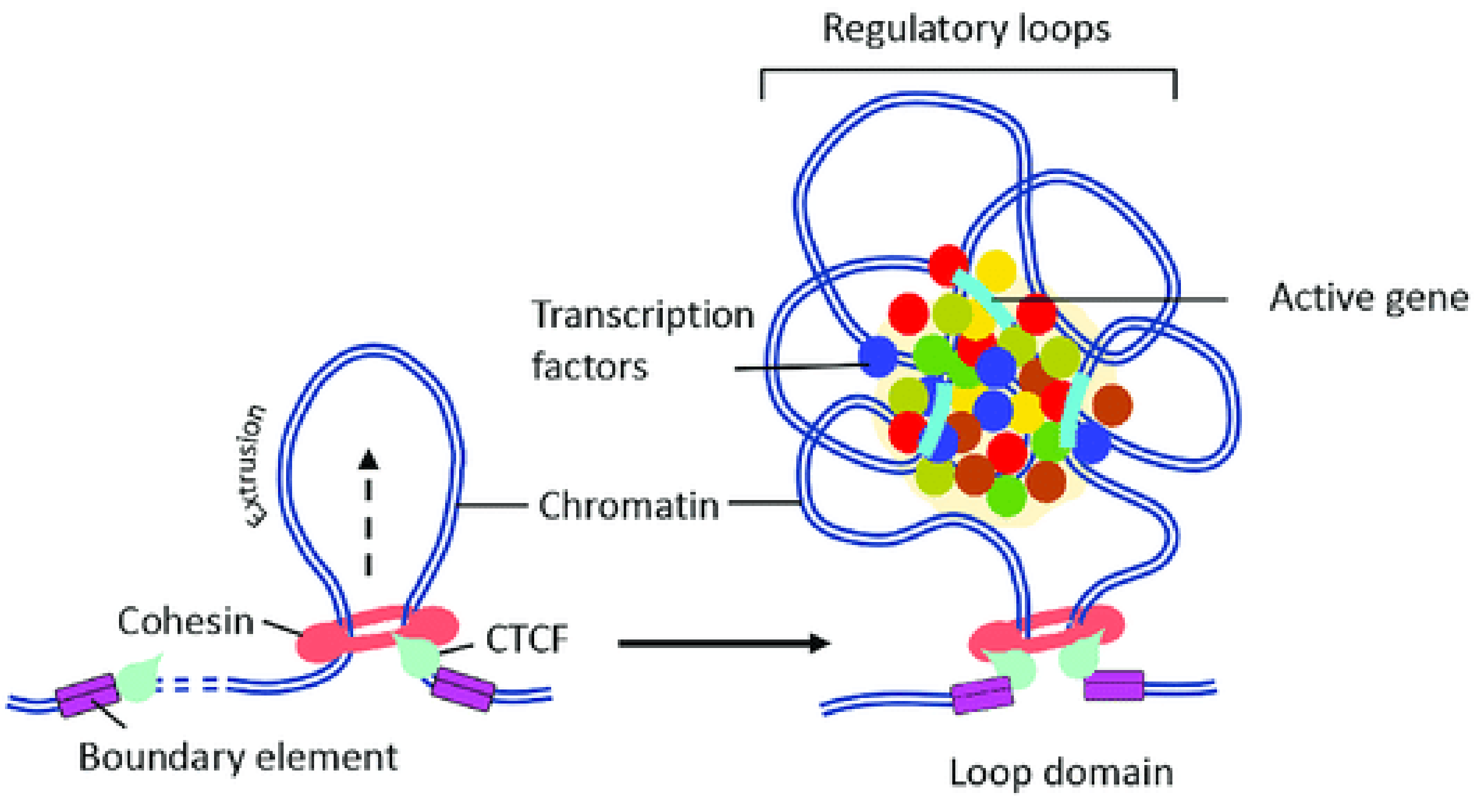
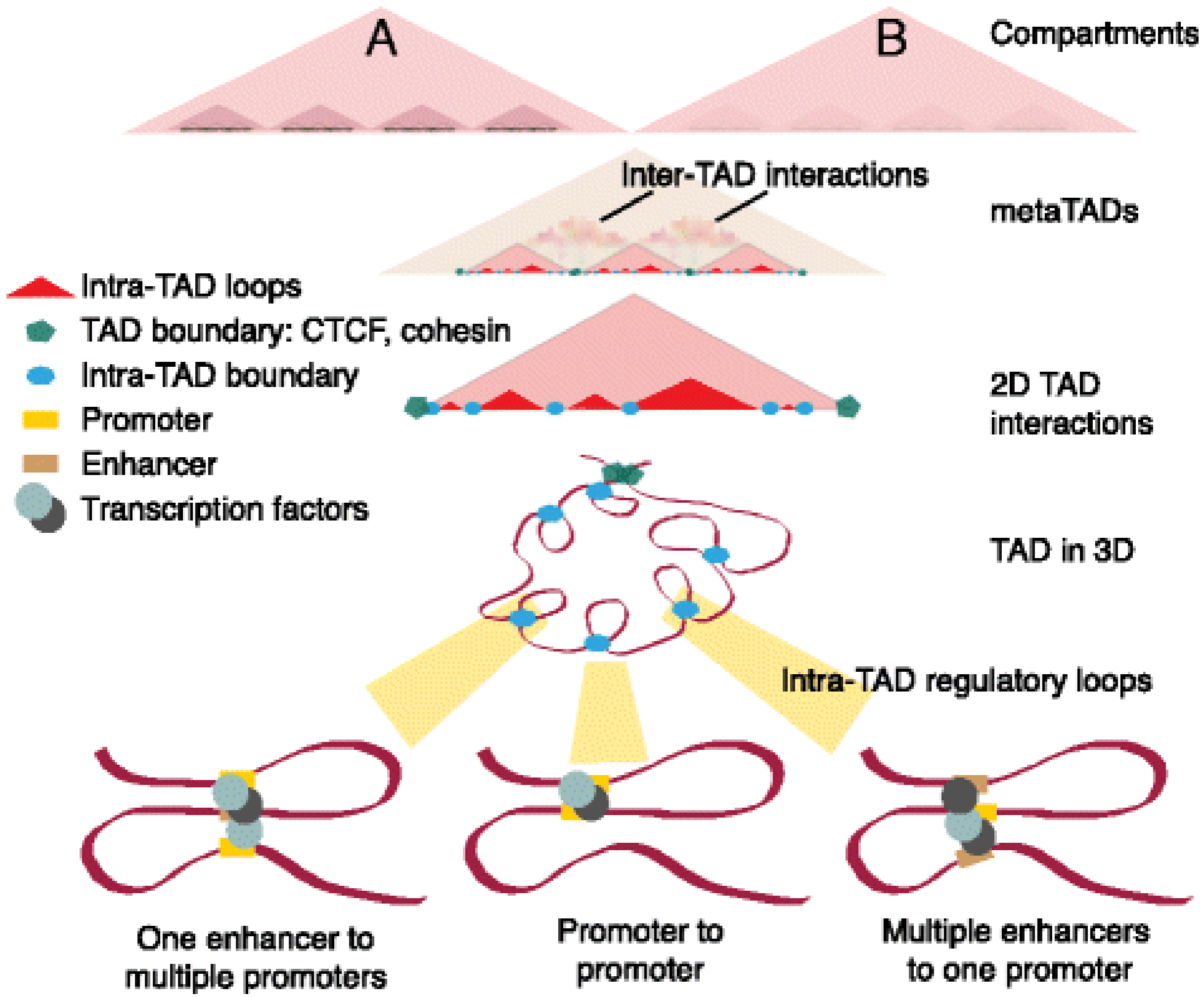
Thesis proposal: Research on CTCF BINDING SITES and implications for Cancer genomics, Genome Deep Structure across Species, the Sound of Genome

