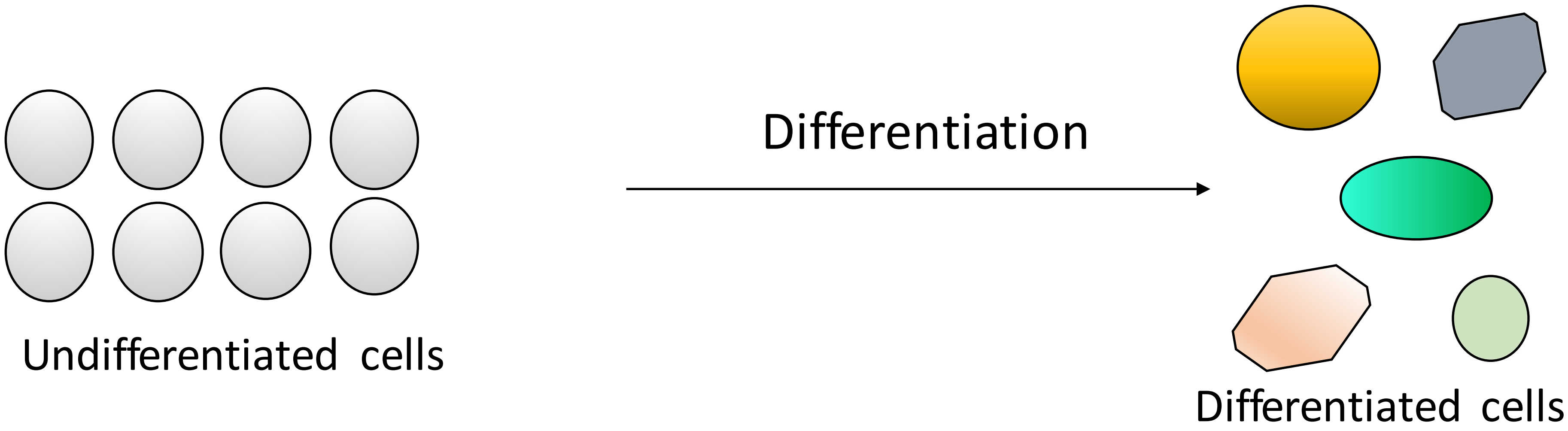
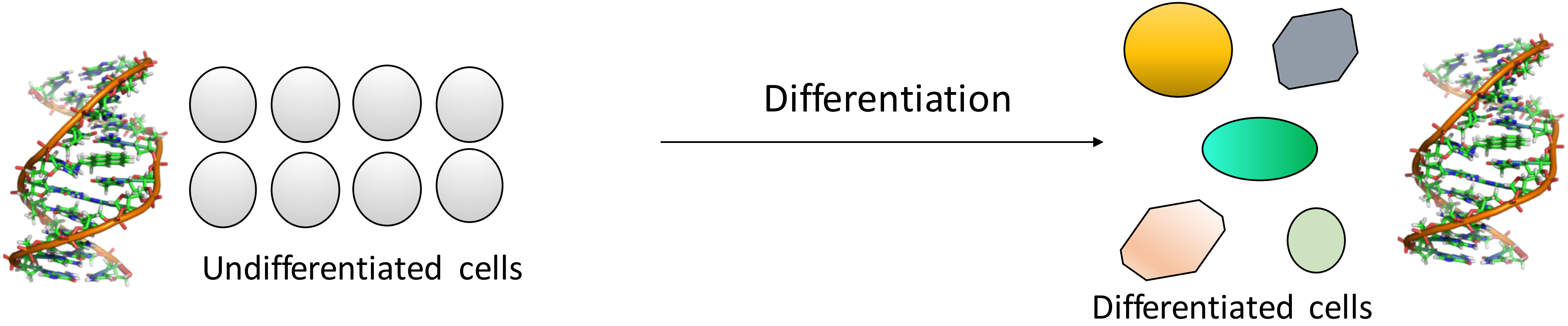
Computational prediction of cis-regulatory elements in eukaryotic genomes.

Dmitry Svetlichnyy

Single input multiple outputs



Single input multiple outputs

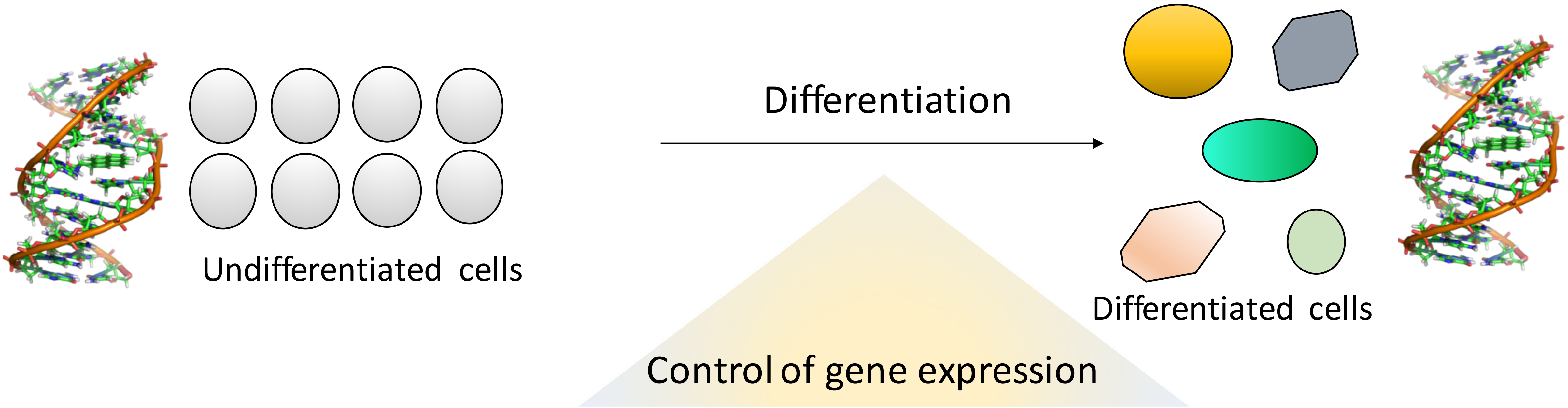


Undifferentiated cells

Differentiation

Differentiated cells

Single input multiple outputs



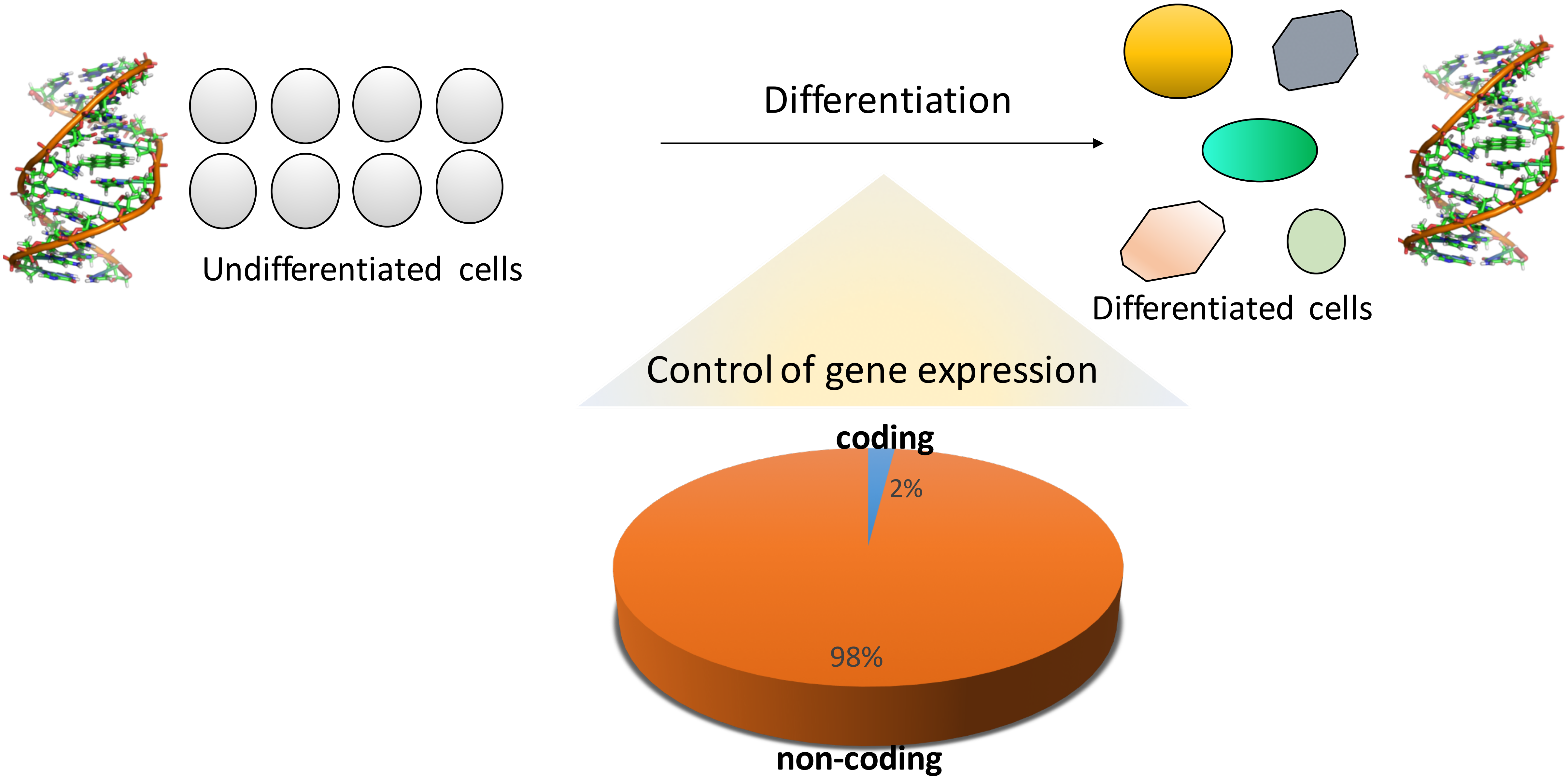
Undifferentiated cells

Differentiation

Differentiated cells

Control of gene expression

Single input multiple outputs

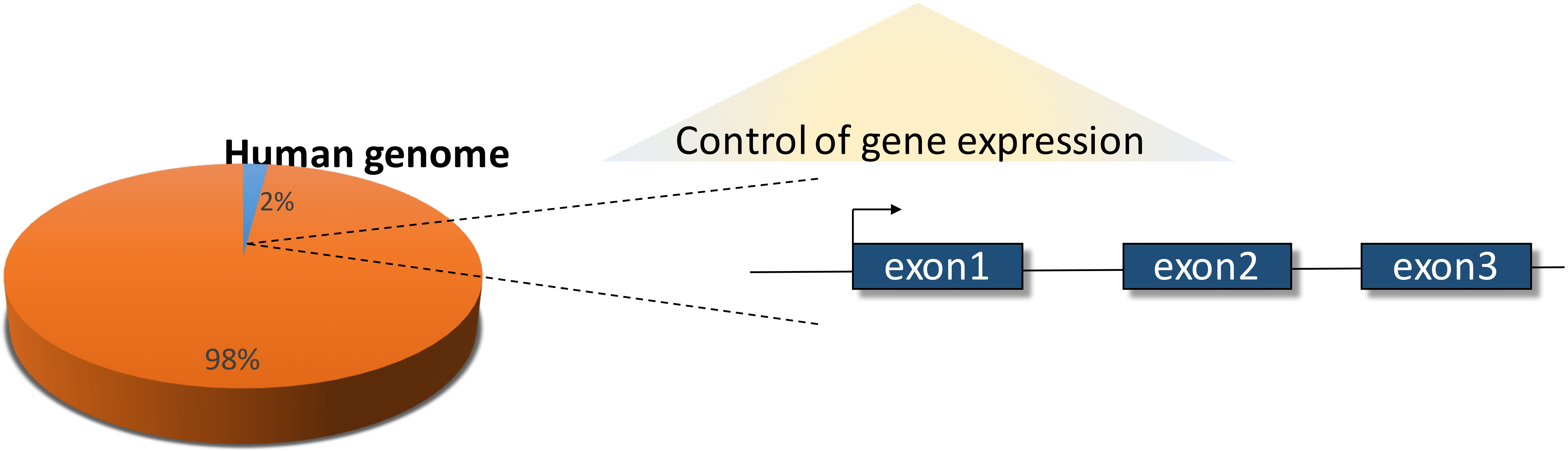


Undifferentiated cells

Differentiation

Differentiated cells

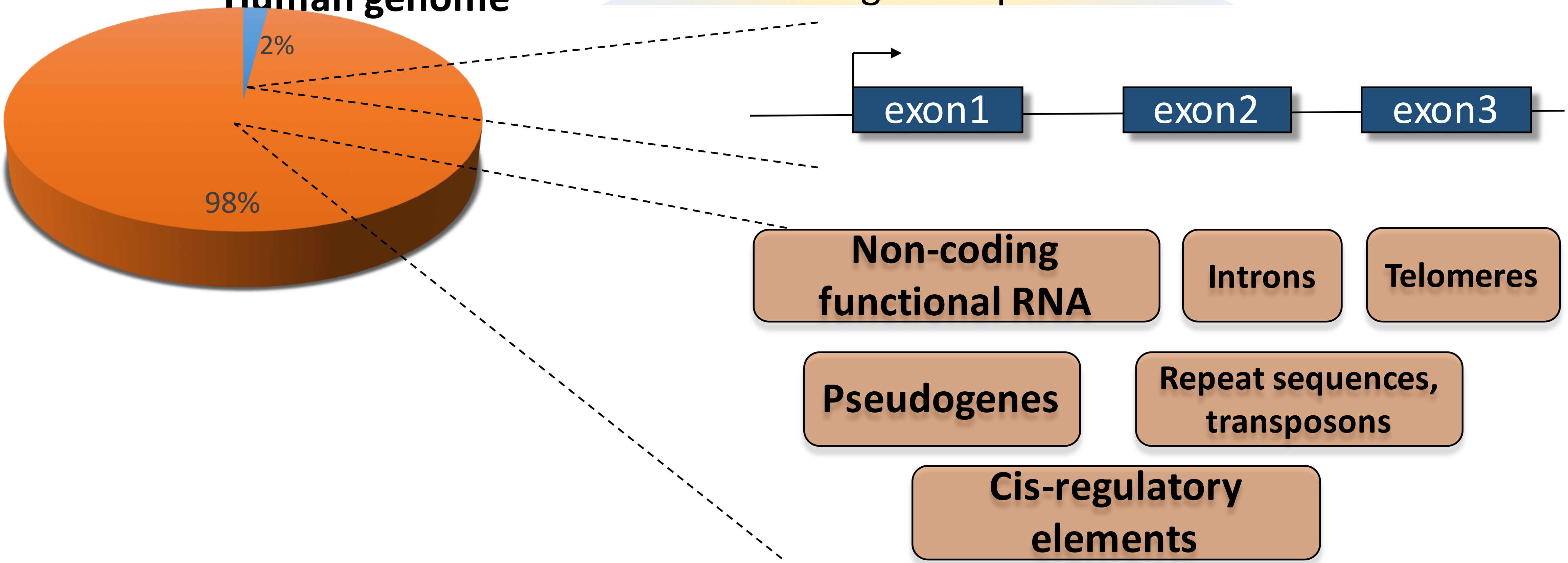
Control of gene expression



Control of gene expression

Human genome

Control of gene expression



Non-coding functional RNA

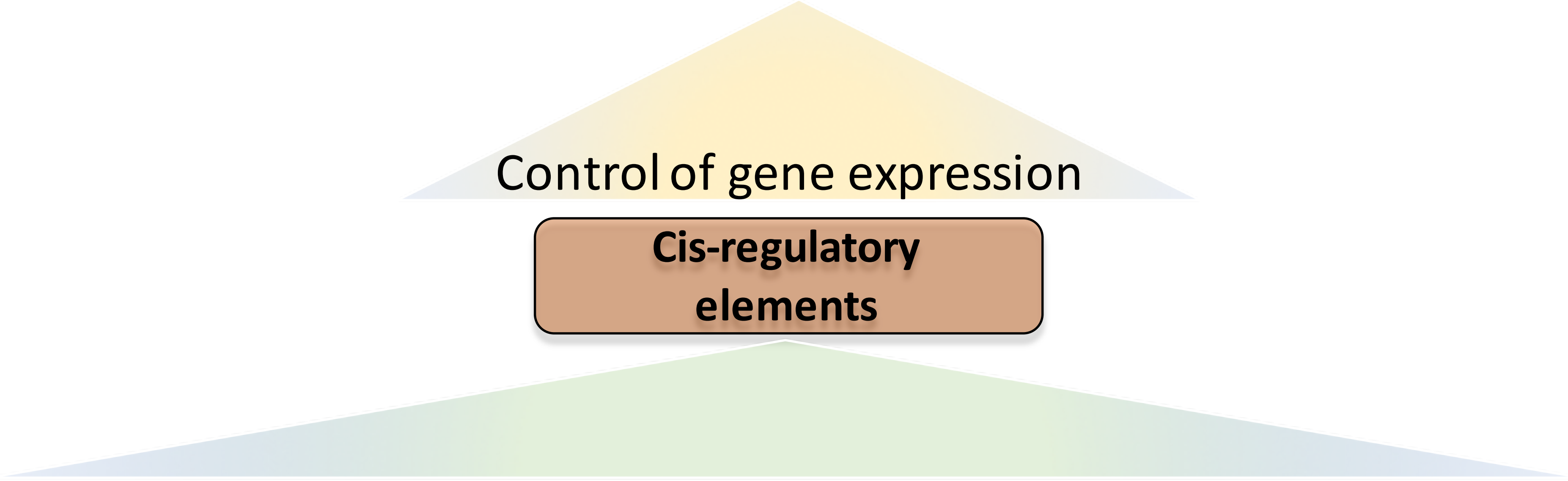
Introns

Telomeres

Pseudogenes

Repeat sequences, transposons

Cis-regulatory elements

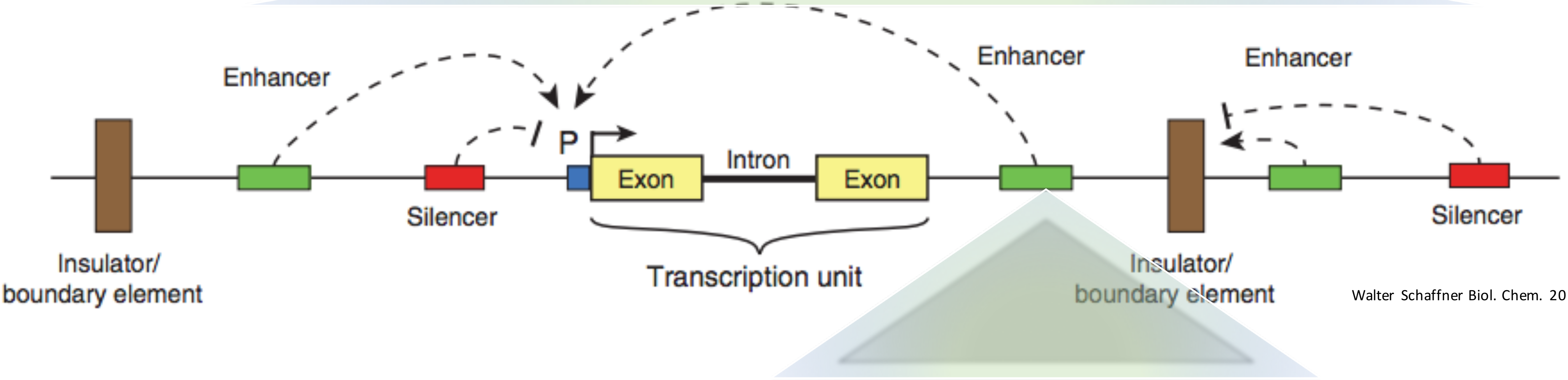


Control of gene expression

Cis-regulatory elements

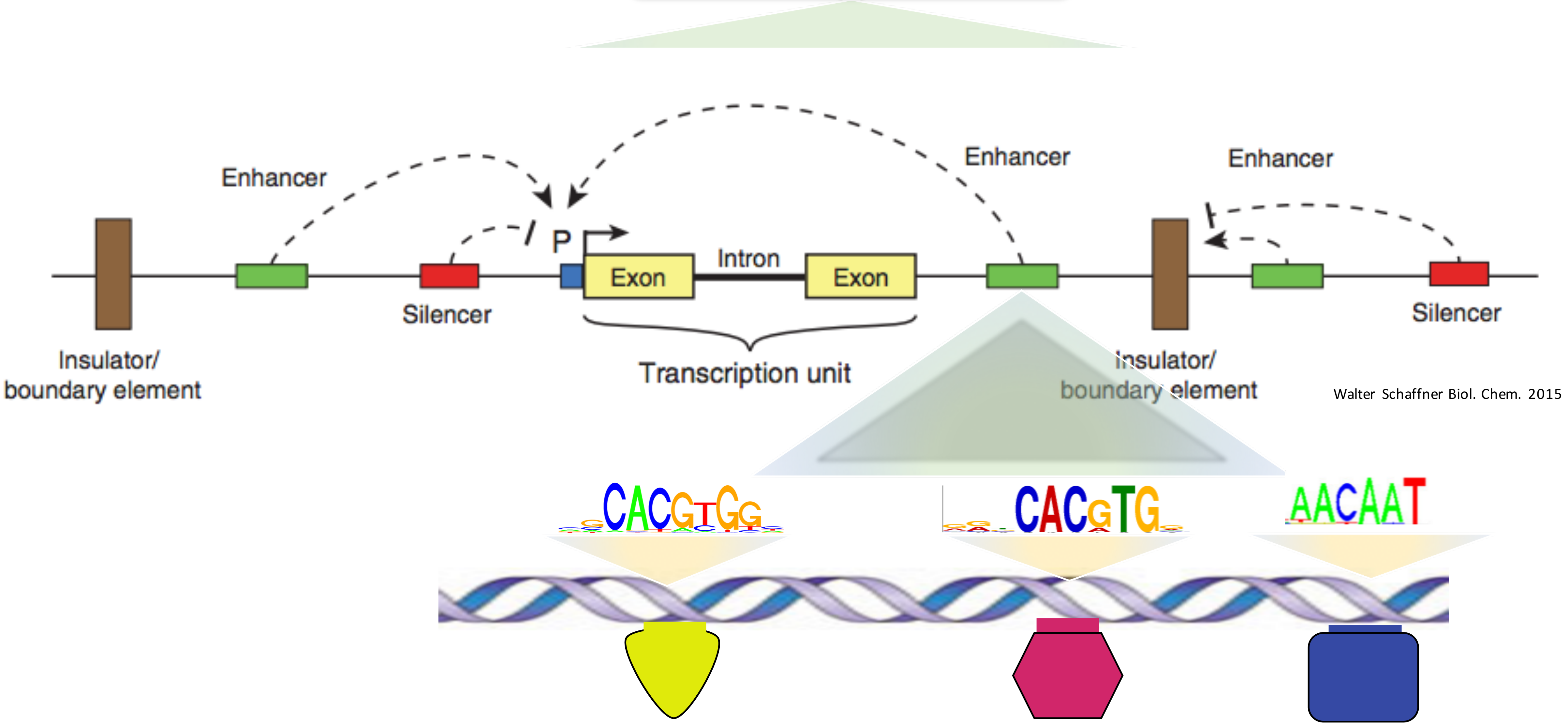
Control of gene expression

Cis-regulatory elements



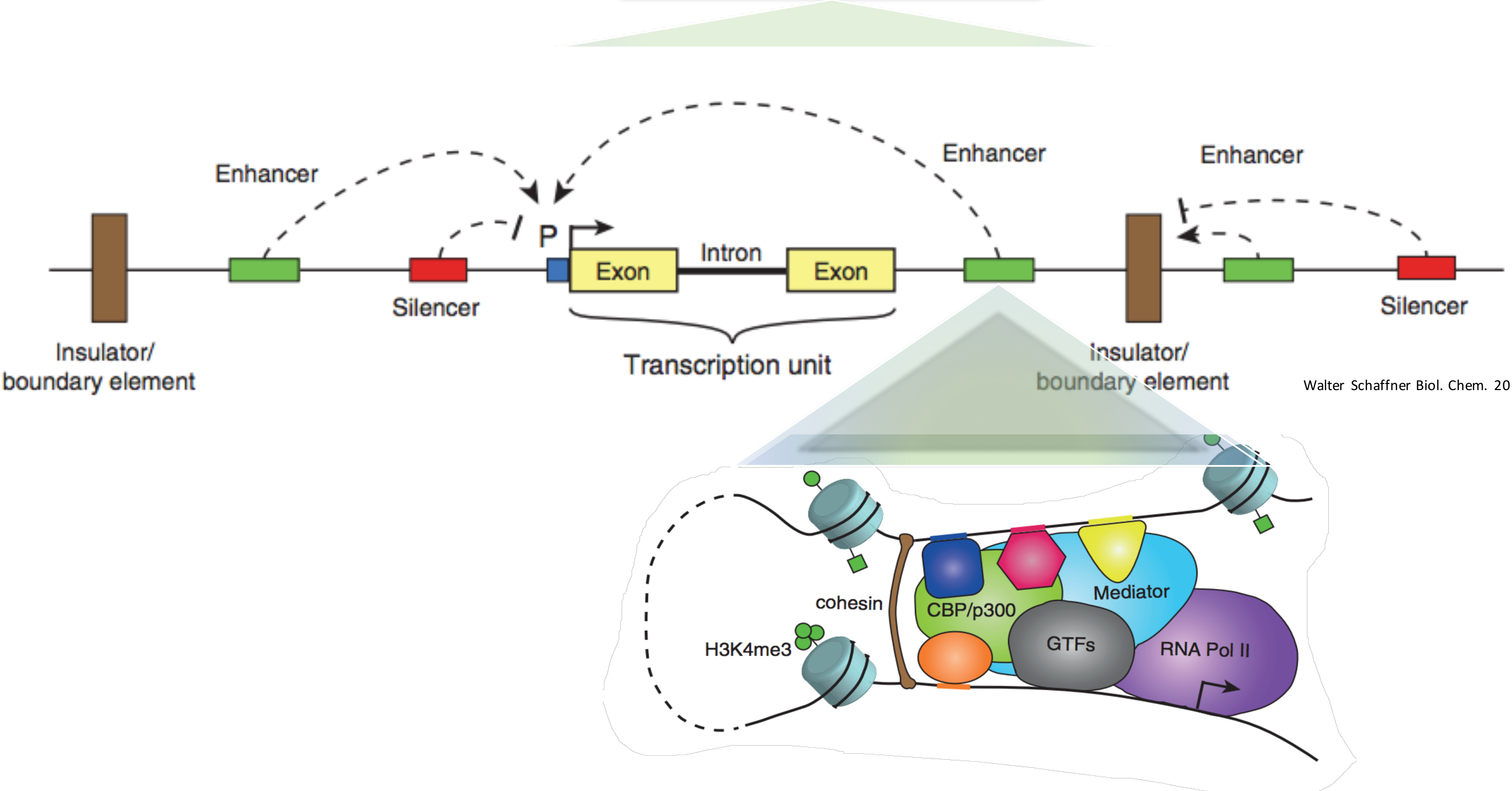
Walter Schaffner Biol. Chem. 2015

Cis-regulatory elements



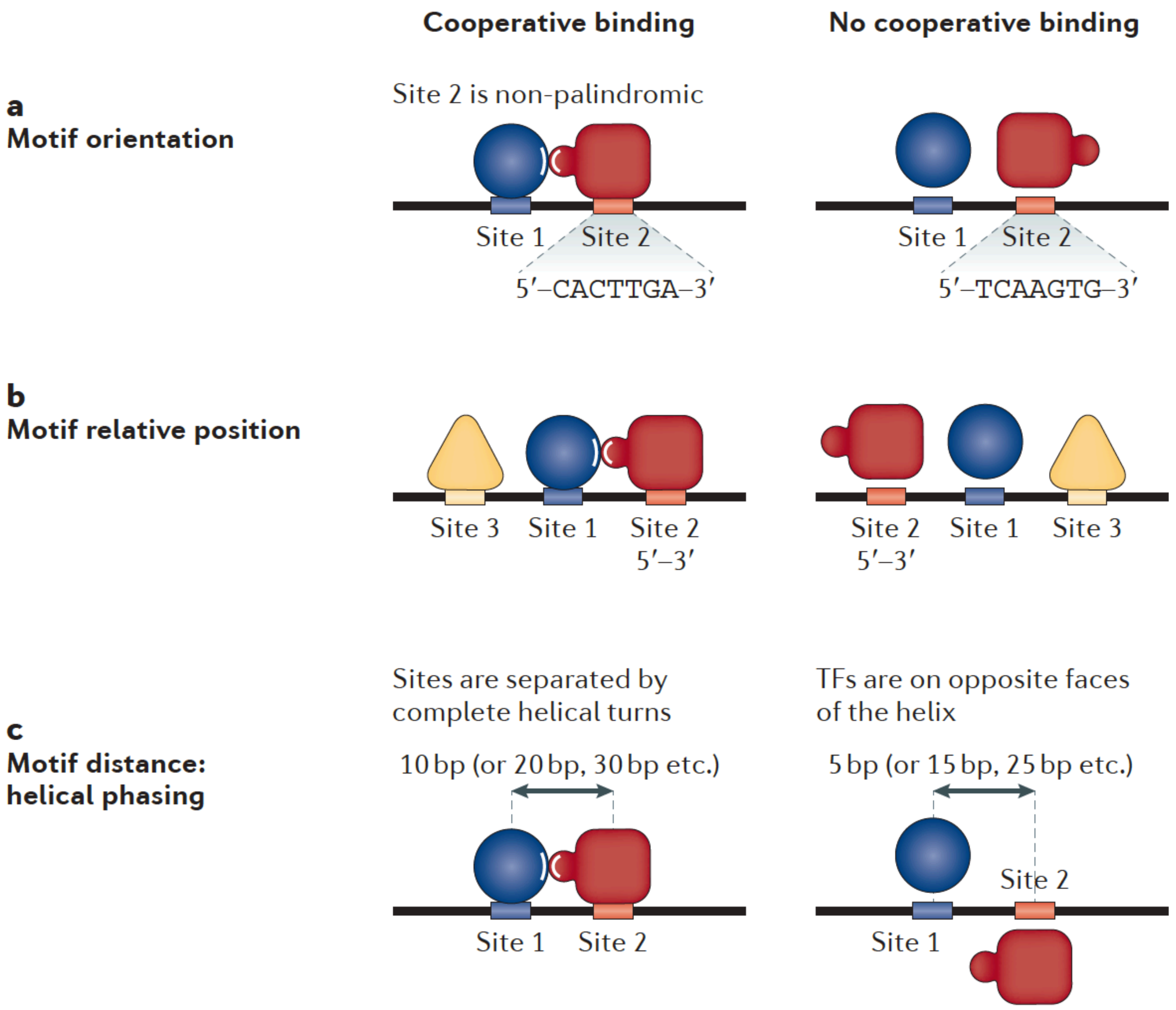
Walter Schaffner Biol. Chem. 2015

Cis-regulatory elements



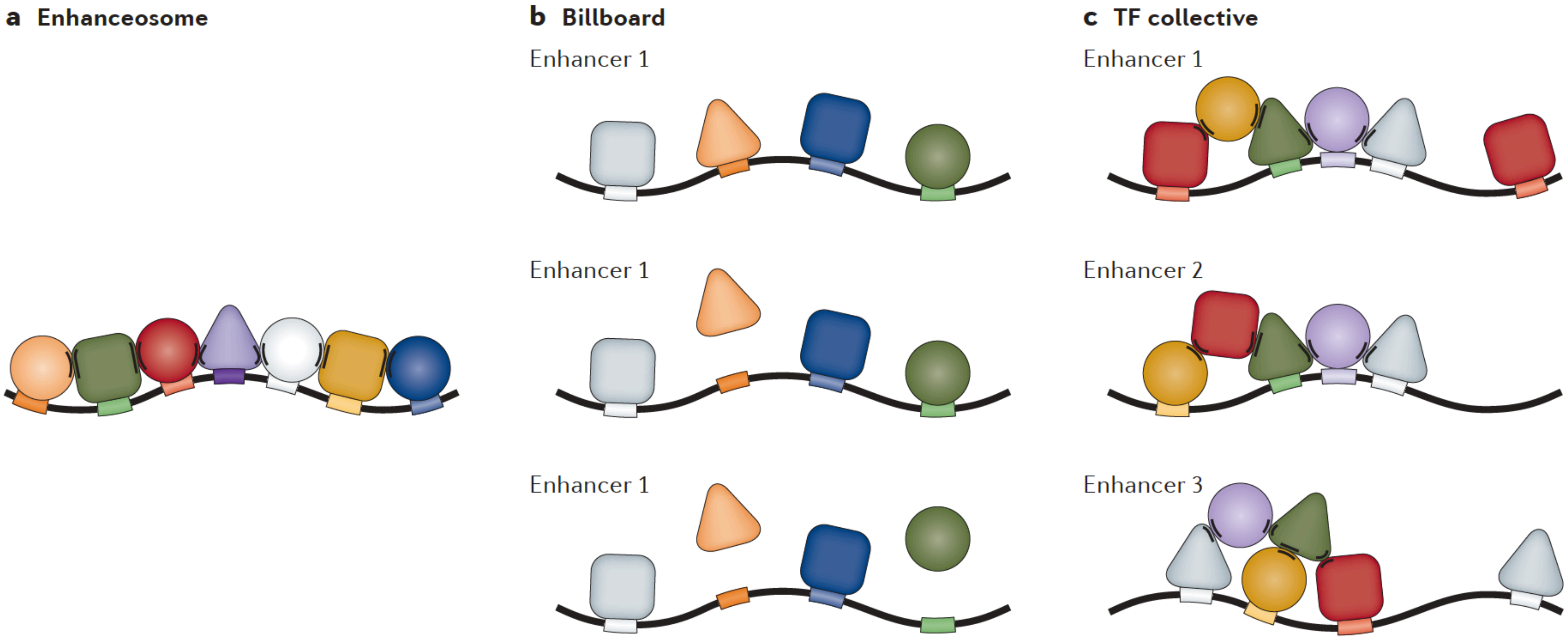
Walter Schaffner Biol. Chem. 2015

Problems



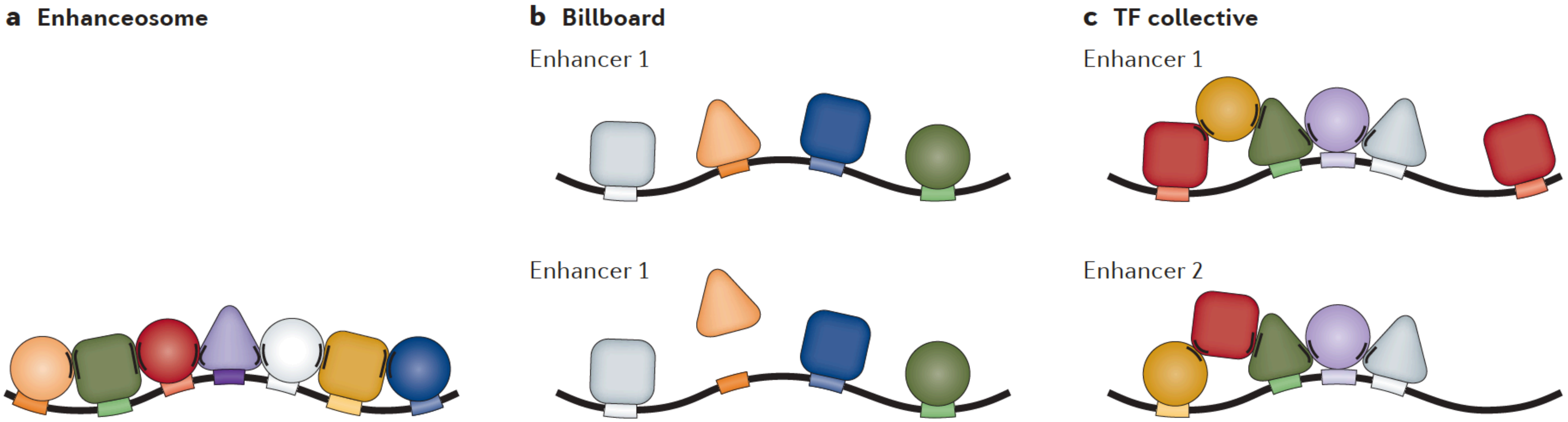
Spitz et al.

Problems



Spitz et al.

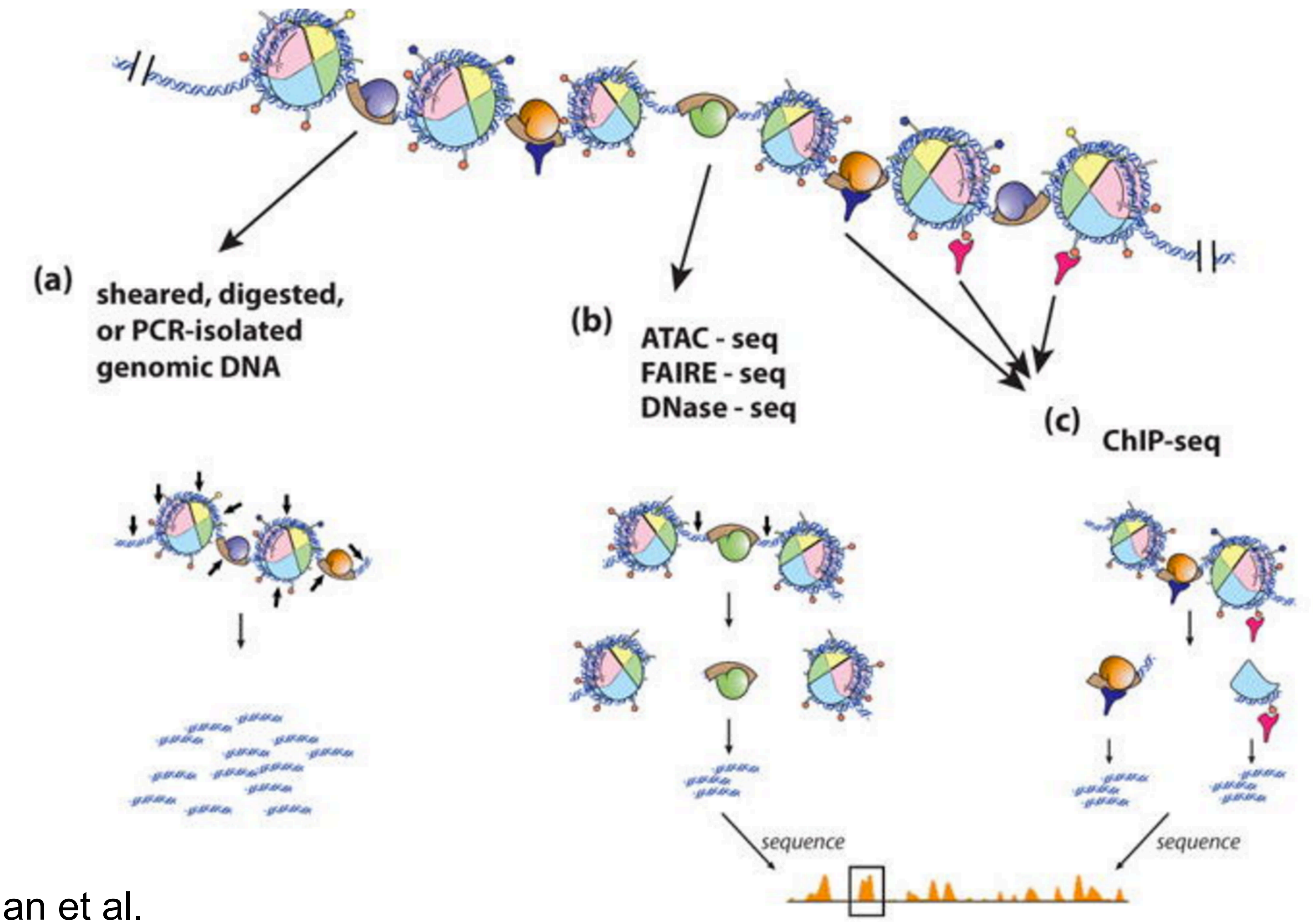
Problems



Protein DNA	• Highly cooperative DNA binding • DNA sequence acts as scaffold	• Cooperative and additive binding	• Cooperative DNA binding • Both DNA and protein may act as scaffold
Protein interface	• Fixed (formed from a higher-order TF–DNA complex)	• Variable	• Variable
Motif	• Fixed motif composition (sites for all factors must be there) • Fixed motif positioning (grammar)	• Fixed motif composition • Flexible motif grammar	• Flexible motif composition (as different TFs directly bind to DNA) • Flexible motif grammar
Output	• Unitary (requires the integrated activity of all TFs)	• Only requires a subset of factors to be active	• Collective (requires most TFs but not clear if it requires all TFs)

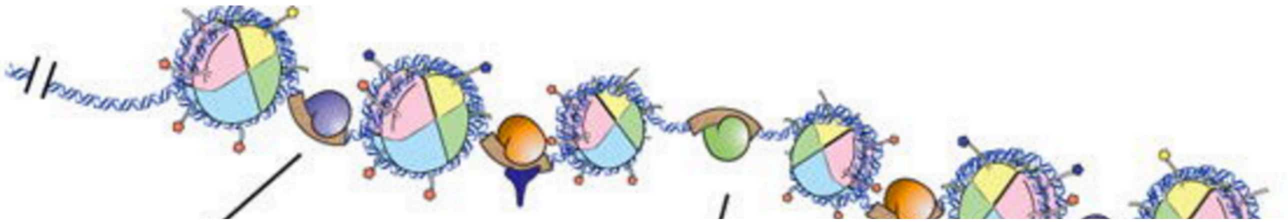
Spitz et al.

