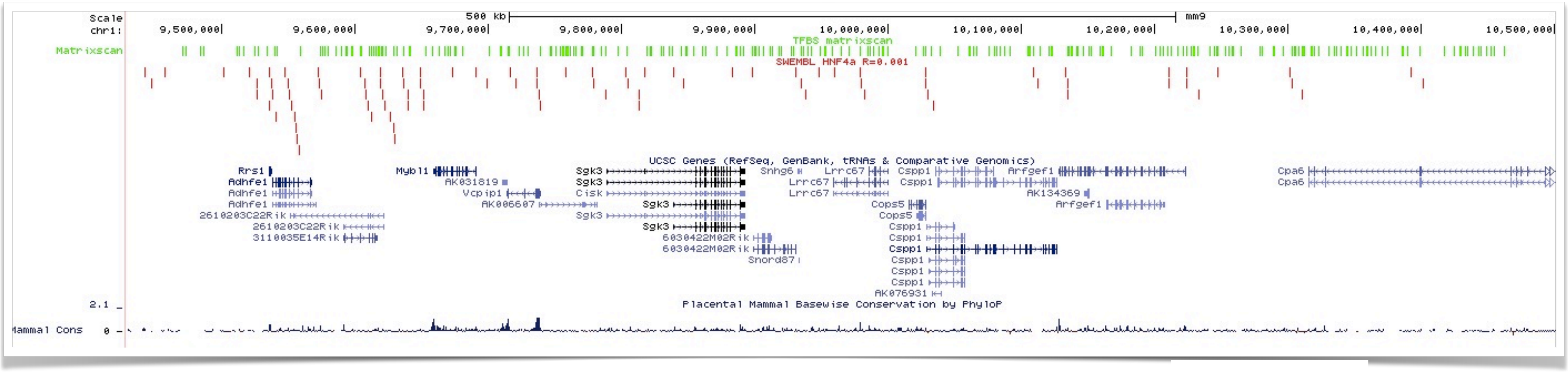
3. Predicting binding sites

- basics of TFBS identification
- defining a background model
- tools
- phylogenetic footprinting
- including "in-vivo features"

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

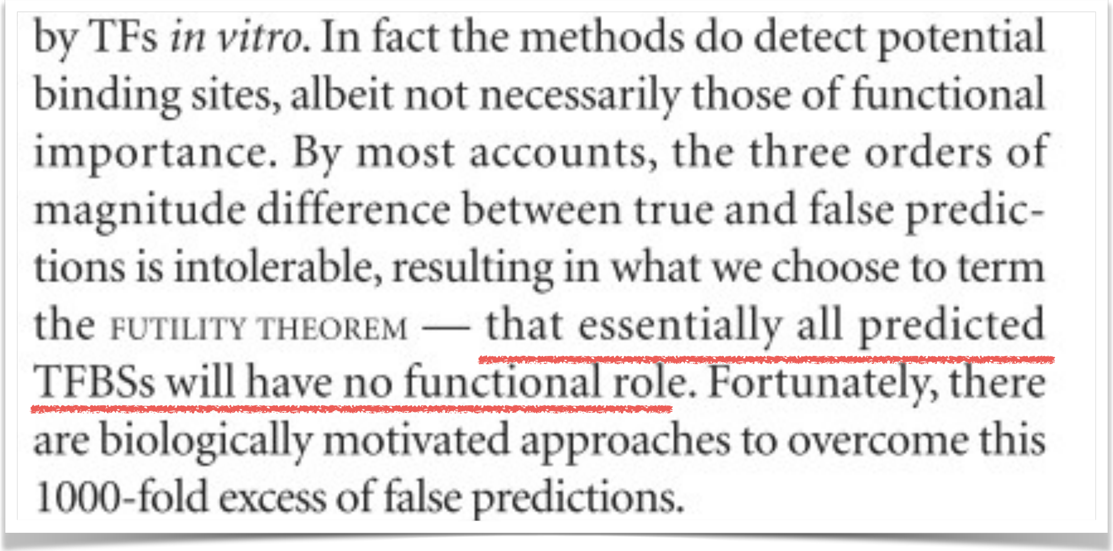
Predicting TFBS on real sequences



RSAT Metazoa
http://rsat.eu/

Improving TFBS predictions

TFBS prediction suffers from a high degree of **false-positive** and **false-negative** predictions



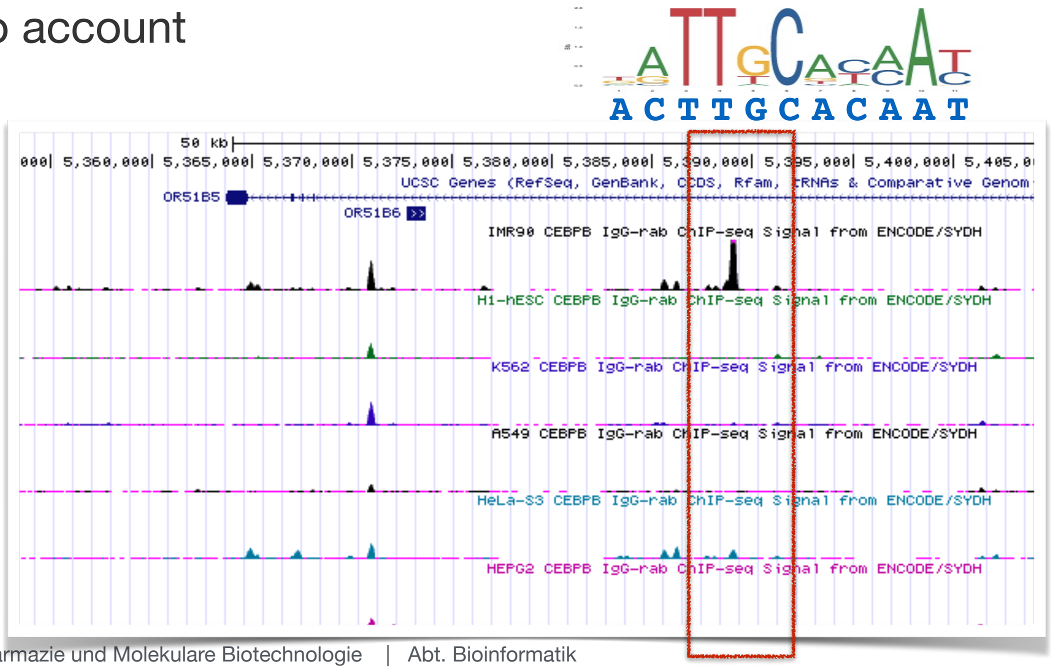
by TFs *in vitro*. In fact the methods do detect potential binding sites, albeit not necessarily those of functional importance. By most accounts, the three orders of magnitude difference between true and false predictions is intolerable, resulting in what we choose to term the FUTILITY THEOREM — that essentially all predicted TFBSs will have no functional role. Fortunately, there are biologically motivated approaches to overcome this 1000-fold excess of false predictions.

