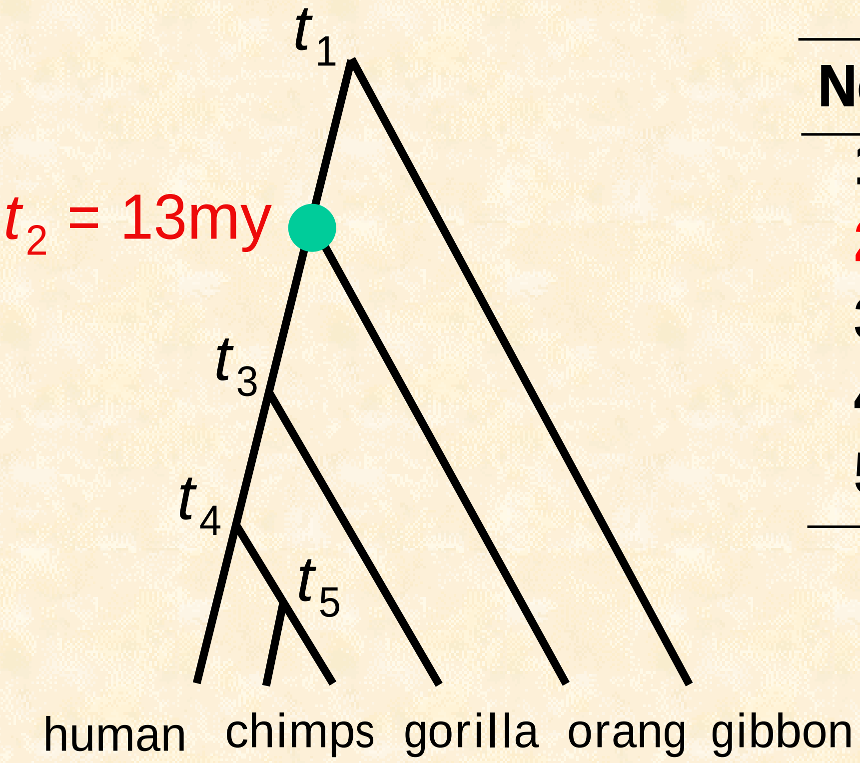
Estimation of species divergence dates with a sloppy molecular clock

Ziheng Yang

Department of Biology
University College London

Date estimation with a clock is easy.

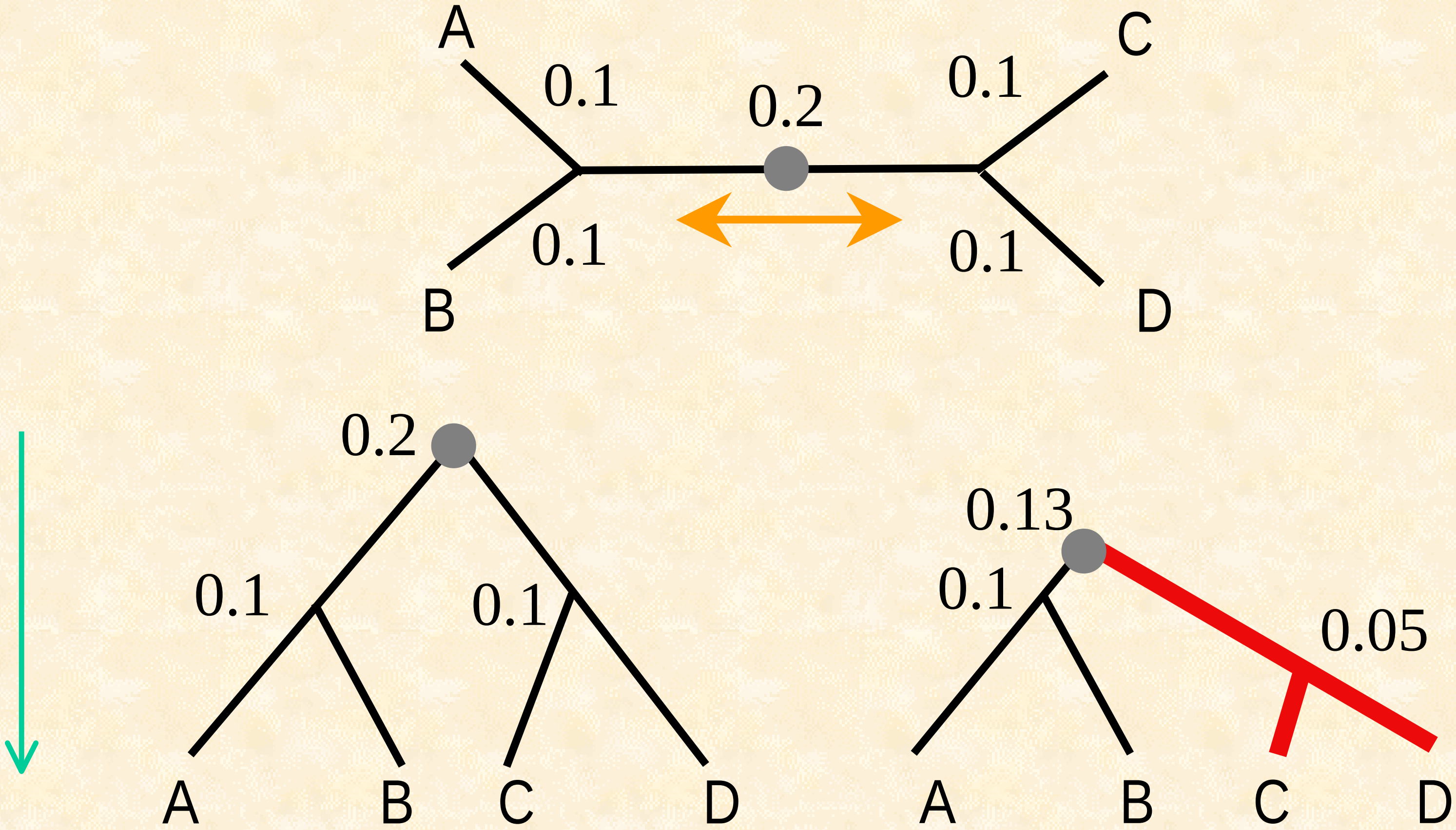


Node	Distance	Time(MY)
1	0.0627	15.8 ± 1.5
2	0.0517	13
3	0.0307	7.7 ± 1.0
4	0.0201	5.0 ± 0.9
5	0.0093	2.4 ± 0.6

(Mutation rate = 0.0517/13)

(From Yang 1996 *TREE* **11**:367-372; Horai et al. data)

Date estimation without clock is tricky.



How do we estimate dates when the clock is violated?

- Remove species or sites so that the clock looks reasonable and then use the clock.
- Account for rate variation when estimating dates.

