

The 15th
Korea-Japan-China Bioinformatics Symposium 2017
@ Seoal 2017.6.21-22

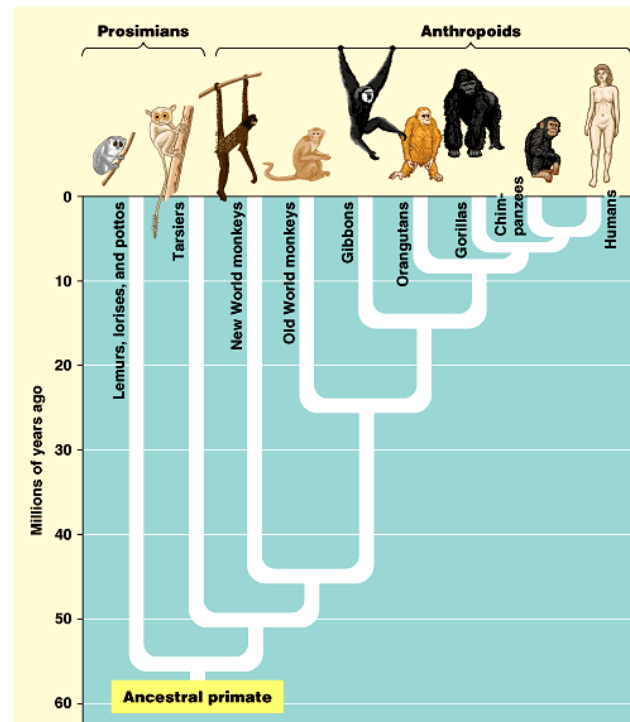
Phylo-demographic study of primates: problems of the mutation rate and an ancestral population size

Yoko Satta

Department of Evolutionary Study of Biosystems
SOKENDAI (The Graduate University for Advanced Studies)
Hayama, JAPAN

Molecular Phylodemography

Molecular Phylogeny + Molecular Demography

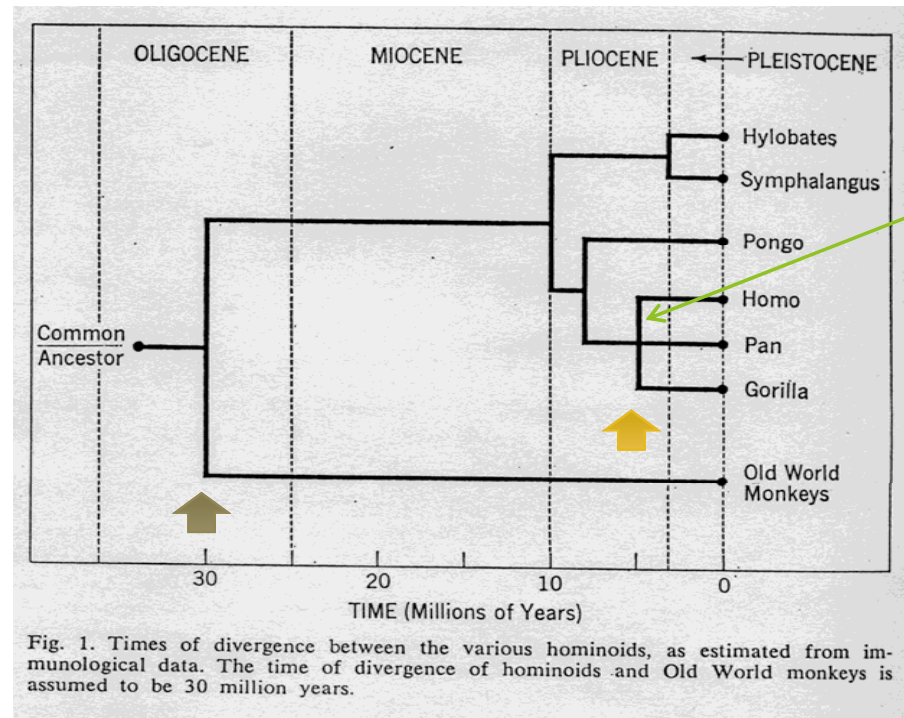


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Milestone of primate phylogeny in 1967

Sarich and Wilson (1967): serological reactions

→ quantitative method to measure the extent of antigen-antibody responses.

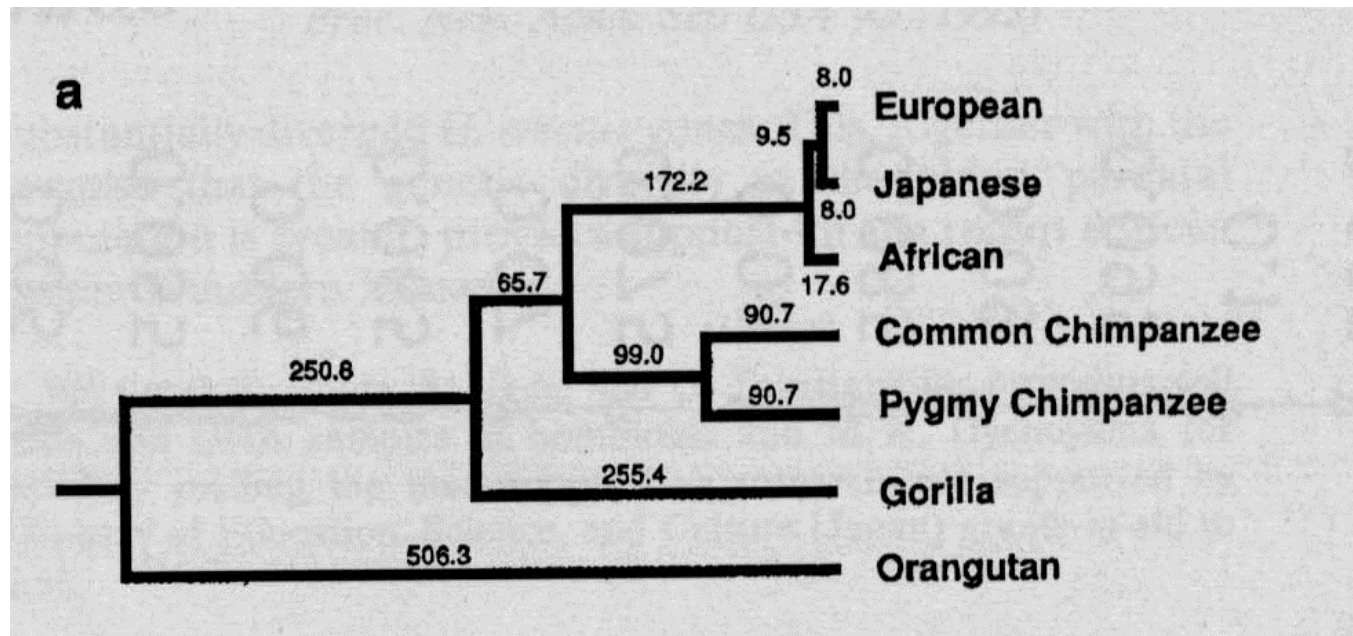


African apes and humans have a more recent common ancestry than was usually considered at that time.

If man and OWMs shared a last common ancestor 30 million years ago (mya), then man and the apes shared a common ancestor 5 mya.



