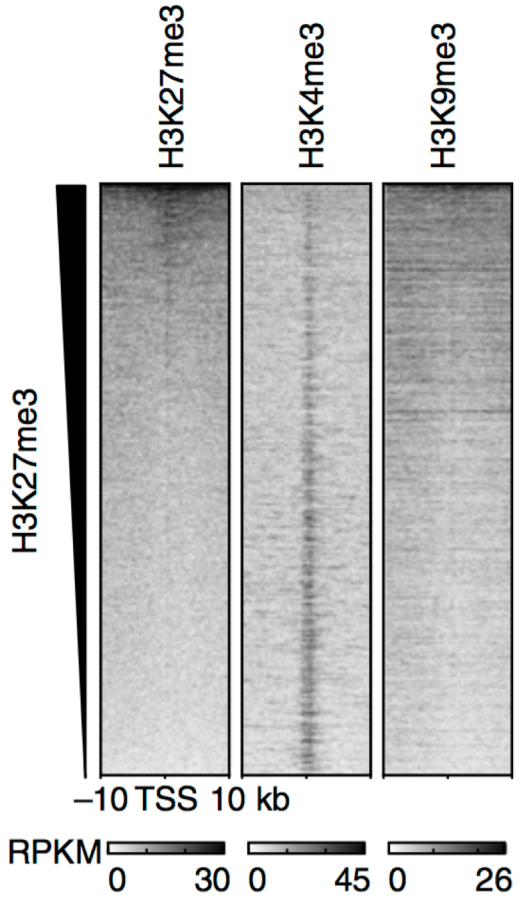
ARTICLE

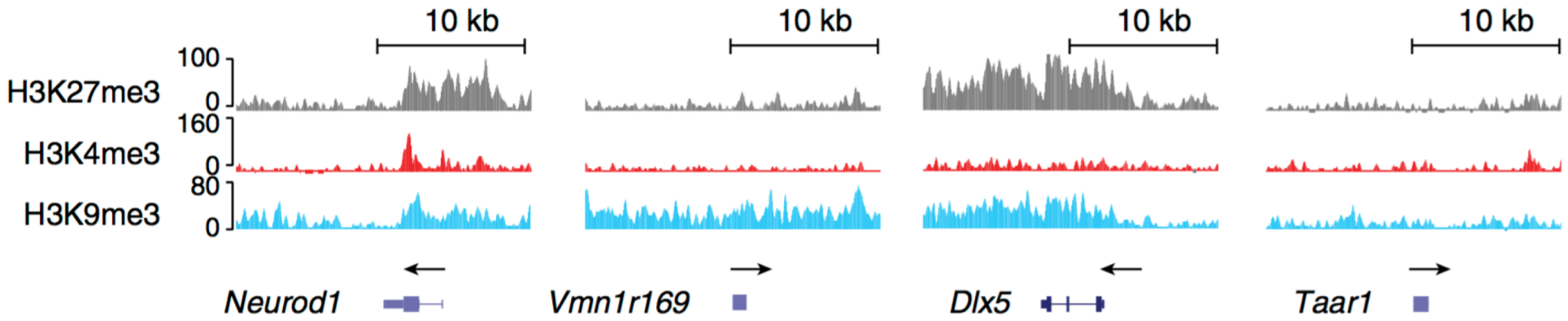
DOI: 10.1038/s41467-018-03538-9 OPEN

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

Erik Södersten[1], Konstantinos Toskas[1], Vilma Rraklli[1], Katarina Tiklova[1], Åsa K. Björklund[2], Markus Ringnér[3], Thomas Perlmann[1] & Johan Holmberg[1]

