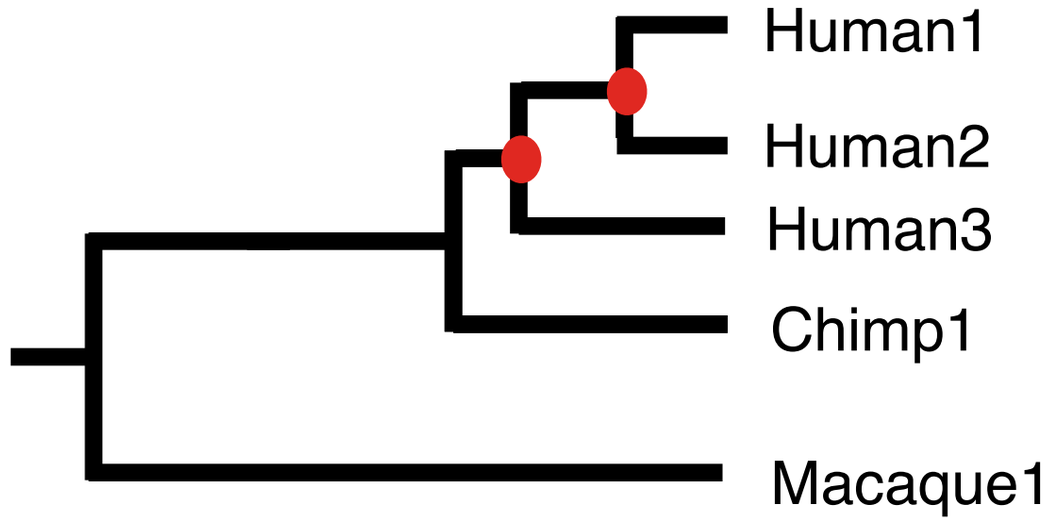
Gene duplication and loss
Part III

Matthew Hahn
Indiana University

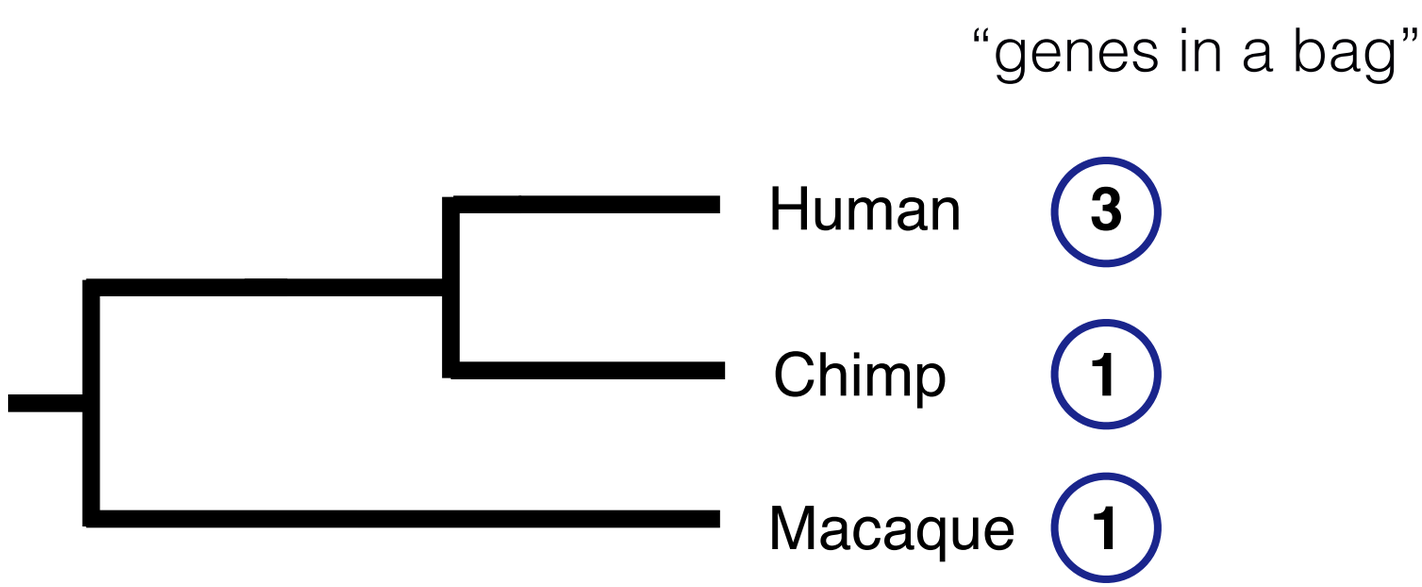
mwh@indiana.edu

Inferring duplications and losses using gene trees



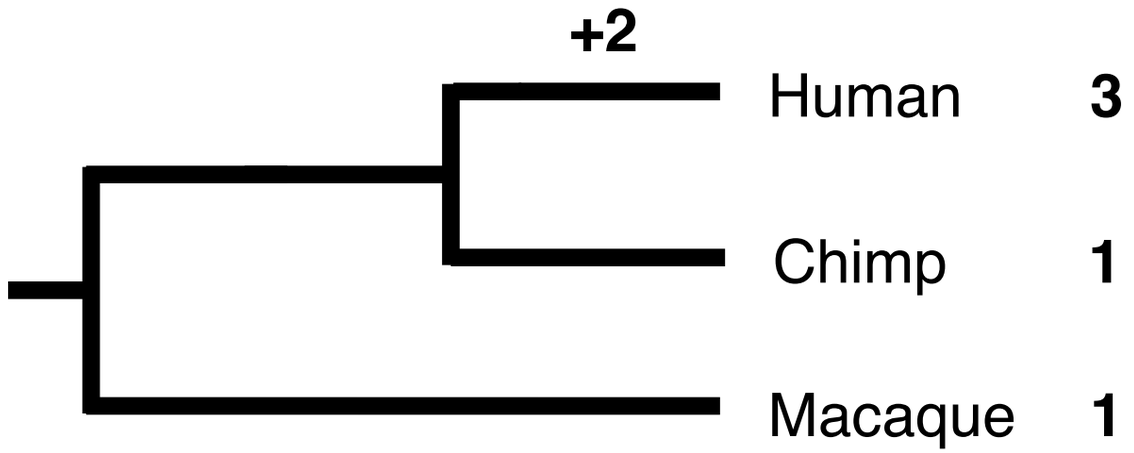
An alternative approach

"Count" methods



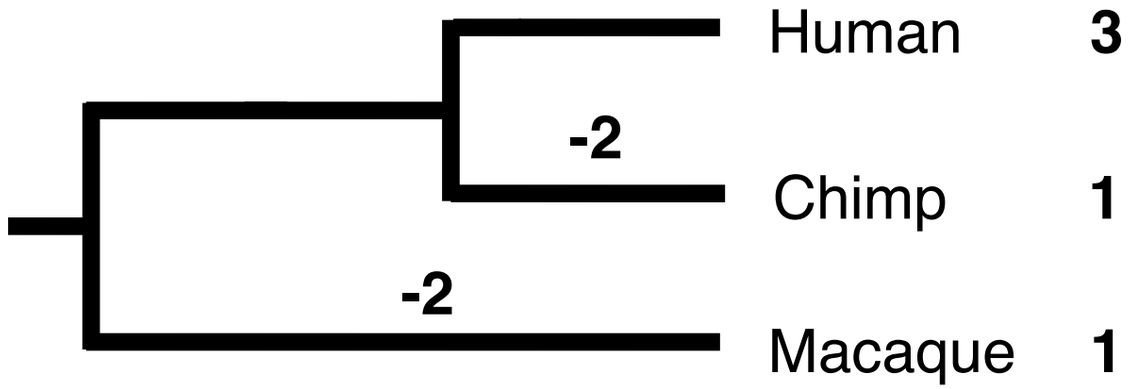
An alternative approach

"Count" methods

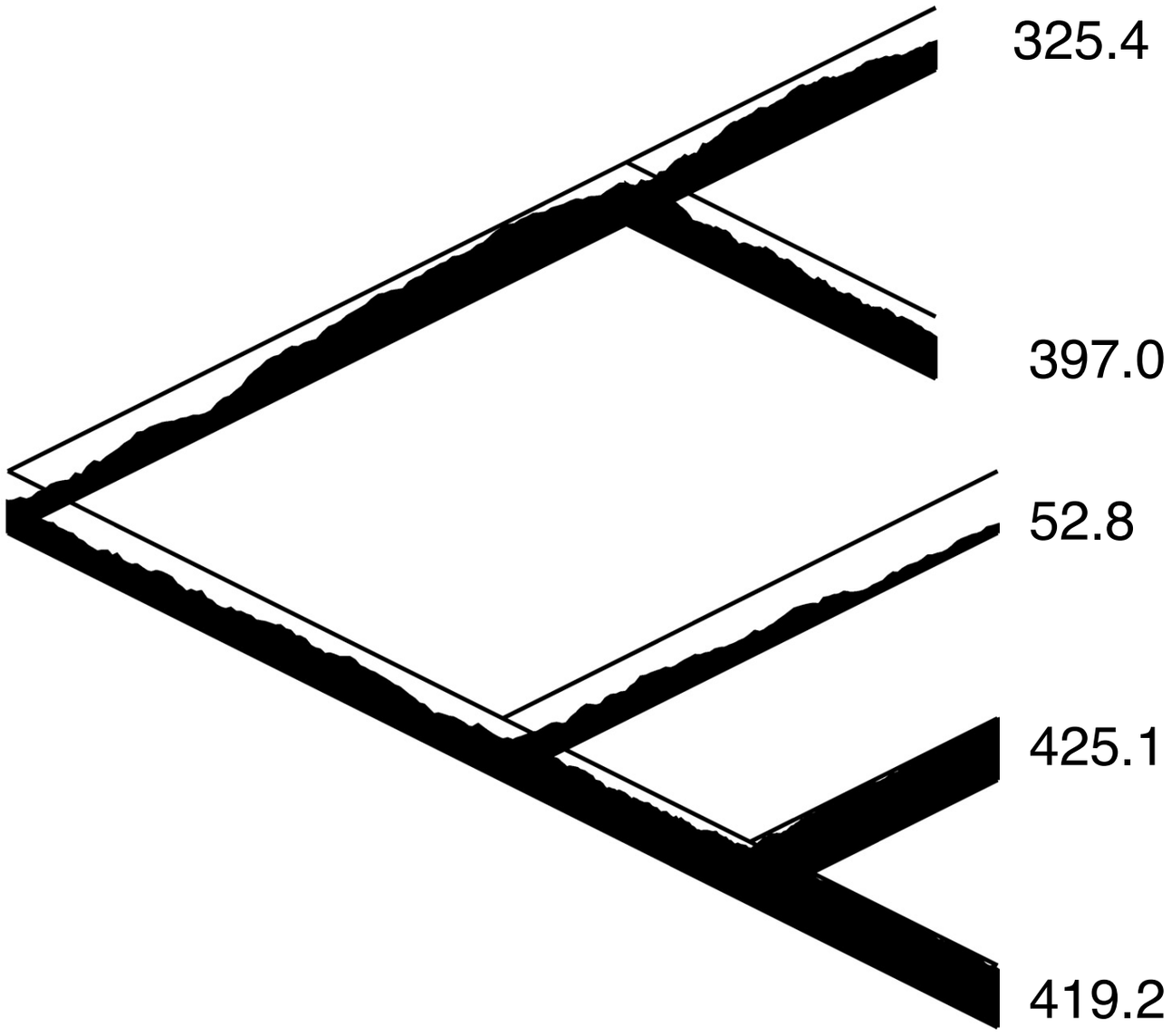


An alternative approach

"Count" methods



Quantitative trait models on trees



Felsenstein (2005)

A model for gene family evolution

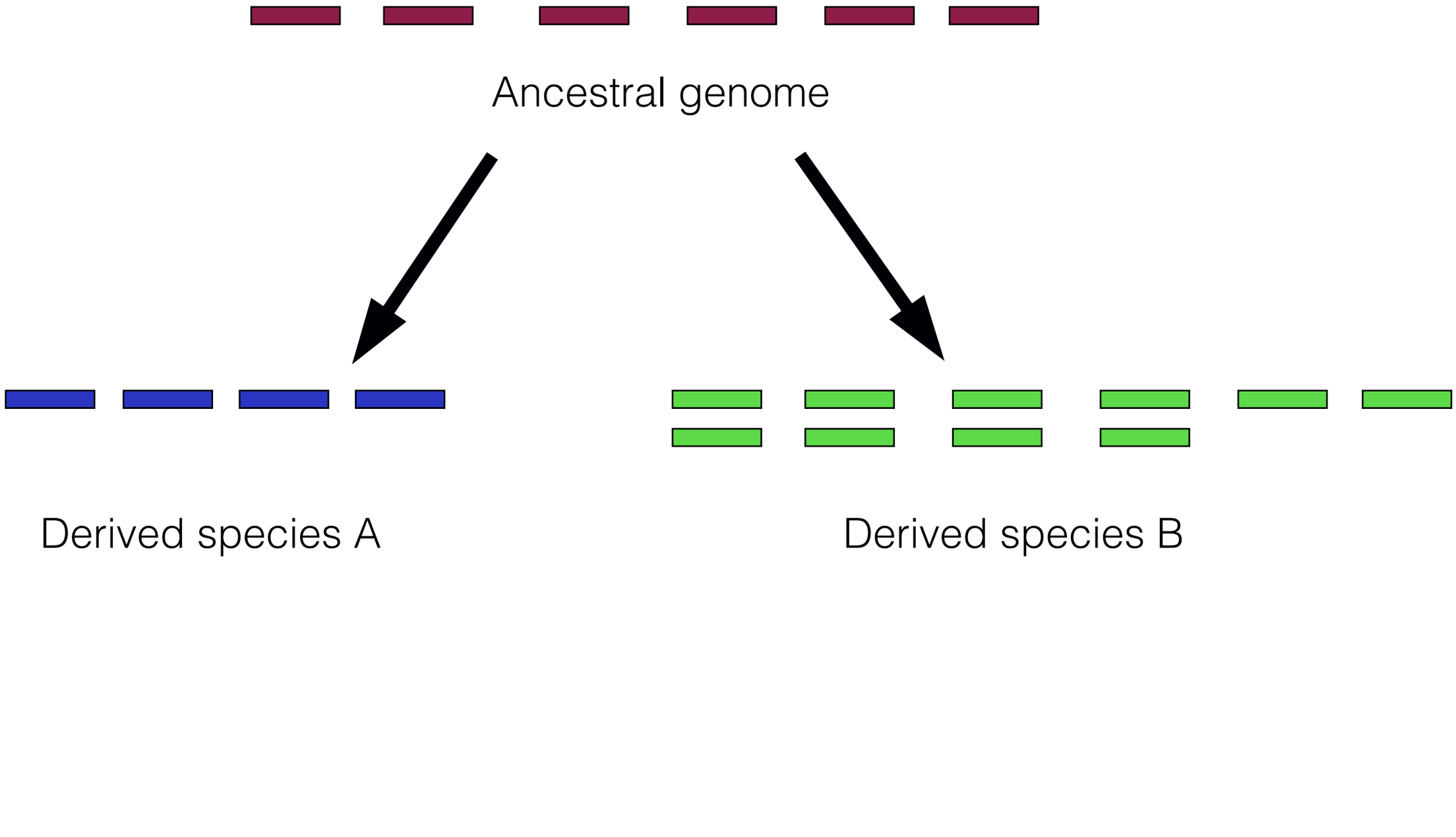
Homogeneous birth and death process

Birth = duplication

Death = loss

Hahn et al. (2005)

A model for gene family evolution



There are no true models, only helpful ones.

--G.E.P. Box

No model, no inference.

--J. Felsenstein

A model for gene family evolution

Birth-Death transition probability (Bailey 1964):

$$P(X(t) = c \mid X(0) = s) = \sum_{j=0}^{\min(s,c)} \binom{s}{j} \binom{s+c-j-1}{s-1} \alpha^{s-j} \beta^{c-j} (1-\alpha-\beta)^j$$

$$\alpha = \frac{\mu(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}, \quad \beta = \frac{\lambda(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}$$

The necessary parameters:

-Current family size

-Ancestral family size

-Time since divergence

-Birth and death rates per gene

A model for gene family evolution

Assuming birth and death rates are equal (Hahn et al. 2005):

$$P(X(t) = c|X(0) = s) = \sum_{j=0}^{\min(s,c)} \binom{s}{j}\binom{s+c-j-1}{s-1}\alpha^{s+c-2j}(1-2\alpha)^j,$$

$$\alpha = \frac{\lambda t}{1+\lambda t}$$

The necessary parameters:

-Current family size

-Ancestral family size

-Time since divergence

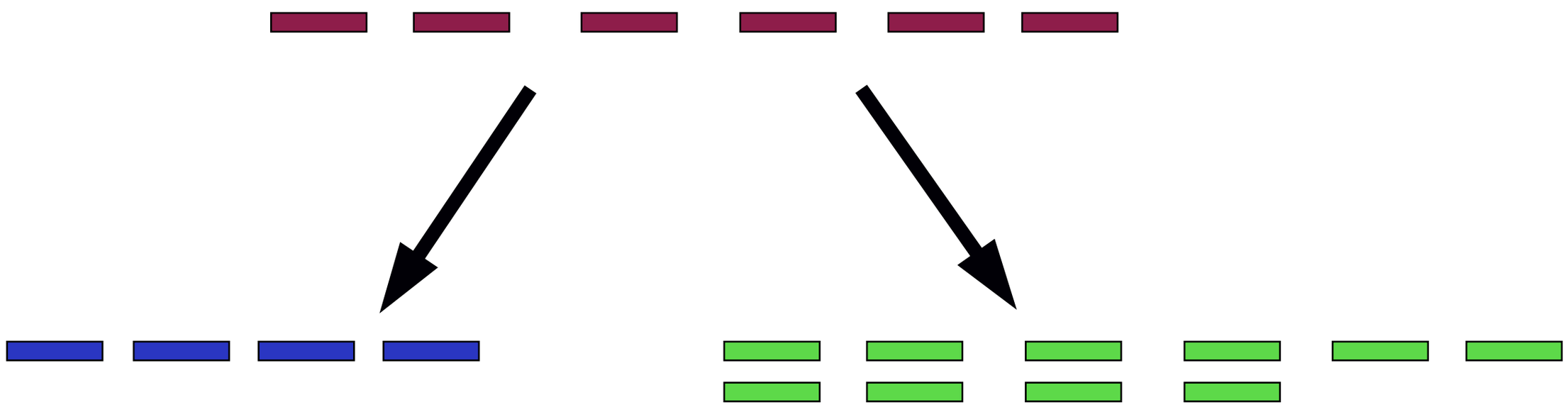
-Birth and death rates per gene

A model for gene family evolution

Assuming birth and death rates are equal:

$$\mathrm{Mean}(X(t)|X(0)=s)=s,$$

$$\mathrm{Var}(X(t)|X(0)=s)=2s\lambda t.$$



A model for gene family evolution

Birth-Death transition probability (Bailey 1964):

$$P(X(t) = c \mid X(0) = s)$$
$$= \sum_{j=0}^{\min(s,c)} \binom{s}{j} \binom{s+c-j-1}{s-1} \alpha^{s-j} \beta^{c-j} (1-\alpha-\beta)^j$$

$$\alpha = \frac{\mu(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}, \quad \beta = \frac{\lambda(e^{(\lambda-\mu)t} - 1)}{\lambda e^{(\lambda-\mu)t} - \mu}$$

The necessary parameters:

-Current family size

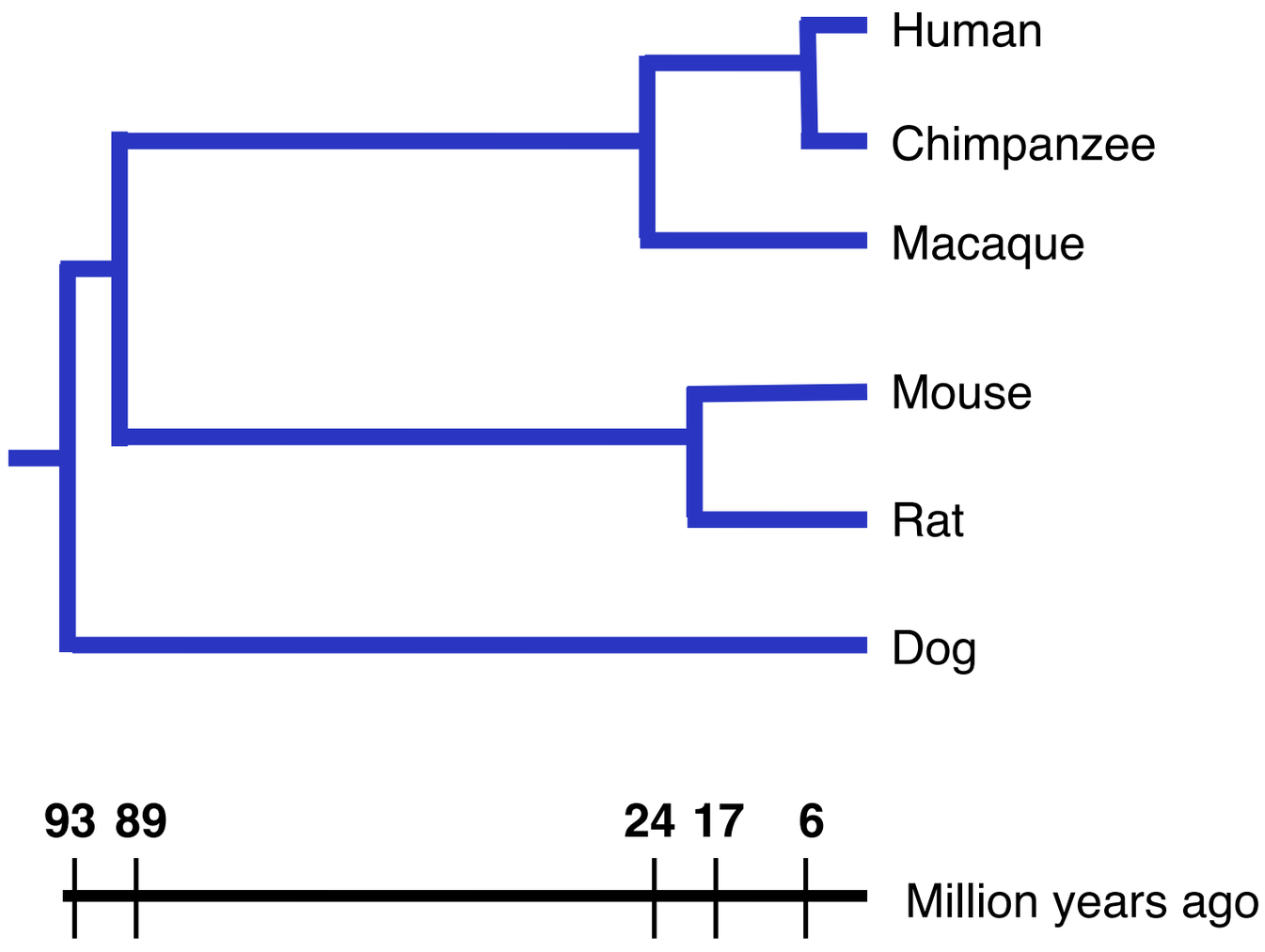
-Ancestral family size

-Time since divergence

-Birth and death rates per gene

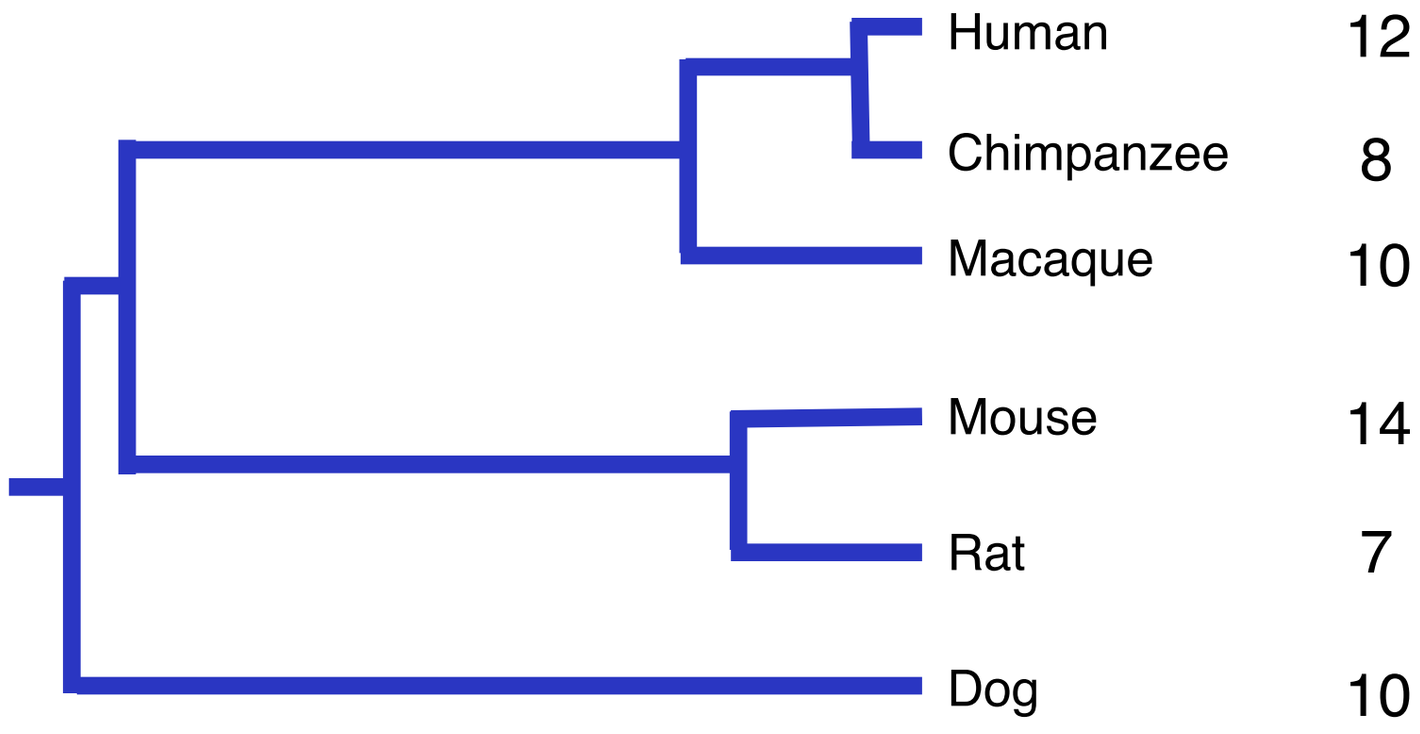
A model for gene family evolution

Time since divergence
(ultrametric tree)



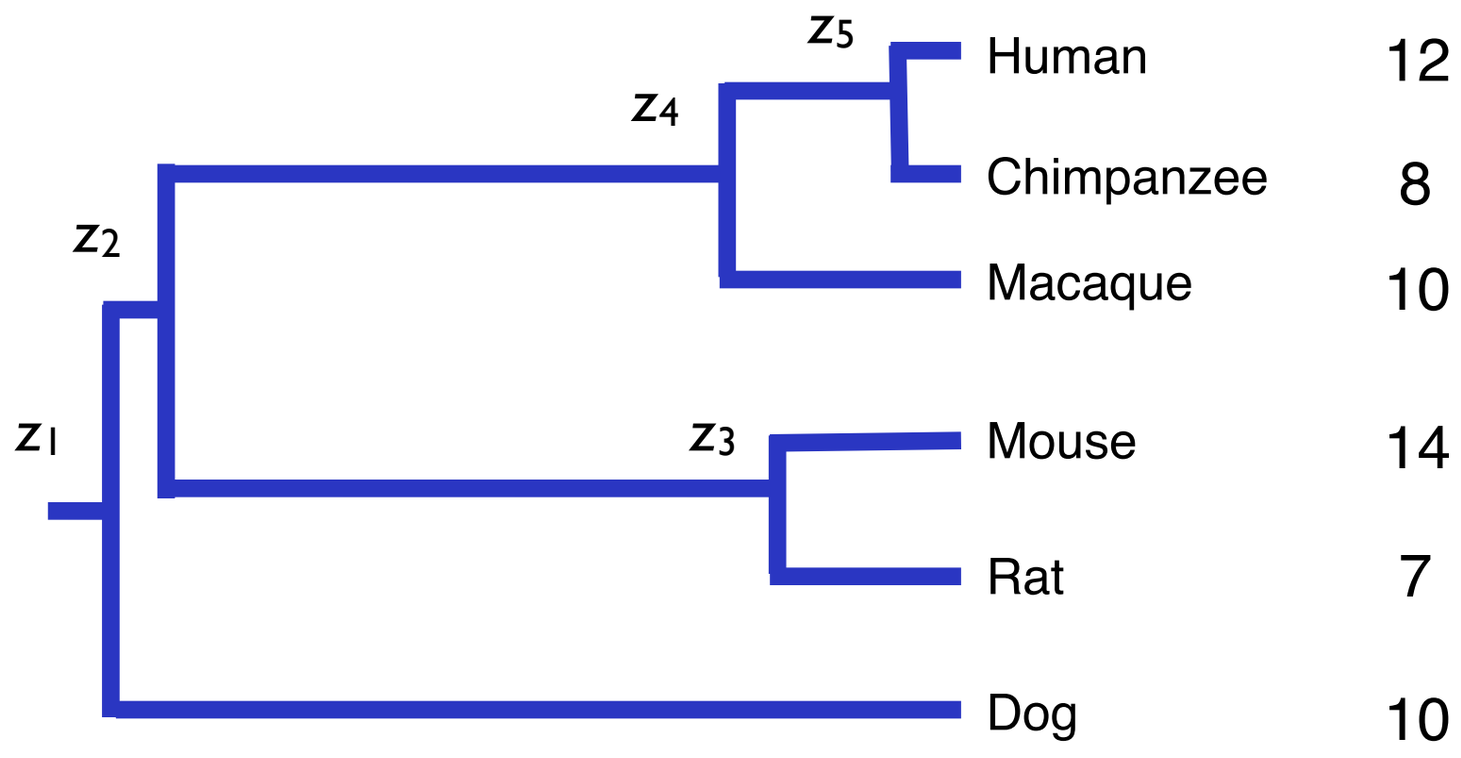
A model for gene family evolution

Current family sizes



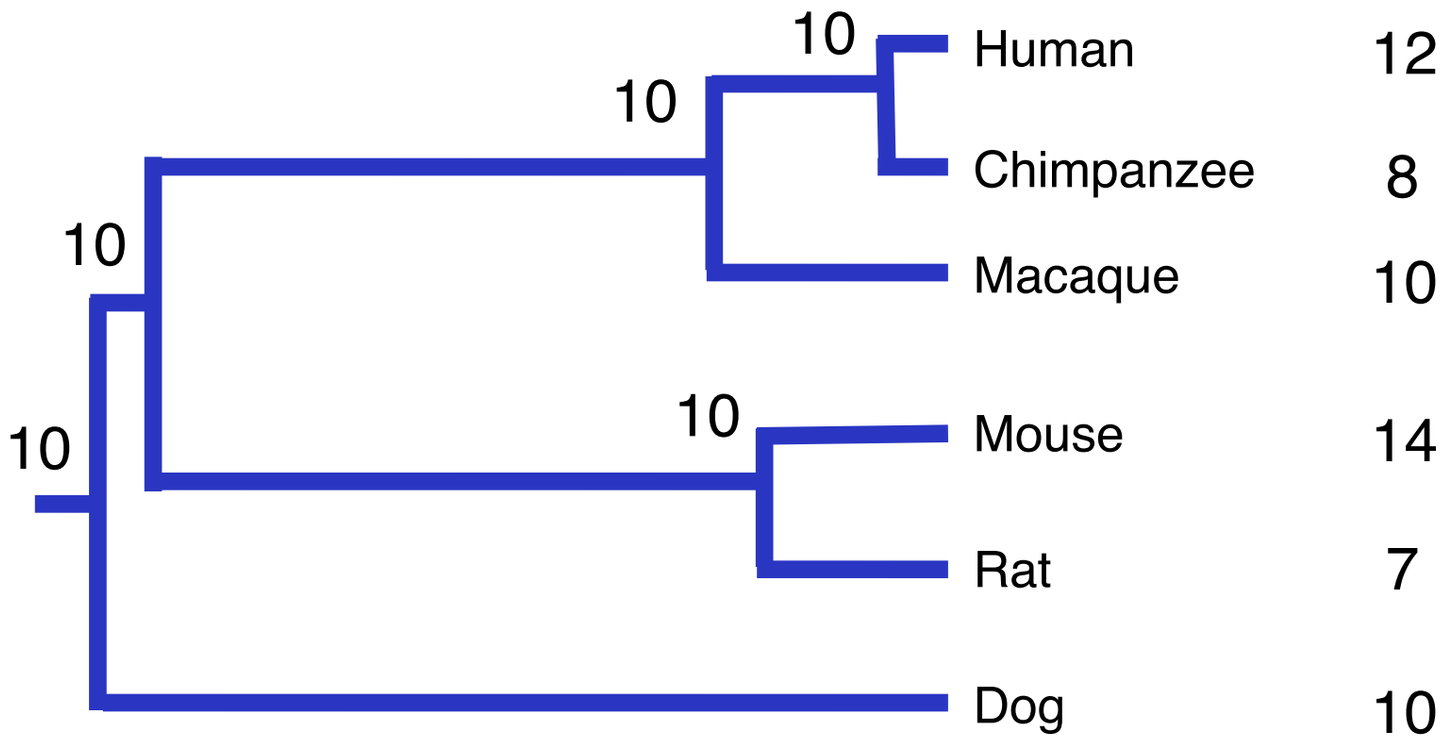
A model for gene family evolution

Ancestral family sizes and birth and death rates (λ, μ) can be inferred using likelihood!



A model for gene family evolution

Ancestral family sizes and birth and death rates (λ, μ) can be inferred using likelihood!

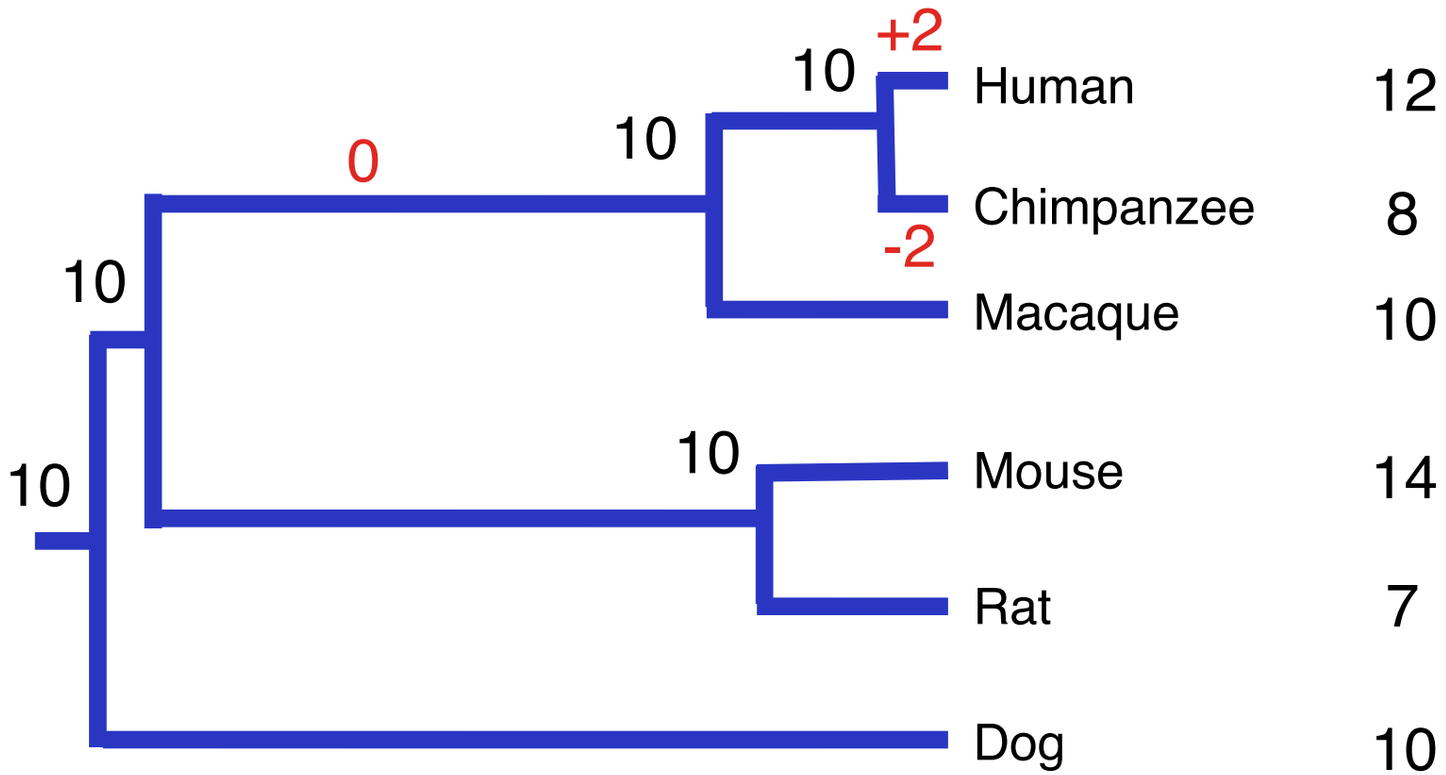


λ=0.002

(assuming birth=death)

