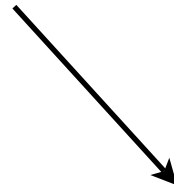
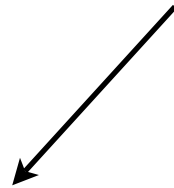


Quantifying gene gain and loss in mammals

Matthew Hahn

Department of Biology
& School of Informatics
Indiana University

The King and Wilson paradox



The King and Wilson paradox

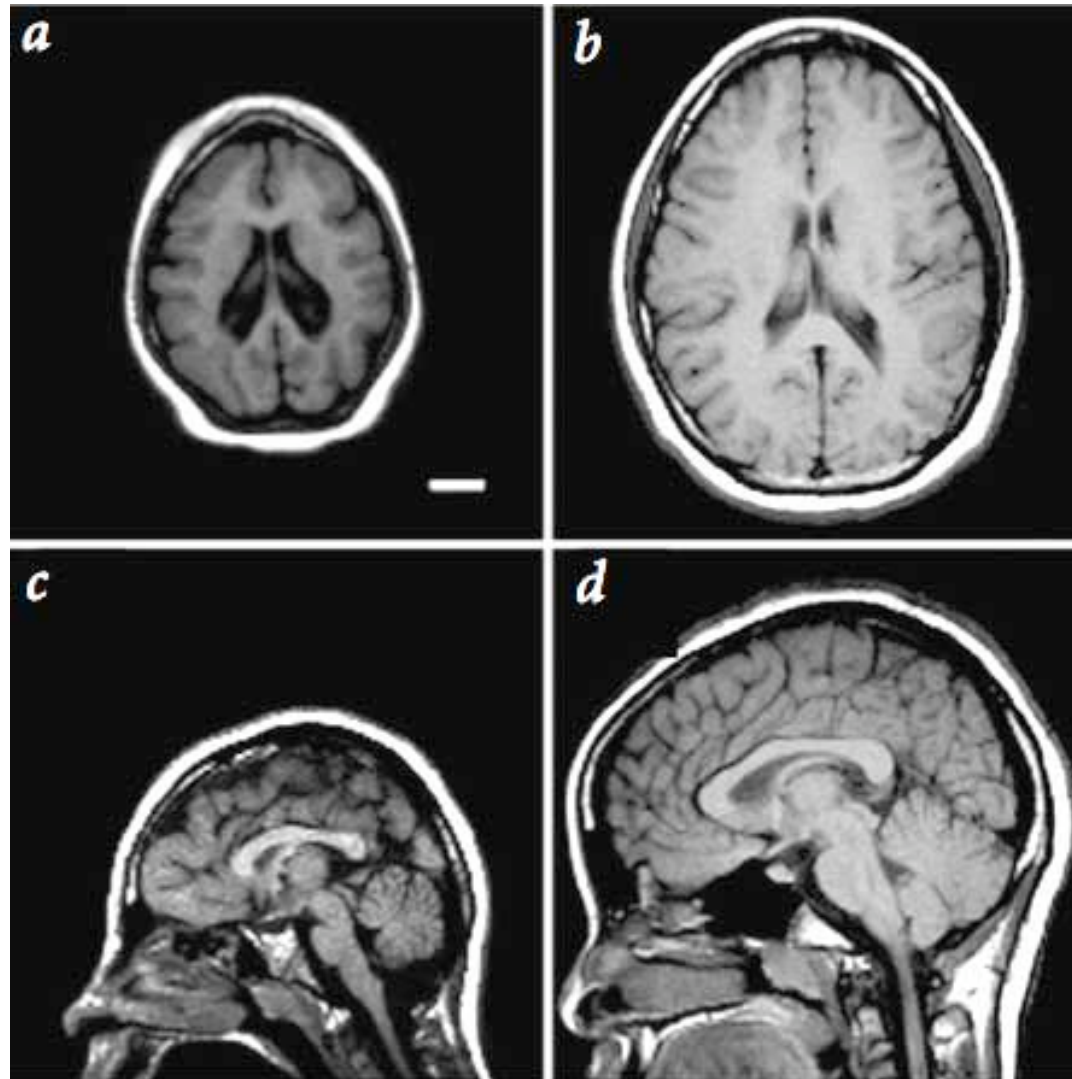
“...the genetic distance between humans and the chimpanzee is probably too small to account for their substantial organismal differences.”

M.-C. King and A. Wilson 1975

Solutions to the paradox

- Coding (Classic)

The *ASPM* protein evolves rapidly and controls brain size



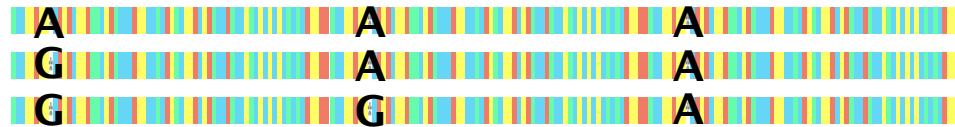
from Bell et al. (2004)

Solutions to the paradox

- Coding (Classic)
- *cis*-Regulatory (King and Wilson)

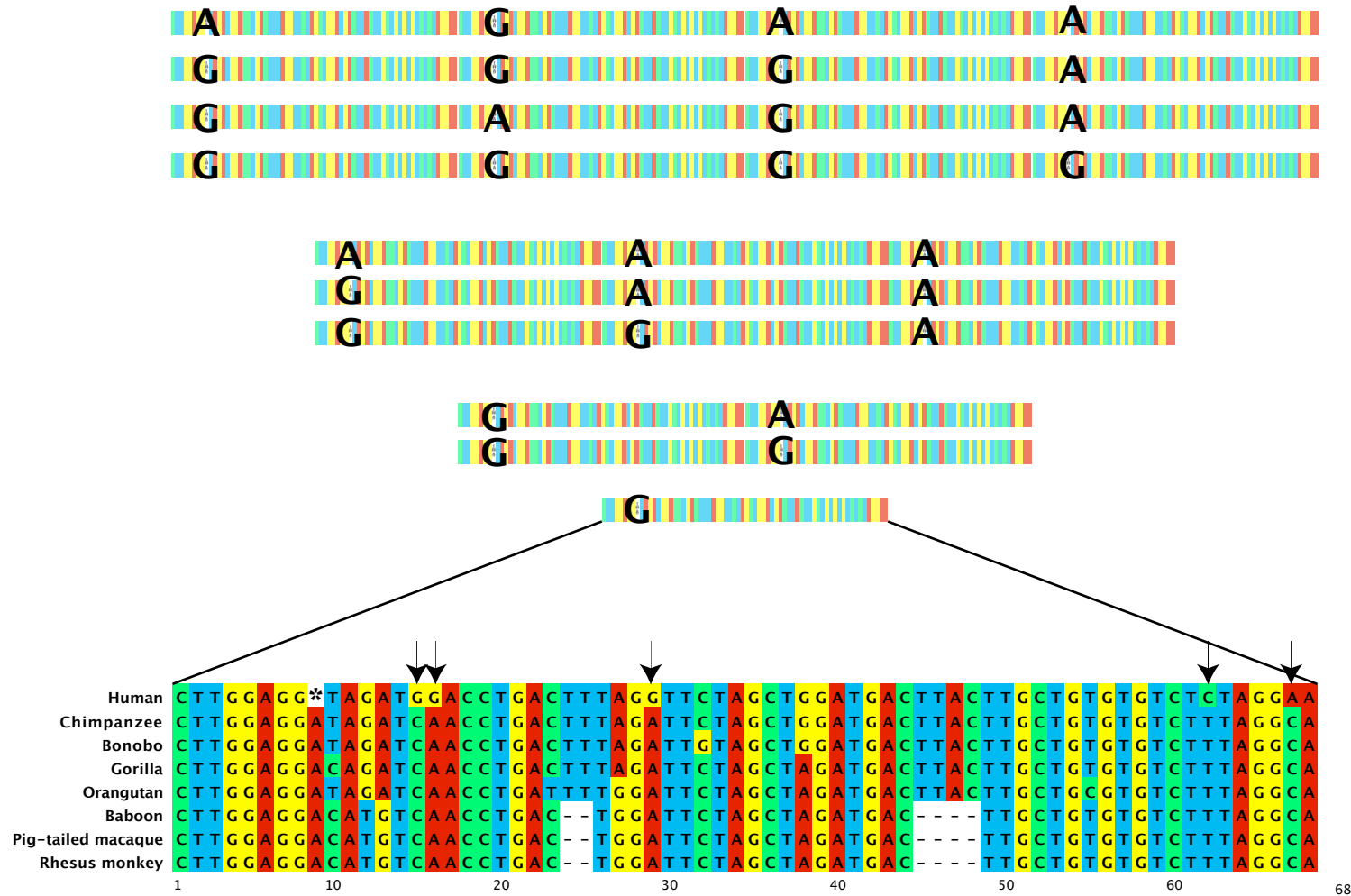
Prodynorphin in humans

Prodynorphin (PDYN) controls the expression of endorphins.



more repeats, more endorphins

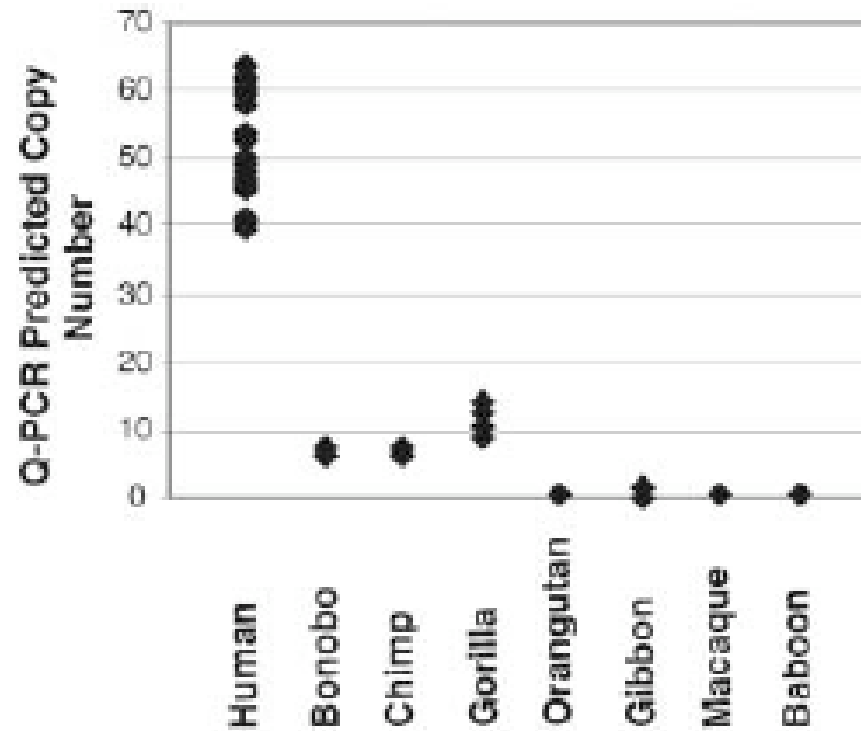
Prodynorphin evolves rapidly in humans



Solutions to the paradox

- Coding (Classic)
- *cis*-Regulatory (King and Wilson)
- Gene duplication (S. Ohno)

