Characterizing the effect of genetic variants within promoters with distal enhancer functions

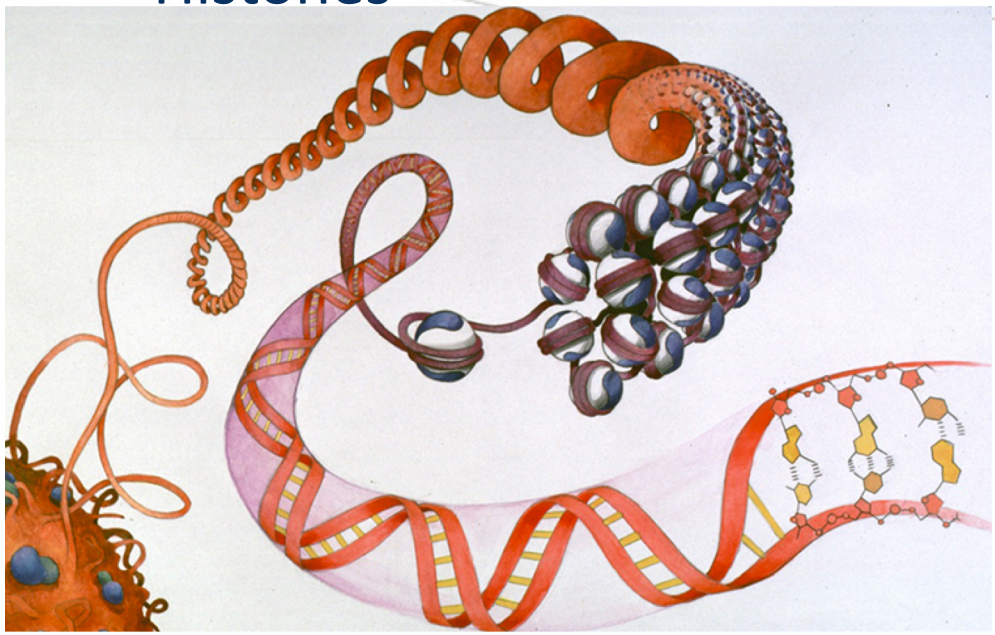
Alejandra Medina Rivera, PhD
Laboratorio Internacional de Investigación en Genoma Humano
Universidad Nacional Autónoma de México
Oaxaca, November 2018

LIIGH UNAM

Regulatory Genome

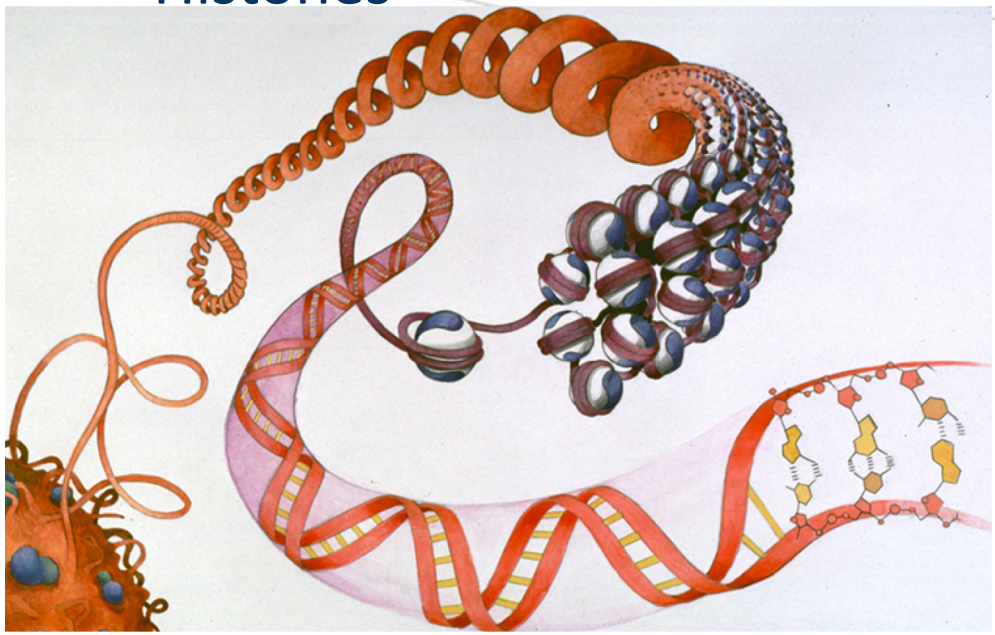
Regulatory Genome

Histones

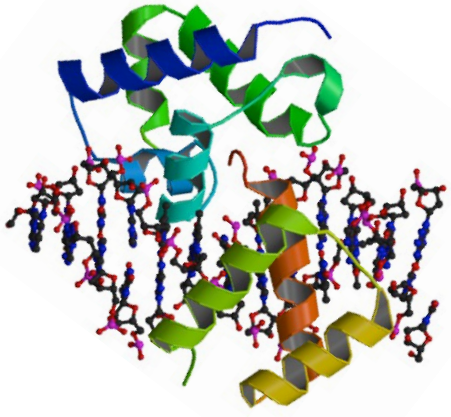


Regulatory Genome

Histones

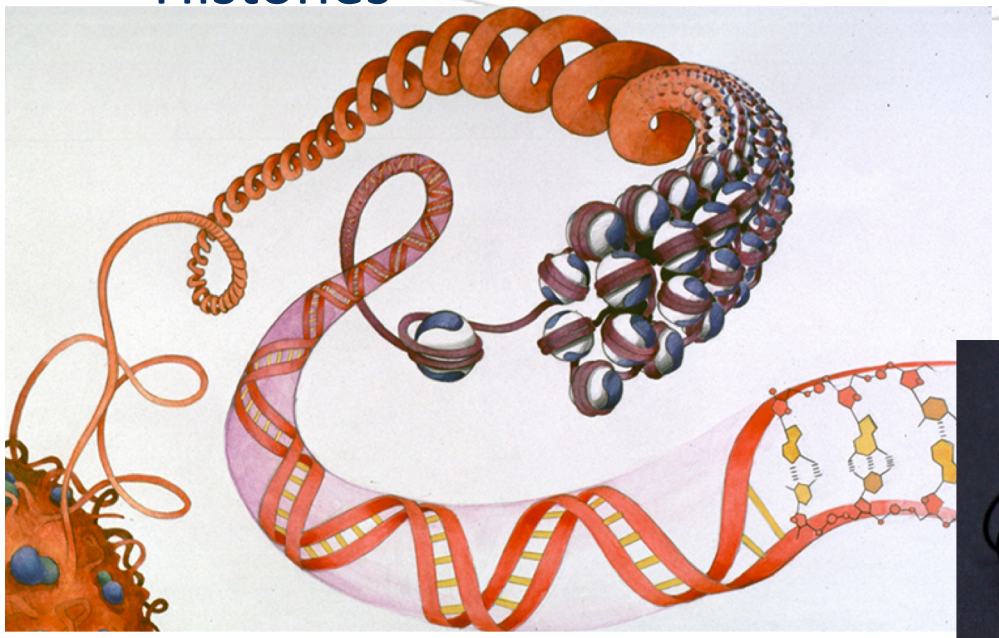


Transcriptional Factors

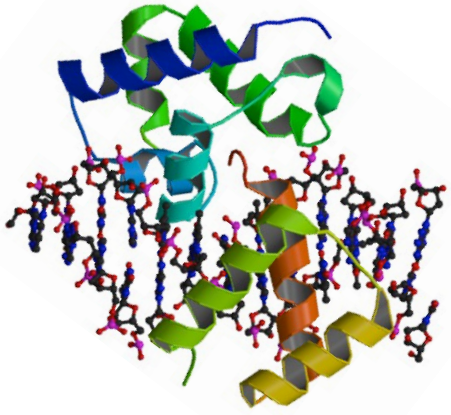


Regulatory Genome

Histones

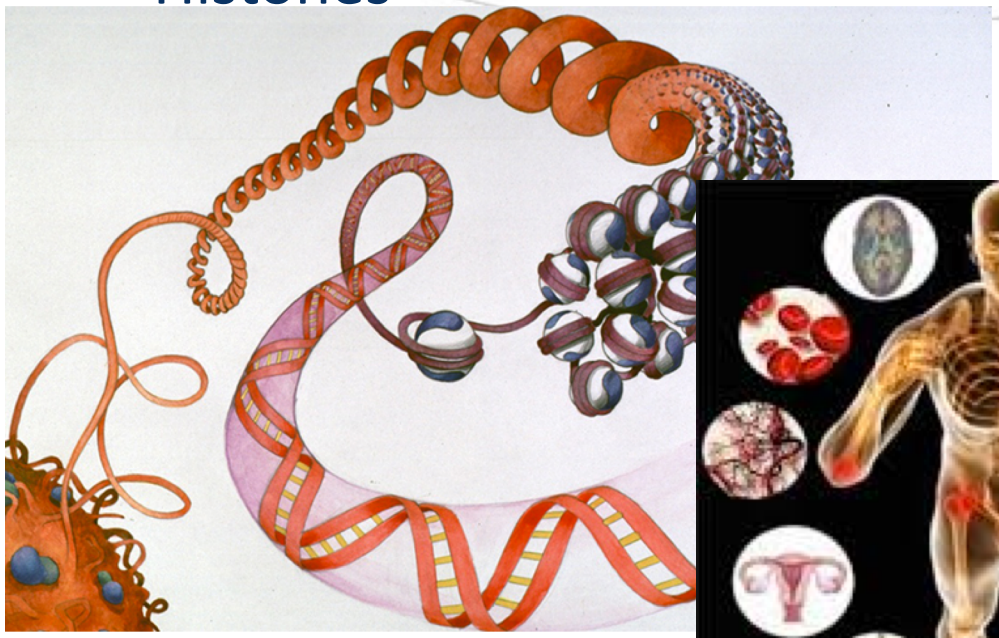


Transcriptional Factors

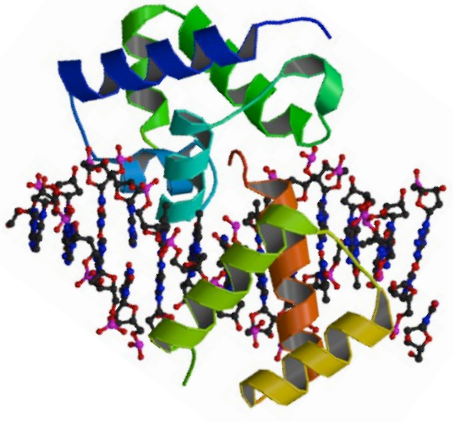


Regulatory Genome

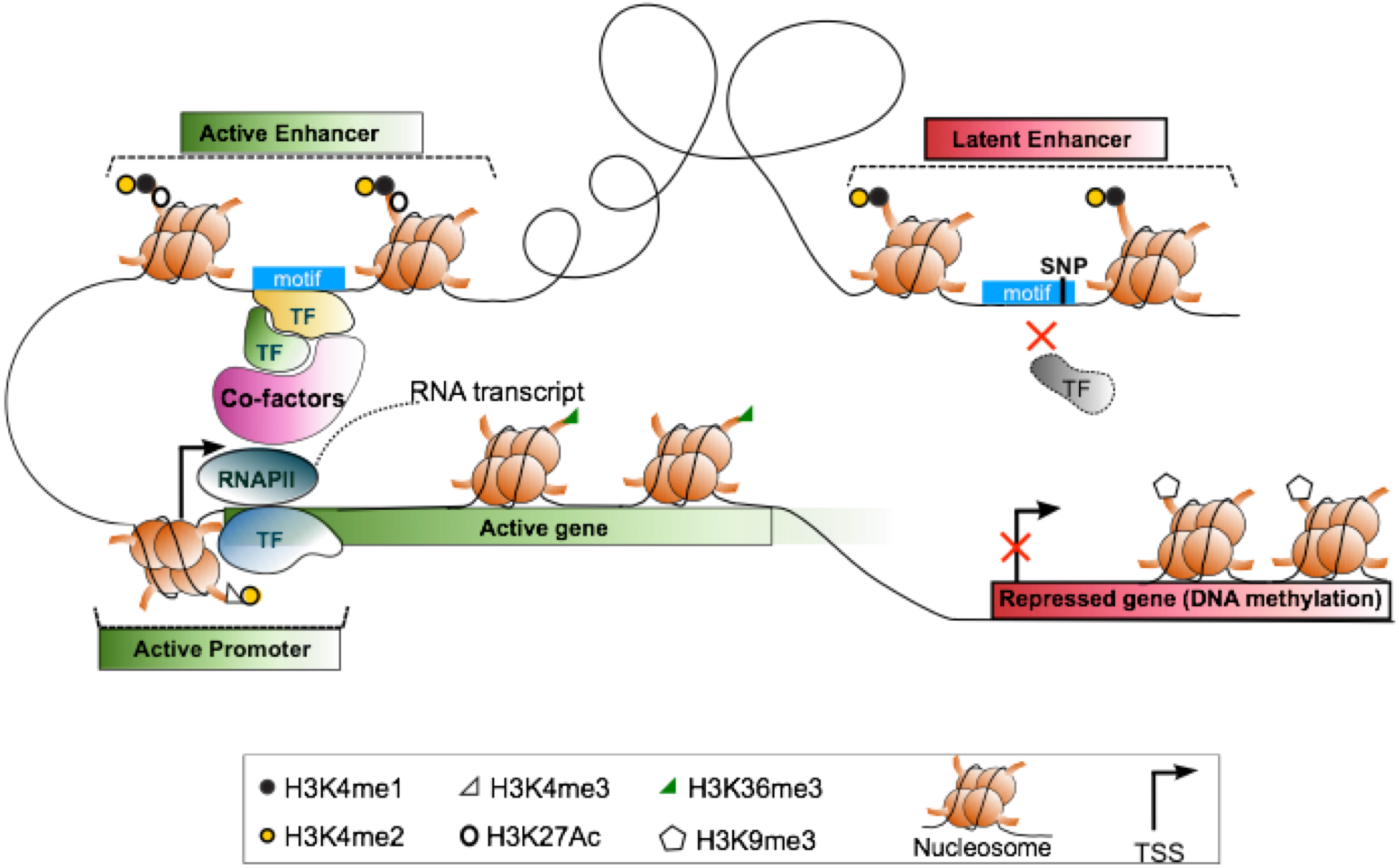
Histones



Transcriptional Factors

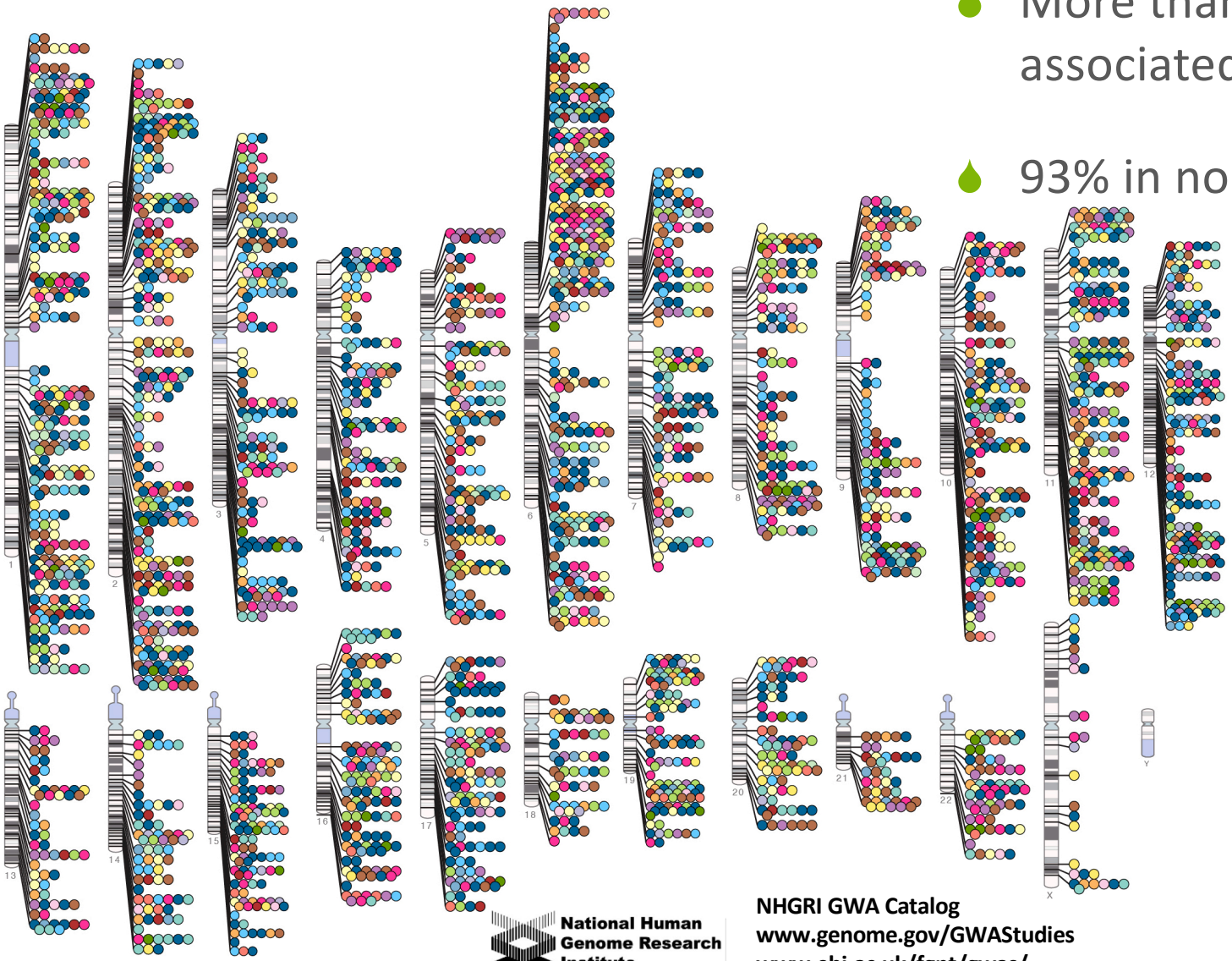


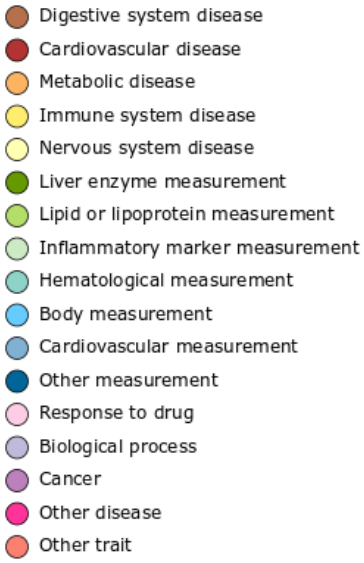
Annotating the Regulatory Genome



Published Genome-Wide Associations since 12/2012
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

- More than 17,000 reported associated SNPs
- 93% in non-coding sequence





National Human Genome Research Institute

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

EMBL-EBI

First characterized Enhancers

Gene	Origin	Size (bp)	Distance from TSS	
Early gene	SV40	196	~200	Banerji, J. et al. (1981); Benoist, C. and Chambon, P. (1981)
Early gene	Cytomegalovirus	406	−524 to −118	Boshart, M. et al. (1985)
Hsp70	*Xenopus*	160	−260 to −100	Bienz, M. and Pelham, H.R. (1986)
Fos	Human	340	−404 to −64	Deschamps, J. et al. (1985)
hMT-IIA	Human	327	−366 to −39	Serfling, E. et al. (1985)
Mmt-IA	Mouse	114 155	−187 to −73 −194 to −39	Serfling, E. et al. (1985)
H2A	Urchin	28	−139 to −111	Grosschedl, R. and Birnstiel, M.L. (1982)
IFNb	Human	40	−77 to −37	Goodbourn, S. et al. (1985)

Medina-Rivera, *Trends in Biochemical Sciences*, 2018

Regulatory Regions: Enhancers vs Promoters

Gene	Origin	Size (bp)	Distance from TSS	
Early gene	SV40	196	~200	Banerji, J. et al. (1981); Benoist, C. and Chambon, P. (1981)
Early gene	Cytomegalovirus	406	−524 to −118	Boshart, M. et al. (1985)
Hsp70	*Xenopus*	160	−260 to −100	Bienz, M. and Pelham, H.R. (1986)
Fos	Human	340	−404 to −64	Deschamps, J. et al. (1985)
hMT-IIA	Human	327	−366 to −39	Serfling, E. et al. (1985)
Mmt-IA	Mouse	114 155	−187 to −73 −194 to −39	Serfling, E. et al. (1985)
H2A	Urchin	28	−139 to −111	Grosschedl, R. and Birnstiel, M.L. (1982)
IFNb	Human	40	−77 to −37	Goodbourn, S. et al. (1985)

Medina-Rivera, *Trends in Biochemical Sciences*, 2018

How to classify Regulatory Regions?

Features (active elements)	Promoter	Enhancer
Intrinsic property	Induce transcription of a heterologous reporter gene	Activate a distal (heterologous) promoter
Transcription initiation	Unidirectional or divergent	Mainly divergent
Ratio between sense and antisense transcripts	Biased toward sense transcription	Equilibrated
Transcription elongation	Produce long polyadenylated transcripts	Some enhancers can produce low levels of polyadenylated transcripts
Histone modifications	H3K27ac (H3K4me1 < H3K4me3)	H3K27ac (H3K4me1 > H3K4me3)
RNAPII and GTF	Present	Present
GpG islands	Majority	Very rare

Medina-Rivera, *Trends in Biochemical Sciences*, 2018

How to assess the enhancer function of promoters?

Major challenge: How to assess the enhancer function of promoters?

Salvatore Spicuglia

TAGC Inserm U1090
technological advances for genomics and clinics

Inserm

LIIGH UNAM

UNIVERSITE MONTPELLIER

GR-EX LABORATORY OF EXCELLENCE GR-EX
THE RED BLOOD CELL: FROM GENESIS TO DEATH

Detection of long-range regulatory regions Enhancers

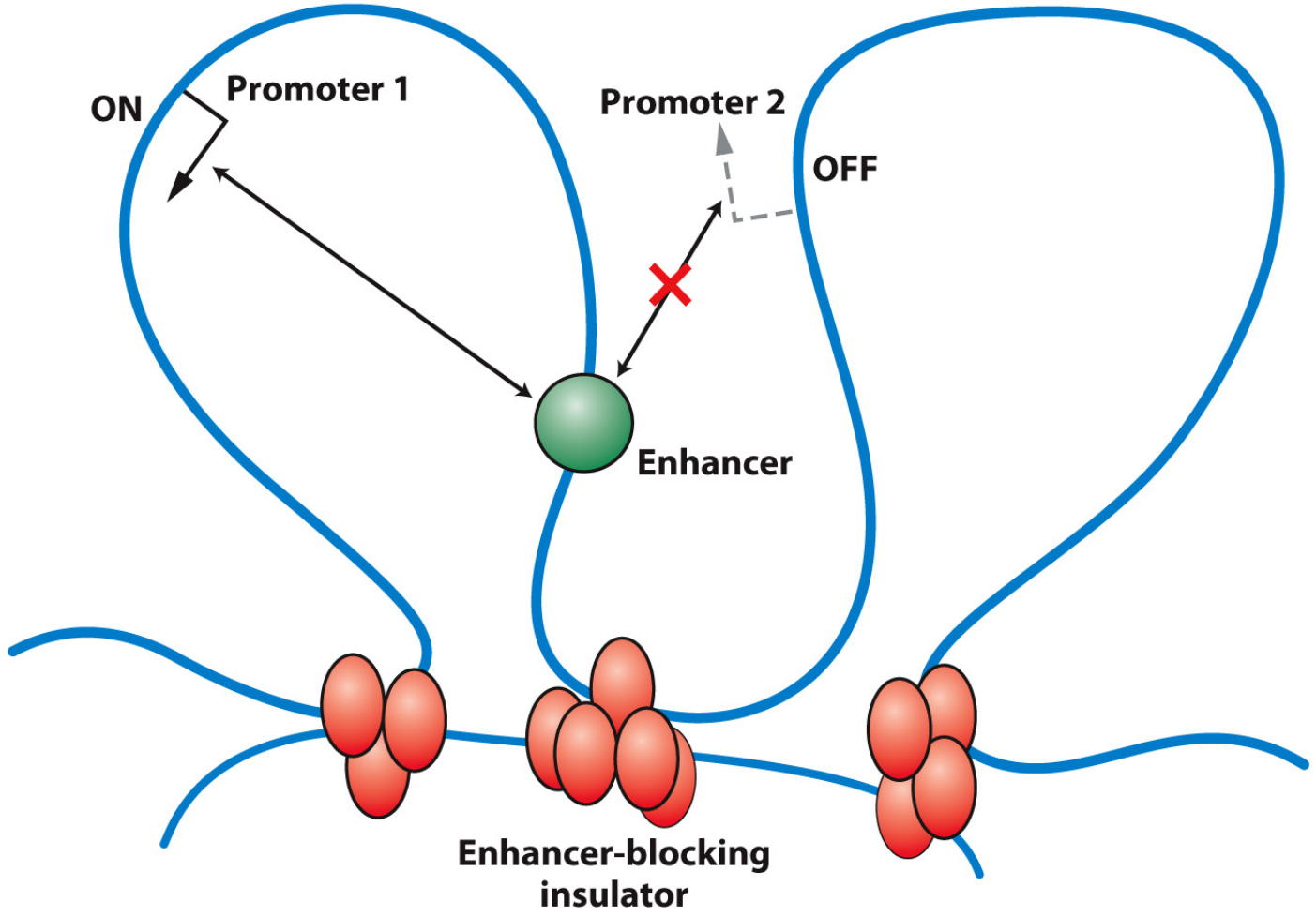
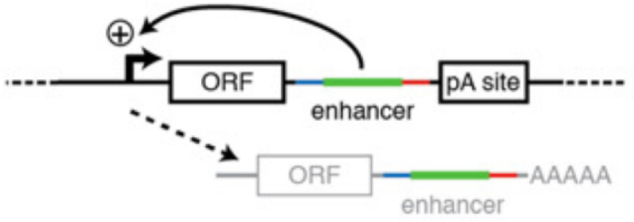


Figure 12-21
Introduction to Genetic Analysis, Eleventh Edition

Detection of long-range regulatory regions Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-.).,

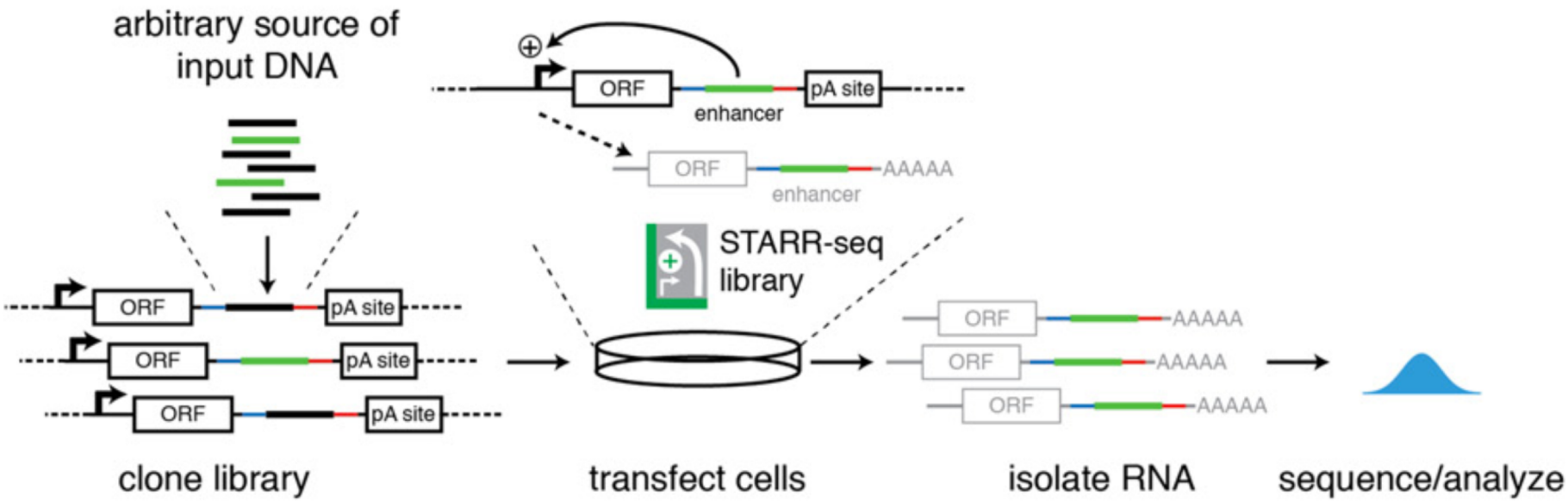


Muerdter, F. et al. (2015).106, 145–150

Detection of long-range regulatory regions Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-.).,

