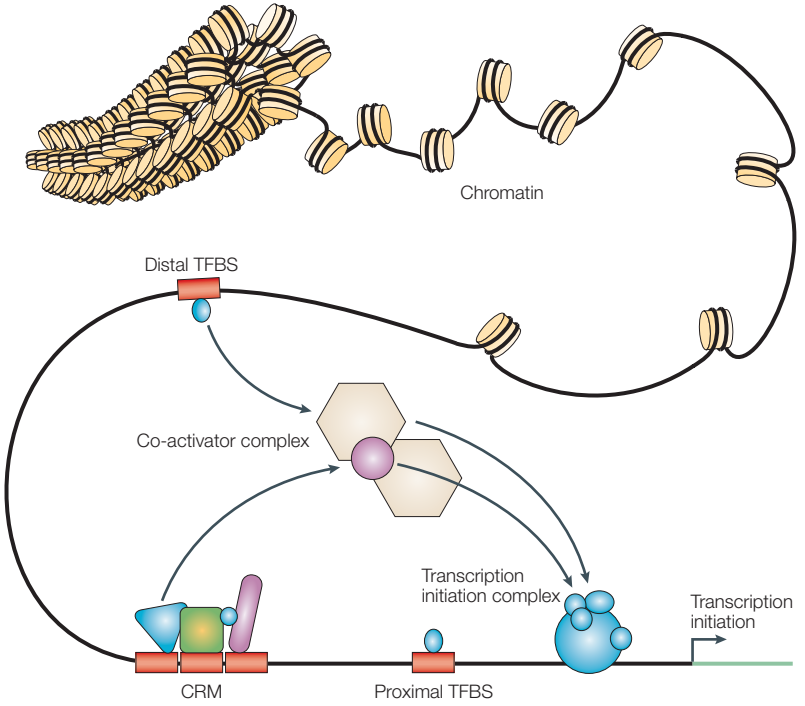
Régulation et conservation

Signaux de régulation de transcription

éléments *trans*-régulateurs (facteurs de transcription) et séquences *cis*-régulatrices (sites de liaison)



TFBS : site de liaison de facteur de transcription ; CRM : module *cis*-régulatoire

Wasserman & Sandelin *Nat Rev Genet* 5 :276 (2004)

Enhancers

animation (enhancer) : http ://www.maxanim.com/genetics/

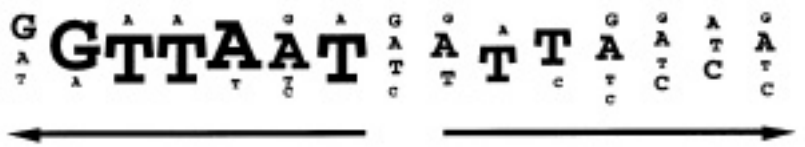
Évaluation de méthodes de recherche

Comment peut-on mesurer le succès de recherche de sites de liaison ?

Validation experimentale [site connu] : vérifier la liaison pour les sites dans le labo

Affinité et prédiction

On trouve beaucoup d'instances par un modèle de site de liaison de type PSSM.

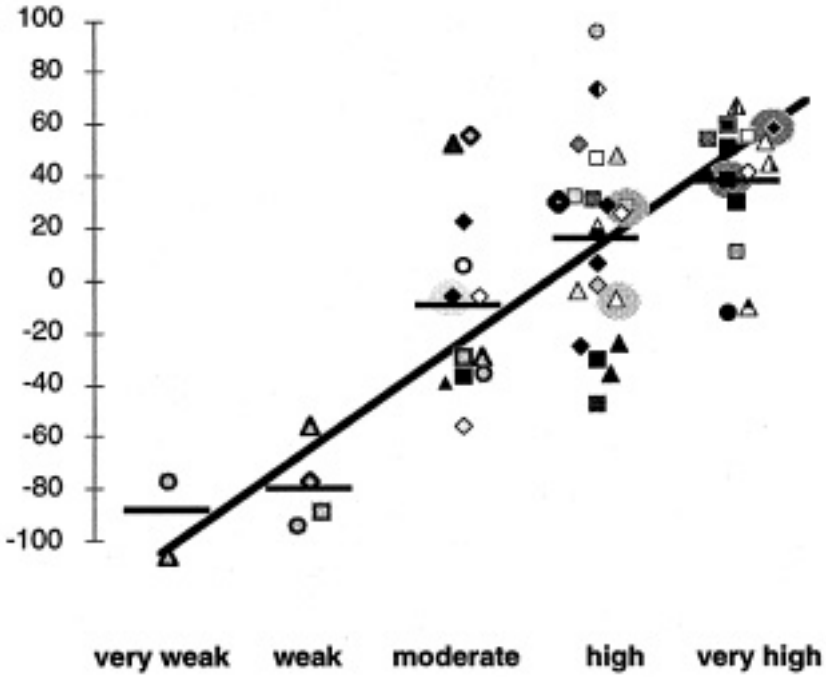


```
A   4-25-25-25 20 16-25 7 14   2-99 12  8  5 10
C -99-99-99-99-99  2-99 2-99 -99  4  1  8 13  7
G  14 20-99-99-99-25-99 4  1 -25-99  3  3-99  2
T   2-99 20 20-25  2 20 7  6  19 17  4  2  2  2
```

Est-ce que c'est le modèle que n'est pas assez spécifique ou plutôt le facteur de transcription ?

Tronche & al *J Mol Biol* 266 :231 (1997)

Affinité et prédiction



C'est vraiment aussi non-spécifique (0.1% des instances avec un rôle régulatoire)

Tronche & al *J Mol Biol* 266 :231 (1997)

Évaluation de méthodes de recherche 2

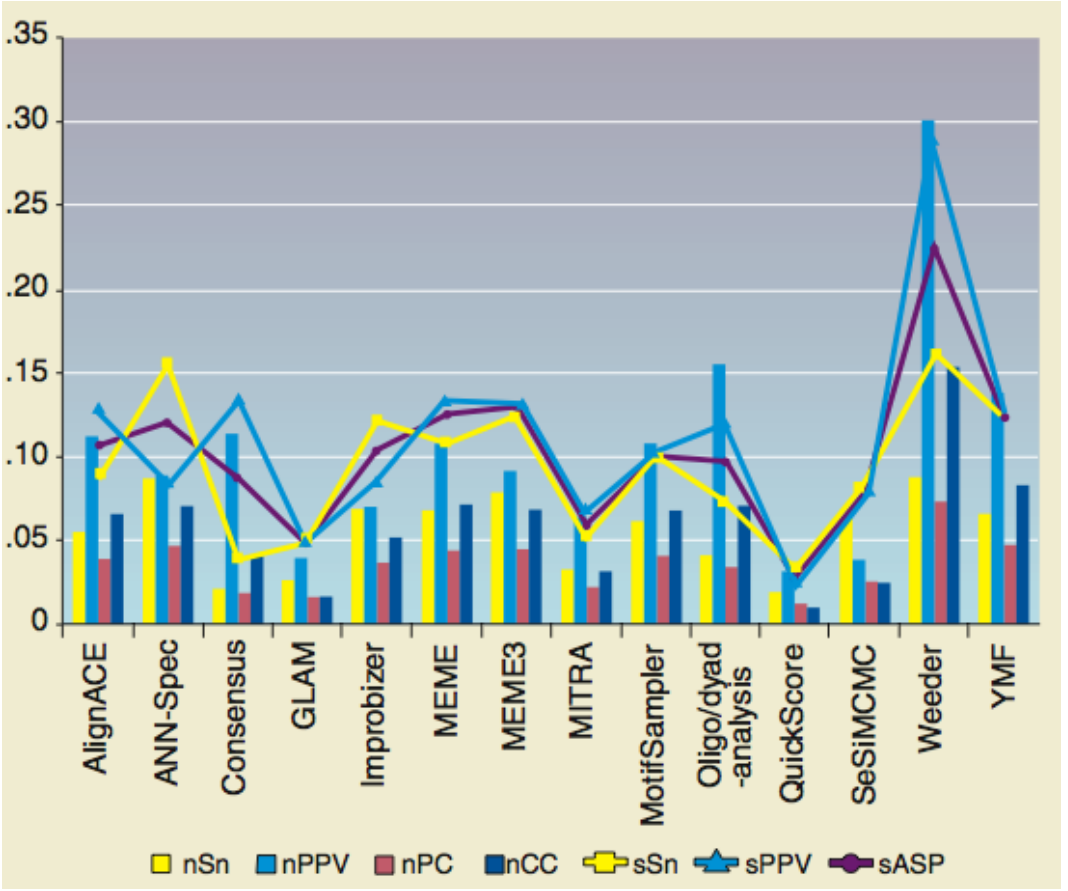
Validation experimentale [prédiction de novo] : utiliser des sites connus, compiler des données de test

nTP, nFN, nTN, nFP (nuclétoides, $\{\text{false}, \text{true}\} \times \{\text{pos}, \text{neg}\}$)
sTP, sFN, sFP (sites)

sensibilité (*sensitivity*) : $\text{xSn} = \frac{\text{xTP}}{\text{xTP}+\text{xFN}}$ avec x=s (site) ou x=n (nucléotide)
valeur prédictive $\text{xPPV} = \frac{\text{xTP}}{\text{xTP}+\text{xFP}}$
coefficient de performance $\text{xPC} = \frac{\text{xTP}}{\text{xTP}+\text{xFP}+\text{xFN}}$

Problème : il y a des inconnus inconnus — on ne connaît pas tous les sites de liaison