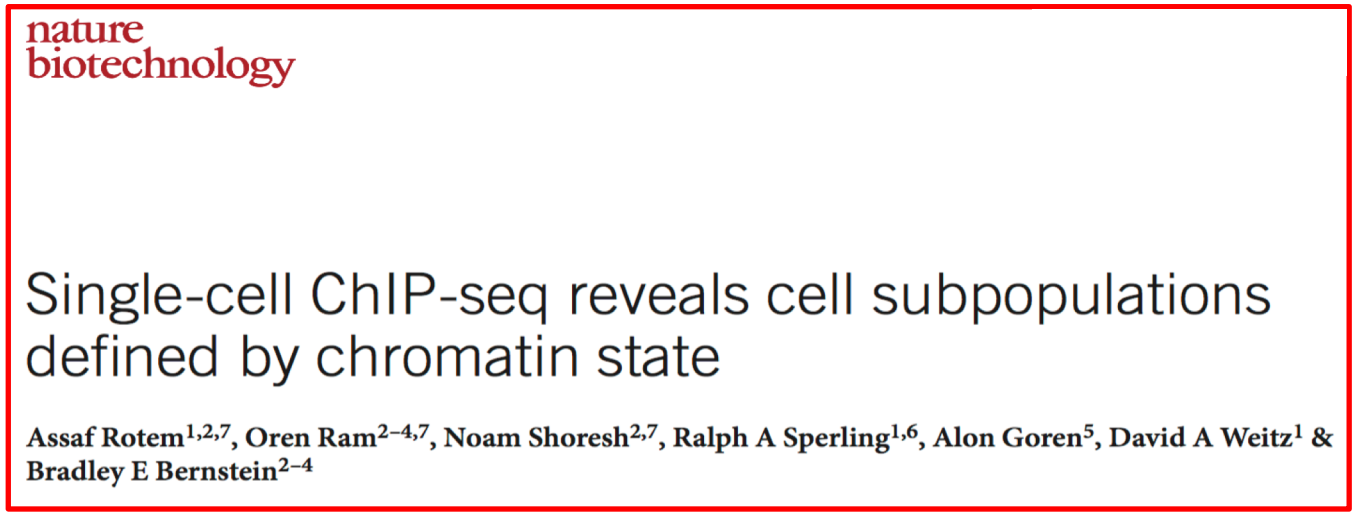
- They were able to get useful data for 3 histone marks.
- Also comparison with RNA-seq data.
- No big changes to analysis
  - Some quality measures might not look as good, e.g. duplication rates
  - QC even more important!



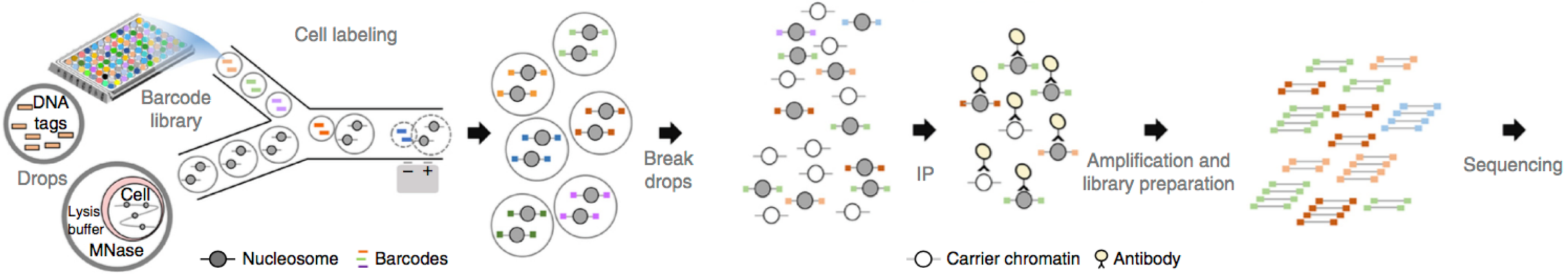


Single cell ChIP-seq

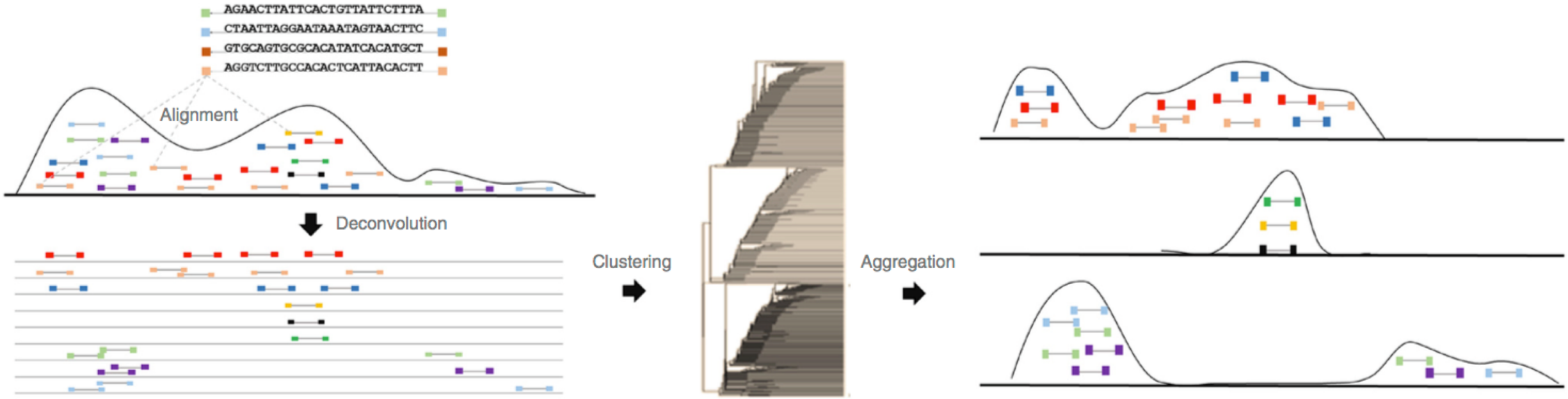
- The signal we get from normal ChIP-seq is an average over all cells in the sample
- This misses heterogeneity
  - Cell types
  - Primed vs unprimed cells
  - Response to stimuli
- With single cell ChIP-seq, we get data for each cell separately
- This is similar to single cell RNA-seq, but much harder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