Chromatin modifications and various classes of CRMs

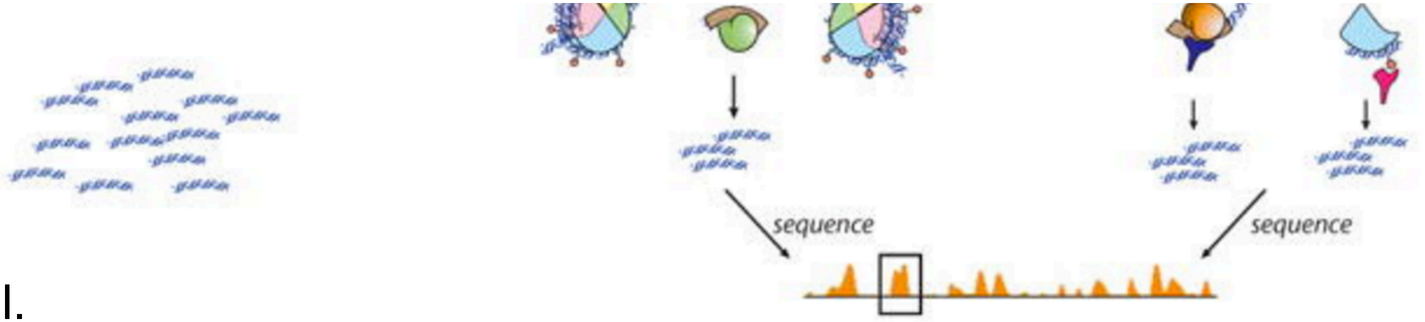


Suryamohan et al.

Chromatin modifications and various classes of CRMs

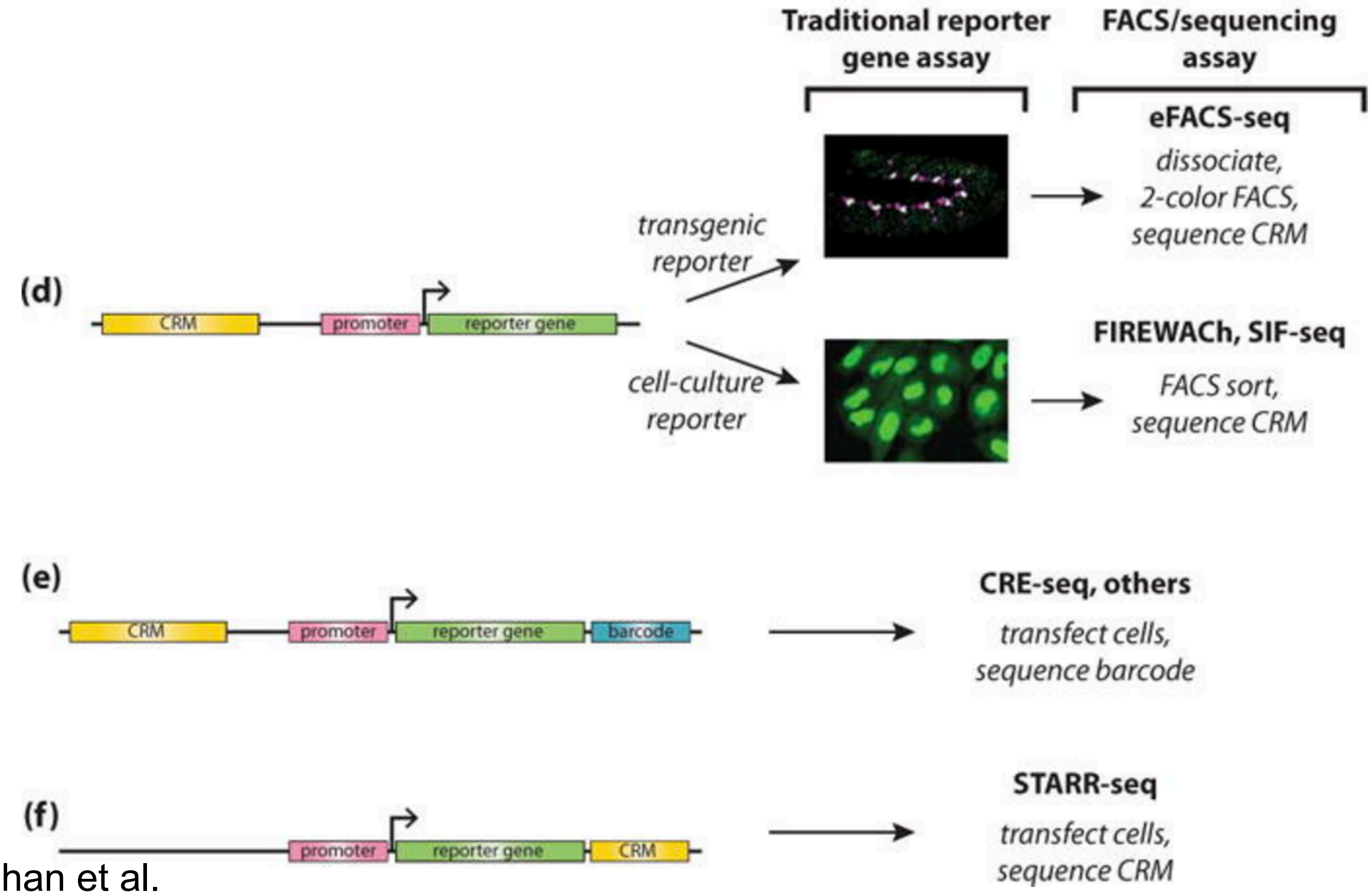


Regulatory element	DHS	H3K4Me1	H3K4Me3	H3K27Ac
Promoter	+	-	+	++
Enhancer	+	+	-	++
Insulator	+	-	+/-	-



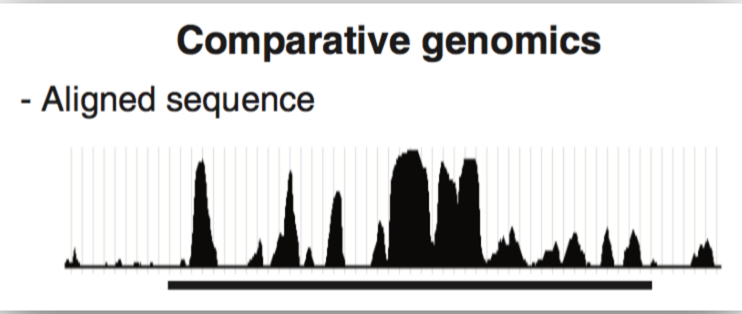
Suryamohan et al.

Chromatin modifications and various classes of CRMs



Suryamohan et al.

Approaches for computational prediction of CRMs



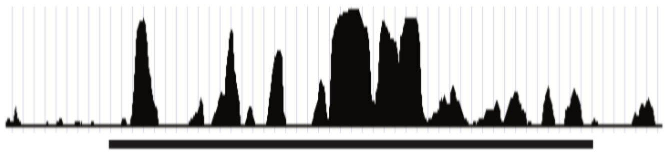
Comparative genomics

- Aligned sequence

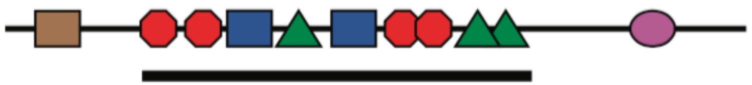
Approaches for computational prediction of CRMs

Comparative genomics

- Aligned sequence



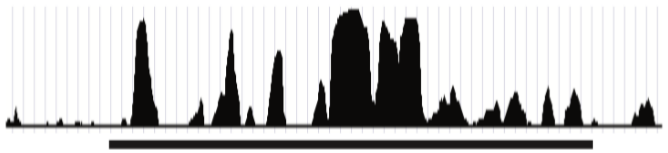
Motif based



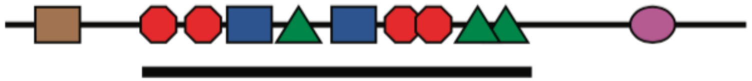
Approaches for computational prediction of CRMs

Comparative genomics

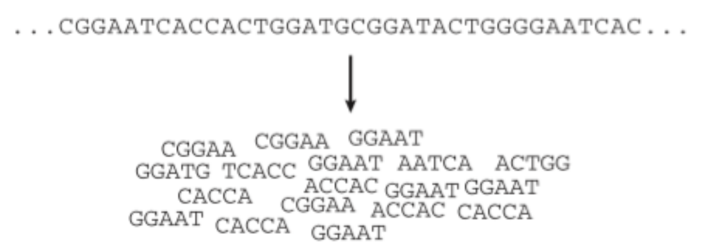
- Aligned sequence



Motif based

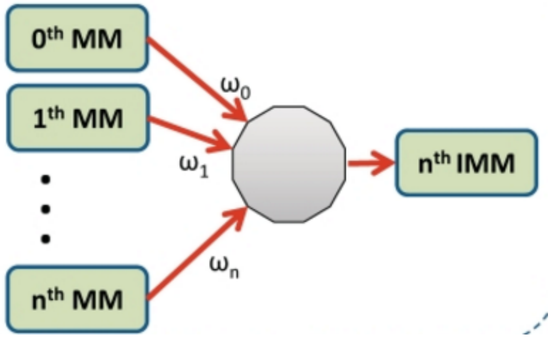


Motif blind


...CGGAATCACCACTGGATGCGGATACTGGGGAATCAC...
CGGAA CGGAA GGAAT
GGATG TCACC GGAAT AATCA ACTGG
CACCA ACCAC GGAAT GGAAT
GGAAT CACCA CGGAA ACCAC CACCA
GGAAT

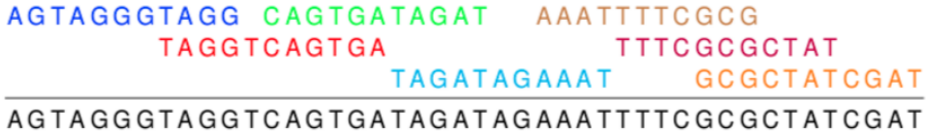
Li et al. Biosystems 2015

Interpolated Markov Model

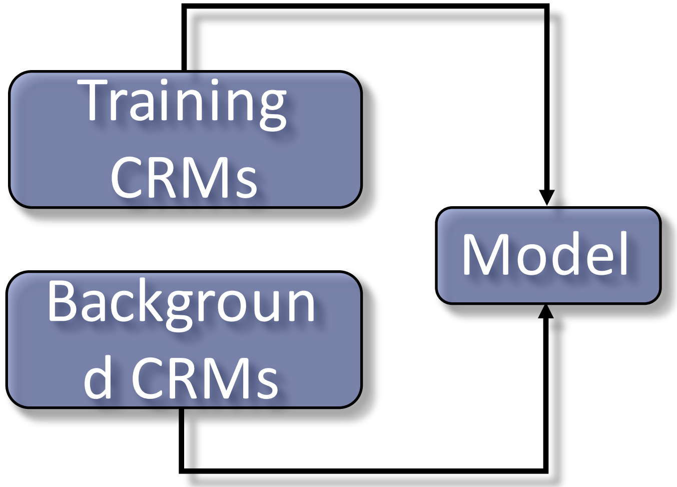


Kazemian et al. NAR 2011

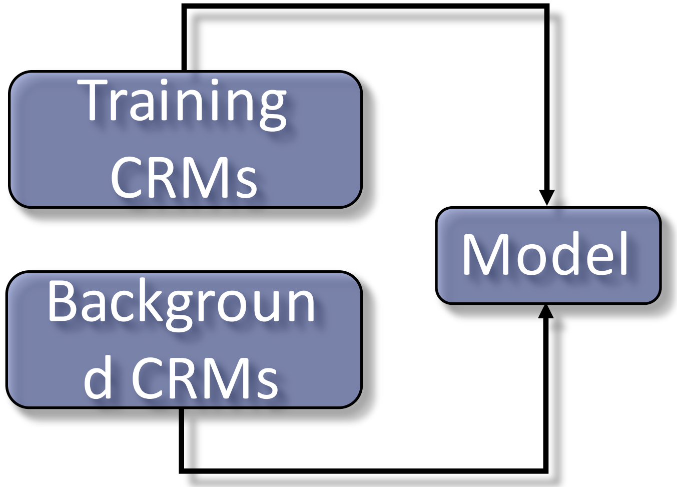
String kernel


AGTAGGGTAGG CAGTGATAGAT AAATTTTCGCG
TAGGTCAGTGA TTTCGCGCTAT
TAGATAGAAAT GCGCTATCGAT
AGTAGGGTAGGTCAGTGATAGATAGAAATTTTCGCGCTATCGAT

Computational prediction of CRMs using training data



Computational prediction of CRMs using training data



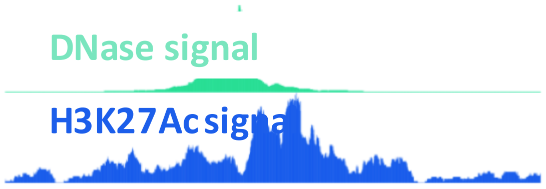
Training CRMs

Background CRMs

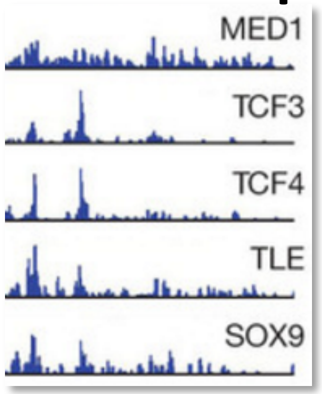
Model

Training CRMs

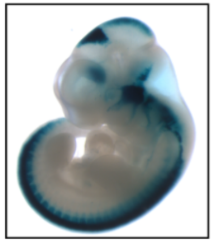
Chromatin profiling



TF ChIP-seq

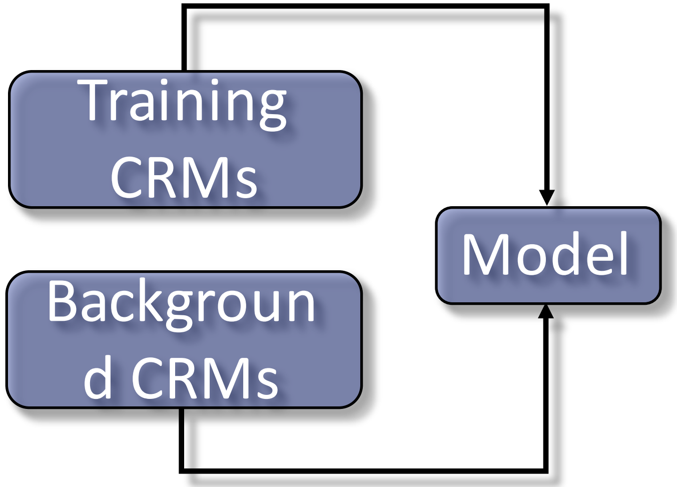


In vivo CRM activity



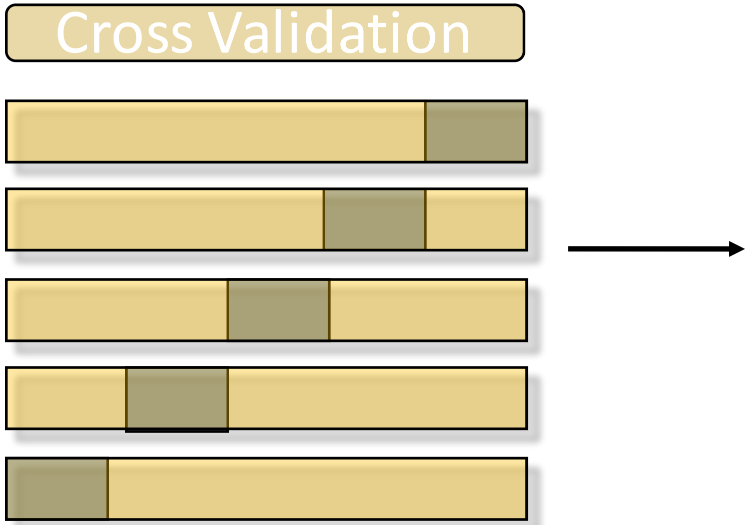
https://www.encodeproject.org/data/annotations/

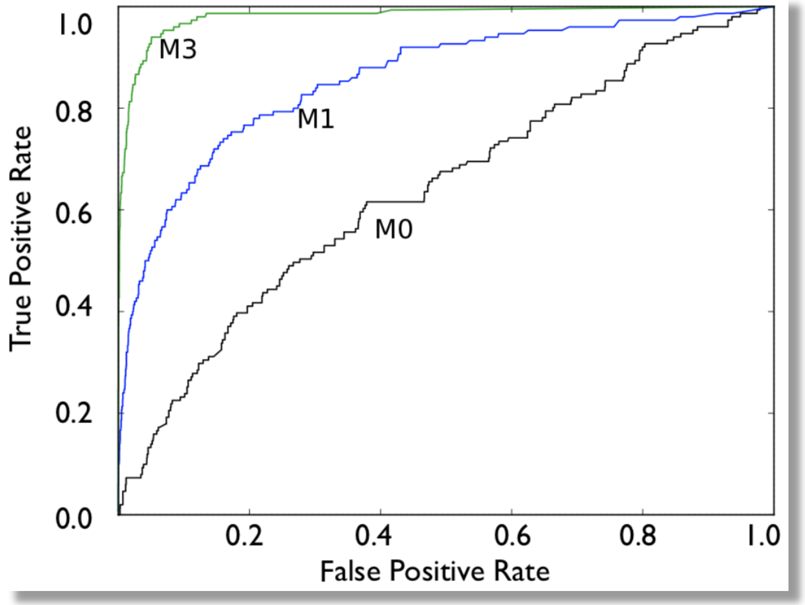
Computational prediction of CRMs using training data


Training CRMs

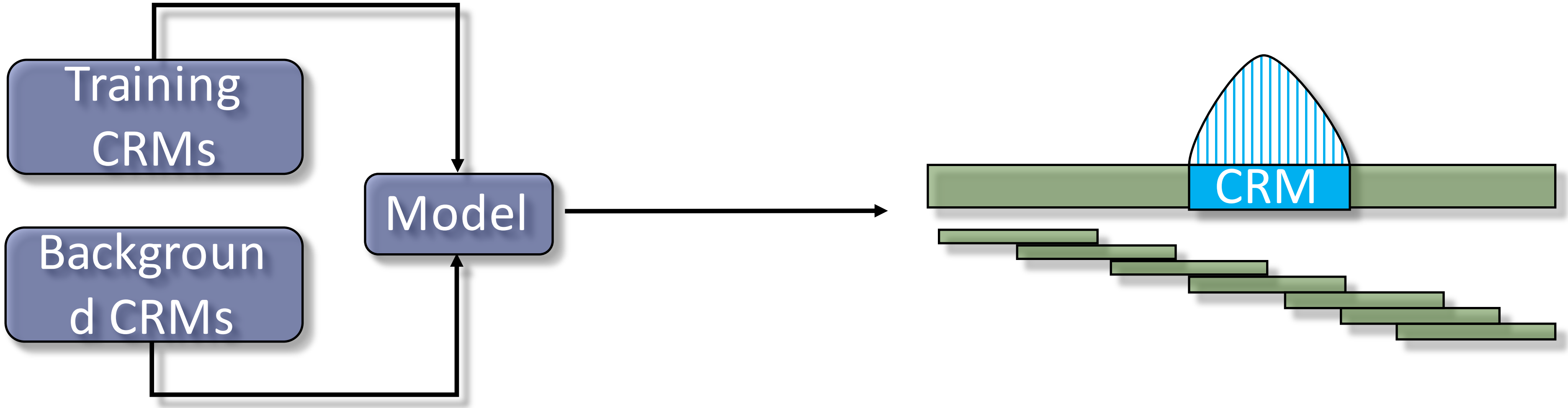
Background CRMs

Model


Cross Validation



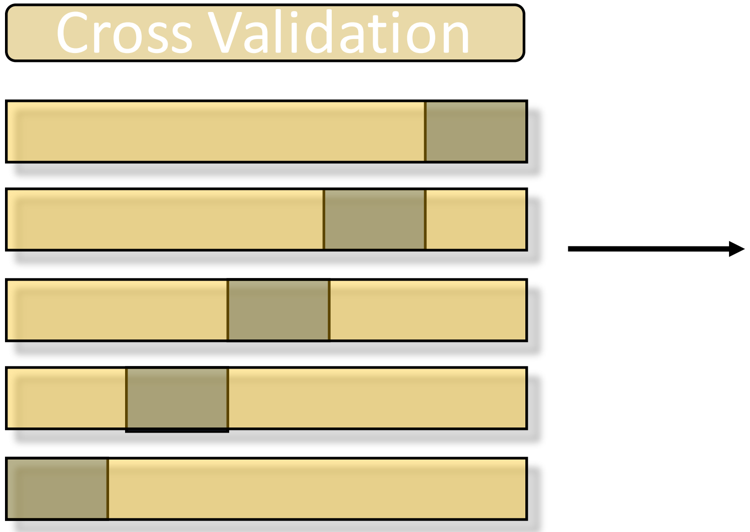
Computational prediction of CRMs using training data



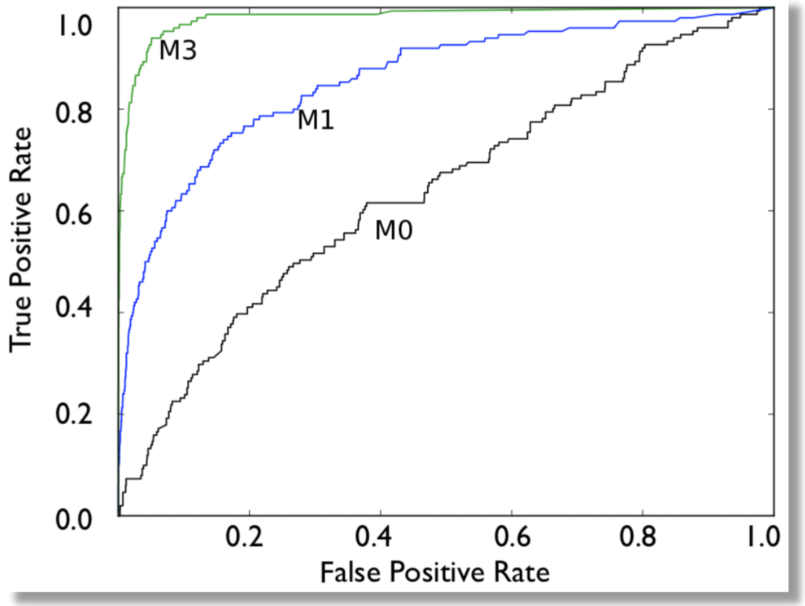
Training CRMs

Backgroun
d CRMs

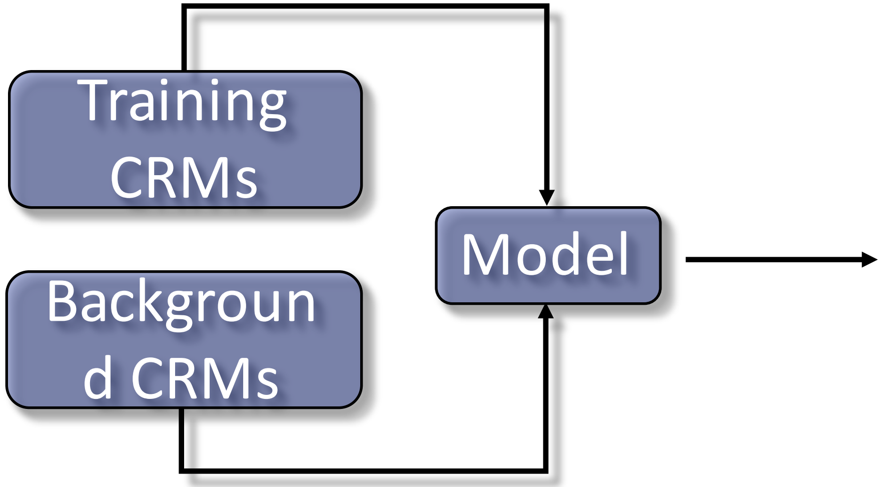
Model



Cross Validation



Computational prediction of CRMs using training data



Training CRMs

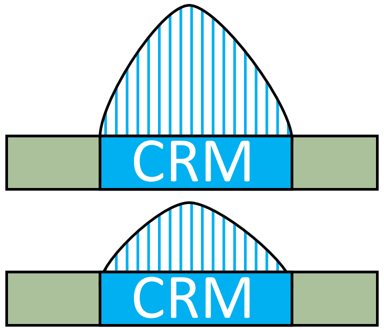
Backgroun
d CRMs

Model

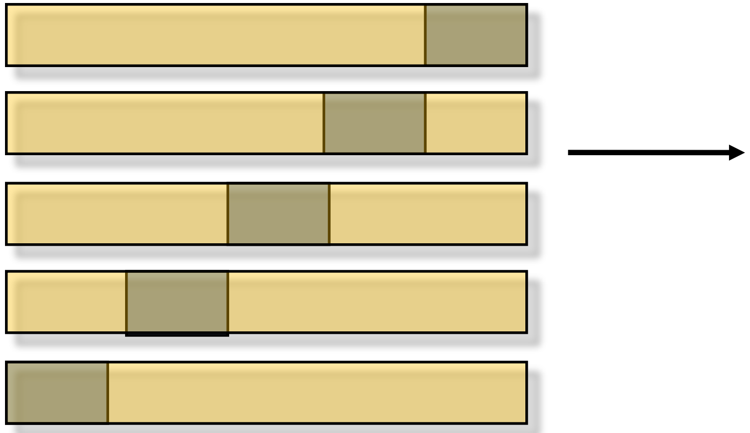
Impact of a mutation

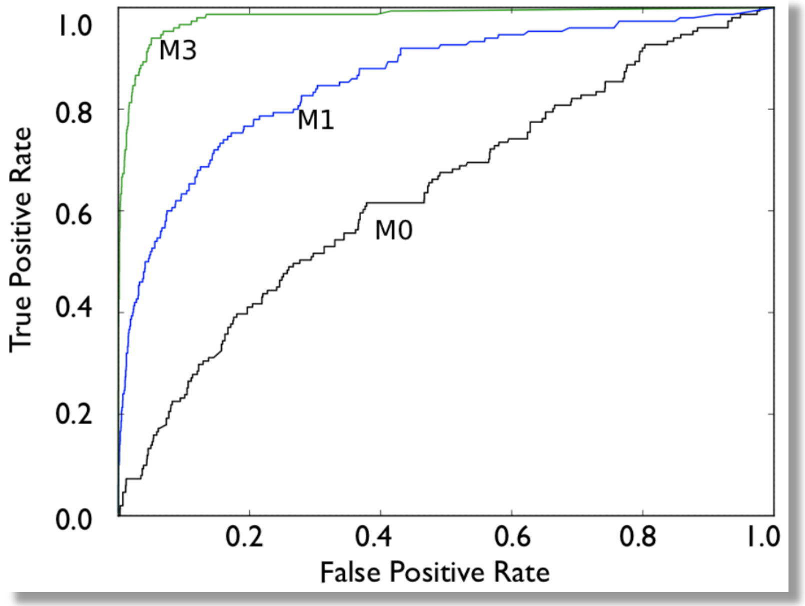
AGCTCACGTGCGTGCGTAGCGAATTGC

AGCTCACGTGCGTGCGTATCGAATTGC



Cross Validation

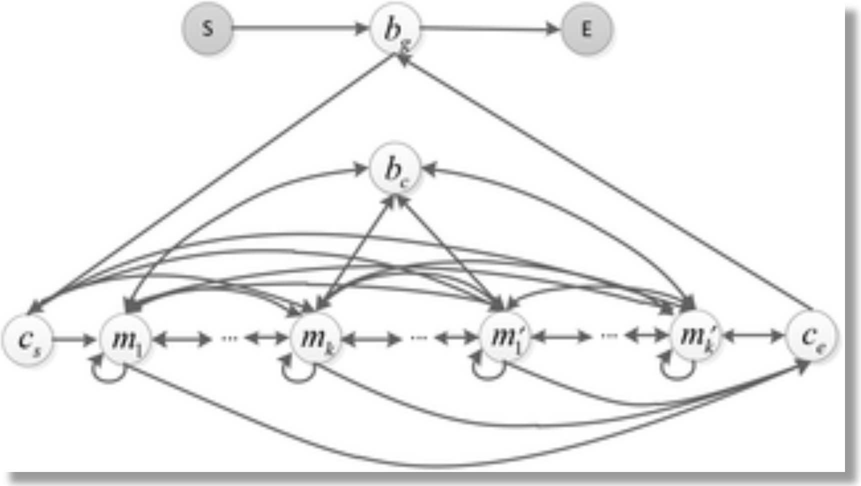




CRM prediction using machine learning methods

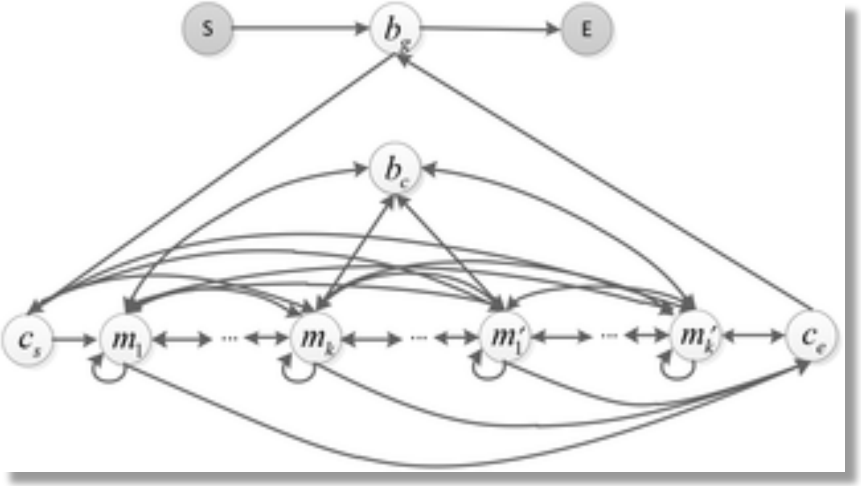
CRM prediction using machine learning methods

