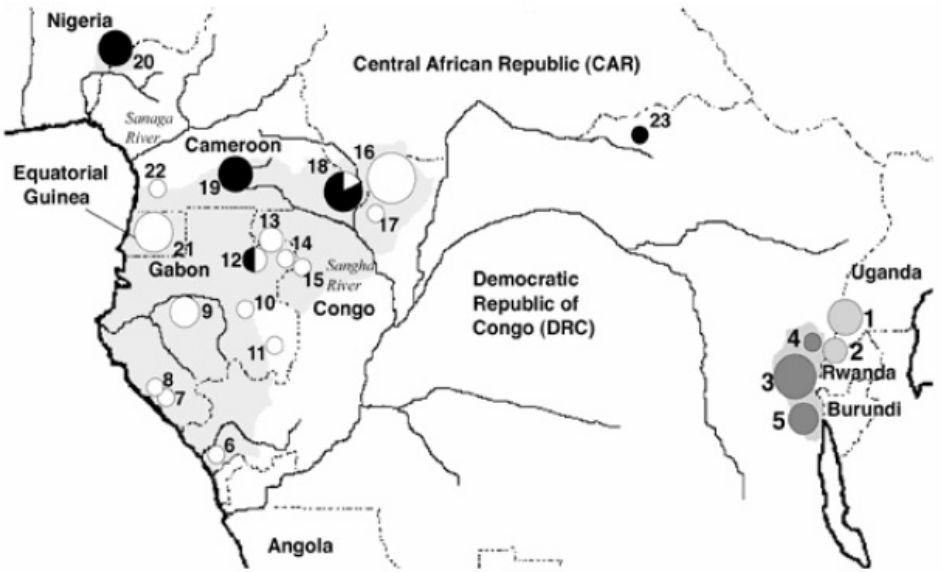
Bayesian Modeling of Numt Evolution with Application to the Estimation of Gorilla Divergence Times

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

June 29, 2008

The Modern Distribution of Gorilla Populations



- There are three subspecies of gorillas:
  - Mountain Gorillas (MTG, sites 1–2)
  - Eastern Lowland Gorillas (ELG, sites 3–5)
  - Western Lowland Gorillas (WLG, sites 6–23)
- Eastern gorillas (ELG and MTG) are physically separated from western gorillas (WLG) by at least 850 km.
- Gorillas live within forests and are not observed crossing open savannah.

Gorilla Phylogeopgraphy

- How did the modern distribution of gorillas arise?
- The *Pleistocene refugia theory*:
  - The Pleistocene was a period of frequent glaciation from about 1.8 million to about 11,000 years ago.
  - During times of maximal glaciation, central Africa would have been arid and forests would have fragmented.
  - Gorilla populations may have been restricted to small refugia.
- Did the current split between eastern and western gorillas originate during the Pleistocene?

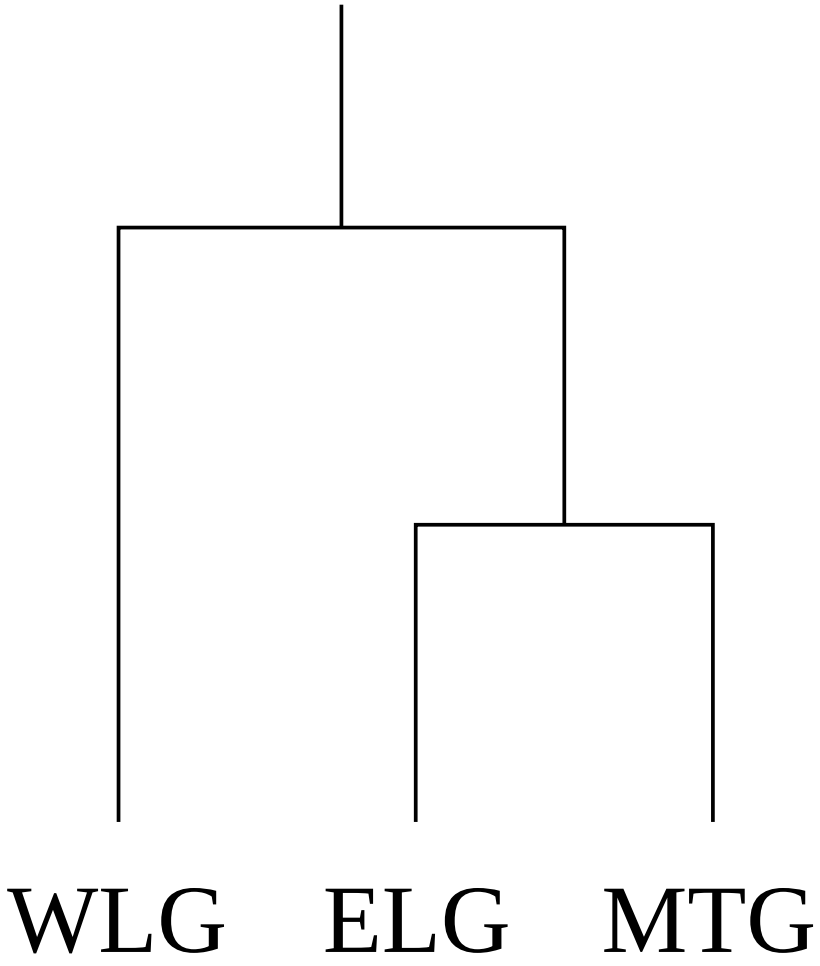
Gorilla Phylogeopgraphy

- How did the modern distribution of gorillas arise?
- The *Pleistocene refugia theory*:
  - The Pleistocene was a period of frequent glaciation from about 1.8 million to about 11,000 years ago.
  - During times of maximal glaciation, central Africa would have been arid and forests would have fragmented.
  - Gorilla populations may have been restricted to small refugia.
- Did the current split between eastern and western gorillas originate during the Pleistocene?

Gorilla Phylogeopgraphy

- How did the modern distribution of gorillas arise?
- The *Pleistocene refugia theory*:
  - The Pleistocene was a period of frequent glaciation from about 1.8 million to about 11,000 years ago.
  - During times of maximal glaciation, central Africa would have been arid and forests would have fragmented.
  - Gorilla populations may have been restricted to small refugia.
- Did the current split between eastern and western gorillas originate during the Pleistocene?

The Phylogeographic Problem



- We wish to address questions related to the phylogeographic history of gorilla populations on the basis of genetic sequence data sampled today.

HV1 sequences

- Many population studies of humans and other primates are based on DNA sequence data from the *first hyper-variable region (HV1)* of the control region of the mitochondrial genome.
- The rate of substitution in HV1 is relatively high.
- HV1 is thought to evolve neutrally.
- The mitochondrial genome is inherited maternally and there is not thought to be any recombination.
- Within the last few years, scientists have developed the technology to extract mtDNA from hairs that gorillas shed in night nests.
- This means that the DNA of gorillas *can be sampled noninvasively from wild populations*.