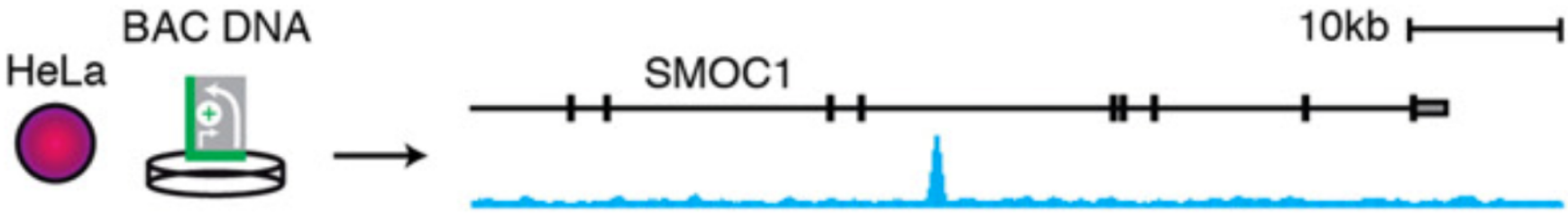


Muerdter, F. et al. (2015).106, 145–150

Detection of long-range regulatory regions Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-.).,

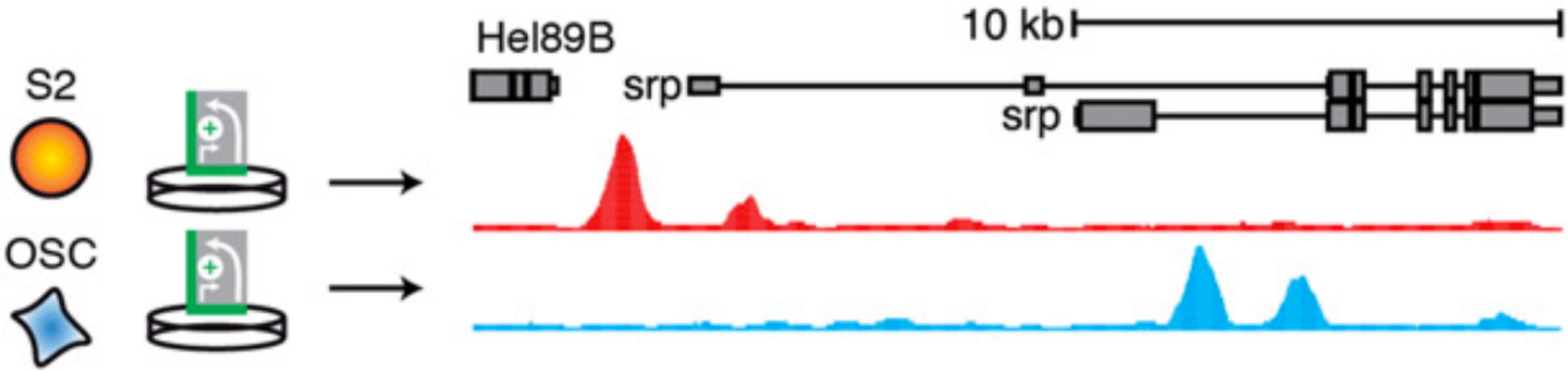


Muerdter, F. et al. (2015).106, 145–150

Detection of long-range regulatory regions Enhancers

- Starr-seq: High-throughput assessment of sequence enhancer potential

Arnold, C.D. et al. (2013). Science (80-.).,



Muerdter, F. et al. (2015).106, 145–150

Major challenge: How to assess the enhancer function of promoters?

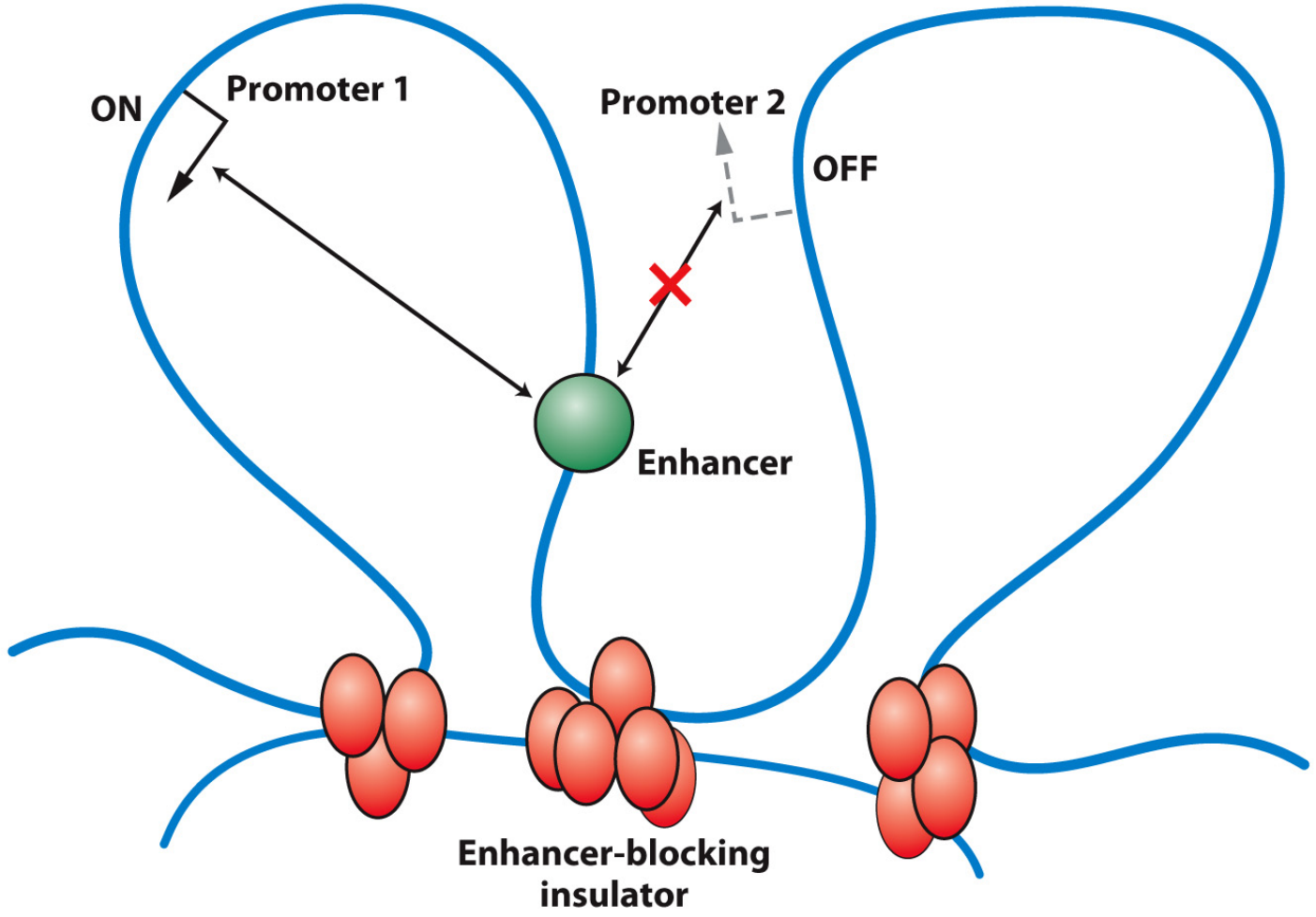


Figure 12-21
Introduction to Genetic Analysis, Eleventh Edition

Major challenge: How to assess the enhancer function of promoters?

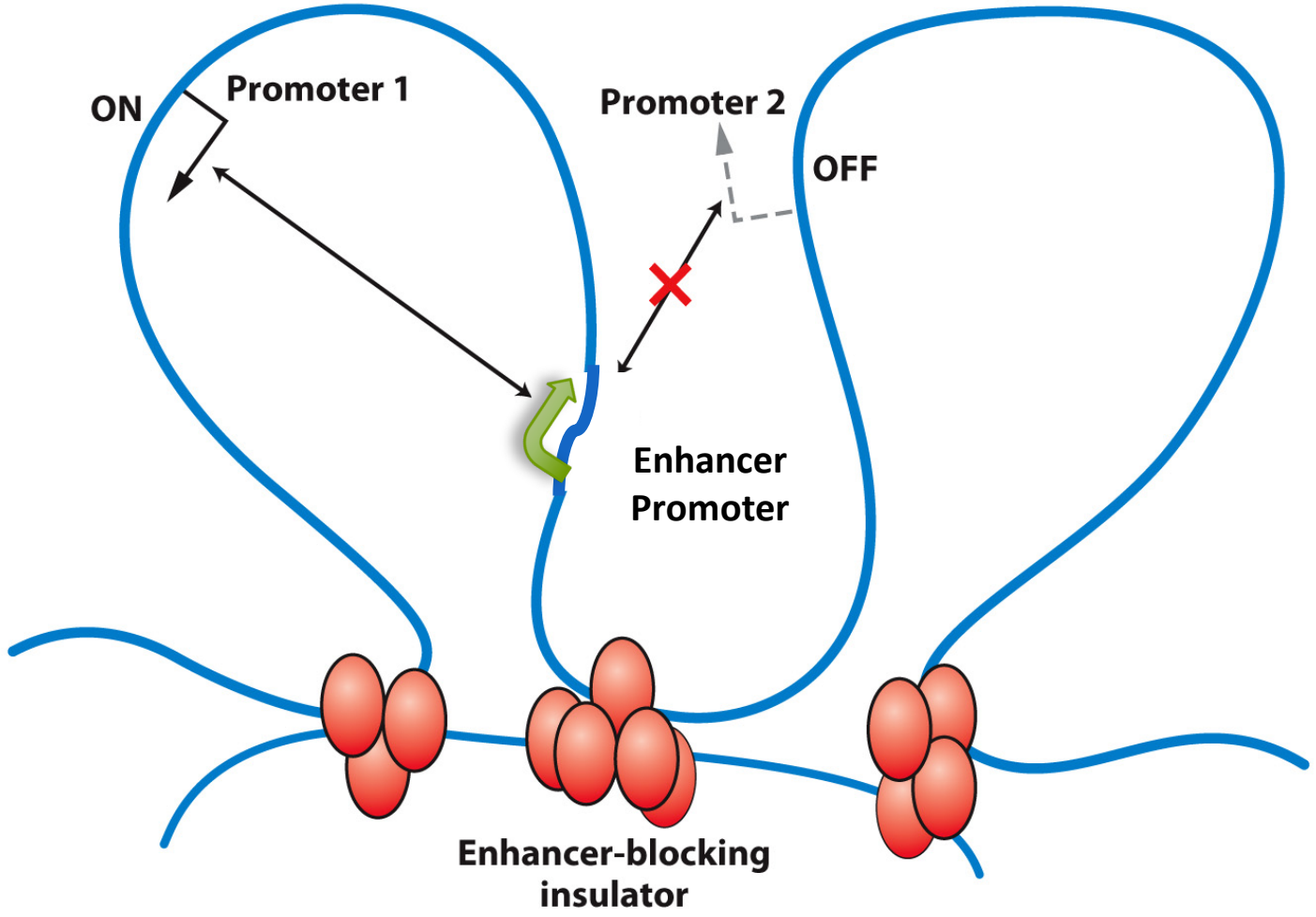
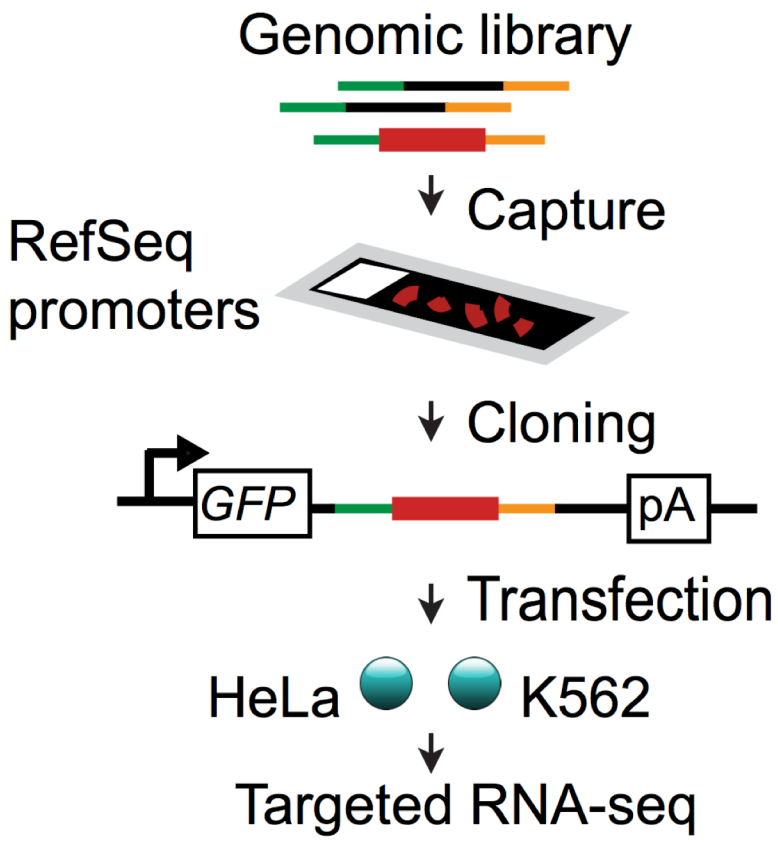


Figure 12-21
Introduction to Genetic Analysis, Eleventh Edition

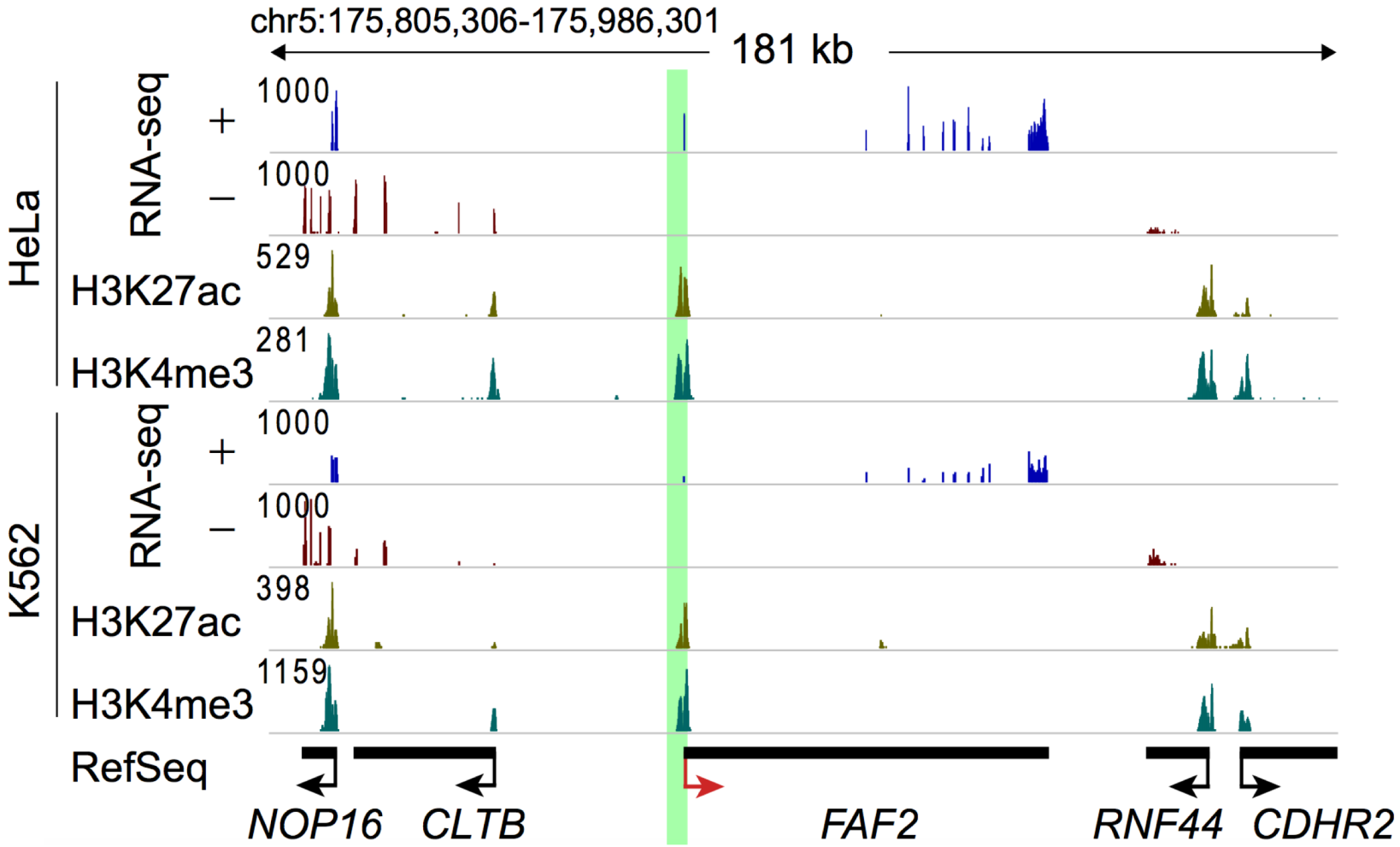
Can promoters work as enhancers? CapStarr-seq

Salvatore Spicuglia



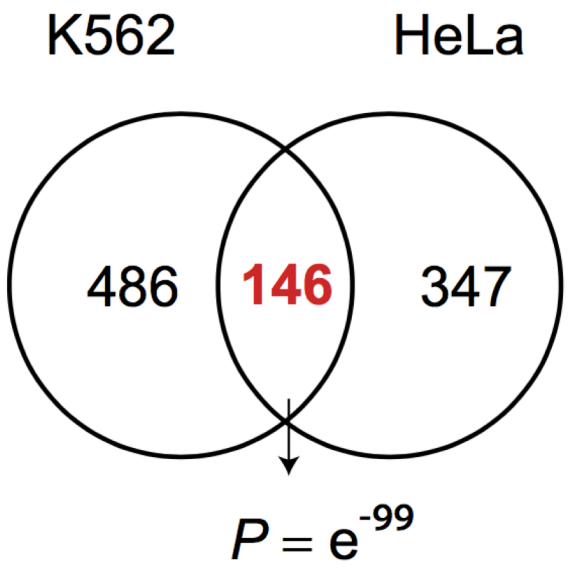
Lan Dao, et al, Nature Genetics. 2017

Can promoters work as enhancers?



Lan Dao, et al, Nature Genetics. 2017

How many promoters work also as enhancers?

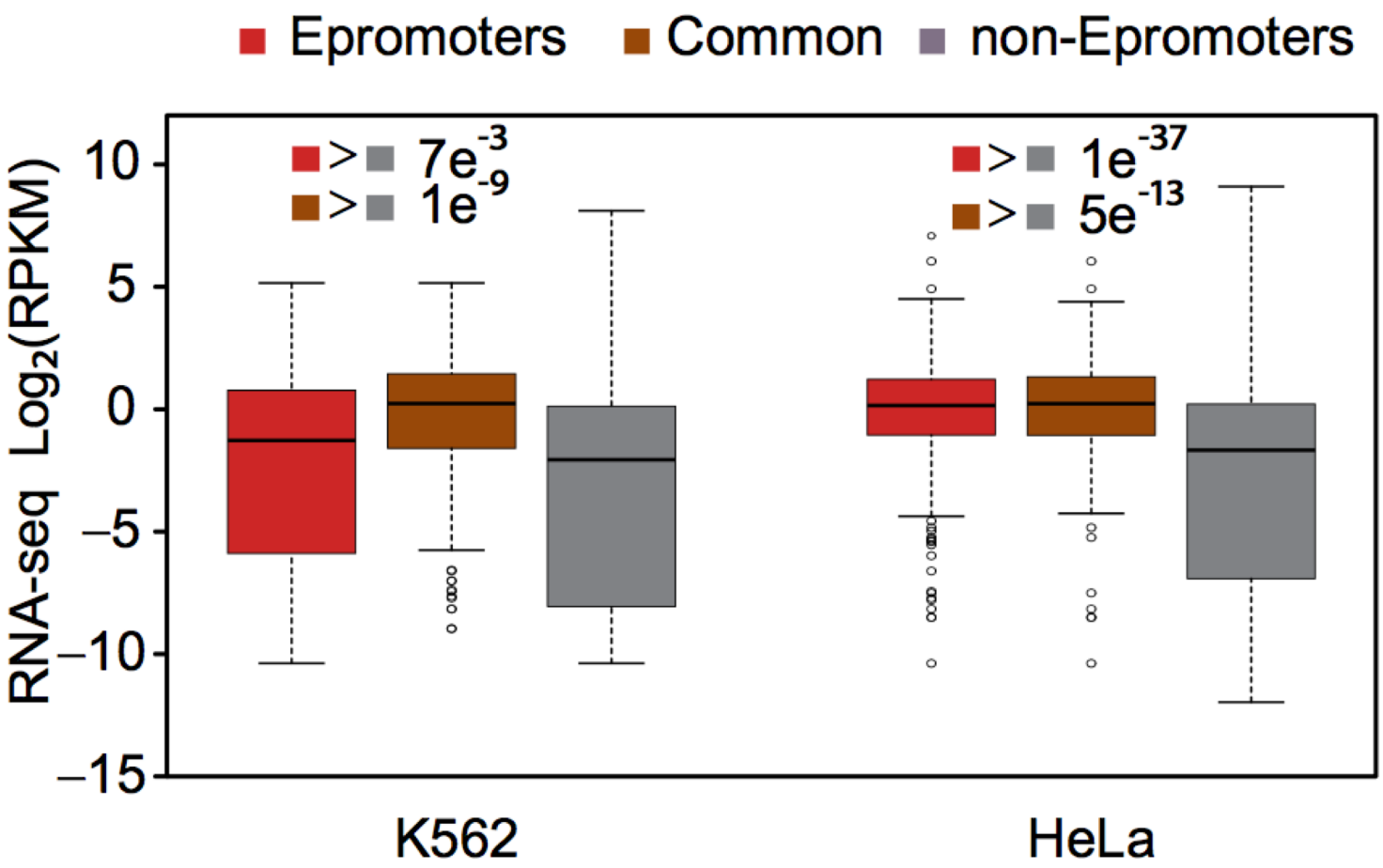


Total of ePromoters:

- K562= 632 (3%)
- HELA= 493 (2.37%)
- Total analyzed promoters= 20,719 (100%)

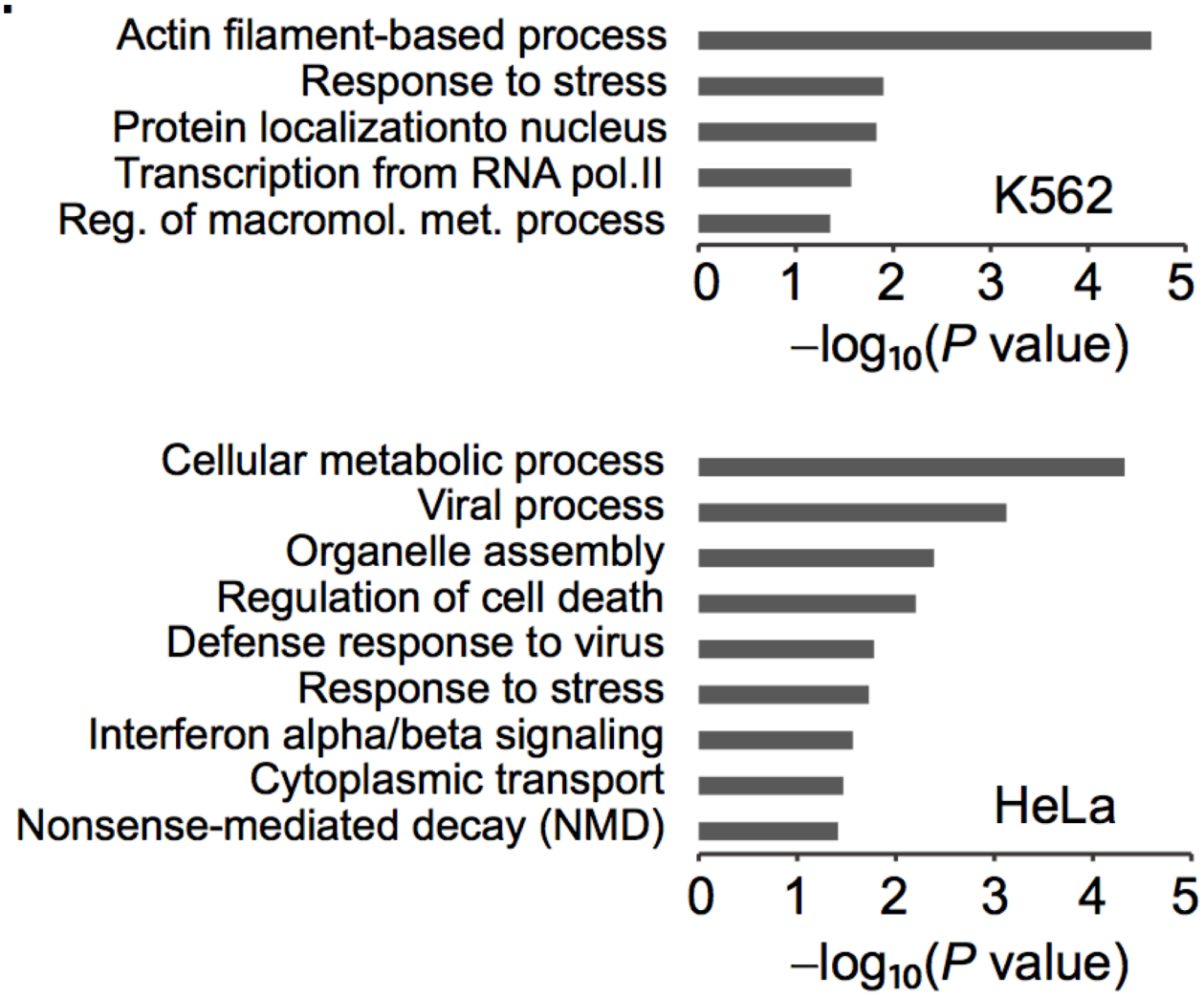
Lan Dao, et al, Nature Genetics. 2017

RNA expression of genes downstream promoters



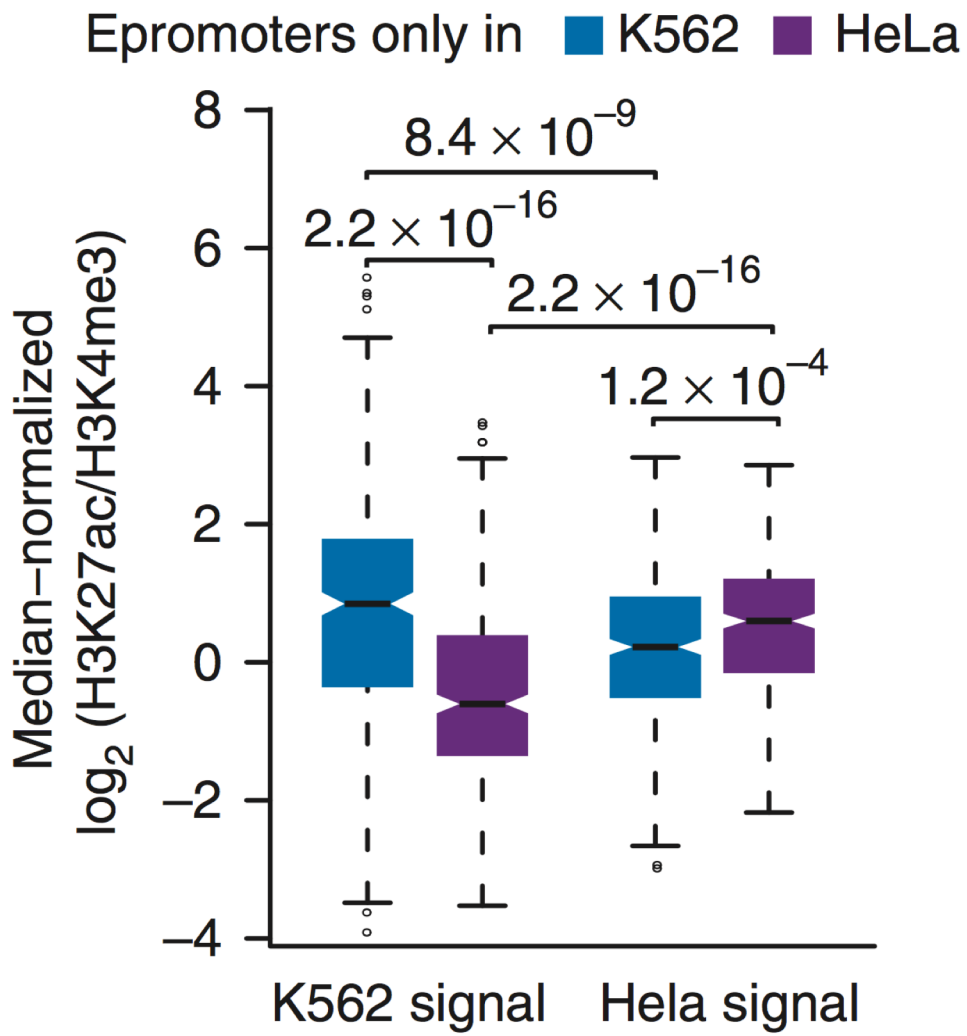
Lan Dao, et al, Nature Genetics. 2017

Functional annotation of genes downstream Epromoters



Lan Dao, et al, Nature Genetics. 2017

Epromoters had a higher H3K27ac/H3K4me3 ratio in the cell type where they were found to be active

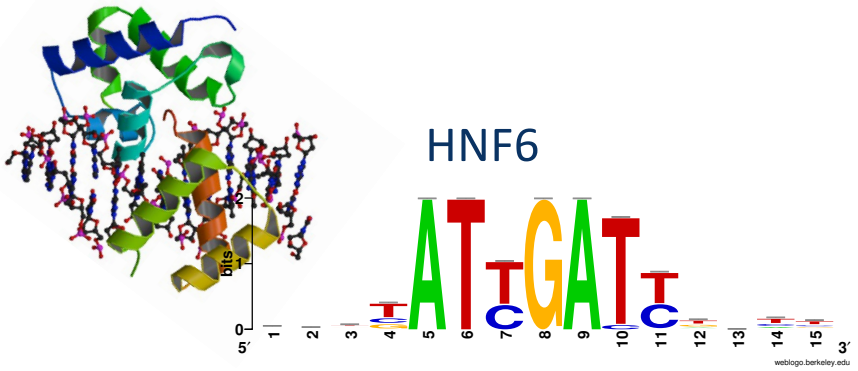


Lan Dao, et al, Nature Genetics. 2017

Which are the regulatory mechanisms affecting Epromoters?

