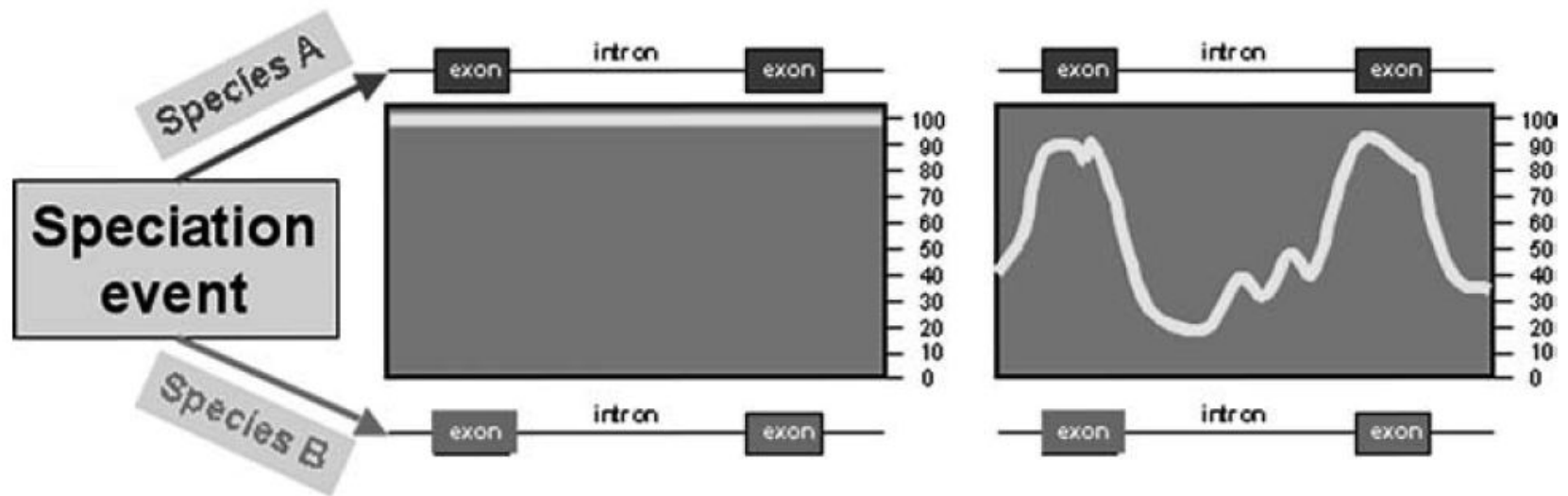


CONSERVATION DE SÉQUENCE

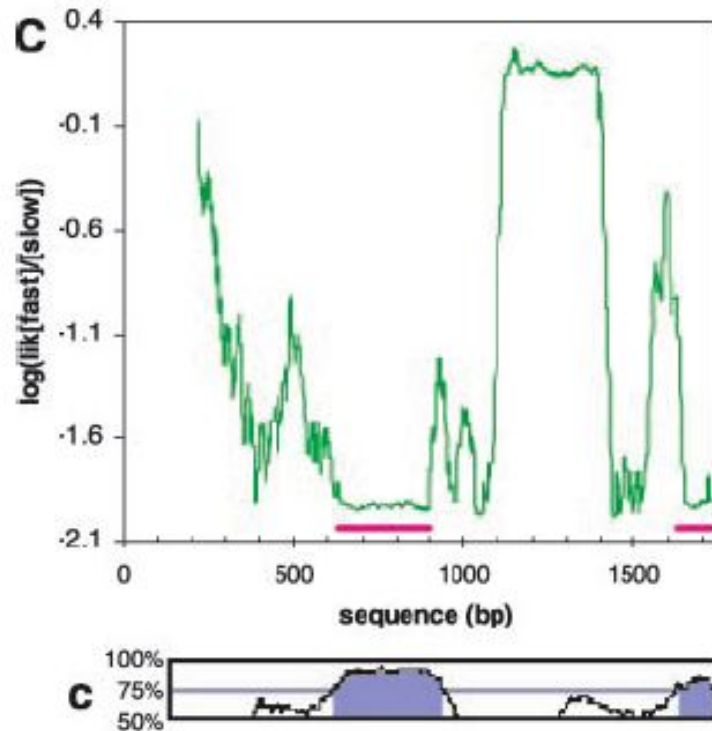
Génomique comparative



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

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

L'ombre de sélection dans une fenêtre



(Ici, on connaît μ , Q , ρ)

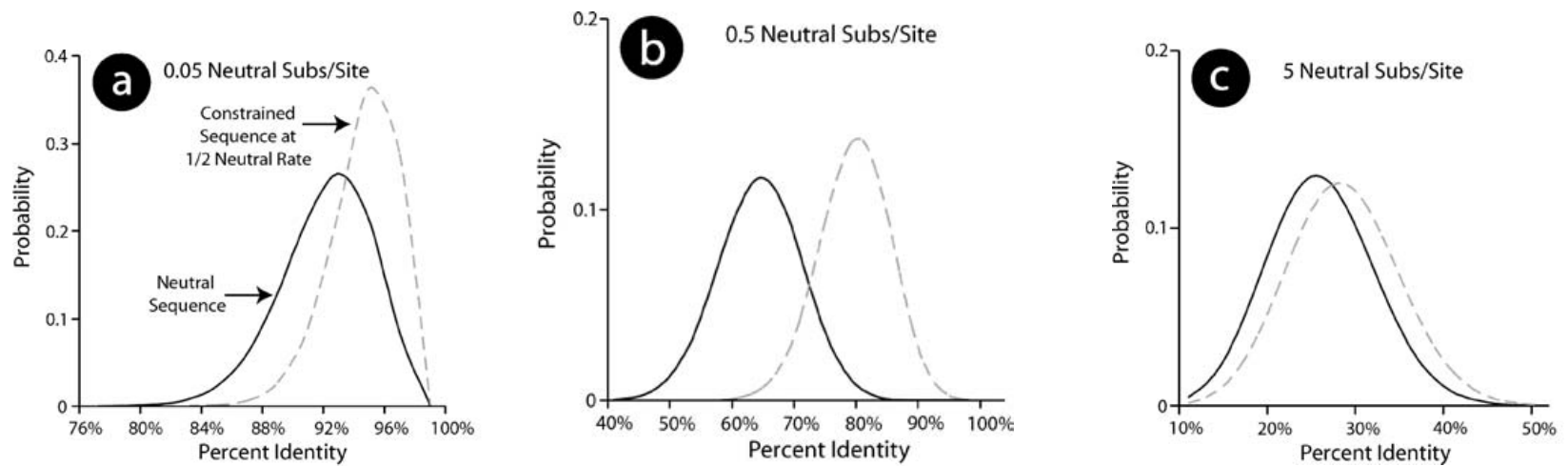
hypothèse 0 = substitutions par $e^{-\mu Q t}$
hypothèse 1 = substitutions par $e^{-\rho \mu Q t}$. Vraisemblances L_0 , L_1 et

$$\text{LLR} = \log \frac{L_1}{L_0}$$

En haut : LLR dans un fenêtre de 50 pb glissant dans l'alignement multiple de 5 primates, avec des exons indiqués. En bas : % identité humain-souris.