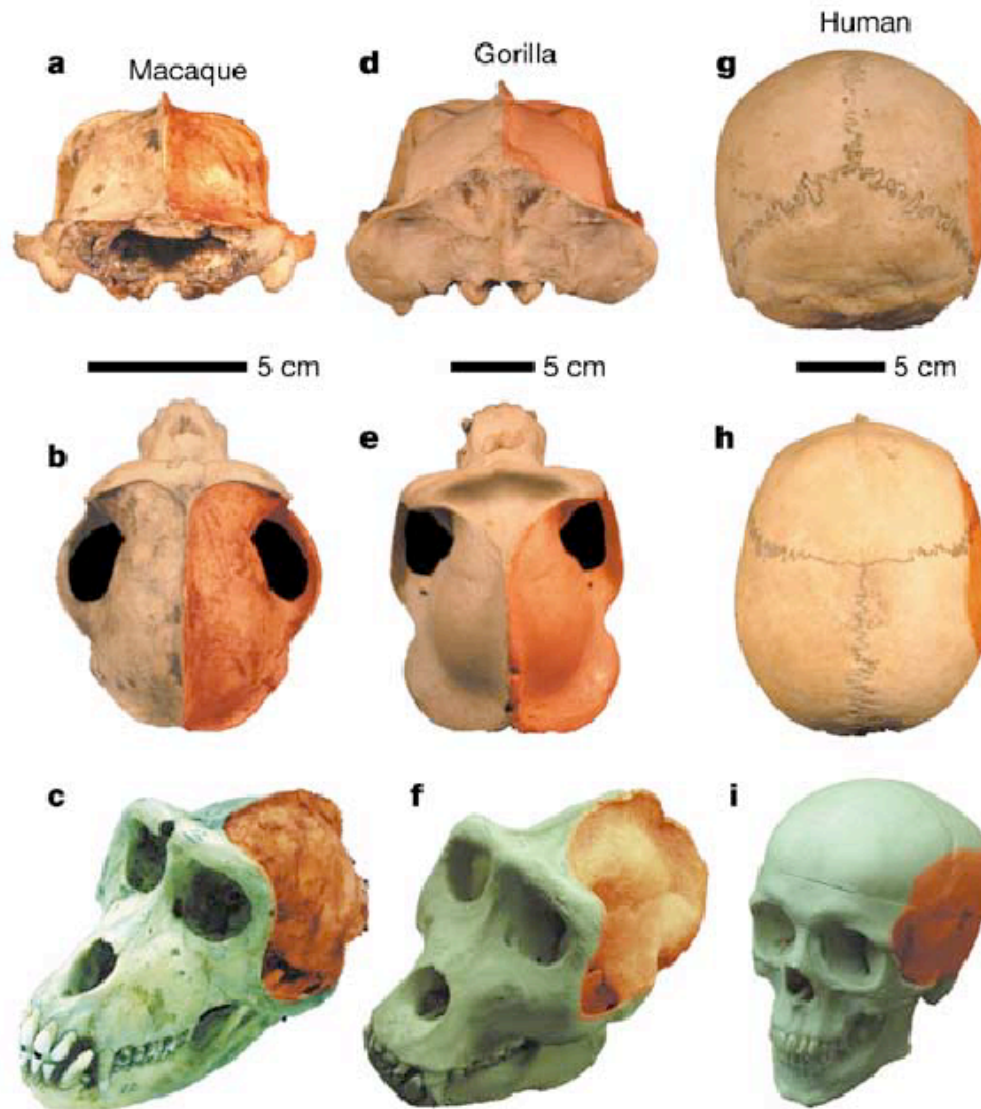
DUF1220 is highly duplicated in humans



Solutions to the paradox

- Coding (Classic)
- *cis*-Regulatory (King and Wilson)
- Gene duplication (S. Ohno)
- Gene loss (“Less is more”)

Loss of myosin associated with cranial enlargement



Solutions to the paradox

- Coding (Classic)
- *cis*-Regulatory (King and Wilson)
- Gene duplication (S. Ohno)
- Gene loss (“Less is more”)

Two aims:

- Quantify the amount of gene gain and loss
- Infer the action of natural selection

Outline

- I. Statistical and computational methods
- II. The amount of gene turnover in mammals
- III. Natural selection on gene duplicates

The evolution of gene families

Gene families are groups of genes that share sequence and functional homology

The evolution of gene families

The size of gene families changes among species.

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>D. melanogaster</i>	<i>H. sapiens</i>	<i>A. thaliana</i>
Homeodomain	9	109	148	267	118
Zinc-finger	121	437	357	706	1049
Nuclear receptor	1	183	25	59	4

from Venter et al. (2001)

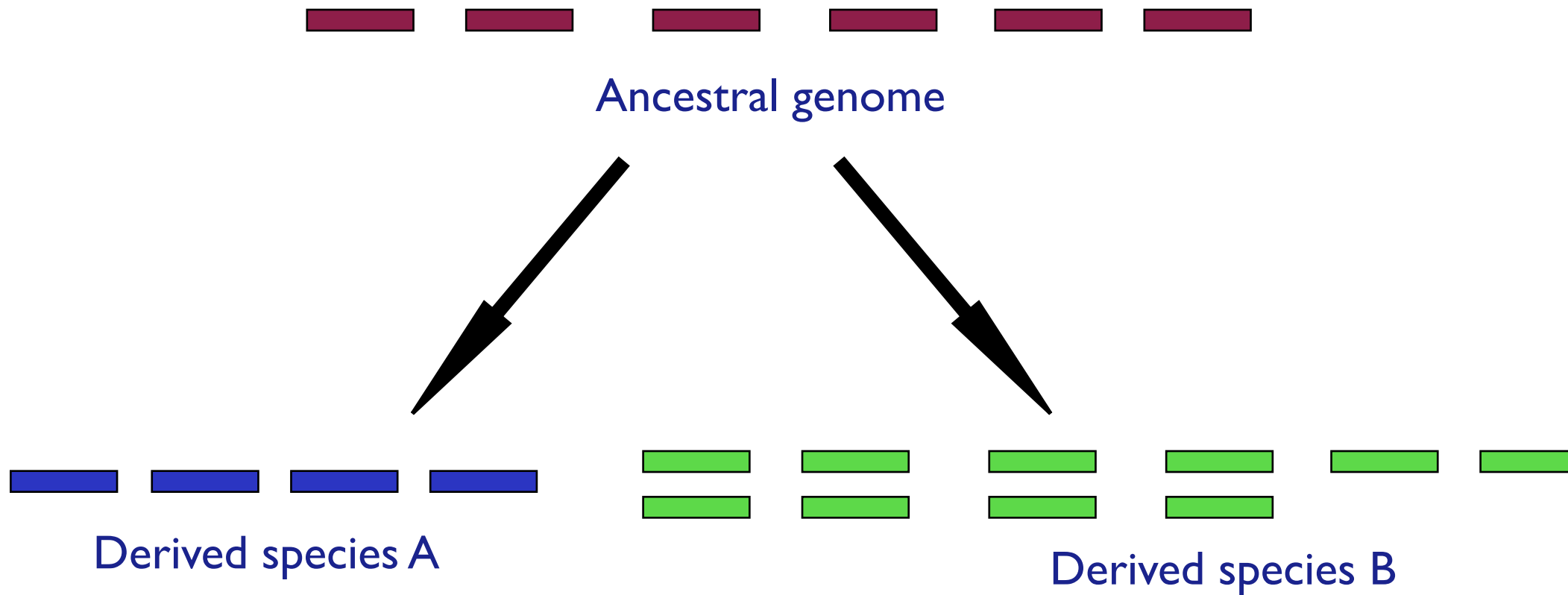
A model for gene family evolution

Homogeneous birth and death process

Birth = duplication

Death = deletion or pseudogenization

Birth-death model of gene family evolution



There are no true models, only helpful ones.

--G.E.P. Box

No model, no inference.

--J. Felsenstein

Birth-death model of gene family evolution

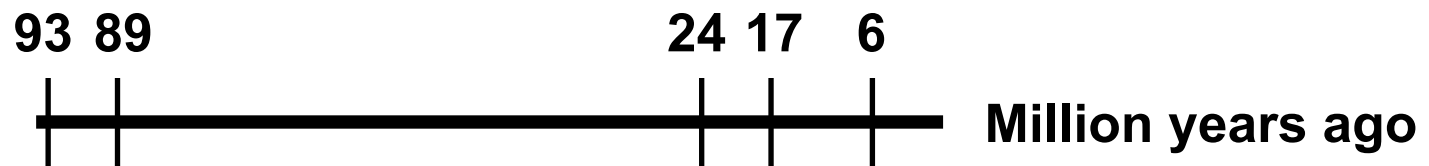
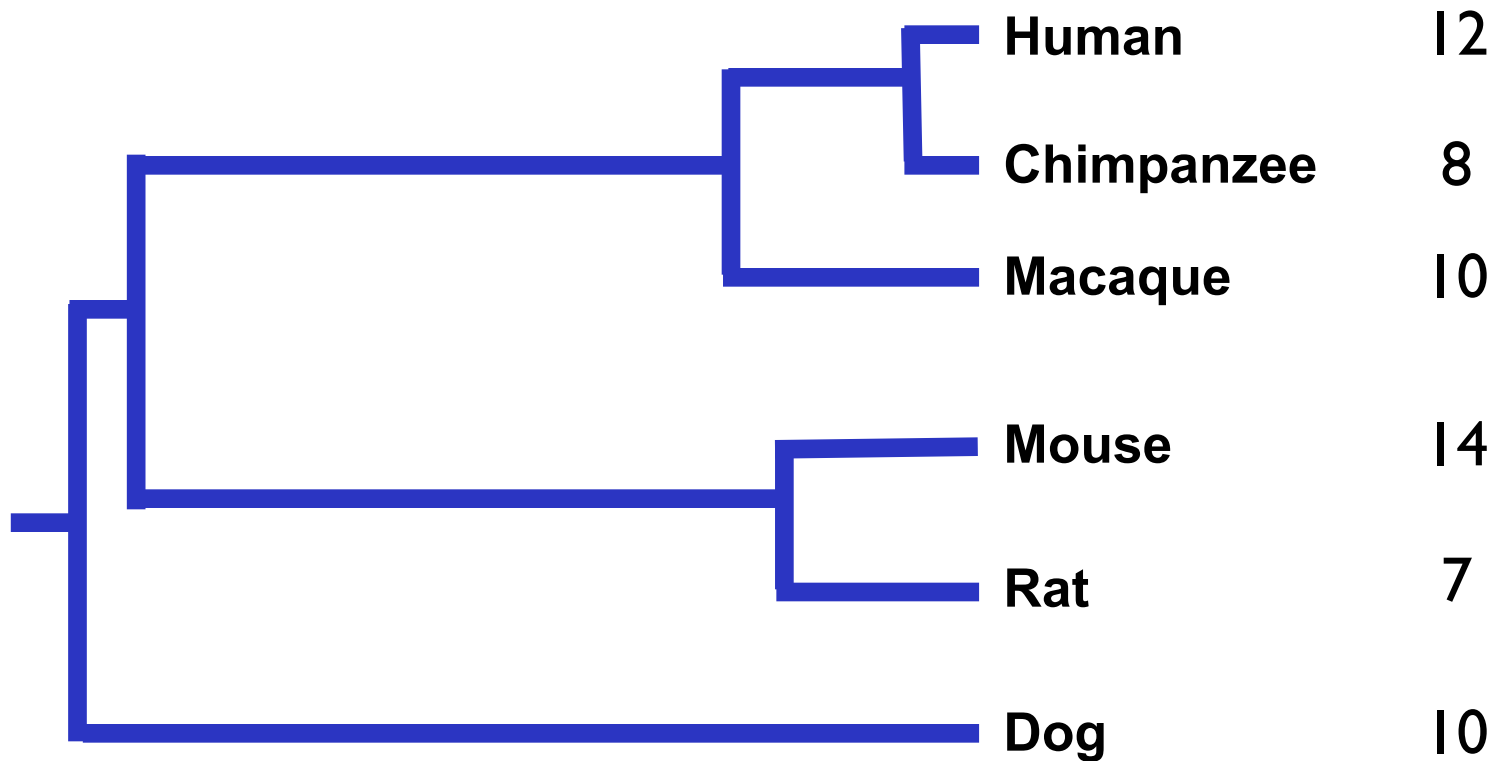
Birth-Death transition probability (Bailey 1964):