Horai et al. (1995): mitochondrial DNA sequence



A large number of nucleotide substitutions were shared between humans and chimpanzees

Rogers (1993): 7 nuclear DNA sequences

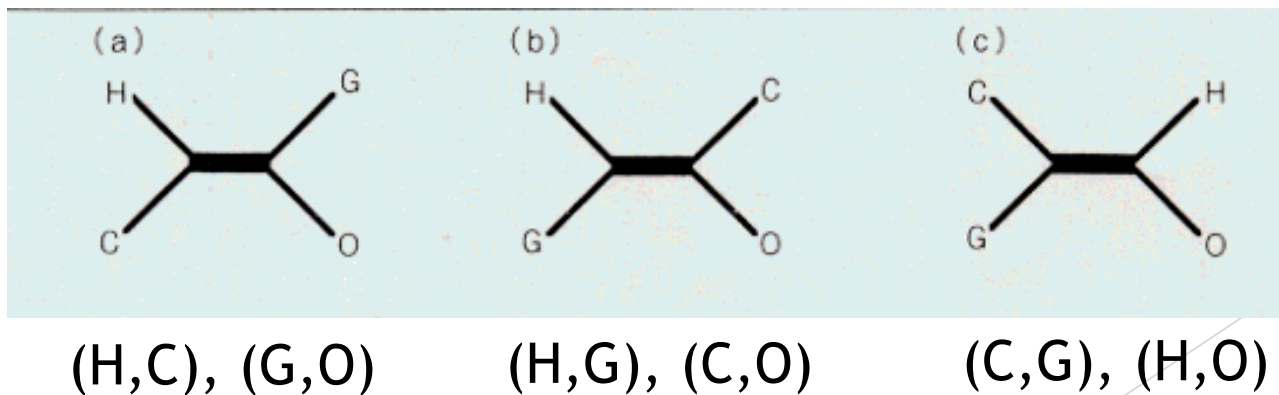
DNA sequence data		
Locus	Phylogeny supported	Reference
(A) Nuclear DNA		
Beta-globin cluster	(H-C)-G	Koop <i>et al.</i> , 1989
Involucrin	H-(C-G)	Djian & Green, 1989
X-chromosomal pseudo-autosomal boundary	H-(C-G)	Ellis <i>et al.</i> , 1990
Y-chromosomal pseudo-autosomal boundary	(H-C)-G	Ellis <i>et al.</i> , 1990
HOX2B	(H-C)-G	Ruano <i>et al.</i> , 1992
Immunoglobulin alpha	?	Kawamura <i>et al.</i> , 1991
Alpha-1,3-galactosyltransferase	(H-C)-G	Galili & Swanson, 1991
(B) Mitochondrial DNA		
mtDNA—5 kb including several genes	(H-C)-G	Horai <i>et al.</i> , 1992 ¹
mtDNA—ribosomal RNA	?	Hixson & Brown, 1986
mtDNA—transfer RNA genes, ND4, ND5	?	Brown <i>et al.</i> , 1992

¹This dataset includes the hominoid sequence published by Ruvolo *et al.* (1991) and uses *Pongo* as the outgroup rather than *Hylobates*.

The nuclear data were not always consistent with the result of mtDNA.

Phylogenetic trees of different loci

Relationship	Satta et al. (2000)	Chen and Li (2001)	O'hUigin et al. (2002)	Kitano et al. (2004)	Total [%]
	# of loci used in the analysis				
	45	53	51	103	252
(H,C), (G,O)	21	(31)	24	34	110 [57]
(H,G), (C,O)	7	(10)	14	14	45 [23]
(C, G), (H,O)	7	(12)	9	10	38 [20]
(H, C, G), O	10	(0)	4	45	59

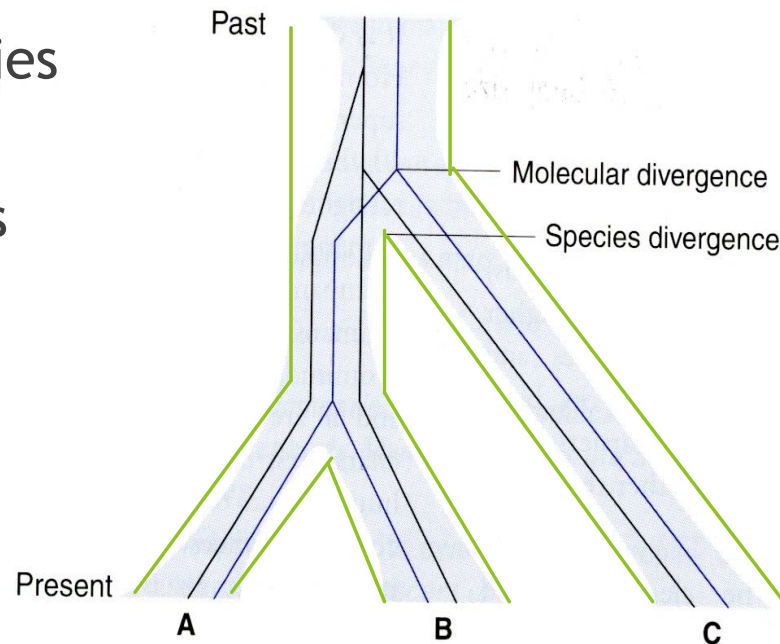


Inconsistency between species tree and gene tree

- ▶ Interval of two successive species divergence (T_s) is short.
- ▶ An ancestral population ($2N_e$) is large.

Inconsistency :
 $2N_e \gg T_s$

Consistency between species tree and gene tree:
 $2N_e \ll T_s$



Ancestral population (N_a)

Coalescence time in an ancestral population (τ)

Species divergence

Species divergence time (t_s)