Puissance statistique

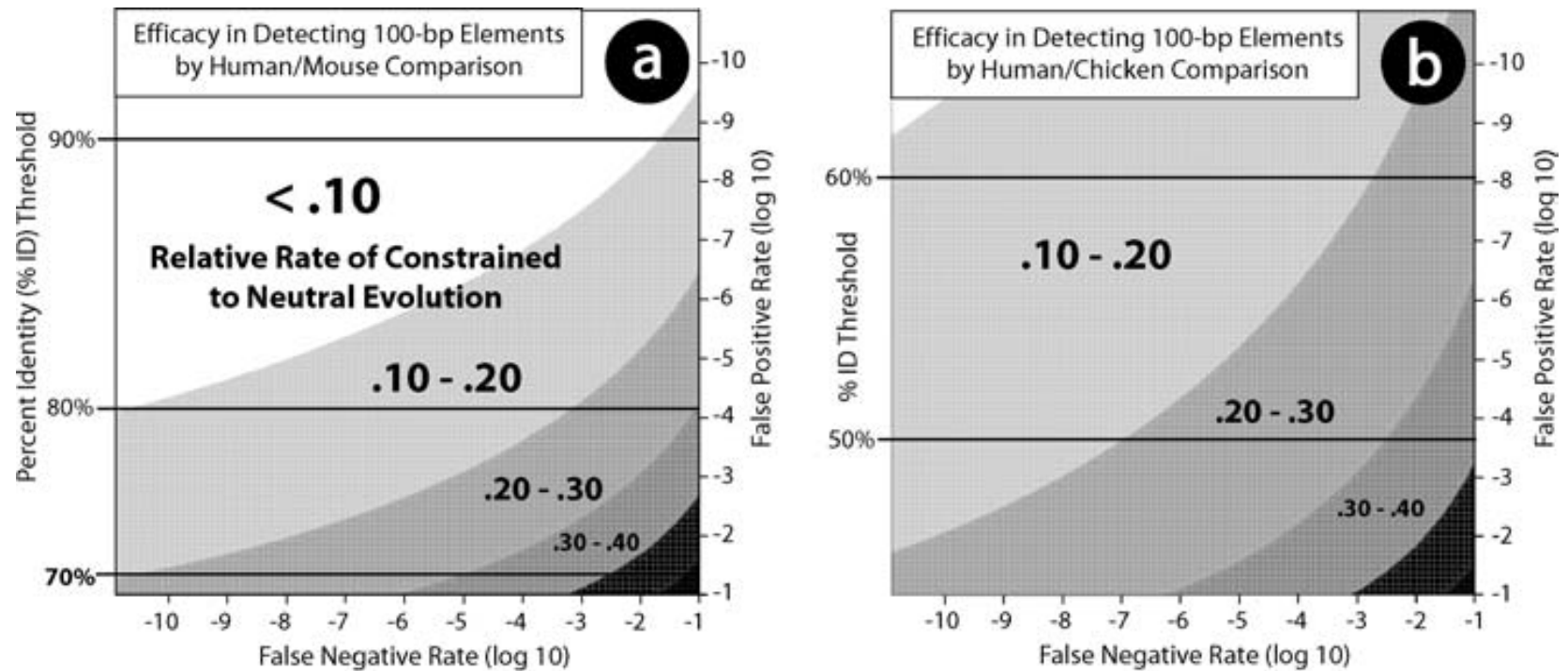
Détection dans une fenêtre de 50 pb selon % identité



Stone, Cooper & Sidow. *Annu. Rev. Genomics Hum Genet.* 6 :143, (2005).

Puissance statistique II

Détection dans une fenêtre de 100 pb selon % identité



Stone, Cooper & Sidow. *Annu Rev Genomics Hum Genet* 6 :143, 2005).

Et avec plus de génomes ?

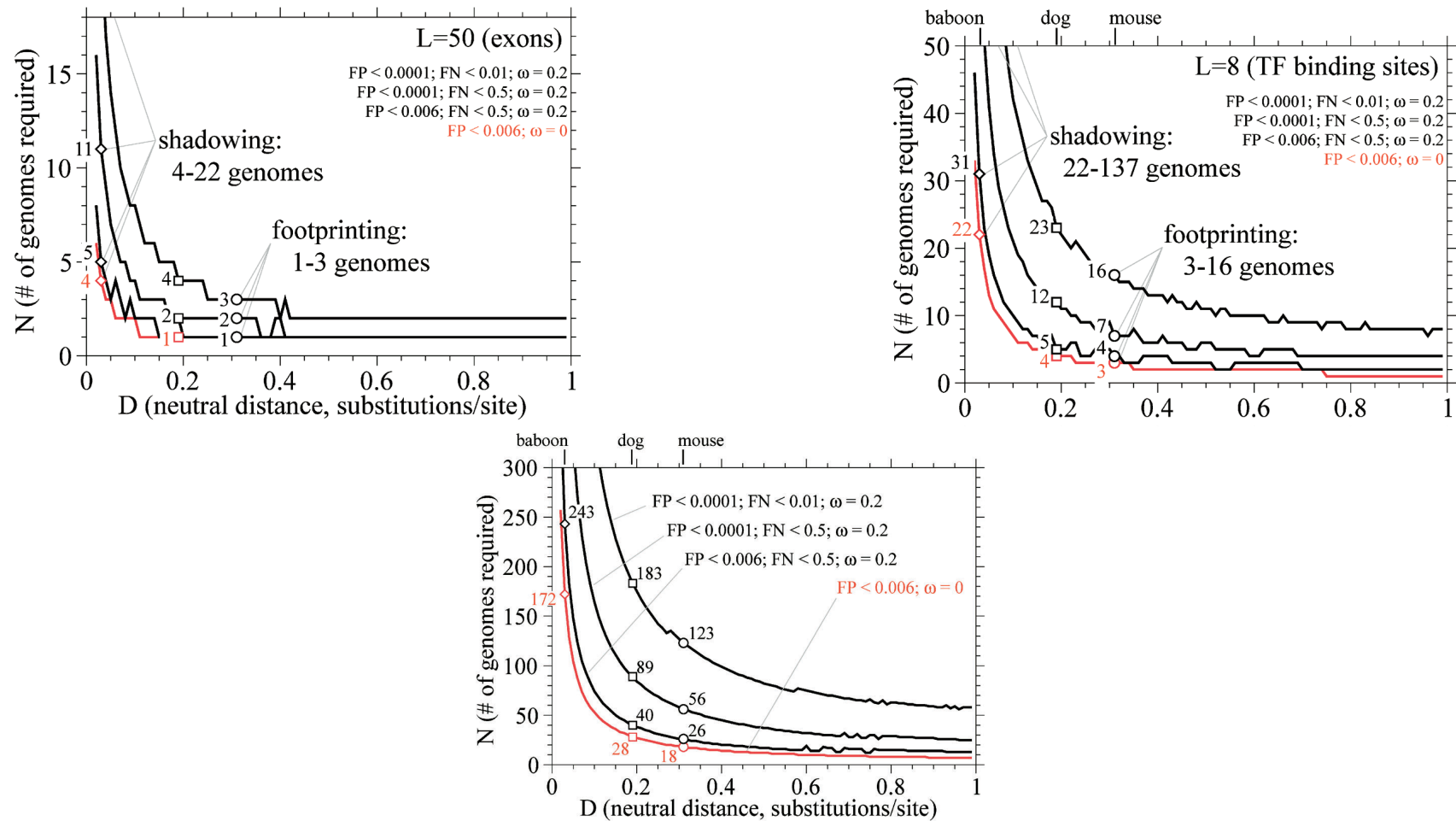
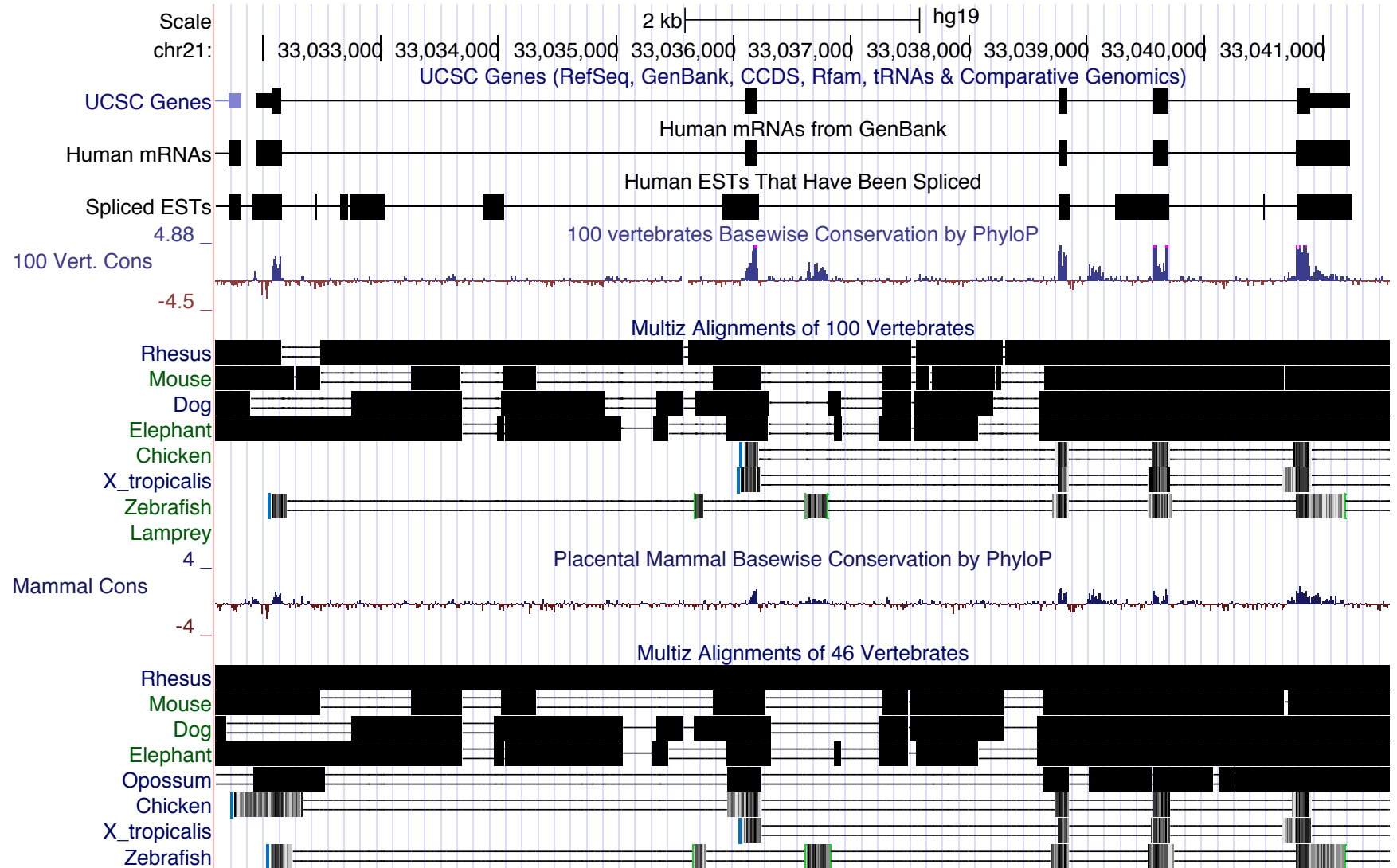