- COHESIN-CTCF BINDINGS IN ACTION --- DNA FOLDING WITHIN 3D SPACE

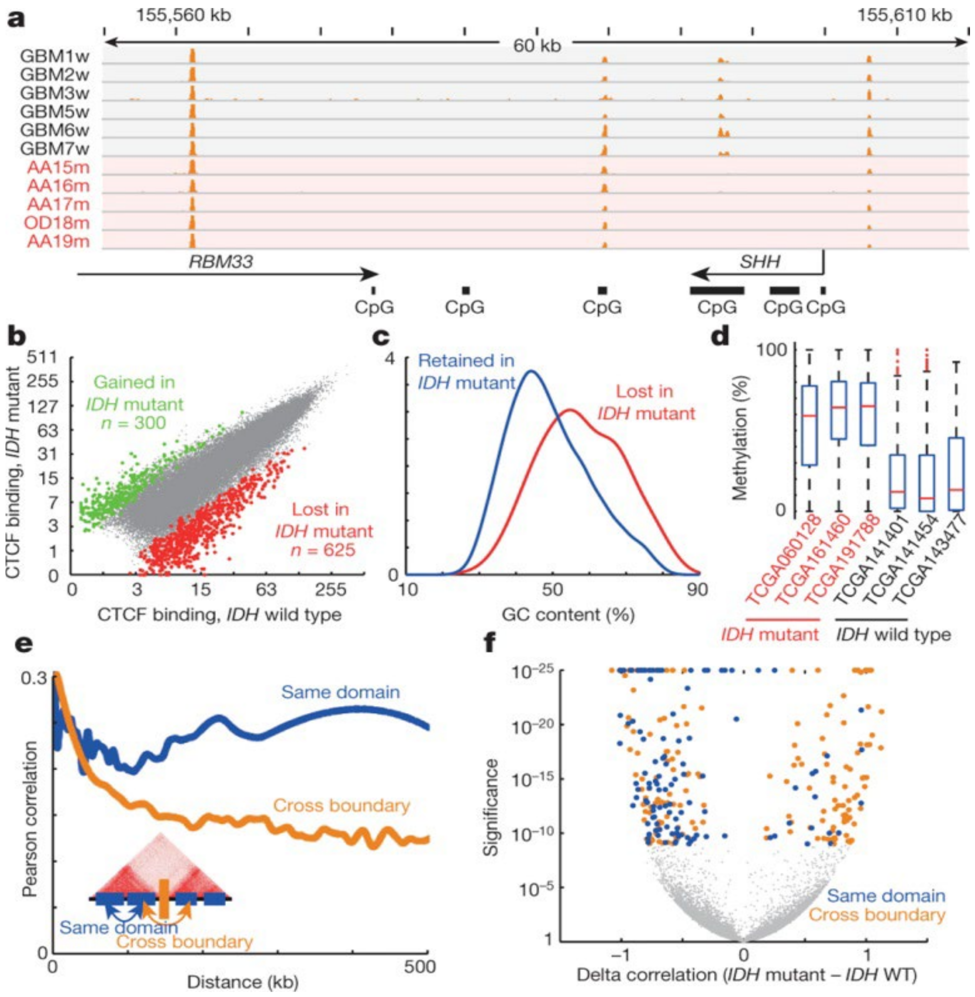


Topological Domains (TADS) and Loops



ORIGIN: Brad Bernstein’s Group, Broad Institute

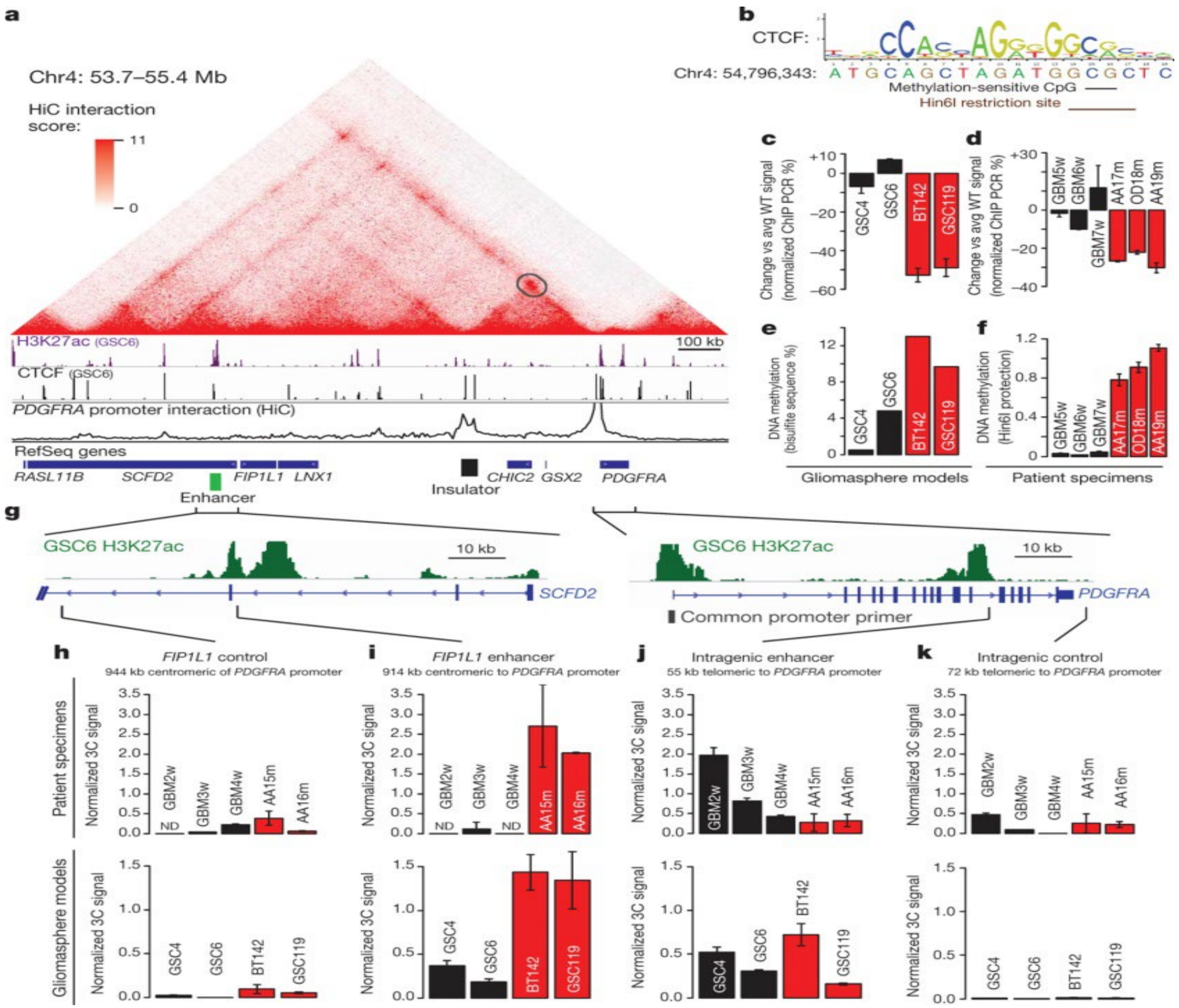
CTCF binding and gene insulation compromised in *IDH* mutant gliomas.



nature

W. A. Flavahan *et al. Nature* 1-5 (2016) doi:10.1038/nature16490

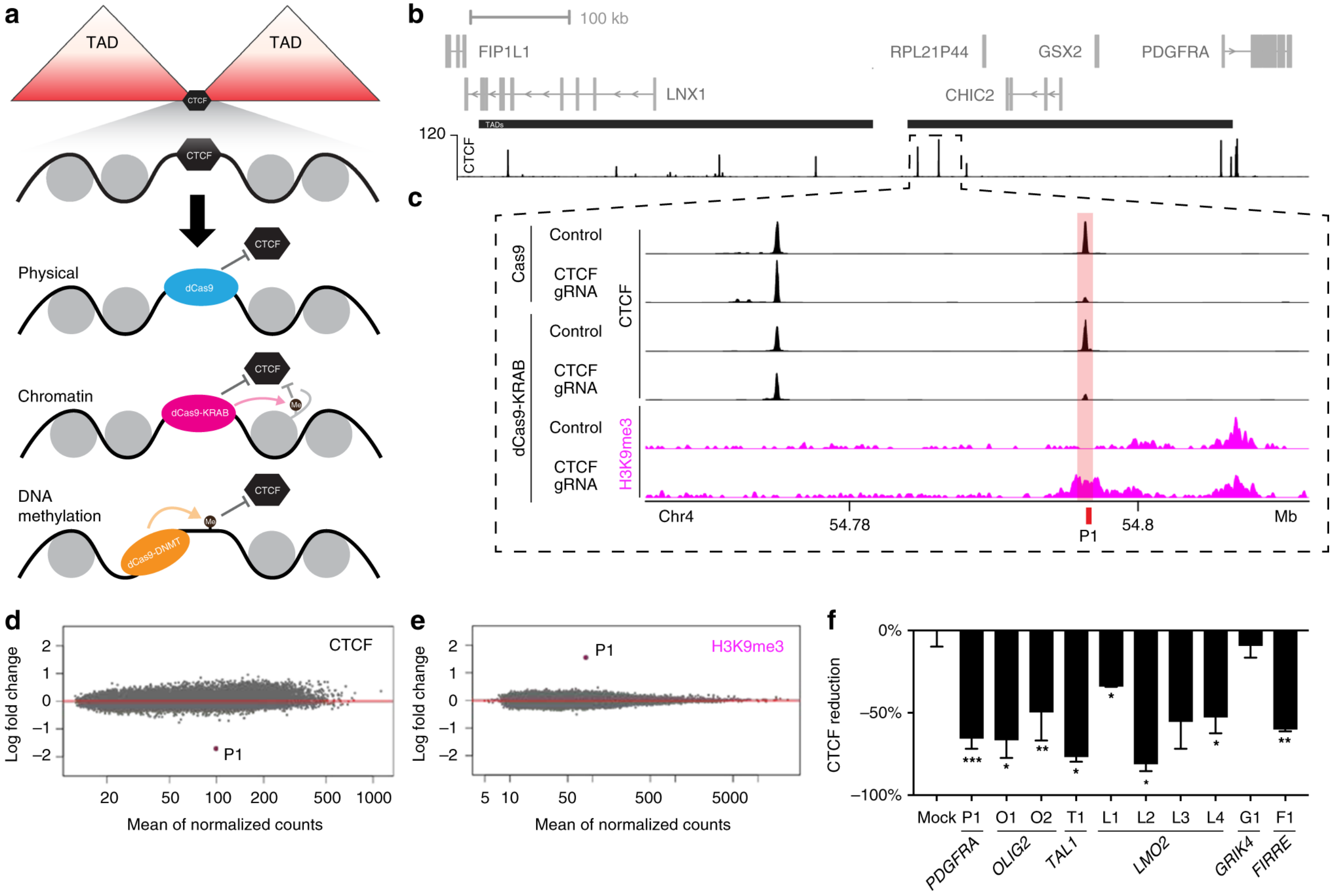
Insulator loss allows *PDGFRA* to interact with a constitutive enhancer.



