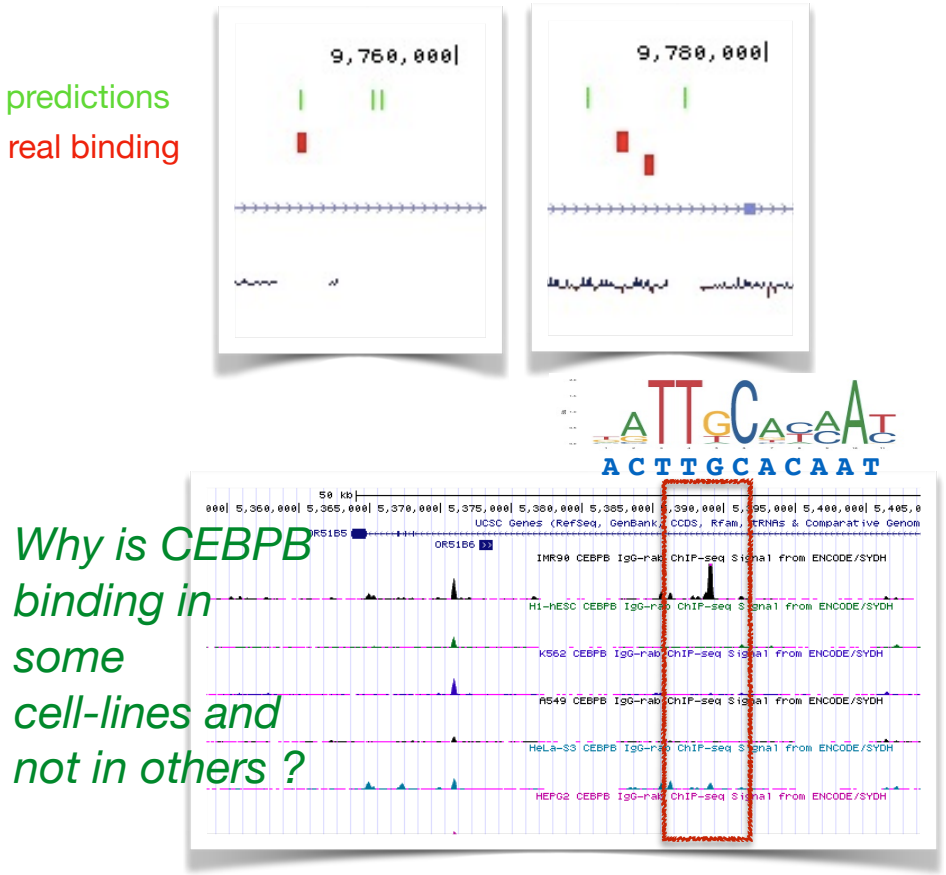
4. improving predictions

Problem of sequence-only predictions

- **Large number of false-positive/ false-negative**
  the sequence looks like a binding site, but the TF is not binding!

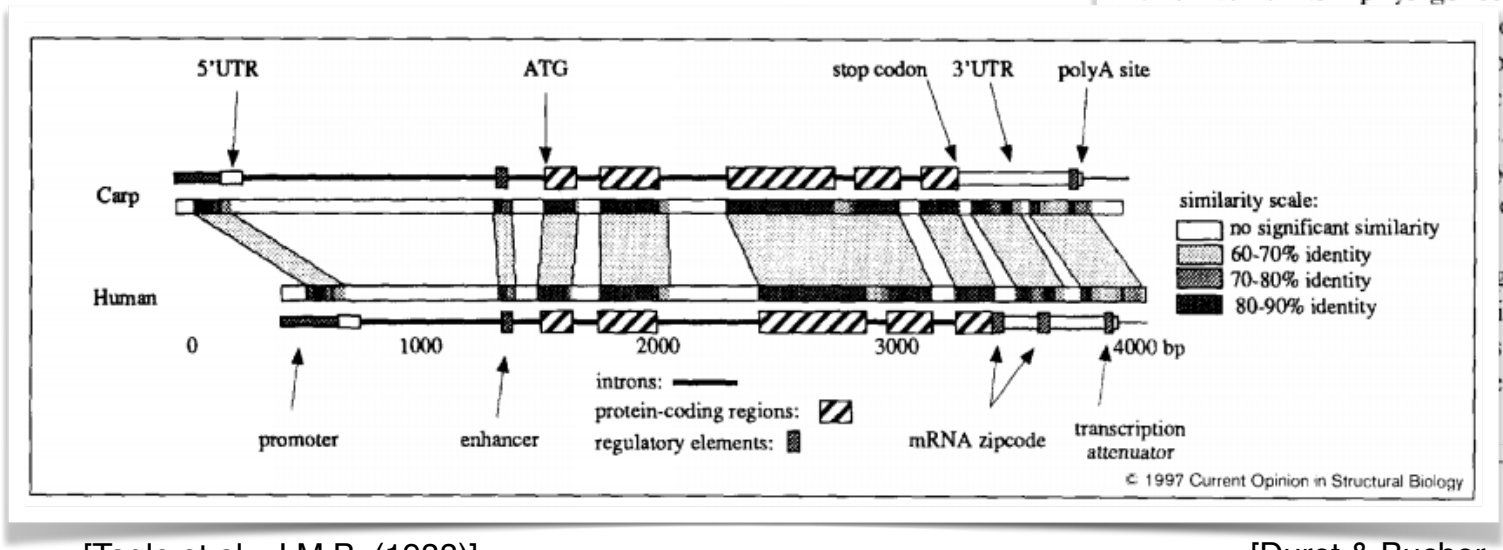
- **Cellular/tissue-context not taken into account**
  a TF might bind in one tissue, but not in another (but the sequence is the same...)



Can we reduce/optimize the search space for regulatory elements?

Phylogenetic footprinting

- Tagle et al. (1988) : study of the promoter of globin genes in vertebrates identifies **conserved regulatory elements**



[Tagle et al., J.M.B. (1988)]

Phylogenetic footprinting

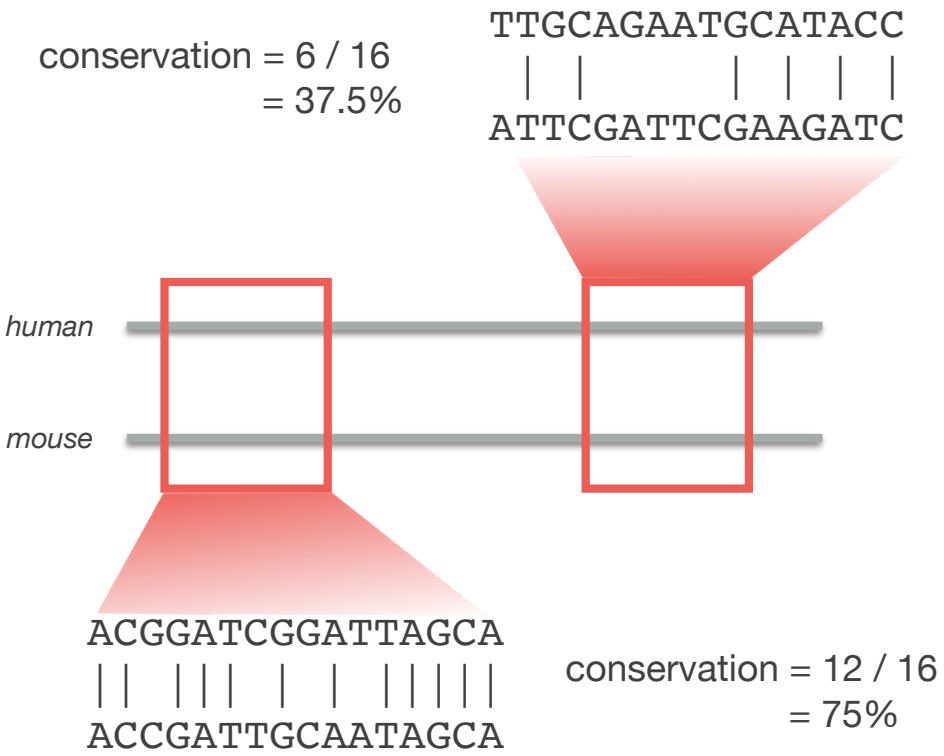
The pattern of mutations that have occurred during evolution is an excellent indicator of functional constraints. Genomes continually undergo mutations, but the outcome of each mutation depends on its phenotypic effect. Mutations that are deleterious are generally eliminated by natural selection, whereas mutations that have no phenotypic effect (neutral mutations) or that are only slightly deleterious can be randomly fixed in the population (genetic drift). The consequence of this is that mutations accumulate much faster at nonfunctional DNA bases than at functionally constrained base positions. Hence, if one detects a sequence that has remained highly conserved during evolution, then it probably means that this sequence is functional (but the reverse proposal is not true: a sequence can be functional albeit nonconserved). Tagle *et al.* [31] proposed the term 'phylogenetic footprinting' to describe the phylogenetic comparisons that reveal

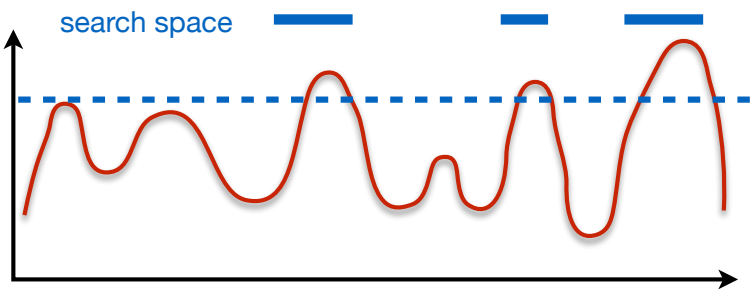
onal elements in homologous
hylogenetic footprinting is
h shows the comparison of
. This comparison shows that
yrs) of divergence (450 Myrs
ete elements in noncoding
conserved. Indeed, these
gions correspond to essential
involved in transcription and
(Fig. 1). Thus, the simple
quences can reveal essential