g = generation time

μ = mutation rate per year

τ : exponential distribution

K (number of substitutions): geometric distribution

$$x = 2\tau g \mu = 4N_a g \mu$$

t_s : constant

K : Poisson distribution

$$y = 2 t_s \mu$$

a

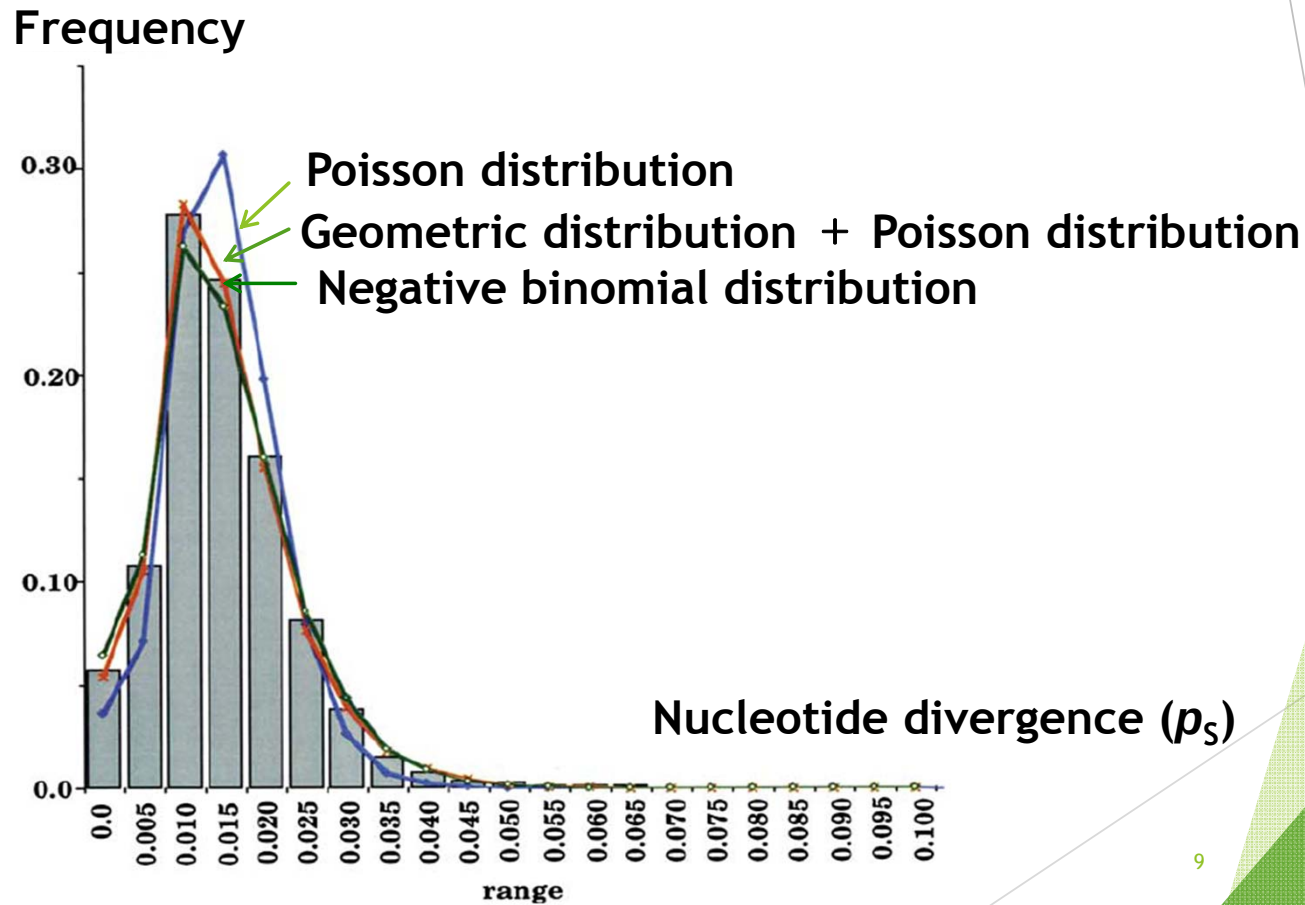
b

a'

b'



Histogram of frequencies of per-site substitutions between humans and chimpanzees at BES 58,158 loci (Satta et al. 2004, JME)



Methods are necessary to estimate the phylogeny incorporating information on demography (the ancestral population size).

Popular methods for inferring demographic histories

PSMC : Pairwise Sequentially Markovian Coalescent for (pseudo-)diploid sequences

(Li & Durbin 2011; Prado-Martinez et al. 2013)

MCMC: Markov Chain Monte Carlo

(Rannala & Yang 2003; Hey & Nielsen 2007; Hara, Imanishi & Satta 2012)

ML: Maximum Likelihood

(Takahata, Satta & Klein 1995; Hey 2010)



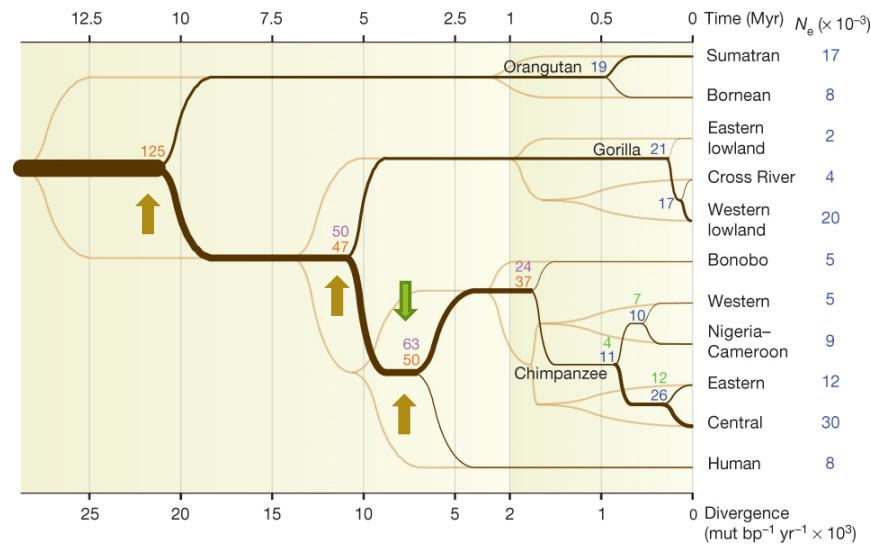
Advantages and disadvantages of the three methods

Methods	Advantages	Disadvantages
PSMC (Li & Durbin)	Applicable to all genomic regions even though partially linked Estimation of population size with contiguous time	Less statistical power for relatively recent history Rather subjective inference of species or population divergence
MCMC (Rannala & Yang)	Applicable to more than two species Simultaneous estimation of multiple ancestral/extant N_e and species divergence times.	Tight linkage within each region Applicable to only non-coding regions Prior dependency
ML (Takahata, Satta & Klein)	Applicable to both coding and noncoding sequences Simultaneous estimation of ancestral N_e and species divergence time	Tight linkage within each region Applicable to only two species Model dependency



Great ape genetic diversity and population history

Inferred population history

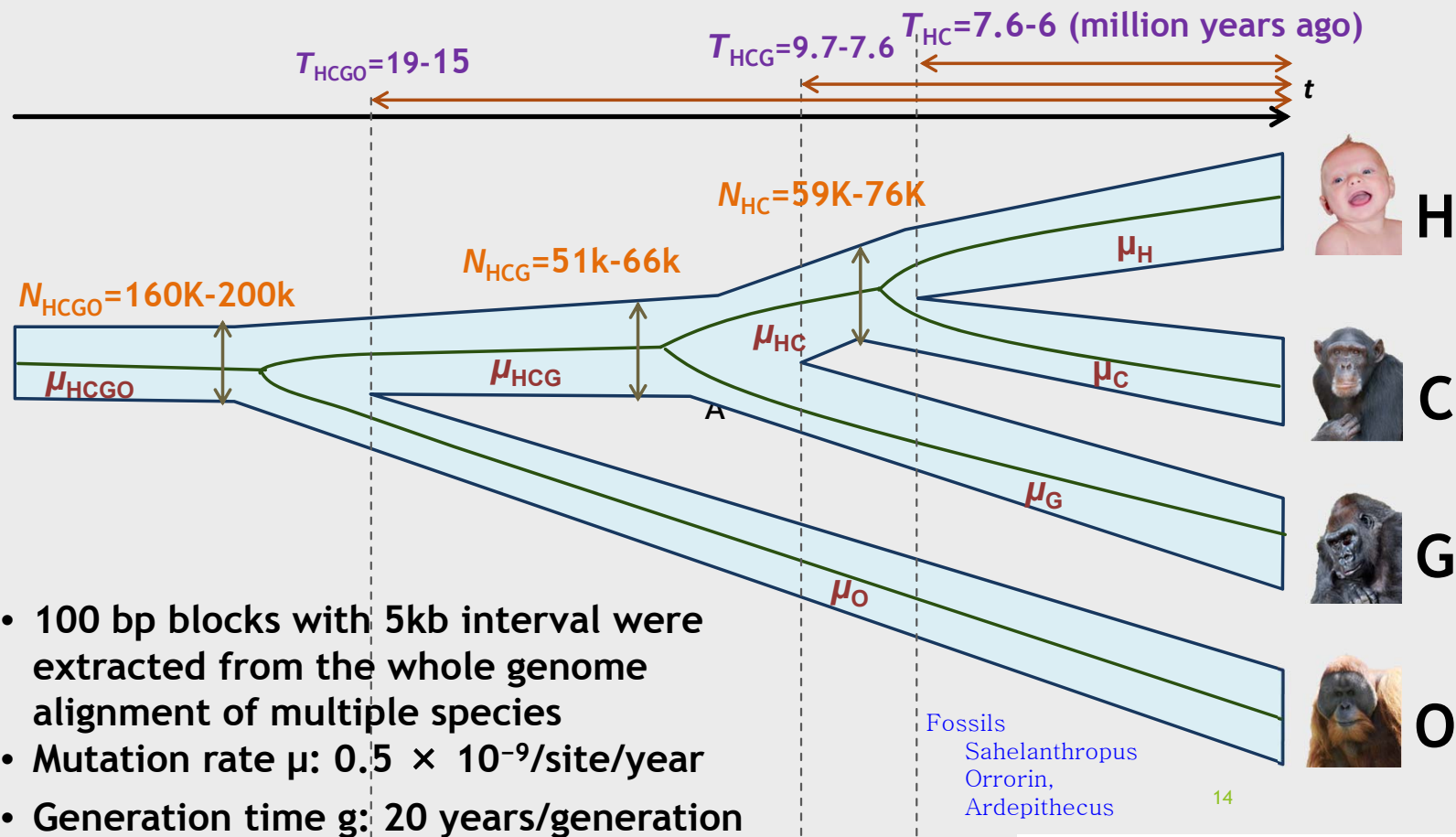


J Prado-Martinez *et al.* *Nature* **000**, 1-5 (2013) doi:10.1038/nature12228

nature



Species divergence time and **ancestral effective population sizes** (MCMC)