HMM

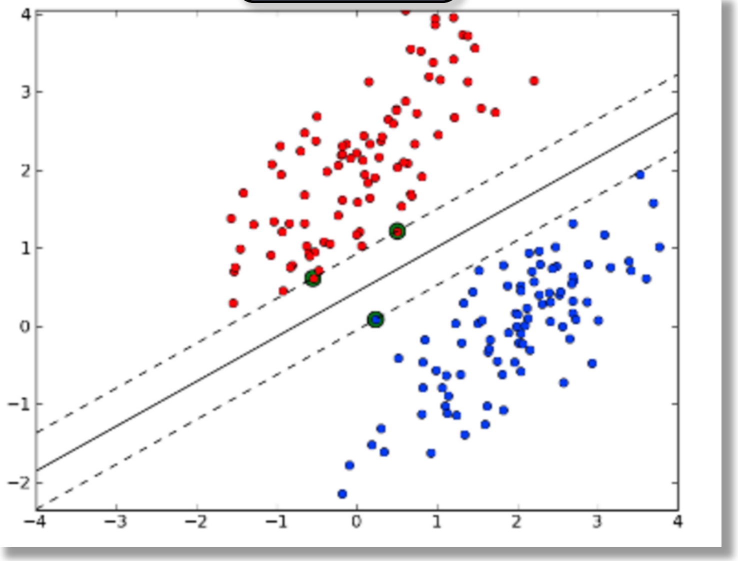


CRM prediction using machine learning methods

HMM

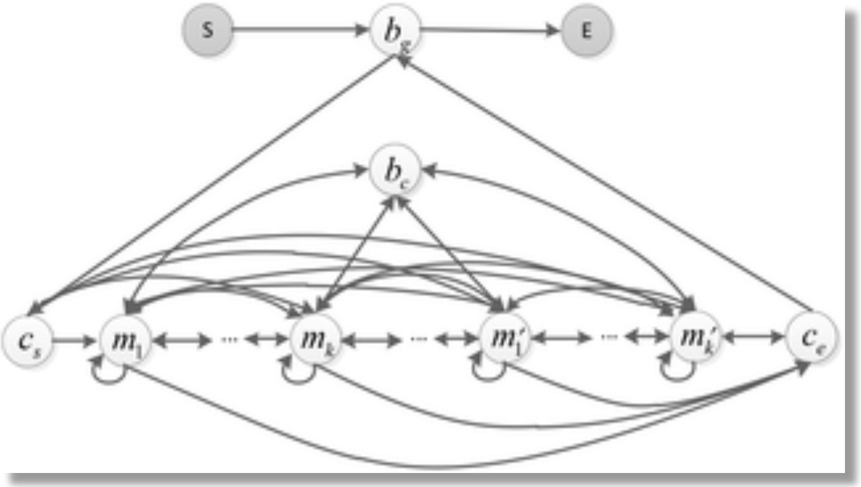


SVM

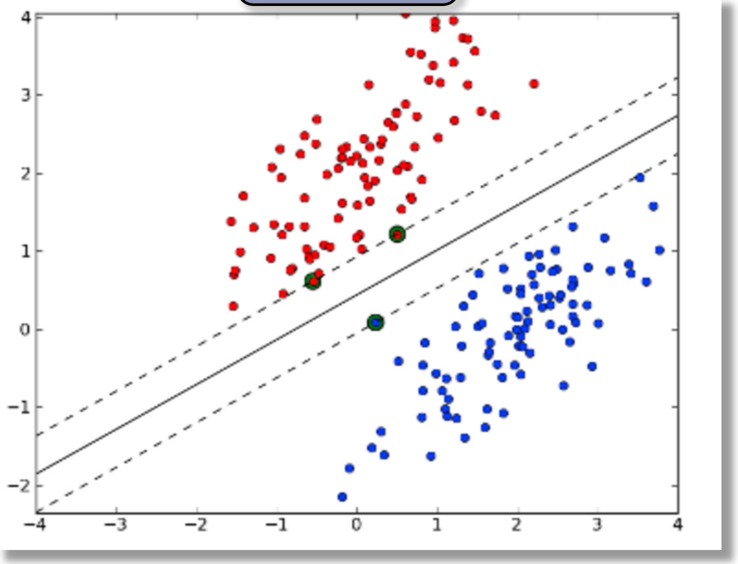


CRM prediction using machine learning methods

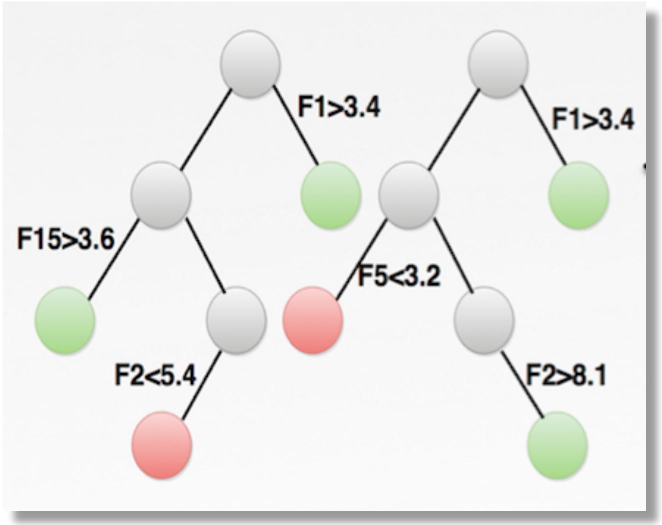
HMM



SVM

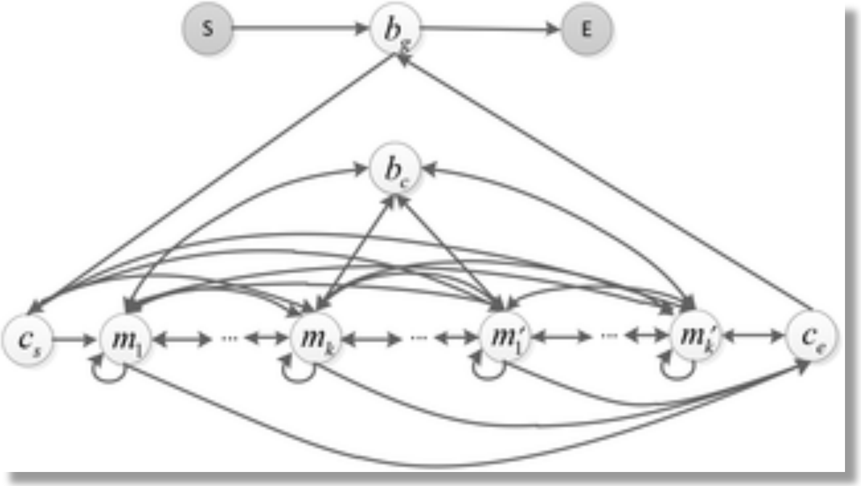


RF

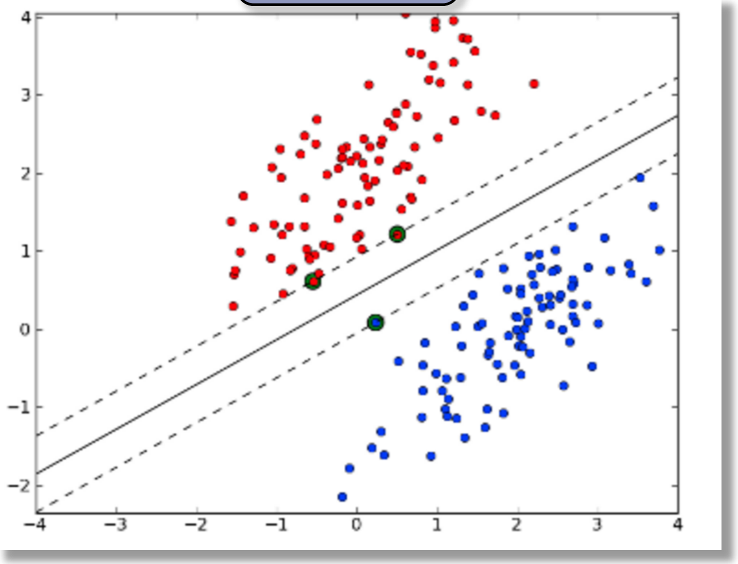


CRM prediction using machine learning methods

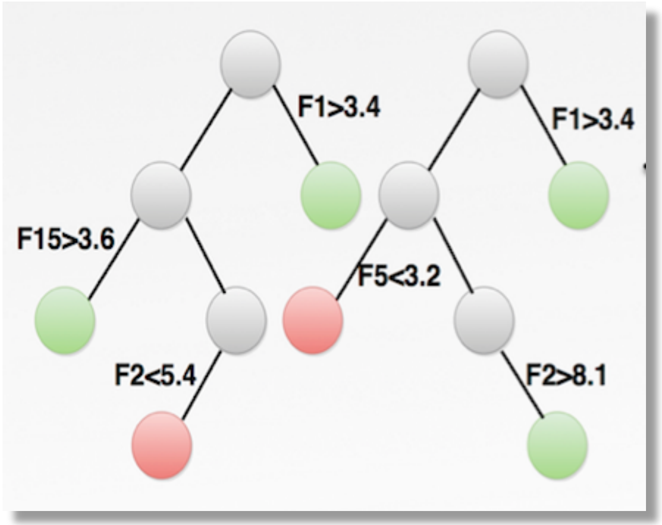
HMM



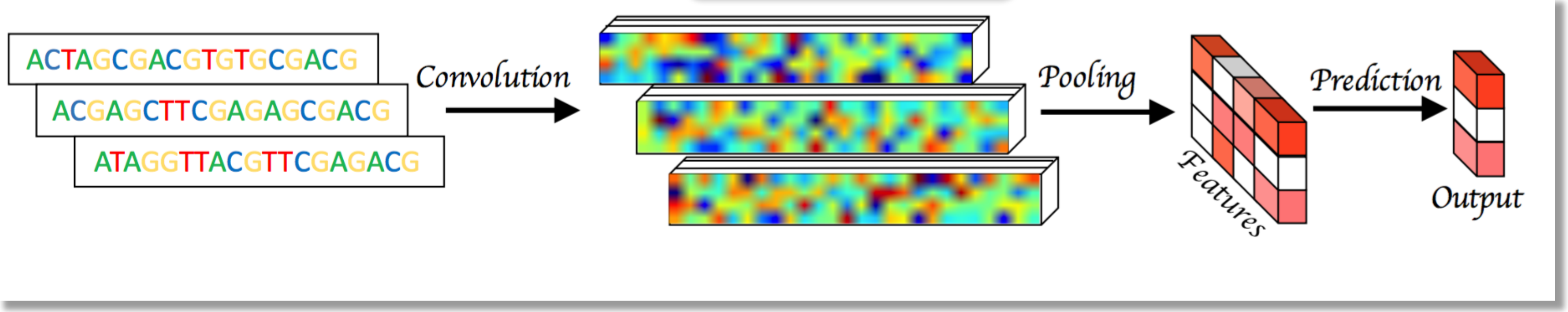
SVM



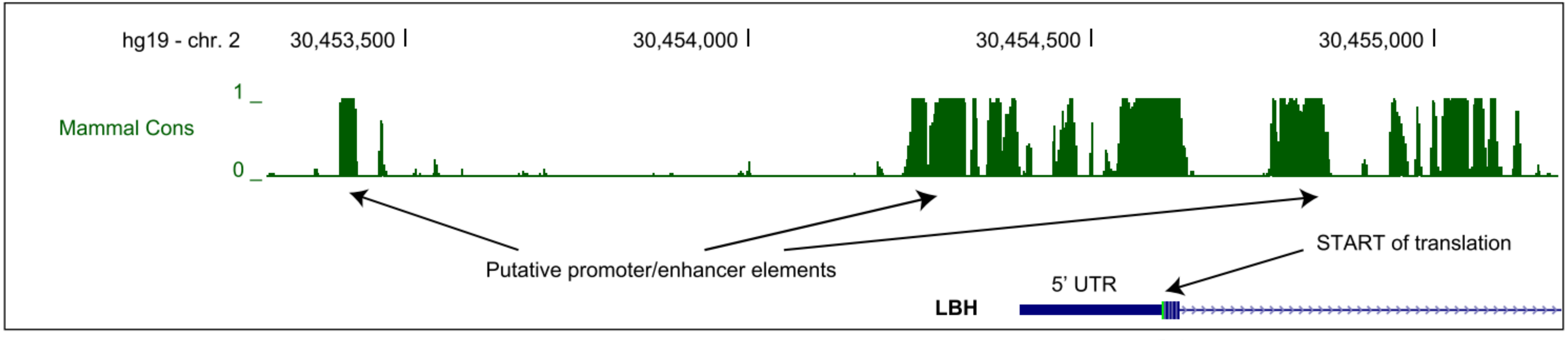
RF



Deep Learning

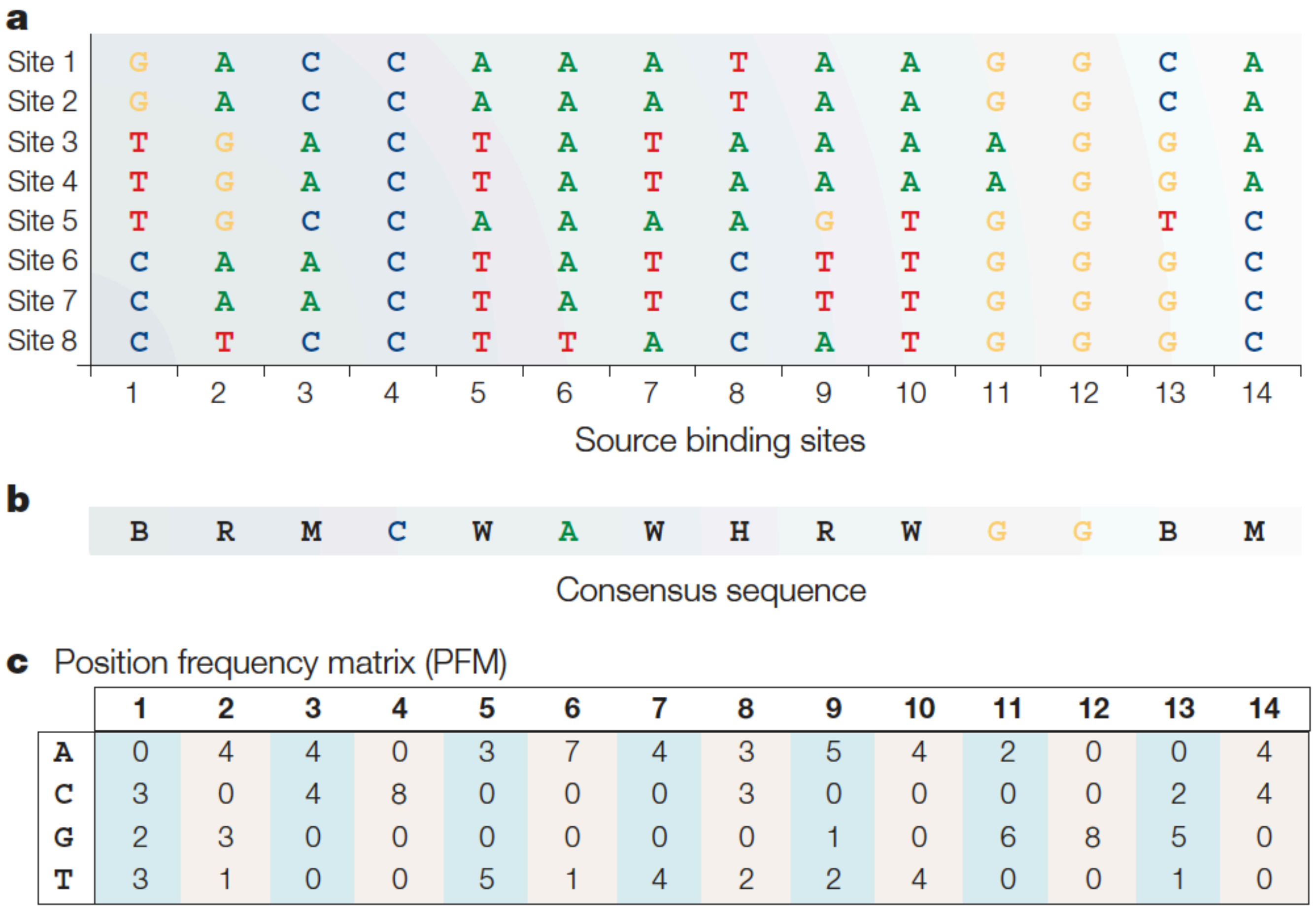


Conservation to find CRMs



Jessica Alfoldi et al. Genome Research 2013

Models for predicting transcription-factor binding sites



a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

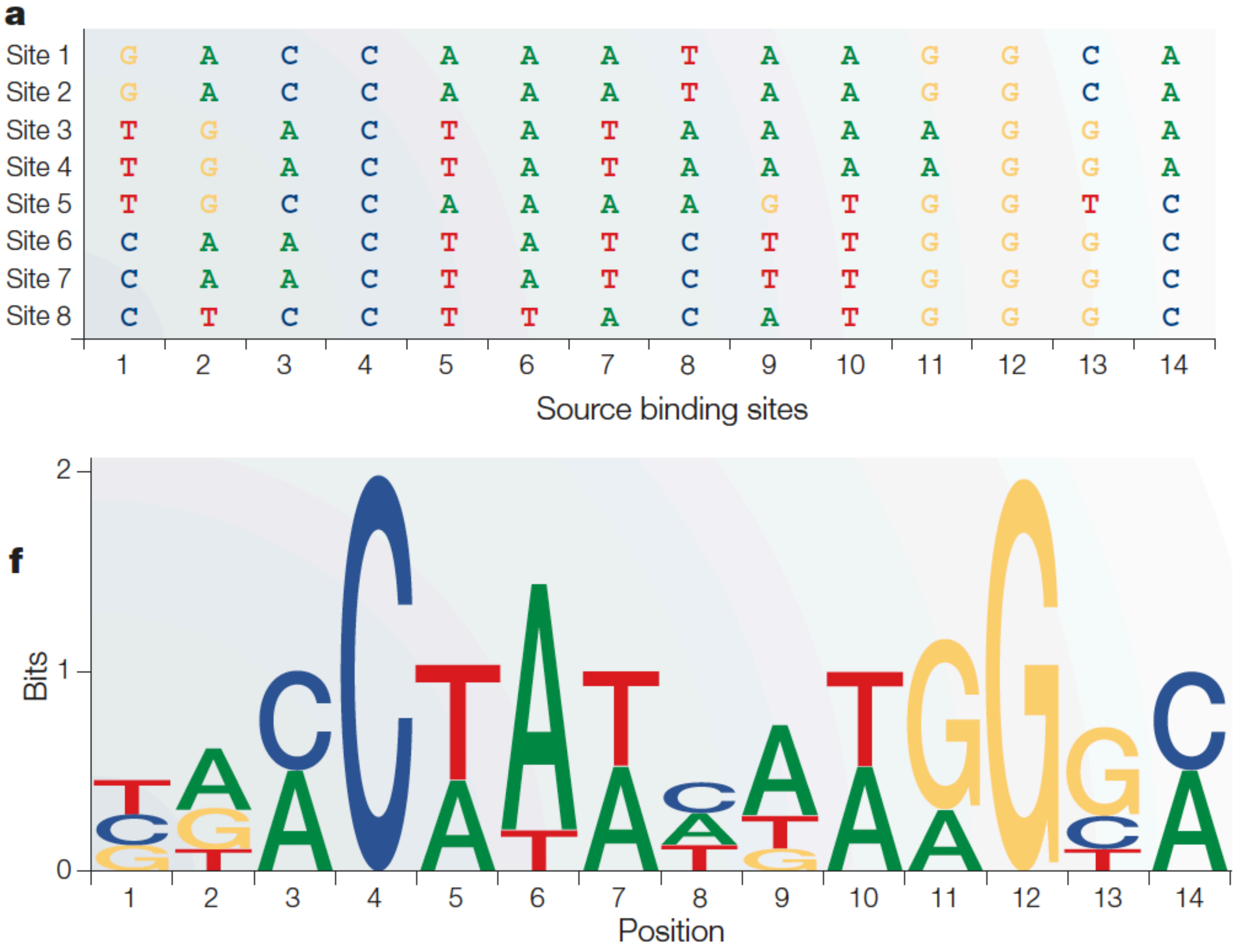
Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Wasserman et al. Nature Reviews 2007

Models for predicting transcription-factor binding sites



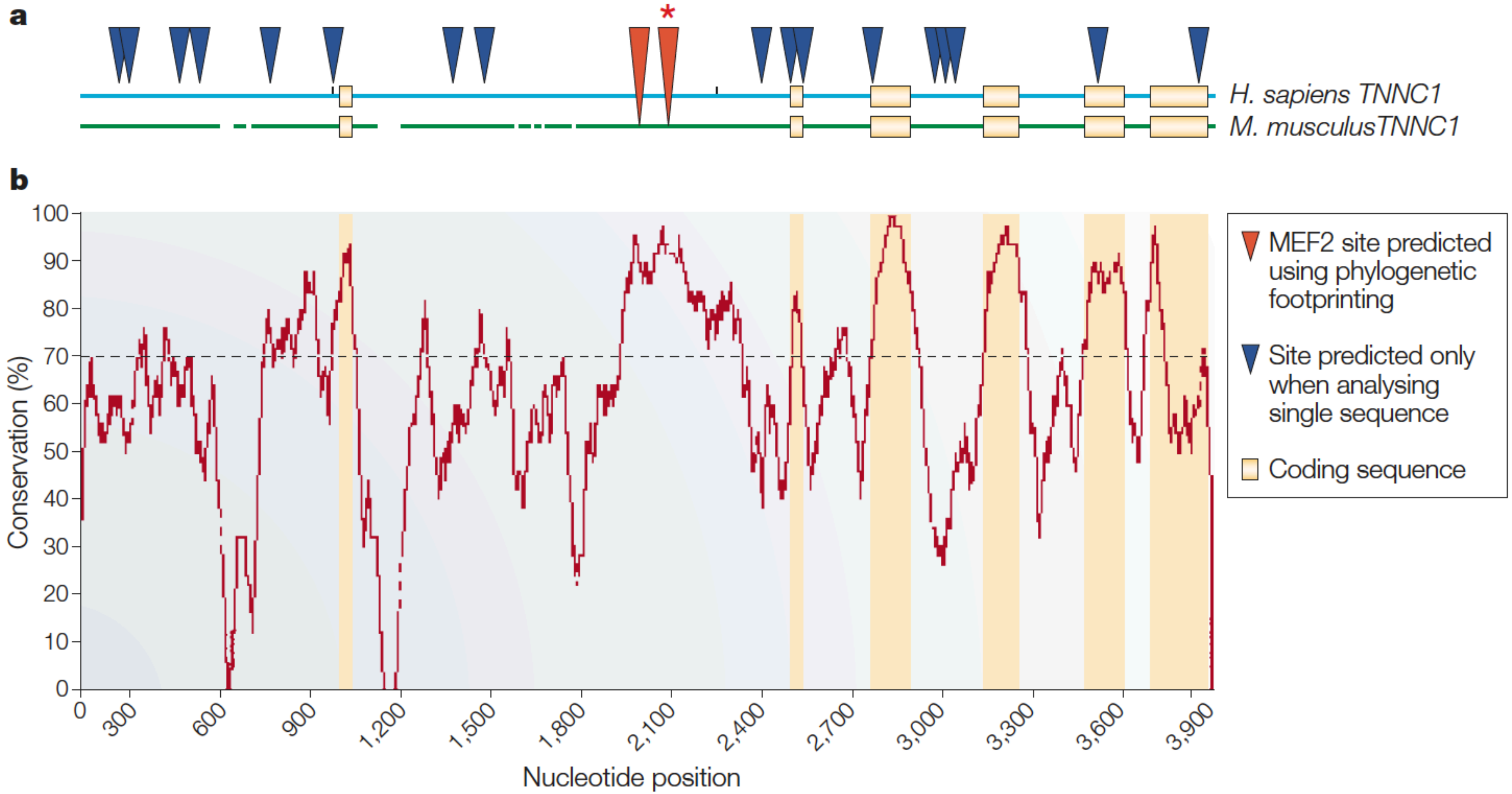
a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

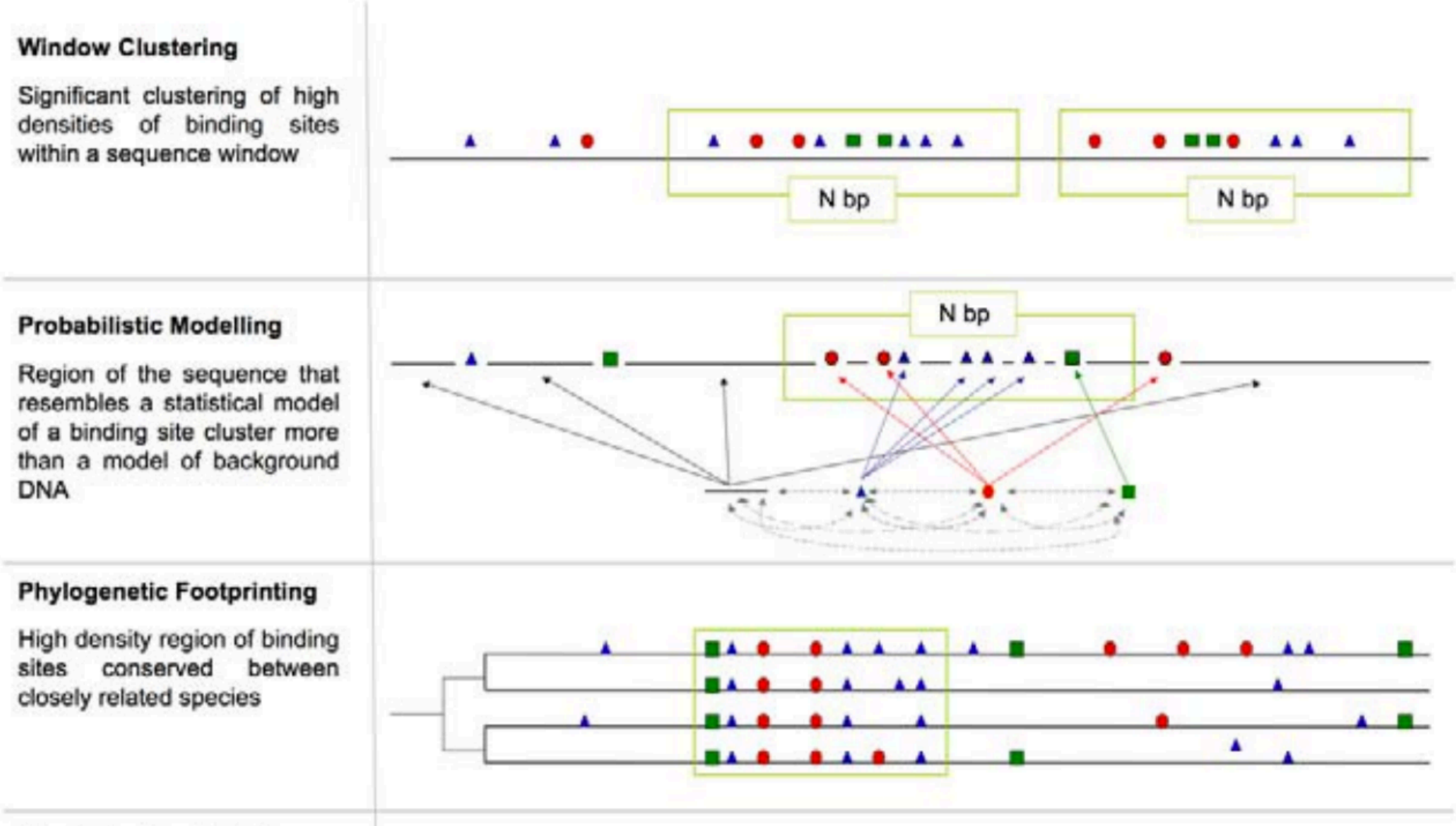
Wasserman et al. Nature Reviews 2007

Models for predicting transcription-factor by Coupling binding-site prediction with phylogenetic footprinting



Wasserman et al. Nature Reviews 2007

Models using clusters of binding sites



Window Clustering

Significant clustering of high densities of binding sites within a sequence window

Probabilistic Modelling

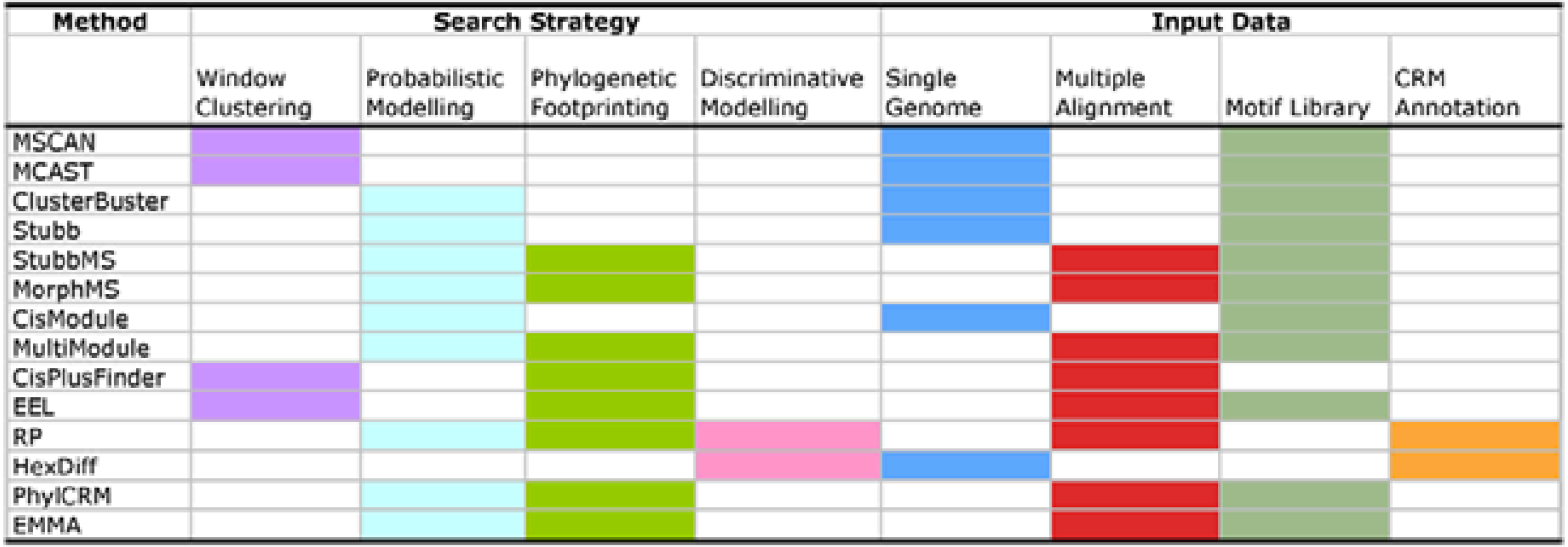
Region of the sequence that resembles a statistical model of a binding site cluster more than a model of background DNA

Phylogenetic Footprinting

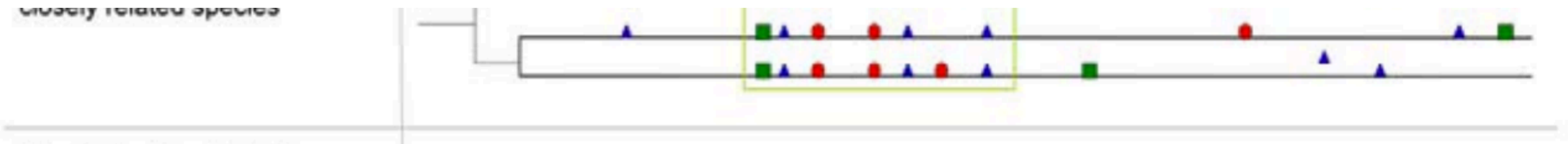
High density region of binding sites conserved between closely related species

Su et al. PLoS Computational Biology 2010

Models using clusters of binding sites



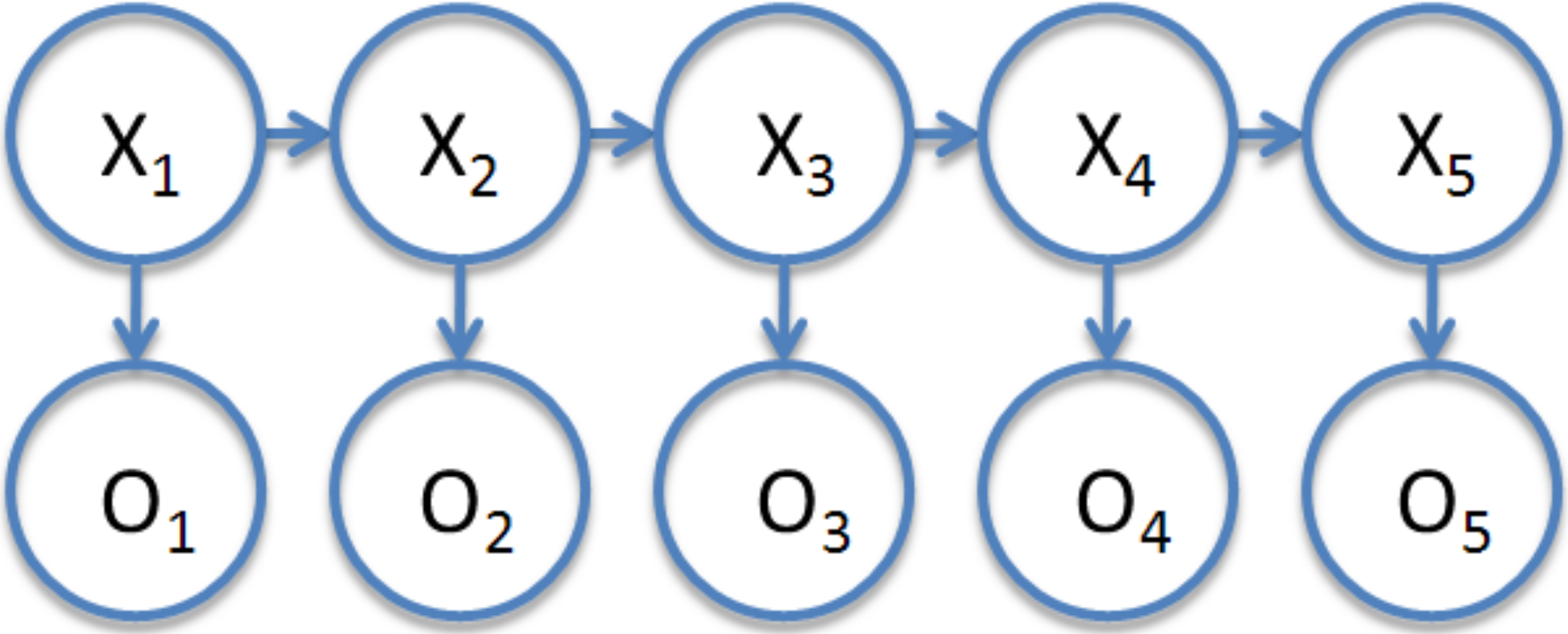
Method	Search Strategy				Input Data			
	Window Clustering	Probabilistic Modelling	Phylogenetic Footprinting	Discriminative Modelling	Single Genome	Multiple Alignment	Motif Library	CRM Annotation
MSCAN	■				■		■	
MCAST	■				■		■	
ClusterBuster		■			■		■	
Stubb		■			■		■	
StubbMS		■	■			■	■	
MorphMS		■	■			■	■	
CisModule		■			■		■	
MultiModule		■	■			■	■	
CisPlusFinder	■		■			■		
EEL	■		■			■	■	
RP		■	■	■		■		■
HexDiff				■	■			■
PhylCRM		■	■			■	■	
EMMA		■	■			■	■	



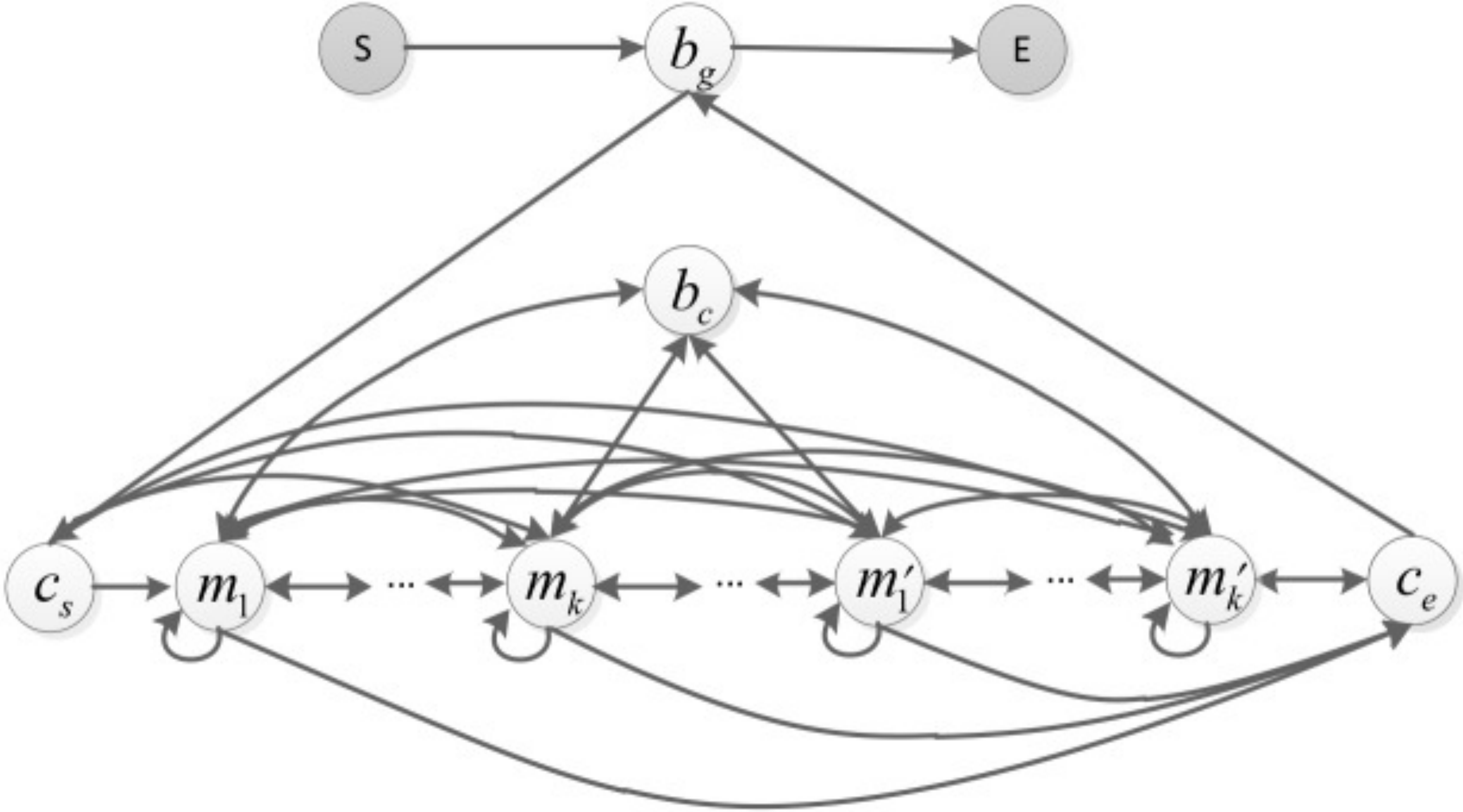
Su et al. PLoS Computational Biology 2010

Models using clusters of binding sites

HMM

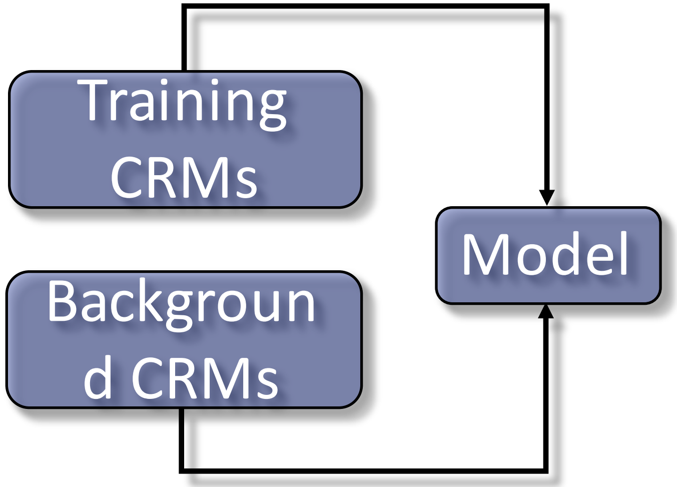


Models using clusters of binding sites



Guo et al.PLoS One 2016

Computational prediction of CRMs using training data



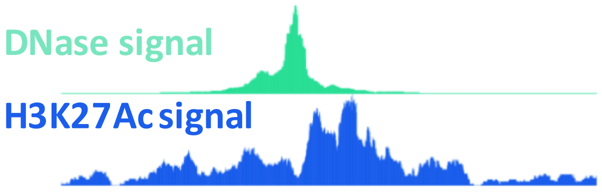
Training CRMs

Background CRMs

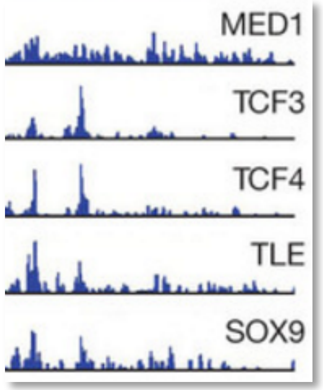
Model

Training CRMs

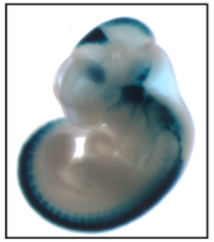
Chromatin profiling



TF ChIP-seq

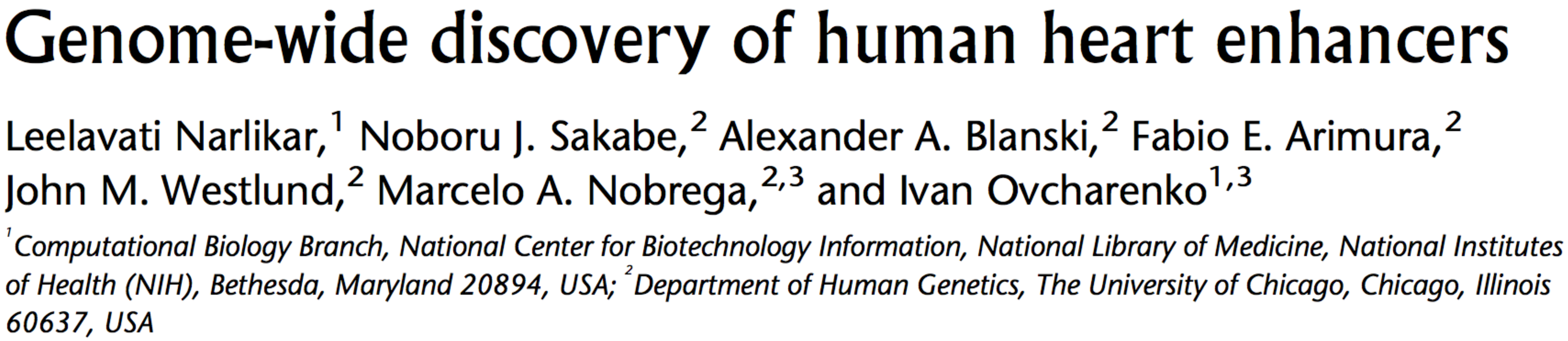


In vivo CRM activit



https://www.encodeproject.org/data/annotations/

Case 1

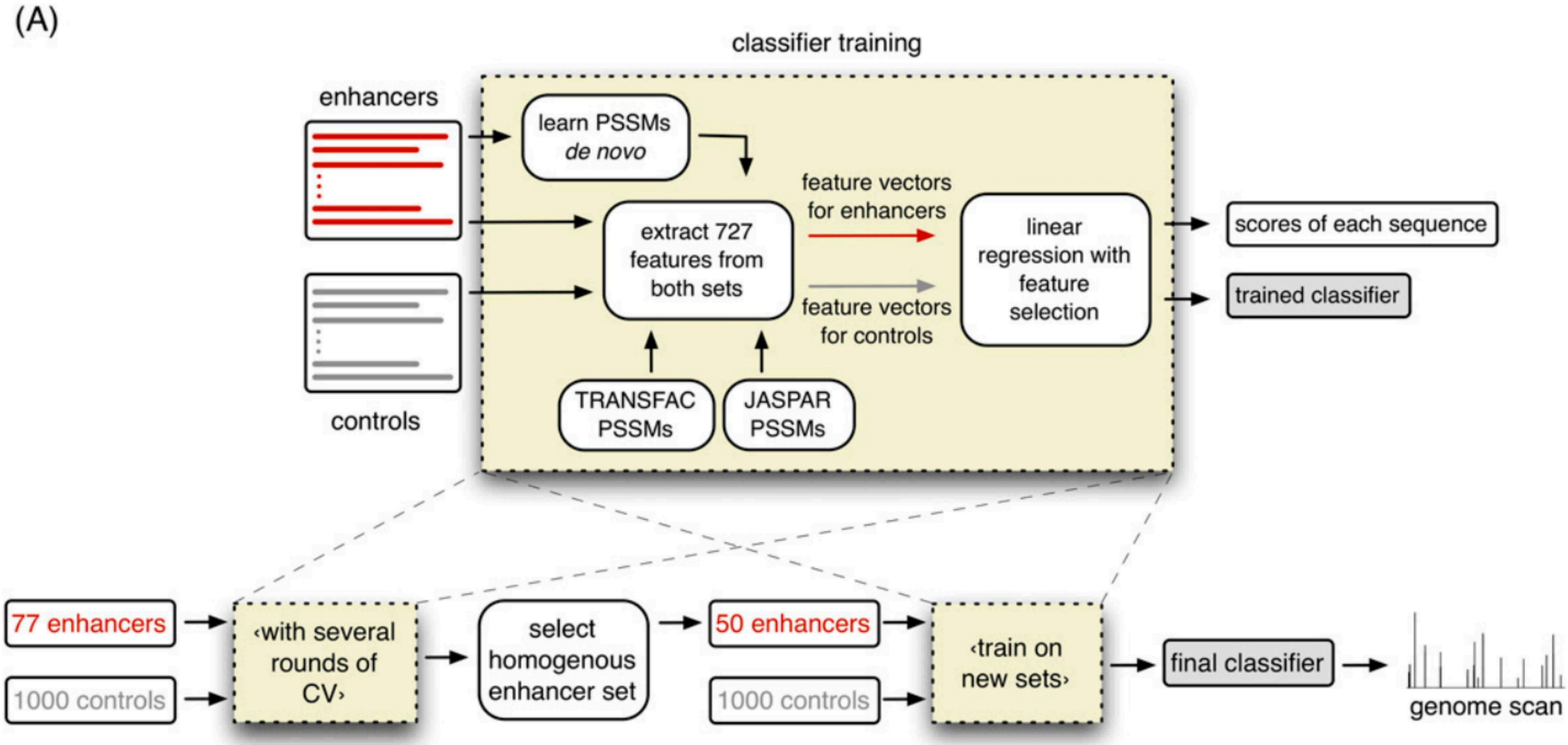


Genome-wide discovery of human heart enhancers

Leelavati Narlikar,[1] Noboru J. Sakabe,[2] Alexander A. Blanski,[2] Fabio E. Arimura,[2] John M. Westlund,[2] Marcelo A. Nobrega,[2,3] and Ivan Ovcharenko[1,3]

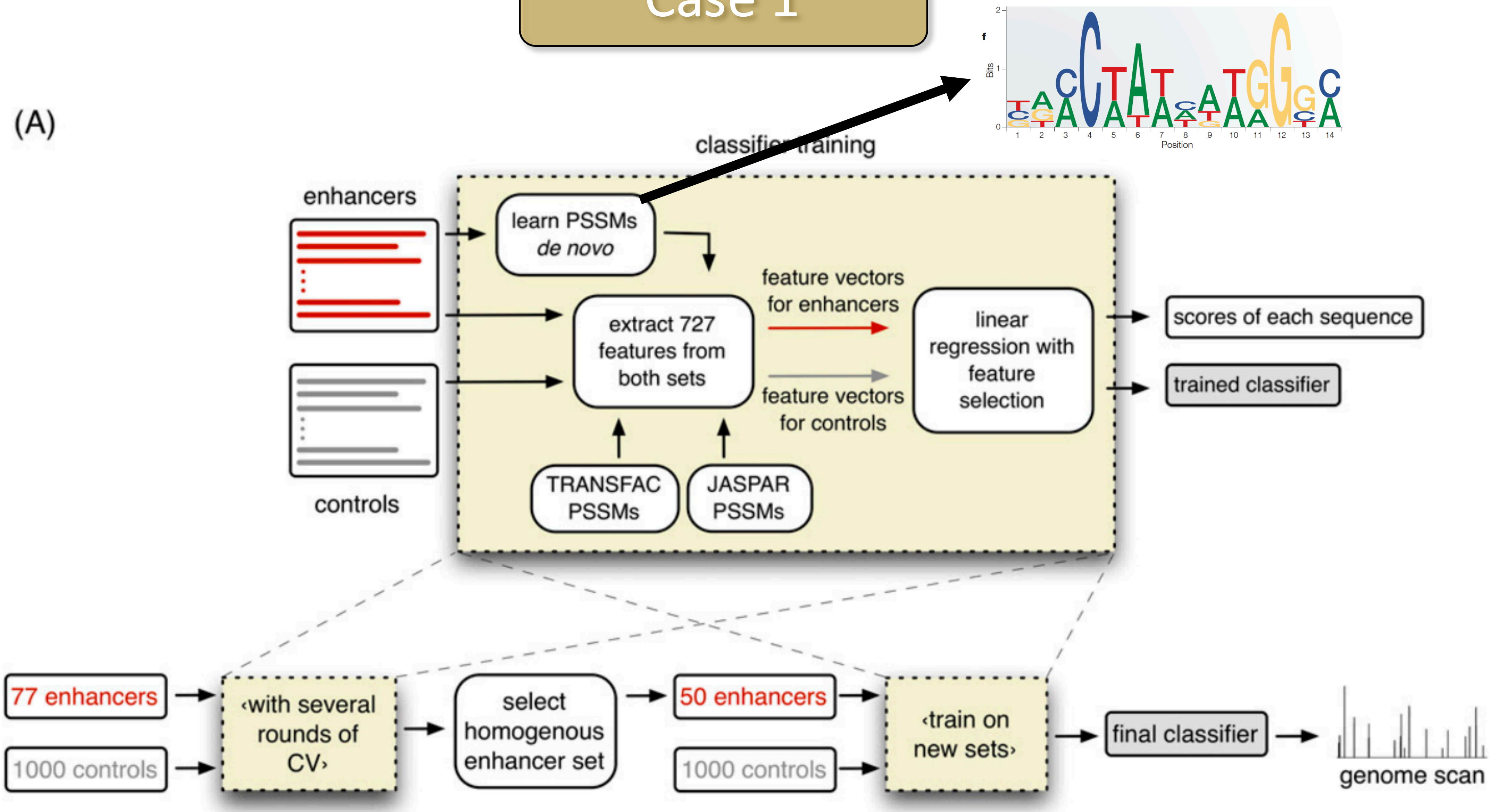
[1] Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, Maryland 20894, USA; [2] Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA

Case 1



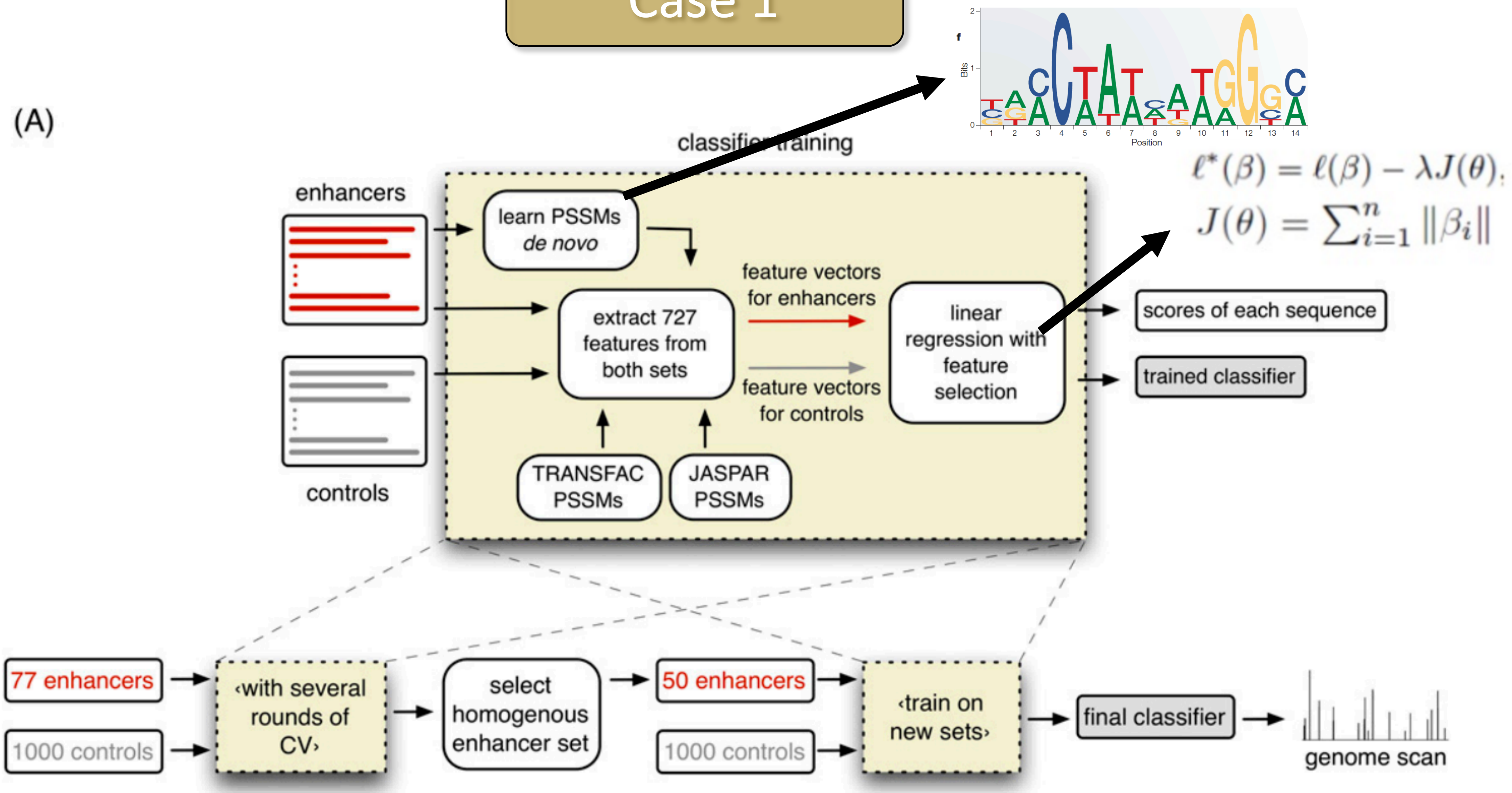
Narlikar et al.

Case 1



Narlikar et al.