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

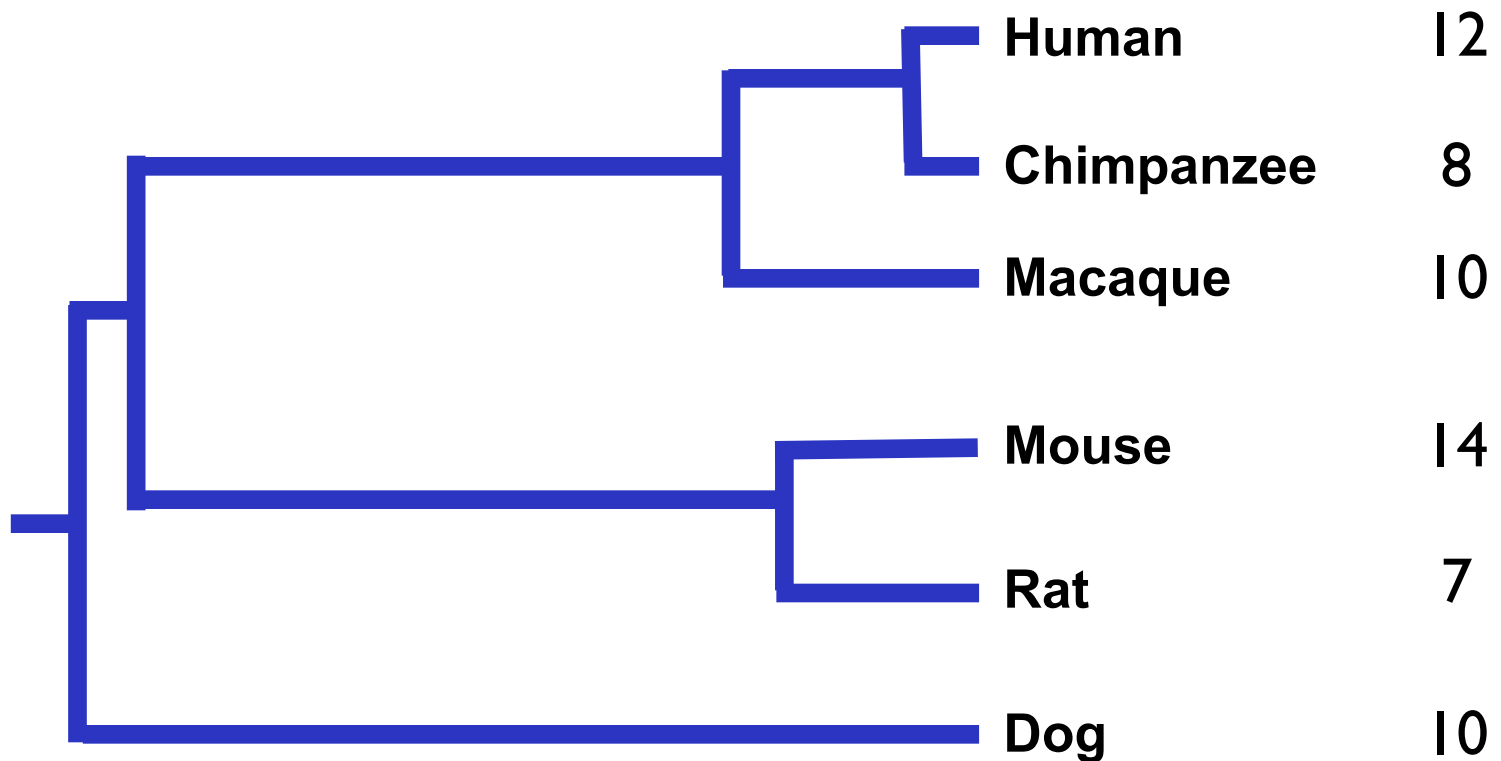
The necessary parameters:

- Current family size
- Ancestral family size
- Time since divergence
- Gain and loss rates

