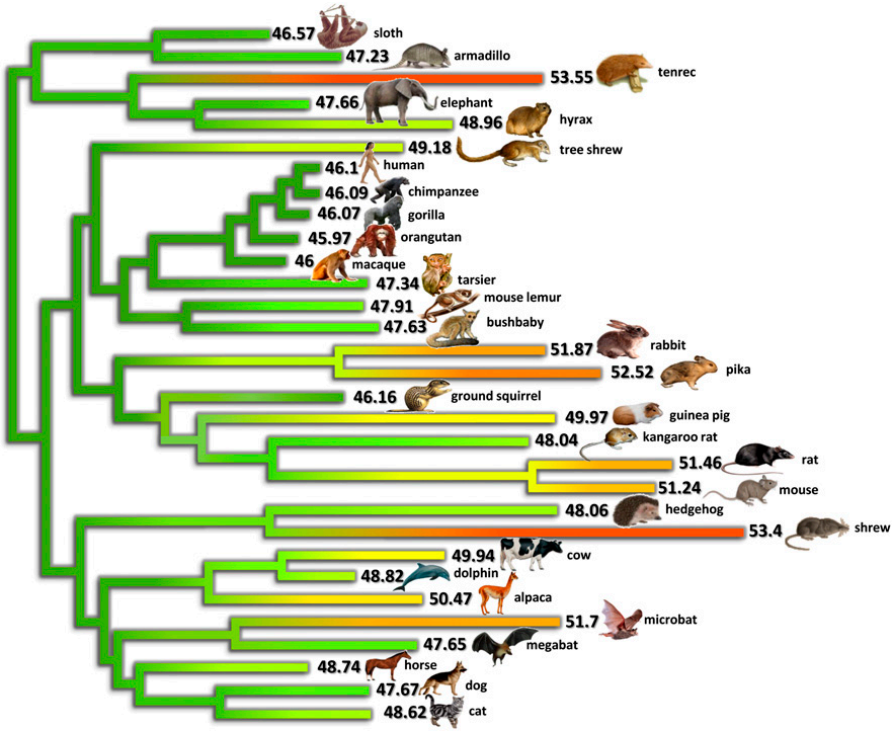
Integrative modeling

The comparative method in evolutionary genomics

Nicolas Lartillot

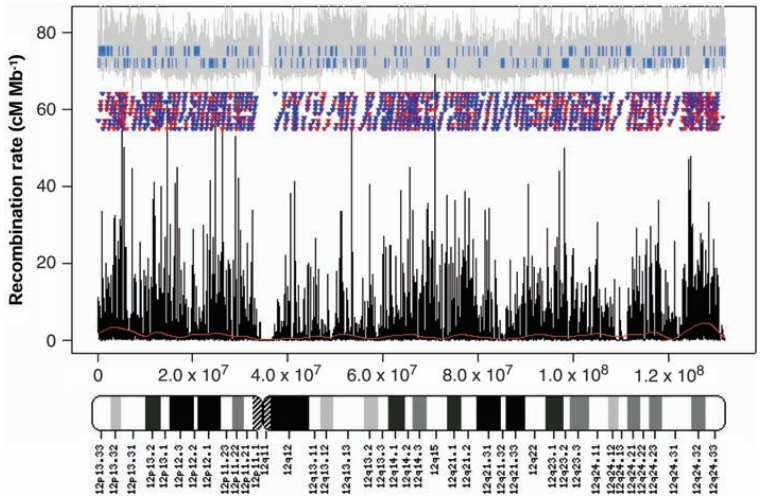
August 30, 2014

GC and recombination



variation among species (GC content)

Romiguier et al, 2010, Genome Res 20:1001



variation along genomes (recombination)

Myers et al, 2005, Science, 310:321

A codon model for dN/dS and GC^*

Nucleotide matrix R

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \frac{\gamma}{2} & \kappa\frac{\gamma}{2} & \frac{1-\gamma}{2} \\ C & \frac{1-\gamma}{2} & - & \frac{\gamma}{2} & \kappa\frac{1-\gamma}{2} \\ G & \kappa\frac{1-\gamma}{2} & \frac{\gamma}{2} & - & \frac{1-\gamma}{2} \\ T & \frac{1-\gamma}{2} & \kappa\frac{\gamma}{2} & \frac{\gamma}{2} & - \end{array}$$

Codon matrix Q (61 x 61)

$$\begin{aligned} Q_{\mathrm{AAA}\to\mathrm{AAC}} &= R_{\mathrm{A}\to\mathrm{C}} \\ Q_{\mathrm{AAA}\to\mathrm{AGA}} &= R_{\mathrm{A}\to\mathrm{G}}\,\omega \\ Q_{\mathrm{AAA}\to\mathrm{AGC}} &= 0 \\ &\dots \end{aligned}$$

- κ: transition-transversion ratio
- γ: equilibrium GC content (GC^*)
- ω: dN/dS

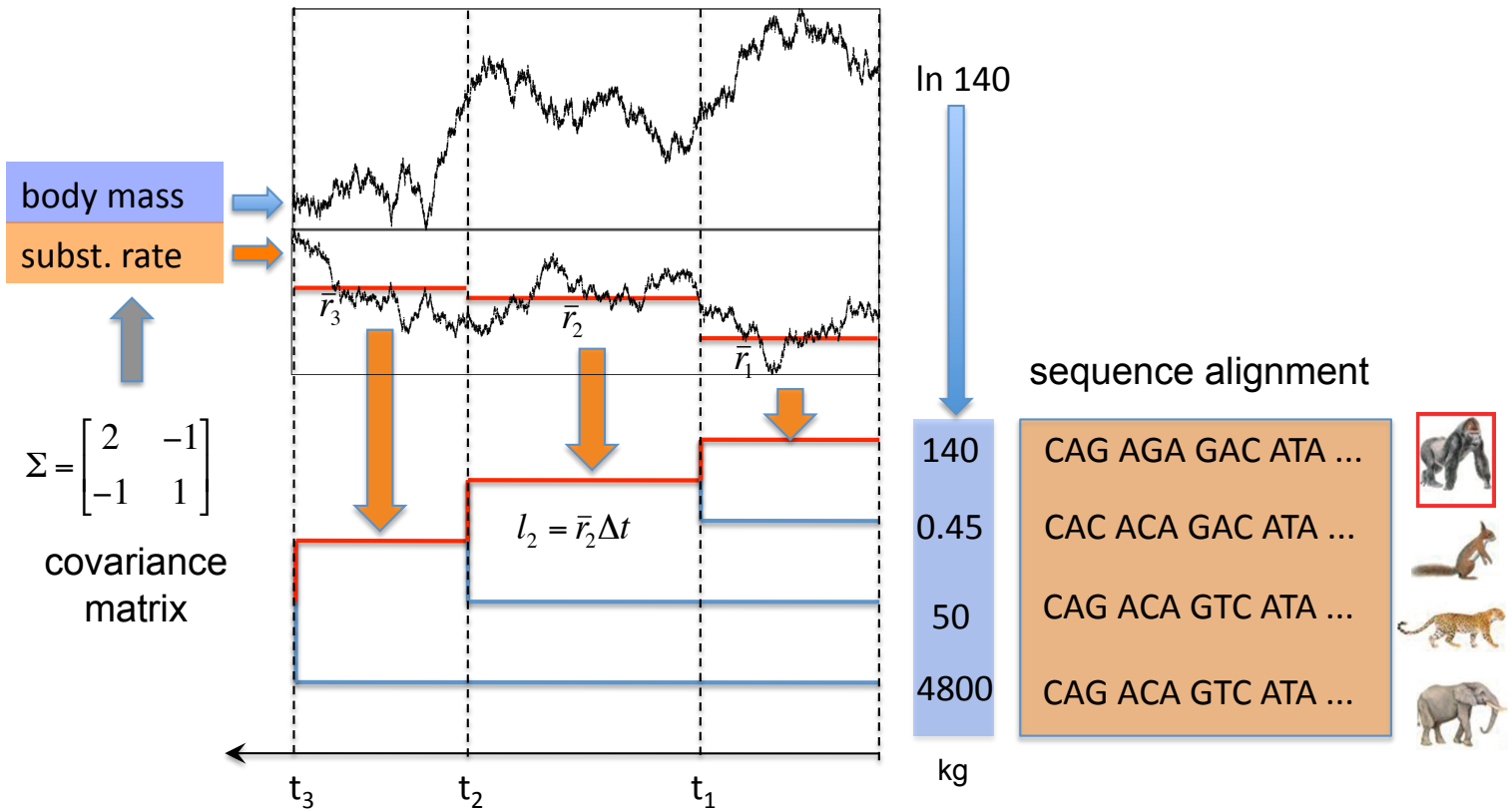
Phylogenetic covariance model

Multivariate Brownian process $X(t)$ (rates and traits):

$$
\begin{aligned}
X_1(t) &= \ln u(t), \\
X_2(t) &= \ln \omega(t), \\
X_3(t) &= \ln \frac{\gamma(t)}{1-\gamma(t)}, \\
X_{l+3}(t) &= \ln C_k(t) \quad k = 1, \ldots, K.
\end{aligned}
$$

- $u(t)$: syonymous substitution rate dS
- $\omega(t)$: dN/dS
- $\gamma(t)$: equilibrium GC content (GC^*)
- $C_k(t)$: kth. quantitative trait

The molecular comparative method



(Lartillot and Poujol, 2011, Molecular Biology and Evolution)

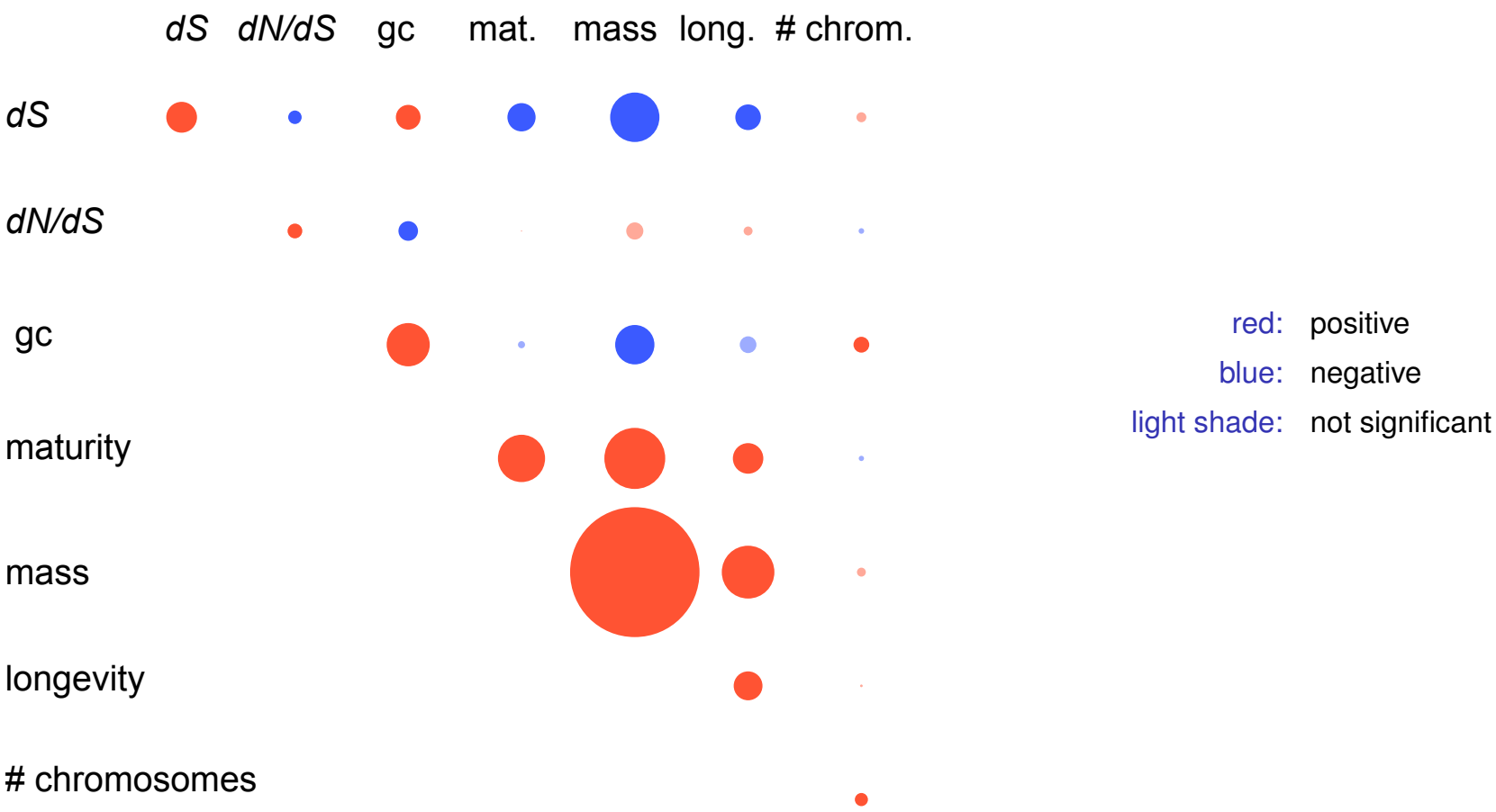
posterior proportional to joint probability:

$$p(\lambda, \mu, \rho)\, p(t \mid \lambda, \mu, \rho)\, p(\Sigma)\, p(X \mid t, \Sigma)\, p(D \mid X, t)$$