Case 1



(A)

classifier training

enhancers

controls

learn PSSMs *de novo*

extract 727 features from both sets

TRANSFAC PSSMs

JASPAR PSSMs

feature vectors for enhancers

feature vectors for controls

linear regression with feature selection

scores of each sequence

trained classifier

77 enhancers

1000 controls

‹with several rounds of CV›

select homogenous enhancer set

50 enhancers

1000 controls

‹train on new sets›

final classifier

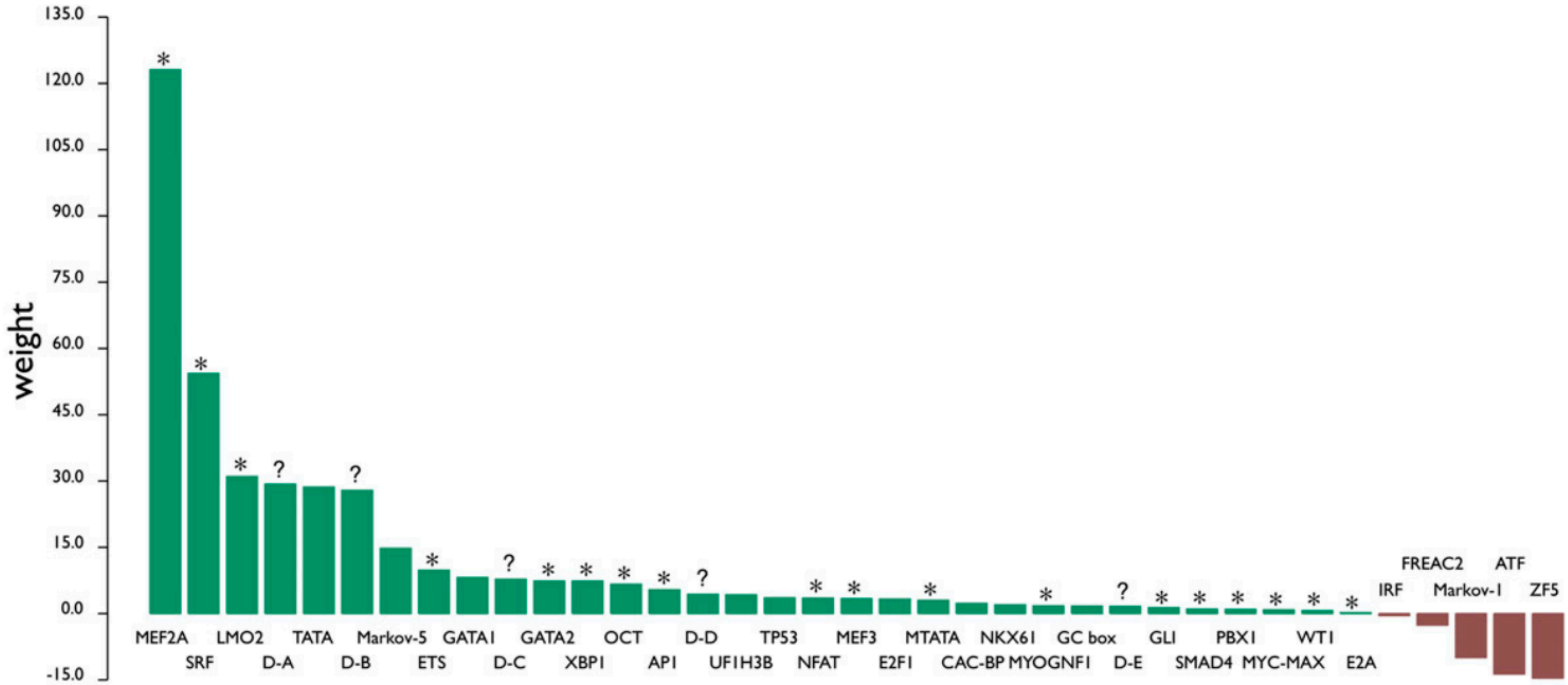
genome scan

$$\ell^*(\beta) = \ell(\beta) - \lambda J(\theta),$$

$$J(\theta) = \sum_{i=1}^{n} \|\beta_i\|$$

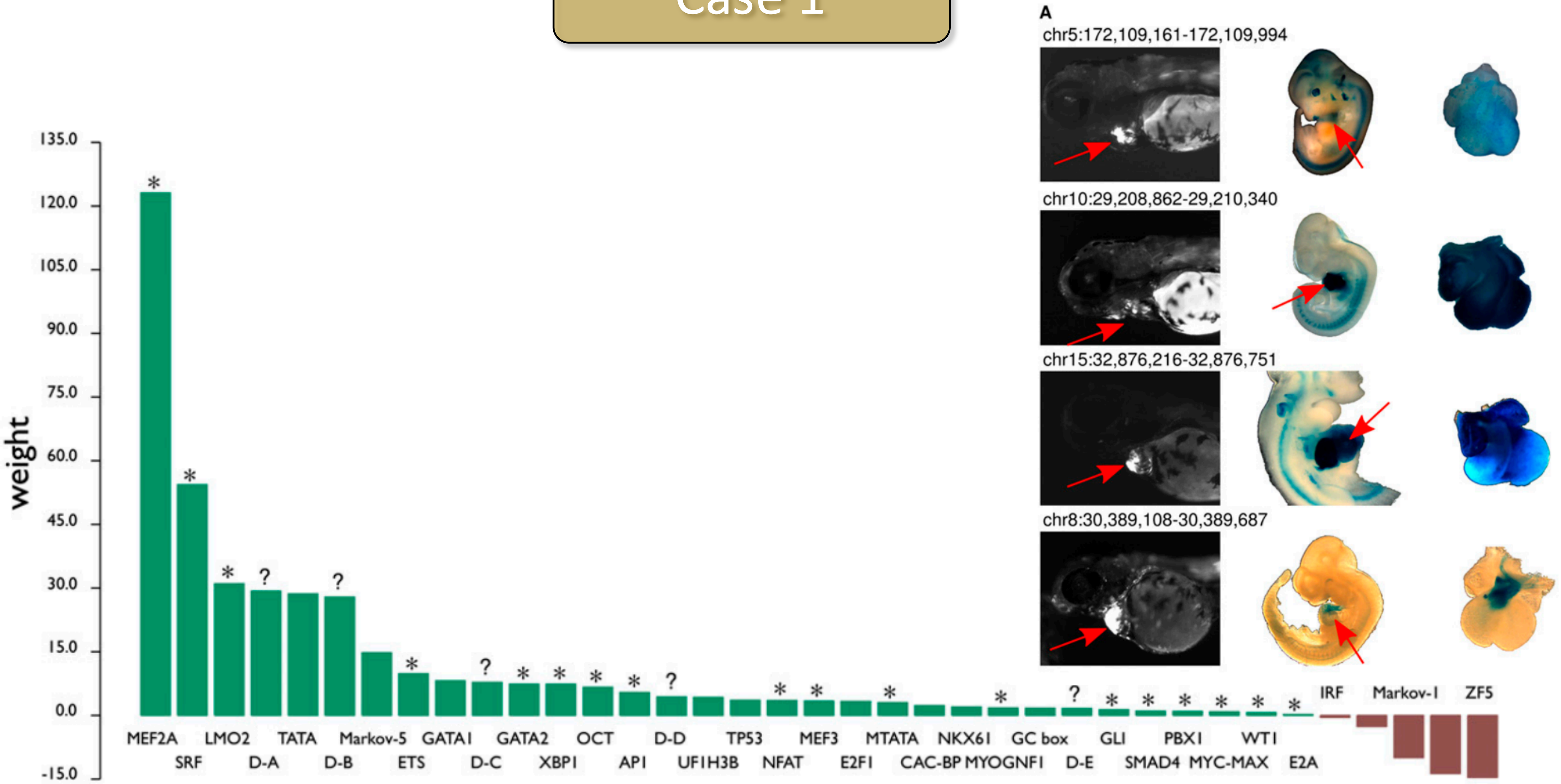
Narlikar et al.

Case 1



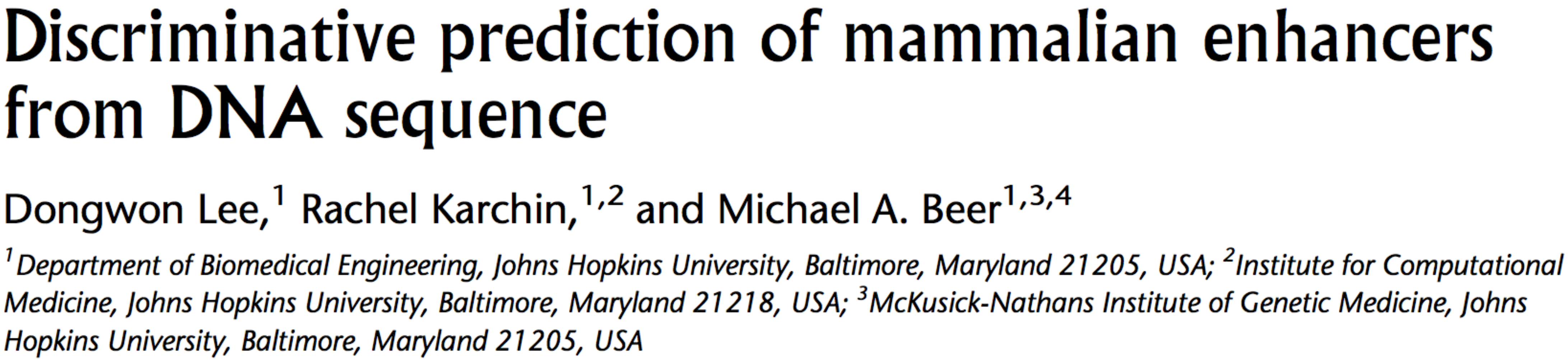
Narlikar et al.

Case 1



Narlikar et al.

Case 2



Discriminative prediction of mammalian enhancers from DNA sequence

Dongwon Lee,[1] Rachel Karchin,[1,2] and Michael A. Beer[1,3,4]

[1] *Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21205, USA;* [2] *Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [3] *McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA*

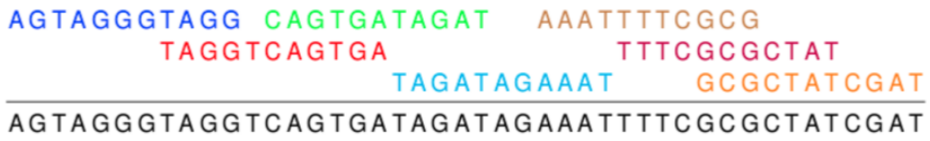
Case 2

Discriminative prediction of mammalian enhancers from DNA sequence

Dongwon Lee,[1] Rachel Karchin,[1,2] and Michael A. Beer[1,3,4]

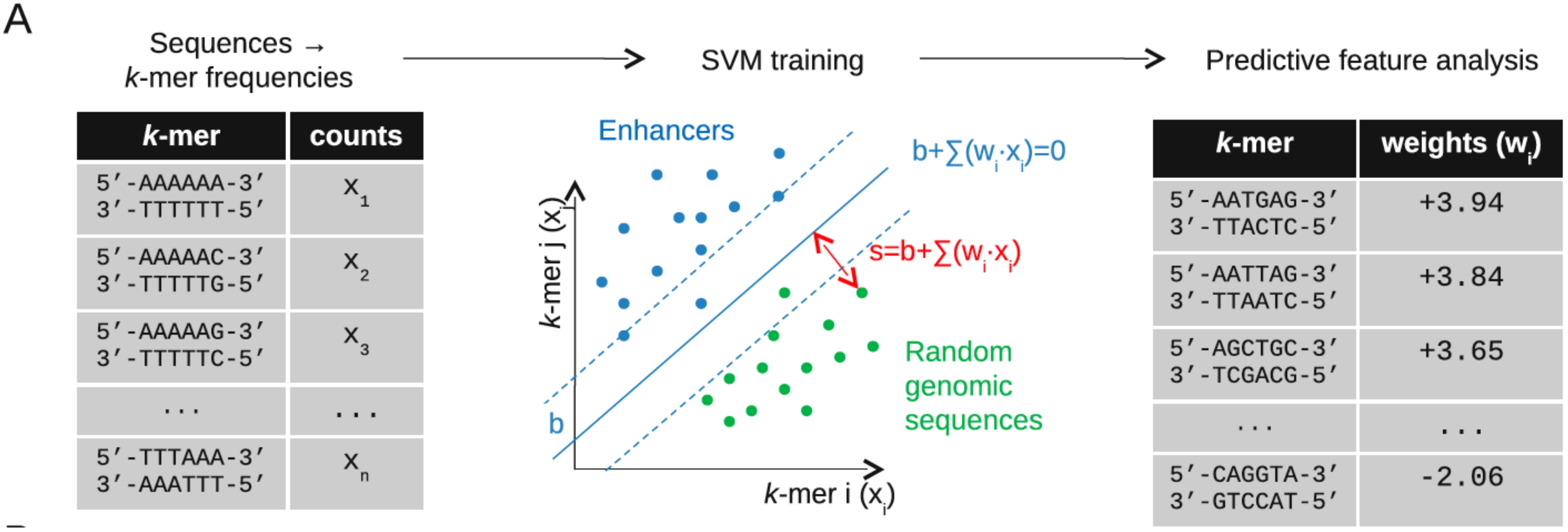
[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21205, USA; [2]Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland 21218, USA; [3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA

String kernel



```
AGTAGGGTAGG CAGTGATAGAT  AAATTTTCGCG
       TAGGTCAGTGA           TTTCGCGCTAT
                  TAGATAGAAAT    GCGCTATCGAT
--------------------------------------------
AGTAGGGTAGGTCAGTGATAGATAGAAATTTTCGCGCTATCGAT
```

Case 2



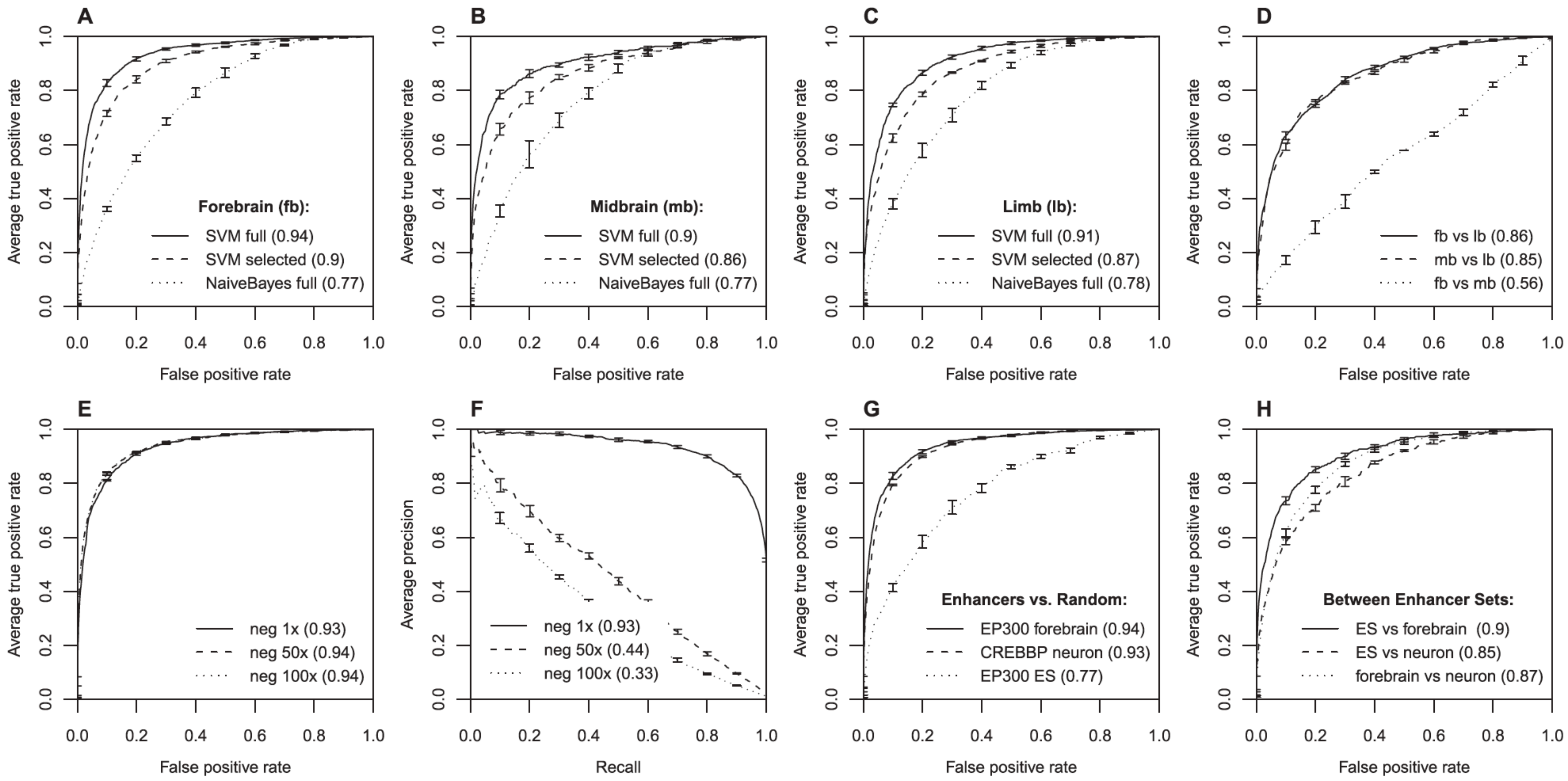
Sequences → *k*-mer frequencies → SVM training → Predictive feature analysis

k-mer	counts
5′-AAAAAA-3′ 3′-TTTTTT-5′	X_1
5′-AAAAAC-3′ 3′-TTTTTG-5′	X_2
5′-AAAAAG-3′ 3′-TTTTTC-5′	X_3
. . .	. . .
5′-TTTAAA-3′ 3′-AAATTT-5′	X_n

k-mer	weights (w_i)
5′-AATGAG-3′ 3′-TTACTC-5′	+3.94
5′-AATTAG-3′ 3′-TTAATC-5′	+3.84
5′-AGCTGC-3′ 3′-TCGACG-5′	+3.65
. . .	. . .
5′-CAGGTA-3′ 3′-GTCCAT-5′	-2.06

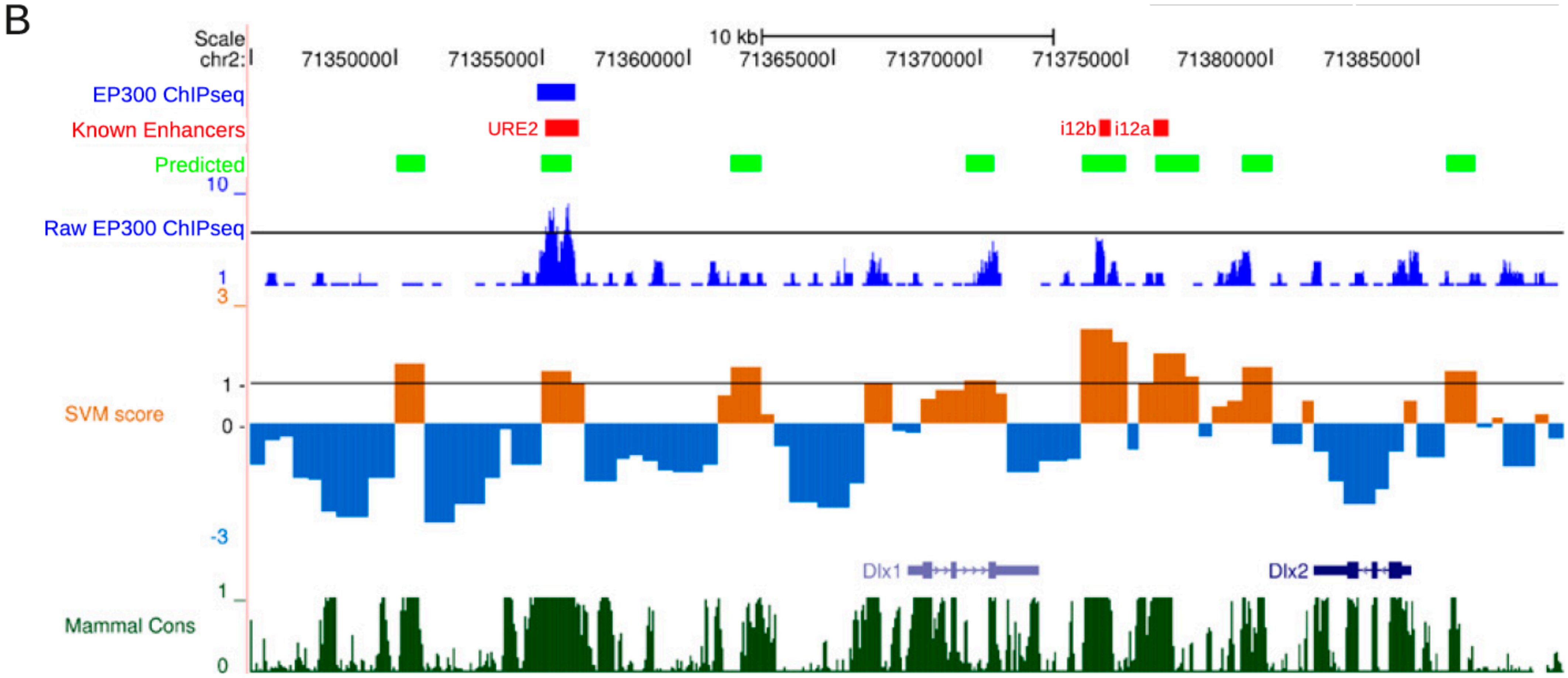
Lee et al.

Case 2



Lee et al.

Case 2



Lee et al.

Case 2

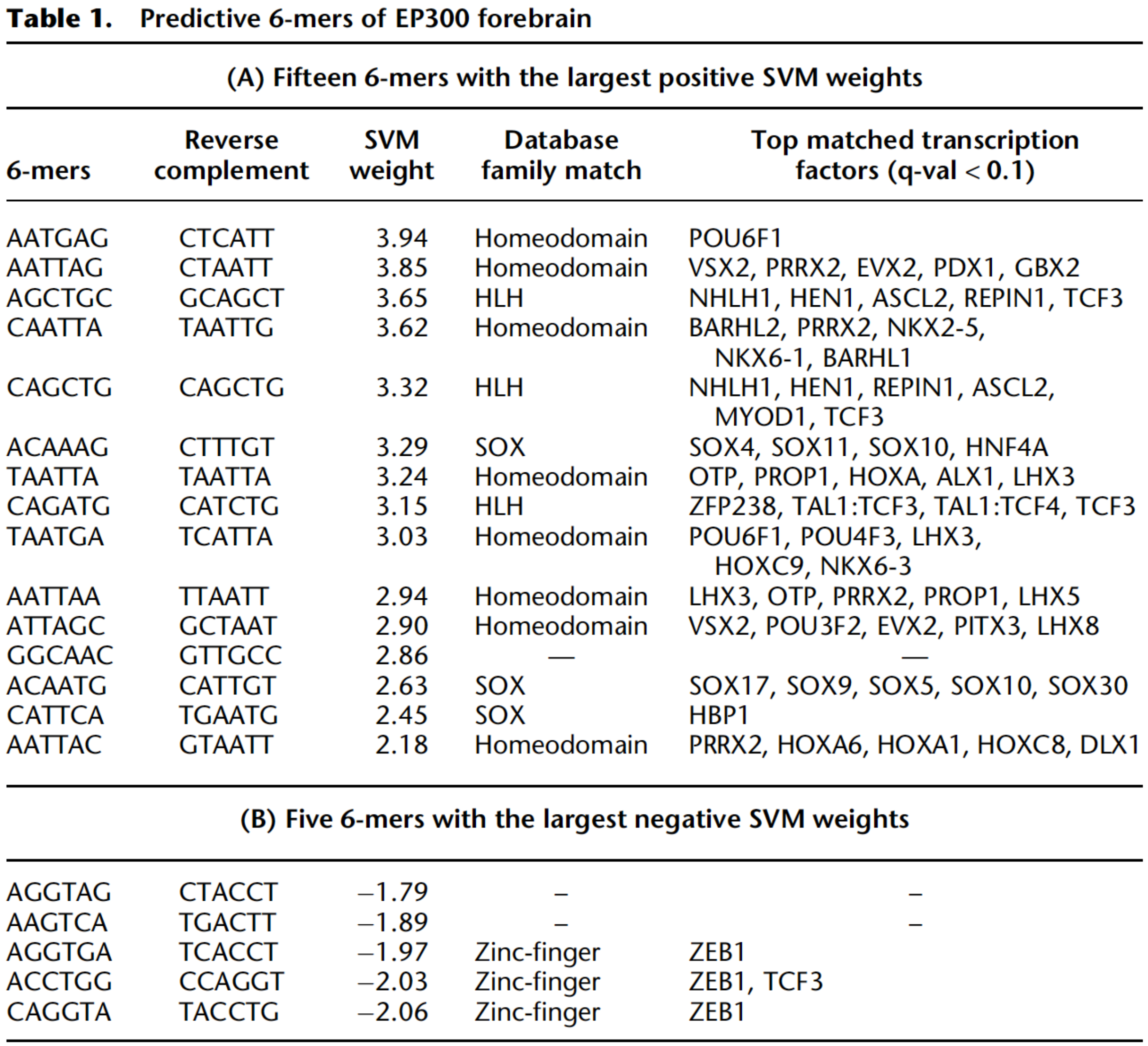
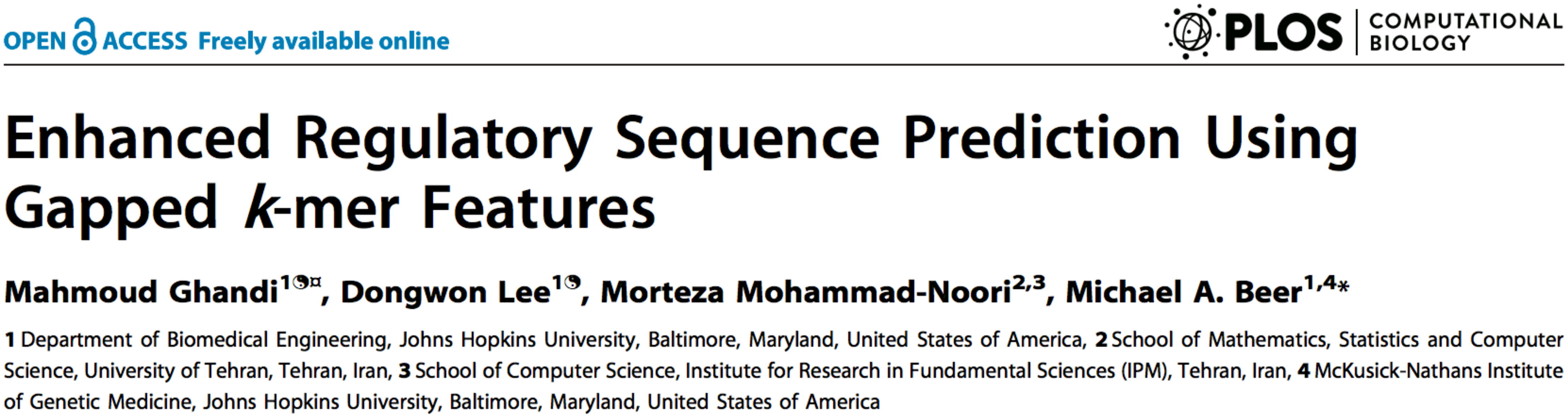


Table 1. Predictive 6-mers of EP300 forebrain

6-mers	Reverse complement	SVM weight	Database family match	Top matched transcription factors (q-val < 0.1)
(A) Fifteen 6-mers with the largest positive SVM weights				
AATGAG	CTCATT	3.94	Homeodomain	POU6F1
AATTAG	CTAATT	3.85	Homeodomain	VSX2, PRRX2, EVX2, PDX1, GBX2
AGCTGC	GCAGCT	3.65	HLH	NHLH1, HEN1, ASCL2, REPIN1, TCF3
CAATTA	TAATTG	3.62	Homeodomain	BARHL2, PRRX2, NKX2-5, NKX6-1, BARHL1
CAGCTG	CAGCTG	3.32	HLH	NHLH1, HEN1, REPIN1, ASCL2, MYOD1, TCF3
ACAAAG	CTTTGT	3.29	SOX	SOX4, SOX11, SOX10, HNF4A
TAATTA	TAATTA	3.24	Homeodomain	OTP, PROP1, HOXA, ALX1, LHX3
CAGATG	CATCTG	3.15	HLH	ZFP238, TAL1:TCF3, TAL1:TCF4, TCF3
TAATGA	TCATTA	3.03	Homeodomain	POU6F1, POU4F3, LHX3, HOXC9, NKX6-3
AATTAA	TTAATT	2.94	Homeodomain	LHX3, OTP, PRRX2, PROP1, LHX5
ATTAGC	GCTAAT	2.90	Homeodomain	VSX2, POU3F2, EVX2, PITX3, LHX8
GGCAAC	GTTGCC	2.86	—	—
ACAATG	CATTGT	2.63	SOX	SOX17, SOX9, SOX5, SOX10, SOX30
CATTCA	TGAATG	2.45	SOX	HBP1
AATTAC	GTAATT	2.18	Homeodomain	PRRX2, HOXA6, HOXA1, HOXC8, DLX1
(B) Five 6-mers with the largest negative SVM weights				
AGGTAG	CTACCT	−1.79	–	–
AAGTCA	TGACTT	−1.89	–	–
AGGTGA	TCACCT	−1.97	Zinc-finger	ZEB1
ACCTGG	CCAGGT	−2.03	Zinc-finger	ZEB1, TCF3
CAGGTA	TACCTG	−2.06	Zinc-finger	ZEB1

Lee et al.

Case 3



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Enhanced Regulatory Sequence Prediction Using Gapped *k*-mer Features

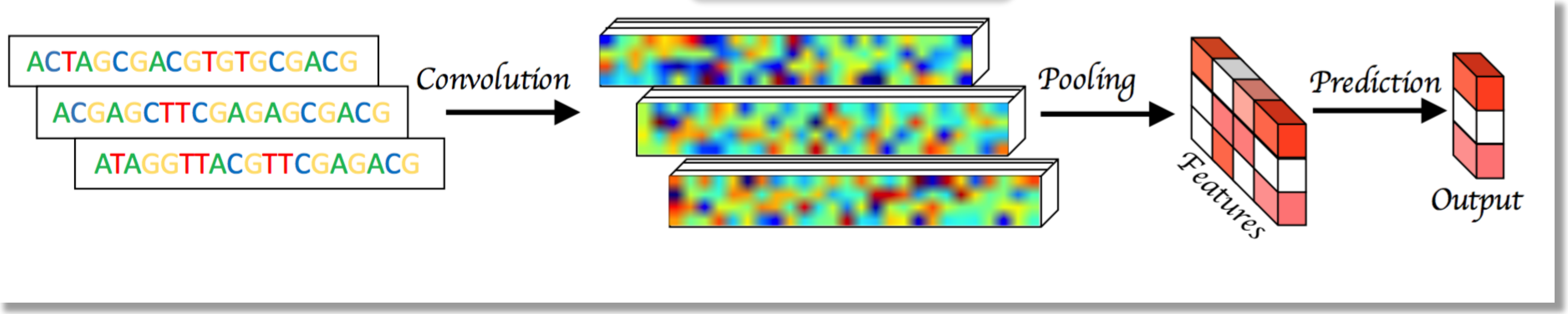
Mahmoud Ghandi[1☯¤], Dongwon Lee[1☯], Morteza Mohammad-Noori[2,3], Michael A. Beer[1,4*]

[1] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America, [2] School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran, [3] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran, [4] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America

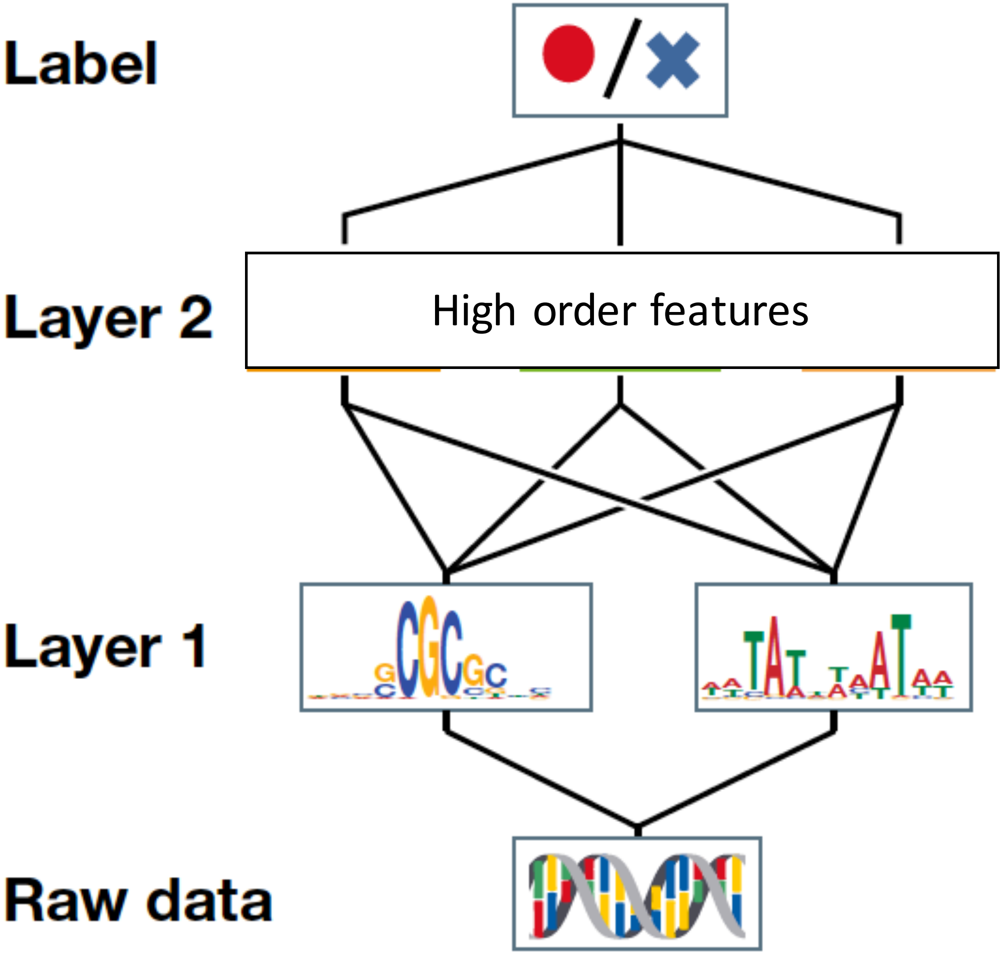
Ghandi et al.

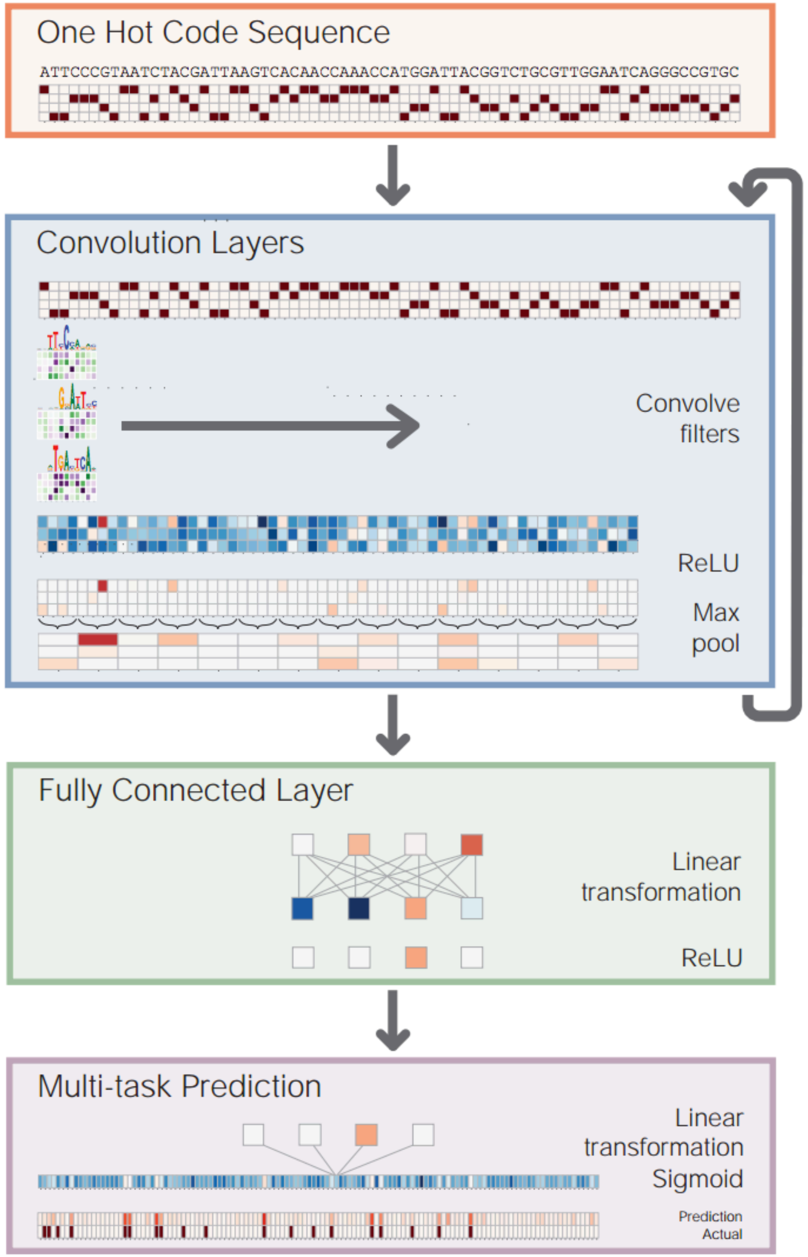
Case 4

Deep Learning



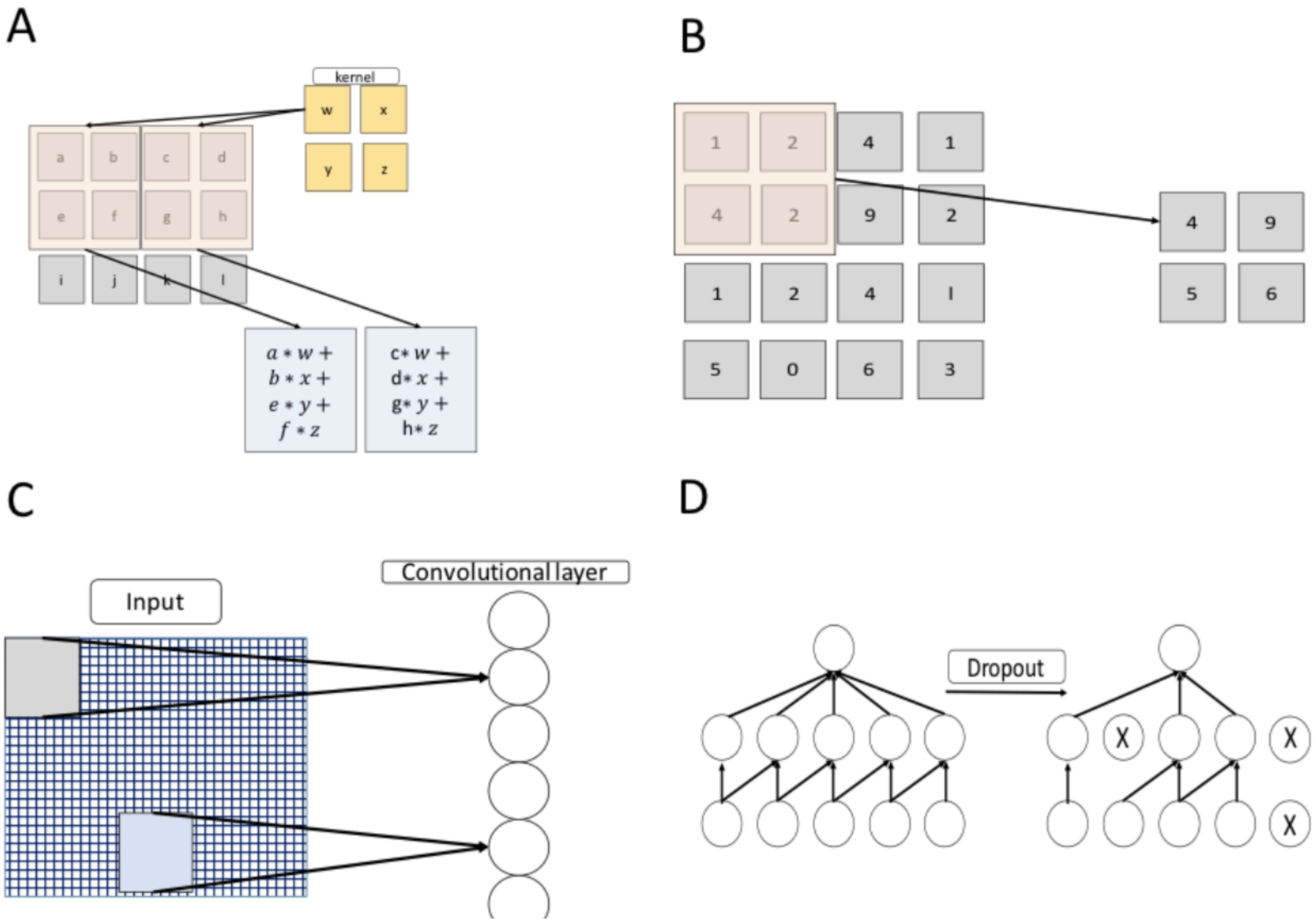
Case 4



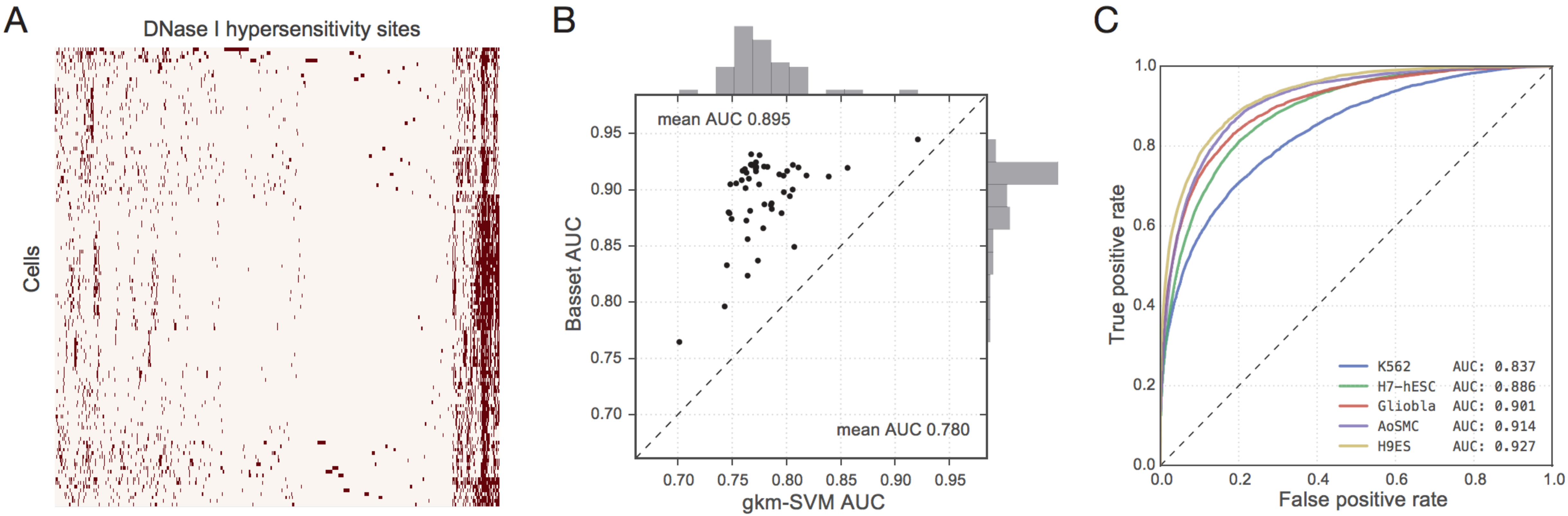


Kelley et al. 2016

Case 4

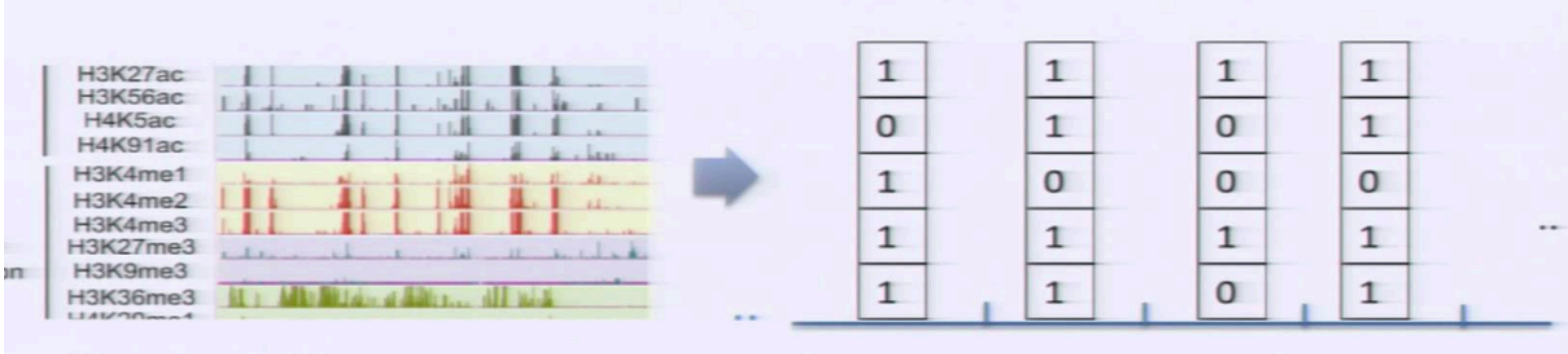

A
kernel
w
x
y
z
a
b
c
d
e
f
g
h
i
j
k
l
$a * w +$
$b * x +$
$e * y +$
$f * z$
c$* w +$
d$* x +$
g$* y +$
h$* z$
B
1
2
4
1
4
2
9
2
1
2
4
l
5
0
6
3
4
9
5
6
C
Input
Convolutional layer
D
Dropout
X
X
X

