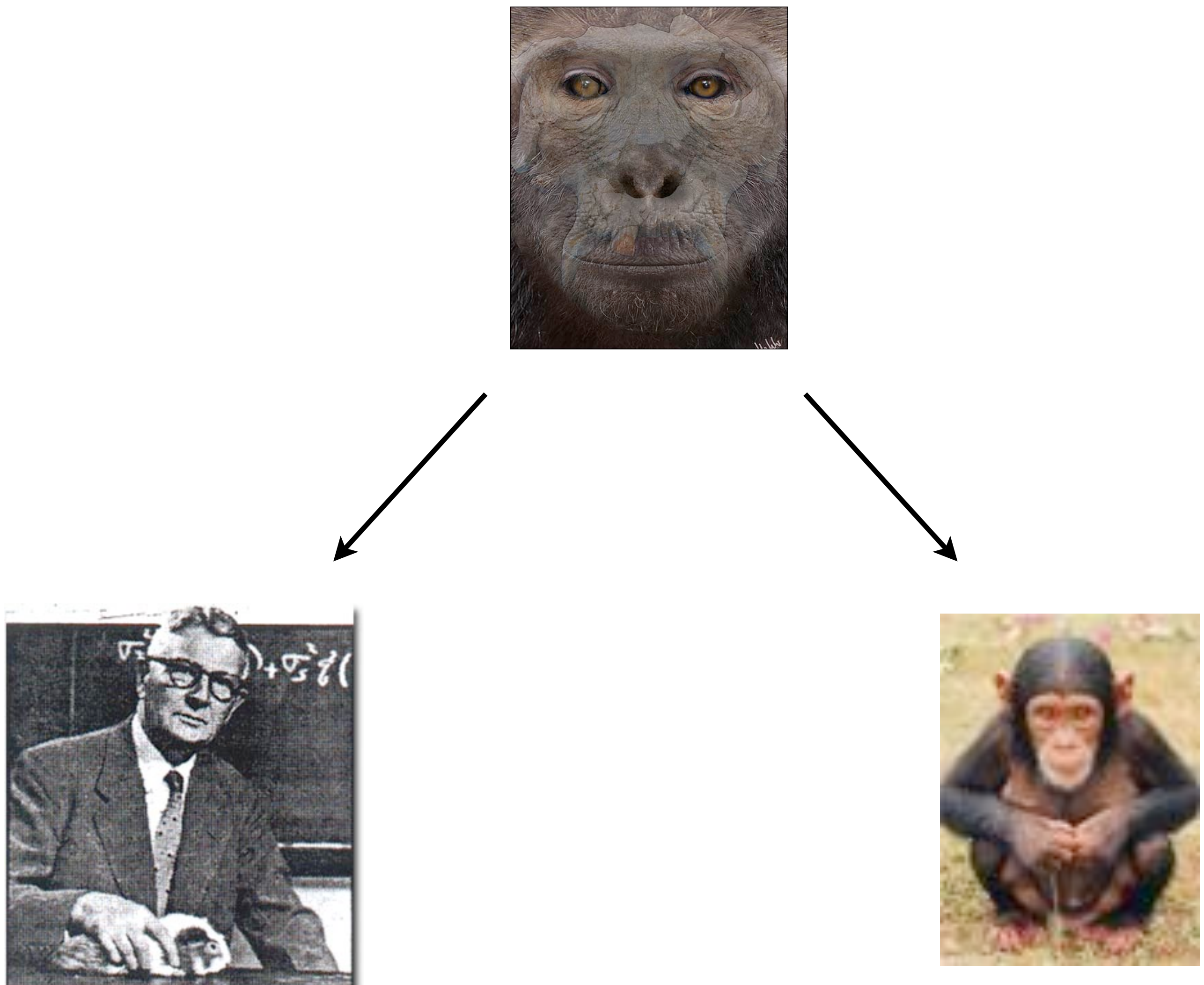
Gene gain and loss in mammals

Matthew Hahn

Department of Biology
& School of Informatics
Indiana University

The King and Wilson paradox

The King and Wilson paradox



The King and Wilson paradox

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

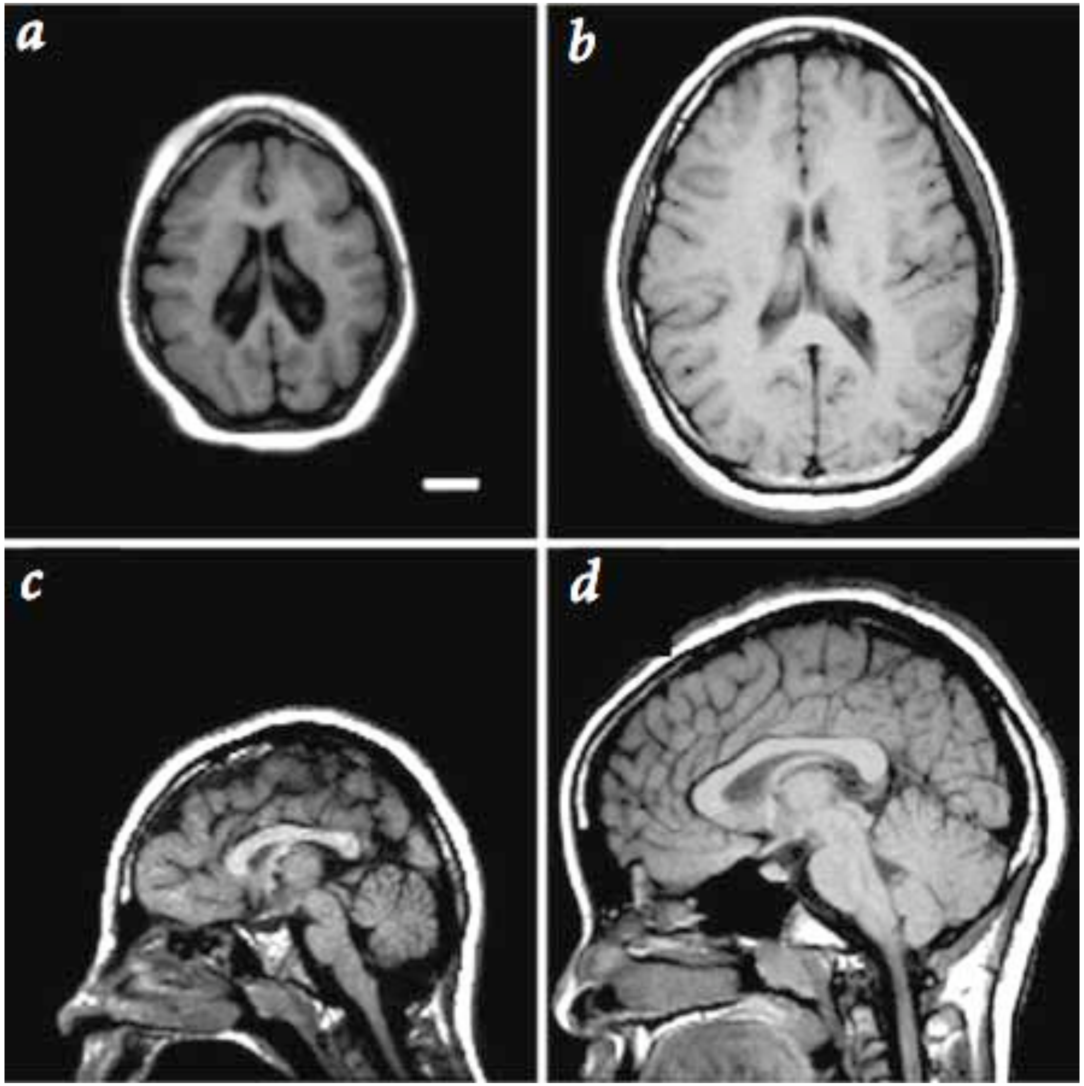
M.-C. King and A. Wilson 1975

Solutions to the paradox

Solutions to the paradox

- Coding (Classic)

The *ASPM* protein evolves rapidly and controls brain size



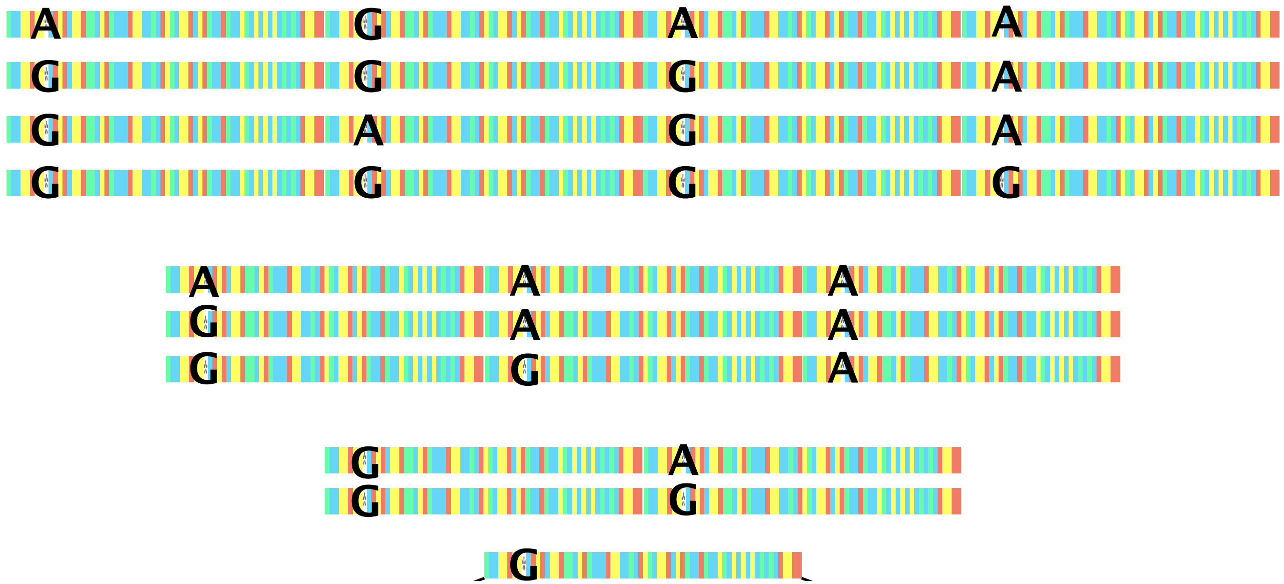
from Bell et al. (2004)

Solutions to the paradox

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

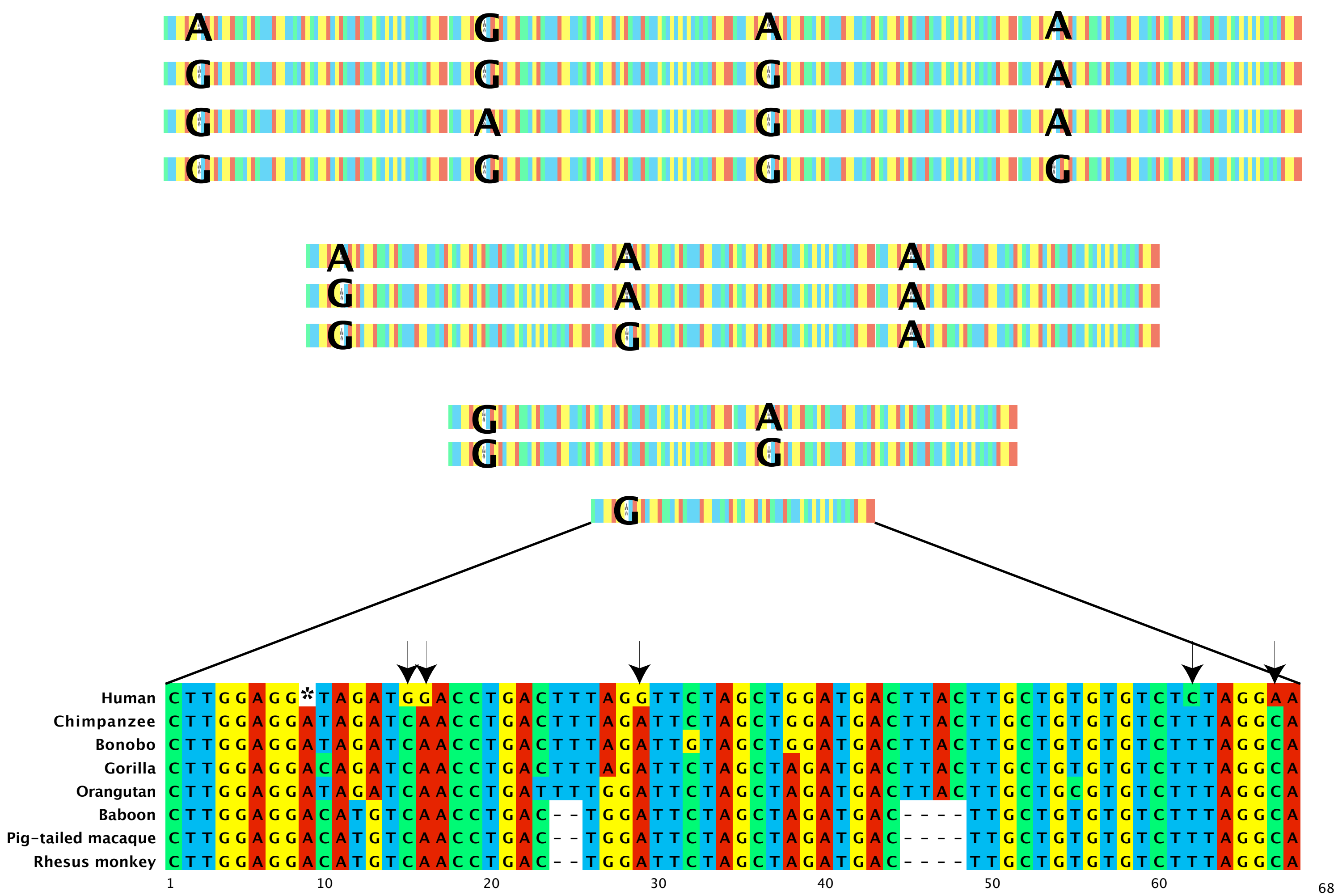
Prodynorphin in humans

Prodynorphin (*PDYN*) controls the expression of endorphins.



more repeats, more endorphins

Prodynorphin evolves rapidly in humans

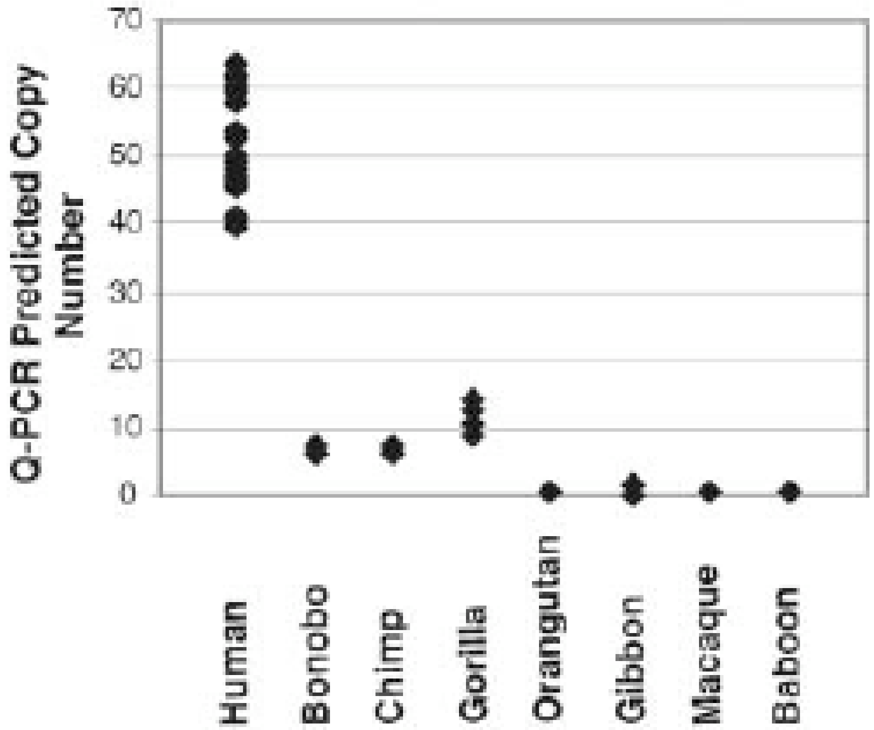


Rockman et al. (2005) *PLoS Biology*

Solutions to the paradox

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

DUF1220 is highly duplicated in humans

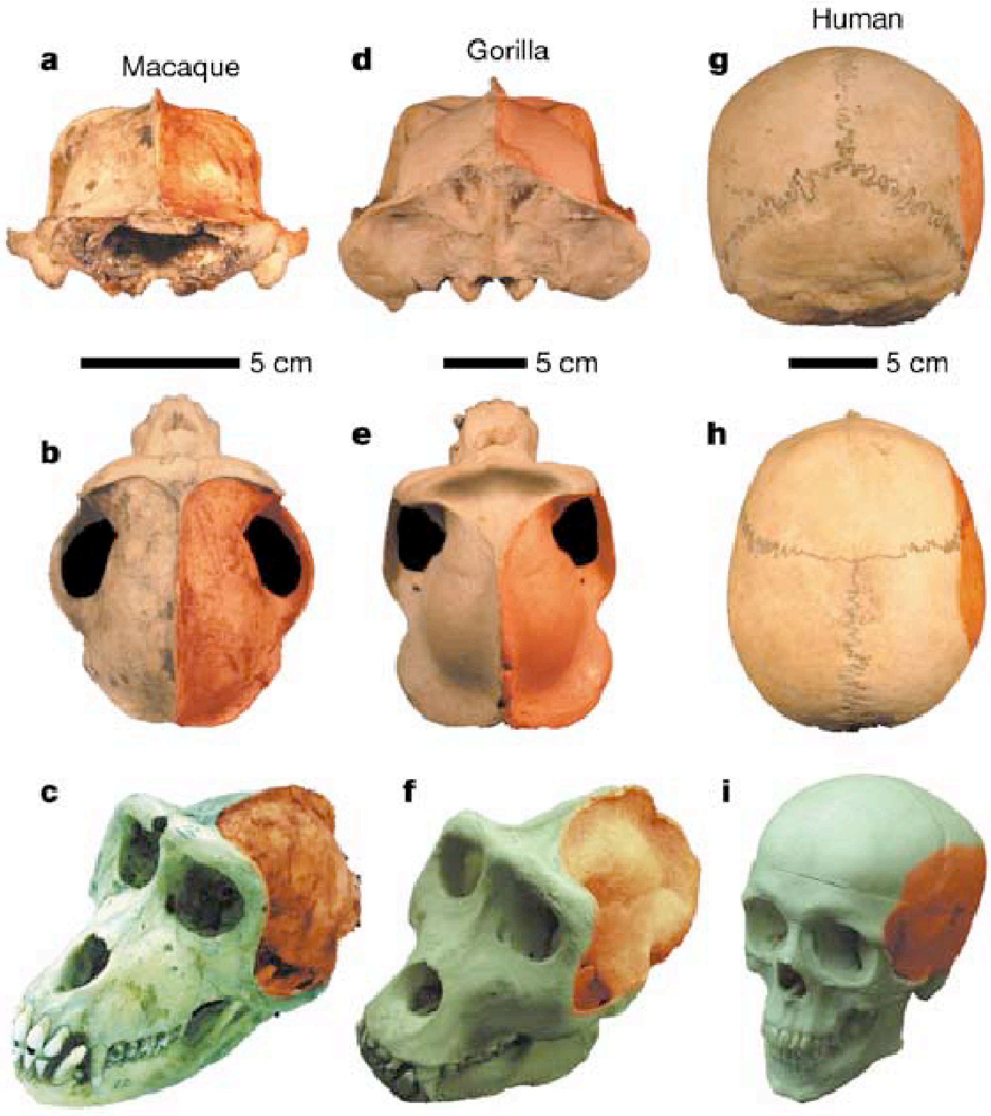


from Popesco et al. (2006)

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



from Stedman et al. (2004)

Solutions to the paradox

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

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 gain and loss
-Infer the action of natural selection

Outline

I. Statistical and computational methods

II. Quantifying gene gain and loss

III. Natural selection on gene duplicates

Preview of results

- Primates gain and lose genes at a rate twice as high as other mammals
- At least 1,415 genes (6% of all genes) are not shared between humans and chimps
- Newly duplicated genes are undergoing adaptive evolution at a high rate

Outline

I. Statistical and computational methods

II. Quantifying gene gain and loss

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.

	S. cerevisiae	*C. elegans*	*D. melanogaster*	*H. sapiens*	*A. thaliana*
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 gain and loss