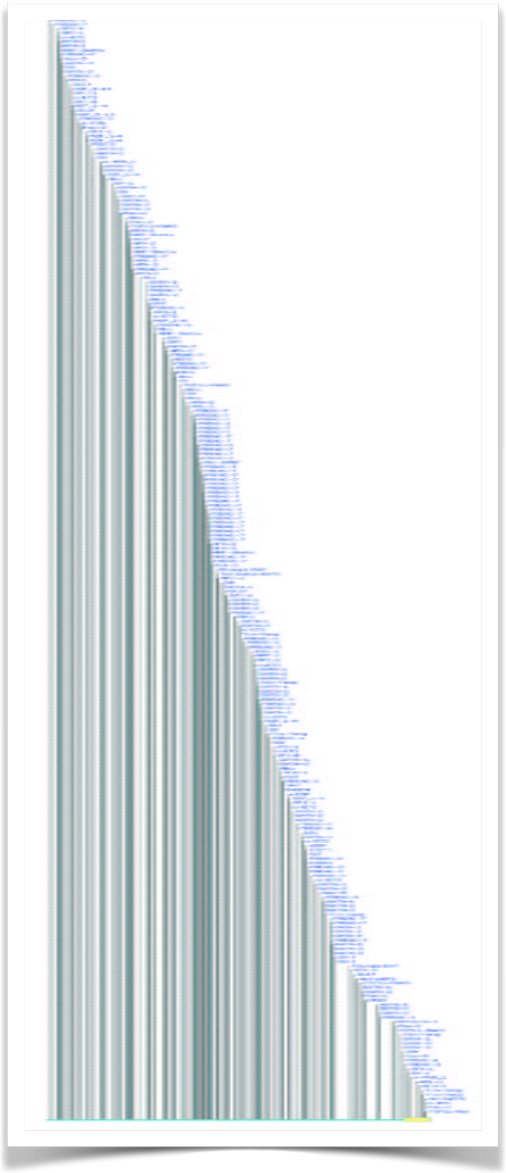
[Duret & Bucher, Curr.Op.Str.Biol. (1997)]

Phylogenetic footprinting

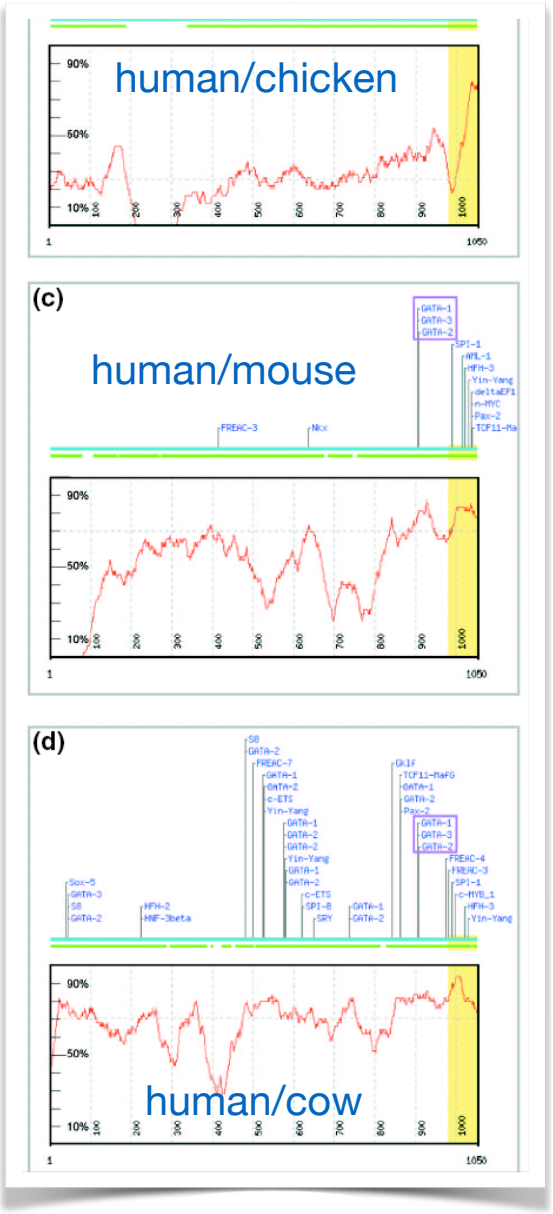
- Starting point : alignment of **2 orthologous regions** (e.g. promoter of orthologous genes)
- Compute the **conservation** inside a sliding window (number of conserved positions divided by length)
- TFBS search using PWM (fixed threshold)
- **Only TFBS inside highly conserved regions** are retained !
- *Choice of organisms to be compared is crucial !*