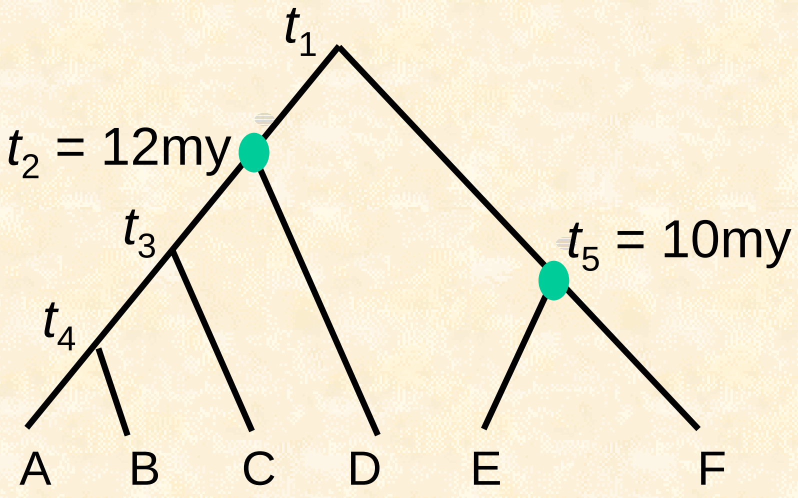
Dealing with evolving evolutionary rates

- Use multiple calibration points
- Combine multiple genes in one analysis
- Use good models of rate evolution and good statistical methodology

Outline

- Likelihood local clock models
- Bayes models of stochastic rate evolution
- Dating divergences in Malagasy mouse lemurs

ML: global clock model with multiple calibrations



((((A, B), C), D), (E, F));

((((A, B), C), D) @0.12, (E, F) @0.1);

((((A, B), C), D) >0.08<0.15, (E, F) @0.1);

Model with fixed dates:

r (rate), t_1, t_3, t_4

($t_2 = 0.12$, $t_5 = 0.1$)

Model not implemented:

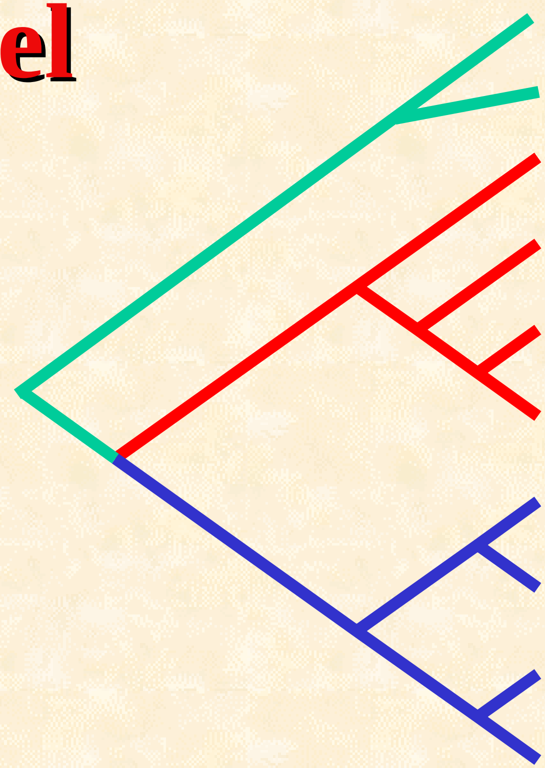
r, t_1, t_2, t_3, t_4

($0.08 < t_2 < 0.15$, $t_5 = 0.1$)

ML: local clock model

assigns rate parameters for
pre-specified branches

$r_0 = 1$
$r_P = 1.41421$
$r_R = 3.14159$

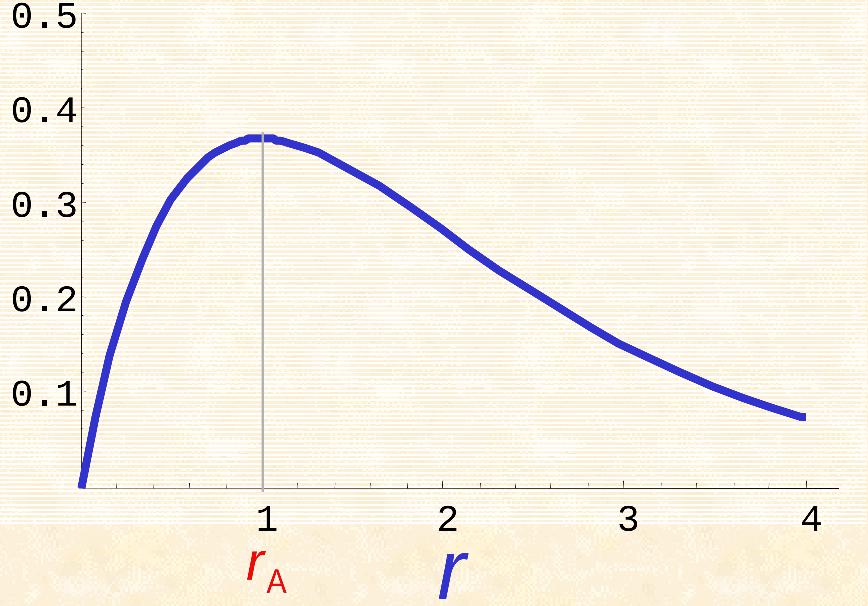


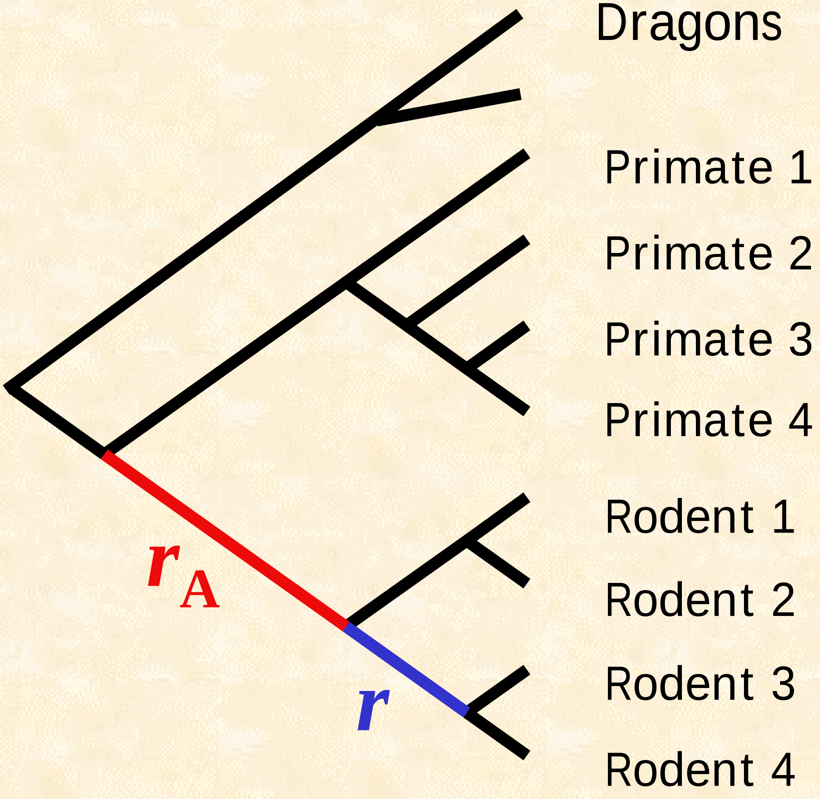
Dragons
Primate 1
Primate 2
Primate 3
Primate 4
Rodent 1
Rodent 2
Rodent 3
Rodent 4