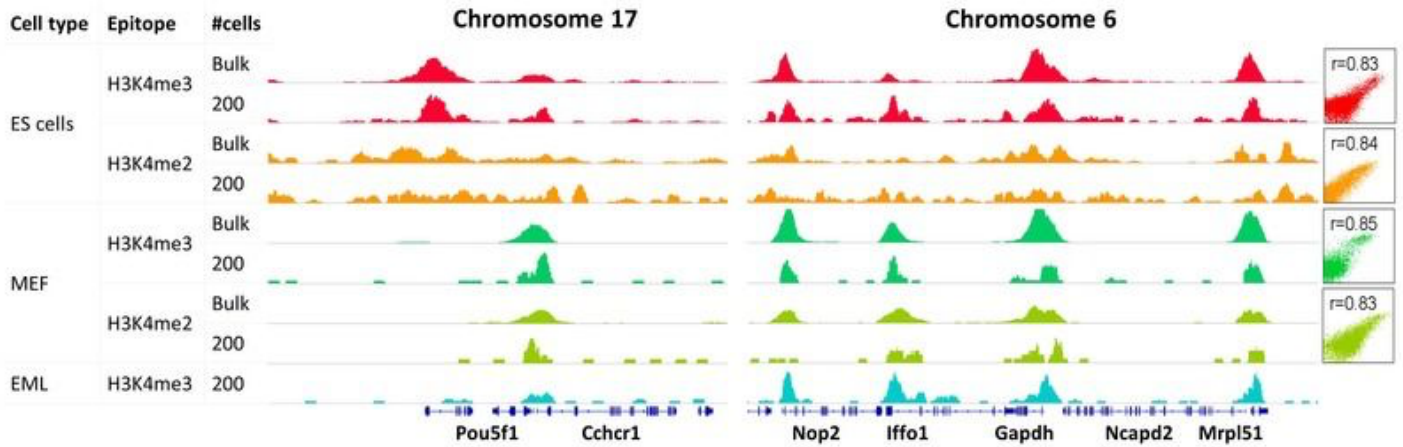
Experiment overview

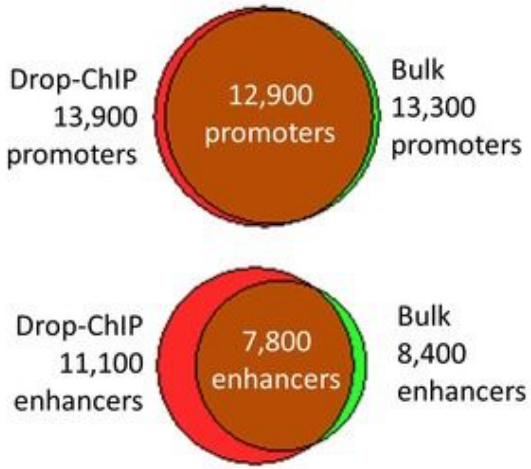


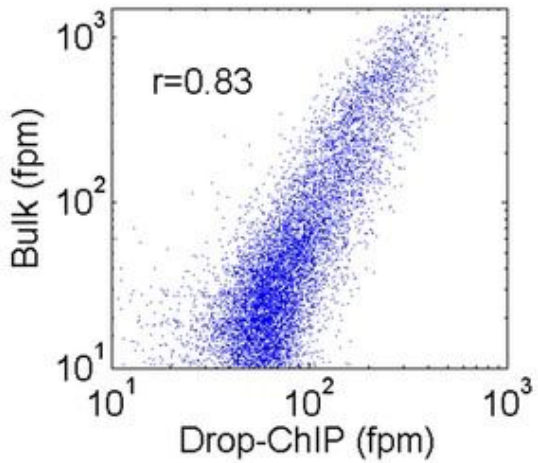