Transcription factors control gene regulation by binding to specific DNA sequences

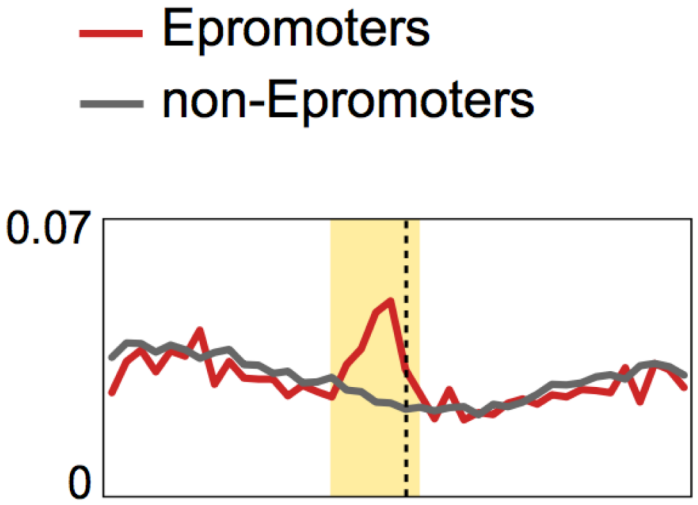
- Transcription Factors interact with DNA binding to sequence specific sites.

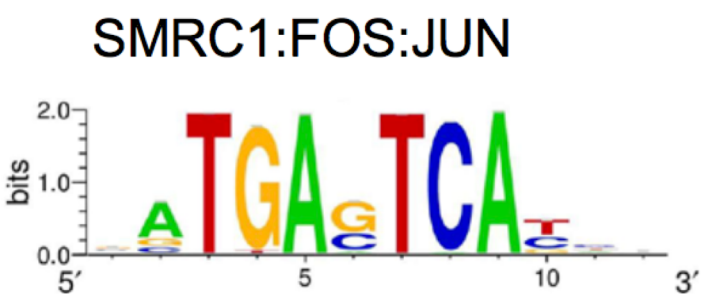


LIIGH UNAM

Transcription Factors enriched motifs in Epromoters

K562





E-value

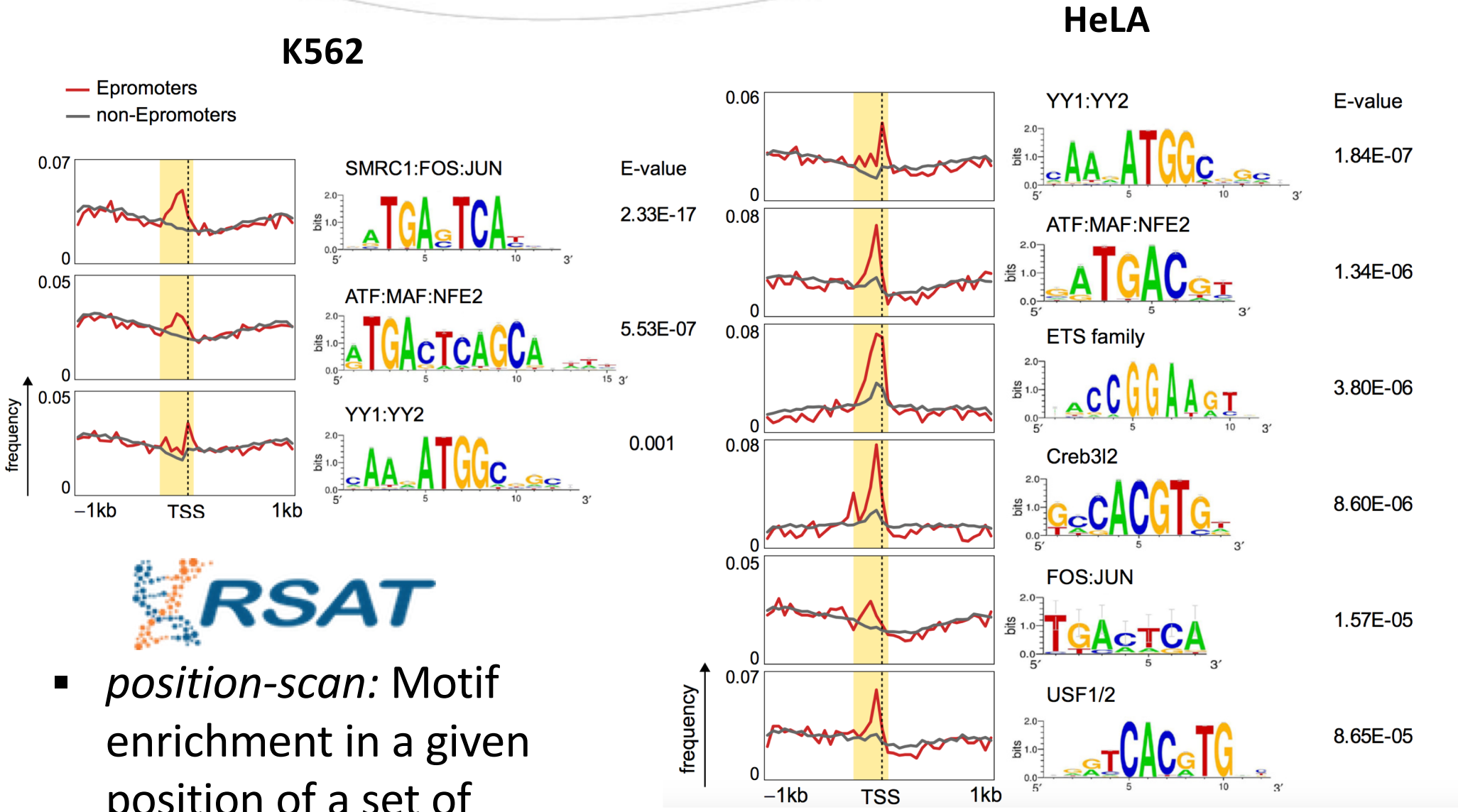
2.33E-17

RSAT

- *position-scan:* Motif enrichment in a given position of a set of sequences

Lan Dao, et al, Nature Genetics. 2017

Transcription Factors enriched motifs in Epromoters

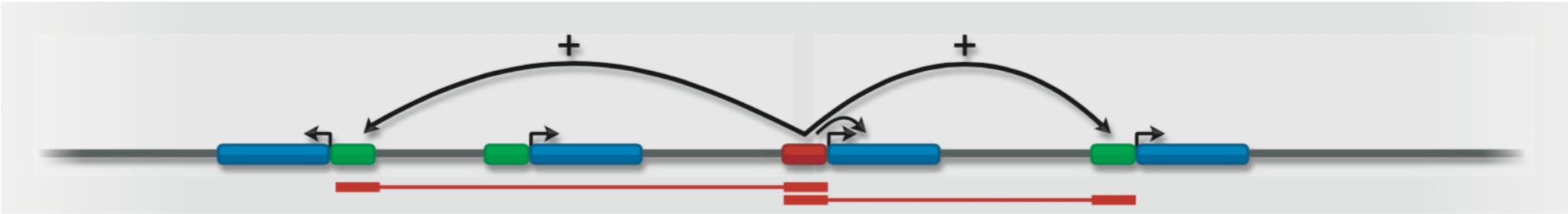


RSAT

- *position-scan:* Motif enrichment in a given position of a set of sequences

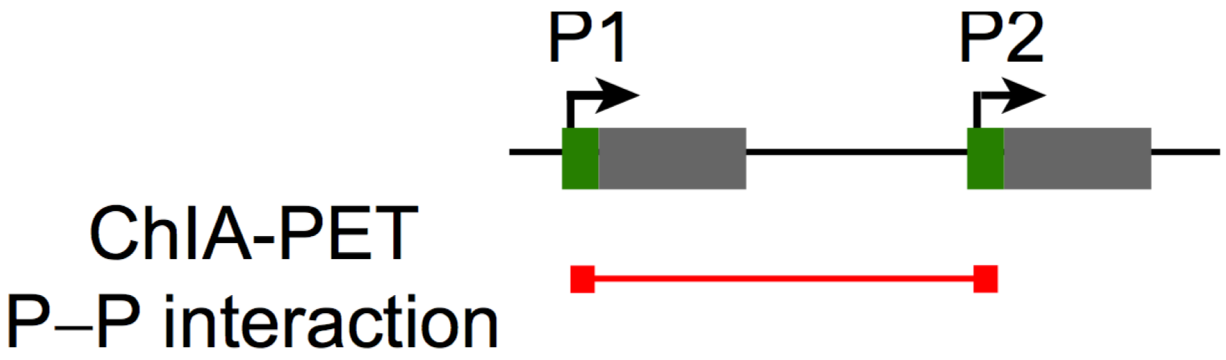
Lan Dao, et al, Nature Genetics. 2017

Which are the genes regulated by Epromoters?



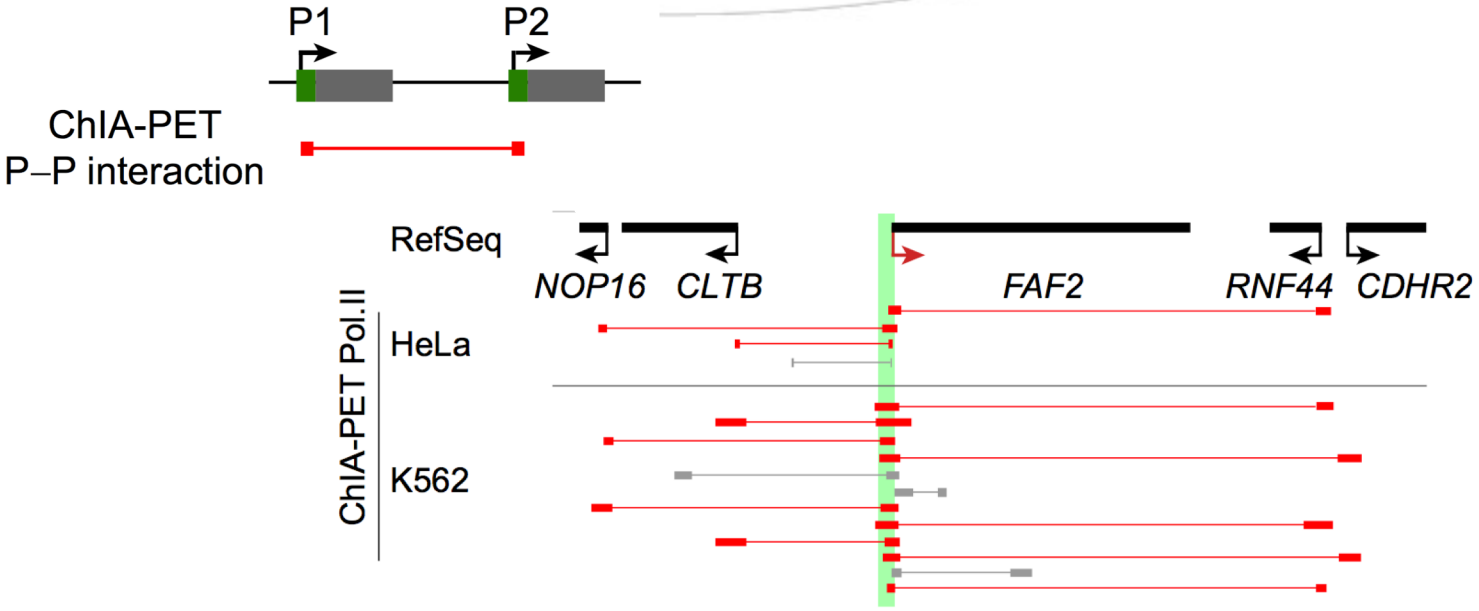
Lan Dao, et al, Nature Genetics. 2017

Which are the genes regulated by Epromoters?



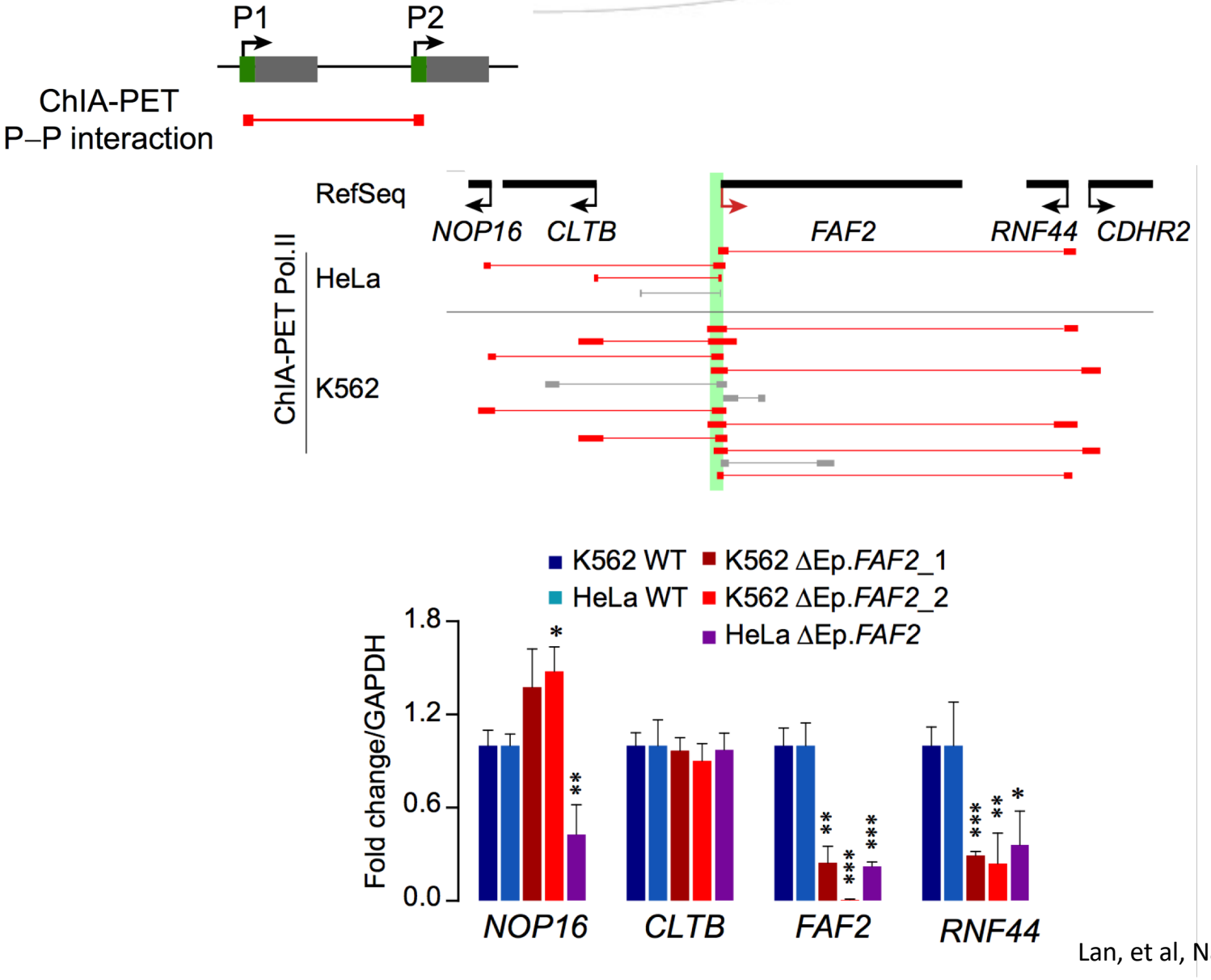
Lan Dao, et al, Nature Genetics. 2017

Which are the genes regulated by Epromoters?



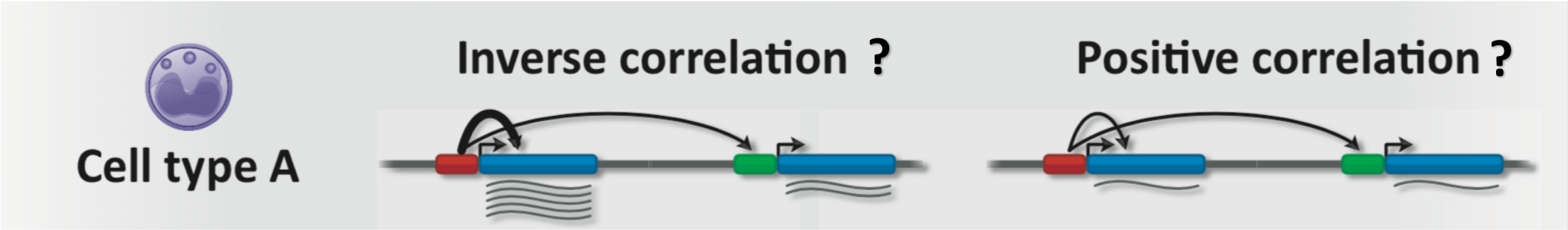
Lan Dao, et al, Nature Genetics. 2017

Which are the genes regulated by Epromoters?



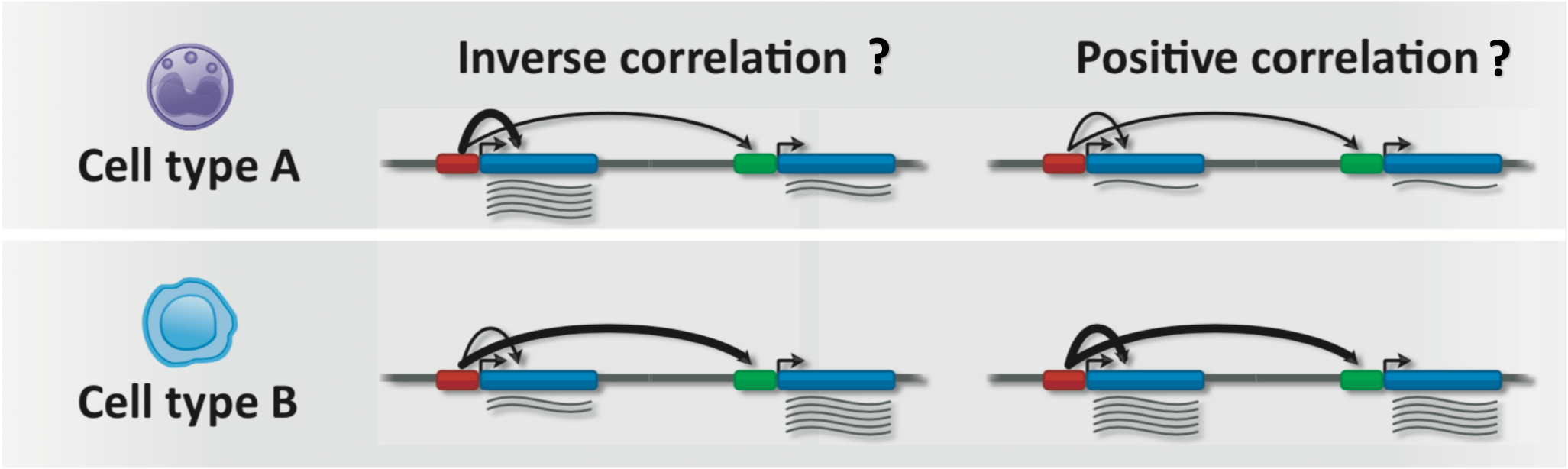
Lan, et al, Nature Genetics, 2017.

What is the effect of Epromoters on long range regulated genes?



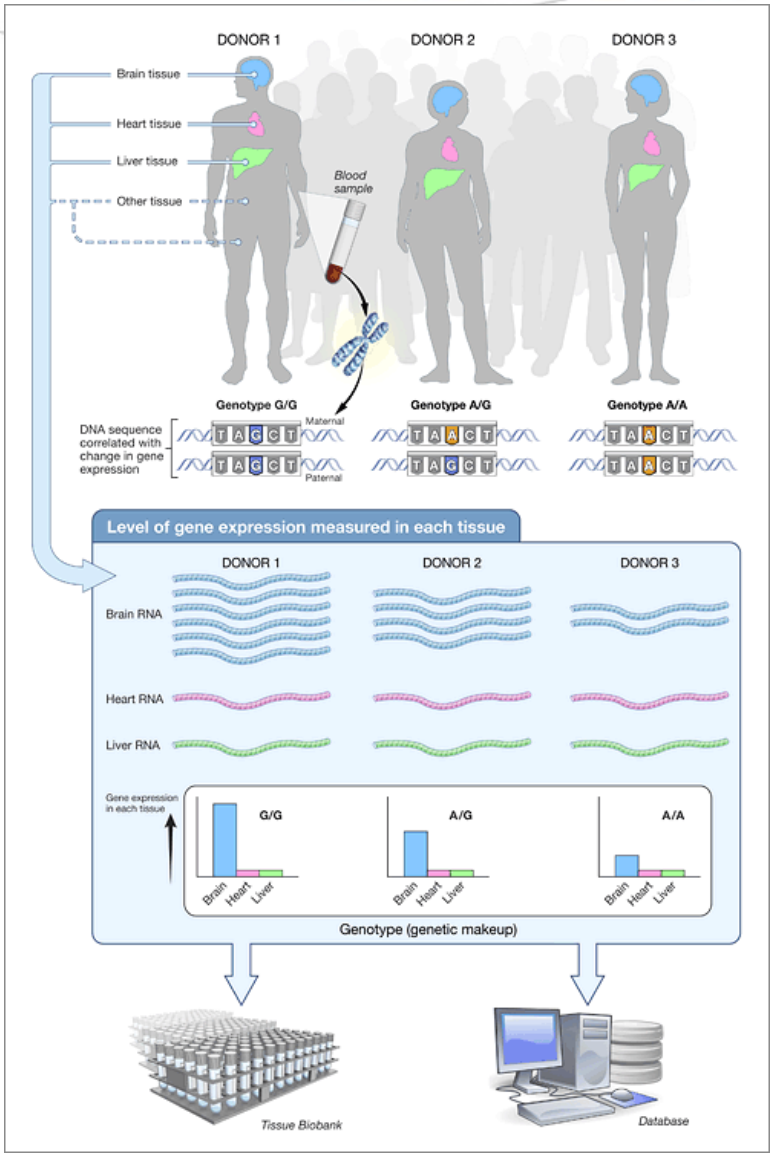
Medina-Rivera, *Trends in Biochemical Sciences*, 2018

What is the effect of Epromoters on long range regulated genes?