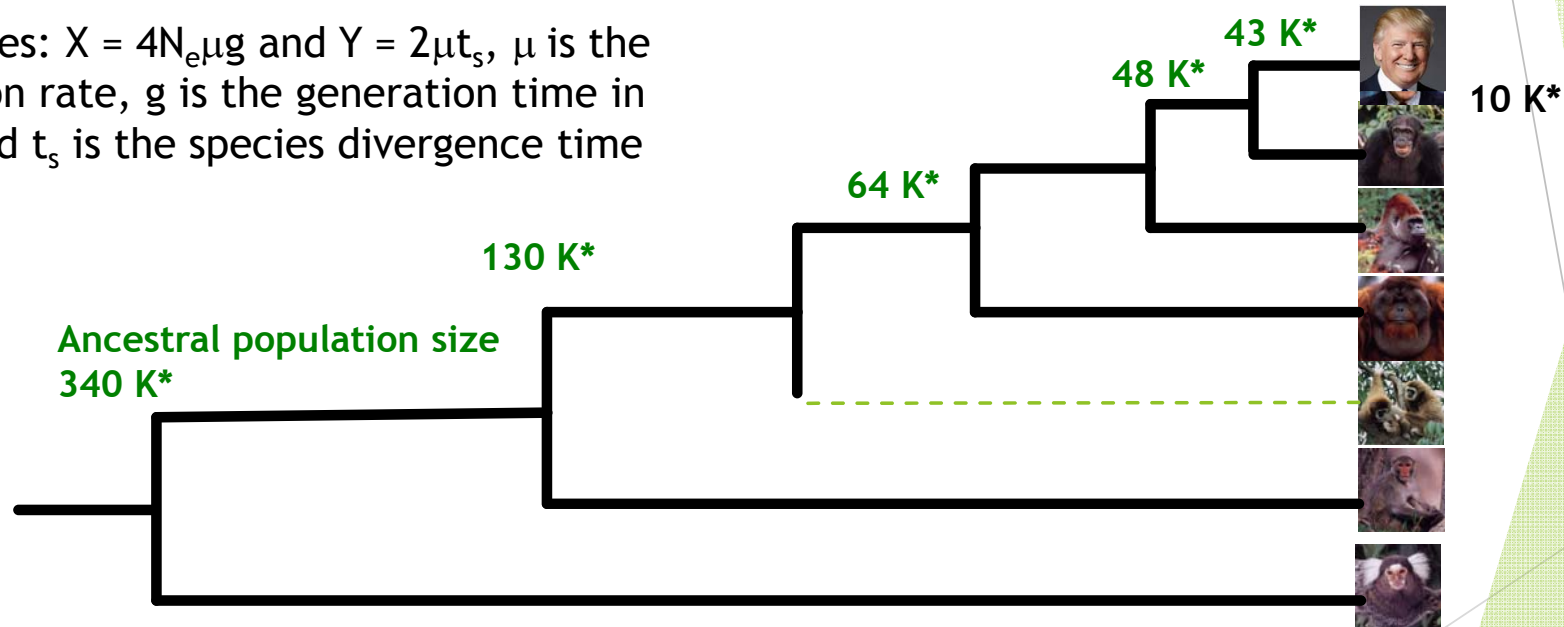
- 100 bp blocks with 5kb interval were extracted from the whole genome alignment of multiple species
- Mutation rate μ : 0.5×10^{-9} /site/year
- Generation time g : 20 years/generation

Hara, Imanishi and Satta (2012)

Primate phylo-demography by ML method (Kim et al. 2010)



Estimates: $X = 4N_e\mu g$ and $Y = 2\mu t_s$, μ is the mutation rate, g is the generation time in year and t_s is the species divergence time



- *Ancestral population size ($\times 10^3$)
- Absolute values are obtained under the assumption of constant substitution rates calibrated at the human and chimpanzee divergence time (7 MYA) : 0.6×10^{-9} per site per year



Comparison of estimates (ancestral population size & species divergence time) among three methods

estimates	method	H-C	H-G	H-O	H-OWM	H-NWM
Ne (10^3)	PSMC	50-63	47-50	125	-	-
	MCMC	59-76	51-66	160-200	-	-
	ML	43	48	64	130	340
T (myr)	PSMC	3-4	5-6	10-11	-	-
	MCMC	6-7.6	7.6-9.7	15-19	-	-
	ML	7**	9	21	42	70

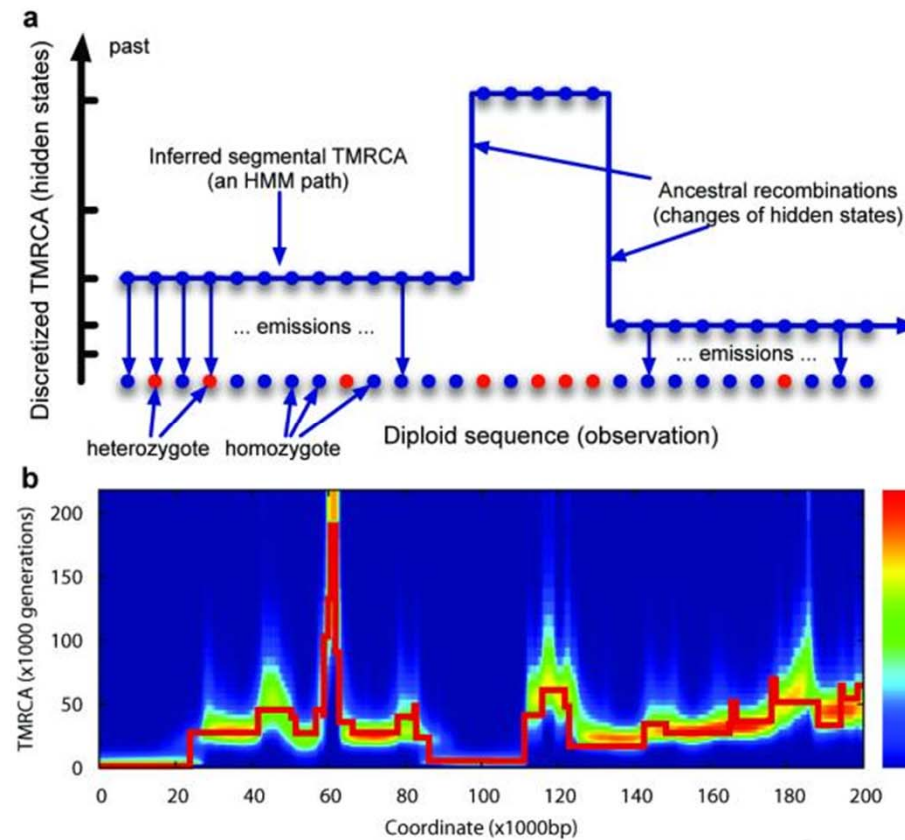
PSMC & MCMC: 0.5×10^{-9} /site/year ** ML: 0.6×10^{-9} /site/year

- ✓ The ancestral population size tends to be estimated larger in a diverged species pairs?
- ✓ Synonymous (neutral) substitution rate, 0.5×10^{-9} /site/year, can be applied to other primates?