Case 4



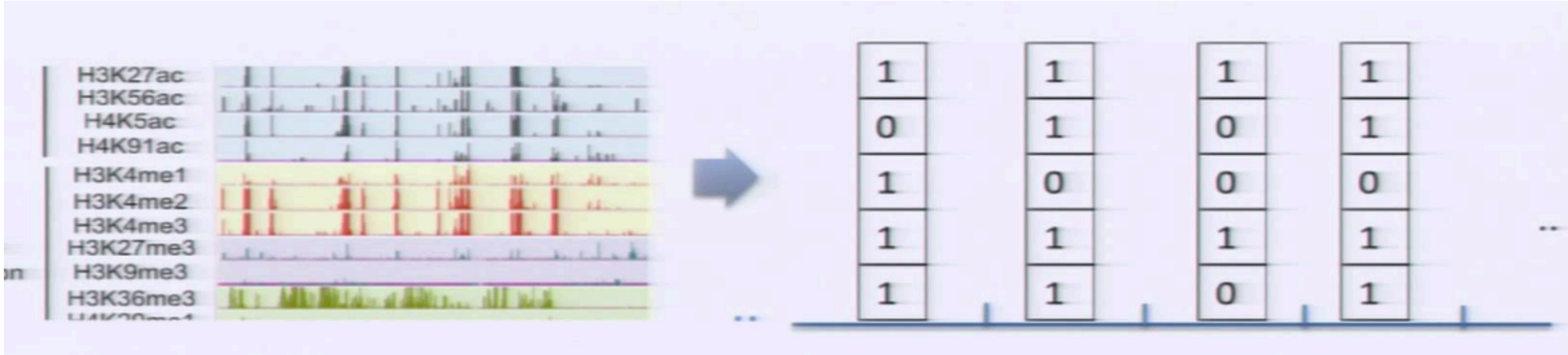
Kelley et al. 2016

Case 5
Models based on chromatin features

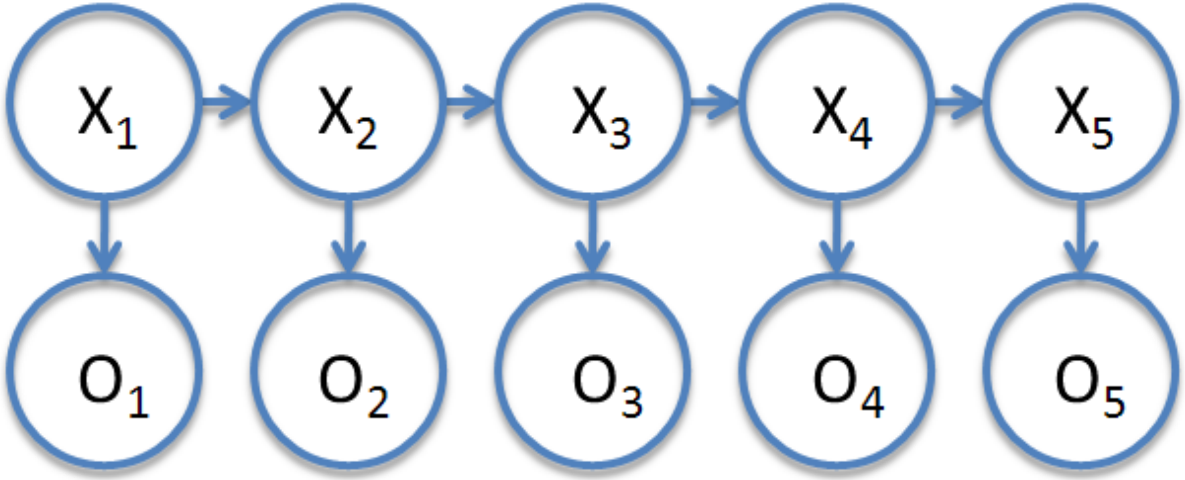


1	1	1	1
0	1	0	1
1	0	0	0
1	1	1	1
1	1	0	1

Case 5
Models based on chromatin features



1	1	1	1
0	1	0	1
1	0	0	0
1	1	1	1
1	1	0	1



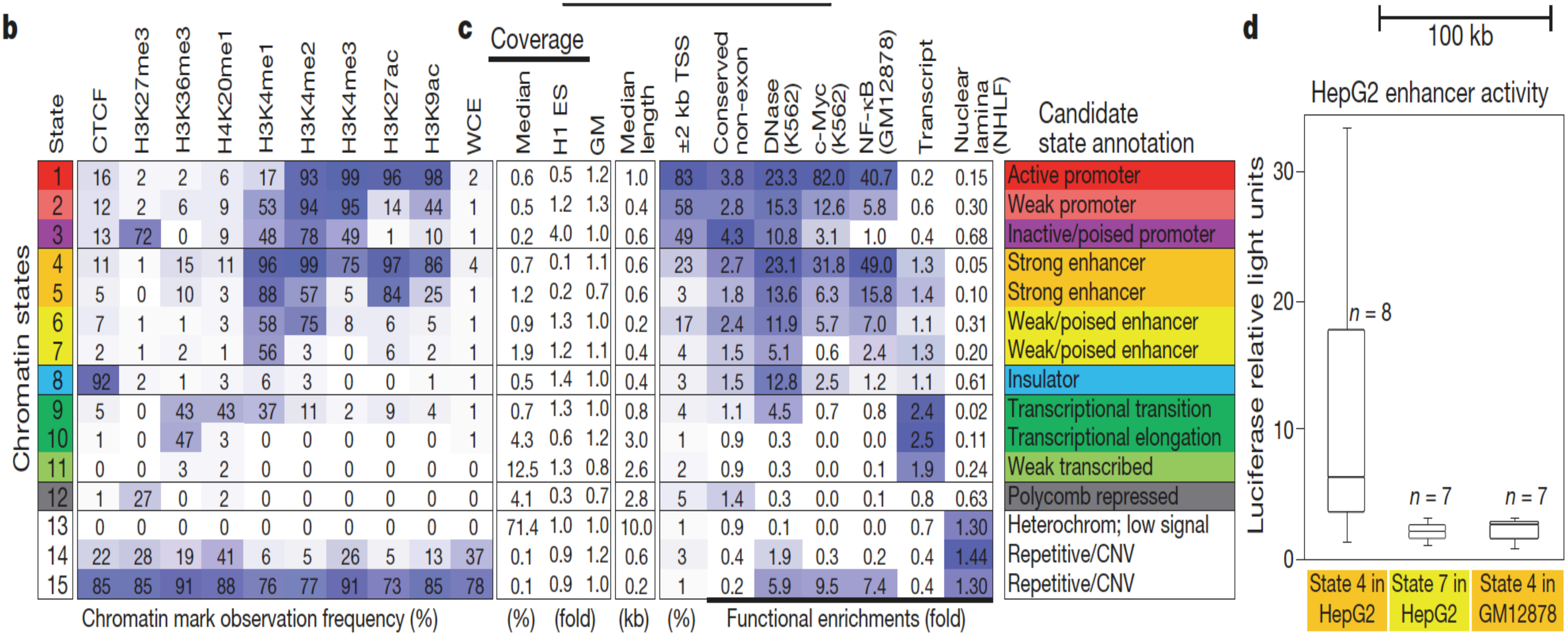
Case 5
Models based on chromatin features

Mapping and analysis of chromatin state dynamics in nine human cell types

Jason Ernst[1,2], Pouya Kheradpour[1,2], Tarjei S. Mikkelsen[1], Noam Shoresh[1], Lucas D. Ward[1,2], Charles B. Epstein[1], Xiaolan Zhang[1], Li Wang[1], Robbyn Issner[1], Michael Coyne[1], Manching Ku[1,3,4], Timothy Durham[1], Manolis Kellis[1,2]* & Bradley E. Bernstein[1,3,4]*

Case 5

Models based on chromatin features



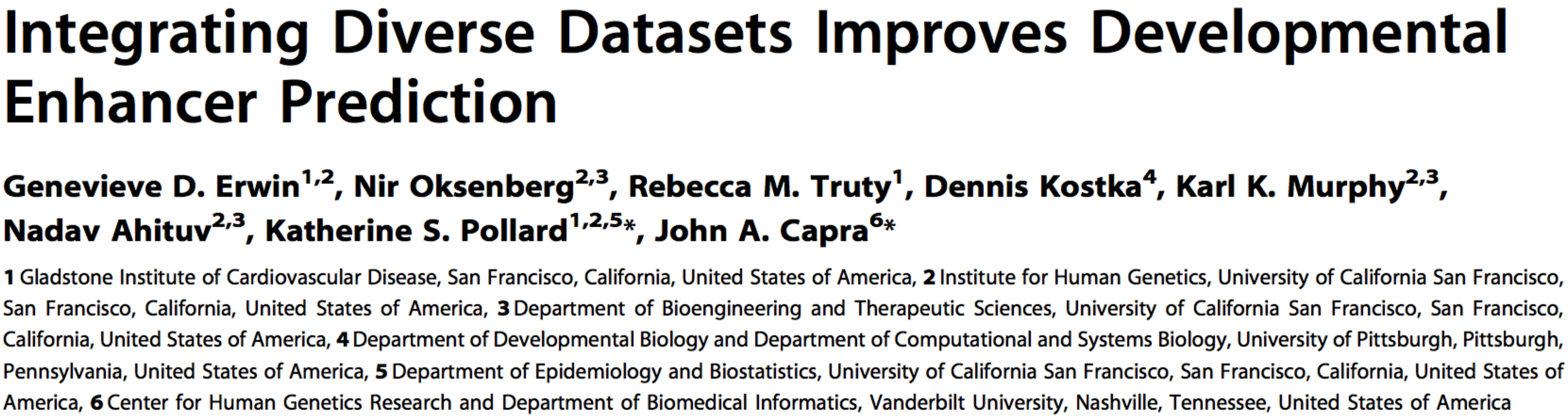
b
c
Coverage
Chromatin states
d
100 kb
HepG2 enhancer activity
Luciferase relative light units
0
10
20
30
n = 8
n = 7
n = 7
State 4 in HepG2
State 7 in HepG2
State 4 in GM12878

State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Median	H1 ES	GM	Median length	±2 kb TSS	Conserved non-exon	DNase (K562)	c-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

Chromatin mark observation frequency (%) (%) (fold) (kb) (%) Functional enrichments (fold)

Ernst et al.

Case 6



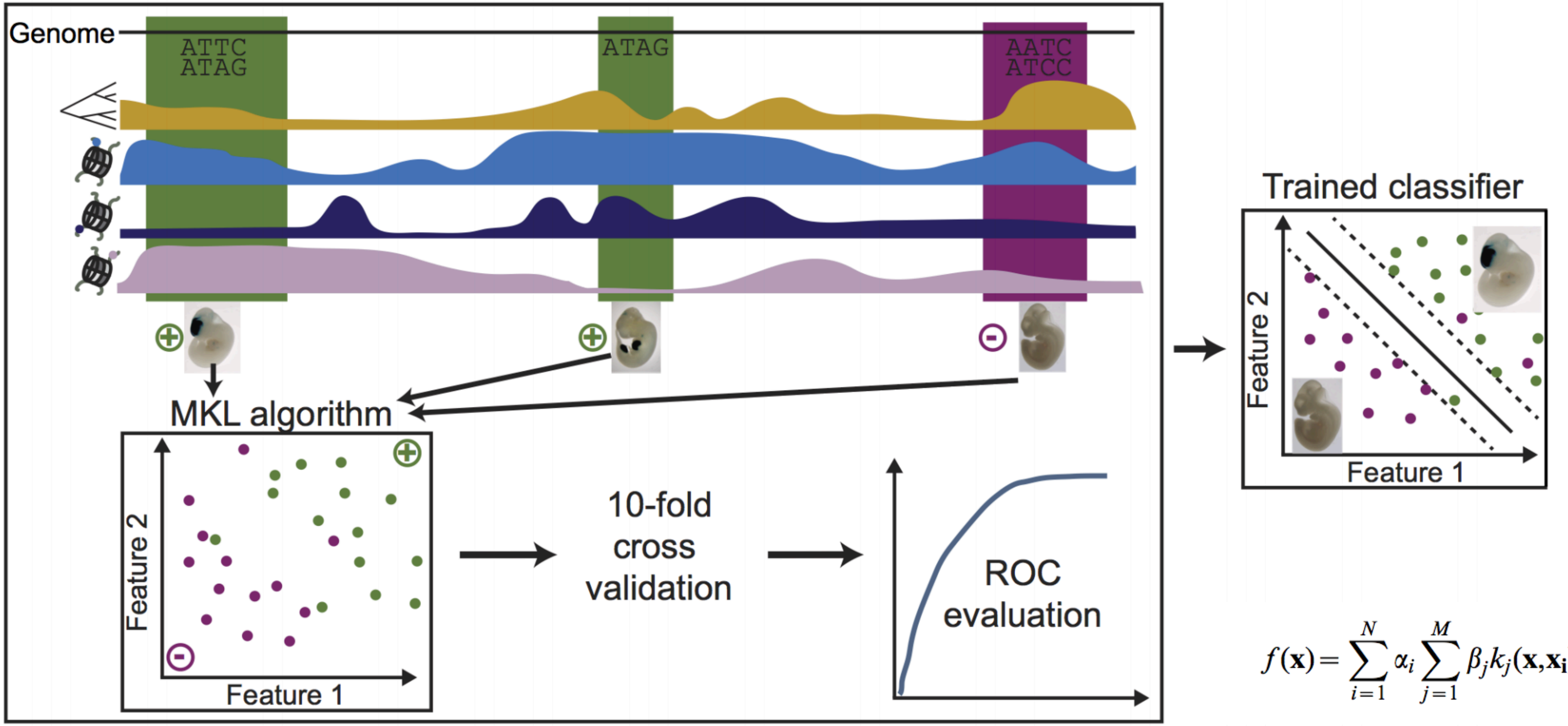
Integrating Diverse Datasets Improves Developmental Enhancer Prediction

Genevieve D. Erwin[1,2], Nir Oksenberg[2,3], Rebecca M. Truty[1], Dennis Kostka[4], Karl K. Murphy[2,3], Nadav Ahituv[2,3], Katherine S. Pollard[1,2,5]*, John A. Capra[6]*

1 Gladstone Institute of Cardiovascular Disease, San Francisco, California, United States of America, **2** Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, **3** Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America, **4** Department of Developmental Biology and Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **5** Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America, **6** Center for Human Genetics Research and Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, United States of America

Case 6

MKL training:



MKL algorithm

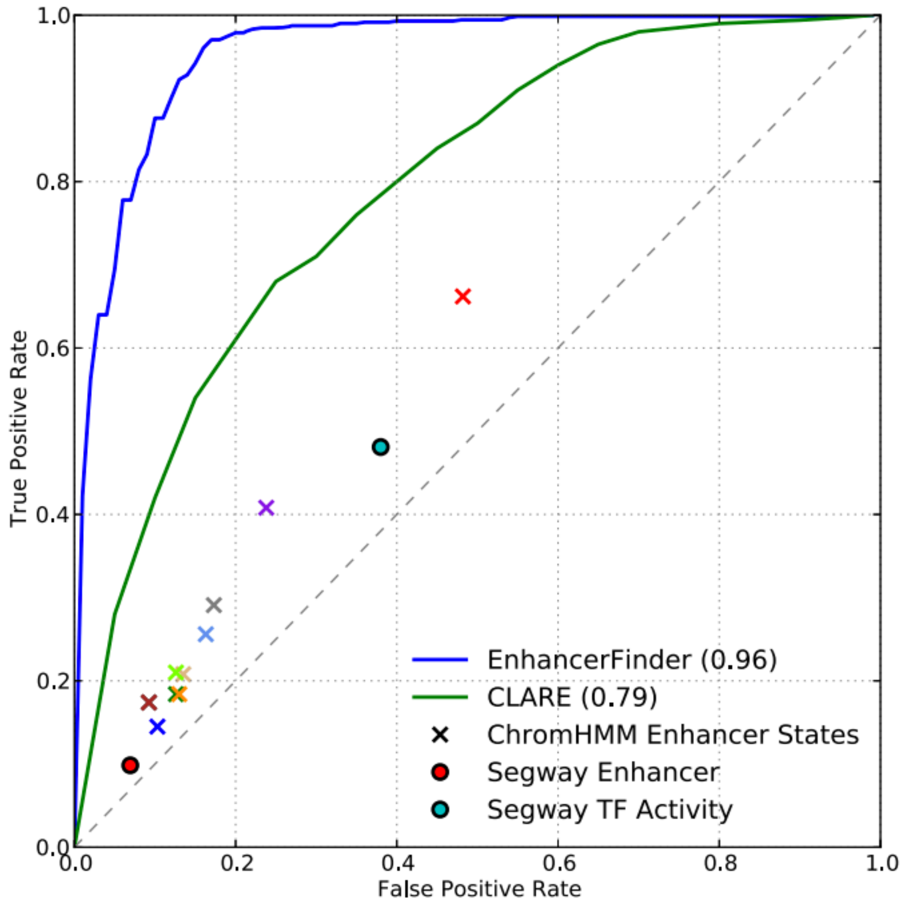
10-fold cross validation

ROC evaluation

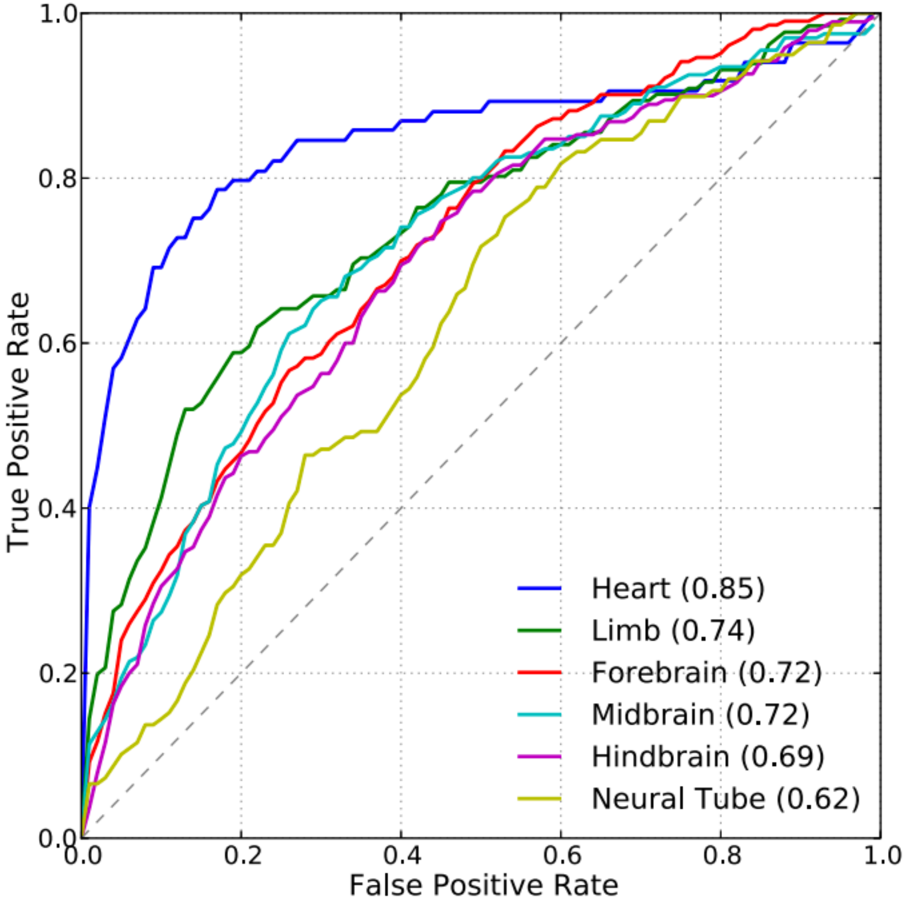
Trained classifier

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \sum_{j=1}^{M} \beta_j k_j(\mathbf{x},\mathbf{x_i}) + b$$

Case 6



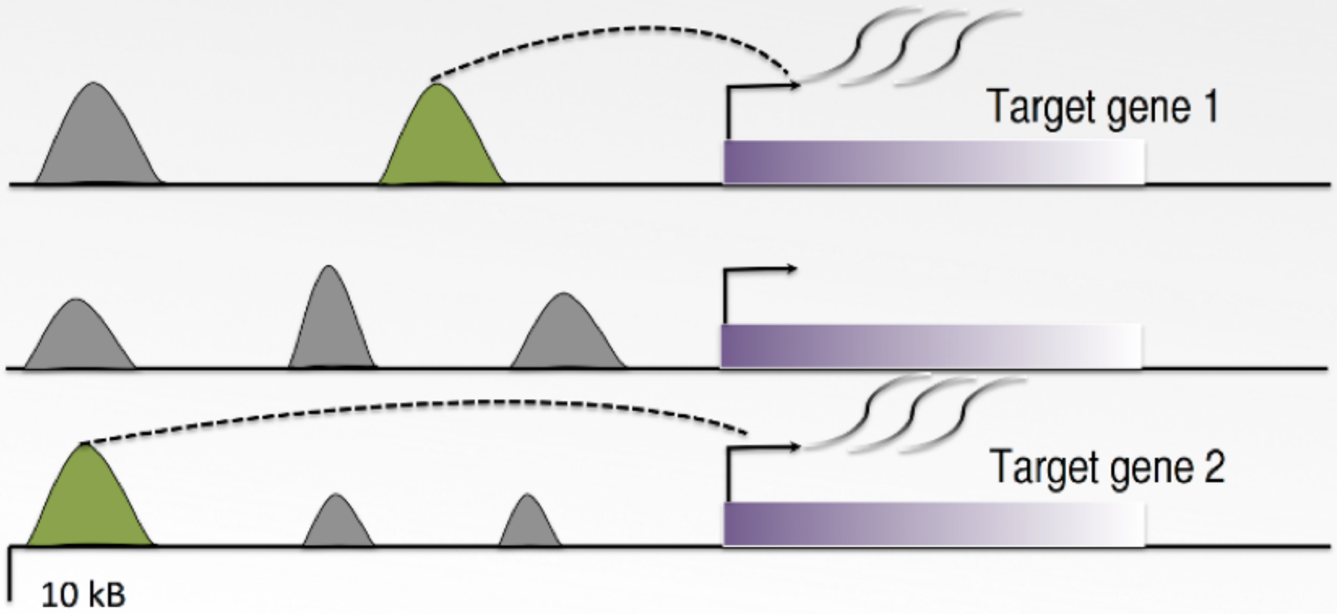
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
EnhancerFinder (0.96)
CLARE (0.79)
ChromHMM Enhancer States
Segway Enhancer
Segway TF Activity

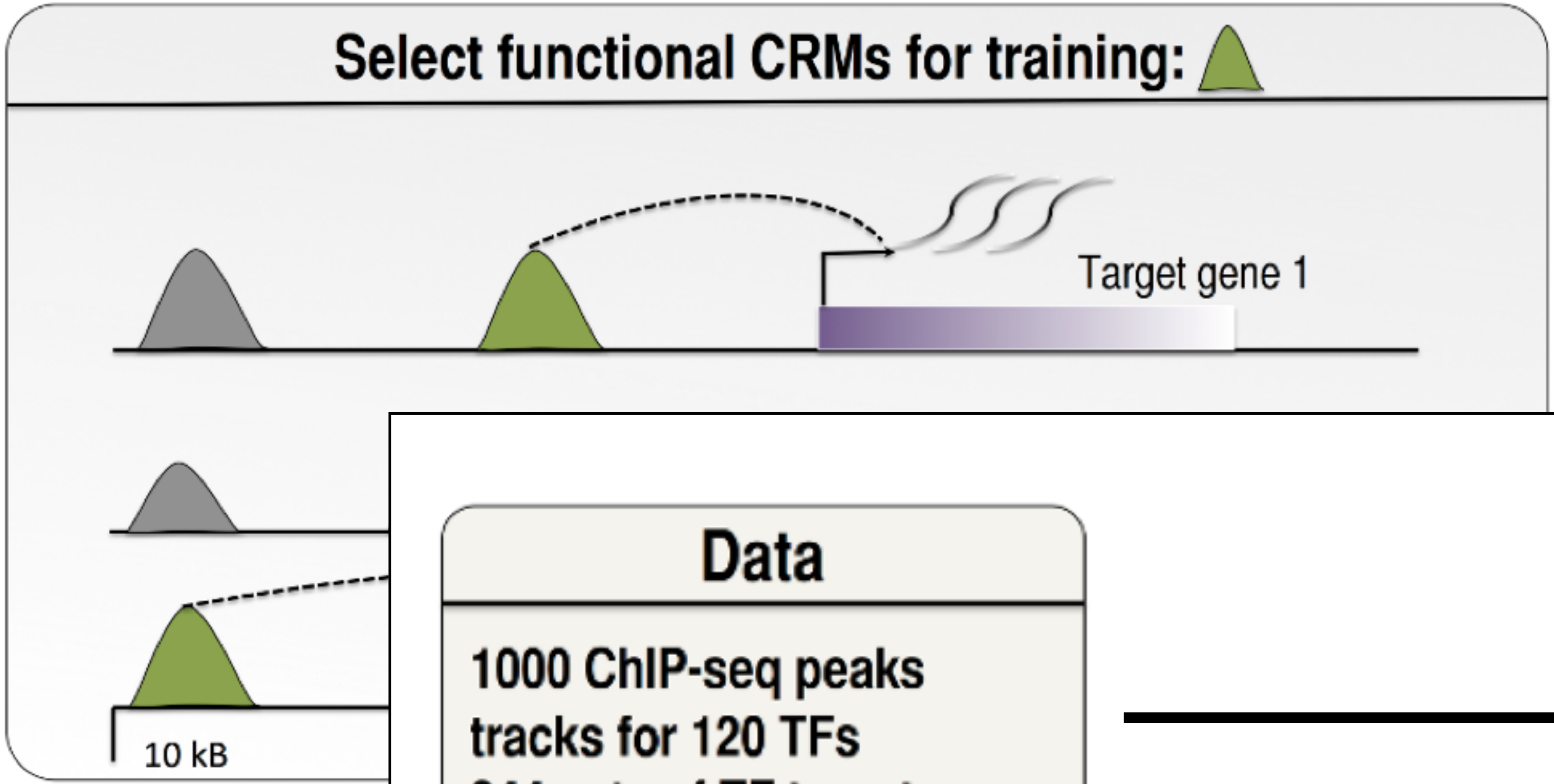


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Heart (0.85)
Limb (0.74)
Forebrain (0.72)
Midbrain (0.72)
Hindbrain (0.69)
Neural Tube (0.62)