Medina-Rivera, *Trends in Biochemical Sciences*, 2018

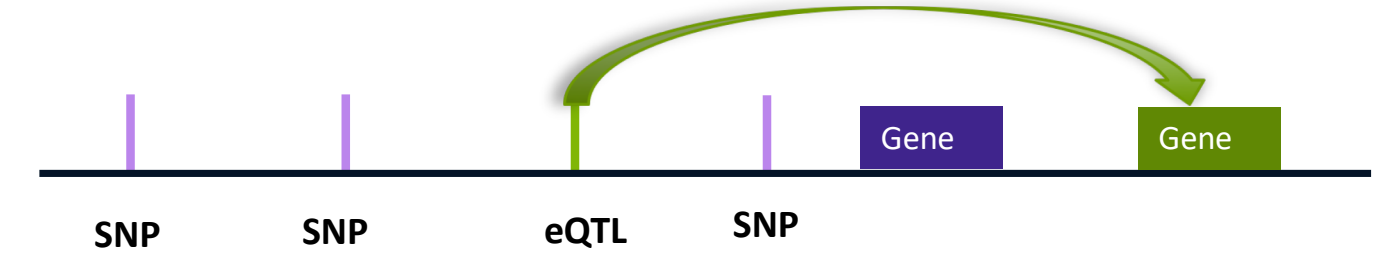
Genotype Tissue Expression Project

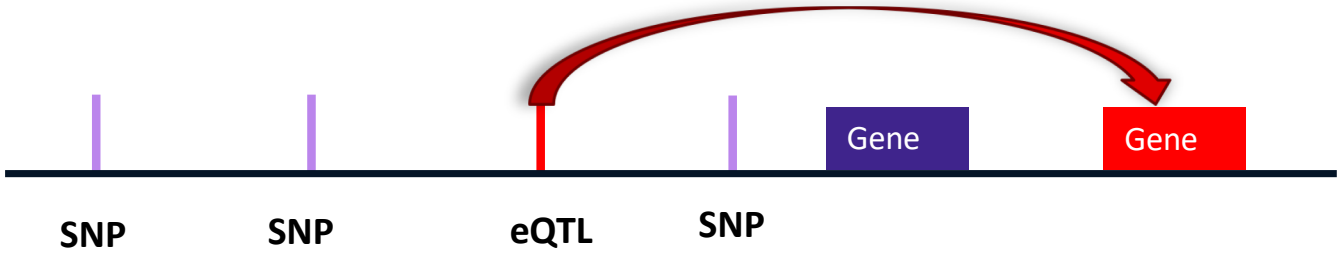


https://commonfund.nih.gov/gtex

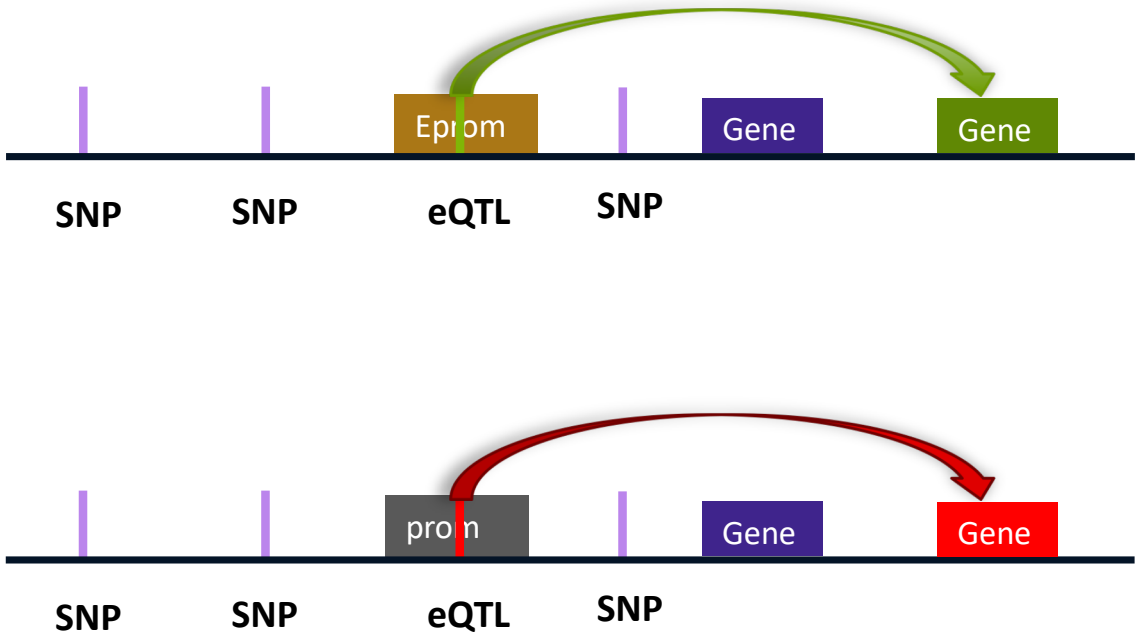
eQTLs can have negative or positive correlation to gene expression

- Beta value is a measure of the effect the variant has on gene expression.
- Beta values is bound between -2 and 2.



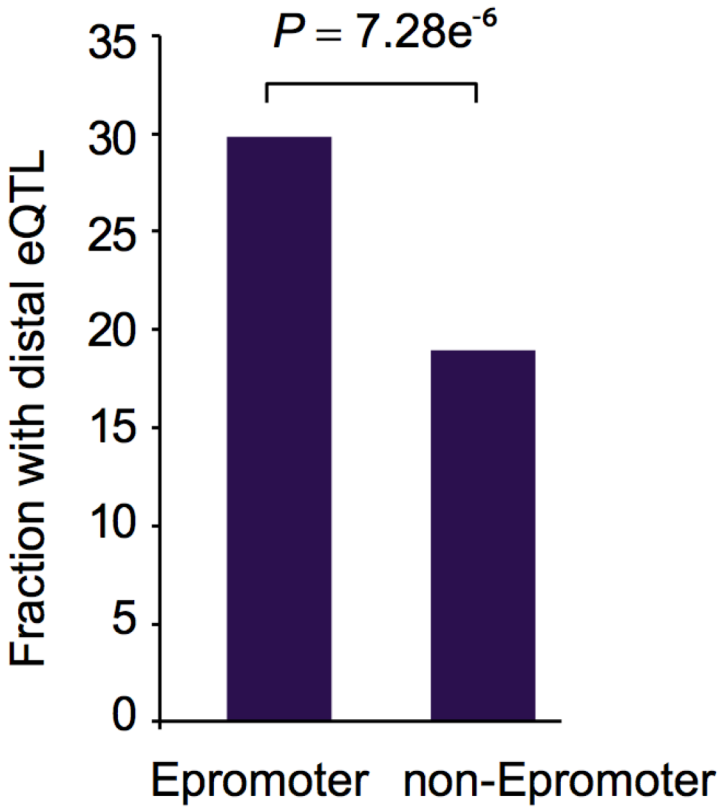


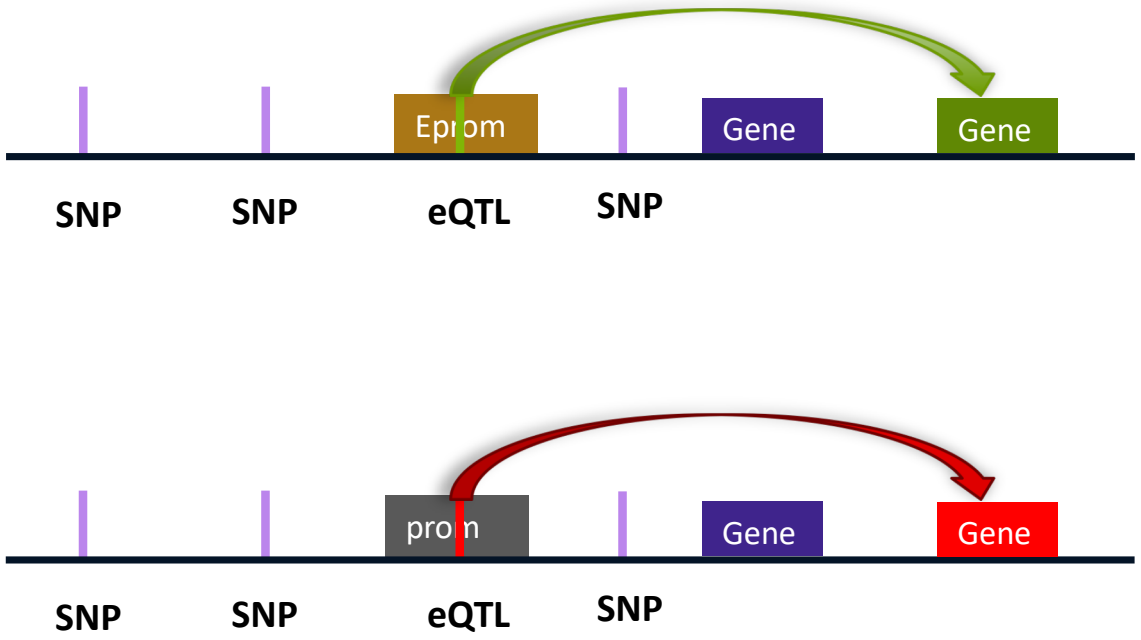
Are eQTLs overlapping Epromoters?



Lan Dao, et al, Nature Genetics. 2017

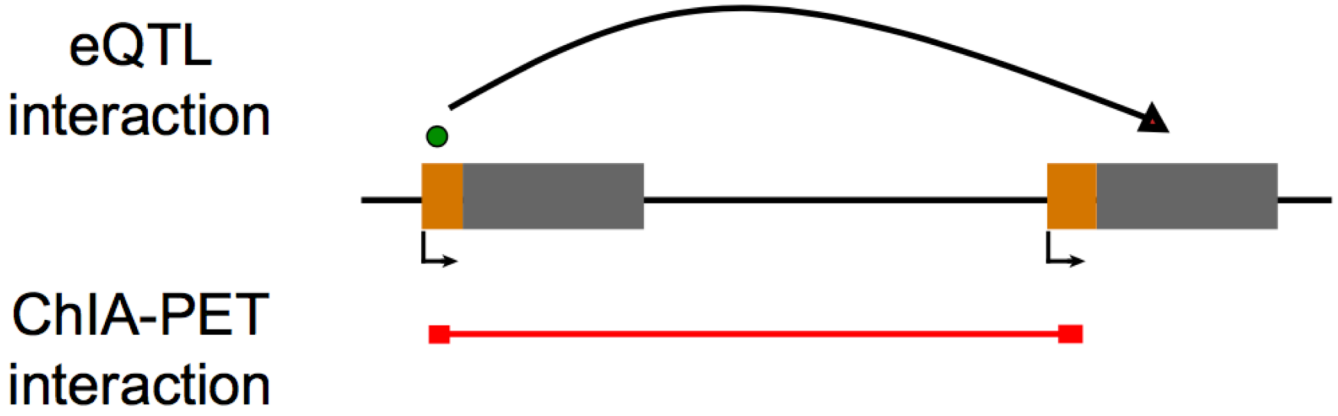
It is more likely to find an eQTL in an Epromoter than in a non-Epromoter





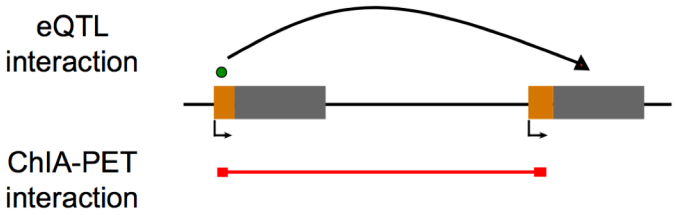
Lan Dao, et al, Nature Genetics. 2017

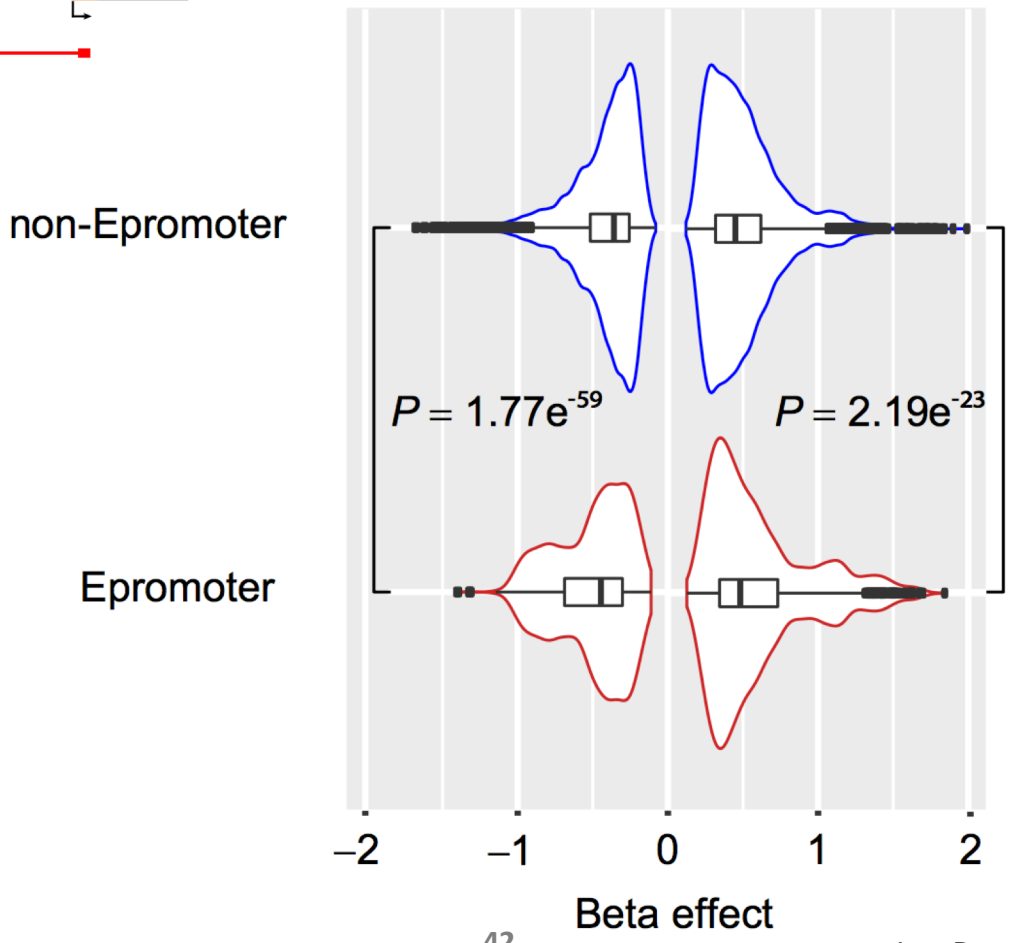
What is the effect of Epromoters on long range regulated genes?



Lan Dao, et al, Nature Genetics. 2017

Epromoter eQTLs have a significantly stronger effect on distal gene expression



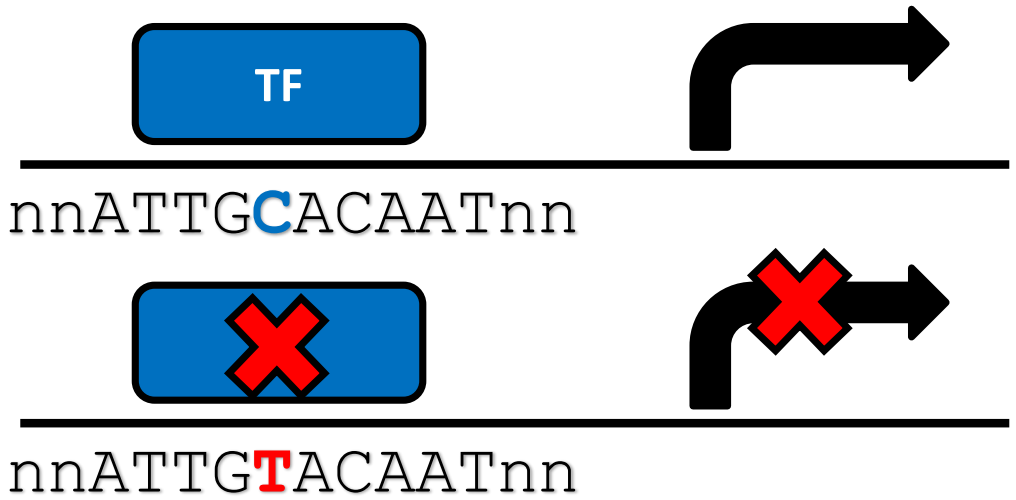


Lan Dao, et al, Nature Genetics. 2017

What is the effect of eQTLs that disrupt TF binding?

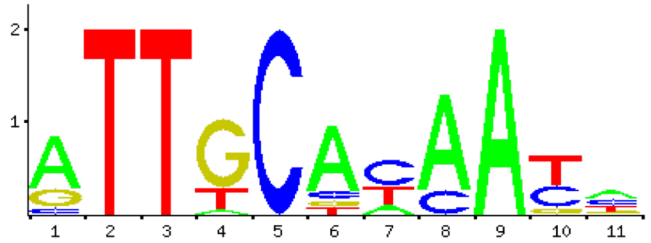
What is the effect of eQTLs that disrupt TF binding?

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms

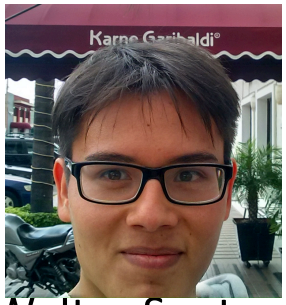


Detecting eQTL effect on TF binding

- *variation-scan:* PSSM based algorithm as part of the RSAT suite
- Flexible tool that can be used to scan any variable with any motif.
- Multiple organisms are available.



nnATTGCACAATnn
nnATTGTACAATnn



Walter Santana

RSAT

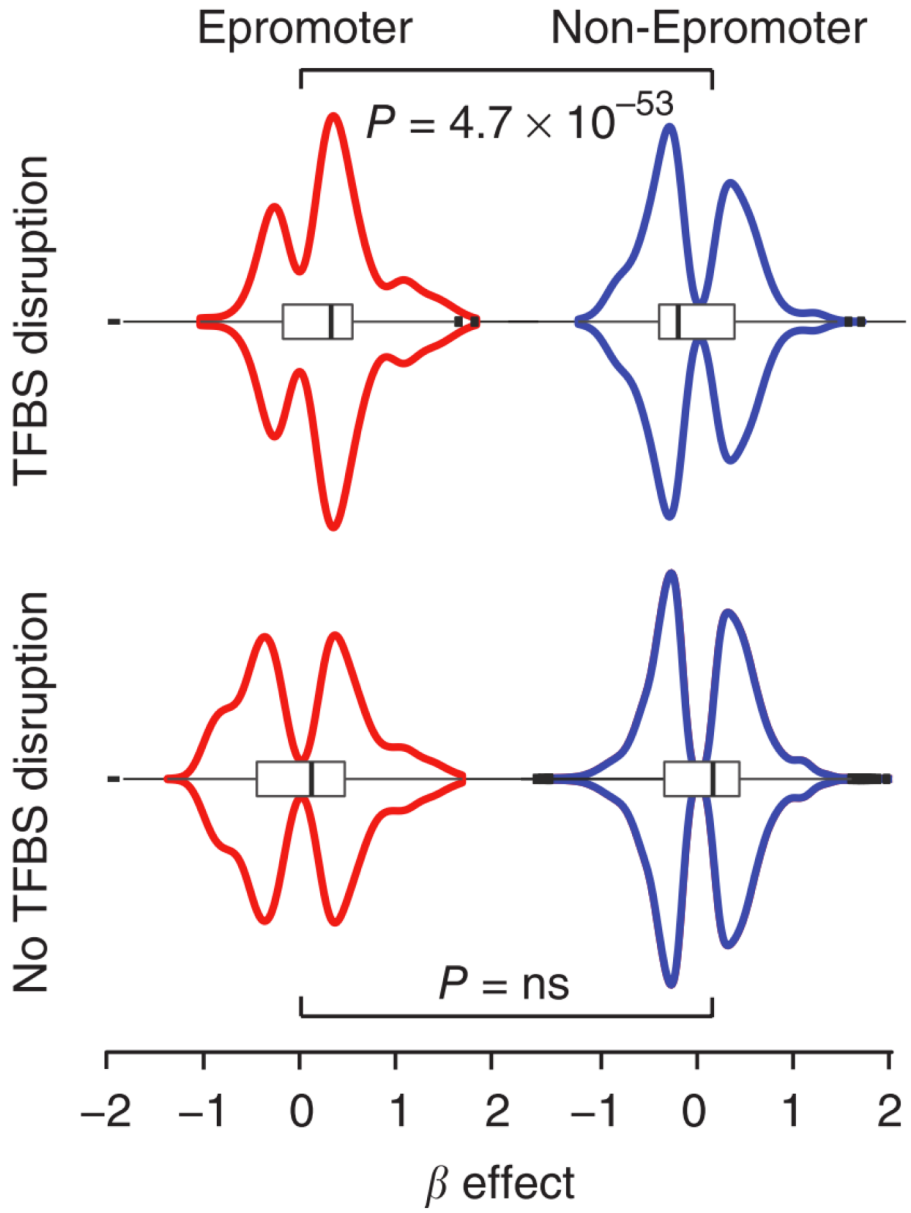
Weight Score

REF: 8.1

MUT: 2.1

DIF: 6.0

eQTL disrupting TF binding tend to have a positive beta-value



Summary

- High-throughput techniques can be used to identify long range regulatory sequences.
- Regulatory sequences can have several functions (Epromoters).
- Variation lying within Epromoters might impact distal gene expression beyond their cognate target gene.
- eQTLs that might disrupt TF binding tend to have a positive beta value in relation to their linked genes.

Future Steps

- Identify ePromoters for additional cell lines: Jukart, CEM, RPMI, GM12878
- Identify ePromoters under inflammatory conditions (IFN- alpha) in K562.
- Leverage public STARR-seq data in HELA with IFN repressors and hESC to recover additional ePromoters.
- Further develop position-scan to become user friendly.
- Establish a link between ePromoter regulated genes and disease (collaboration with Nancy J Cox's laboratory).
- Identify regulatory variants laying in ePromoters related to diseases that might be disrupting TF binding sites.

Acknowledgments:

Regulatory Genomics Lab

Karen Nuñez

Ana Villaseñor

Walter Santana

Monica Padilla

Lucia Ramirez

Diego Terol

Marisol Alvarez

INSERM

Salvatore Spicuglia

Lan Dao

Ariel Galindo

Jacques van Helden

Jaime Castro

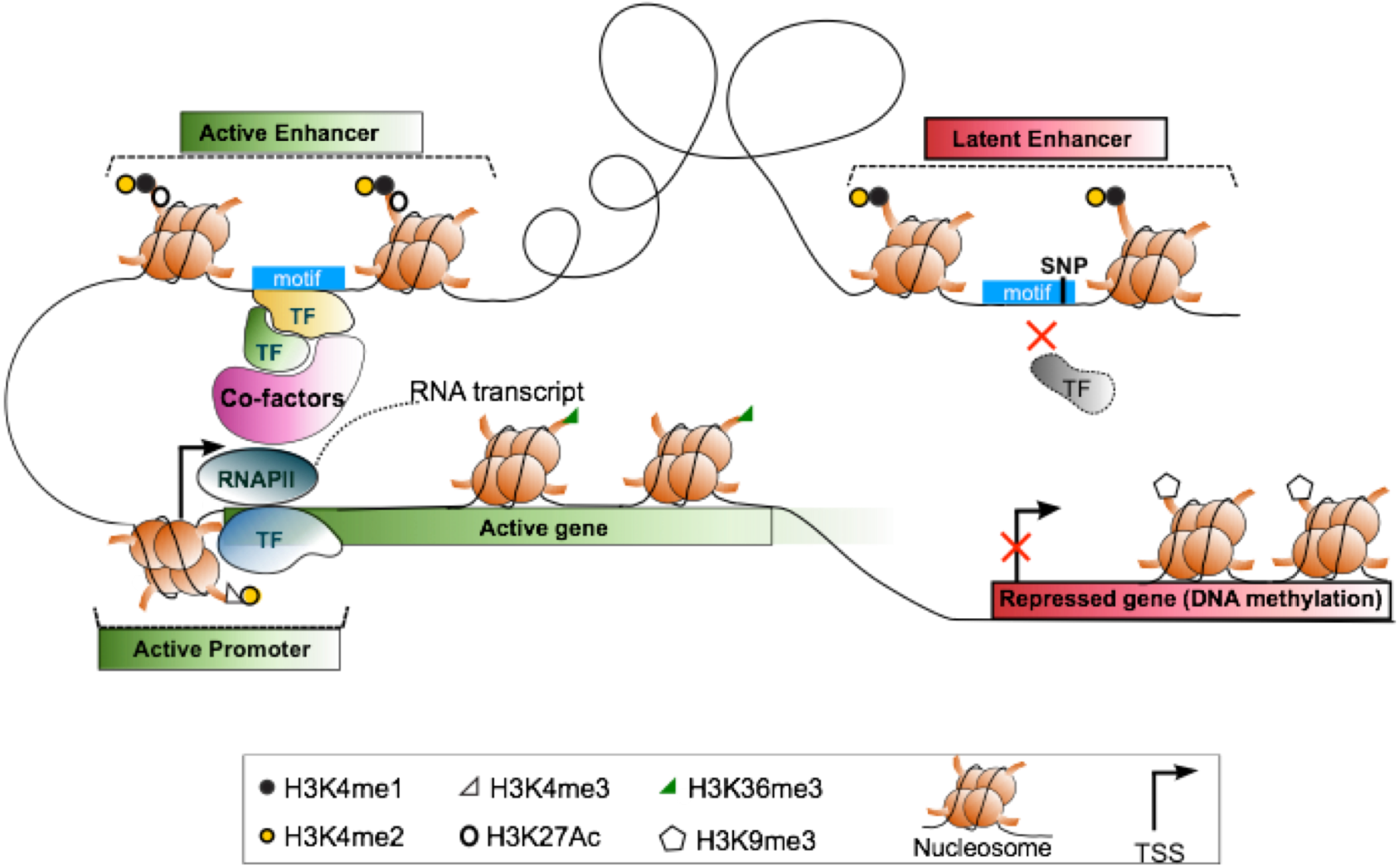
CONACYT

LIIGH UNAM

RSAT

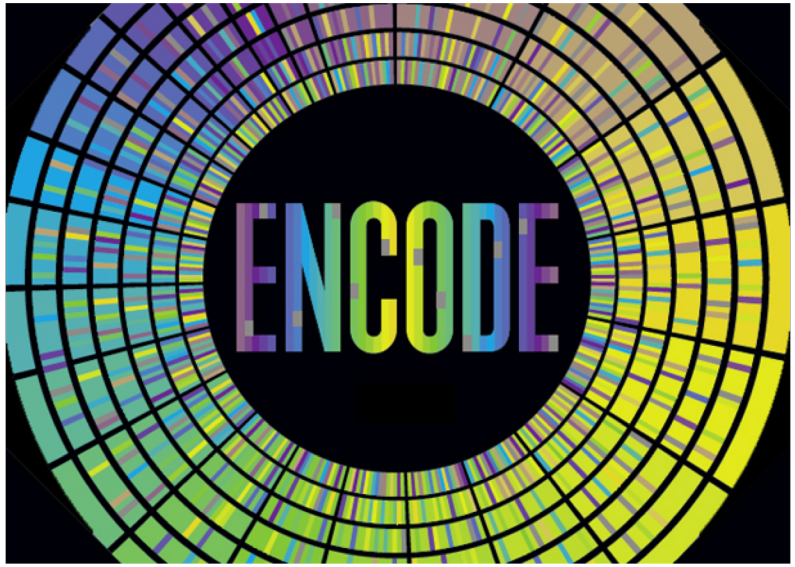
UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

Challenging biological concepts



The Non-Coding Genome

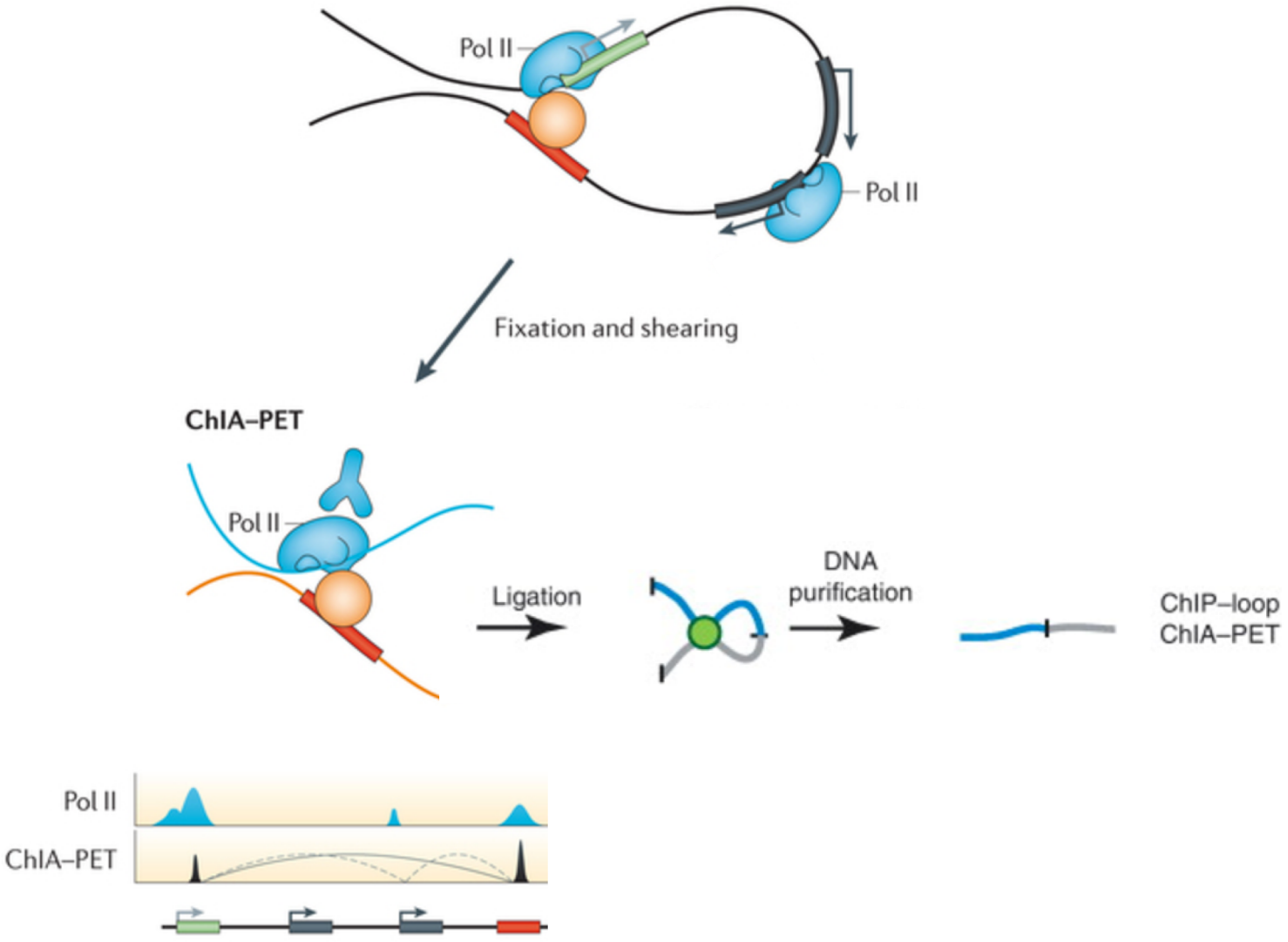
- Around 80% of the genome has shown to have some biochemical interaction.
- Most non-coding region have a regulatory function.