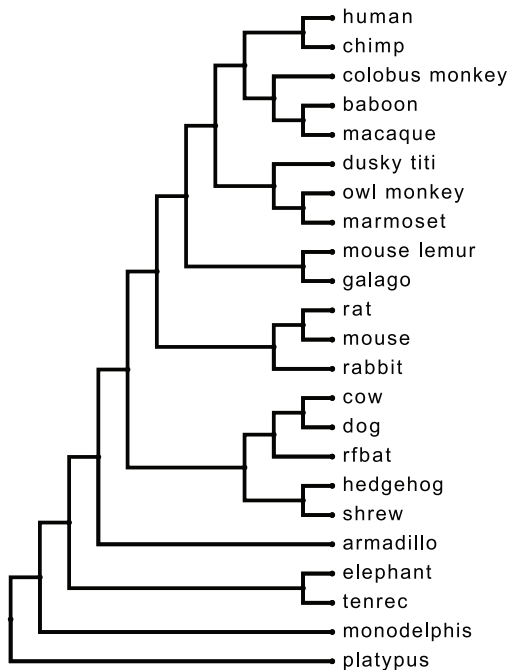
Figure 1. Number of Genomes Required for Single Nucleotide Resolution

On a bcp de génomes



Inférer la conservation

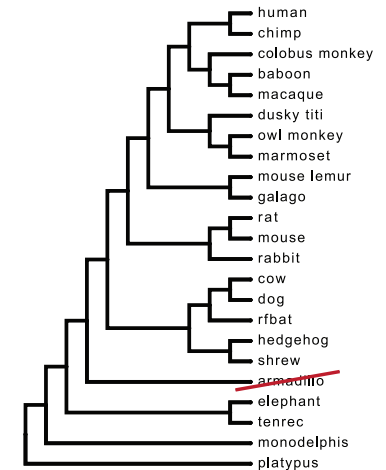
Phylogenetic Tree



Multiple Sequence Alignment

human	AATACGG	A	ACTTCATTCATT
chimp	AATATGG	A	ACTTCATTCATT
colobus monkey	AGTATGG	A	ACTTCATTCATT
baboon	AGTATGG	A	ACTTCATTCATT
macaque	AGTATGG	A	ACTTCATTCATT
dusky titi	AGTATGG	A	ACTTCATTCATT
owl monkey	AGTATGG	A	ACTTCATTCATT
marmoset	AGTATGG	A	ACTTCATTCATT
mouse lemur	AGTACGG	A	ACTTCATTCATT
galago	AGTACGG	A	ACTTCATTCATT
rat	AGTATGG	A	ACATCGTTCATT
mouse	AGTATGG	A	ACATCTTTCATT
rabbit	AGTATGG	A	ACATCATTCATT
cow	AGTATGG	A	ACATCATTCATT
dog	AGTACGG	A	ACATCATTCATT
rfbat	AGTATGG	A	ACATCGTTCATT
hedgehog	AGTATGG	A	ACATCATTCATT
shrew	AGTATGG	G	ACATCCTTCATT
armadillo	-----	-	-----
elephant	AGTATGG	A	ACATCGTTCATT
tenrec	AGTATGG	A	ACATCGTTCATT
monodelphis	AGTATGG	G	ACATCTTTCATT
platypus	AGTATGG	A	ACGTCATTCATT