Kishino & Hasegawa 1990 *Methods Enzymol*. 183:550-570
Rambaut & Bromham 1998 *Molecular Biology and Evolution* 15:442-448
Yoder & Yang 2000 *Molecular Biology and Evolution* 17:1081-1090

Stochastic model of rate change

The rate r of a branch (node) is a random variable centred around the ancestral rate r_A





Thorne & Kishino 2002 *Systematic Biology* 51:689-702
Thorne et al. 1998 *Molecular Biology and Evolution* 15:1647-1657
Kishino et al. 2001 *Molecular Biology and Evolution* 18:352-361

Bayes MCMC algorithm for date estimation

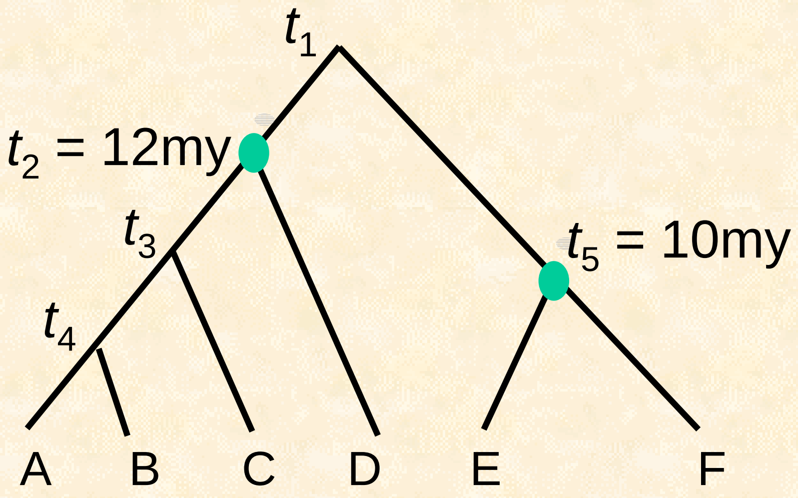
$$f(\mathbf{t},\mathbf{r},\theta \mid D) = \frac{f(\overset{\checkmark}{D \mid \mathbf{t},\mathbf{r}})\, f(\mathbf{r} \mid \theta,\mathbf{t})\, f(\mathbf{t} \mid \theta)\, f\overset{\checkmark}{(\theta)}}{f(\overset{\checkmark}{D})}$$

t: times
r: rates
θ: parameters
D: data

Details in Thorne's implementation changed

- Likelihood calculation is approximate
- Gamma prior for rate at root
- Gamma prior for age of rate
- Hard bounds for fossil calibration nodes

Prior for times f(t) from the birth-death process with species sampling



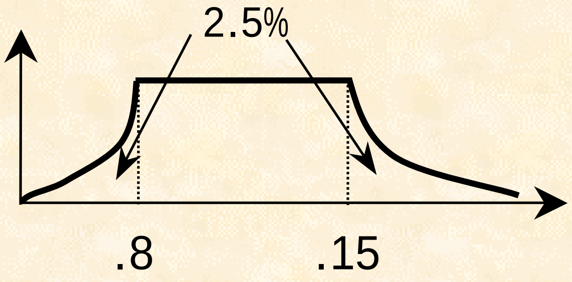
$\mathbf{t}_\mathrm{C} = \{t_2,\ t_5\}$ calibration nodes

$\mathbf{t}_{\overline{\mathrm{C}}} = \{t_1,\ t_3,\ t_4\}$ all other nodes

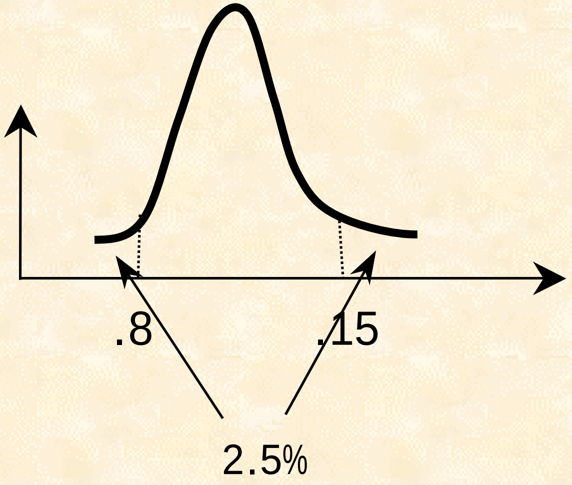
$$f(\mathbf{t}_\mathrm{C}, \mathbf{t}_{\overline{\mathrm{C}}}) = f_\mathrm{BD}(\mathbf{t}_{\overline{\mathrm{C}}} \mid \mathbf{t}_\mathrm{C})\, f(\mathbf{t}_\mathrm{C})$$

Prior for fossil calibration dates

>0.8<0.15 Soft bounds between (0.8, 0.15)



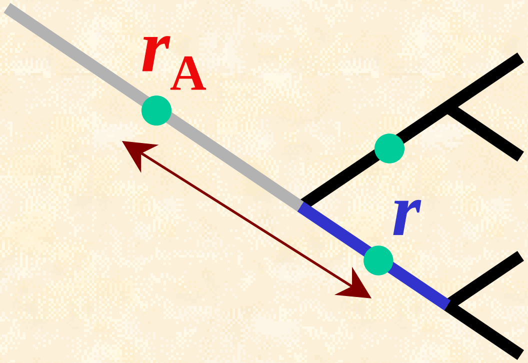
>0.8=0.12<0.15 gamma density with peak at 0.12



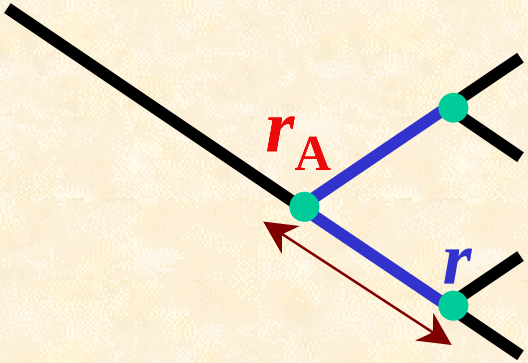
<0.8 Uniform for $t < 0.8$, power decay for $t > 0.8$

>0.8 Power decay to 0 when $t < 0.8$, uniform improper when $t > 0.8$

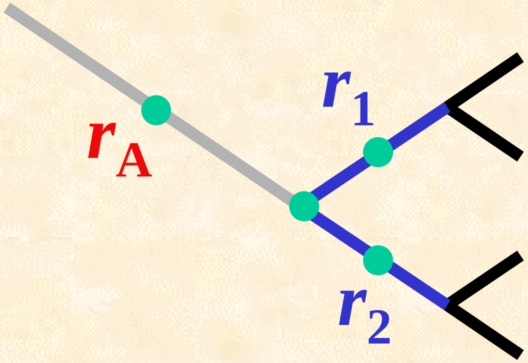
Models of rate change



Rates for branches (Thorne et al. 1998)