Hahn et al. (2005) *Genome Research*

A model for gene gain and loss

Homogeneous birth and death process

Hahn et al. (2005) *Genome Research*

A model for gene gain and loss

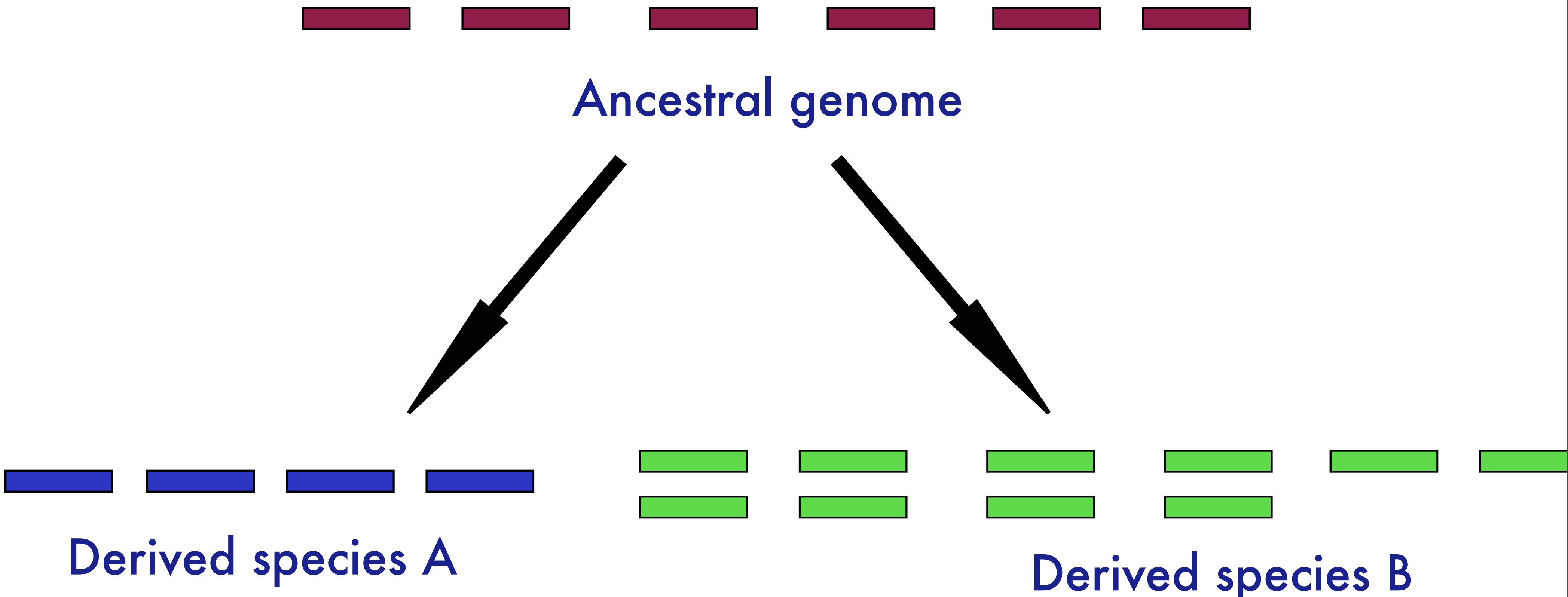
Homogeneous birth and death process

Birth = duplication

Death = deletion or pseudogenization

Hahn et al. (2005) *Genome Research*

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

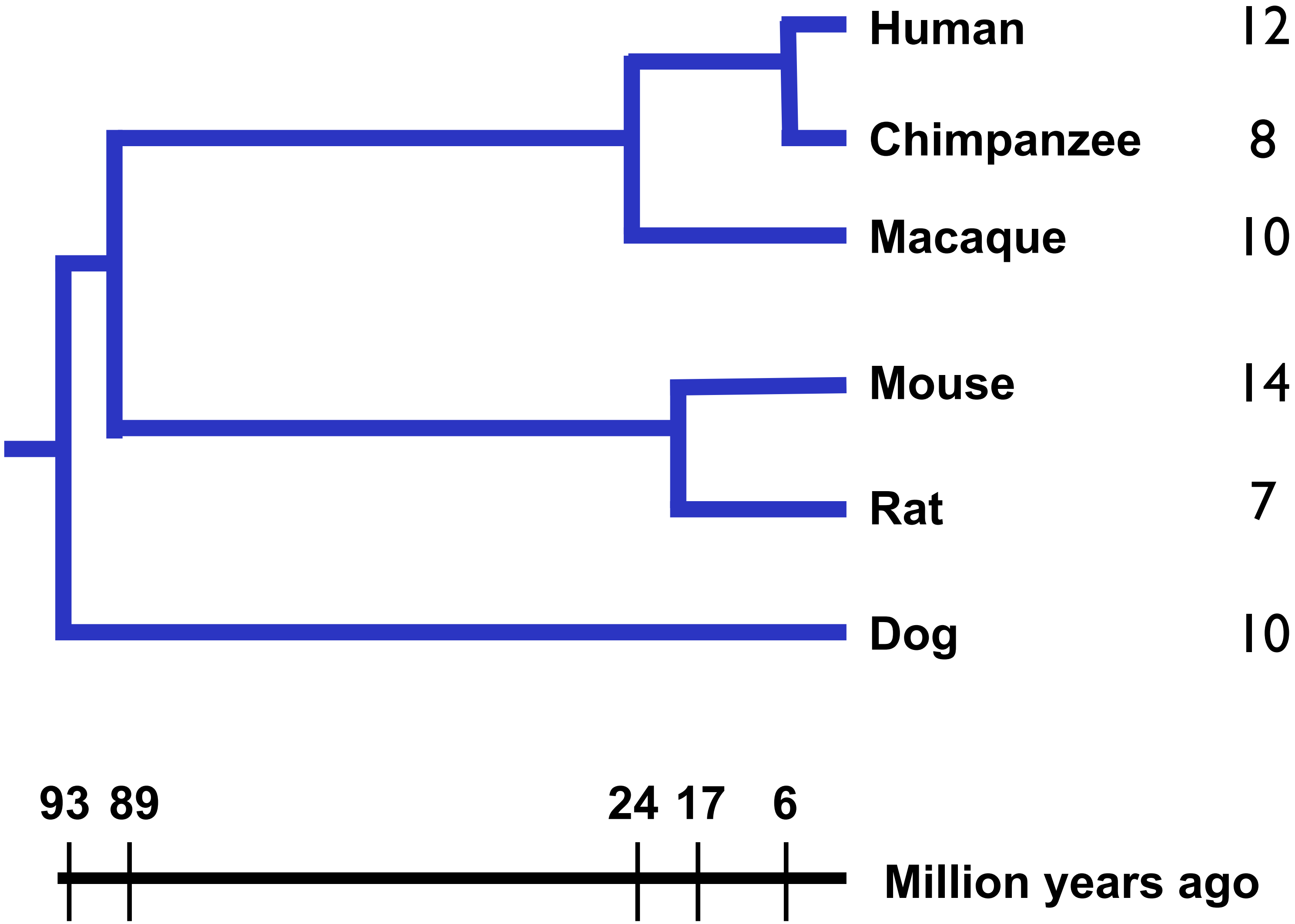
-Time since divergence

-Ancestral family size

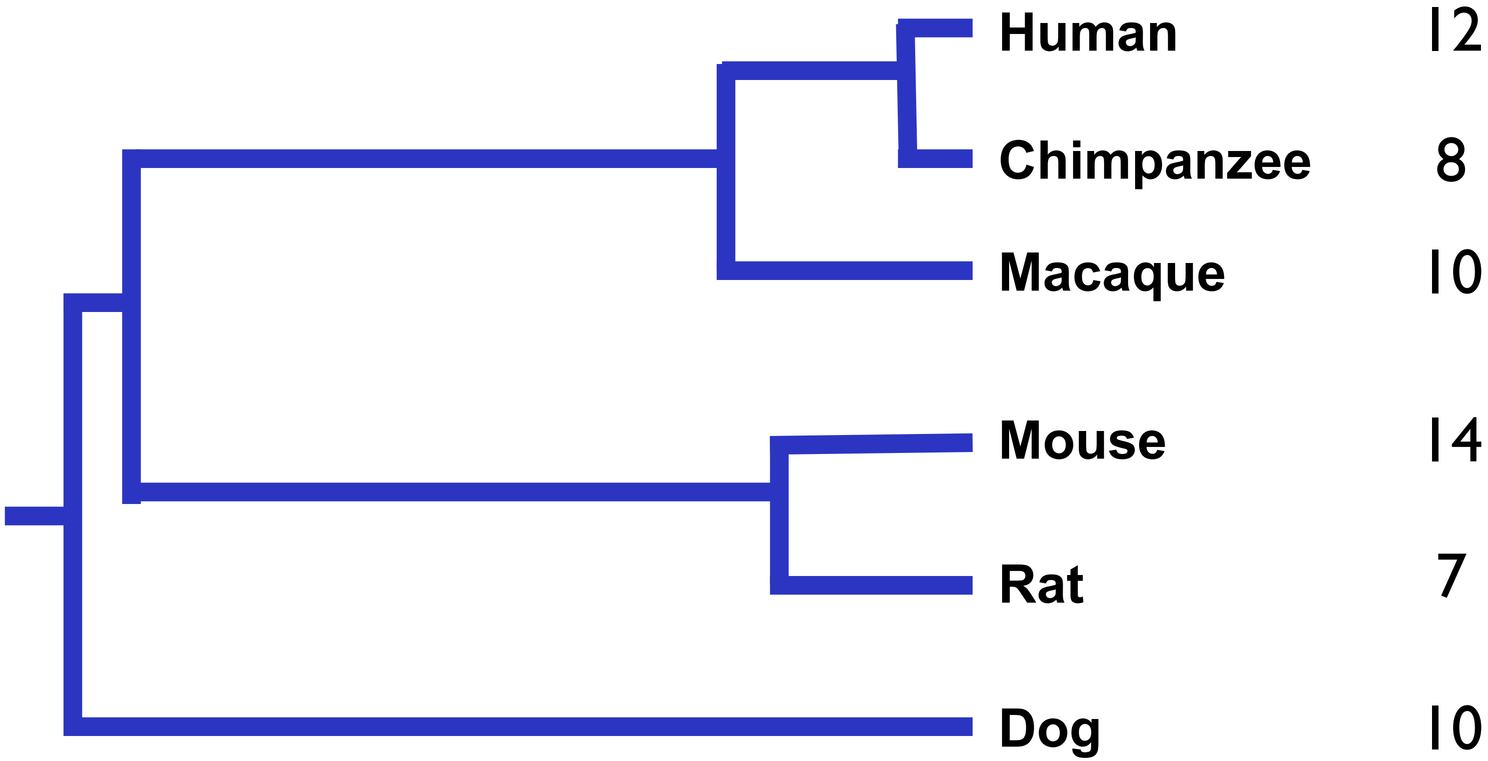
-Gain and loss rates

Hahn et al. (2005) *Genome Research*

Probabilistic graphical models

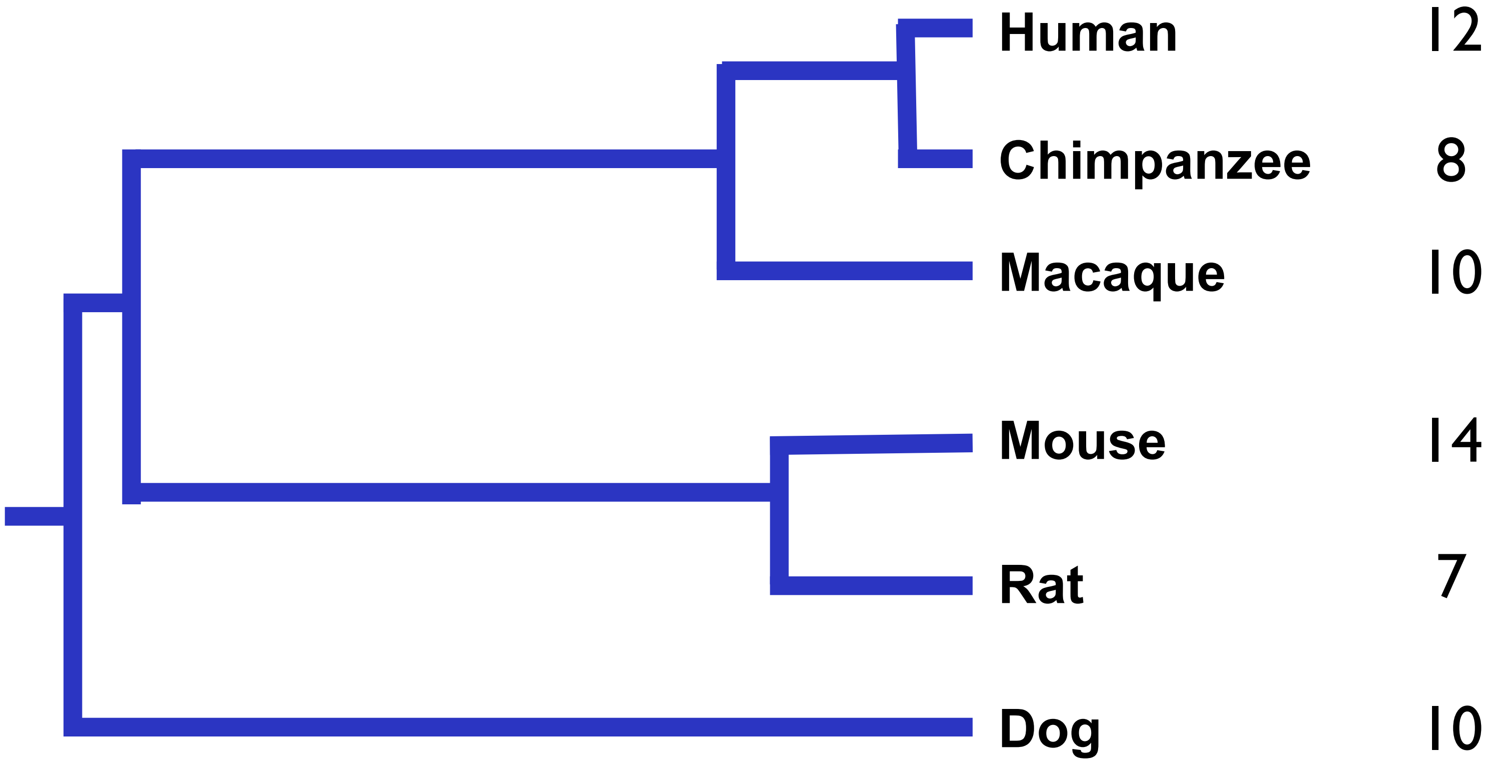


Inferring rates of gain and loss

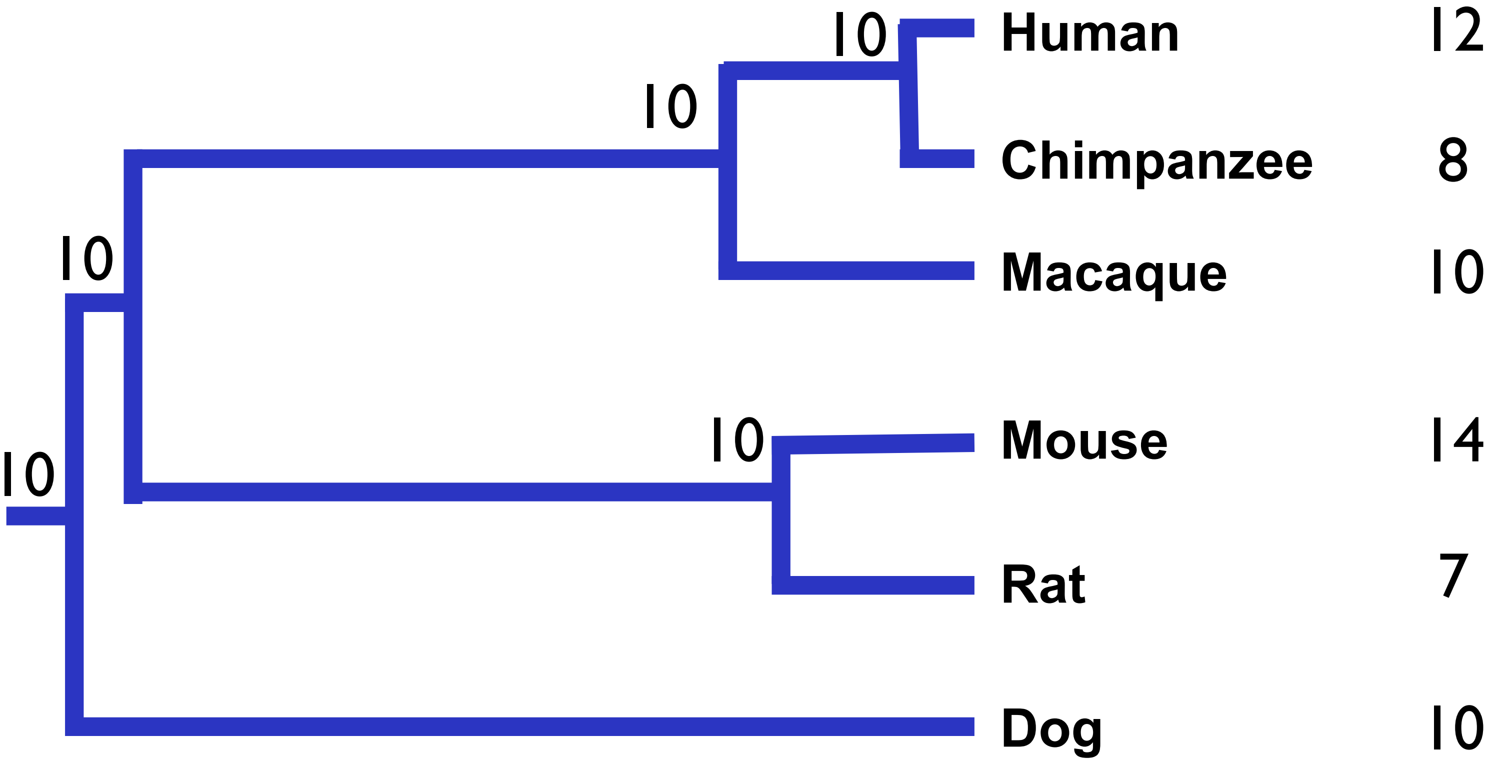


We estimate the average rate across all families

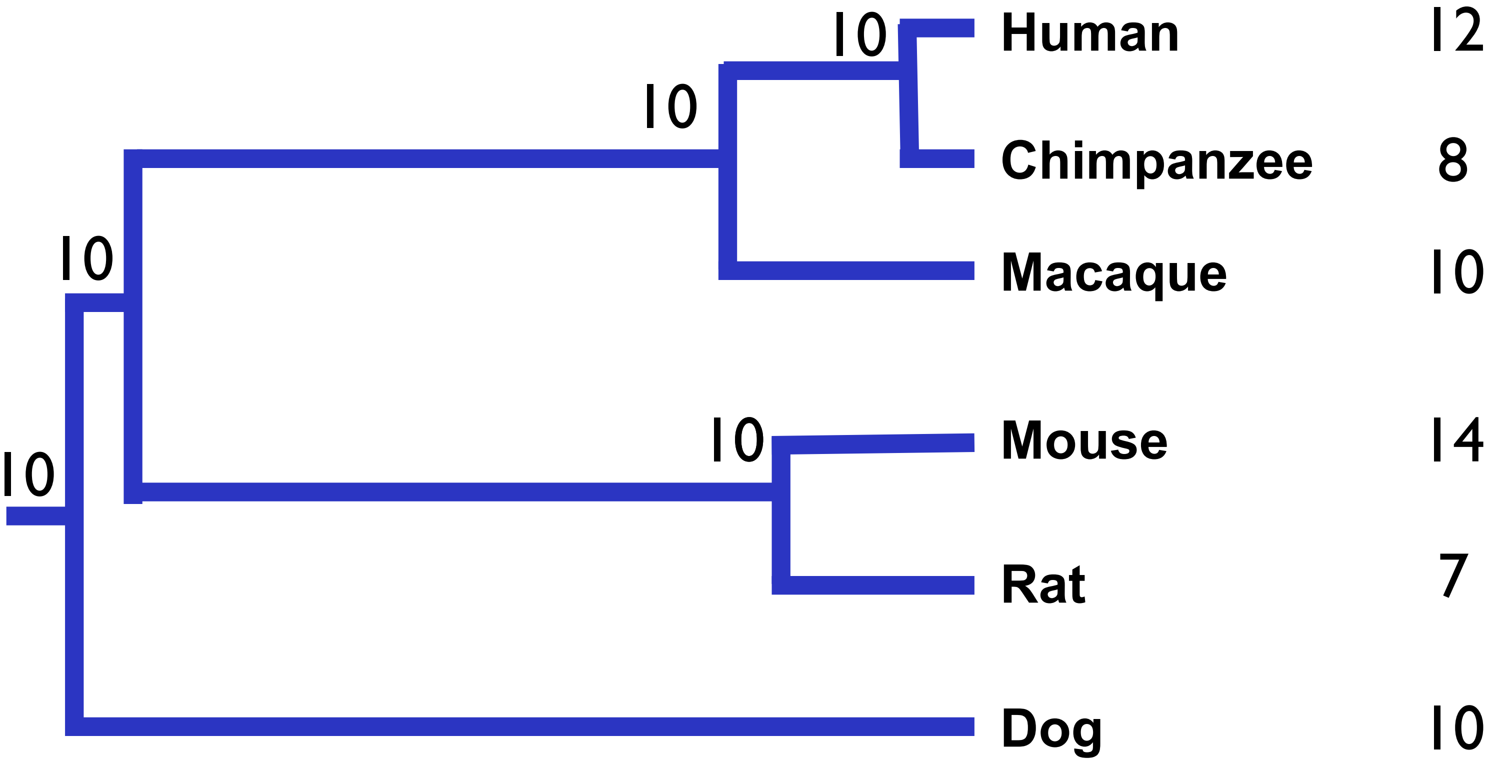
Inferring ancestral states



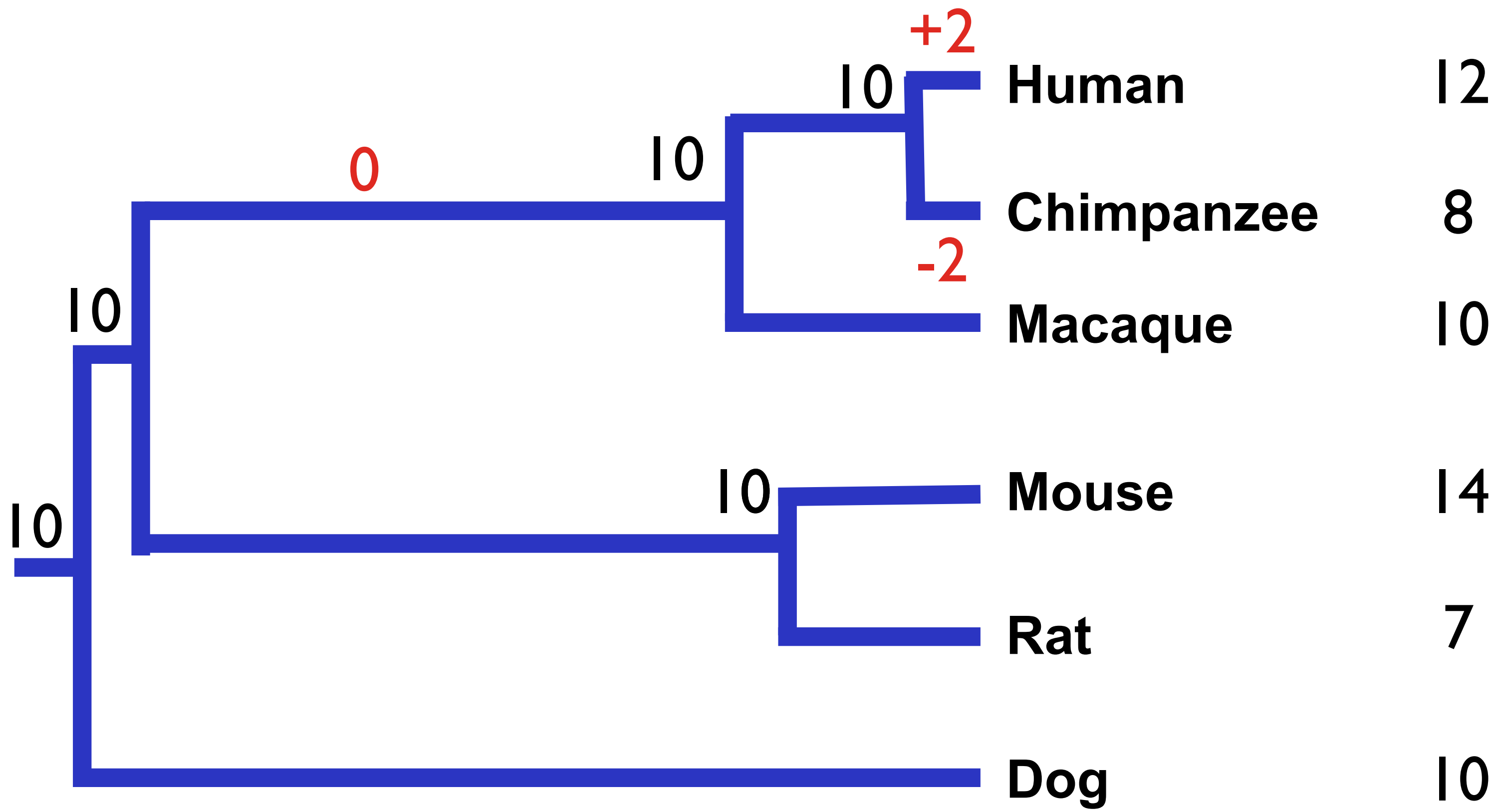
Inferring ancestral states



Inferring gains and losses

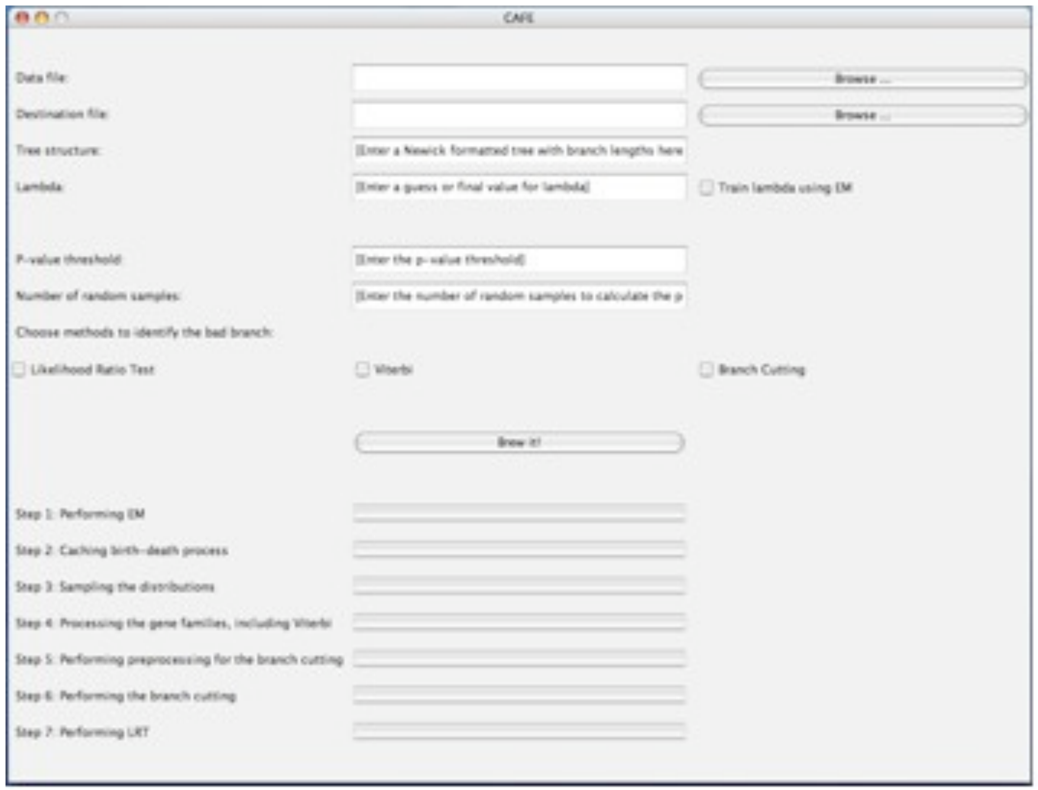


Inferring gains and losses



CAFE

(Computational Analysis of gene Family Evolution)



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

De Bie et al. (2006) *Bioinformatics*

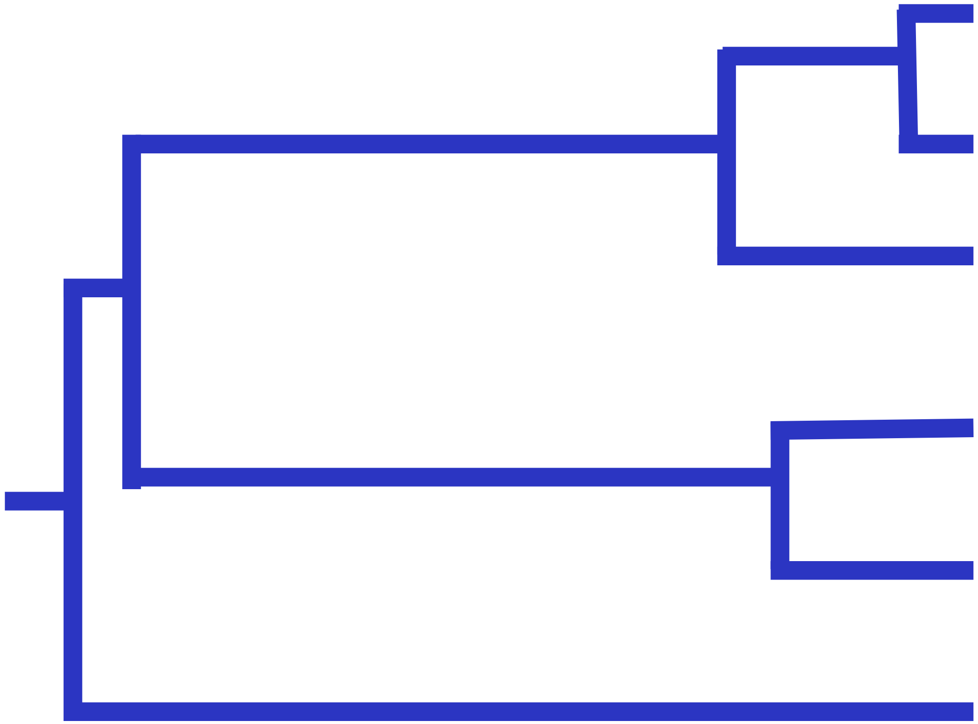
Outline

I. Statistical and computational methods

II. Quantifying gene gain and loss

III. Natural selection on gene duplicates

Genome size in mammals



	# genes
Human	22,209
Chimpanzee	21,625
Macaque	22,149
Mouse	25,613
Rat	21,952
Dog	18,201

(Data from Ensembl v41)