Consortium EP et al. *Nature*. 489(7414):57. (2012)

How is regulation being affected by genetic variants?

Promoter interactions detected with Pol II ChIA-PET



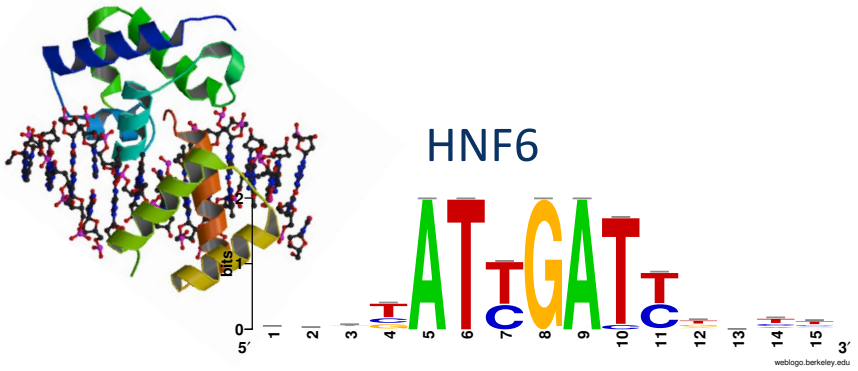
Shlyueva, et al, Nature Reviews Genetics, 2014

Major challenges to connect the regulatory genome to human pathologies

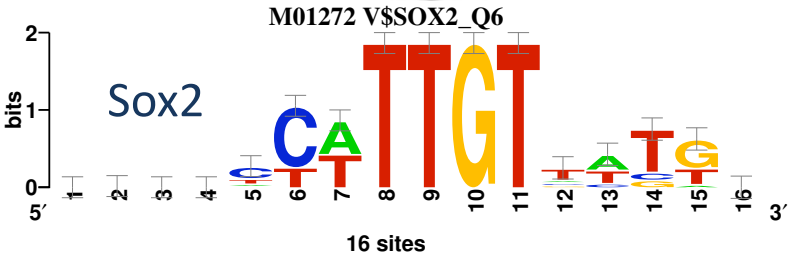
- Detect long-range regulatory regions (enhancers) and connect them to the regulated genes.
- Evaluate the effect genetic variants within regulatory sequences have on gene expression.

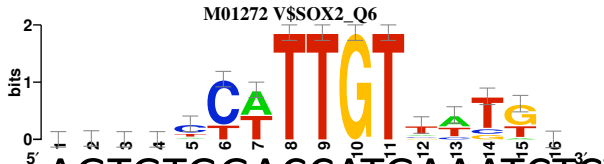
Transcription factors control gene regulation by binding to specific DNA sequences

- Transcription Factors interact with DNA binding to sequence specific sites.



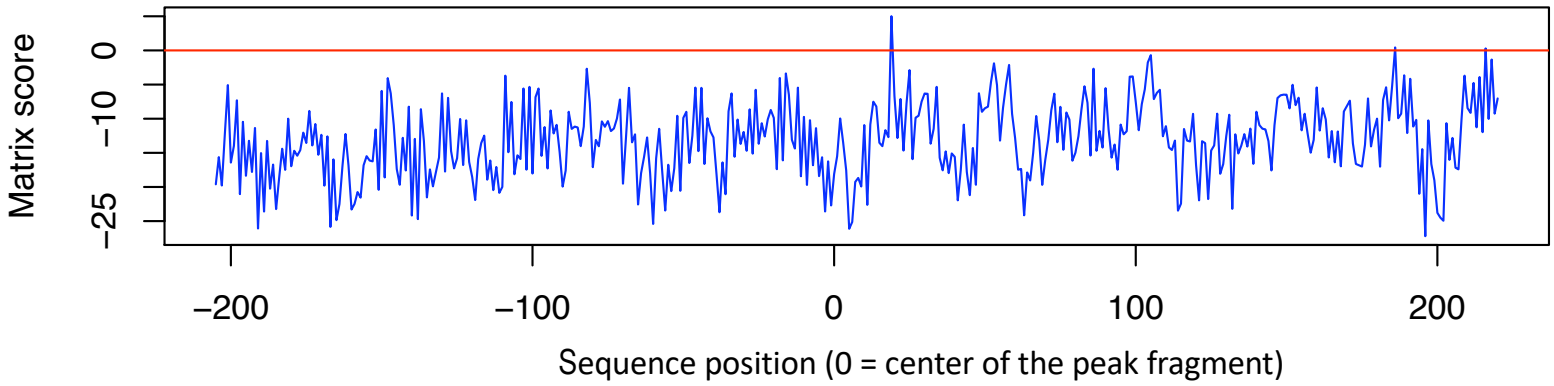
Scanning a peak sequence with a TF binding motif





...AGTGTGGACCATGAAATCTCTAACAGGTTCTGATACTGAGGGA...

Scanning

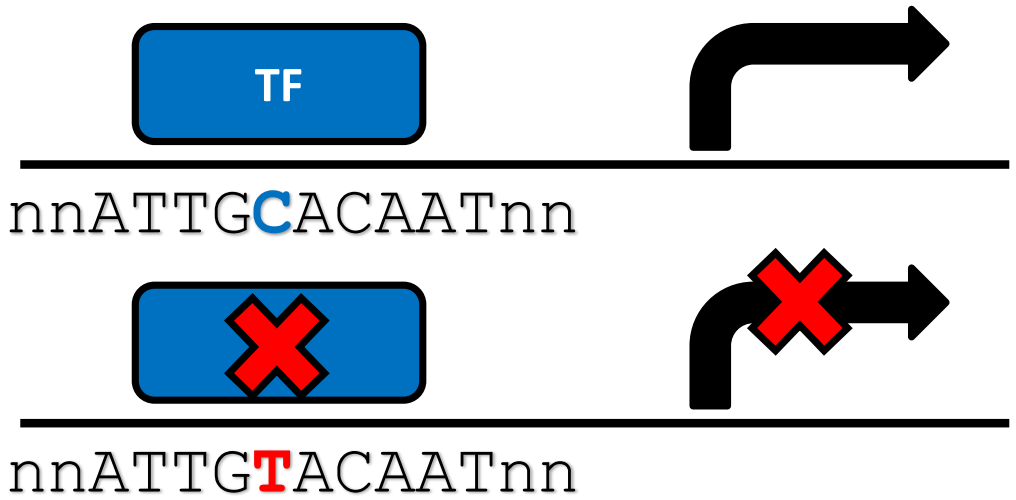


Detecting protein disruption of binding at base pair resolution

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms

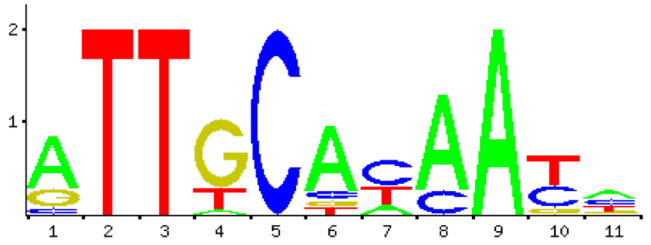
Detecting protein disruption of binding at base pair resolution

- Genetic variants in gene regulatory regions are known to cause many diseases
- Can be informative of transcription factors and regulatory mechanisms



Detecting protein disruption of binding at base pair resolution*: variation-scan*

- *variation-scan:* PSSM based algorithm as part of the RSAT suite
  - Flexibility with species, SNPs, motifs
  - High-throughput



nnATTGCACAATnn
nnATTGTACAATnn

Weight Score

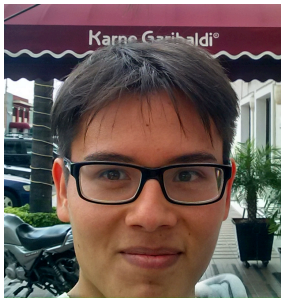
REF: 8.1

MUT: 2.1

DIF: 6.0

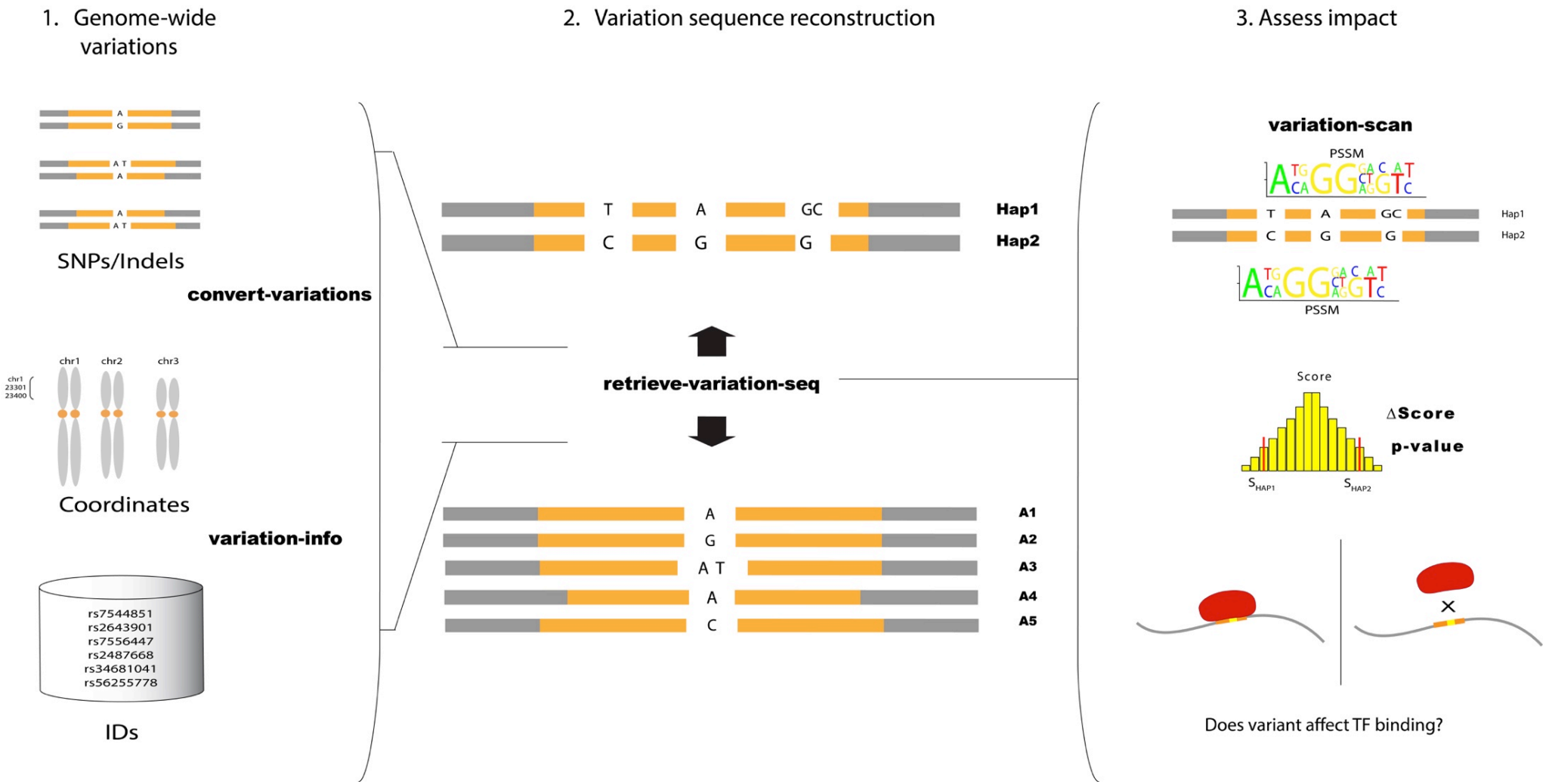
RSAT

Jacques van Helden

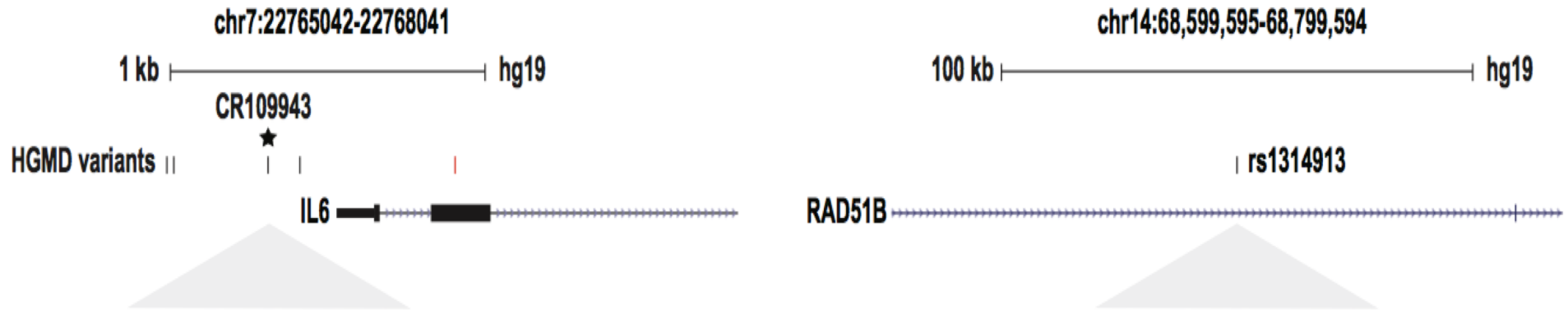


Walter Santana

Detecting protein disruption of binding at base pair resolution



Regulatory mutations alter JUN binding motifs in promoter region of *IL6* and intronic region of *RAD51B*.



Reference SNP 5'GTGCTGAGTC[A]CTAATAAAAG 3'
Alternative SNP 5'GTGCTGAGTC[C]CTAATAAAAG 3'

p-value fold change = 67.44

Reference SNP 5' GTTCTGACT[C]ATTCAGGCCT 3'
Alternative SNP 5' GTTCTGACT[T]ATTCAGGCCT 3'

p-value fold change = 36.36

Michael Wilson

DNA mutations in protein coding genes can explain many human diseases

Mutation Type	HGMD entries
Missense/nonsense	82176
Small deletions	22610
Splicing	13641
Gross deletions	10968
Small insertions	9423
Regulatory	2884
Gross insertions/duplications	2600
Small indels	2173
Complex rearrangements	1504
Repeat variations	434
Mutation total	**148413**
Gene total	**6137**

DNA mutations in protein coding genes can explain many human diseases

Mutation Type	HGMD entries
Missense/nonsense	82176
Small deletions	22610
Splicing	13641
Gross deletions	10968
Small insertions	9423
Regulatory	2884
Gross insertions/duplications	2600
Small indels	2173
Complex rearrangements	1504
Repeat variations	434
Mutation total	**148413**
Gene total	**6137**

DNA mutations in protein coding genes can explain many human diseases

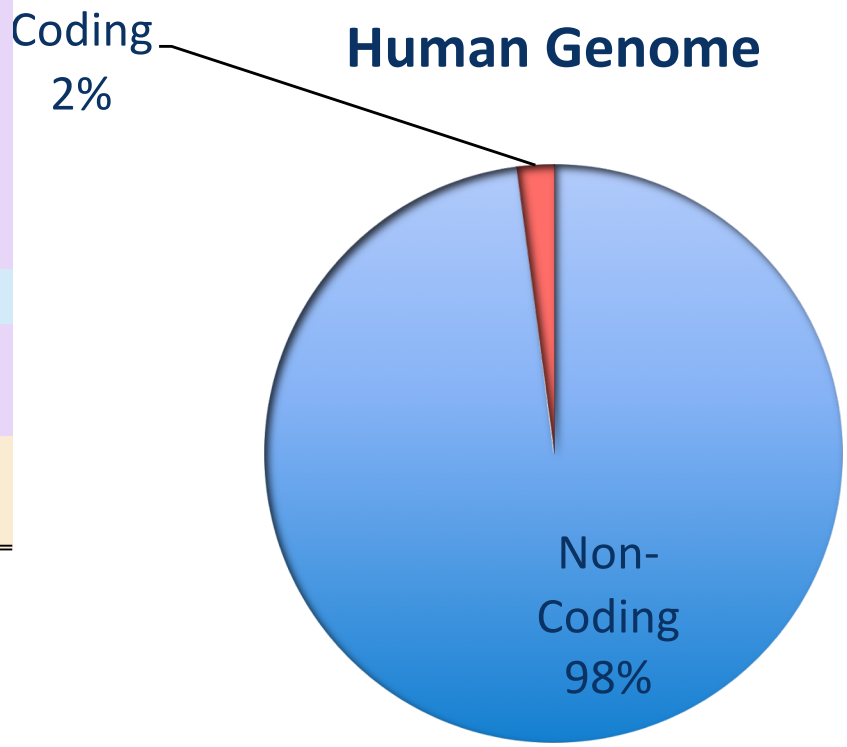
Mutation Type	HGMD entries
Missense/nonsense	82176
Small deletions	22610
Splicing	13641
Gross deletions	10968
Small insertions	9423
Regulatory	2884
Gross insertions/duplications	2600
Small indels	2173
Complex rearrangements	1504
Repeat variations	434
Mutation total	**148413**
Gene total	**6137**

DNA mutations in protein coding genes can explain many human diseases

Mutation Type	HGMD entries
Missense/nonsense	82176
Small deletions	22610
Splicing	13641
Gross deletions	10968
Small insertions	9423
Regulatory	2884
Gross insertions/duplications	2600
Small indels	2173
Complex rearrangements	1504
Repeat variations	434
Mutation total	**148413**
Gene total	**6137**

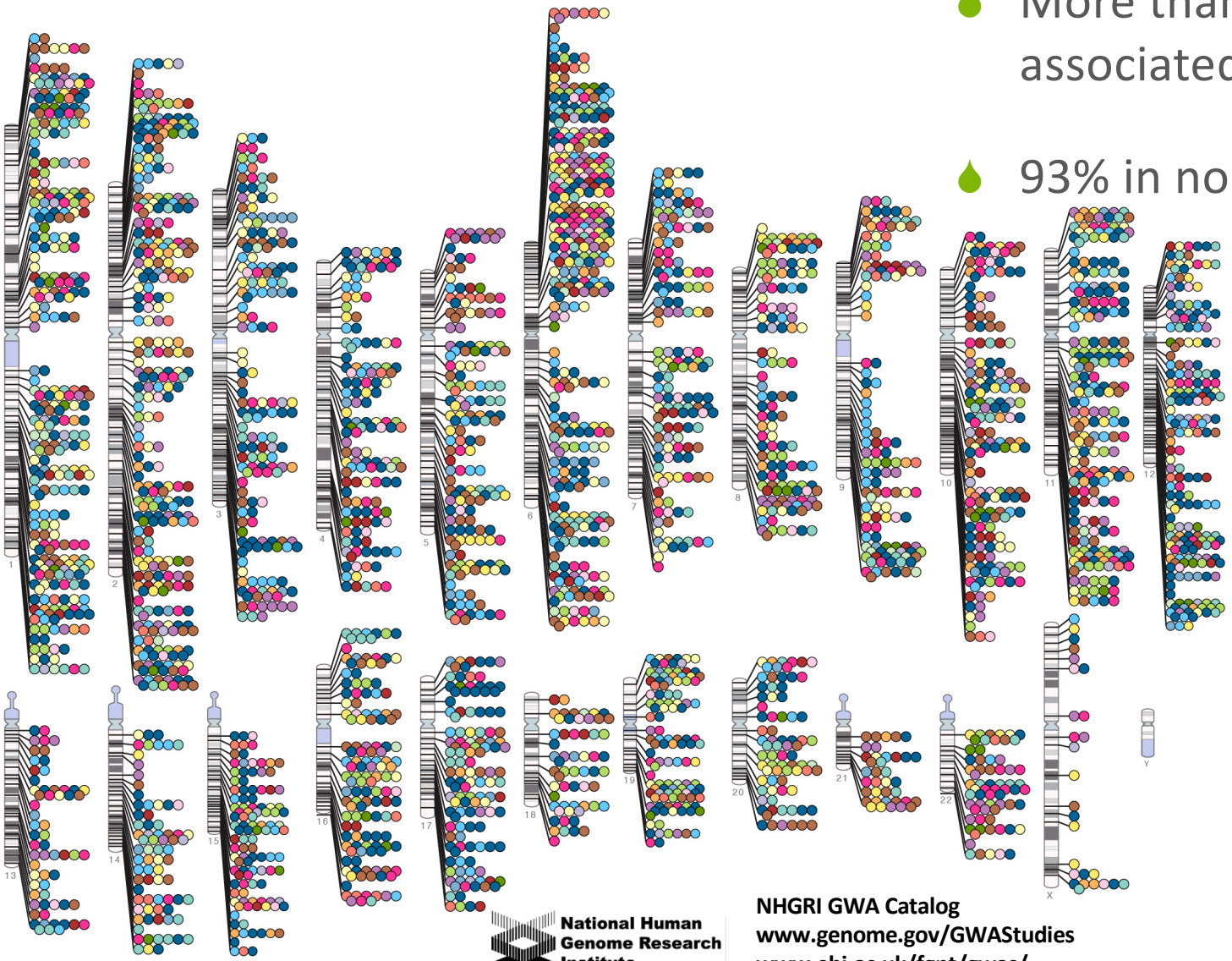
DNA mutations in protein coding genes can explain many human diseases

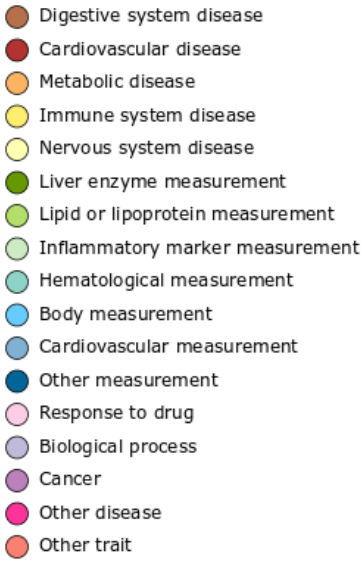
Mutation Type	HGMD entries
Missense/nonsense	82176
Small deletions	22610
Splicing	13641
Gross deletions	10968
Small insertions	9423
Regulatory	2884
Gross insertions/duplications	2600
Small indels	2173
Complex rearrangements	1504
Repeat variations	434
Mutation total	**148413**
Gene total	**6137**



Published Genome-Wide Associations since 12/2012
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

- More than 17,000 reported associated SNPs
- 93% in non-coding sequence





National Human Genome Research Institute

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

EMBL-EBI

Weight score difference has proven not to be enough to reliably detect variations affecting TF binding in a genome wide scale

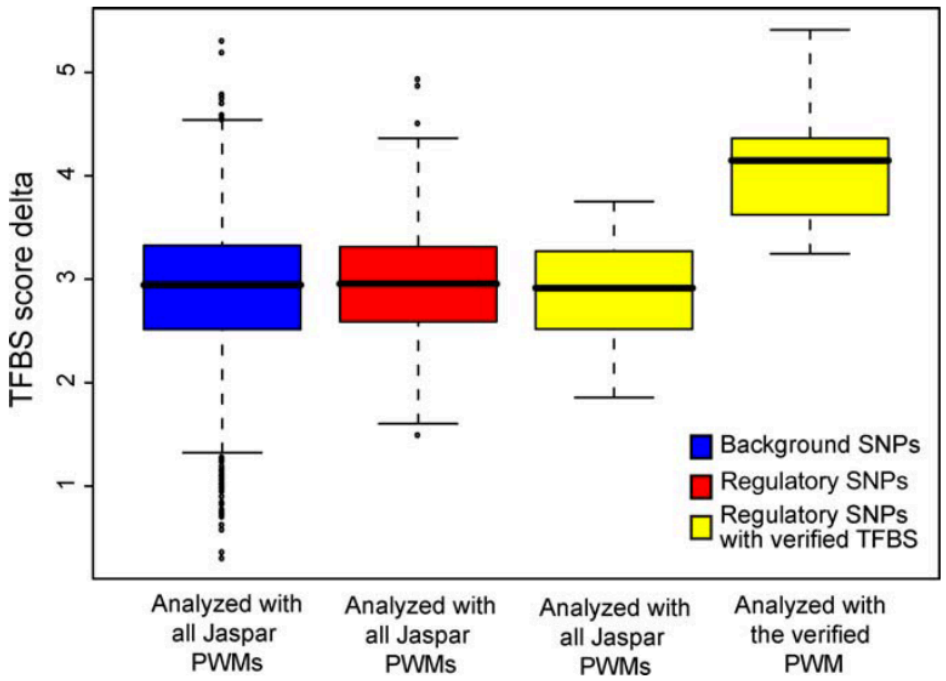
Weight score difference has proven not to be enough to reliably detect variations affecting TF binding in a genome wide scale

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

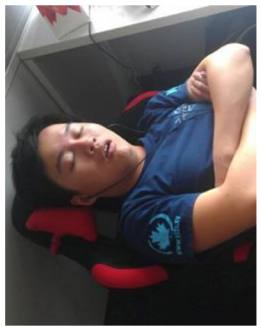
In Silico Detection of Sequence Variations Modifying Transcriptional Regulation

Malin C. Andersen[1,2], Pär G. Engström[3,4], Stuart Lithwick[5,6], David Arenillas[6], Per Eriksson[2], Boris Lenhard[3], Wyeth W. Wasserman[6*], Jacob Odeberg[1,2,7*]

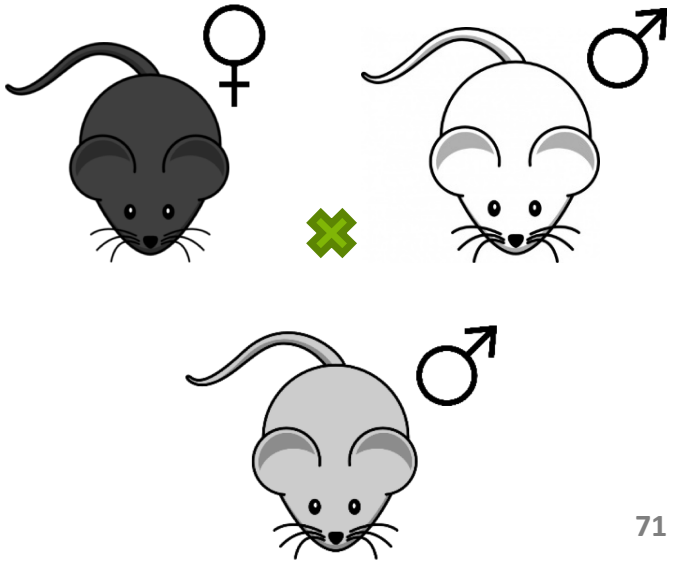


Detecting protein disruption of binding at base pair resolution: *Heterozygous ChIP-seq*