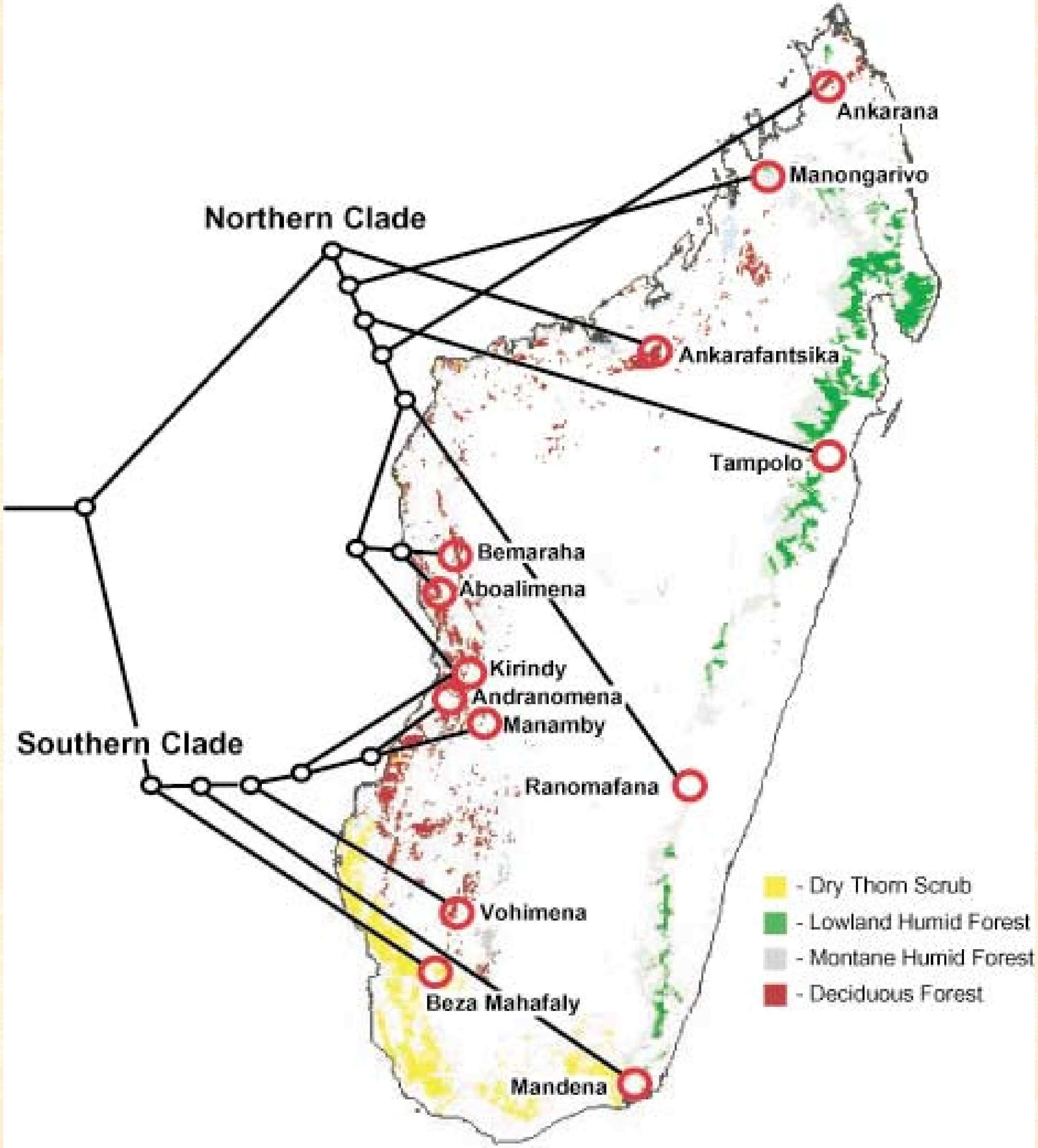


Rates for nodes (Kishino et al. 2001)