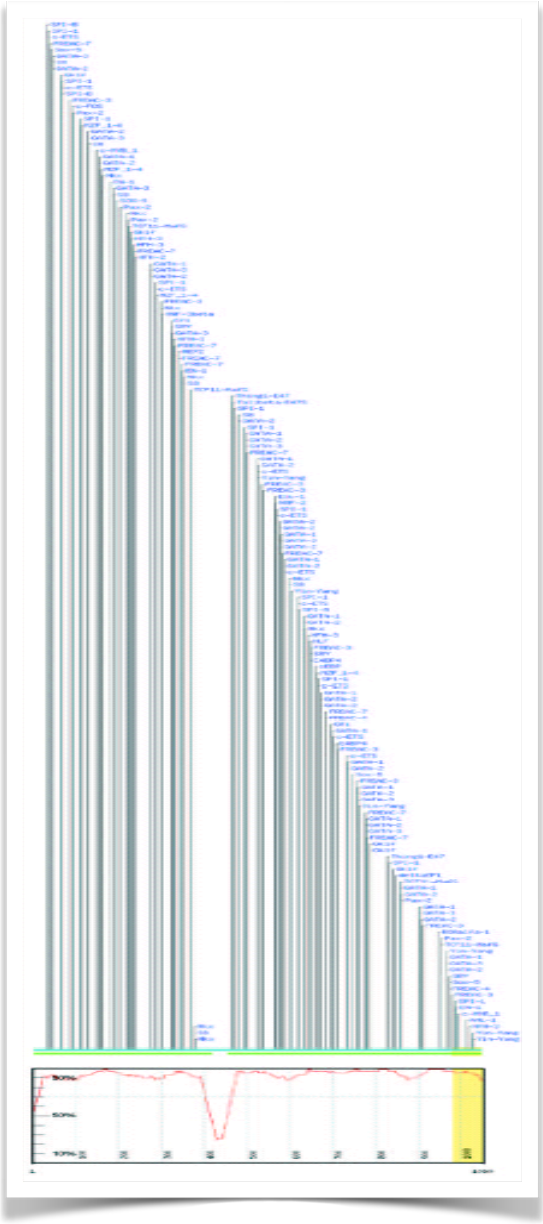






Human globin gene promoter alone



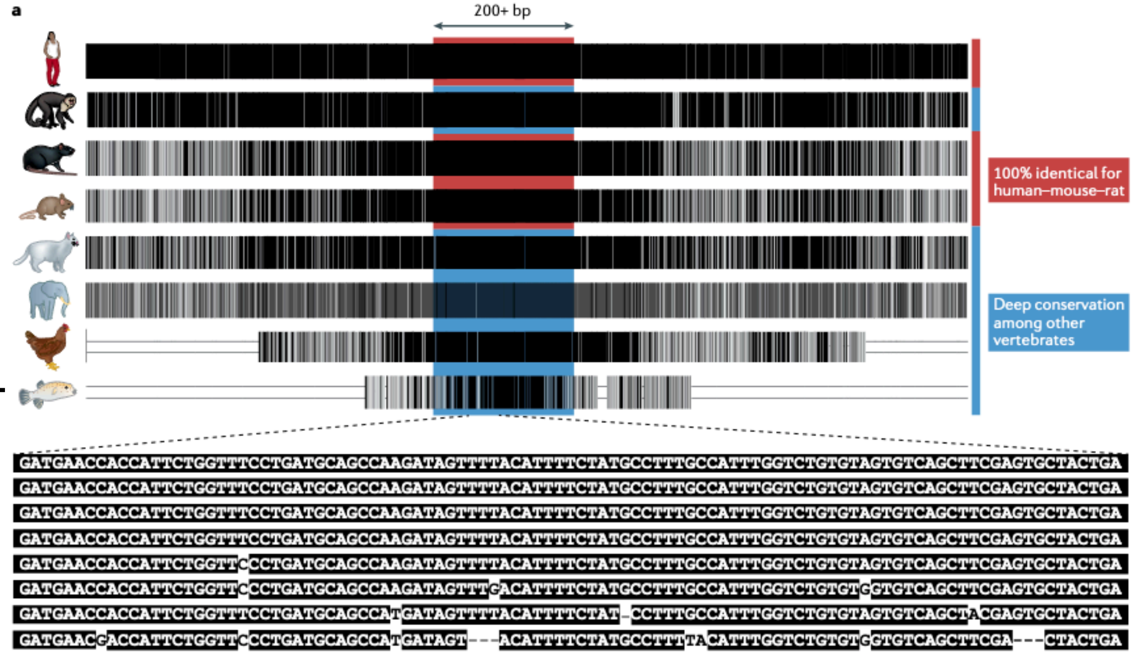


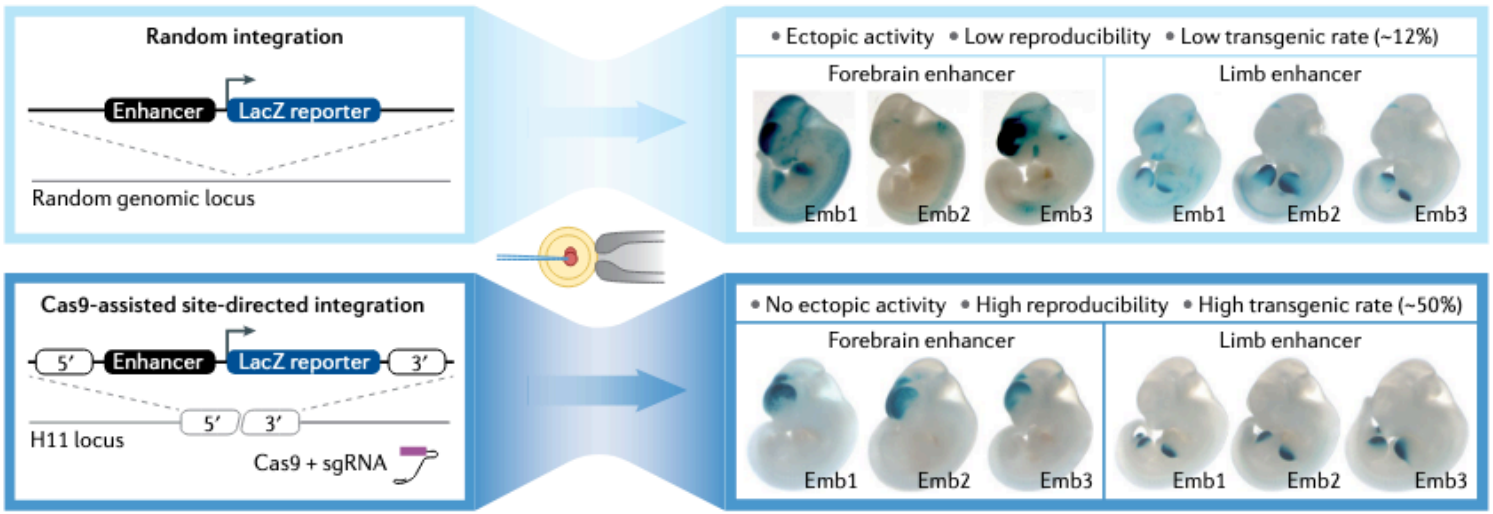
Alignment of human/macaque globin promoter

Deeply conserved non-coding elements

Ultra-conserved elements
perfect conservation over 200 bp between human and mouse/rat

~450 million years

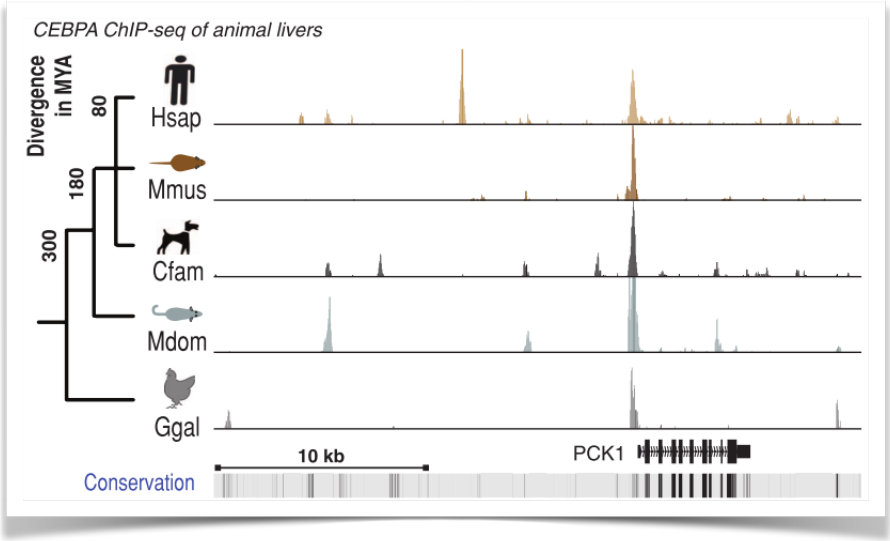


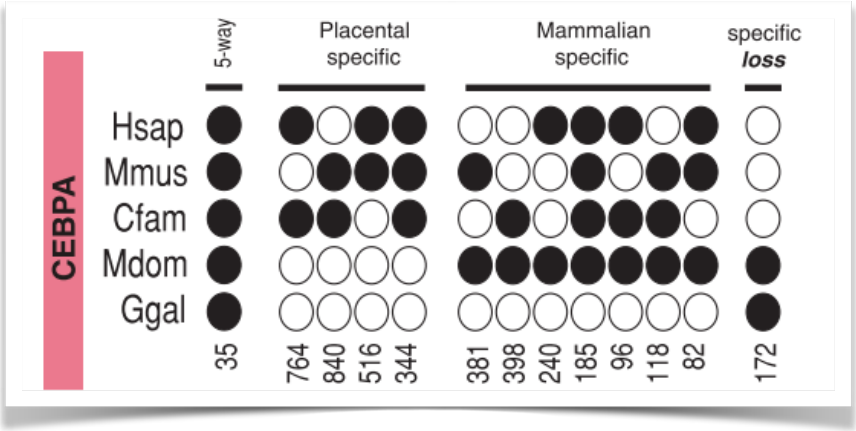


[Snetkova, Nature Rev.Gen. (2020)]

Are TFBS really conserved ?

- Study of TFBS for 2 liver specific TF : CEBPa and HNF4a
- 5 species
  - 3 placental mammals (human, mouse dog)
  - opossum + chicken
- ChIP-seq against both factors in all species
- Take home message : **a minority of binding events are shared by all species ; most are species/clade specific**

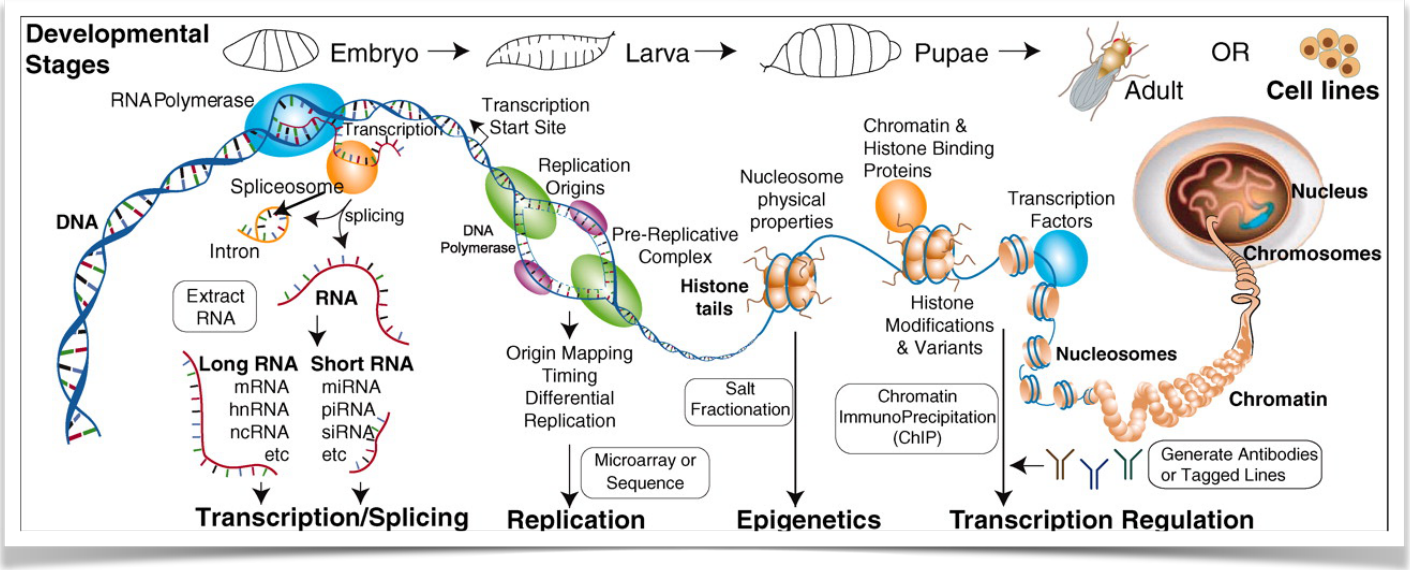




[Schmidt, Wilson Ballester et al., Science (2010)]

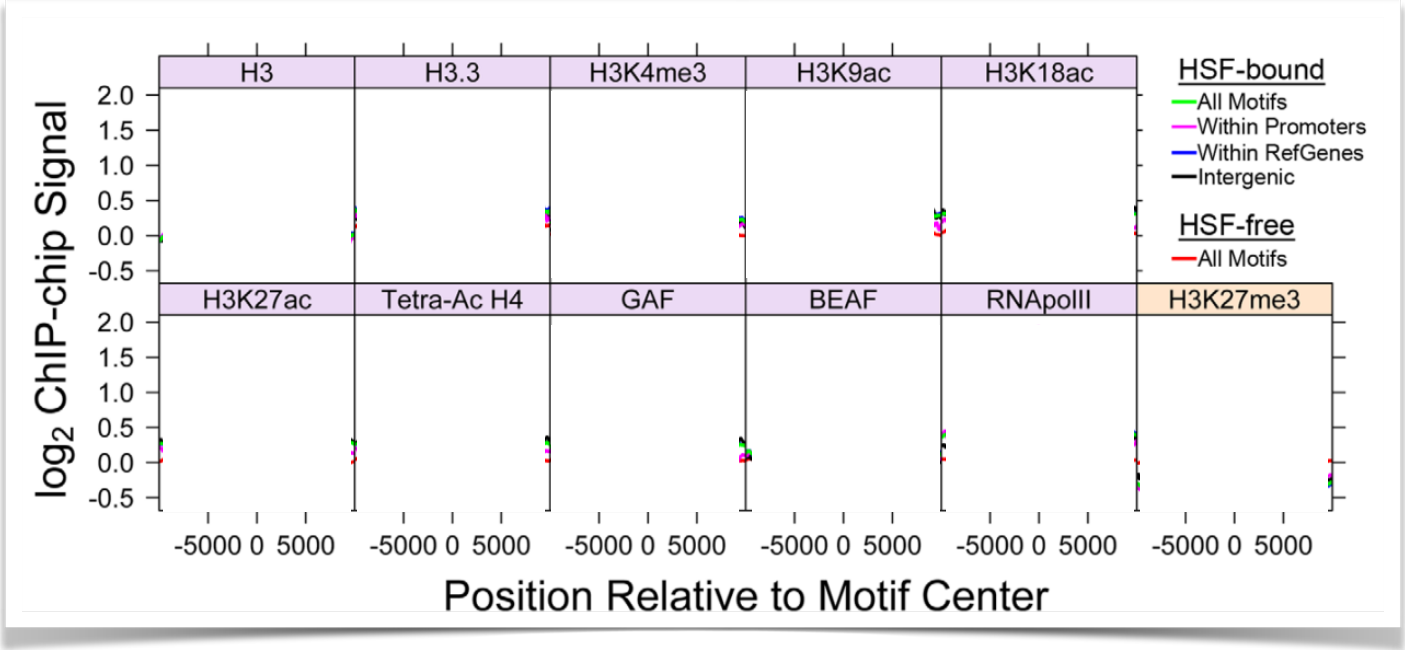
In-vivo regulatory features

- Particular in-vivo chromatin states seem to be correlated with gene activation
  - **p300/CBP** binding (acethyl-transferase) [Visel et al., Nature (2010)]
  - methylation of histone 3 Lysine 4 (**H3K4me1**) [Heintzman et al., Nature (2007)]
  - acetylation of histone 3 Lysine (**H3K9ac, H3K27ac**,...)
  - presence of **Pol II** binding at active enhancers
  - presence of modified form of H3 → **H3.3**
  - DNA accessibility (**DNAse hypersensitive** sites; **ATAC-seq**) [Pique-Regi et al., Gen. Res (2010); Buenrostro (2014)]