Data

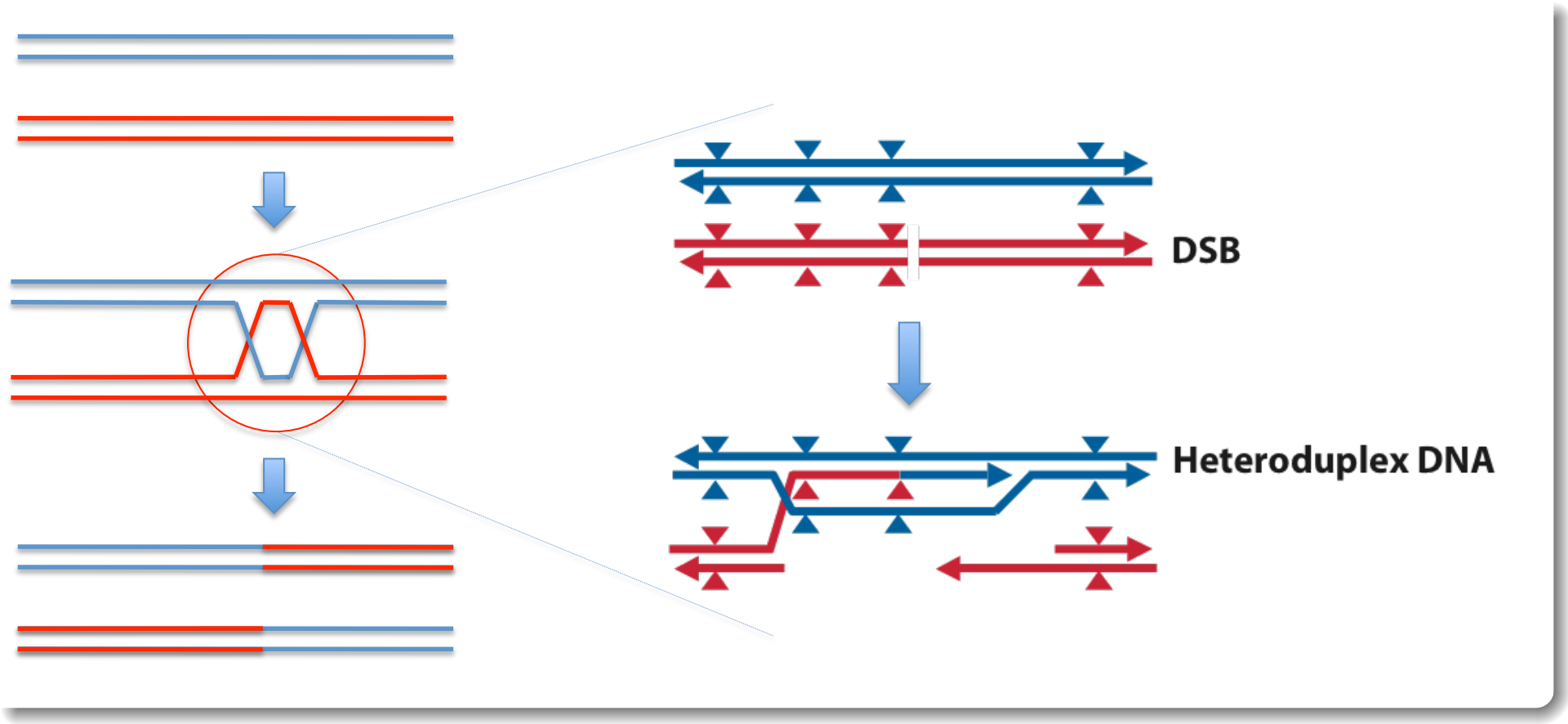
- 17 single-exon genes in 73 placental mammals
- adult body mass
- female age at sexual maturity (generation time)
- maximum recorded lifespan (longevity)
- number of chromosomes