The rate of gene gain and loss

We estimate λ (the gain/loss rate) to be 0.0017 /gene/my across the whole tree

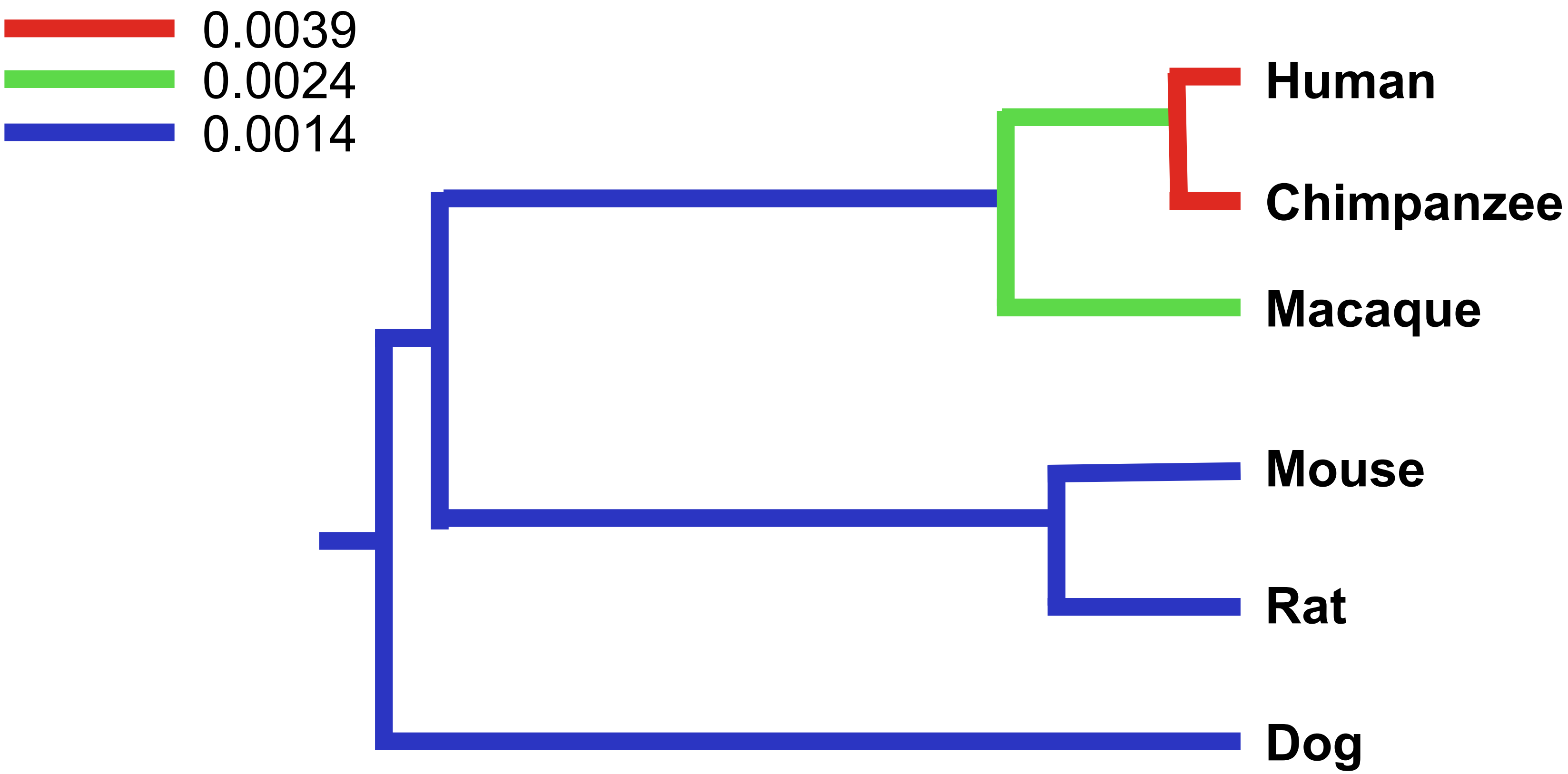
The rate of gene gain and loss

We estimate λ (the gain/loss rate) to be 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:

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

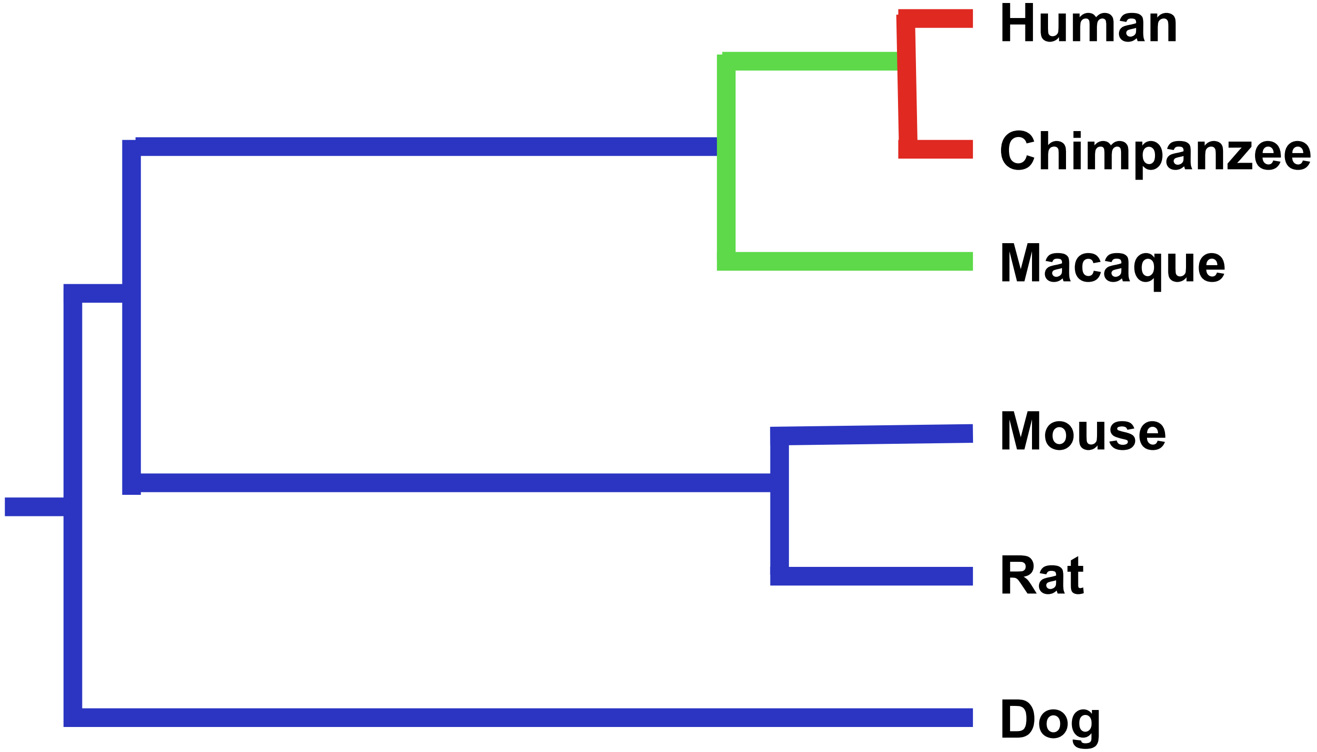
The rate of gene gain and loss



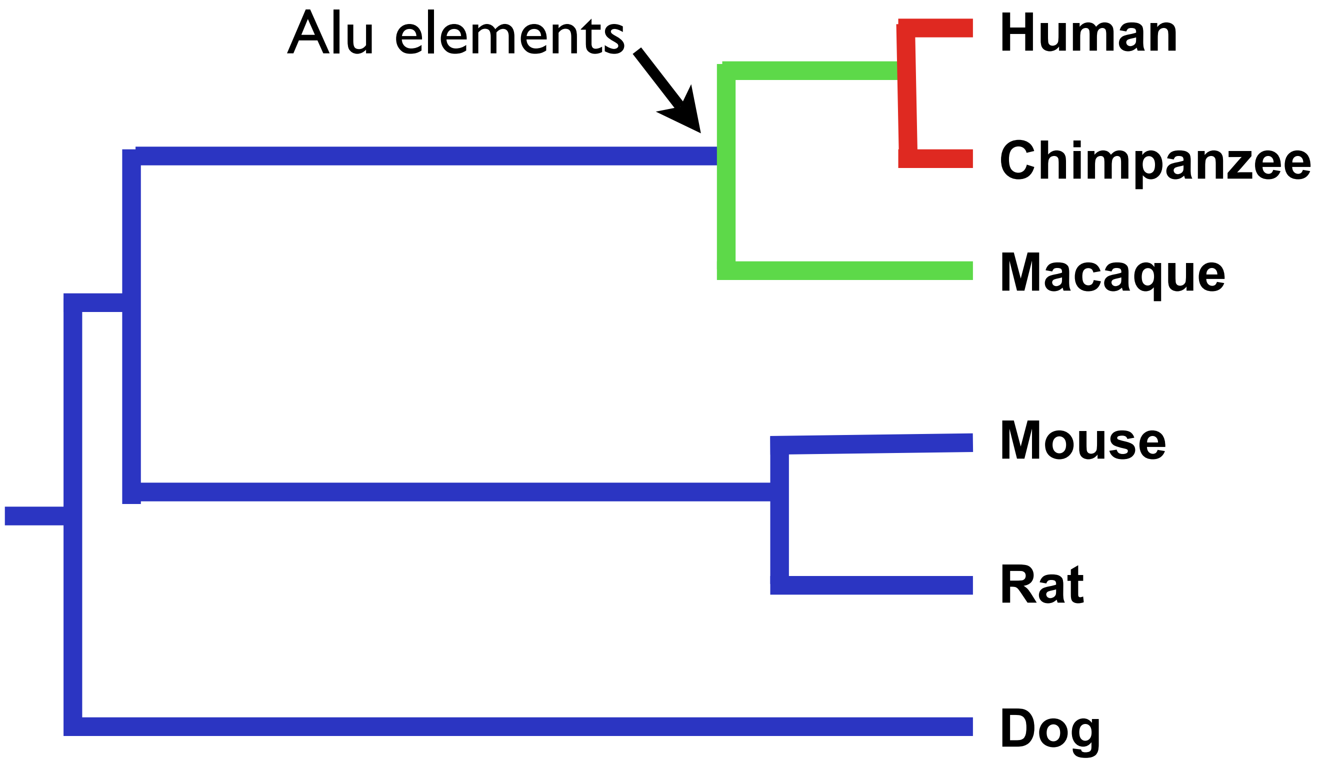
The rate of gain and loss in primates is 2-3 times higher than the rest of the mammals

Hahn et al. (2007) *Genetics*

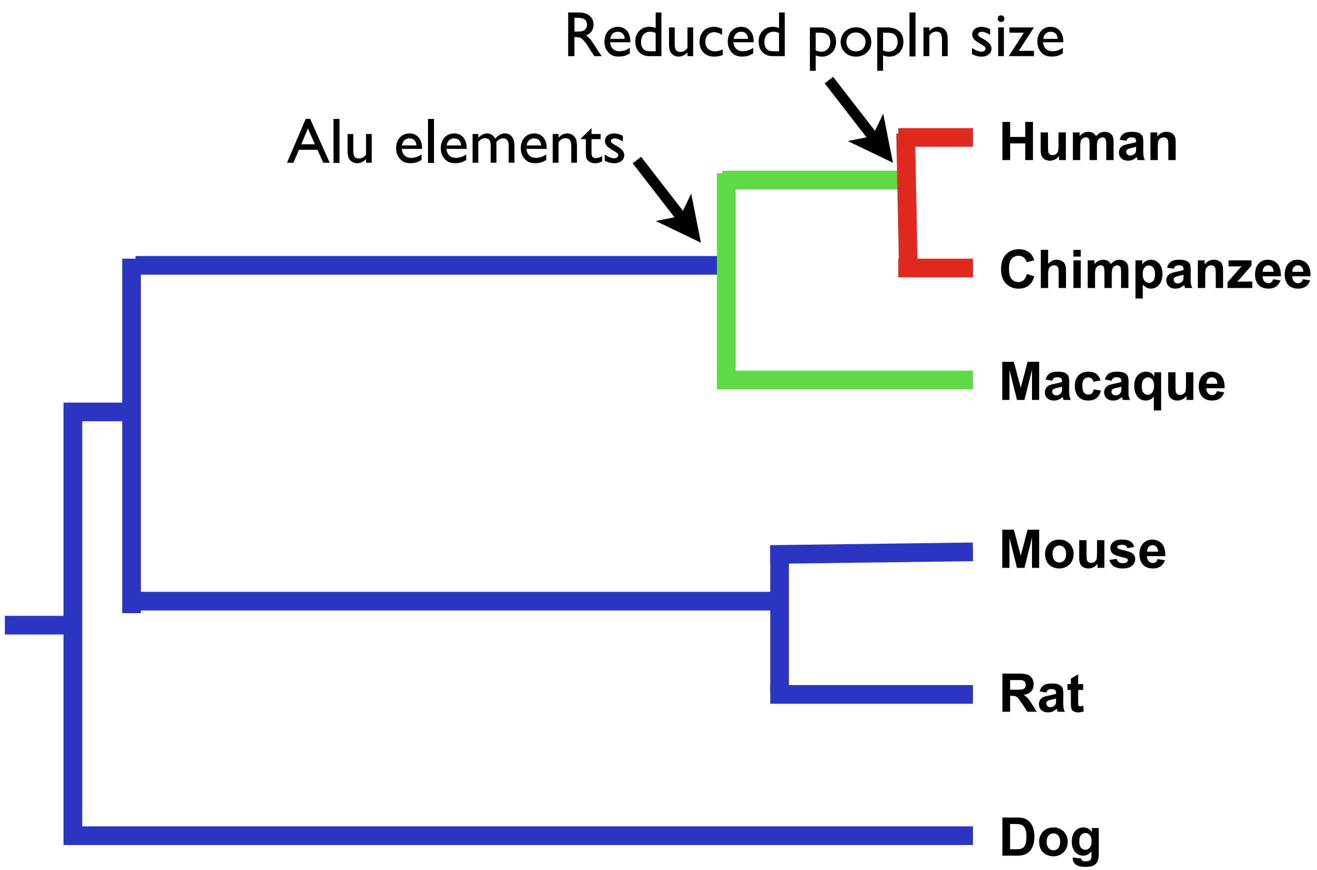
Accelerated rate of gene gain and loss in primates



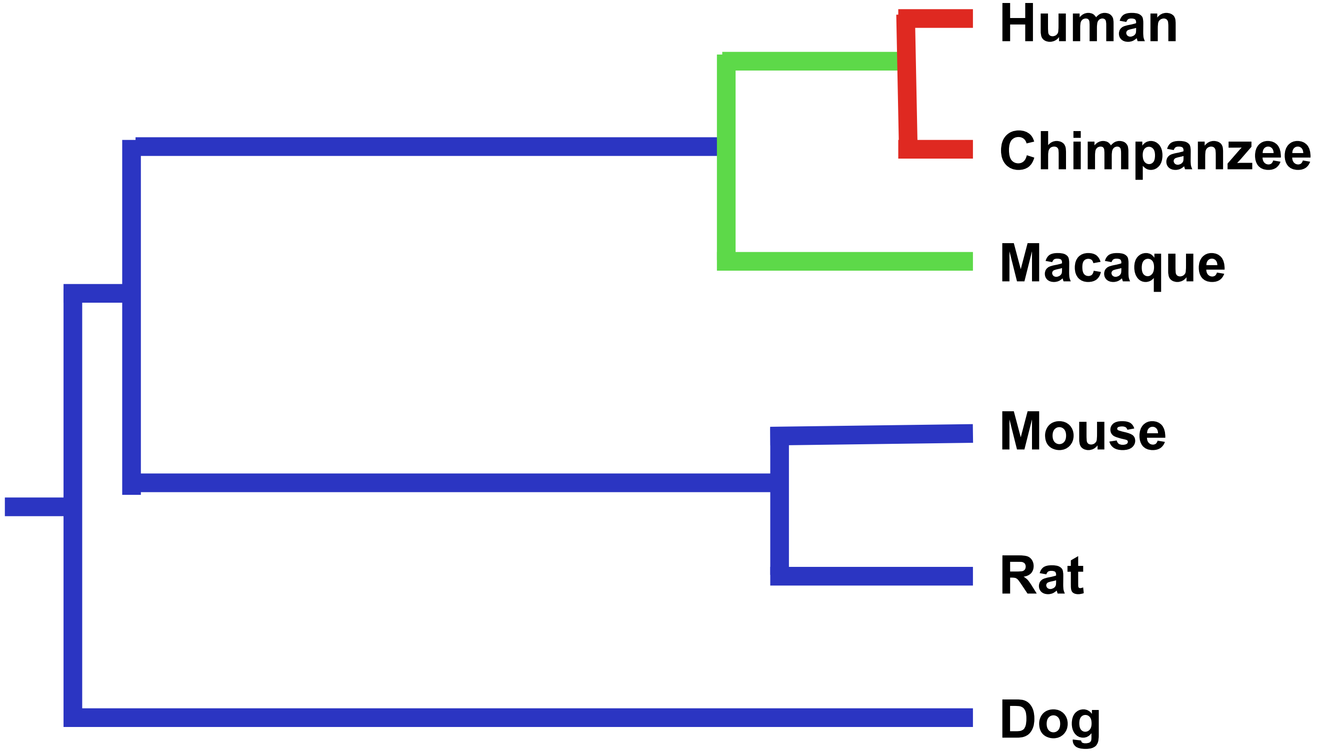
Accelerated rate of gene gain and loss in primates



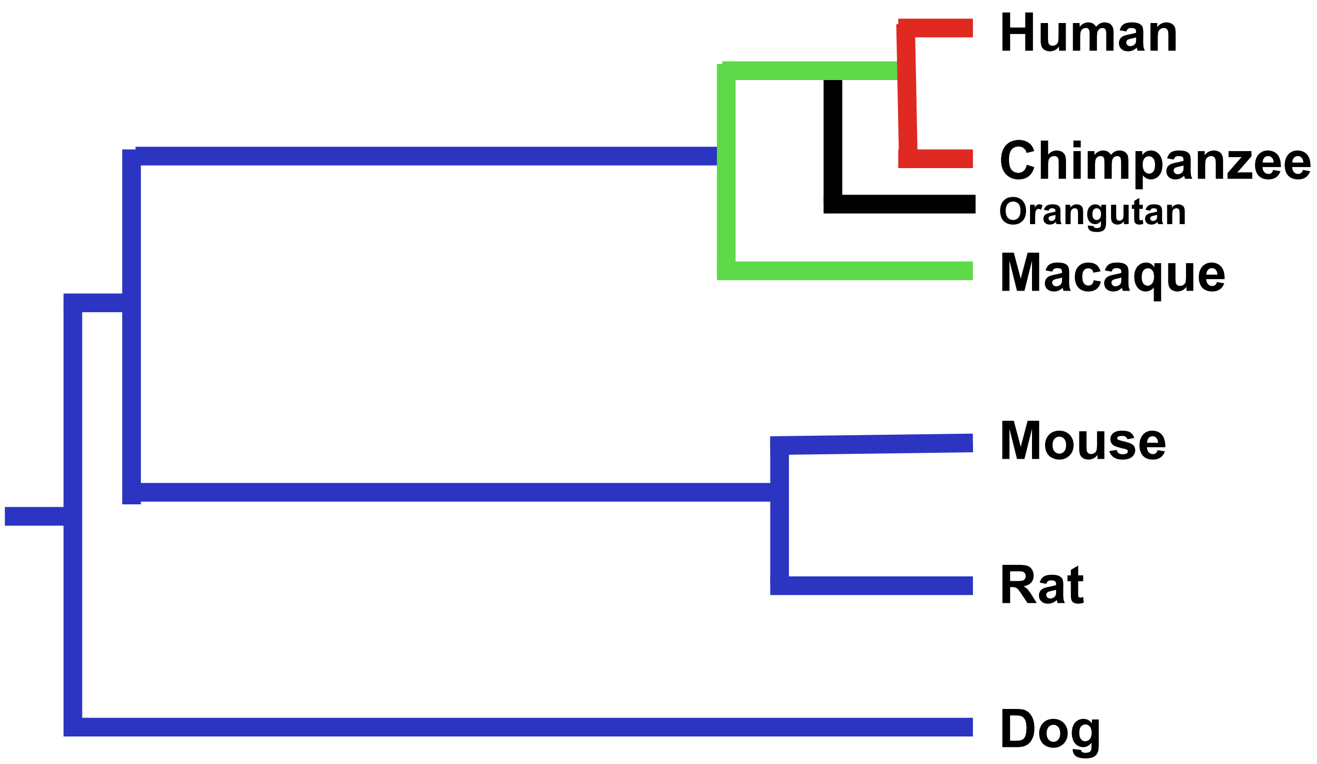
Accelerated rate of gene gain and loss in primates



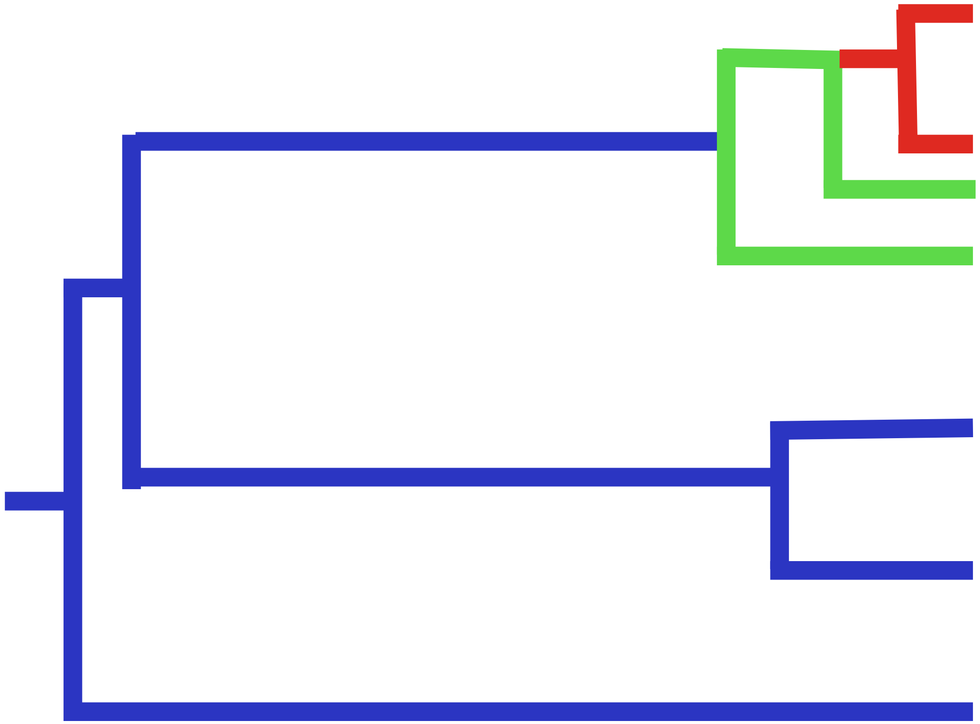
Accelerated rate of gene gain and loss in primates



