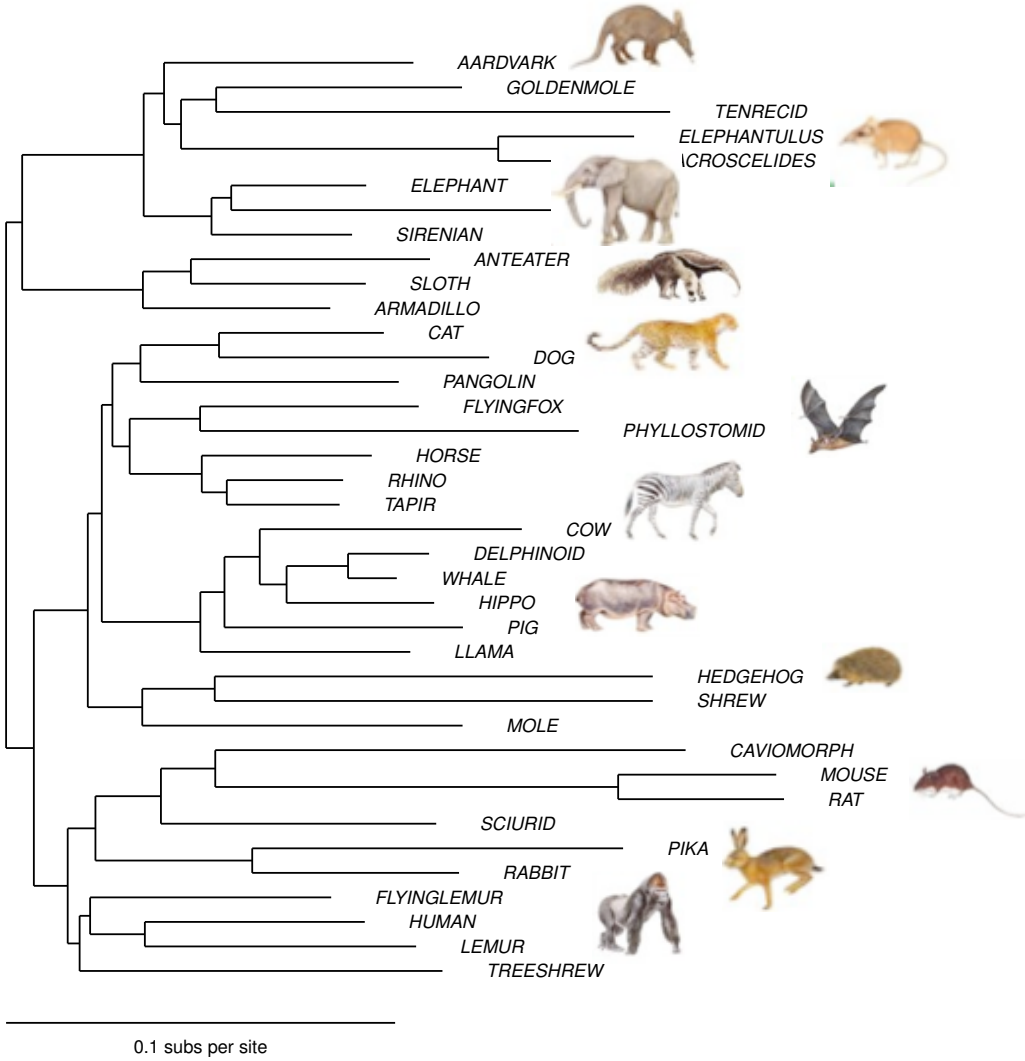
Rates, dates and traits.

The comparative method in evolutionary genomics

Nicolas Lartillot

June 19, 2014

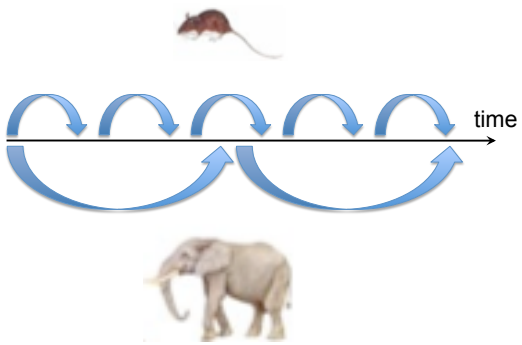
Variation of the substitution rate among lineages



Variation of the substitution rate among lineages

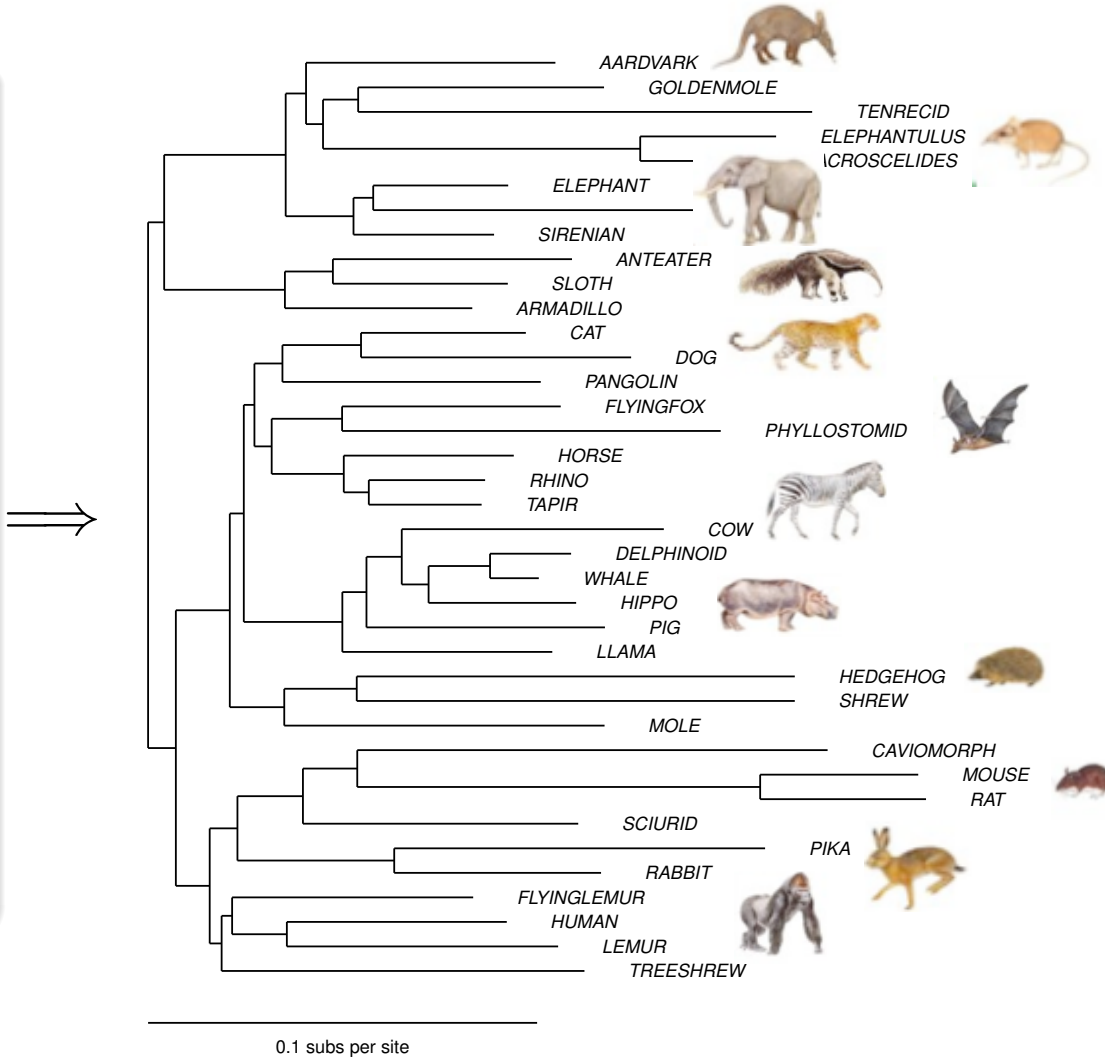
Possible causes

- generation-time effect

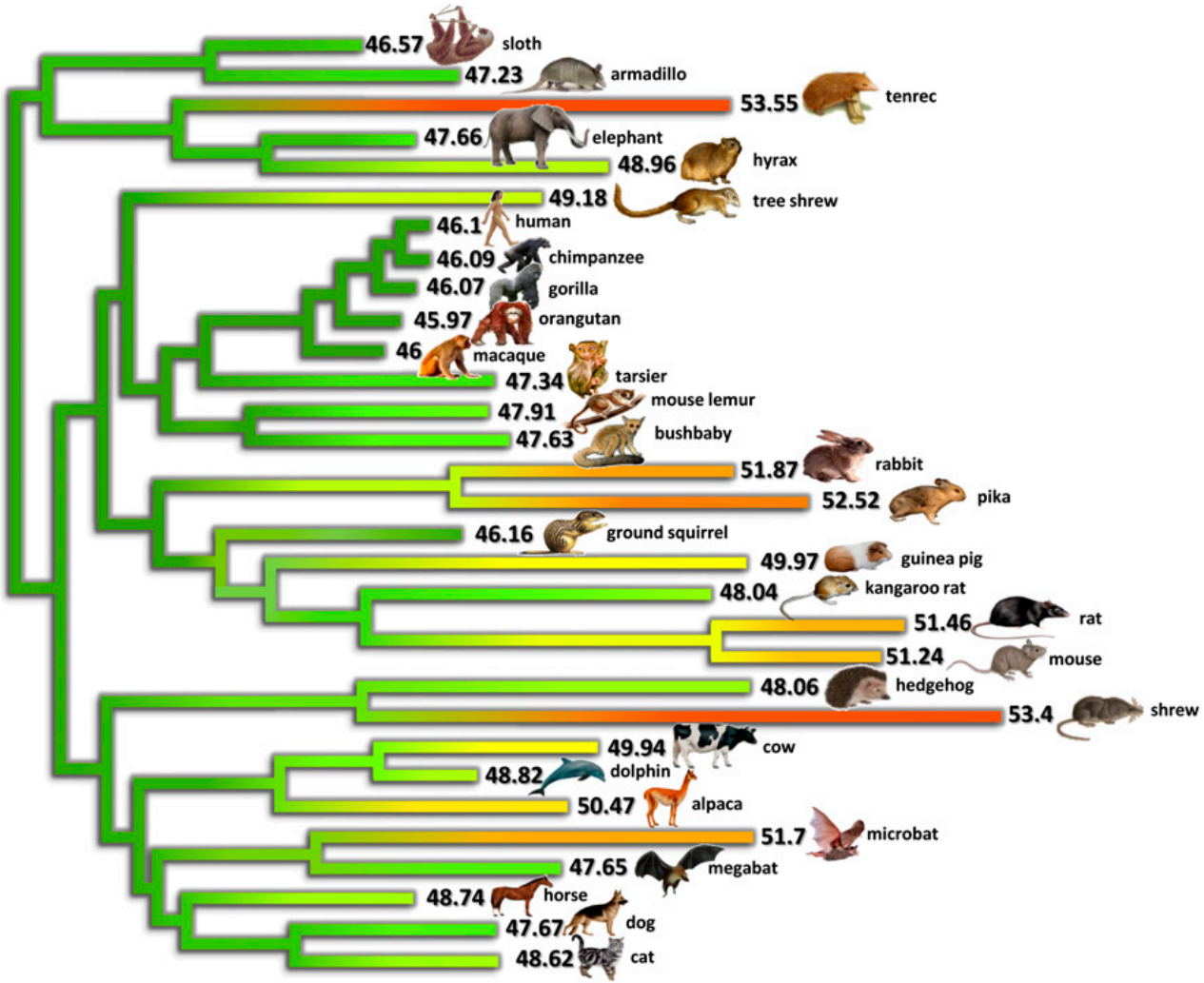


- metabolic rate effects
- selection for longevity

(reviewed in Lanfear et al, 2010)



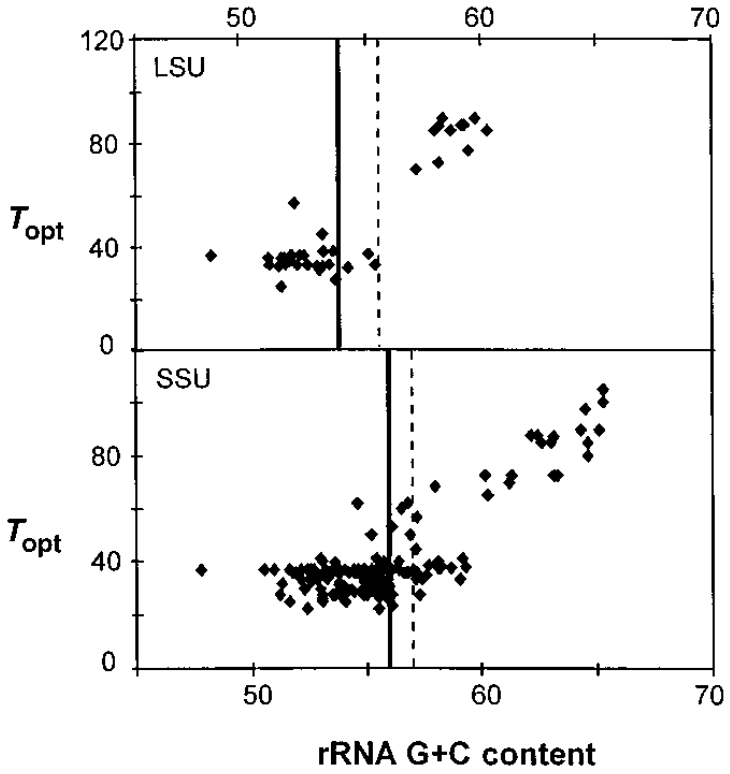
Variation in GC content



variation in GC content

Romiguier et al, 2010, Genome Res 20:1001

rRNA and proteome composition vs. temperature



Galtier et al 1999

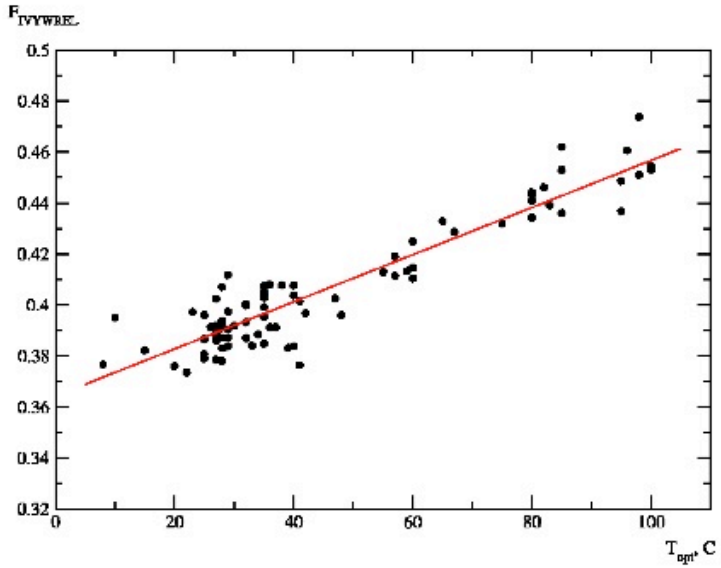
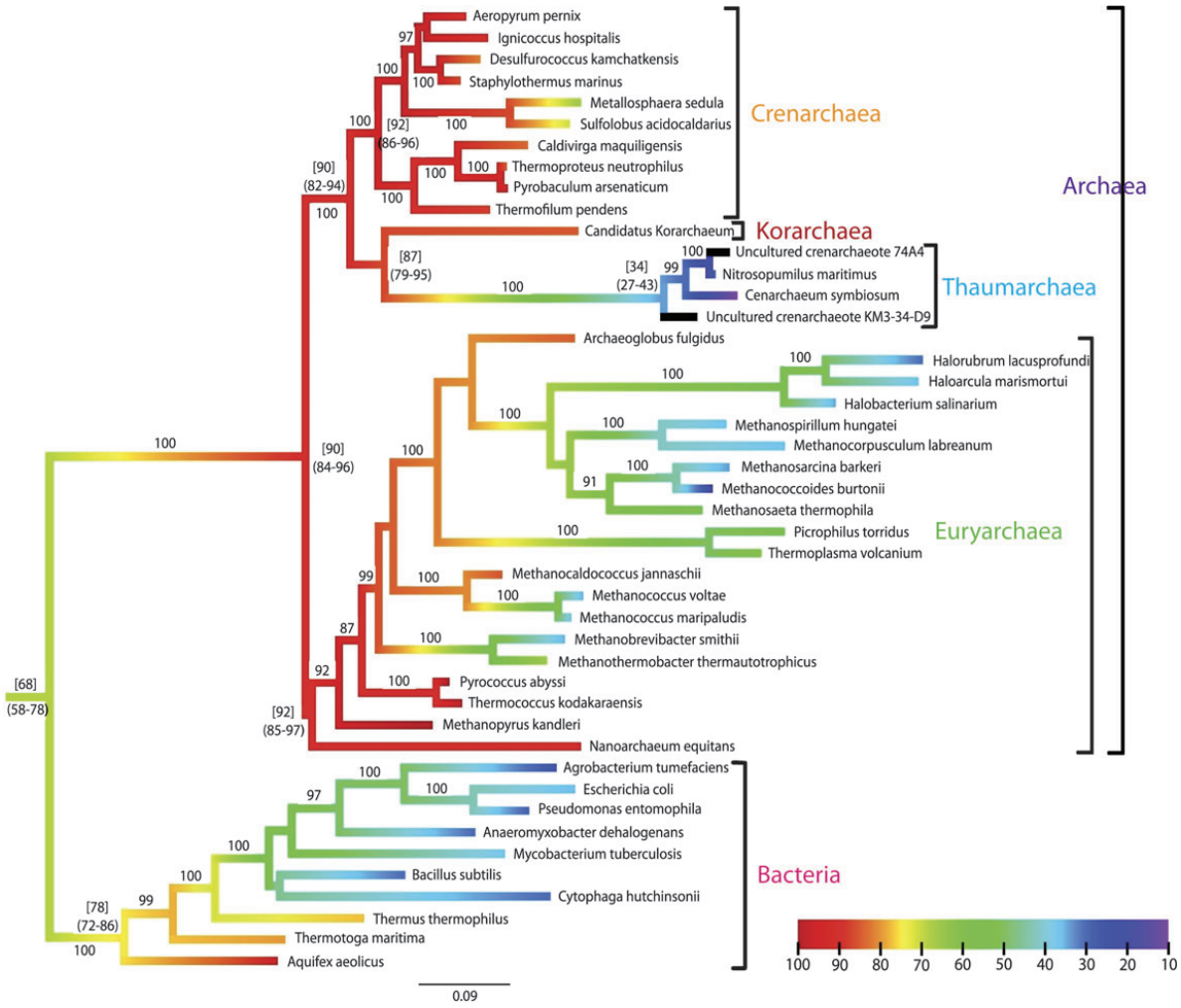
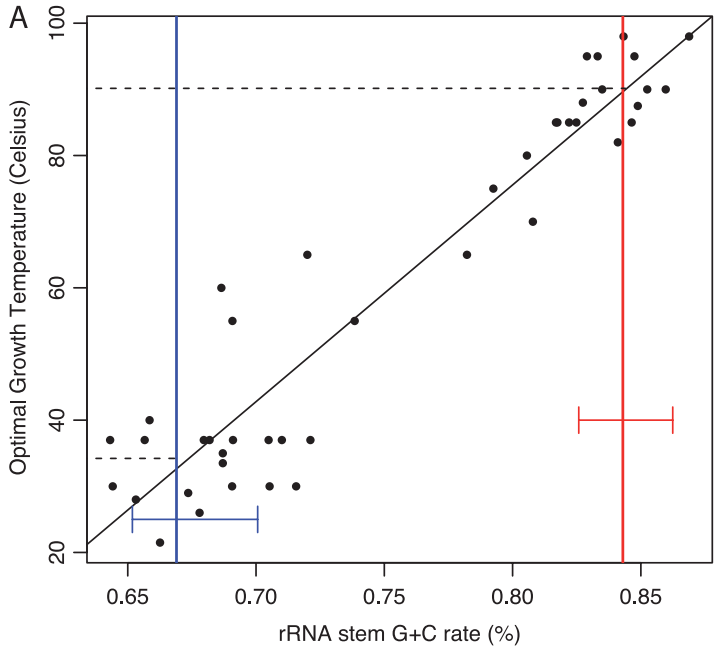


Figure: Procaryotic protein content in amino acids IVYWREL correlated to the species optimal growth temperatures (OGT) [Zeldovich et al., 2007].