Analysis overview



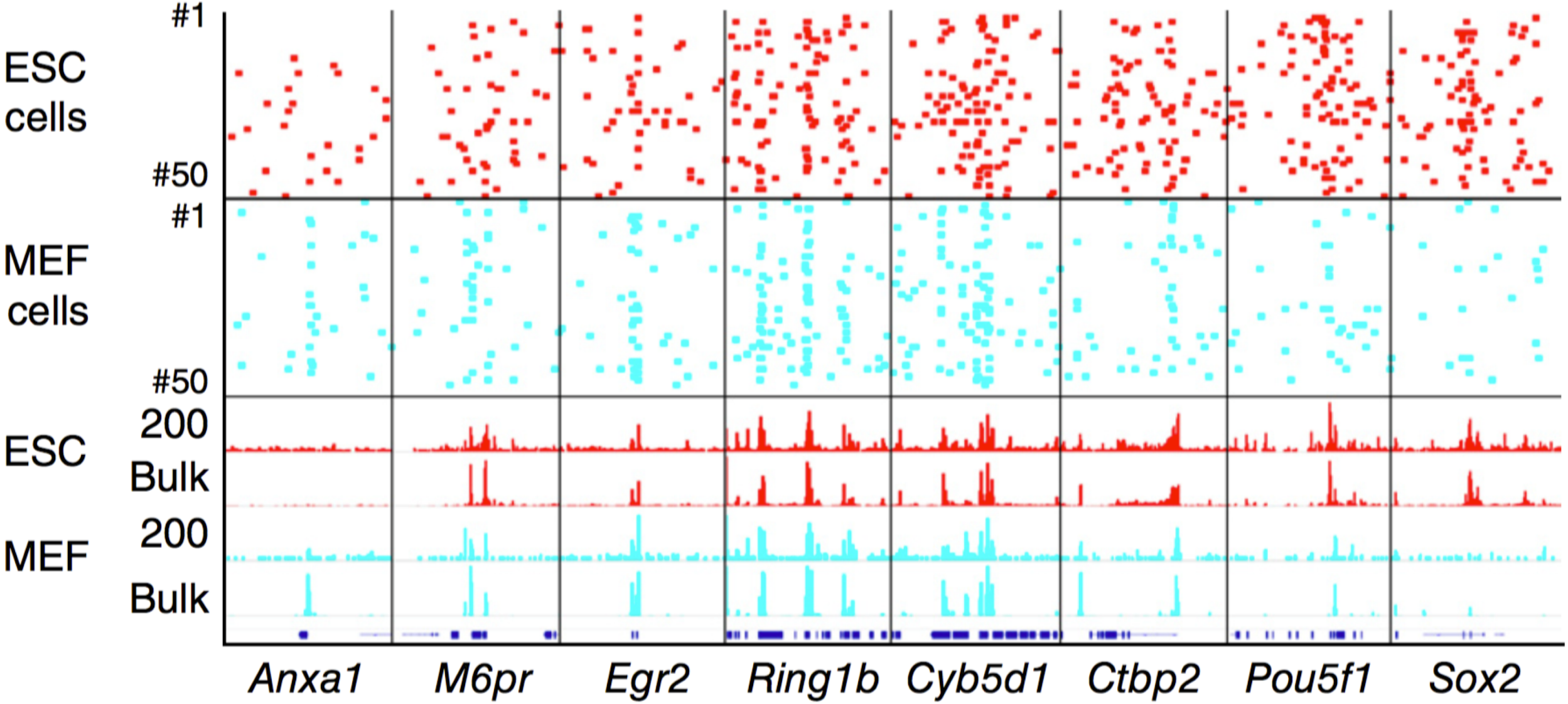
Aggregated single cell vs bulk data



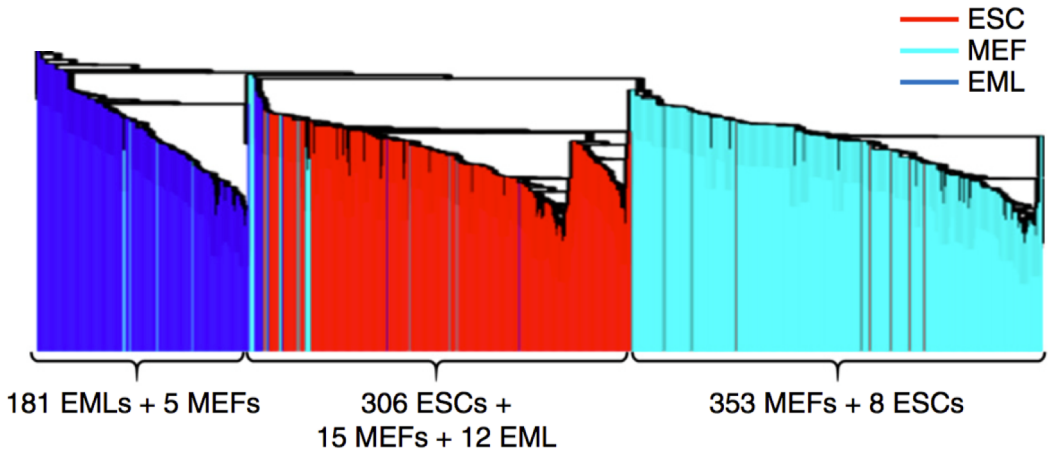