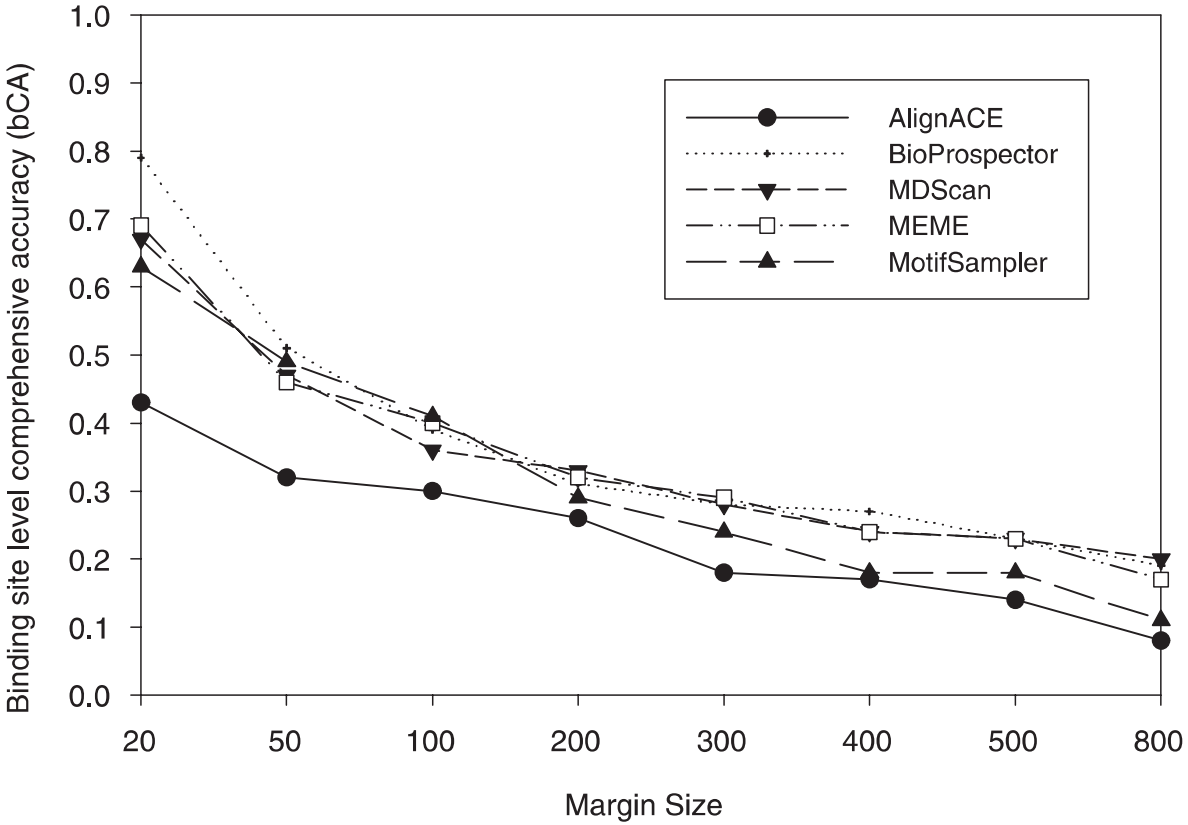
Évaluation de méthodes 3



(humain, souris, mouche, levure ; pas comparative)

Tompa & al. *Nat Biotech* 23 :137 (2005)

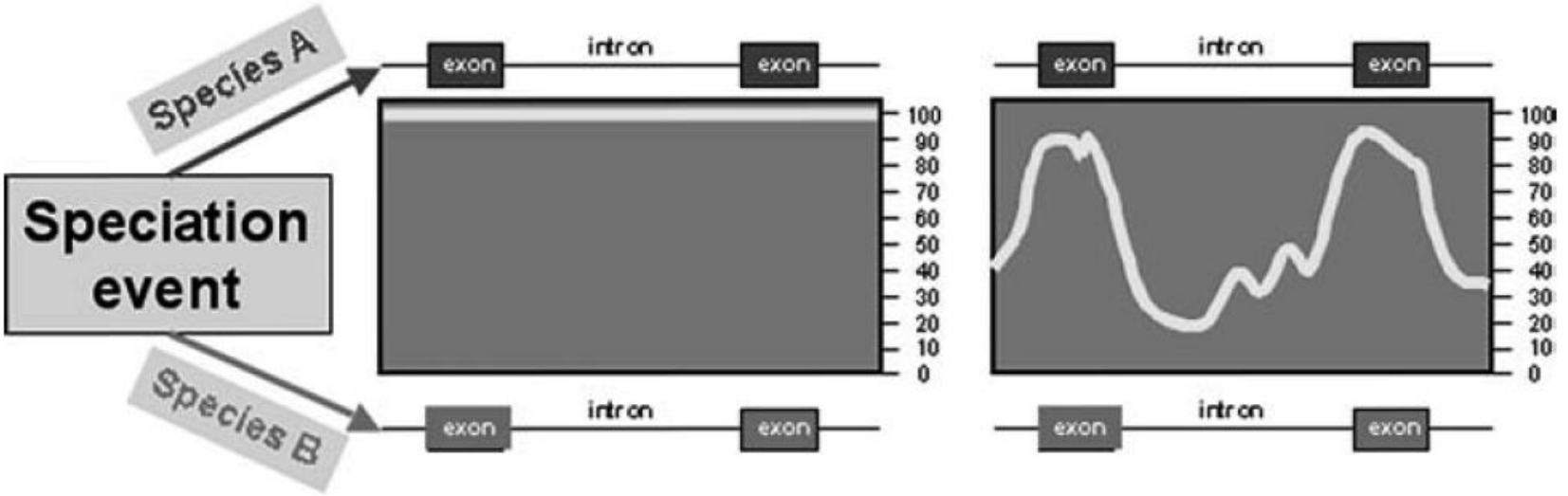
Évaluation de méthodes 4



(E. coli ; succès en fonction de la longueur de séquence)

Hu & al *Nucleic Acids Res* 33 :4899 (2005)

Génomique comparative

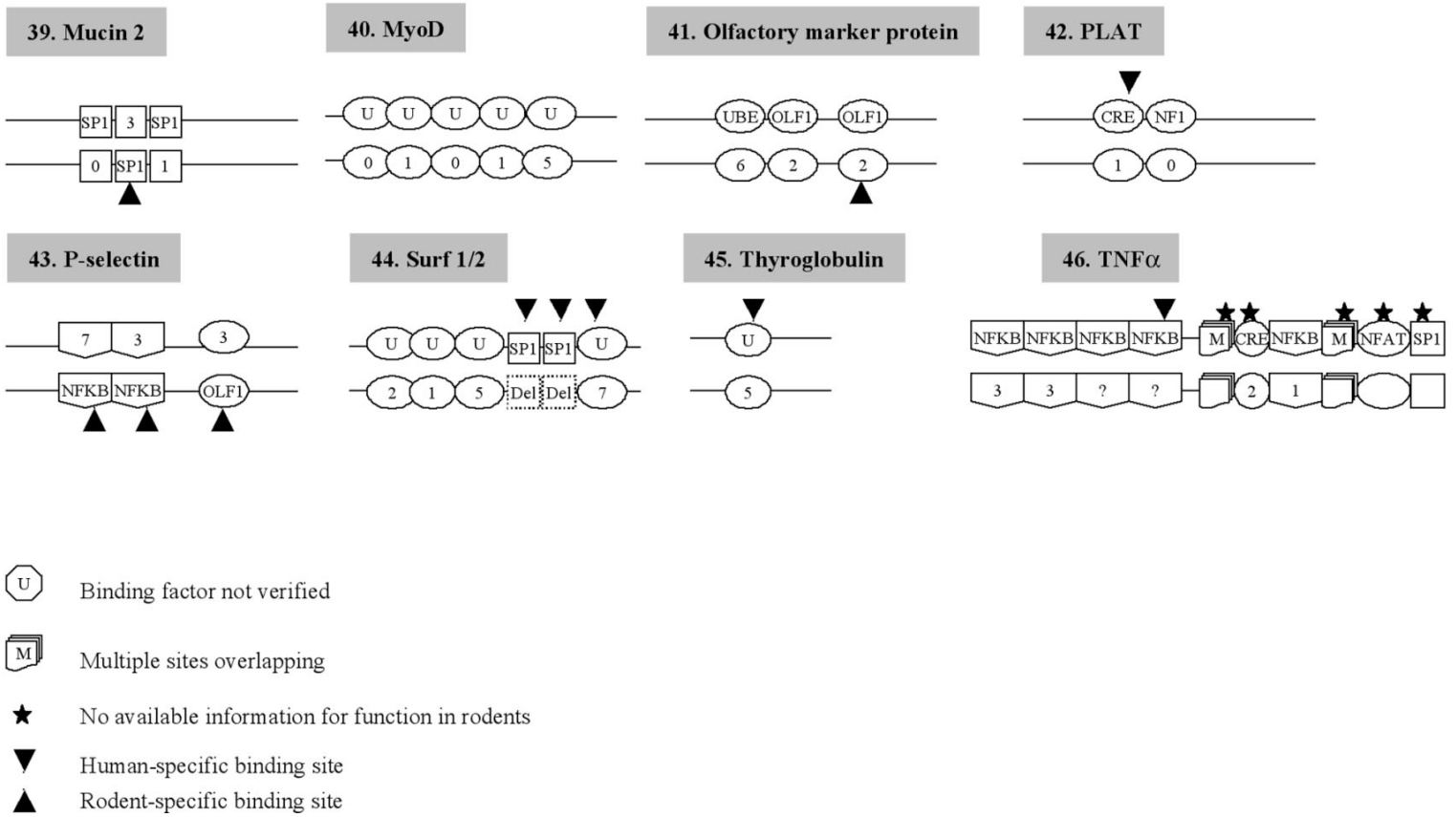


Principe de génomique comparative : éléments fonctionnels sont plus conservés (séléction négative) que les éléments non-fonctionnels (évolution neutre)

Miller & al. *Annu Rev Genomics Hum Genet* 5 :15 (2004)

Méthodes comparatives ?