HV1 sequences

- Many population studies of humans and other primates are based on DNA sequence data from the *first hyper-variable region (HV1)* of the control region of the mitochondrial genome.
- The rate of substitution in HV1 is relatively high.
- HV1 is thought to evolve neutrally.
- The mitochondrial genome is inherited maternally and there is not thought to be any recombination.
- Within the last few years, scientists have developed the technology to extract mtDNA from hairs that gorillas shed in night nests.
- This means that the DNA of gorillas *can be sampled noninvasively from wild populations*.

Numt Sequences

- Sometimes, the PCR reaction to sequence HV1 (or other mitochondrial DNA) produces multiple sequences from one individual.
- This can be caused by the existence of *Numts*, or *nuclear DNA of mitochondrial origin*.
- Numts are thought to have arisen from the introgression of mitochondrial DNA into the nuclear genome in the past.
- From analysis of complete human and chimpanzee genomes, more than 400 Numts have been discovered. (“A Comparative Analysis of Numt Evolution in Human and Chimpanzee”, Hazkani-Covo and Graur, 2006).
- Only a small fraction of these would amplify using primers for HV1.
- Most of these Numts are shared between human and chimps, but several dozen have originated in each species after their split.
- Gorilla HV1 samples include putative Numts.

Numt Sequences

- Sometimes, the PCR reaction to sequence HV1 (or other mitochondrial DNA) produces multiple sequences from one individual.
- This can be caused by the existence of *Numts*, or *nuclear DNA of mitochondrial origin*.
- Numts are thought to have arisen from the introgression of mitochondrial DNA into the nuclear genome in the past.
- From analysis of complete human and chimpanzee genomes, more than 400 Numts have been discovered. ("A Comparative Analysis of Numt Evolution in Human and Chimpanzee", Hazkani-Covo and Graur, 2006).
- Only a small fraction of these would amplify using primers for HV1.
- Most of these Numts are shared between human and chimps, but several dozen have originated in each species after their split.
- Gorilla HV1 samples include putative Numts.

Numt Sequences

- Sometimes, the PCR reaction to sequence HV1 (or other mitochondrial DNA) produces multiple sequences from one individual.
- This can be caused by the existence of *Numts*, or *nuclear DNA of mitochondrial origin*.
- Numts are thought to have arisen from the introgression of mitochondrial DNA into the nuclear genome in the past.
- From analysis of complete human and chimpanzee genomes, more than 400 Numts have been discovered. (“A Comparative Analysis of Numt Evolution in Human and Chimpanzee”, Hazkani-Covo and Graur, 2006).
- Only a small fraction of these would amplify using primers for HV1.
- Most of these Numts are shared between human and chimps, but several dozen have originated in each species after their split.
- Gorilla HV1 samples include putative Numts.

Numt Sequences

- Sometimes, the PCR reaction to sequence HV1 (or other mitochondrial DNA) produces multiple sequences from one individual.
- This can be caused by the existence of *Numts*, or *nuclear DNA of mitochondrial origin*.
- Numts are thought to have arisen from the introgression of mitochondrial DNA into the nuclear genome in the past.
- From analysis of complete human and chimpanzee genomes, more than 400 Numts have been discovered. (“A Comparative Analysis of Numt Evolution in Human and Chimpanzee”, Hazkani-Covo and Graur, 2006).
- Only a small fraction of these would amplify using primers for HV1.
- Most of these Numts are shared between human and chimps, but several dozen have originated in each species after their split.
- Gorilla HV1 samples include putative Numts.

Objectives

- We wish to develop a model *based on population genetics* for the relationship among HV1 and Numt sequences sampled from gorilla populations.
- Within the framework of this model, we wish to:
  - estimate divergence times of African gorilla populations;
  - estimate how many introgressions occurred and when they occurred in the history of the sampled data;
  - estimate effective population sizes, and evolutionary parameters such as substitution rates and transition-transversion ratio.

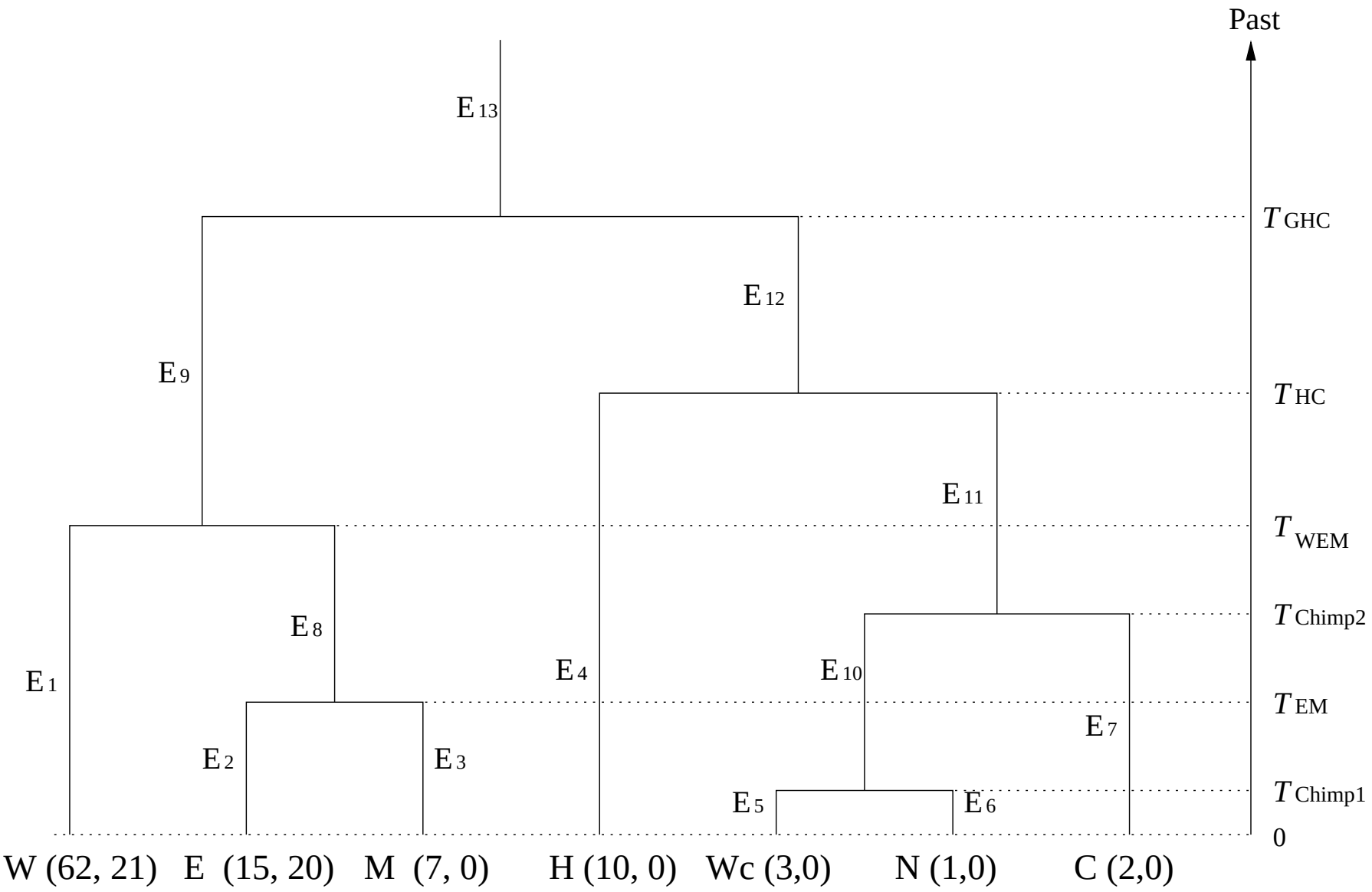
Data

- We analyze an alignment of 141 sequences with 236 sites.
- We include human and chimpanzee sequences in order to calibrate divergence times.
- There are:
  - 125 Gorilla sequences (Anthony *et al.*, 2006);
  - 10 Human sequences (sampled from Ingman *et al.*, 2000);
  - 6 Chimpanzee sequences (from Hu *et al.*, 2001; Thalmann *et al.*, 2004);