A model for gene family evolution

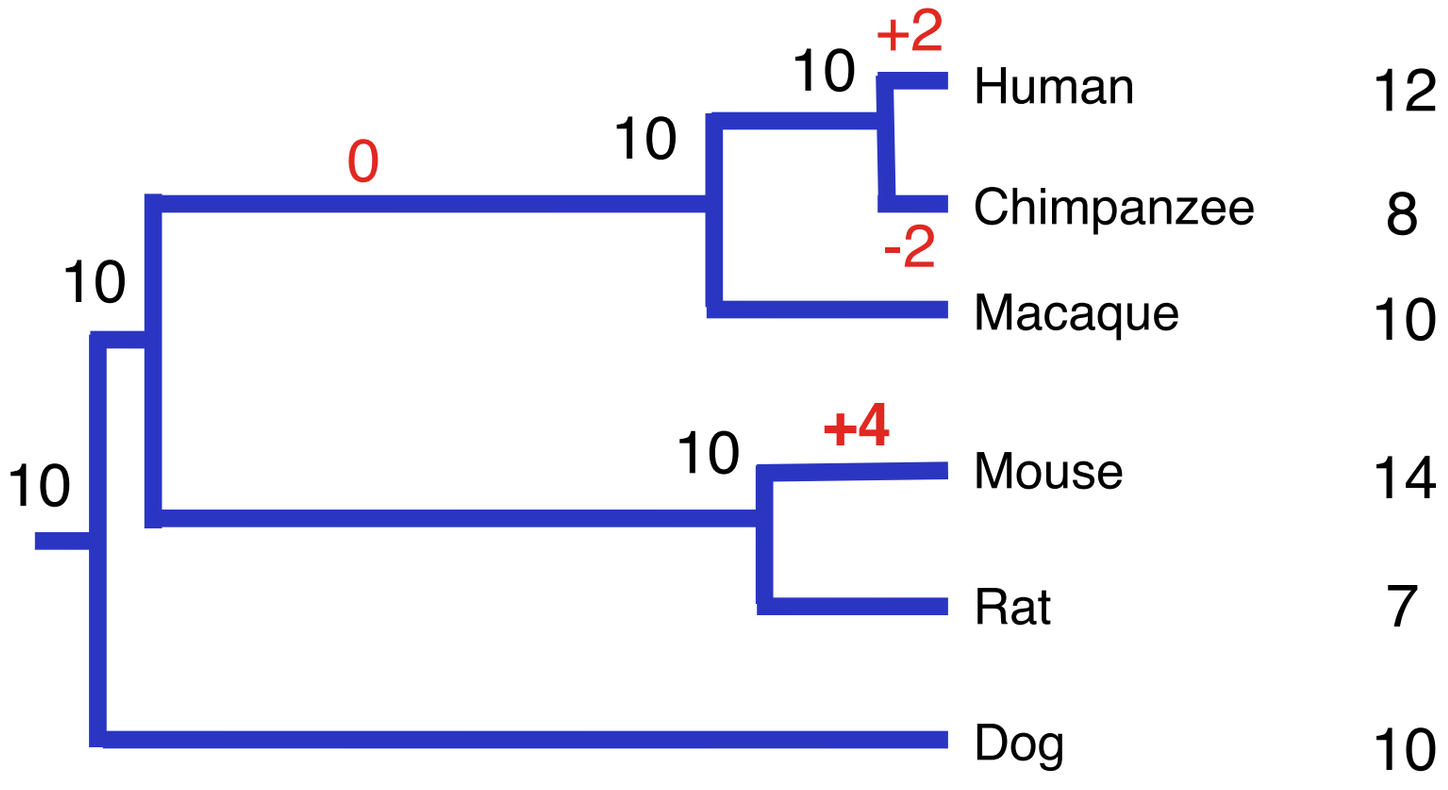
With this information we can infer changes in gene family size…



λ=0.002

A model for gene family evolution

…and identify rapidly evolving families/lineages



λ=0.002

We use Monte Carlo simulations to find families with low *P*-values and branches with large numbers of changes

A model for gene family evolution

Assumptions:

-Single-gene changes at a time*

-All families have a representative at the root*

-birth=death**

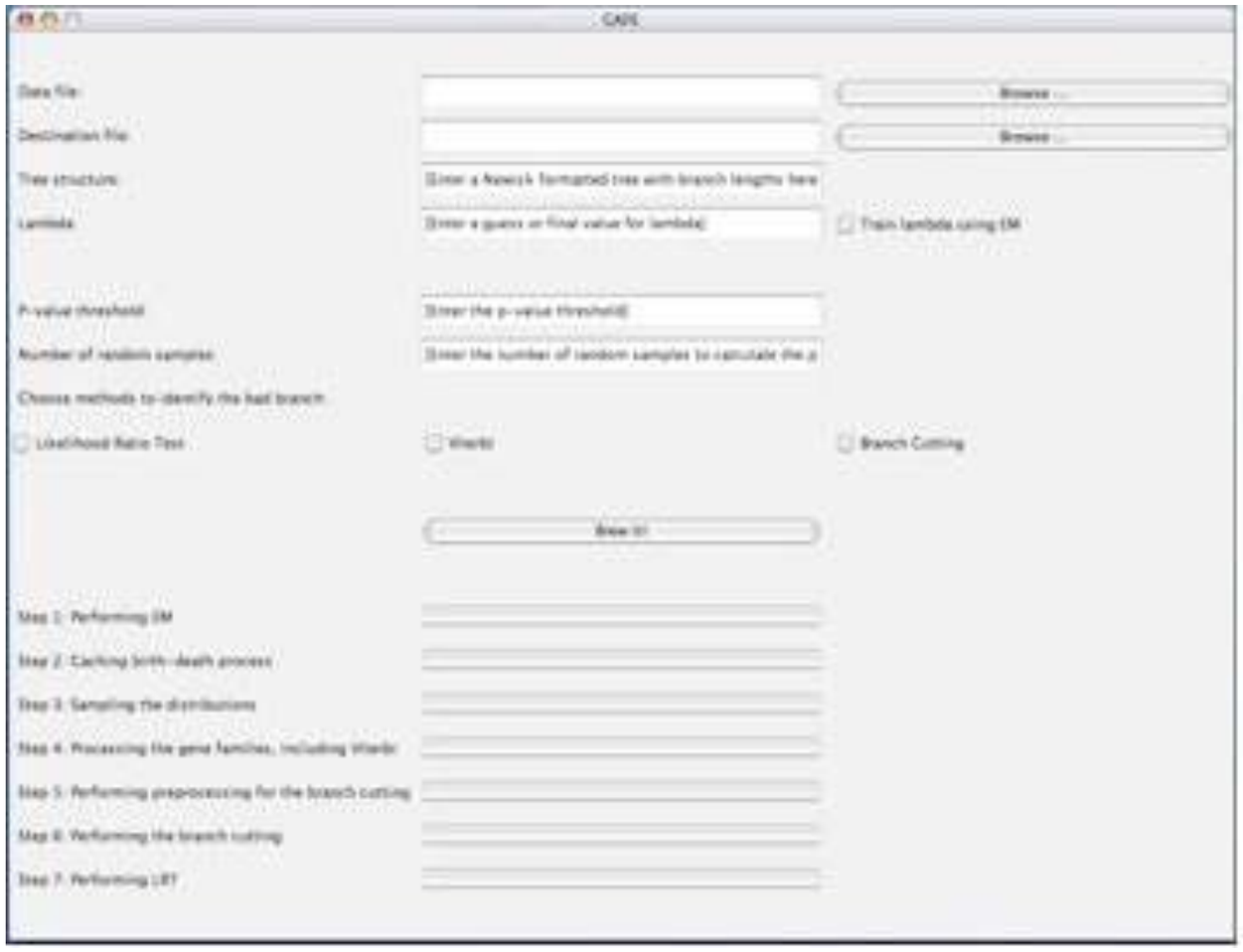
-All lineages have same rate**

-No error in our observations**

-All families have equal per-gene transition probabilities***

CAFE

(Computational Analysis of gene Family Evolution)

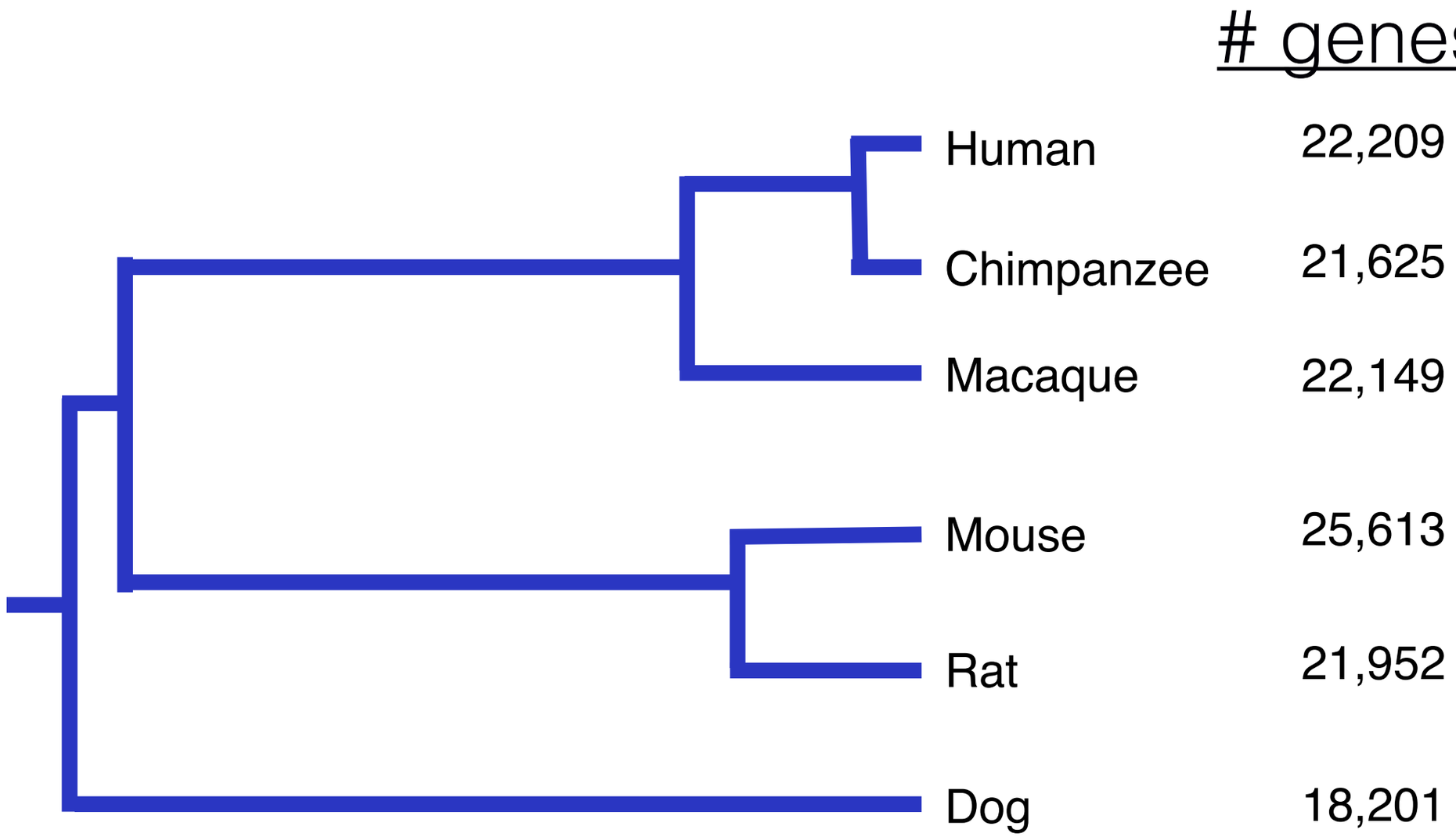


De Bie et al. (2006)

Using CAFE

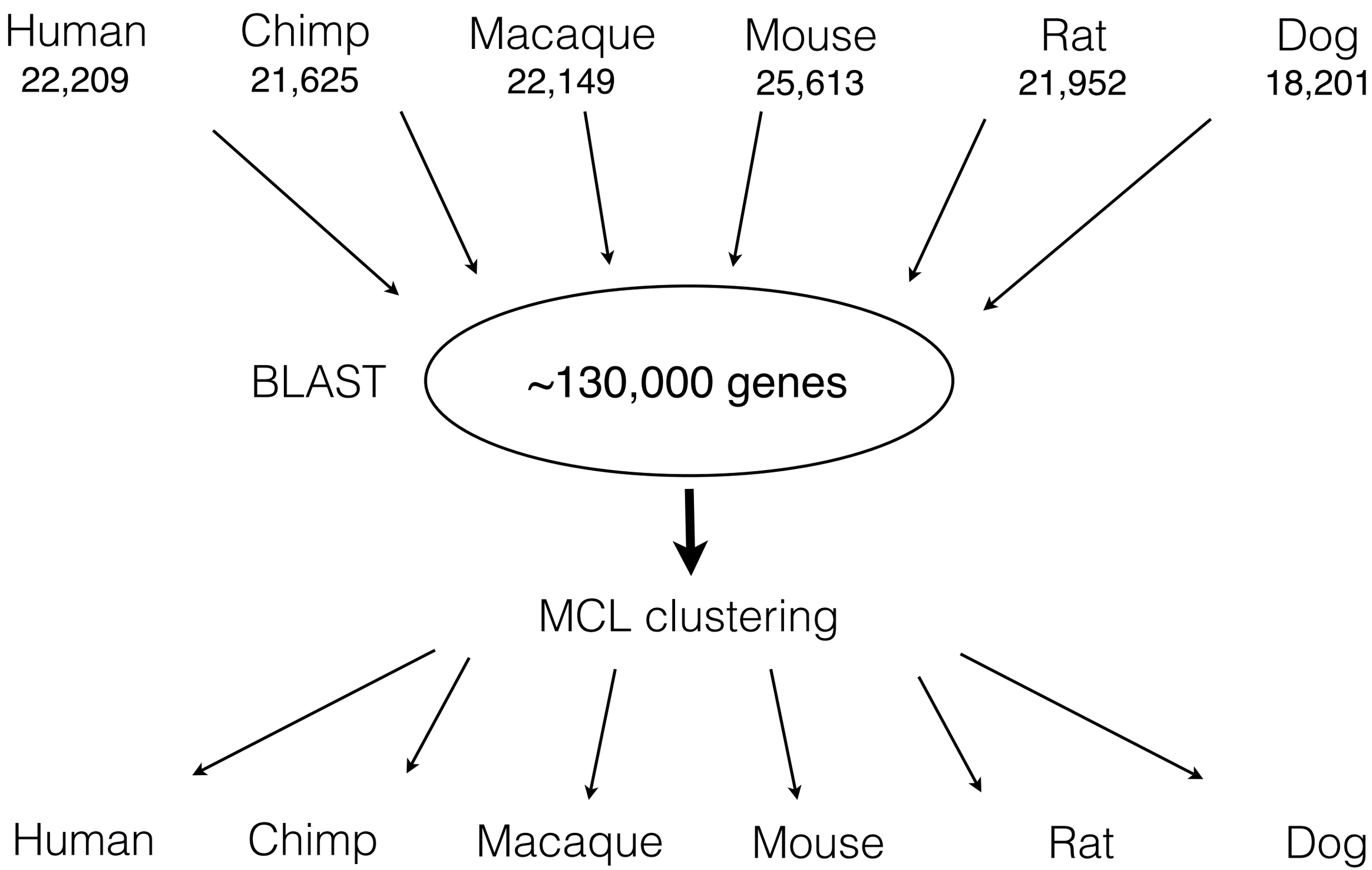
Now let's work through a dataset and elaborations on the basic model…

Genome size in mammals

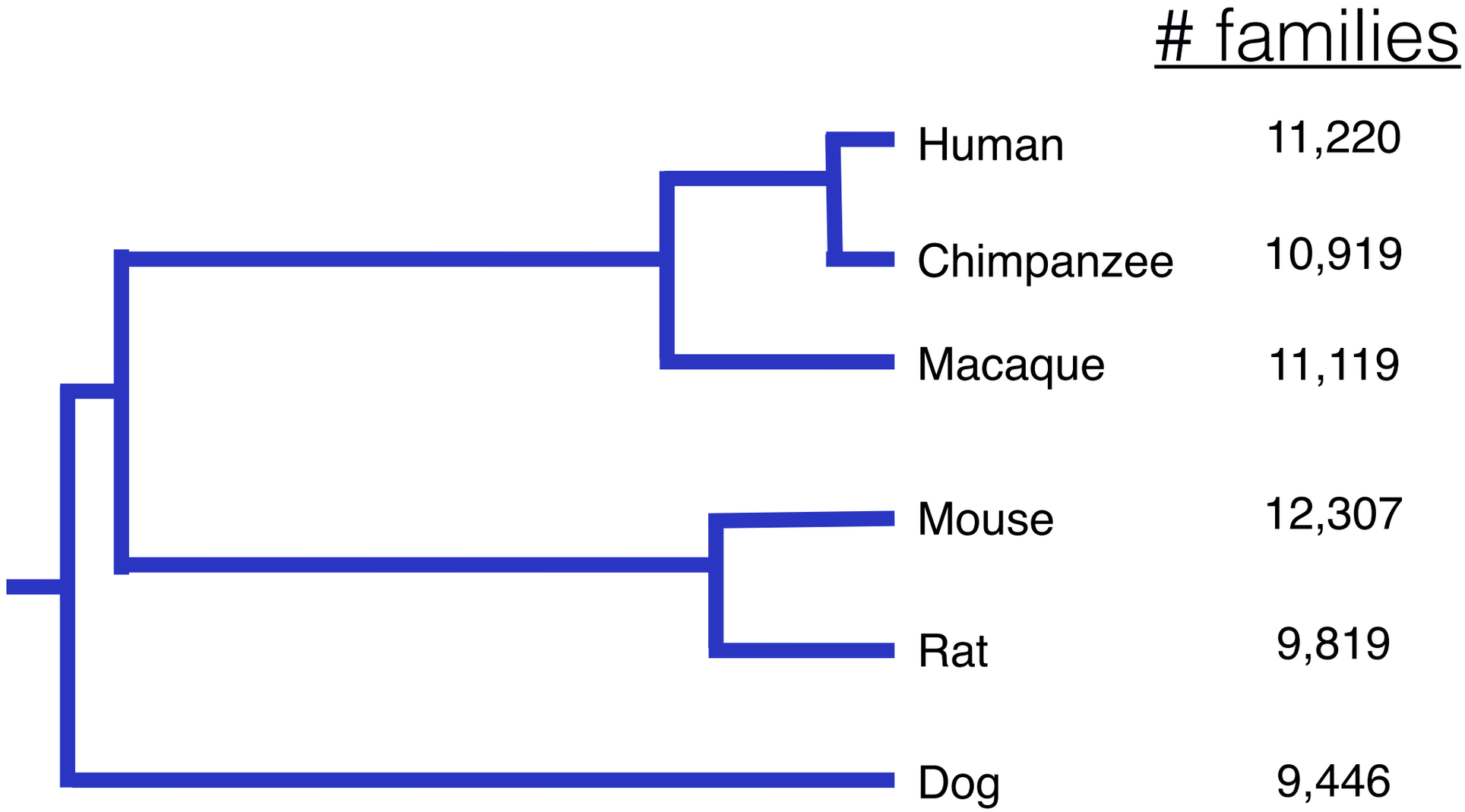


All data from Ensembl

Constructing gene families



Gene families in mammals



Annotation artifacts

Some gene families are present in only one copy, in only one species.

Human	Chimp	Mouse	Rat	Dog
476	549	2,673	620	240

We remove these from further analyses.

Present at the root?

9,990 gene families were inferred to have been present in the mammalian most recent common ancestor (MRCA).