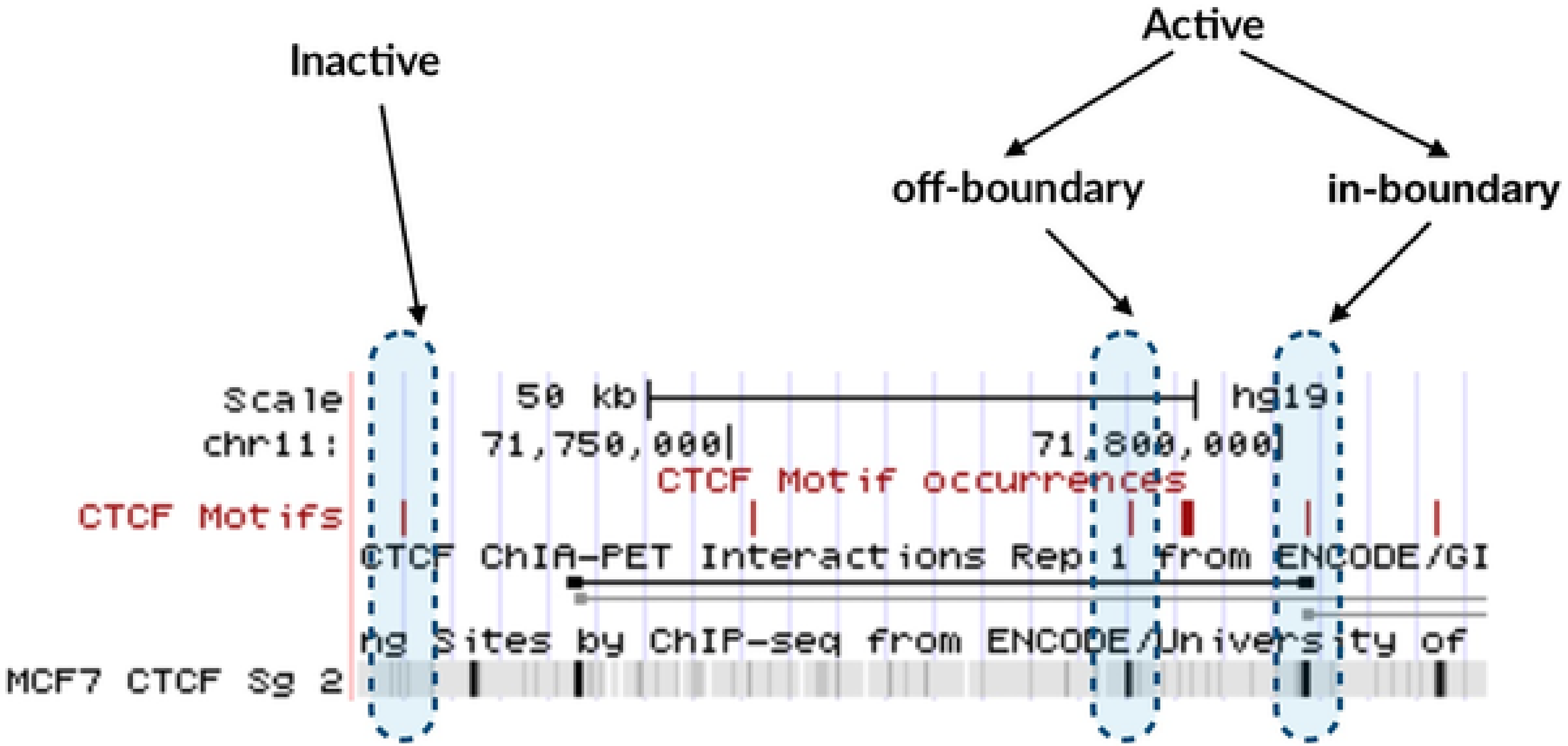
nature

W. A. Flavahan *et al. Nature* 1-5 (2016) doi:10.1038/nature16490

Epigenome editing strategies for the functional annotation of CTCF insulators



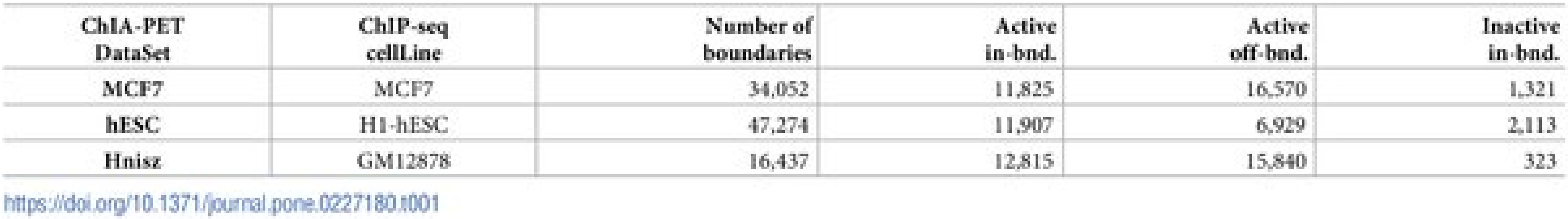
First Attempt: Big Data, lots of opportunities (so far, unexploited)



Classification of CTCF motifs, within a short portion of chromosome 11.
Motifs are classified as active (confirmed by a CTCF ChIP-seq peak) and inactive (not confirmed). Active motifs are further divided into in-boundary and off-boundary according to whether they overlap a boundary, as defined by a ChIA-PET experiment.

Pinoli P, Stamoulakatou E, Nguyen AP, Rodríguez Martínez M, Ceri S (2020) Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLOS ONE 15(1): e0227180. https://doi.org/10.1371/journal.pone.0227180
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227180

Table 1. Summary statistics of the number of boundaries and motifs.