[Wasserman & Sandelin, Nat.Rev.Gen (2004)]

Improving TFBS predictions

- **Limitations**
  - quality of the matrix (PWMs constructed from few sites are not discriminative, low information content !)
  - difficulty to predict **low affinity** binding events
  - correct choice of the **background** model
  - **in-vivo context** is not taken into account

Why is CEBPB binding in some cell-lines and not in others ?

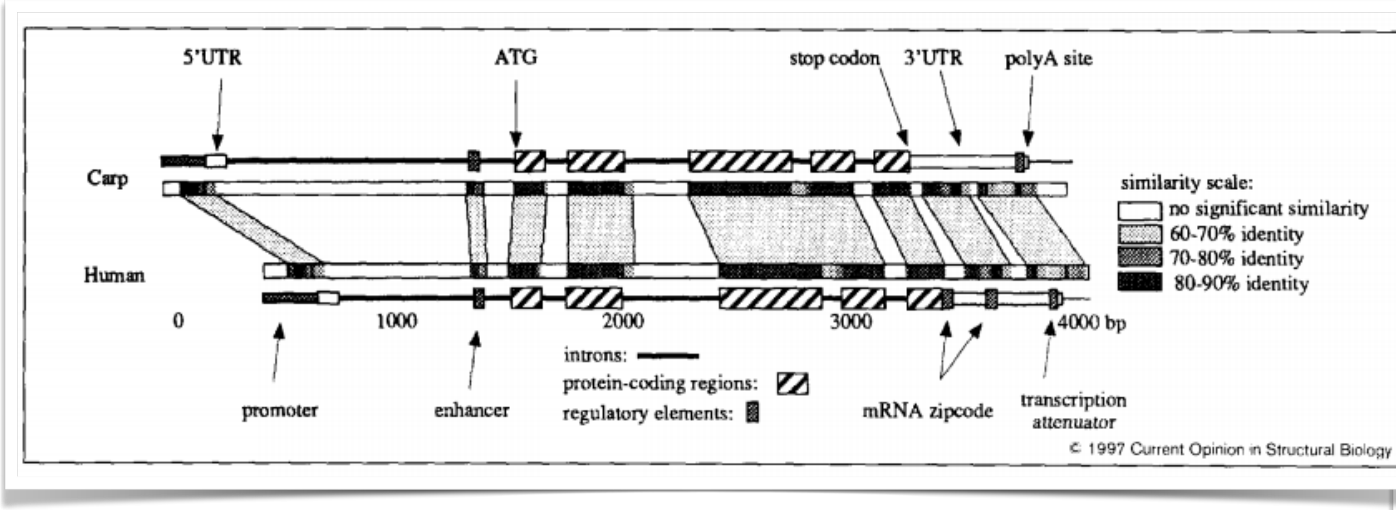

ATTGCACAAT
A C T T G C A C A A T
50 kb
000| 5,360,000| 5,365,000| 5,370,000| 5,375,000| 5,380,000| 5,385,000| 5,390,000| 5,395,000| 5,400,000| 5,405,00
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genom
OR51B5
OR51B6
IMR90 CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH
H1-hESC CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH
K562 CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH
A549 CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH
HeLa-S3 CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH
HEPG2 CEBPB IgG-rab ChIP-seq Signal from ENCODE/SYDH

Improving TFBS predictions

- How can we improve TFBS predictions ?
  - functional sites are believed to be under **selective pressure** → detect "footprints" of evolution in the genome ("phylogenetic footprinting”)
  - TF binding is influenced by the **chromatin state** (accessibility, histone modifications, …)