Motifs are not always binding events

- Compare **in-silico** TFBS to **in-vivo** binding event using ChIP data
- some bona fide motifs are not bound in-vivo : why ?
- example in Drosophila : heat-shock factor (HSF)
  - 464 ChIP peaks containing a HSF-motif ($p < 0.001$)
  - 708 unbound motifs (with $p < 5e\text{-}6$)

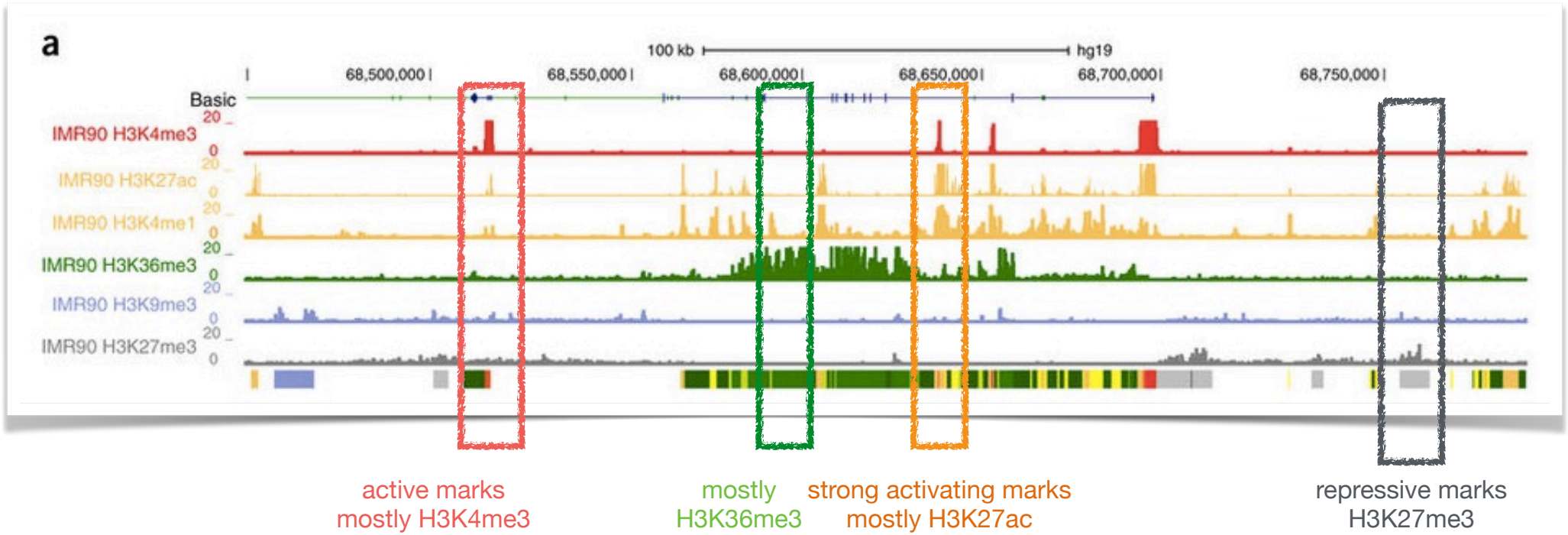


Bound sites have:

- **high levels of lysine acetylation**
- **high levels of polII binding**
- **low levels of H3K27me3 (repressive mark related to polycomb repression)**

[Guertin et al., PLoS Gen. (2010)]

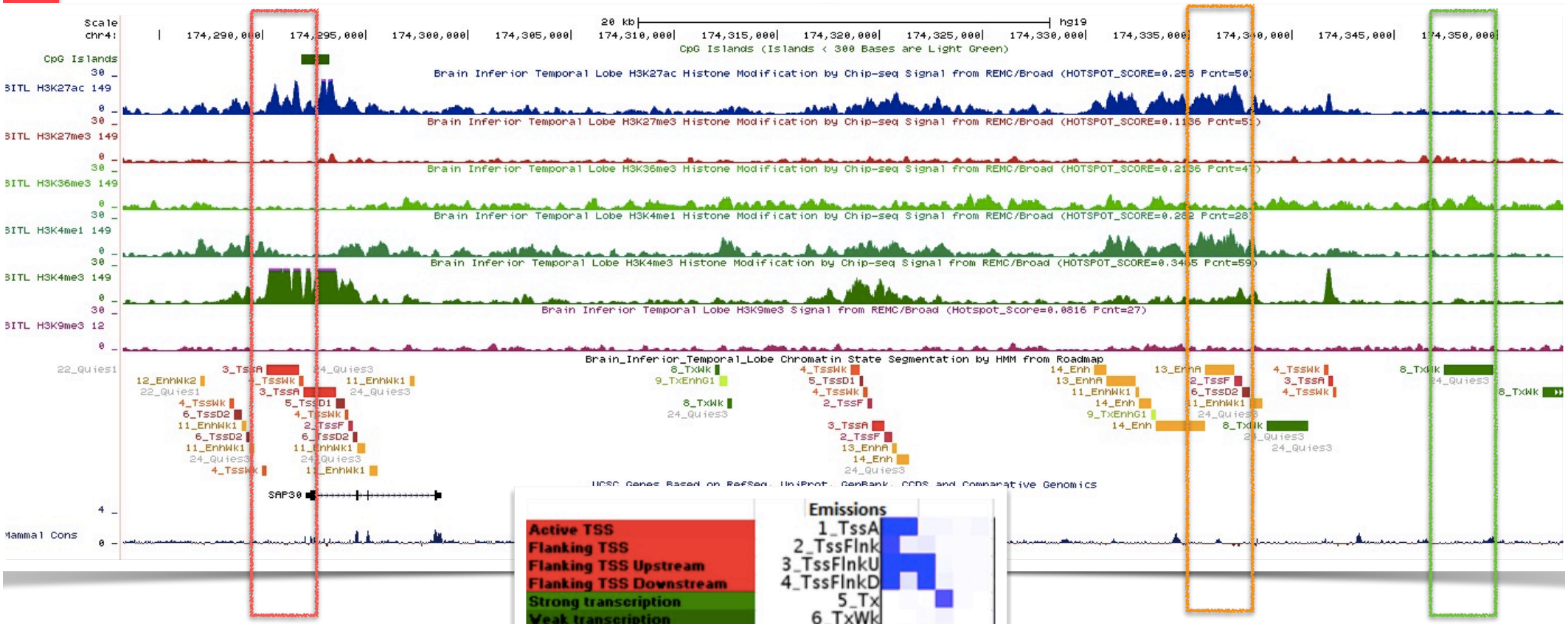
Histone code



Histone modifications appear to occur in specific combinations related to functional impact → **combinatorial chromatin states**

How can we define/annotate these **chromatin states** ?
→ **Hidden Markov model**