Inferring rates of gain and loss

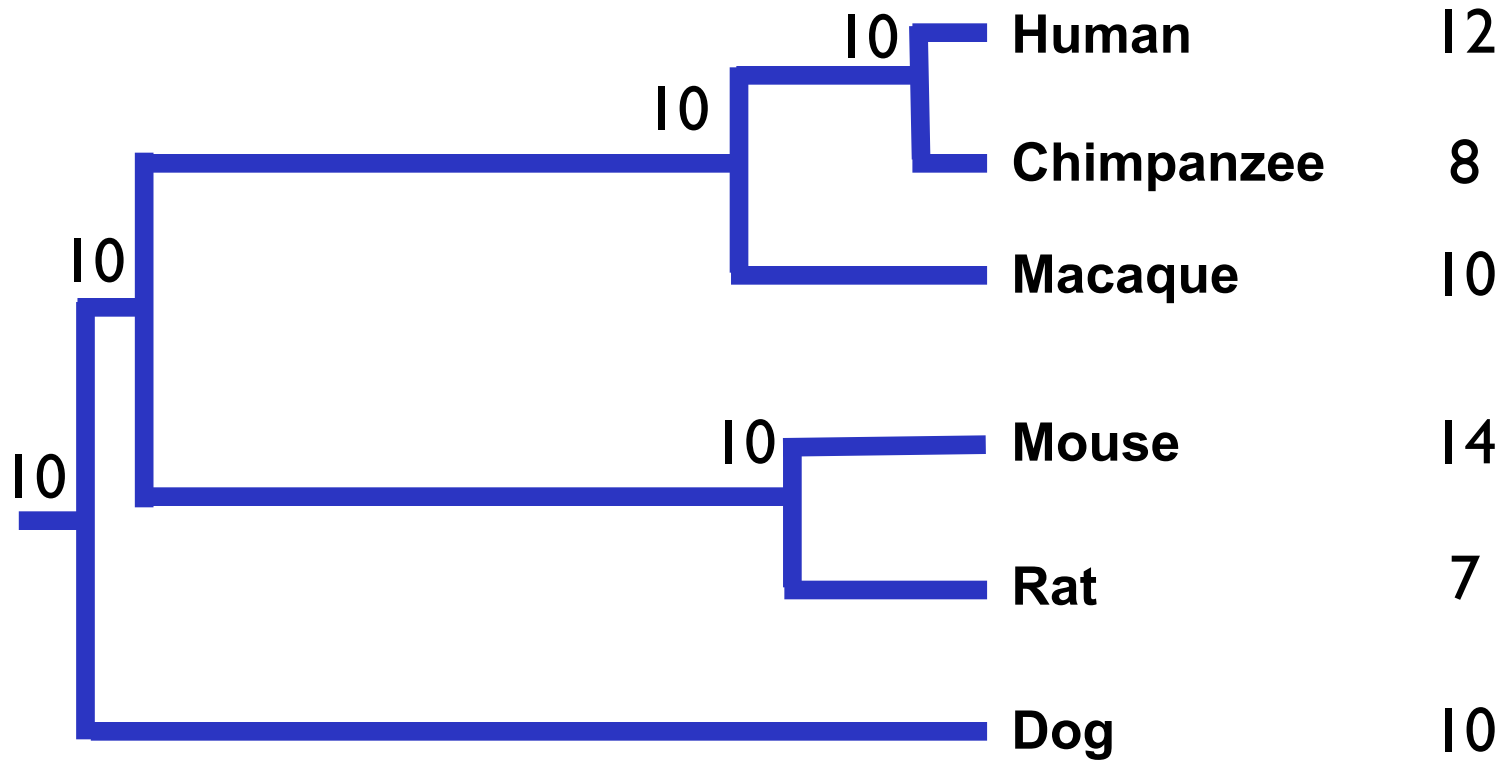


Inferring rates of gain and loss

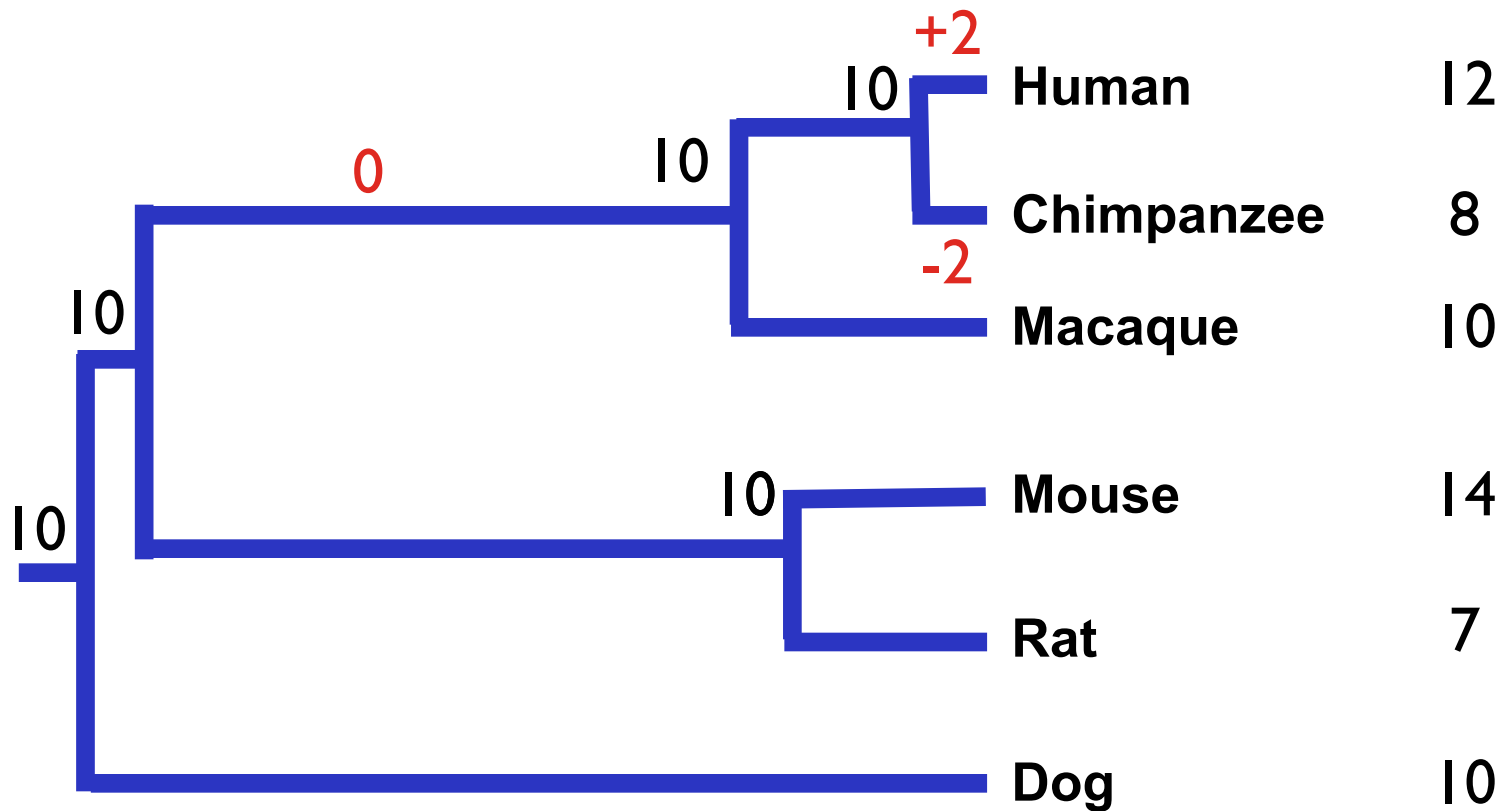


We estimate the average rate across all families

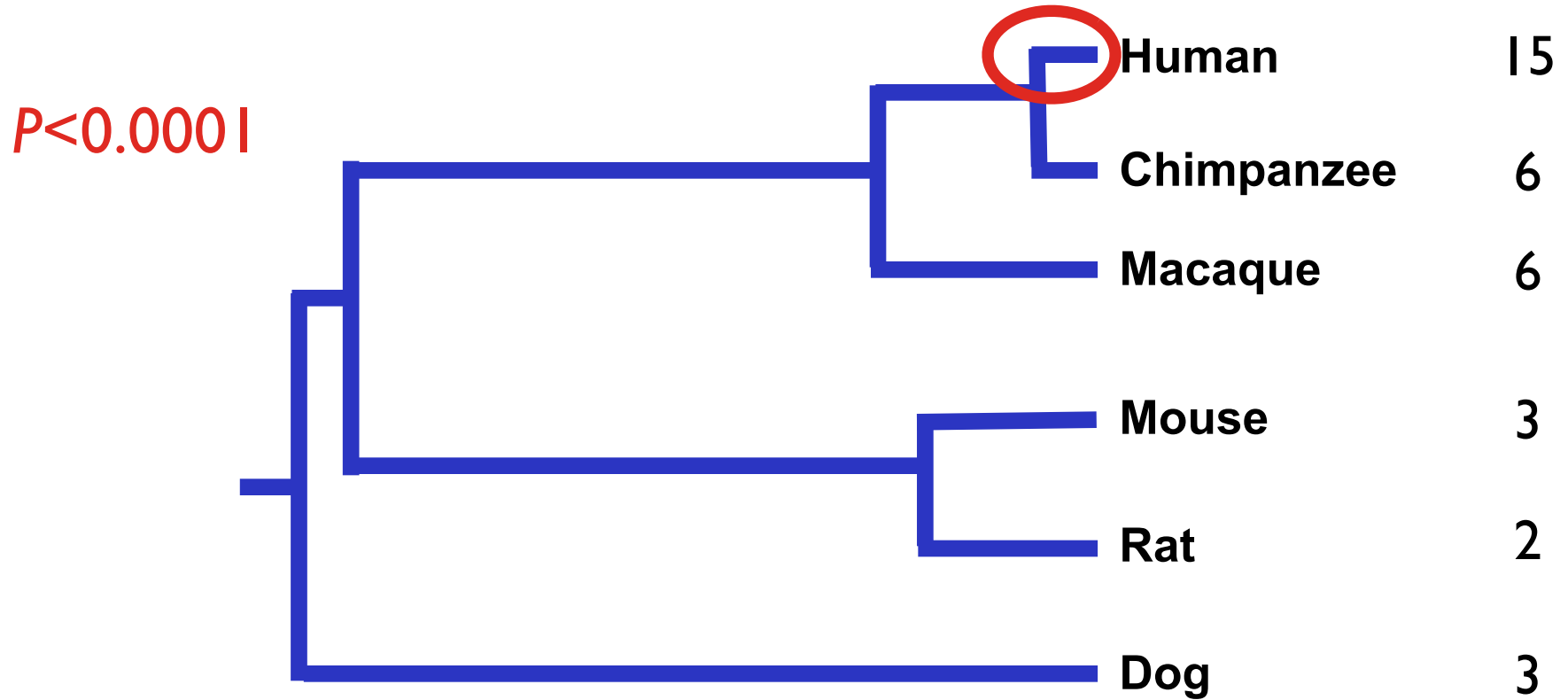
Inferring ancestral states



Inferring changes in gene family size



Identifying rapidly-evolving families



CAFE

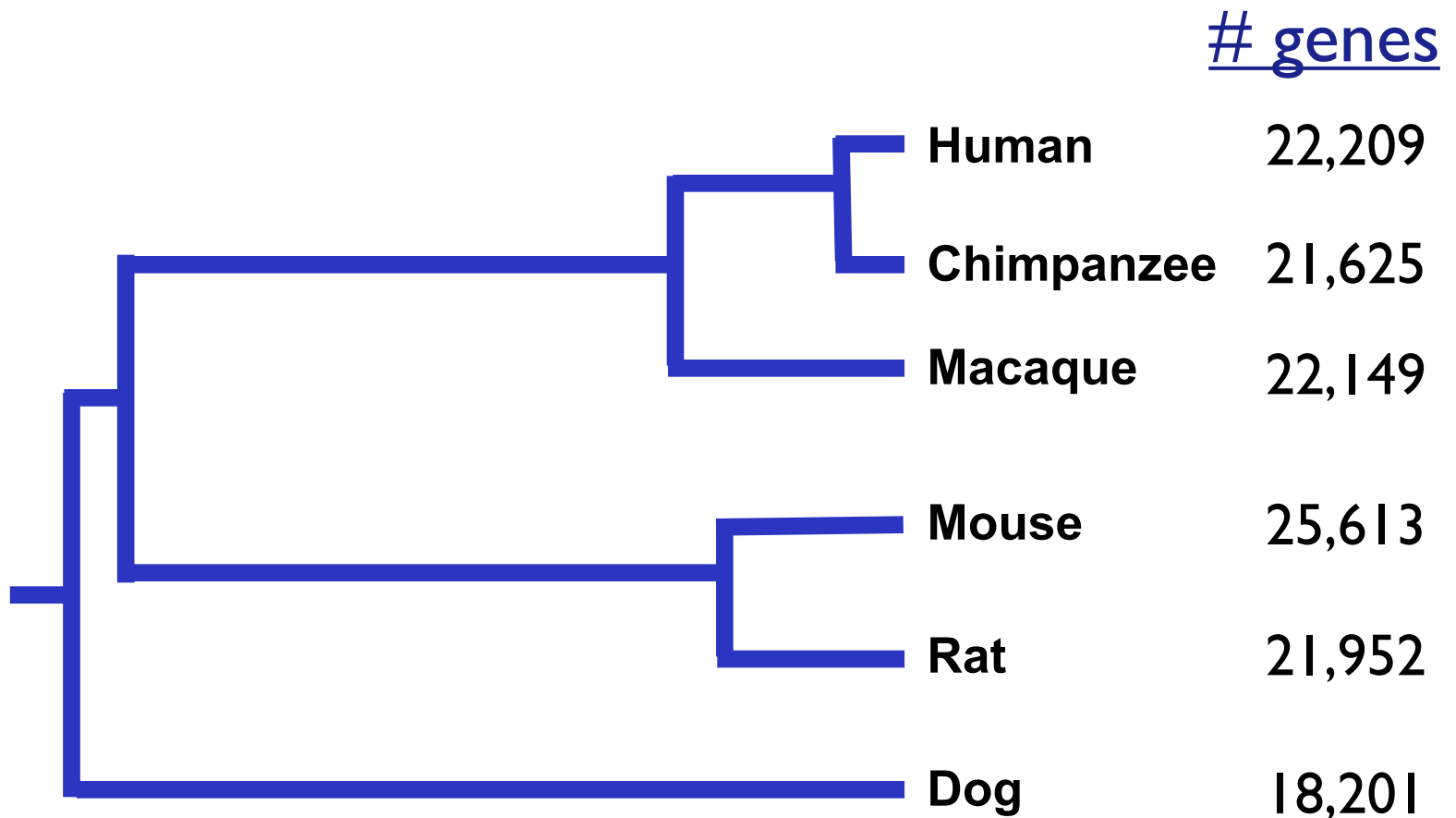
(Computational Analysis of gene Family Evolution)

The screenshot shows the CAFE software interface with the following fields and options:

- Data file: [Text input field] [Browse ...]
- Destination file: [Text input field] [Browse ...]
- Tree structure: [Text input field] [Enter a Newick formatted tree with branch lengths here]
- Lambda: [Text input field] [Enter a guess or final value for lambda] Train lambda using EM
- P-value threshold: [Text input field] [Enter the p-value threshold]
- Number of random samples: [Text input field] [Enter the number of random samples to calculate the p]
- Choose methods to identify the best branch:
 - Likelihood Ratio Test
 - Stochastic
 - Branch Cutting
- [Show list]
- Step 1: Performing EM [Progress bar]
- Step 2: Caching birth-death process [Progress bar]
- Step 3: Sampling the distributions [Progress bar]
- Step 4: Processing the gene families, including Stochastic [Progress bar]
- Step 5: Performing preprocessing for the branch cutting [Progress bar]
- Step 6: Performing the branch cutting [Progress bar]
- Step 7: Performing LRT [Progress bar]