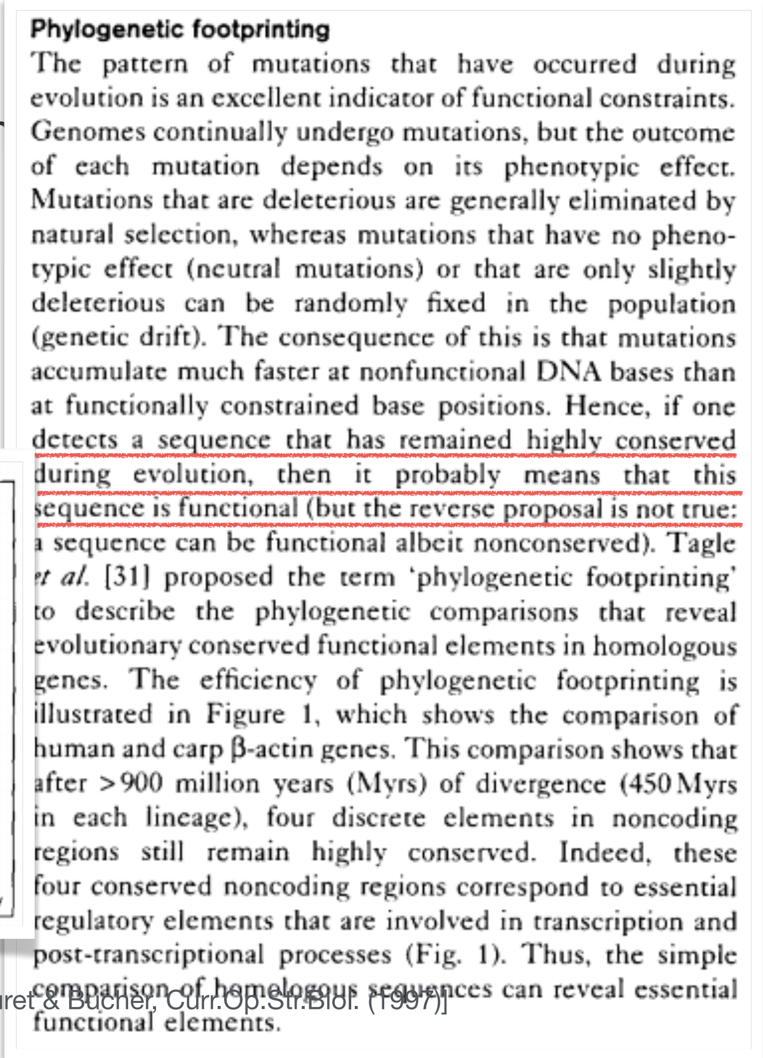
Phylogenetic footprinting

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



© 1997 Current Opinion in Structural Biology

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

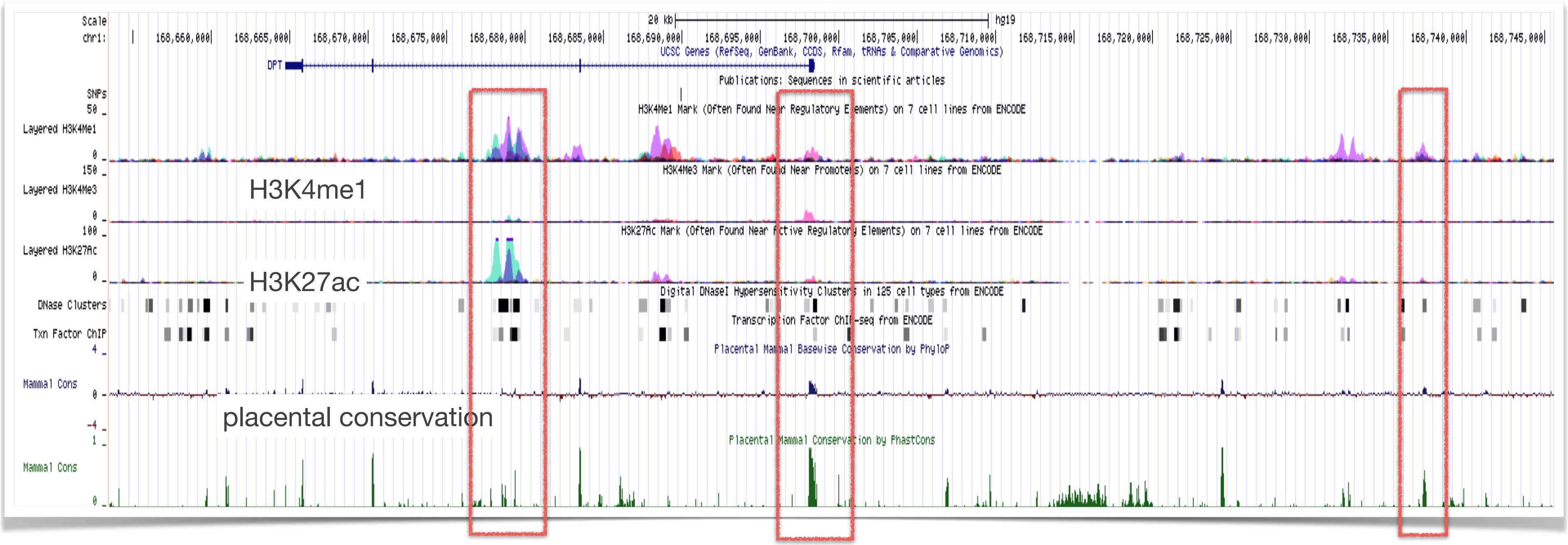


Phylogenetic footprinting

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 evolutionary conserved functional elements in homologous genes. The efficiency of phylogenetic footprinting is illustrated in Figure 1, which shows the comparison of human and carp β-actin genes. This comparison shows that after >900 million years (Myrs) of divergence (450 Myrs in each lineage), four discrete elements in noncoding regions still remain highly conserved. Indeed, these four conserved noncoding regions correspond to essential regulatory elements that are involved in transcription and post-transcriptional processes (Fig. 1). Thus, the simple comparison of homologous sequences can reveal essential functional elements.

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

Phylogenetic footprinting


Scale
20 kb
hg19
chr1:
168,660,000
168,665,000
168,670,000
168,675,000
168,680,000
168,685,000
168,690,000
168,695,000
168,700,000
168,705,000
168,710,000
168,715,000
168,720,000
168,725,000
168,730,000
168,735,000
168,740,000
168,745,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
DPT
Publications: Sequences in scientific articles
SNPs
50
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
H3K4me1
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
H3K27ac
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
placental conservation
1
Mammal Cons
0
Placental Mammal Conservation by PhastCons