Zeldovich et al 2007

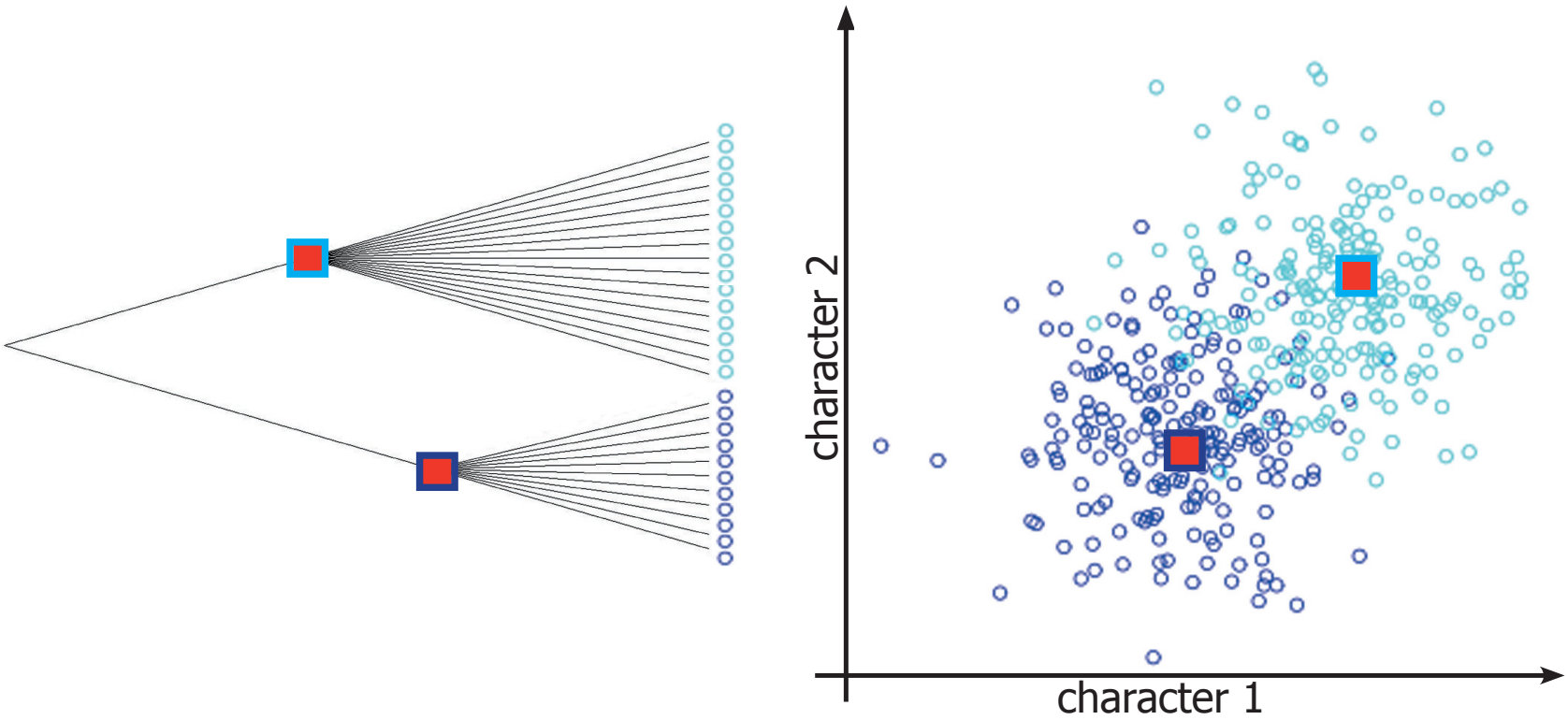
Ancestral growth temperatures inferred using rRNA





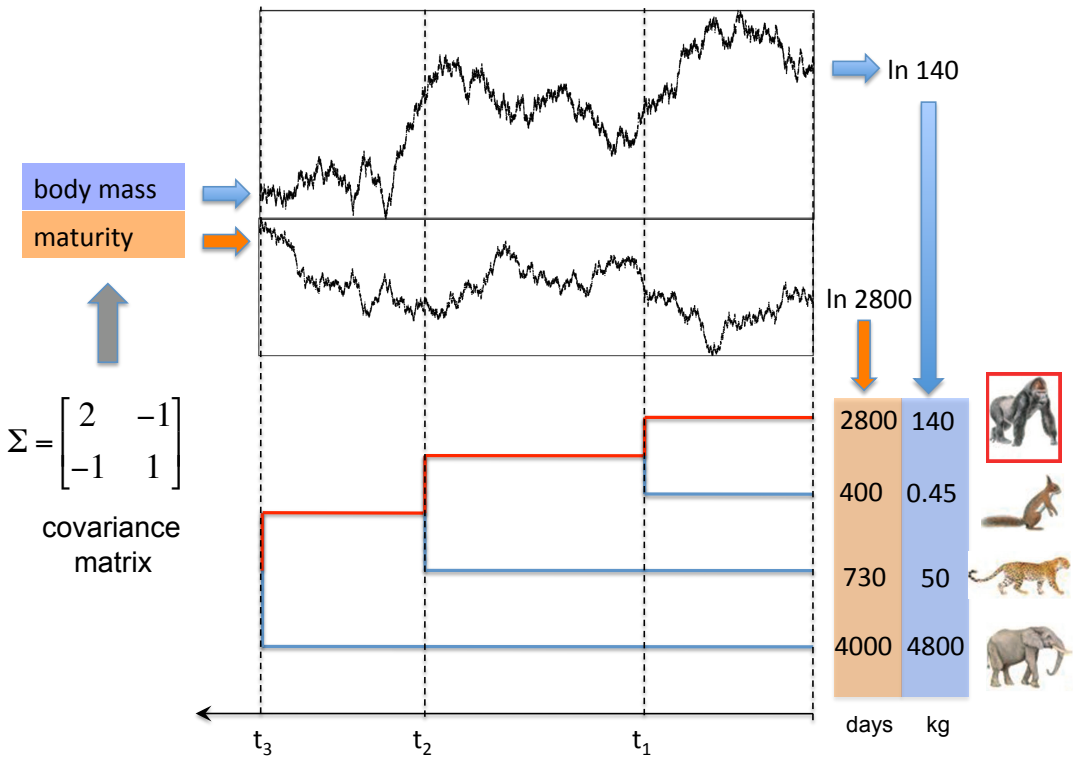
Groussin et al, 2011, Mol Biol Evol 28:2661

The problem of phylogenetic inertia



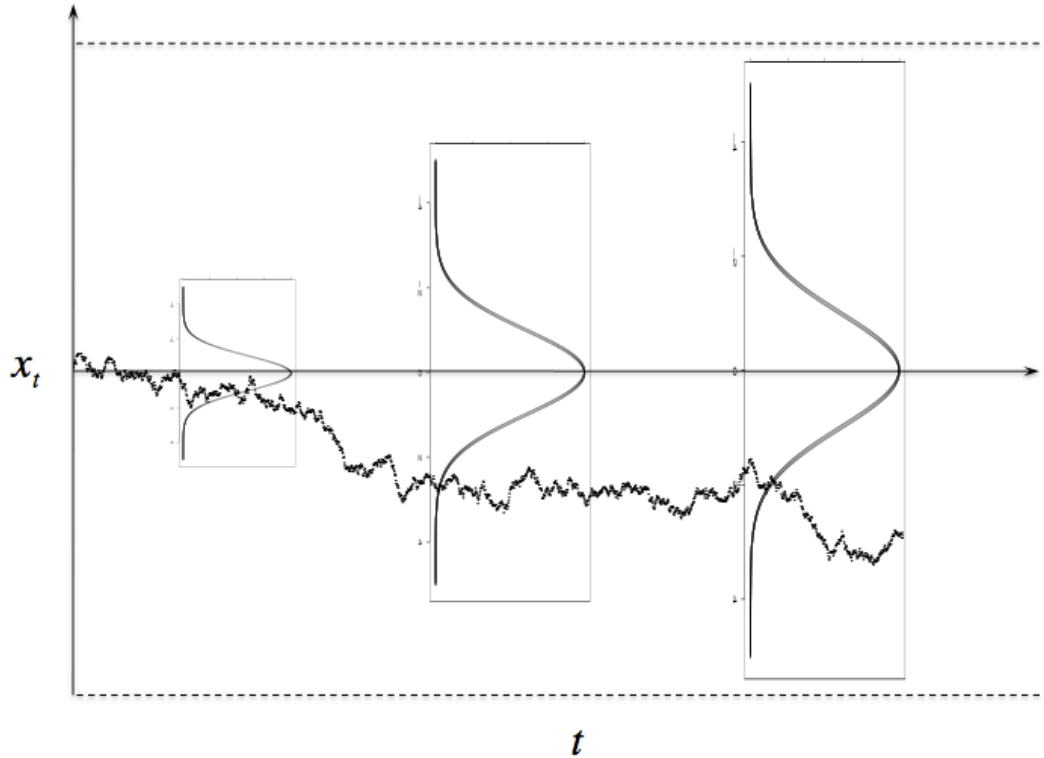
Felsenstein, 1985, Am Nat 125:1

The comparative method



- Assume traits follow bivariate Brownian motion
- data $X = (x_{ik})$, $i = 1..N_{taxa}$, $k = 1, 2$
- parameter Σ (3 independent parameters)
- maximize likelihood $L(\Sigma) = p(X \mid \Sigma)$
- estimate Σ_{12}, test whether $\Sigma_{12} < 0$, etc

Brownian process

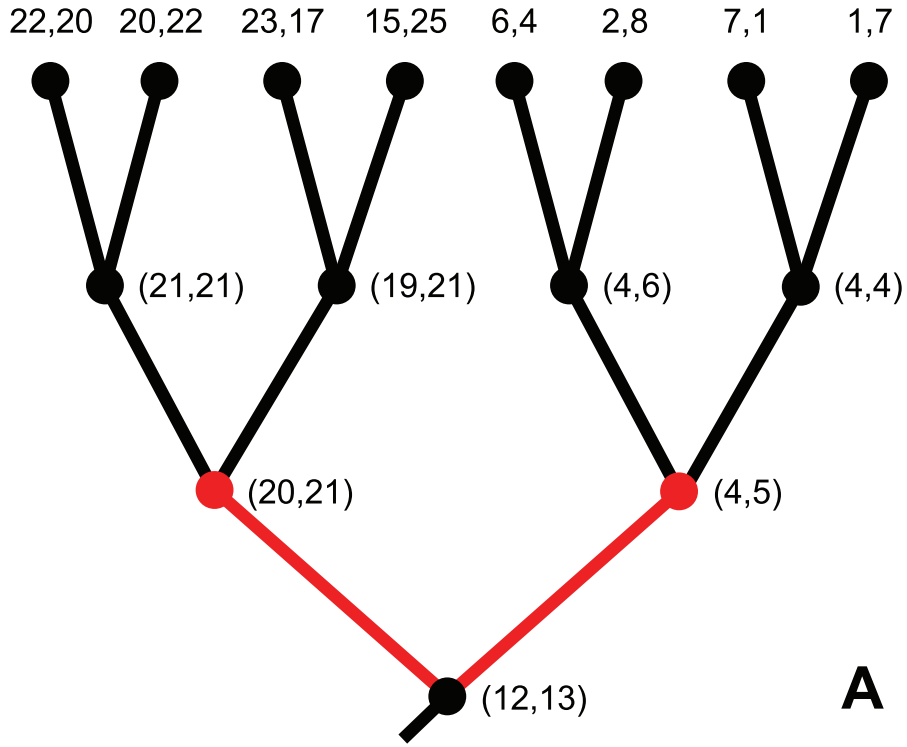


$$x_t \sim Normal(x_0, \sigma^2 t)$$

or, for a bivariate or multivariate process:

$$X_t \sim Normal(X_0, t\Sigma)$$

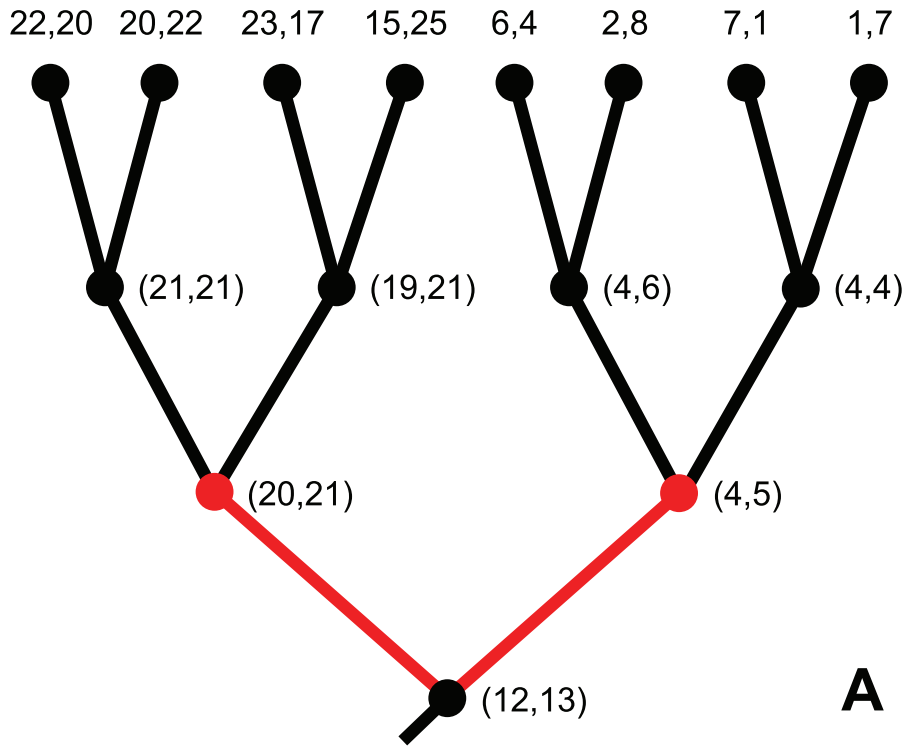
The comparative method



Algorithm

- take two sister species i and j, with traits X_i and X_j
- T: time since their last common ancestor
- $\Delta X = X_j - X_i \sim Normal(0, 2T\Sigma)$.
- define normalized *contrast* $\Delta Y = \Delta X / \sqrt{2T}$: $\Delta Y \sim Normal(0, \Sigma)$.
- do it for all tips and then recursively up to the root

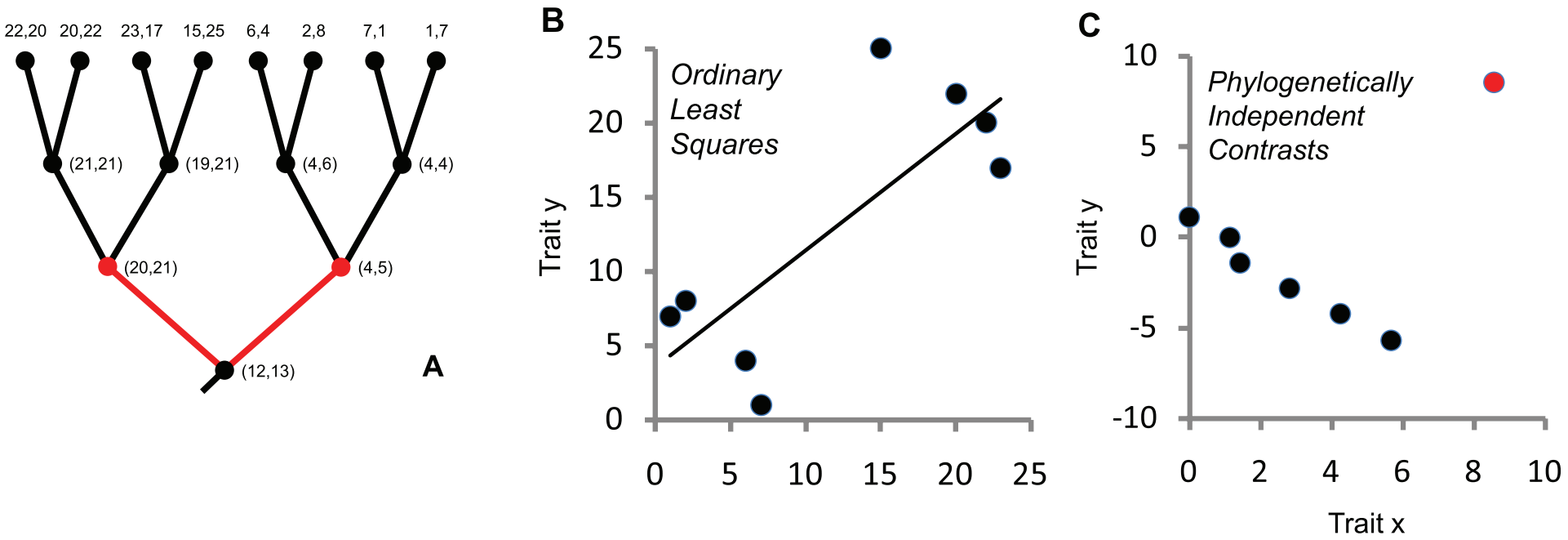
The comparative method



- P taxa $\rightarrow P-1$ normalized contrasts ΔY_j, for $j = 1..P-1$
- normalized contrasts are iid: $\Delta Y_j \sim Normal(0, \Sigma)$
- or equivalently, likelihood reduces to

$$p(X \mid \Sigma) \quad \propto \quad \prod_{j=1}^{P-1} Normal(\Delta Y_j; 0, \Sigma)$$

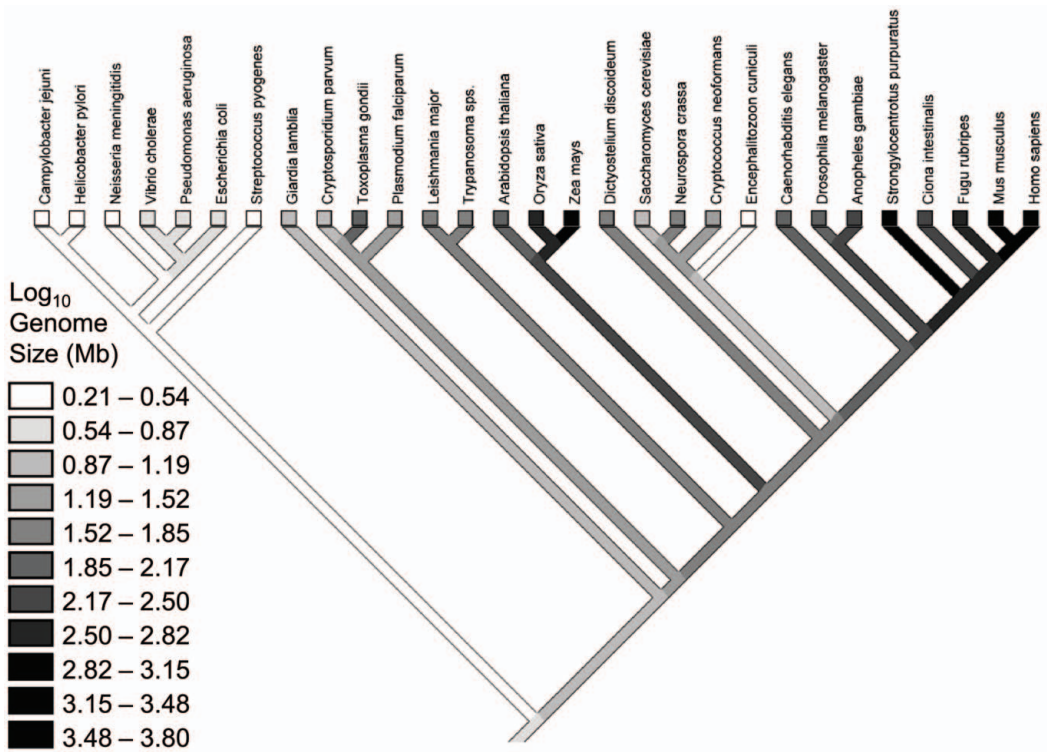
Independent Contrasts

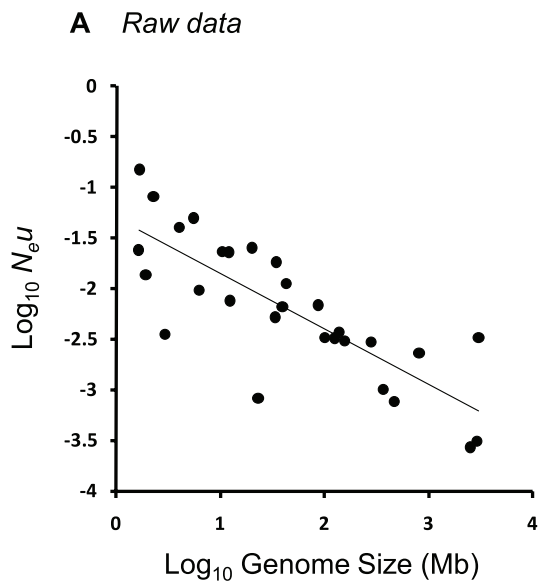


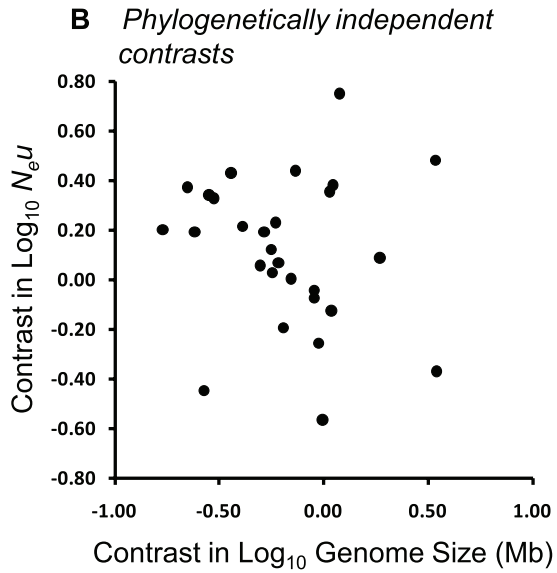
Whitney and Garland Jr, 2010, PLoS Genetics 6:e1001080

- contrasts are statistically independent
- equivalent to asking whether traits show correlated *variations*