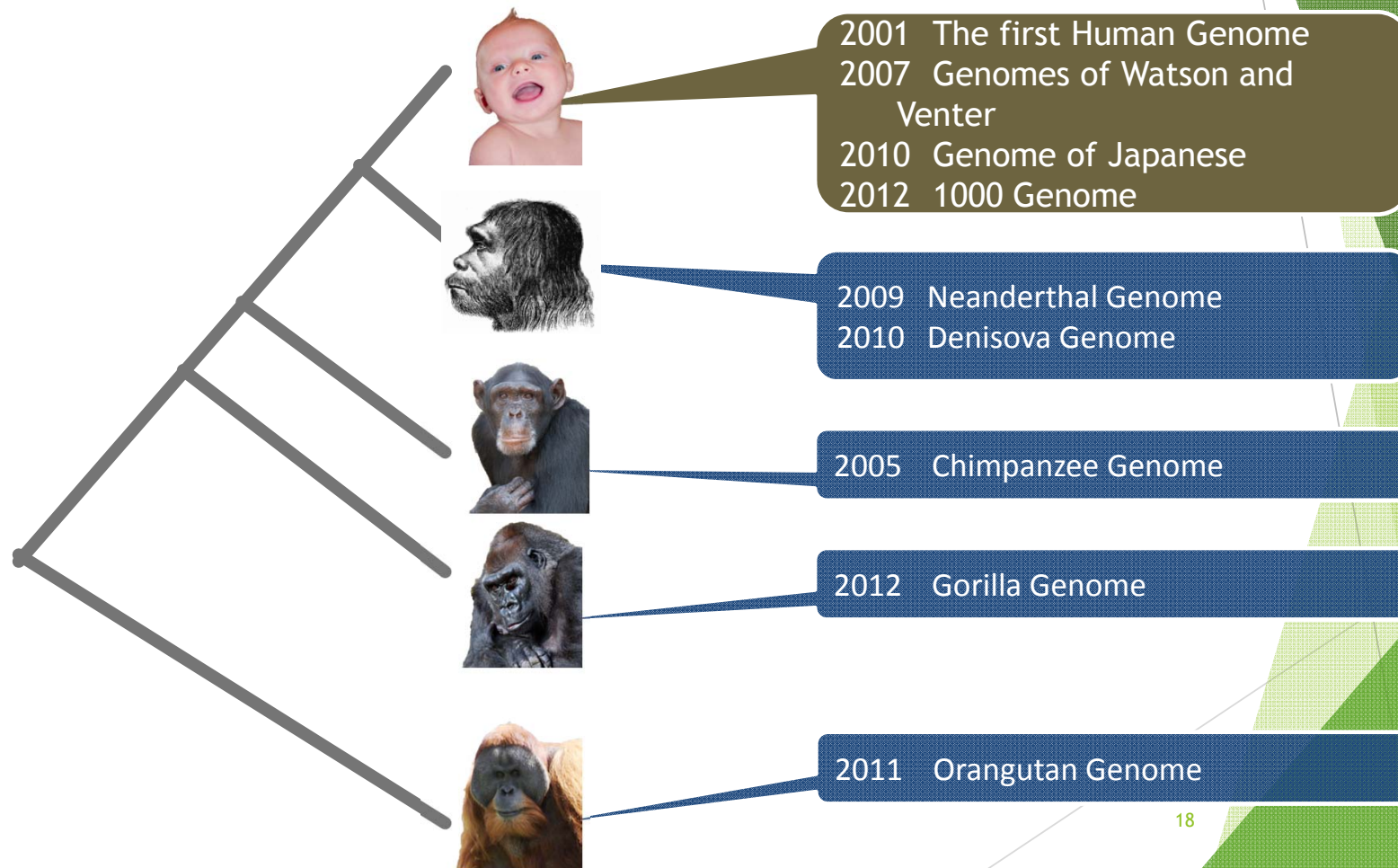
PSMC

Inference of human population history from the whole genome sequences of a single individual

Li and Durbin (2011) Nature 475: 493



Phylogenetic relationship of humans and Hominoids



18

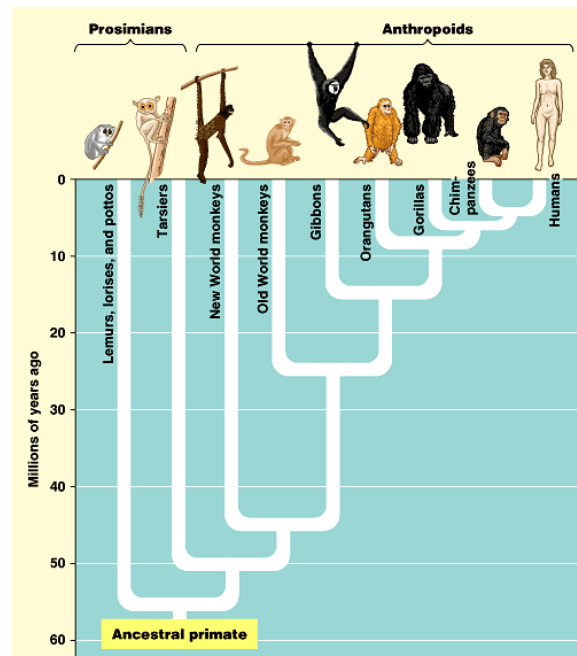
Problems raised

The entire primate phylogeny

Human and chimpanzee divergence time : 6-7 myr, Others?

Synonymous (neutral) substitution rate?

The ancestral population size tends to be estimated larger in a diverged species pairs?



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

- For the relatively distantly related species, OWMs, NWMs, and prosimians, the ML method for CDS is suitable.
 - ✓ The alignment for the CDS sequences is more reliable than that for non-coding sequences.

ML Application to CDS sequences in relatively distantly related primates

► Application of the method to OWMs, NWMs, and prosimians

- For the relatively distantly related species, OWMs, NWMs, and prosimians, the ML method for CDS is suitable.

The alignment for the CDS sequences is more reliable than that for non-coding sequences.

► Canonical Exon sequences from 17 species

Chr 22: (vs. *Hosa*)

Number of loci (**n**): 419 - 446

Average number of synonymous sites **per locus (L)**: 376.2 - 446.6

Canonical Exon sequences from 17 species

Chr 22: (vs. *Hosa*)

Number of loci (**n**): *Mimu* 419 - *Patr* 446

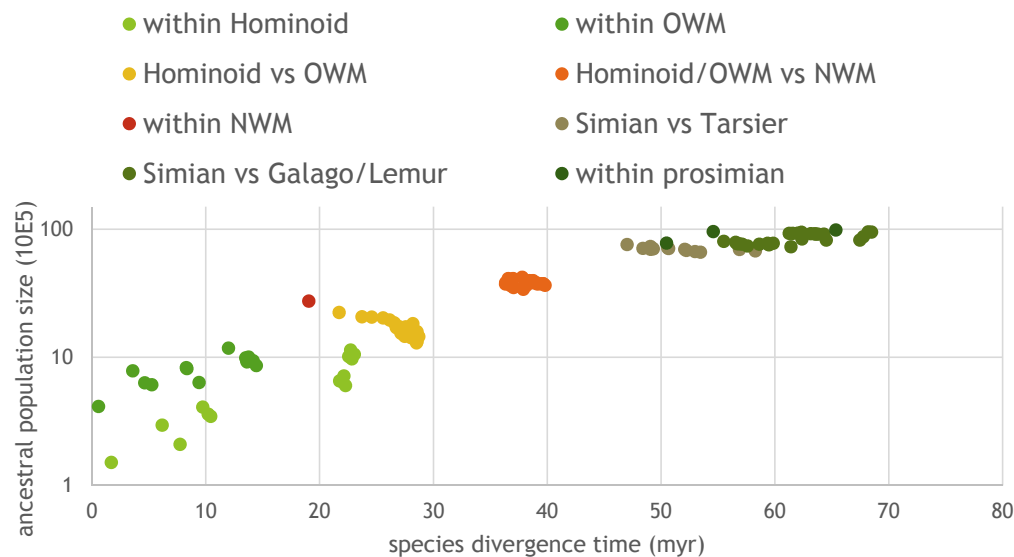
Average number of synonymous sites **per locus (L)**:

Mimu 376.2 - *Paan* 446.6

451 loci on chromosome 22 (TSML)						
species pair	n (L)	d	s.e	$x = 4N_e\mu_g$	$y = 2\mu t_s$	y/2
Hosa-Patr	446 (436.6)	0.01665	0.01417	0.00852	0.00797	0.00398
Hosa-Papa	432 (430.4)	0.01634	0.01170	0.00644	0.00981	0.00490
Hosa-Gogo	430 (429.1)	0.02460	0.03917	0.01062	0.01320	0.00660
Hosa-Poab	437 (437.2)	0.04768	0.01027	0.01946	0.02840	0.01420
Hosa-Nole	432(421,3)	0.05888	0.05291	0.02988	0.02965	0.01482
Hosa-Mamu	439 (430.1)	0.08568	0.03070	0.03079	0.05509	0.02754
Hosa-Mafa	444 (445.7)	0.08652	0.03907	0.03284	0.05399	0.02700
Hosa-Paan	443 (446.6)	0.08734	0.03788	0.03445	0.05319	0.02659
Hosa-Chsa	445 (443.1)	0.08590	0.03017	0.03113	0.05514	0.02757
Hosa-Nala	432 (388.0)	0.09013	0.04534	0.03839	0.05199	0.02599
Hosa-Rhro	433 (439.1)	0.09209	0.09141	0.03734	0.05330	0.02665
Hosa-Caja	434 (436.1)	0.15629	0.07244	0.06094	0.09439	0.04719
Hosa-Sabo	427 (426.4)	0.15949	0.09542	0.06354	0.09278	0.04639
Hosa-Tasy	429 (415.4)	0.29241	0.09822	0.13911	0.15530	0.07765
Hosa-Miru	419 (376.2)	0.28981	0.11239	0.13366	0.15672	0.07836
Hosa-Otga	424 (424.8)	0.33250	0.12380	0.16075	0.16940	0.08470

Estimation of ancestral population size (N_e) and the speciation time (t)

Ancestral population size (y axis) vs.
species divergence time (x axis)

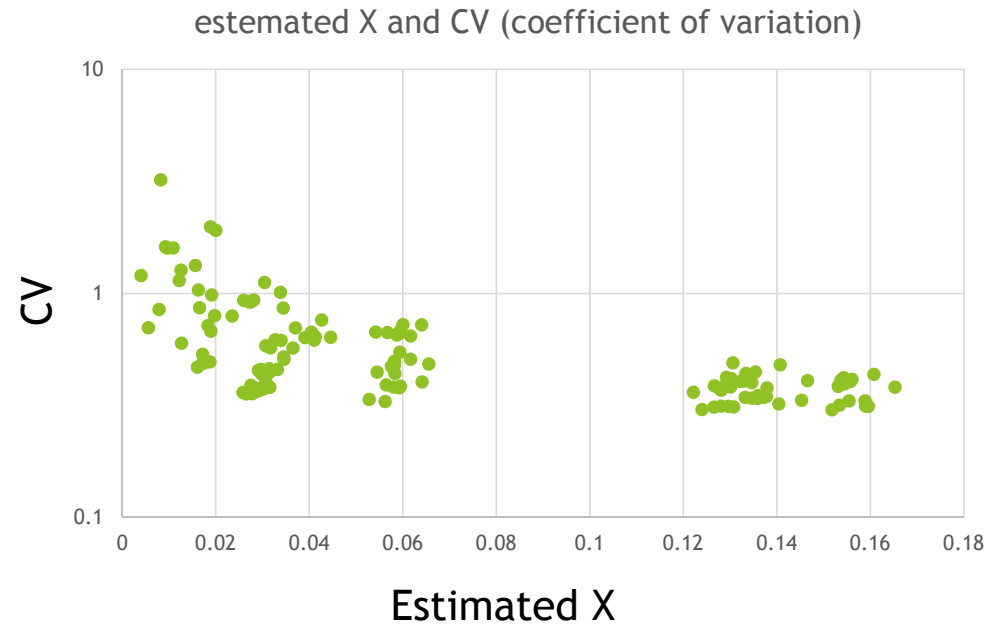


Ancestral population sizes tend to be larger in divergent species pairs.
The estimated size is too large.

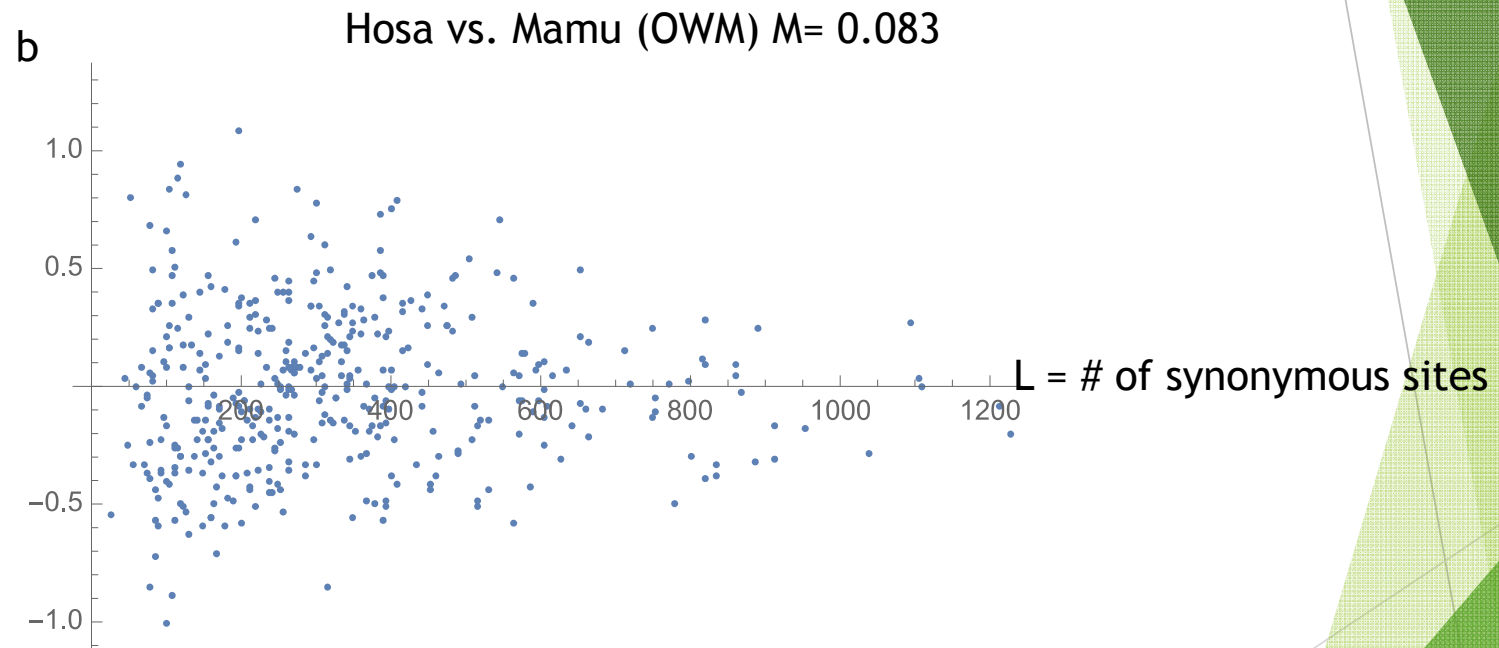
Comparison of ML estimates of X and Y between other studies and present one

	Takahata (2001)			Kim et al. (2010)		Present study		
	d (%)	X	Y	X	Y	X	Y	d
Hosa vs chimp	1.75	0.45	1.32	0.35	0.82	0.85	0.80	1.67
gorilla	2.04	0.50	1.60	0.39	1.06	1.06	1.32	2.46
orang	4.03	0.92	3.16	0.52	2.46	1.95	2.84	4.77
gibbon	4.88	0.91	4.01	-	-	3.00	3.00	5.89
OWM	7.72	0.40	7.35	1.03	4.84	3.3	5.4	8.57
NWM	13.1	1.05	12.1	2.73	8.0	6.1	9.4	15.6
Lemur	27.7	0.40	26.0	-	-	13.4	15.7	29.0

Cause of large estimated X is the variation of synonymous substitution rates among loci ??