Genome evolution in mammals

- 5,285 of 9,990 gene families have changed in size
- We estimate that the MRCA contained 19,523 genes in these families

Demuth et al. (2006)

The rate of gene gain and loss

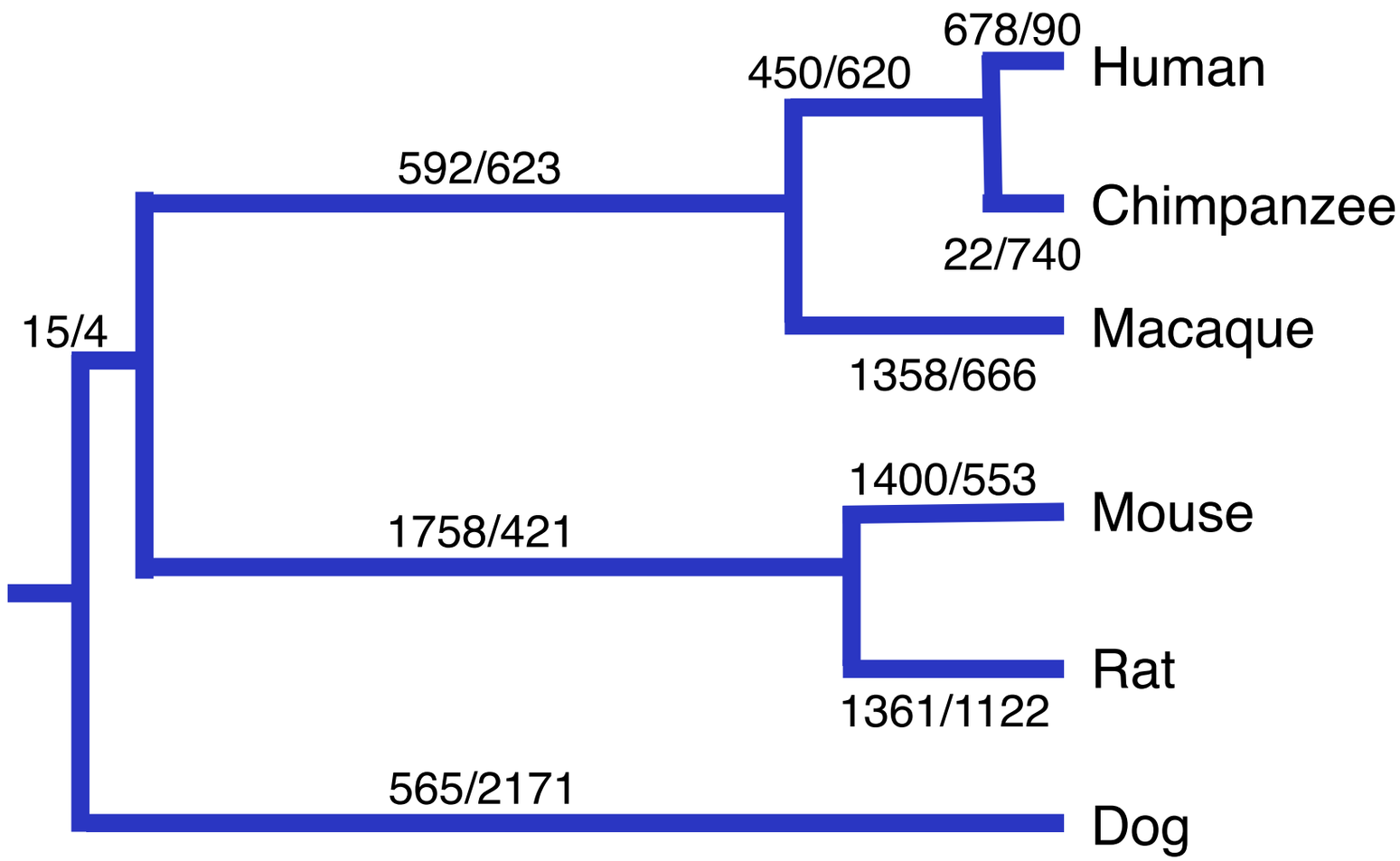
We estimate λ=0.0017 /gene/my
across the whole tree

This number is very similar to estimates by other groups for just the rate of gene duplication.

mouse and rat: 0.0013-0.0026 (Lynch and Conery 2003; Gibbs et al. 2004)

human: 0.009 (Lynch and Conery 2003)

Gene gain and loss in the great apes



Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

- 740 genes have been lost

$$\begin{array}{r} + \\ \hline 1415 \end{array}$$

1,415 human genes not found in chimps!

Reality check

Do these numbers make sense?

Back-of-the-envelope calculations

Lynch and Conery (2000, 2003):

rate of gene duplication in humans is 0.009/gene/my

22,000 x 0.009 dupes/gene/my x 5 million years

= 990 new gene duplicates

If genomes are not constantly expanding, expect approximately equal number of gene losses (990).

= 1,980 human genes not shared with chimps [1,415]

Differences between human and chimp

There are a large number of differences between humans and chimps (6% at the gene level).

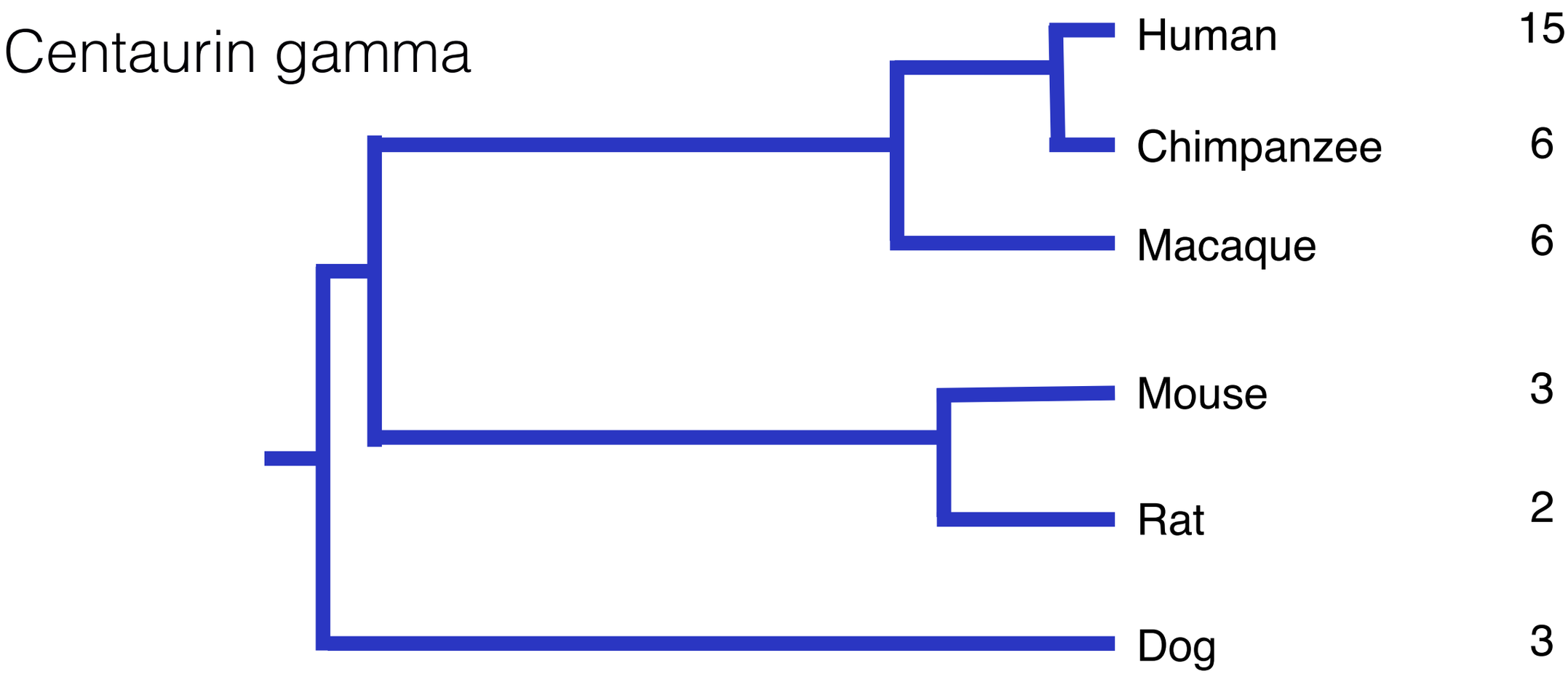
Losses and gains of genes are occurring in the primates at high rates.

The genomic revolving door

Accelerated evolution of gene families

Of the 9,990 families inferred to be present in the mammalian MRCA, we found 180 with $P<0.0001$.

Accelerated evolution of gene families



Accelerated evolution of gene families

The most common biological functions assigned to the significant families include:

immune defense

brain and neuronal development

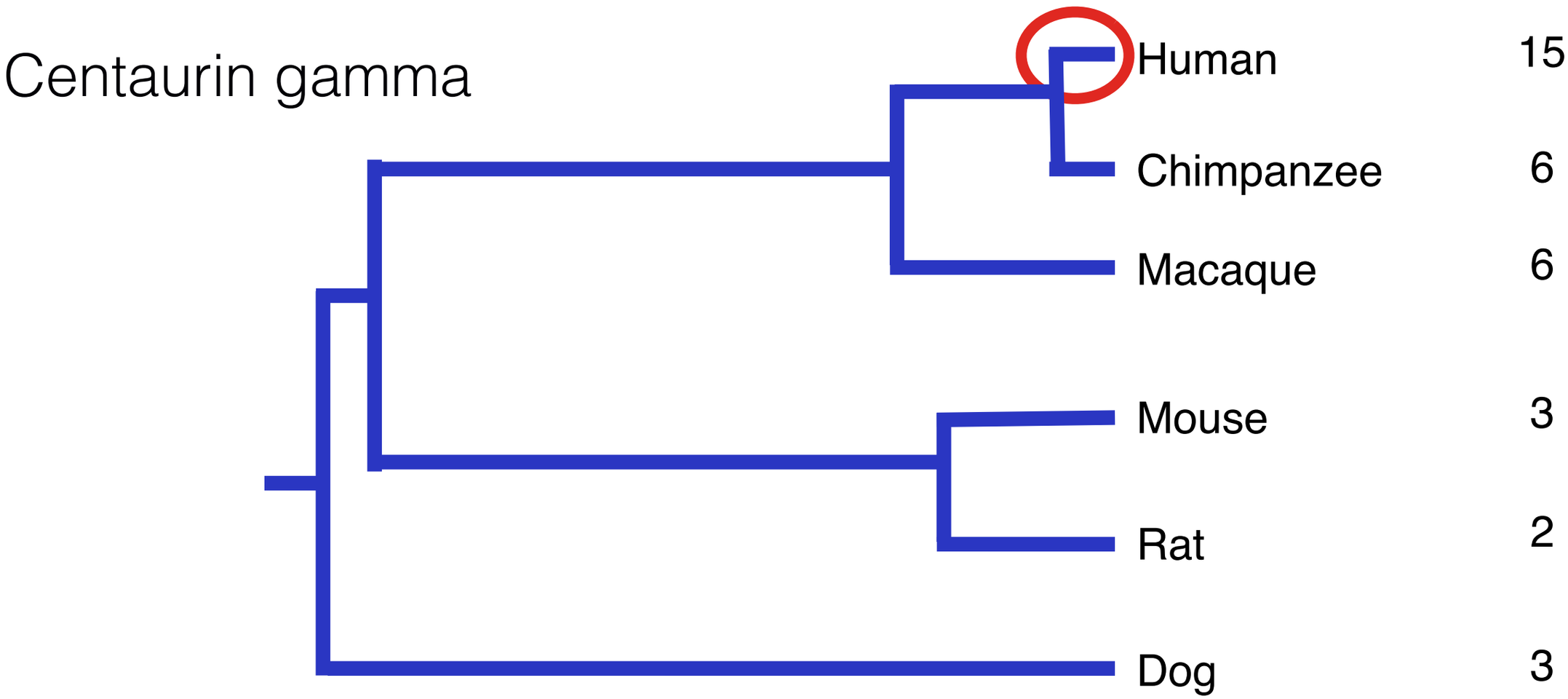
intercellular transport

Interestingly, these are the same functions that evolve rapidly at the nucleotide level in primates.

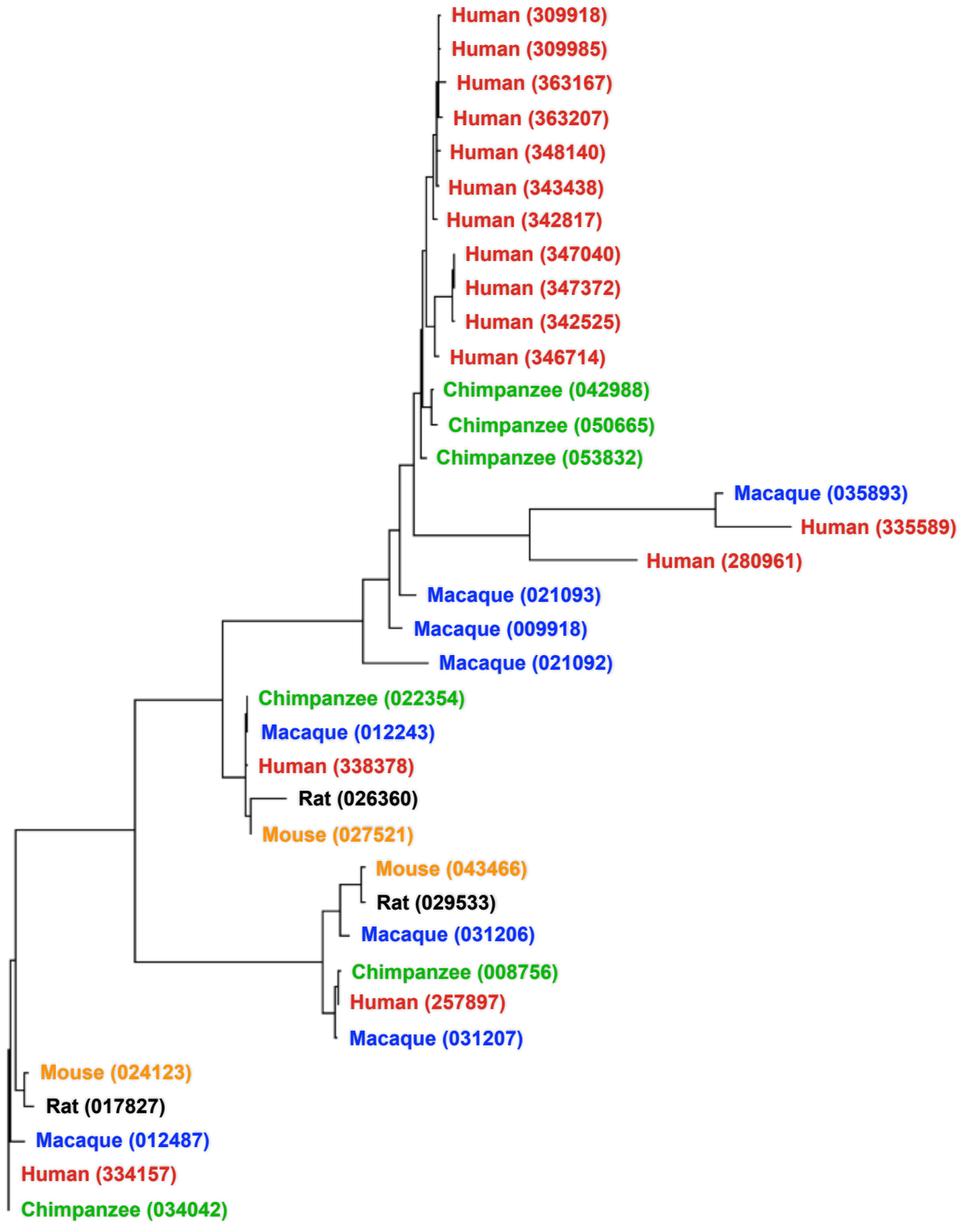
Accelerated evolution of gene families

Using CAFE, one can identify which branches of the tree show the most unlikely changes in family size.

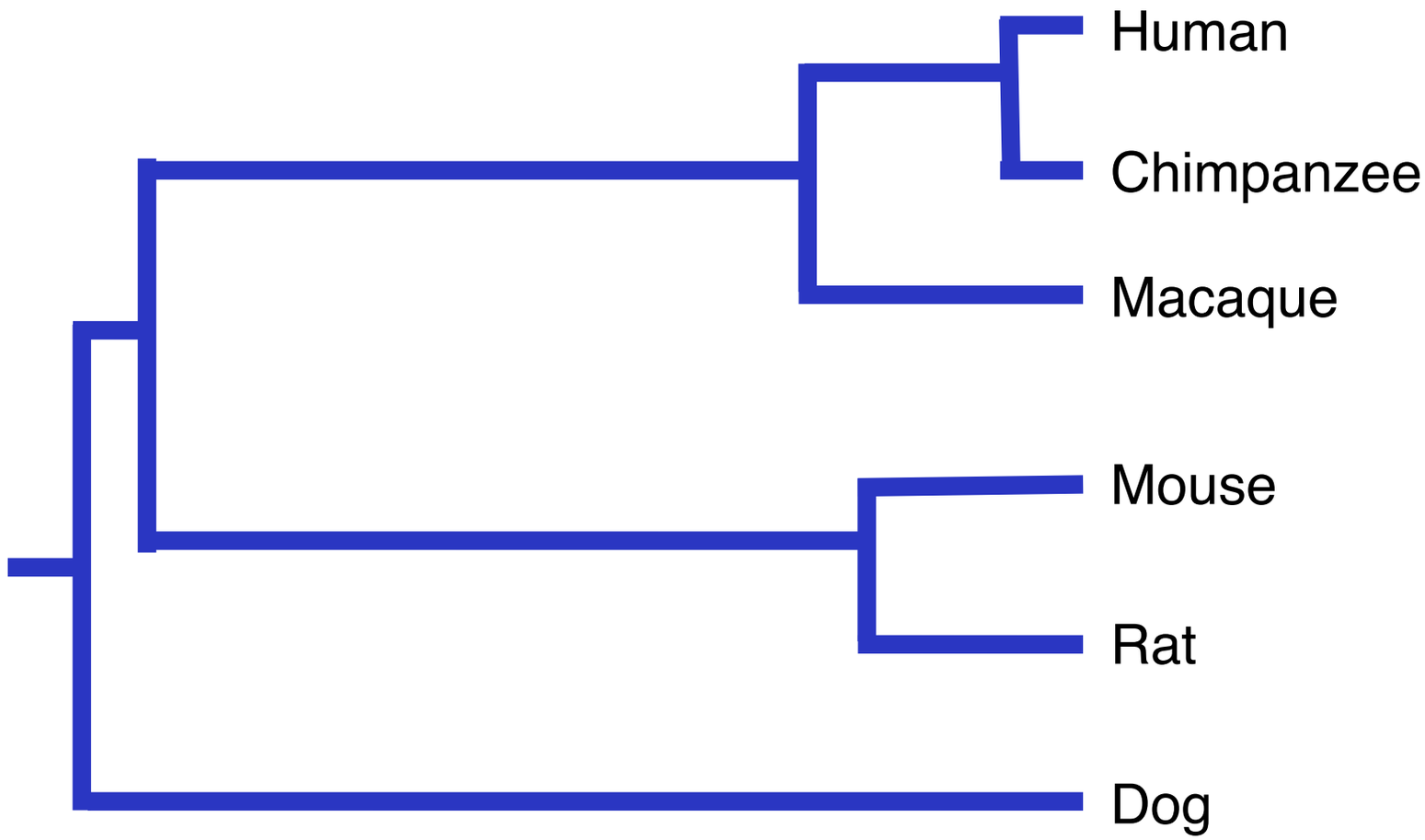
Accelerated evolution of gene families



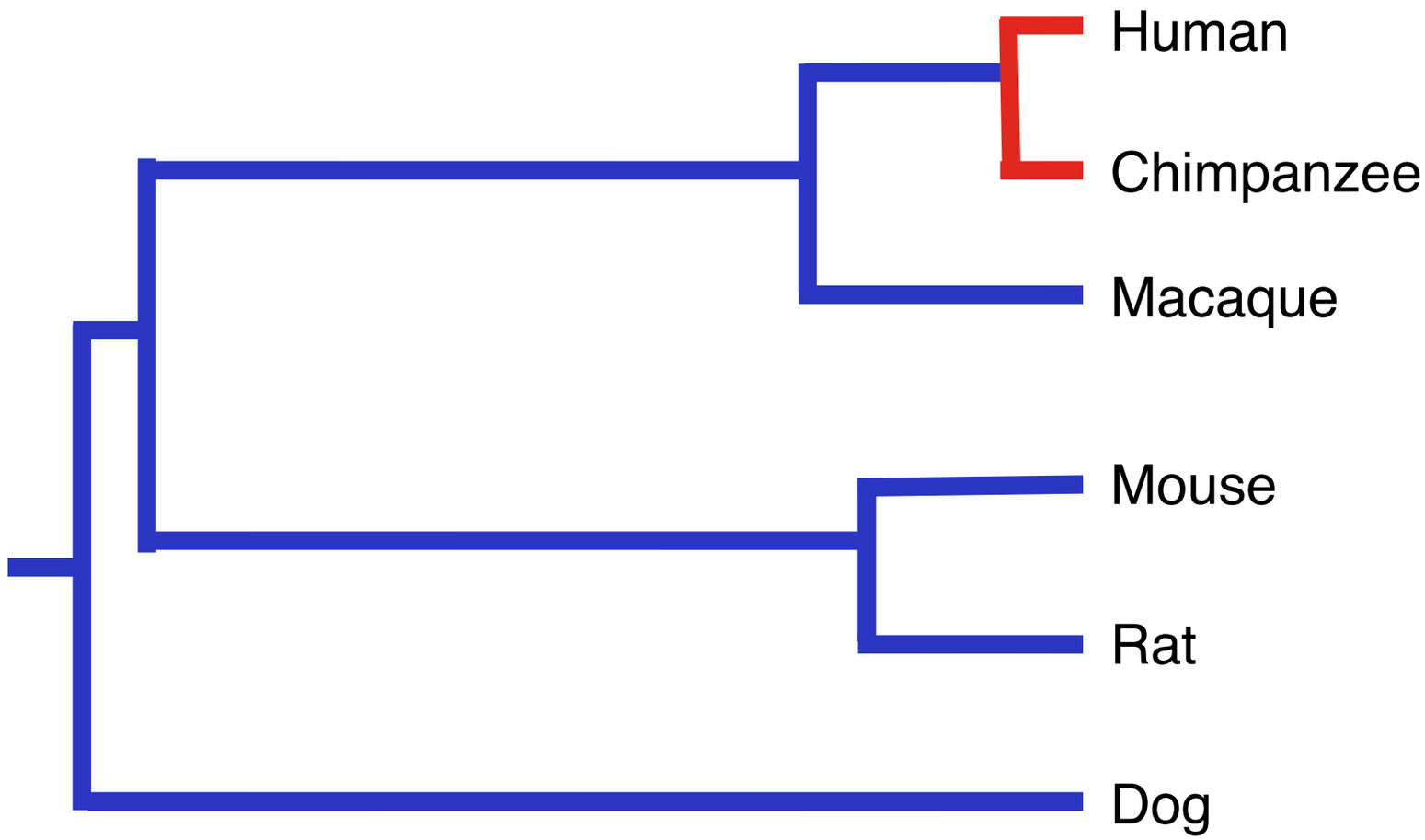
Large expansion of Centaurin gamma in humans