ML Tree Scaling Factor $r = 0.7$

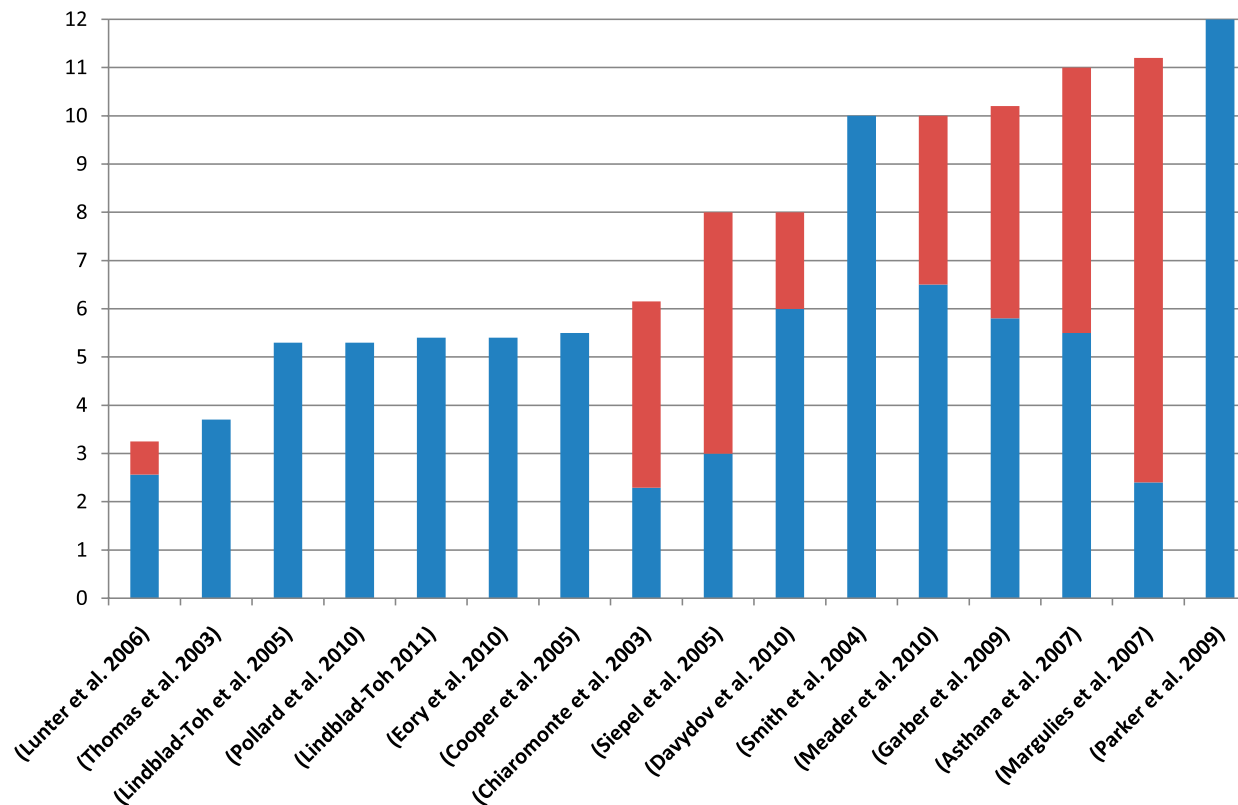


Questions : (1) classes de taux fixées ou non ; (2) estimation du taux neutre ; (3) groupage de positions consécutives ; (4) statistique pour détecter la conservation ; (5) alignement / trous / séquences manquantes

Davydov & al. *PLoS Comput Biol* 6 :e1001025 (2010)

Contraintes fonctionnelles dans le génome humaine

Selon la méthode d'inférence et les données, on arrive à des conclusions différentes sur la fraction α de régions conservées. . .

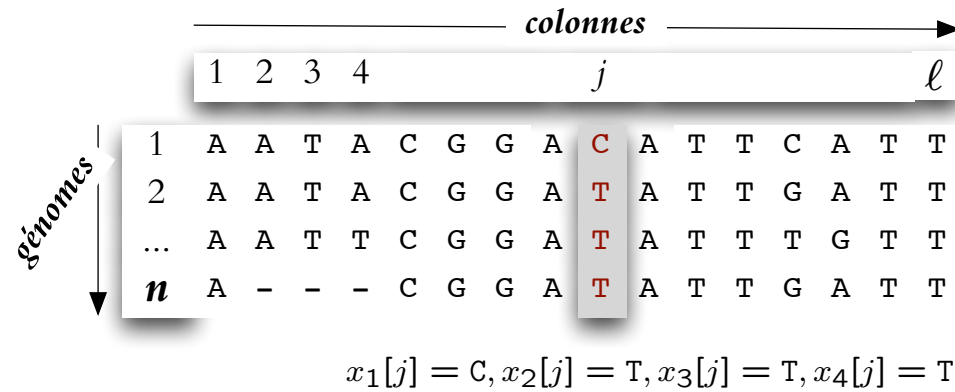


Estimates of α_{sel} from 16 studies ranked by increasing values. Lower and upper bound values are indicated in blue and red, respectively.

Ponting & Hardison *Genome Res* 21 :1769 (2011)

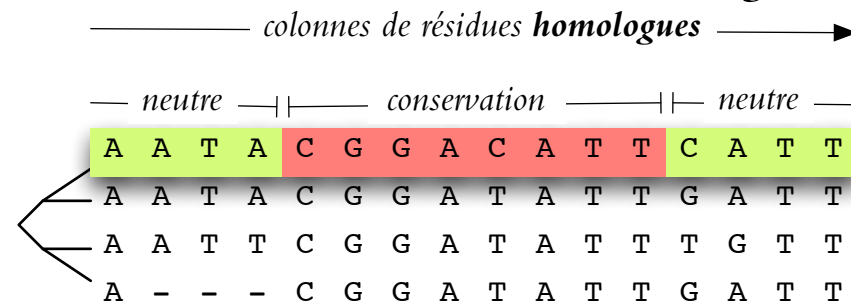
La méthode phylogénétique

On a un alignement multiple...



colonne = nucléotides homologues

⇒ inférer la variation de contraintes évolutives dans le génome



Substitutions sur la phylogénie

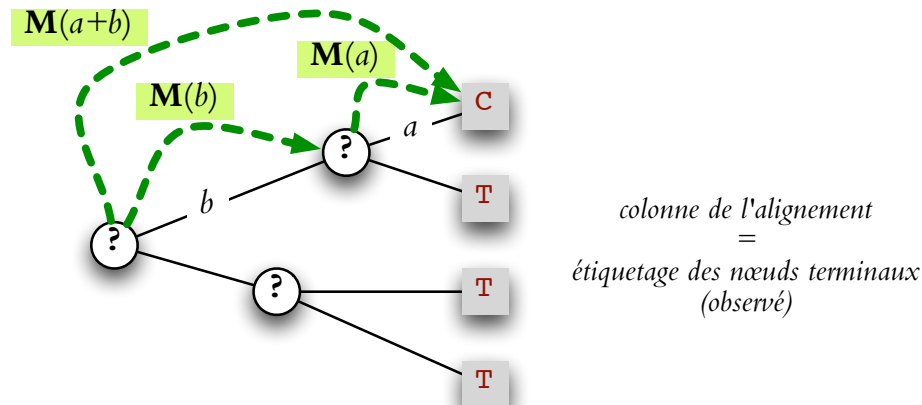
on veut capturer la variation du taux de substitution

⇒ employer un modèle phylogénétique d'évolution

procéssus de Markov à temps continu avec matrice de taux $\mathbf{Q}$:

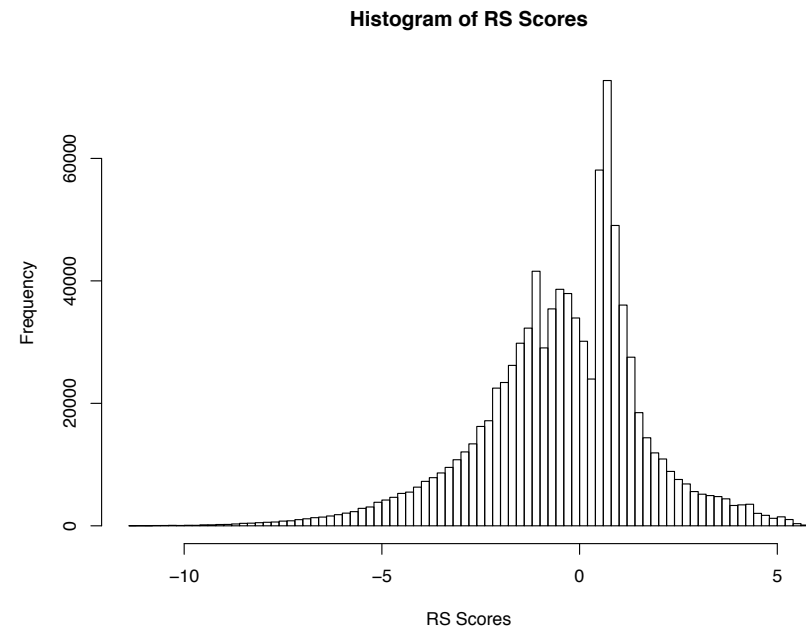
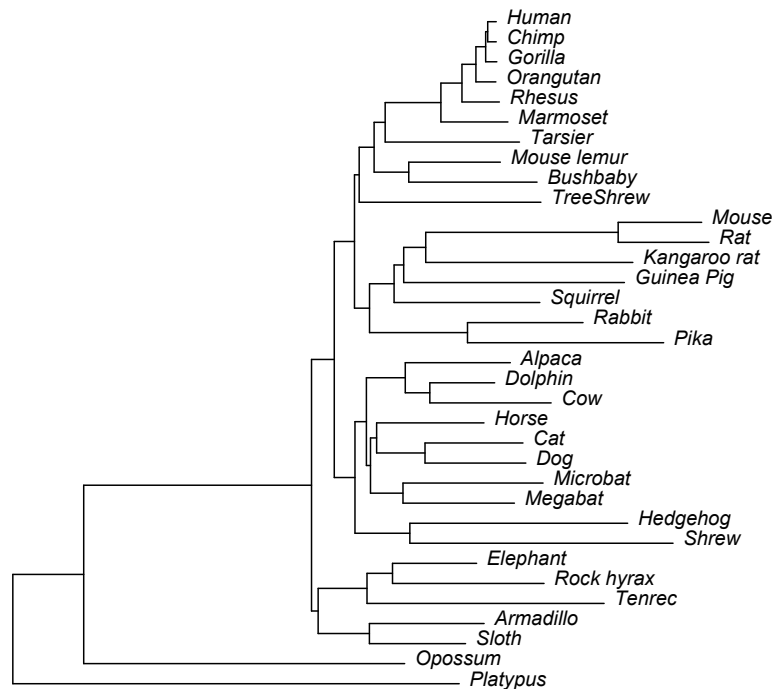
$$\mathbf{M}(t) = \begin{bmatrix} p_{A \rightarrow A}(t) & p_{A \rightarrow C}(t) & p_{A \rightarrow C}(t) & p_{A \rightarrow G}(t) \\ p_{C \rightarrow A}(t) & p_{C \rightarrow C}(t) & p_{C \rightarrow C}(t) & p_{C \rightarrow G}(t) \\ p_{G \rightarrow A}(t) & p_{G \rightarrow C}(t) & p_{G \rightarrow C}(t) & p_{G \rightarrow G}(t) \\ p_{T \rightarrow A}(t) & p_{T \rightarrow C}(t) & p_{T \rightarrow C}(t) & p_{T \rightarrow G}(t) \end{bmatrix} = \exp(\mathbf{Q}t)$$

$\mathbf{M}(t)$ décrit les probabilités de substitution sur une arête de longueur t



Taux variable — GERP++

- (1) arbre neutre : longueur vient de sites «dégénérés» (p.e., codon **GTN** = valine)
- (2) définir la vraisemblance $L(r)$ pour facteur d'échelle $r > 0$ — appliquer $M(rt)$ sur arête de longueur t
- (3) maximiser la fonction $L(r)$ pour choisir r ; score $RS = (1 - r) \sum_{uv \in \text{arêtes}} t_{uv}$

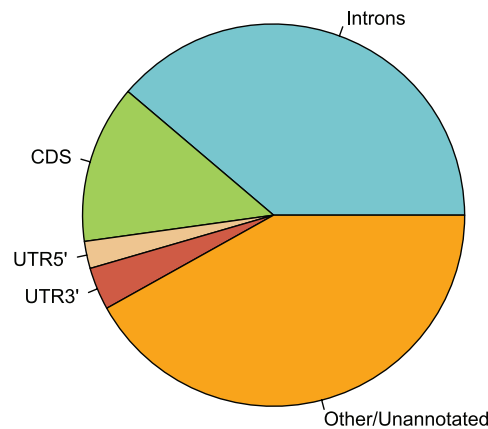


Davydov & al. *PLoS Comput Biol* 6 :e1001025 (2010)

Distribution d'éléments conservés

Beaucoup d'éléments dans les régions introniques et intergéniques : régulation, gènes ARN

B. Composition of Constrained Elements

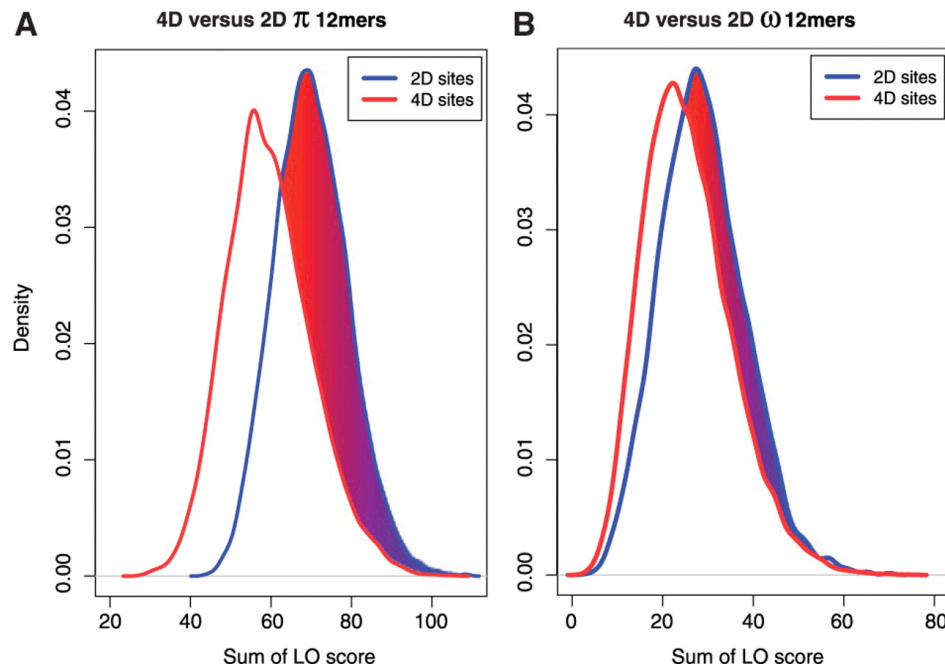


Davydov & al. *PLoS Comput Biol* 6 :e1001025 (2010)