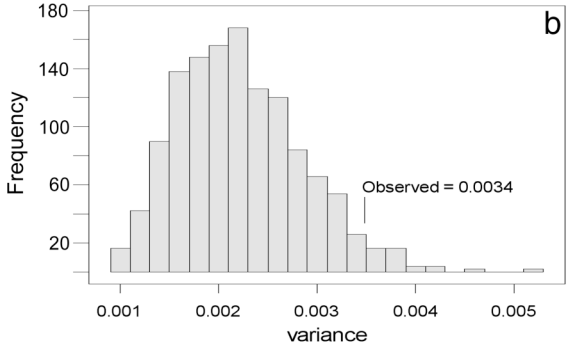
Est-ce que les signaux *cis* évoluent plus lentement ?



Dermitzakis & Clark *Mol Biol Evol* 19 :1114 (2002)

Évolution de sites de liaison

1. le taux d'évolution est variable



(comparaison de la variance de divergence entre séquences aléatoires et vraies)

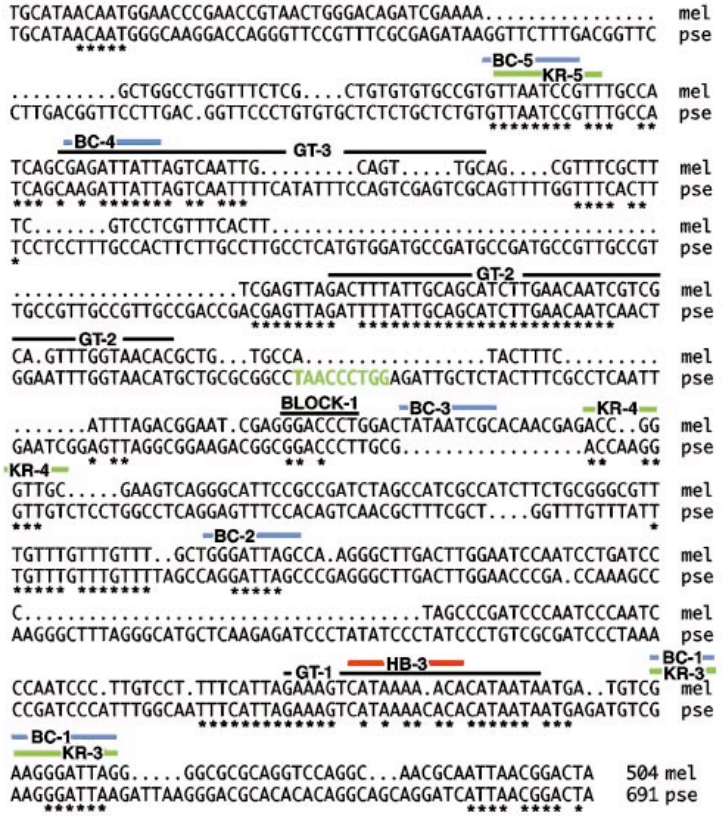
2. divergence (Kimura 2P) entre humain et souris TF : 0.27 ± 0.18, synonyme : 0.47 ± 0.17, non-syn : 0.09 ± 0.1, background : 0.4 ± 0.18

3. turnover : à peu près $1/3$ des sites sont spécifique à un des espèces (humain ou souris)

Dermitzakis & Clark *Mol Biol Evol* 19 :1114 (2002)

Évolution de sites de liaison 2

enhancer du gène *eve* dans Drosophila



⋆ préservé parmi 13 espèces, . trou

Ludwig & al. *Nature* 403 :564 (2000)

Évolution de sites de liaison 3

L'expression de *eve* est la même . . .

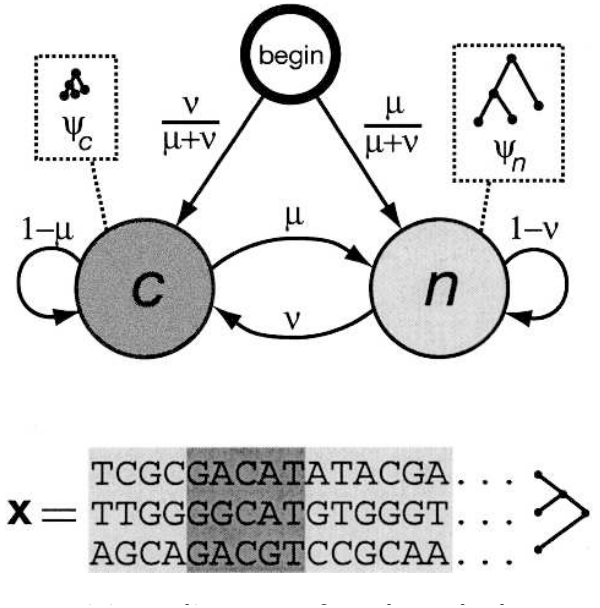
Est-ce que les différences dans l'enhancer ont des conséquences fonctionnels ?

Oui : si on remplace par des séquences chimeras (Dmel+Dpse), l'expression change.

Donc les mutations dans l'élément *cis* sont accompagnées par des mutations dans les éléments *trans*

Phylo-HMM

phastCons

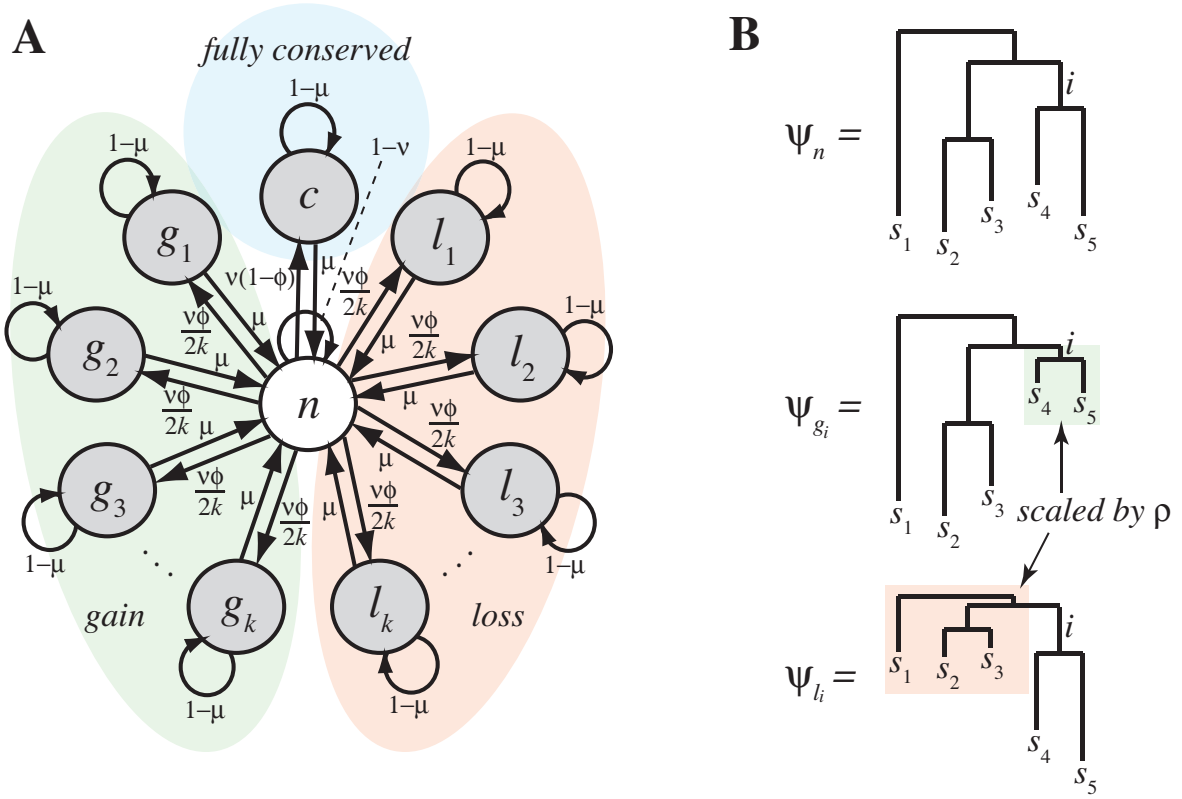


émissions : colonnes de l'alignement multiple, avec probabilités de transition $e^{\mathbf{Q}t}$ (neutre) ou $e^{\rho\mathbf{Q}t}$ (sélection négative avec $\rho < 0$)

Siepel & al *Genome Res* 15 :1034 (2005)