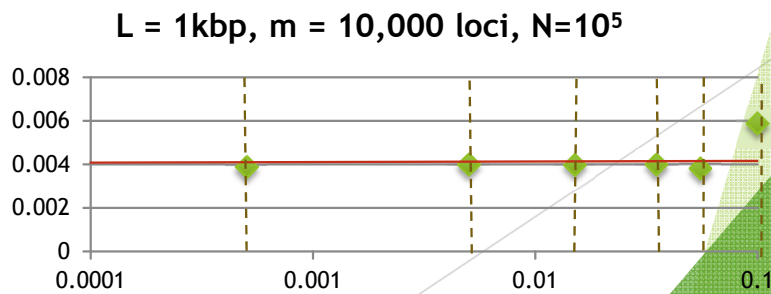
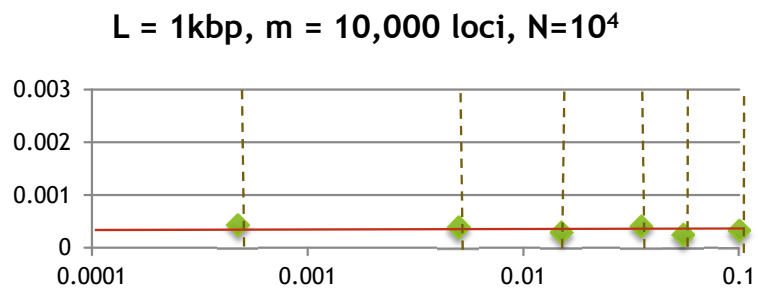
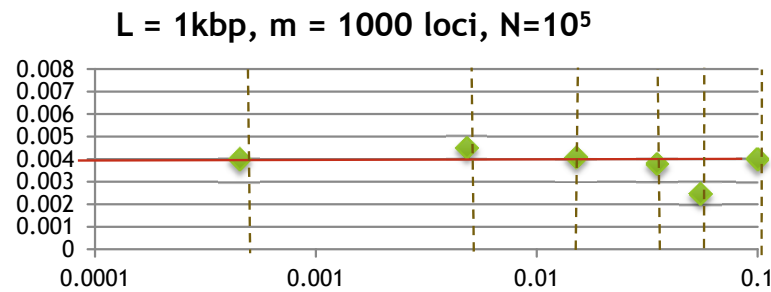
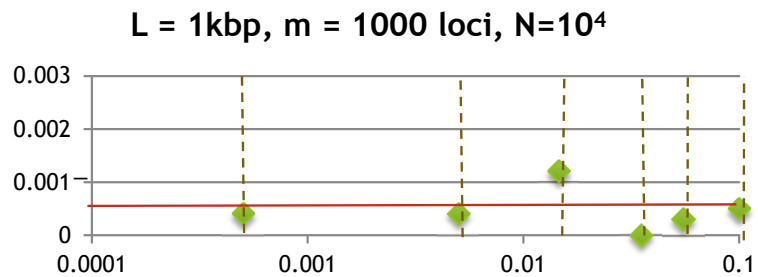
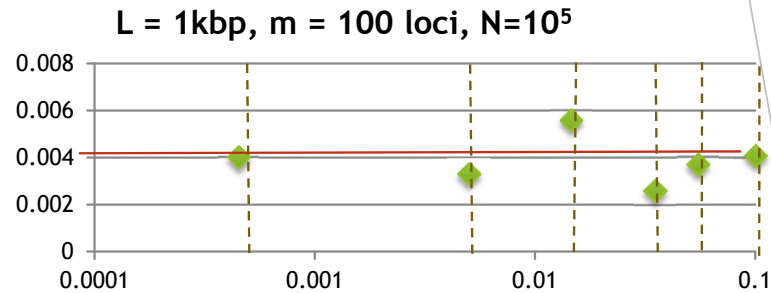
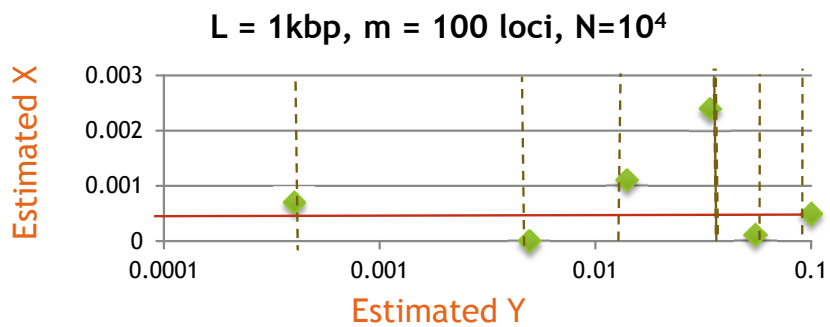


The deviation of synonymous divergence at each locus from the mean



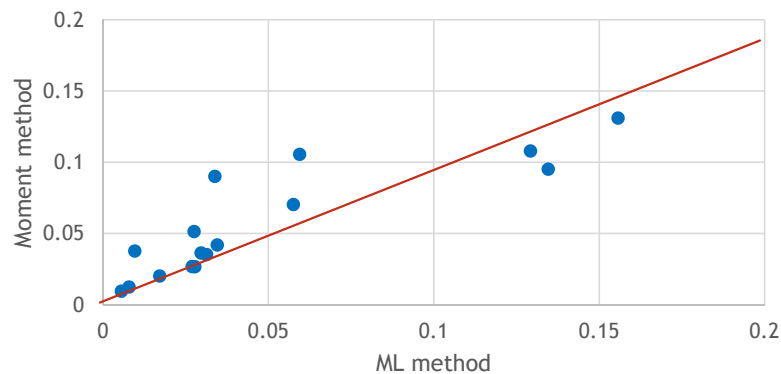
Measured by an index of $b = (d_i - M) - 0 / (\text{Var}(d_i - M))^{1/2}$