Estimated covariance matrix

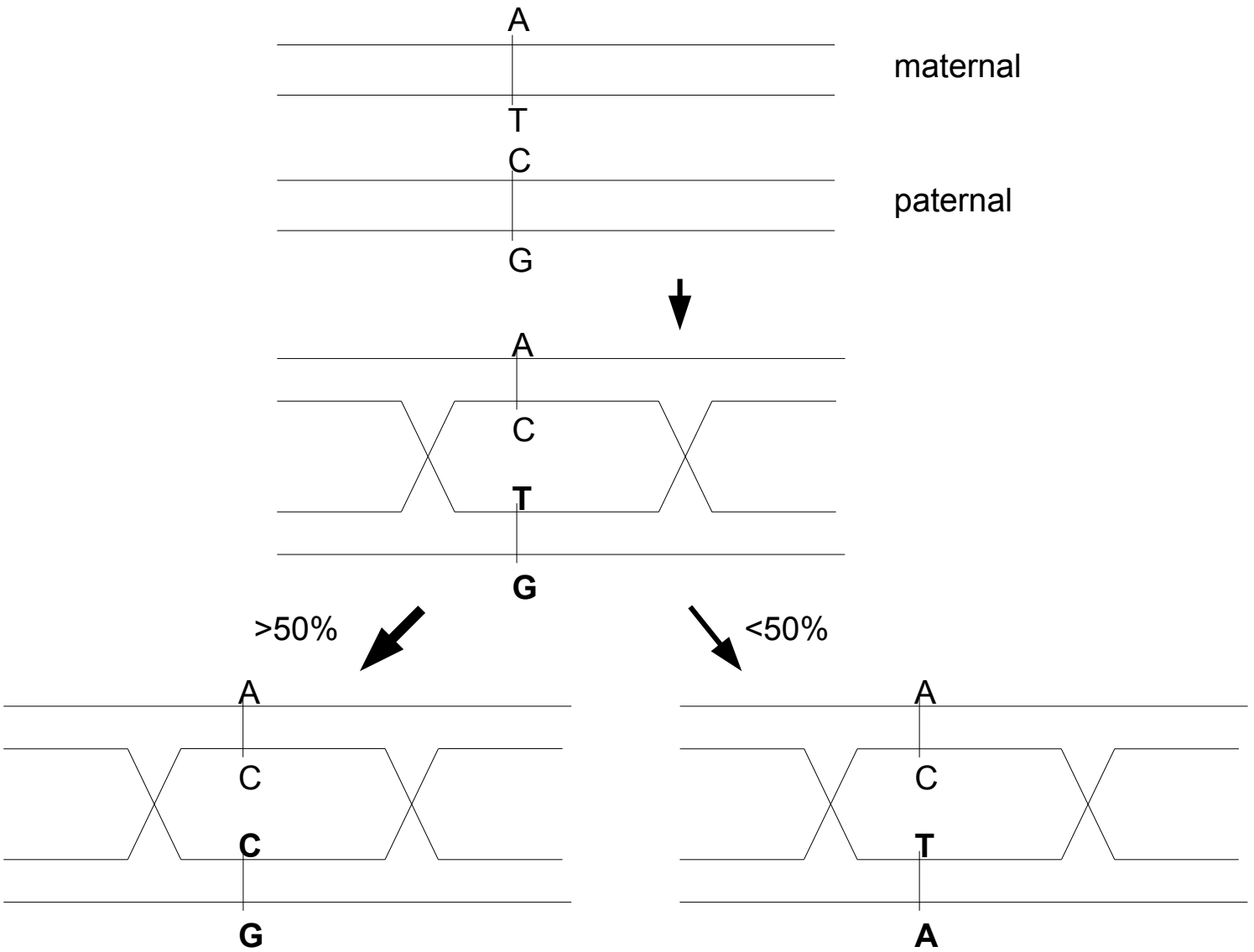


Lartillot, 2012, Mol Biol Evol 30:356

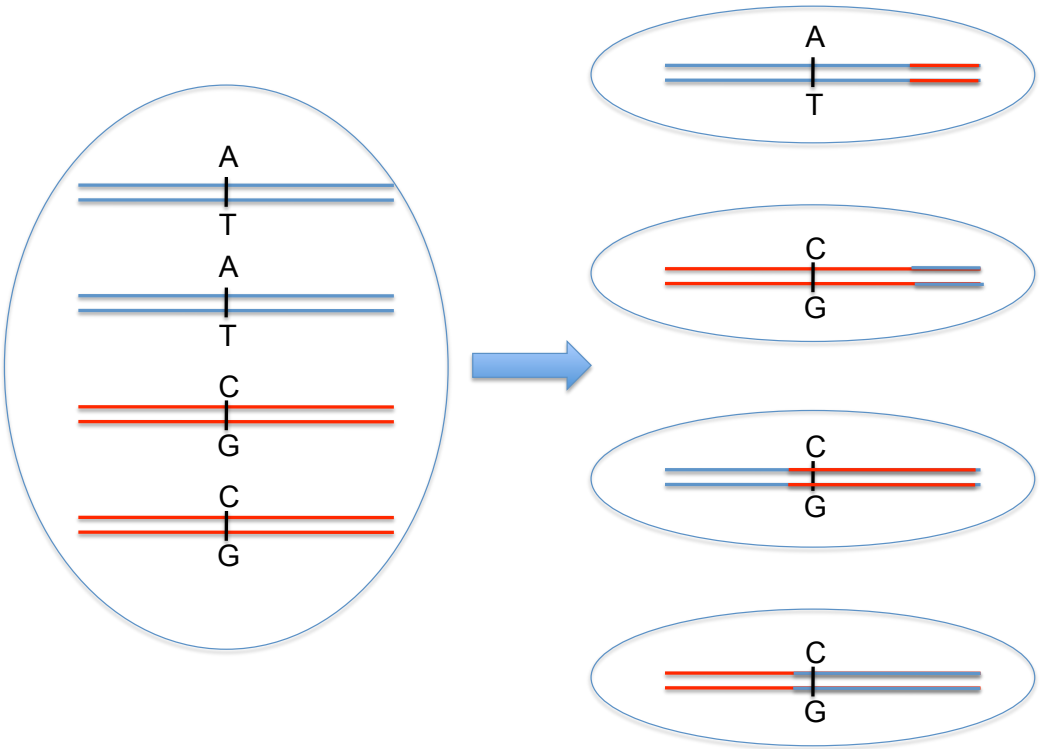
Biased conversion during meiosis



Biased conversion during meiosis



Biased gene conversion (BGC) during meiosis



$$x_{GC} = \frac{1+b}{2}$$

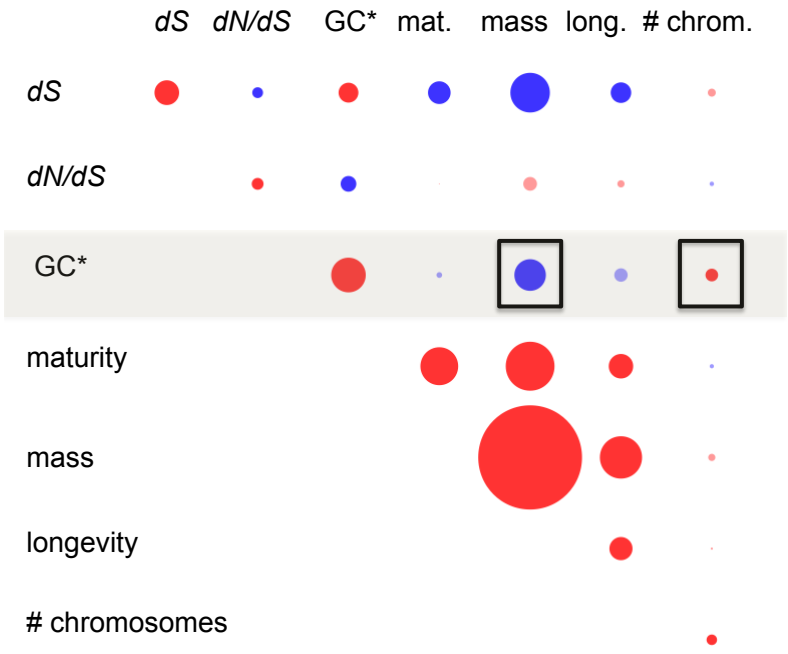
$$x_{AT} = \frac{1-b}{2}$$

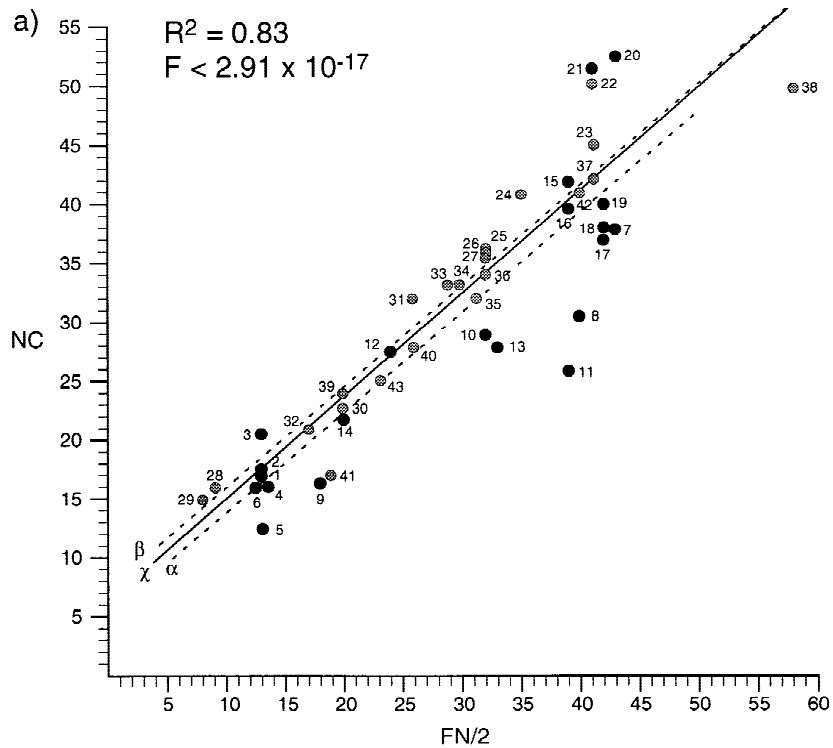
GC overtransmission

- meiotic distorsion bias $b \iff$ like positive selection for GC
- b proportional to local recombination rate ($b = b_0 r$)

Duret and Galtier 2009, Annu Rev Genomics Hum Genet 2009, 10:285

Biased gene conversion explains variation in *GC**



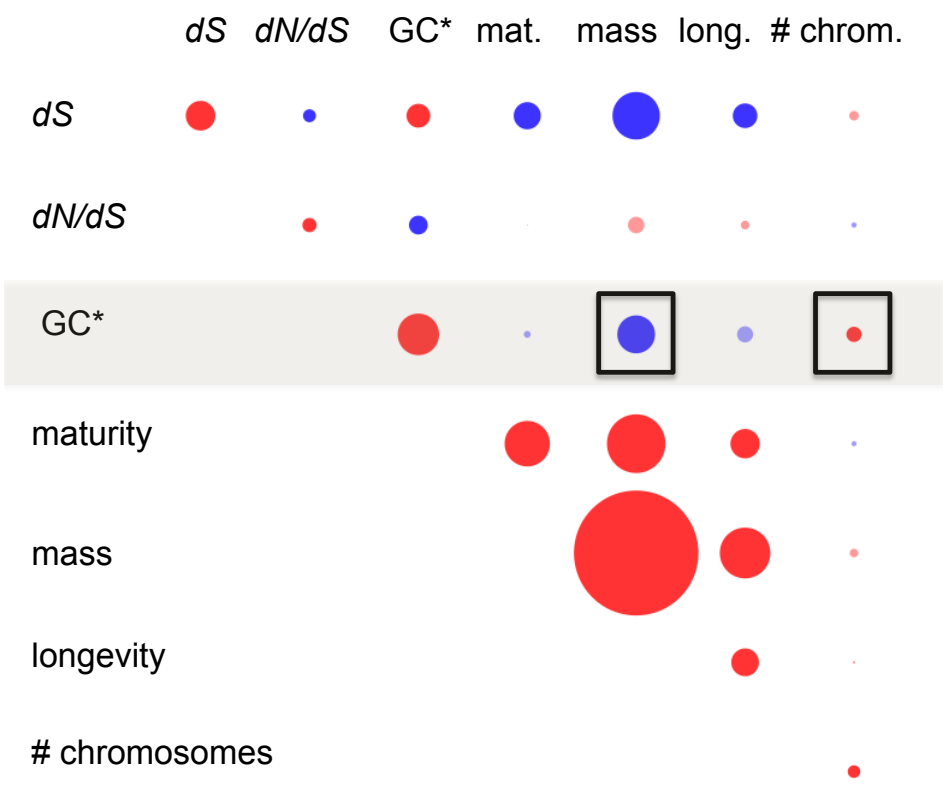


de Villena and Sapienza, 2001, Mamm Genome 12:318

Positive correlation *GC** / chromosome number

- $\sim$ 1 recombination event per chromosome arm per meiosis
- more fragmented karyotype = smaller chromosomes
  = higher recombination rate = stronger gene conversion

Biased gene conversion explains variation in *GC**



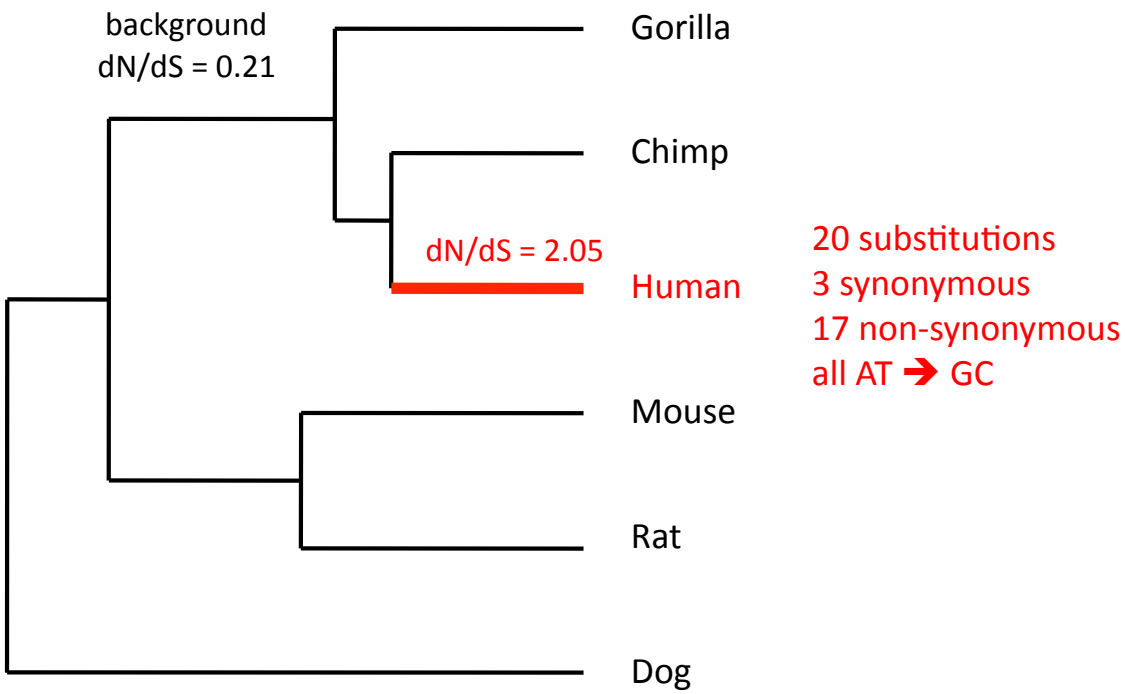
red: positive
blue: negative
light shade: not significant

Negative correlation *GC** / body mass

- larger animals = smaller population = less efficient selection
- also less efficient BGC (lower *GC**)

Romiguier and Galtier 2010, Genome Res 20:1001

Selection and GC-biased gene conversion

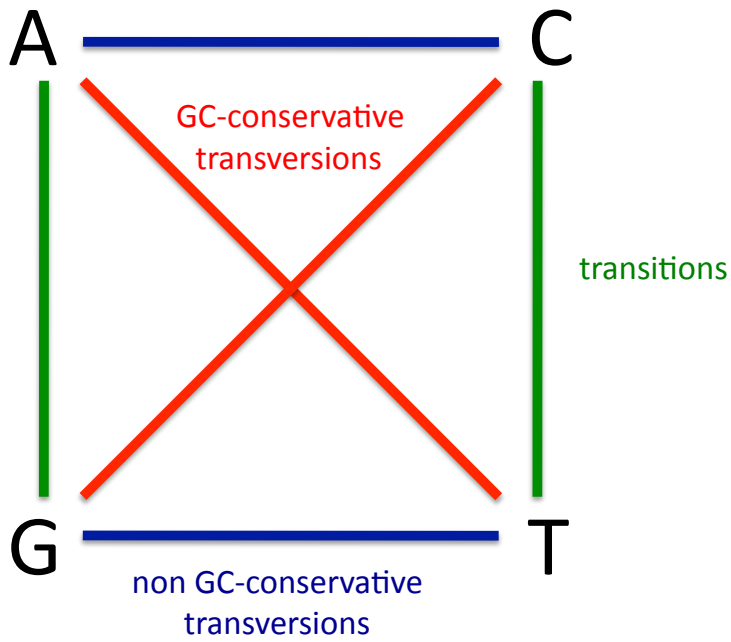


ADCYAP1 gene, Ratnakumar et al, 2010, Phil Trans R. Soc. B 365:2571

gBGC interferes with selection

- gBGC episodes may result in elevated dN/dS (Galtier et al, 2009)
- could explain lack of correlation between dN/dS and body mass

Estimating selection independently of gBGC



Method

- separately estimate dS and dN/dS on
  - GC-conservative transversions
  - non GC-conservative transversions
  - transitions
- GC-conservative dN/dS should be immune from gBGC