Composition variable — SiPhy

Utiliser une matrice de taux, varier échelle ω (comme r avant) et distribution stationnaire π
score dans une fenêtre de longueur k

$$\text{LO} = \log \frac{\max_{\pi, \omega} \mathbb{P}\{x_1[j..j+k-1], x_2[j..j+k-1], \dots, x_n[j..j+k-1] \mid \pi, \omega\}}{\mathbb{P}\{x_1[j..j+k-1], x_2[j..j+k-1], \dots, x_n[j..j+k-1] \mid \pi_0, \omega_0\}}$$



meilleure séparation par composition que par vitesse
(4D = 3e position du codon dégénéré avec 4 choix; 2D = 3e position du codon dégénéré avec 2 choix)

Garber & al. *Bioinformatics* 25 :i64 (ISMB 2009)

Effet de voisinage — SCONE

On veut capturer les dépendances entre nucléotides consécutifs :

(1) considérer des triples de nucléotides consécutifs, calculer les taux neutres pour $AAA \rightarrow AAA$, $AAA \rightarrow AAT$, ...

(2) dans colonne j : reconstruire les caractères ancestraux par parcimonie dans colonnes $j - 1$ et $j + 1$, utiliser probabilité de substitutions

$$p_{x[j] \rightarrow y[j]} = \frac{\mathbb{P}\{x[j-1..j+1] \rightarrow y[j-1..j+1]\}}{\sum_{a,b} \mathbb{P}\{x[j-1..j+1] \rightarrow ay[j]b\}}$$

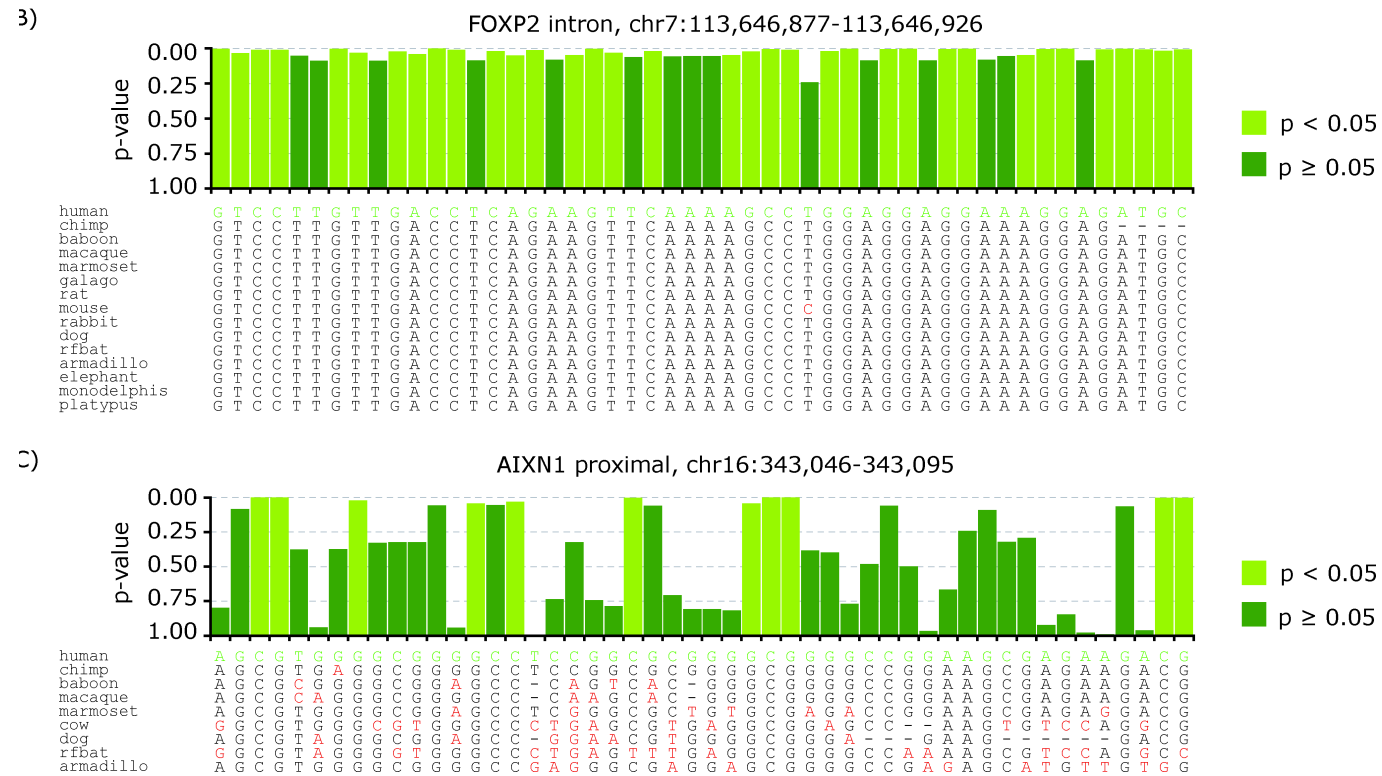
score = espérance postérieure avec m catégories d'échelle ω_k (pondération par vraisemblance $L(\omega_k)$)

$$\mathbb{E}\omega = \frac{\sum_{k=1}^m \omega_k \cdot L(\omega_k)}{\sum_{k=1}^m L(\omega_k)}$$

P-value : vérifier distribution de $\mathbb{E}\omega$ dans évolution simulée (10000 fois) sur l'arbre neutre

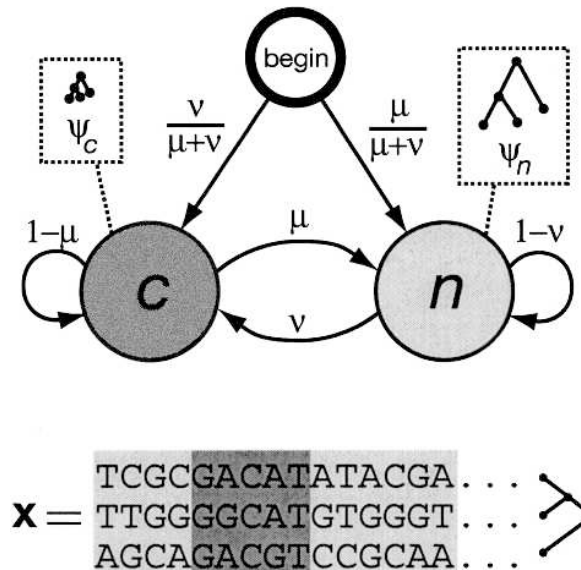
SCONE

(intron et une région intergénique)