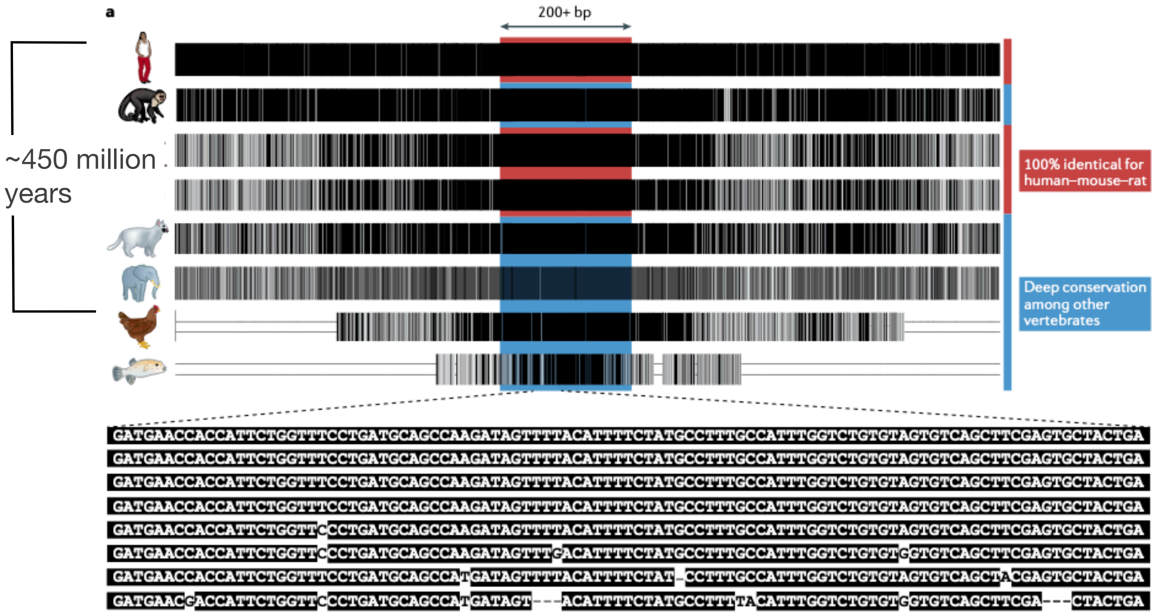
potential regulatory region weakly conserved

conserved active promoters

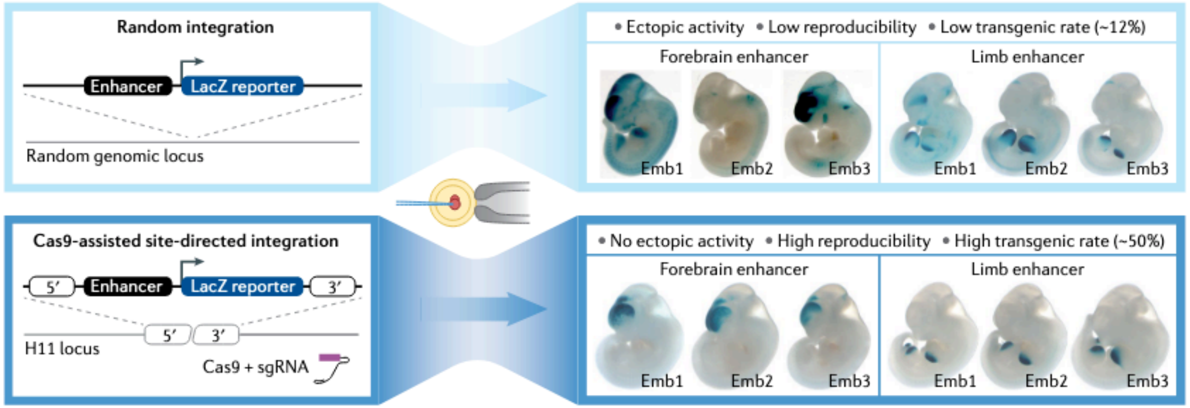
highly conserved regulatory elements

Deeply conserved non-coding elements

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


a
200+ bp
~450 million years
100% identical for human–mouse–rat
Deep conservation among other vertebrates

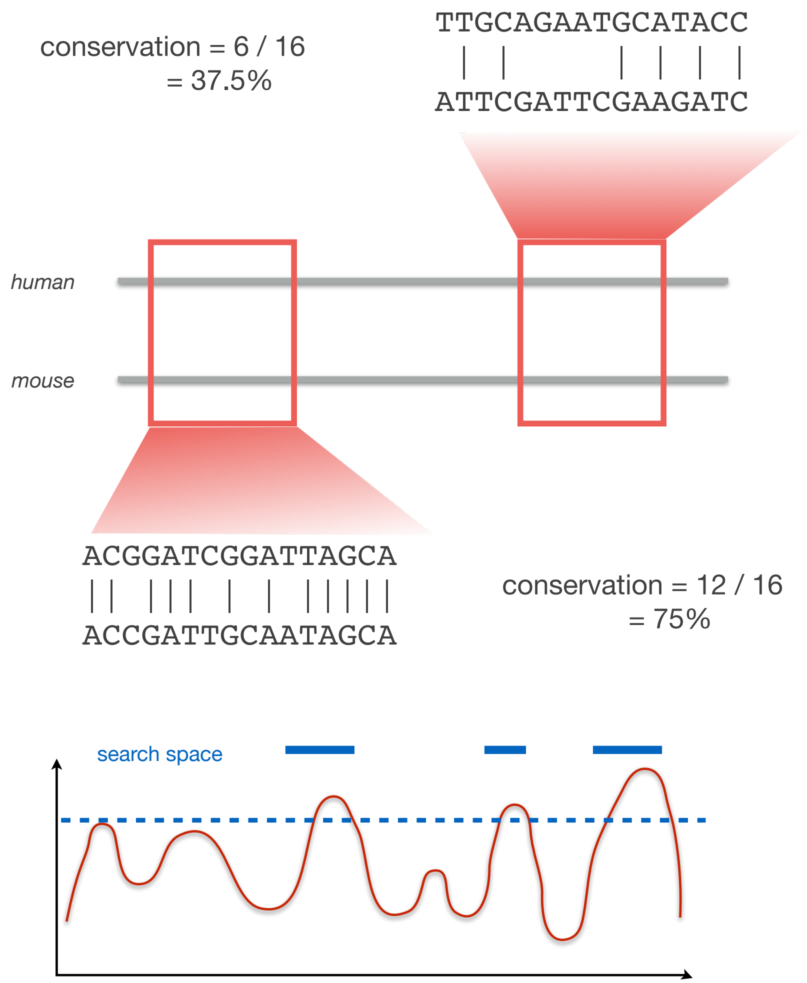
```
GATGAACCACCATTCTGGTTTCCTGATGCAGCCAAGATAGTTTTACATTTTCTATGCCTTTGCCATTTGGTCTGTGTAGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTTCCTGATGCAGCCAAGATAGTTTTACATTTTCTATGCCTTTGCCATTTGGTCTGTGTAGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTTCCTGATGCAGCCAAGATAGTTTTACATTTTCTATGCCTTTGCCATTTGGTCTGTGTAGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTTCCTGATGCAGCCAAGATAGTTTTACATTTTCTATGCCTTTGCCATTTGGTCTGTGTAGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTCCCTGATGCAGCCAAGATAGTTTTACATTTTCTATGCCTTTGCCATTTGGTCTGTGTAGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTCCCTGATGCAGCCAAGATAGTTTGACATTTTCTATGCCTTTGCCATTTGGTCTGTGTGGTGTCAGCTTCGAGTGCTACTGA
GATGAACCACCATTCTGGTTTCCTGATGCAGCCATGATAGTTTTACATTTTCTAT-CCTTTGCCATTTGGTCTGTGTAGTGTCAGCTACGAGTGCTACTGA
GATGAACGACCATTCTGGTTCCCTGATGCAGCCATGATAGT---ACATTTTCTATGCCTTTTACATTTGGTCTGTGTGGTGTCAGCTTCGA---CTACTGA
```