Rates for branches derived from rates for nodes

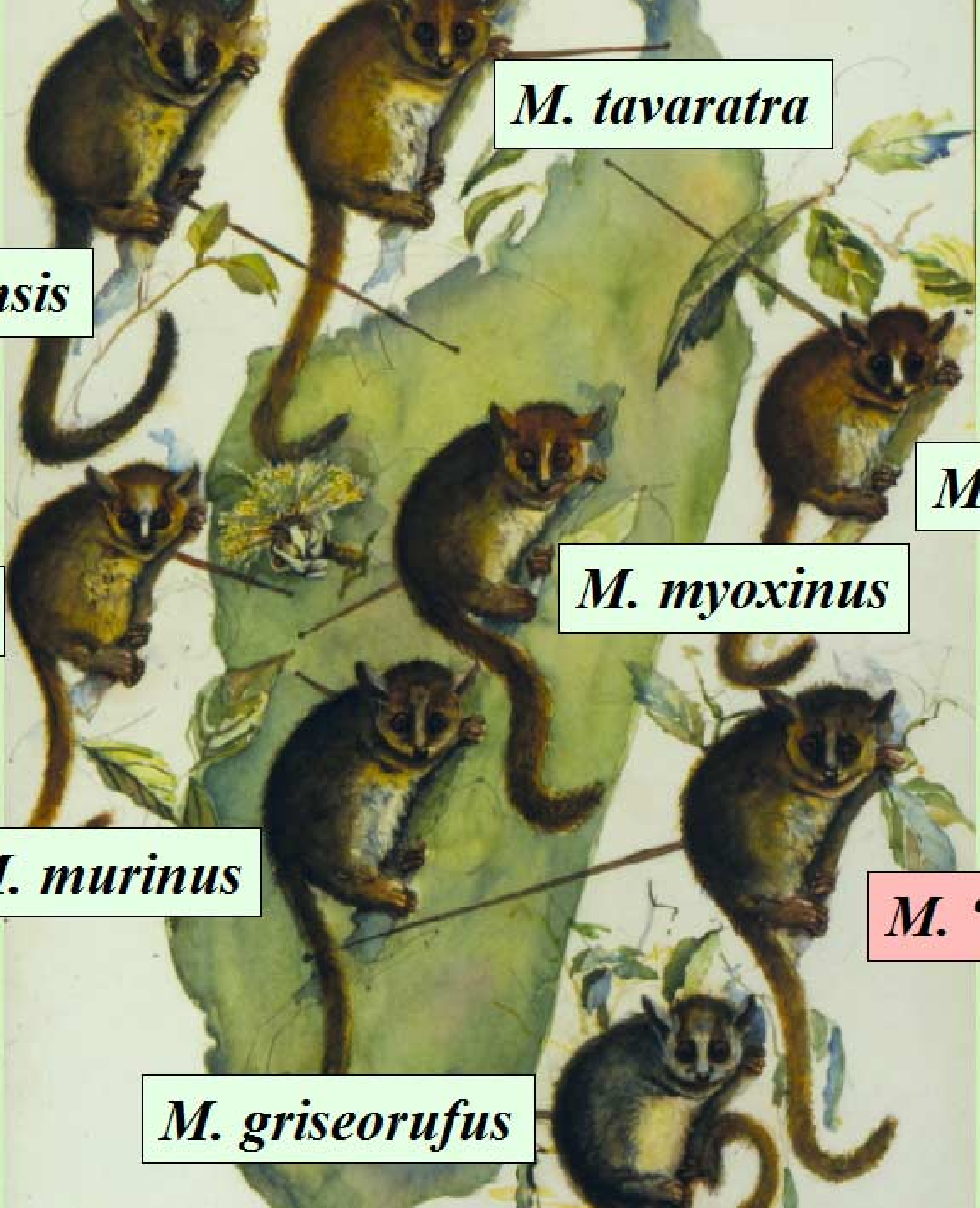
Dating mouse lemur divergences



1795: *Microcebus murinus*

1970s: 2 species (*murinus*, long-eared grey west; *rufus*, short-eared reddish east)