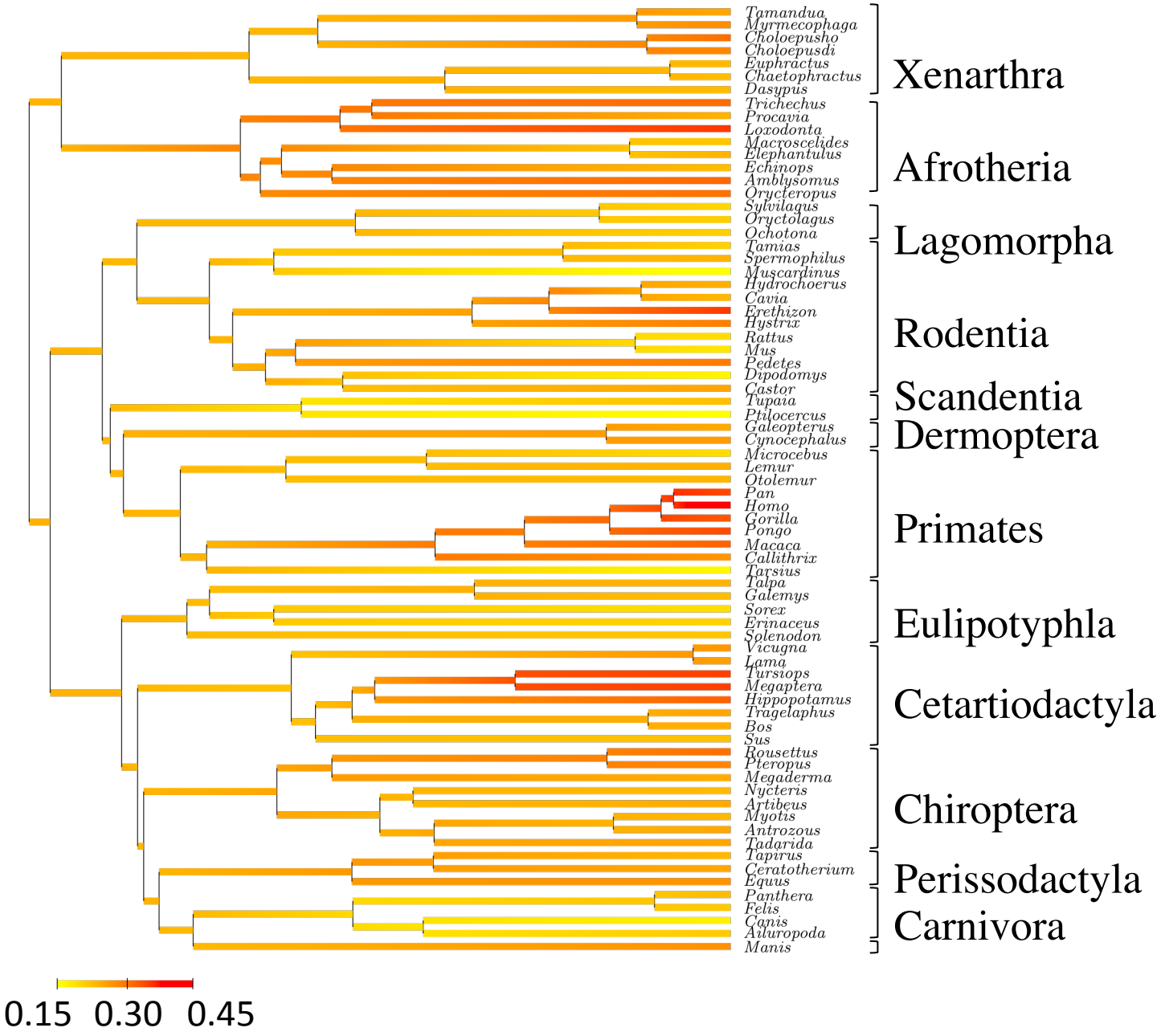
dN/dS: correlation(r^2) with life-history traits

	maturity	mass	longevity
regular dN/dS (all substitutions)	0.27^{*}	0.13	0.27
transitions	0.26	0.15	0.24
non GC-conservative transversions	0.14	0.22	0.26
GC-conservative transversions	0.46^{**}	0.45^{**}	0.59^{**}

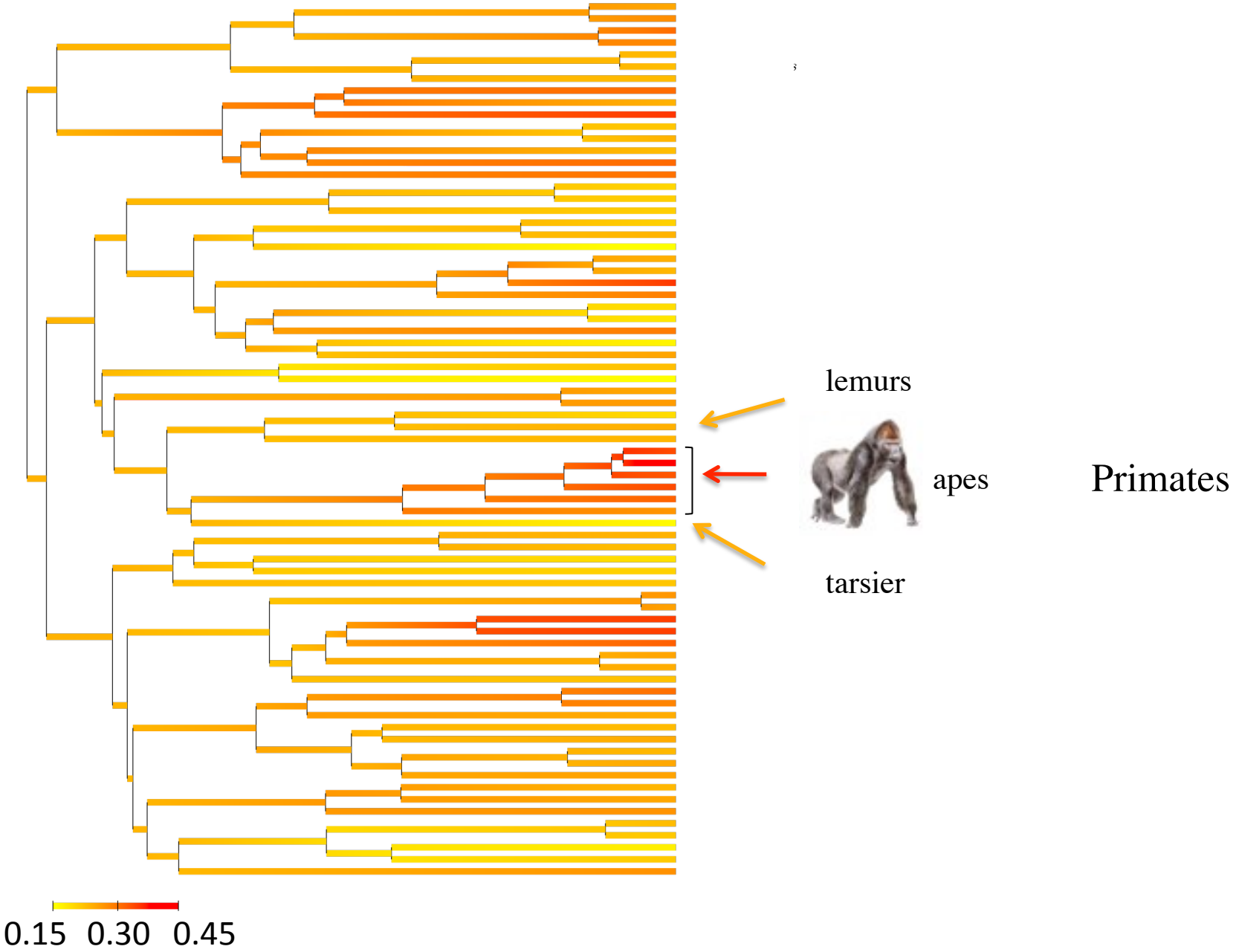
(posterior probability of a positive correlation $> 0.95^{*}$ or $> 0.975^{**}$)