Chromatin states

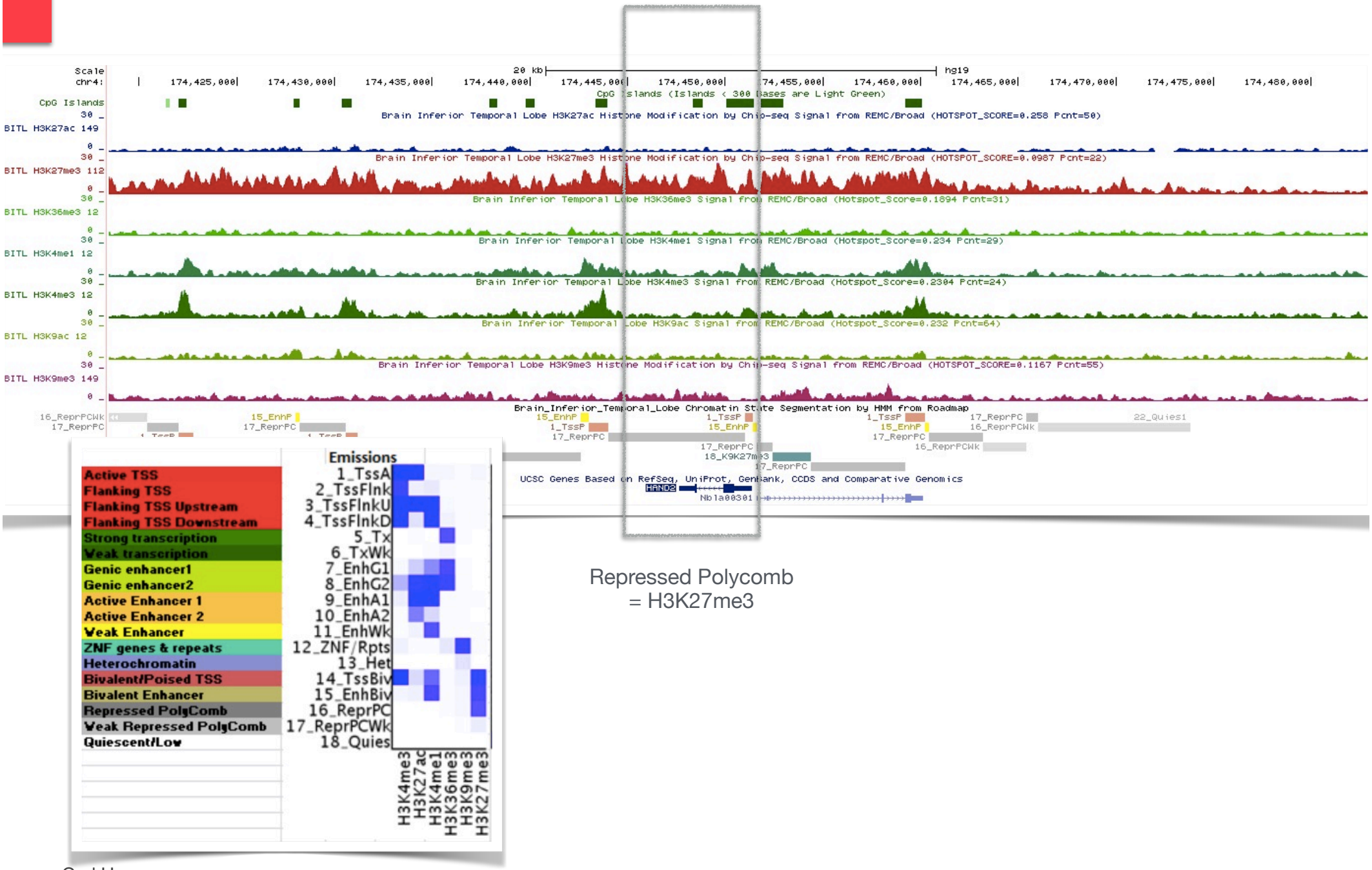


Active Transcription Start Site
= H3K4me3 + H3K27ac

Active Enhancer
= H3K4me1 + H3K27ac

Transcribed region
=H3K36me3

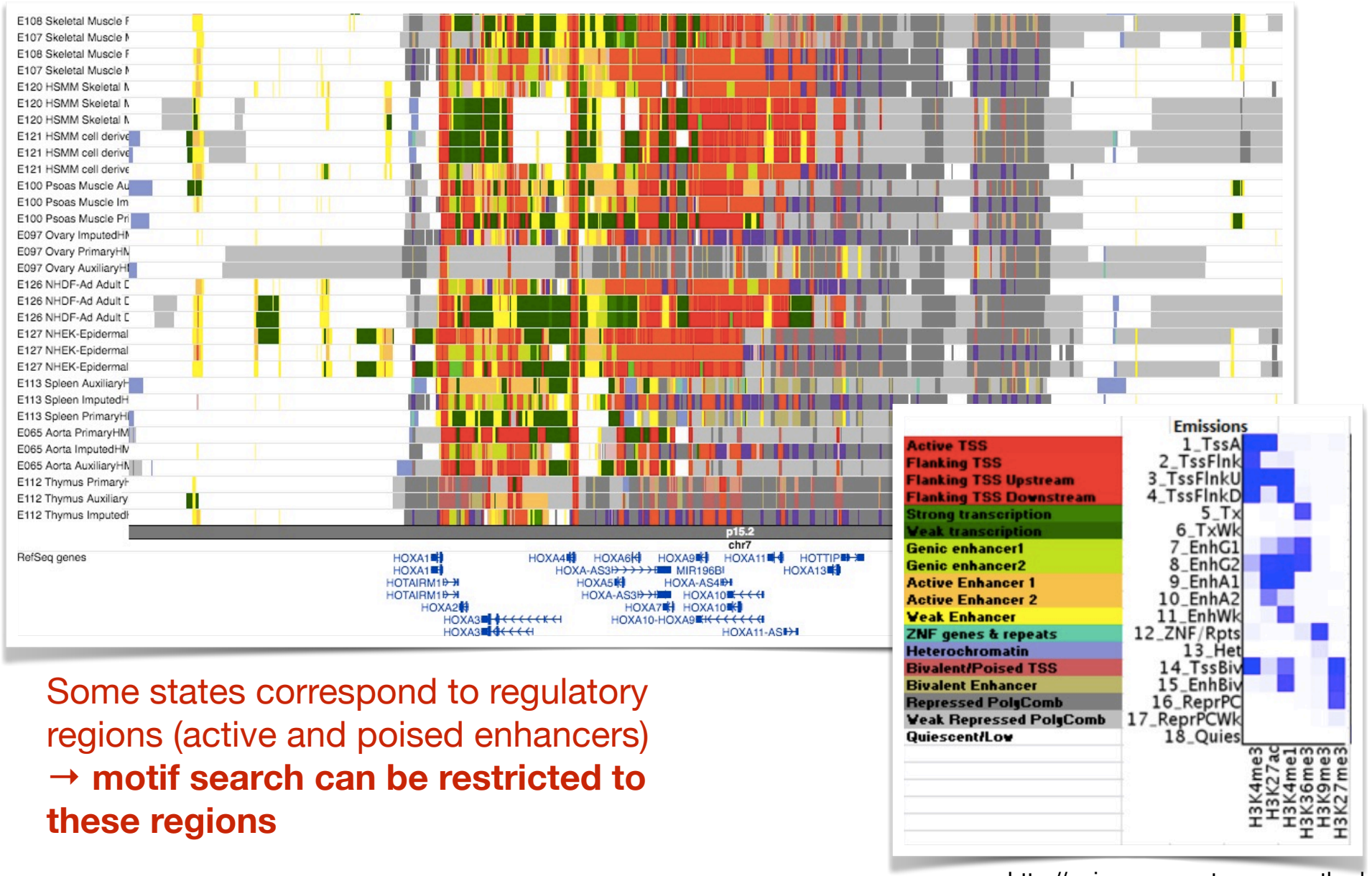
Chromatin states



Repressed Polycomb
= H3K27me3

Carl Herrmann

Roadmap chromatin segmentation in different human adult tissues



Some states correspond to regulatory regions (active and poised enhancers)
→ **motif search can be restricted to these regions**

http://epigenomegateway.wustl.edu

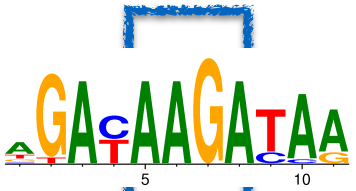
5. Motif discovery

Finding unknown motifs in sequences

Carl Herrmann

Pattern matching vs. Motif discovery

Consider a particular TF of interest (Evi1)



Where are TFBS ?
What are the potential target genes ?

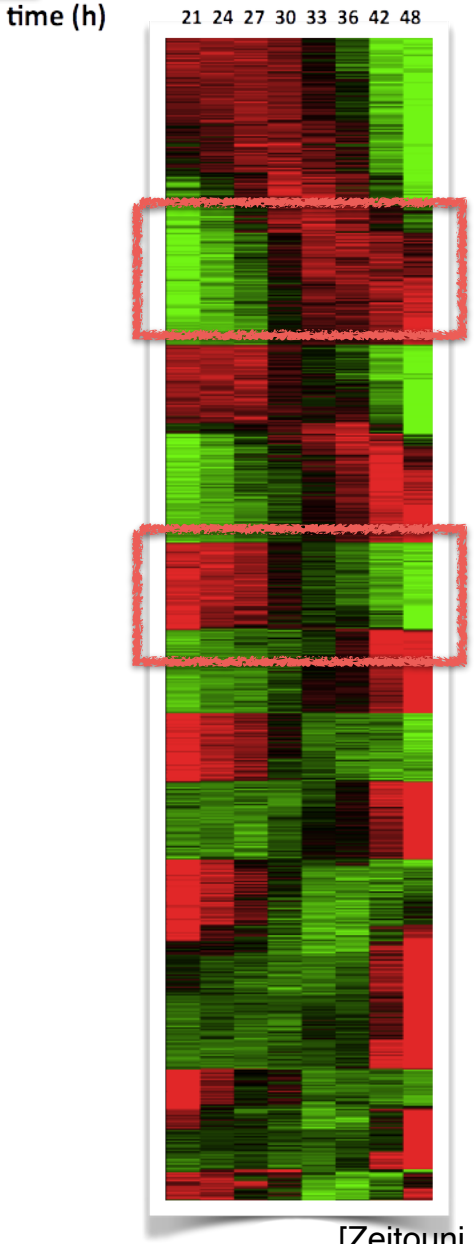
Pattern Matching

Consider a set of regions of interest (e.g. promoters of co-expressed genes)

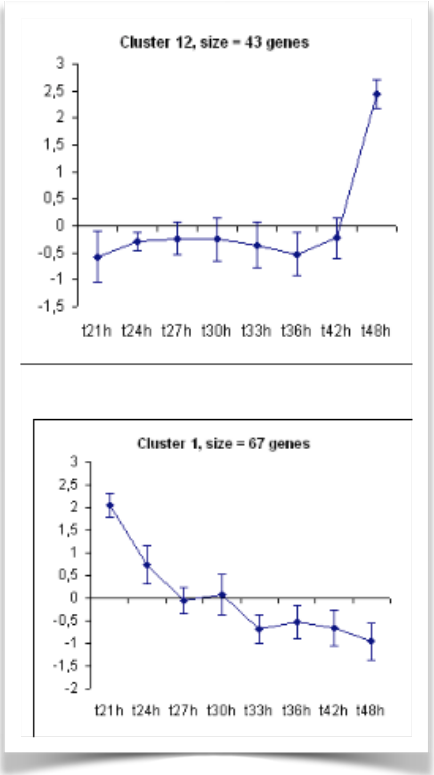
What is their potential common regulator ?