- F1 mice with inbred parents
- All SNP positions known – reduce false positives and search space
- Each allele compared within same organism and genomic context



Minggao Liang



BL6 (Reference) Genome

✚

5 million SNPs

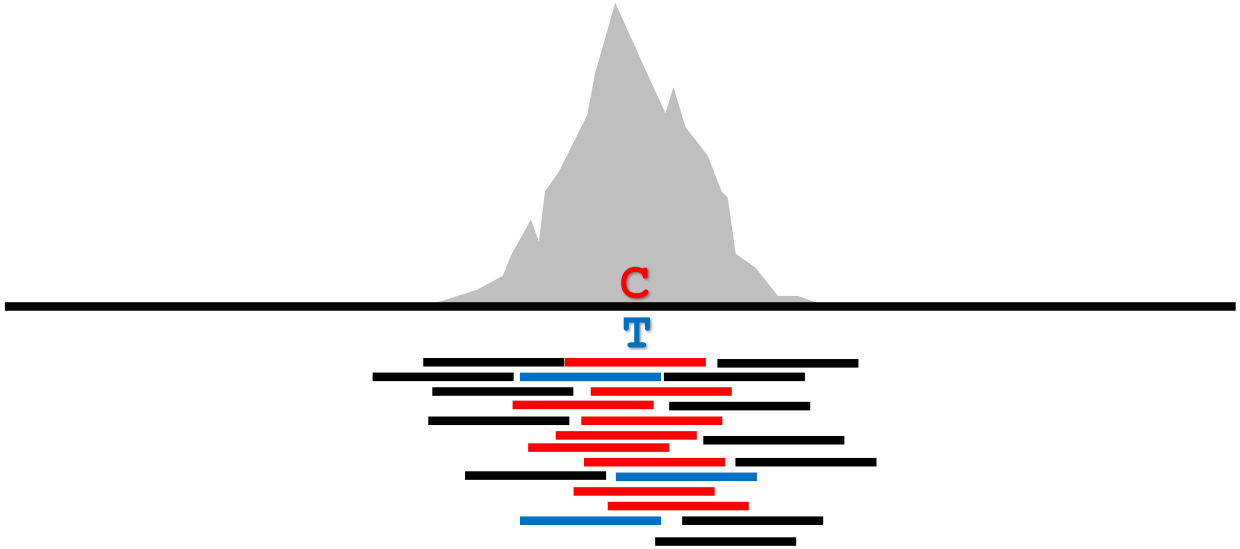
AJ Parental Genome

Detecting protein disruption of binding at base pair resolution: *Heterozygous ChIP-seq*

HNF6
CEBPA

Processing	Purpose	Remaining mapped reads
mm9 alignment		33.0 million
Filter against Encode Blacklist regions	• Remove known sequencing artifact regions	31.9 million
MACS	• Peak calling	
WASP	• Extract reads overlapping known alleles • Remove reads with alignment bias • Remove duplicates	1.68 million 1.50 million 0.97 million
ALEA	• Alignment to BL6 Parental genome (REF) • Alignment to A/J Parental genome (ALT)	0.43 million 0.47 million

Detecting protein disruption of binding at base pair resolution*: Heterozygous ChIP-seq*



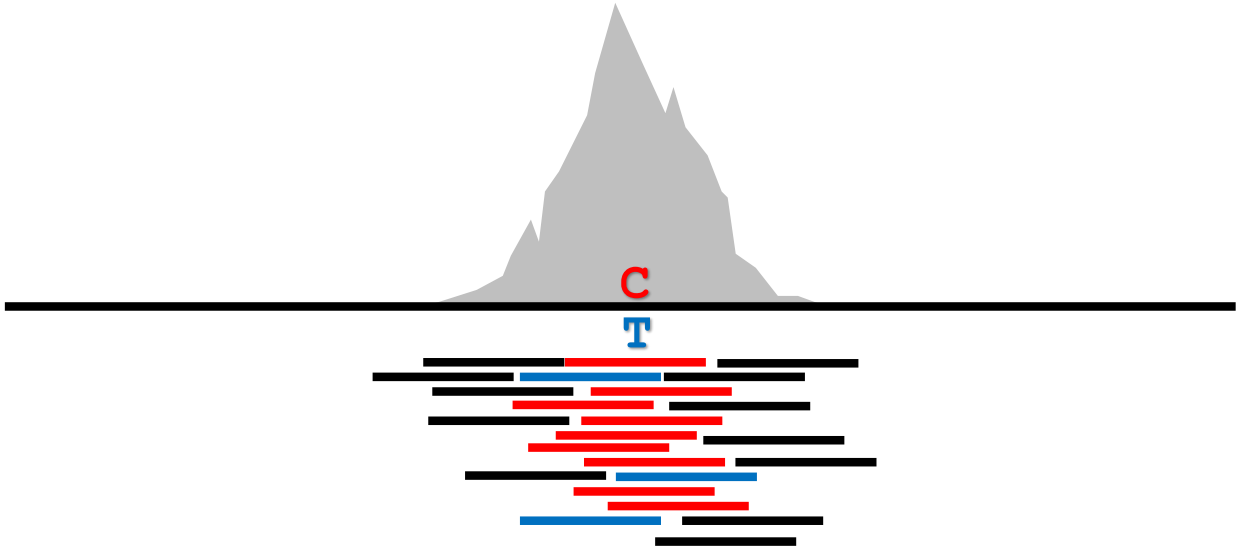
Count Reads:

REF: 9

ALT: 3

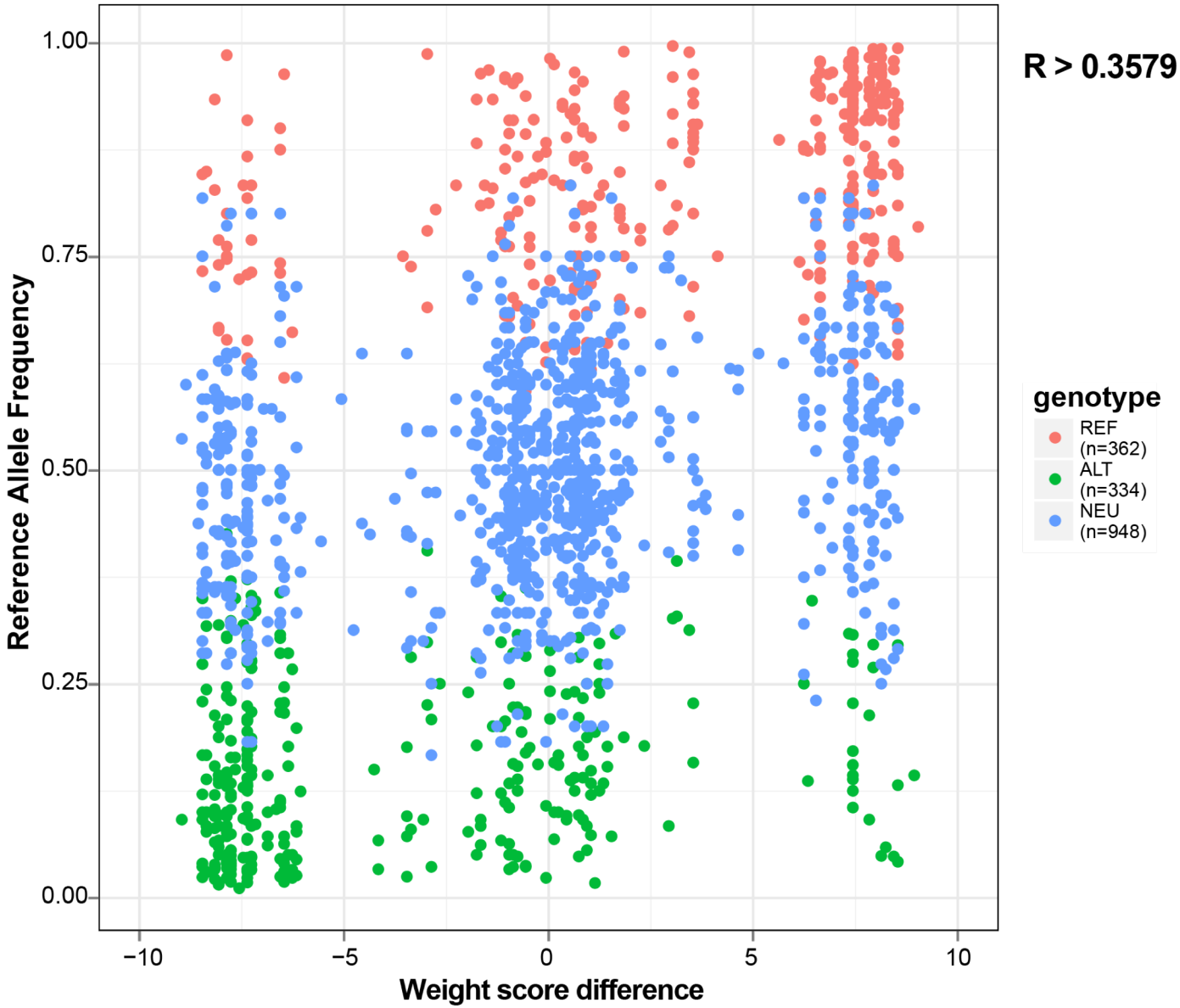
RAF: 0.75

Detecting protein disruption of binding at base pair resolution*: Heterozygous ChIP-seq*



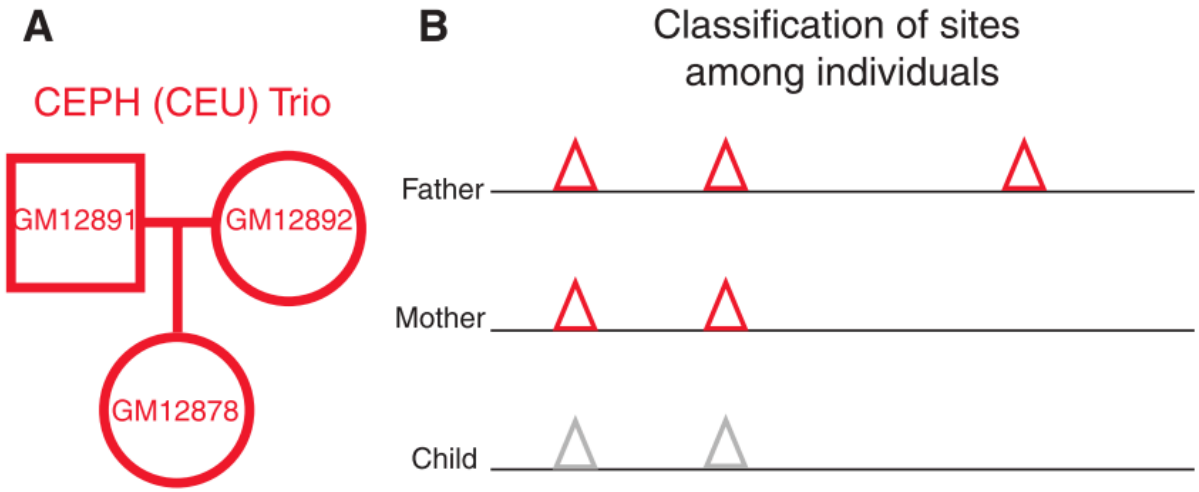
Count Reads:	Weight score:
REF: 9	REF: 7.0
ALT: 3	ALT: 2.0
RAF: 0.75	DIF: 5.0

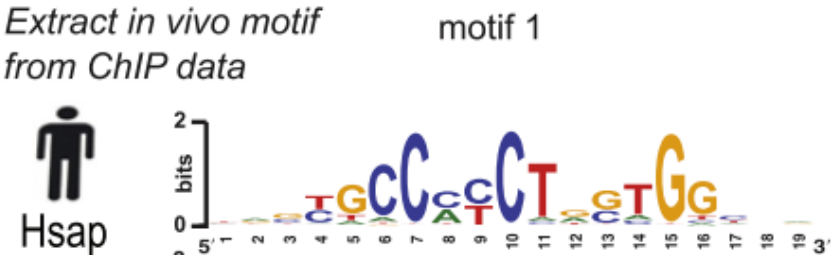
Detecting protein disruption of binding at base pair resolution*: Heterozygous ChIP-seq*

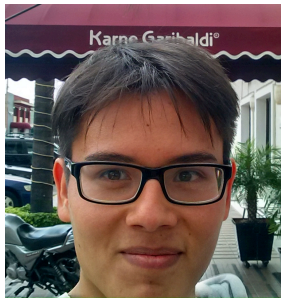


Detecting protein disruption of binding at base pair resolution*: Heterozygous ChIP-seq*

- Human trio families with genotypes





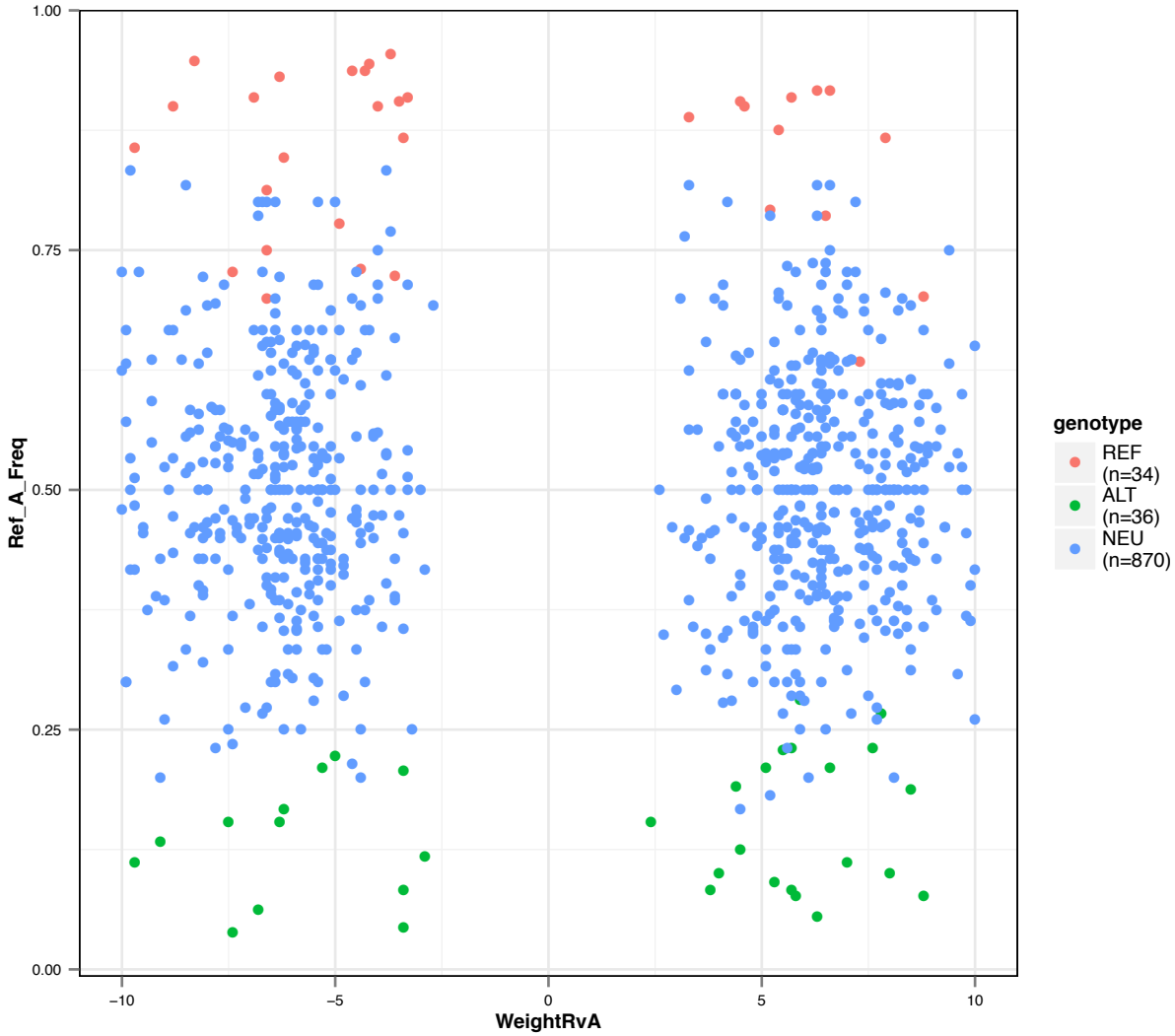


Walter Santana

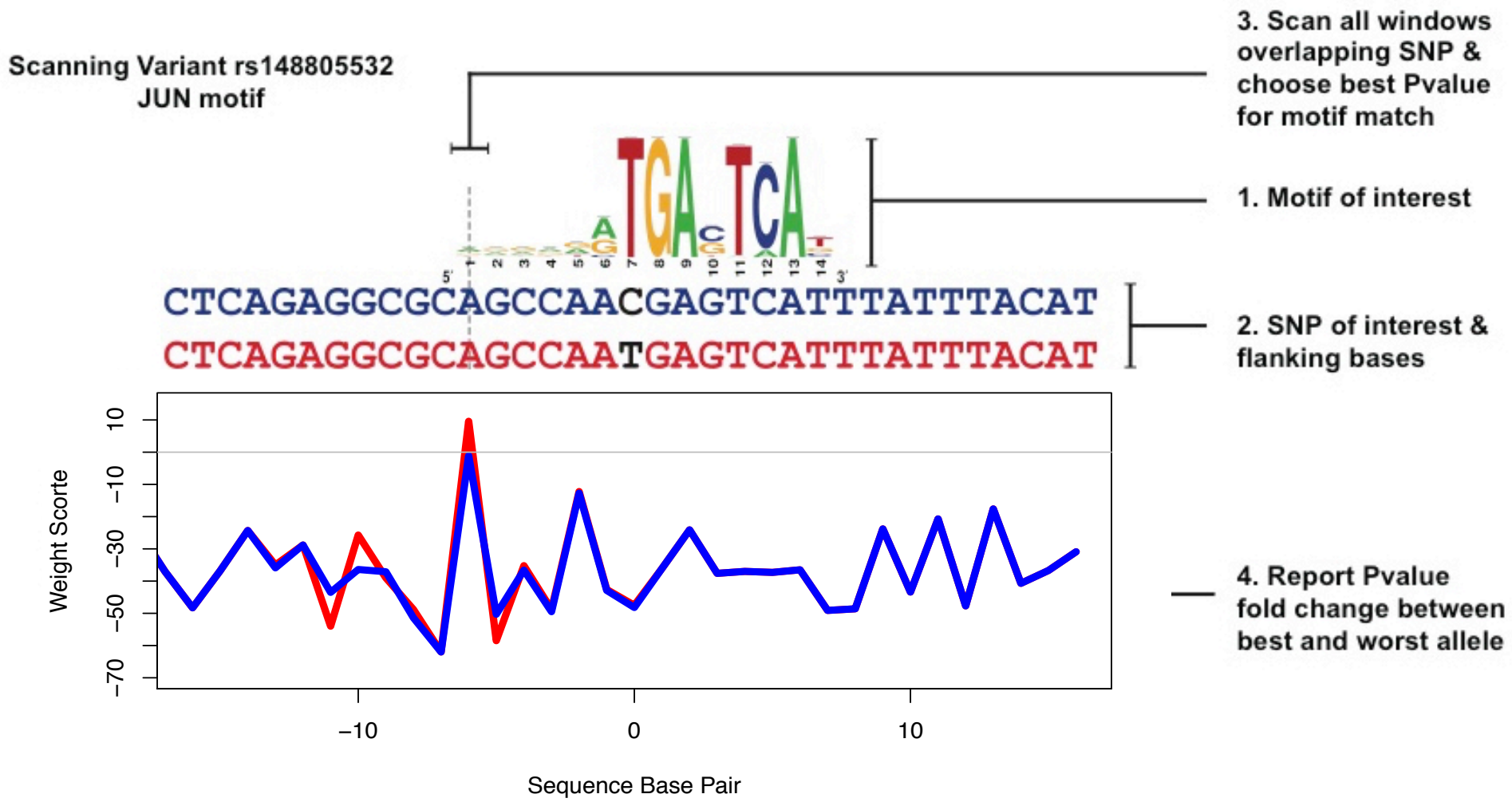
McDaniell, R.et al (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science (New York, N.Y.),

Detecting protein disruption of binding at base pair resolution*: Heterozygous ChIP-seq*

- Human trio families with genotypes



Detecting protein disruption of binding at base pair resolution*: variation-scan*



Transcription Factor Sox2

Sox2
-Regulator in neuronal development.
-Required for both the maintenance of neural stem cells and the differentiation of specific neuron sub-types

16 Annotated binding sites in Transfac

1	g	c	c	c	t	c	a	t	t	g	t	t	a	t	g	c
2	a	a	a	c	t	c	t	t	t	g	t	t	t	g	g	a
3	t	t	c	a	c	c	a	t	t	g	t	t	c	t	a	g
4	g	a	c	t	c	t	a	t	t	g	t	c	t	c	t	g
5	g	a	t	a	t	c	t	t	t	g	t	t	t	c	t	t
6	t	g	c	a	c	c	t	t	t	g	t	t	a	t	g	c
7	a	a	t	t	c	c	a	t	t	g	t	t	a	t	g	a
8	a	a	a	c	t	c	t	t	t	g	t	t	t	g	g	a
9	a	t	g	g	a	c	a	t	t	g	t	a	a	t	g	c
10	a	g	g	c	c	t	t	t	t	g	t	c	c	t	g	g
11	t	g	t	g	c	t	t	t	t	g	t	n	n	n	n	n
12	c	t	c	a	a	c	t	t	t	g	t	a	a	t	t	t
13	g	c	a	g	c	c	a	t	t	g	t	g	a	t	g	c
14	c	a	c	c	c	c	t	t	t	g	t	t	a	t	g	c
15	t	t	t	t	c	t	a	t	t	g	t	t	t	t	t	a
16	a	a	a	g	g	c	a	t	t	g	t	g	t	t	t	c

a	6	7	4	4	2	0	8	0	0	0	0	2	7	0	1	4
c	2	2	6	5	9	12	0	0	0	0	0	2	2	2	0	6
g	4	3	2	4	1	0	0	0	0	16	0	2	0	2	9	3
t	4	4	4	3	4	4	8	16	16	0	16	9	6	11	5	2

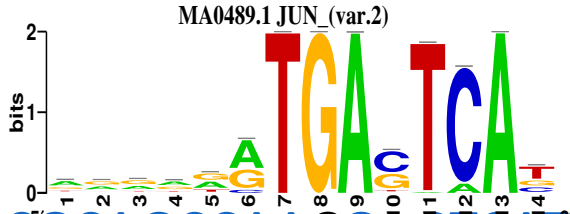
Detecting protein disruption of binding at base pair resolution: Variation Scan

CTGTCCCCCTGACTCAGAGGCGCAGCCAA**C**GAGTCATTTATTTACATGTAACTACAAGA

CTGTCCCCCTGACTCAGAGGCGCAGCCAA**T**GAGTCATTTATTTACATGTAACTACAAGA

Collaboration with Dr. van Helden, Université d'Aix-Marseille (AMU).

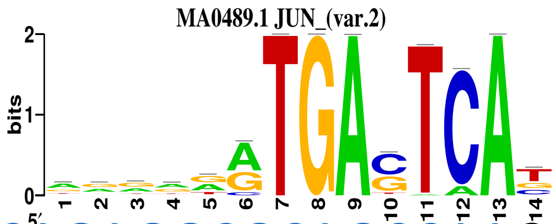
Detecting protein disruption of binding at base pair resolution: Variation Scan



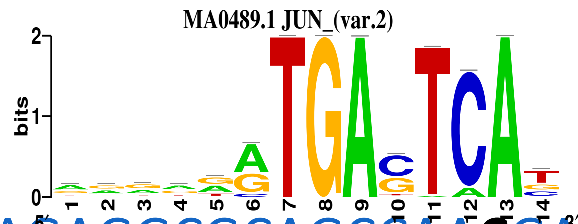
CTGTCCCCCTGACTCAGAGGCGCAGCCAACGAGTCATTTATTTACATGTAACTACAAGA

CTGTCCCCCTGACTCAGAGGCGCAGCCAATGAGTCATTTATTTACATGTAACTACAAGA

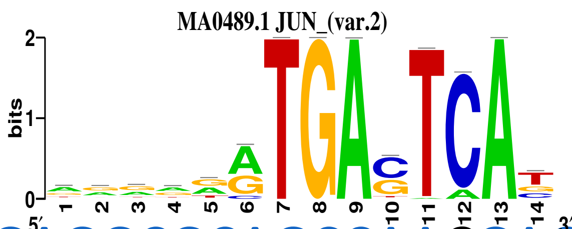
Detecting protein disruption of binding at base pair resolution: Variation Scan