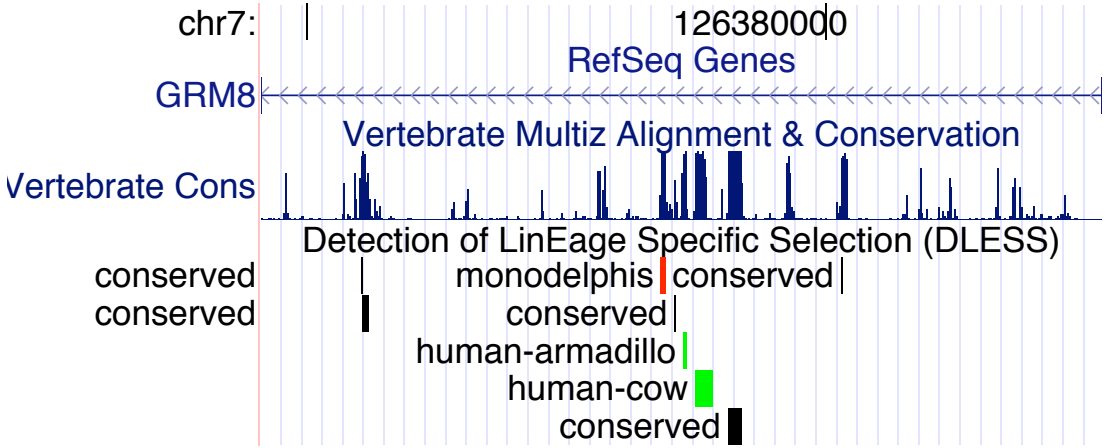
Turnover

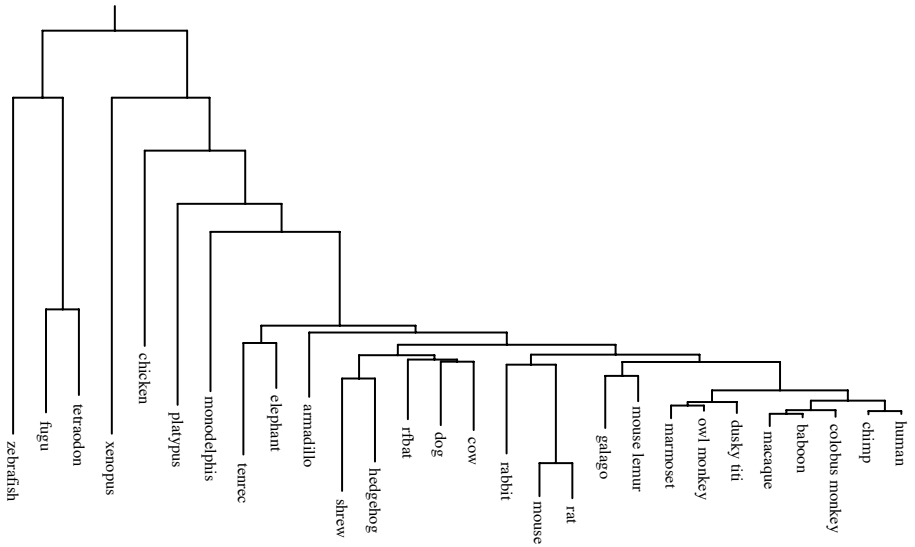
Modélisation par phylo-HMM



Siepel & al. *RECOMB 2006*

DLESS ∈ goldenPath/hg17





Siepel & al. *RECOMB 2006*

Regulatory potential

ordre 5, alphabet de 10 symboles pour alignment de humain-souris-rat

Symb=1		2	3		4	5	6	7	8	9	10
ACG	G-C	AAA	AAC	TAG	ACC tv	AA-	CCC	GGG	TTT	ATT tv	CAA tv
AC-	G-G		AAG	TAT	AGA ts	AGG ts			-CC	CTT ts	CGG tv
AGC	G-T		AAT	TC-	A--	CTC ts				GGA ts	GCC tv
ATG	TAC		ACA	TGA	CC-	TCT ts				TGG tv	GG-
AT-	TA-		ACT	TGT	GAA ts	TT-				T--	G--
A-C	TCA		AGT	TTA	GAG ts					-T-	TAA tv
A-G	TCC		AG-	TTG	TTC ts					--C	-A-
A-T	TCG		ATA	T-C	T-T						-GG
CAT	TGC		ATC	--A	-AA						-G-
CA-	TG-		A-A		--G						-TT
CCT	T-A		CAC		--T						
CGA	T-G		CAG								
CGT	-AC		CCA								
CG-	-AG		CCG								
CTG	-AT		CGC								
C-A	-CA		CTA								
C-C	-CG		CT-								
C-G	-CT		C-T								
GAC	-C-		C--								
GA-	-GA		GAT								
GCA	-GC		GCG								
GC-	-GT		GCT								
GTA	-TA		GGC								
GTT	-TC		GGT								
GT-	-TG		GTC								
G-A			GTG								

Kolbe & al *Genome Res* 14 :700 (2004)

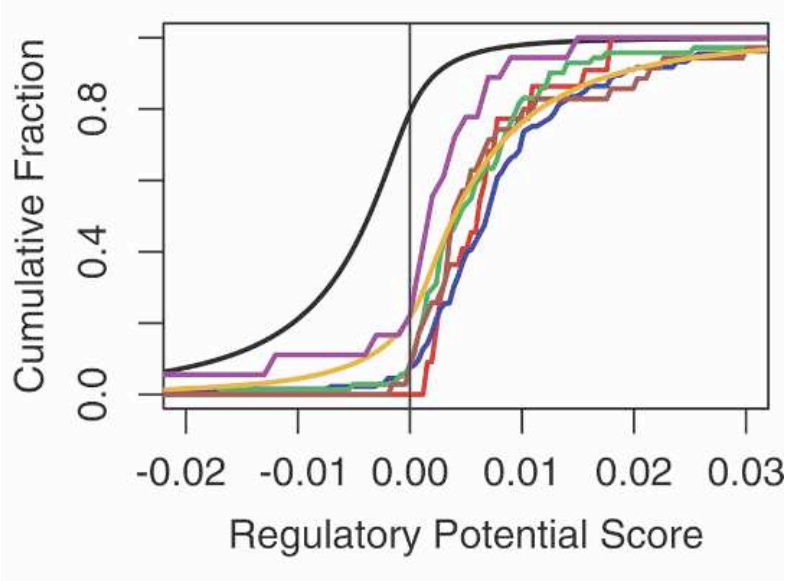
Regulatory potential 2

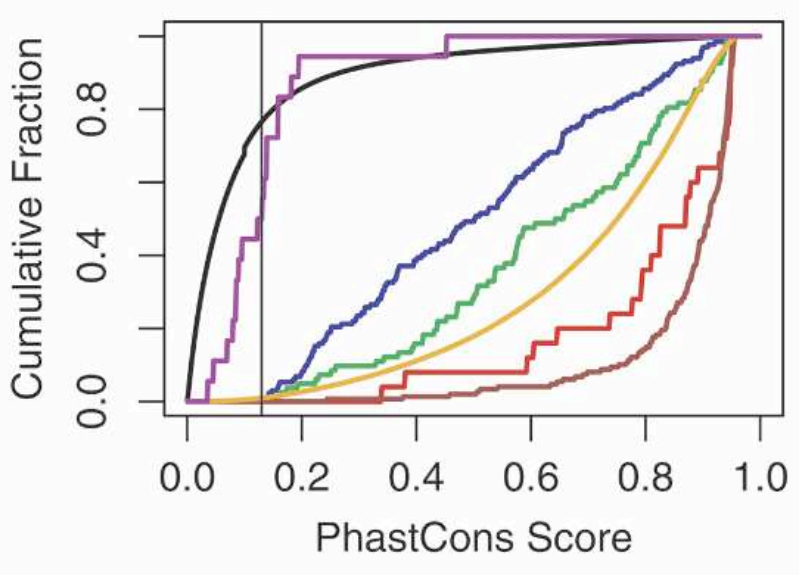
RP score : $\mathsf{RP} = \sum_i \log \frac{p_{\mathsf{REG}}\left(S[i] \middle| S[i-k..i-1]\right)}{p_{\mathsf{AR}}\left(S[i] \middle| S[i-k..i-1]\right)}$

p_{REG} : paramètres estimés d'un échantillon de sites *cis*-régulatoires

p_{AR} : paramètres estimés d'un échantillon d'éléments ancients répétés

Performance





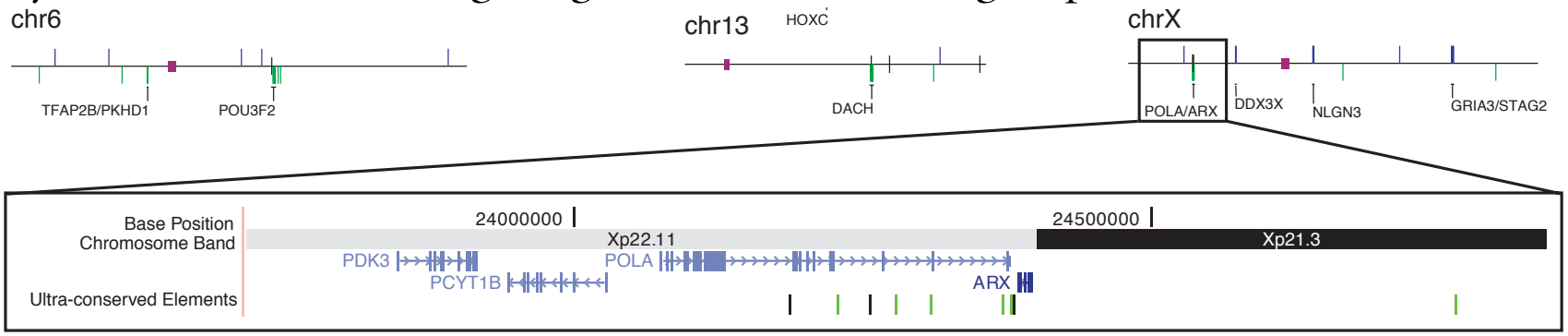
King & al *Genome Res* 15 :1051 (2005)

Conservation extrème

On peut chercher des cas d'évolution ralentie partout dans le génome (et non pas seulement près de gènes)

Élément ultra-conservé : 100% d'identité, longueur au moins 200pb (p.e., entre humain-souris)

Il y en a 481, tout au long du génome, souvent en groupes



Peut–être en exons, en introns, ou en régions intergéniques

Bejerano & al *Science* 304 :1321 (2004)

Conservation extrème 2

Quelle est la fonction des éléments ultra-conservés ?