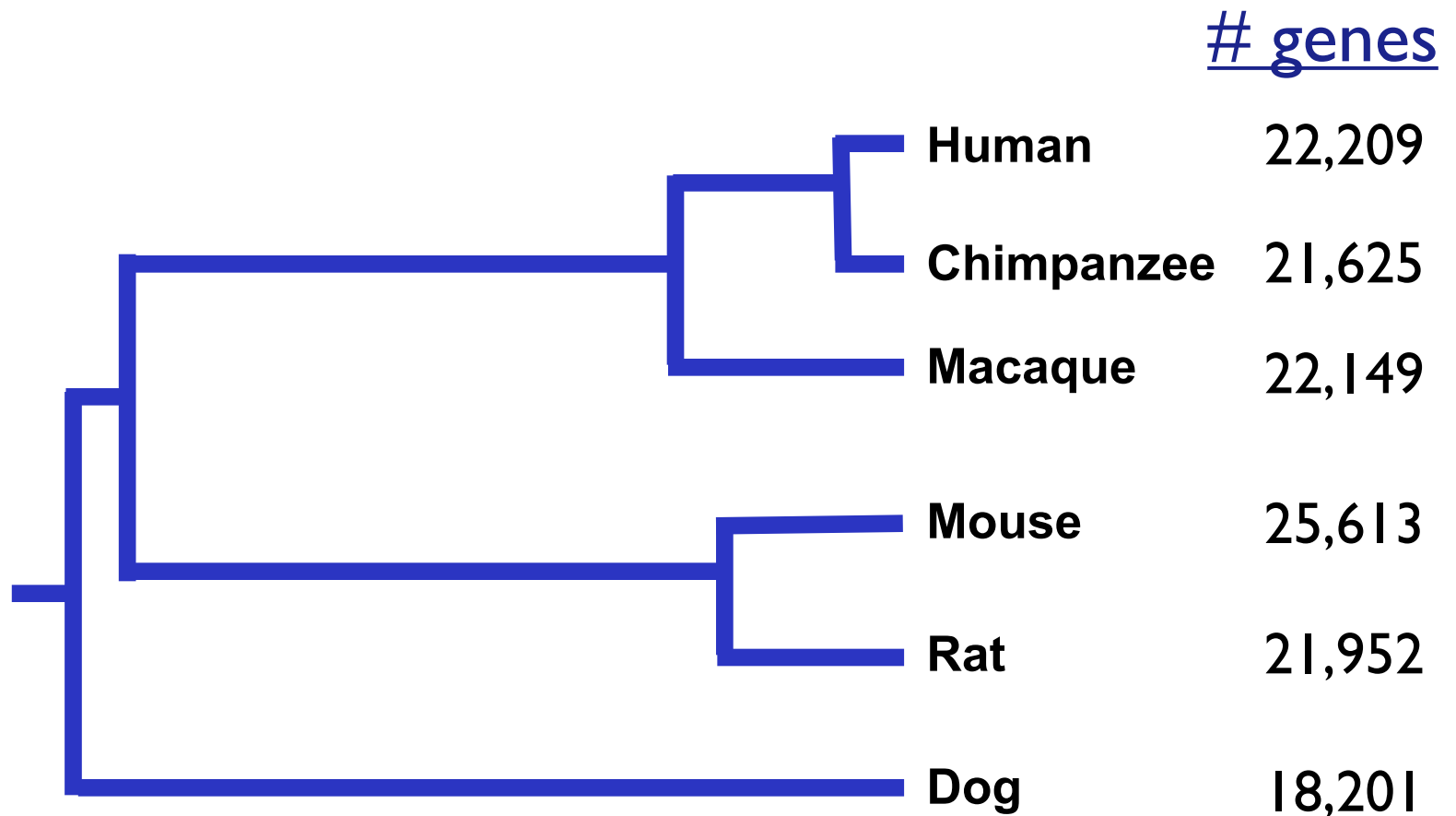
www.bio.indiana.edu/~hahnlab/Software.html

Genome size in mammals



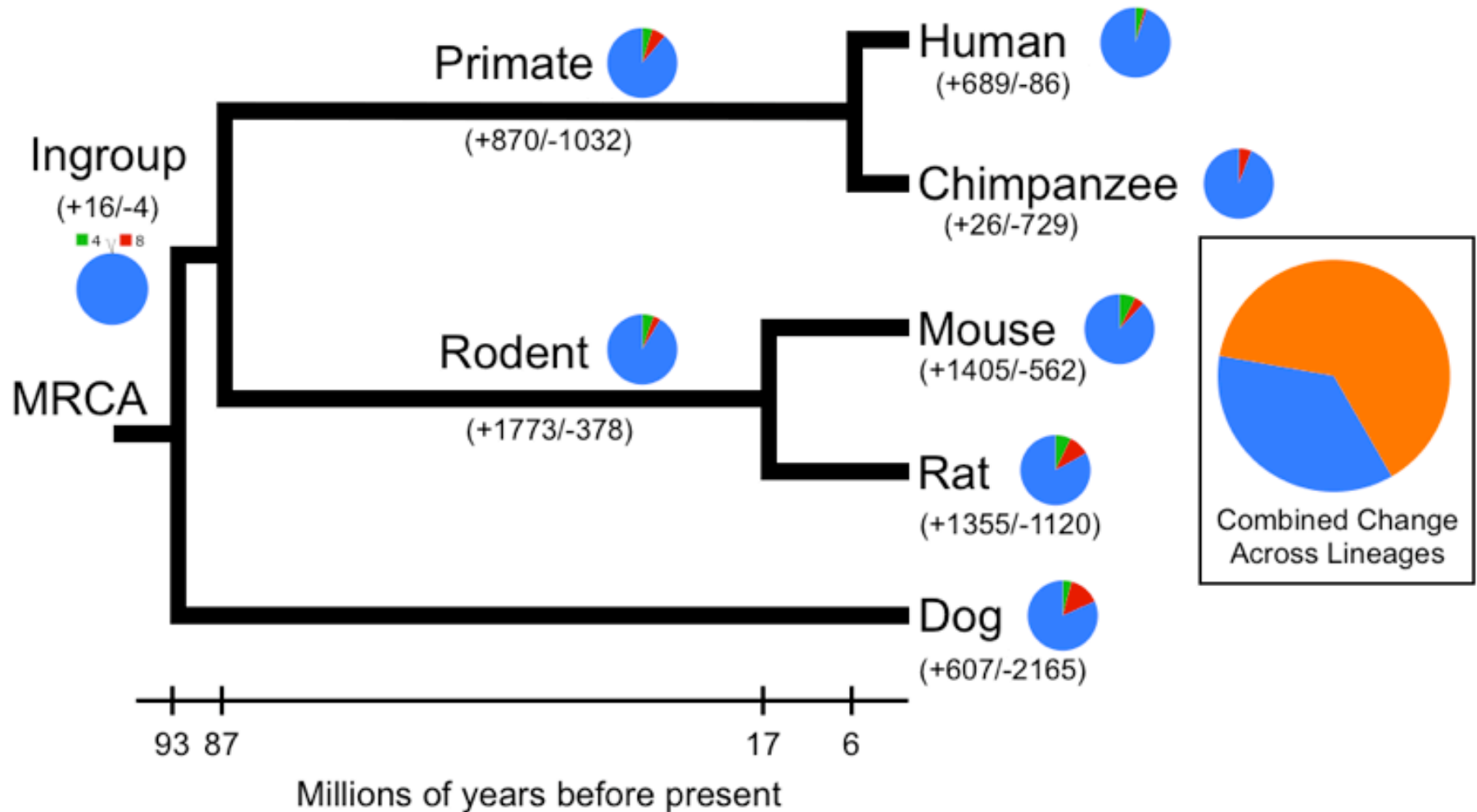
All data from Ensembl v41 (October 2006)