Accelerated rate of gene gain and loss in primates

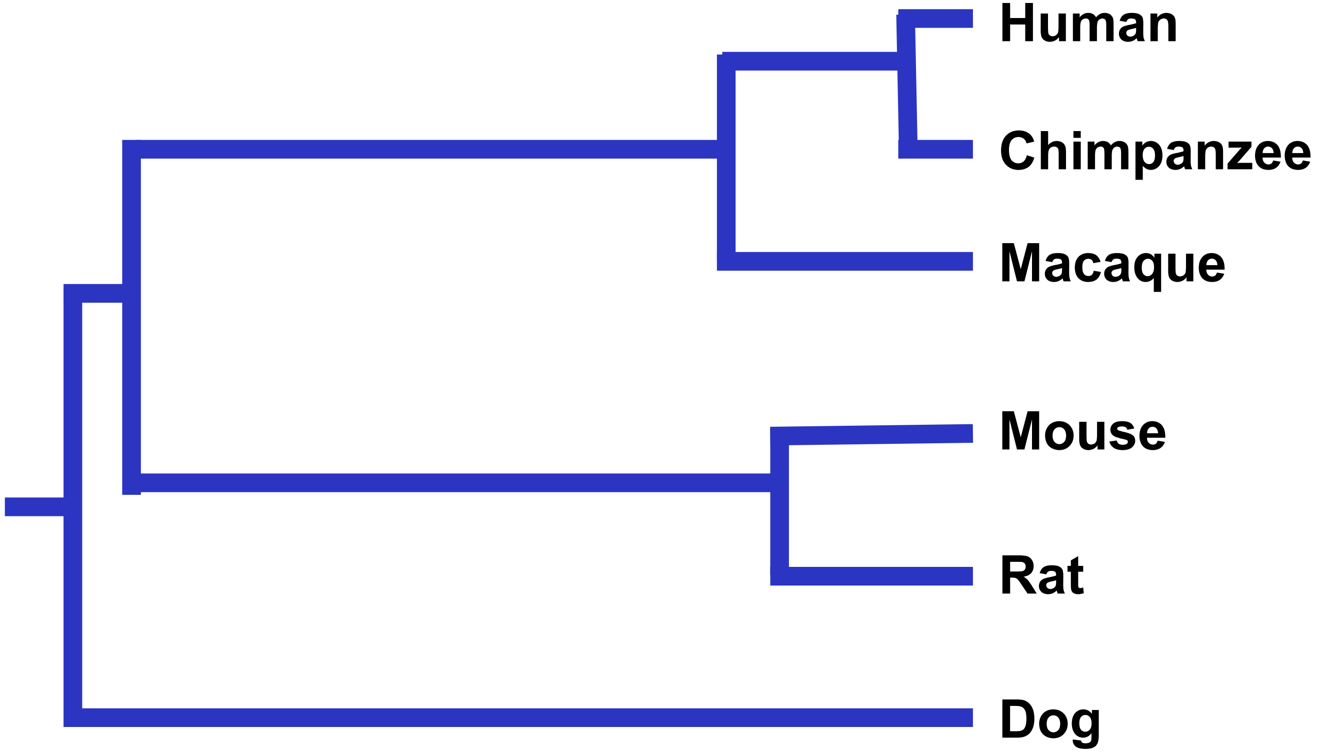


Accelerated rate of gene gain and loss in hominids



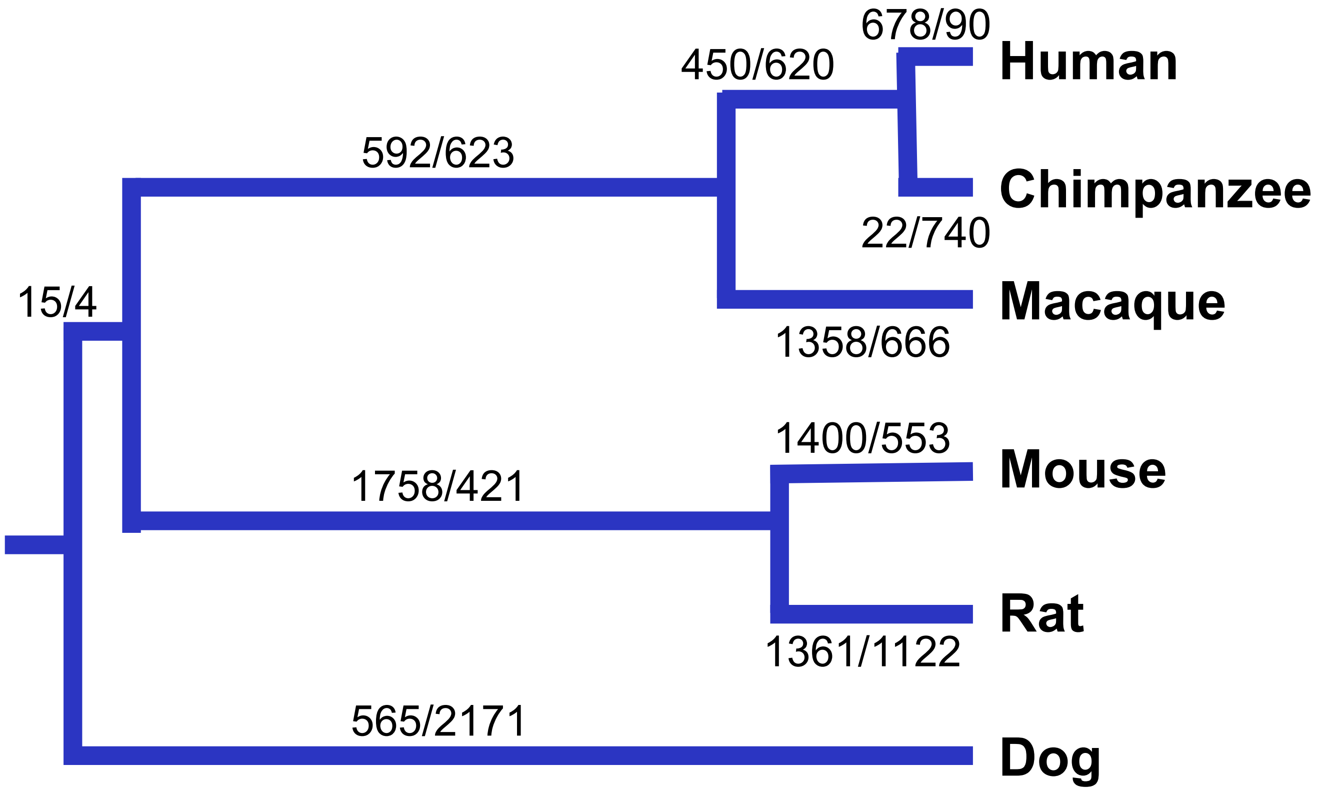
Locke et al. (*in prep.*)

Gene gain and loss in mammals



Demuth et al. (2006) *PLoS ONE*
Gibbs et al. (2007) *Science*

Gene gain and loss in mammals



Demuth et al. (2006) *PLoS ONE*
Gibbs et al. (2007) *Science*

Gene gain and loss in the great apes

In humans:

In chimpanzees:

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

- 740 genes have been lost

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

- 740 genes have been lost +

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

- 740 genes have been lost +

1415

Gene gain and loss in the great apes

In humans:

- 675 genes have been gained

In chimpanzees:

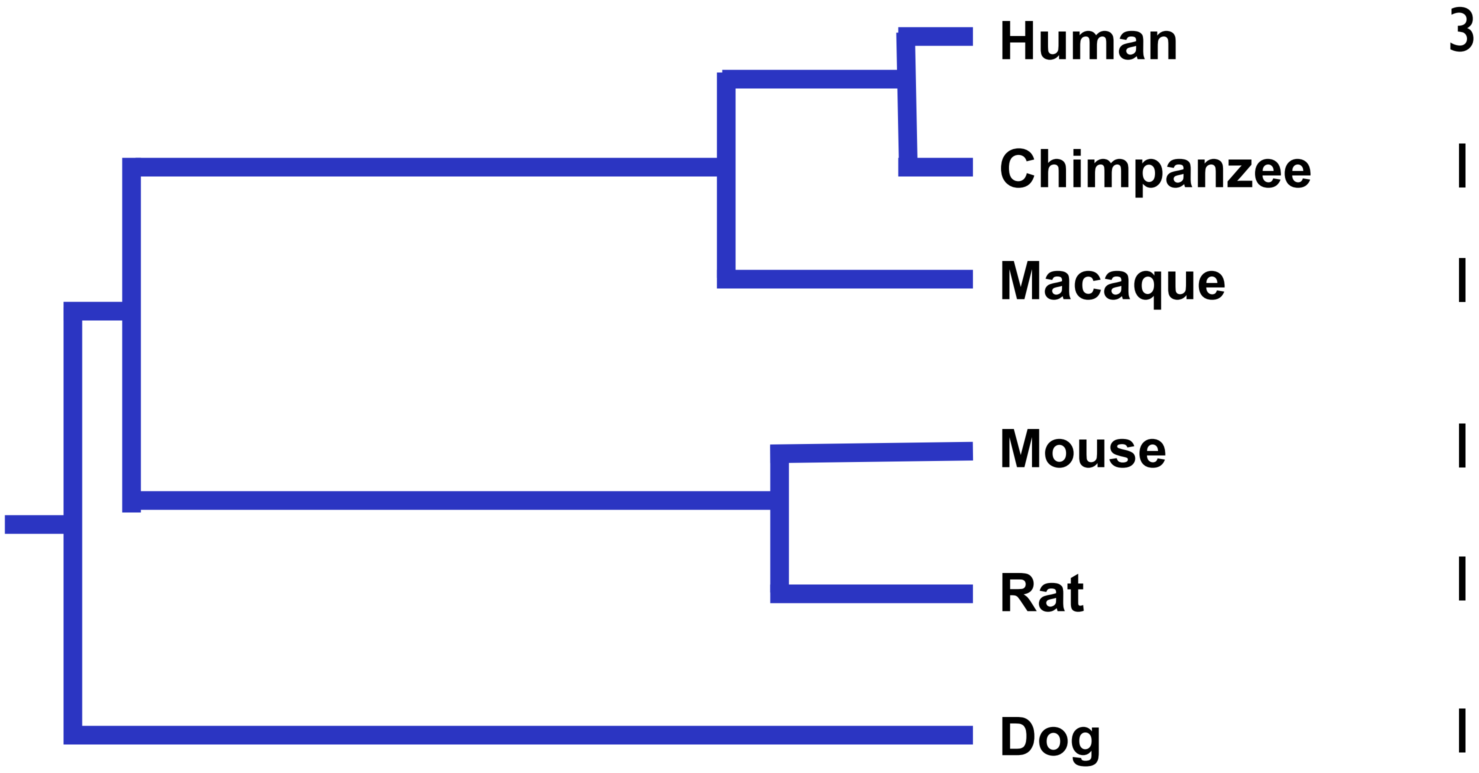
- 740 genes have been lost +

1415

1,415 genes not shared between humans and chimps!

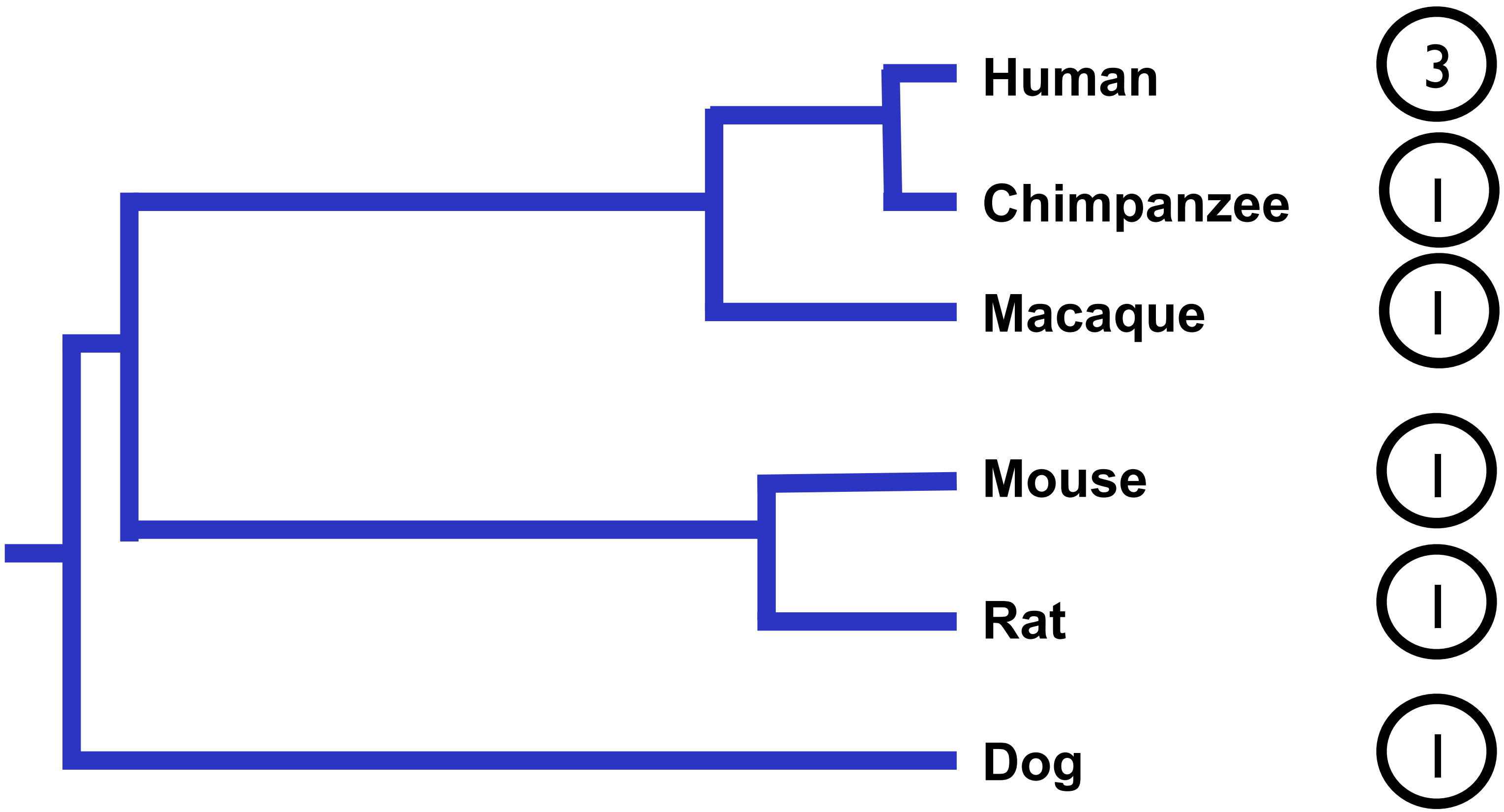
An alternative method for estimating gain and loss

An alternative method for estimating gain and loss



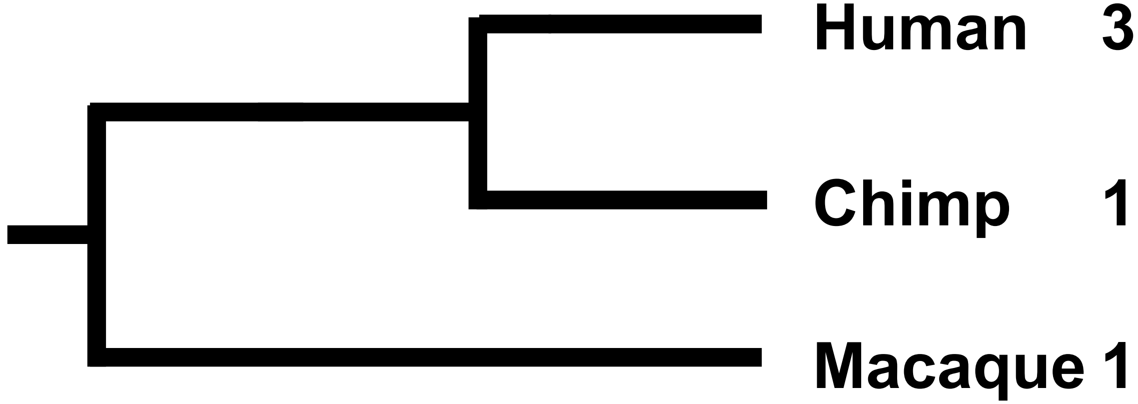
An alternative method for estimating gain and loss

"Genes in a bag"



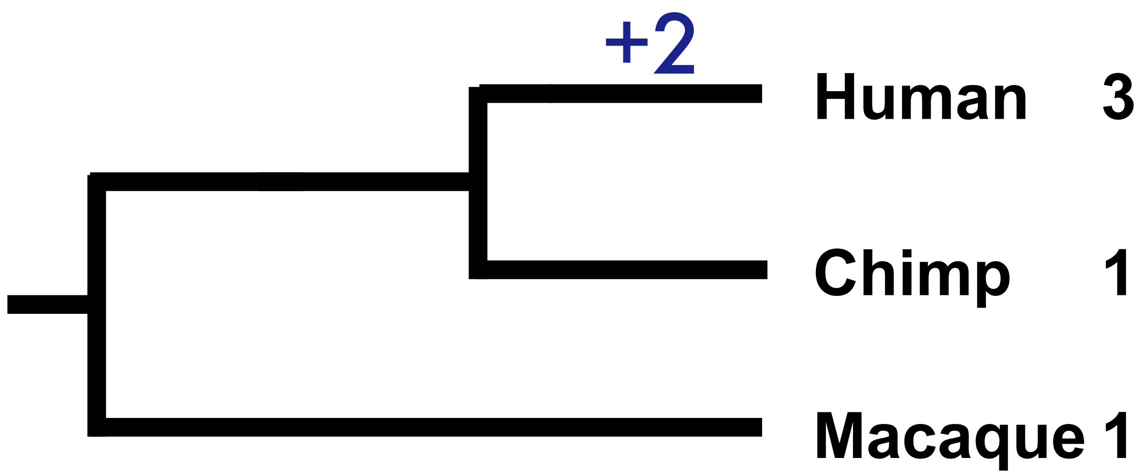
Species trees vs. gene trees

Species trees vs. gene trees



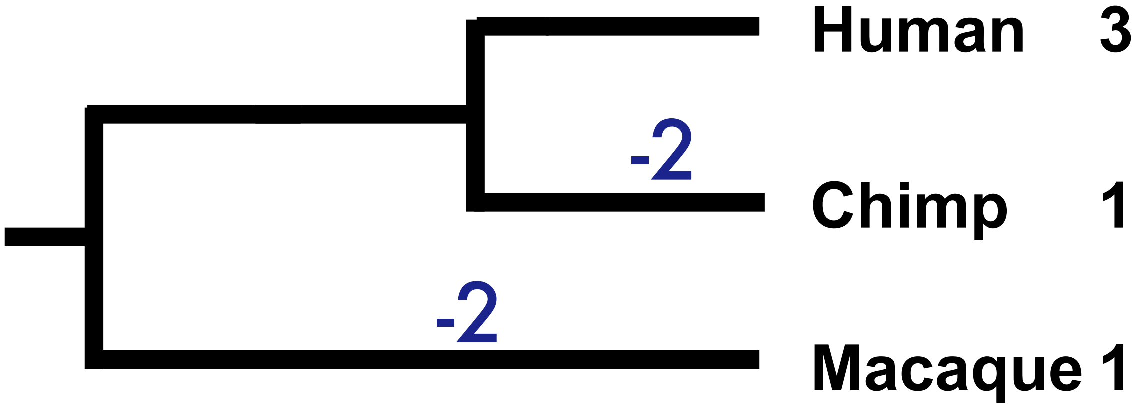
Species tree

Species trees vs. gene trees



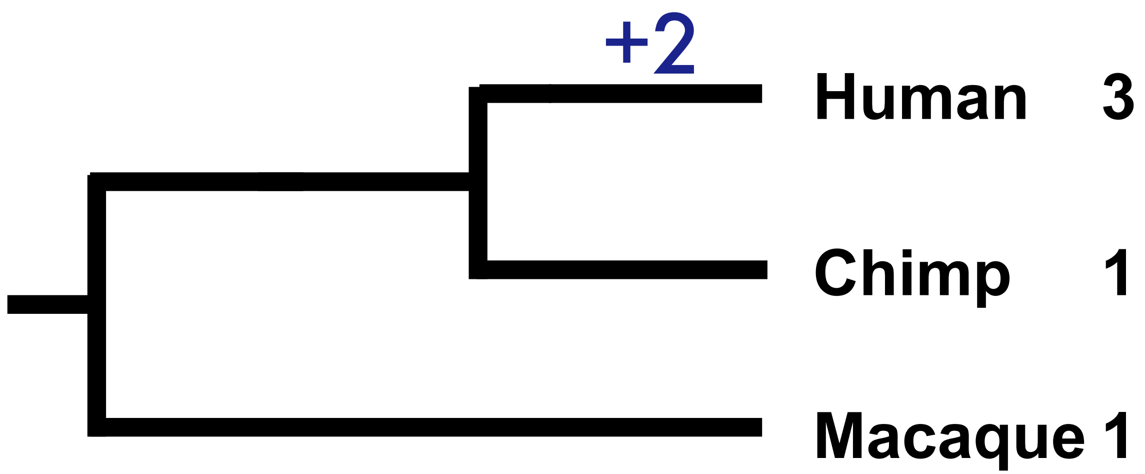
Species tree

Species trees vs. gene trees



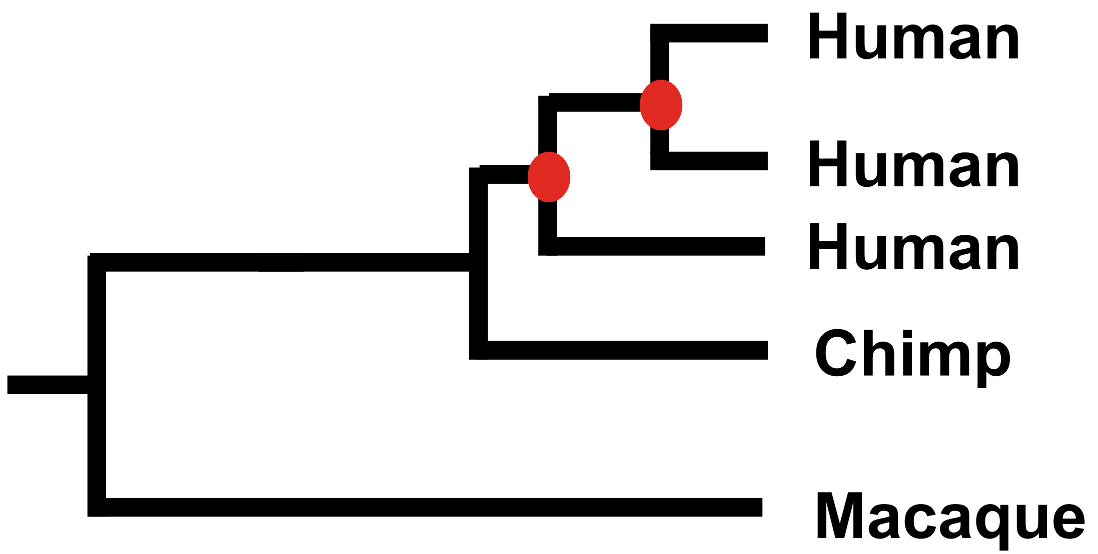
Species tree

Species trees vs. gene trees



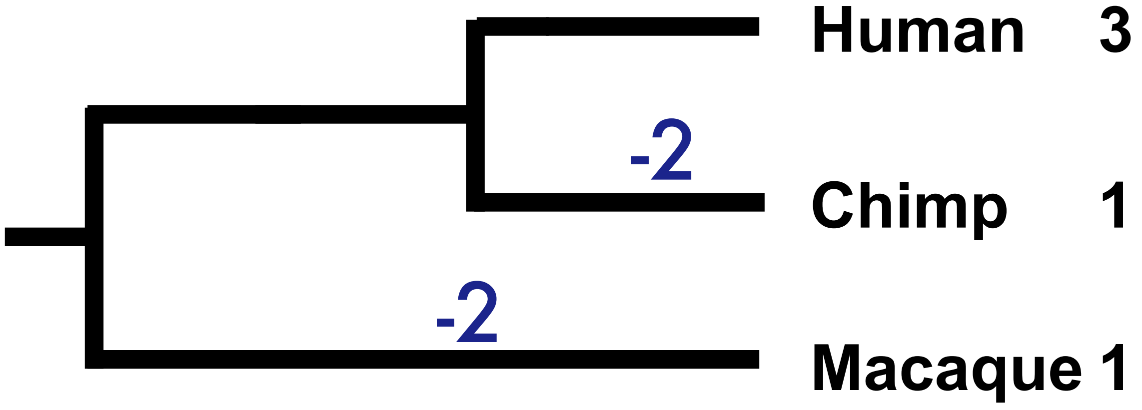
Species tree

Species trees vs. gene trees



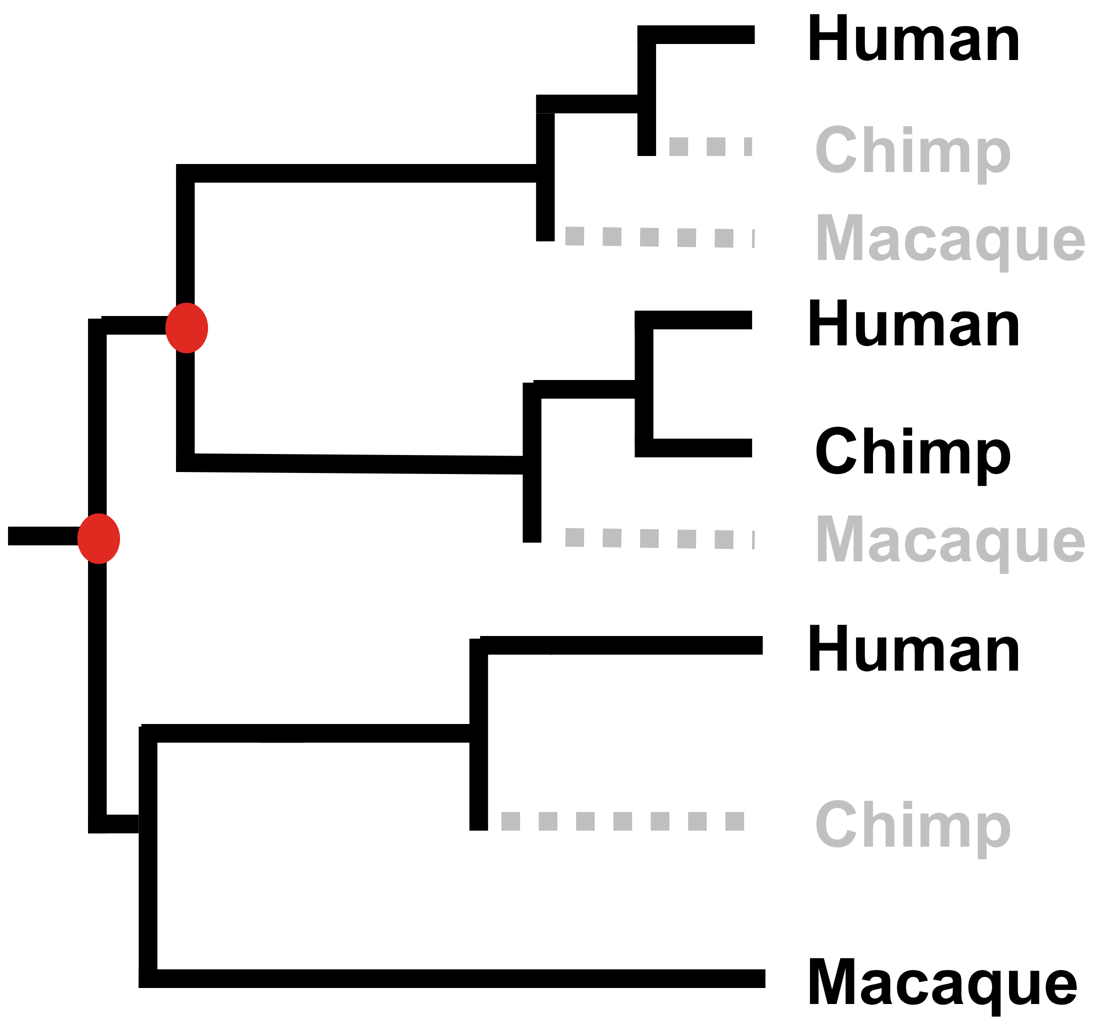
Gene tree

Species trees vs. gene trees

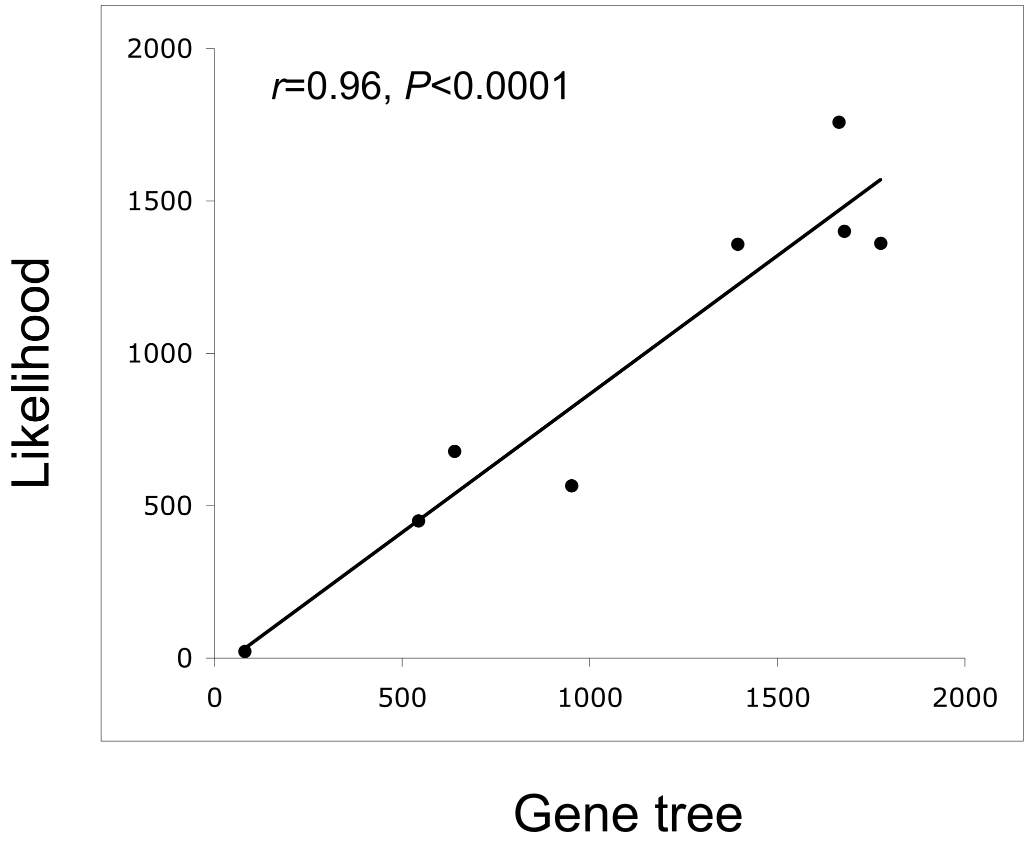


Species tree

Species trees vs. gene trees



Likelihood vs. Reconciliation



Hahn (2007) *Genome Biology*

Rapidly expanding gene families

Rapidly expanding gene families

The most common biological functions assigned to individual rapidly expanding families include:

immune defense

brain and neuronal development

intercellular transport

Rapidly expanding gene families

The most common biological functions assigned to individual rapidly expanding families include:

immune defense

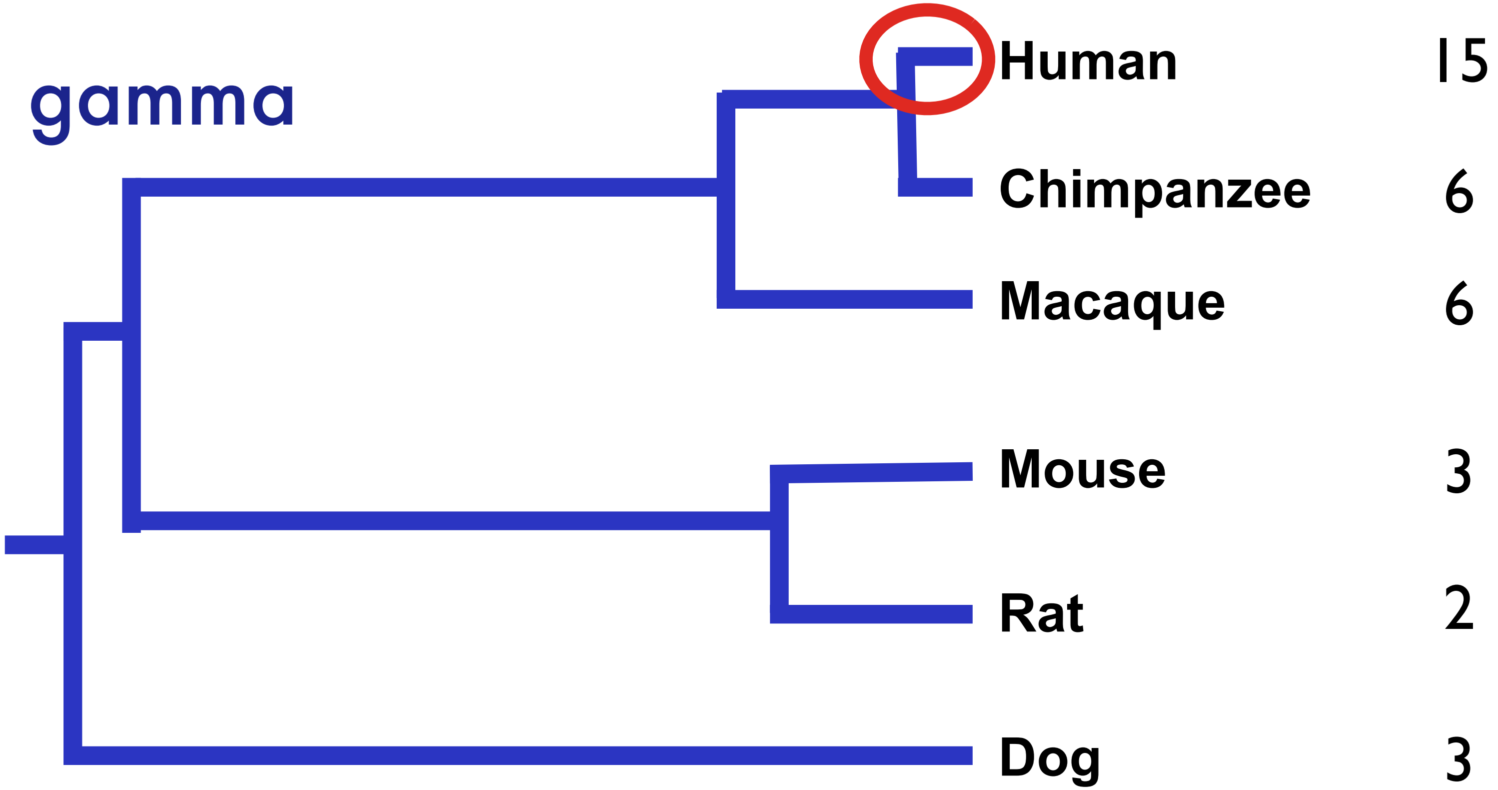
brain and neuronal development

intercellular transport

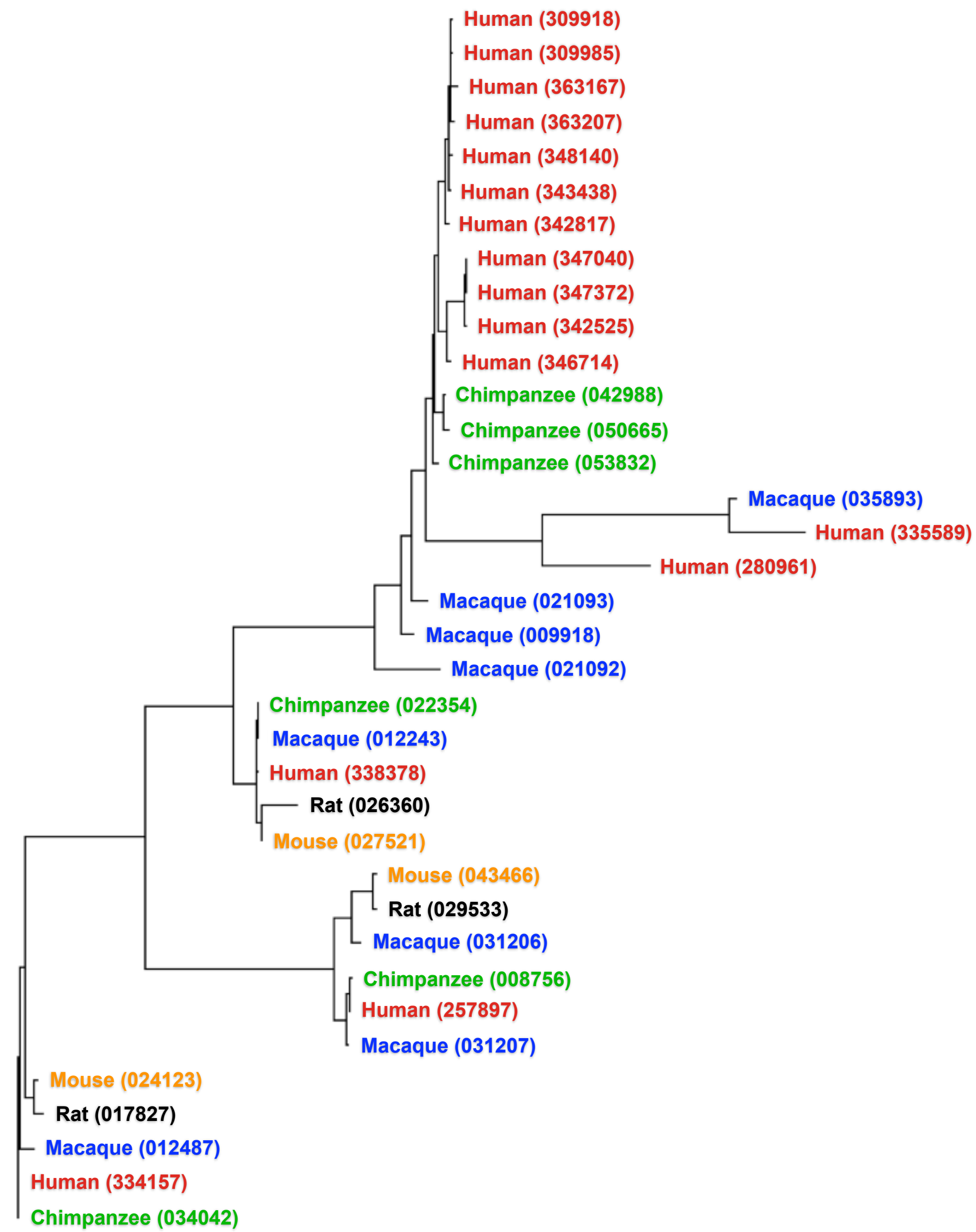
Interestingly, these are the same functions that evolve rapidly at the nucleotide level in primates.

Accelerated evolution of gene families

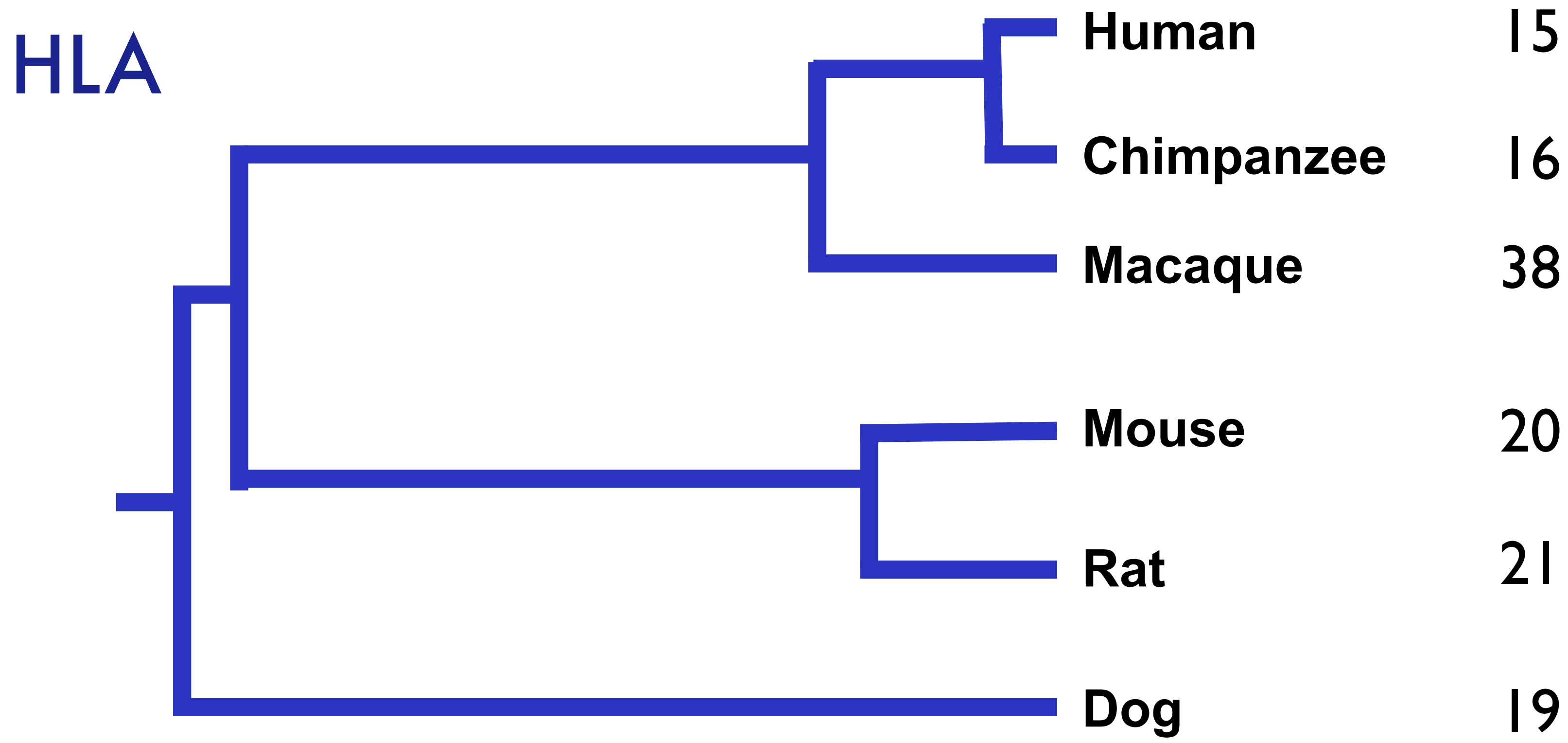
Centaurin gamma



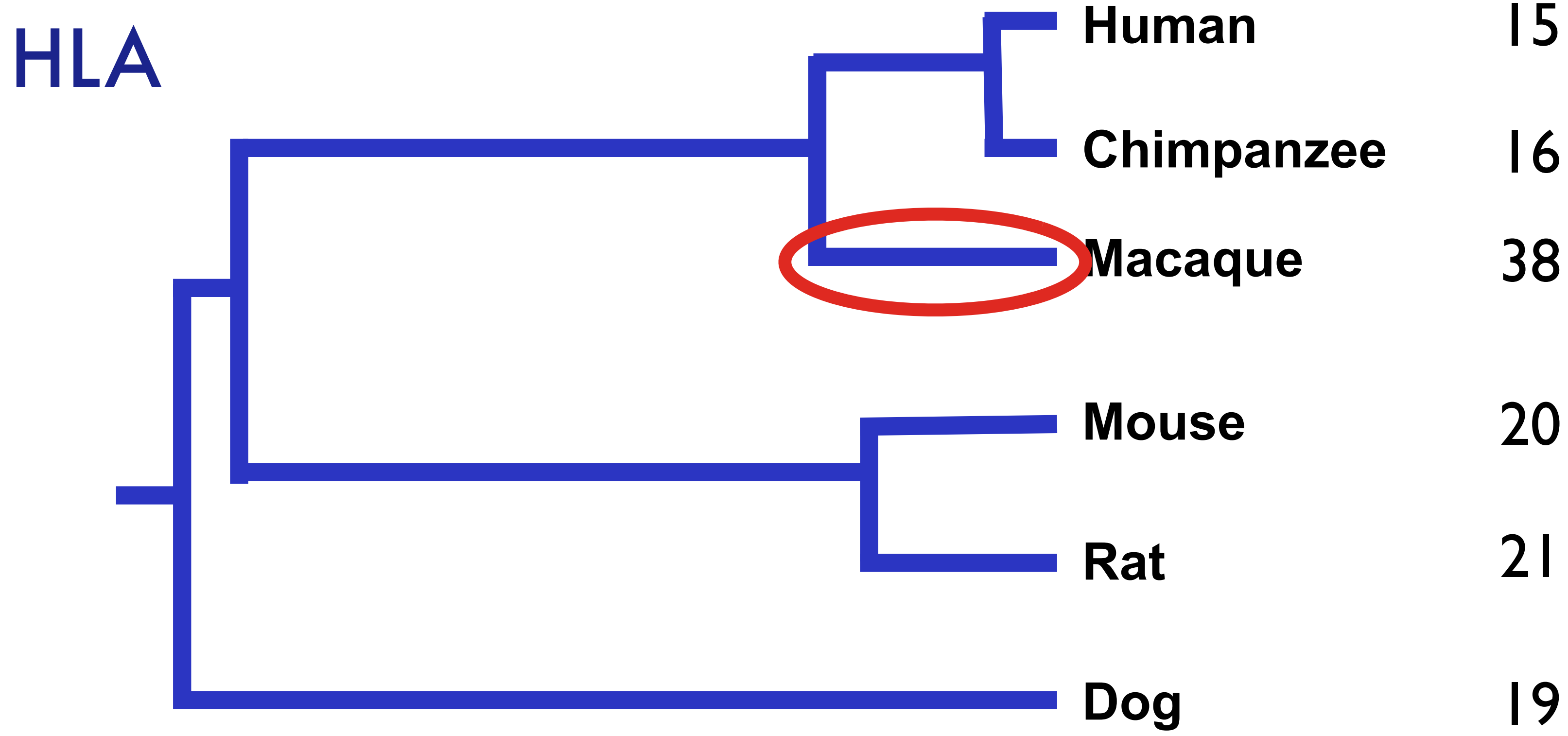
Large expansion of Centaurin gamma in humans



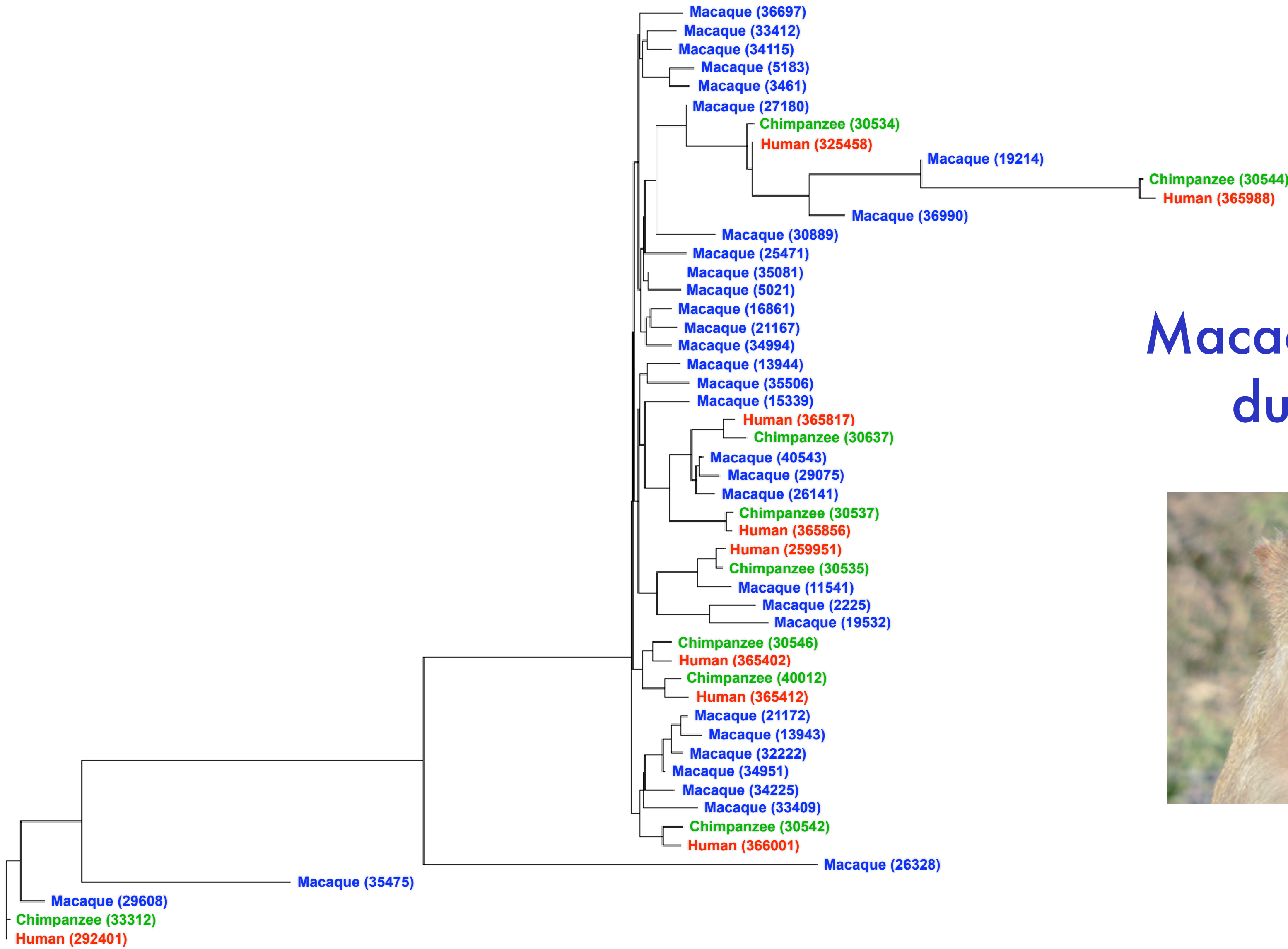
Large expansion of HLA genes in Rhesus macaque



Large expansion of HLA genes in Rhesus macaque



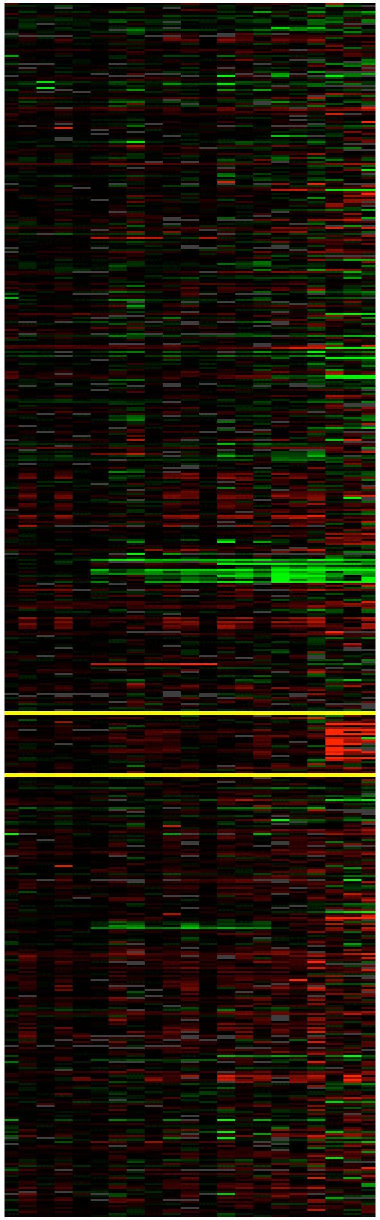
Large expansion of HLA genes in Rhesus macaque