Effet de voisinage — PhyloHMM

phastCons

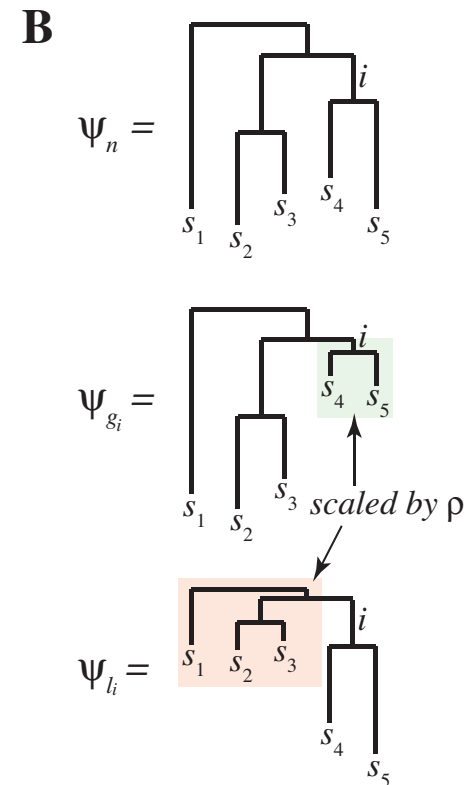
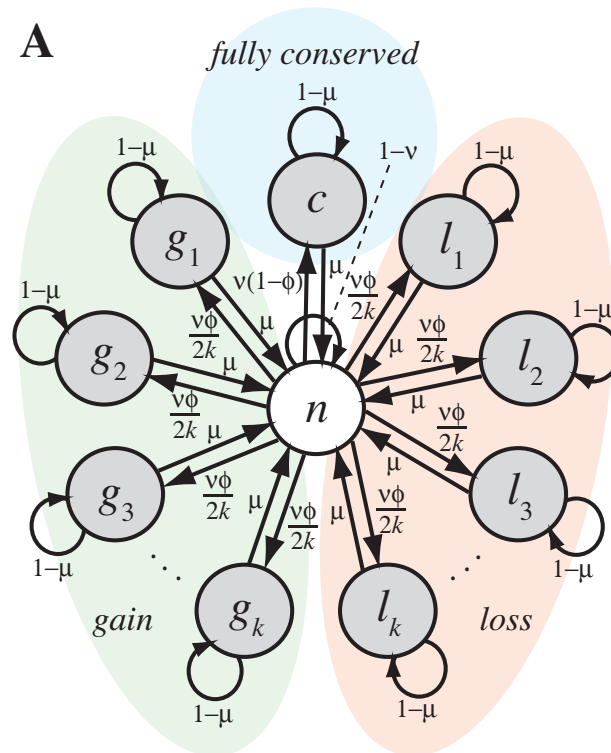


émissions : colonnes de l'alignement multiple, avec probabilités de transition e^{Qt} (neutre) ou $e^{\rho Qt}$ (sélection négative avec $\rho < 0$)

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

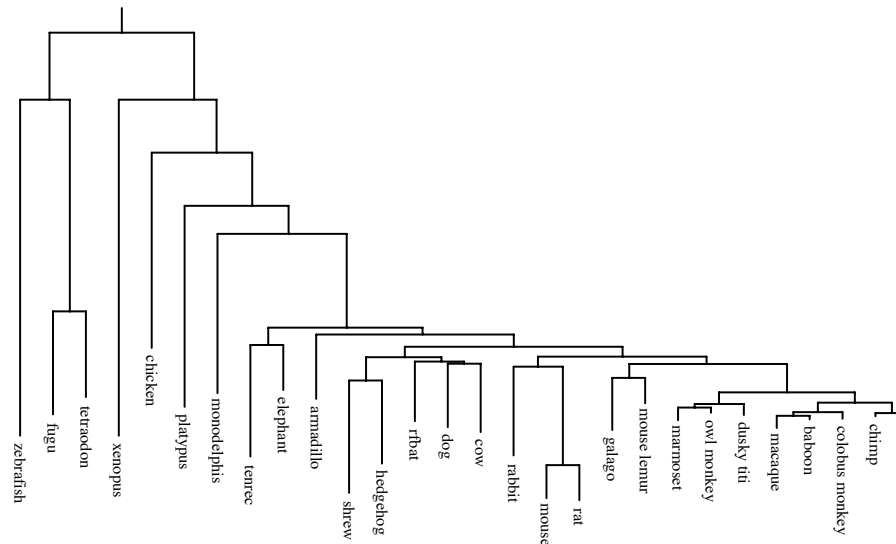
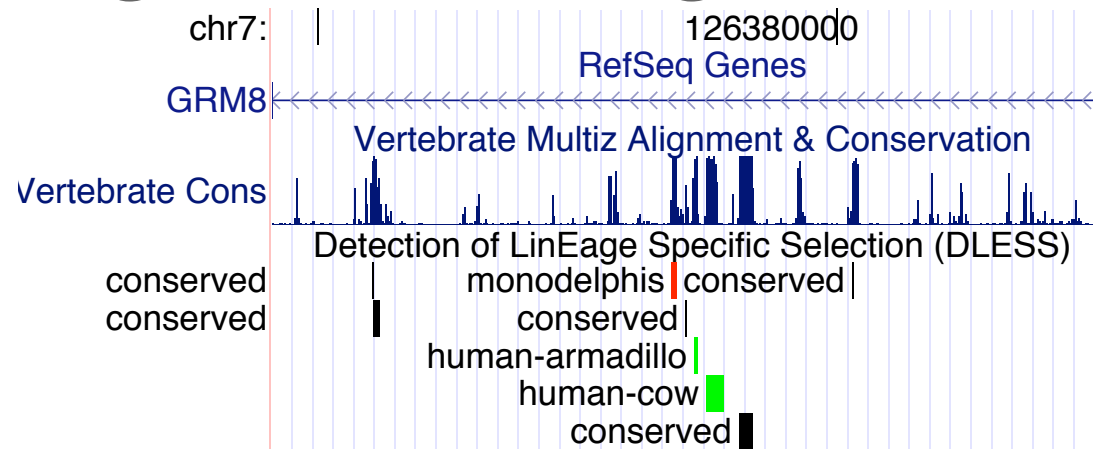
Turnover

Modélisation par phylo-HMM



Siepel & al. RECOMB 2006

DLESS ∈ goldenPath/hg17



Siepel & al. *RECOMB* 2006

Accélération — HAR

Human Accelerated Regions : plus rapide dans l'humain mais conservé ailleurs

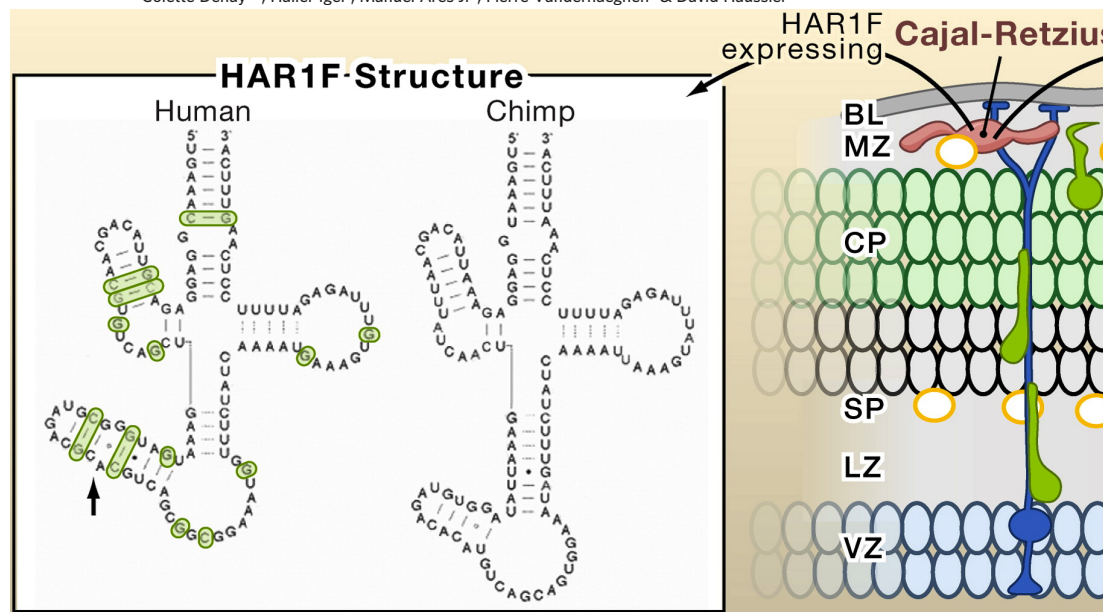
Vol 443|14 September 2006|doi:10.1038/nature05113