2000 (Yoder et al.): 9 species



M. tavaratra

M. ravelobensis

M. sambariensis

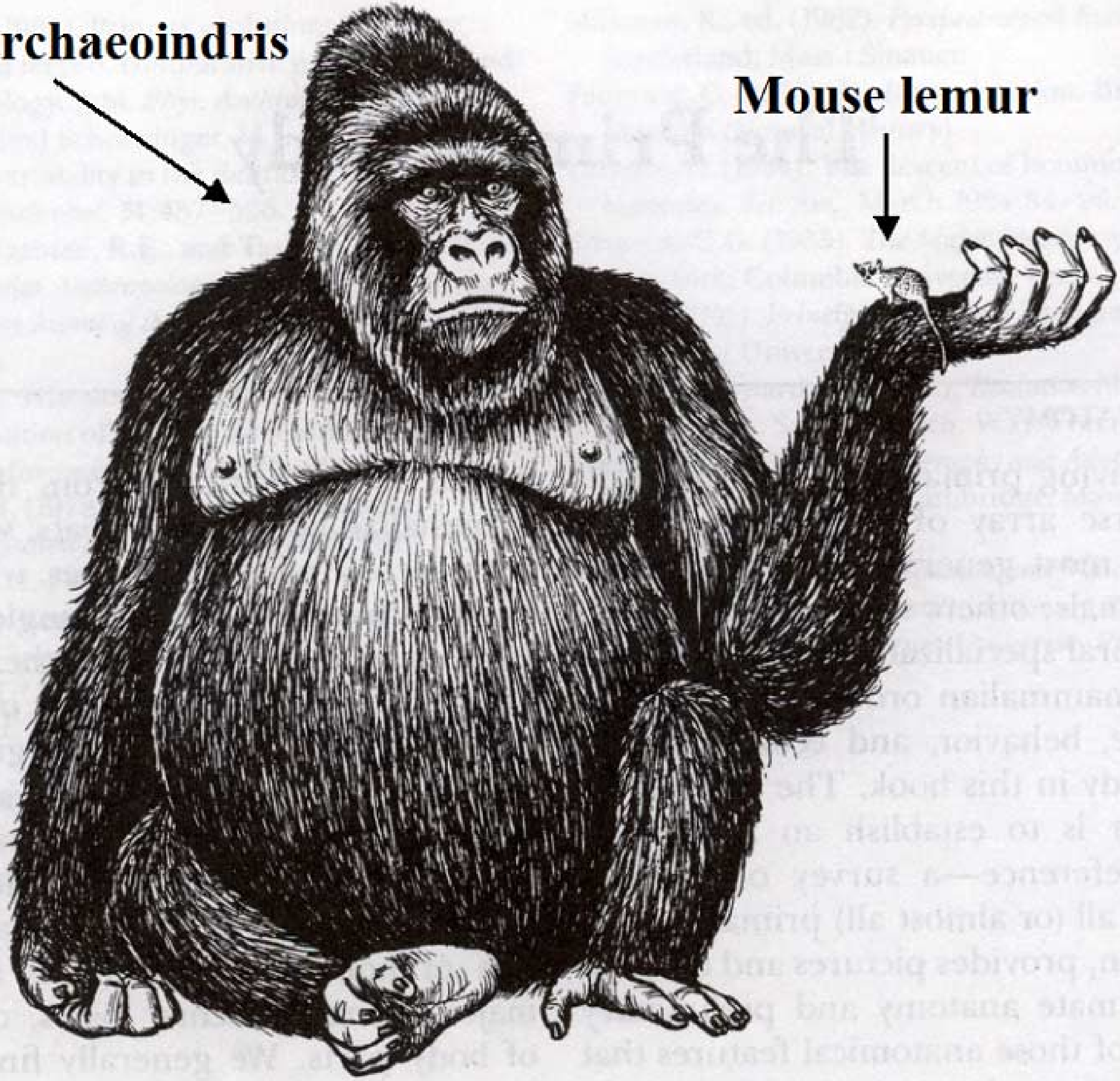
M. berthae

M. myoxinus

M. murinus

M. “rufus”

M. griseorufus



Archaeoindris

Mouse lemur

C_1 8 — 12

C_2 32 — 38

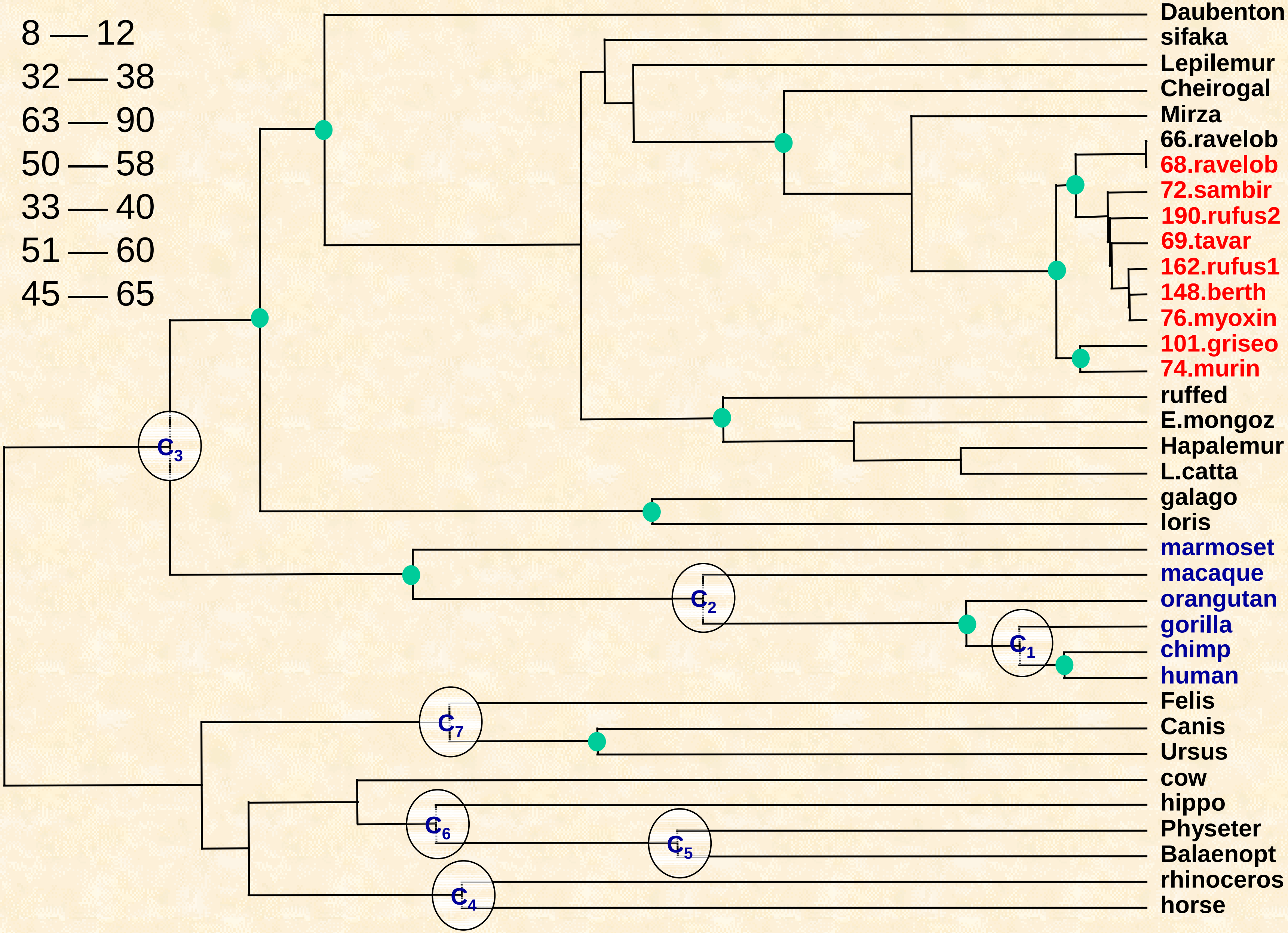
C_3 63 — 90

C_4 50 — 58

C_5 33 — 40

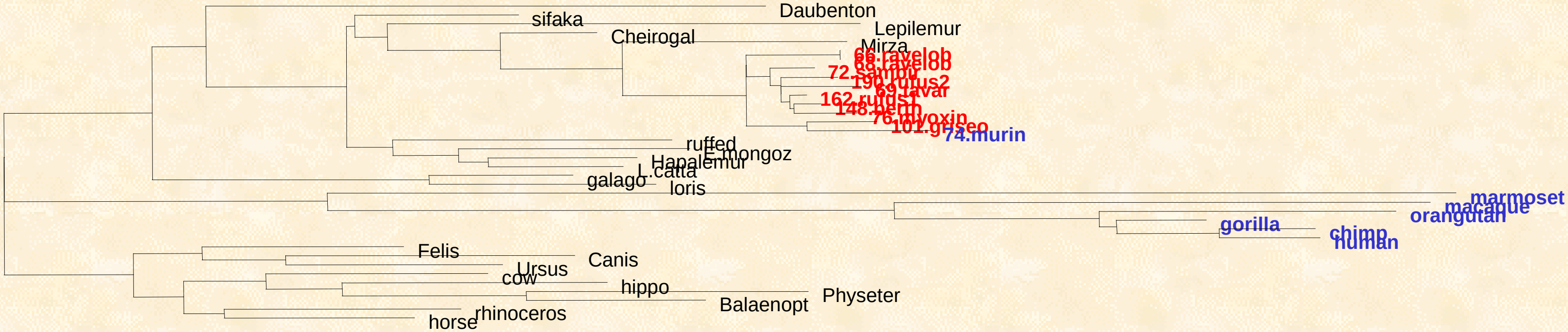
C_6 51 — 60

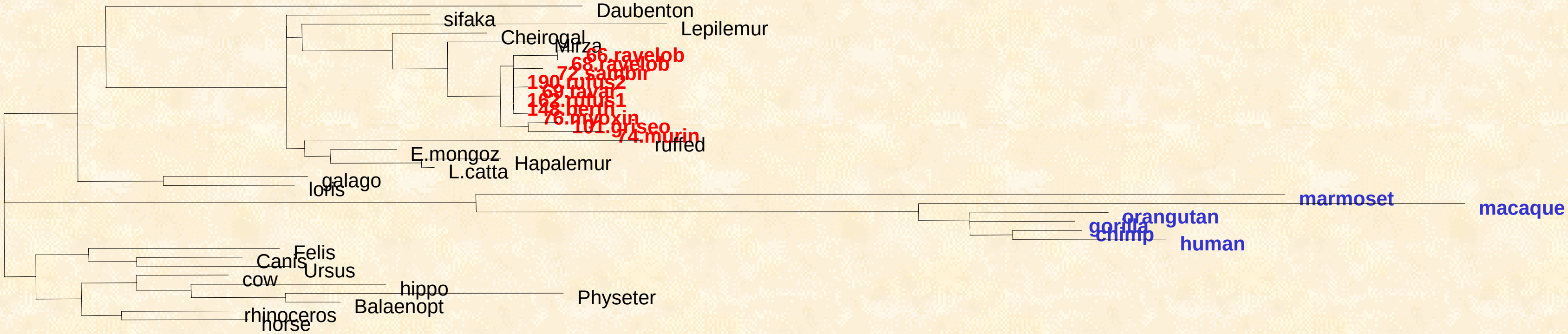
C_7 45 — 65

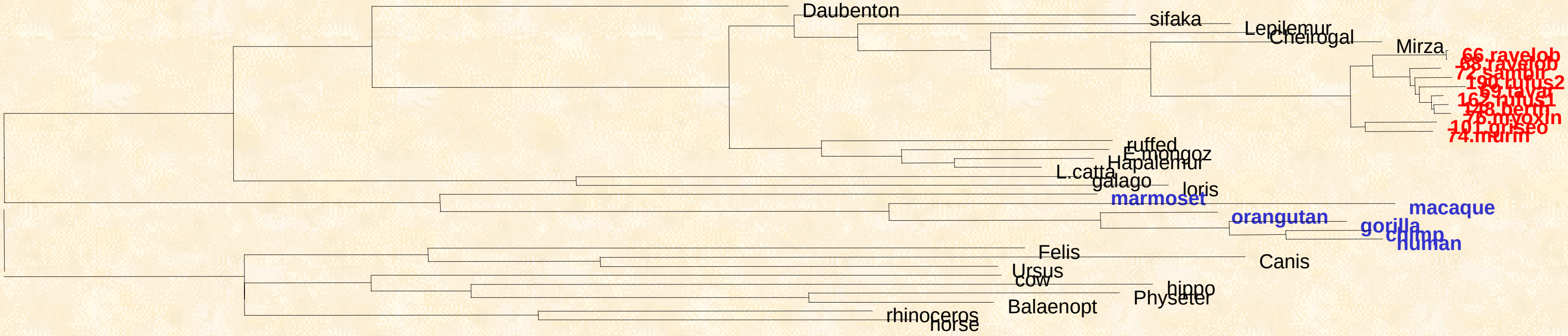


Parameter estimates (F84G) for COII and cytochrome b

Position	L	π_T	π_C	π_A	π_G	κ	α	S
1	606	0.23	0.26	0.29	0.22	3.5	0.29	3.4
2	606	0.40	0.25	0.22	0.13	2.7	0.17	1.2
3	606	0.21	0.34	0.40	0.04	19.0	1.17	55.3
All	1818	0.28	0.28	0.31	0.13	4.4	0.29	10.8