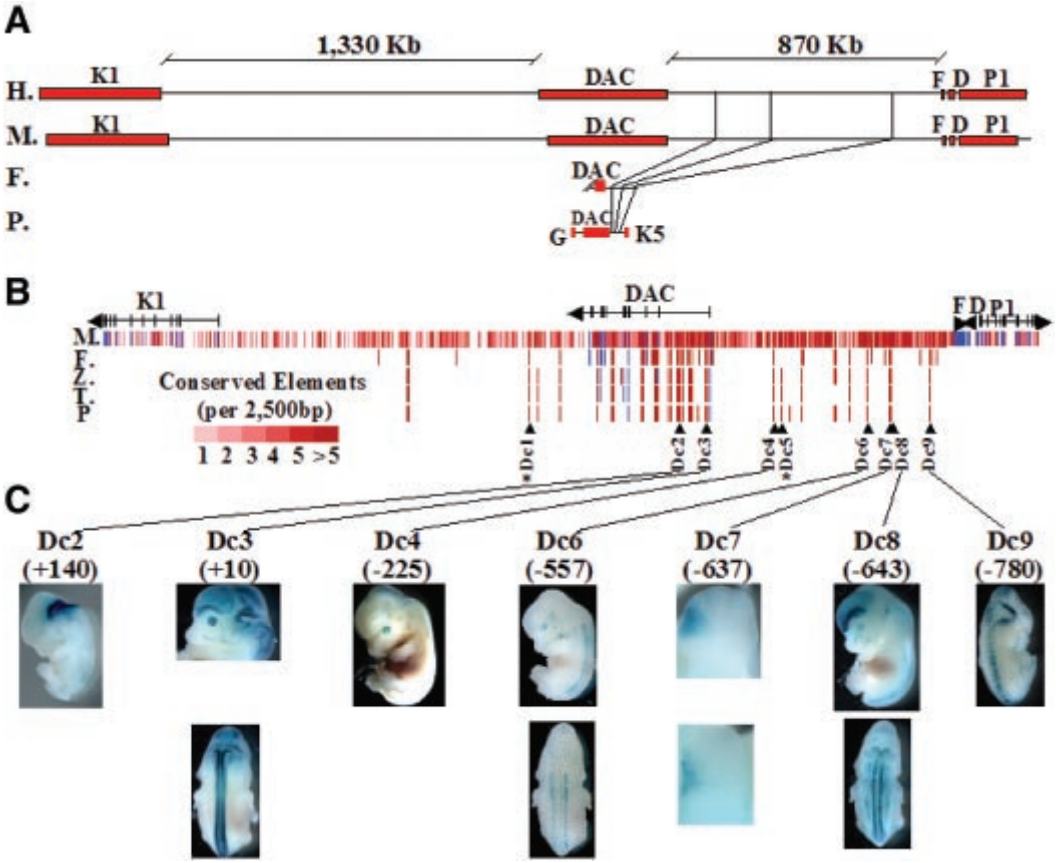
Possibilités : gènes ARN ou sites *cis*-régulatoires

Beaucoup d'entre eux s'alignent avec poulet et même poisson (Fugu)

Ils se trouvent souvent dans des déserts de gènes

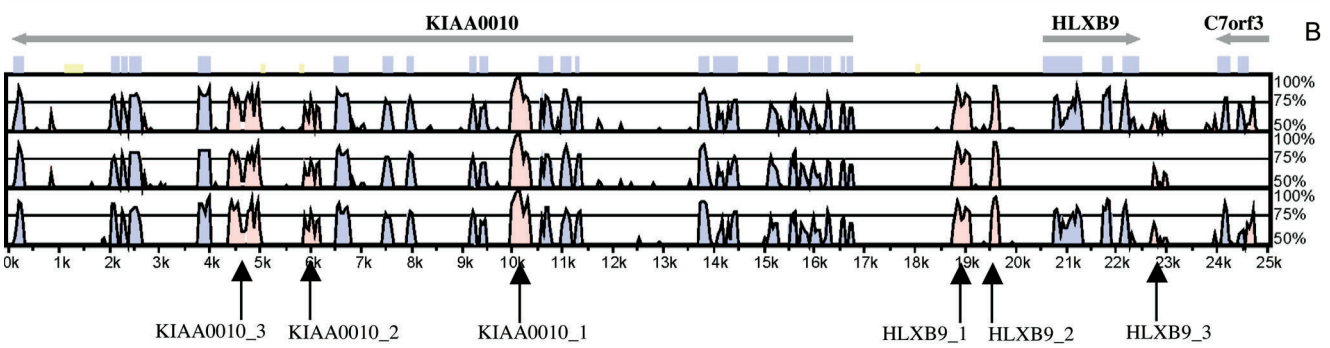
Régulation de développement

Éléments non-codants conservés (p.e. humain-souris au moins 70% identité et longueur 100pb)



Nobrega & al *Science* 302 :413 (2003)

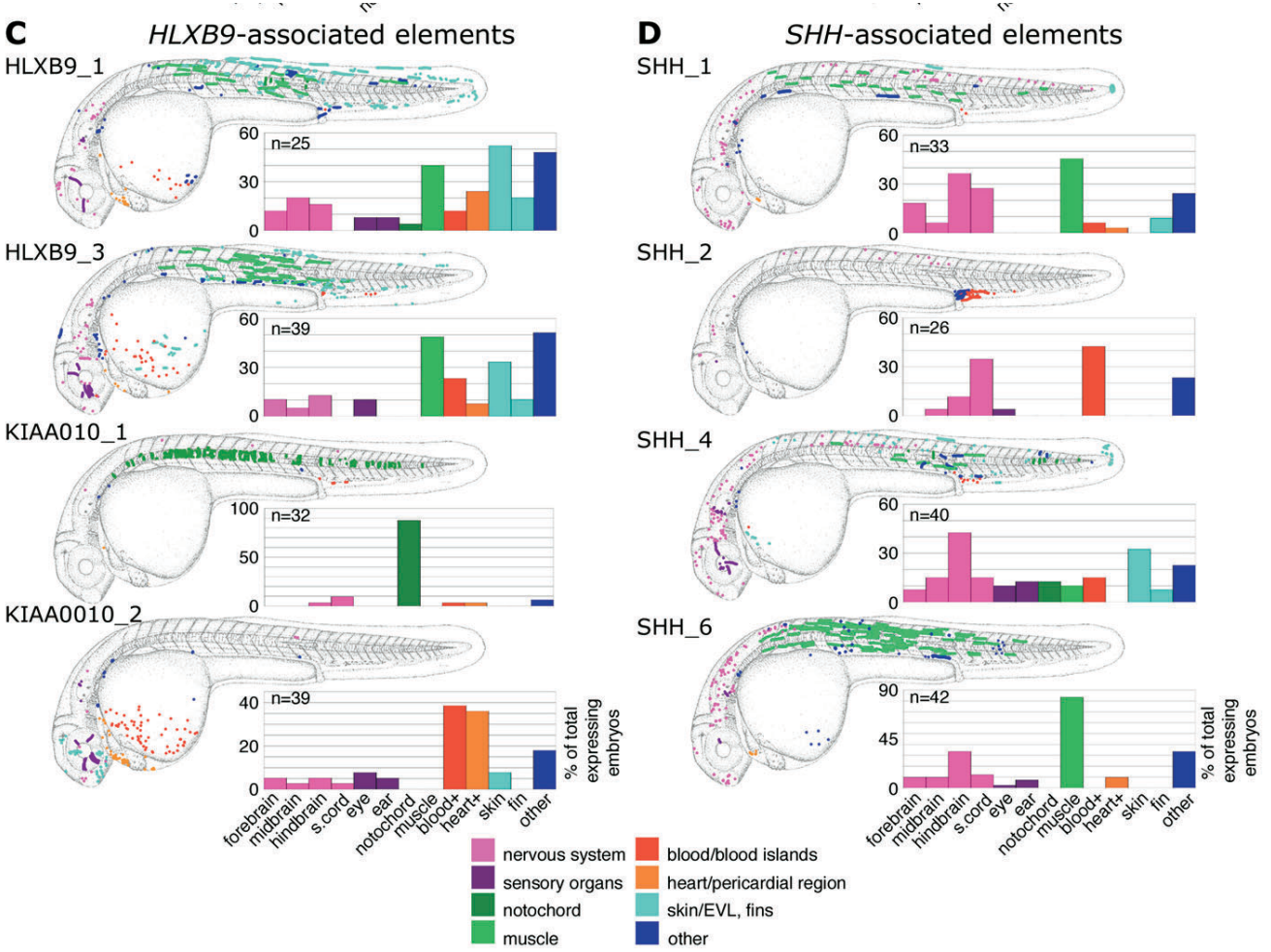
Régulation de développement 2



(humain, souris, rat)

Woolfe & al *PLoS Biology* 3 :e7 (2005)