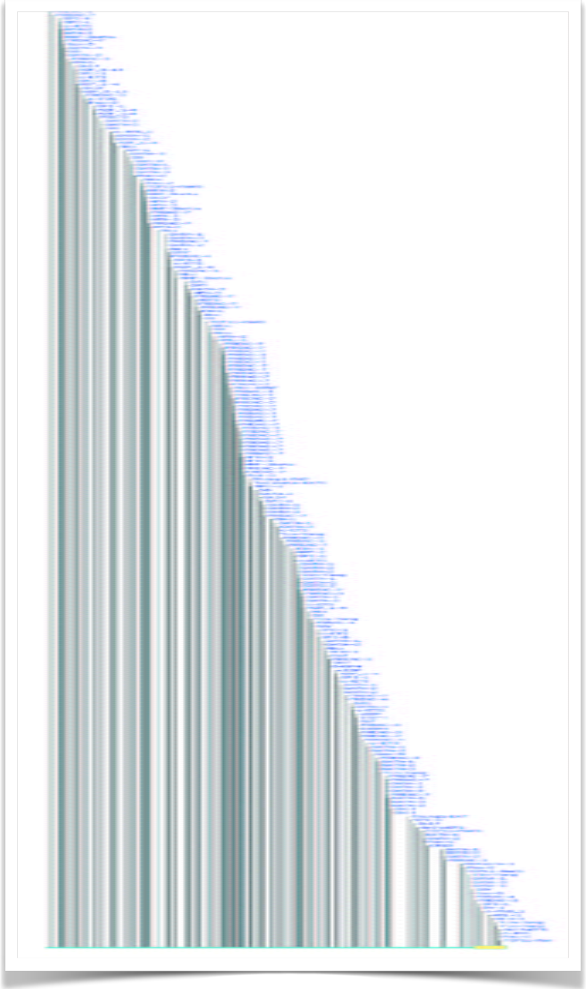


[Snetkova, Nature Rev.Gen. (2020)]