nature

ARTICLES

An RNA gene expressed during cortical development evolved rapidly in humans

Katherine S. Pollard^{1,*,†}, Sofie R. Salama^{1,2,*}, Nelle Lambert^{4,5}, Marie-Alexandra Lambot⁴, Sandra Coppens⁴, Jakob S. Pedersen¹, Sol Katzman¹, Bryan King^{1,2}, Courtney Onodera¹, Adam Siepel^{1,†}, Andrew D. Kern¹, Colette Dehay^{6,7}, Haller Igel³, Manuel Ares Jr³, Pierre Vanderhaeghen⁴ & David Haussler^{1,2}



Amadio & Walsh *Cell* 126 :1033 (2006)

Pas trop grande différence avec assez de génomes ...

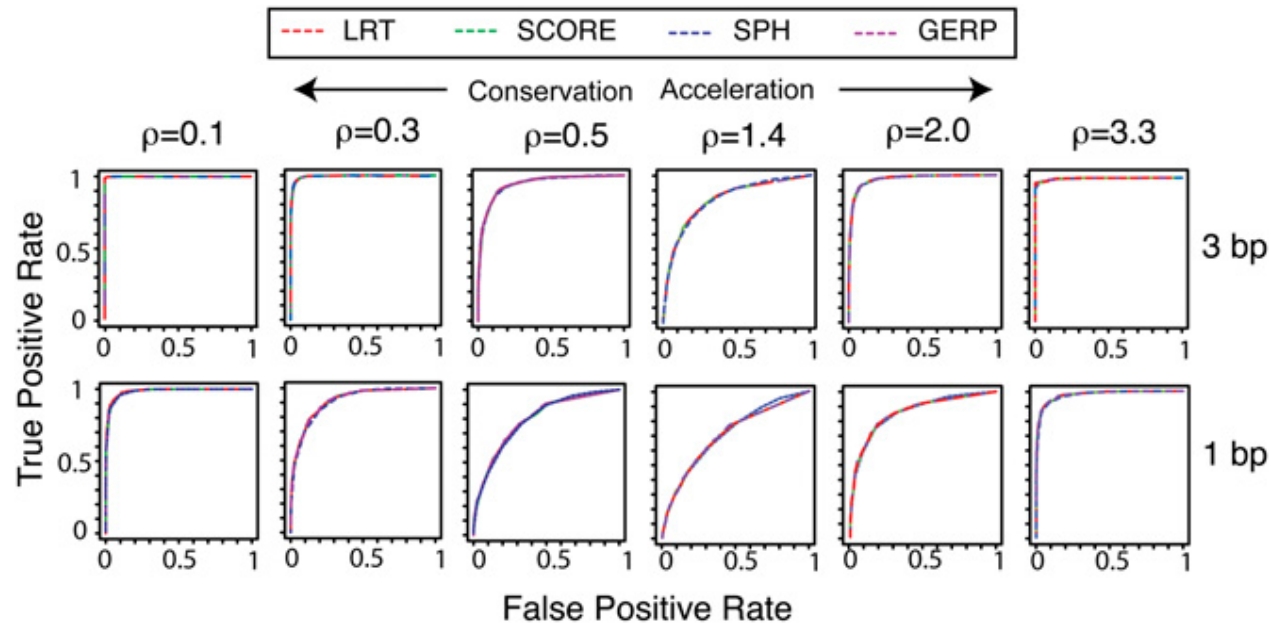


Figure 1. Receiver operating characteristic (ROC) curves showing false-positive versus true-positive rates for the all-branch tests implemented in phyloP: (red) LRT, (green) SCORE, (blue) SPH, and (purple) GERP. Individual plots show results for simulated data sets with either 3-bp (*top*) or 1-bp (*bottom*) elements generated from models with a range of deviations ρ from the neutral rate $\rho = 1.0$ (columns).

Accord qualitatif sur sélection

Table 1. Publications describing estimates of α_{sel} and α_{sel}^0

Publication	Method (α_{sel} and α_{sel}^0 estimation)	α_{sel} (%) lower	α_{sel} (%) higher	Substitutions or indels or topography	Neutral model/standard	Whole or partial genome	Multiple, or pair of, genomes	Local or global neutral rate
Lunter et al. (2006)	NIM (α_{sel})	2.56	3.25	Indels	Randomly placed indels	Whole	Pair	Local
Thomas et al. (2003)	MCSs (α_{sel})	3.7	3.7	Substitutions	4D sites	Partial	Multiple	Local
The ENCODE Project Consortium (2007) (ENCODE)	Two of three methods (α_{sel})	4.9	4.9	Substitutions	4D sites/most aligned sites	Partial (ENCODE)	Multiple	Global
Lindblad-Toh et al. (2005)	Substitutions (α_{sel})	5.3	5.3	Substitutions	ARs	Whole	Pair	Local
Pollard et al. (2010)	Various (α_{sel})	5.3	5.3	Substitutions	Various	Partial (ENCODE)	Multiple	Global
Lindblad-Toh et al. (2011)	SiPhy (α_{sel})	5.4	5.4	Patterns	ARs	Whole	Multiple	Global
Eory et al. (2010)	Substitutions (α_{sel})	5.4	5.4	Substitutions	ARs	Whole	Pair	Local
Cooper et al. (2005)	GERP (α_{sel})	5.5	5.5	Substitutions	Most aligned sites	Partial	Multiple	Local
Chiaromonte et al. (2003)	Substitutions (α_{sel})	2.29	6.15	Substitutions	ARs	Whole	Pair	Local
Siepel et al. (2005)	phastCons (α_{sel})	3	8	Substitutions	none	Whole	Multiple	Global
Davydov et al. (2010)	GERP++ (α_{sel})	6	8	Substitutions	Most aligned sites	Whole	Multiple	Global
Smith et al. (2004)	Substitutions/Turnover (α_{sel} and α_{sel}^0)	10	10	Substitutions	Simulation	Partial	Multiple pairs	Local
Meader et al. (2010)	NIM (α_{sel} and α_{sel}^0)	6.5	10	Indels	Randomly placed indels	Whole	Multiple pairs	Local
Garber et al. (2009)	SiPhy (α_{sel})	5.8	10.2	Patterns	ARs	Partial (ENCODE)	Multiple	Global
Asthana et al. (2007)	SCONE (α_{sel})	5.5	11	Substitutions within trinucleotides	Most aligned sites	Partial (ENCODE)	Multiple	Global
Parker et al. (2009)	Topography (Chai; α_{sel})	12	12	Topography	Most aligned sites	Partial (ENCODE)	Multiple	Global

(4D) Four-fold degenerate; (ARs) ancestral repeats; (ENCODE) Encyclopedia of DNA Elements project; (GERP) Genomic Evolutionary Rate Profiling; (MCSs) multispecies conserved sequence; (NIM) neutral indel model; (SCONE) Sequence CONservation Evaluation.