Genome size in mammals

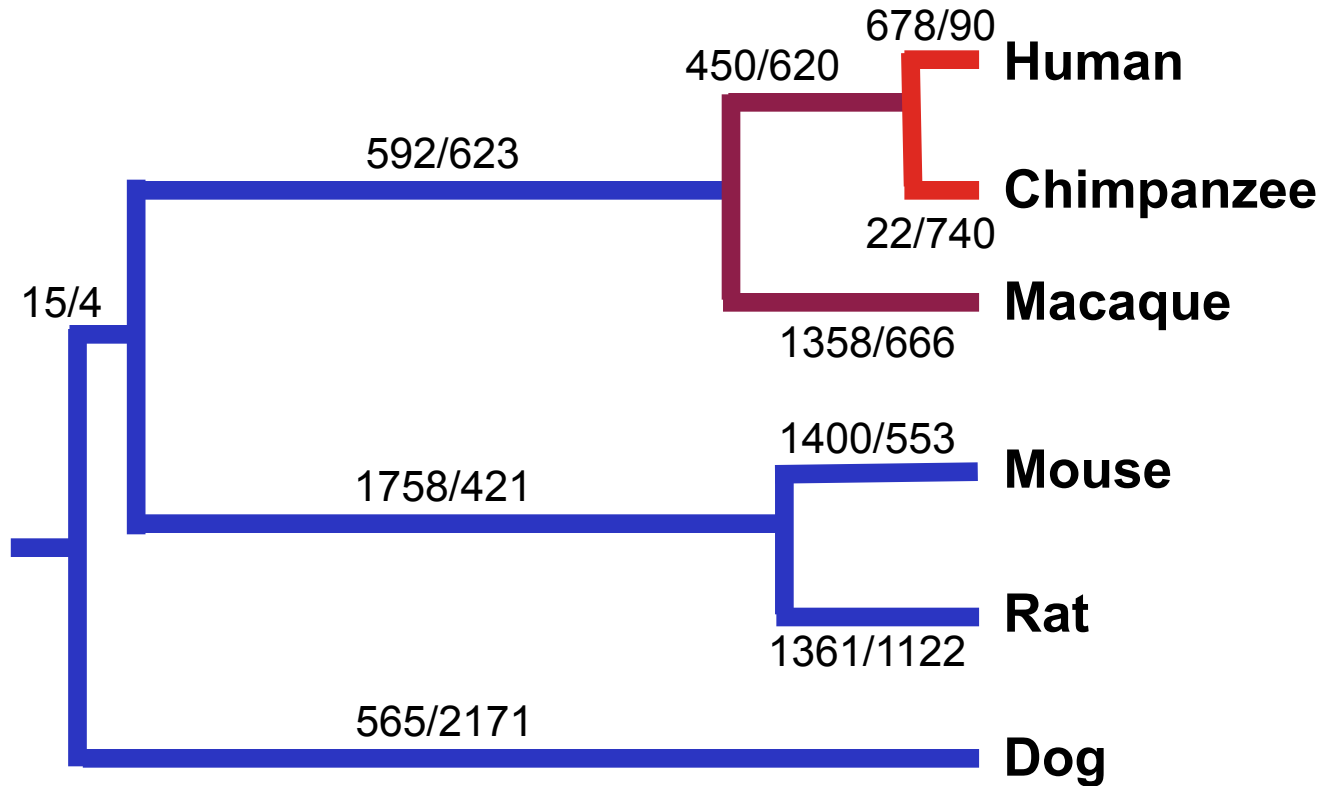


9,990 families found in the mammalian MRCA

Gene gain and loss in mammals



Gene gain and loss in mammals



Gibbs et al. (2007) *Science*
Hahn et al. (in press) *Genetics*

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

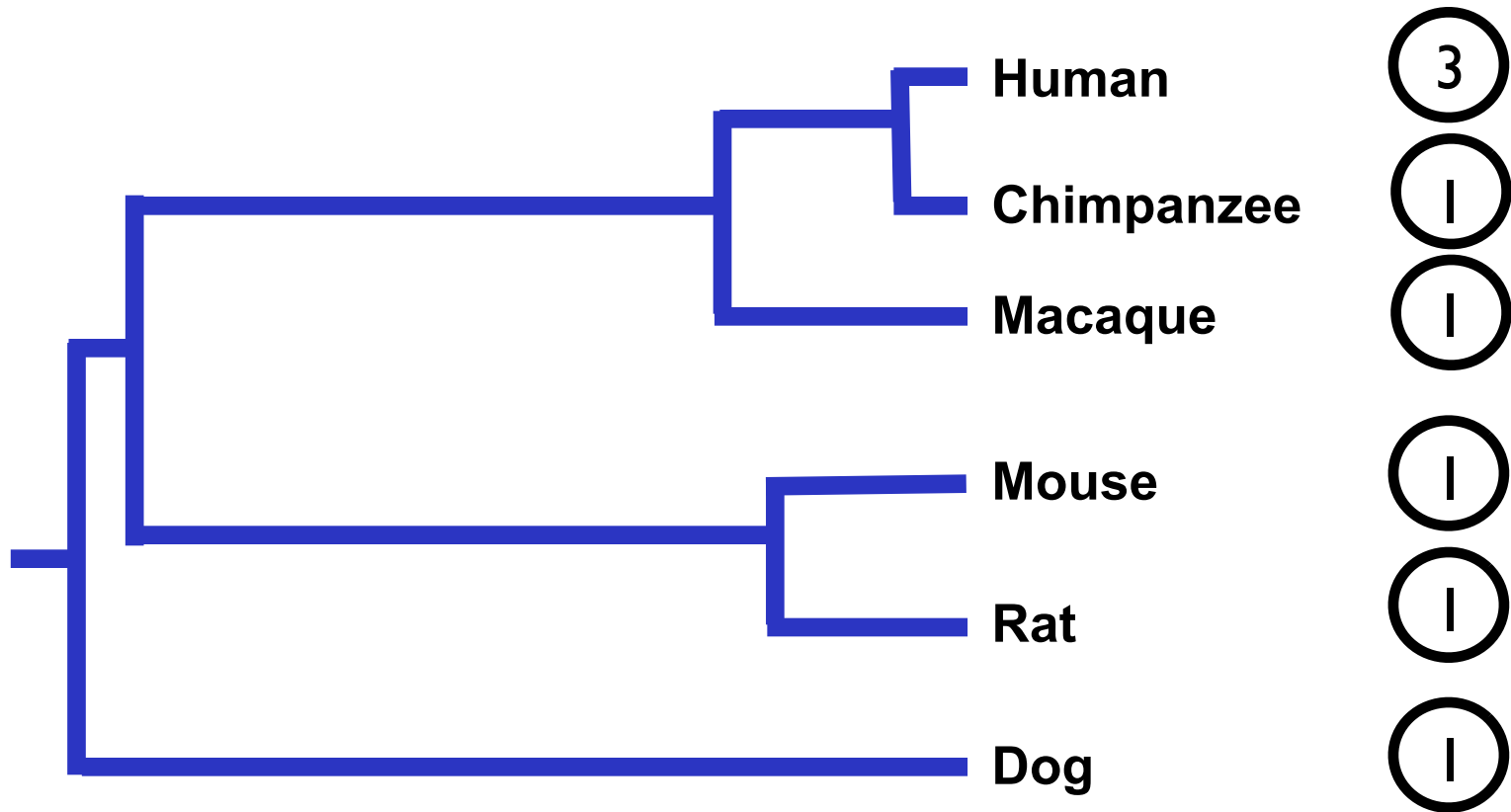
In chimpanzees:

- 740 genes have been lost
- $$\begin{array}{r} + \\ \hline 1415 \end{array}$$

1,415 human genes not found in chimps!

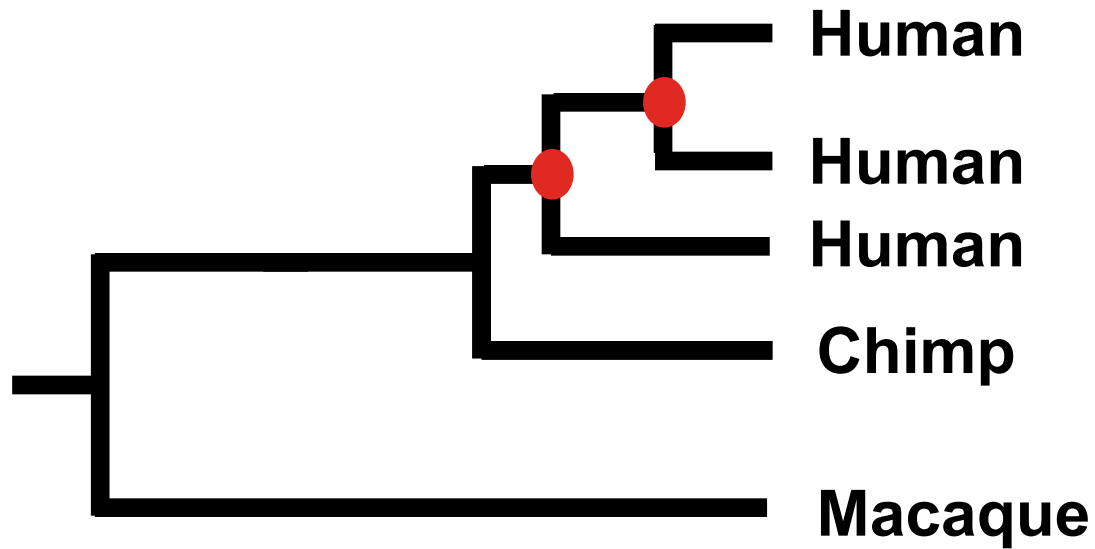
An alternative method for estimating gain and loss

“Genes in a bag”