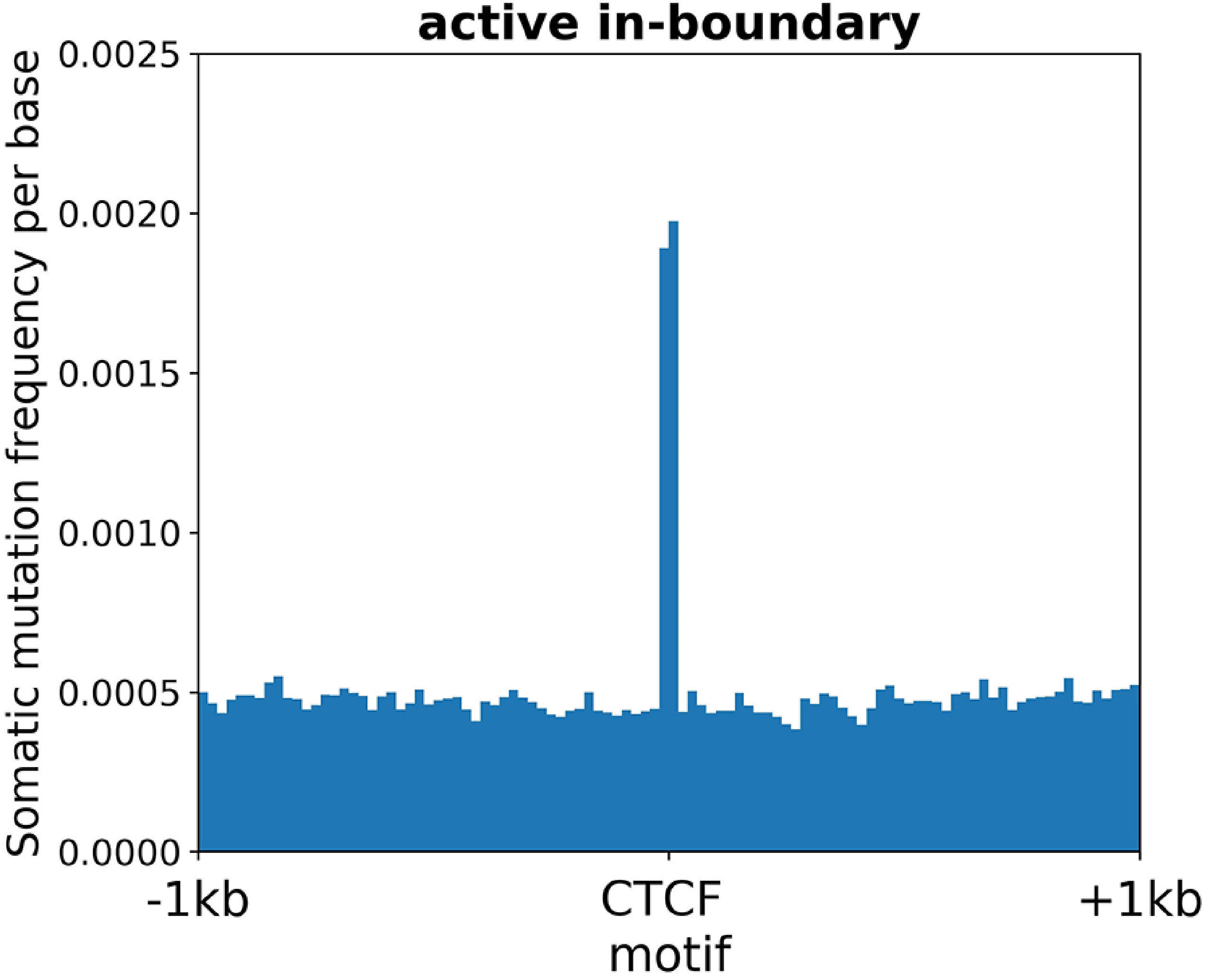


ChIA-PET DataSet	ChIP-seq cellLine	Number of boundaries	Active in-bnd.	Active off-bnd.	Inactive in-bnd.
MCF7	MCF7	34,052	11,825	16,570	1,321
hESC	H1-hESC	47,274	11,907	6,929	2,113
Hnisz	GM12878	16,437	12,815	15,840	323

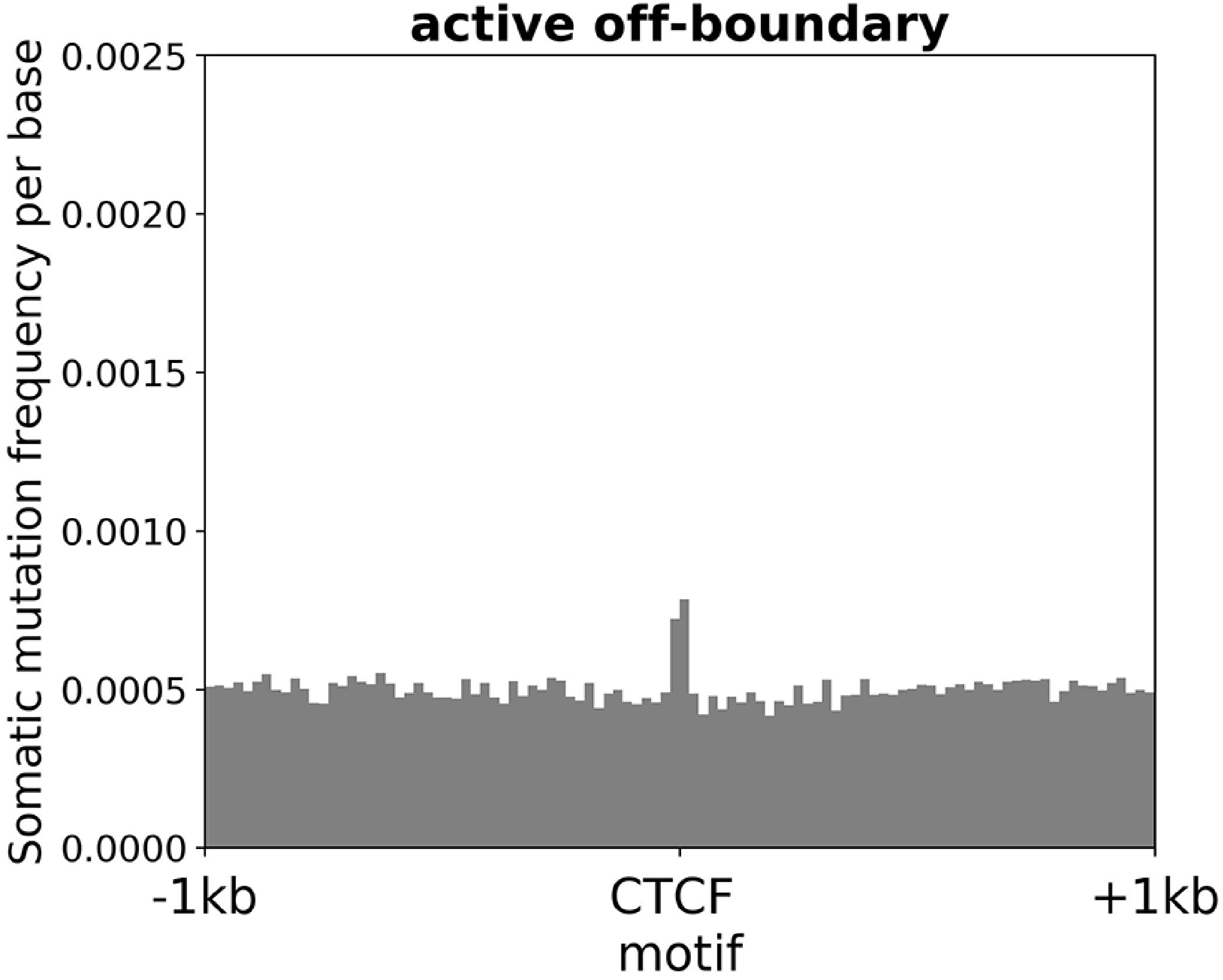
https://doi.org/10.1371/journal.pone.0227180.t001

Pinoli P, Stamoulakatou E, Nguyen AP, Rodríguez Martínez M, Ceri S (2020) Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLOS ONE 15(1): e0227180. https://doi.org/10.1371/journal.pone.0227180
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227180