Data Table

Group	Symbol	HV1	Numt
Western lowland gorilla	W	62	21
Eastern lowland gorilla	E	15	20
Mountain gorilla	M	7	0
Western common chimpanzee	$\mathrm{W_C}$	3	0
Central common chimpanzee	C	2	0
Nigerian chimpanzee	N	1	0
Human	H	10	0
Total		100	41

The Population Tree



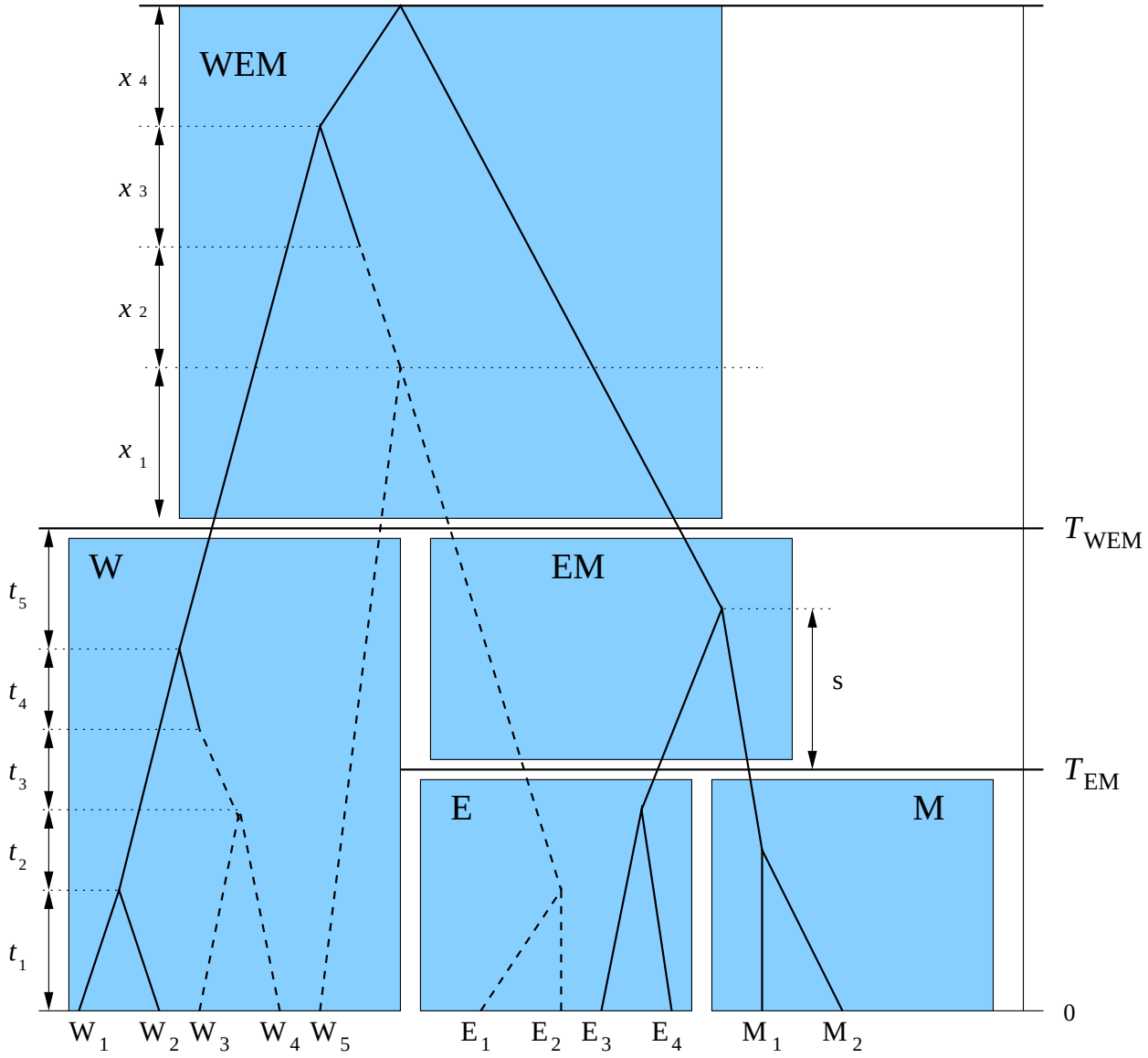
The Hybrid Coalescent Process

- We have a population tree relating populations.
- There are sequences of two types at the tips of the population tree.
- We model three separate processes in each edge of the tree:
  - an HV1 coalescent process with rate θ_{HV1};
  - a Numt coalescent process with rate θ_{Numt};
  - a Numt transfer process (introgression in reverse time) with rate η.
- This provides a likelihood model for a sequence tree within a population tree framework.
- We model coalescent rates directly, but can infer *effective population sizes*.

The Hybrid Coalescent Process

- We have a population tree relating populations.
- There are sequences of two types at the tips of the population tree.
- We model three separate processes in each edge of the tree:
  - an HV1 coalescent process with rate θ_{HV1};
  - a Numt coalescent process with rate θ_{Numt};
  - a Numt transfer process (introgression in reverse time) with rate η.
- This provides a likelihood model for a sequence tree within a population tree framework.
- We model coalescent rates directly, but can infer *effective population sizes*.

Example


WEM
W
EM
E
M
x_4
x_3
x_2
x_1
t_5
t_4
t_3
t_2
t_1
s
T_{WEM}
T_{EM}
0
W_1 W_2 W_3 W_4 W_5
E_1 E_2 E_3 E_4
M_1 M_2

Substitution Model

- Separate HKY Models for HV1 and Numts

$$Q = \mu\phi \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix}$$

where

- π_{A}, π_{C}, π_{G}, π_{T}: relative frequencies for each nucleotide base.
- κ is the transition-transversion parameter.
- $\phi = 1/(2(\kappa(\pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}}) + \pi_{\mathrm{R}}\pi_{\mathrm{Y}}))$, a scaling parameter.
- μ is the number of substitutions per site per million years.