Daubenton
sifaka
Lepilemur
Cheirogal
Mirza
66.ravelob
68.ravelob
72.sambir
190.rufus2
69.tavar
162.rufus1
148.berth
76.myoxin
101.griseo
74.murin
ruffed
E.mongoz
Hapalemur
L.catta
galago
loris
marmoset
macaque
orangutan
gorilla
chimp
human
Felis
Canis
Ursus
cow
hippo
Physeter
Balaenopt
rhinoceros
horse


Daubenton
sifaka
Lepilemur
Cheirogal
Mirza
66.ravelob
68.ravelob
72.sambir
190.rufus2
69.tavar
162.rufus1
148.berth
76.myoxin
101.griseo
74.murin
ruffed
E.mongoz
L.catta
Hapalemur
galago
loris
marmoset
macaque
orangutan
gorilla
chimp
human
Felis
Canis
Ursus
cow
hippo
Physeter
Balaenopt
rhinoceros
horse


Daubenton
sifaka
Lepilemur
Cheirogal
Mirza
66.ravelob
68.ravelob
72.sambir
190.rufus2
69.tavar
162.rufus1
148.berth
76.myoxin
101.griseo
74.murin
ruffed
E.mongoz
Hapalemur
L.catta
galago
loris
marmoset
macaque
orangutan
gorilla
chimp
human
Felis
Canis
Ursus
cow
hippo
Physeter
Balaenopt
rhinoceros
horse

ML estimates, global clock (F84G)

Node	all	pos1	pos2	pos3	Comb
40 dog/bear	43.3	37.2	29.1	42.6	38.9
43 human/chimp	8.0	6.8	8.4	6.9	7.1
45 Hominoid	17.8	20.2	15.8	15.8	17.3
47 Anthropoi	61.7	65.4	60.5	58.1	62.0
48 Lorisiform	36.0	23.6	28.8	42.7	33.6
51 Lemuridae	31.0	34.3	35.7	23.6	28.0
59 mouse lemurs	13.1	13.8	10.0	7.2	8.5
61 Cheirogaleidae	30.7	30.0	22.4	26.3	26.5
65 Lemuriform	59.0	59.5	63.9	53.7	57.8
66 Strepsirrhine	62.9	62.6	63.9	65.7	63.4

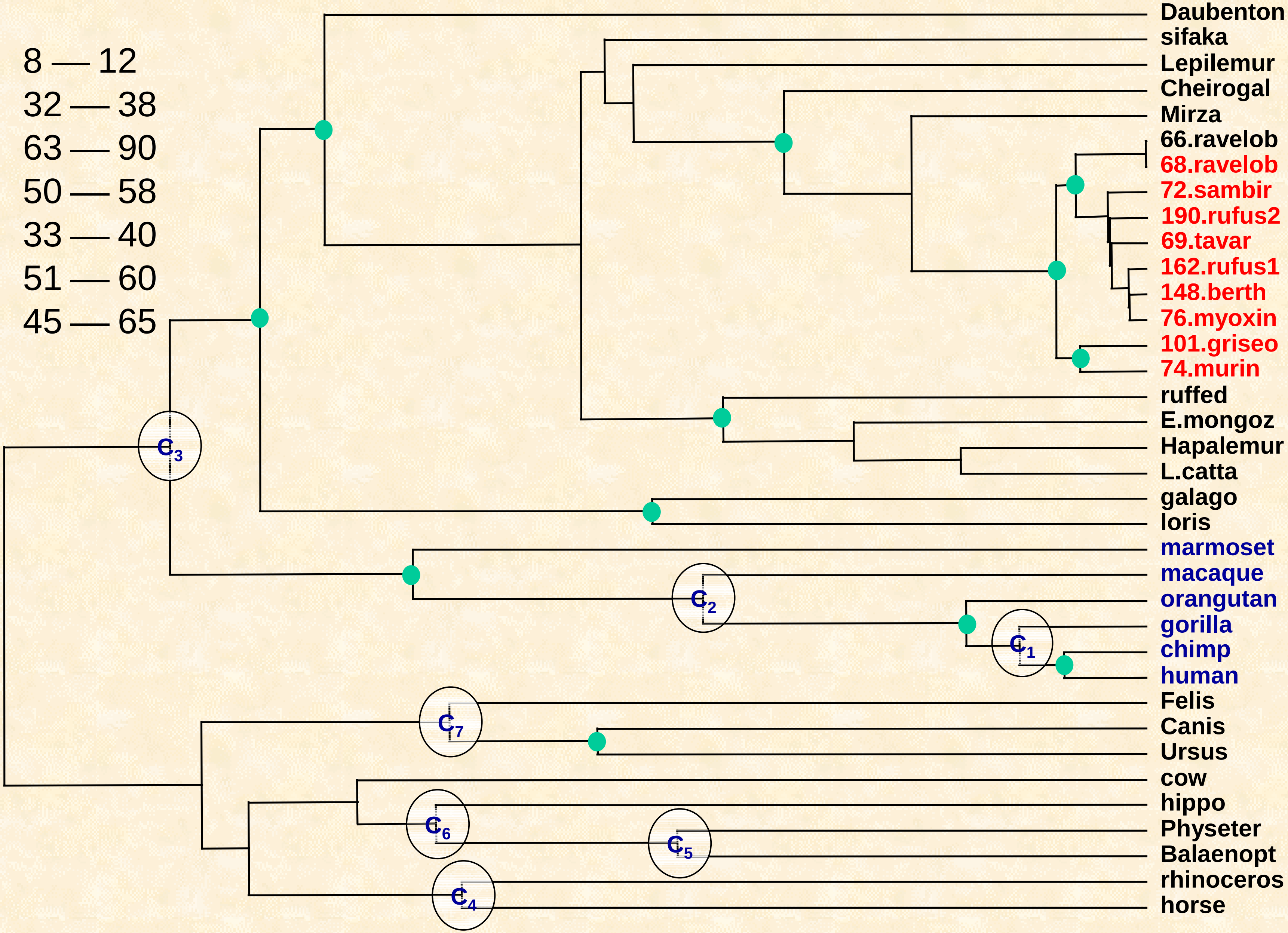
ML estimates, local clock models (F84G)