Lartillot, 2012, Mol Biol Evol, 30:356

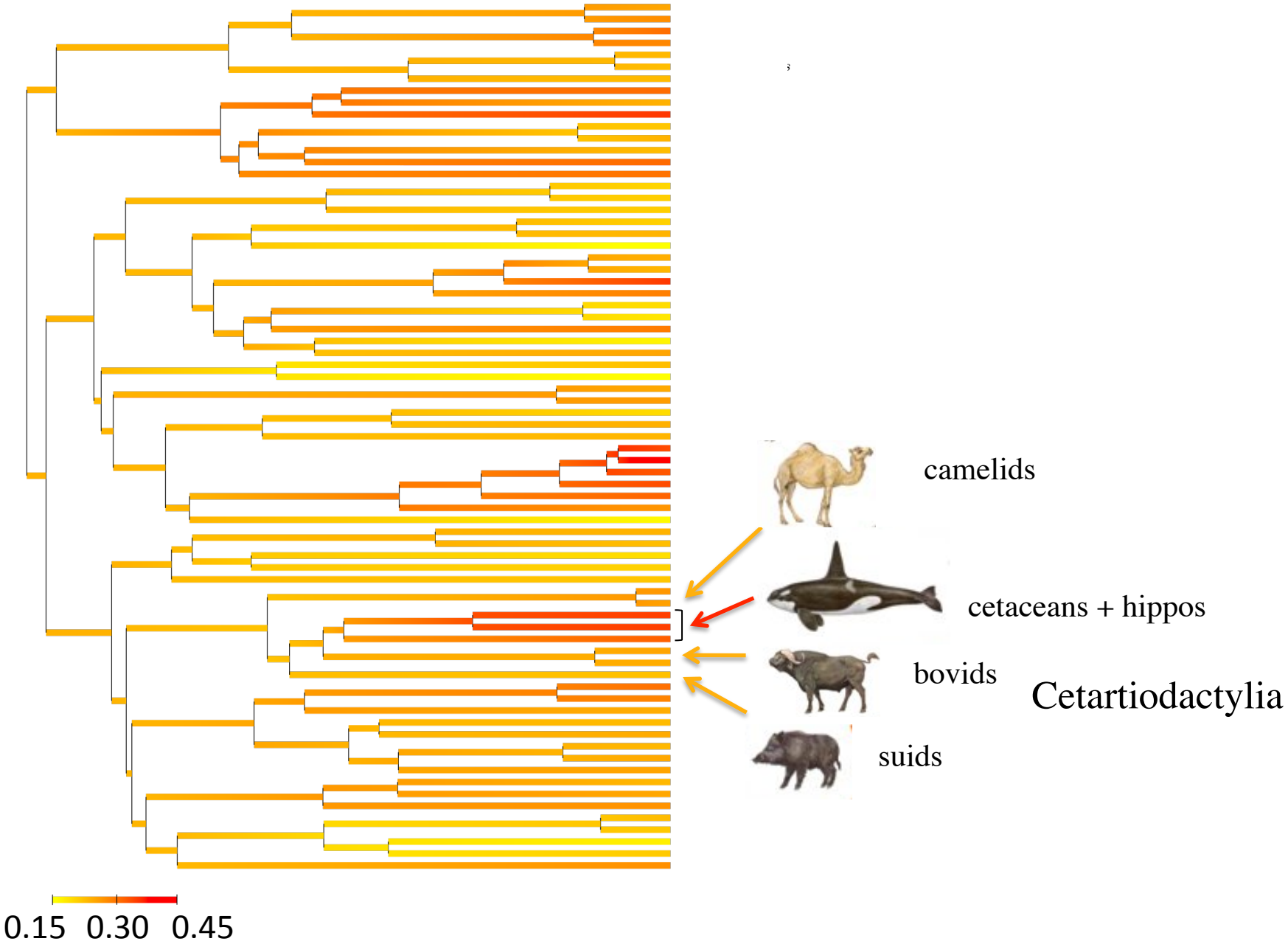
GC-conservative *dN*/*dS*



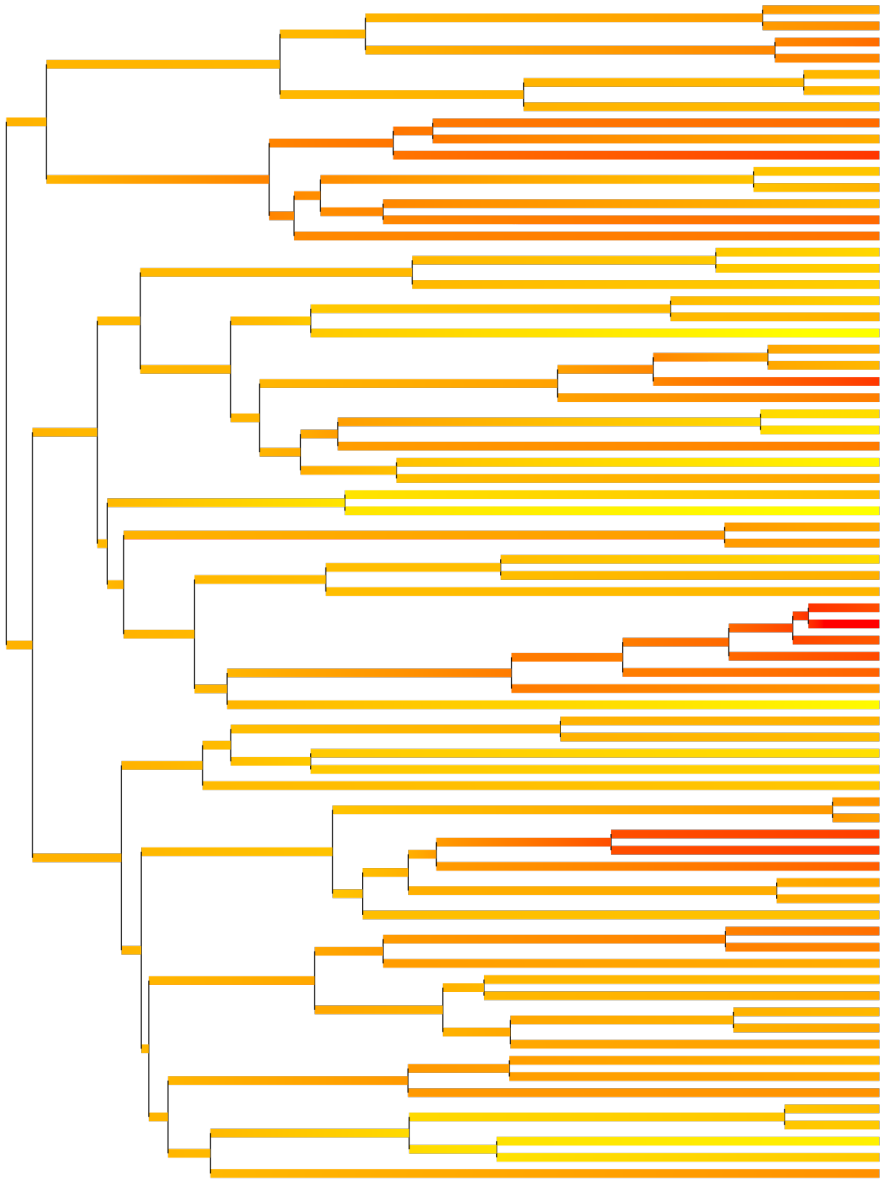
GC-conservative *dN*/*dS*

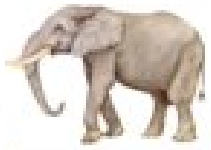


GC-conservative *dN*/*dS*



GC-conservative *dN*/*dS*





elephant

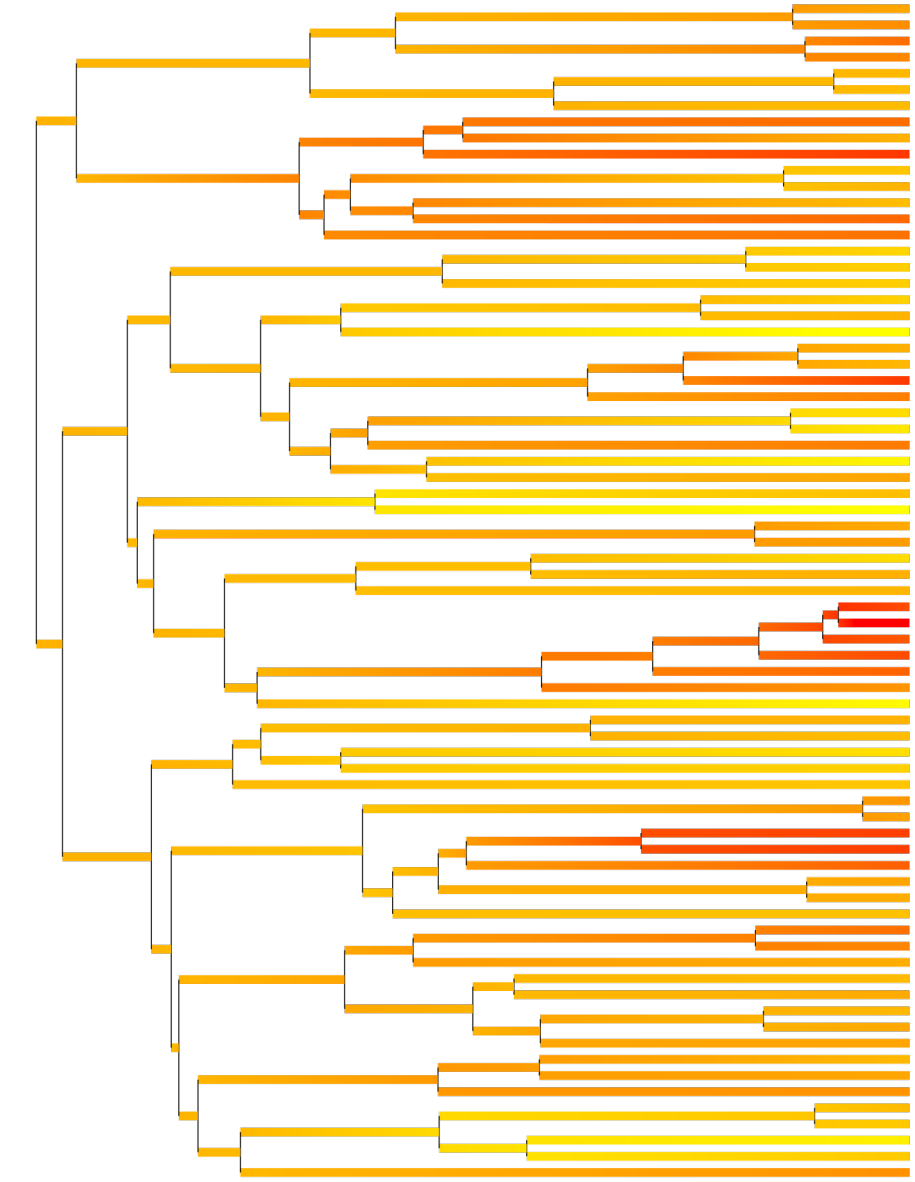
elephant shrews

aardvark

Afrotheria

0.15 0.30 0.45

GC-conservative *dN*/*dS*



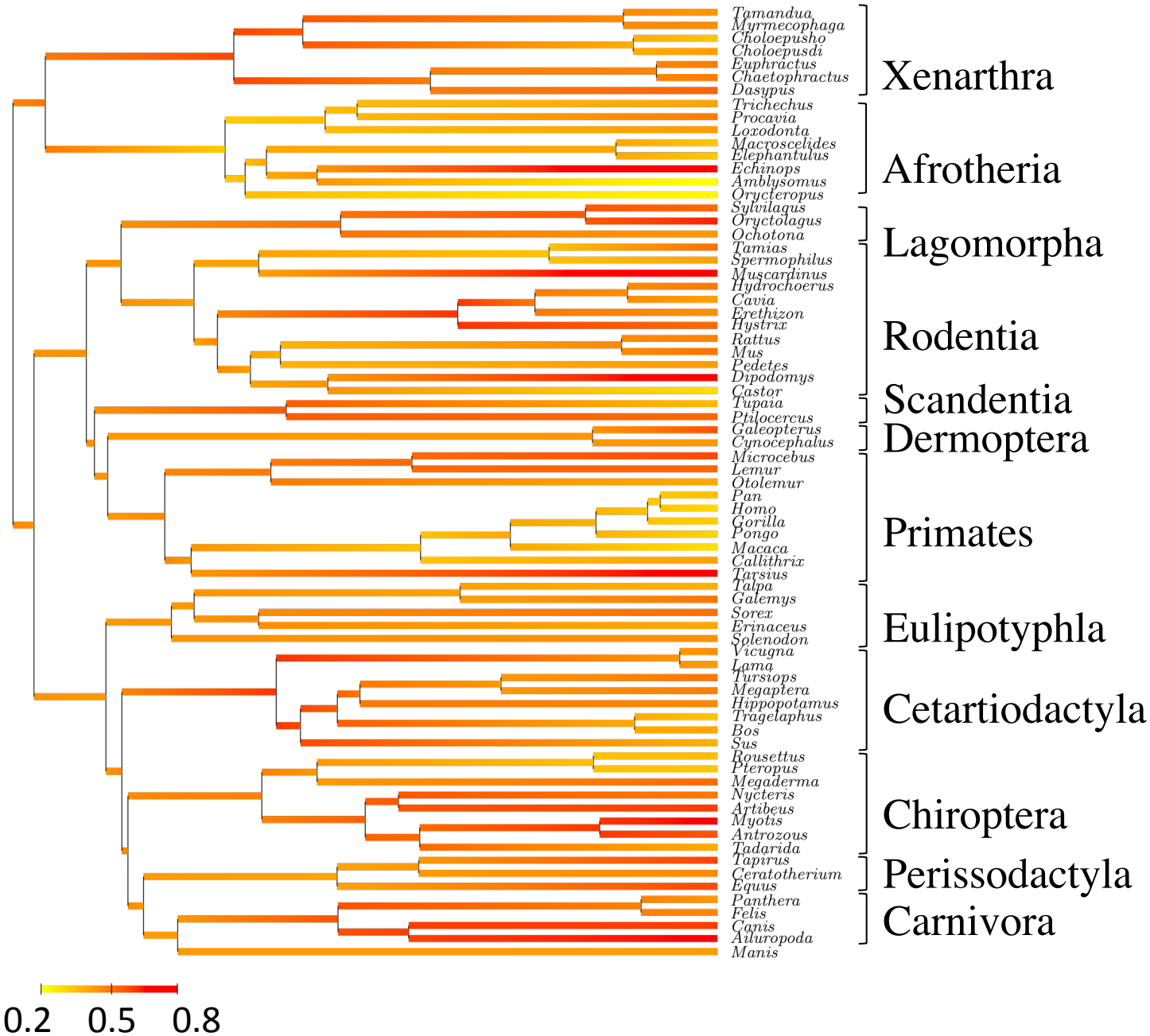
megabats

microbats

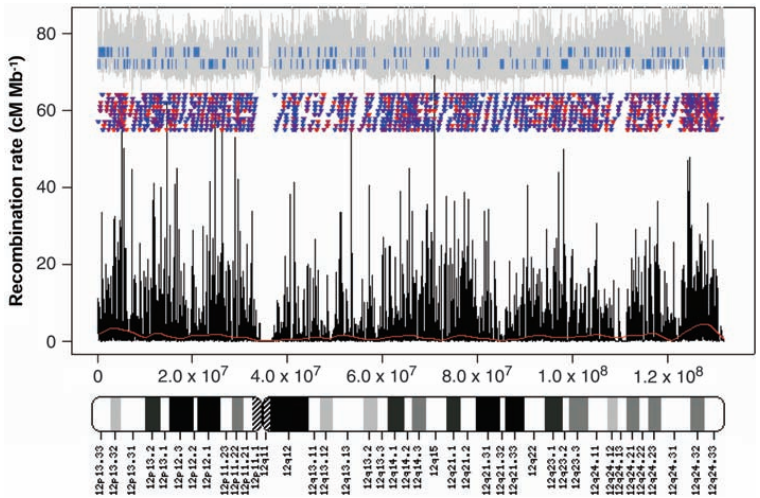
Chiroptera

0.15 0.30 0.45

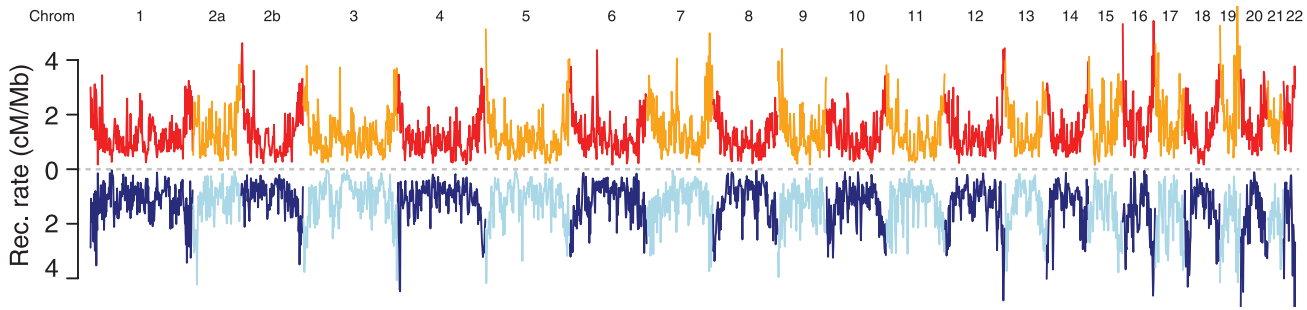
Equilibrium GC



Recombination landscapes



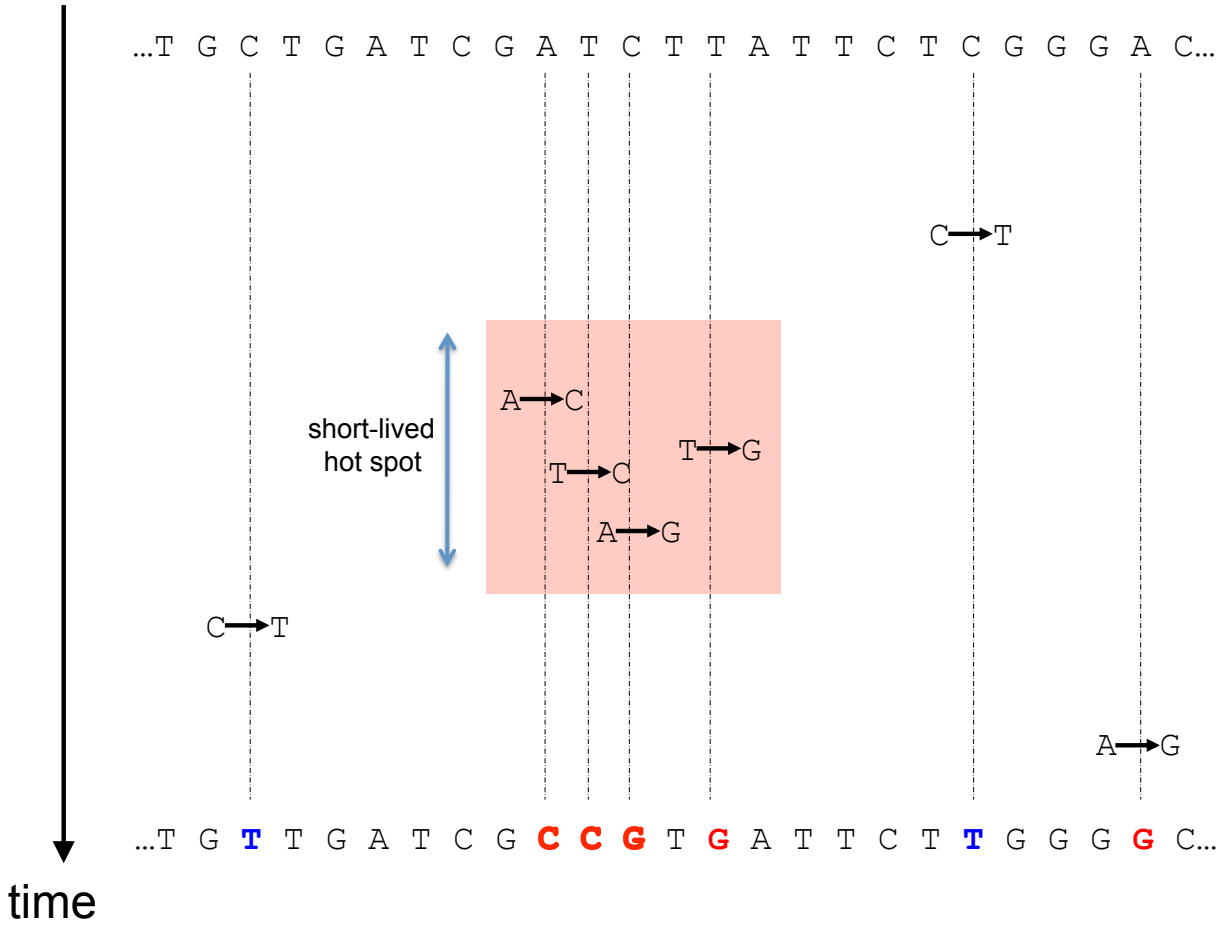
Myers et al, 2005, Science, 310:321



Auton et al, 2012, Science, 336:193

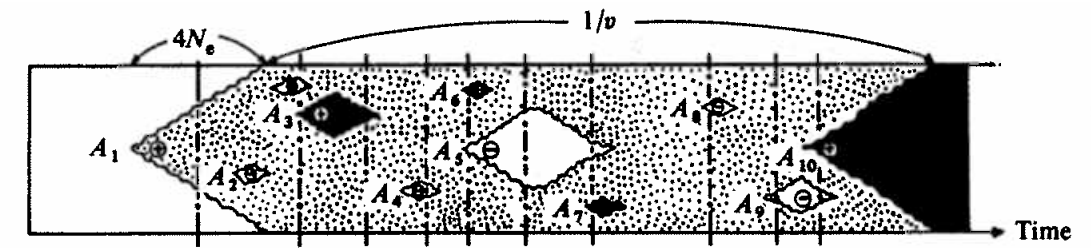
- fine-scale recomb. rates: high variance (coef of var $\simeq 2$)
- high rate of turnover (short lived recombination hotspots)
- broad-scale recomb. rates: determined by karyotype

Genomes record past recombination landscapes



- gBGC links recombination and substitution patterns
- model-based analysis of evolutionary dynamics of recombination

Mutation selection model



Kimura 1982

Substitution process (low mutation approx.)

Substitution rate = mutation rate x fixation probability

$$r = 2Nu\,p = u\,2Np = uP$$

u: mutation rate

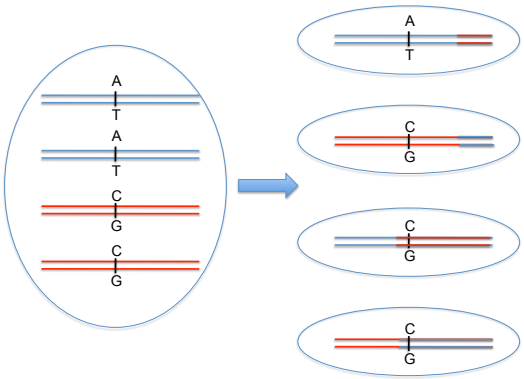
p: fixation probability

N: effective population size

$P = 2Np$: relative fixation probability (relative to neutral)

Fixation probability in the presence of BGC

GC overtransmission



$$x_{GC} = \frac{1+b}{2}$$

$$x_{AT} = \frac{1-b}{2}$$