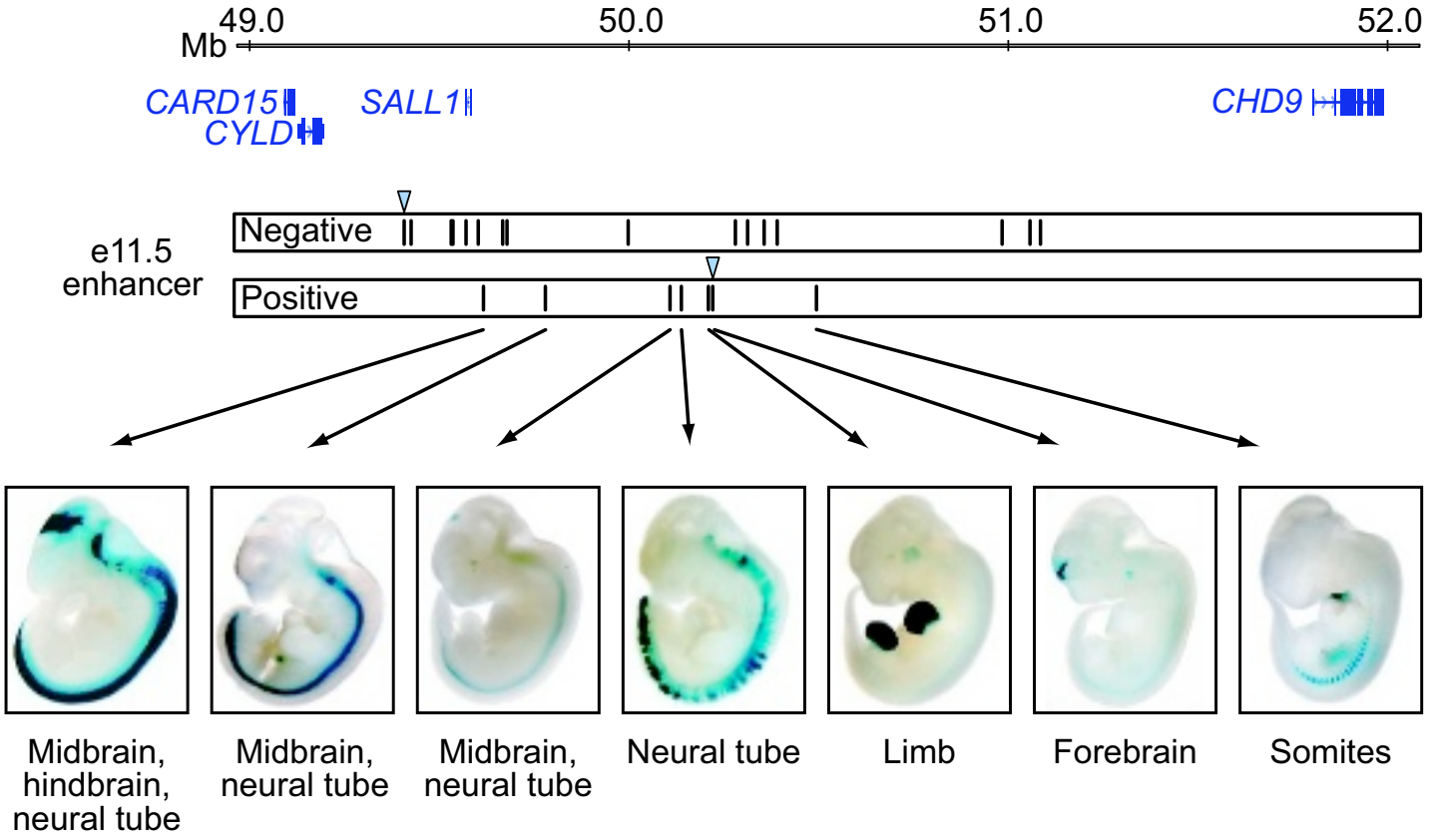
Régulation de développement 3



Woolfe & al *PLoS Biology* 3 :e7 (2005)

Régulation de développement 4

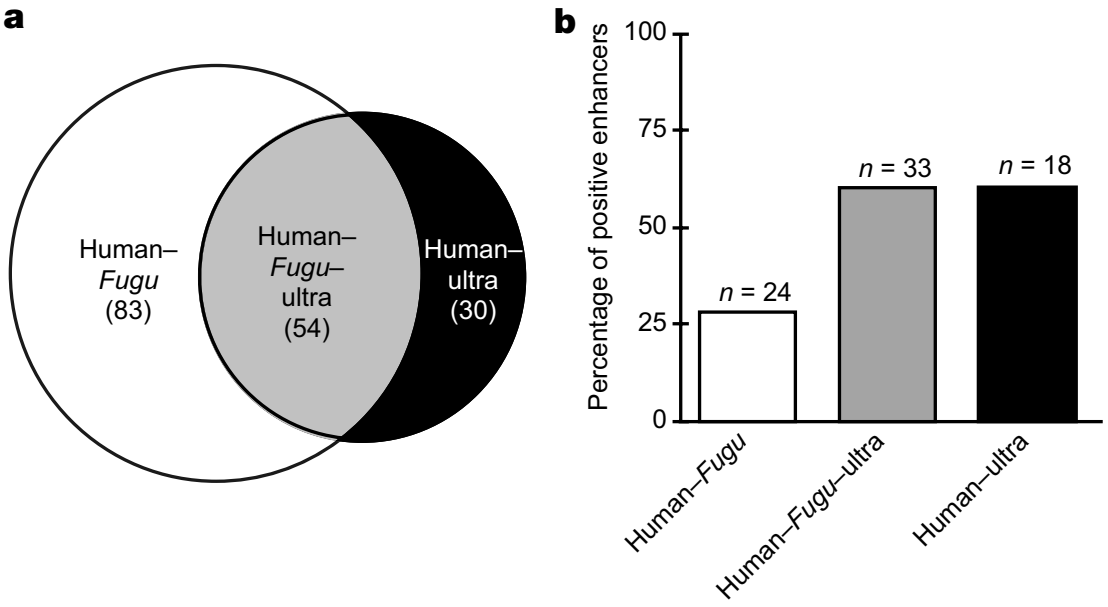
Dernier exemple (démonstration de rôle régulatoire)



Pennacchio & al *Nature* doi :10.1038/nature05295

Régulation de développement 5

à peu près 45% des éléments conservés entre humain et Fugu ou ultra-coonservé entre humain, souris et rat ont démontré d'être des enhancers

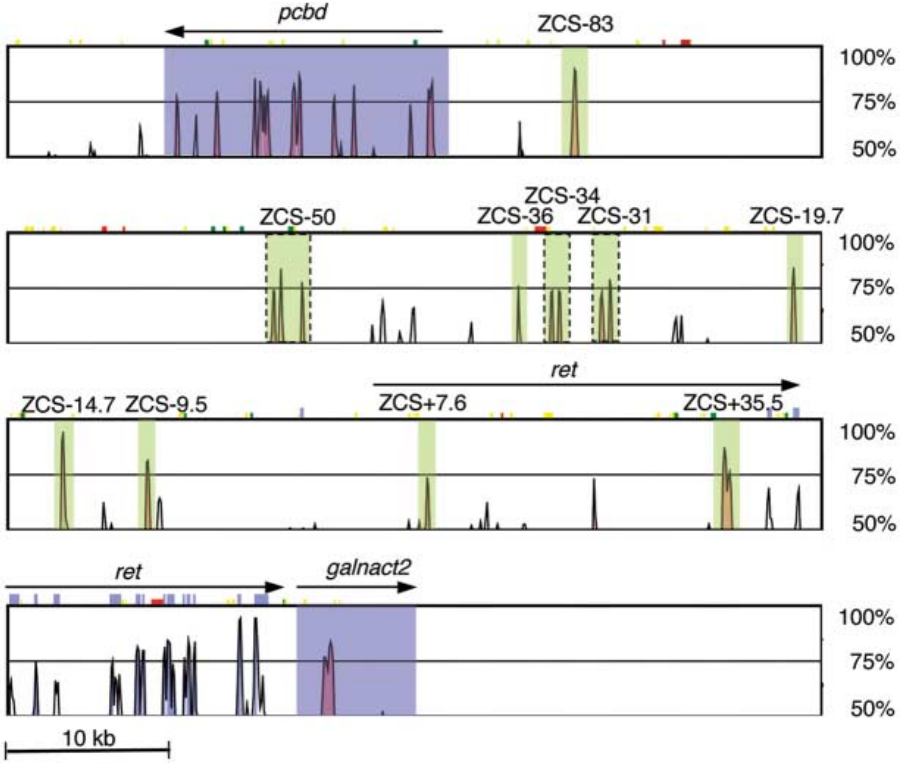


extraction des plus fréquents motifs (énumeration de 5-mers) + scoring d'autres éléments fréquents par l'occurrence de ces motifs
→ mieux qu'utiliser seulement la conservation

Pennacchio & al *Nature* doi :10.1038/nature05295

Conservation ?

expérience de Fisher & al (2006) : séquences conservés en poissons, et en mammifères (mais pas entre les deux !) dans la région régulatoire du gène *ret*



Fisher & al *Science* 312 :276 (2006)

Conservation??

les séquences humaines implantées dans des embryos de poisson ont contrôlé l'expression du gène

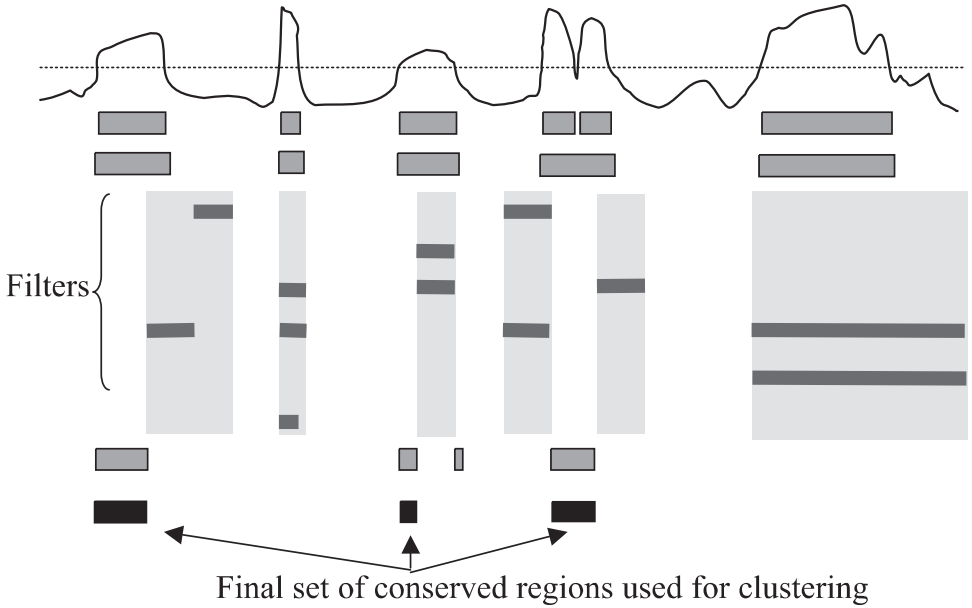
Constructs	Brain	SC	CG	ENS	NTC	OLF	Retina	Heart	IM/PND	Fin bud
ZCS-83	+	+	+			+				
ZCS-50	+	+	+		+			+		+
ZCS-36	+								+*	
ZCS-34	+								+	
ZCS-31	+								+	
ZCS-19.7	+	+	+	+			+		+	
ZCS-14.7	+	+								
ZCS-9.5	+	+	+							
ZCS+7.6									+	
ZCS+35.5		+	+	+				+		
HCS-32	+	+	+					+		+
HCS-30									+	
HCS-23					+					
HCS-12		+							+*	
HCS-8.7									+	
HCS-7.4										
HCS-5.2	+					+			+	
HCS+9.7				+					+	
HCS+16	+									
HCS+19	+	+								

Fisher & al *Science* 312 :276 (2006)