Genome size and effective population size revisited

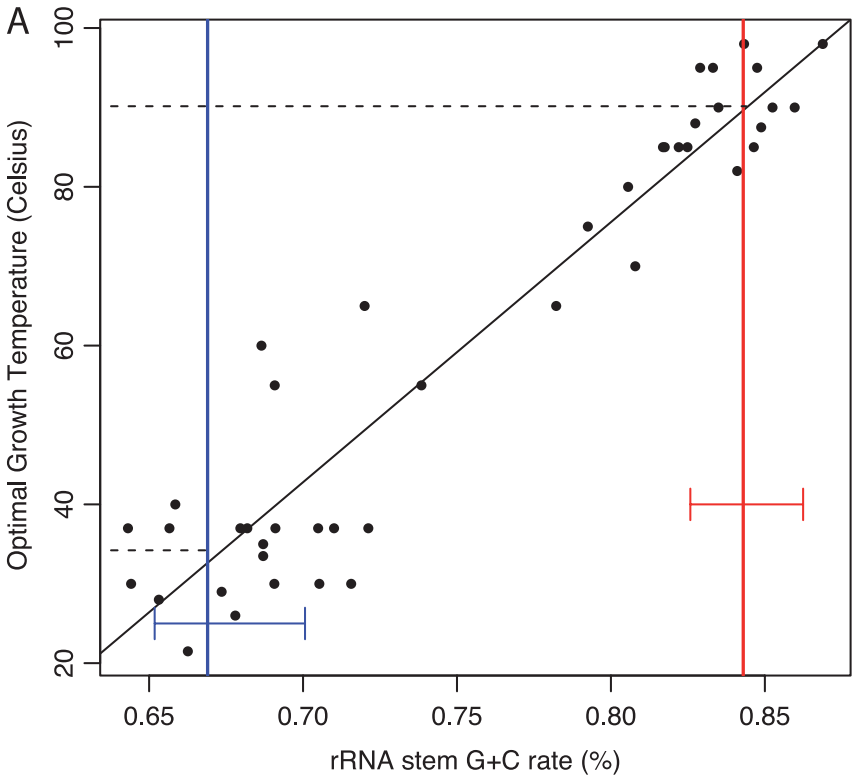




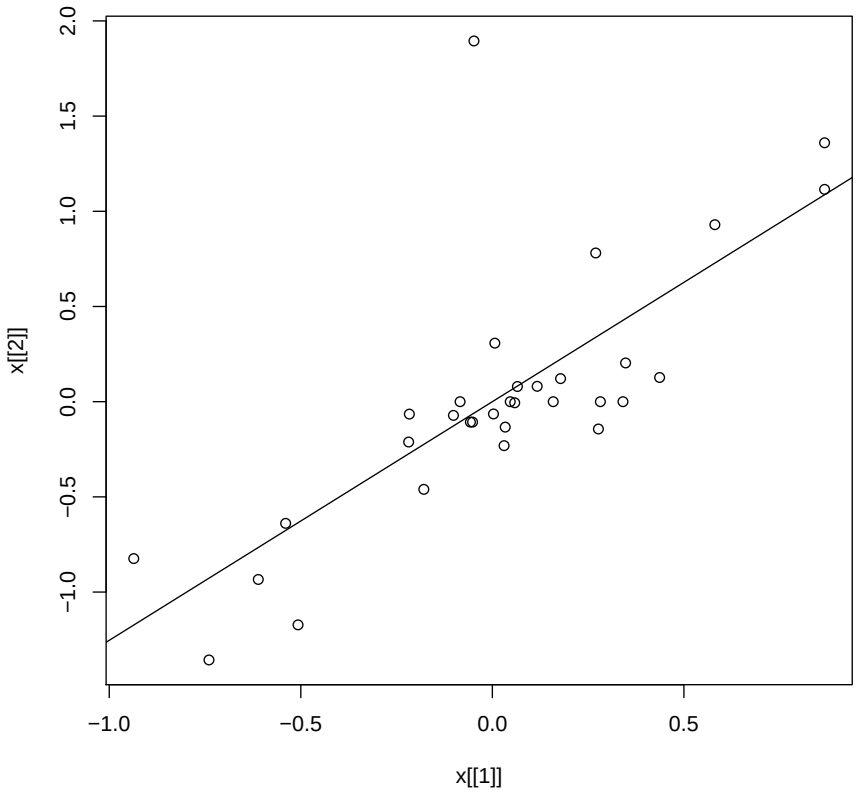


Whitney and Garland Jr, 2010, PLoS Genetics 6:e1001080

Phylogenetically-corrected regression in Archaea

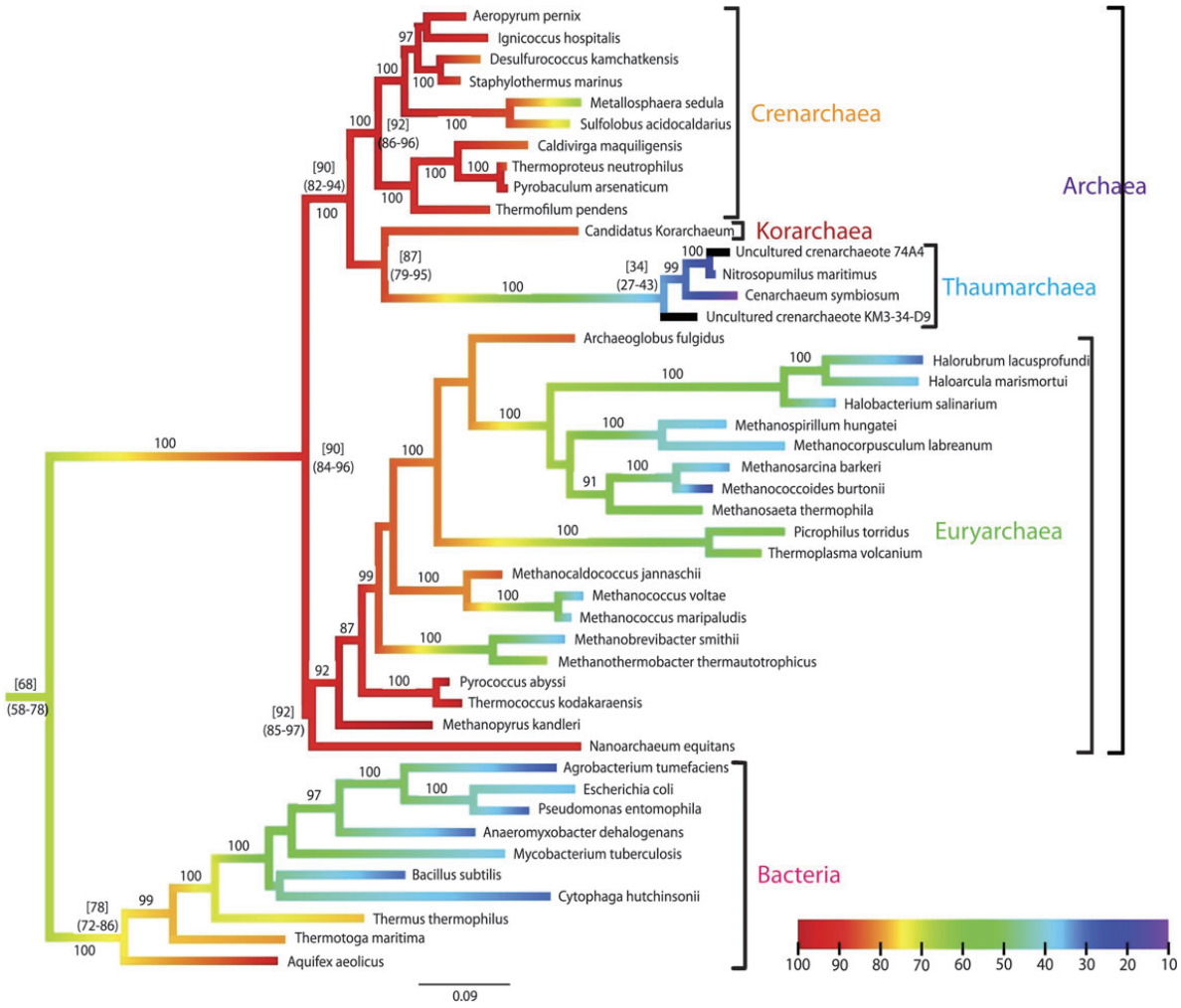


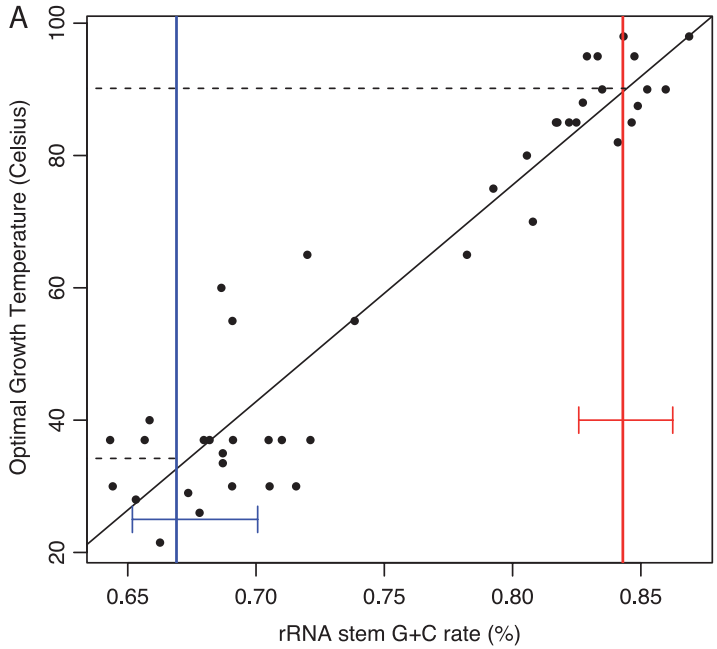
raw regression



normalized contrasts

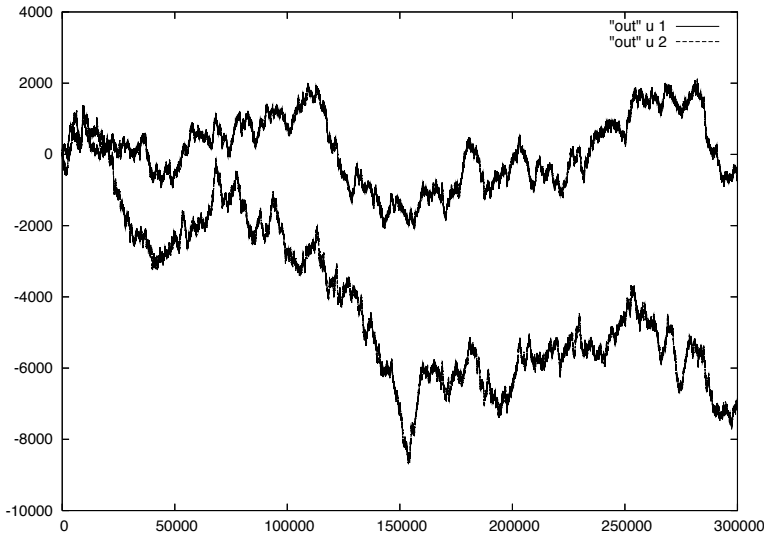
Ancestral growth temperatures inferred using rRNA



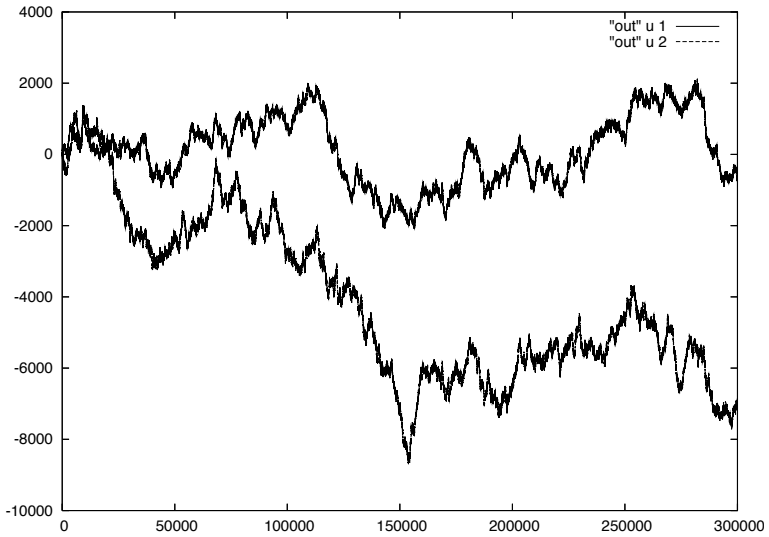


Groussin et al, 2011, Mol Biol Evol 28:2661

Brownian model



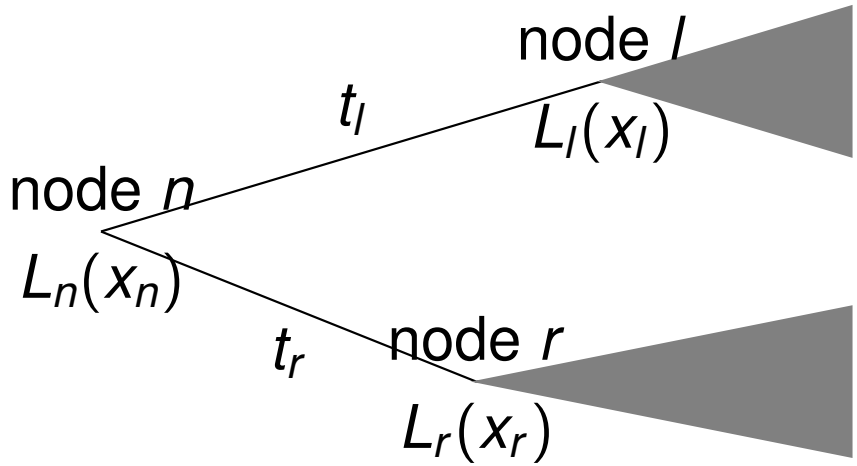
Brownian model



Regressing the trait against the predictor

- trait x (e.g. temperature)
- predictor y (e.g. GC content)
- trait evolution over time t: $\Delta x \sim N(0, \eta^2 t)$
- predictor evolution: $\Delta y = \alpha \Delta x + \epsilon$, with $\epsilon \sim N(0, \kappa^2 t)$
- $\Delta x \mid \Delta y, \alpha, \eta, \kappa \sim N(m, \lambda^2)$

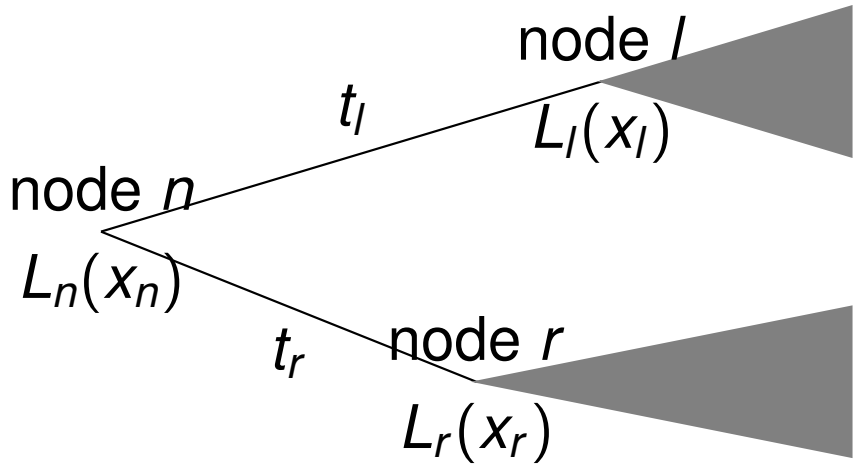
A phylogenetic Kalman filter



Regressing the trait against the predictor

- trait x (e.g. temperature)
- predictor y (e.g. GC content)
- trait evolution along branch l: $\Delta x_l = x_l - x_n \sim N(0, \eta^2 t_l)$
- predictor evolution: $\Delta y_l = y_l - y_n = \alpha \Delta x_l + \epsilon_l$, with $\epsilon_l \sim N(0, \kappa^2 t_l)$
- $\Delta x \mid \Delta y, \alpha, \eta, \kappa \sim N(m, \lambda^2)$

A phylogenetic Kalman filter



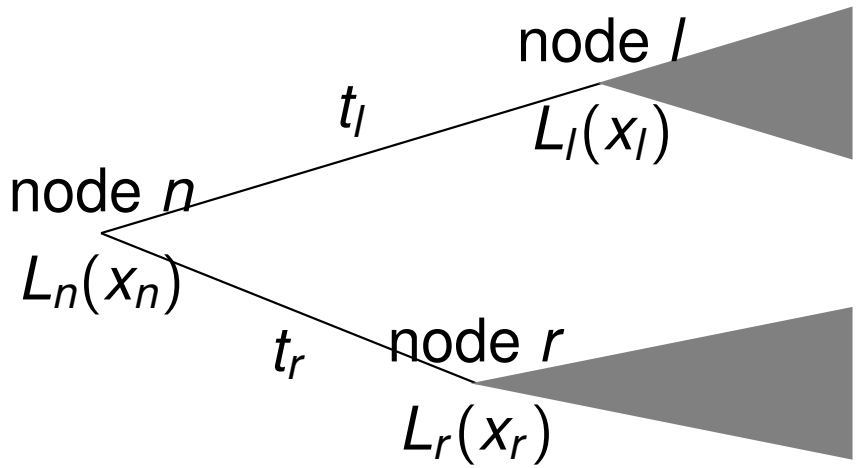
Conditional likelihoods

- at node n, and for $x_n \in \mathbb{R}$, $L_n(x)$ is Gaussian:

$$L_n(x_n) \quad = \quad K_n e^{-\frac{1}{2\sigma_n^2}(x_n-\mu_n)^2}$$

- calculate K_n, σ_n, μ_n as functions of K_l, K_r, σ_l, σ_r, μ_l, μ_r ...
- backward: likelihood computation; forward: stochastic traceback
- combined with conjugate sampling of covariance matrix
- Lartillot, 2014. Bioinformatics, 30:486-496 (see also Ho and Ané, Syst Biol 2014)

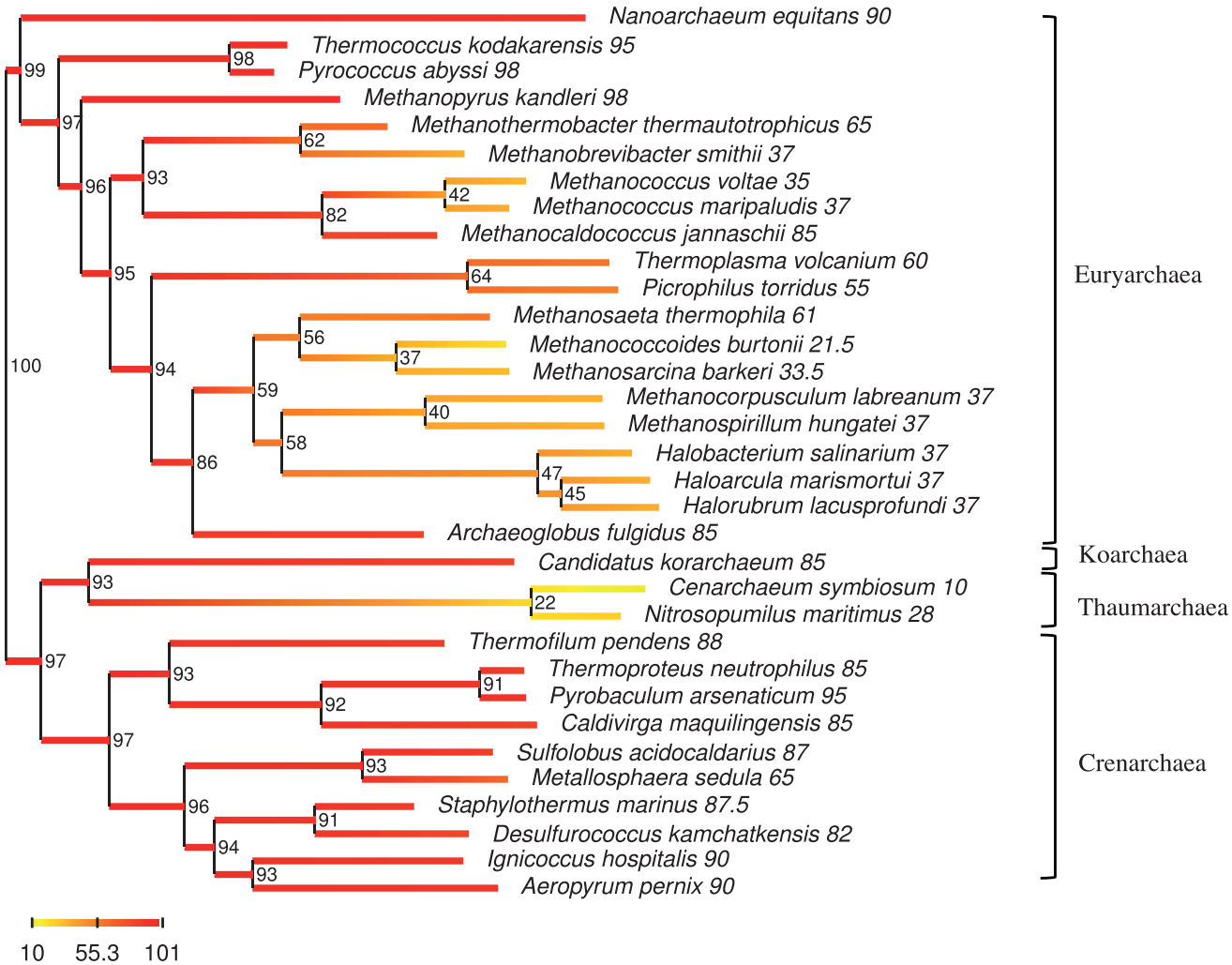
A phylogenetic Kalman filter



$$L_n(x_n) = \left[\int_{-\infty}^{\infty} p(x_n \to x_l \mid t_l) L_l(x_l)\, dx_l\right] \left[\int_{-\infty}^{\infty} p(x_n \to x_r \mid t_r) L_r(x_r)\, dx_r\right]$$