Relative fixation probability: $2N_ep$

mutation from AT to GC

$$2N_ep = \frac{B}{1-e^{-B}} > 1$$

mutation from GC to AT

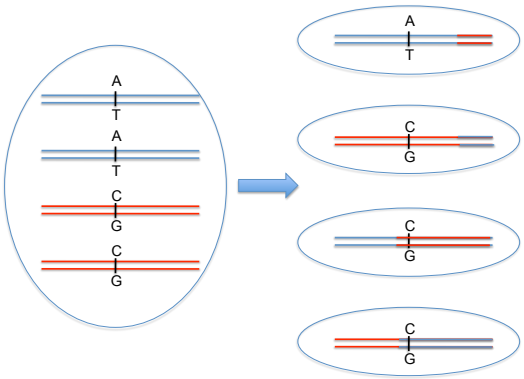
$$2N_ep = \frac{-B}{1-e^{B}} < 1$$

N_e: effective population size

$B = 4N_eb$: scaled conversion coefficient

Fixation probability in the presence of BGC

GC overtransmission



$$x_{GC} = \frac{1+b}{2}$$

$$x_{AT} = \frac{1-b}{2}$$

Relative fixation probability: $2N_ep$

mutation from AT to GC

$$2N_ep = \frac{B}{1-e^{-B}} > 1$$

mutation from GC to AT

$$2N_ep = \frac{-B}{1-e^{B}} < 1$$

N_e: effective population size

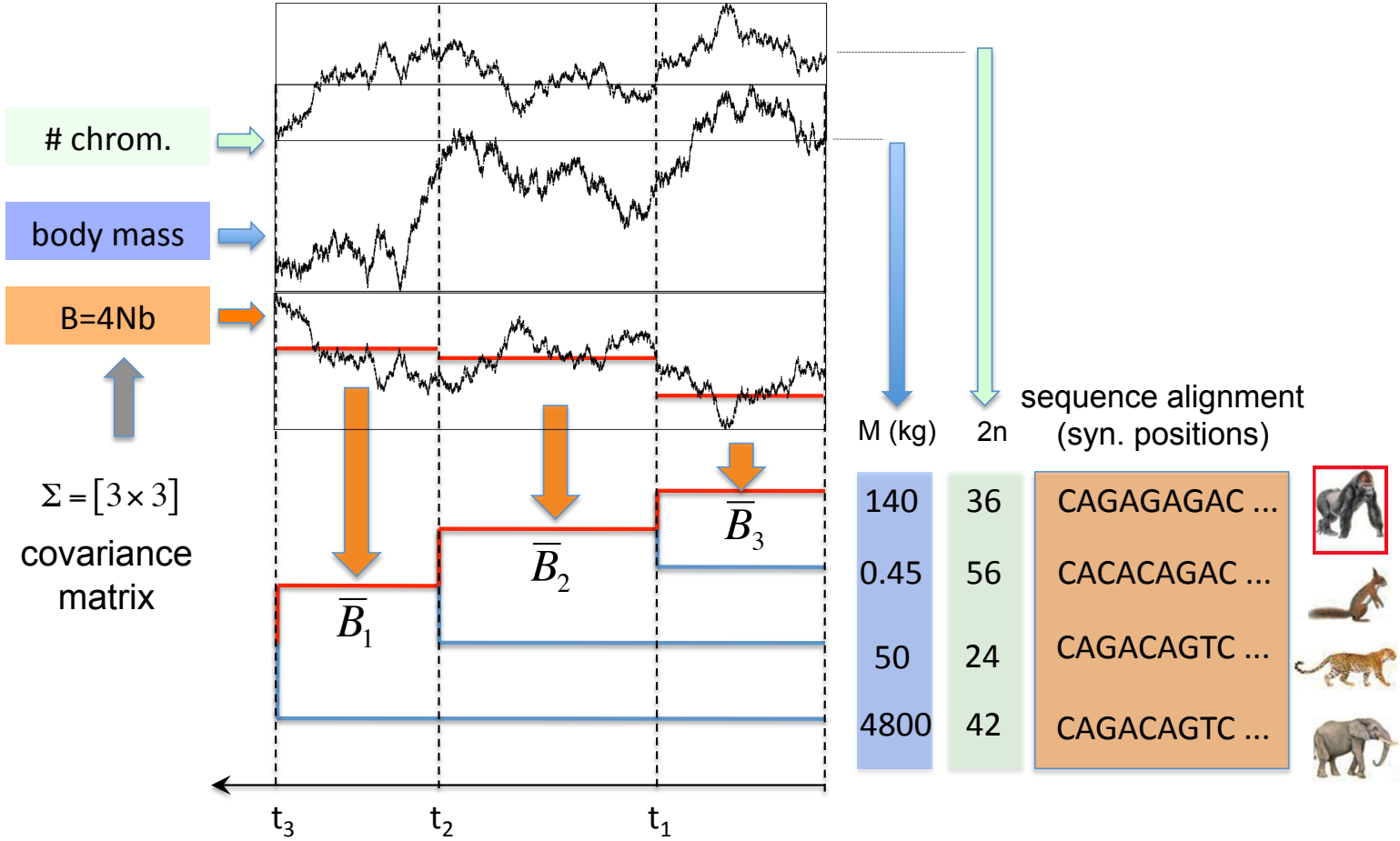
$B = 4N_eb$: scaled conversion coefficient

A mechanistic phylogenetic covariance model

substitution rate = mutation rate x fixation probability

$$\begin{pmatrix} - & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & - & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & - & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & - \end{pmatrix} + B \quad \Longrightarrow \quad \begin{pmatrix} - & \mu_{AC}\frac{B}{1-e^{-B}} & \mu_{AG}\frac{B}{1-e^{-B}} & \mu_{AT} \\ \mu_{CA}\frac{-B}{1-e^{B}} & - & \mu_{CG} & \mu_{CT}\frac{-B}{1-e^{B}} \\ \mu_{GA}\frac{-B}{1-e^{B}} & \mu_{GC} & - & \mu_{GT}\frac{-B}{1-e^{B}} \\ \mu_{TA} & \mu_{TC}\frac{B}{1-e^{-B}} & \mu_{TG}\frac{B}{1-e^{-B}} & - \end{pmatrix}$$

$B = 4N_e\, b$: scaled conversion coefficient



Lartillot, 2013, Molecular Biology and Evolution, in press

Overall modeling strategy

- only 4-fold degenerate third codon positions
- modeling joint variation in B, body mass (M) and karyotype ($2n$)
- modeling among-gene variation (recombination seascapes)