Inventoire

Groupage d'éléments non-codant conservés

1. identification

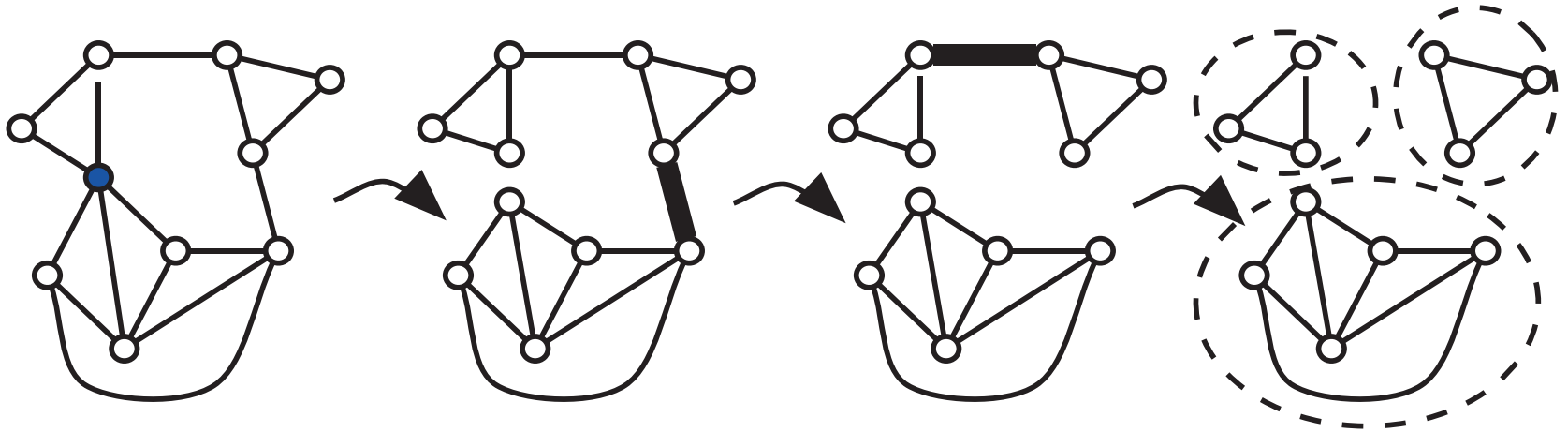


Human-mouse-rat sequence conservation	Fraction of genome	Number of regions
Top ~5%	5.11%	1055823
Extend/merge	5.835%	969857
Remove repeats	5.393%	959820
Remove coding exons	4.541%	1074181
Remove cds-like	4.239%	1084945
Remove non-syntenic	4.086%	1072148
Remove seg. dups.	4.006%	1043450
Remove pseudogenes	4.001%	1042608
Remove <50bp	3.727%	699647

Bejerano & al *Bioinformatics* 20 :i40 (2004)

Inventoire 2

2. groupage



Bejerano & al *Bioinformatics* 20 :i40 (2004)

Inventoire 3

Groupes

- gènes ARN

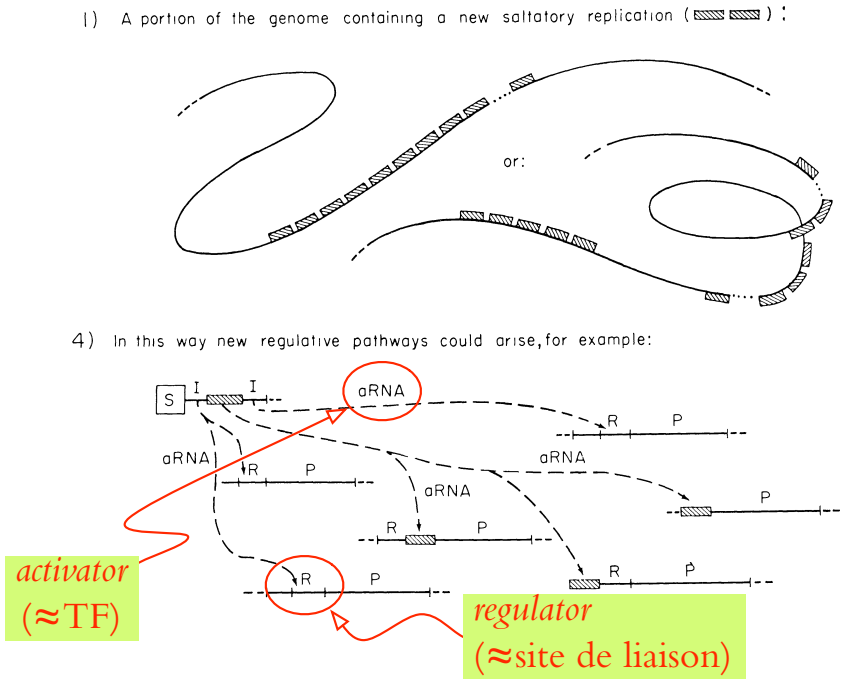
- nouveaux gènes codant

- éléments de régulation de transcription ou épissage

Origines

D'où viennnent les éléments *cis*-régulatoires ?

une théorie ancienne : Britten & Davidson (1971)

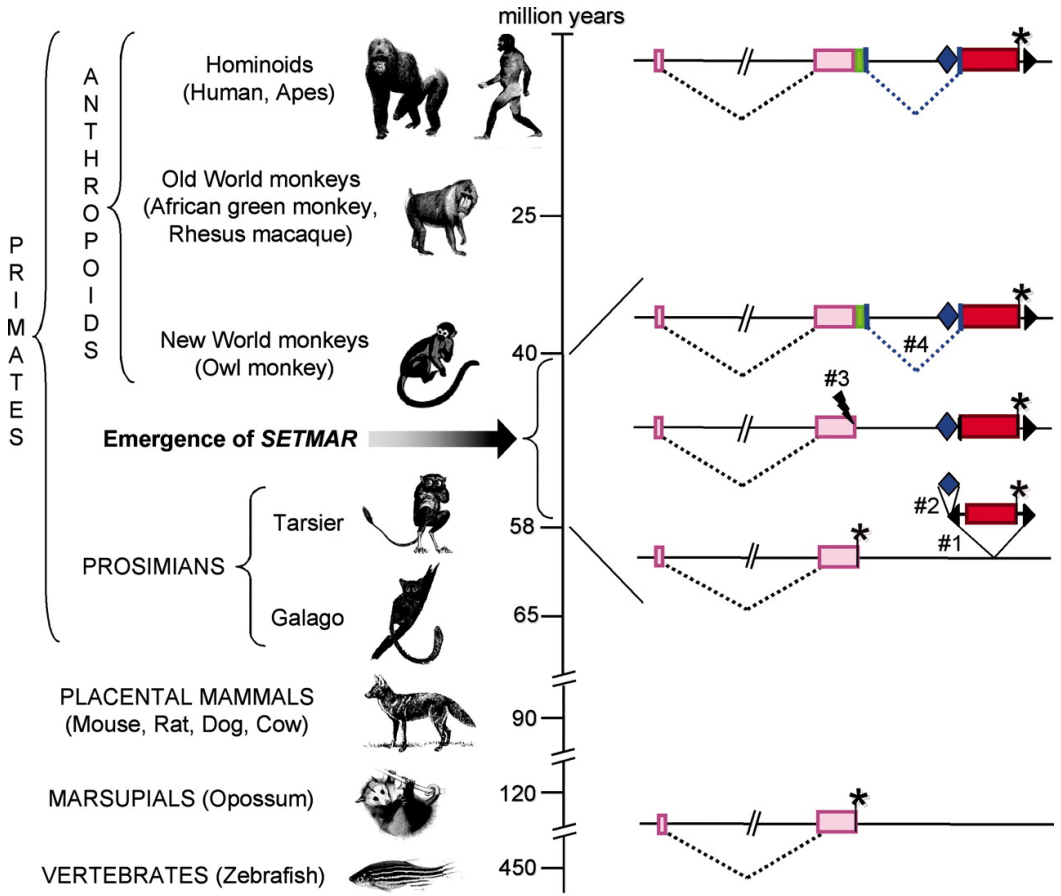


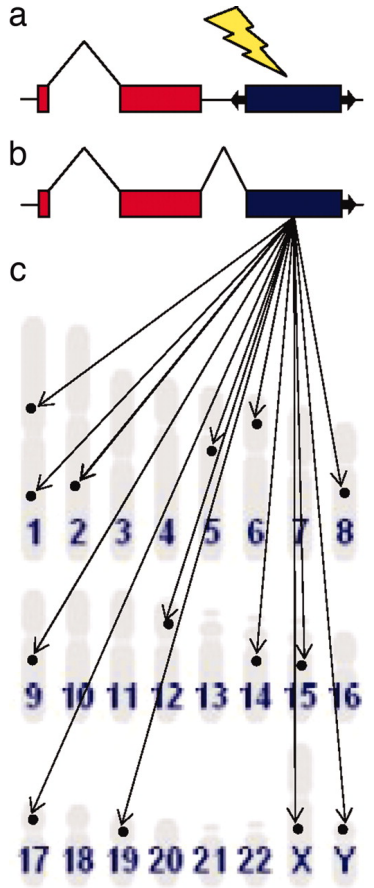
Britten & Davidson *Q Rev Biol* 46 :111 (1971)