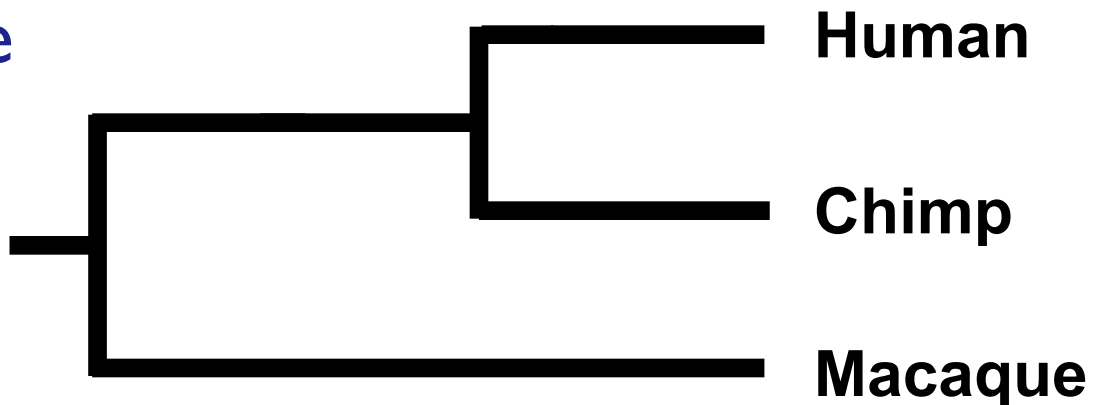


Gene tree-Species tree reconciliation

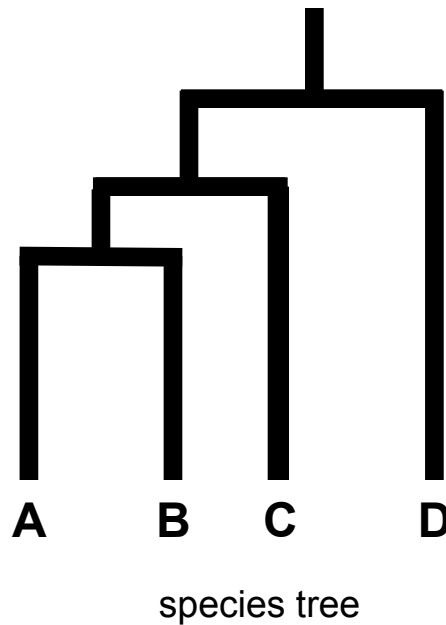
Gene tree



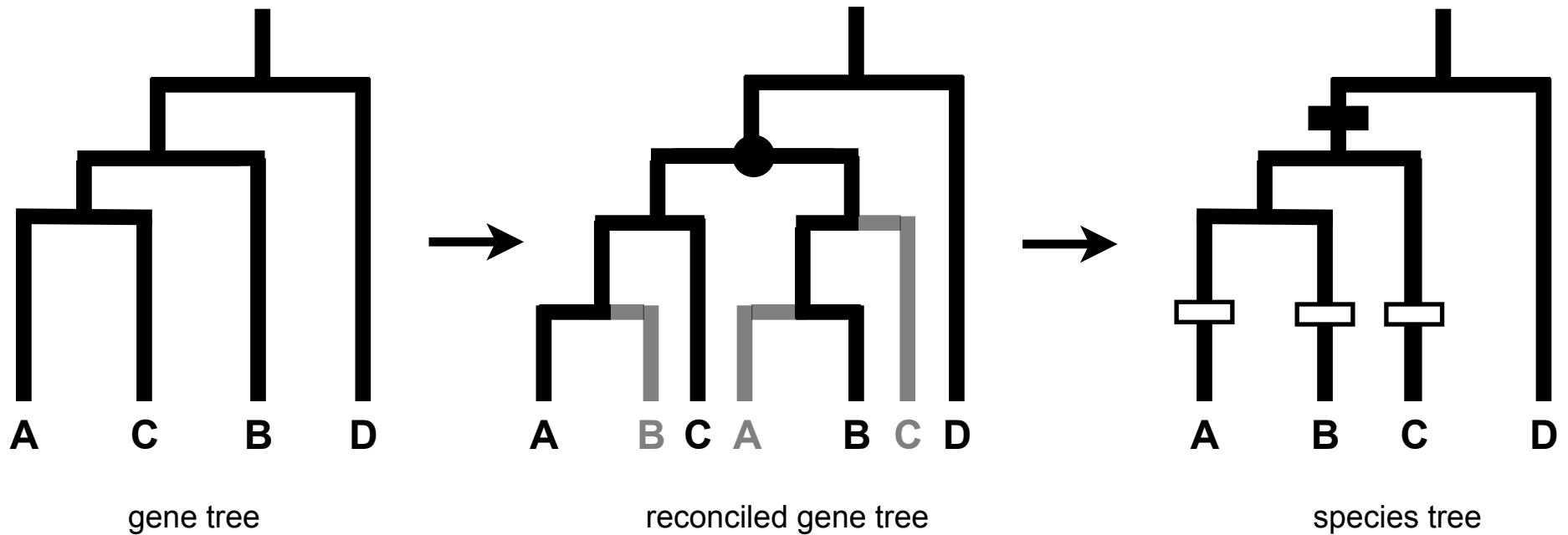
Species tree



Bias in tree reconciliation



Bias in tree reconciliation



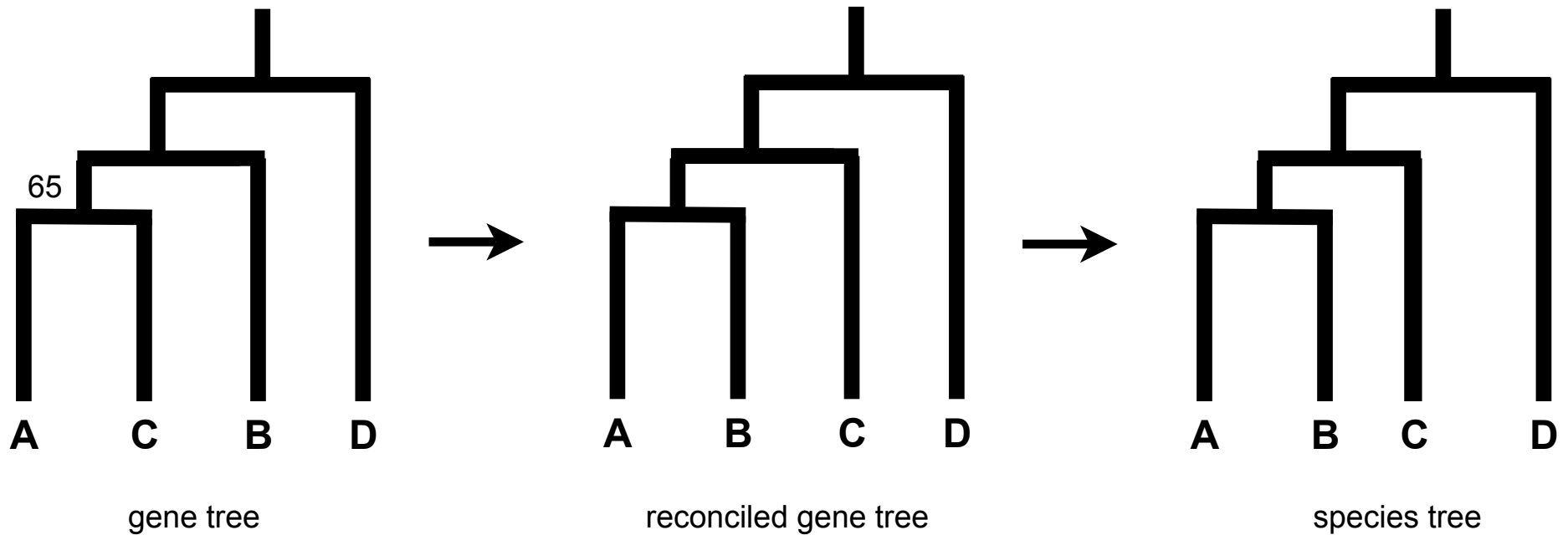
● duplication

▬ loss

▬ gain

□ loss

Using bootstrap cutoff can alleviate bias



● duplication

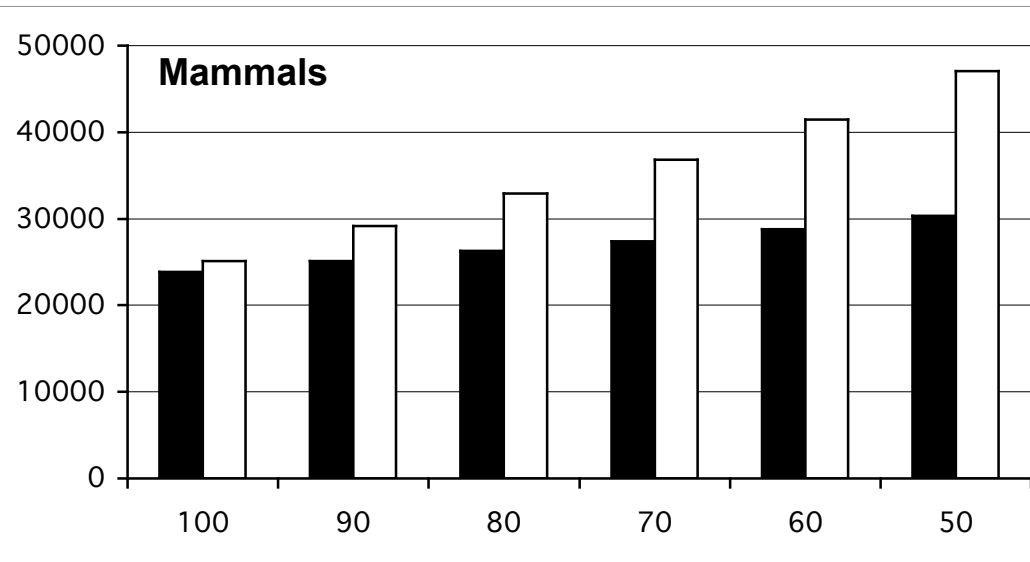
■ loss

■ gain

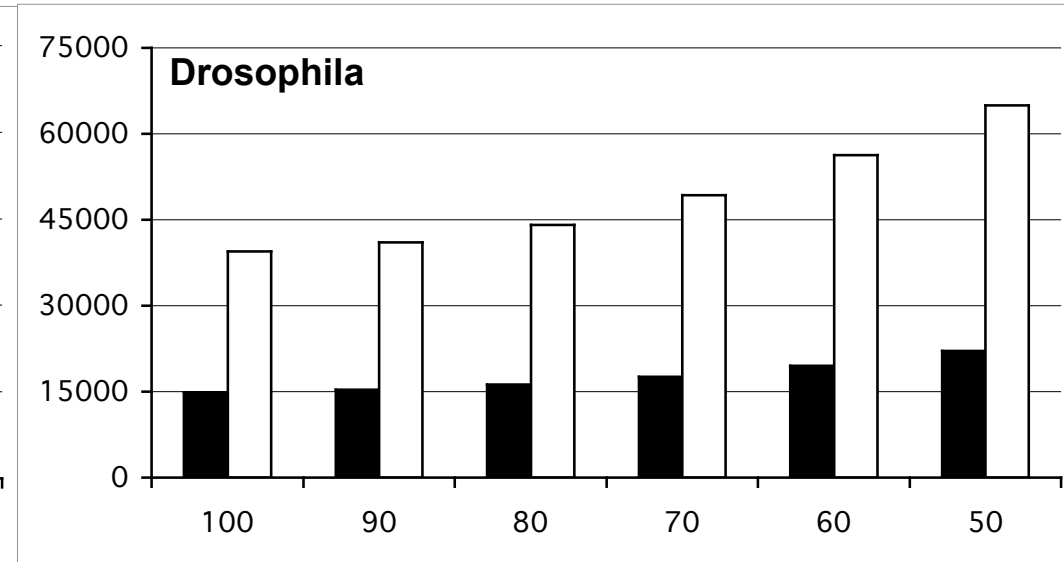
□ loss

Bias increases with decreasing bootstrap cutoff

a)



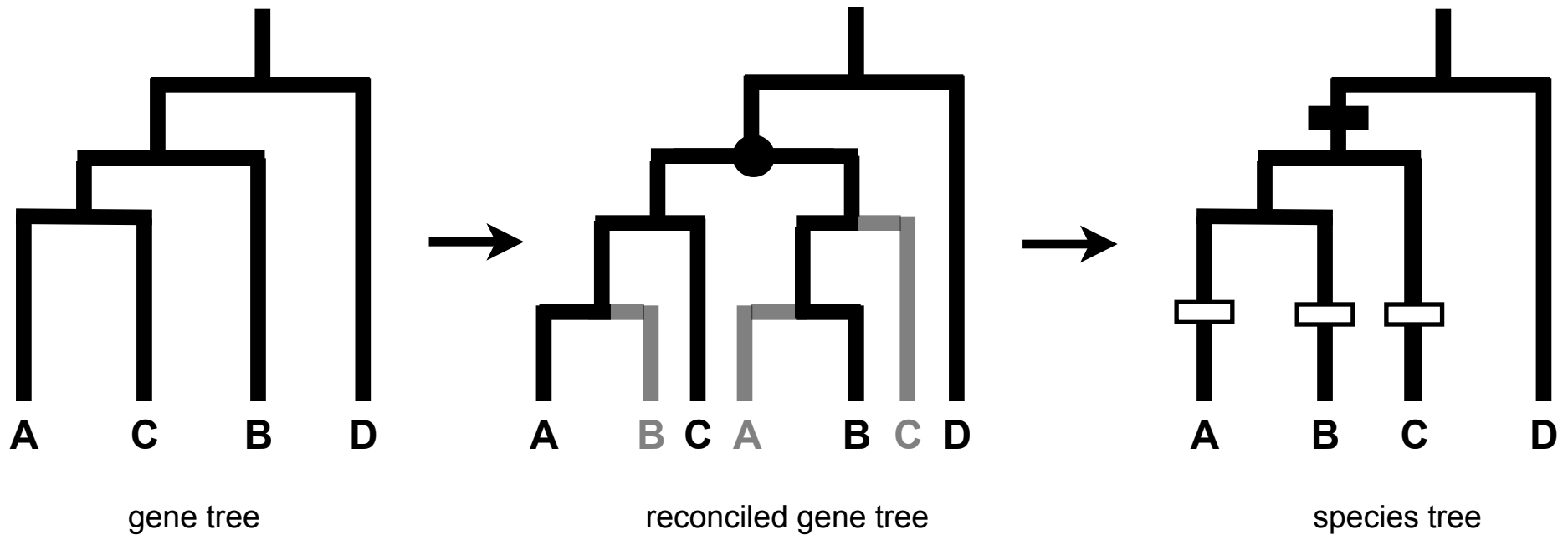
b)