- analytical integrals: all factors are Gaussian, result is Gaussian

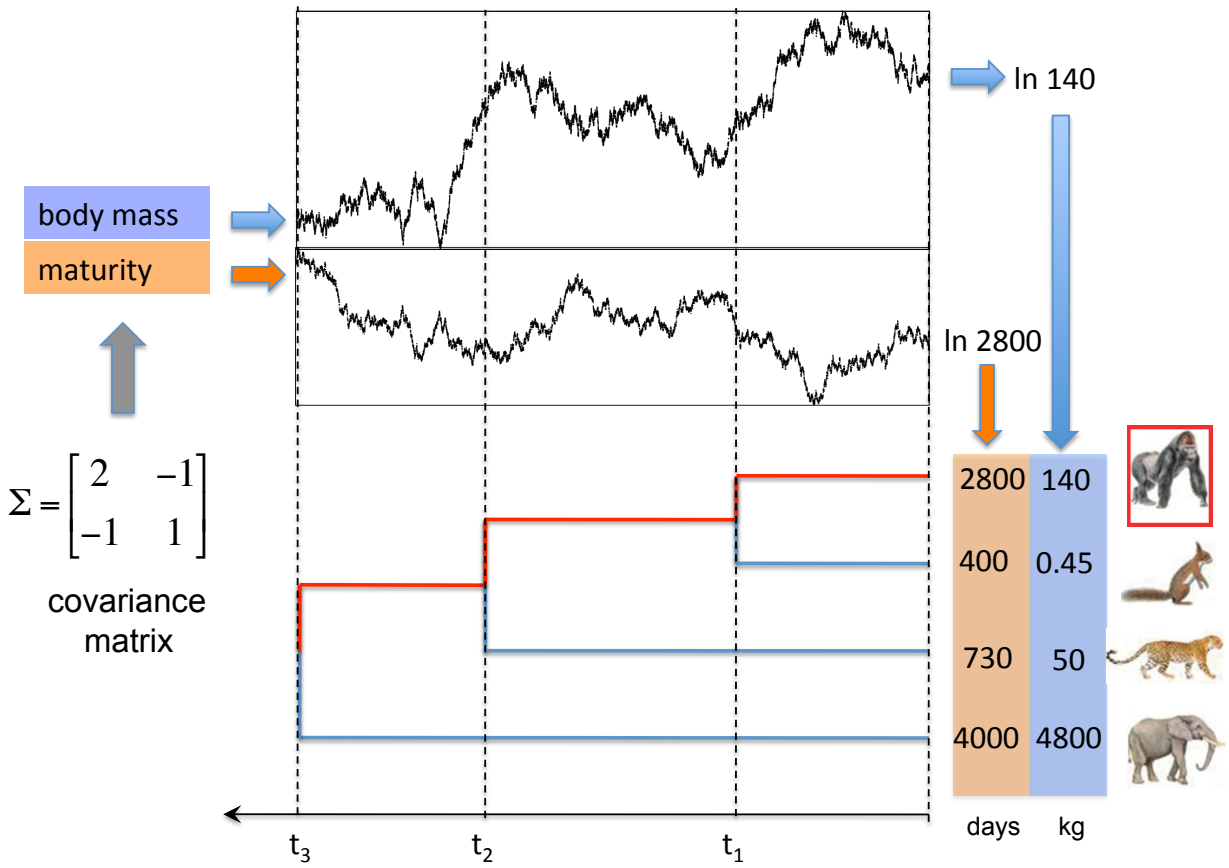
archaeal rRNA dataset



Inferred temperature for archaeal ancestor

- 95% credible interval: (91,110) Celsius.
- without molecular information: (60,96) Celsius

The comparative method – Summary

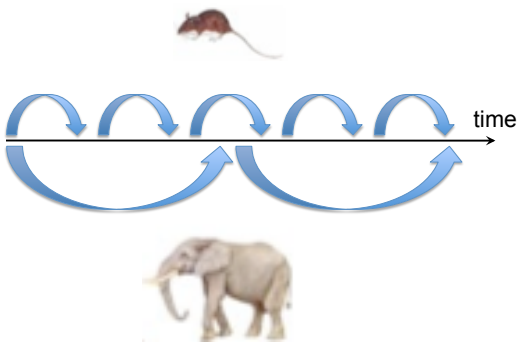


- independent contrast $\Longleftrightarrow$ traits follow bivariate Brownian motion
- more generally: statistical models of the evolutionary *processes*
- a large variety of questions: bursts, trends, jumps, correlations.

Variation of the substitution rate among lineages

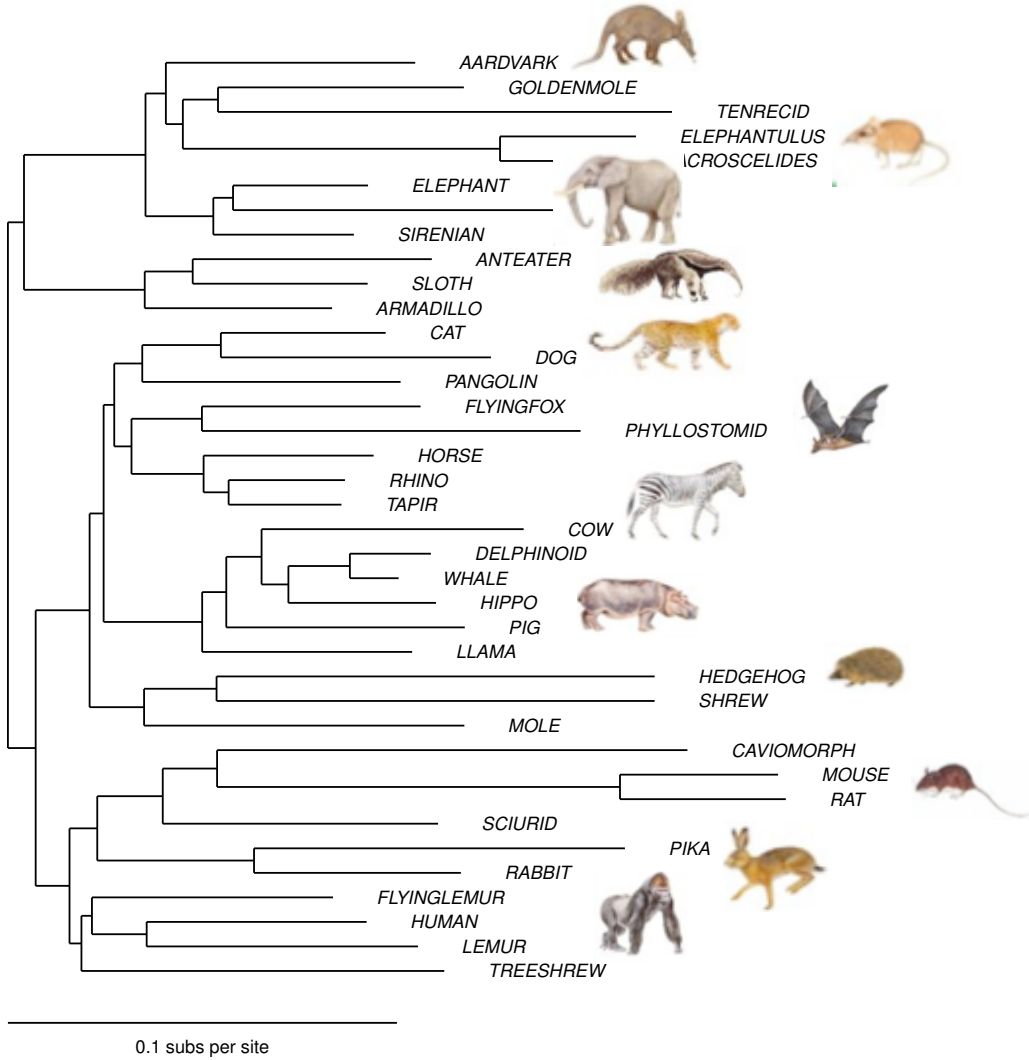
Possible causes

- generation-time effect

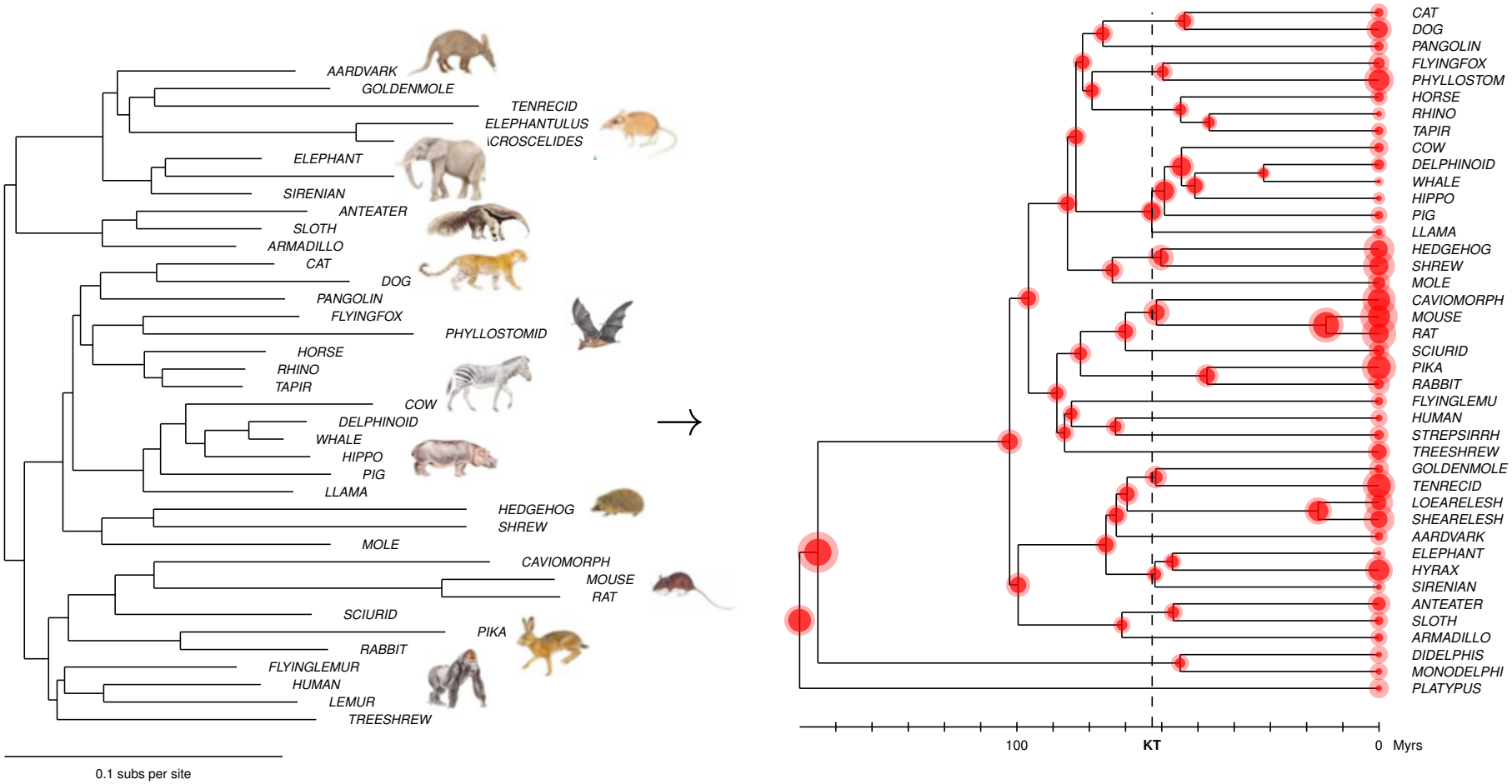


- metabolic rate effects
- selection for longevity

(reviewed in Lanfear et al, 2010)

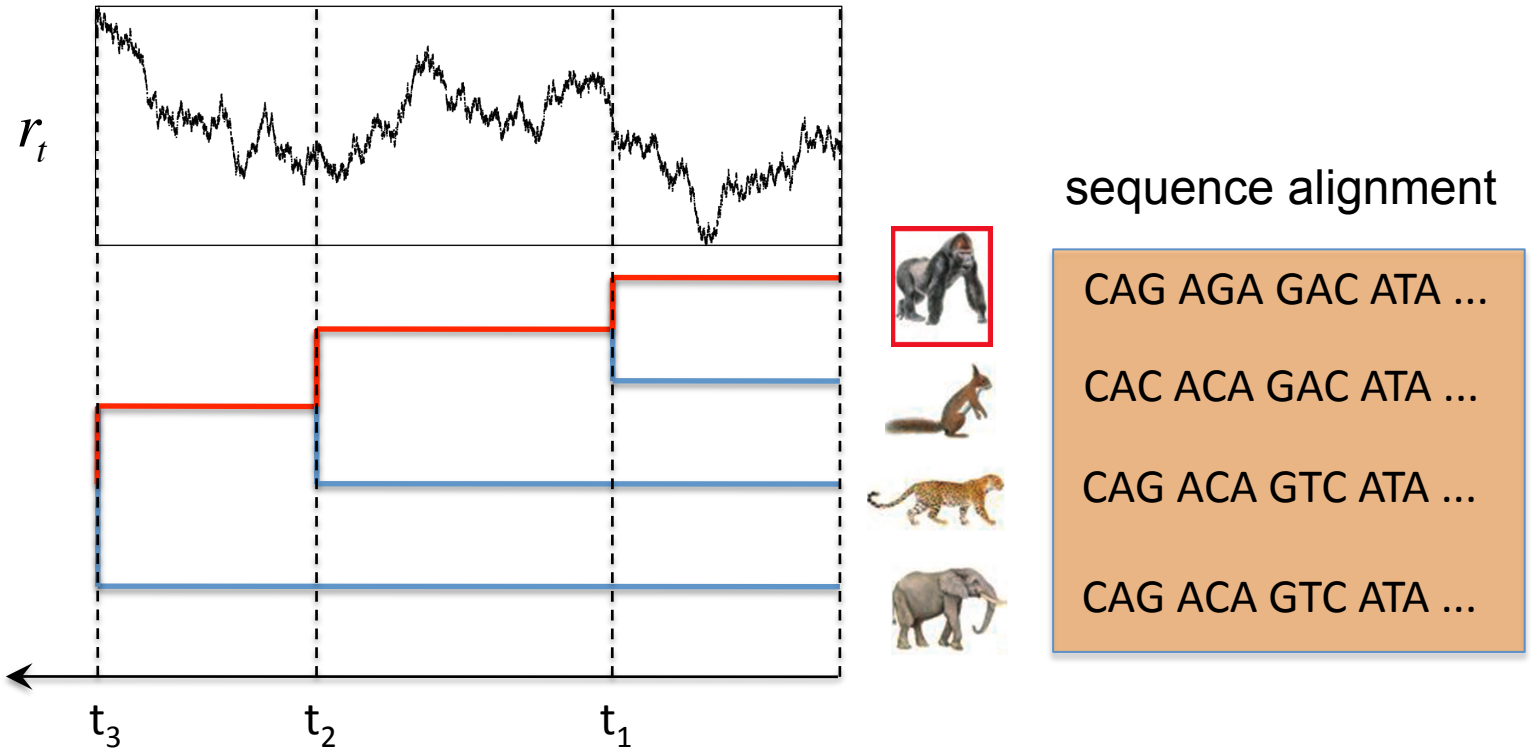


The relaxed molecular clock



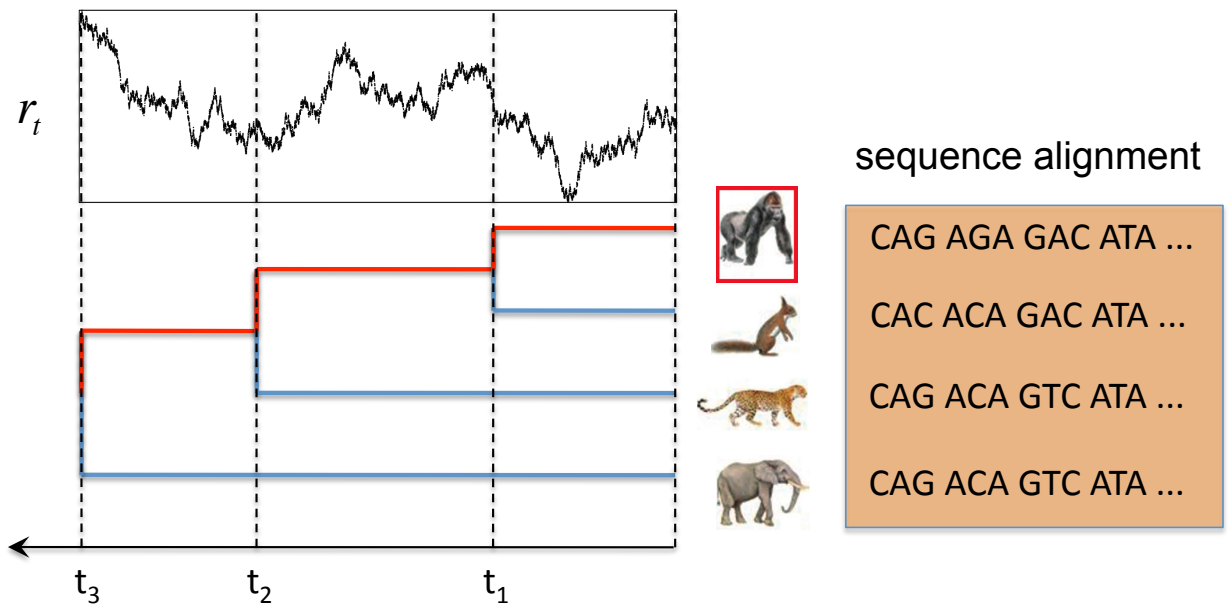
Branch lengths = times x rates

The Brownian relaxed molecular clock



- substitutions occur at rate r_t
- r_t modeled as Brownian motion along branches
- Brownian model induces rate *autocorrelation* across branches
- joint estimation of rates and times by Bayesian MCMC

Estimating divergence times: the relaxed clock model



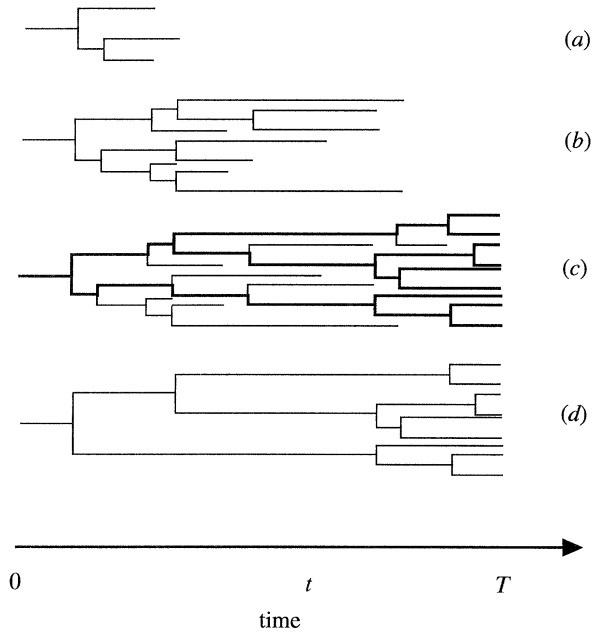
Data and constraints

- multiple alignment D (here, nuclear coding genes in mammals)
- tree topology T, and fossil calibrations Φ

Principle of the method

- build a hierarchical model, with a prior on its parameters
- sample from the posterior distribution using MCMC algorithms

Diversification process

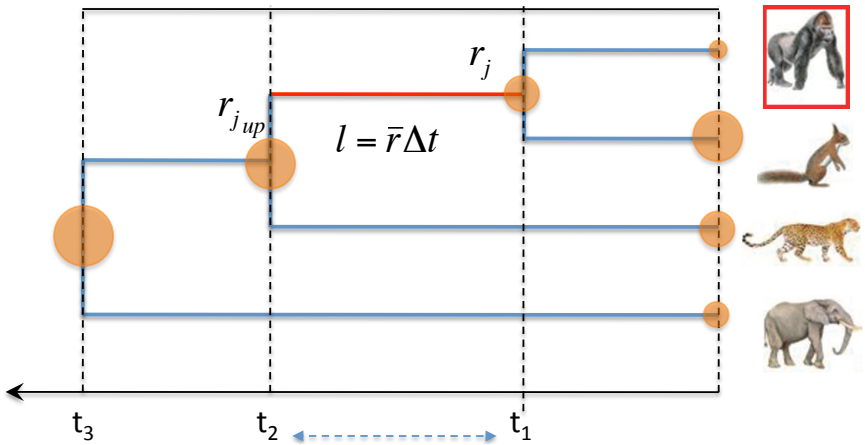


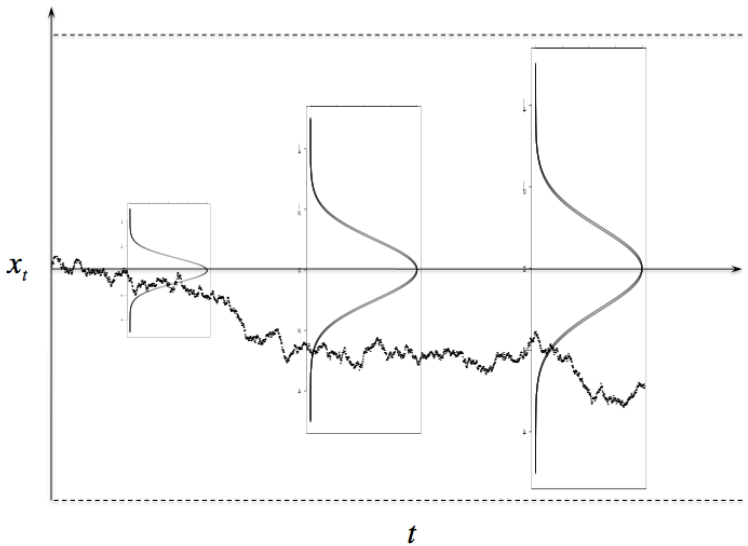
Nee et al, 1994

birth-death with subsampling

- speciation rate λ, extinction rate μ, sampling fraction ρ
- t: vector of divergence times
- gives you a probability distribution on times: $p(t \mid \lambda, \mu, \rho)$