Motif discovery

Motif discovery for co-expressed genes



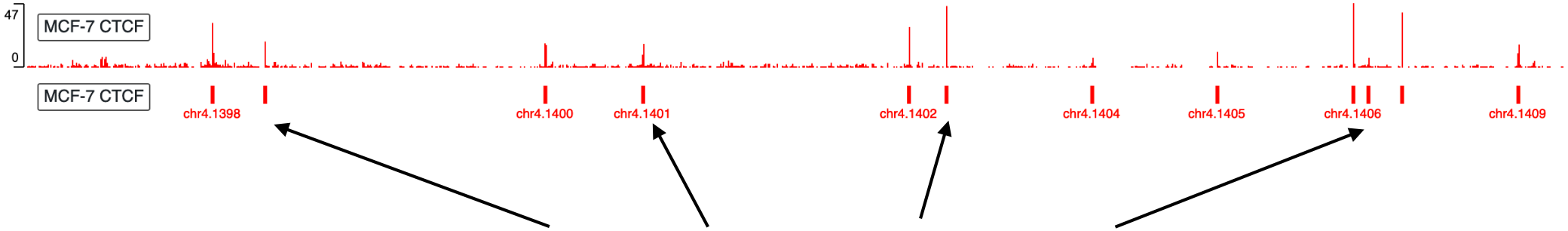
[Zeitouni (2007)]



- clusters of co-expressed genes during cardiac remodelling in Drosophila
- ***Are these cluster of genes co-regulated ?***
- ***If so, what is their common regulator ?***

ChIP-seq

ChIP-seq for CTCF in MCF-7 cell line



all these ChIP-seq peak sequences should contain a common motif (CTCF, but maybe other motifs too?)

```
>MA0139.1       CTCF    1
acttgcttgtaagcatgggggggggggggtttgatgtgggttgggtttttatgtaCTACAGGTGGTcaagtatttatggtaccgtacaatattcatggtggctggcagtaatgtacgaaatacatagcggttgttgatgggtgagtcaatacttgggtggtaccca
>MA0139.1       CTCF    2
tatgtactgttaagggtgggtaggtttgttggtatcctagtgggtgaggggtggctttggagttgcagttgatgtgtgatagttgaaggttgattgctgtacttgcttgtaagcatgggggggggggggtttgatgtgggttgggtttttatgtaCTACAGGTGGT
>MA0139.1       CTCF    3
ggatgacccccctcagataggggtcccttgaccaccatcctccgtgaaatcaatatcccgCACAAGAGTGCtactctcctcgctccgggcccataacacttgggggtagctaaagtgaactgtatccgacatctggttcctacttcagggccataaagcctaaat
>MA0139.1       CTCF    4
atggcttccatgaggctctaagctccccaagctgggttcaccttccgtcgctacttcgccctttcttccaggttgtcctgggcactgatggtCACCAGGTGGCgatgttgcccaggaggacgctctgttgccatctcatggccaaaattcaacactacactcttc
>MA0139.1       CTCF    5
aatccagggcaccgagaaagtccagcagcgcctgagcctgctgcacttgcagtagcagggactggcCGGGAGATGGCagccctgtcccaggtcgggtggtgcctcagccgcagttcctgtcacggccatcgggaggcagtgcaggaacactcactgatgtgtagc
>MA0139.1       CTCF    6
nnnnnnnnnnnnnnnnnnnnnnnnnnnaggagagaagacaagattctgggtgaacaggaacagtaagttggctcaggactagggtaagccggtaggaggcgccatttctccacaatccagcggcagcatggccctctgaagcCACTTGAGGGCgtgttacgtccga
>MA0139.1       CTCF    7
aggtgccggggactgaggctacccctgctgtgggctgcagcttgctctcttctcaacttcccagctcccacctgggggcgccagagggcacatatactgctgtactcgtttgtgcCACCAGAGGGCagcctccgcacatcgatctctactgcttccgagaactgt
>MA0139.1       CTCF    8
ggacactacccgctgggctacccagaccggagcgcggcgcgagaggcgtgagtcgatgacctctagcgagggaggccacctagcggccaggccaagtctggaggatacttccgttgggcCGCCAGGGGGCagtaacgccagatcatctgtctcccggaagtaatg
>MA0139.1       CTCF    9
cttggcaagggggcgccgctcaccagctggggtggatgaggagcgtgaagacagctcgctccagtggccgcgggggcgctcgcccactcagagcatggcCGCCAGATGGCgccggctcgcccgctctccacaaggctcttccaaccccttaccctgtcccacac
>MA0139.1       CTCF    10
accgagtccttgtccccagcagaaaccgtggtccctttgctgttccagatgctctggaccctgggtactggctccccctggcggtcatgataggaagcgcccttcccctctggctcCTGTAGAGGGCgtcgcggatgcattgatcctttagttccttcactgtcc
>MA0139.1       CTCF    11
tcggcccaaacctctggcggccgcgtcactcggctccaccagtctgtgccgcctgcaagcgcccctagcggcctcgcagccatcgcgcctcagcagcgccgtggcgtcctgttcggcagtgtcCATTAGAGGGCgccagcgcgcctcagaagagggaccgtgacg
>MA0139.1       CTCF    12
cagtgcctaagtgtctagagagctgagtgggttgcctcggatcggccaggcctggatatcagcgaacagatgcggatttcCACCAGGGGGCgctaggacgccacgacctttcactatgtgcaccaggggcgctgccgcccgccttggccatcccgggctgggaa
>MA0139.1       CTCF    13
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnngcagcggaggtggcCGCCAGGGGACgctcgggagccccgtgggccgtcaggctgtggtgcagcagctcctcctcctggctggagccgaaactgcacaggccgcacttccagggc
>MA0139.1       CTCF    14
ttacggatcggcttactccgcggagttggcctcatttctgcagtcggcgctccctgtagtttctcctctcgaaCGCCAGGTGGAgcaaccggccggataccgccacagccctggcaggcggcgctgtgatgcctgagctgatcctgtatgttgcaatcactctat
>MA0139.1       CTCF    15
ttcagcagcccagcggatttcagcgtttgacaacctggaagttctgctgccccctcacggtgggagaaaagcggtgggtcctgcacaggccgaagaacCACTAGAGGGCagcagcacacgggctccataagacttgctgcgtgtcgaccctggctcatacagtgc
>MA0139.1       CTCF    16
cccgccaagccggccccagcaagccgcaggaagccgtccgctgaacctcaggttcttgggttttcctatgccccctggaggcatctgcgaacctctcggcatcgctccaccacacctcccacCAGCAGGTGGCactgttgccatctccataacatggactggcct
```

Gene regulation

Follow their sight…

… you'll find their common regulator !

Motif discovery using word counting

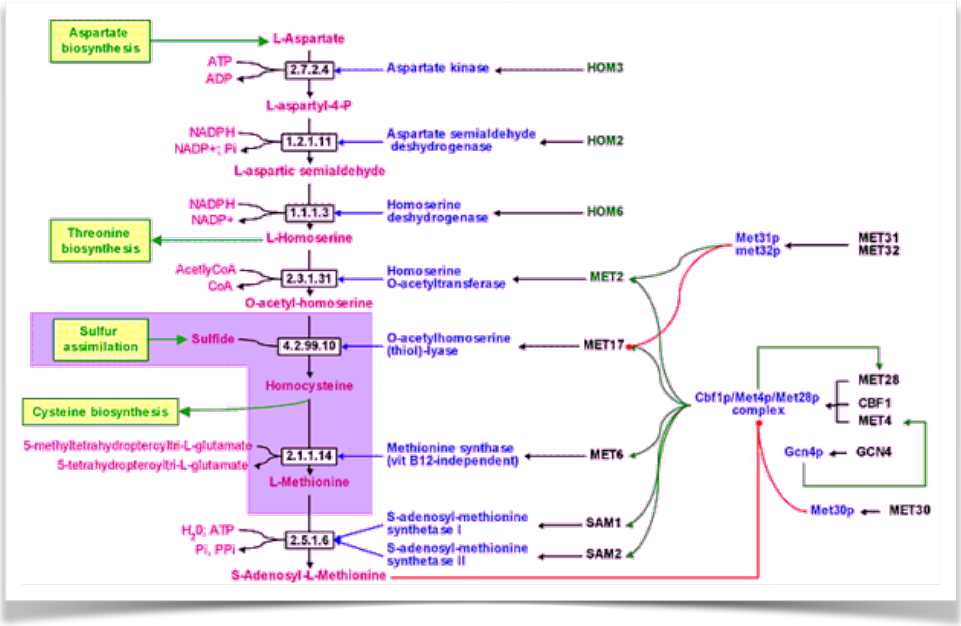
Idea:
motifs corresponding to binding sites are generally **repeated**
→ **capture this statistical signal**

- **Algorithm**
  - count **observed number of occurrences** of all k-mers in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the **expected number of occurrences** in background model
  - build a theoretical background model (MM)
  - empirical based on observed k-mer frequencies
  - **statistical significance** of the deviation observed (P-value/E-value)