The rate of gene gain and loss

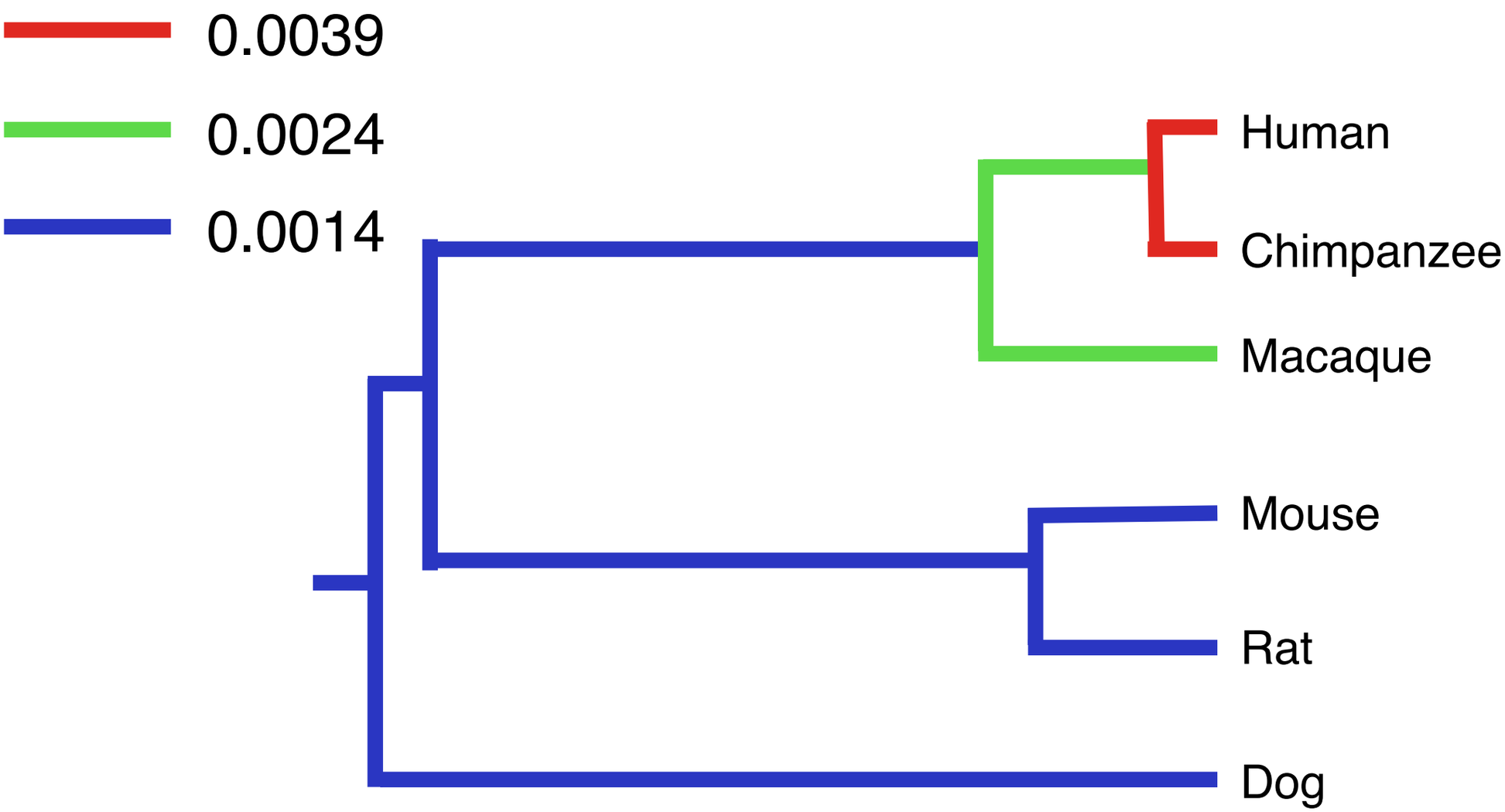


The rate of gene gain and loss



A 2-parameter model fits the data significantly better

The rate of gene gain and loss

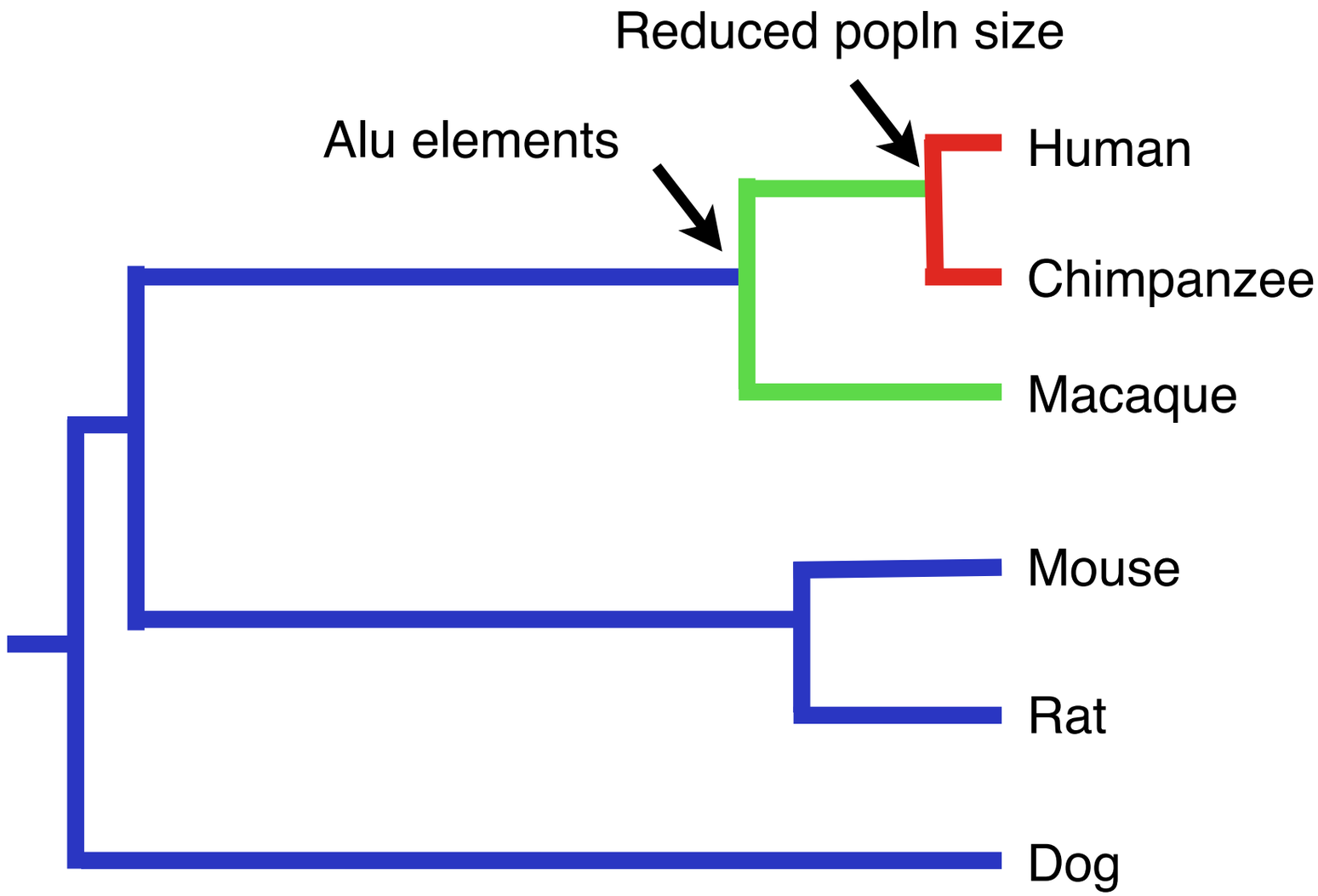


The rate of gain and loss in primates is 2-3 times higher than the rest of the mammals

Accelerated rate of gene gain and loss in primates

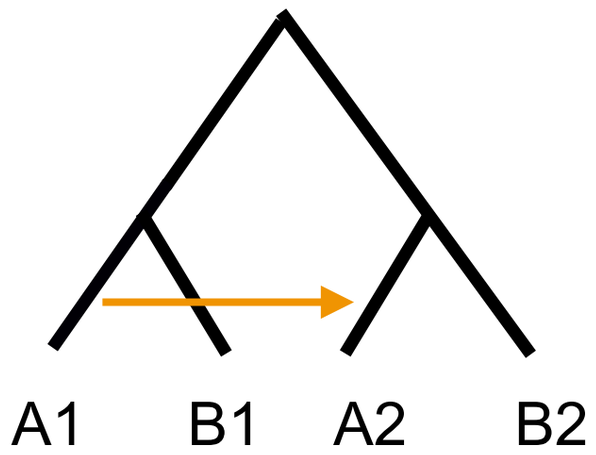
Why?

Accelerated rate of gene gain and loss in primates

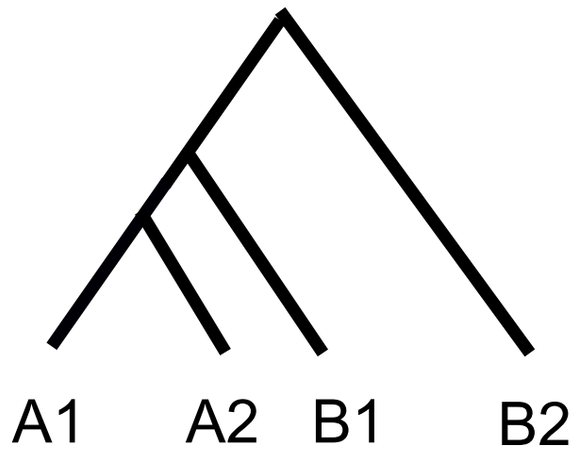


Gene conversion

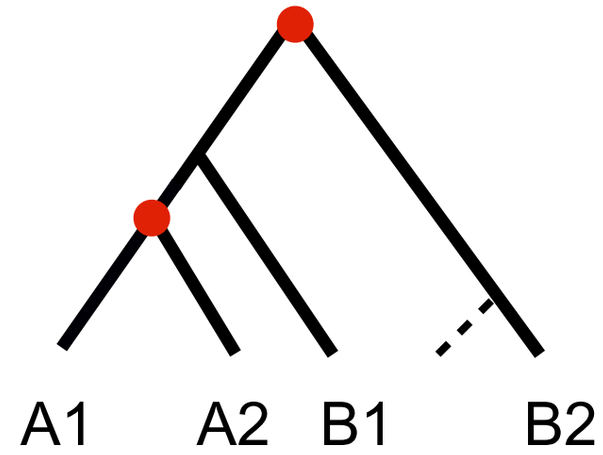
If there is gene conversion, reconciliation can result in extra duplications and losses



gene tree
(before conversion)



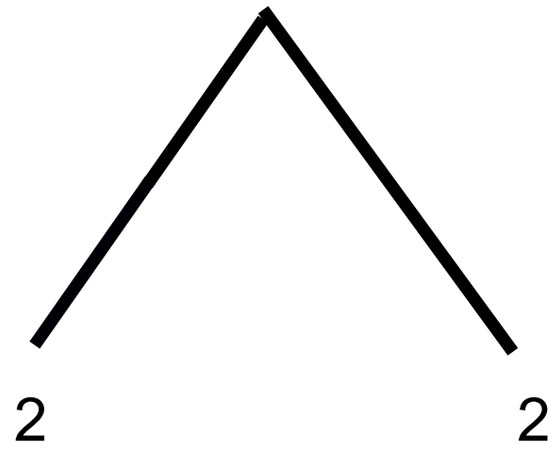
gene tree
(after conversion)



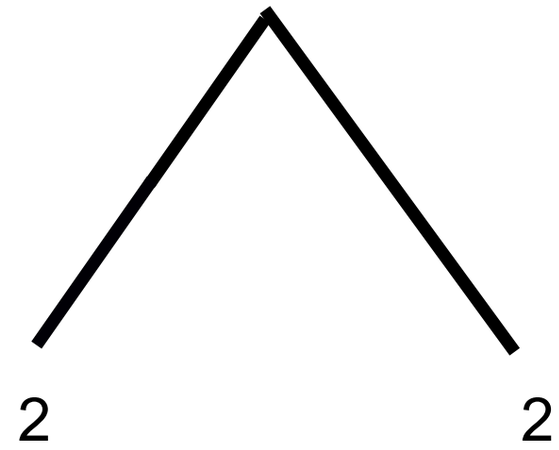
reconciled gene tree

Gene conversion

But with “count” models, there is no change due to conversion:

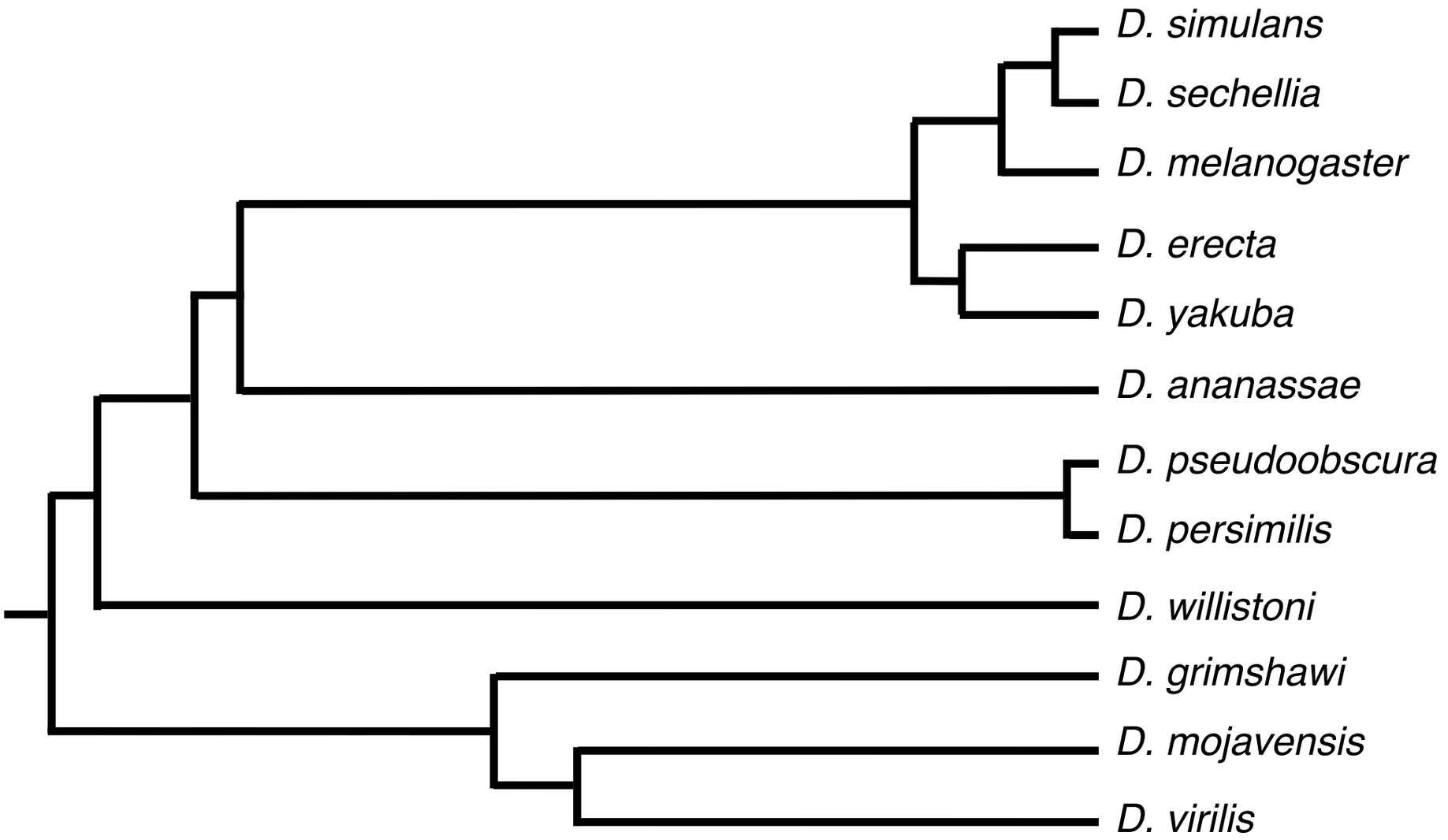


data
(before conversion)