Case 7
Prediction of functional TFBS using bag of motifs

Select functional CRMs for training: ▲



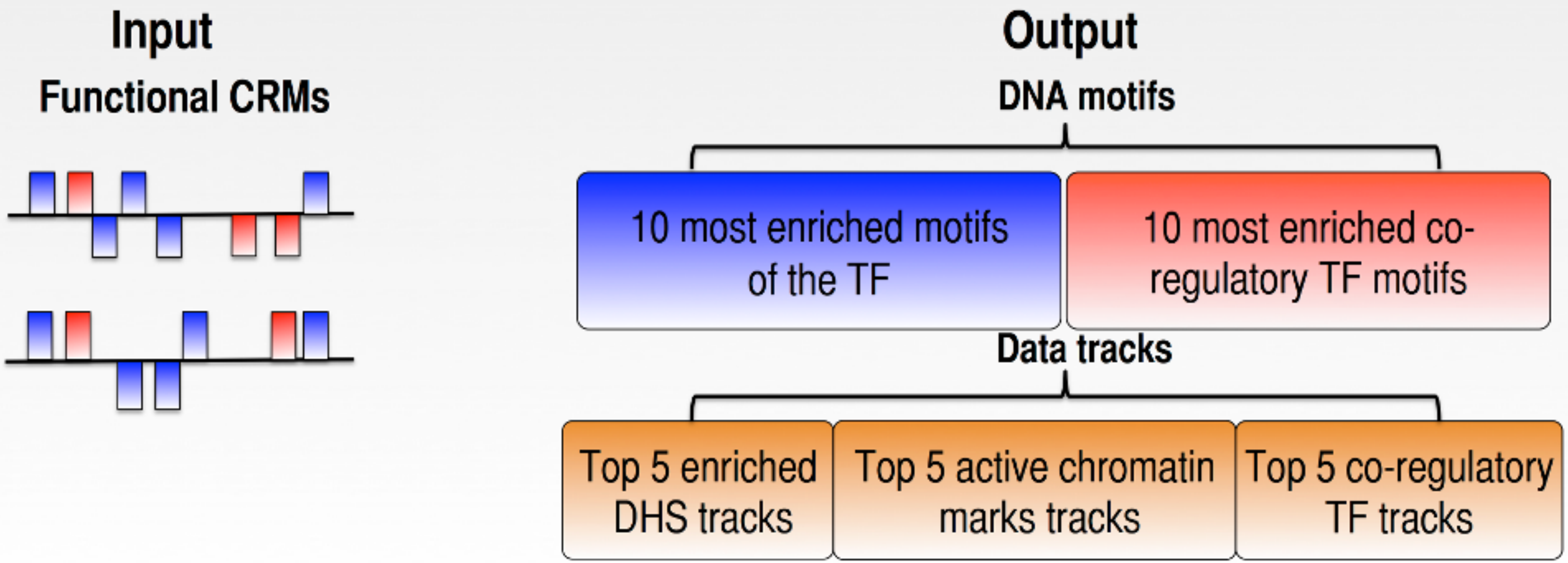


Data

1000 ChIP-seq peaks tracks for 120 TFs
344 sets of TF target genes

→ **45 training sets**

Feature selection for training data



Input

Functional CRMs

Output

DNA motifs

10 most enriched motifs of the TF

10 most enriched co-regulatory TF motifs

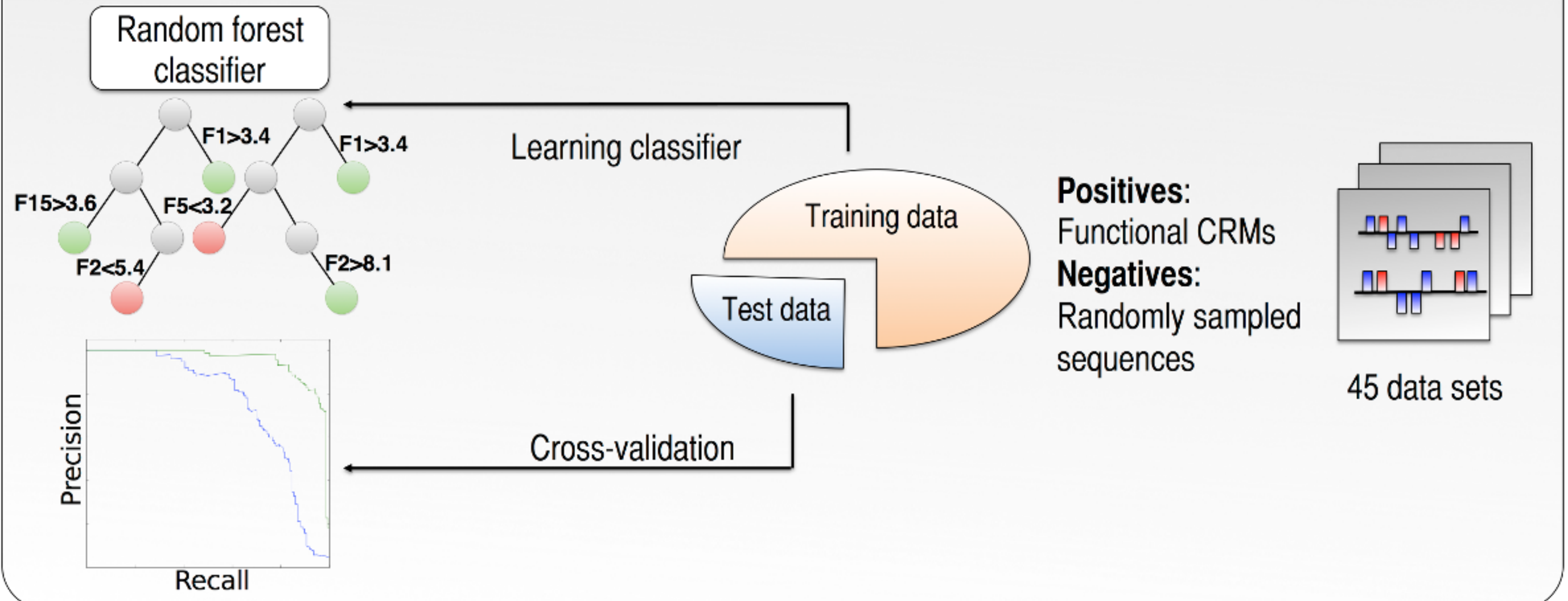
Data tracks

Top 5 enriched DHS tracks

Top 5 active chromatin marks tracks

Top 5 co-regulatory TF tracks

Machine learning



Random forest classifier

Learning classifier

Training data

Test data

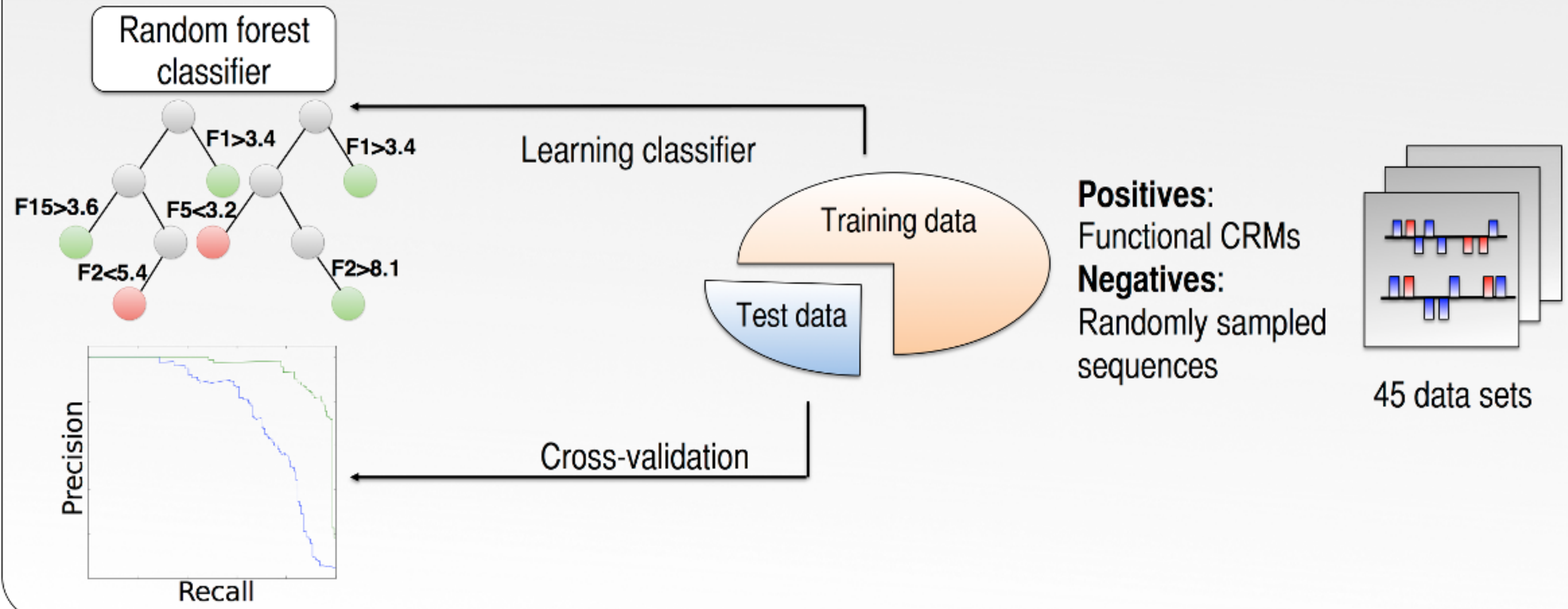
Cross-validation

Positives: Functional CRMs

Negatives: Randomly sampled sequences

45 data sets

Machine learning



Random forest classifier

Learning classifier

Training data

Test data

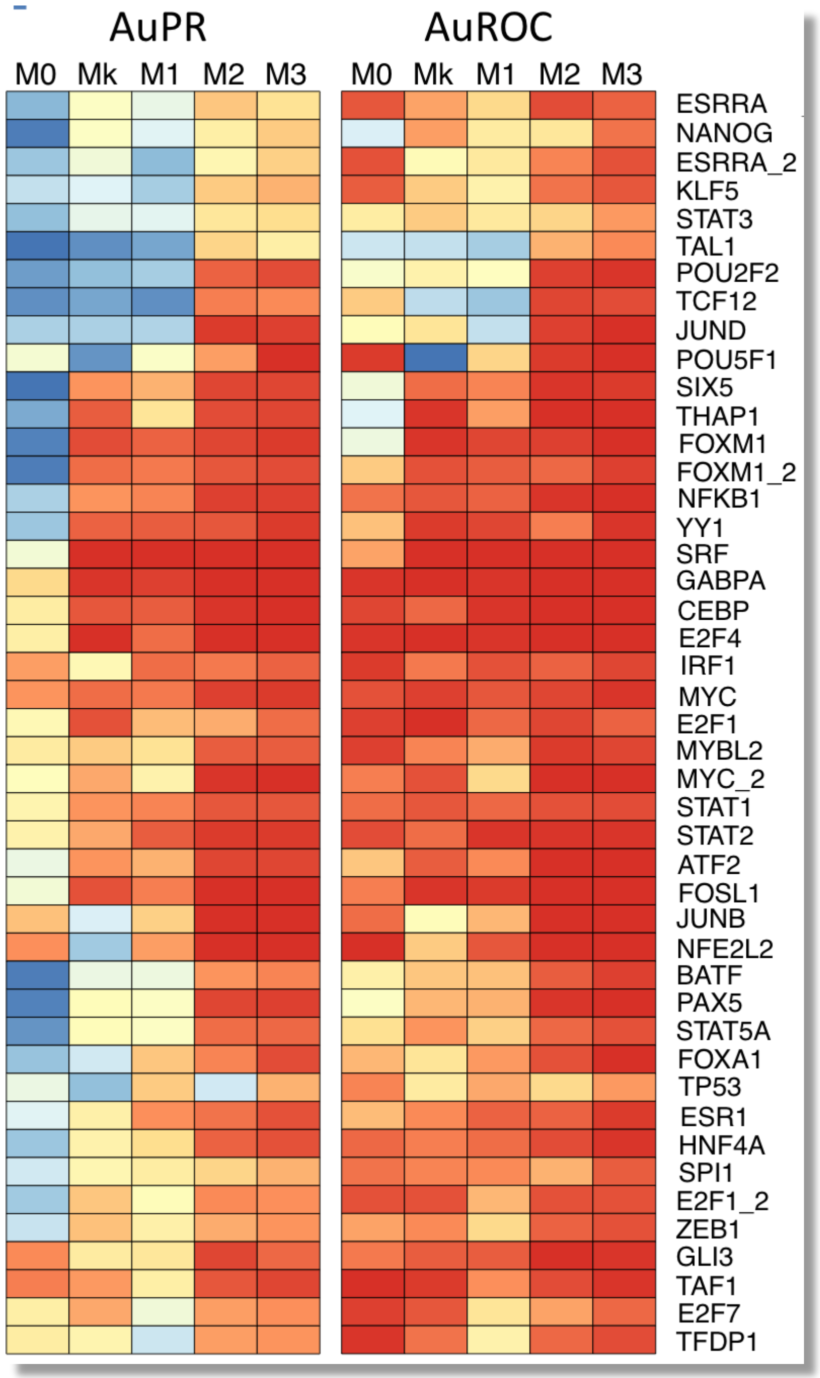
Cross-validation

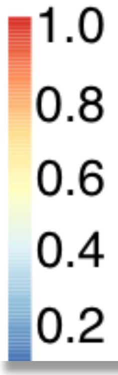
Positives: Functional CRMs

Negatives: Randomly sampled sequences

45 data sets

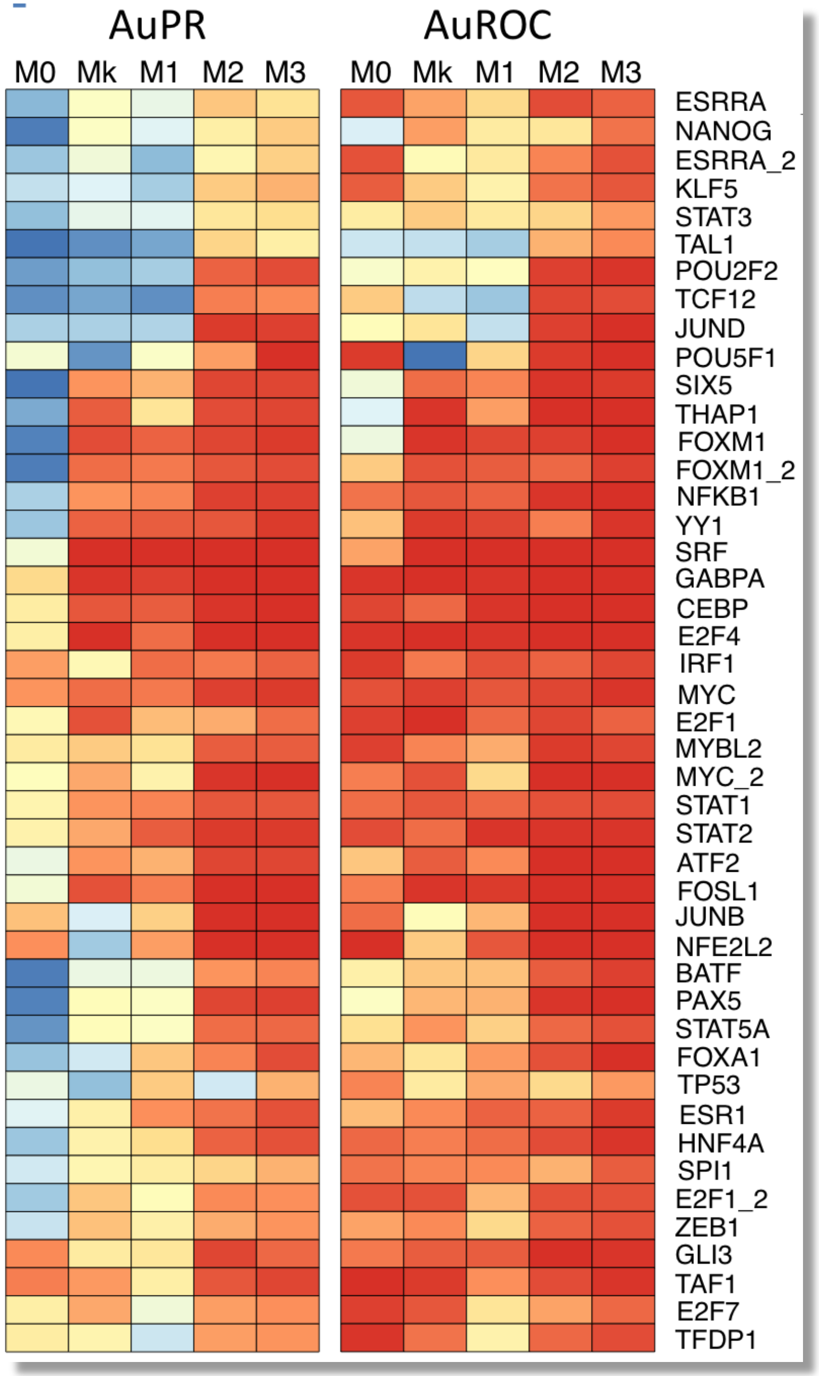
Cross-validation

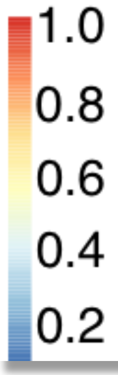




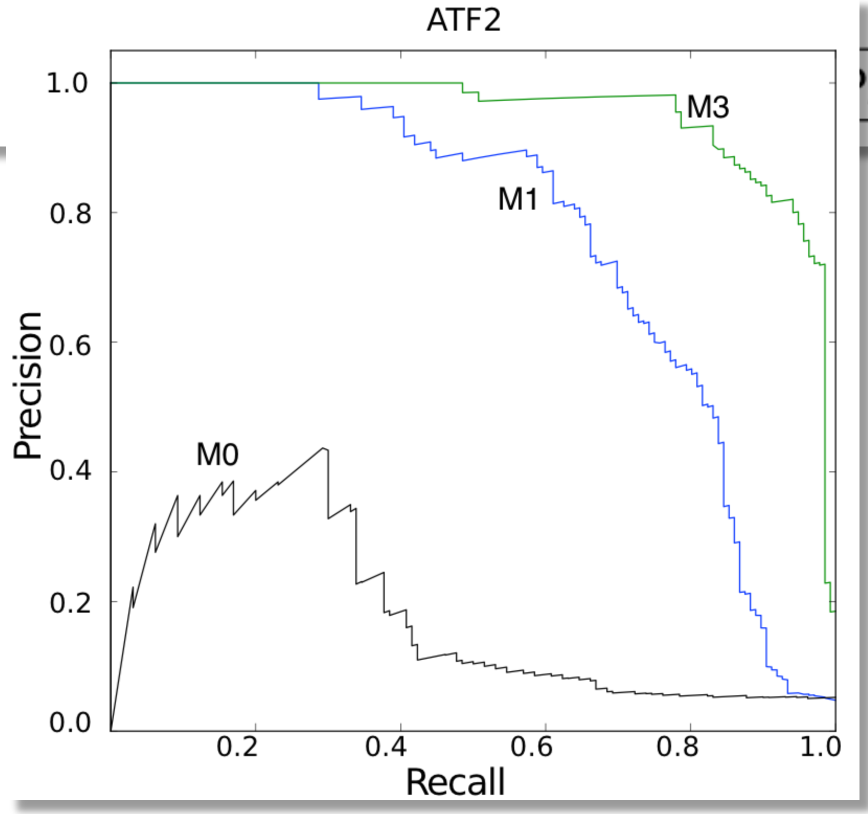
M0	Single PWM
Mk	kmer-SVM
M1	RF with multiple PWMs
M2	RF with ChIP/DHS
M3	RF with multiple PWMs and ChIP/DHS

Cross-validation



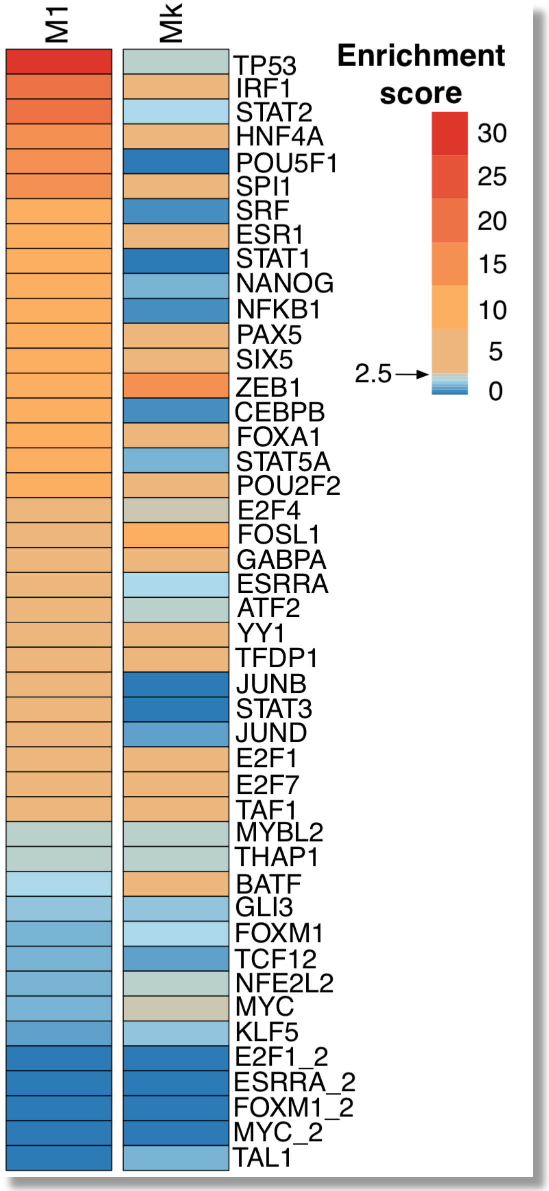


M0	Single PWM
Mk	kmer-SVM
M1	RF with multiple PWMs
	/DHS



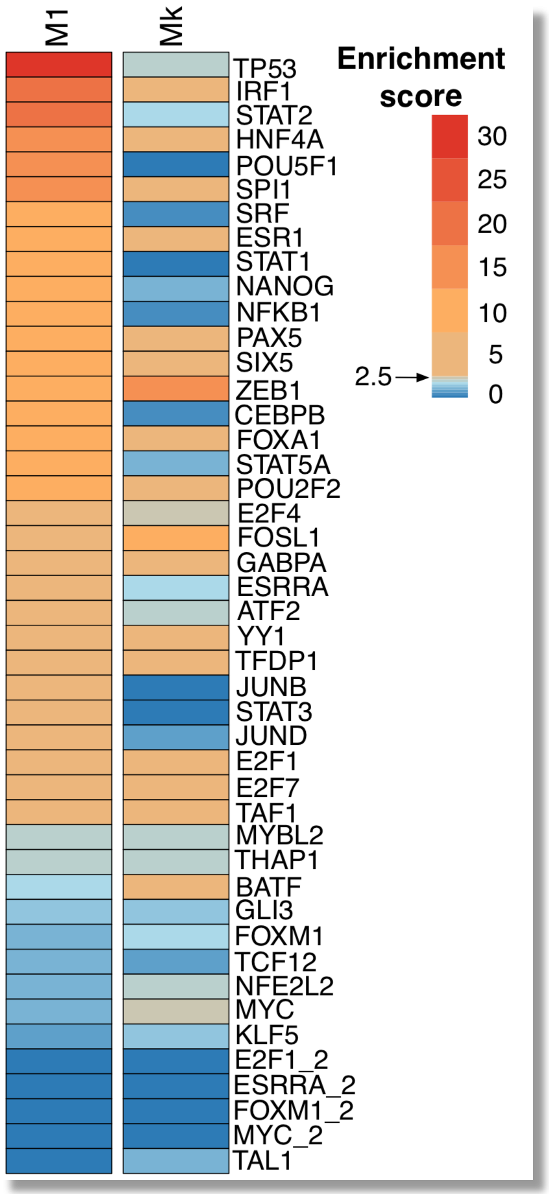
M1 RF MODELS CAN PREDICT BINDING OF TFs

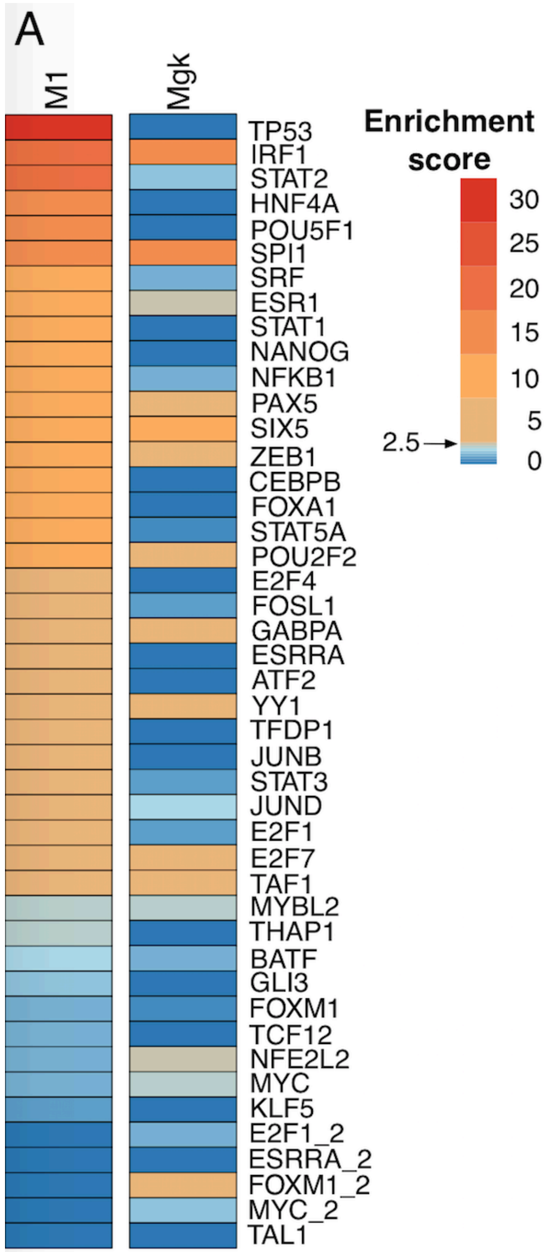
TF ChIP-seq



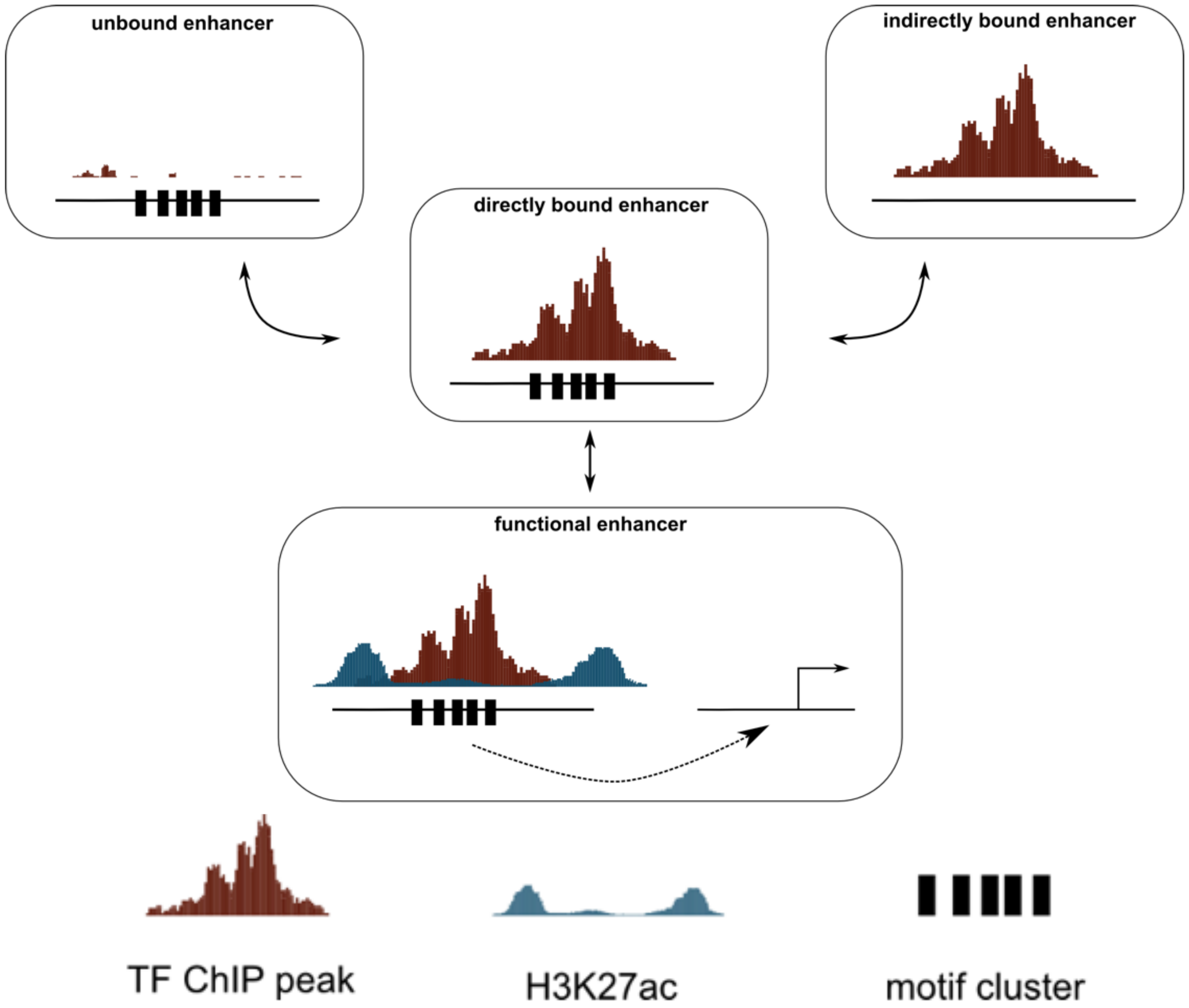
TF ChIP-seq

M1 RF MODELS CAN PREDICT BINDING OF TFs





Open questions in enhancer identification



unbound enhancer

directly bound enhancer

indirectly bound enhancer

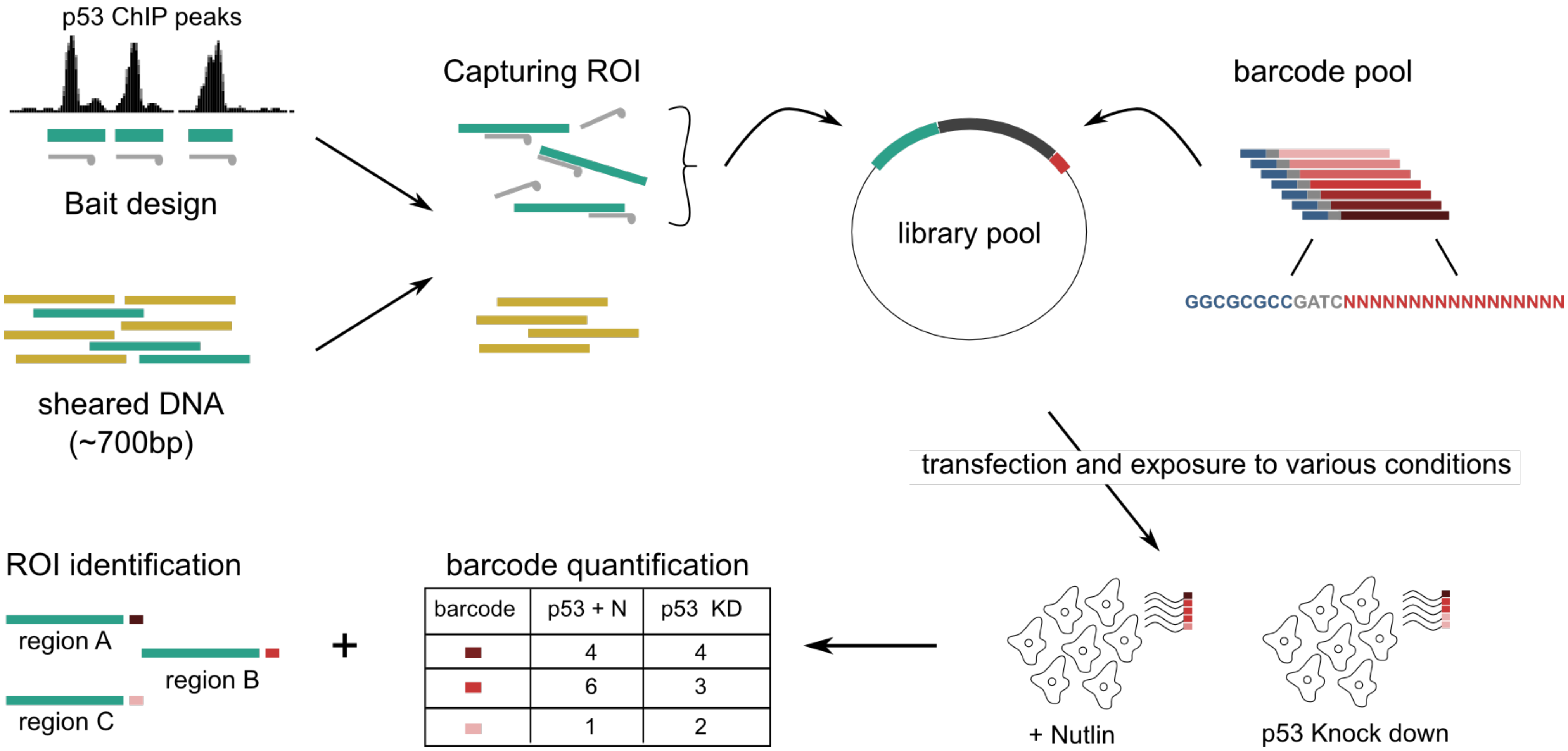
functional enhancer

TF ChIP peak

H3K27ac

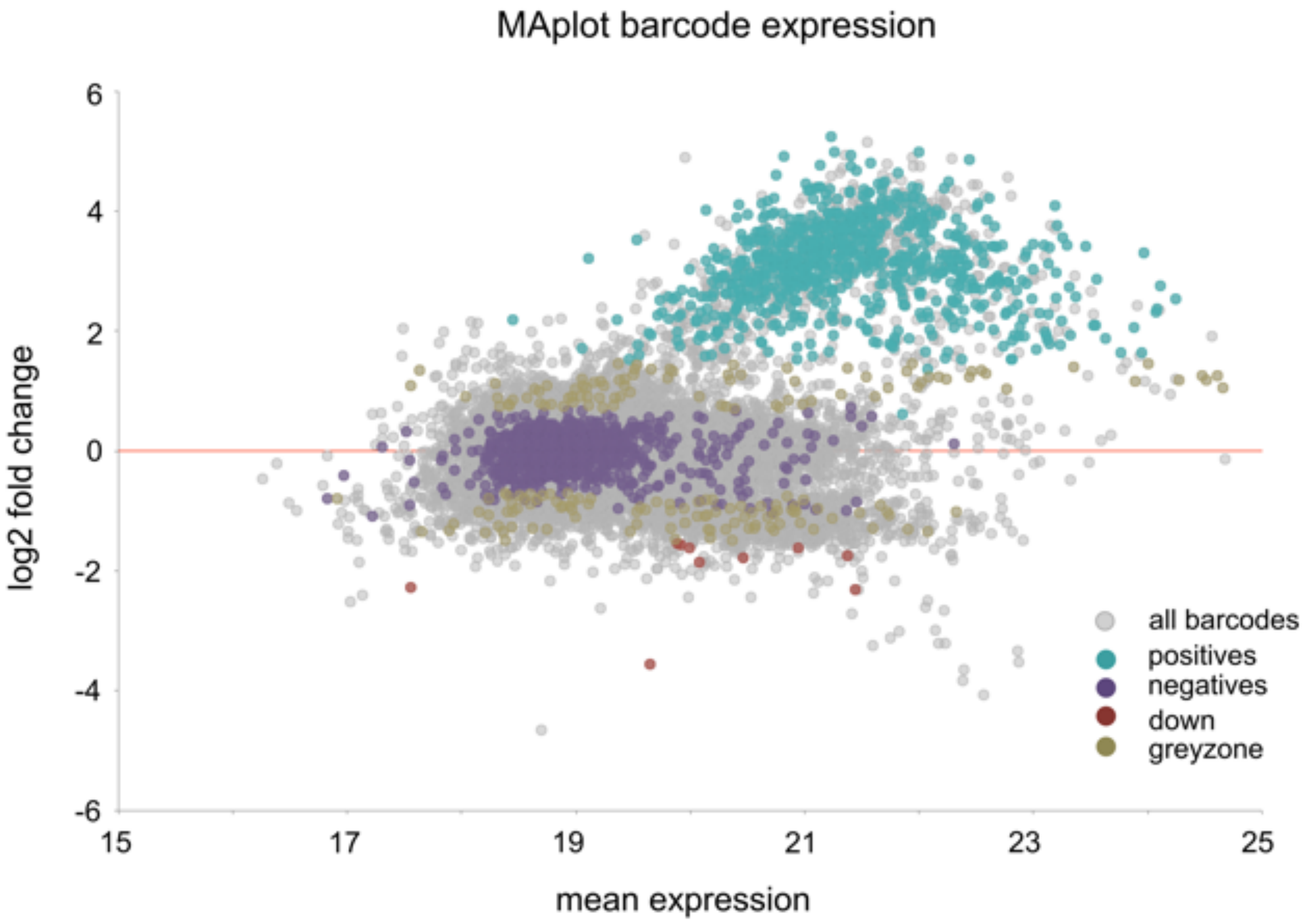
motif cluster

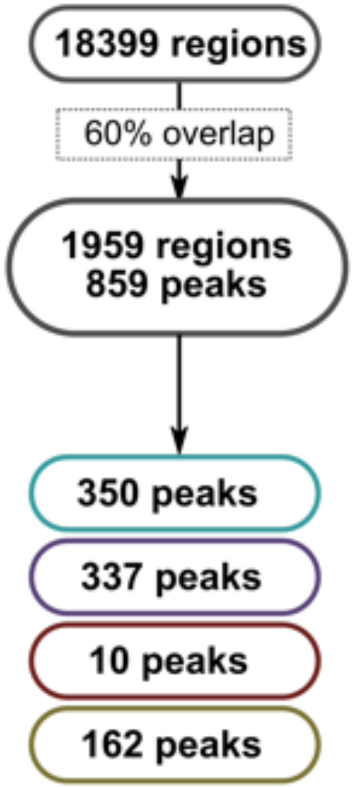
CHEQ-seq: Captured High-throughput Enhancer testing by Quantitative sequencing