Example

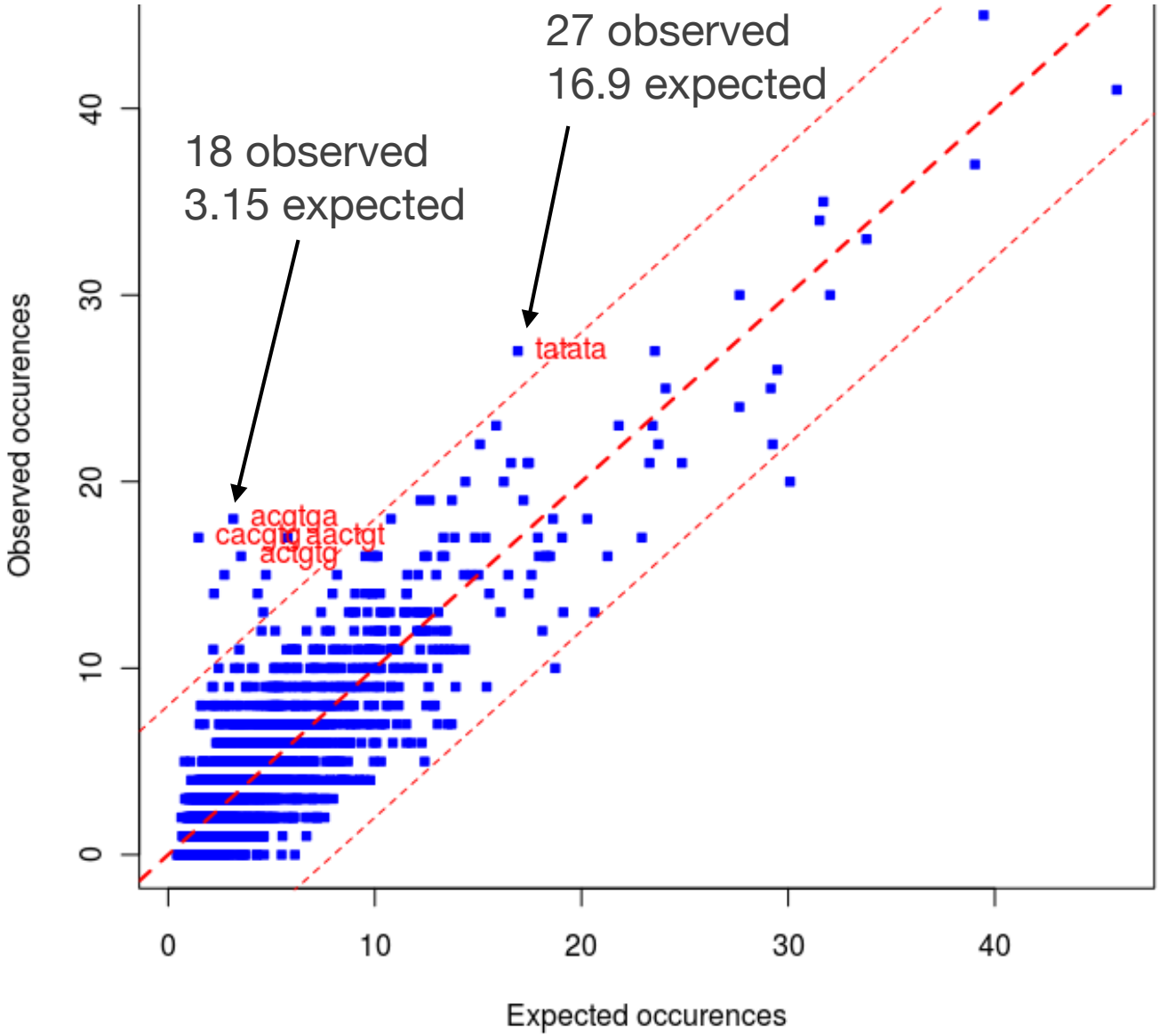
- Are they co-regulated ?
- Do they share common regulatory motifs ?
- Principle
  - Count occurrences of k=6 mers in the 800 bp upstream of the TSS (!! on both strands !!)
  - 9000 possible positions
  - compare observed and expected number of occurences

Methionine synthesis pathway in *S. cerevisae*



19 genes from *S. cerevisiae* involved in methionine biosynthesis pathway

Motif discovery using word counting



How to evaluate expected number of occurrences ?

Could these be statistical fluctuations ?

Example

seq	identifier	exp_freq	occ	exp_occ	Pvalue	E-value	occ_sig	rank
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e-09	8.98	1
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e-05	4.82	2
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e-04	3.68	3
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e-04	3.39	4
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e-03	2.68	5
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e-02	1.42	6

- **P-value** : what is the risk you take by rejecting the null hypothesis for one particular event (i.e. consider it to be significant)
- but you are testing 2080 possible hexanucleotides *("multiple testing")*
- if you are taking 2080 times a risk of p=1e-4, on average, in 2080*1e-4=0.208 of these cases, you will be wrong
  → ***E-value***

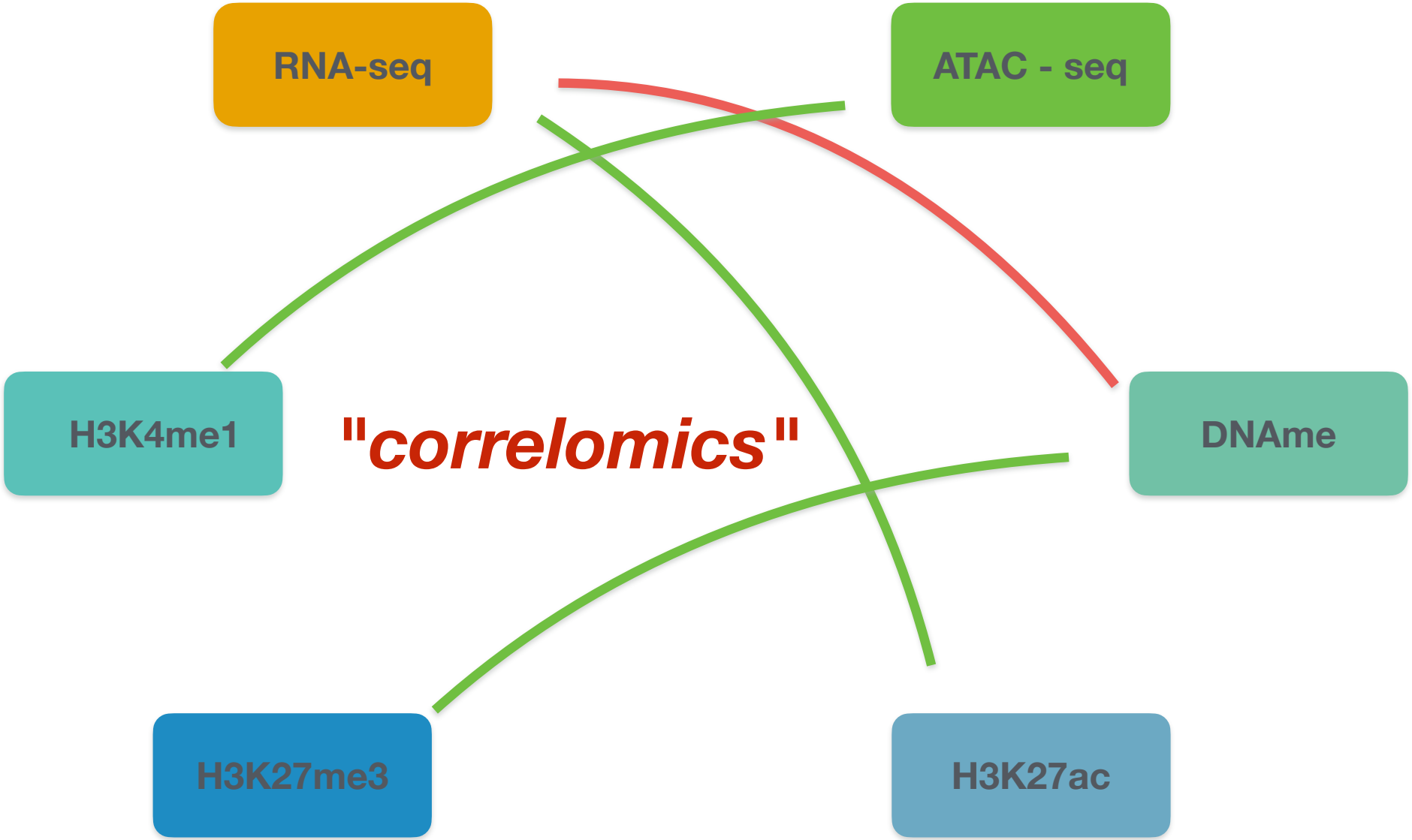
From words to logo

seq	identifier	exp_freq	occ	exp_occ	occ_P	occ
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e
aactgt	aactgt|acagtt	0.0006168655788	17	5.78	0.00011	2.4e
agtcat	agtcat|atgact	0.0005039616969	15	4.73	0.00012	2.6e
tagtca	tagtca|tgacta	0.0004613751449	14	4.33	0.00017	3.5e
agccac	agccac|gtggct	0.0002599968758	10	2.44	0.00023	4.7e
cgtgac	cgtgac|gtcacg	0.0001695417189	8	1.59	0.00025	5.2e
cgcgca	cgcgca|tgcgcg	0.0001715224888	8	1.61	0.00027	5.6e
acgtgc	acgtgc|gcacgt	0.0002276443015	9	2.13	0.00038	7.9
gactca	gactca|tgagtc	0.0002319359695	9	2.18	0.00043	9.0e