Data

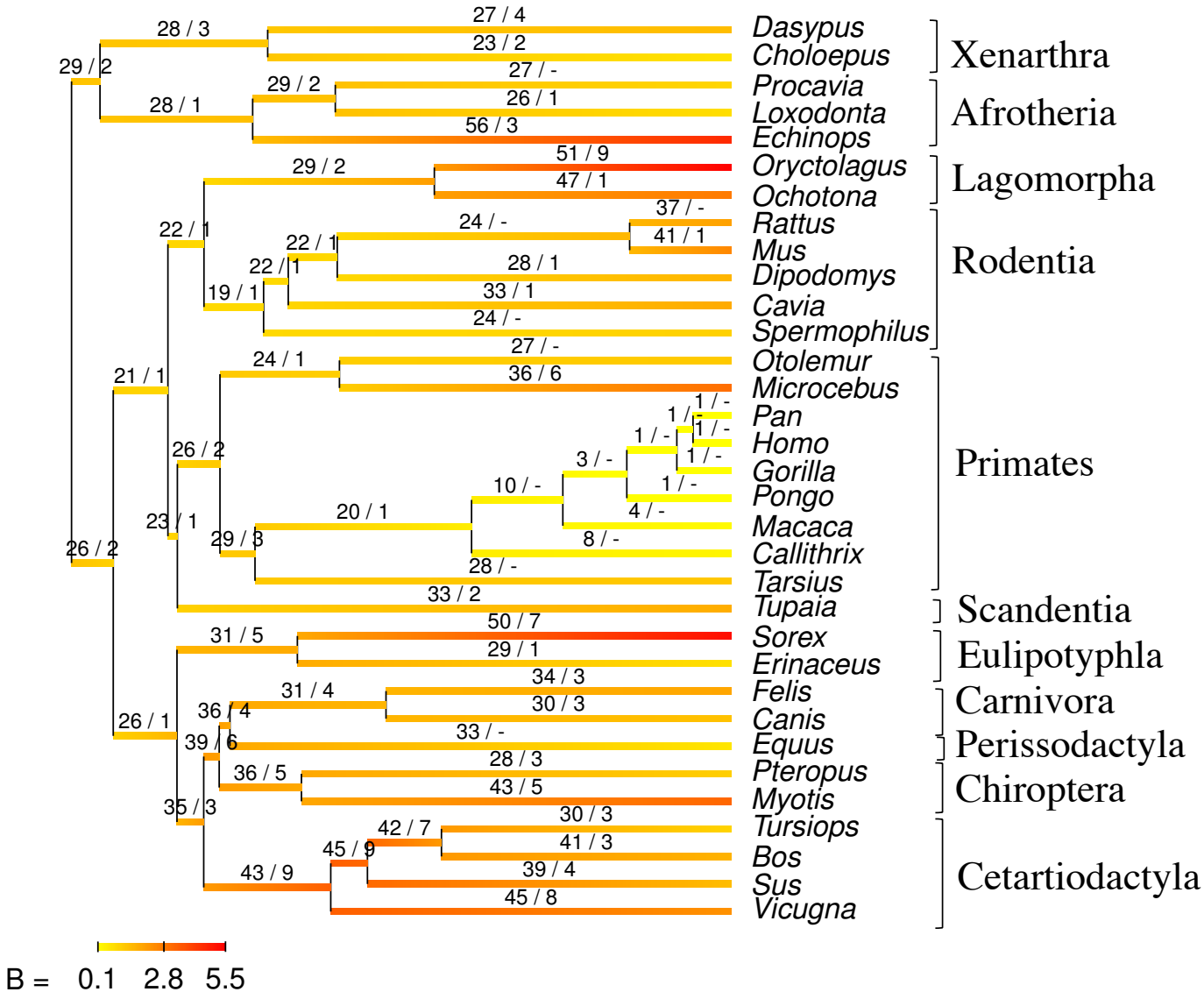
Exon-rich dataset

- 180 exons from Orthomam, with at least 30 taxa
- 1000 exons (30 jacknife replicates of 100 exons)
- only 4-fold degenerate positions
- analysis replicated using non-CpG positions

Taxon-rich dataset

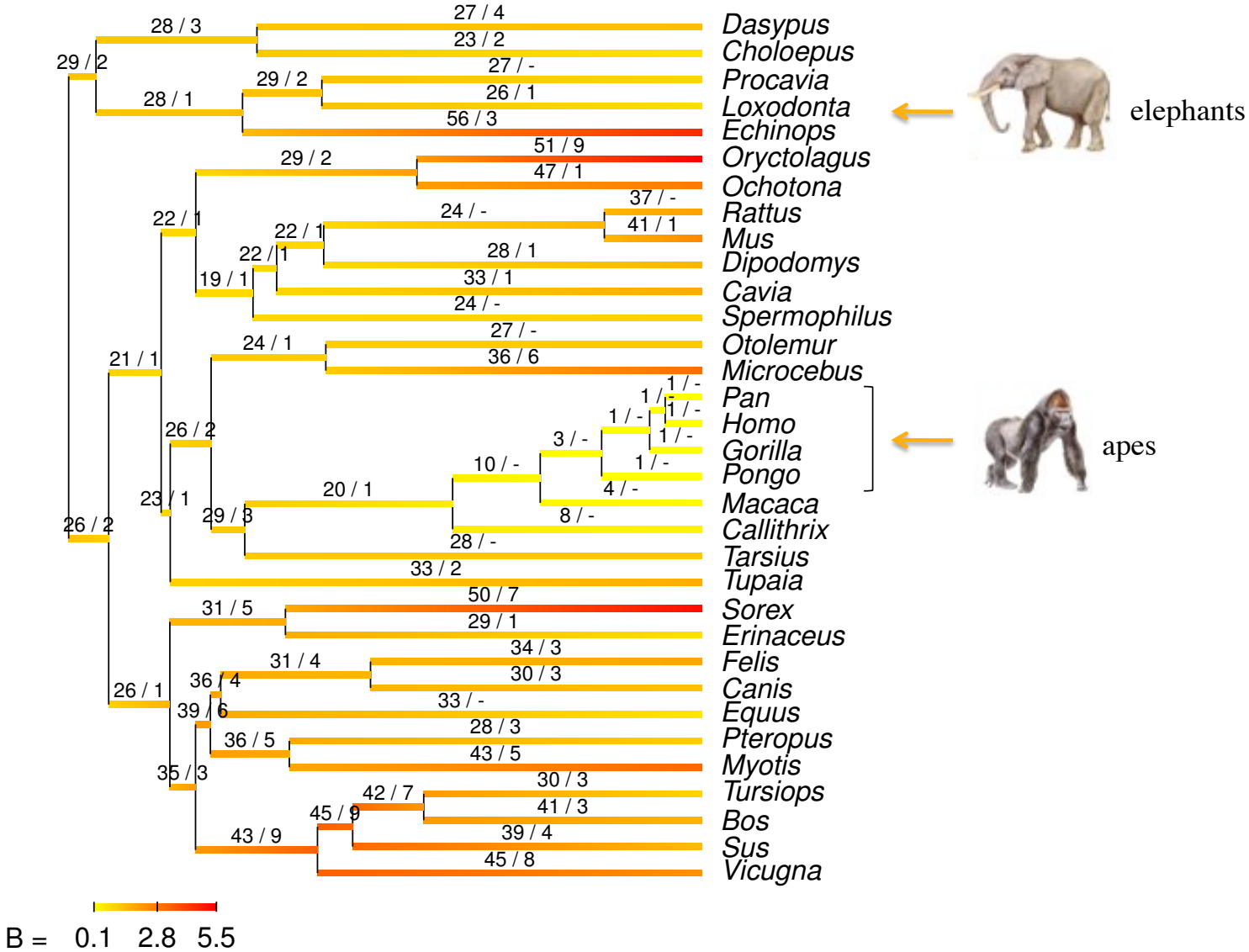
- 17 single-exon genes 73 mammals

Reconstructed history of $B = 4N_e b$



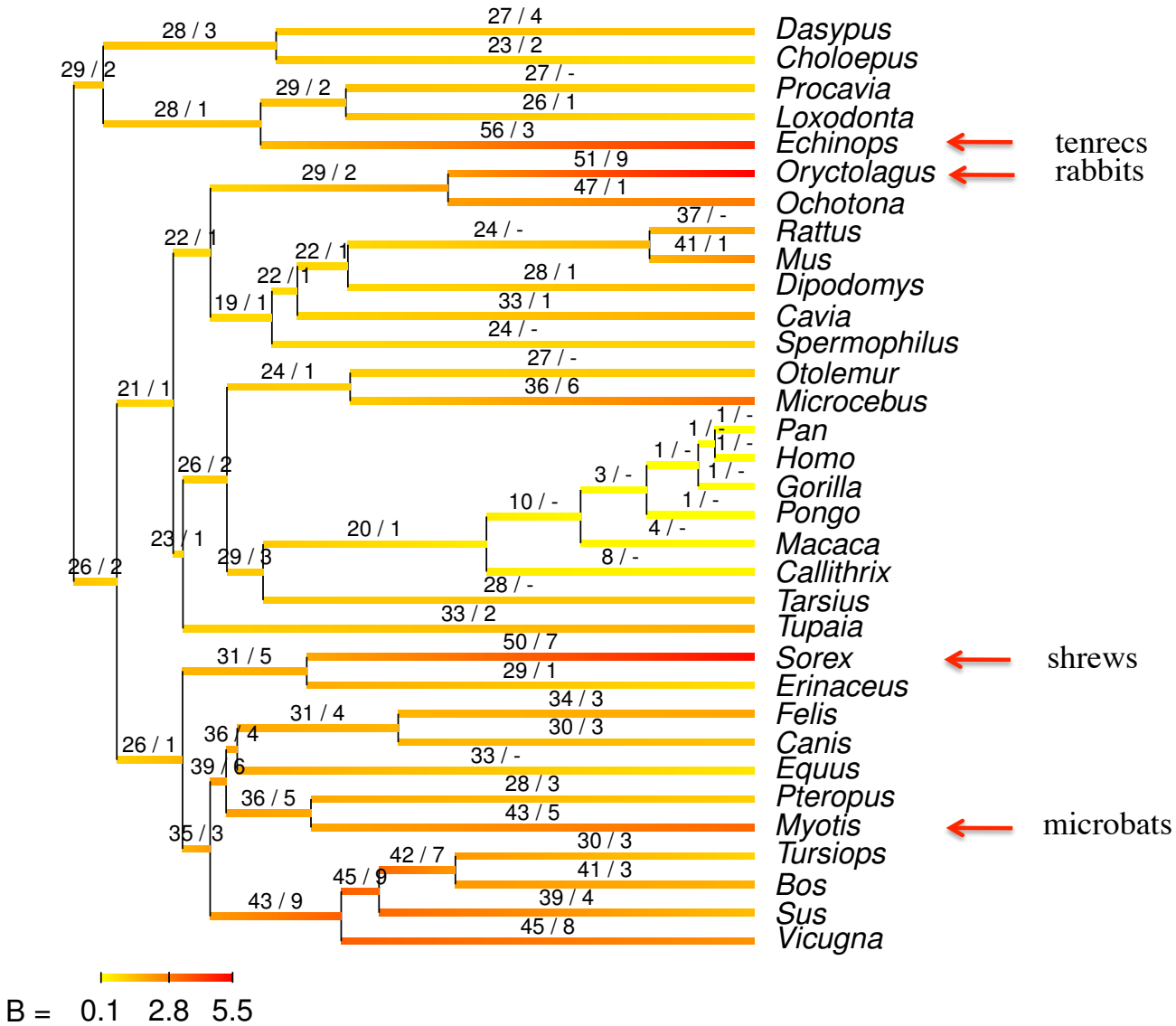
Lartillot, 2013, Molecular Biology and Evolution, 30:489