Macaque-specific duplicates

Empirical evidence: HLA

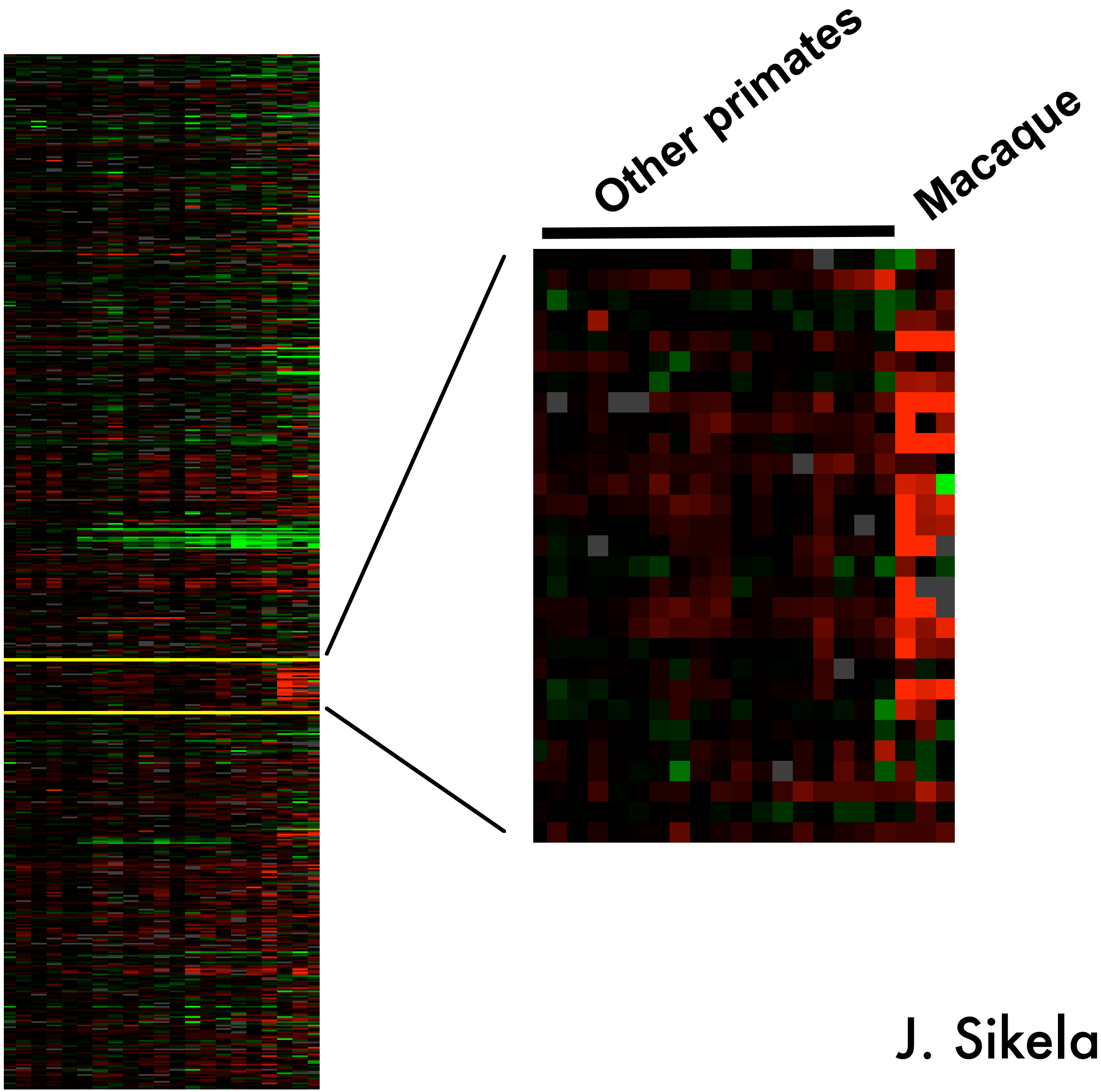
aCGH data



J. Sikela UCHSC

Empirical evidence: HLA

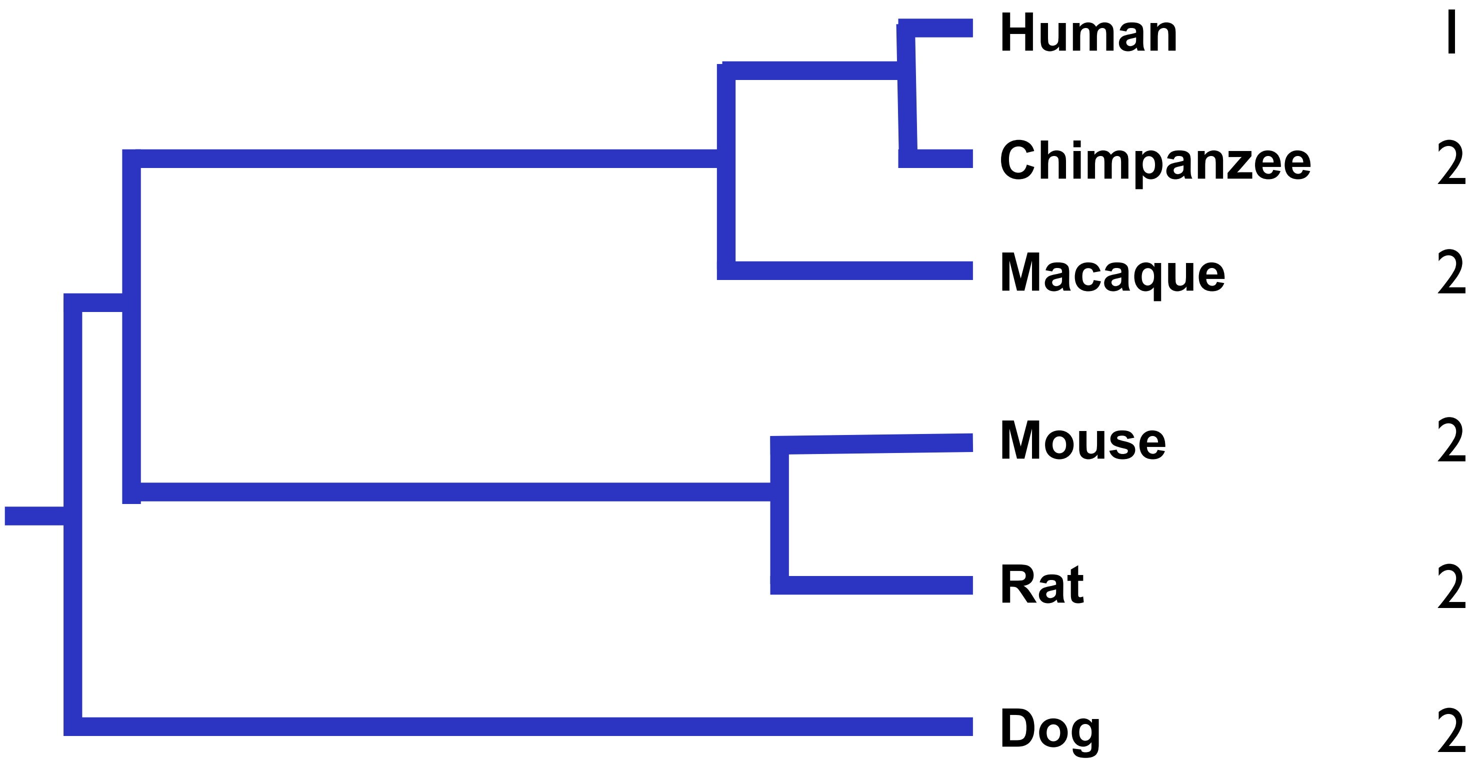
aCGH data



J. Sikela UCHSC

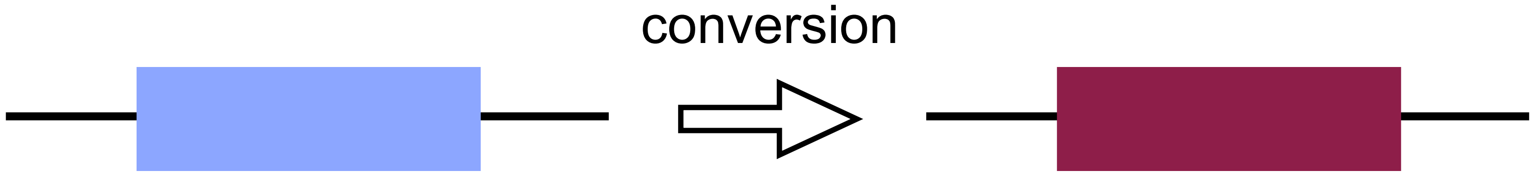
Gene losses

Gene losses

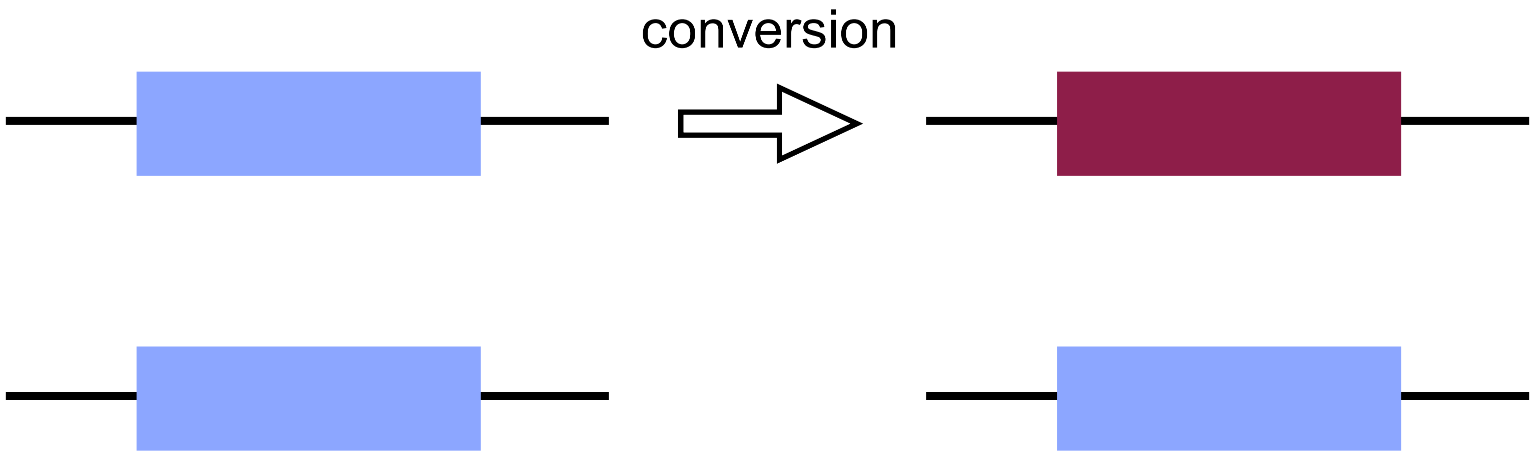


Gene losses

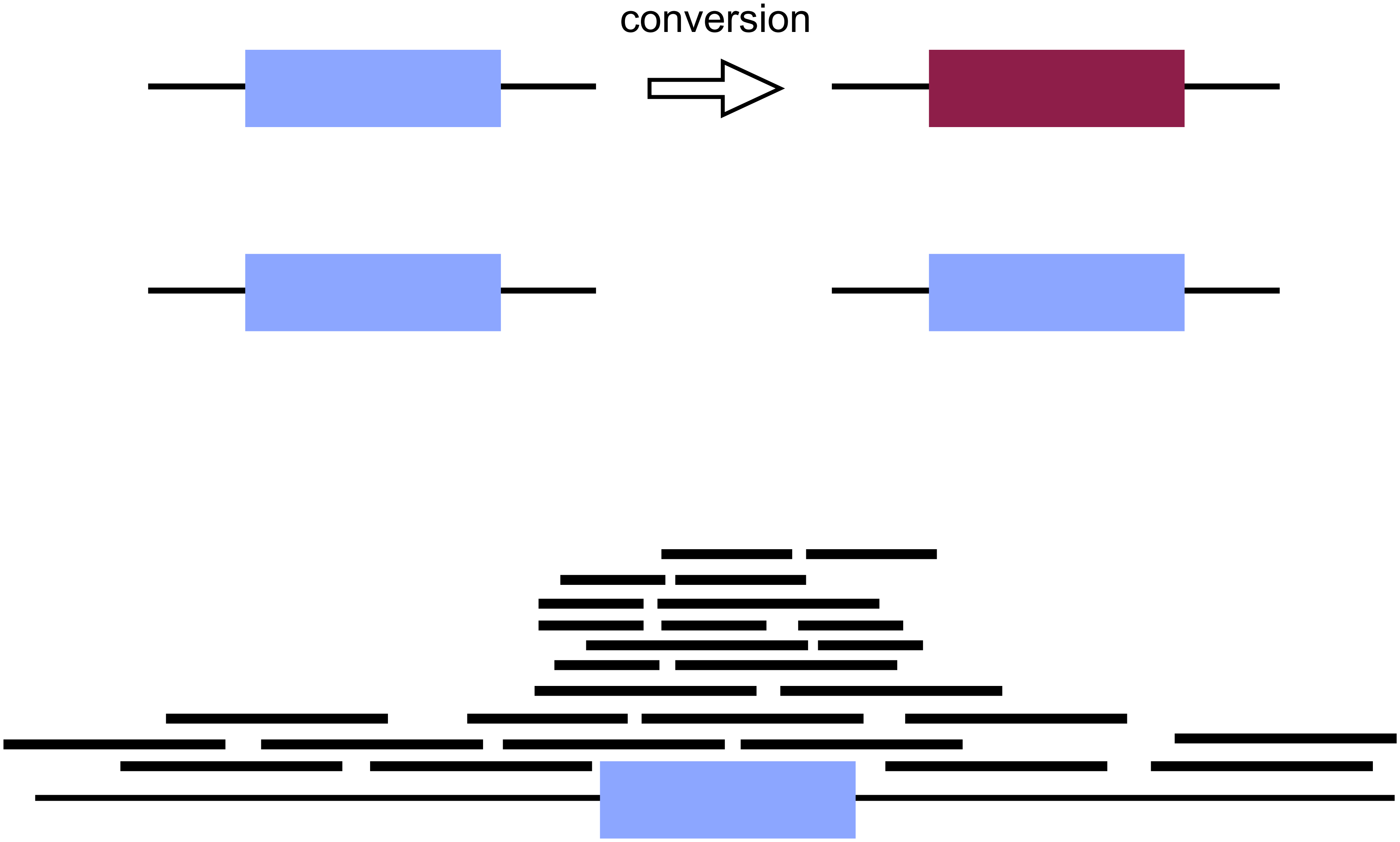
Gene losses



Gene losses



Gene losses



Gene losses

Checked for the presence of 424 genes “lost” from humans

Costello et al. (2008) *RECOMB-CG*
Schrider and Hahn (unpublished)

Gene losses

Checked for the presence of 424 genes “lost” from humans

203 genes not found

221 genes found

Costello et al. (2008) *RECOMB-CG*
Schrider and Hahn (unpublished)

Gene losses

Checked for the presence of 424 genes "lost" from humans

203 genes not found

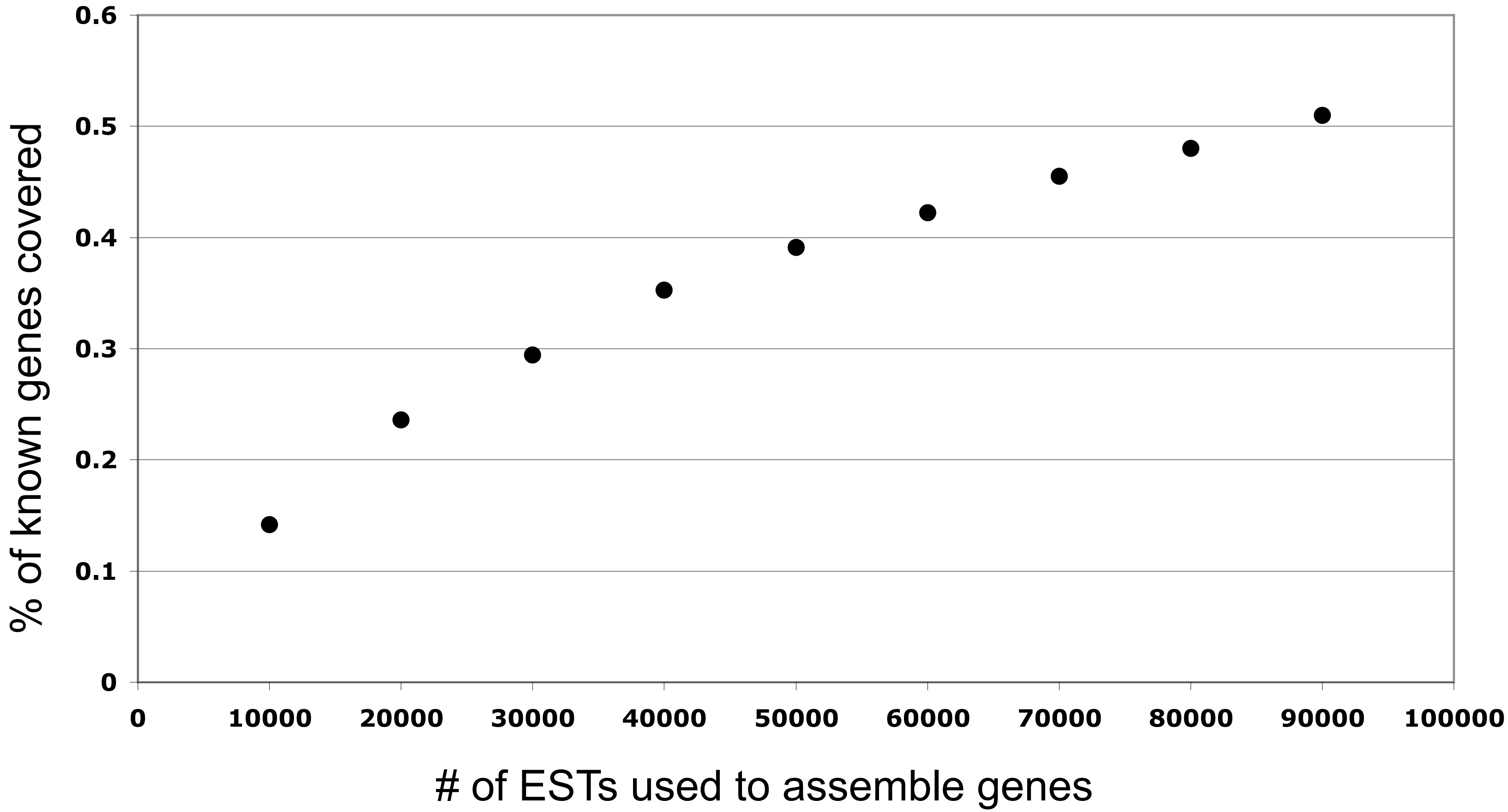
221 genes found

191 lost

12 in assembly gaps

Costello et al. (2008) *RECOMB-CG*
Schrider and Hahn (unpublished)

Quantifying gain and loss in low-coverage genomes



Differences between human and chimp

Differences between human and chimp

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

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

Human-chimp divergence

Human-chimp divergence

Polymorphism ($4N\mu$)

Divergence ($2T\mu$)

Human-chimp divergence

	Polymorphism (4Nμ)	Divergence (2Tμ)
Nucleotides	0.10%	1.23%

Human-chimp divergence

	Polymorphism ($4N\mu$)	Divergence ($2T\mu$)
Nucleotides	0.10%	1.23%
Copy Number		

Human-chimp divergence

	Polymorphism (4Nμ)	Divergence (2Tμ)
Nucleotides	0.10%	1.23%
Copy Number		6.40%

Human-chimp divergence

	Polymorphism ($4N\mu$)	Divergence ($2T\mu$)
Nucleotides	0.10%	1.23%
Copy Number	0.55%	6.40%

McCarroll et al. (2008) *Nature Genetics*

The King and Wilson paradox

Humans and chimps are 1% 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% different at the nucleotide level

AACGCATCGATCGATCAGCTACGACG----------

----------TCGATCAACTACGACGAACGCATCGA

Differences between human and chimp

Do any of these gains or losses matter?

Outline

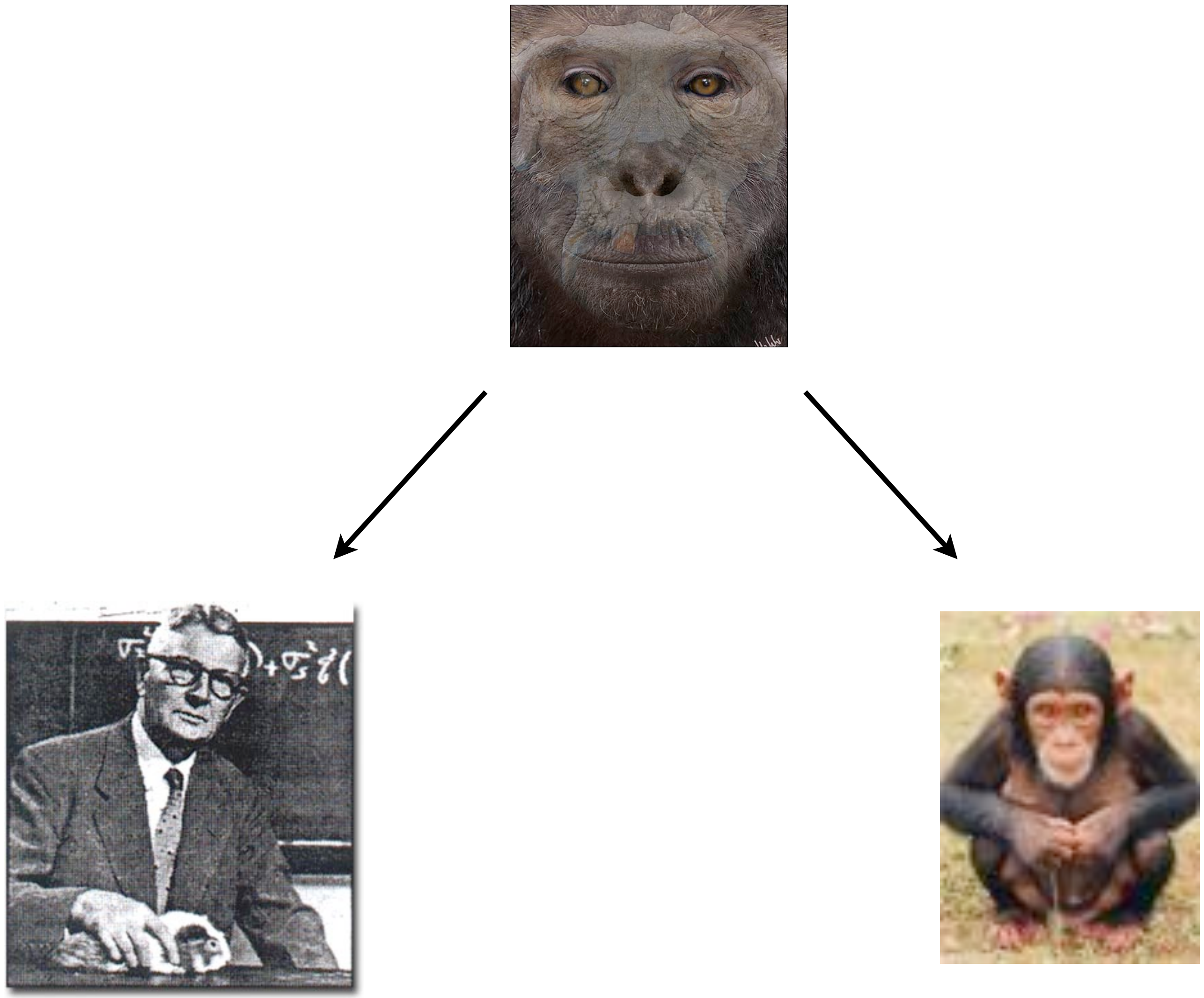
I. Statistical and computational methods

II. Quantifying gene gain and loss

III. Natural selection on gene duplicates

Genome scans for positive selection

Genome scans for positive selection



Genome scans for positive selection

Positive selection in humans:

Genome scans for positive selection

Positive selection in humans:

	<u>*P*<0.05</u>
Nielsen et al. 2005	35/13,653 (0.2%)

Genome scans for positive selection

Positive selection in humans:

	P<0.05	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0

Genome scans for positive selection

Positive selection in humans:

	$P<0.05$	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0
Bakewell et al. 2007 Gibbs et al. 2007	154/13,888 (1.1%)	2