barcode	p53 + N	p53 KD
■	4	4
■	6	3
■	1	2

CHEQ-seq shows distinct sets of active enhancers




18399 regions

60% overlap

1959 regions
859 peaks

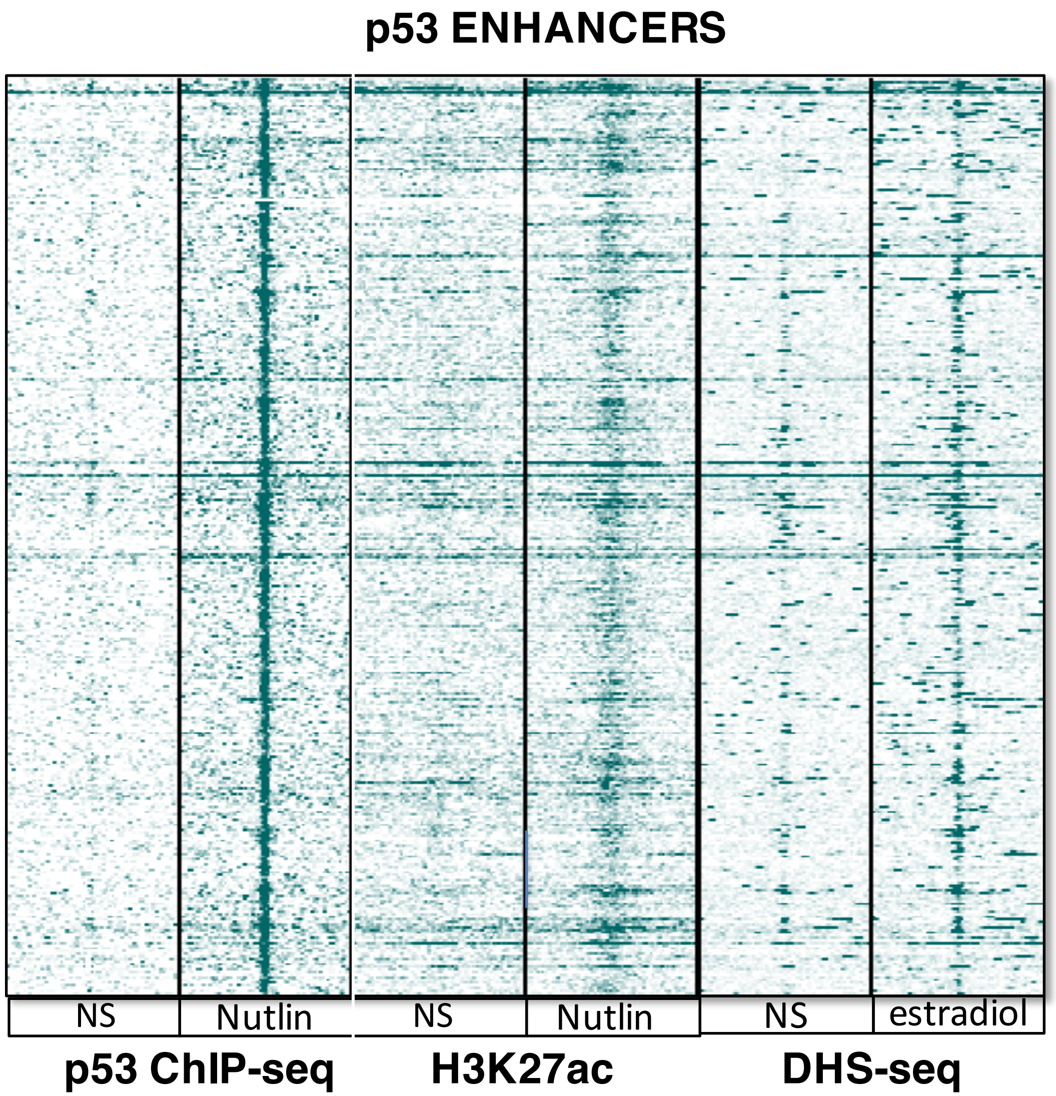
350 peaks

337 peaks

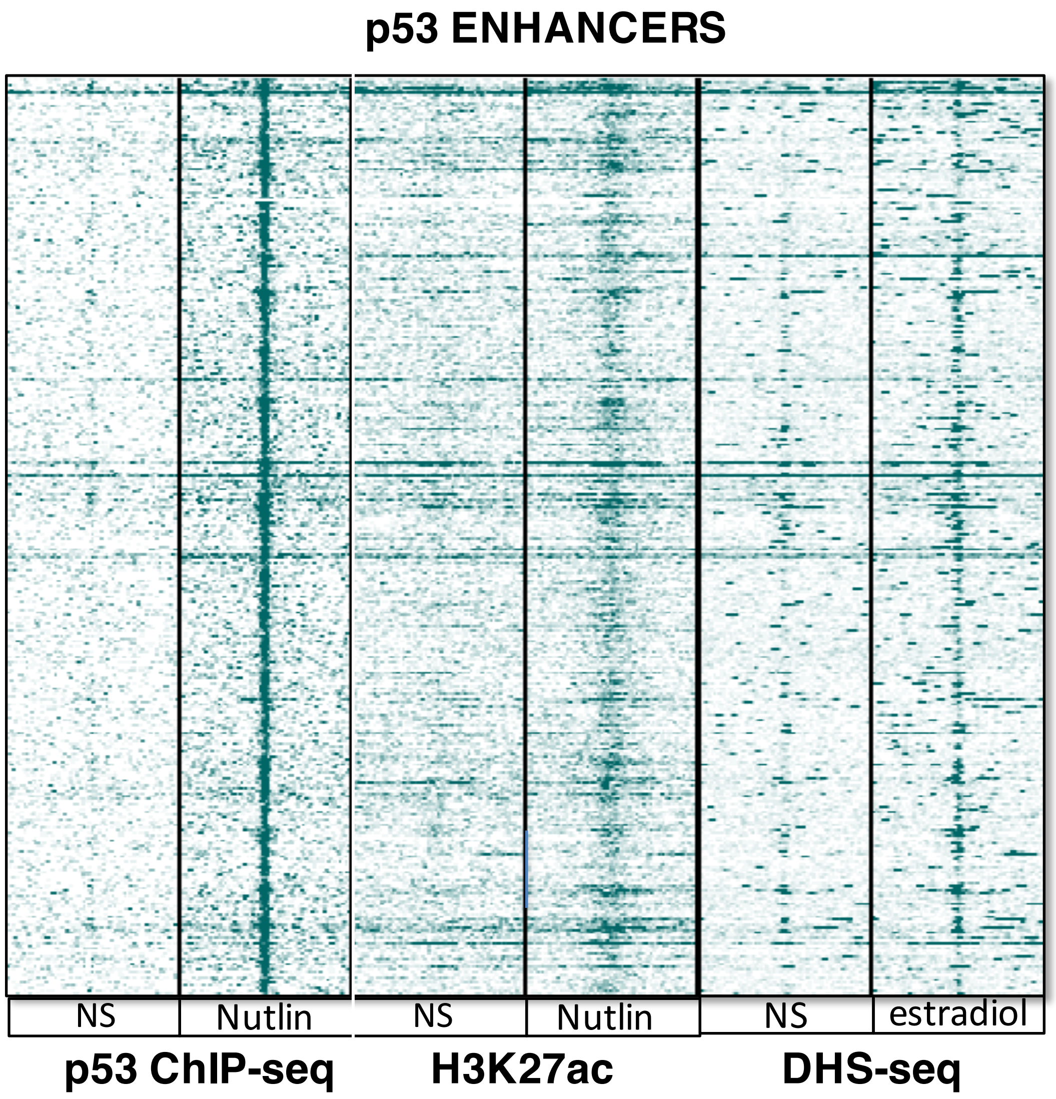
10 peaks

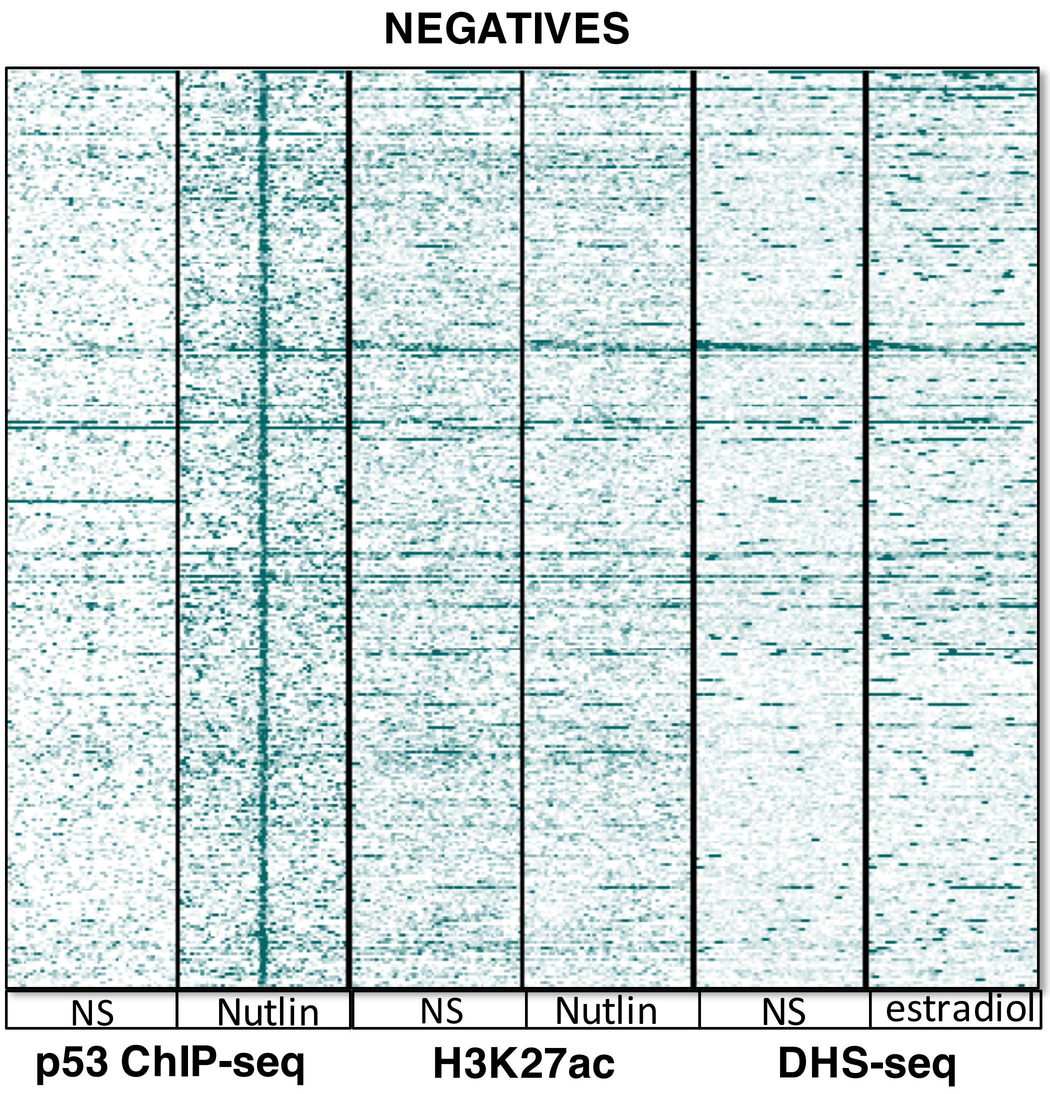
162 peaks

CHEQ-seq positives show marks of functional enhancers

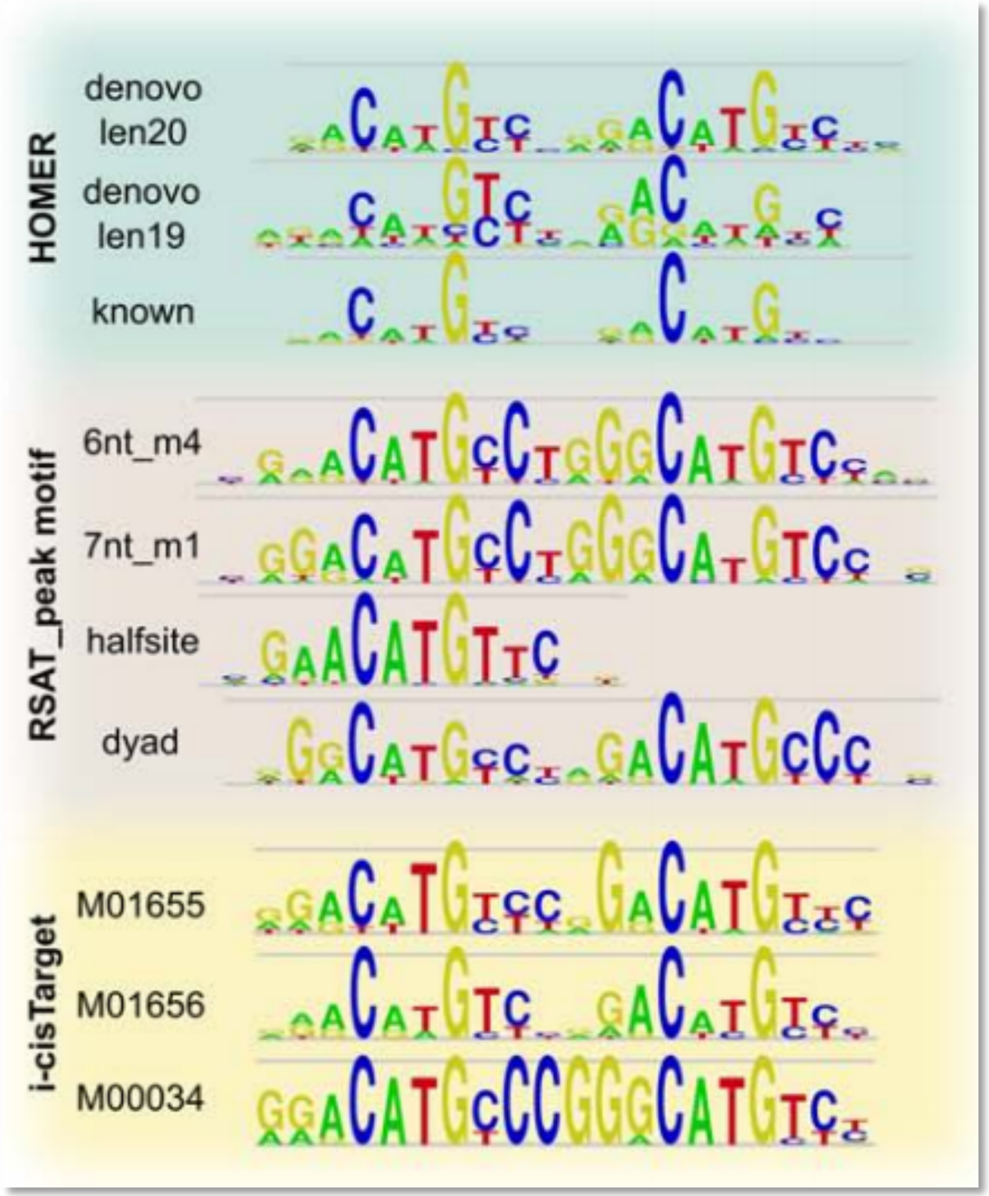


CHEQ-seq positives show marks of functional enhancers

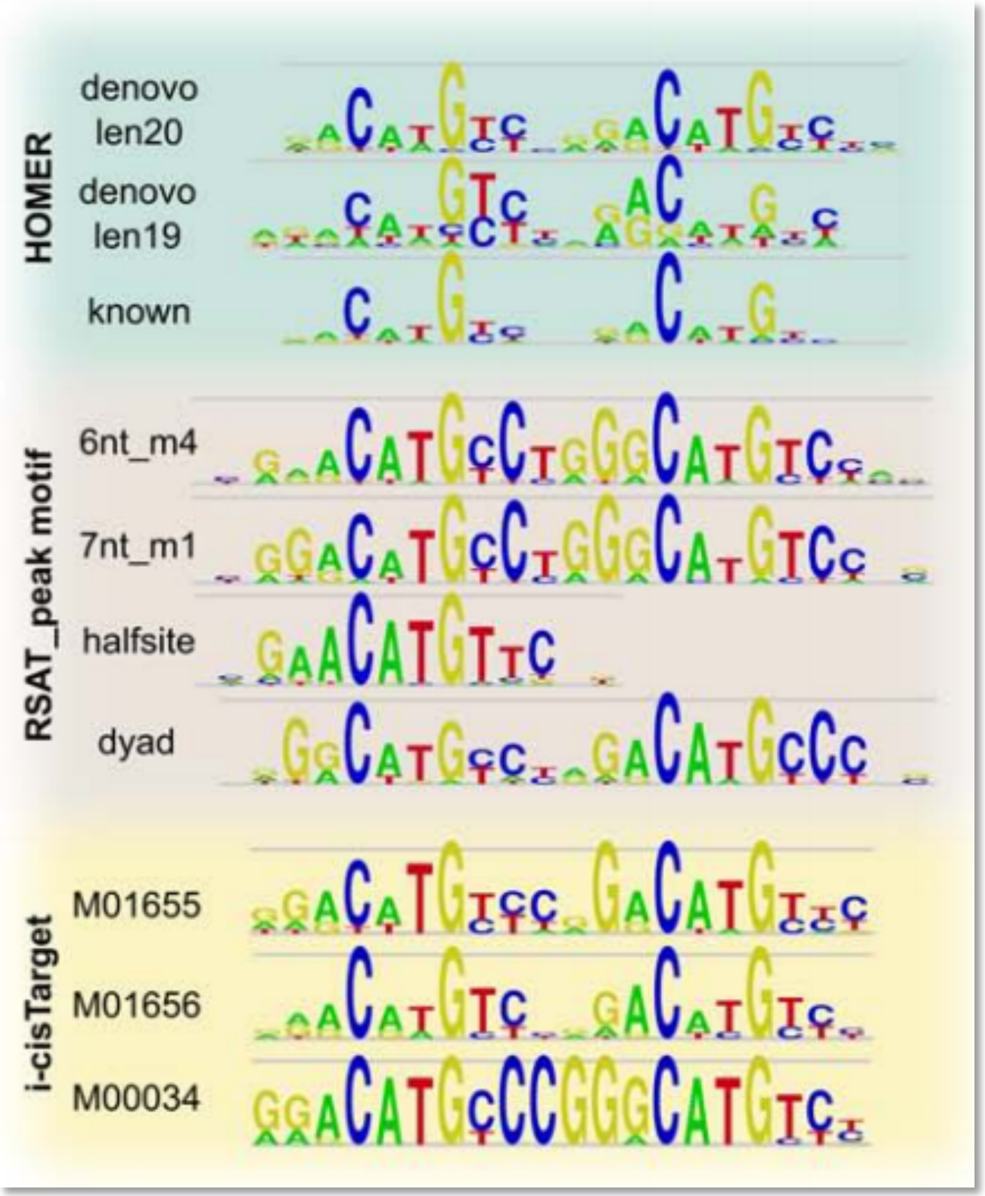


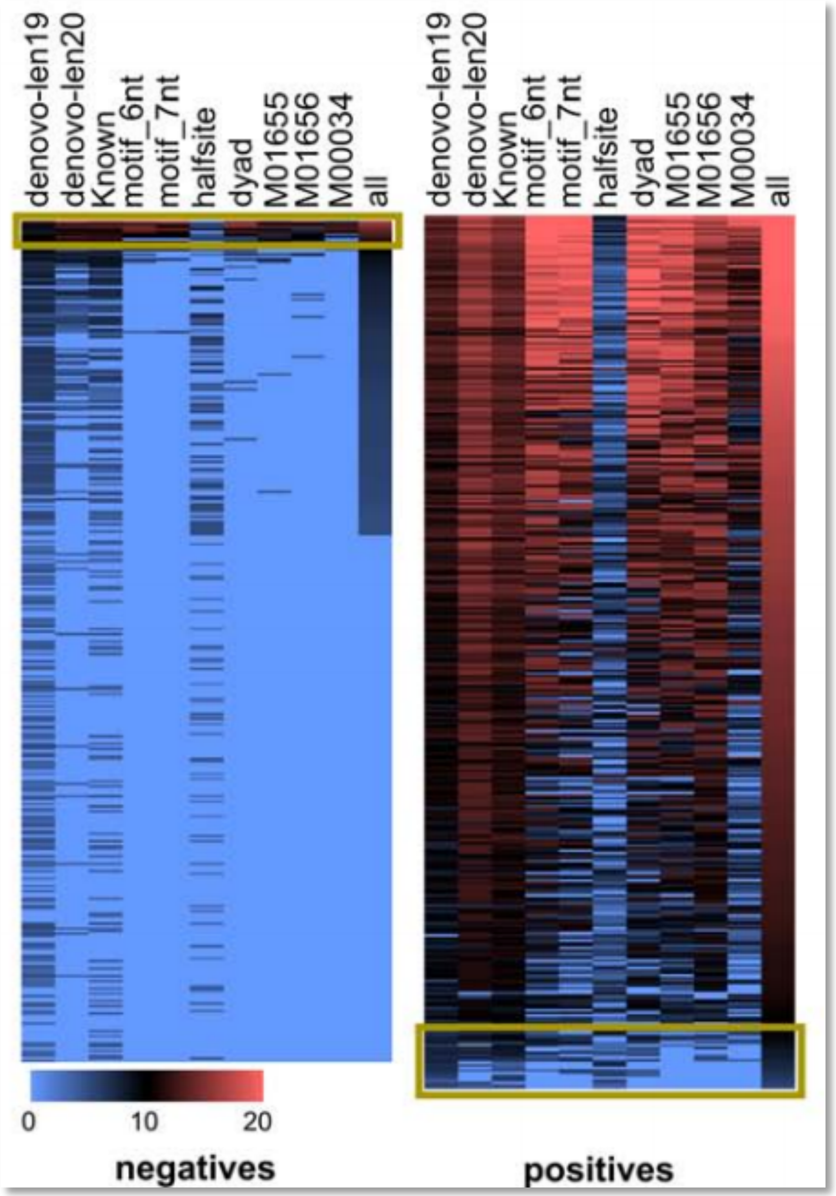


Enhancer classification using motifs

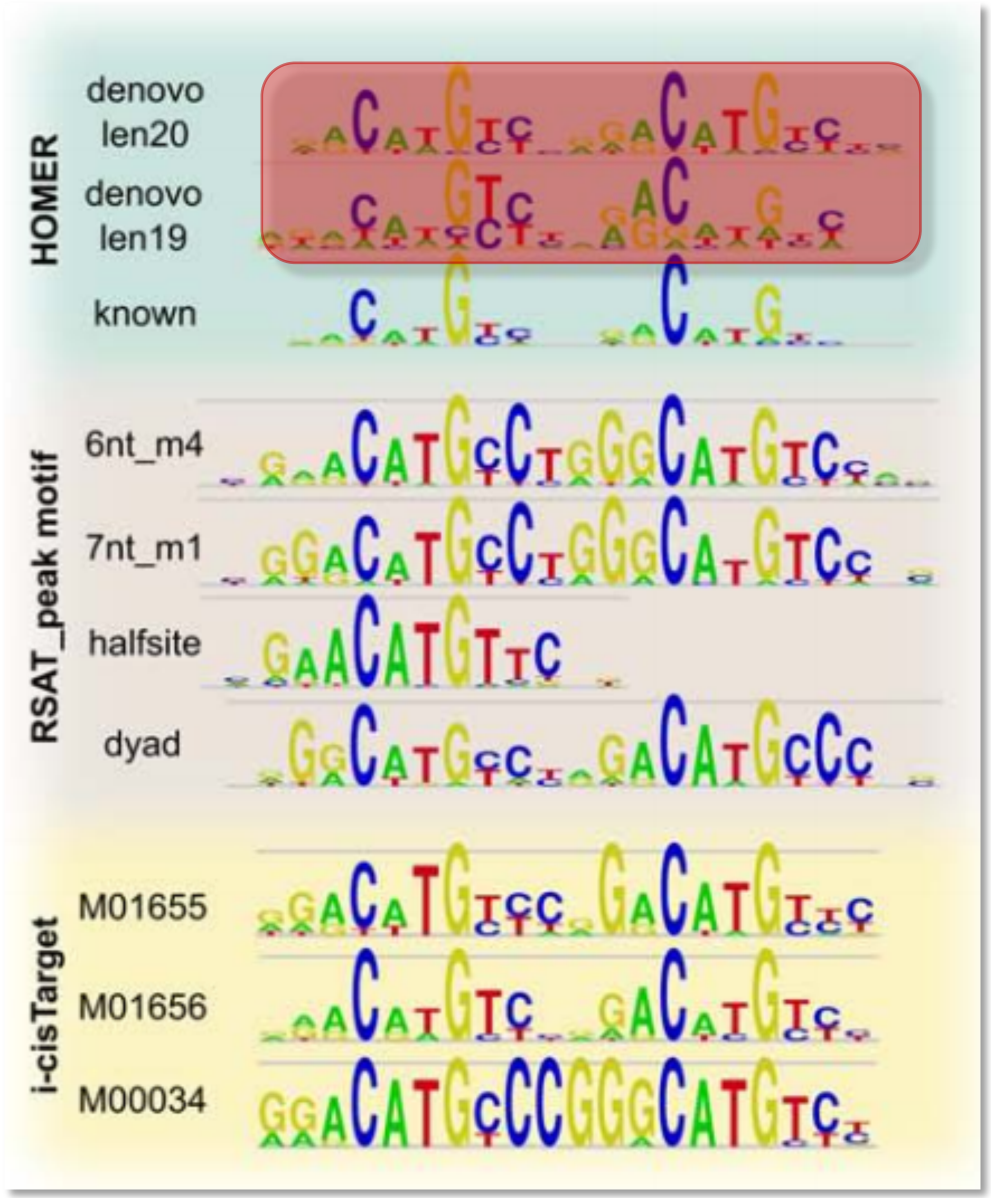


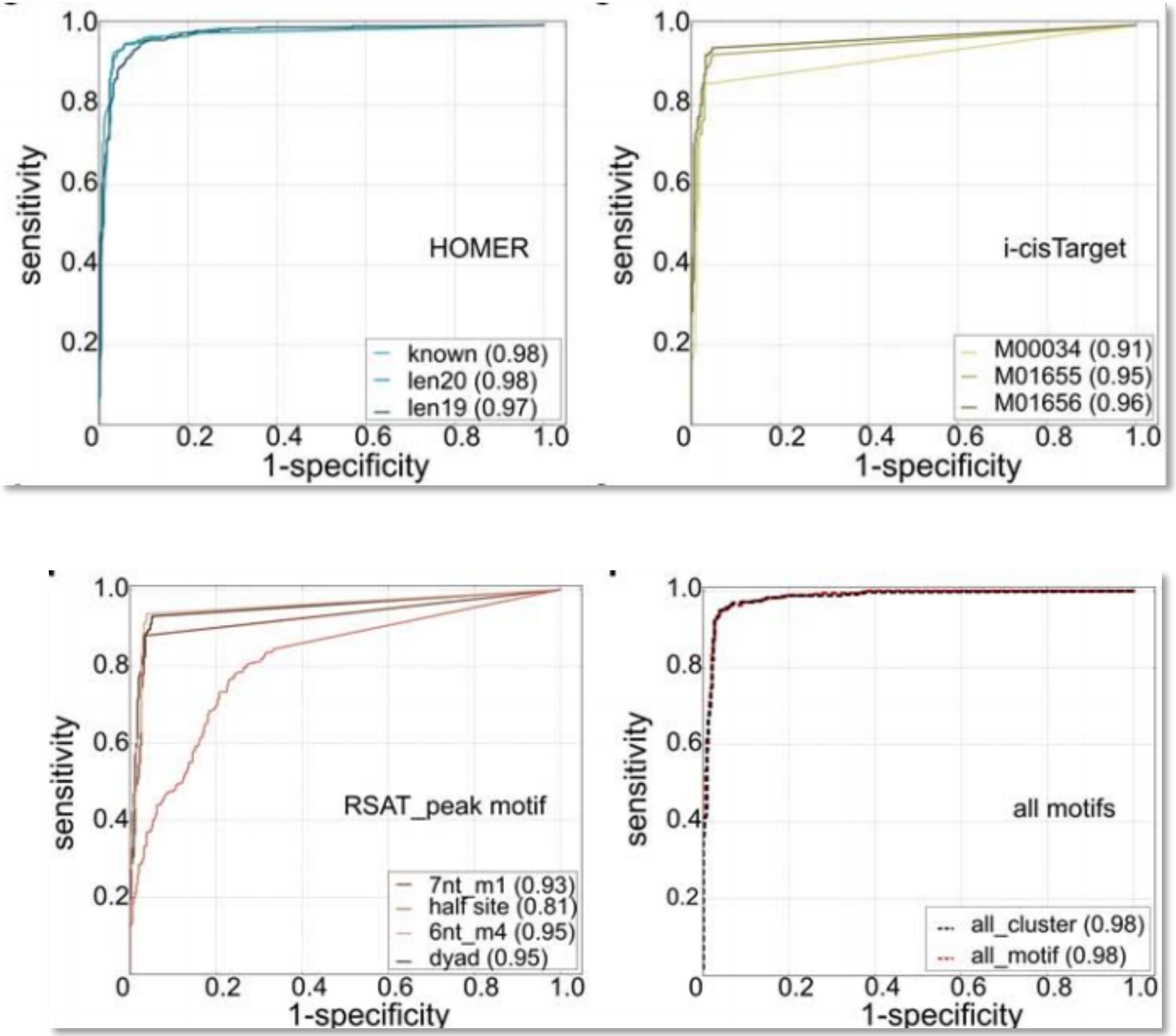
Enhancer classification using motifs



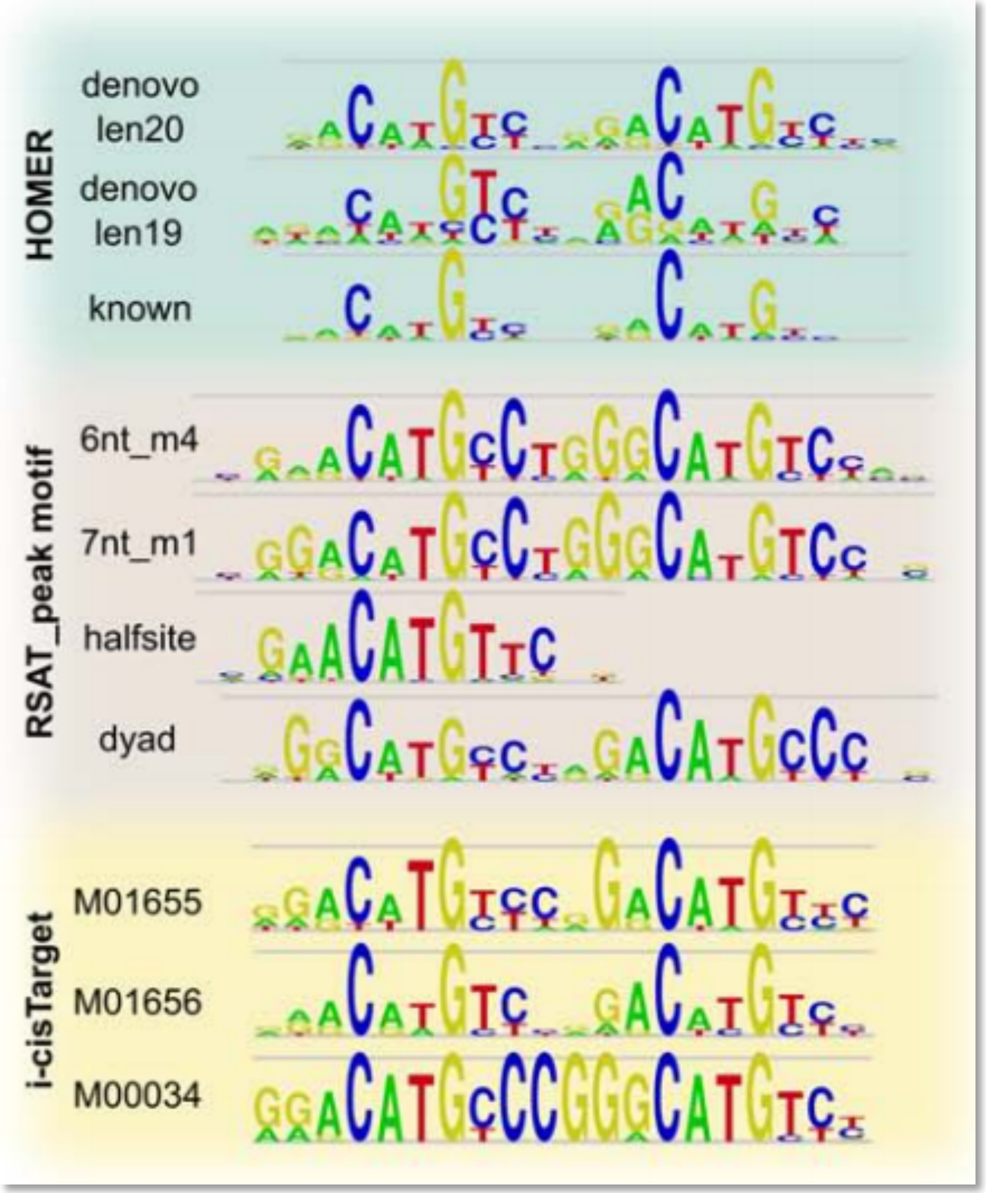


Enhancer classification using motifs





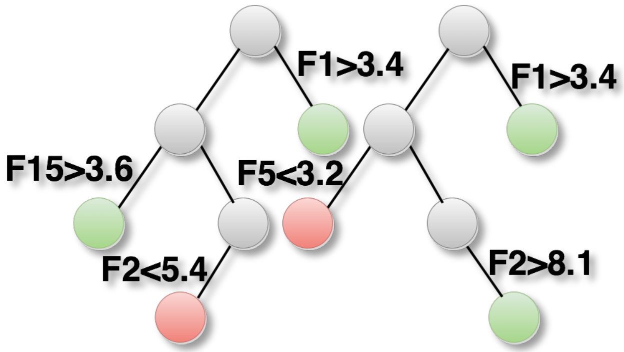
Enhancer classification using motifs



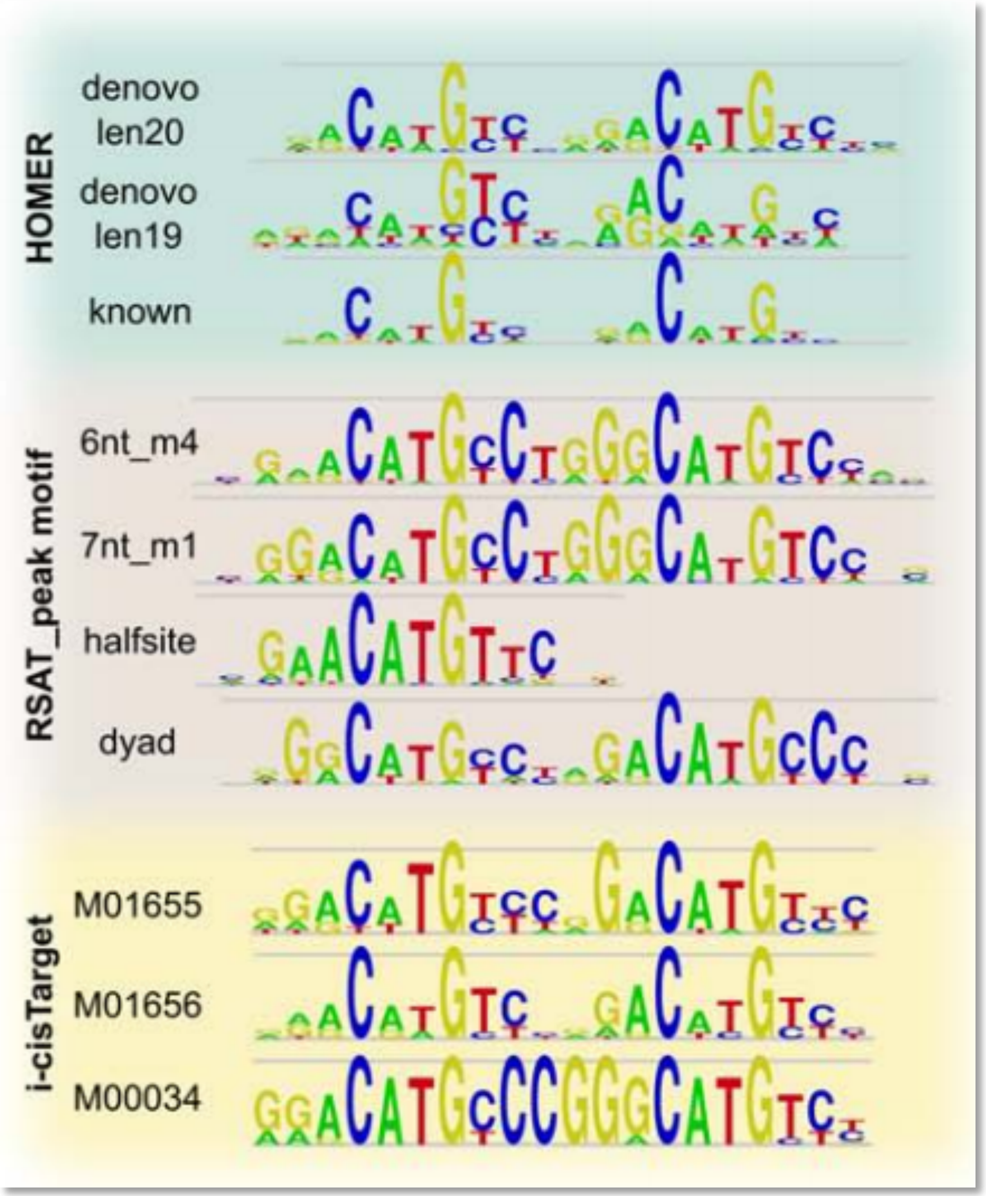
Sequence 1 $[X_{1,1} \quad X_{1,2} \cdots\cdots\cdots X_{1,p}]$

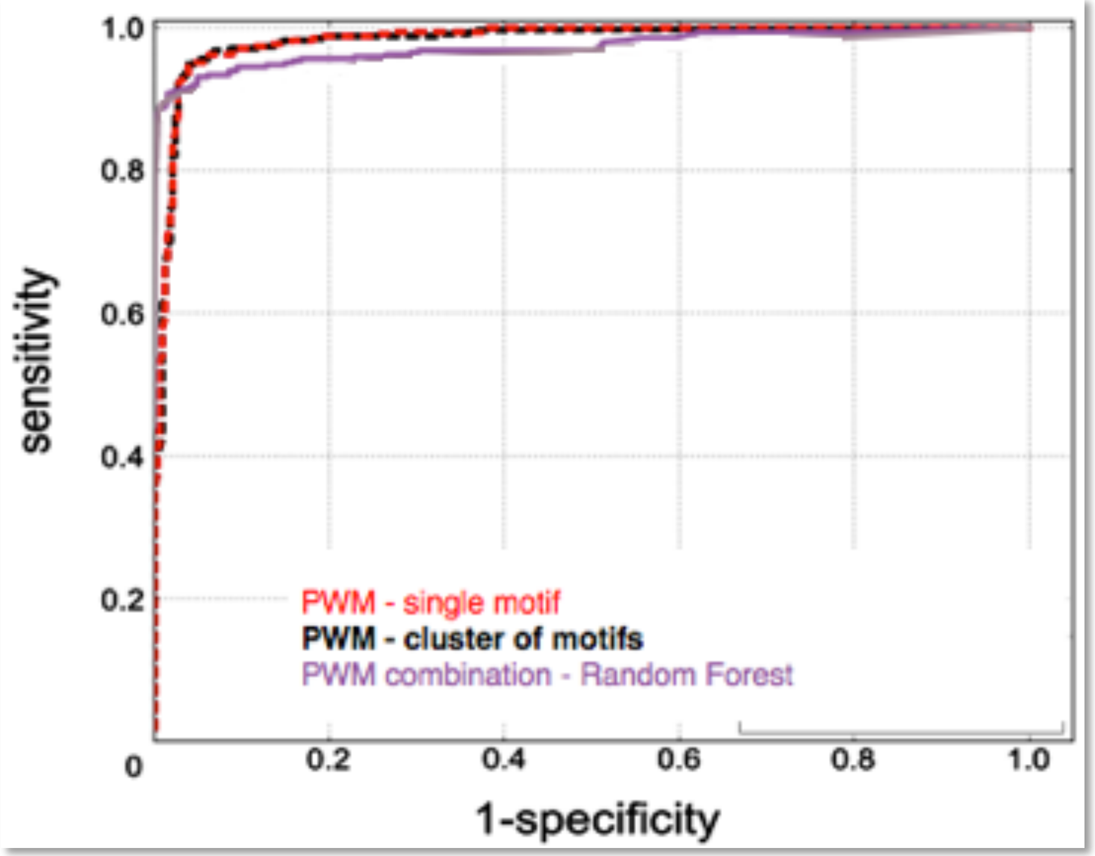
Sequence 2 $[X_{2,1} \quad X_{2,2} \cdots\cdots\cdots X_{2,p}]$

Sequence 3 $[X_{n,1} \quad X_{n,2} \cdots\cdots\cdots X_{n,p}]$

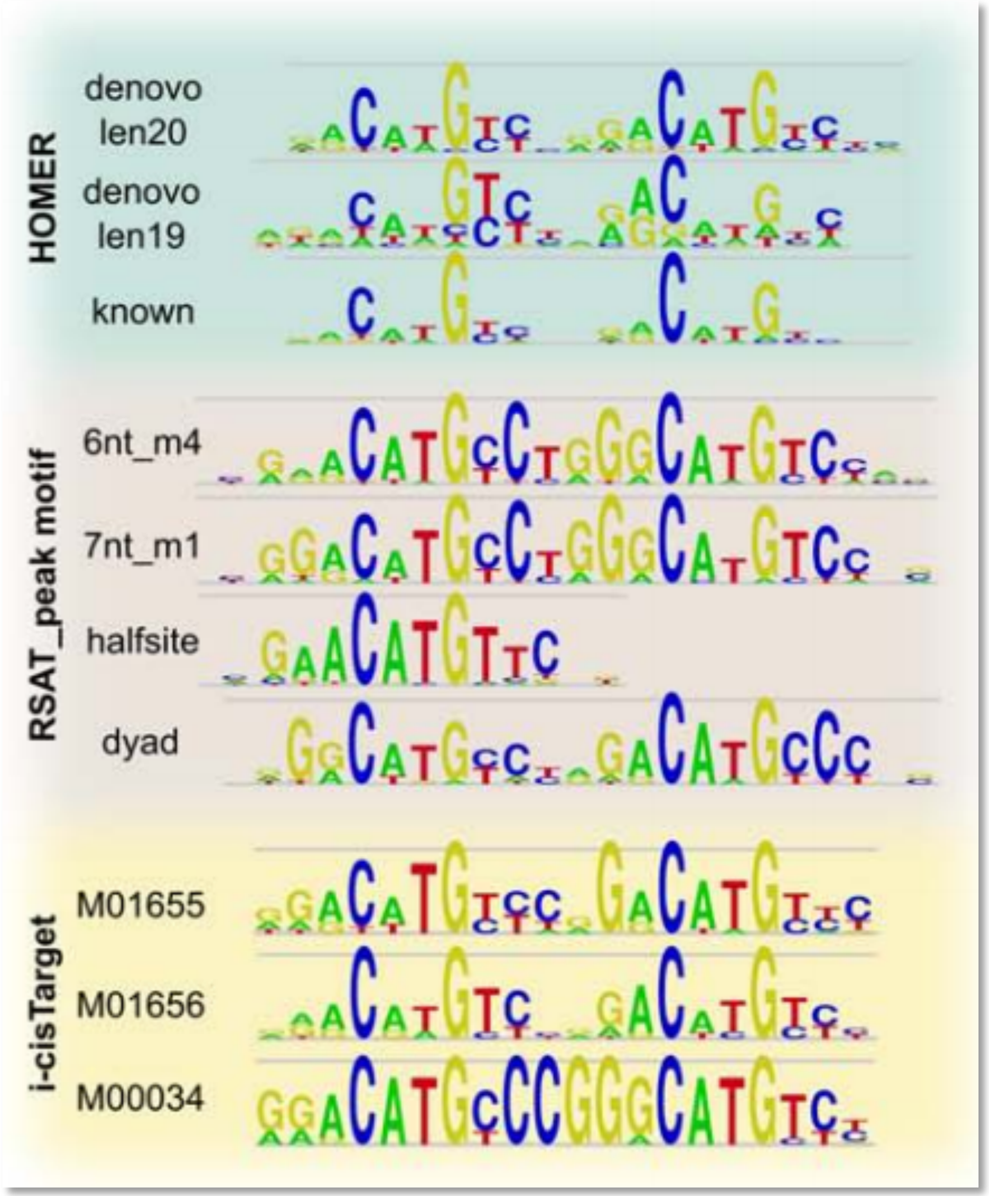


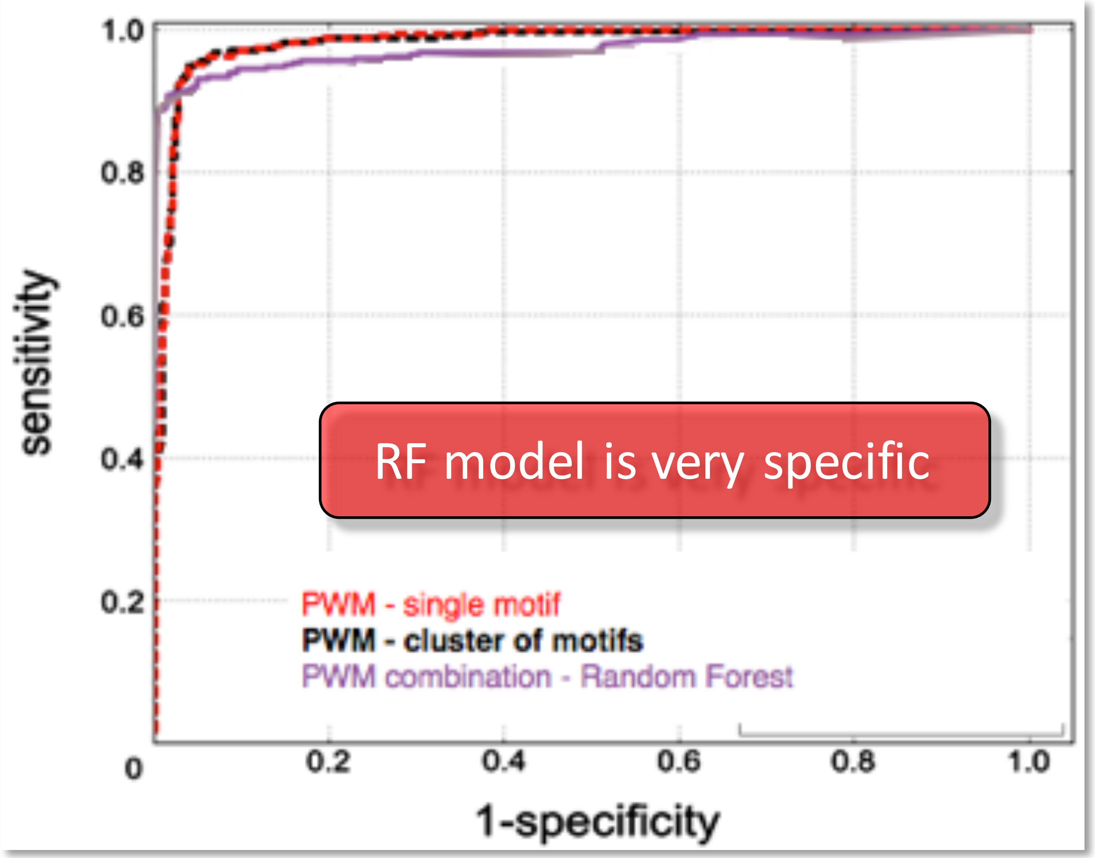
Enhancer classification using motifs





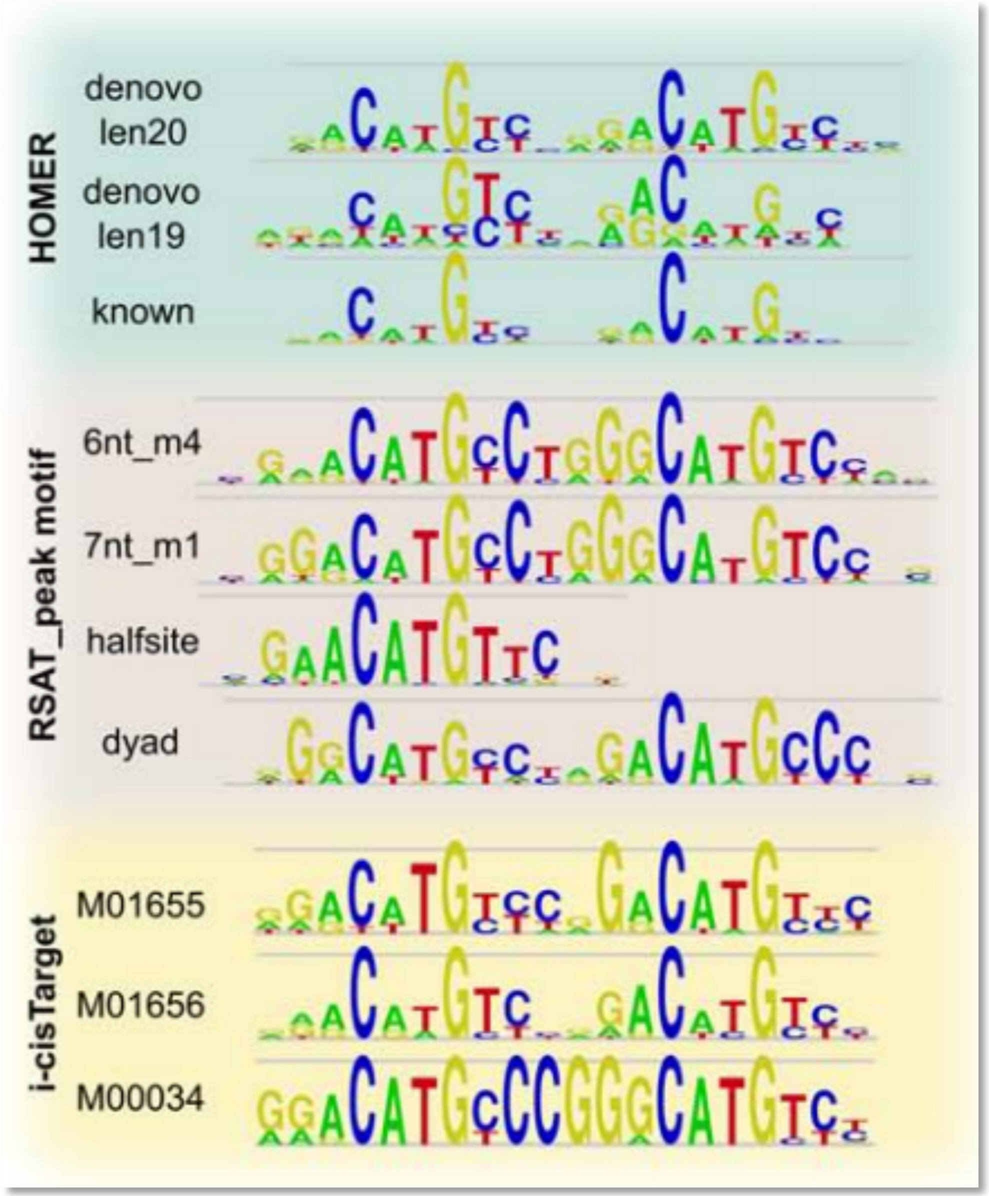
Enhancer classification using motifs

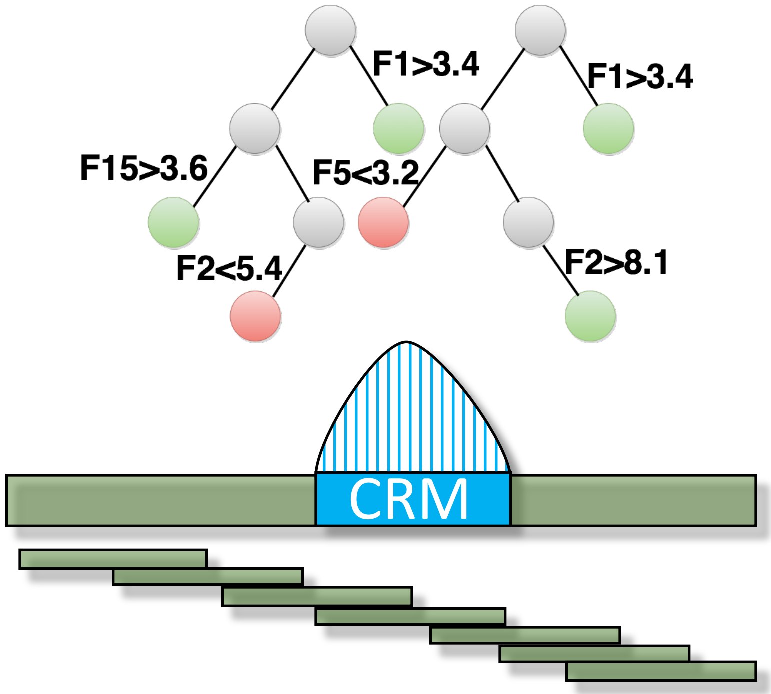




RF model is very specific

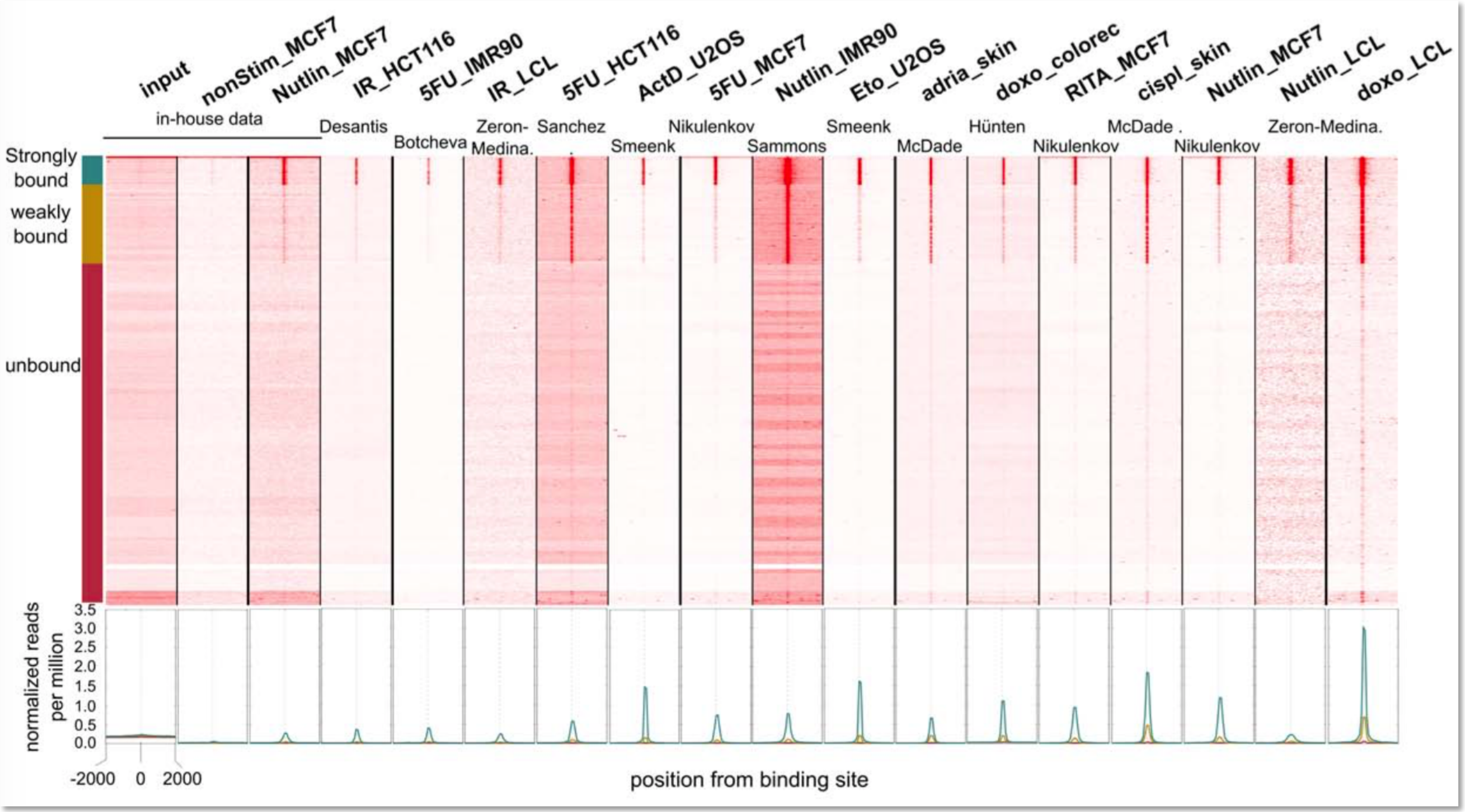
Enhancer classification using motifs



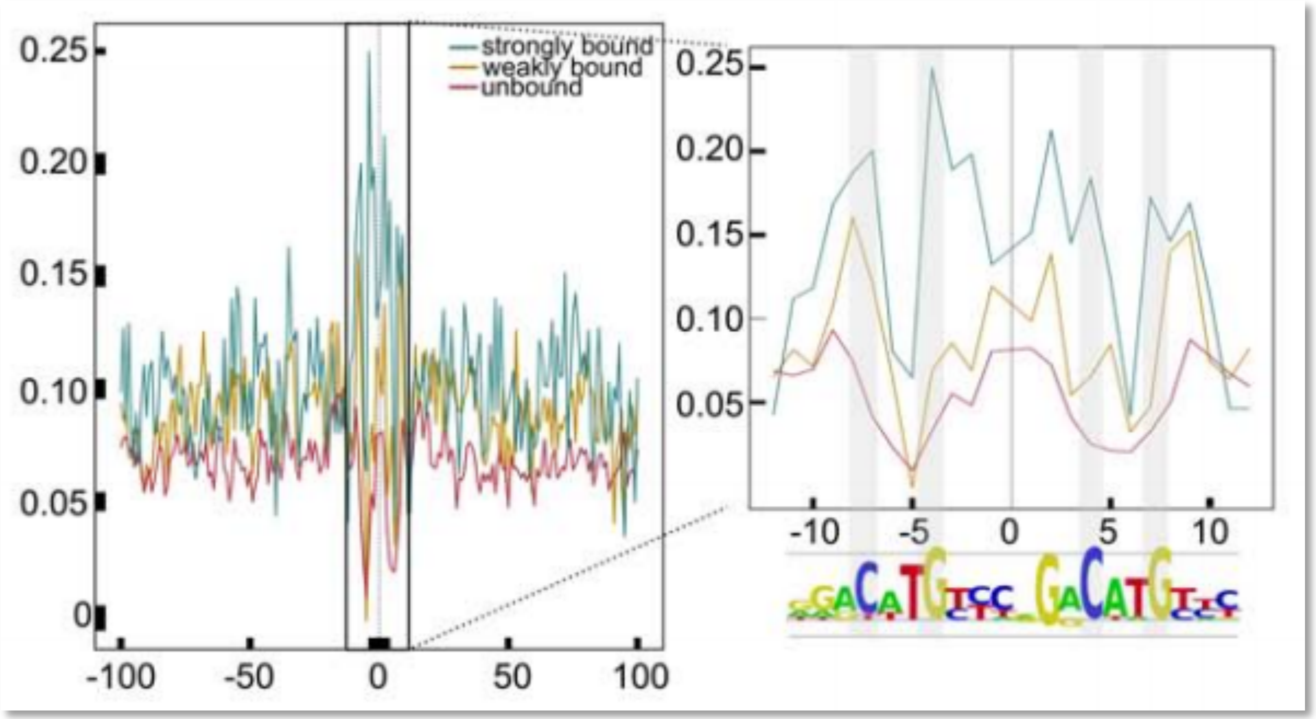


21K TP53 CRMs

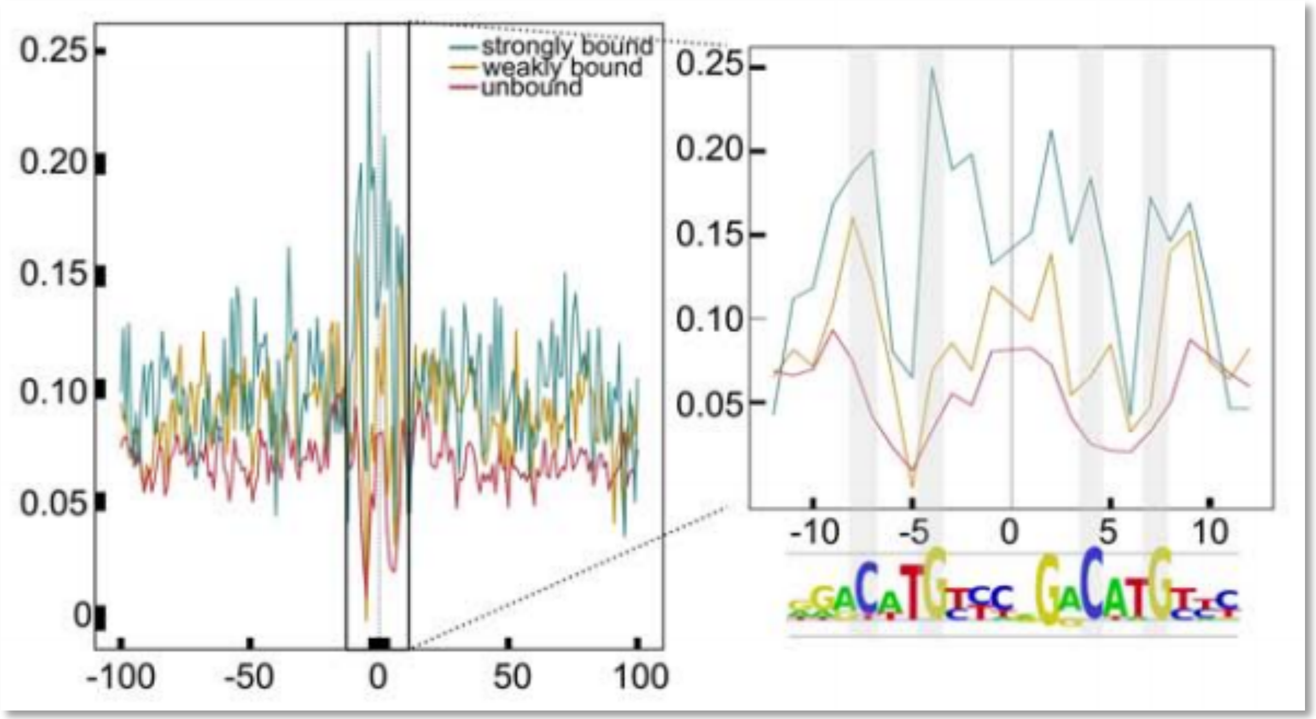
Discriminate bound versus unbound sites in the genome