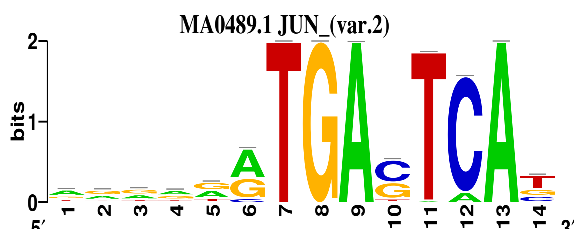
CTGTCCCCTGACTCAGAGGCGCAGCCAACGAGTCATTTATTTACATGTAACTACAAGA



CTGTCCCCTGACTCAGAGGCGCAGCCAACGAGTCATTTATTTACATGTAACTACAAGA

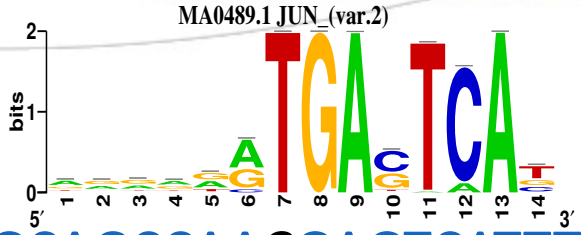


CTGTCCCCTGACTCAGAGGCGCAGCCAACGAGTCATTTATTTACATGTAACTACAAGA



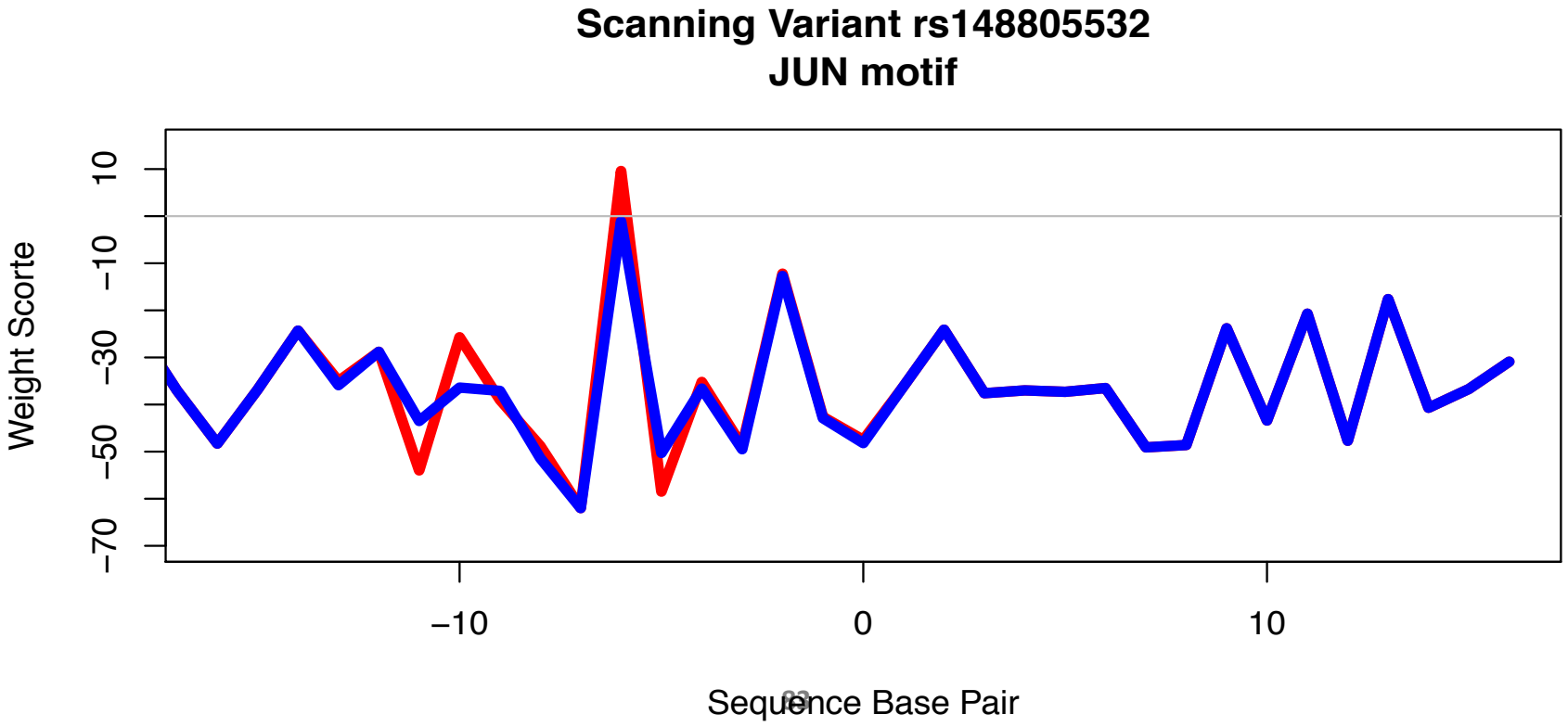
CTGTCCCCTGACTCAGAGGCGCAGCCAACGAGTCATTTATTTACATGTAACTACAAGA

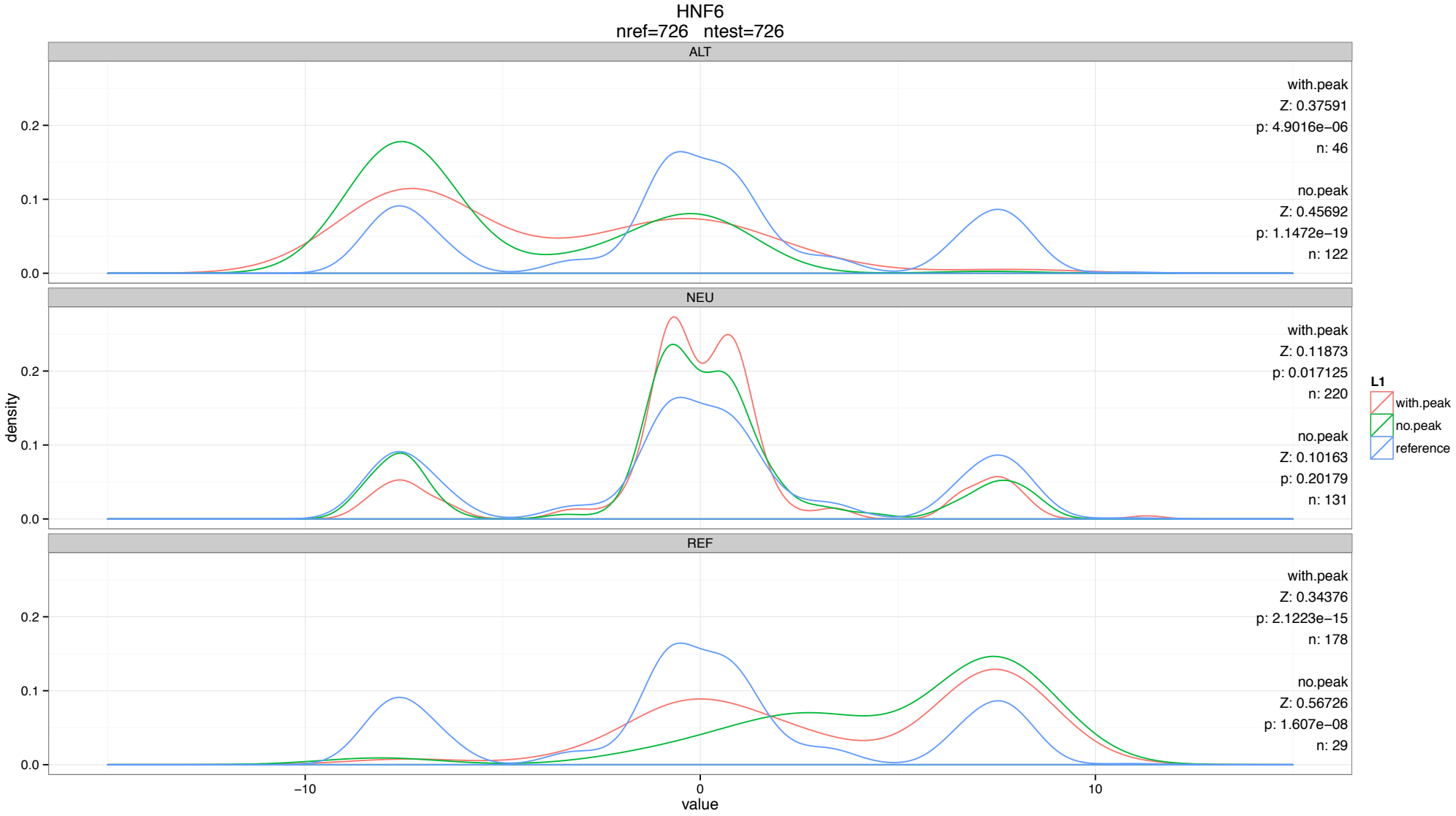
Detecting protein disruption of binding at base pair resolution: Variation Scan



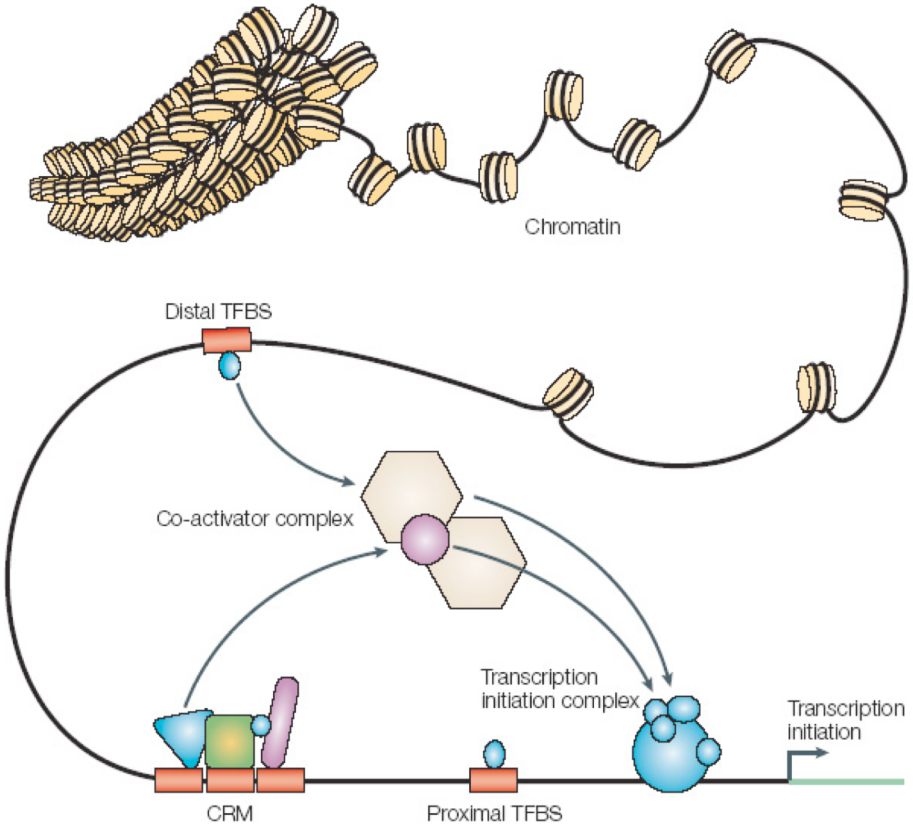
CTGTCCCCCTGACTCAGAGGCGCAGCCAA**C**GAGTCATTTATTTACATGTAACTACAAGA

CTGTCCCCCTGACTCAGAGGCGCAGCCAA**T**GAGTCATTTATTTACATGTAACTACAAGA

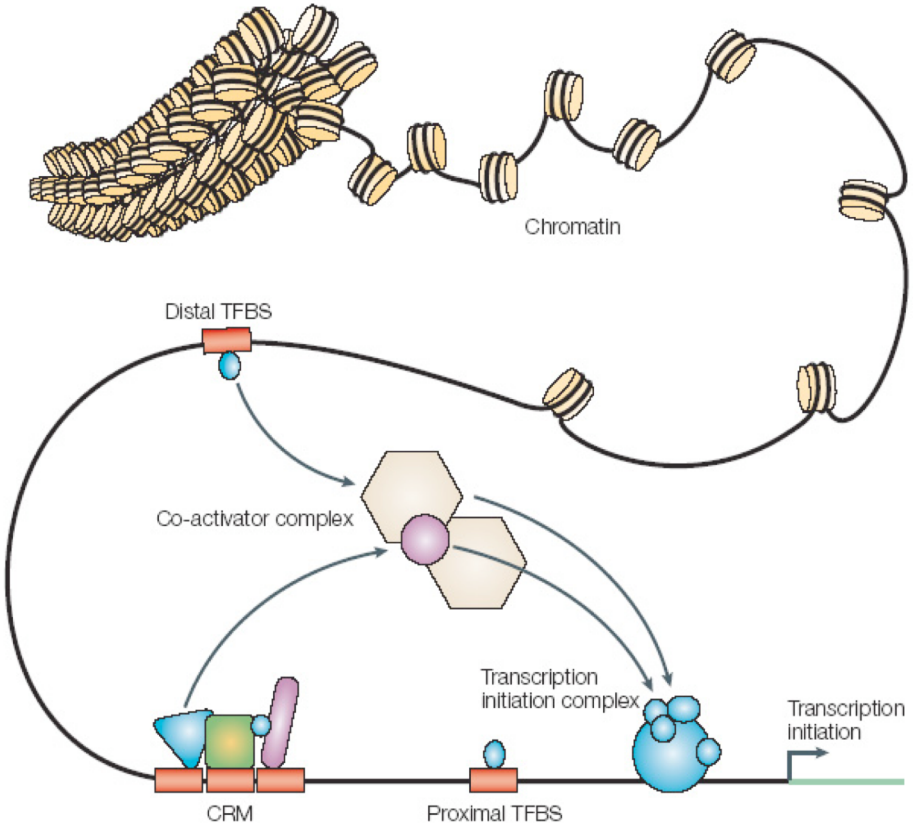



HNF6
nref=726 ntest=726
ALT
with.peak
Z: 0.37591
p: 4.9016e−06
n: 46
no.peak
Z: 0.45692
p: 1.1472e−19
n: 122
NEU
with.peak
Z: 0.11873
p: 0.017125
n: 220
no.peak
Z: 0.10163
p: 0.20179
n: 131
REF
with.peak
Z: 0.34376
p: 2.1223e−15
n: 178
no.peak
Z: 0.56726
p: 1.607e−08
n: 29
density
0.0
0.1
0.2
−10
0
10
value
L1
with.peak
no.peak
reference

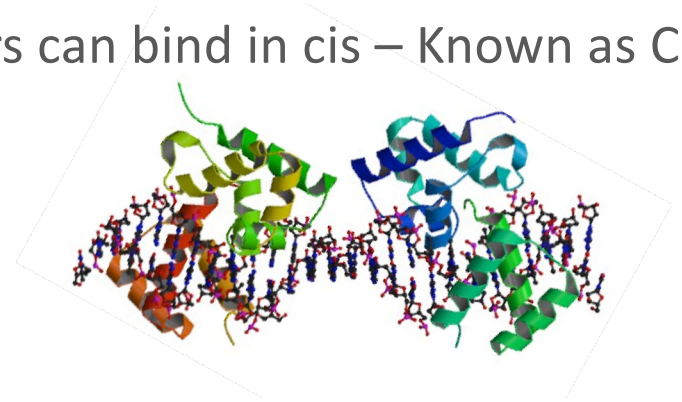
Multiple transcription factors can cooperatively regulate gene regulation



Multiple transcription factors can cooperatively regulate gene regulation

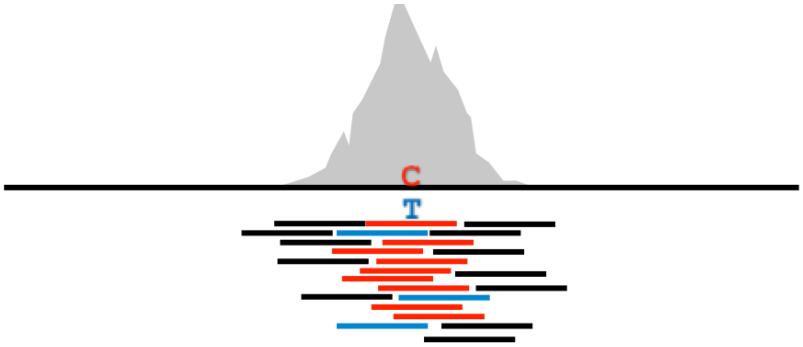


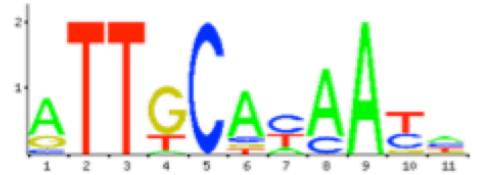
- Transcription Factors can bind in cis – Known as Cluster of Regulatory Modules (CRMs)



Can we detect when alleles disrupt cofactors?

- Why do we expect co-factors?
  - TFs in Eukaryotes tend to bind in clusters of regulatory modules.





nnATTGCACAATnn
nnATTGTACAATnn

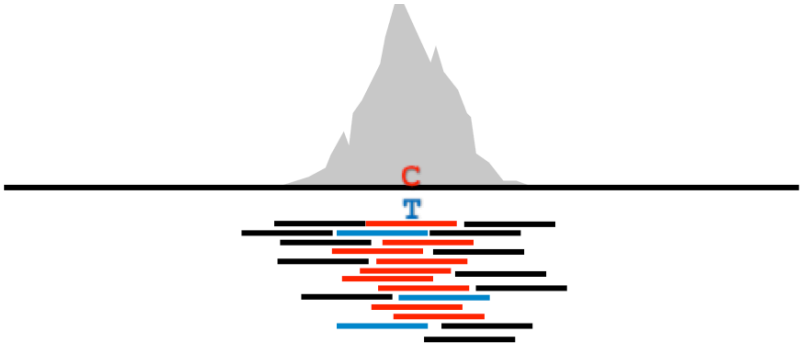
Count Reads:
REF: 9
ALT: 3
RAF: 0.75

Weight score:
REF: 7.0
ALT: 2.0
DIF: 5.0

Weight Score
REF: 8.1
MUT: 2.1
DIF: 6.0

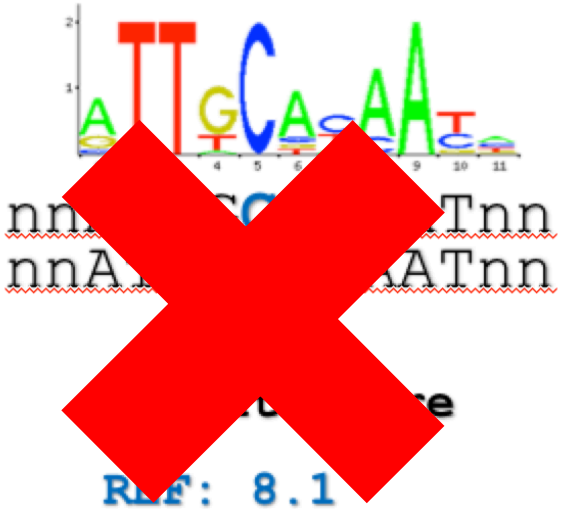
Can we detect when alleles disrupt cofactors?

- Why do we expect co-factors?
  - TFs in Eukaryotes tend to bind in clusters of regulatory modules.



Count Reads:
REF: 9
ALT: 3
RAF: 0.75

Weight score:
REF: 7.0
ALT: 2.0
DIF: 5.0

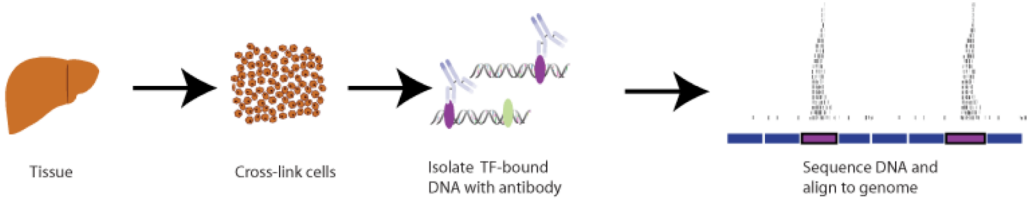


REF: 8.1

MUT: 2.1

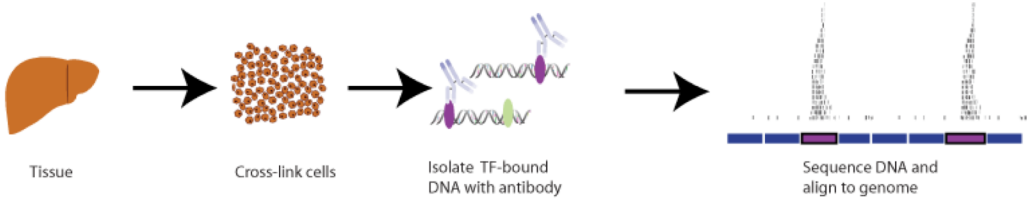
DIF: 6.0

Experimentally determine genome wide binding of 4 TFs relevant for liver from five mammalian species.



Collaboration:
Benoit Ballester (INSERM)
Duncan Odom (Cambridge U)
Paul Flicek (EBI)

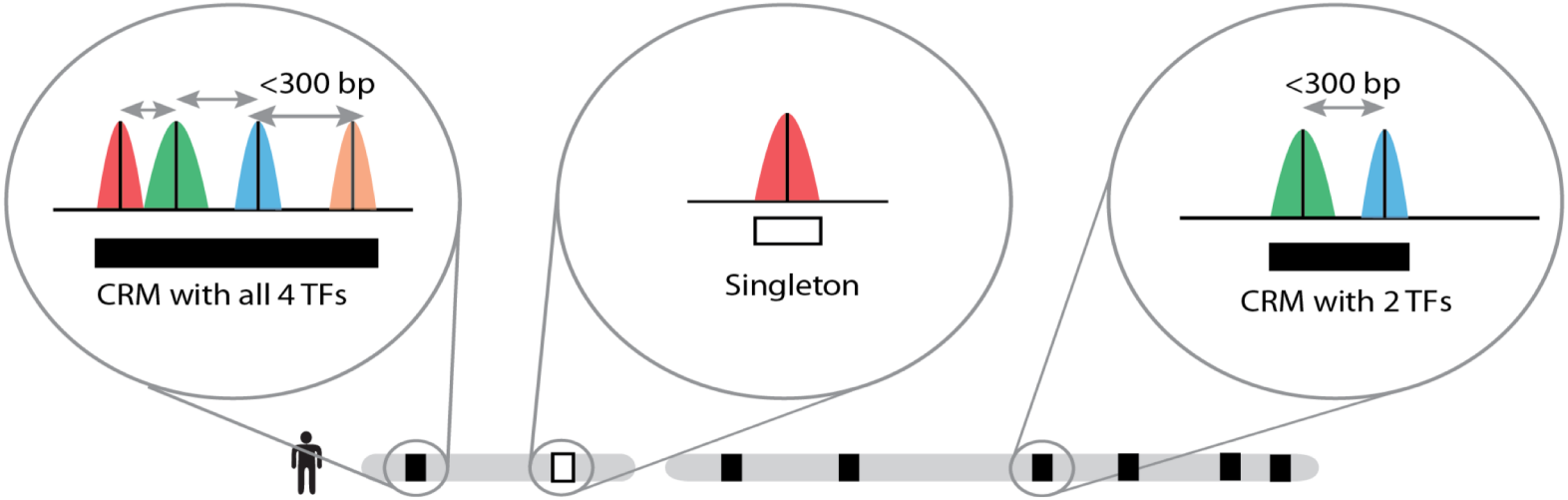
Experimentally determine genome wide binding of 4 TFs relevant for liver from five mammalian species.



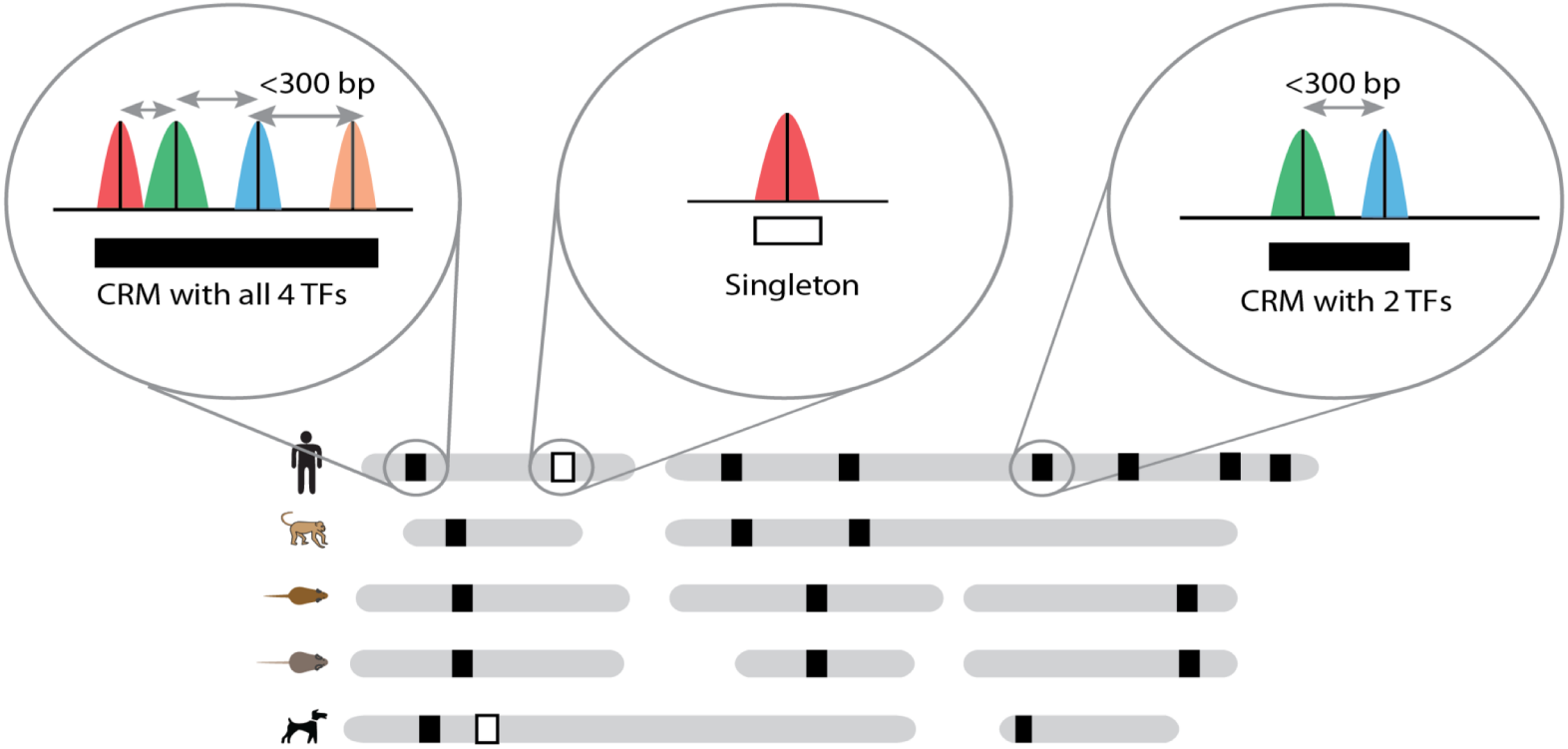
Collaboration:
Benoit Ballester (INSERM)
Duncan Odom (Cambridge U)
Paul Flicek (EBI)

Regulator	Mouse knockout phenotype	Liver phenotype	TF PFAM Class
FoxA1	No gastrulation	no liver formation	Forkhead
HNF4a	No gastrulation	disrupted liver gene expr.	Nuclear Receptor
ONECUT1	Poor hepatocyte, b-cell proliferation	no gall bladder	Onecut domain
CEBPA	Disturbed liver architecture	No glycogen storage	bZIP

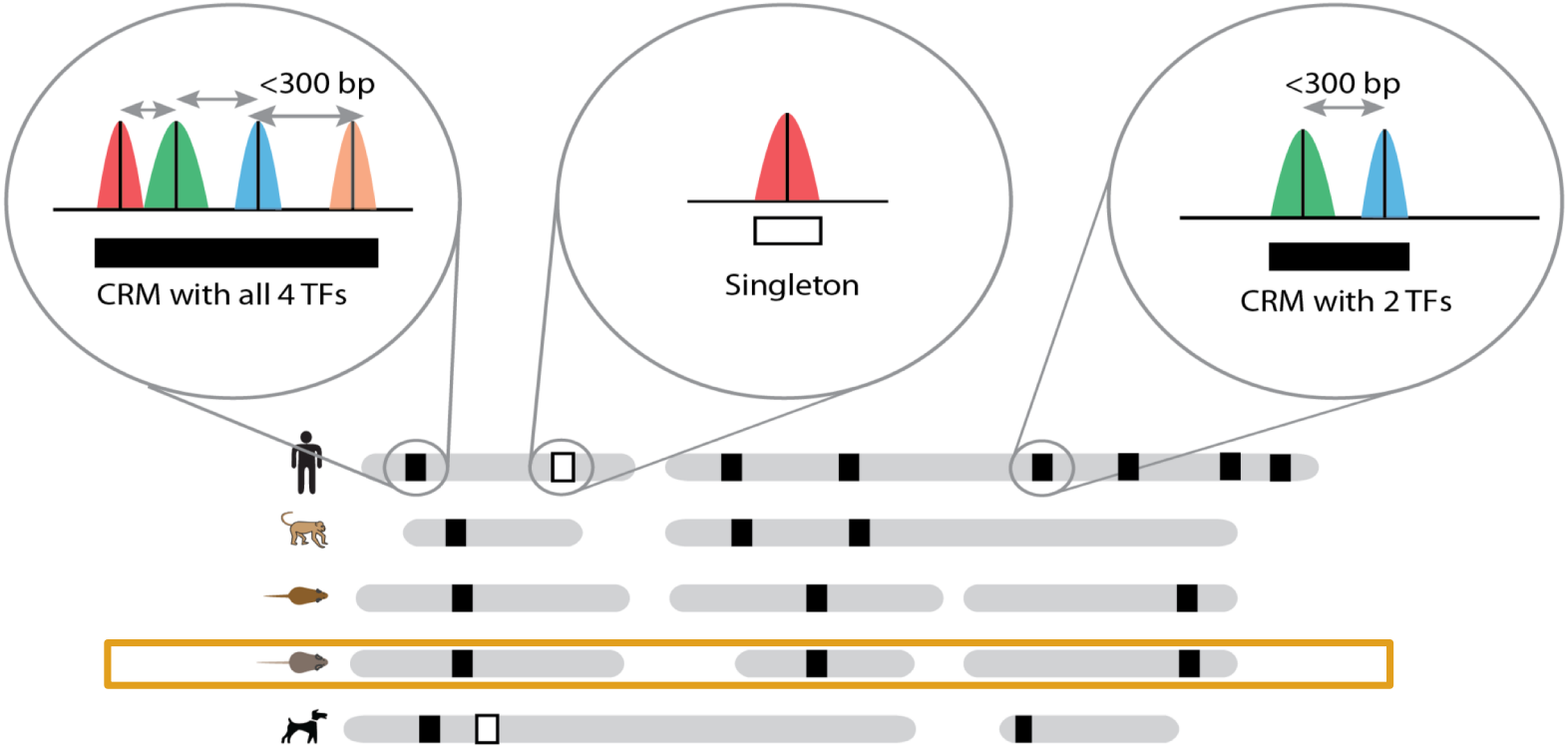
Transcription factors cluster together to regulate genes
CRMs