■ gains

□ losses

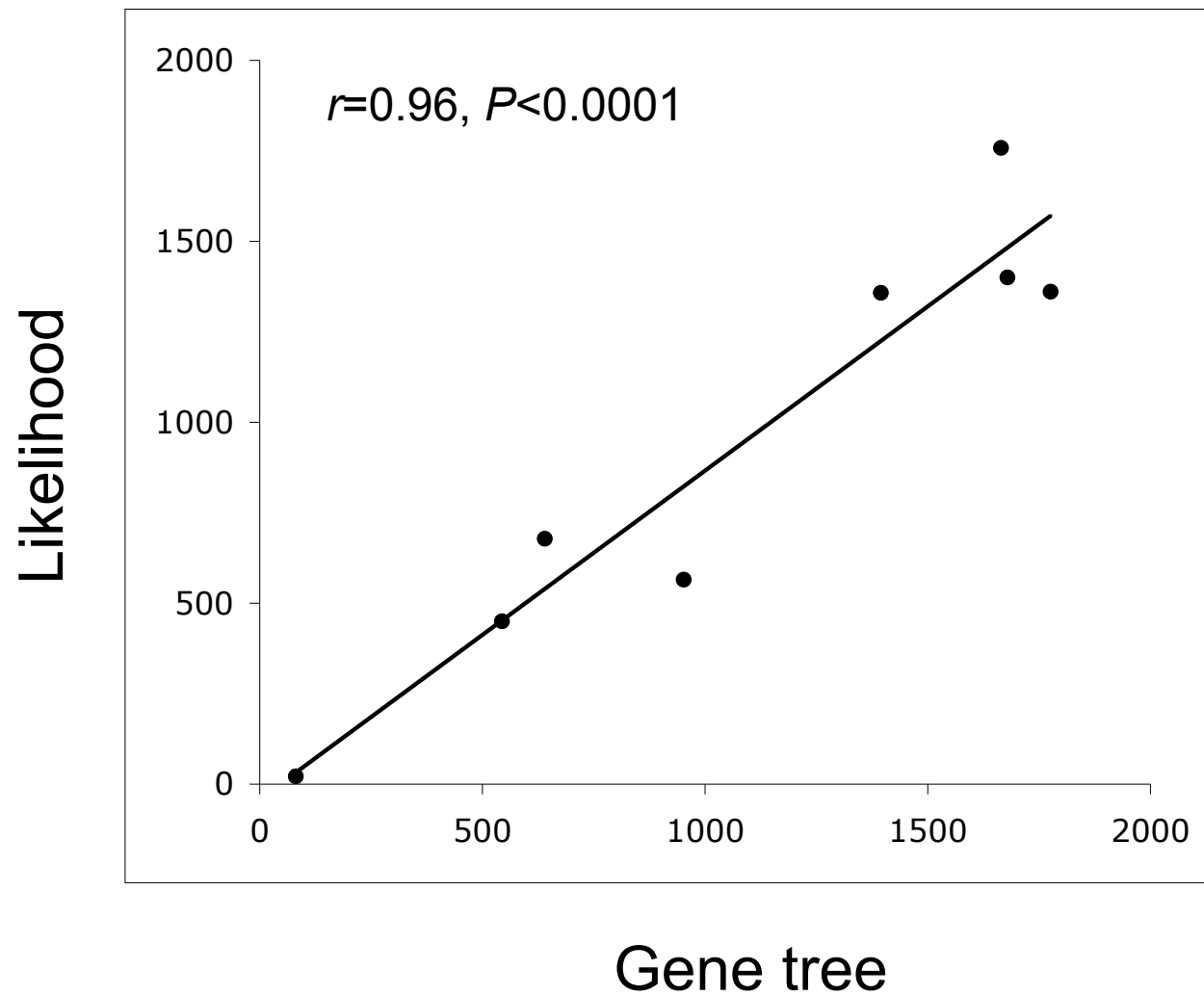
Only some branches show bias



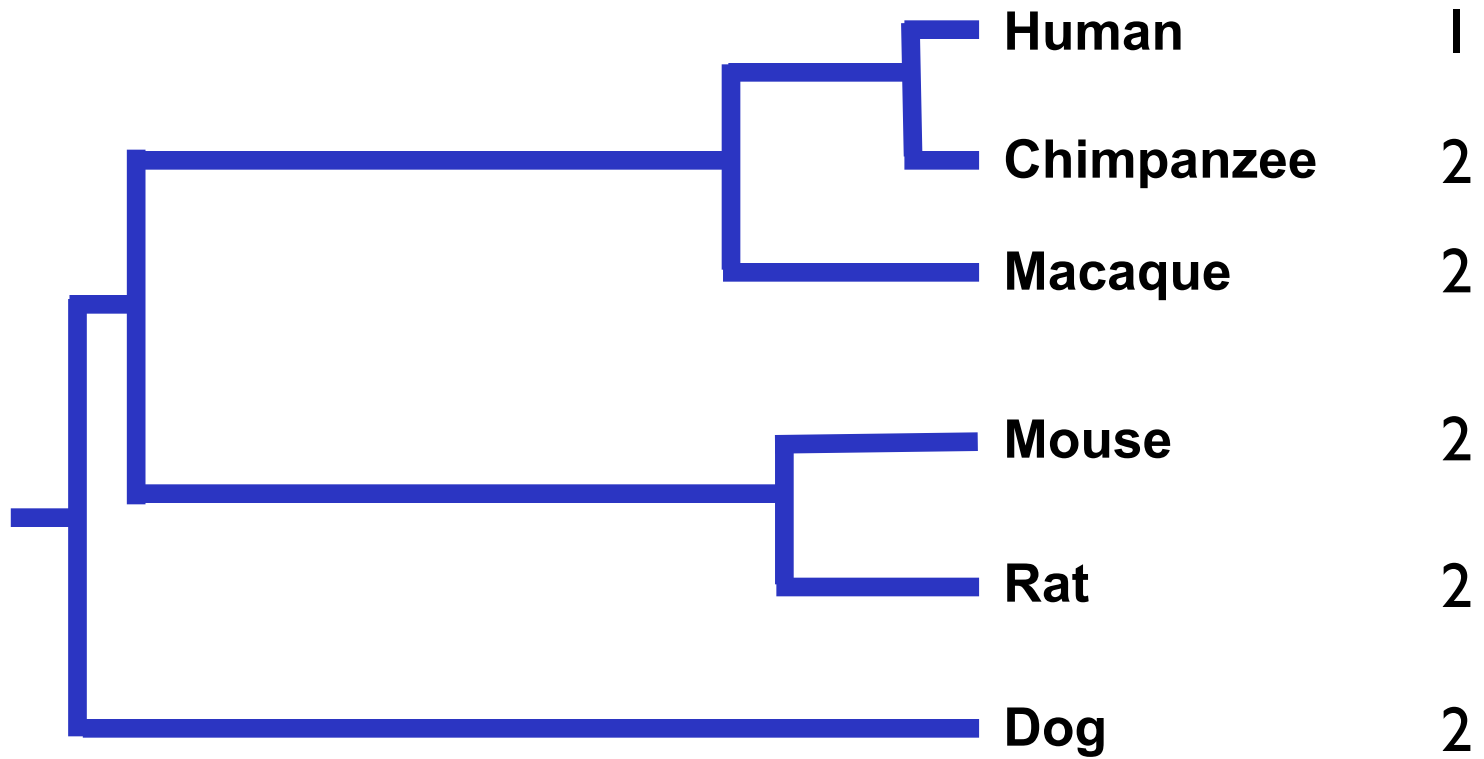
● duplication
| loss

█ gain
□ loss

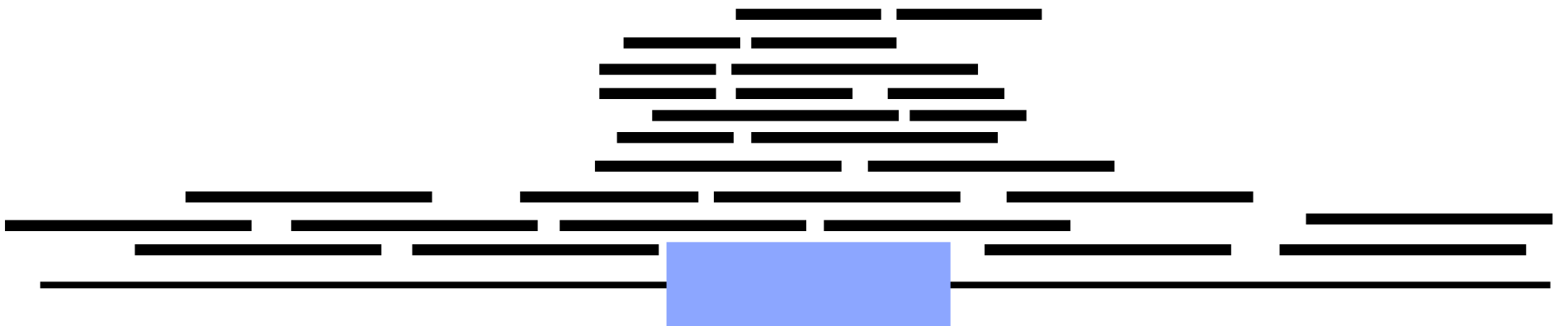
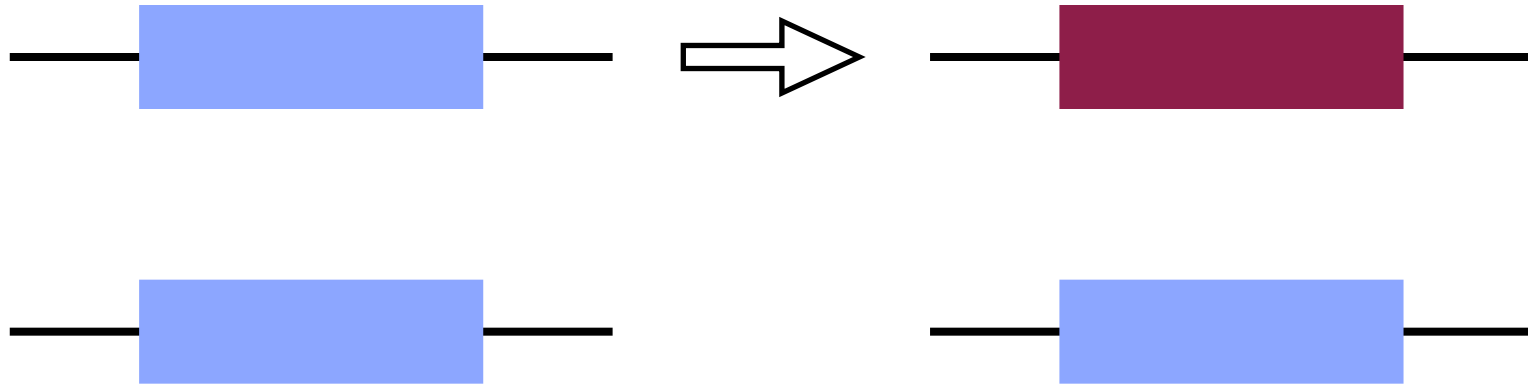
Likelihood vs. Reconciliation



Loss of genes



Loss of genes



Differences between human and chimp

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



The genomic revolving door

The King and Wilson paradox

Humans and chimps are ~1.5% different at the nucleotide level

But the number of genic differences is much larger than equally distant pairs of non-primates

The King and Wilson paradox

Humans and chimps are ~1.5% different at the nucleotide level

AACGCATCGATCGATCAGCTACGACG-----
-----TCGATCACTACGACGAACGCATCGA

Conclusions

11 April 1975, Volume 188, Number 4184

SCIENCE

Evolution at Two Levels in Humans and Chimpanzees

Their macromolecules are so alike that regulatory mutations may account for their biological differences.

Mary-Claire King and A. C. Wilson

evidence concerning the molecular basis of evolution at the organismal level. We suggest that evolutionary changes in anatomy and way of life are more often based on changes in the mechanisms controlling the expression of genes than on sequence changes in proteins. We therefore propose that regulatory mutations account for the major biological differences between humans and chimpanzees.

Similarity of Human and Chimpanzee Genes

To compare human and chimpanzee genes, one compares either homologous

“Evolution at multiple levels”

- change the protein sequence
- change the way genes are expressed
- change the number of proteins

Thanks

Jeff Demuth



Sang-Gook Han



Mira Han