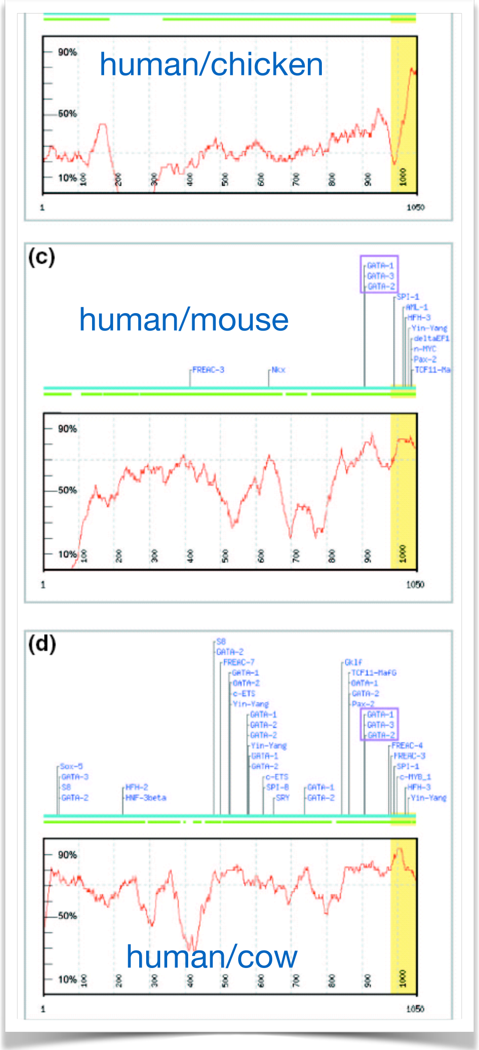
Phylogenetic footprinting

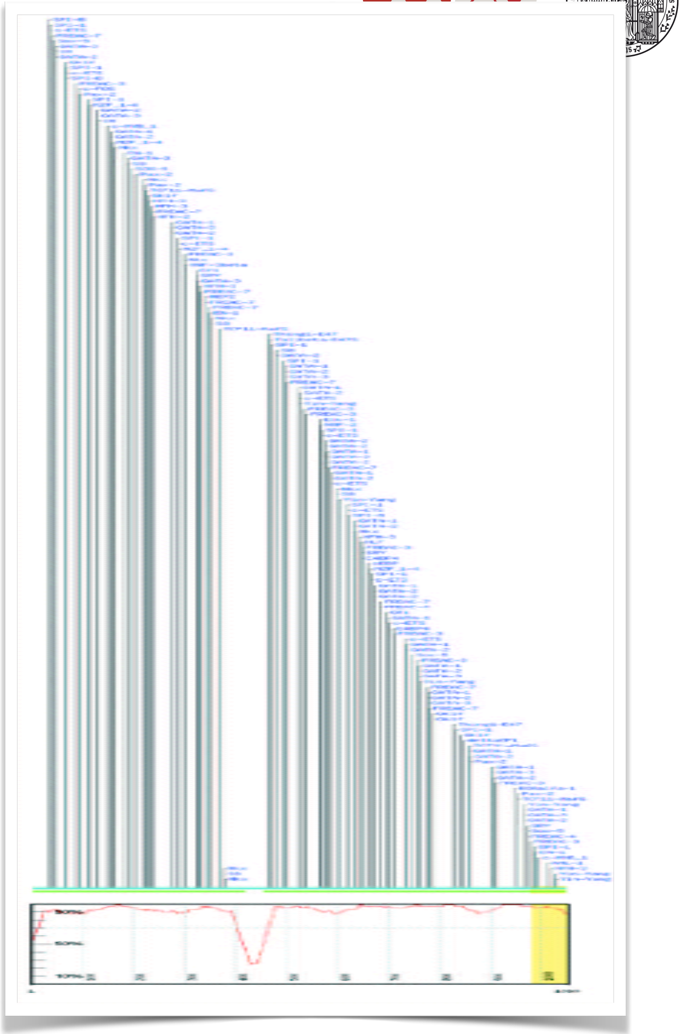
- Starting point : alignment of **2 orthologous regions**
  (e.g. promoter of orthologous genes)
- Compute the **conservation** inside a sliding window (number of conserved positions divided by length)
- TFBS search using PWM (fixed threshold)
- **Only TFBS inside highly conserved regions** are retained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





Human globin gene promoter alone





Alignment of human/macaque globin promoter

Model of neutral evolution

- DNA sequences can evolve due to natural occuring mutations (replication errors, ..) [Kimura]
  → neutral evolution (no positive/negative pressure)
- they evolve towards an **equilibrium state** $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$
- The rate at which bases mutate into other bases can be described by a rate matrix Q

$$Q = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix} \qquad q_{ii} = -\sum_j q_{ij}$$

Substitution frequency *i* → *j* : $r_{ij} = q_{ij} \cdot \pi_i$ $p_i =$ frequency of base *i*

Model of neutral evolution

JC69 Model [Jules & Cantor, 1969]

$$Q = \begin{pmatrix} -\frac{3}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & -\frac{3}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & \frac{1}{4}\alpha & -\frac{3}{4}\alpha & \frac{1}{4}\alpha \\ \frac{1}{4}\alpha & \frac{1}{4}\alpha & \frac{1}{4}\alpha & -\frac{3}{4}\alpha \end{pmatrix}$$

K80 Model [Kimura, 1980]

$$Q = \begin{pmatrix} q_{AA} & 1 & \kappa & 1 \\ 1 & q_{CC} & 1 & \kappa \\ \kappa & 1 & q_{GG} & 1 \\ 1 & \kappa & 1 & q_{TT} \end{pmatrix}$$

$$q_{ii} = -\kappa - 2$$

Assumption in both models : $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$

$$\kappa = \frac{\text{transition}}{\text{transversion}} > 1$$

Model of TFBS evolution

- HKY85 model

[Hasegawa, Kishino, Yano (1985)]

- $P(t) = (p_T, p_C, p_A, p_G)$ Probability of each base at time *t*:

$$Q = \begin{pmatrix} -\lambda_A & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\lambda_C & \pi_G & \boxed{\kappa\pi_T} \\ \kappa\pi_A & \pi_C & -\lambda_G & \pi_T \\ \pi_A & \kappa\pi_C & \boxed{\pi_G} & -\lambda_T \end{pmatrix}$$

C → T

T → G

- κ = transition/transversion > 1
- π = nucleotide frequency at equilibrium

Model of TFBS evolution

- HKY85 model

[Hasegawa, Kishino, Yano (1985)]

- *P(t)* = (p_T,p_C,p_A,p_G) Probability of each base at time *t*:

$$Q = \begin{pmatrix} -\lambda_A & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\lambda_C & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -\lambda_G & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -\lambda_T \end{pmatrix}$$

- κ = transition/transversion > 1
- π = nucleotide frequency at equilibrium

$$p_T(t+\Delta t) = p_T(t) \overset{\mathtt{T} \to \mathtt{N}}{\boxed{-\lambda_T p_T(t)\Delta t}} + \overset{\mathtt{C} \to \mathtt{T}}{\boxed{\kappa\pi_T p_C(t)\Delta t}} + \overset{\mathtt{A} \to \mathtt{T}}{\boxed{\pi_T p_A(t)\Delta t}} + \overset{\mathtt{G} \to \mathtt{T}}{\boxed{\pi_T p_G(t)\Delta t}}$$

$$P(t+\Delta t) = P(t) + P(t)Q\Delta t \quad \Rightarrow \quad P(t) = P(0)\ e^{Qt}$$

$$e^{Qt} = 1 + Qt + \frac{Q^2}{2}t^2 + \cdots + \frac{Q^n}{n}t^n + \cdots$$

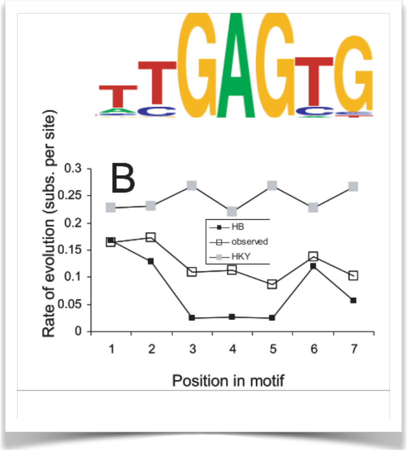
Model of TFBS evolution

- Assumption: TFBS are **functional regulatory elements**, which should be under **negative selection** (lower mutation rate)
- Halpern-Bruno Model
  - positions with **high degeneracy** (low IC) evolve **more rapidly** (but lower than neutral)
  - positions with **low degeneracy** (high IC) evolve **more slowly**

$$Q = \begin{pmatrix} -\lambda_A & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -\lambda_C & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -\lambda_G & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -\lambda_T \end{pmatrix}$$