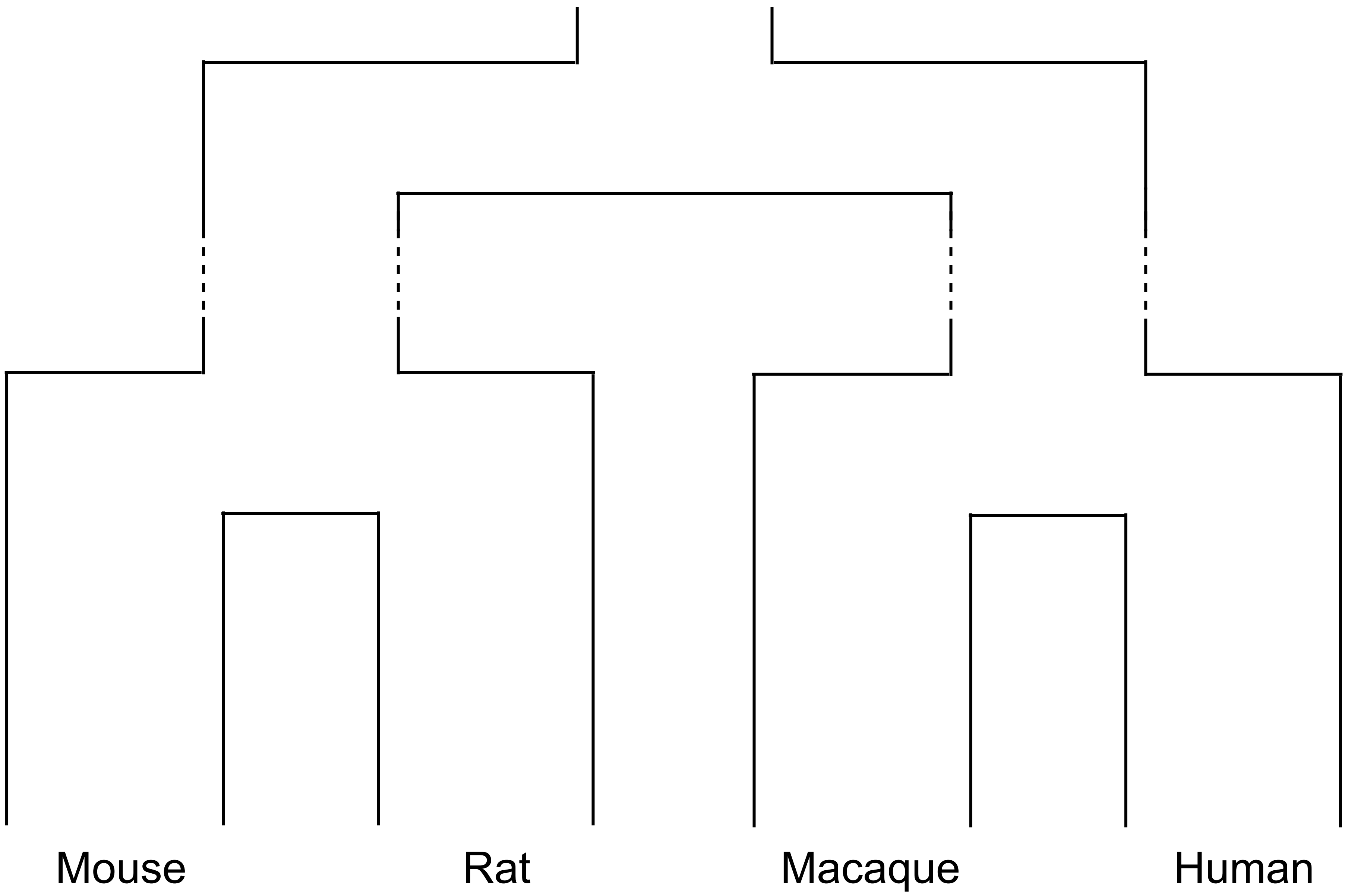
Genome scans for positive selection

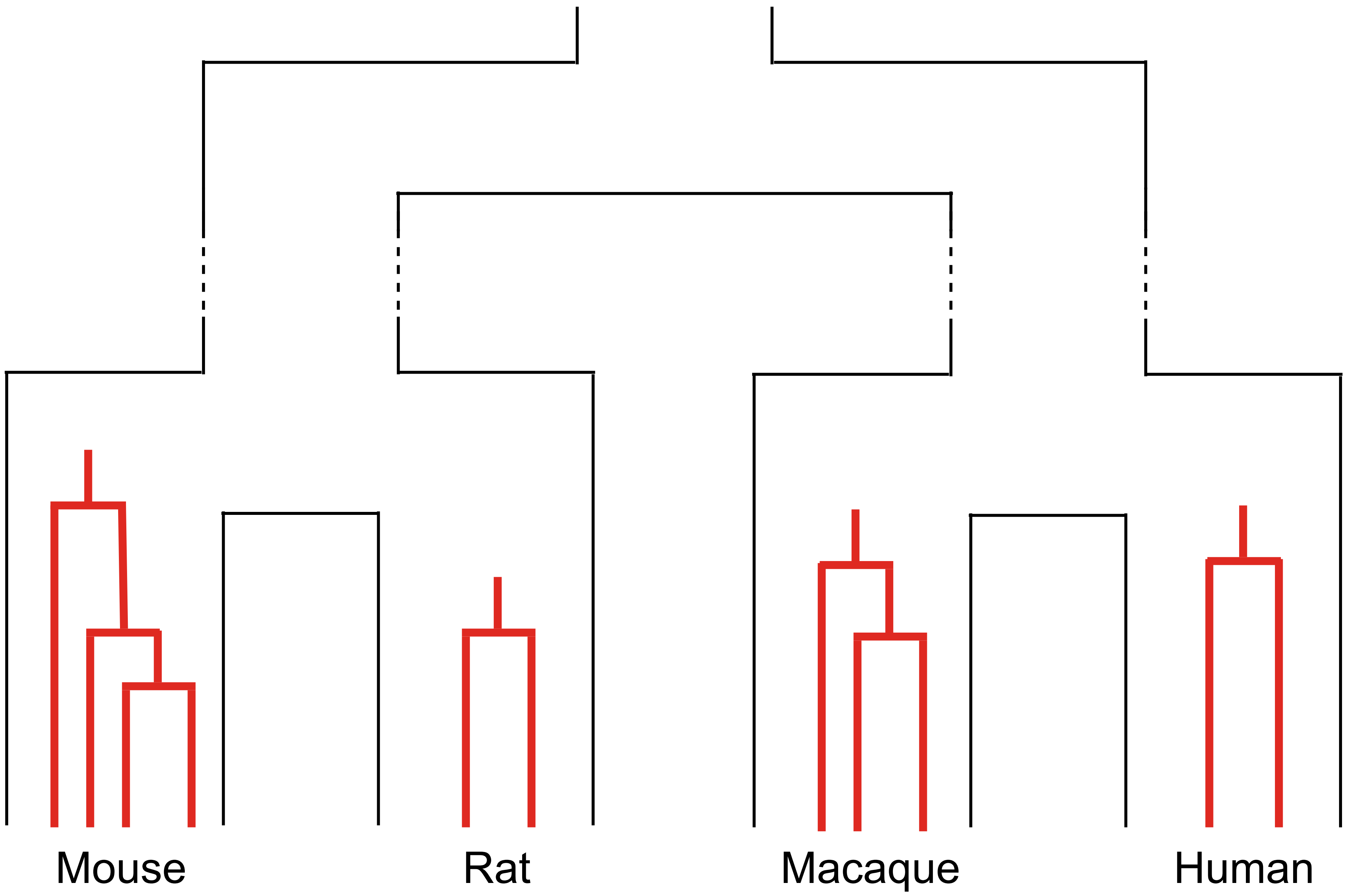
Positive selection in humans:

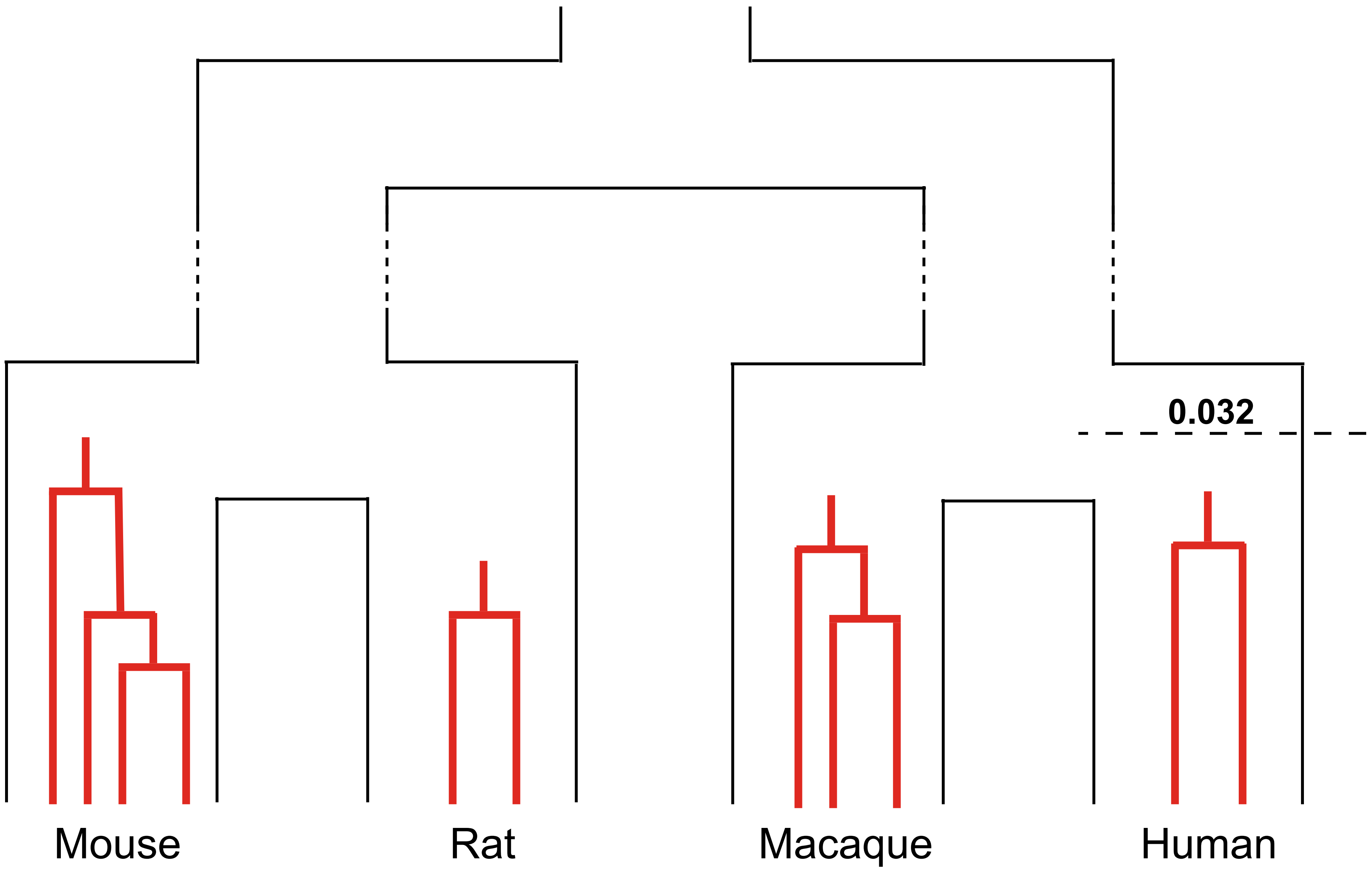
	P<0.05	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0
Bakewell et al. 2007 Gibbs et al. 2007	154/13,888 (1.1%)	2

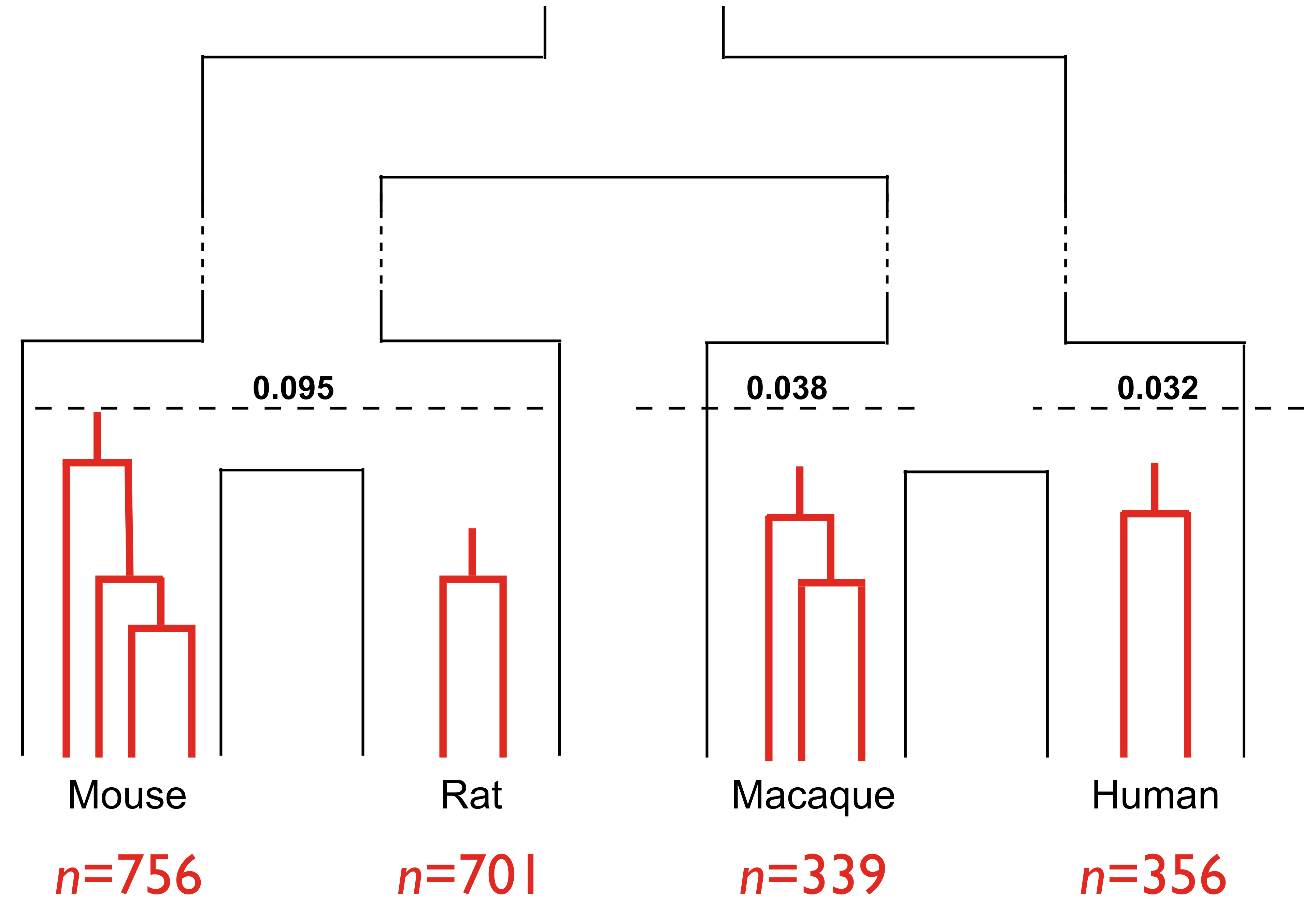
These analyses do not consider duplicates!

Genome scan for positive selection on duplicates


Mouse
Rat
Macaque
Human


Mouse
Rat
Macaque
Human


0.032
Mouse
Rat
Macaque
Human


0.095
0.038
0.032
Mouse
Rat
Macaque
Human
n=756
n=701
n=339
n=356

Test for dN/dS>1

M1a:

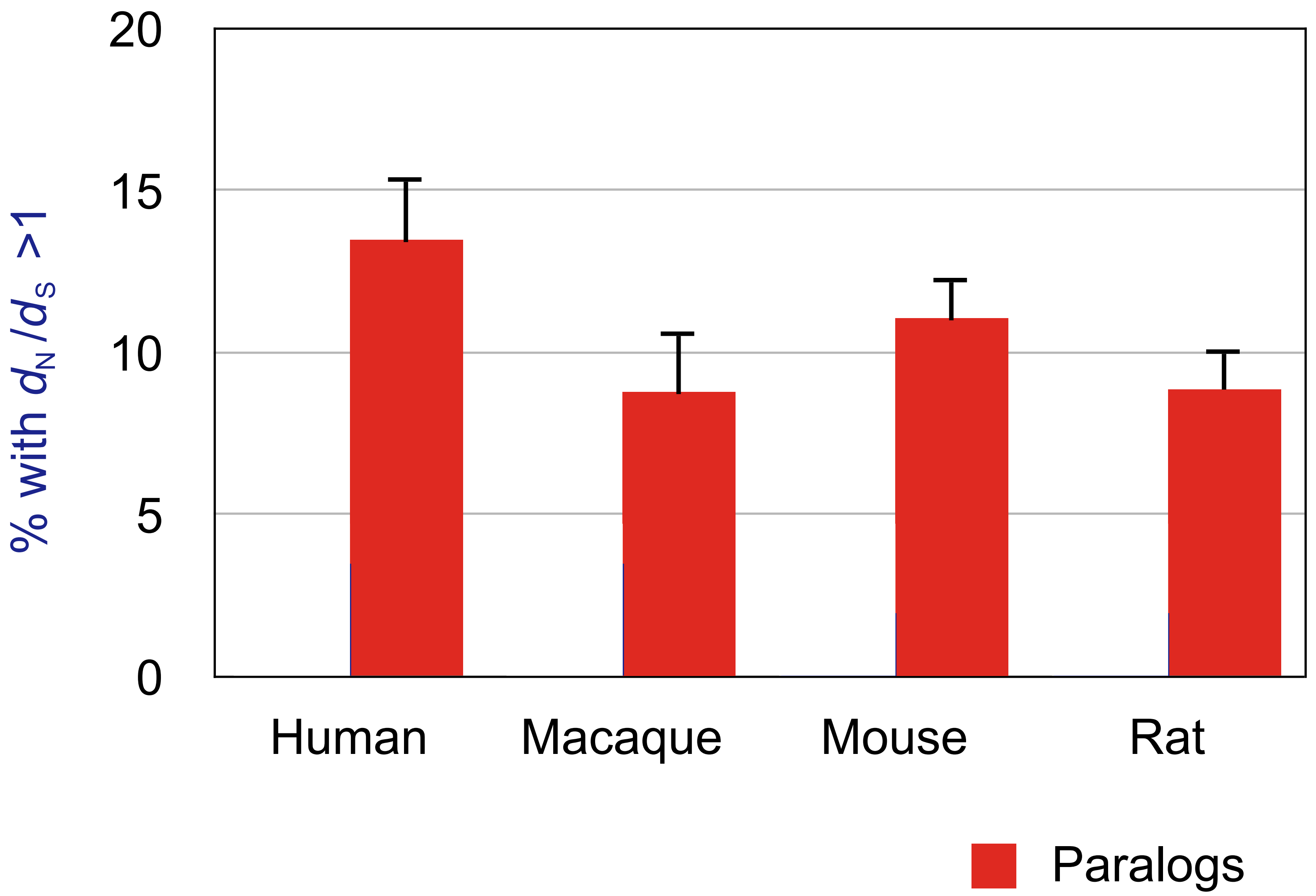
dN/dS<1
dN/dS=1

M2a:

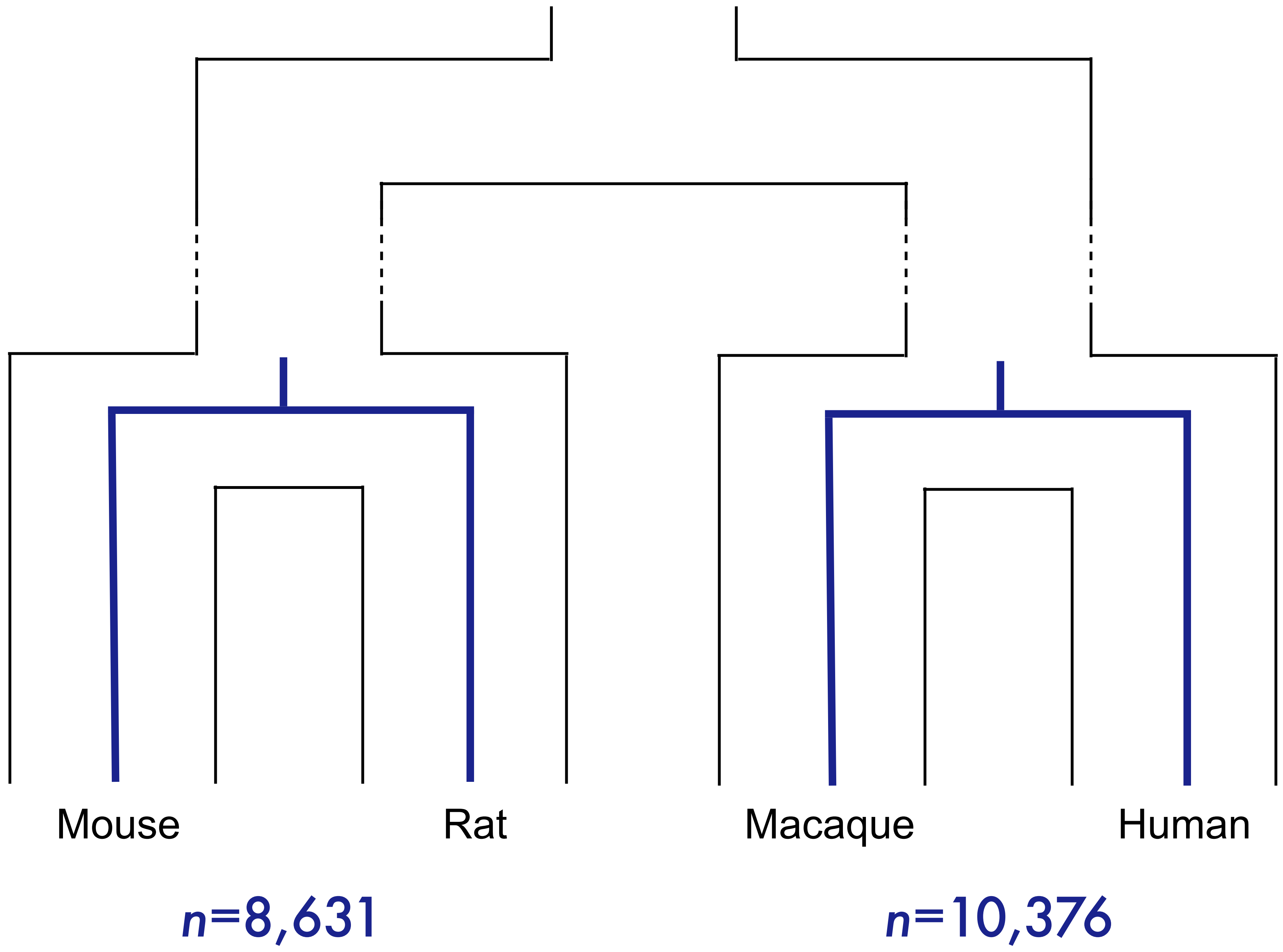
dN/dS<1
dN/dS=1
dN/dS>1

Compare M1a vs. M2a in likelihood ratio test using PAML

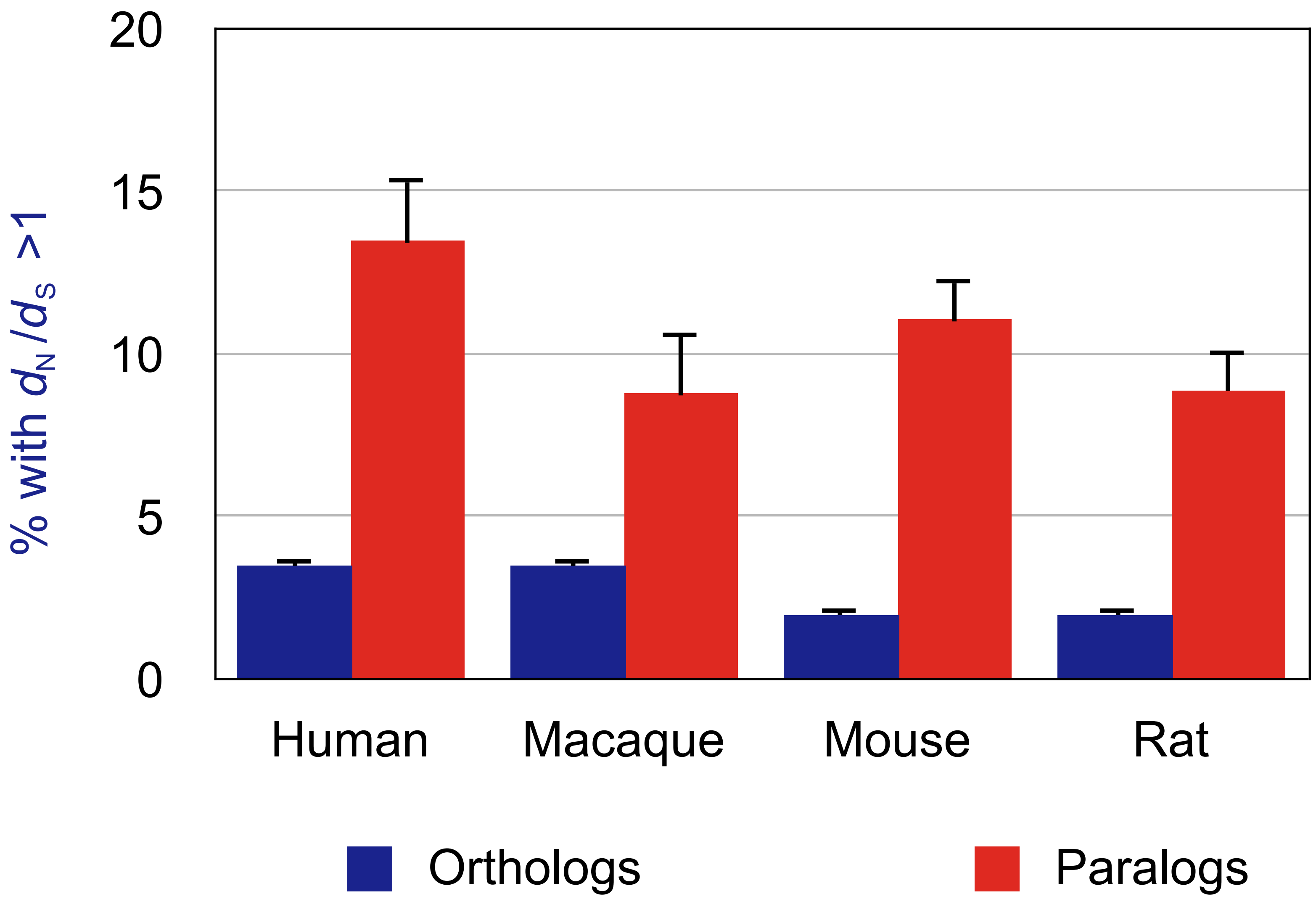
All duplicates



Han et al. (*submitted*)