Brownian process





$$x_t = \ln r_t$$
$$x_t \sim N(x_0, \sigma^2 t)$$

gives you a probability distribution on rates: $p(r \mid t, \sigma^2)$

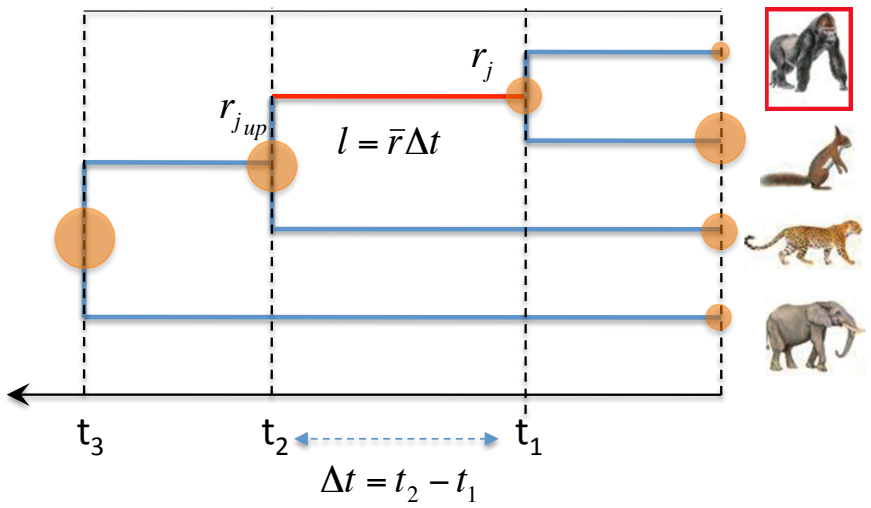
Model of sequence evolution by point substitutions

Substitution rate matrix Q (4 x 4)

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

- κ: transition-transversion ratio
- γ: equilibrium GC (GC^*)

Substitution process (rate matrix Q)



sequence alignment

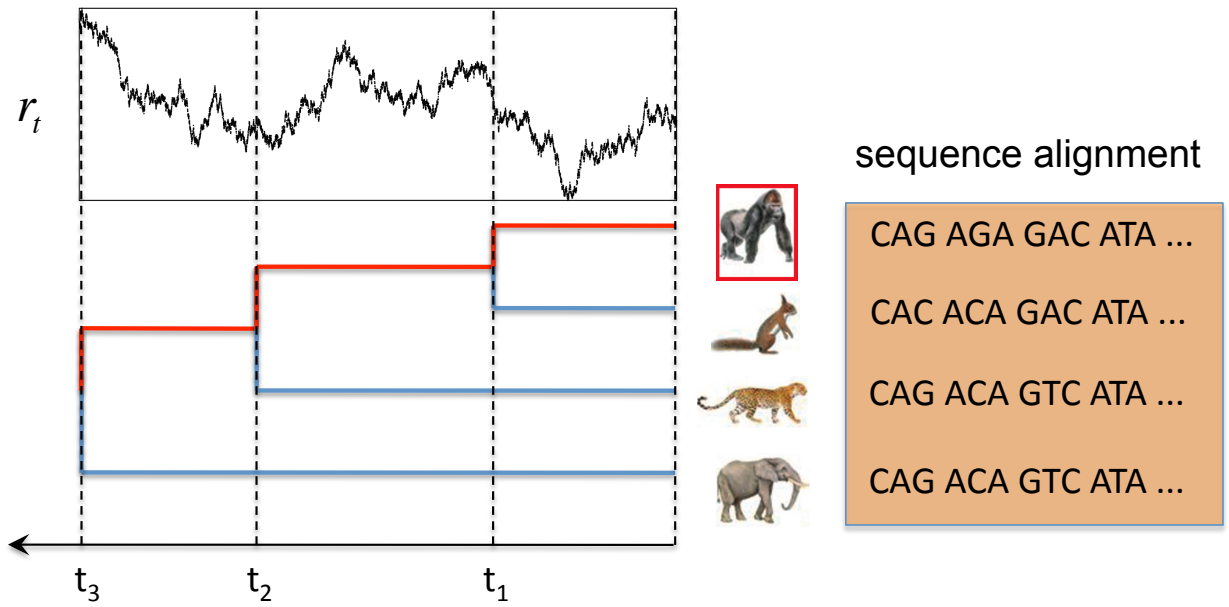
CAG AGA GAC ATA ...
CAC ACA GAC ATA ...
CAG ACA GTC ATA ...
CAG ACA GTC ATA ...

- branch length approximated as:

$$l = \bar{r}\Delta t, \quad \text{where} \quad \bar{r} = \frac{r_j + r_{j_{up}}}{2}$$

- length l: expected number of point substitutions along the branch
- gives you a probability distribution on sequences: $p(D \mid r, t, Q)$

Complete model

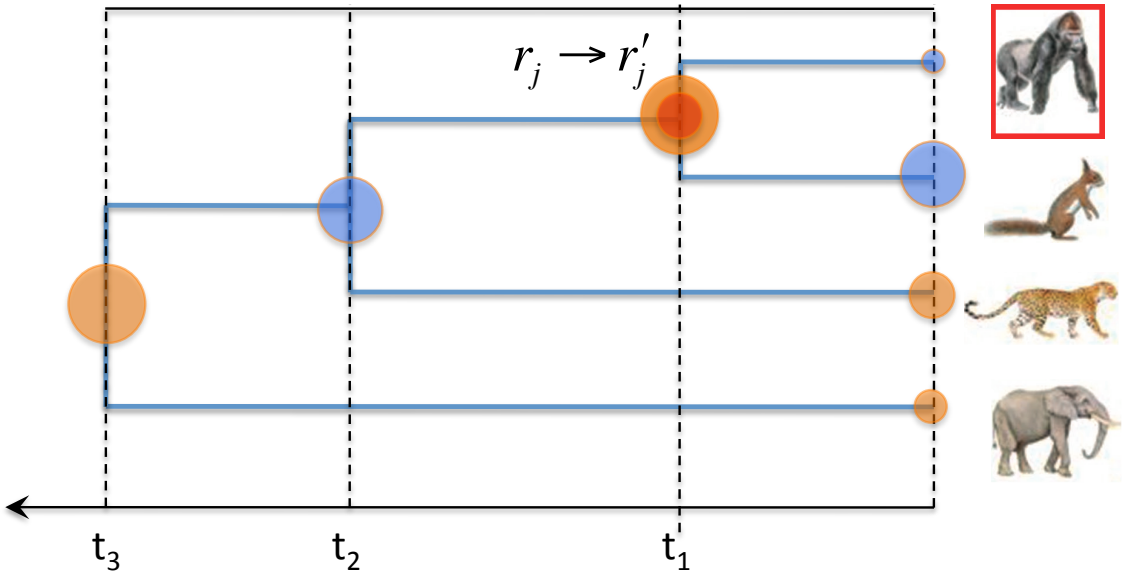


- diversification process (e.g. birth-death, parameters λ, μ, ρ)
- substitution rate: Brownian log-normal process (variance σ^2)
- substitution process (4x4 substitution matrix Q)
- complete model configuration: $\theta = (\lambda, \mu, \rho, \sigma, t, r)$

posterior distribution proportional to joint probability:

$$p(\lambda)p(\mu)p(\rho)p(\sigma^2) \quad p(t \mid \lambda, \mu, \rho) \quad p(r \mid t, \sigma^2) \quad p(D \mid r, t, Q)$$

Bayesian inference and Monte Carlo sampling



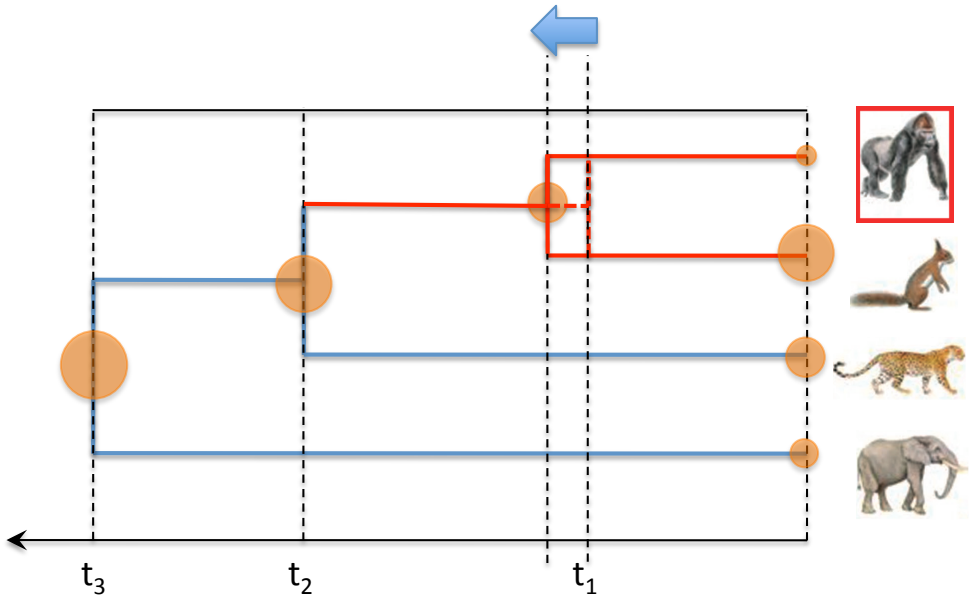
Metropolis Hastings on rates

$$\alpha = \frac{p(D \mid r', t, Q)\, p(r' \mid t, \sigma^2)}{p(D \mid r, t, Q)\, p(r \mid t, \sigma^2)}$$

$\alpha > 1$: accept move

$\alpha < 1$: accept with prob. α

Bayesian inference and Monte Carlo sampling



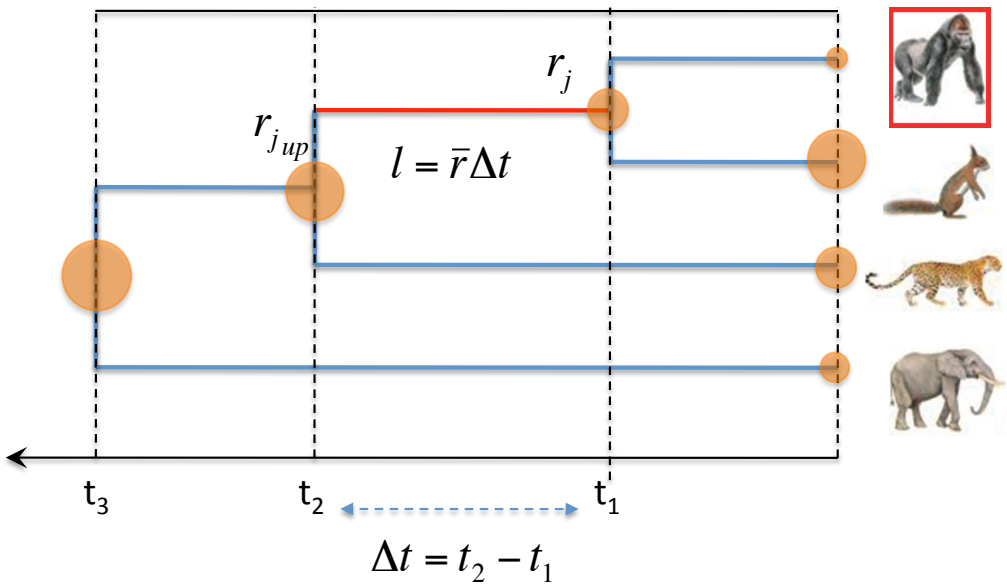
Metropolis Hastings on divergence times

$$\alpha = \frac{p(D \mid r, t', Q)\, p(r \mid t', \sigma^2)\, p(t' \mid \lambda, \mu, \rho)}{p(D \mid r, t, Q)\, p(r \mid t, \sigma^2)\, p(t \mid \lambda, \mu, \rho)}$$

$\alpha > 1$: accept move

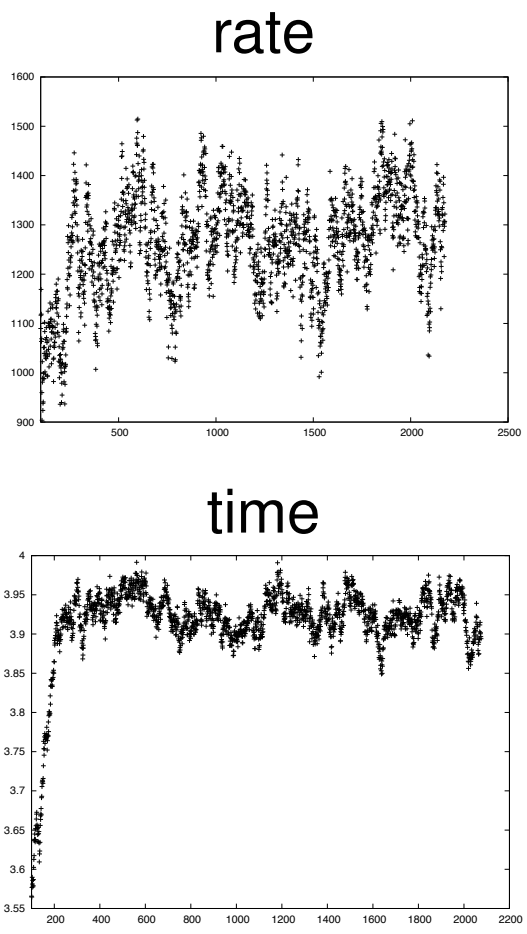
$\alpha < 1$: accept with prob. α

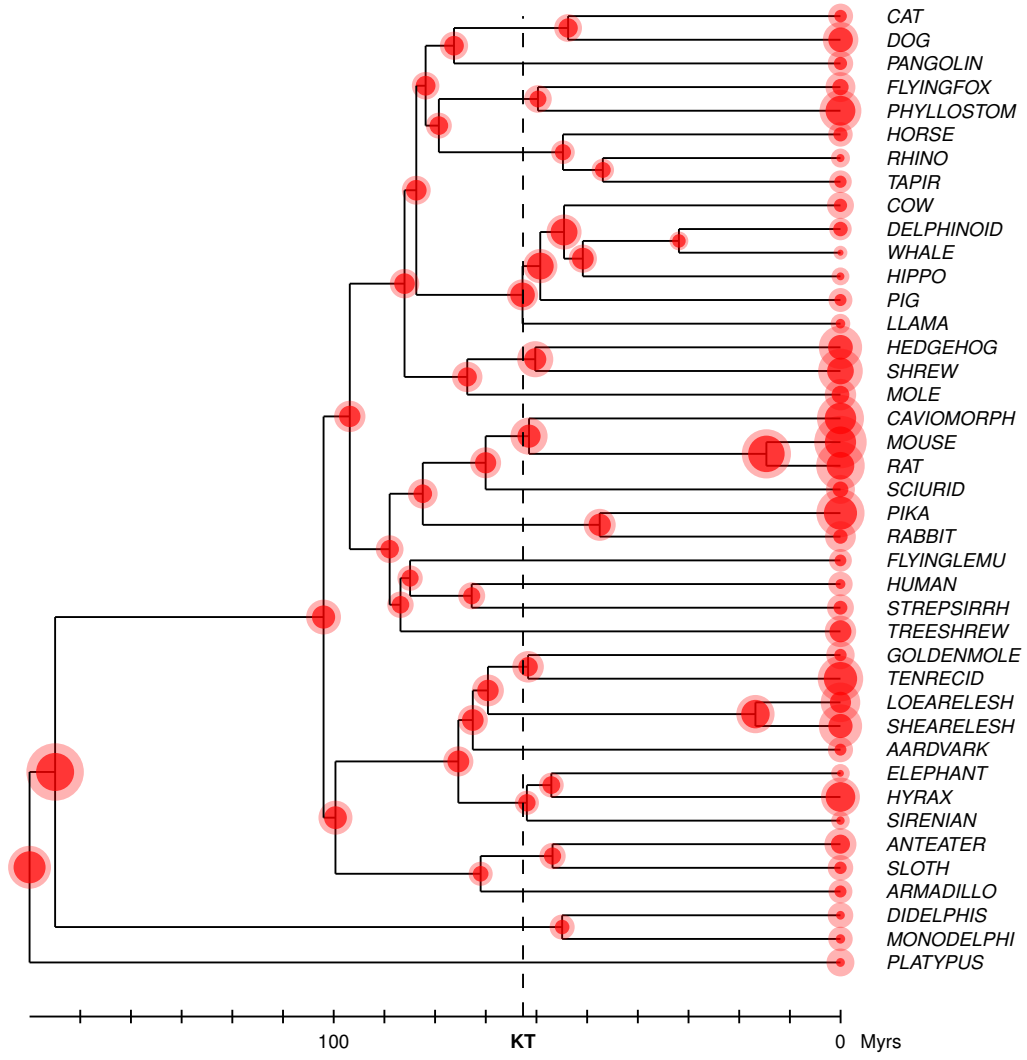
Bayesian inference and Monte Carlo sampling



Metropolis Hastings on σ^2, λ, μ, ρ

Posterior mean times and rates

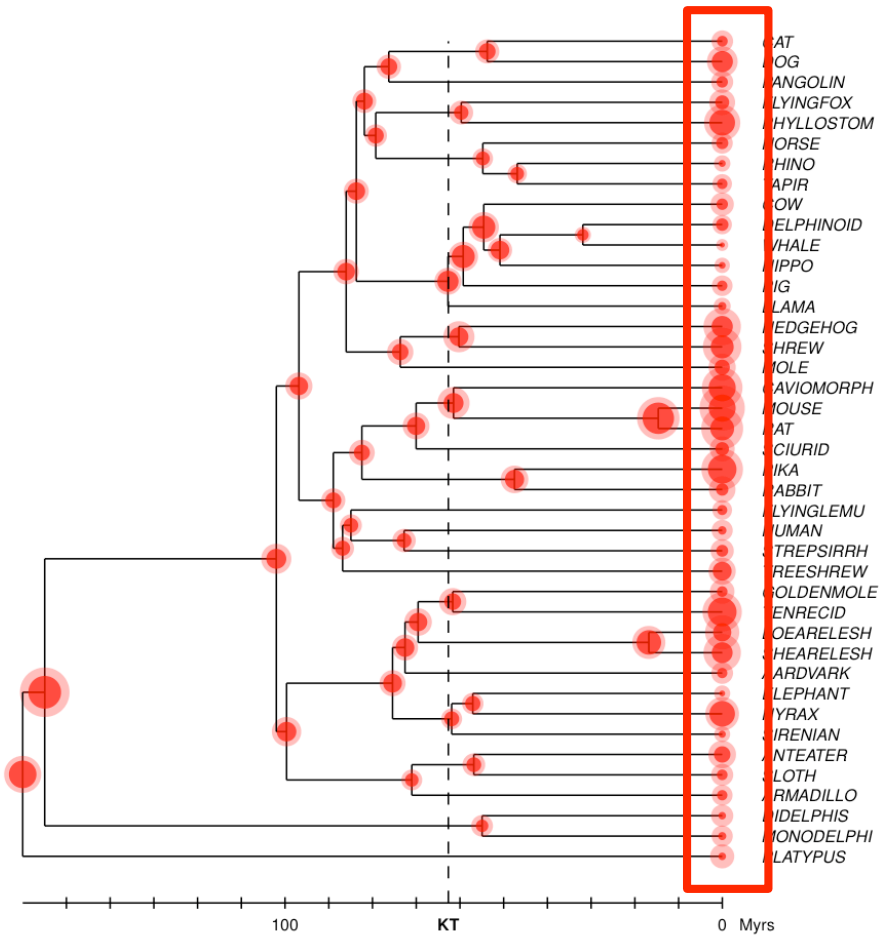


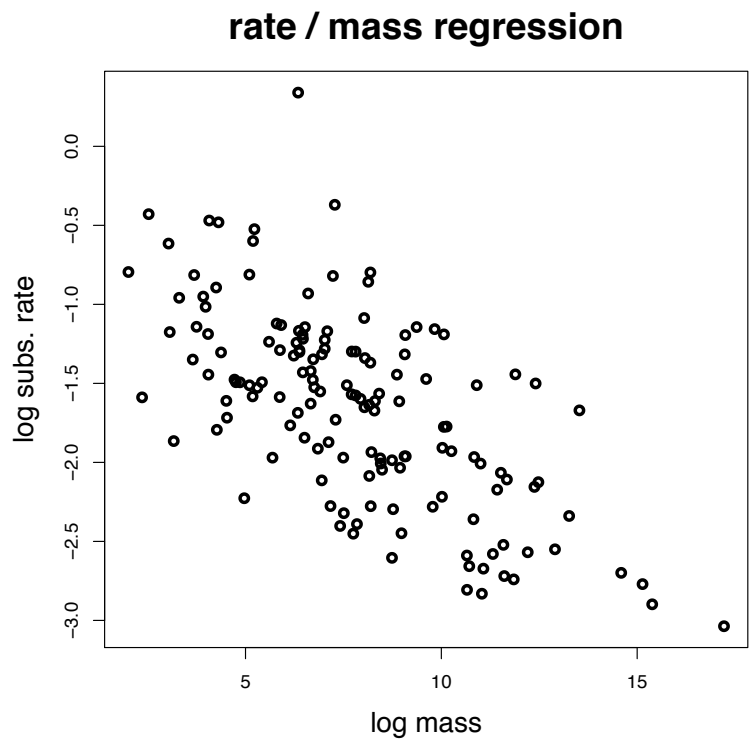


(Thorne et al 1998, Lepage et al 2007, Rannala and Yang 2007)