- κ = transition/transversion > 1
- π = nucleotide frequency at equilibrium
- Rows normalized to zero

$$R_{ia\to b} = Q_{ab} \cdot \frac{\ln\left(\frac{f_{ib}Q_{ba}}{f_{ia}Q_{ab}}\right)}{1 - \frac{f_{ia}Q_{ab}}{f_{ib}Q_{ba}}}$$



f_{ia} = frequency of base *a* at position *i*

$$f_{ia} \sim \pi_a \Rightarrow R_{ia\to b} \sim Q_{ab}$$
$$f_{ia} \sim 1 \Rightarrow R_{ia\to b} \sim 0$$

[Moses et al., Gen.Biol. 2004 ; Moses et al., PLoS CompBiol, 2006]

Model of TFBS evolution

- Given 2 sites in 2 species
  - what is the likelihood that they evolved according to the **neutral** mutation rate ?
  - what is the likelihood that they evolved according to the **constrained model** of TFBS evolution ?

→ ***likelihood ratio test***

Is this a binding site or rather not ?

X_i

human

ACGTTGCT**AGGCTAG**GCTAGGAGC

ACGTTGCT**ACGCAAG**GCAACGCGG

Y_i

mouse

$$T_i = \log \frac{P(X_i, Y_i | motif,\ t,\ R_{mot})}{P(X_i, Y_i | background,\ t,\ Q)}$$

t = evolutionary distance

Model of TFBS evolution

- Test statistics : log-likelihood

$$T_i = \log \frac{P(X_i, Y_i | motif,\ t,\ R_{mot})}{P(X_i, Y_i | background,\ t,\ Q)}$$

- Sum over all possible ancestor states A_i

$$T_i = \log \frac{\sum_{A_i} P(X_i|A_i, t_{AX}, R_{mot}) P(Y_i|A_i, t_{AY}, R_{mot}) P(A_i|motif)}{\sum_{A_i} P(X_i|A_i, t_{AX}, Q) P(Y_i|A_i, t_{AY}, Q) P(A_i|background)}$$

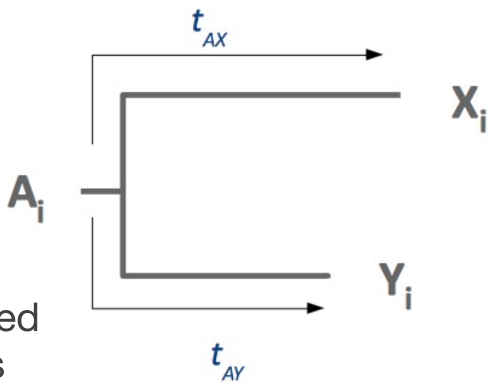
Is this a binding site or rather not ?

X_i *human*

ACGTTGCT**AGGCTAG**GCTAGGAGC

ACGTTGCT**ACGCAAG**GCAACGCGG

Y_i *mouse*



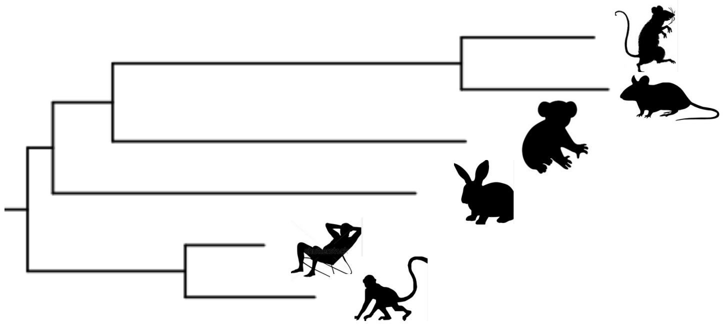
$$P(X_i|A_i, t, R_{mot}) \sim e^{tR_{mot}}$$

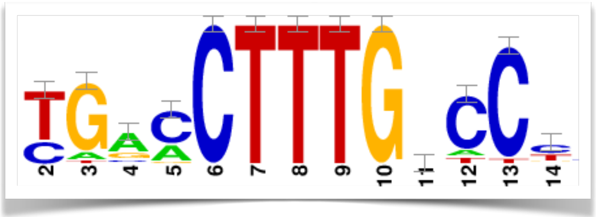
$$P(X_i|A_i, t, Q) \sim e^{tQ}$$

null distribution of T can be computed exactly or empirically by simulations
→ p-value (Null hypothesis: not a TFBS)

Model of TFBS evolution

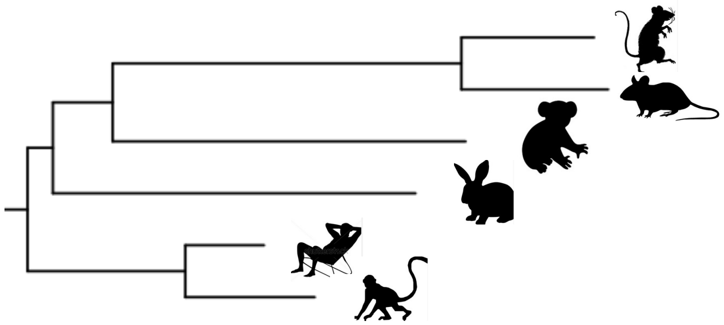
- Using alignment of 6 species
  - mouse (reference species)
  - rat
  - guinea pig
  - rabbit
  - human
  - marmoset (monkey)
- Run MONKEY on ~1 Mb regions of mouse chromosome 1 and orthologous regions to predict **HNF4a** TFBS