Bayesian Estimation of Parameters

- State space ($\boldsymbol{T}$, Θ, G)
  - $\boldsymbol{T}$=(T_{EM}, T_{WEM}, T_{Chimp1}, T_{Chimp2}, T_{HC}, T_{GHC}).
  - Θ=(θ_1,..., θ_{13},λ_θ, η, μ_{HV1}, μ_{Numt}, κ_{HV1}, κ_{Numt}).
  - G= Gene genealogy determined by the hybrid coalescent process.
- Target distribution

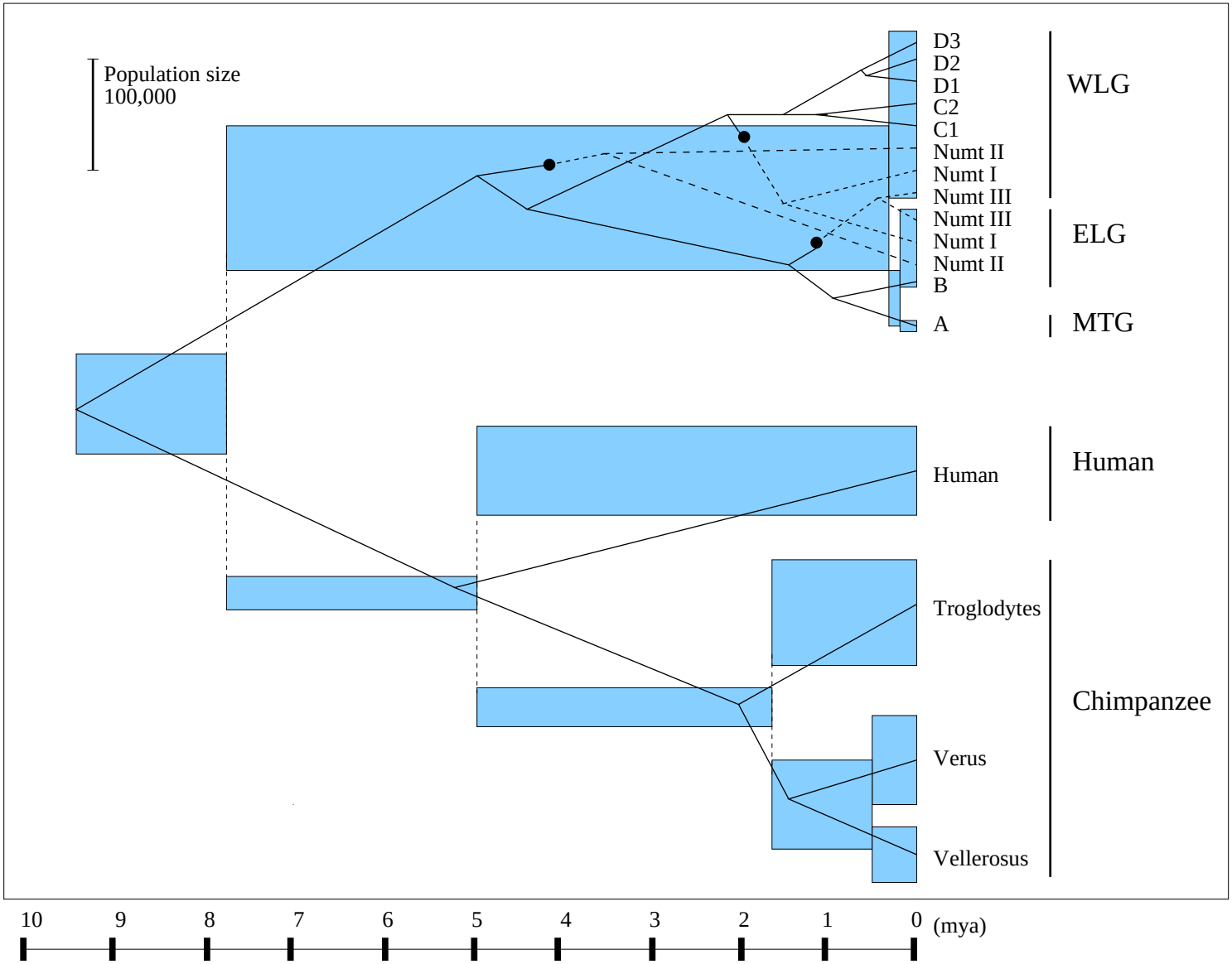
$$f(\boldsymbol{T}, \Theta, G \mid D) = \frac{f(D \mid \boldsymbol{T}, \Theta, G) f(G \mid \boldsymbol{T}, \Theta) f(\boldsymbol{T}) f(\Theta)}{\int_{(\boldsymbol{T}, \Theta, G)} f(D \mid \boldsymbol{T}, \Theta, G) f(G \mid \boldsymbol{T}, \Theta) f(\boldsymbol{T}) f(\Theta)}$$

- Use MCMC over this space.

MCMC Approaches

- Update population divergence time $\boldsymbol{T}$ and gene genealogy G:
  - Multiply a constant to the whole trees.
  - Update population divergence times.
  - Update histories in a population edge.
    - Generate new event times in a population edge.
    - Relocate an event in a population edge.
    - Change a pair of coalescent events.
  - Update the total number of transfer events
- Update parameters in Θ:
  - Propose a new value by multiplying a scale factor from Gamma(2,2) distribution to the current value.

Results (100 HV1 + 41 Numt Sequences)



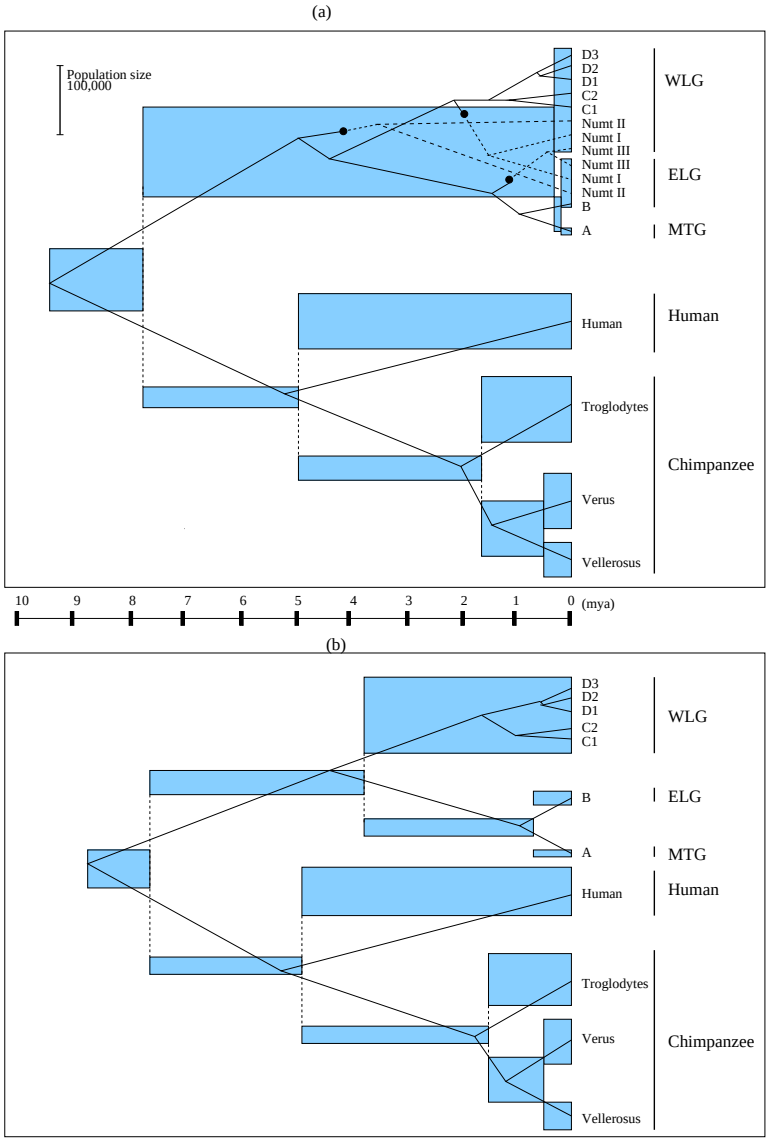
Robustness?