Pan-cancer analysis of somatic mutations and epigenetic

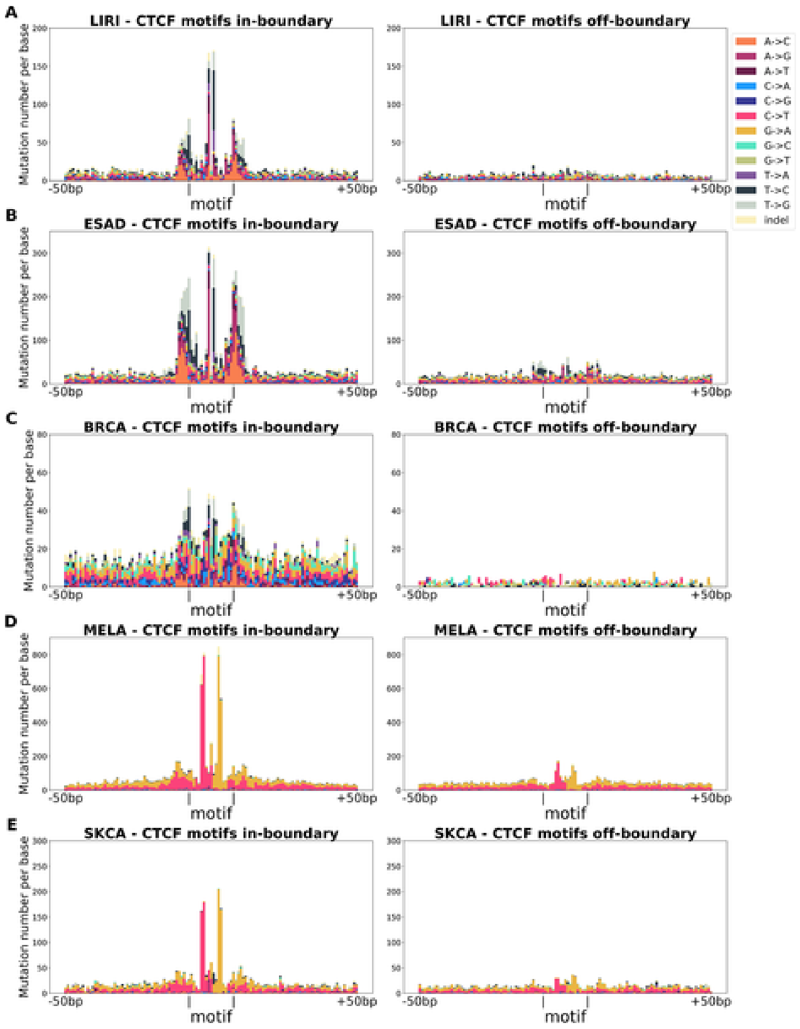


(a) hESC in-boundary, ESAD mutations



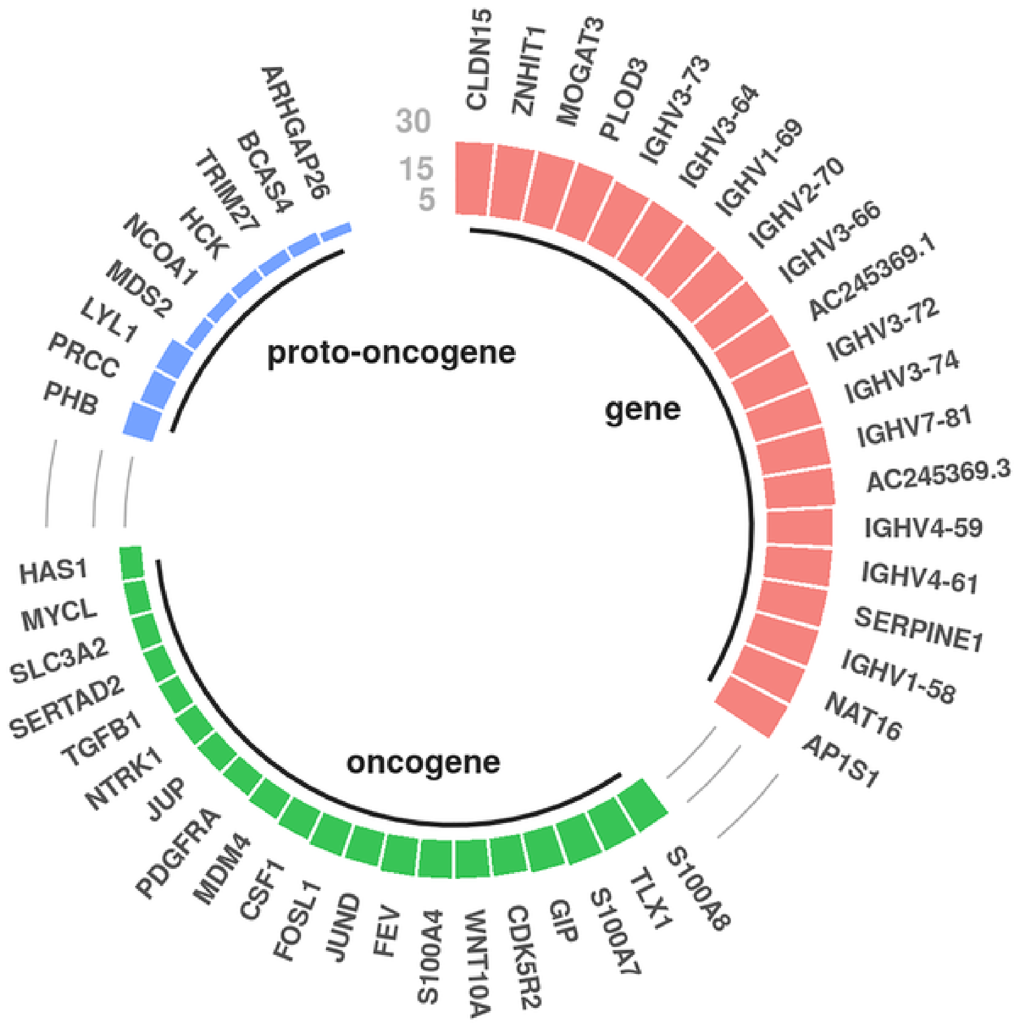
(b) hESC off-boundary, ESAD mutations

Fig 5. Mutations in active in-boundary CTCF motifs and flanking regions (19 bp ±50 bp).



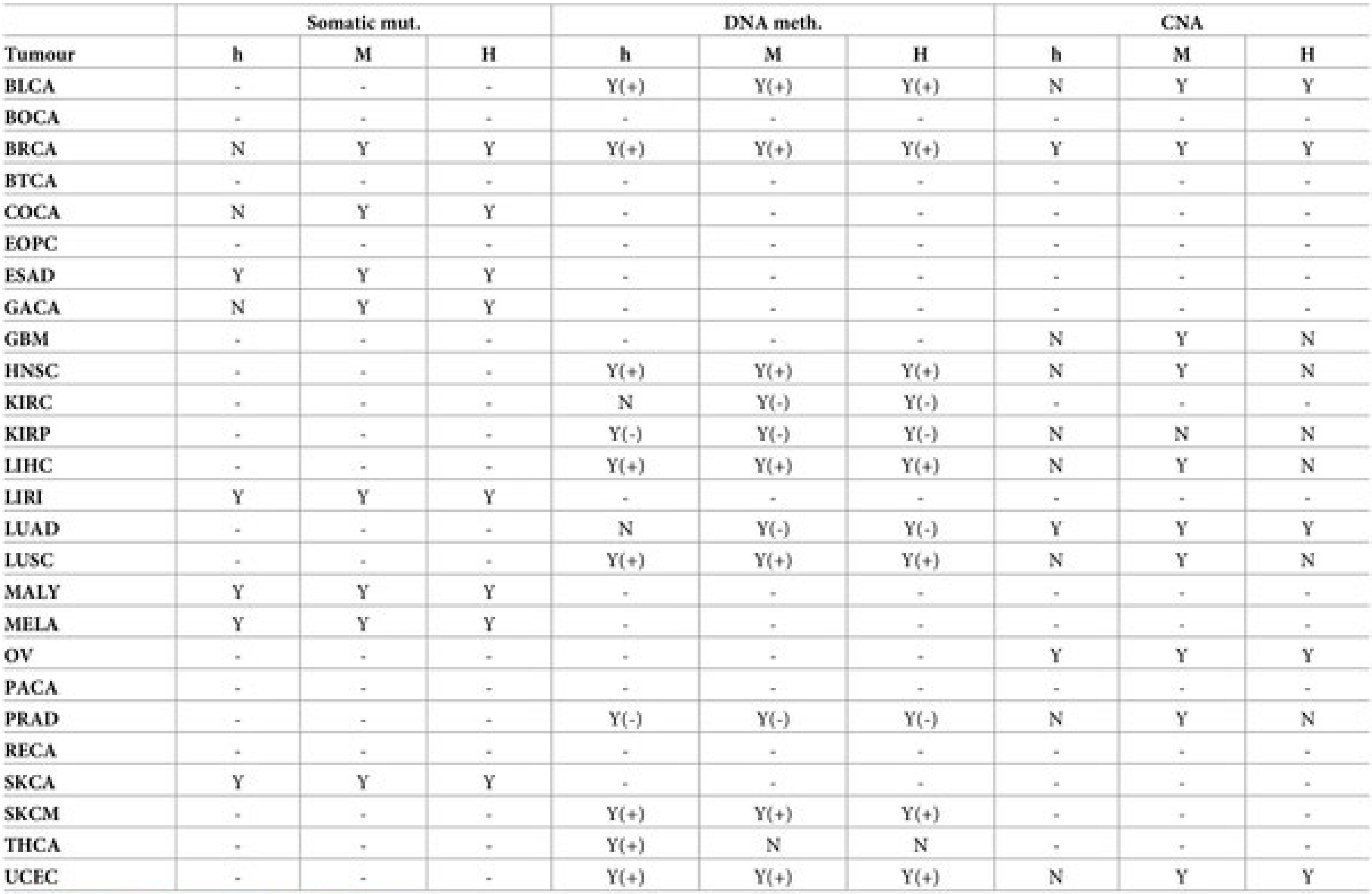
Pinoli P, Stamoulakatou E, Nguyen AP, Rodríguez Martínez M, Ceri S (2020) Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLOS ONE 15(1): e0227180. https://doi.org/10.1371/journal.pone.0227180
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227180

Fig 4. Genes close to mutated CTCF in-boundary motifs in melanoma.



Pinoli P, Stamoulakatou E, Nguyen AP, Rodríguez Martínez M, Ceri S (2020) Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLOS ONE 15(1): e0227180. https://doi.org/10.1371/journal.pone.0227180
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227180

Table 5. Significant alterations in cancer types.