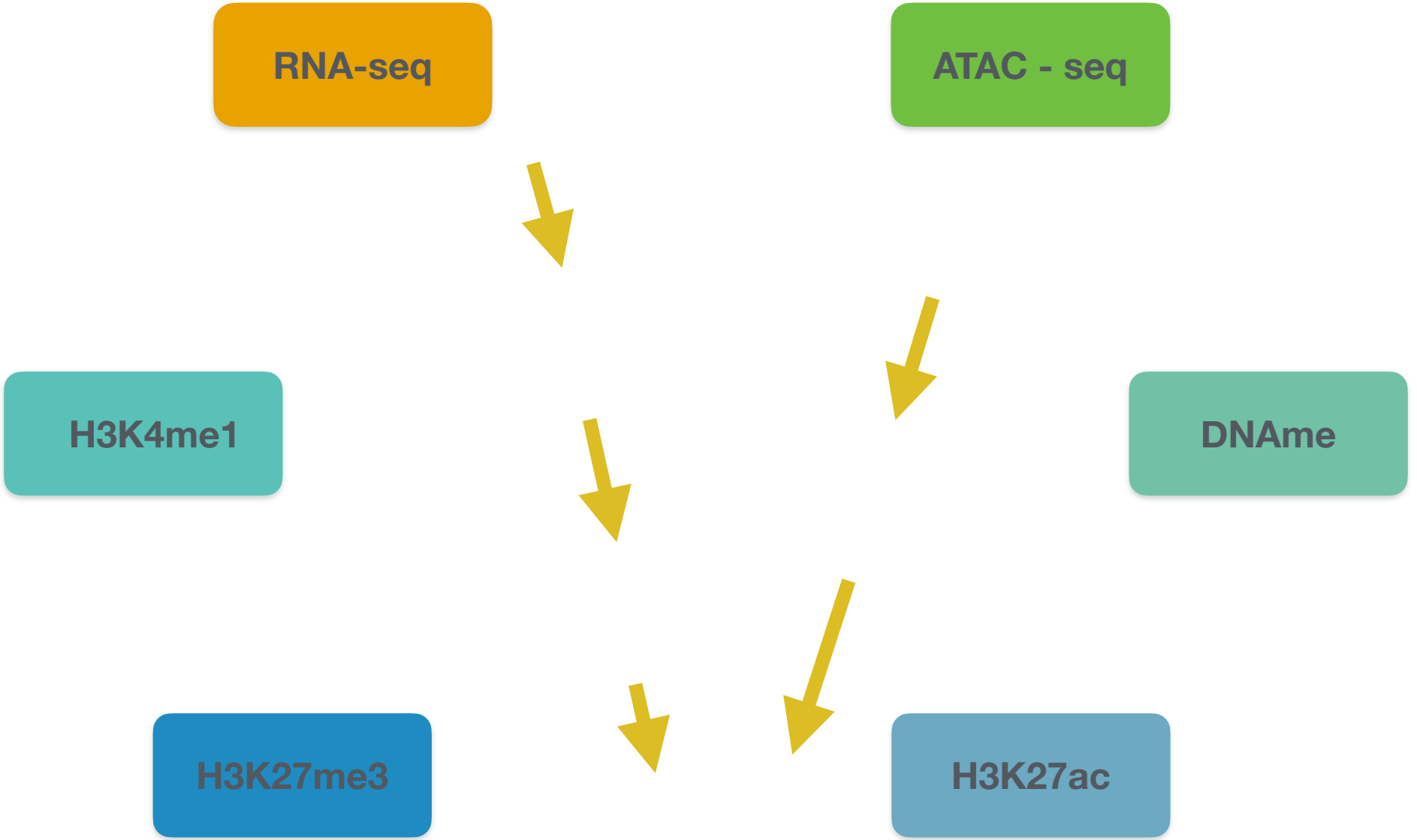

```
;assembly # 1    seed:
;   alignt          rev
gtcacg....       ....cg
.tcacgt...       ...acg
..cacgtg..       ..cacg
...acgtga.       .tcacg
....cgtgac       gtcacg
gtcacgtgac       gtcacg

;assembly # 2    seed:
;   alignt          rev
agccac....       ....gt
.gccaca...       ...tgt
..ccacag..       ..ctgt
...cacagt.       .actgt
....acagtt       aactgt
agccacagtt       aactgt

;assembly # 3    seed:
;   alignt          rev
gtcacg....       ....cg
.tcacgt...       ...acg
..cacgtg..       ..cacg
...acgtgc.       .gcacg
....cgtgca       tgcacg
```

6. chromatin networks

Current challenges



Current challenges



Genomics application

DNA methylation

Gene expression

H3K27ac

H3K4me1

H3K4me3

H3K36me3

H3K9me3

H3K27me3

MYCN

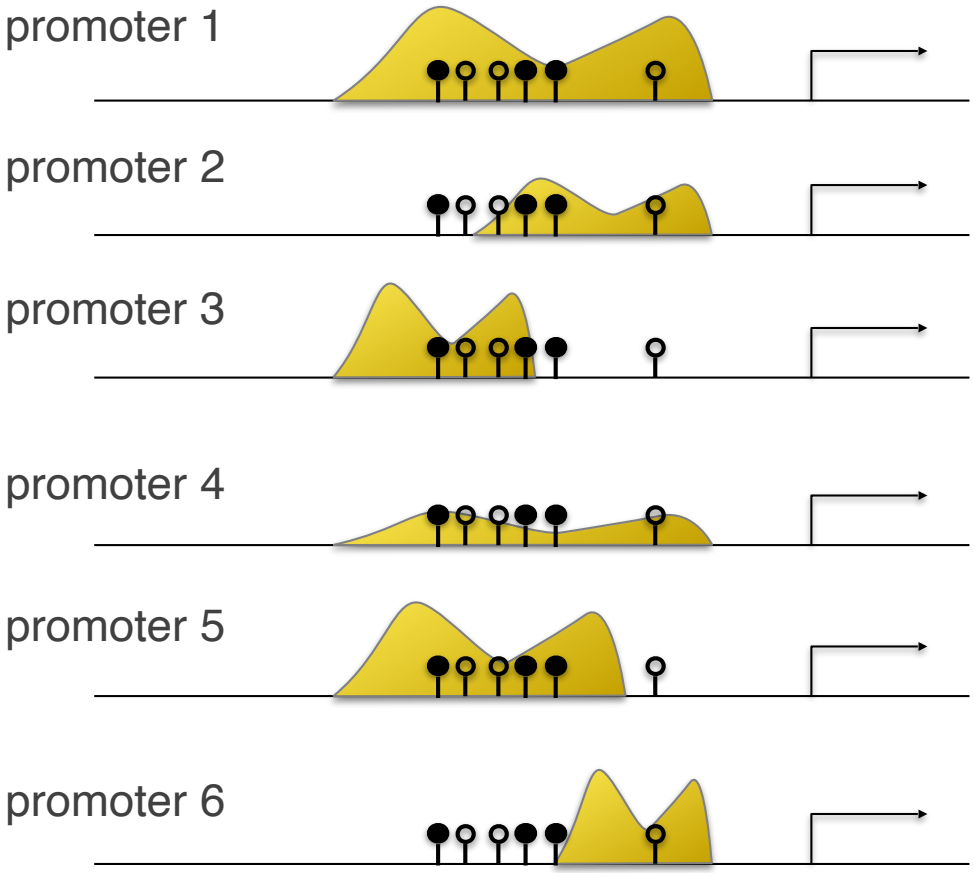
EZH2

DNMT1

DNMT3

- Various neuroblastoma cell lines
- normal conditions / treated (inhibition)
- state at gene promoters represent the observations of the random variables

Learning Network Structures



DNAme	K27ac
0.57	128.8
0.45	75.2
0.89	98.3
0.21	21.3
0.18	86.2
0.41	67.3

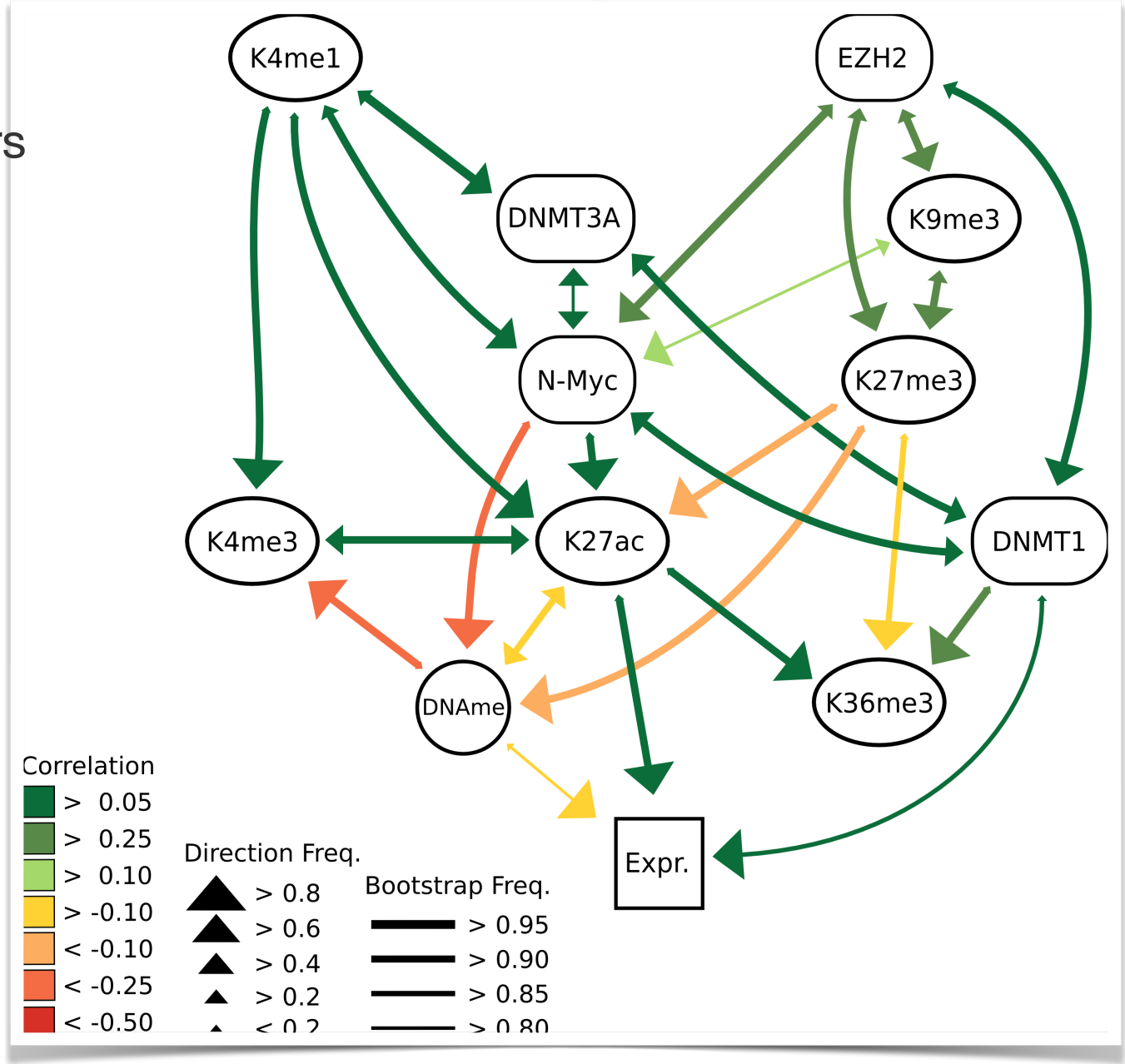
DNAme	K27ac
mid	5
mid	3
high	4
low	2
low	4
mid	3

3 states

5 states

Promoter BN

- non-CGI Promoters
- (*n* = 5139)



7. conclusion

Conclusions

- Transcription regulation is a **complex process** with an interplay of **multiple components**
- **Transcription factors** play a central role, usually in combination with other TF inside **enhancers**
- Tissue / context specificity of the activity of regulatory elements is given by the **cell-specific chromatin state**: open/accessible or closed/compact
- **Many data types available** to build integrative models of regulatory activity
- **Single-cell genomics** is becoming the new challenge in regulatory genomics