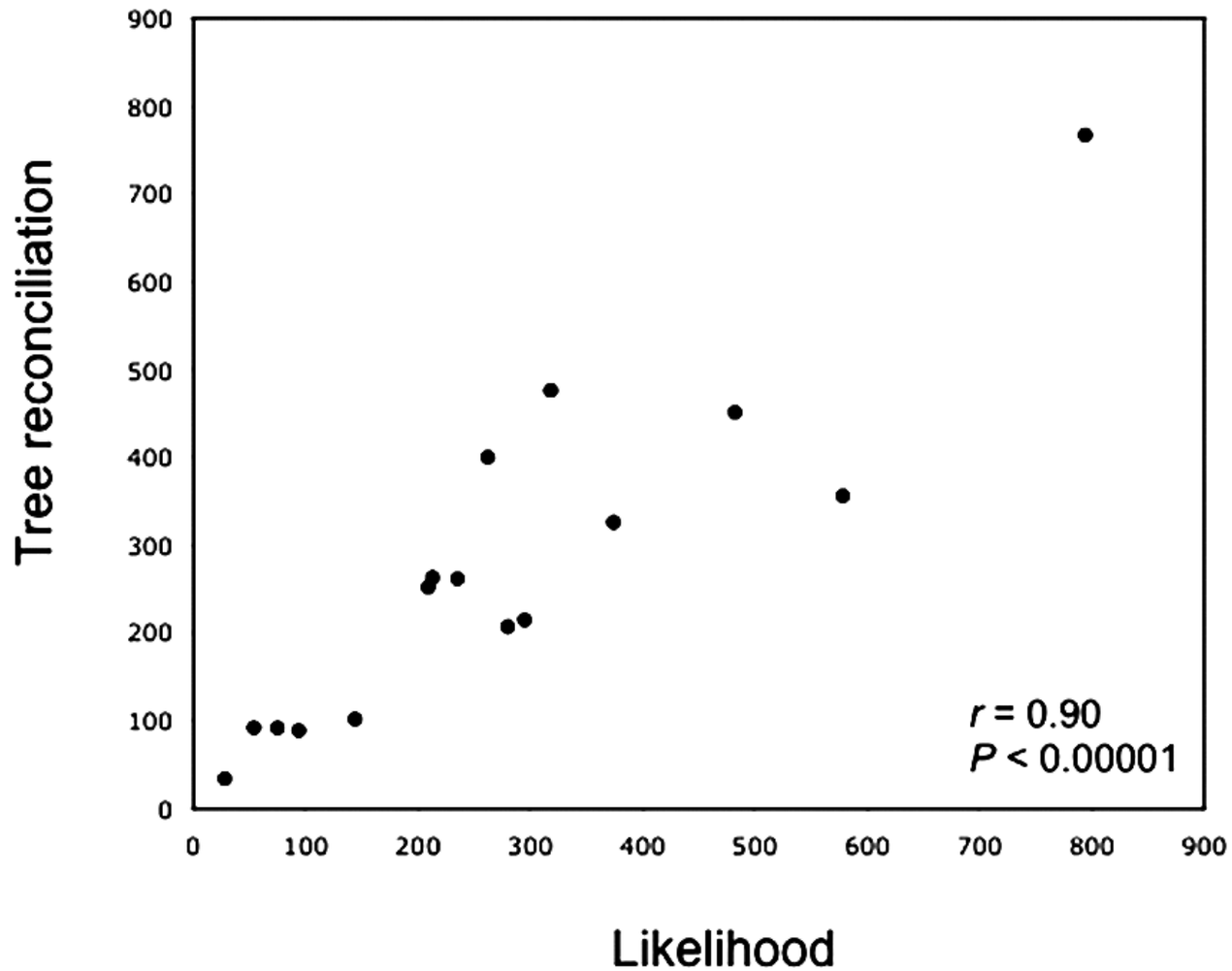


data
(after conversion)

12 *Drosophila* genomes

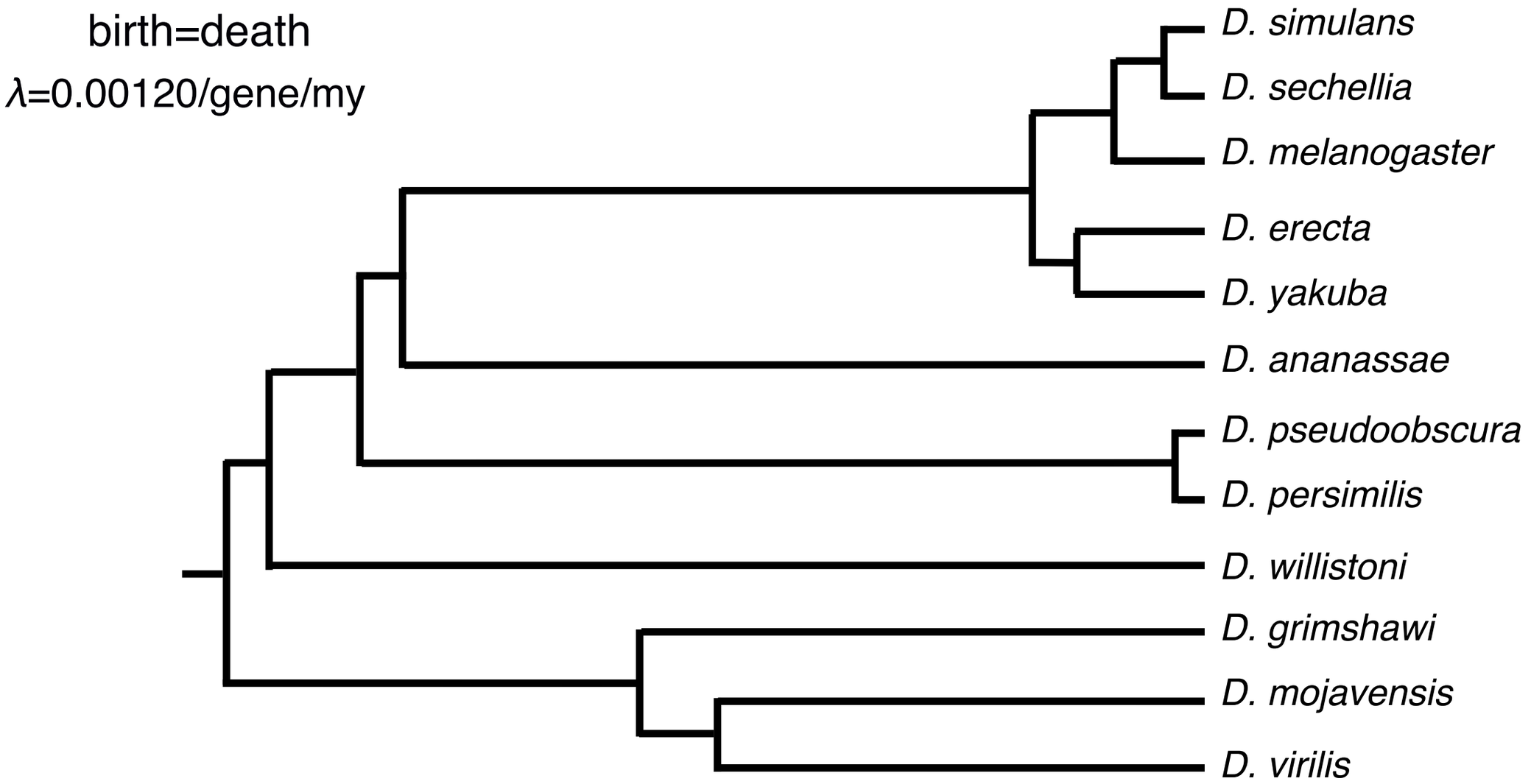


Gene conversion



Hahn et al. (2007)

Differences in rates of gain and loss

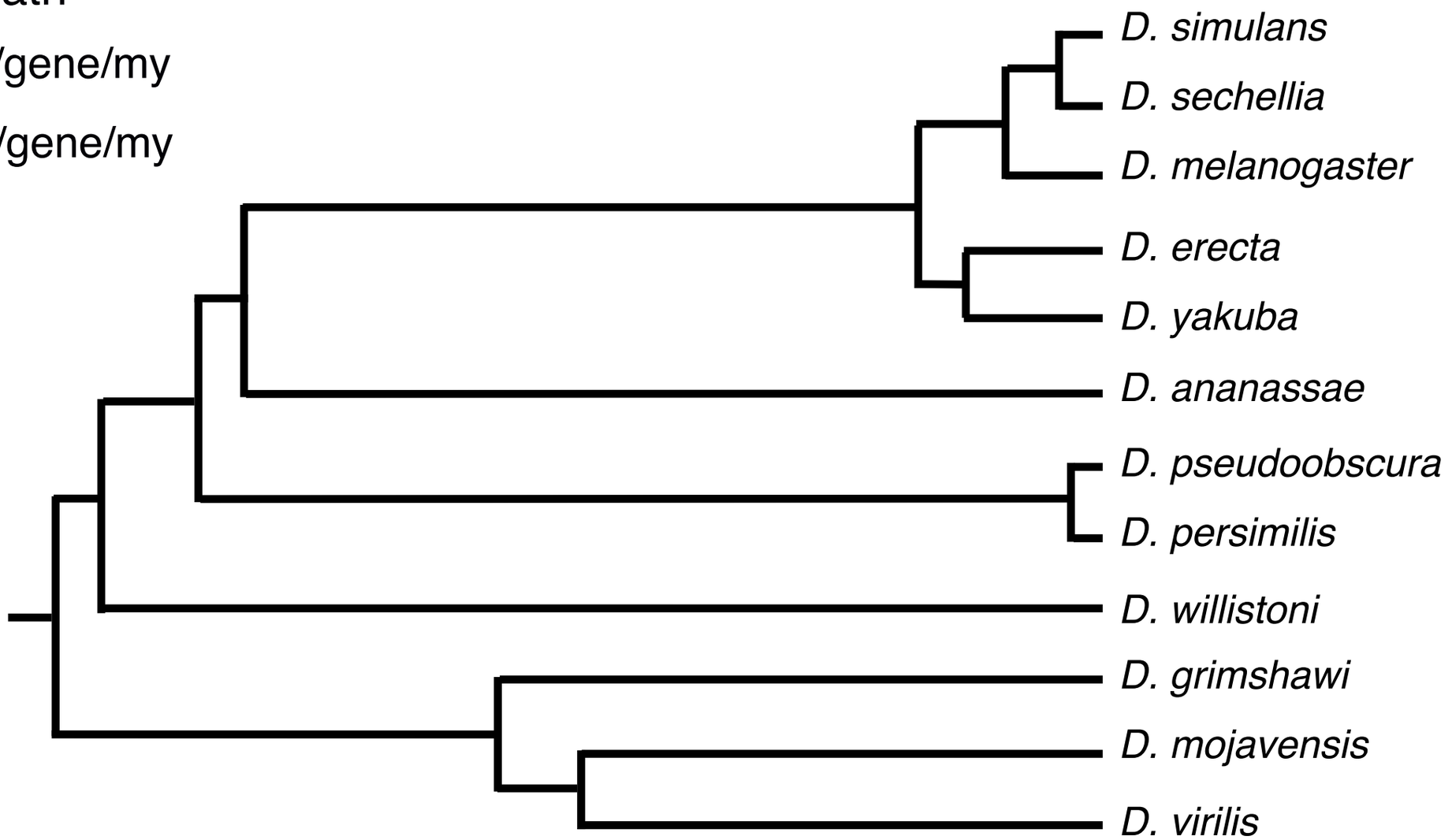


Differences in rates of gain and loss

birth≠death

λ=0.00108/gene/my

μ=0.00134/gene/my



When genomes go bad

-2X Sanger: **very bad**, vastly undercounts genes

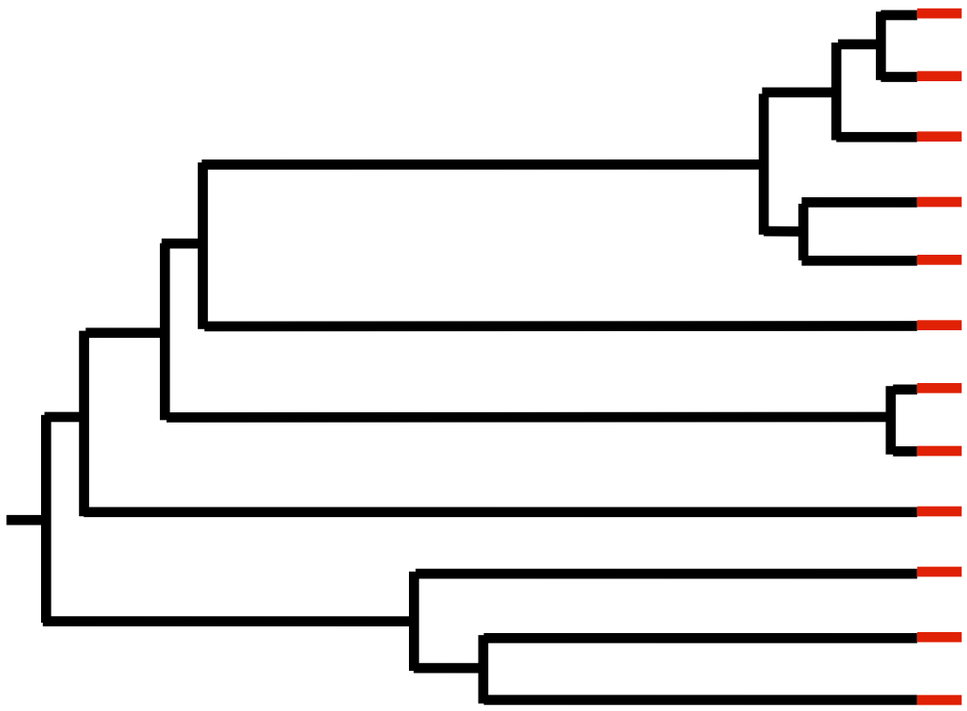
-12X 454: **pretty bad**, slightly overcounts

-82X Illumina: **bad**, but equally over- and undercounts

The best of these (Illumina) still has ~40% of families with errors

(and don’t think your transcriptome assembly is any better!)