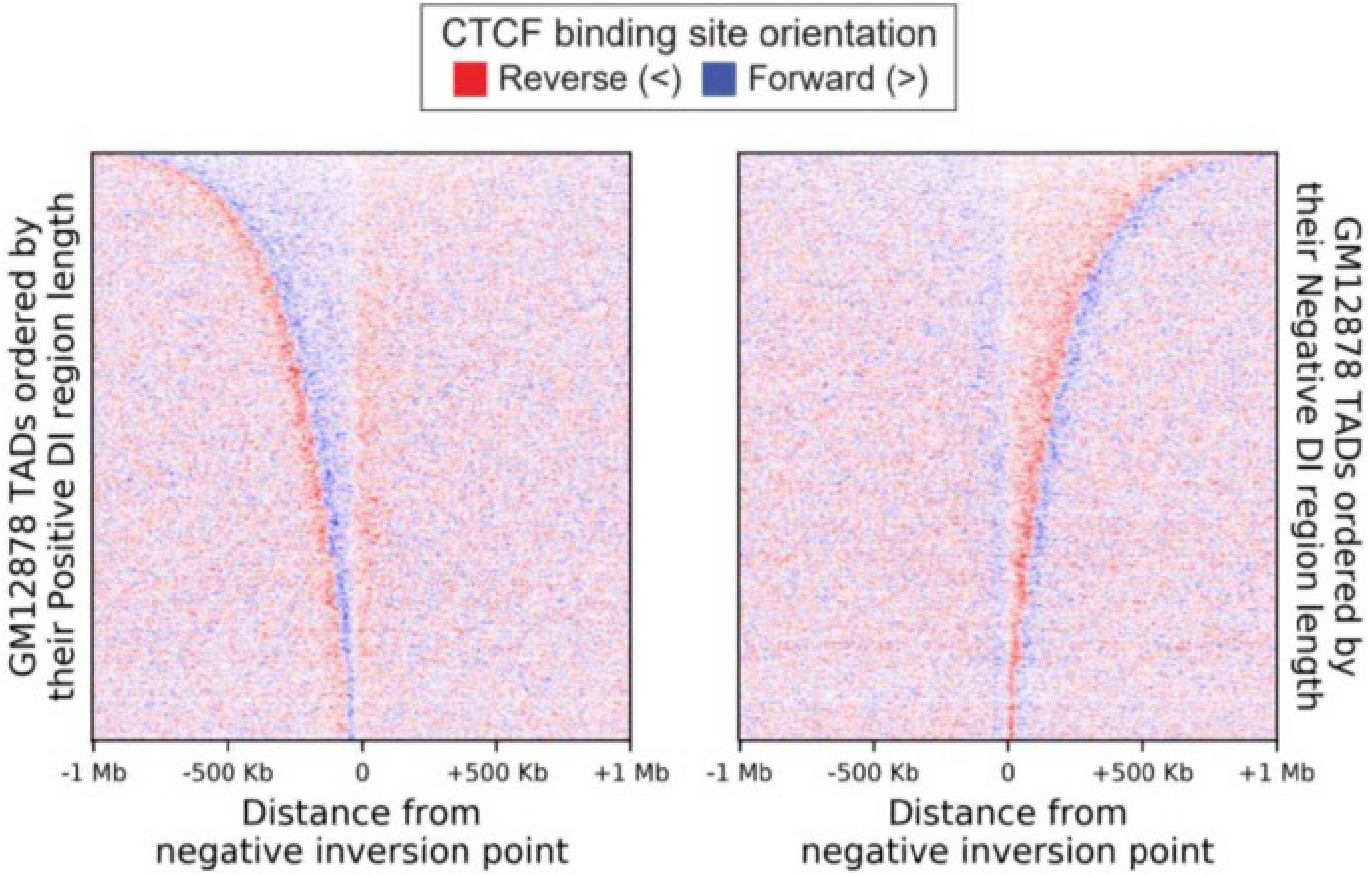


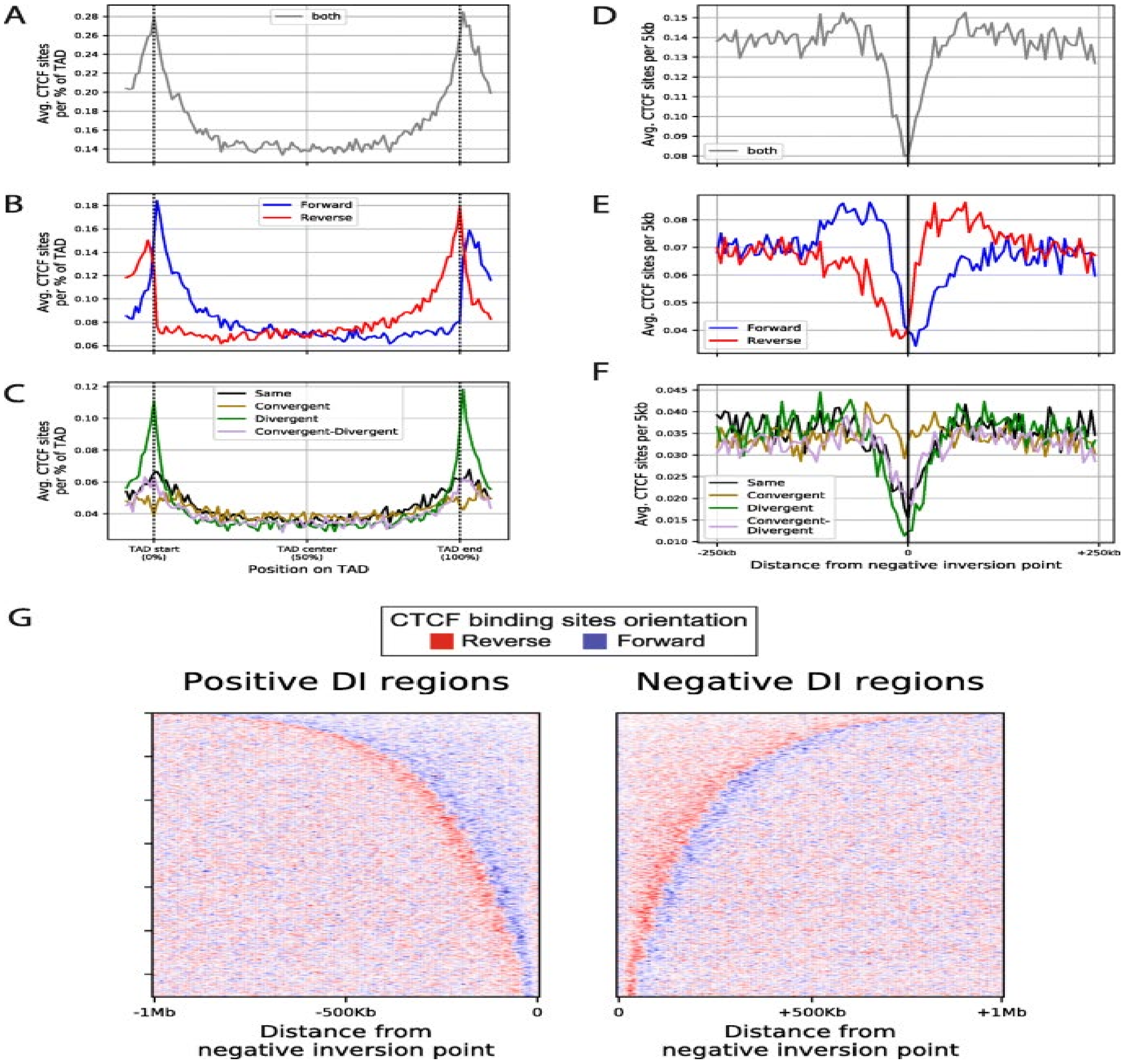
	Somatic mut.			DNA meth.			CNA		
Tumour	h	M	H	h	M	H	h	M	H
BLCA	-	-	-	Y(+)	Y(+)	Y(+)	N	Y	Y
BOCA	-	-	-	-	-	-	-	-	-
BRCA	N	Y	Y	Y(+)	Y(+)	Y(+)	Y	Y	Y
BTCA	-	-	-	-	-	-	-	-	-
COCA	N	Y	Y	-	-	-	-	-	-
EOPC	-	-	-	-	-	-	-	-	-
ESAD	Y	Y	Y	-	-	-	-	-	-
GACA	N	Y	Y	-	-	-	-	-	-
GBM	-	-	-	-	-	-	N	Y	N
HNSC	-	-	-	Y(+)	Y(+)	Y(+)	N	Y	N
KIRC	-	-	-	N	Y(-)	Y(-)	-	-	-
KIRP	-	-	-	Y(-)	Y(-)	Y(-)	N	N	N
LIHC	-	-	-	Y(+)	Y(+)	Y(+)	N	Y	N
LIRI	Y	Y	Y	-	-	-	-	-	-
LUAD	-	-	-	N	Y(-)	Y(-)	Y	Y	Y
LUSC	-	-	-	Y(+)	Y(+)	Y(+)	N	Y	N
MALY	Y	Y	Y	-	-	-	-	-	-
MELA	Y	Y	Y	-	-	-	-	-	-
OV	-	-	-	-	-	-	Y	Y	Y
PACA	-	-	-	-	-	-	-	-	-
PRAD	-	-	-	Y(-)	Y(-)	Y(-)	N	Y	N
RECA	-	-	-	-	-	-	-	-	-
SKCA	Y	Y	Y	-	-	-	-	-	-
SKCM	-	-	-	Y(+)	Y(+)	Y(+)	-	-	-
THCA	-	-	-	Y(+)	N	N	-	-	-
UCEC	-	-	-	Y(+)	Y(+)	Y(+)	N	Y	Y