- The analysis of the full data set implies a fairly recent divergence time between eastern and western gorillas (95% credible region 110,000 to 610,000 million years ago).
- However, the earliest east/west coalescents are among Numt sequences.
- How robust are the estimated times if the Numt data is removed?

Robustness?

- The analysis of the full data set implies a fairly recent divergence time between eastern and western gorillas (95% credible region 110,000 to 610,000 million years ago).
- However, the earliest east/west coalescents are among Numt sequences.
- How robust are the estimated times if the Numt data is removed?

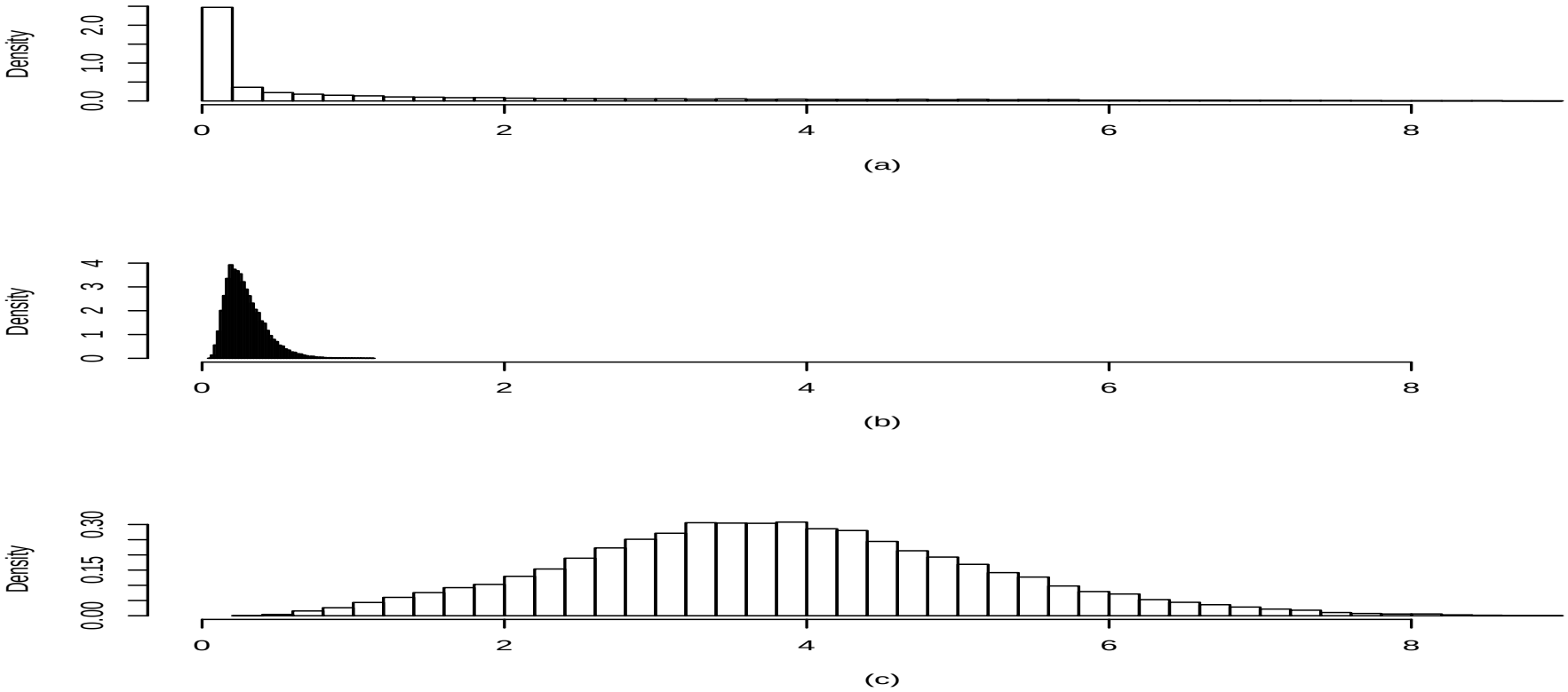
HV1 + Numt (top) versus HV1 only (bottom)



Estimated Population Divergence Times

Time	Numt-HV1 case			only-HV1 case		
	mean	median	95% C.R.	mean	median	95% C.R.
T_{EM}	0.190	0.170	(0.045, 0.430)	0.67	0.63	(0.14, 1.50)
T_{WEM}	0.29	0.26	(0.11, 0.61)	3.8	3.8	(1.3, 6.7)
T_{Chimp1}	0.53	0.38	(0, 1.8)	0.52	0.38	(0, 1.7)
T_{Chimp2}	1.60	1.60	(0.26, 2.90)	1.50	1.50	(0.36, 2.70)
T_{HC}	4.9	4.8	(4.0, 5.9)	4.9	4.9	(4.0, 5.9)
T_{GHC}	7.7	7.9	(5.4, 9.0)	7.6	7.8	(5.3, 8.9)

East-West Split Time Posterior Distributions



Conjecture

- What is a biologically plausible reason for the discordance in the results?
- We conjecture that *differences in male and female gorilla migratory behavior* could be the reason for the difference.
- If there is some male migration between east and west long after female migration ends, then we would expect that nuclear genes would coalesce much more recently than mitochondrial genes.
- We address this indirectly by modeling separate divergence times for HV1 and Numt sequences in an extended model.

Conjecture

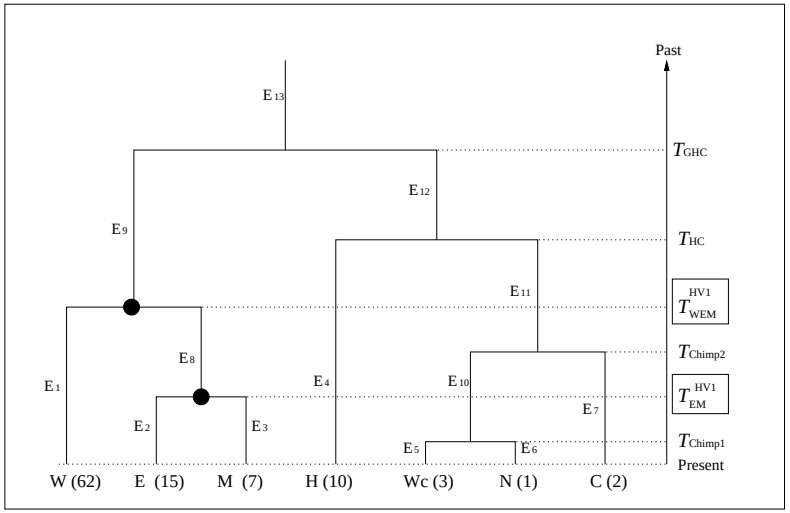
- What is a biologically plausible reason for the discordance in the results?
- We conjecture that *differences in male and female gorilla migratory behavior* could be the reason for the difference.
- If there is some male migration between east and west long after female migration ends, then we would expect that nuclear genes would coalesce much more recently than mitochondrial genes.
- We address this indirectly by modeling separate divergence times for HV1 and Numt sequences in an extended model.