Error increases estimated rate of gain and loss

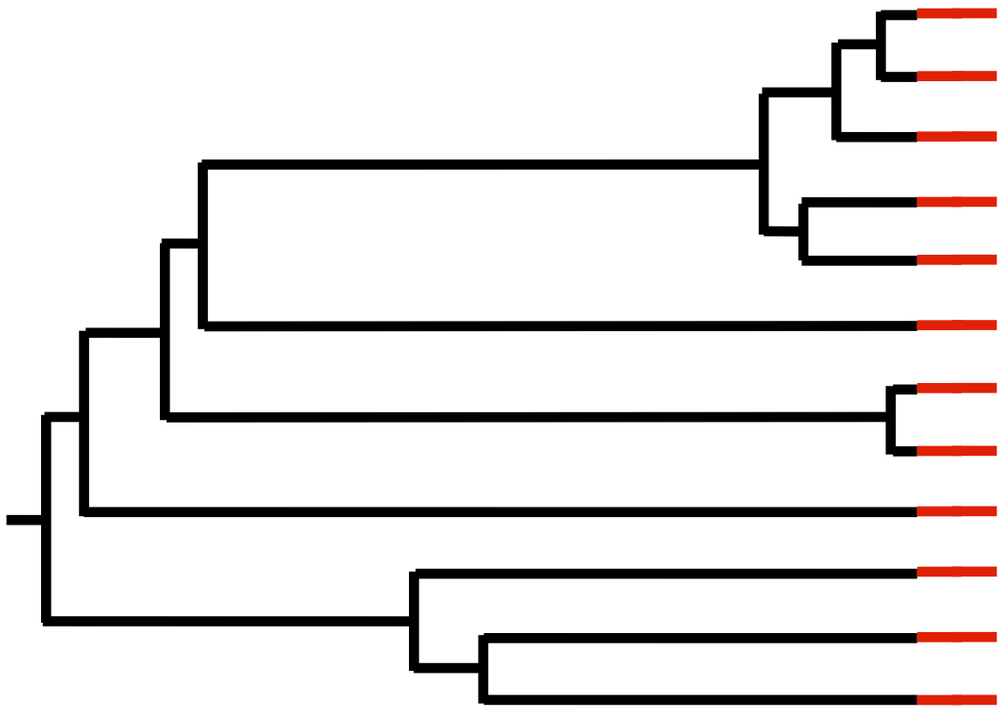


simulated rate: **λ=0.0012**

add **10%** error to data →

after adding error: **λ=0.0027**

Error increases estimated rate of gain and loss

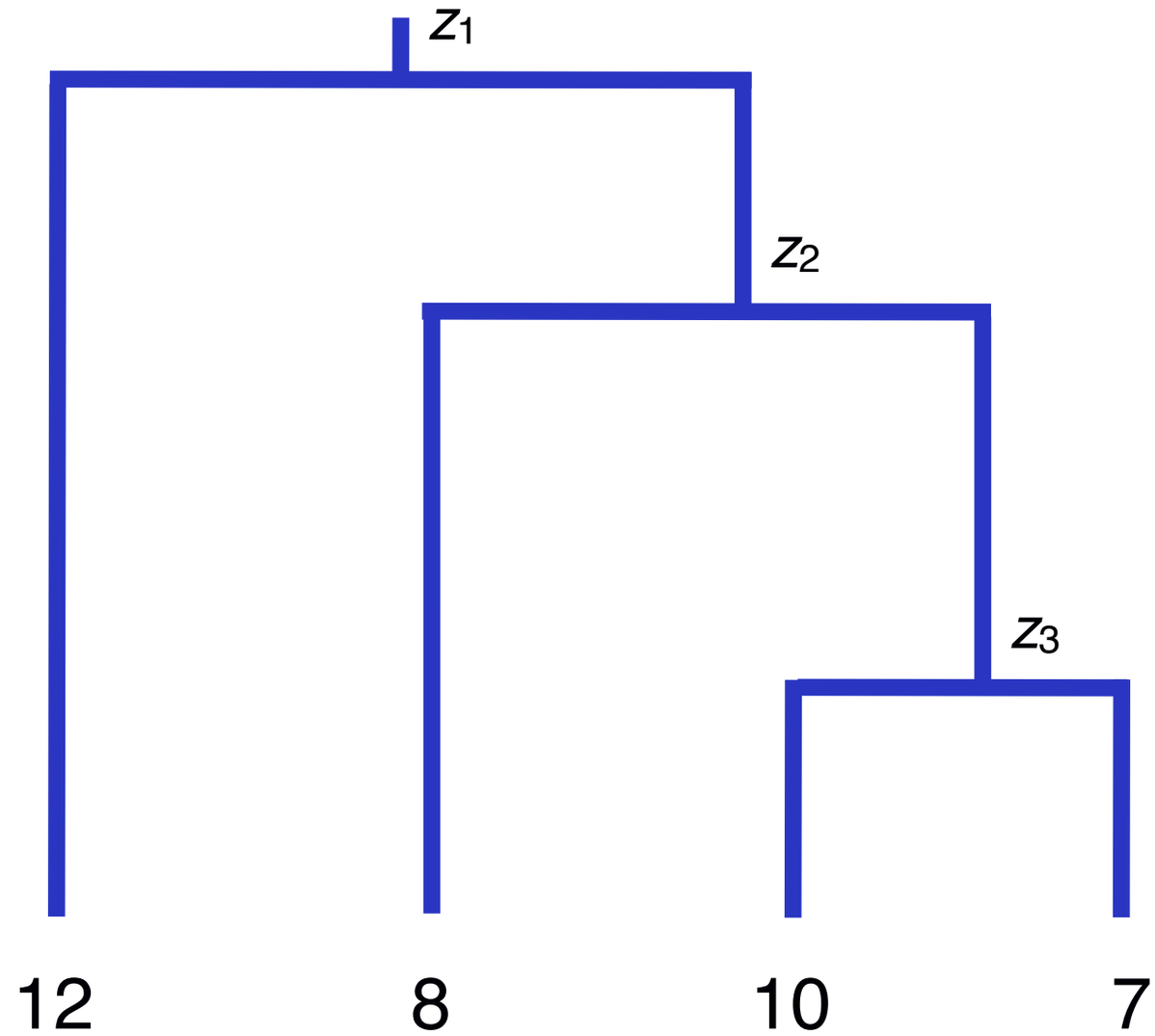


simulated rate: **λ=0.0012**

add **40%** error to data →

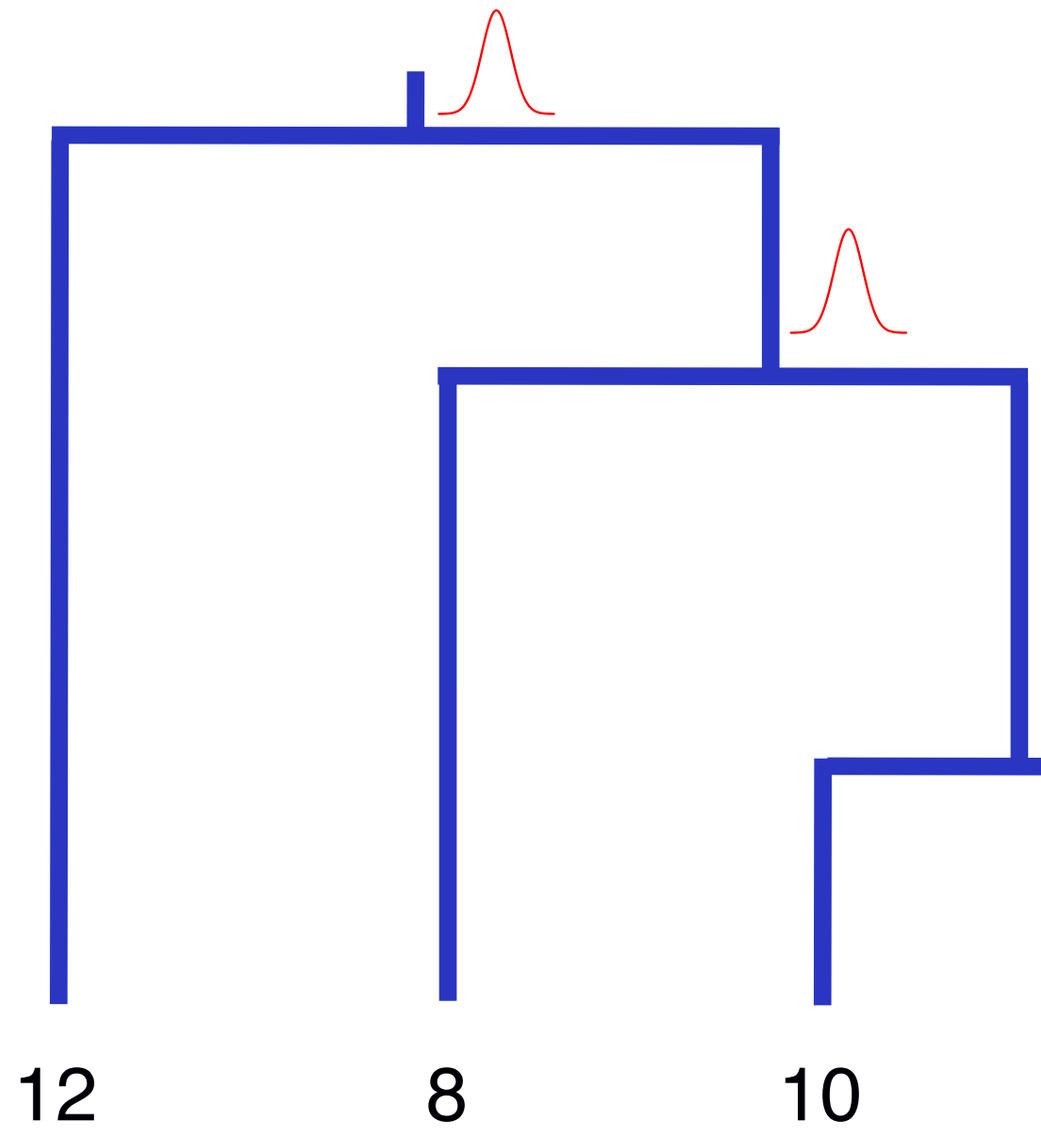
after adding error: **λ=0.0085**

A model for gene gain and loss



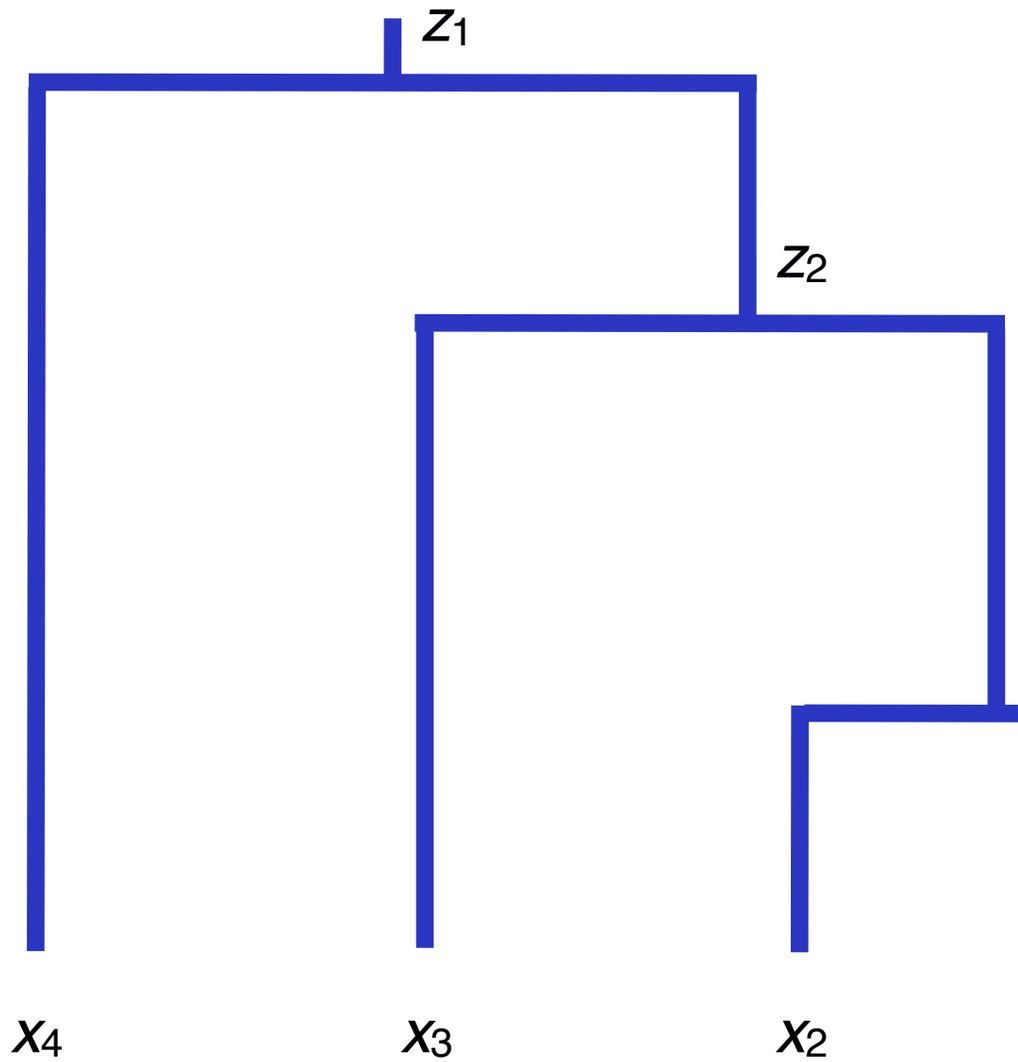
Z: hidden variables

A model for gene gain and loss



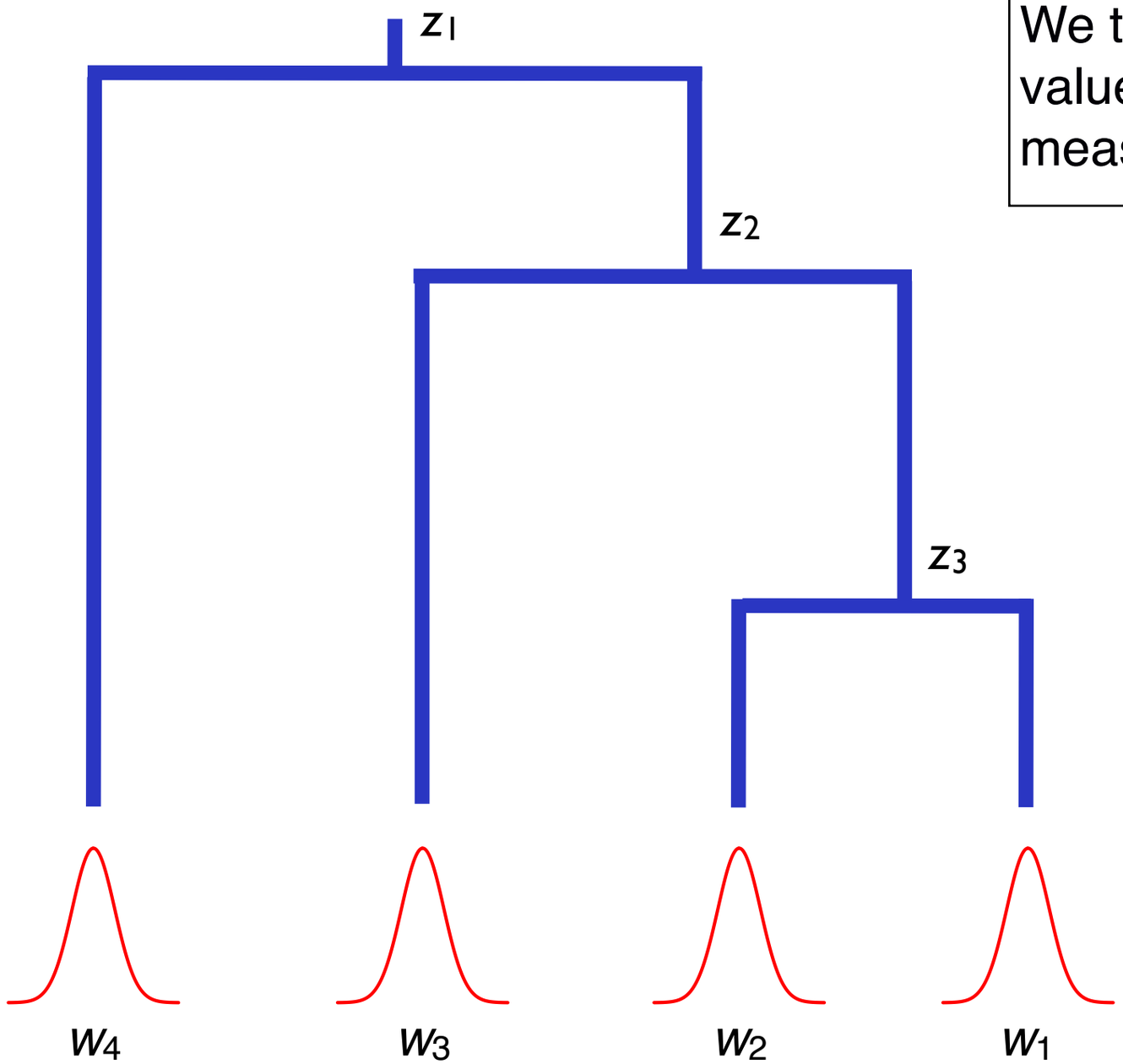
Z: hidden variables

Modeling genome error



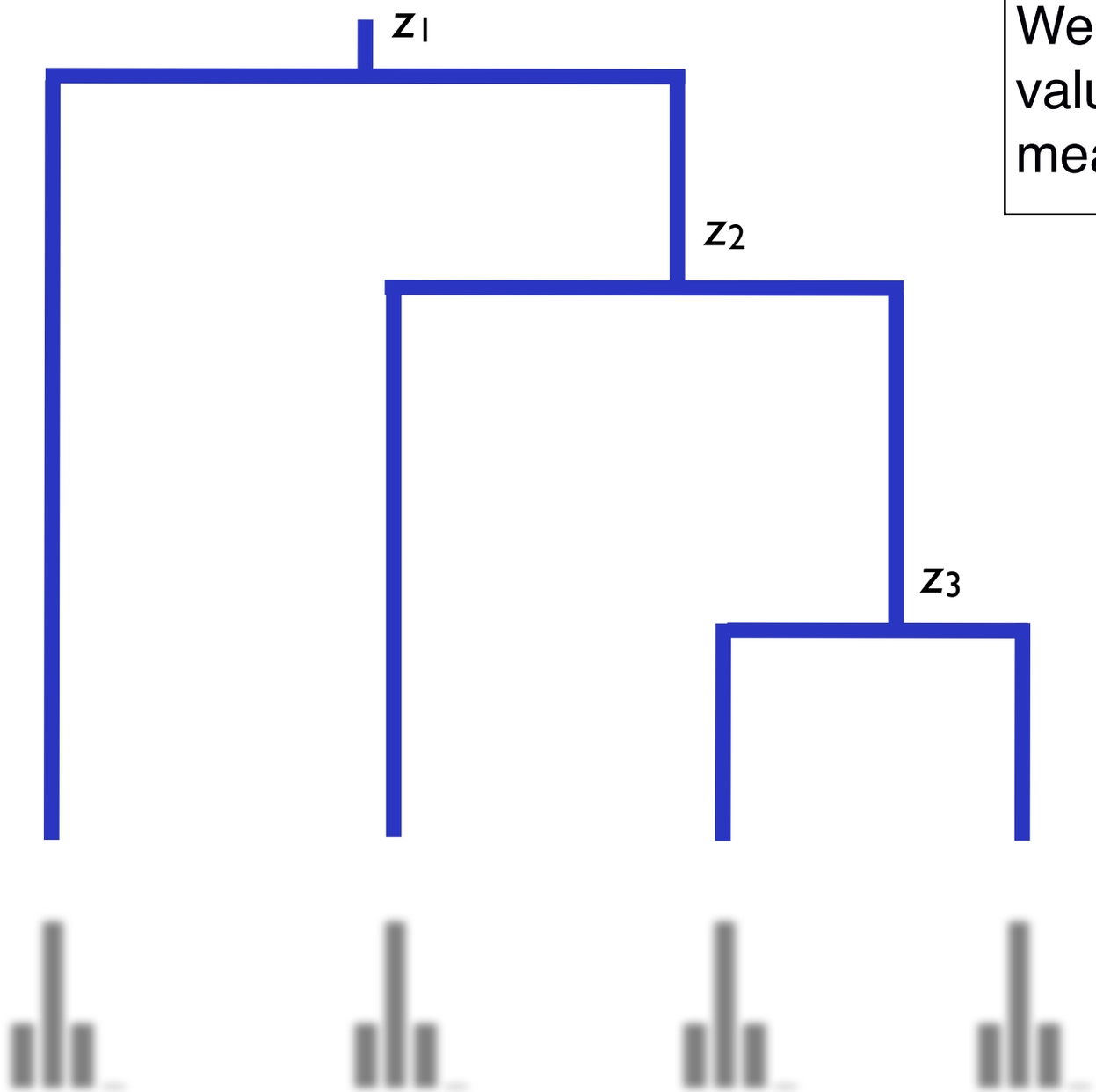
Z: hidden variables
X: observed variables

Modeling genome error



We treat the observed values as error-prone measures

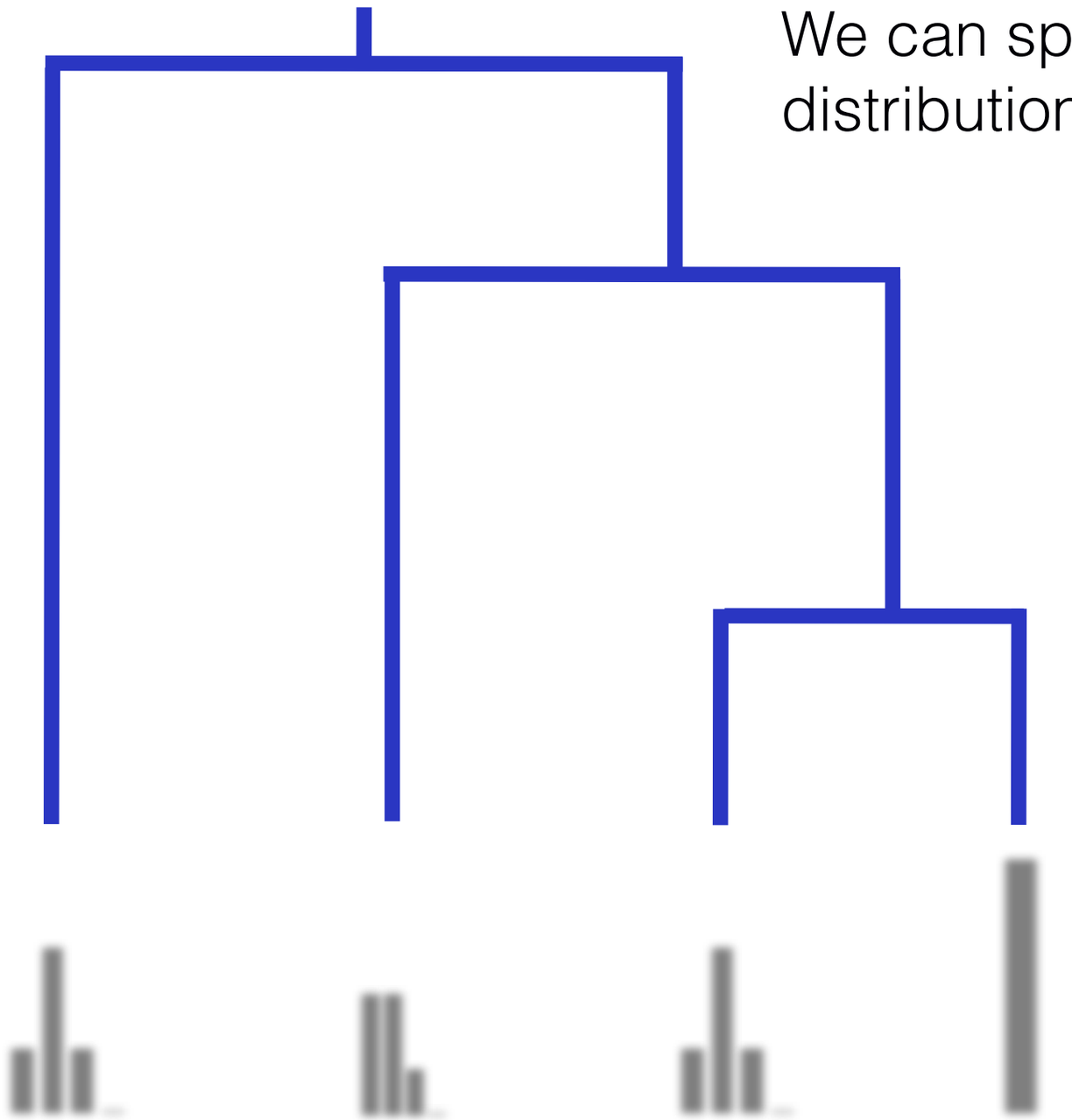
Modeling genome error



We treat the observed values as error-prone measures

Modeling genome error

CAFE 3.0



We can specify different error distributions for different species

Modeling genome error

Original model

$$\lambda, \mu = \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N} P(X_n \mid \lambda, \mu, T)\right)$$

$$= \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N} P(X_n \mid Z_n, \lambda, \mu, T) P(Z_n \mid \lambda, \mu, T)\right)$$

$$= \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{z_{n1}=0}^{M}\sum_{z_{n2}=0}^{M}\cdots\sum_{z_{nu}=0}^{M} P(X_n \mid Z_n \\ = (z_{n1}, z_{n2} \ldots z_{nu}), \lambda, \mu, T) \\ \times P(Z_n = (z_{n1}, z_{n2} \ldots z_{nu}) \mid \lambda, \mu, T)\end{array}\right\}\right).$$