https://doi.org/10.1371/journal.pone.0227180.t005

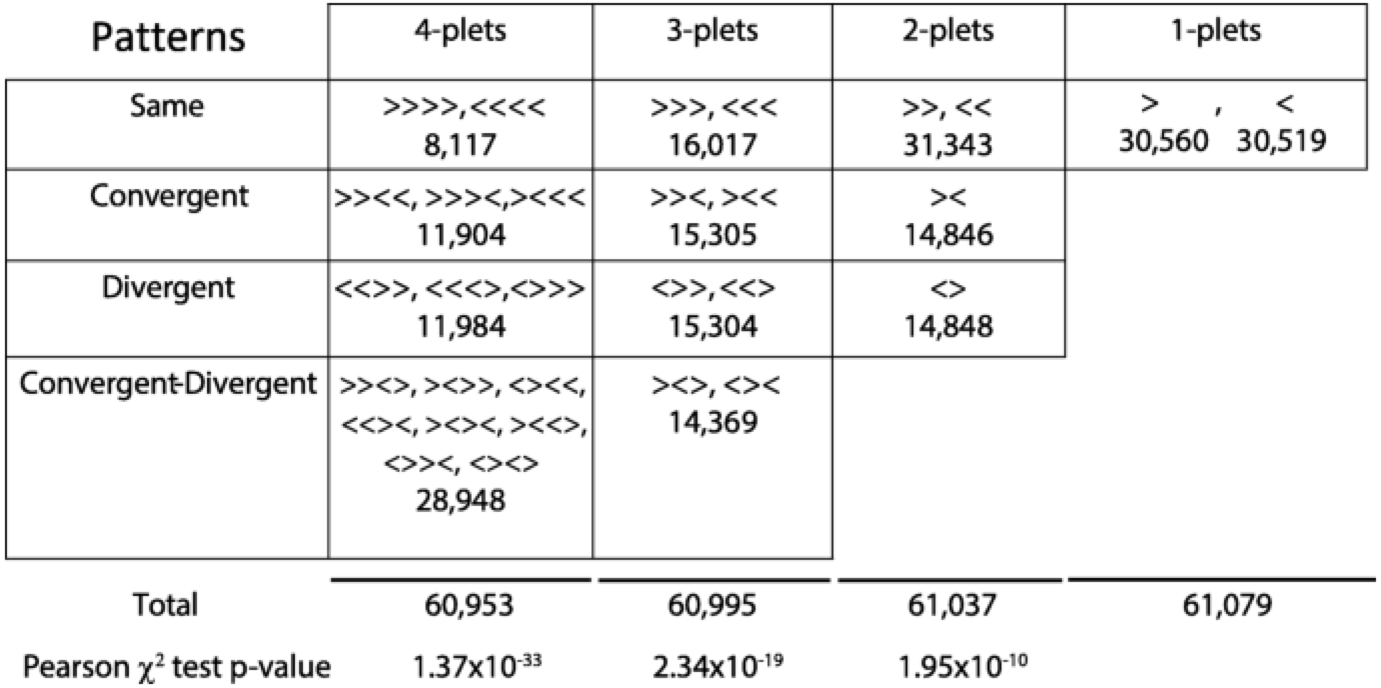
Pinoli P, Stamoulakatou E, Nguyen AP, Rodríguez Martínez M, Ceri S (2020) Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLOS ONE 15(1): e0227180. https://doi.org/10.1371/journal.pone.0227180
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0227180

Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries
Nanni, Ceri, Logie – Genome Biology 2020




A
Avg. CTCF sites per % of TAD
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
both
B
Avg. CTCF sites per % of TAD
0.18
0.16
0.14
0.12
0.10
0.08
0.06
Forward
Reverse
C
Avg. CTCF sites per % of TAD
0.12
0.10
0.08
0.06
0.04
Same
Convergent
Divergent
Convergent-Divergent
TAD start (0%)
TAD center (50%)
TAD end (100%)
Position on TAD
D
Avg. CTCF sites per 5kb
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
both
E
Avg. CTCF sites per 5kb
0.08
0.07
0.06
0.05
0.04
Forward
Reverse
F
Avg. CTCF sites per 5kb
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
Same
Convergent
Divergent
Convergent-Divergent
-250kb
0
+250kb
Distance from negative inversion point
G
CTCF binding sites orientation
Reverse
Forward
Positive DI regions
Negative DI regions
-1Mb
-500Kb
0
0
+500Kb
+1Mb
Distance from negative inversion point
Distance from negative inversion point

CTCF site clusters



Patterns	4-plets	3-plets	2-plets	1-plets
Same	>>>>,<<<< 8,117	>>>, <<< 16,017	>>, << 31,343	> , < 30,560 30,519
Convergent	>><<, >>><,><<< 11,904	>><, ><< 15,305	>< 14,846	
Divergent	<<>>, <<<>,<>>> 11,984	<>>,<<> 15,304	<> 14,848	
Convergent-Divergent	>><>, ><>>, <><<, <<><, ><><, ><<>, <>><, <><> 28,948	><>, <>< 14,369		
Total	60,953	60,995	61,037	61,079
Pearson χ^2 test p-value	1.37x10^{-33}	2.34x10^{-19}	1.95x10^{-10}	

Classification of CTCF site clusters by relative orientation. CTCF mono-plet, di-plet, tri-plet and tetra-plet adjacent binding sites in all possible patterns of relative orientation. Patterns are divided into four classes: *Same* (all sites oriented in the same direction), *Convergent* (sites pointing towards each other), *Divergent* (sites pointing away from each other) and, for tri-plets and tetra-plets, the class *Convergent + Divergent*.

Confirming work on mouse tissues
txs to: Lucia Morato Gomez
Work with L. Nanni, C. Logie, P. Pinoli

Patterns	4-plets	3-plets	2-plets	1-plets
Same	>>>> , <<<< 48,646 48,112 **96,758**	>>> , <<< 95,187 94,589 **189,776**	>> , << 188,788 188,265 **377,053**	> , < **381,740 381,221**
Convergent	>><<, >>><, ><<< 45,512 46,538 46,474 **138,524**	>>< , ><< 93,596 93,670 **187,266**	>< **192,944**	
Divergent	<<>>, <<<>, <>>> 45,573 46,475 46,537 **138,585**	<>> , <<> 93,596 93,671 **187,267**	<> **192,944**	
Convergent-Divergent	>><>, ><>>, <><<, <<><, ><><, ><<>, <>><, <><> 48,080 48,020 48,158 48,096 51,246 47,193 47,057 51,184 **389,034**	><> , <>< 99,267 99,345 **198,612**		
Total	**762,901**	**762,921**	**762,941**	**762,961**
p-value	**3.68x10^{-90}**	**1.39x10^{-98}**	**6x10^{-23}**	

Figure 14: Classification of CTCF site clusters by relative orientation for motifs not found under chIP seq peaks. p values calculated using the Pearson chi-square test.