Phylogenetic history of population-genetic regimes

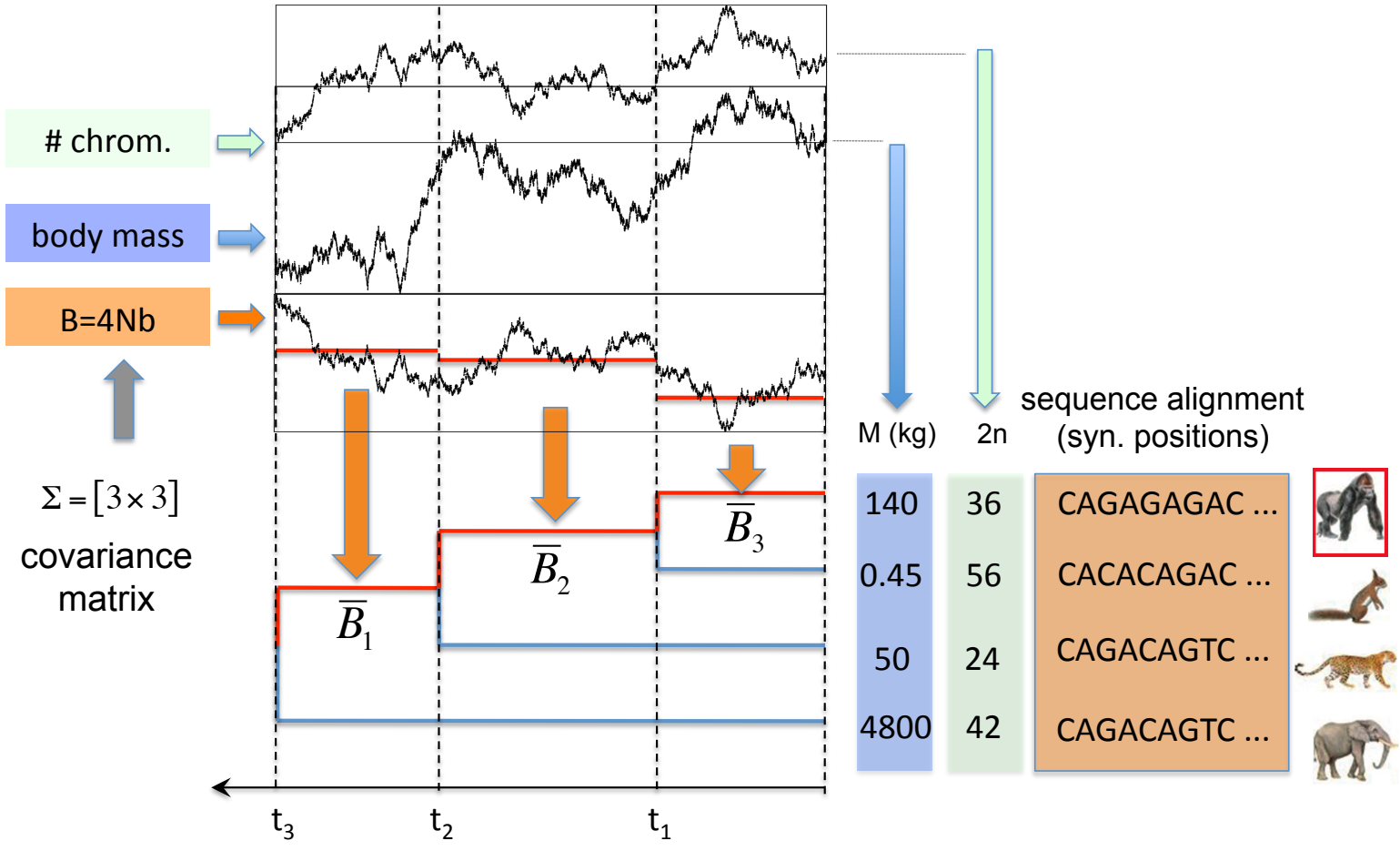


large mammals, small N_e: drift dominates ($B < 1$)

Phylogenetic history of population-genetic regimes



small mammals, large N_e: BGC dominates ($B > 1$)



Lartillot, 2013, Molecular Biology and Evolution, in press

Allometry and covariance

- (ln B, ln M, ln n) follow a trivariate Brownian motion
- $B \sim M^{\gamma} n^{\alpha}$ for some coefficients of allometry γ and α

Estimated allometric scaling of $B = 4N_e b$

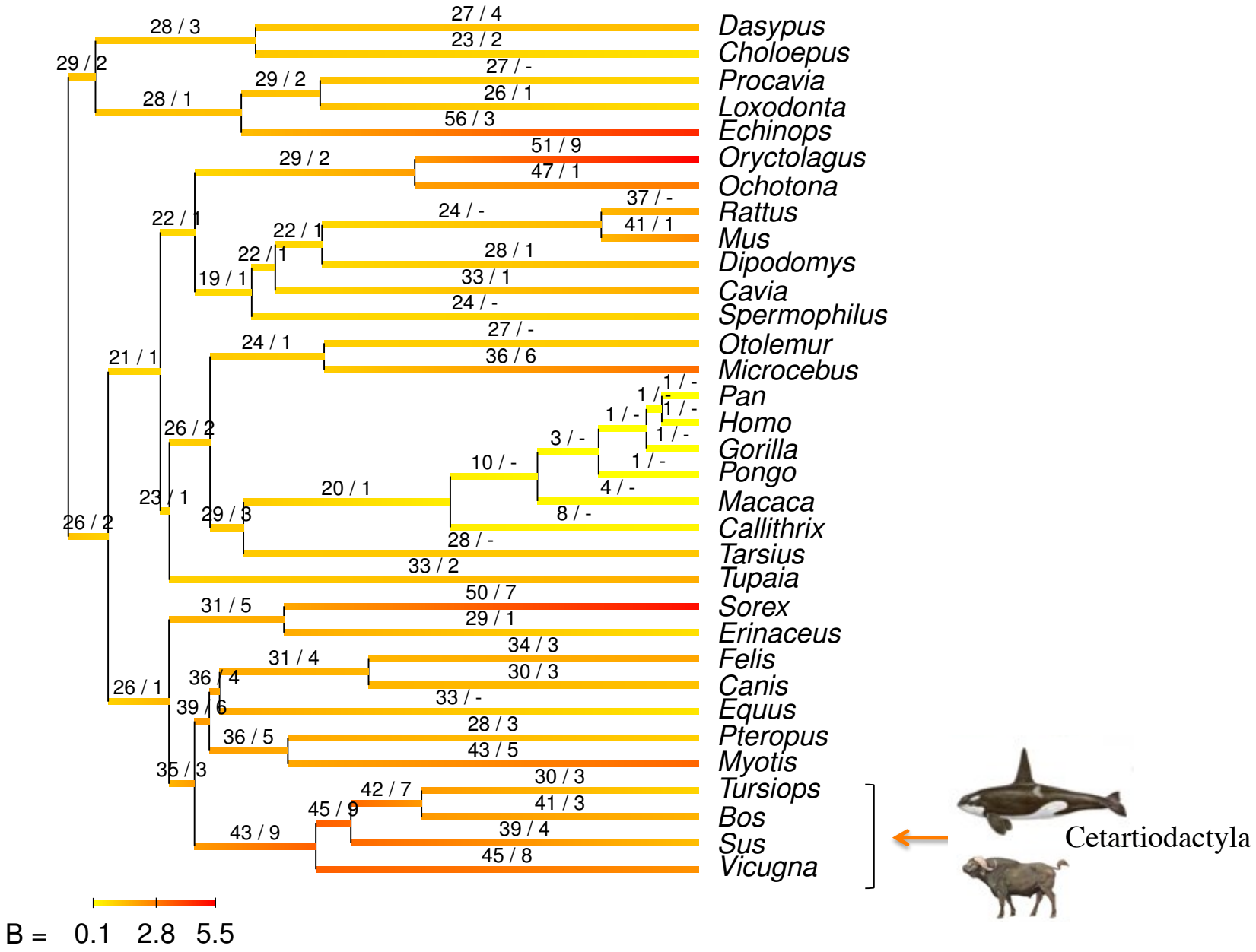
$$B \sim M^{\gamma} n^{\alpha}$$

M: body mass (prediction: $\gamma < 0$)

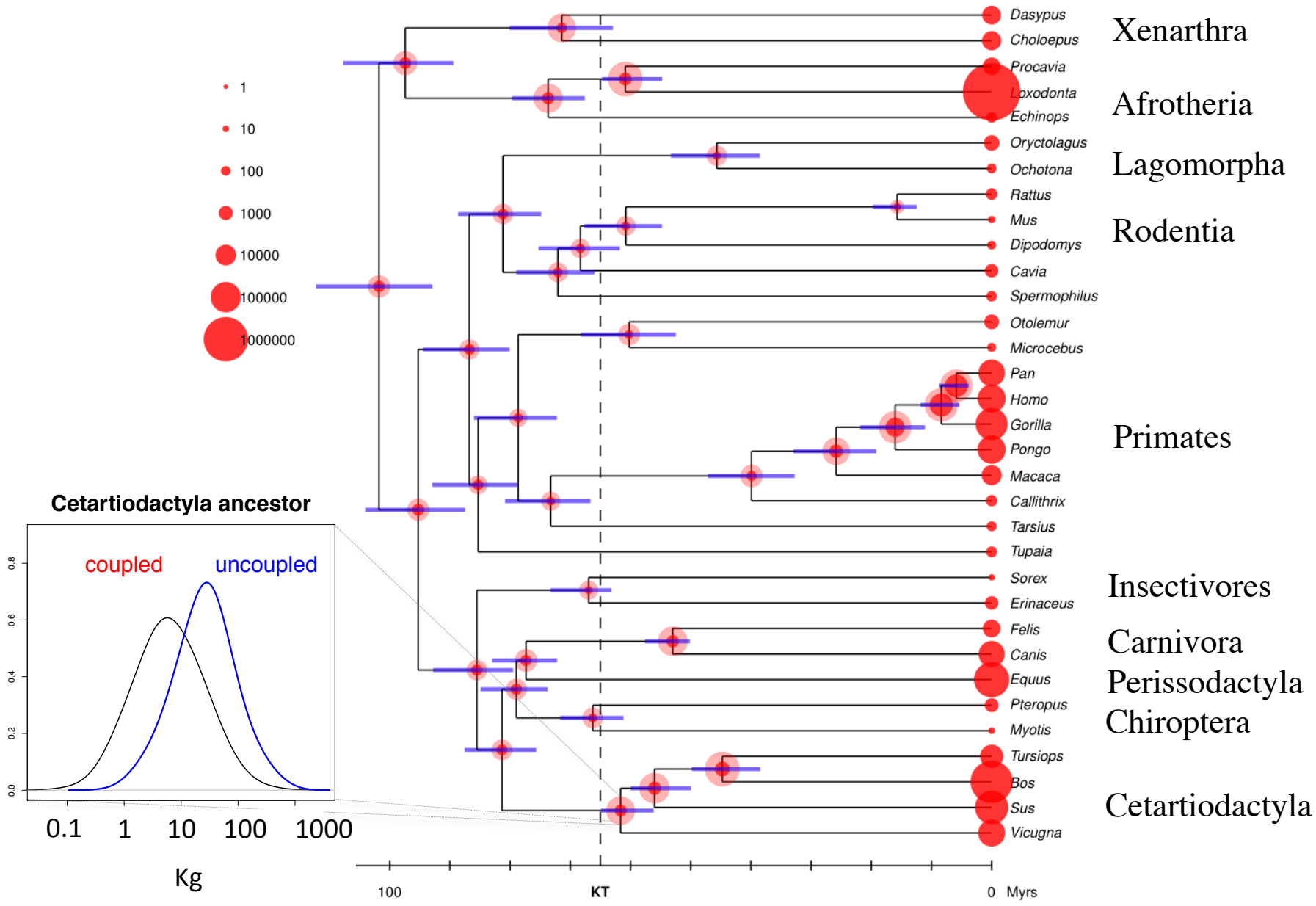
n: number of chromosomes (prediction: $\alpha = 1 > 0$)

	γ		α	
73 taxa 17 genes	-0.11**	(-0.19, -0.03)	1.28**	(0.54, 2.03)
33 taxa 1000 exons	-0.28*	(-0.52, -0.01)	0.21	(-1.20, 1.56)

Reconstructed history of $B = 4N_e b$



Divergence times and body mass evolution



last common ancestor: between 150 g and 3.5 kg

Reconstructing past population-genetic regimes

- mutation rate per generation u (substitution rate)
- effective population size N_e (dN/dS, GC)
- scaled conversion coefficient $B = 4N_e b$ (GC)
- evolutionary dynamics of recombination landscapes (GC)
- useful for understanding mechanisms of genome evolution

Acknowledgments

- Raphael Poujol (coevol software)
- Frédéric Delsuc (rates, dates and traits)
- Mathieu Groussin, Manolo Gouy
- Nicole Uwimana, Benoit Nabholz
- Benjamin Horvilleur (Brownian paths)
- many others...

Software availability *(coevol)*

- www.phylobayes.org