Comparison of ML estimate accuracy in simulated data sets

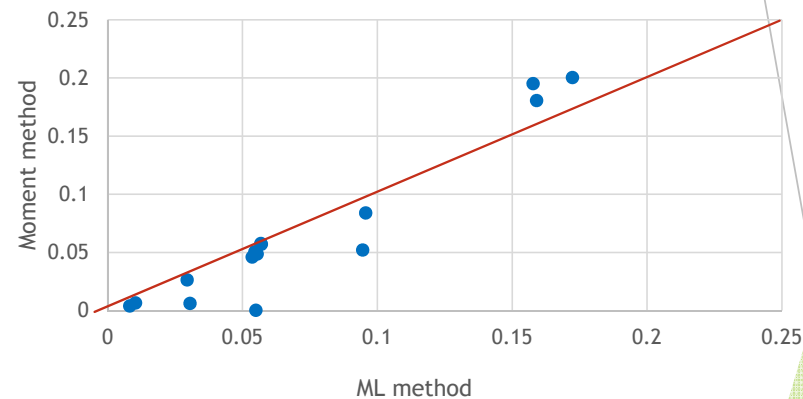


Comparison between moment method and ML method

X estimate:
ML (x axis) vs. moment (y Axis) methods



Y estimate:
ML (x axis) vs. moment (y axis) methods

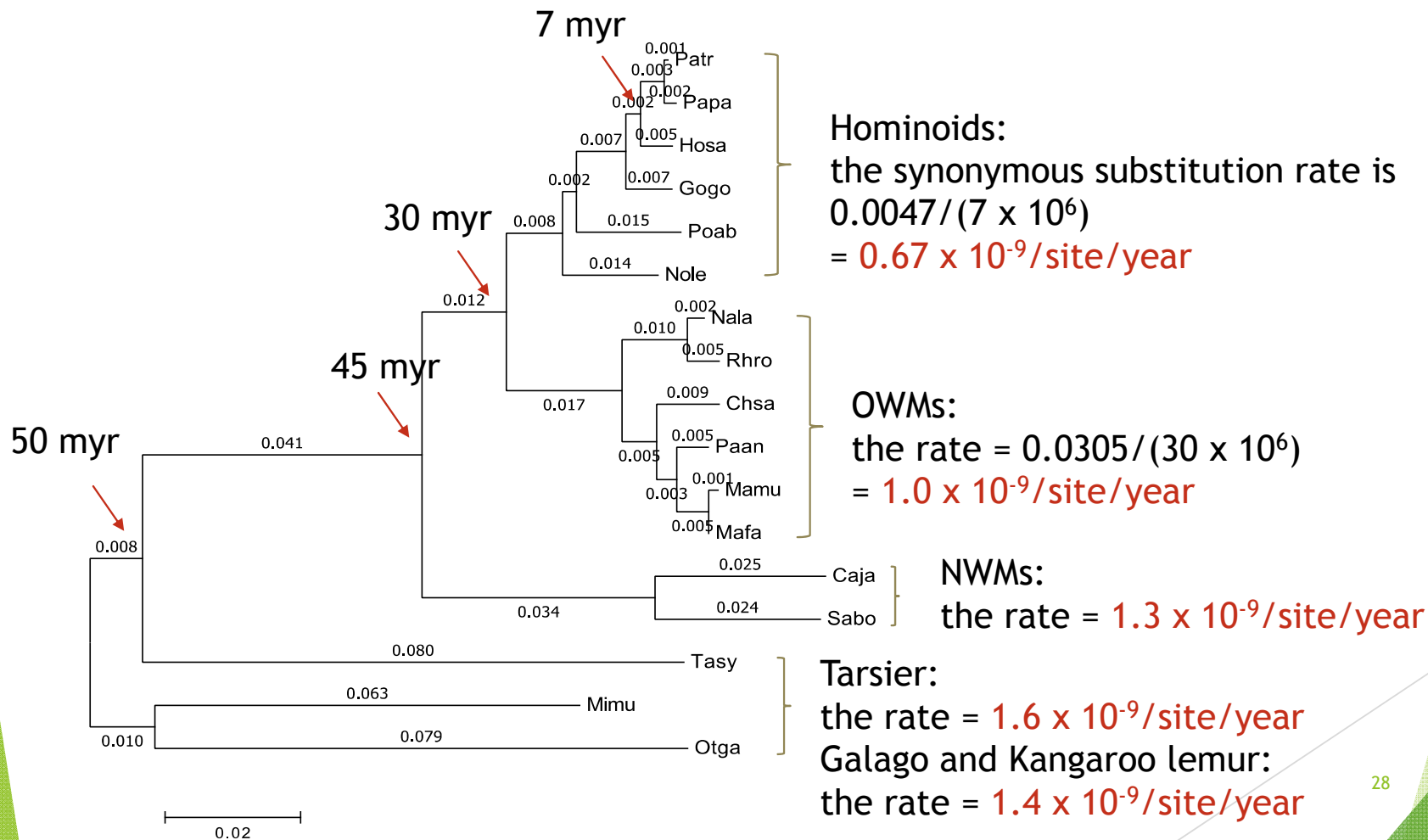


For the moment method (Takahata 1986), $X = (V - M/n)^{1/2}$, $Y = M - X$,

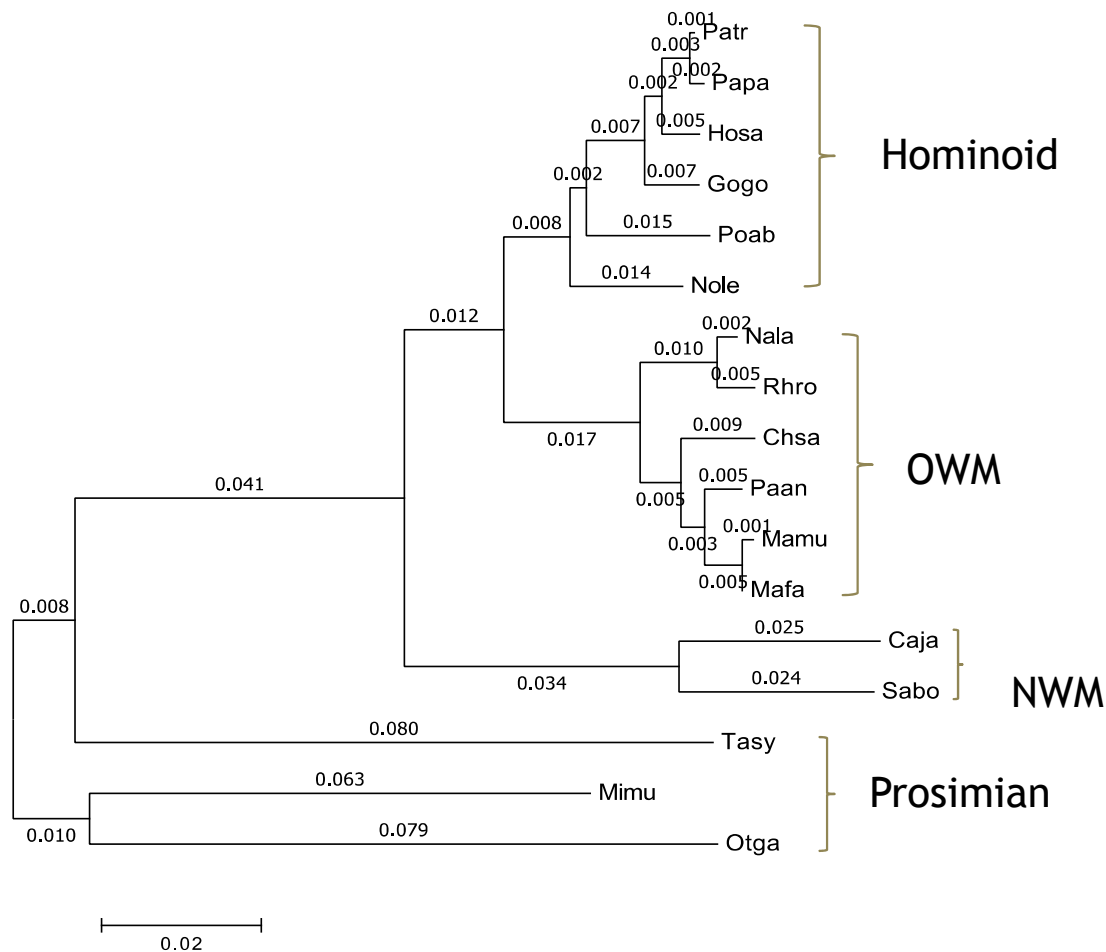
where M and V is the average and variance of the divergence.

$X = 4N_e\mu g$ and $Y = 2\mu t_s$, μ is the mutation rate, g is the generation time in year and t_s is the species divergence time

NJ tree based on ML-y estimates



ML-x and -y estimates



Compared to OWMs, NWMs, prosimians, all the branch lengths leading to hominoids are relatively short.

One possibility is that the synonymous substitution rate in hominoids slowed down compared to other primates.

The larger ancestral population size previously obtained for distantly related species may result from relatively higher substitution rates in the OWM and NWM species.

Summary

1. To get an accurate phylogeny, it is necessary to take proper account of polymorphism in the ancestral population.
2. In previous studies, the ancestral population size was inferred from the estimated ancestral polymorphism and the size tended to increase as distantly related primate species were compared.
3. Can the slower synonymous (or neutral) substitution rate be applied to other primates?
4. Here I have demonstrated that the large estimates for the ancestral population sizes are not caused by the ML method.
5. Data seems to be heterogeneous in synonymous divergence. This is caused by the rate heterogeneity among loci?