Conjecture

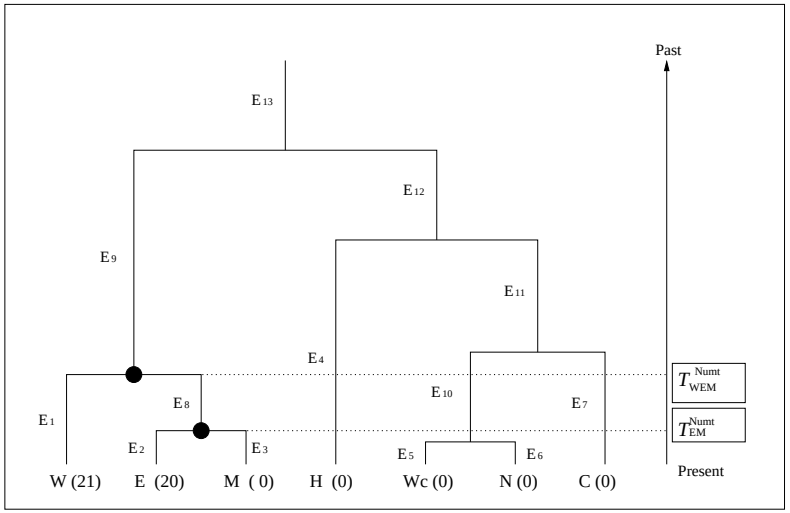
- What is a biologically plausible reason for the discordance in the results?
- We conjecture that *differences in male and female gorilla migratory behavior* could be the reason for the difference.
- If there is some male migration between east and west long after female migration ends, then we would expect that nuclear genes would coalesce much more recently than mitochondrial genes.
- We address this indirectly by modeling separate divergence times for HV1 and Numt sequences in an extended model.