Multidivtime, PhyloBayes, MCMCtree, Beast

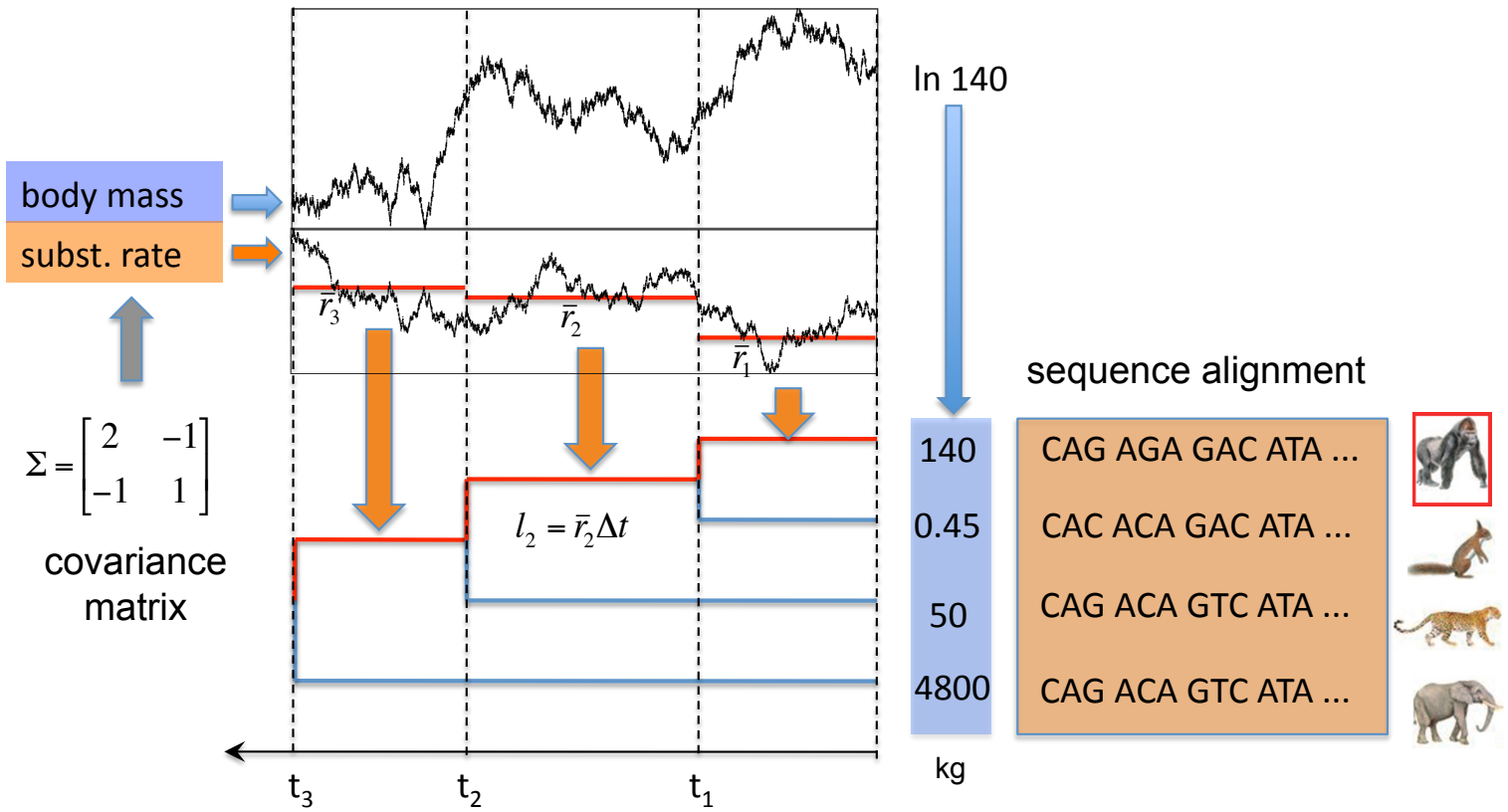
Correlating rates and traits





- sequential method: error propagation problems
- circularity in the way phylogenetic inertia is dealt with
- suggests a more direct *integrative* approach

Coupling trait evolution and substitution process

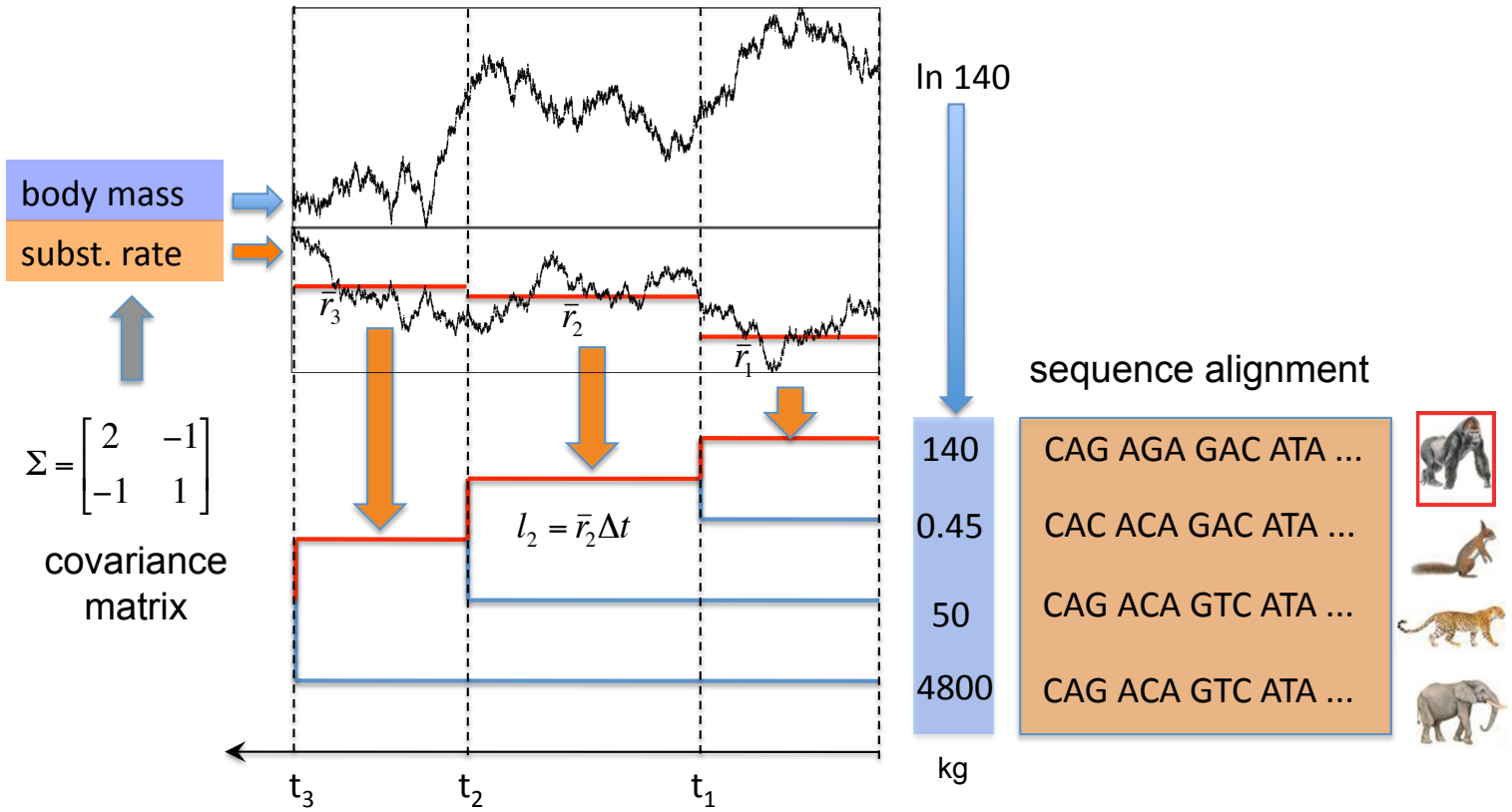


Lartillot and Poujol, 2011, Mol Biol Evol, 28:729

Hierarchial Bayesian model (parameter estimation by MCMC)

- diversification process *t* (birth-death, parameters λ, μ, ρ)
- Brownian multivariate process *X* (covariance matrix Σ)
- time-dependent codon model *Q*

Coupling substitution process with life-history evolution



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

posterior proportional to joint probability:

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

Generalization

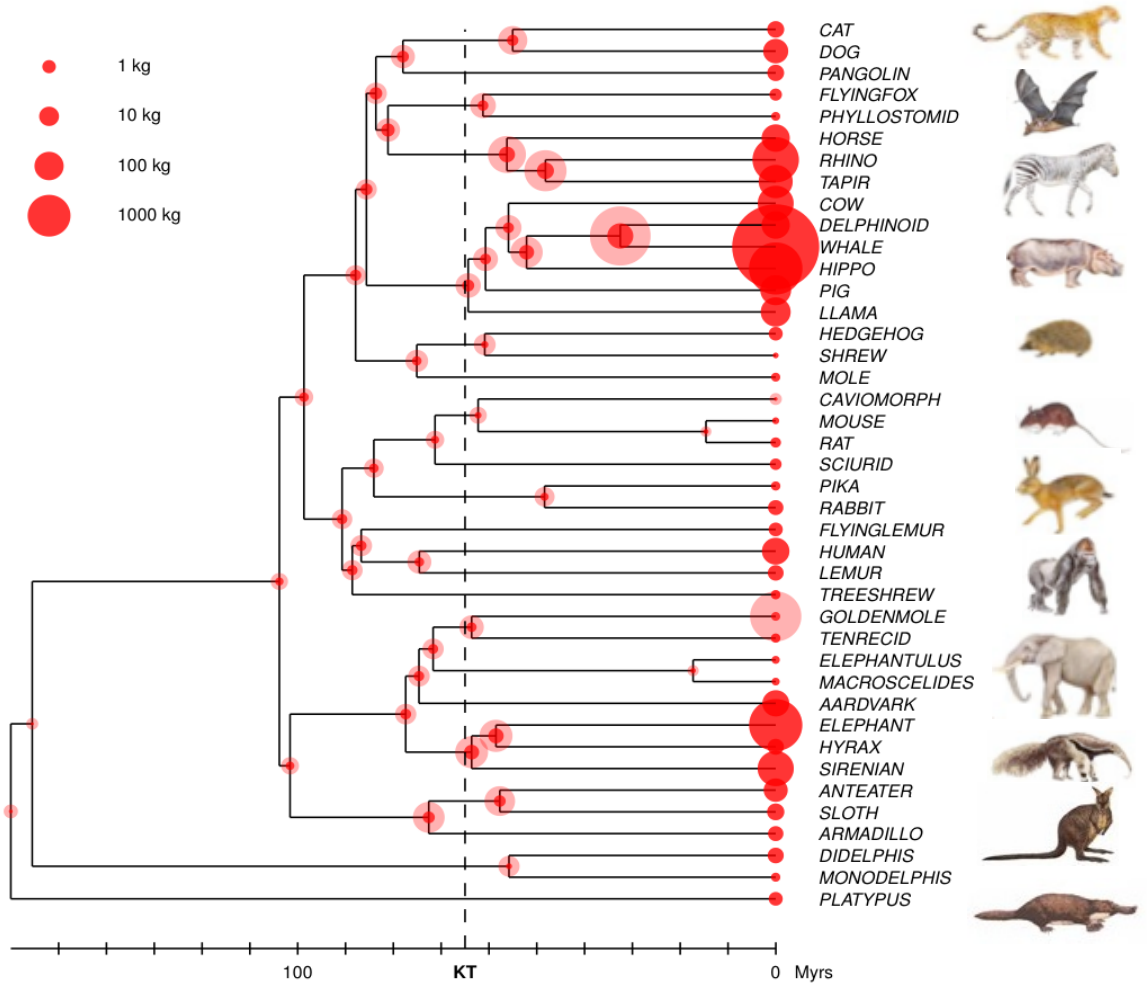
time-dependent substitution parameters

- rate of synonymous substitution (r)
- non-synonymous / synonymous ratio (ω)
- equilibrium GC composition (γ)

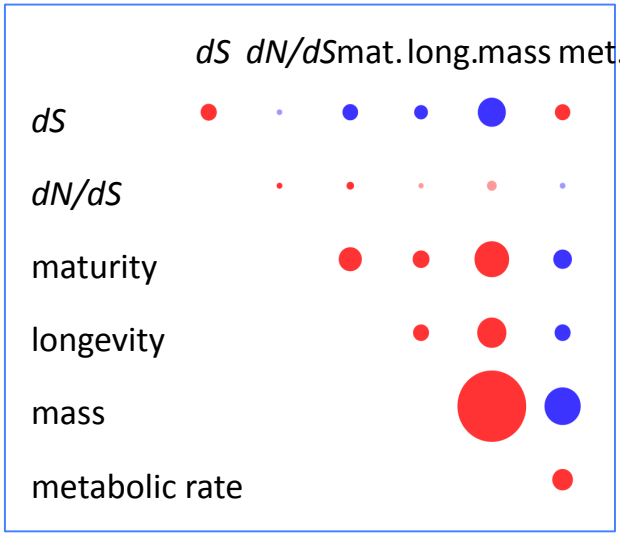
time-dependent quantitative traits

- sexual maturity (proxy of generation time)
- adult body mass
- maximum recorded lifespan (proxy of longevity)
- metabolic rate
- genome size
- karyotypic number (number of chromosomes $2n$)

Joint inference of rates, dates and traits



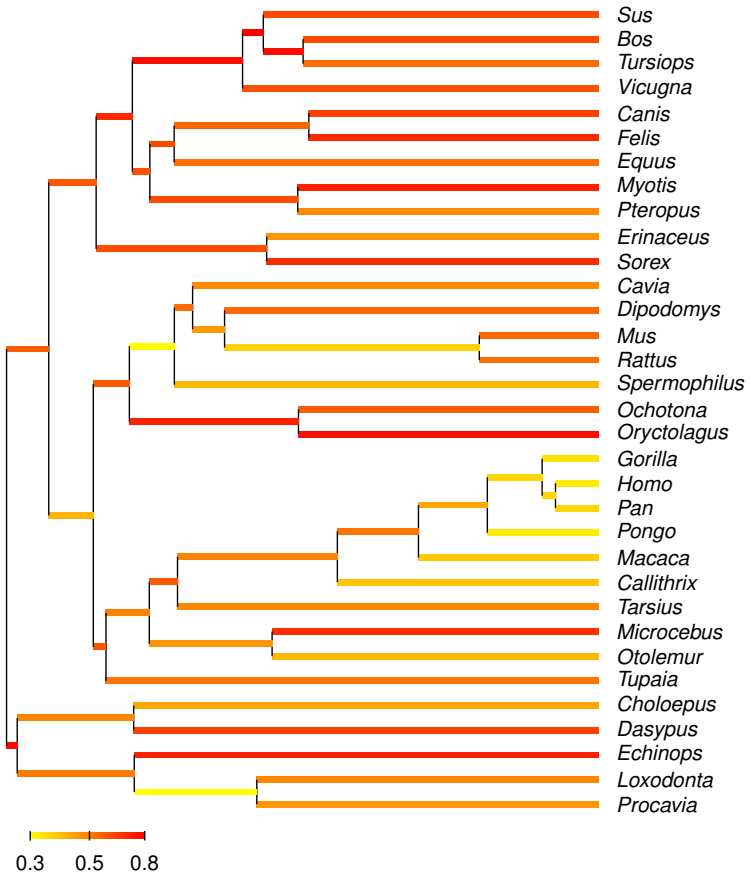
Alignment of 13 genes
4800 coding positions