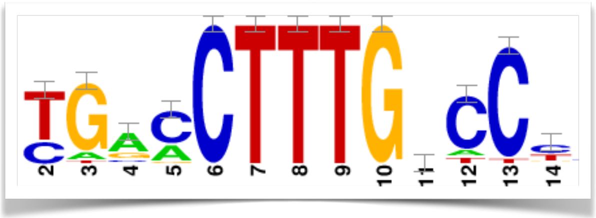




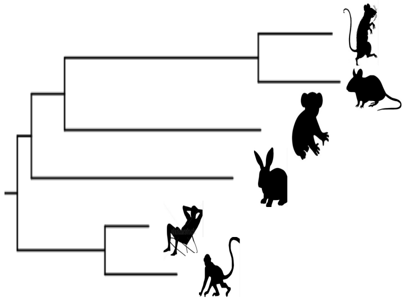
Model of TFBS evolution

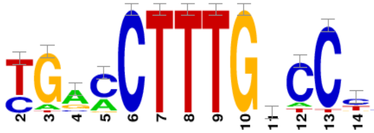
- 2 P-values are computed :
  - p-value of TFBS prediction on **mouse genome alone**
  - **combined p-value** including all genomes
- 113 TFBS requiring that
  Pval(mouse) < 1e-4
  (should be a TFBS in mouse !)
- 78 TFBS with
  - Pval(mouse) < 1e-4
  - AND Pval(combined) < 1e-4





Model of TFBS evolution





GTACCCTTTGTCAAA

ATGCCCTTTGT**T**AAA

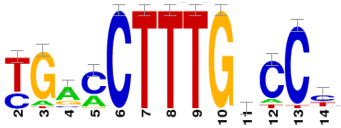
A**AA**TC**AC**TT**A**T**T**AAA

TTATTCTTTGT**A**GAA

A**C**ACTCTTTGT**T**AAA

ATACTCTTTGG**T**AAA

Pval(mouse) = 0.00104
Pval(combined) = 0.0538



CTGGA**G**TT**C**GACCTG

CTGGAGTTCGACCTG

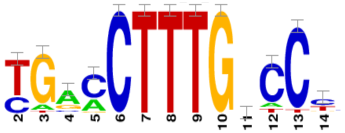
CTGGAGTTTGA**T**CTG

CTGGAGTTTGACCTG

CTGCAGTTTGACCTG

CTGGAGTTTGACCTG

Pval(mouse) = 0.0046
Pval(combined) = 1.1e-06



CTGGAGTTCGACCTG

CTGGAGTTCGACCTG

CTGGAGTTTGATCTG

CTGGAGTTTGACCTG

CTGCAGTTTGACCTG

CTGGAGTTTGACCTG

Pval(mouse) = 0.0046
Pval(combined) = 1.1e-06

Good TFBS in mouse
but weakly conserved

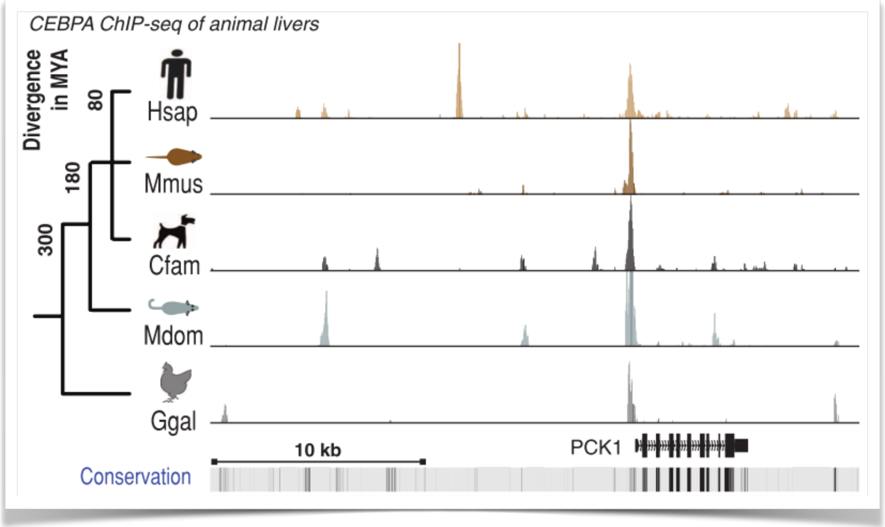
New mouse specific TFBS ?

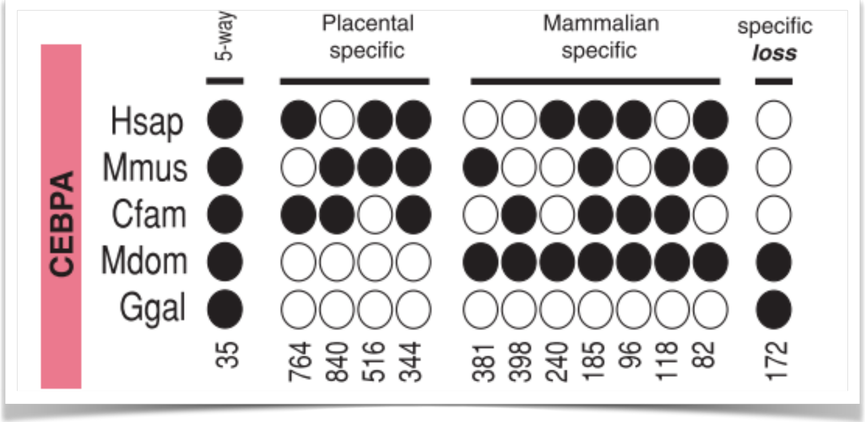
Fair TFBS in mouse
but strongly conserved

Low affinity binding site ?

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**





CEBPA	5-way	Placental specific				Mammalian specific							specific *loss*
Hsap	●	●	○	●	●	○	○	●	●	●	○	●	○
Mmus	●	○	●	●	●	●	○	○	●	○	●	●	○
Cfam	●	●	●	○	●	○	●	○	●	●	●	○	○
Mdom	●	○	○	○	○	●	●	●	●	●	●	●	●
Ggal	●	○	○	○	○	○	○	○	○	○	○	○	●
	35	764	840	516	344	381	398	240	185	96	118	82	172

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