A Genome-differentiated Population Tree Model

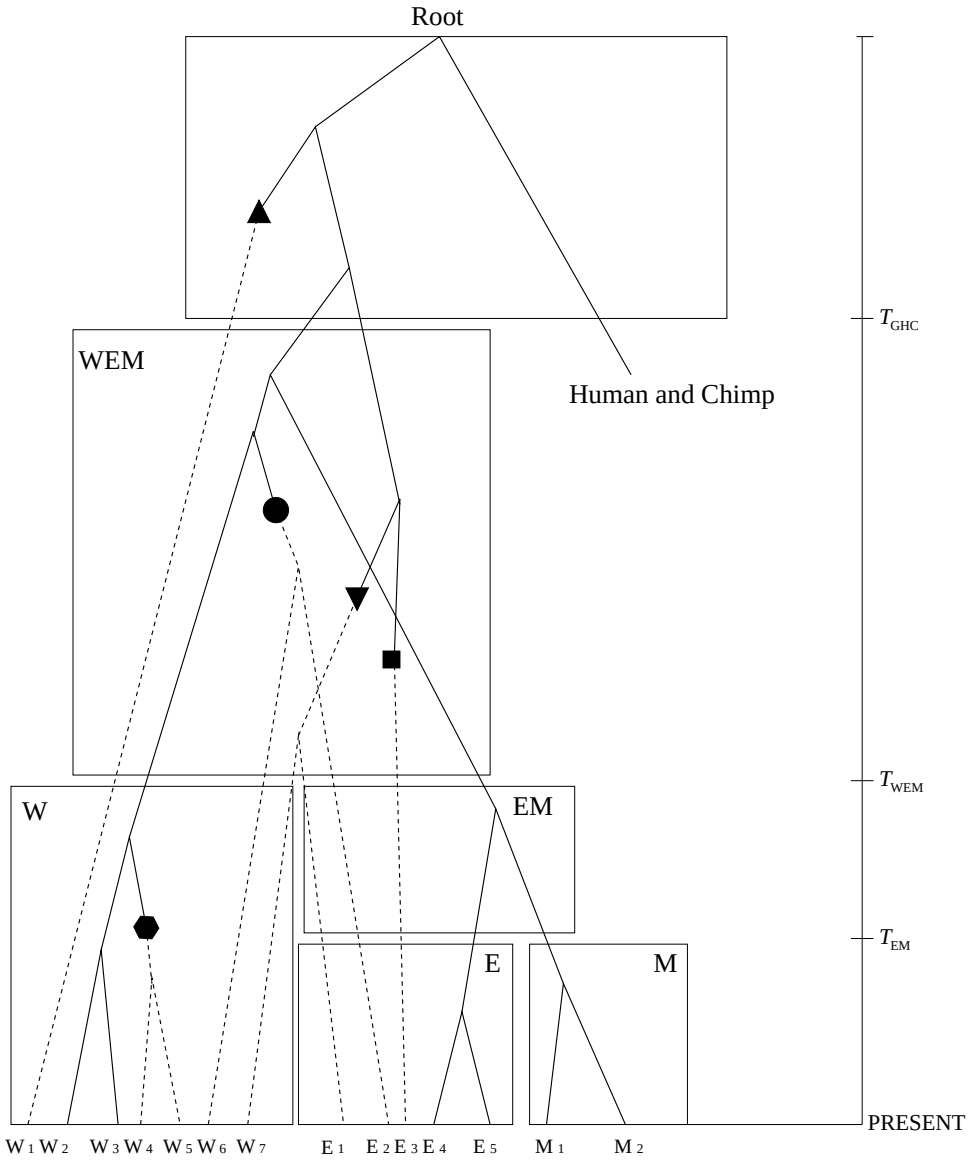
(a) HV1 Population Tree



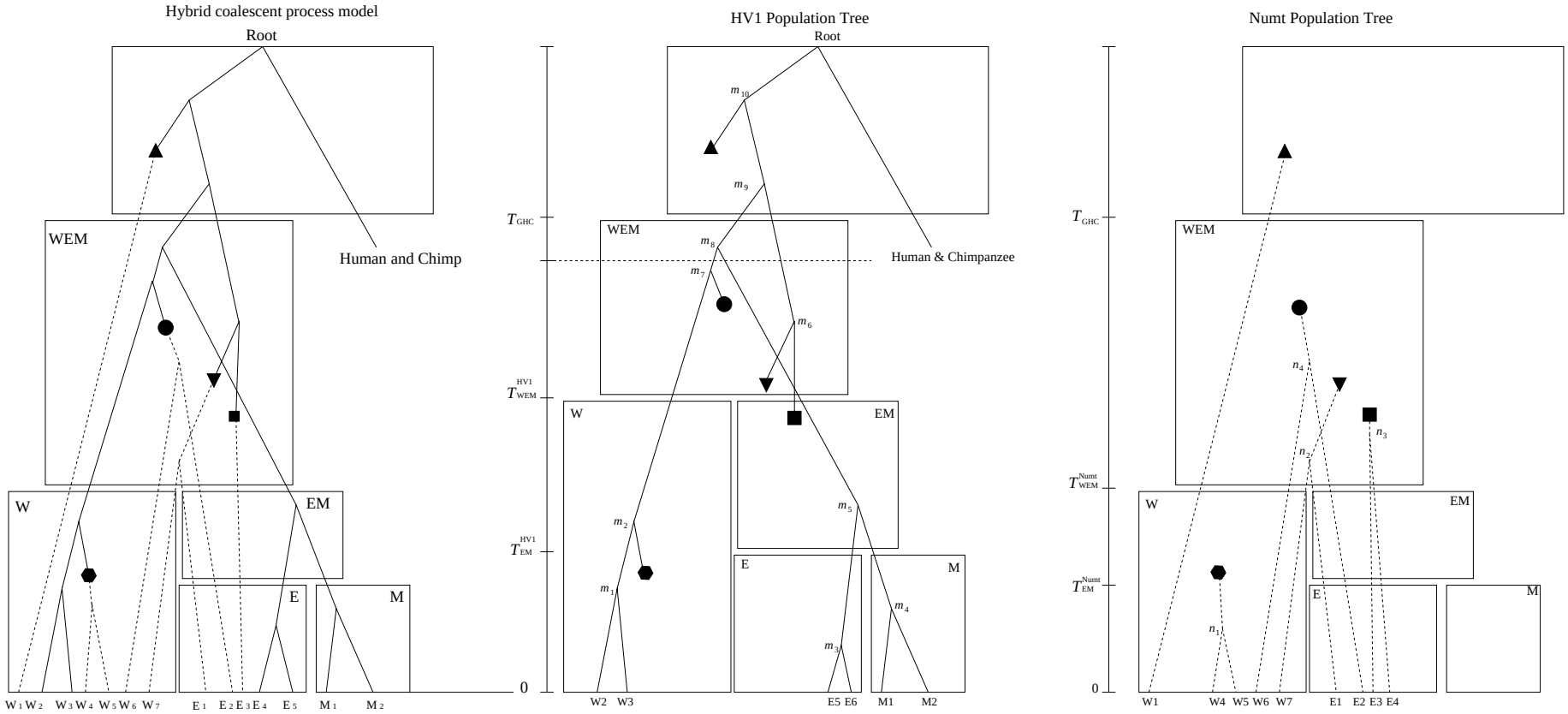
(b) Numt Population Tree



Example Gene Genealogy (old)



Example Gene Genealogy (old and new)



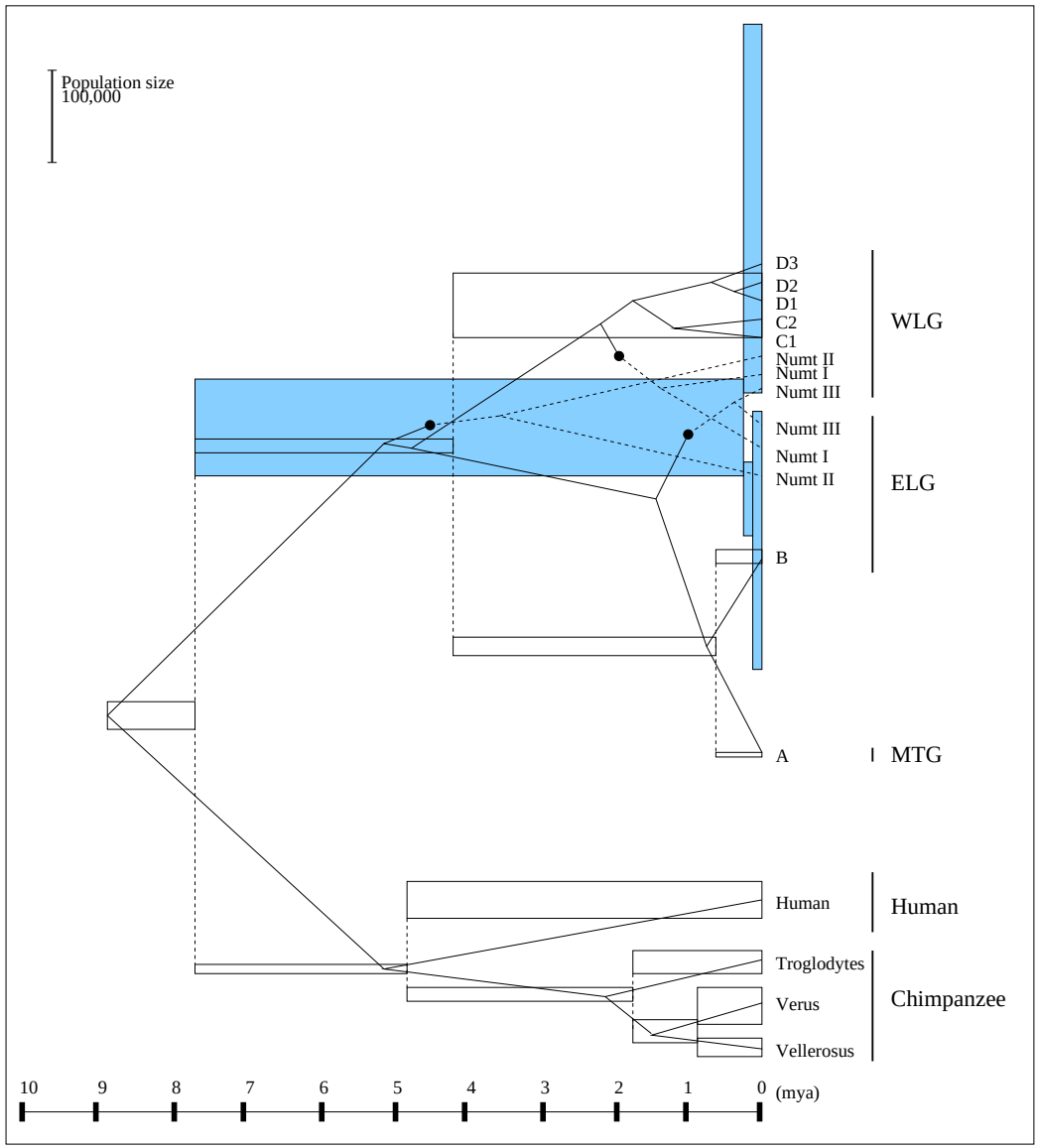
Bayesian Estimation of Parameters

- State space ($\boldsymbol{T}$, Θ, G)
  - $\boldsymbol{T}$=($T_{\mathrm{EM}}^{\mathrm{HV1}}$, $T_{\mathrm{EMW}}^{\mathrm{HV1}}$, T_{Chimp1}, T_{Chimp2}, T_{HC}, T_{GHC}, $T_{\mathrm{EM}}^{\mathrm{Numt}}$, $T_{\mathrm{EMW}}^{\mathrm{Numt}}$).
  - Θ=($\left\{\theta_i^{\mathrm{HV1}}\right\}$, $\left\{\theta_j^{\mathrm{Numt}}\right\}$, λ_{HV1}, λ_{Numt}, η, μ_{HV1}, μ_{Numt}, κ_{HV1}, κ_{Numt})
  - G= Gene genealogy over the genome-differentiated population trees.
- Target distribution