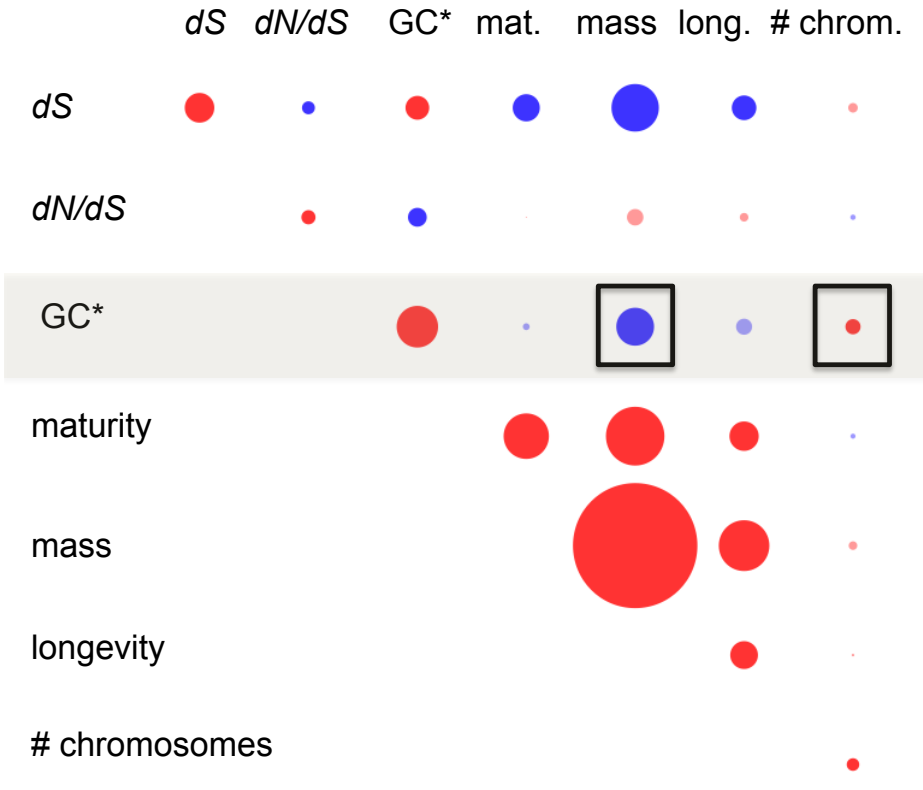


red: positive correlation
blue: negative correlation

Lartillot and Delsuc, 2012, Evolution 66:1773

Equilibrium GC (GC^*) in nuclear genomes



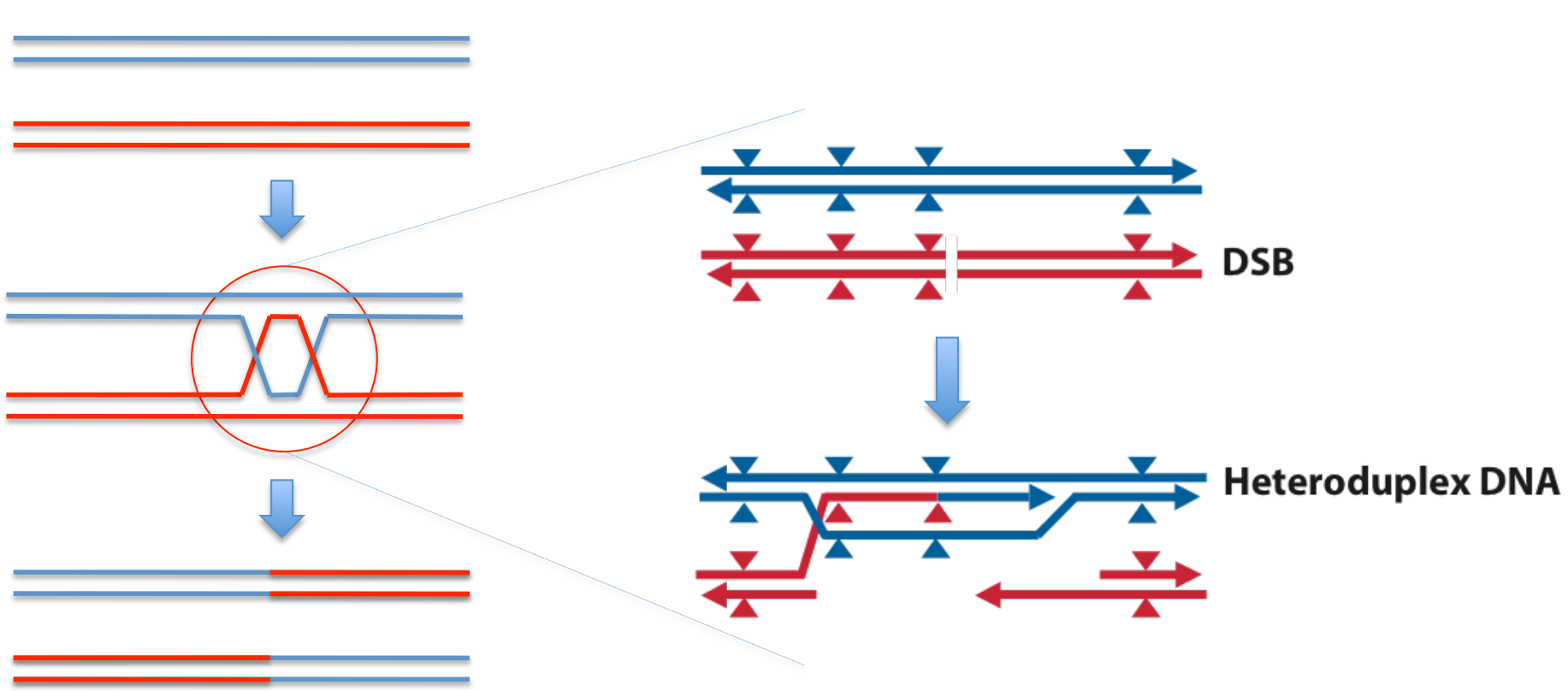


red: positive
blue: negative
light shade: not significant

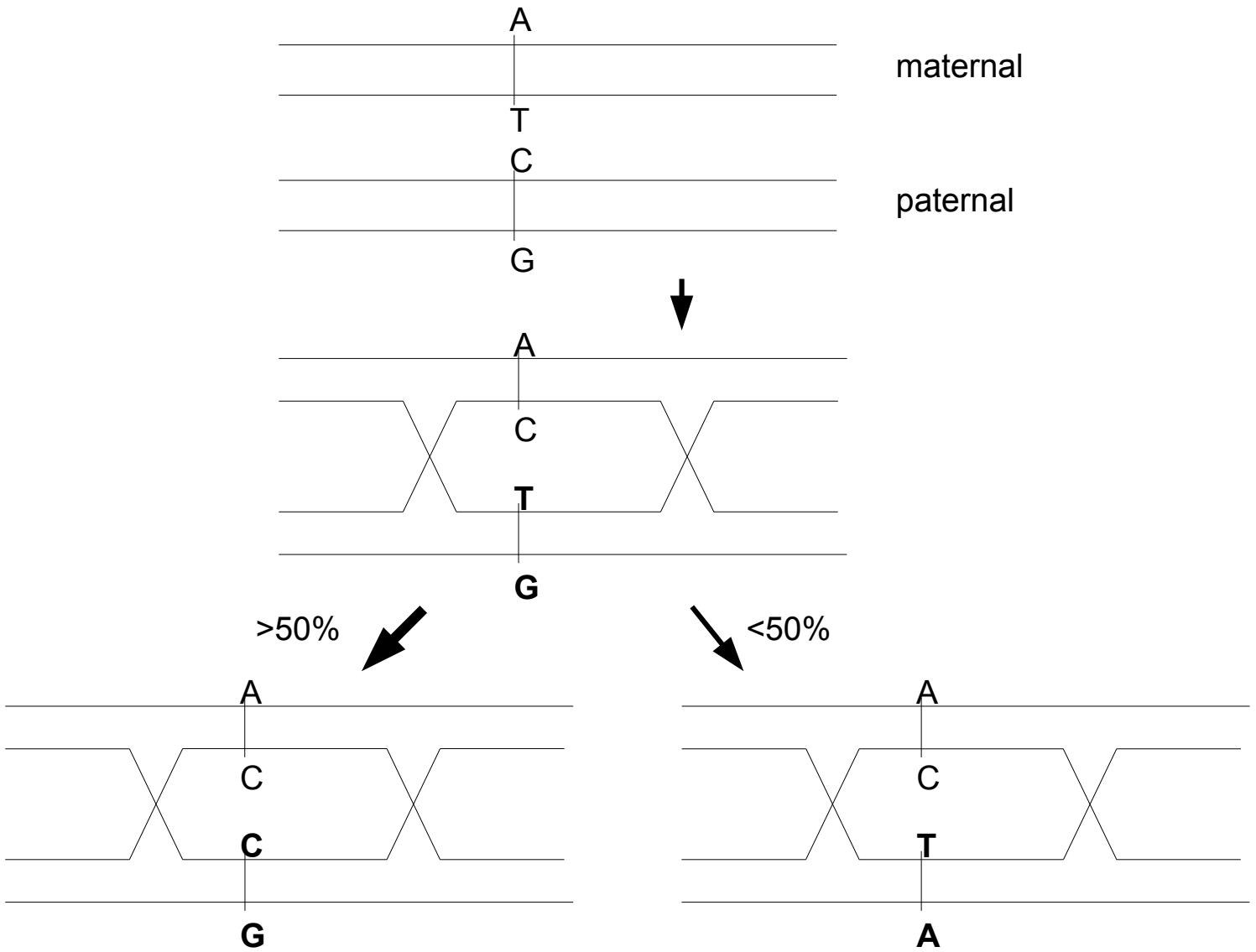
Lartillot, 2013, Molecular Biology and Evolution, 30:356

- negative correlation between GC^* and body size
- positive correlation between GC^* and number of chromosomes

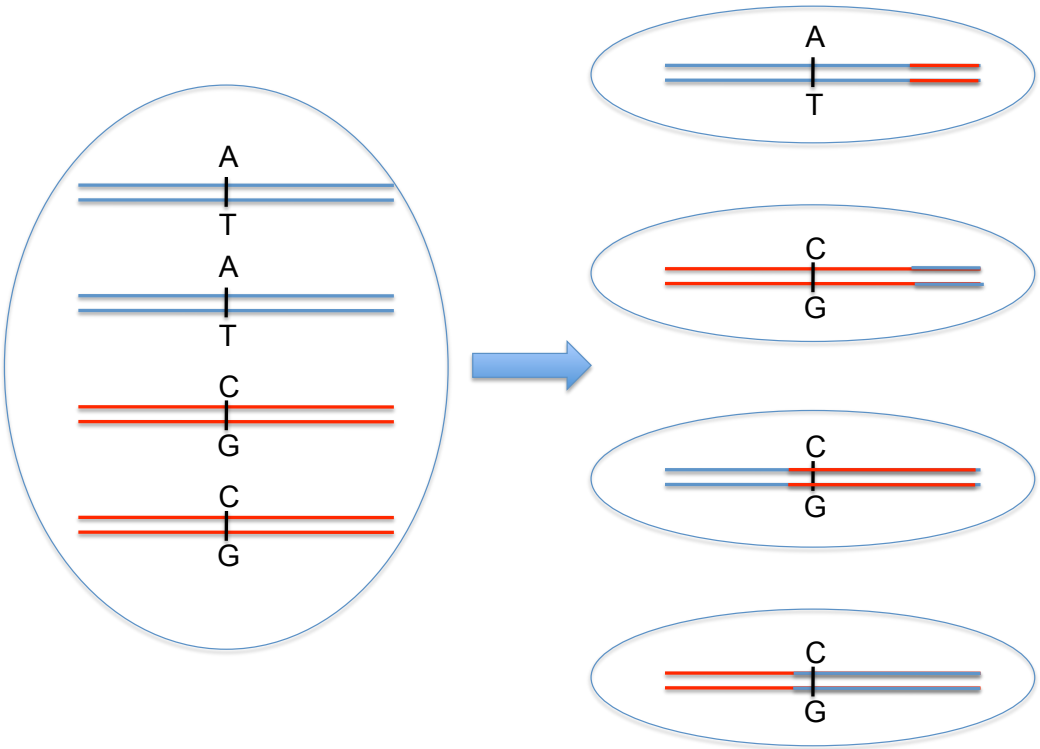
Biased conversion during meiosis



Biased conversion during meiosis



Biased gene conversion (BGC) during meiosis



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

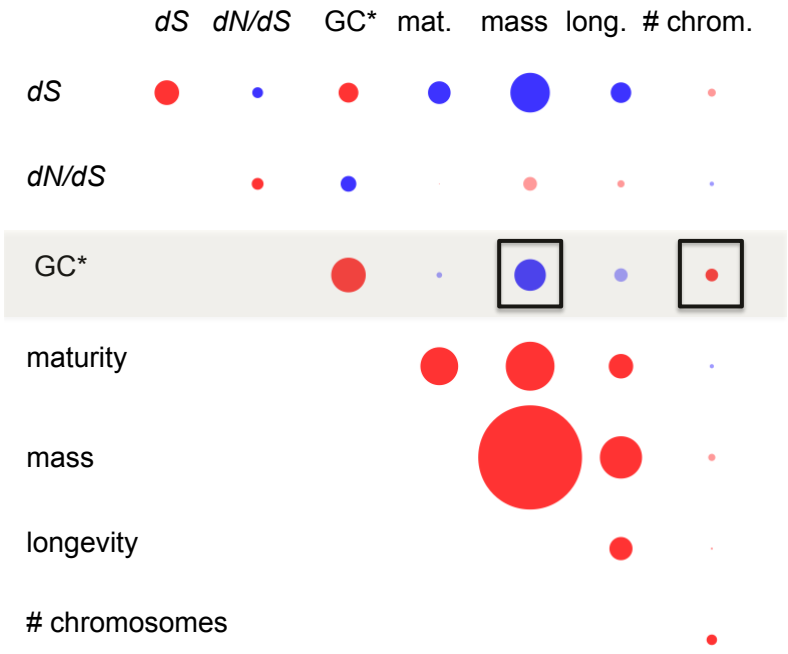
$$x_{AT} = \frac{1-b}{2}$$

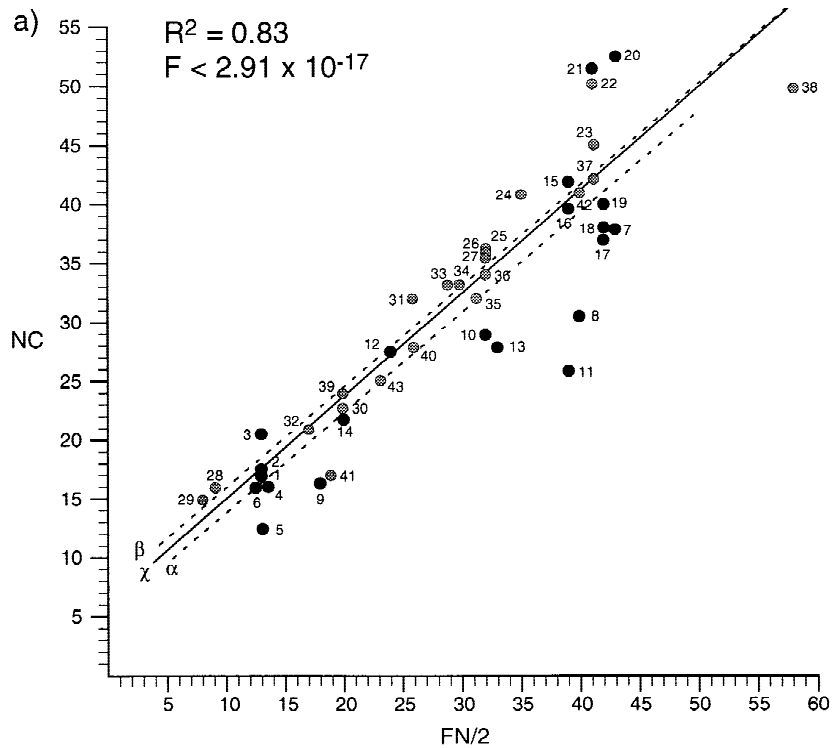
GC overtransmission

- 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**



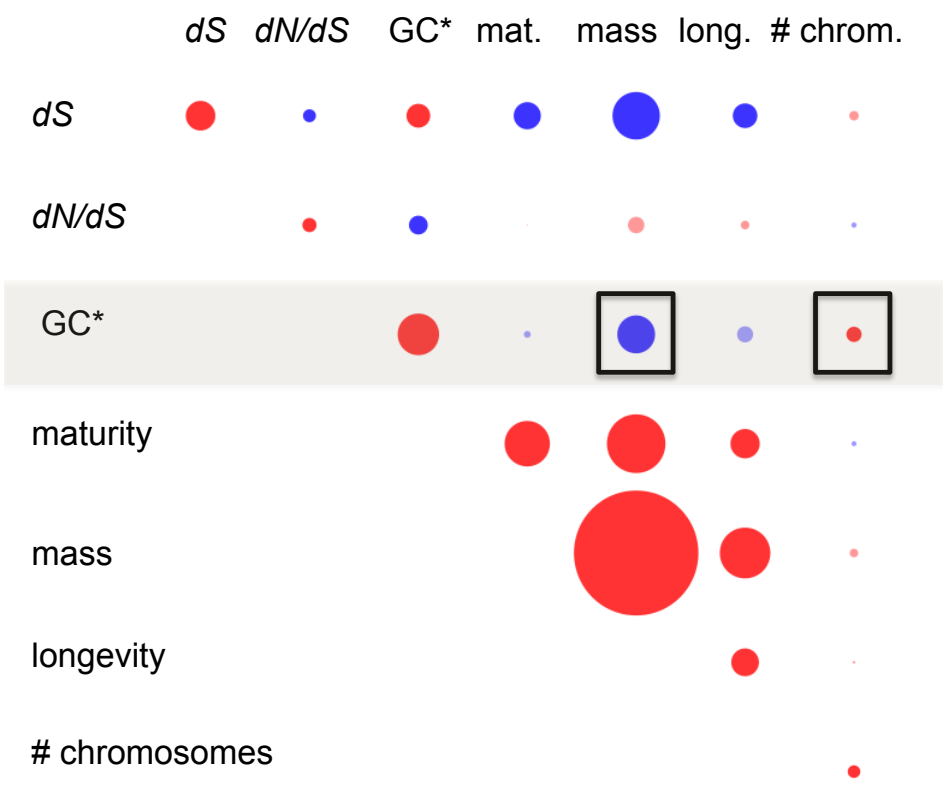


de Villena and Sapienza, 2001, Mamm Genome 12:318

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**



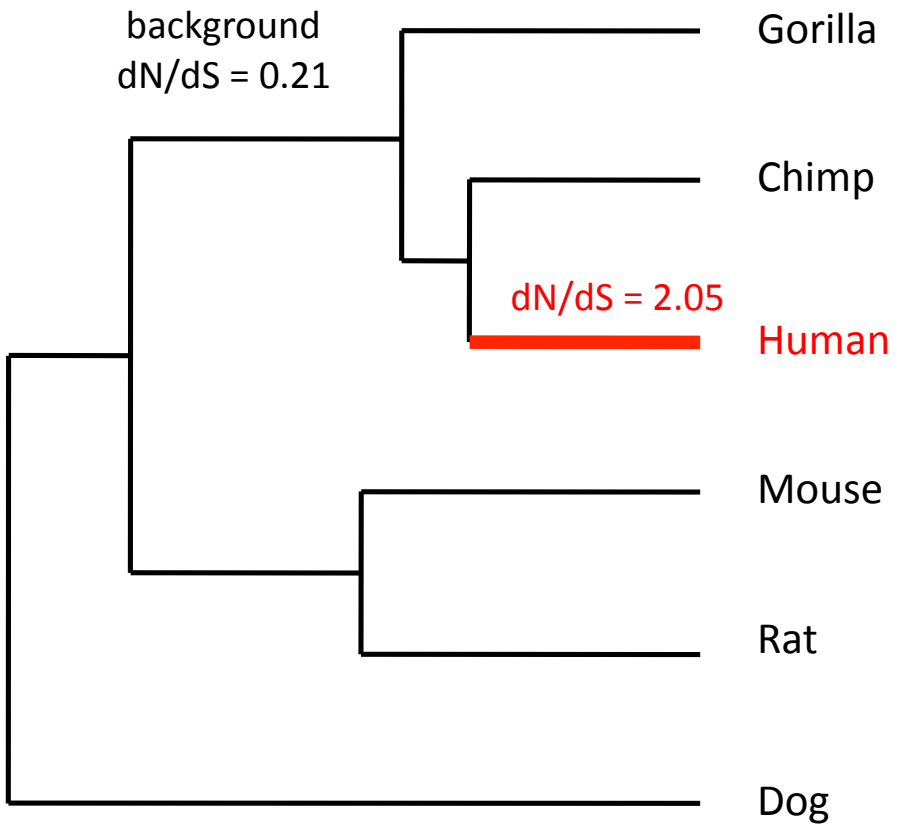
red: positive
blue: negative
light shade: not significant

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

GC-biased gene conversion interferes with selection

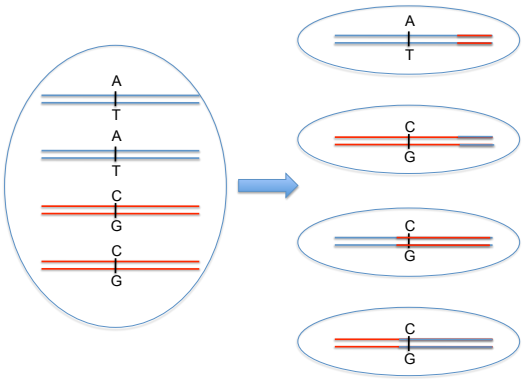


20 substitutions
3 synonymous
17 non-synonymous
all AT ➔ GC

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

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

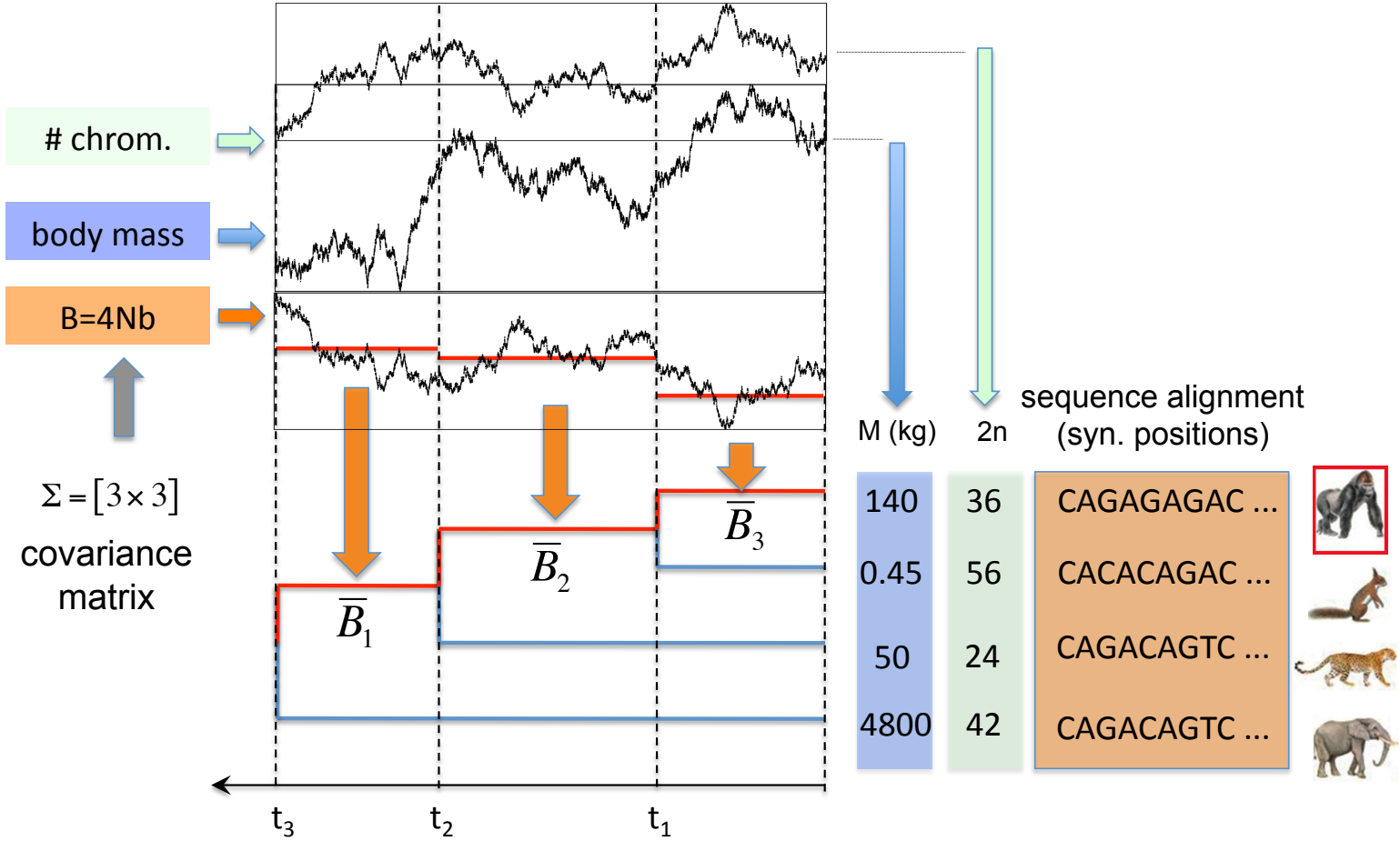
$B = 4N_eb$: scaled conversion coefficient