Node	all	pos1	pos2	pos3	Comb
40 dog/bear	46.4	43.0	36.6	44.4	43.2
43 human/chimp	7.3	5.6	7.6	6.4	6.5
45 Hominoid	14.5	14.6	10.4	14.7	14.2
47 Anthropoid	58.2	61.9	55.1	54.8	57.8
48 Lorisiform	41.1	29.3	36.9	45.6	38.9
51 Lemuridae	36.3	43.8	46.2	25.8	33.3
59 mouse lemurs	9.1	11.4	9.3	4.5	7.2
61 Cheirogaleidae	31.1	33.8	26.9	25.7	28.6
65 Lemuriform	64.7	69.0	71.7	56.7	64.8
66 Strepsirrhine	68.8	71.8	71.7	68.8	69.9

Bayes estimates of divergence times (F84G)

Node	all	pos1	pos2	pos3	Comb
40 dog/bear	49.9	43.3	34.7	46.4	45.2 (34.9, 55.8)
43 human/chimp	8.1	7.1	5.9	6.9	7.1 (5.1, 9.3)
45 Hominoid	16.2	15.6	19.2	15.2	15.2 (12.1, 18.6)
47 Anthropoid	59.4	60.4	56.0	60.5	61.1 (50.1, 73.0)
48 Lorisiform	41.7	29.9	40.4	45.1	40.5 (29.3, 53.0)
51 Lemuridae	37.0	45.0	45.4	30.0	35.3 (26.2, 46.1)
59 mouse lemurs	13.1	23.1	25.8	8.5	10.0 (6.4, 15.4)
61 Cheirogaleidae	31.1	38.0	38.6	29.5	30.3 (21.9, 40.5)
65 Lemuriform	65.4	66.8	63.7	60.5	66.9 (55.2, 78.3)
66 Strepsirrhine	70.4	72.0	69.1	71.5	73.3 (62.2, 83.6)

C_1	8 — 12
C_2	32 — 38
C_3	63 — 90
C_4	50 — 58
C_5	33 — 40
C_6	51 — 60
C_7	45 — 65



Summary

- For the mouse lemur data,
  - Models (JC vs. F84G) have a big impact.
  - Codon positions are very different.
  - Likelihood local clock models and Bayes MCMC calculation. produced similar estimates in combined analysis.
  - Bayes analysis was robust to prior parameters.
- It is advantageous to use multiple calibrations and multiple genes in a combined analysis.
- Date estimation without the clock remains tricky.

“Remarkable species diversity in Malagasy mouse lemurs (primates, Microcebus)”, Yoder *et al.* 2000. *Proc. Natl. Acad. Sci. U.S.A.* **97:**11325-11330.

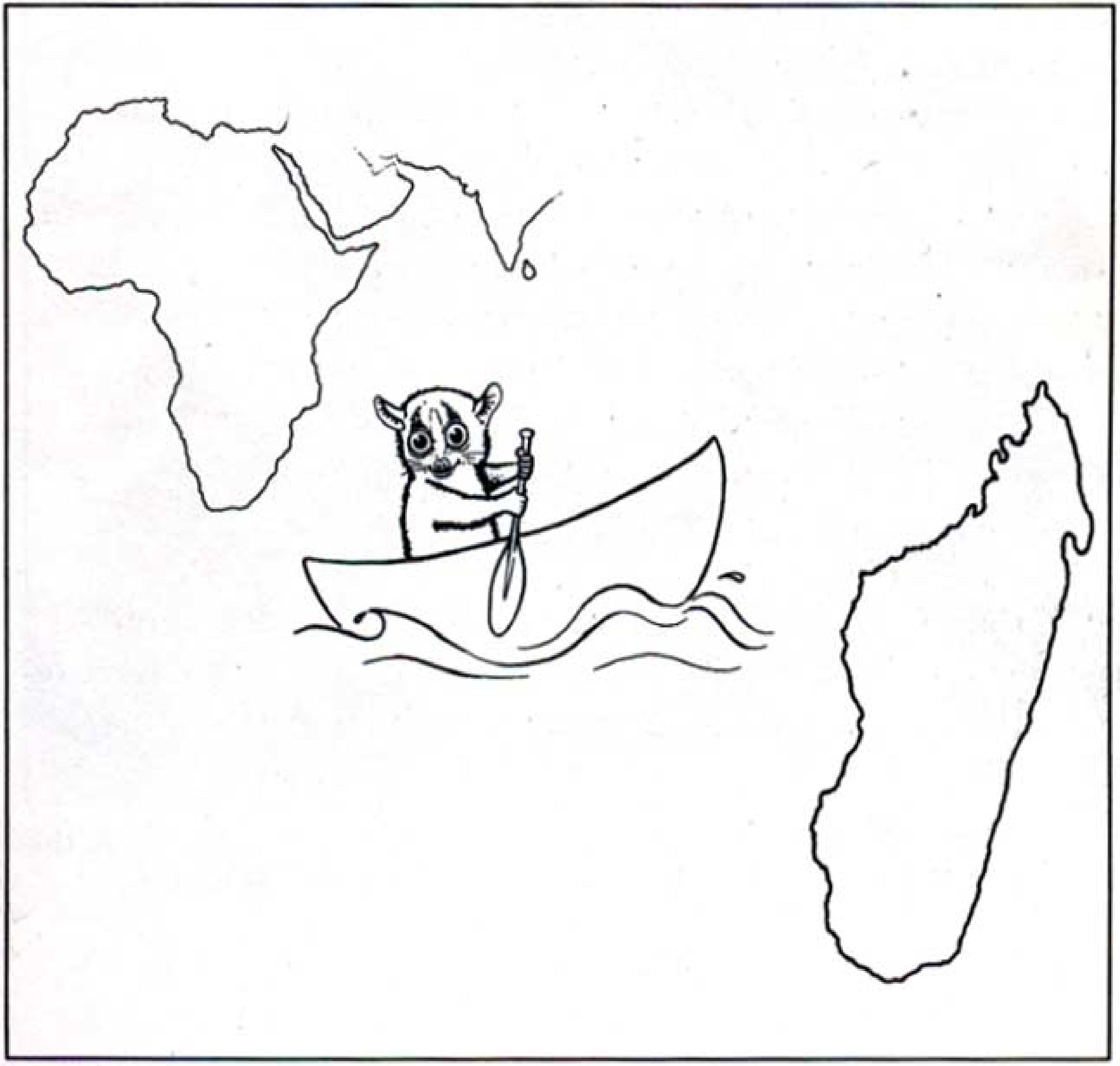


ILLUSTRATION BY STEPHEN D. NASH

Fig. 2.2: An ancestral lemur makes its way across to Madagascar.

Acknowledgments

Anne Yoder, Yale Univ.

Bruce Rannala, Univ. Alberta

BBSRC