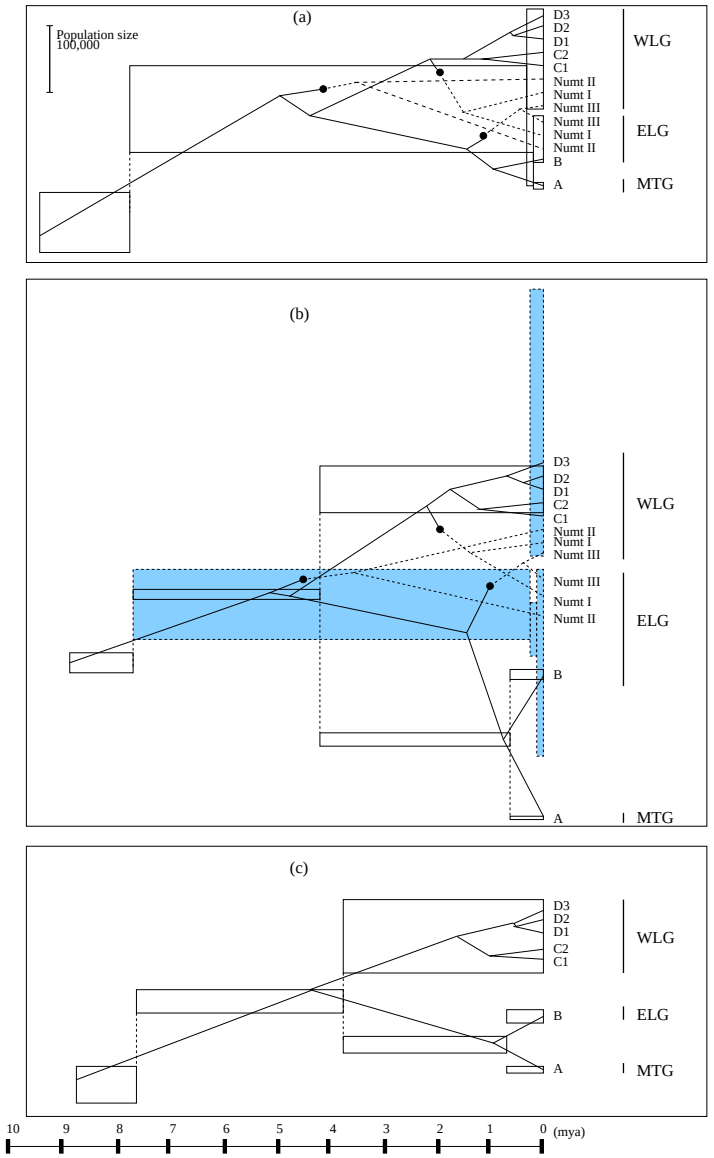
$$f(\boldsymbol{T}, \Theta, G \mid D) = \frac{f(D \mid \boldsymbol{T}, \Theta, G) f(G \mid \boldsymbol{T}, \Theta) f(\boldsymbol{T}) f(\Theta)}{\int_{(\boldsymbol{T}, \Theta, G)} f(D \mid \boldsymbol{T}, \Theta, G) f(G \mid \boldsymbol{T}, \Theta) f(\boldsymbol{T}) f(\Theta)}$$

- MCMC approaches: *more complex than the hybrid coalescent process model*

Summary


Population size
100,000
D3
D2
D1
C2
C1
Numt II
Numt I
Numt III
WLG
Numt III
Numt I
Numt II
ELG
B
A
MTG
Human
Human
Troglodytes
Verus
Vellerosus
Chimpanzee
10
9
8
7
6
5
4
3
2
1
0
(mya)

Comparison


(a)
Population size
100,000
D3
D2
D1
C2
C1
Numt II
Numt I
Numt III
Numt III
Numt I
Numt II
B
A
WLG
ELG
MTG
(b)
D3
D2
D1
C2
C1
Numt II
Numt I
Numt III
Numt III
Numt I
Numt II
B
A
WLG
ELG
MTG
(c)
D3
D2
D1
C2
C1
B
A
WLG
ELG
MTG
10 9 8 7 6 5 4 3 2 1 0 (mya)

Estimated Population Divergence Times

Time	mean	median	95% C.R.
T_{EM}^{HV1}	0.59	0.53	(0.12, 1.4)
T_{WEM}^{HV1}	4.2	4.2	(1.4, 6.9)
T_{Chimp1}	0.86	0.83	(0.0034, 2.0)
T_{Chimp2}	1.8	1.8	(0.75, 3.00)
T_{HC}	4.8	4.8	(4.0, 5.9)
T_{GHC}	7.7	7.9	(5.4, 9.0)
T_{EM}^{Numt}	0.15	0.12	(0.0024, 0.45)
T_{WEM}^{Numt}	0.25	0.23	(0.053, 0.61)

Summary

- Discordant east-west split times for Numt and HV1 sequences.
- HV1 east-west split time may predate the Pleistocene.
- Numt east-west split time probably falls within the Pleistocene.
- Male-mediated gene-flow may have persisted much longer after female east/west migration stopped.
- There are likely three separate Numt loci in this data set.

Future work

- Use more human data to better estimate HV1 substitution rates and parameters.
- Extend the current population tree to incorporate real gorilla populations.
- Topological uncertainty in the population trees.
- Changes in population sizes over time.
- Explicit migration in a population tree.
- Multiple loci DNA data.
- Subdivisions in each Numt population.
- Recombination in Numt sequences.

Acknowledgments

- The data set we analyze was constructed by *Nicola Anthony*, a biologist at the University of New Orleans.
  - She and co-workers collected many hair samples from the field;
  - She also obtained sequence data from other investigators;
  - I met Nicola when she was at UW—Madison for the semester after Hurricane Katrina which closed her university for a semester.
- The great majority of this programming and analysis was done by *Joungyoun Kim* who completed her Ph.D. with me in June, 2008.
- We were supported by the NIH.

Acknowledgments

- The data set we analyze was constructed by *Nicola Anthony*, a biologist at the University of New Orleans.
  - She and co-workers collected many hair samples from the field;
  - She also obtained sequence data from other investigators;
  - I met Nicola when she was at UW—Madison for the semester after Hurricane Katrina which closed her university for a semester.
- The great majority of this programming and analysis was done by *Joungyoun Kim* who completed her Ph.D. with me in June, 2008.
- We were supported by the NIH.

Acknowledgments

- The data set we analyze was constructed by *Nicola Anthony*, a biologist at the University of New Orleans.
  - She and co-workers collected many hair samples from the field;
  - She also obtained sequence data from other investigators;
  - I met Nicola when she was at UW—Madison for the semester after Hurricane Katrina which closed her university for a semester.
- The great majority of this programming and analysis was done by *Joungyoun Kim* who completed her Ph.D. with me in June, 2008.
- We were supported by the NIH.