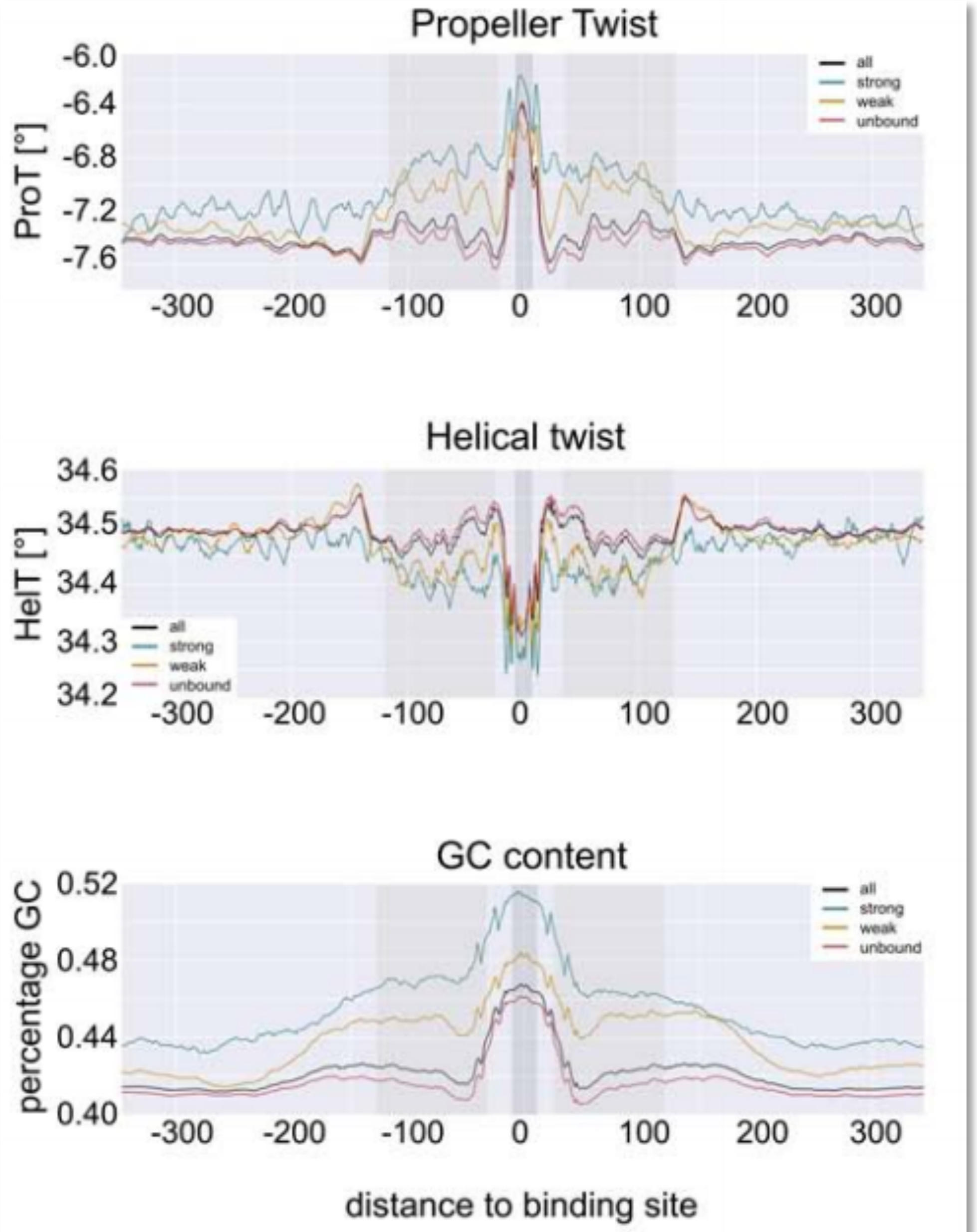


RF model to classify strongly bound vs unbound regions

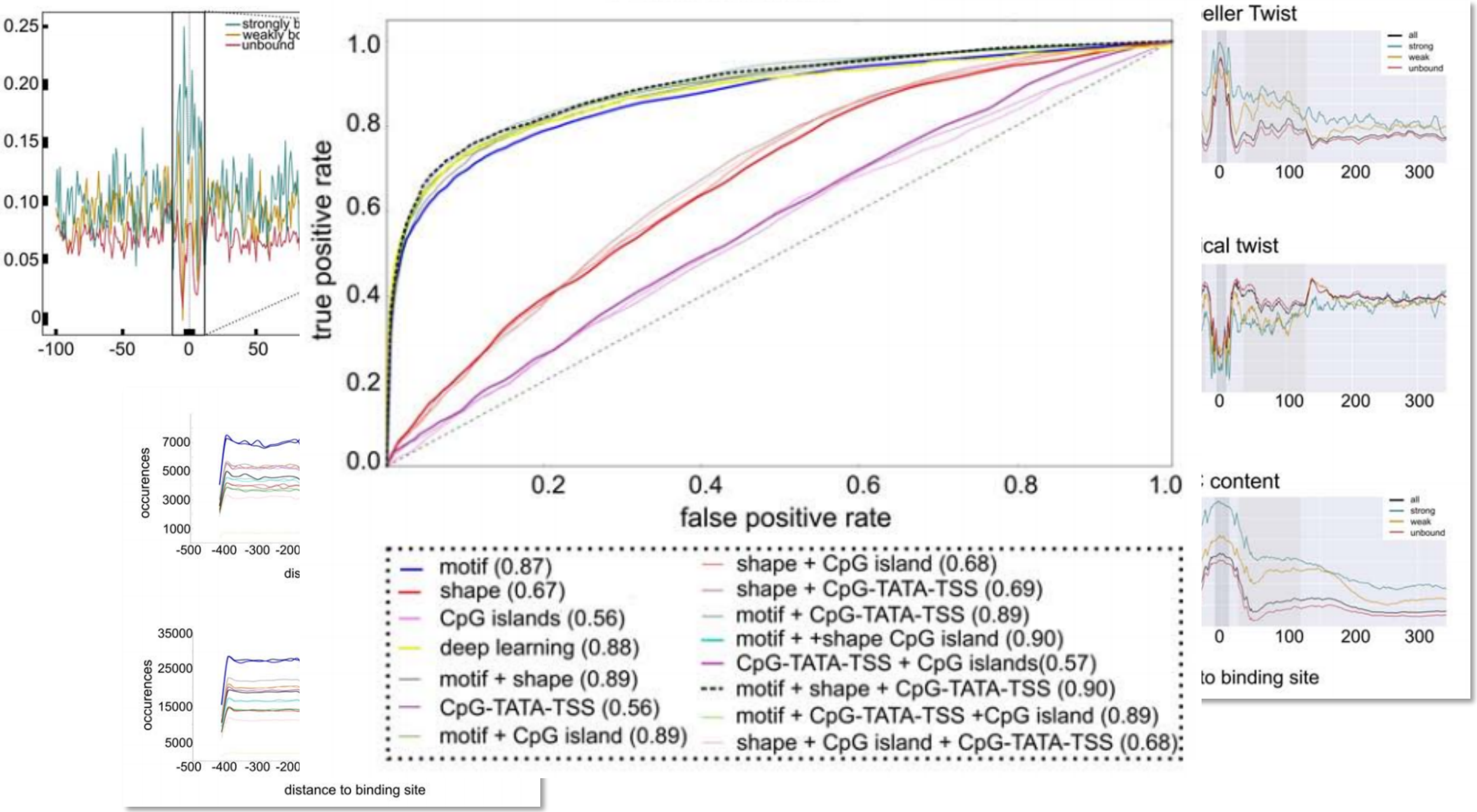


RF model to classify strongly bound vs unbound regions

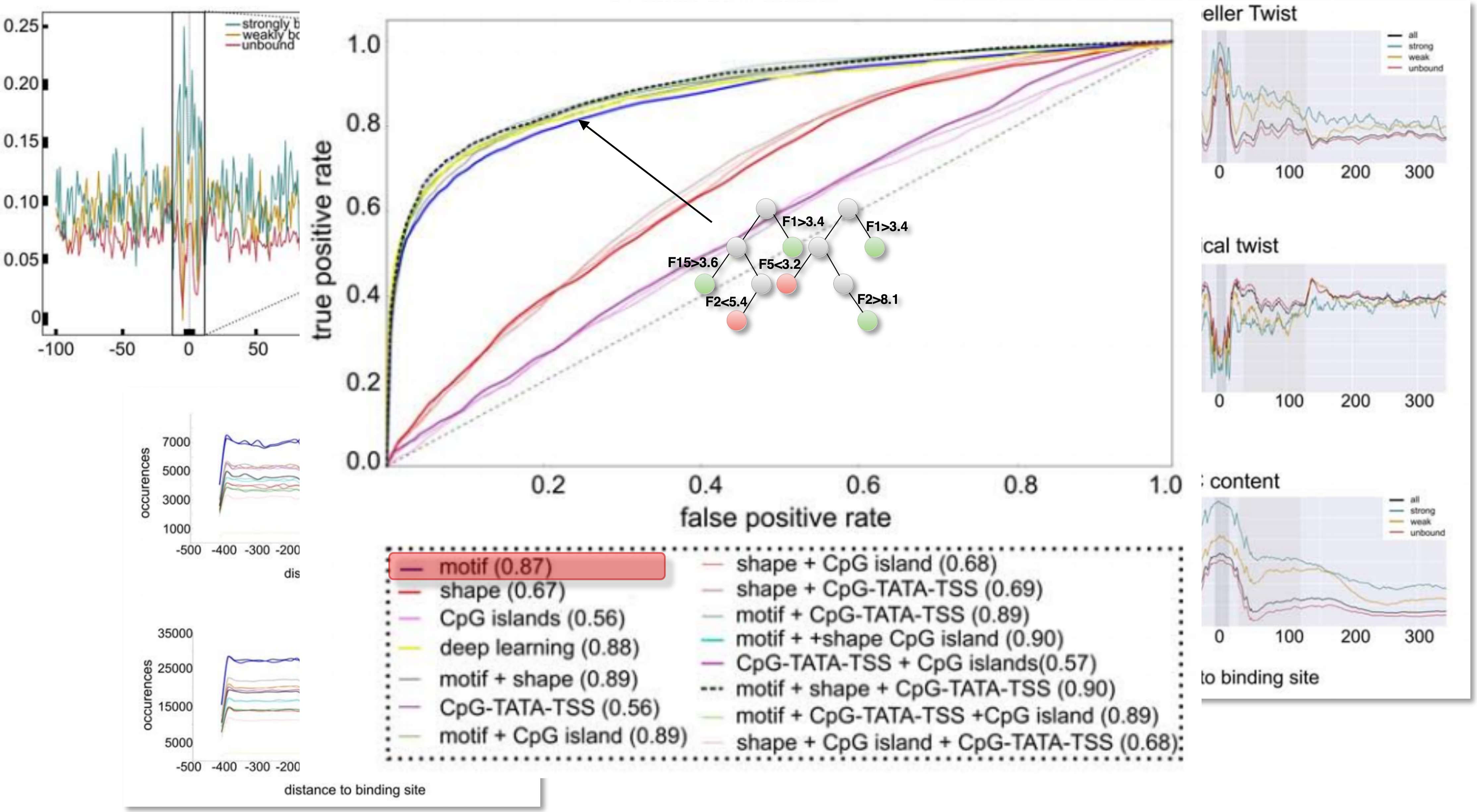




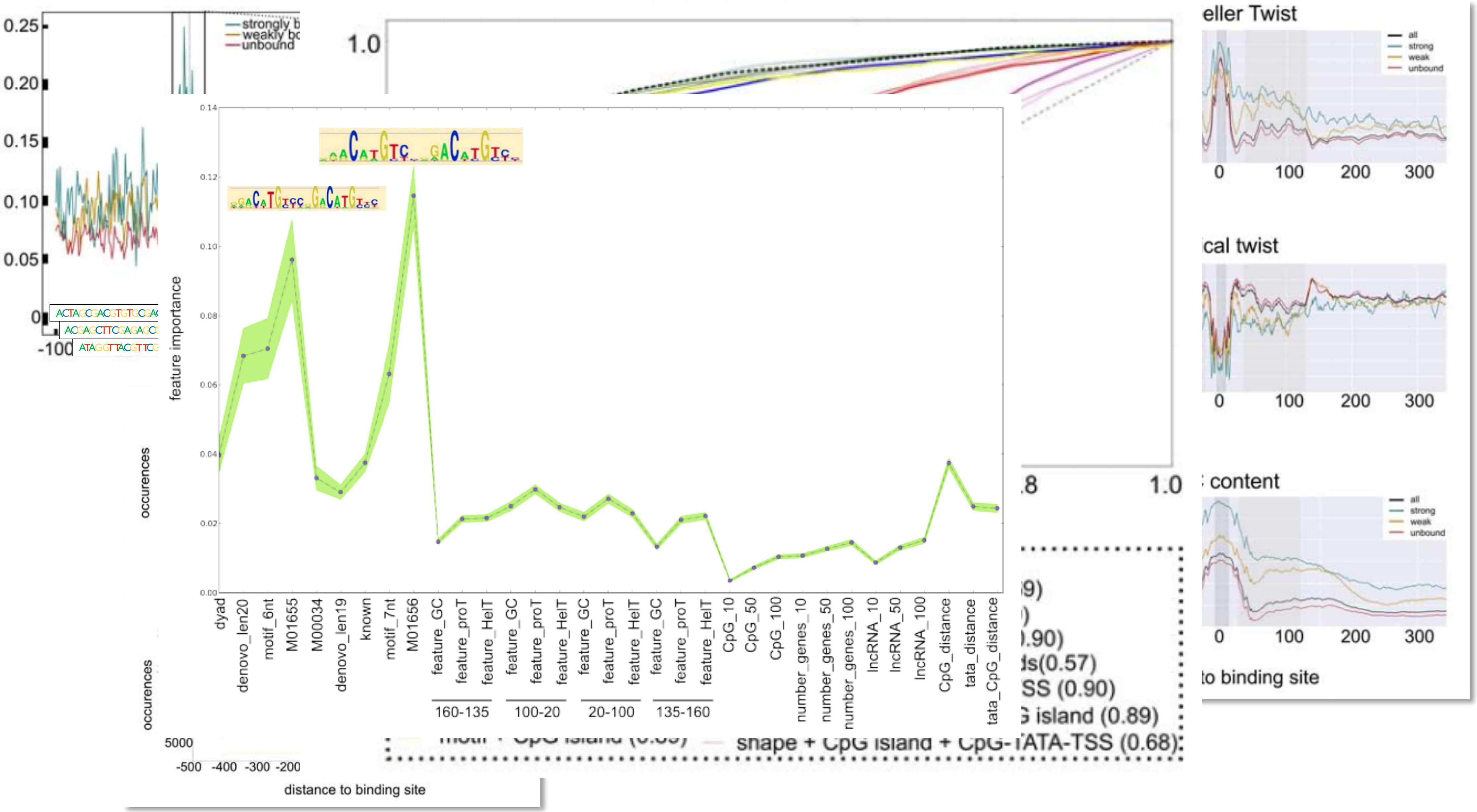
RF model to classify strongly bound vs unbound regions



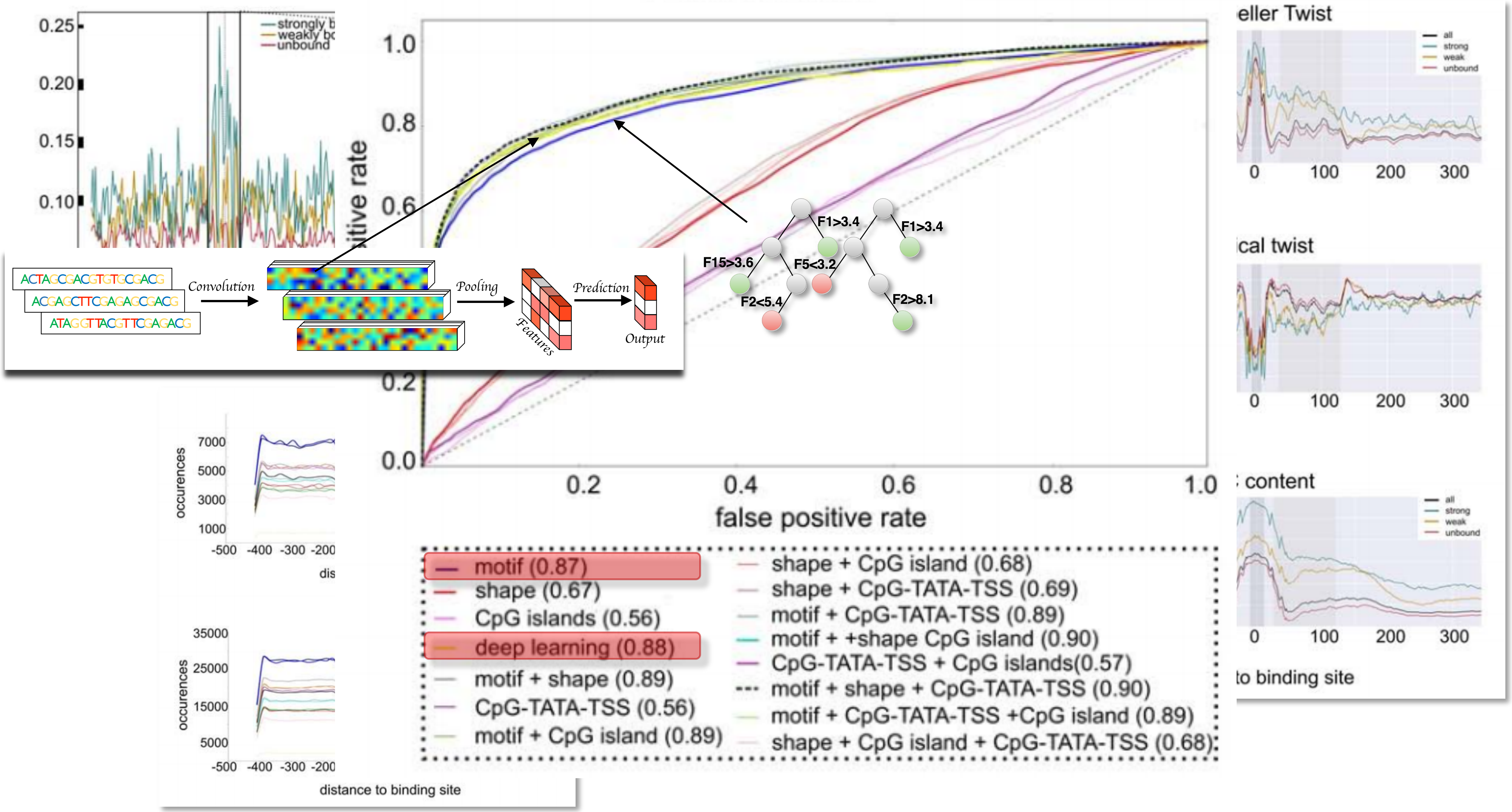
RF model to classify strongly bound vs unbound regions



RF model to classify strongly bound vs unbound regions

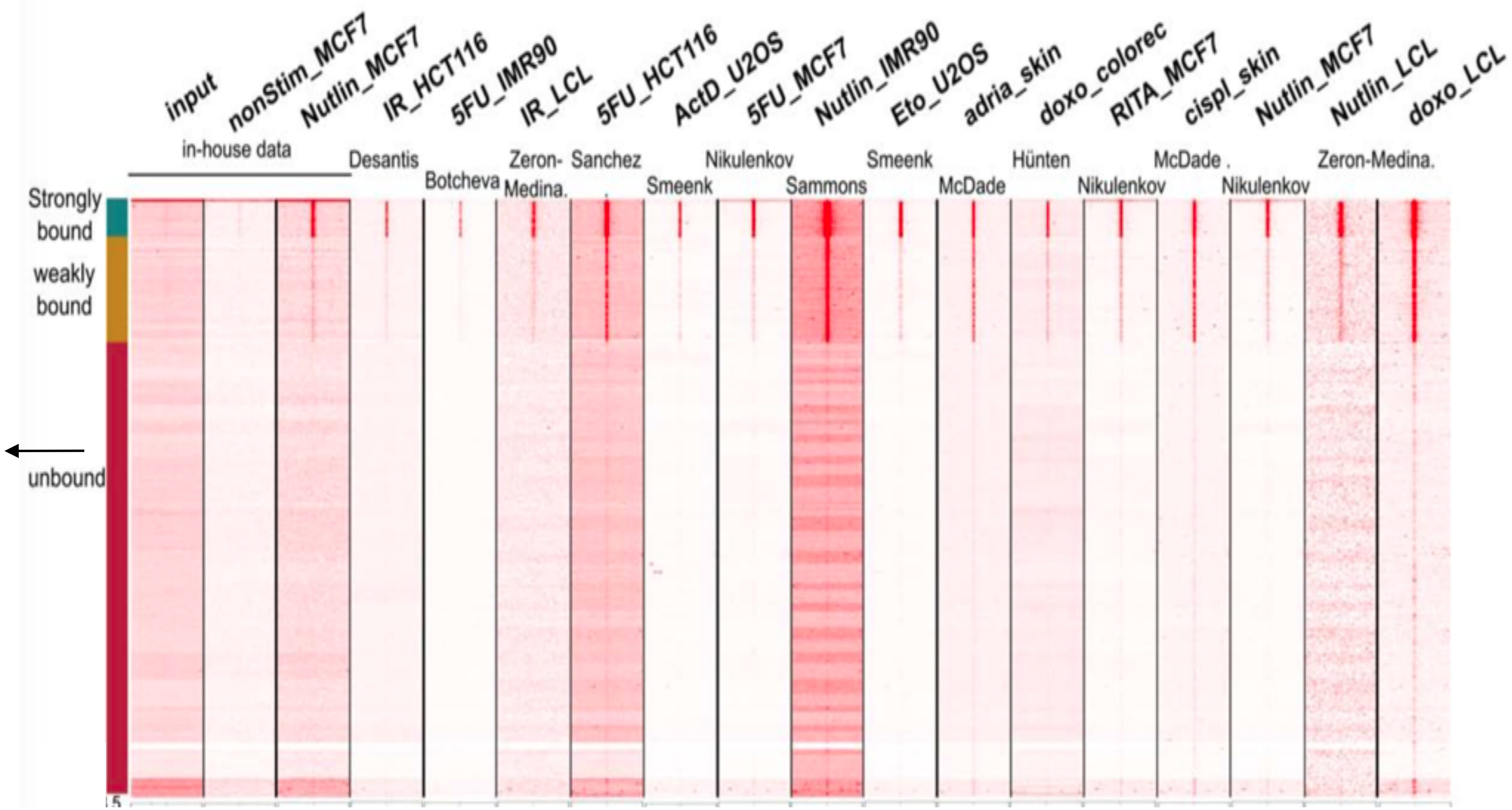


RF model to classify strongly bound vs unbound regions

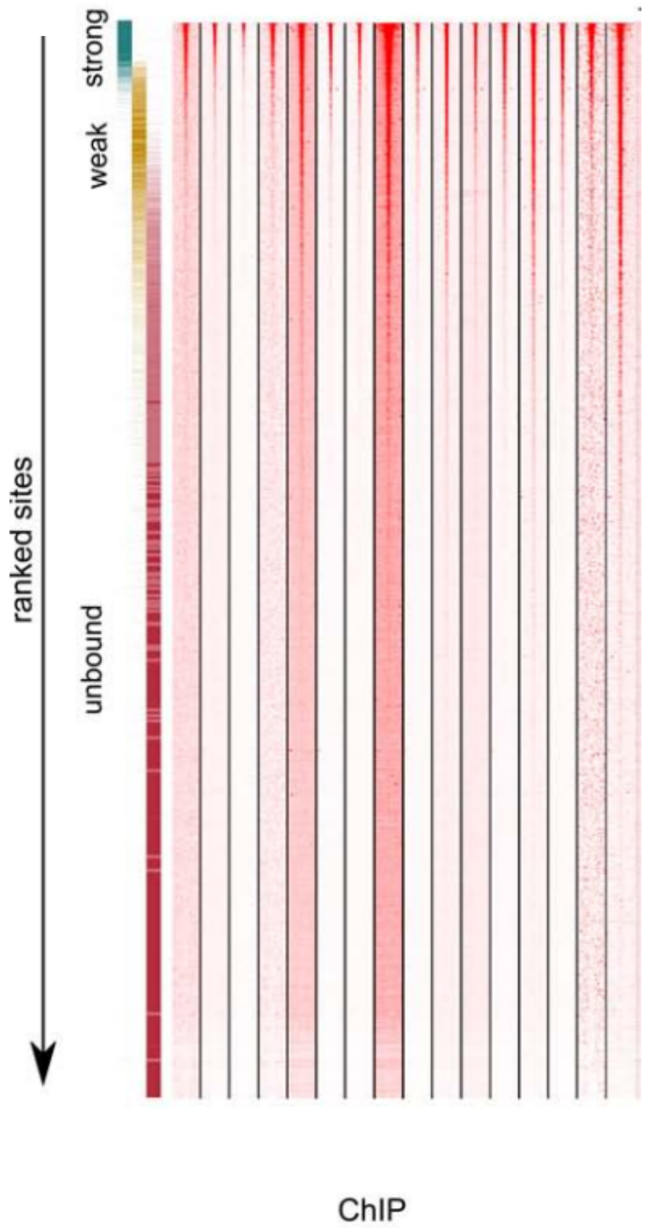


Strength of binding site predicts quantitative TP53 binding

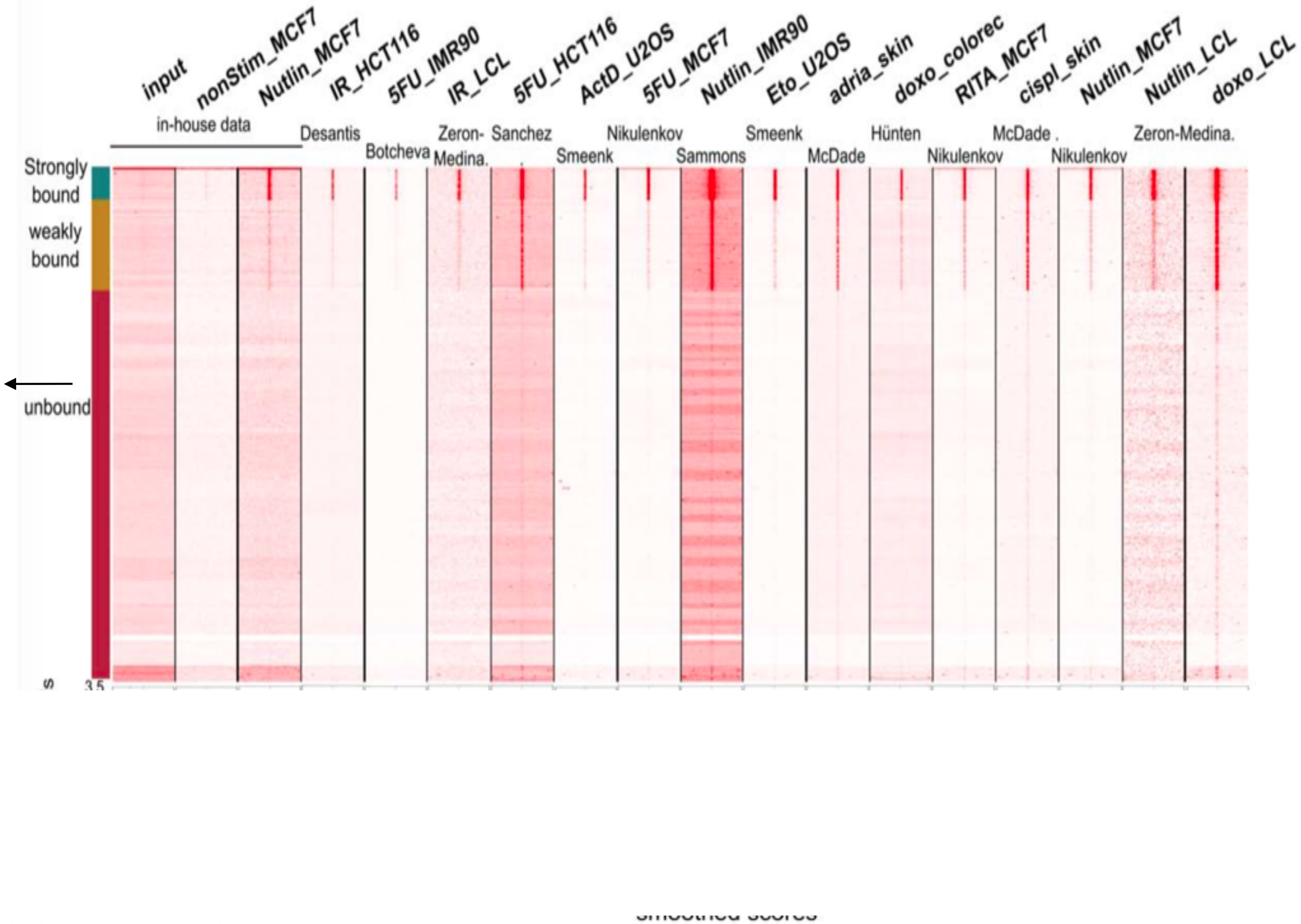
Rank aggregation



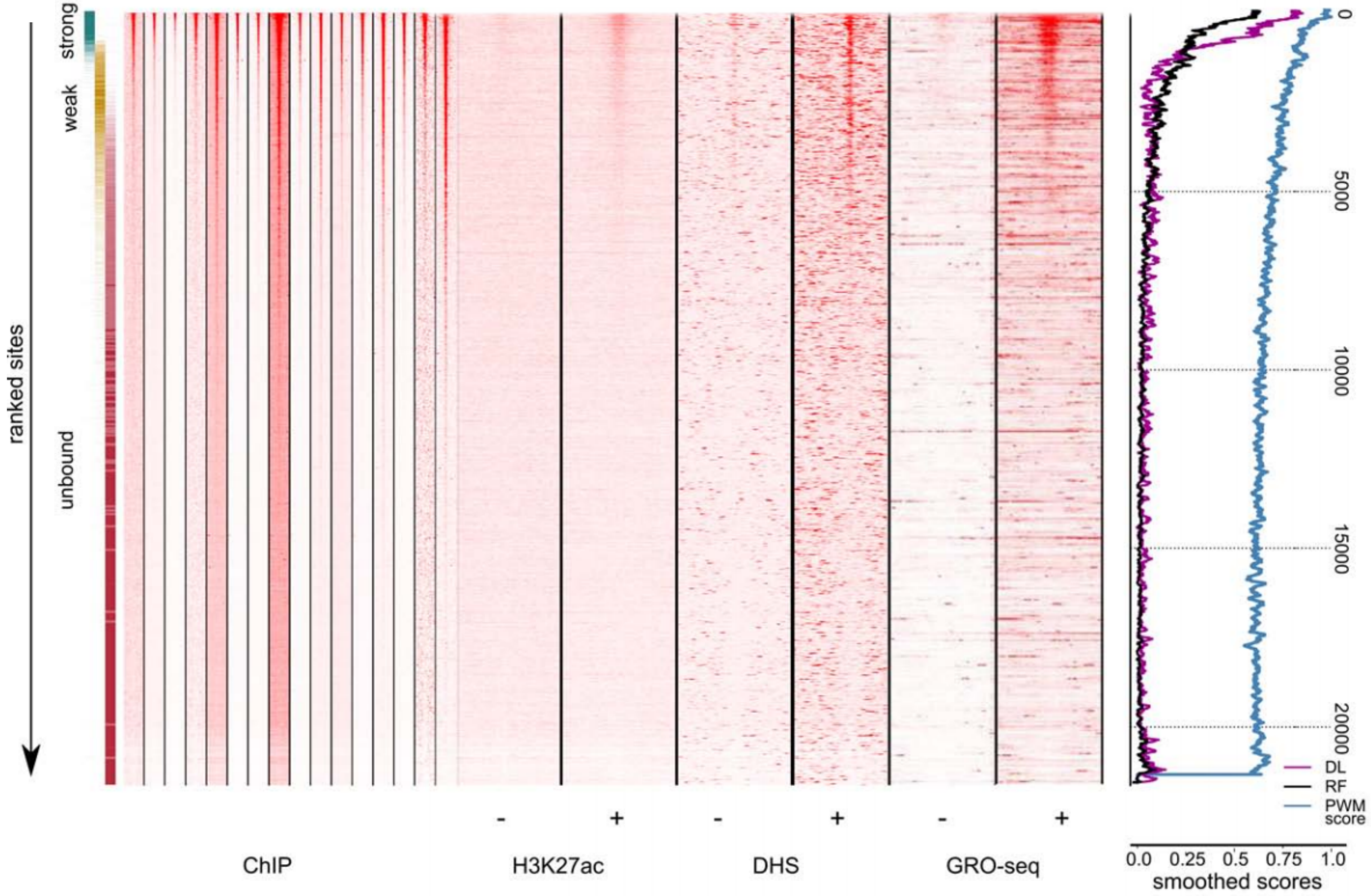
Strength of binding site predicts quantitative TP53 binding



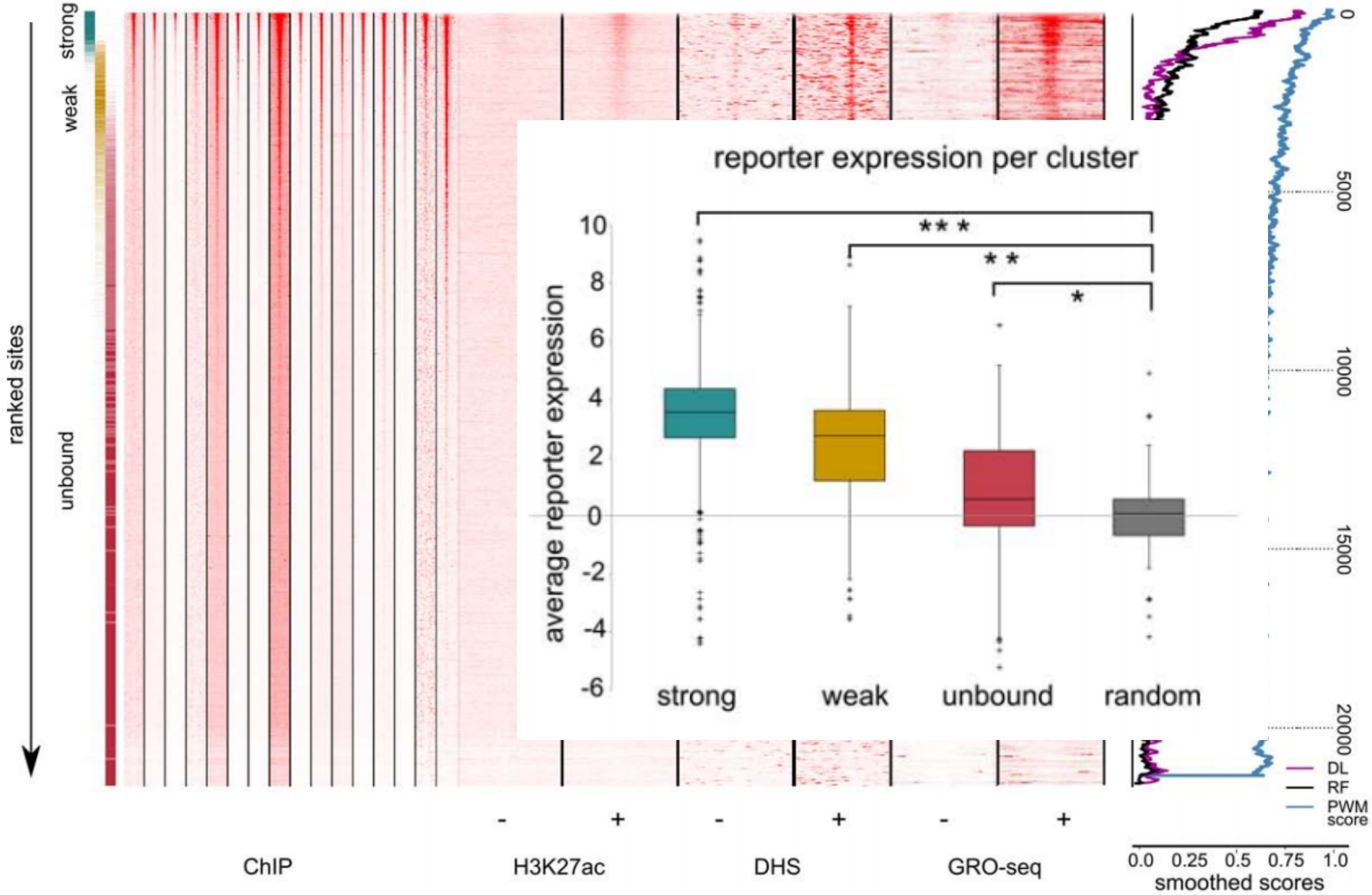
Rank aggregation



Strength of binding site predicts quantitative TP53 binding



Strength of binding site predicts quantitative TP53 binding



Conclusions

1) Only a small subset of experimentally determined binding events represents TP53 responsive elements

2) TP53 binds the DNA strictly as a tetramer, to a duplicate of the consensus palindromic responsive element

3) Strength of binding site predicts quantitative TP53 binding

4) TP53 acts on its own, without co-regulatory transcription factors that bind to the same enhancer

5) TP53 is activator but not a repressor