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

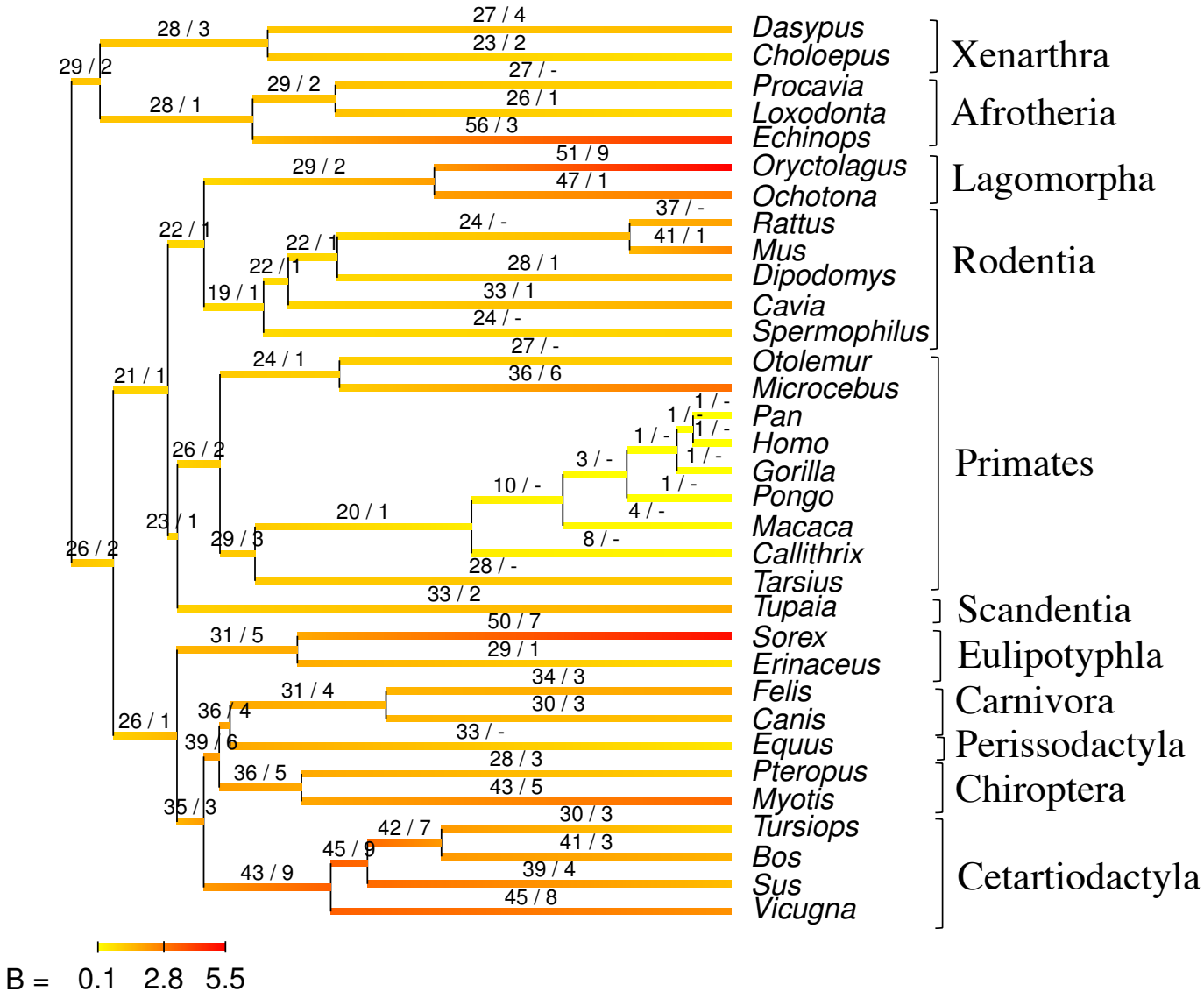
Exon-rich dataset

- 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

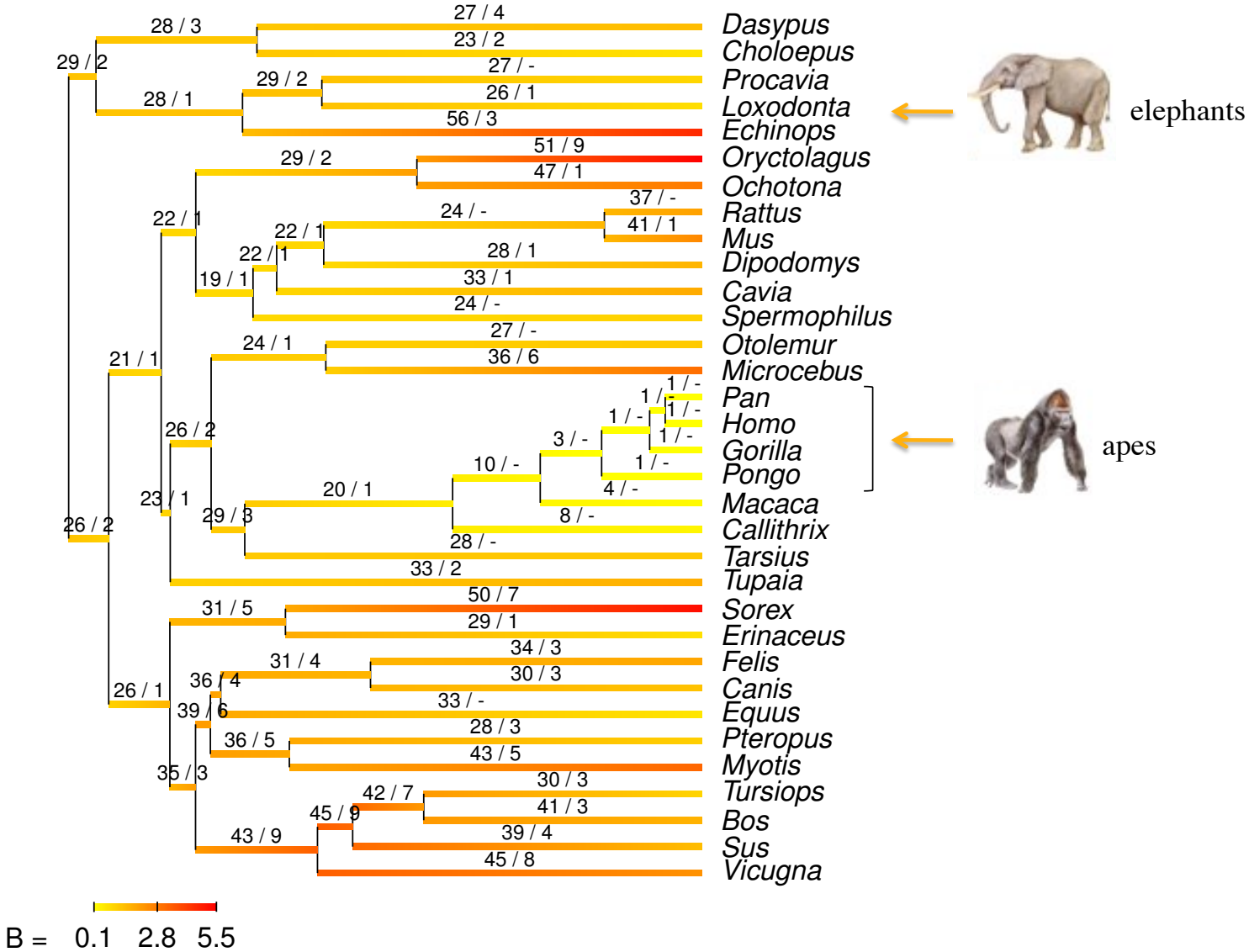
- 17 single-exon genes 73 mammals

Reconstructed history of $B = 4N_e b$



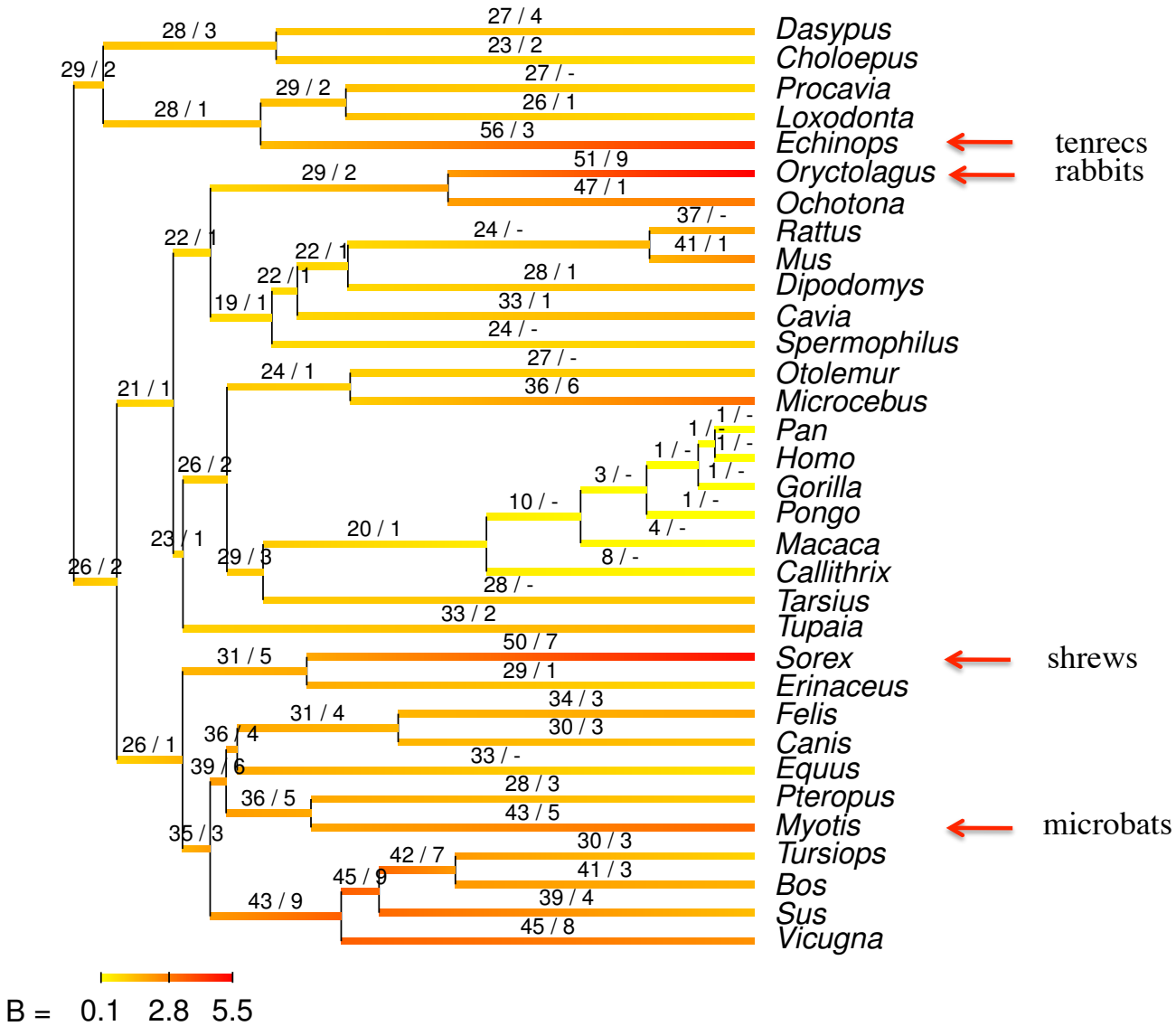
Lartillot, 2013, Molecular Biology and Evolution, 30:489

Phylogenetic history of population-genetic regimes

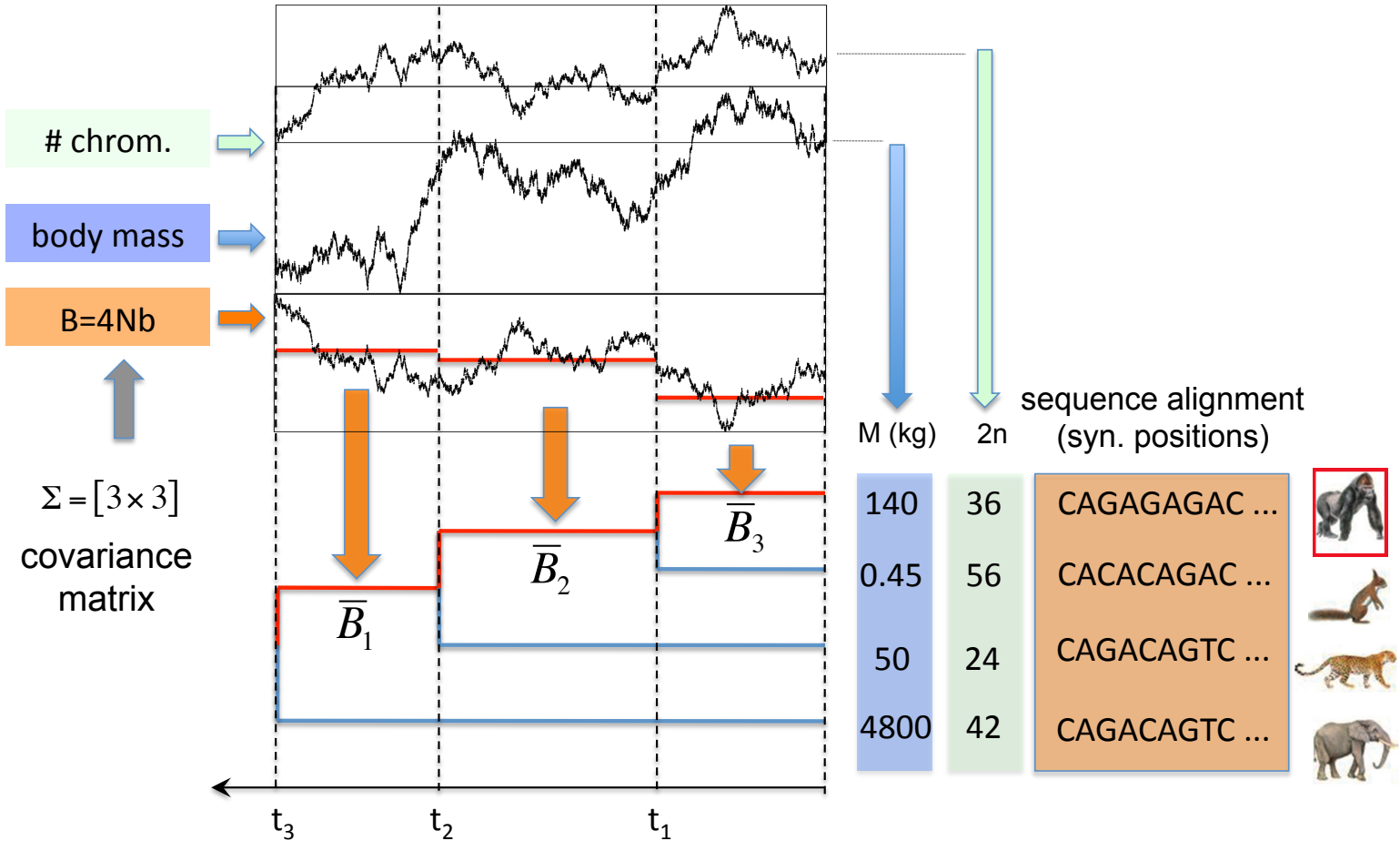


large mammals, small N_e: drift dominates ($B < 1$)

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 $B = 4N_e b$

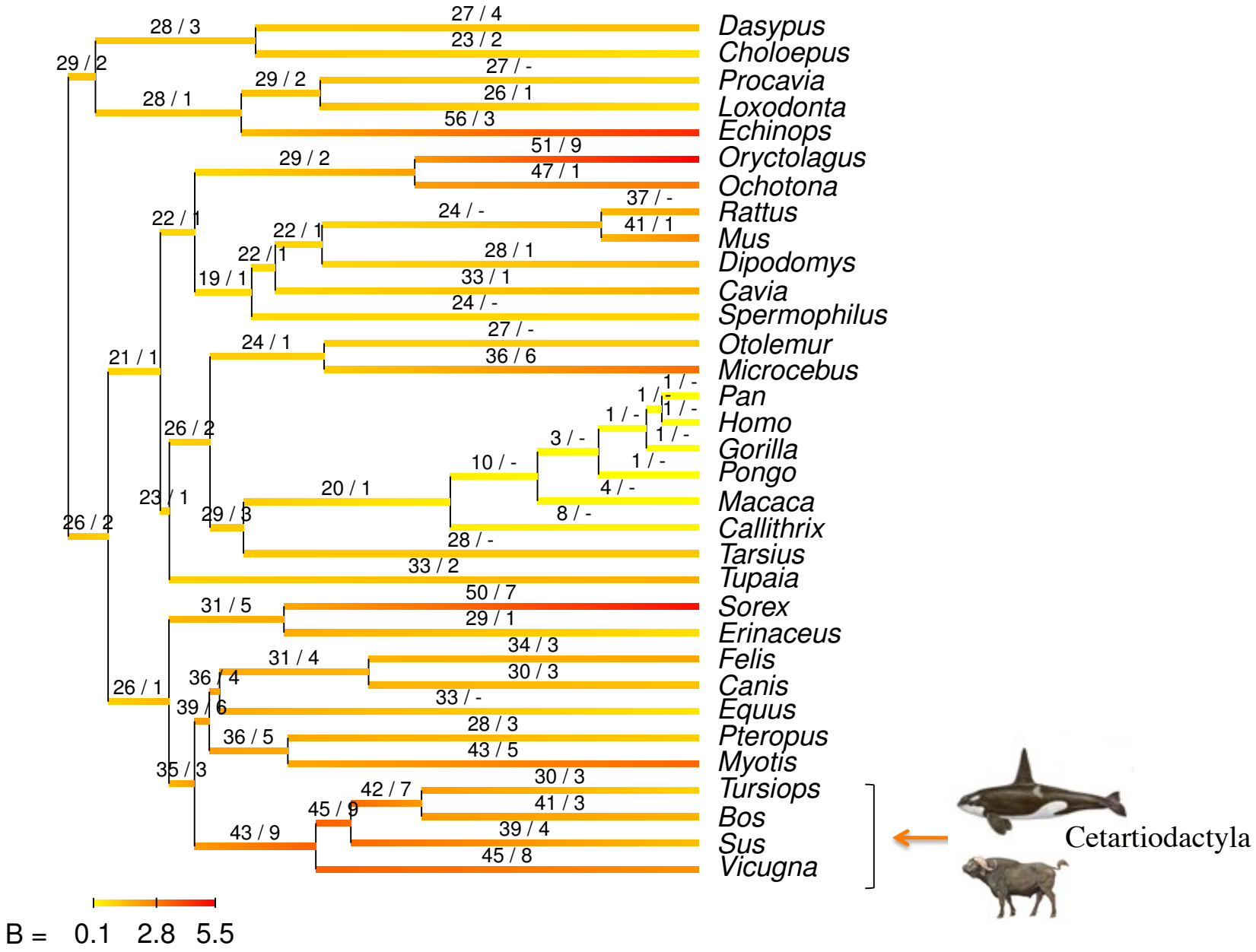
$$B \sim M^{\gamma} n^{\alpha}$$

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

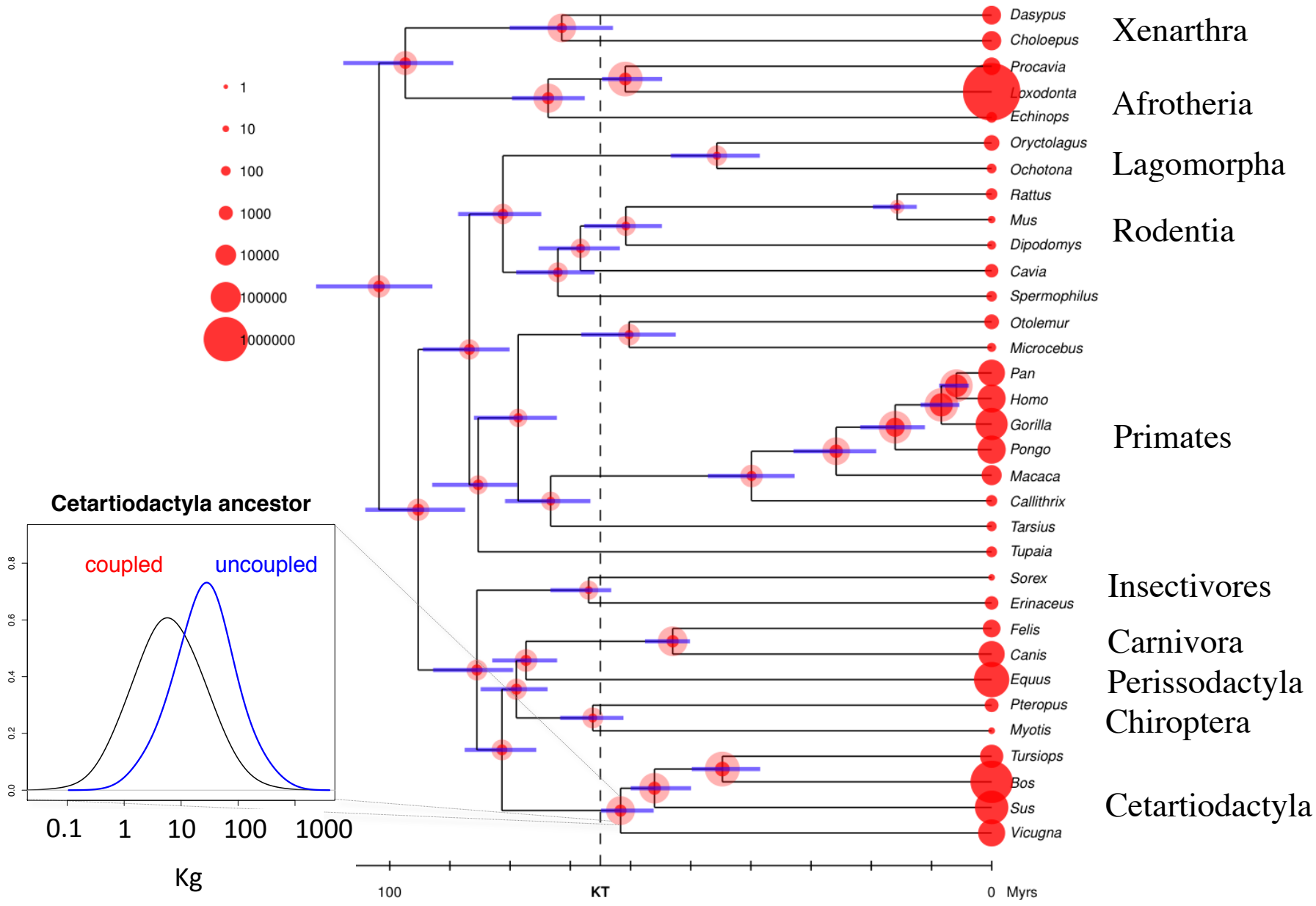
n: number of chromosomes (prediction: $\alpha = 1 > 0$)

	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

Molecular dating and diversification studies

Diversification studies: current approach

- likelihood (time-calibrated tree T, diversification parameters θ):

$$L(\theta) \quad = \quad p(T \mid \theta)$$

Diversification studies: integrative approach

- use diversification model as your prior on divergence times:

$$p(D \mid r, T)\, p(r \mid T)\, p(T \mid \theta)\, p(\theta)$$

- compare models based on their marginal likelihoods
- avoids circularity and overconfidence
- dating and diversification: two sides of a same coin

Integrative models for macroevolutionary studies

Toward a unified probabilistic framework for

- reconstructing divergence times
- fitting / testing diversification models
- fitting / testing models of trait evolution
- understanding driving forces of molecular evolution
- correlating diversification / traits / substitution patterns
- see also total evidence dating (Ronquist et al, Syst Biol 61:973)

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