TAD boundaries

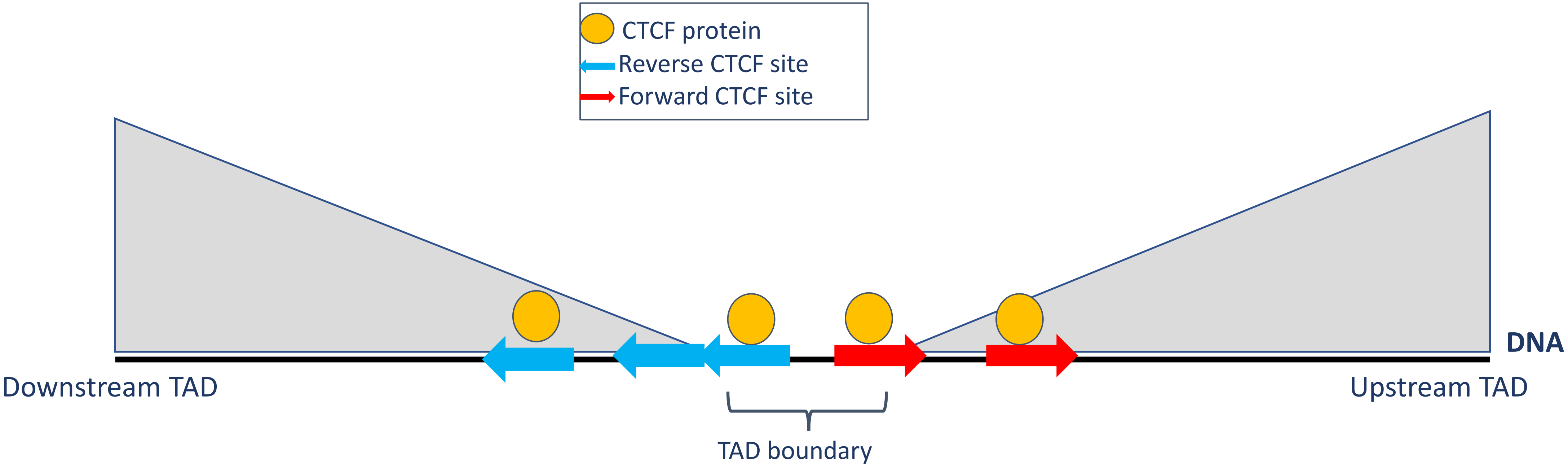


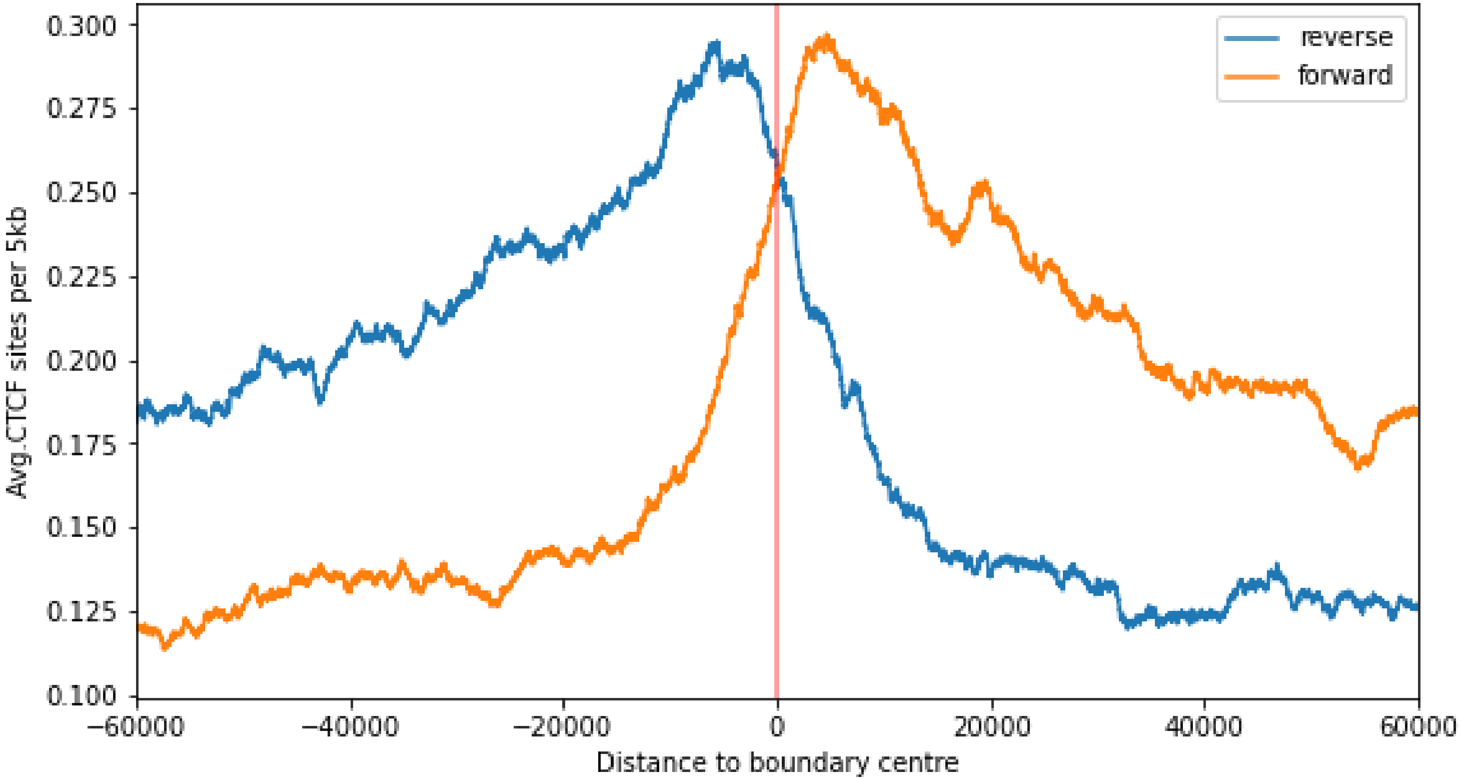
Figure 15: Current proposed model for CTCF relative orientation pattern at TAD boundaries

CTCF motifs under boundaries (1)

POLITECNICO MILANO 1863

Orientation analysis for non-intersecting CTCF motifs found under a peak with a single motif occurrence under boundaries – 100 percentile

Figure 19: Distribution of reverse (blue) and forward (orange) oriented CTCF motifs in 5kb bins across Louvain boundaries respecting their boundary centre



37,640 motifs
16,138 boundaries

Research Questions – some will be answered by Lucia’s work, some by future MS students

- Conservation of CTCF «strong boundaries» across species (humans and mice)
  - Possibly: separation between «constituents» TADS vs more «transient» loops (tissue-specific, possibly causing cancer)
- Integration with epigenetic data – gene promoters and enhancers (along the direction of work at Bernstein lab, to be studied, and taking advantage of the PLOS partial results).
- Cancer-specific studies, with possible indication for DNA-lab experiments (e.g. use CRISP technology to artificially remove bindings and measure gene expression for connected pairs enhancer-promoter)

The sound of genome

- Going beyond humans and mouses: is there an «higher level organization»?
- Can we predict missing properties (using ML)?
- CTCF binding directions respond to a sort of «grammar»; is this grammar also responsible of the other epigenetic signals?
- Can we discover the «sound of genome»? The figure looks like a music score, can we create a sound from genomic tracks?

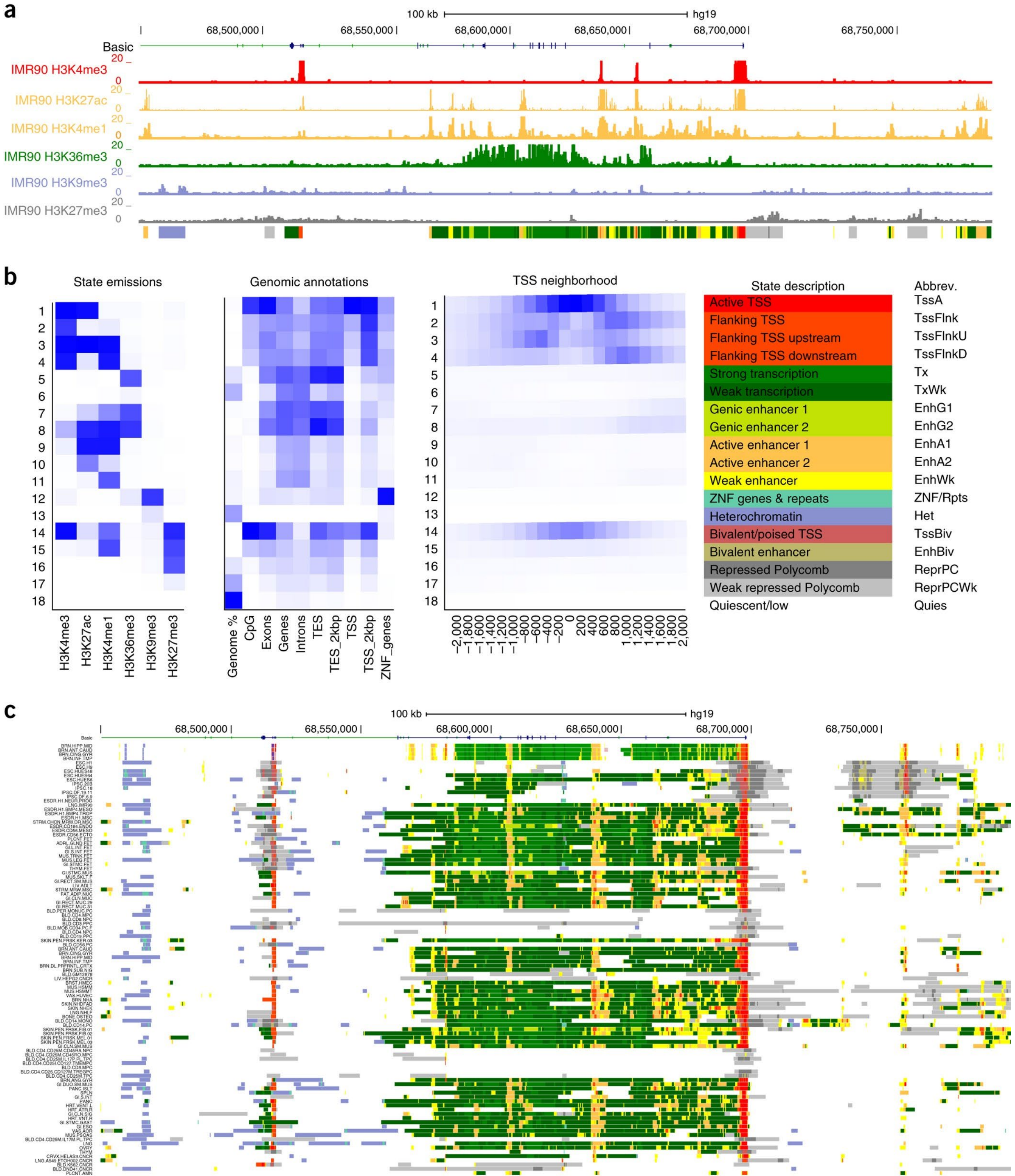


Figure 1: Overview of ChromHMM.

From: Chromatin-state discovery and genome annotation with ChromHMM (Nature Methods)

On the sound of genome… call for MS theses

- If you want to know more, send mail to Stefano.Ceri@Polimi.it
- Project will start in September 2023. Theses will be created based upon active collaborations at the time of first contact.
- Potential players:
  - Colin Logie (nejmegen)
  - Luca Nanni (human technopole)
  - Augusto Sarti (Sound Engineering Master Director, POLIMI)

    + some students from sound engineering
  - Luca Francesconi (composer)
  - Several PhDs