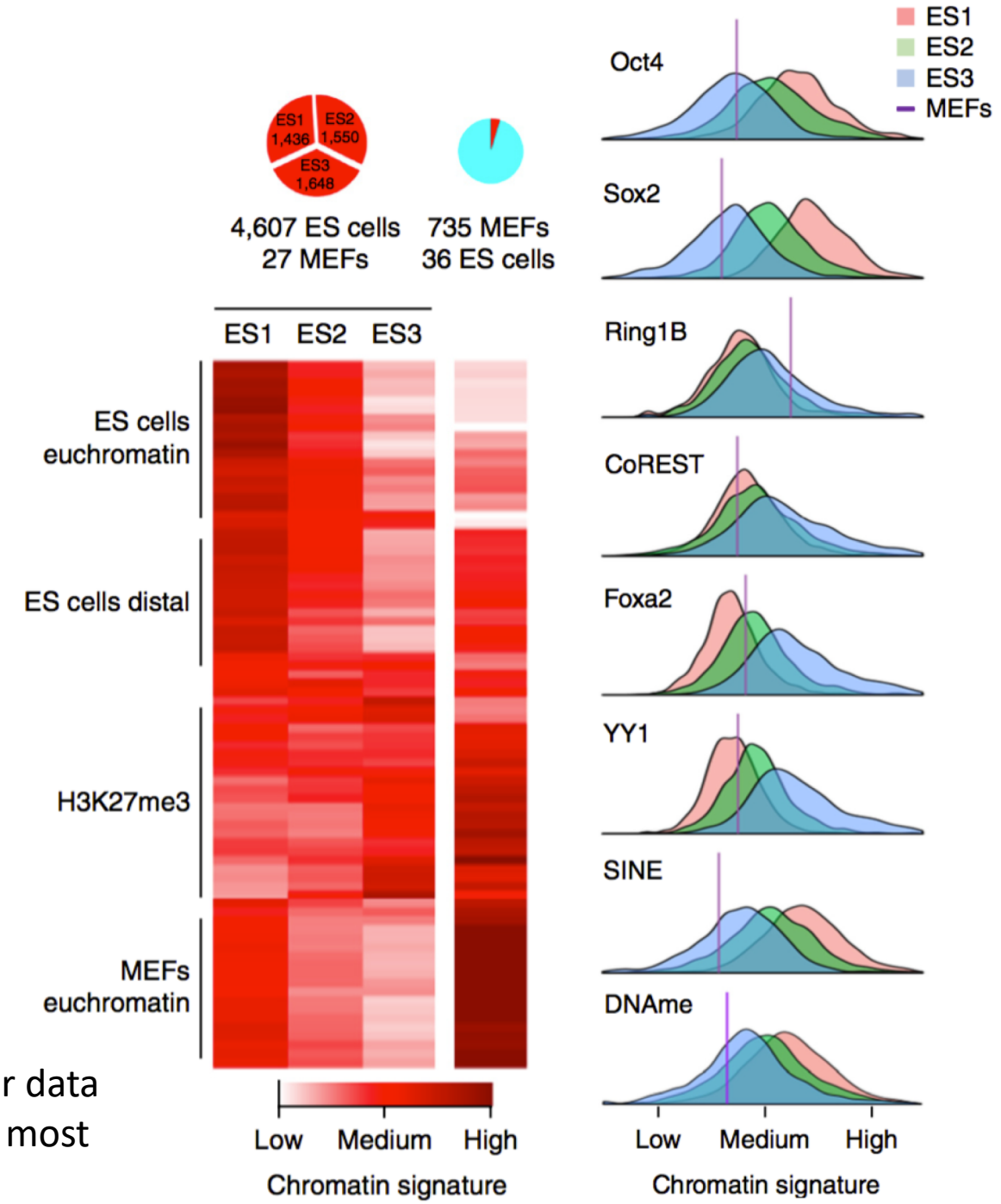
Data from individual cells



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