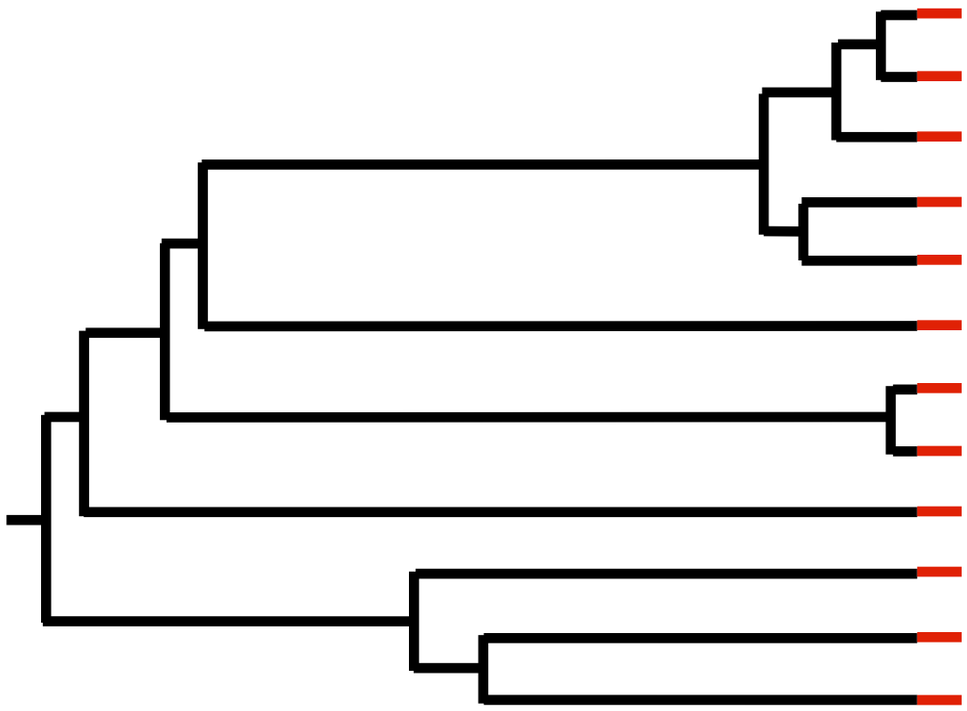
Model with error

$$\lambda, \mu = \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M} P(W_n \mid X_n \\ = (x_{n1}, x_{n2}, \ldots x_{ns})) \\ \times P(X_n, \lambda, \mu, T)\end{array}\right\}\right)$$

$$= \text{argmax}_{\lambda,\mu}\left(\prod_{n=1}^{N}\left\{\begin{array}{l}\sum_{x_{n1}=0}^{M}\sum_{x_{n2}=0}^{M}\cdots\sum_{x_{ns}=0}^{M} P(W_{n1} \mid X_{n1} \\ = x_{n1}) P(W_{n2} \mid X_{n2} = x_{n2}) \\ \quad \times \ldots P(W_{ns} \mid X_{ns} = x_{ns}) \\ \quad \times P(X_n, \lambda, \mu, T)\end{array}\right\}\right).$$

Han et al. (2013) *Molecular Biology and Evolution*

Error increases estimated rate of gain and loss

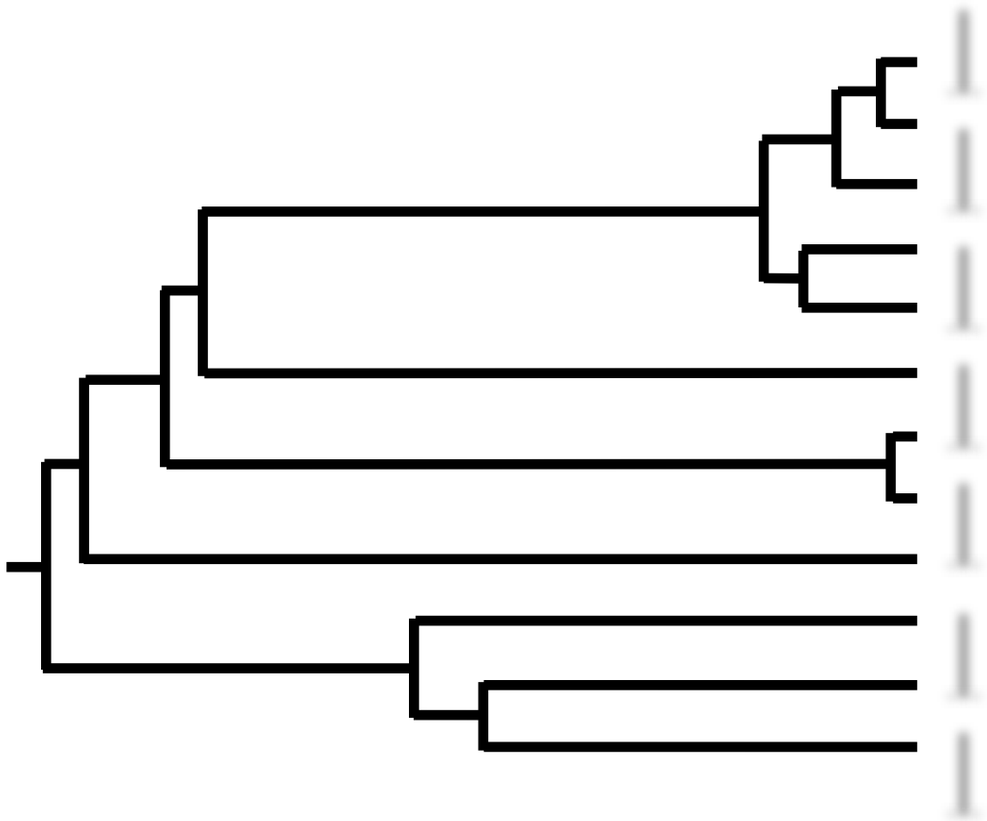


simulated rate: **λ=0.0012**

add **10%** error to data

after adding error: **λ=0.0027**

Accounting for error corrects rate estimate



simulated rate: **λ=0.0012**

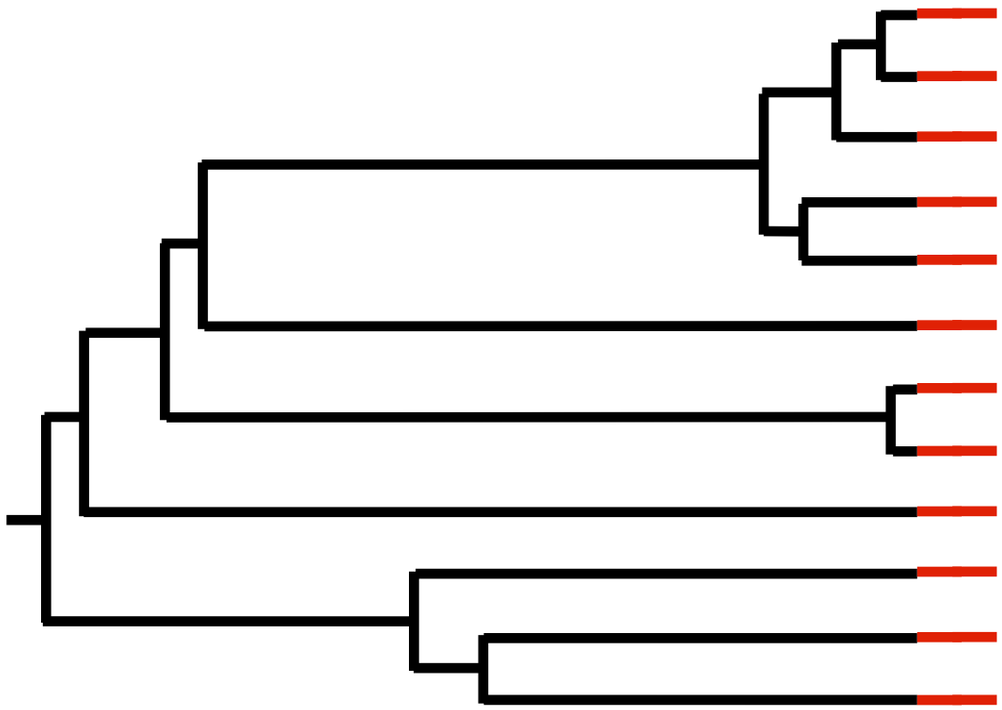
add **10%** error to data →

after adding error: **λ=0.0027**

↓

using correct error model: λ=0.0012

Error increases estimated rate of gain and loss

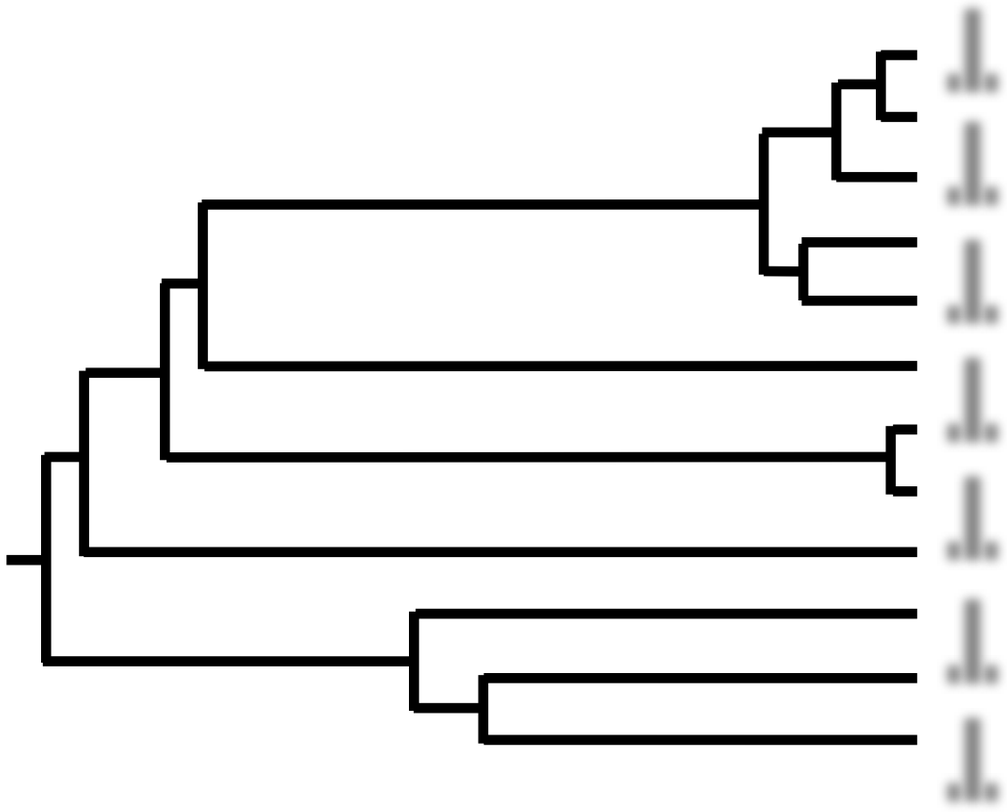


simulated rate: **λ=0.0012**

add **40%** error to data →

after adding error: **λ=0.0085**

Accounting for error corrects rate estimate



simulated rate: **λ=0.0012**

add **40%** error to data →

after adding error: **λ=0.0085**

using correct error model: λ=0.00124

Estimating the correct error model

Thus far, we have assumed the correct error model is known.

Can we estimate it when it is not known?

Estimating the correct error model

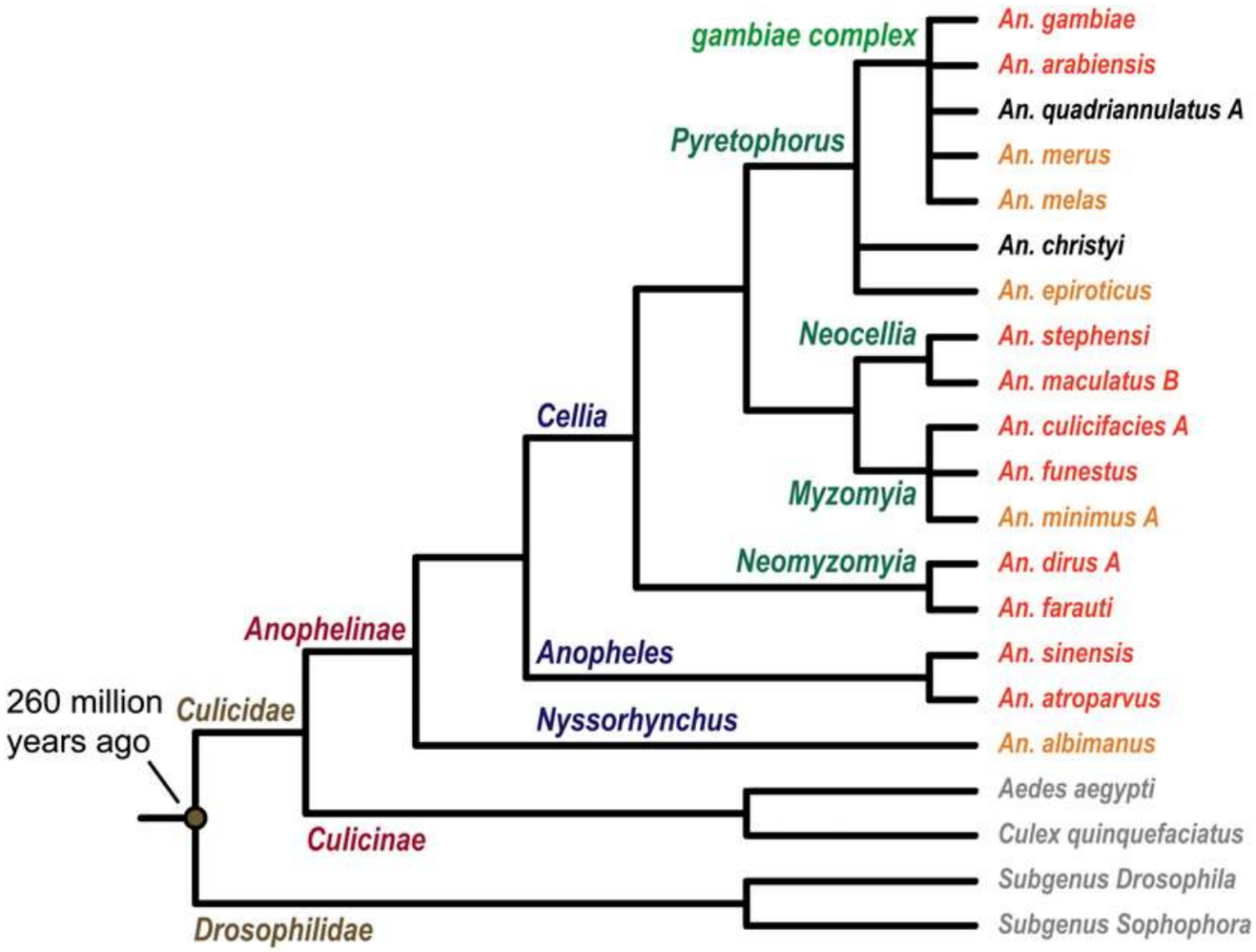
simulated error: 0.1

simulated error: 0.4

Estimating the error model from data

Using the real 12 Drosophila genomes

Accounting for error corrects rate estimate



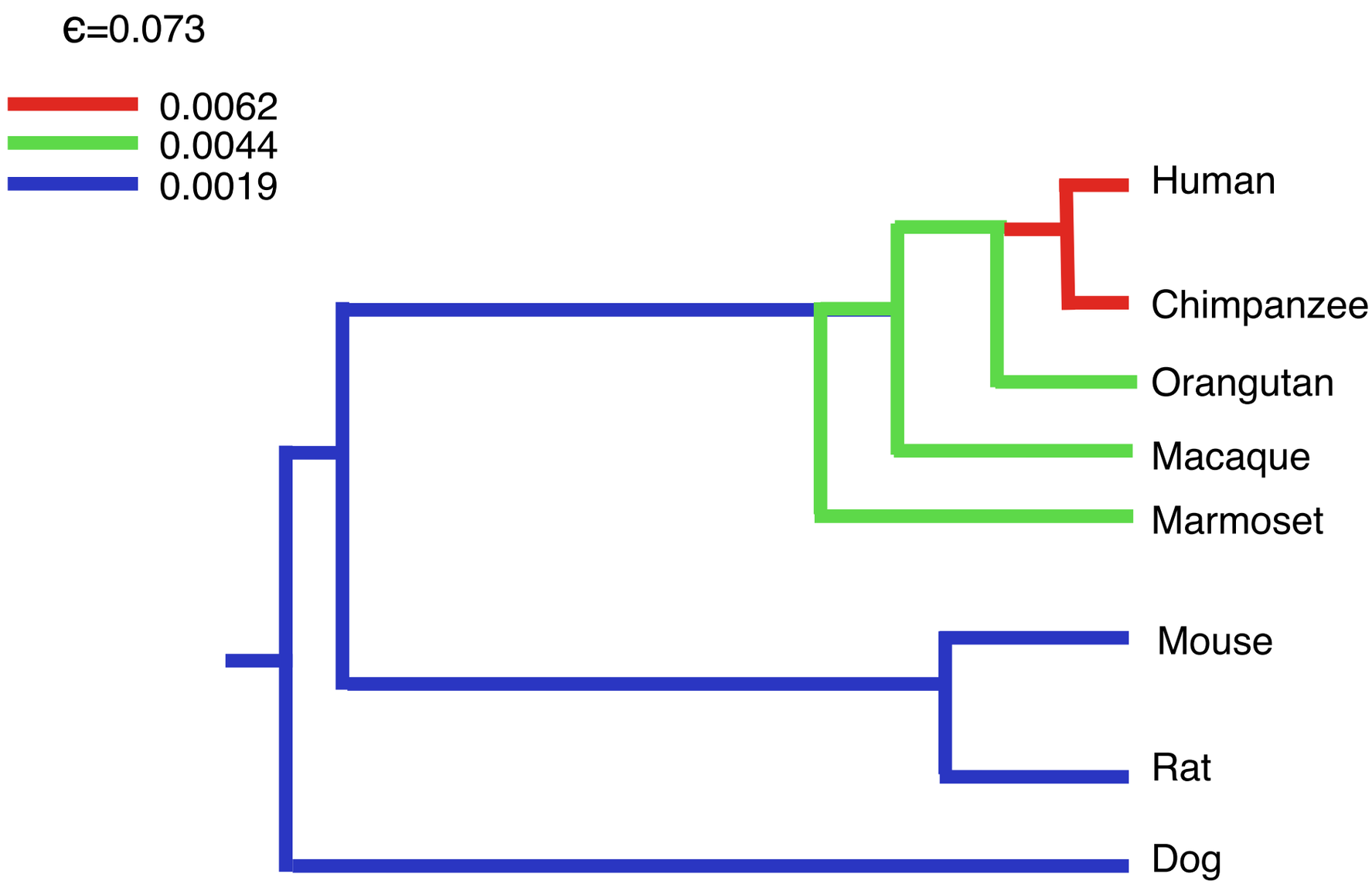
uncorrected λ=0.0032

ϵ=0.104

corrected λ=0.0018

Neafsey et al. (2015) *Science*

Is there really a hominoid rate acceleration?



Han et al. (2013) *Molecular Biology and Evolution*