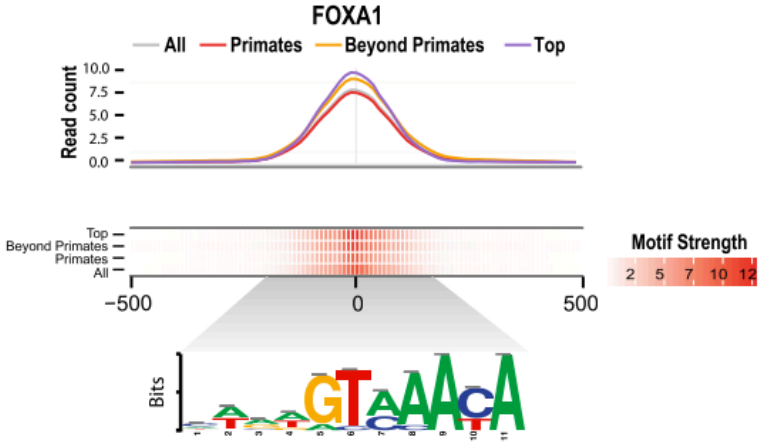
Transcription factors cluster together to regulate genes
CRMs

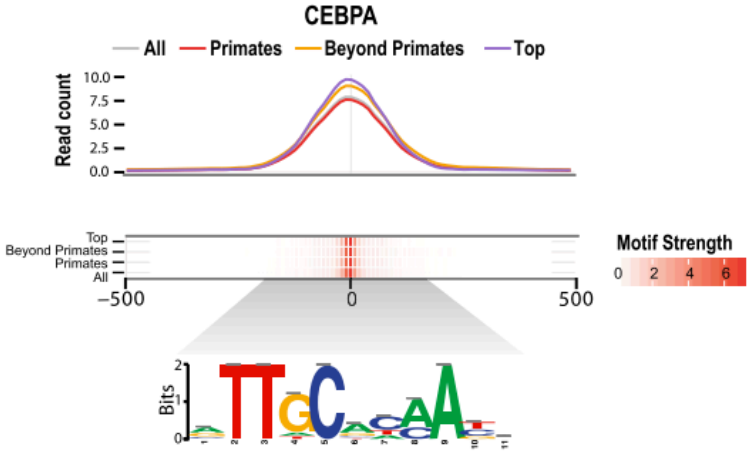


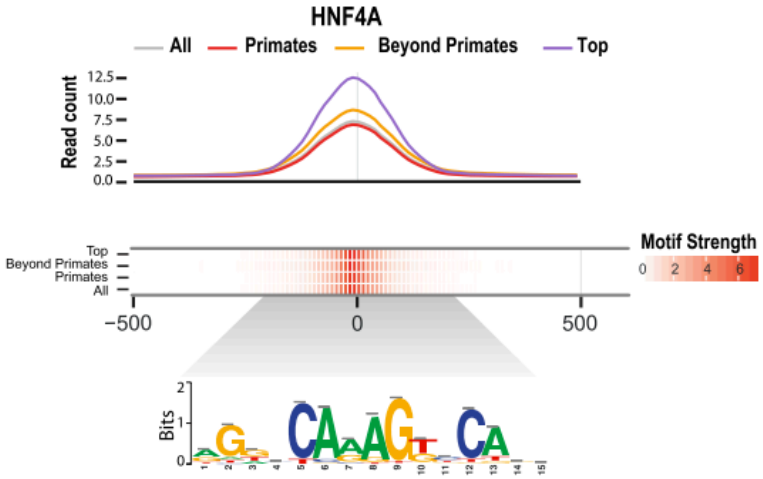
Transcription factors cluster together to regulate genes
CRMs

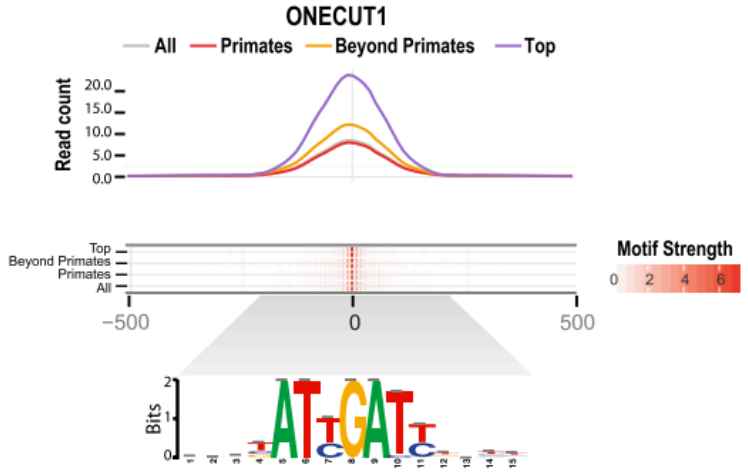


Transcription factors cluster together to regulate genes CRMs



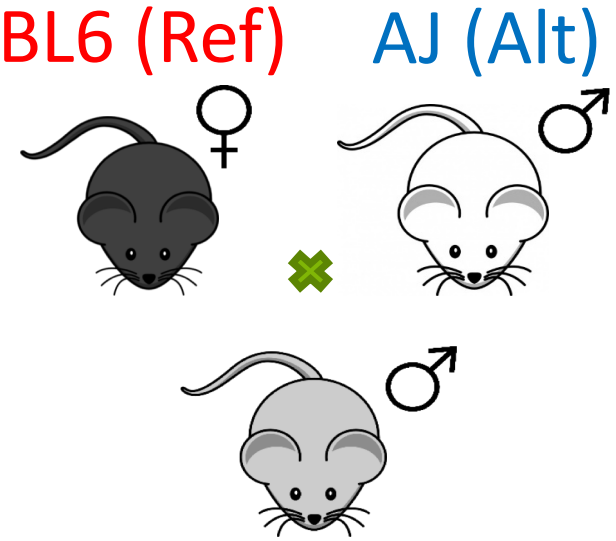






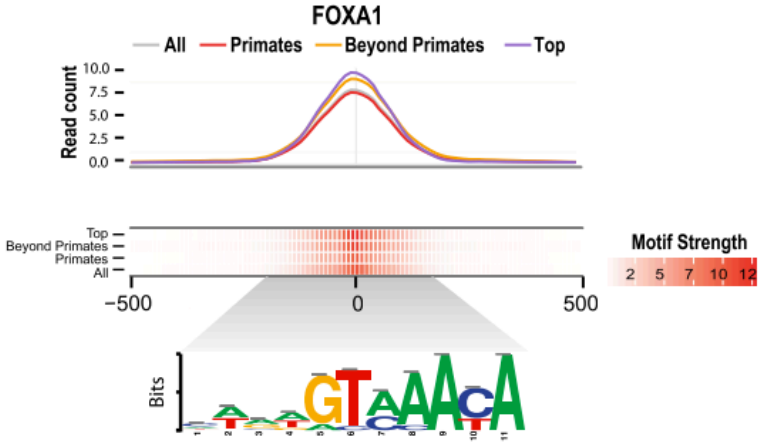
Heterozygous ChIP-seq: Can we detect co-factor effects?

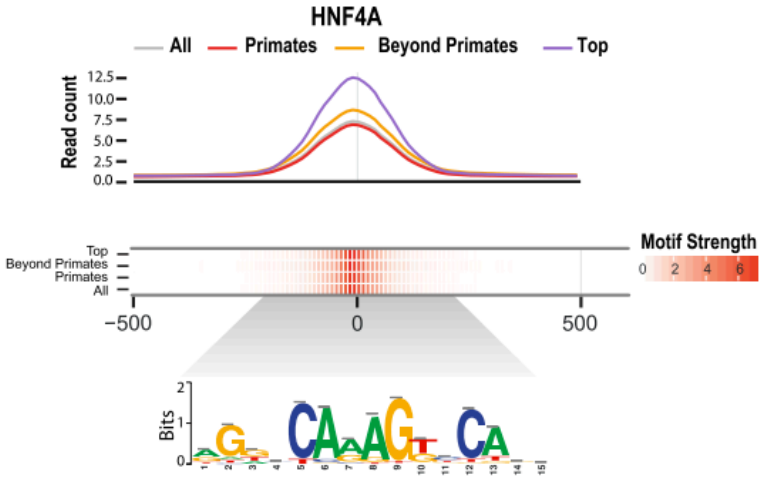
- F1 mice with inbred parents

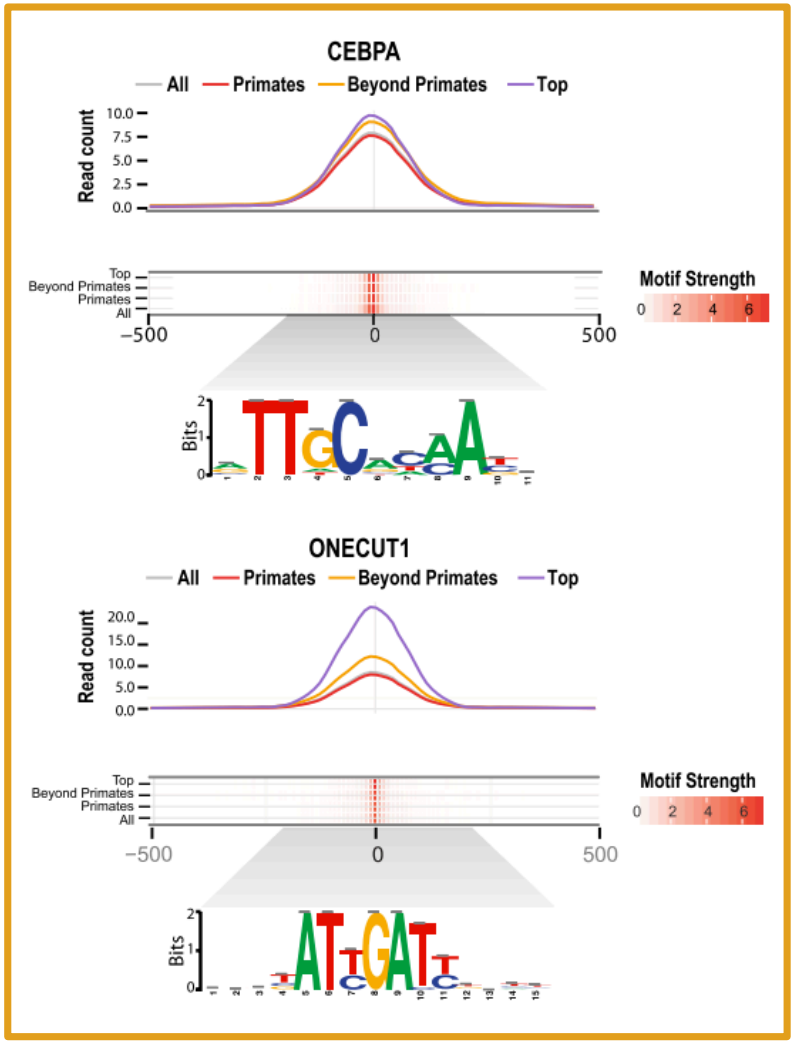


HNF6
CEBPA

Transcription factors cluster together to regulate genes CRMs

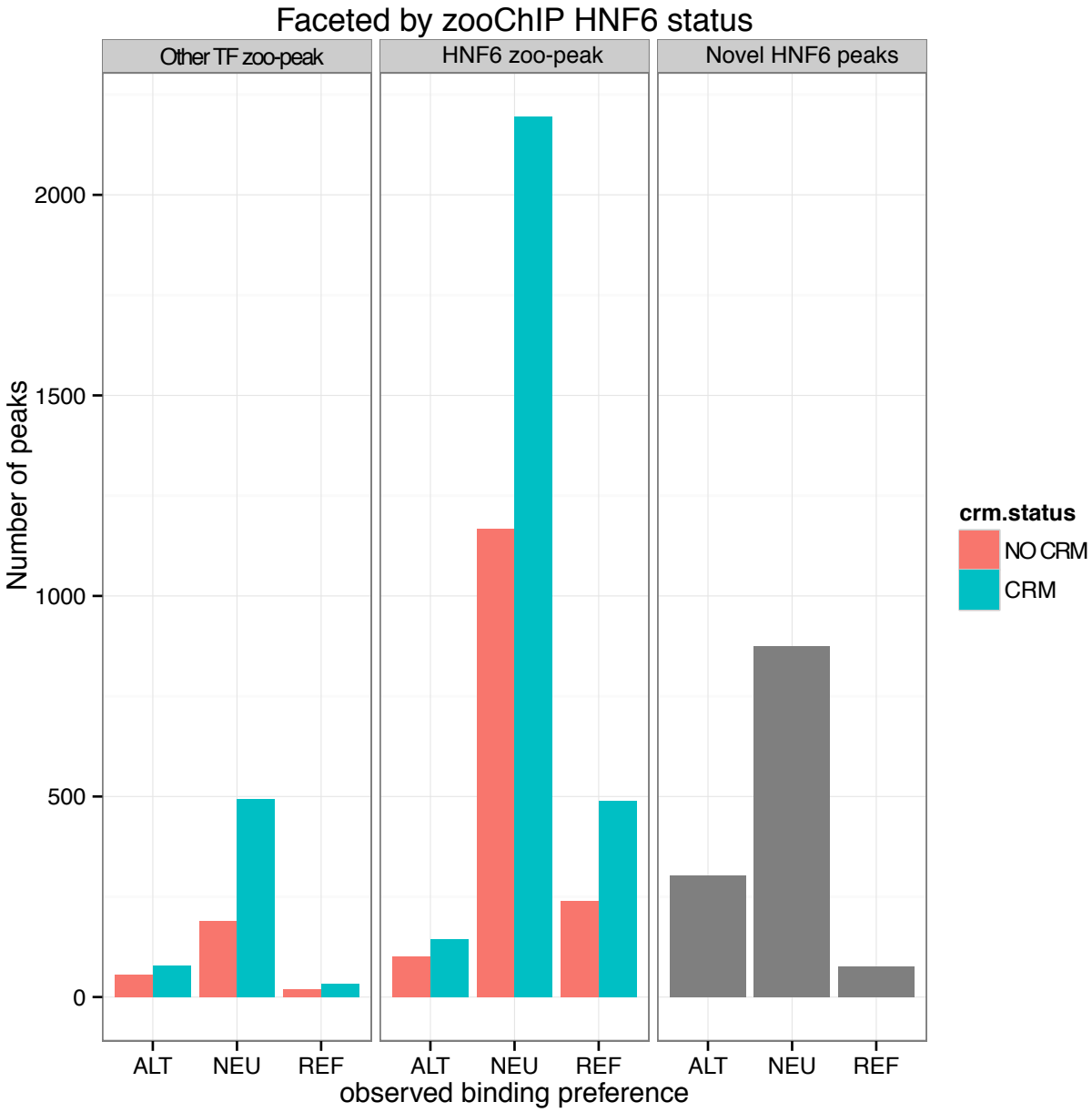




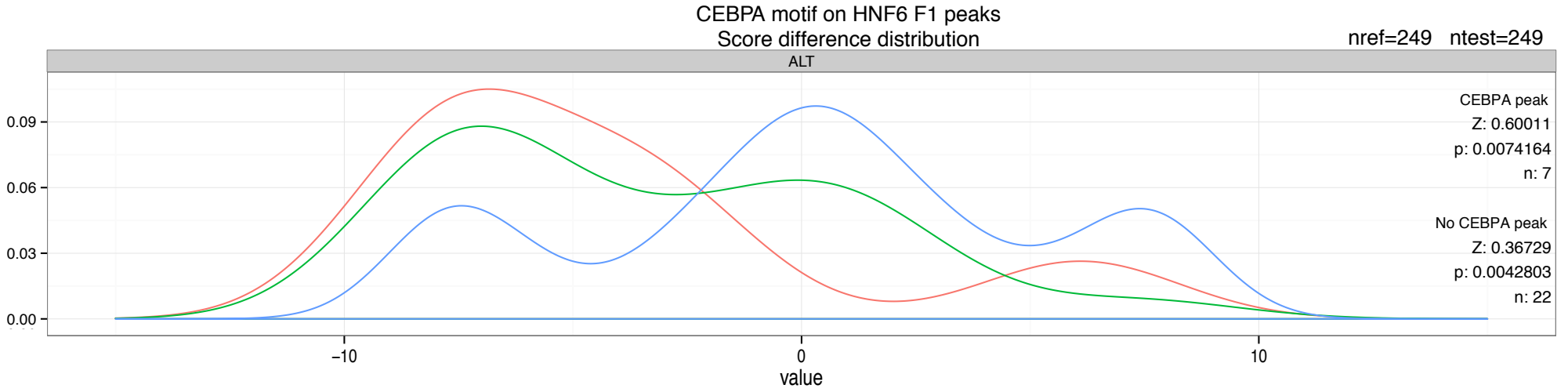


Heterozygous ChIP-seq: Can we detect co-factor effects? HNF6 F1 data set

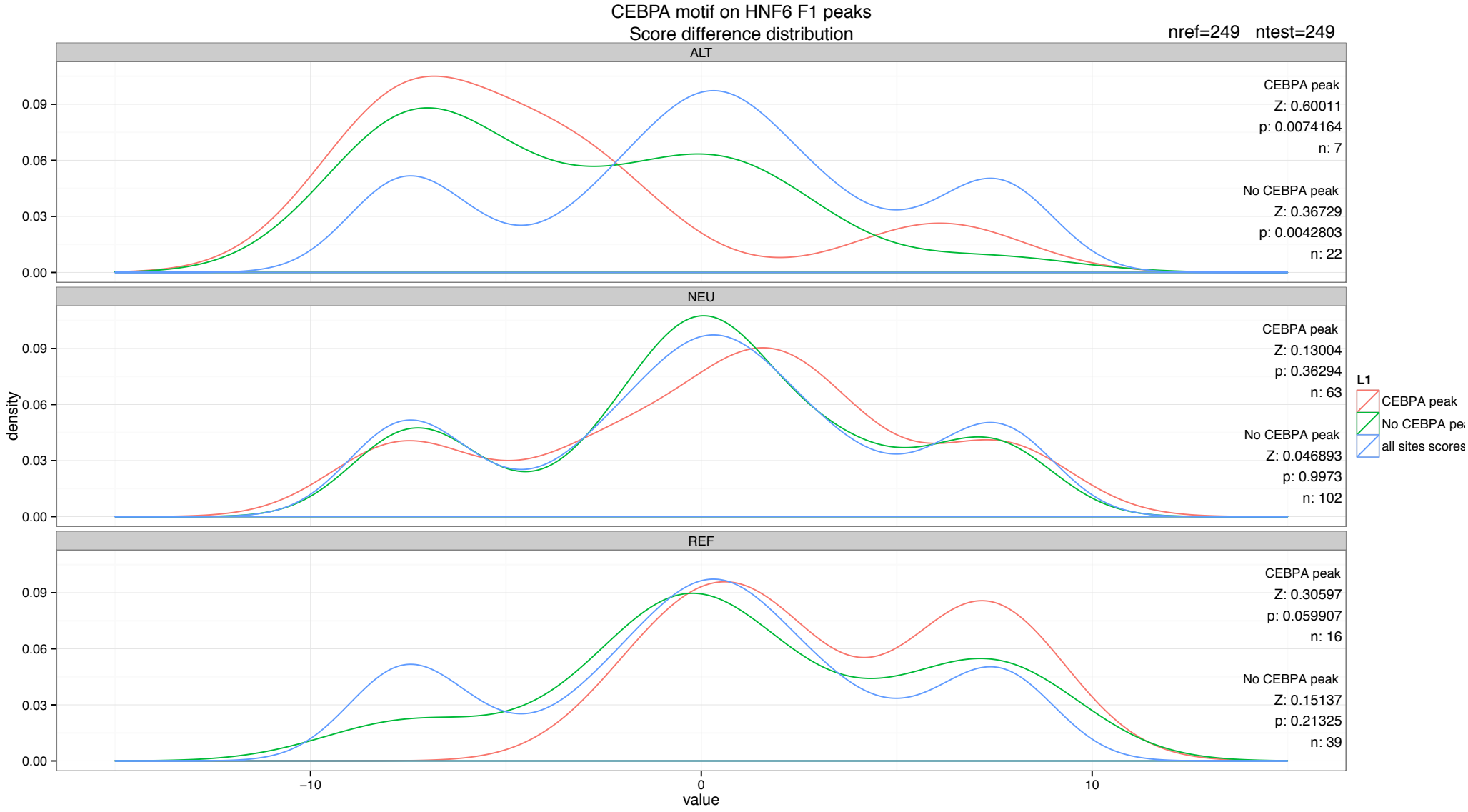
- 6458 Peaks with SNPs
- 2123 sites are novel HNF6 binding.
- 3433 overlap an zoo-Chip HNF6.
- 868 are gains at existing binding sites for other TFs.
- 1253 represent novel binding sites that don't overlap any peak from the Zoo-Chip mouse data.



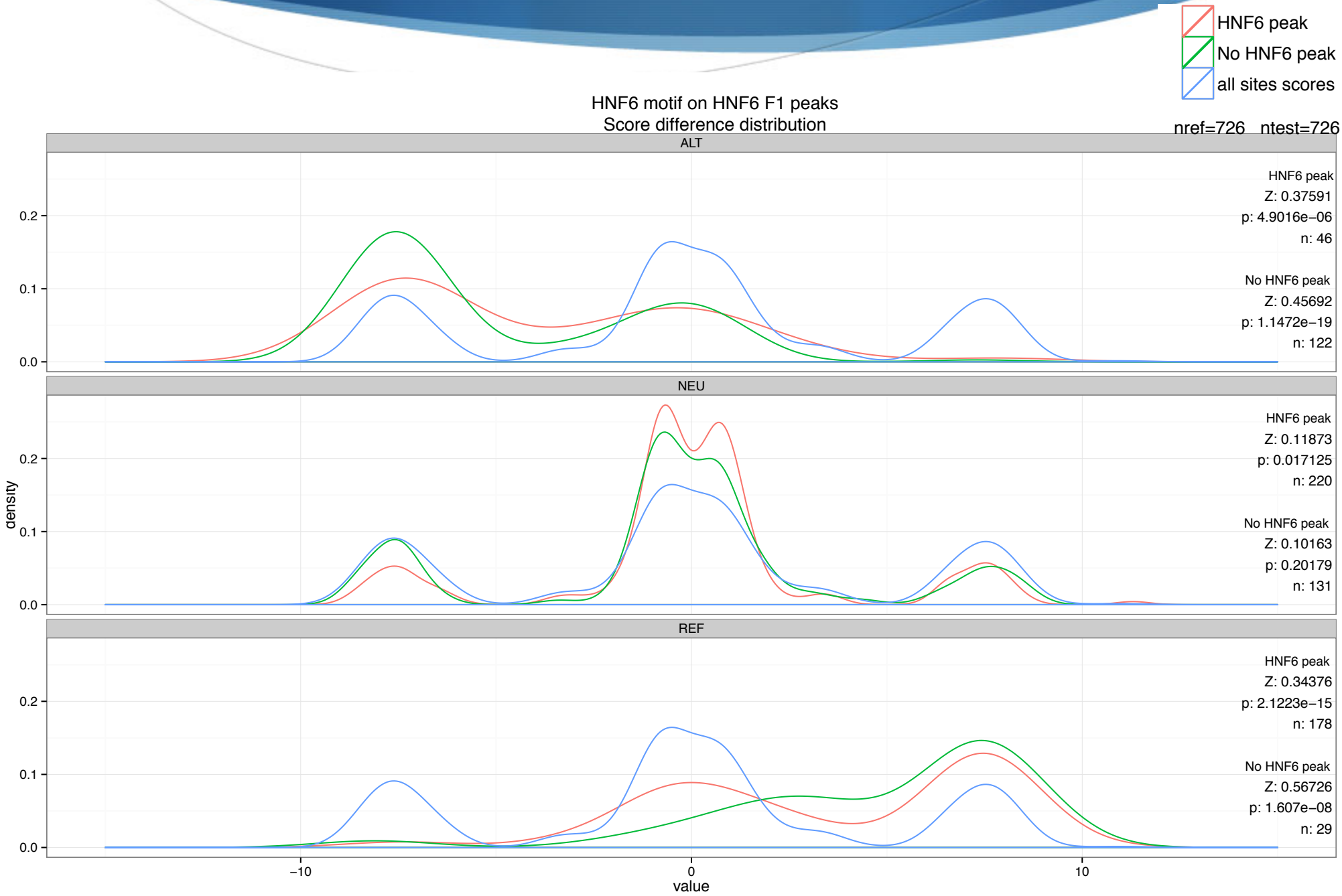
Heterozygous ChIP-seq: Can we detect co-factor effects?



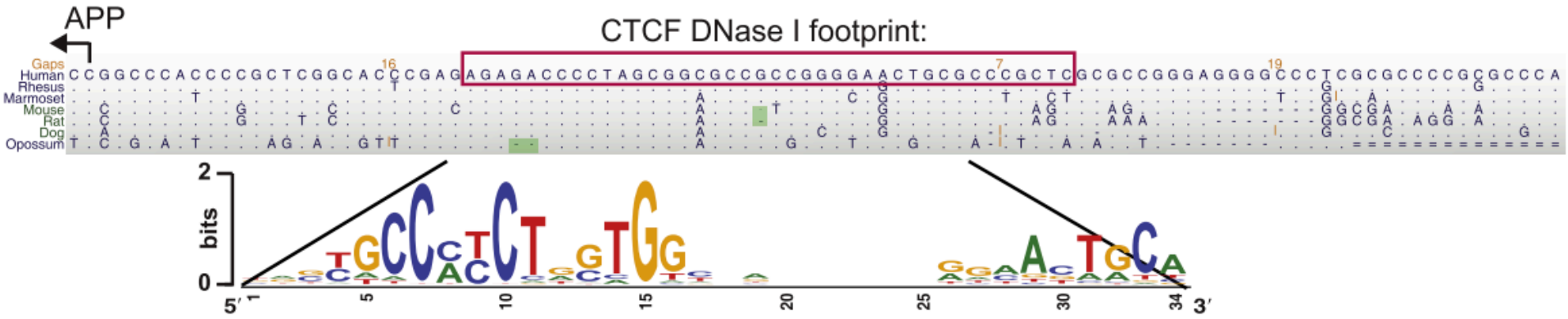
Heterozygous ChIP-seq: Can we detect co-factor effects?

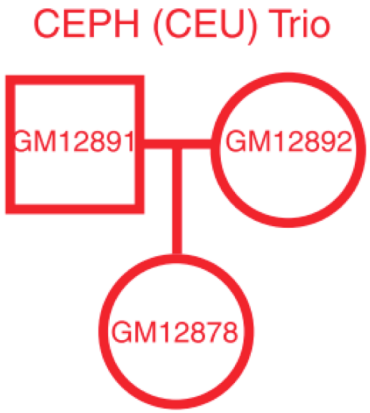


Heterozygous ChIP-seq: Can we detect co-factor effects?



Heterozygous ChIP-seq: CTCF trio





Can we detect allele effects on secondary motifs?

Schmidt, D., Schwalie, P. C., Wilson, M., Ballester, B., Gonçalves, A.et al (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. CELL