Transposition

Animation : http ://www.maxanim.com/genetics/Transposition/Transposition.swf

Validation récente

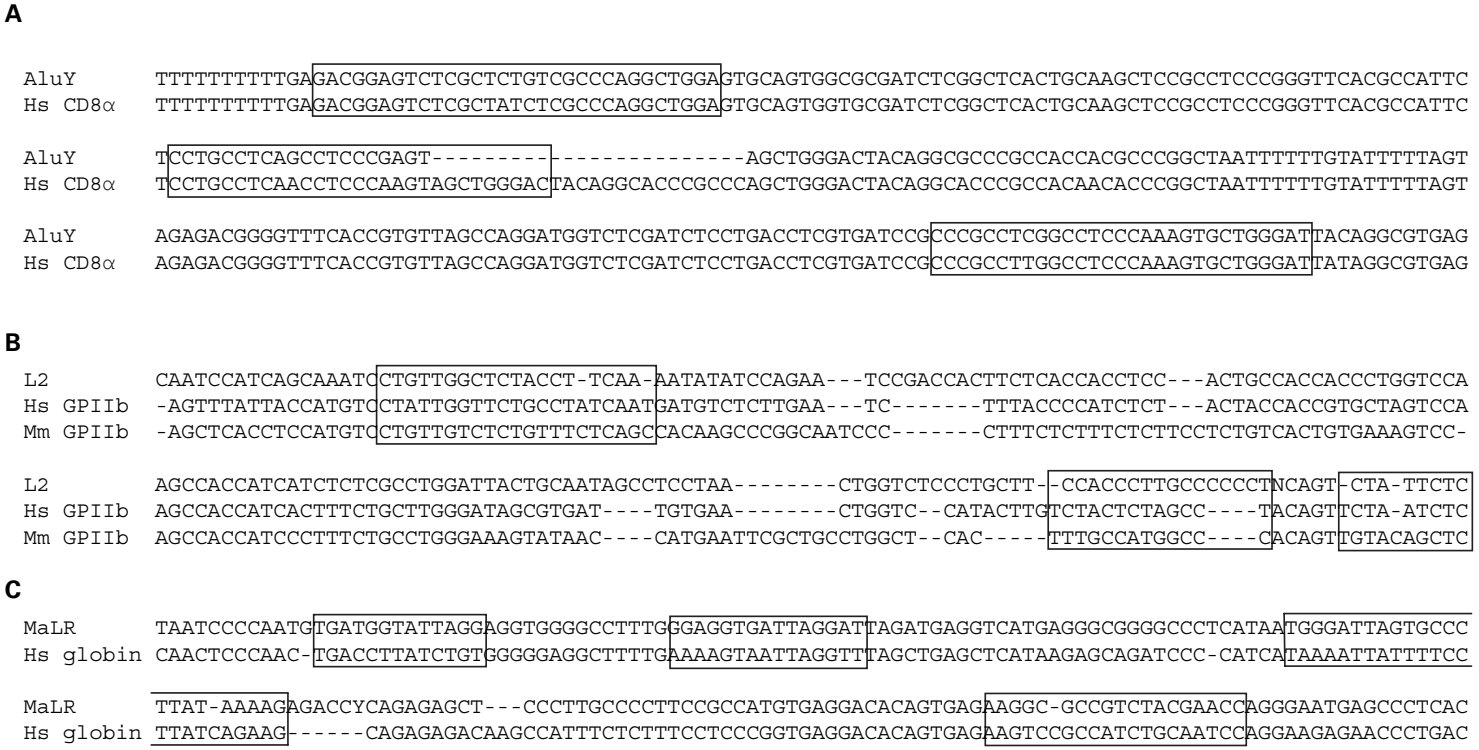




Cordaux & al. *PNAS* 103 :8101 (2006) ; Jordan *PNAS* 103 :7941 (2006)

D'autres exemples

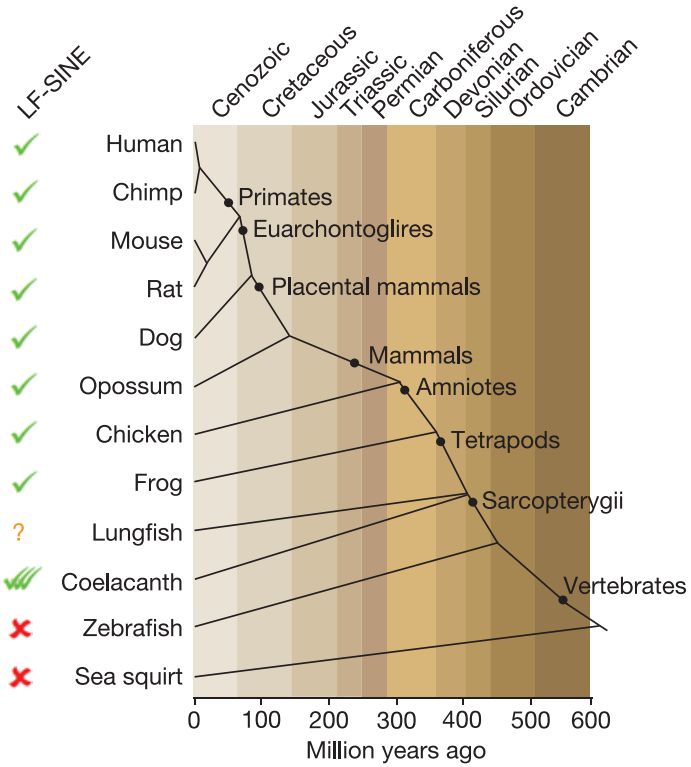
éléments *cis*-regulatoires provenant de l'insertion d'éléments répétés



Mariño-Ramírez & al *Cytogenet Genome Res* 110 :333 (2005)

Un élément ancien

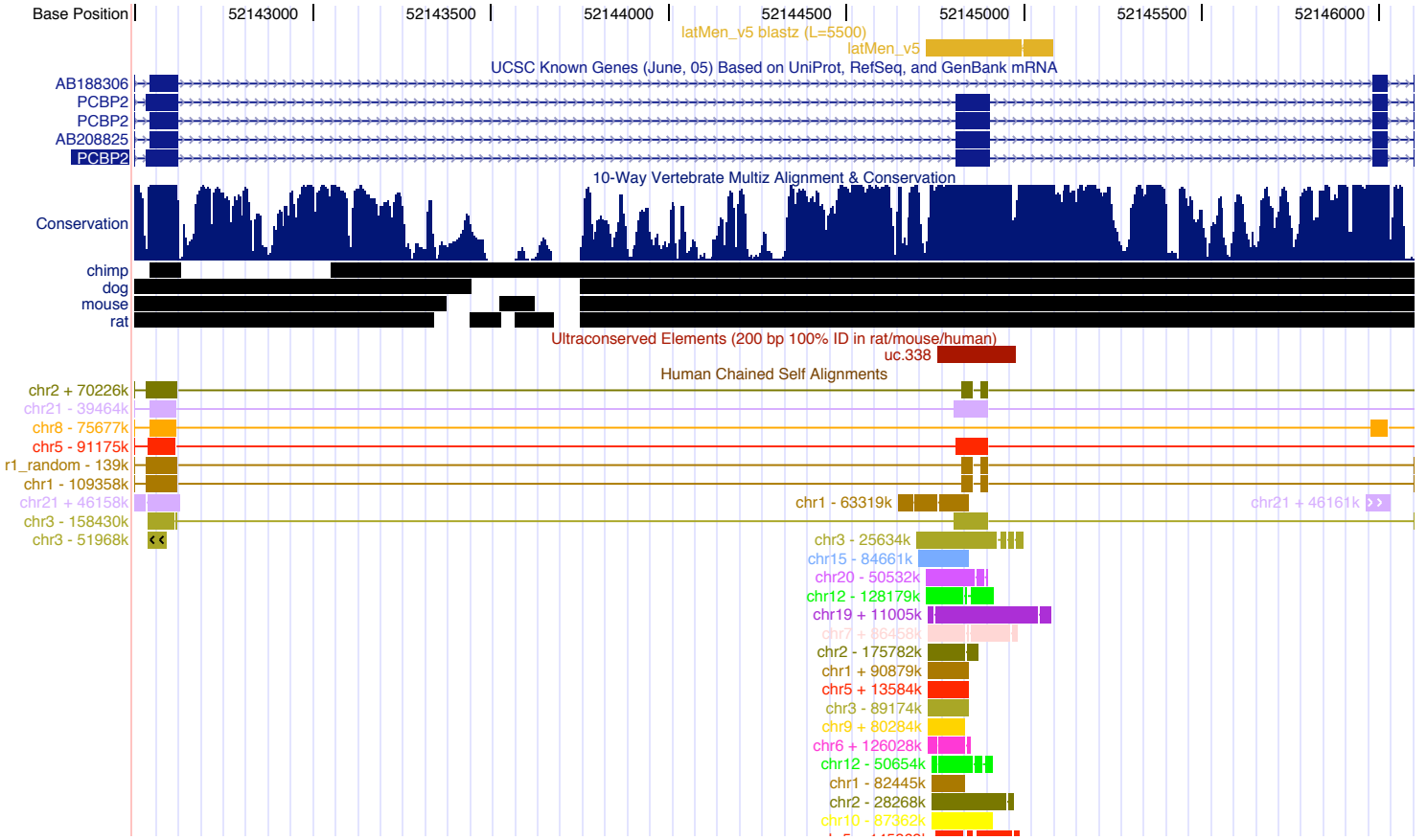
Un élément répété (LF-SINE) identifié dans *Latimeria menadœnsis*



Pouyaud & al *CR Académie des Sciences III* 322 :261 (1998) ; Bejerano & al *Nature* 441 :87 (2006)

Un élément ancien 2

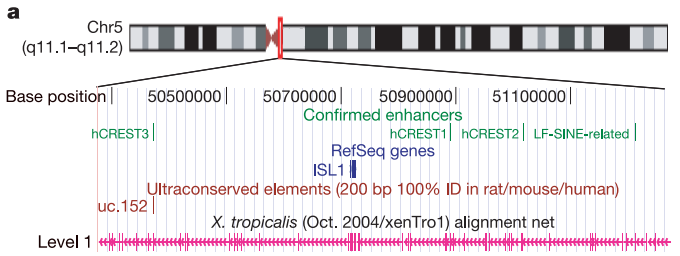
chevauchant un élément exonique ultra-conservé



Bejerano & al *Nature* 441 :87 (2006)

Un élément ancien 3

dans un désert intergénique proche de gène ISL1 impliqué dans le développement de neurones



c'est en fait un enhancer (démontré en embryos de souris transgéniques)

Bejerano & al *Nature* 441 :87 (2006)