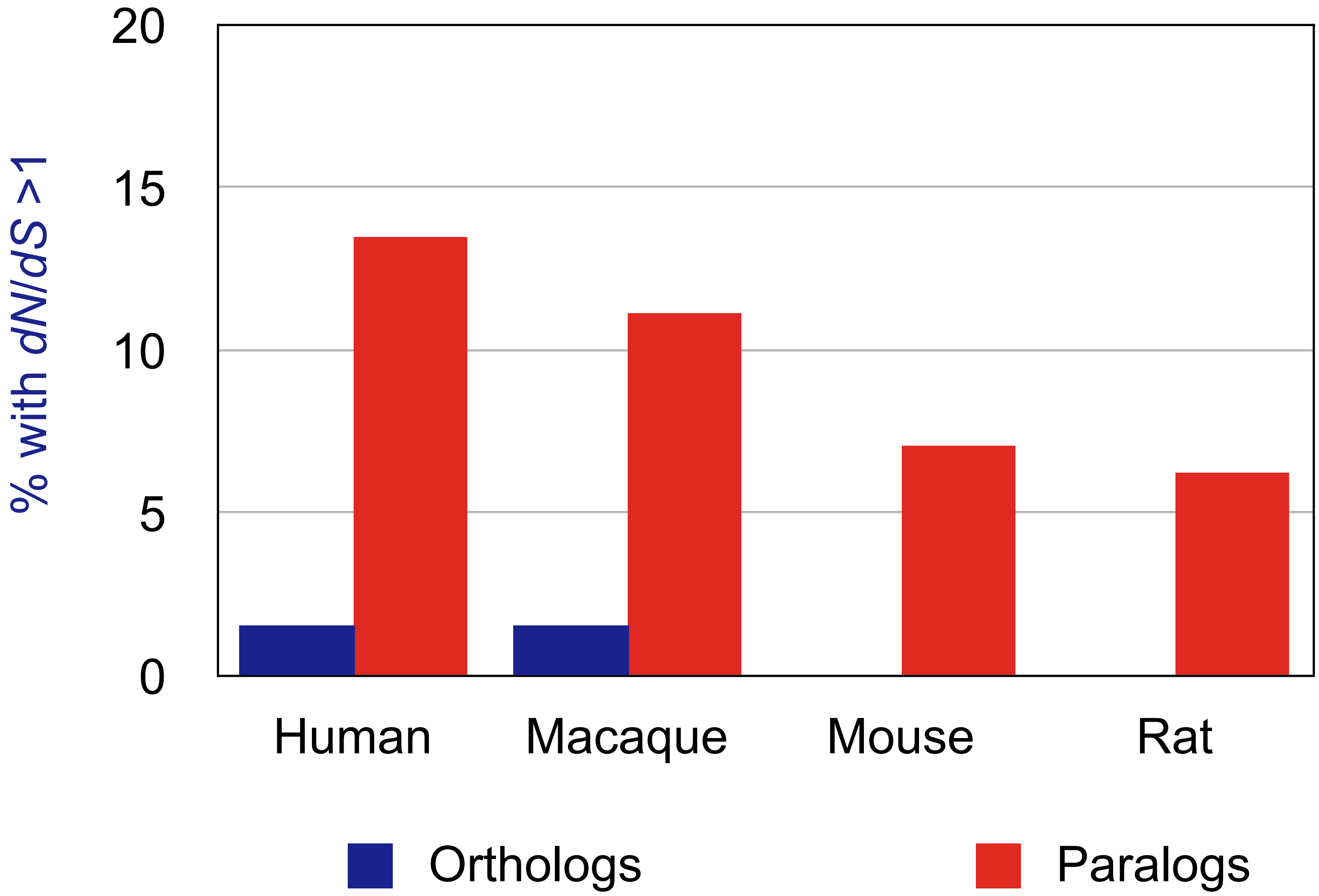

Mouse
Rat
Macaque
Human
n=8,631
n=10,376

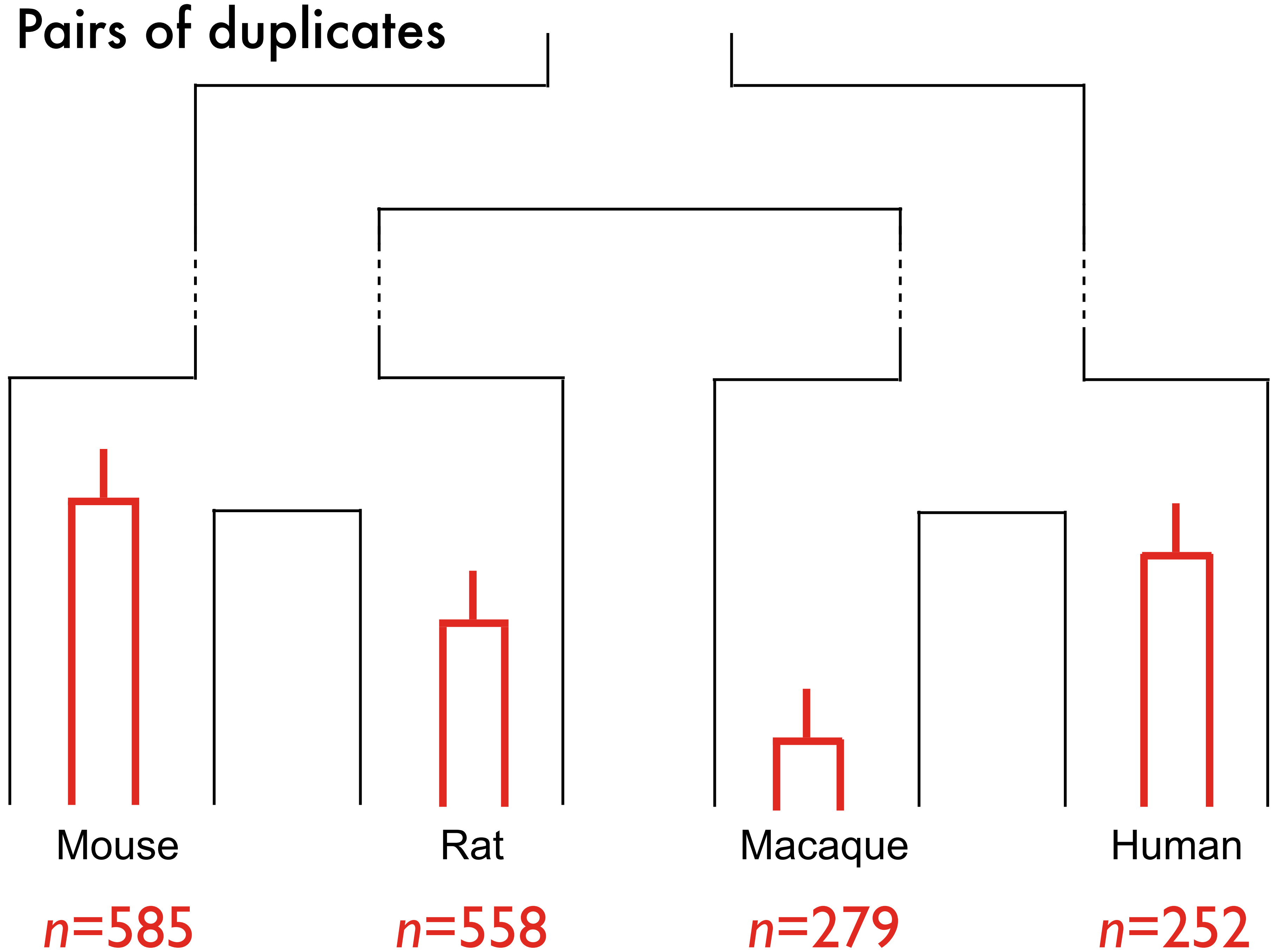
All duplicates + orthologs



Han et al. (*submitted*)

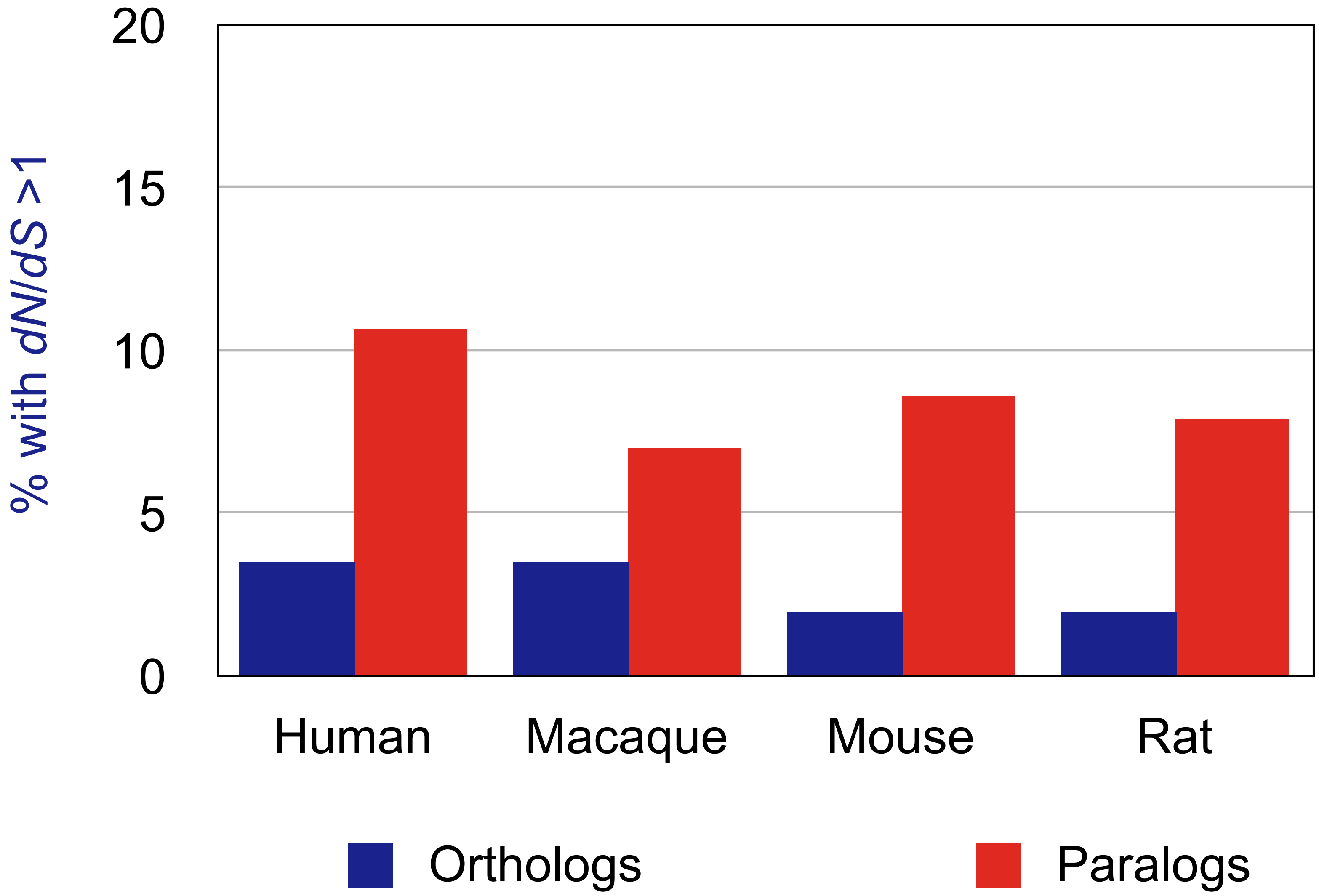
All duplicates (Nei-Gojobori)





Pairs of duplicates
Mouse
Rat
Macaque
Human
n=585
n=558
n=279
n=252

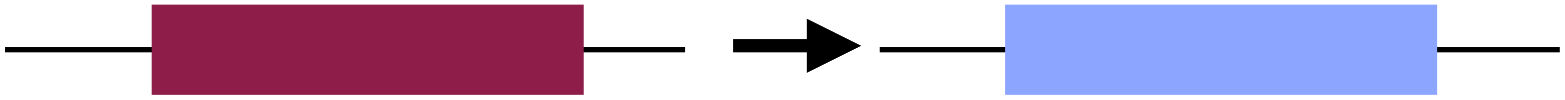
Pairs of duplicates



Gene conversion

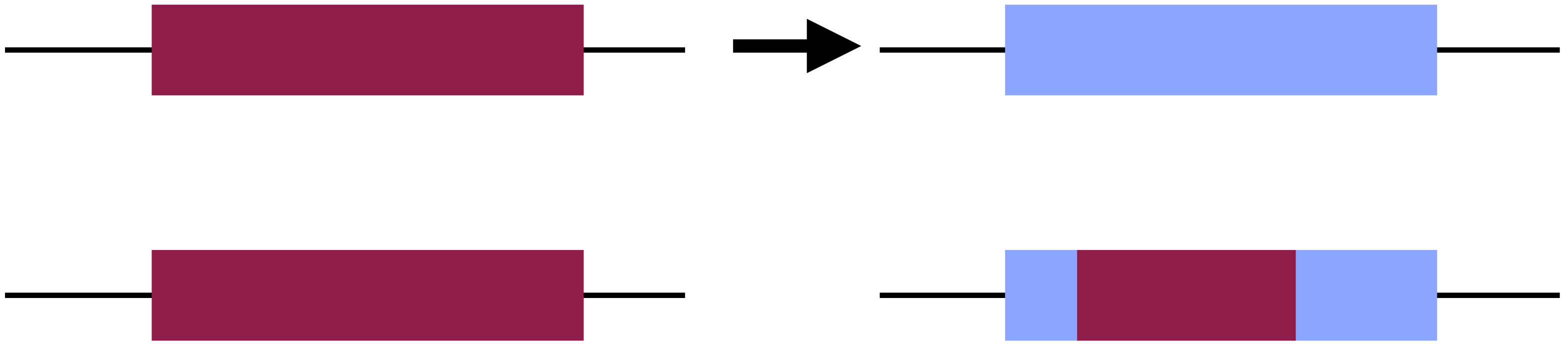
Casola and Hahn (*submitted*)

Gene conversion



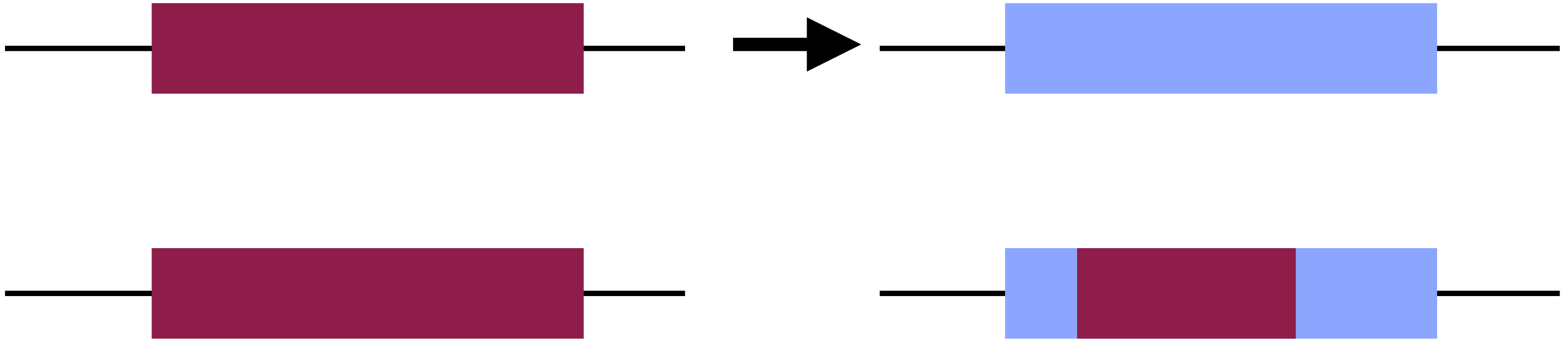
Casola and Hahn (*submitted*)

Gene conversion



Casola and Hahn (*submitted*)

Gene conversion



Gene conversion may cause false positives in tests for selection

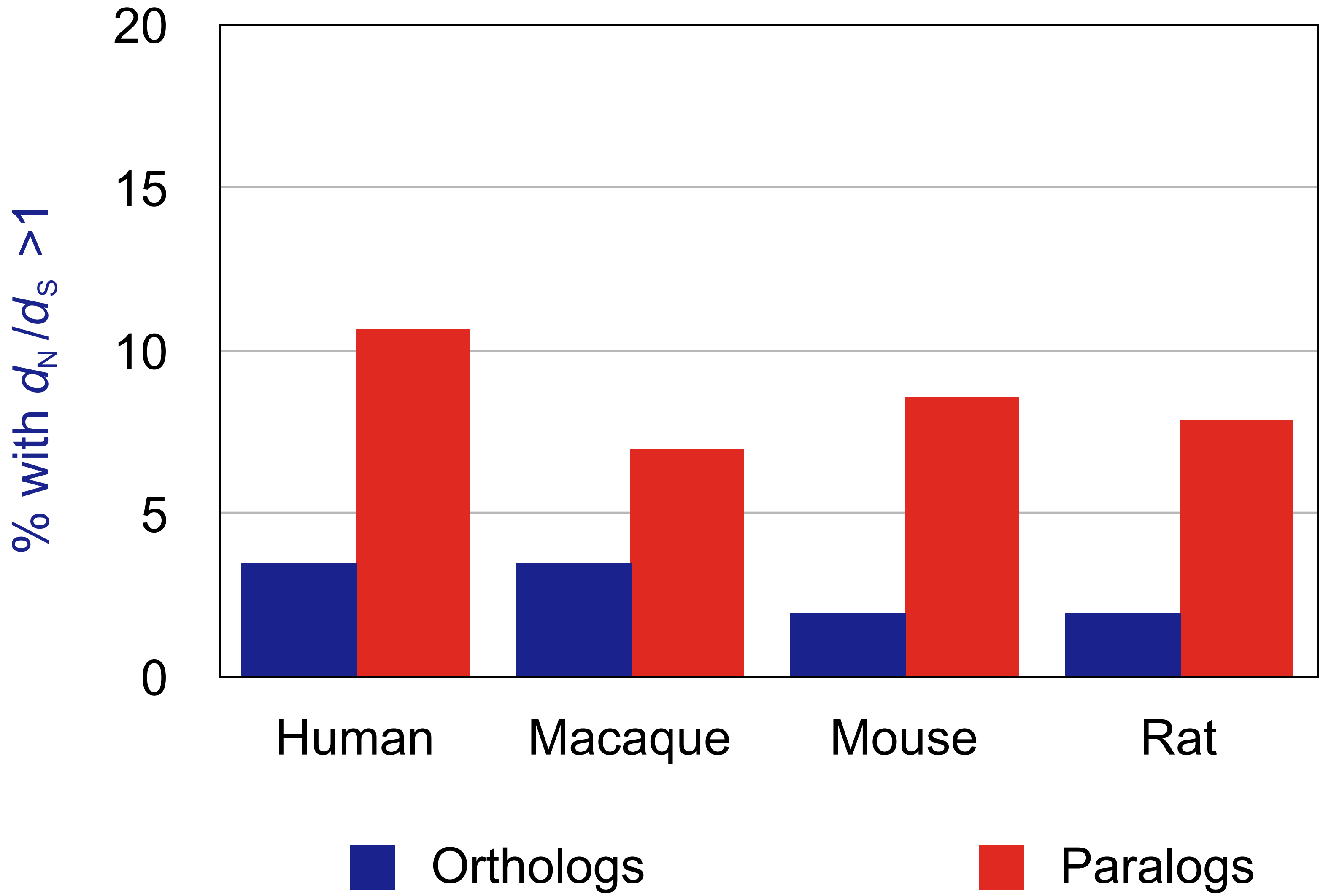
Casola and Hahn (*submitted*)

Gene conversion

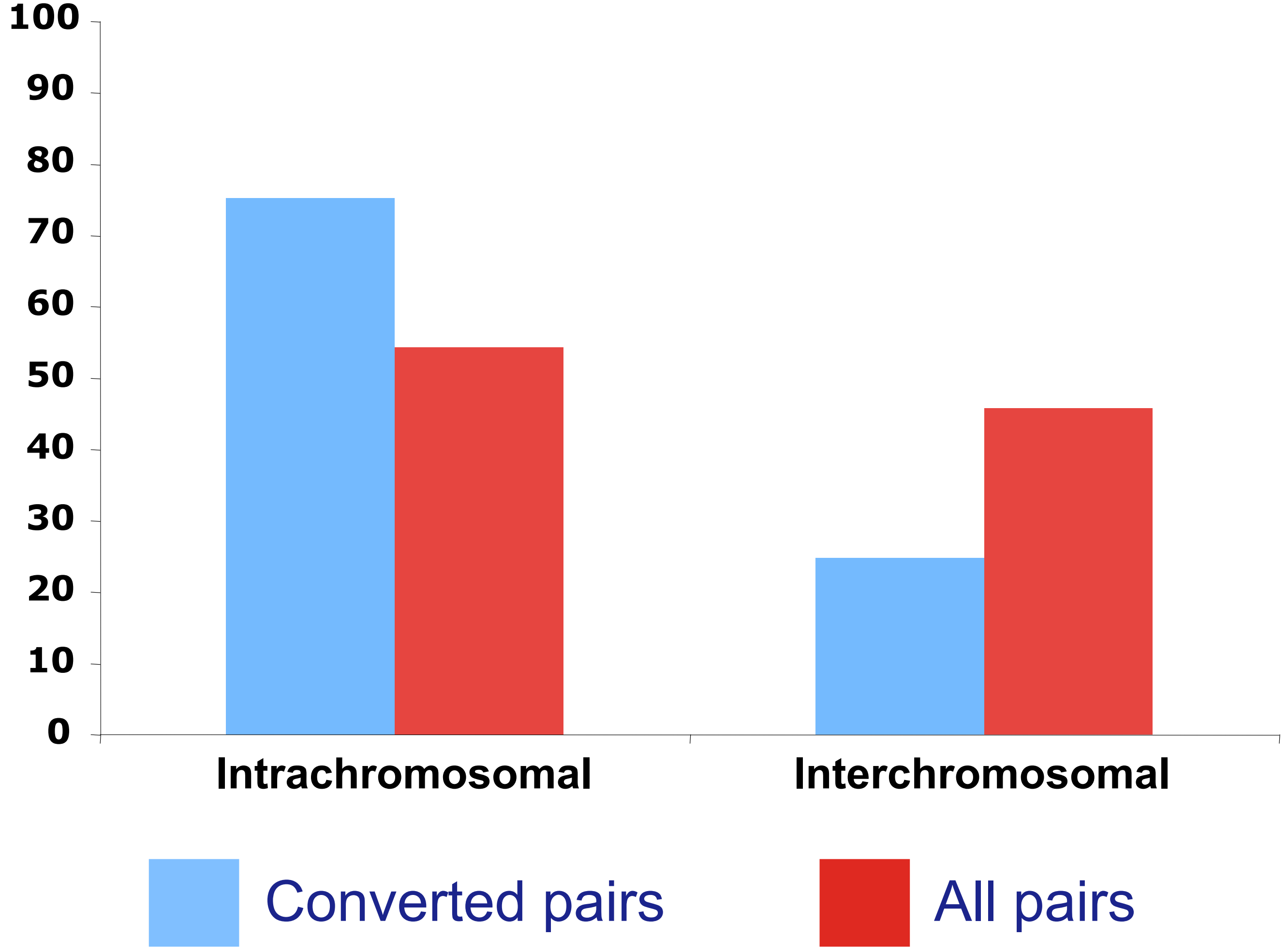
(<5% of paralogs have undergone conversion)

Gene conversion

(<5% of paralogs have undergone conversion)



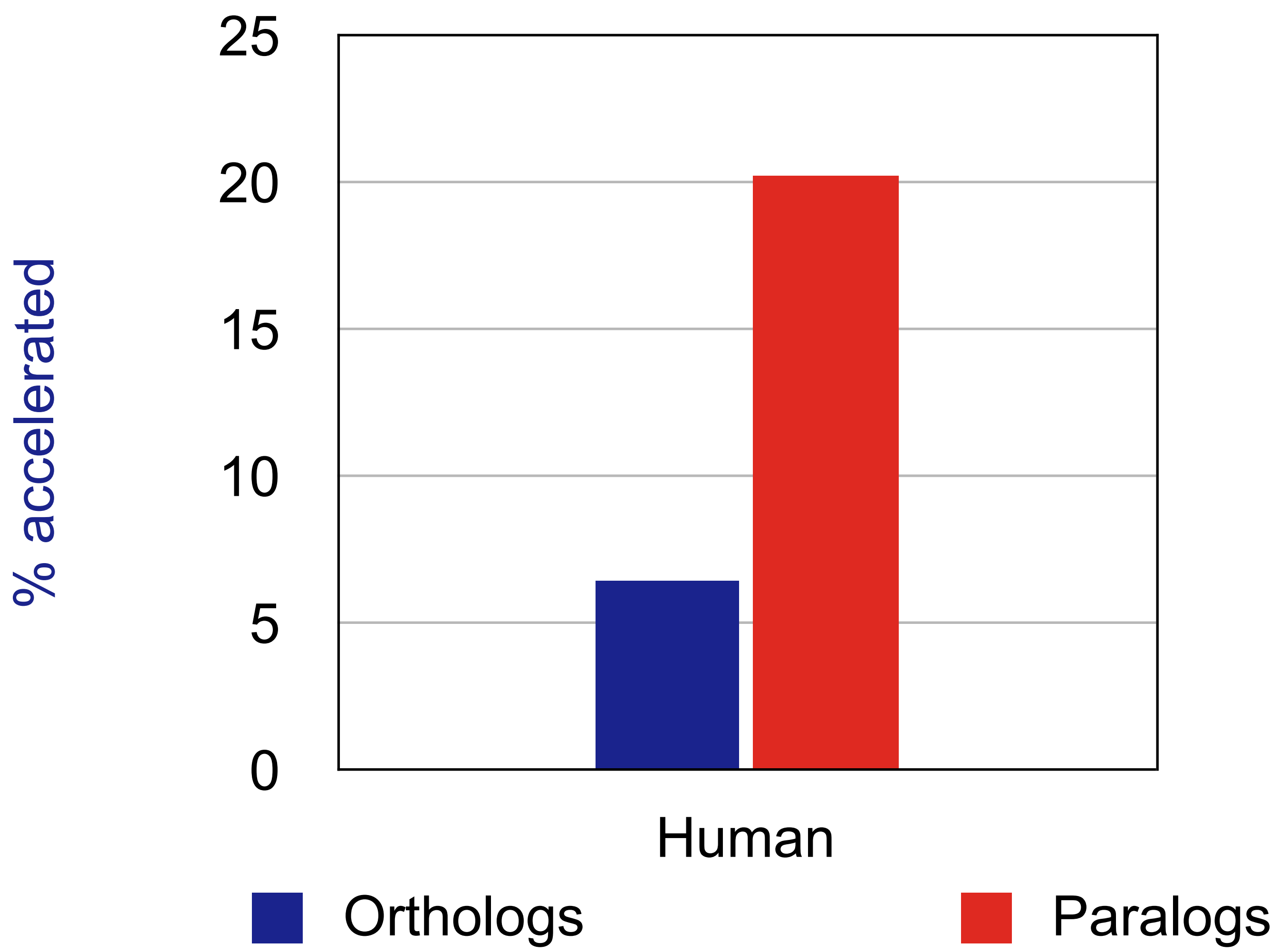
Gene conversion



McGrath et al. (*submitted*)

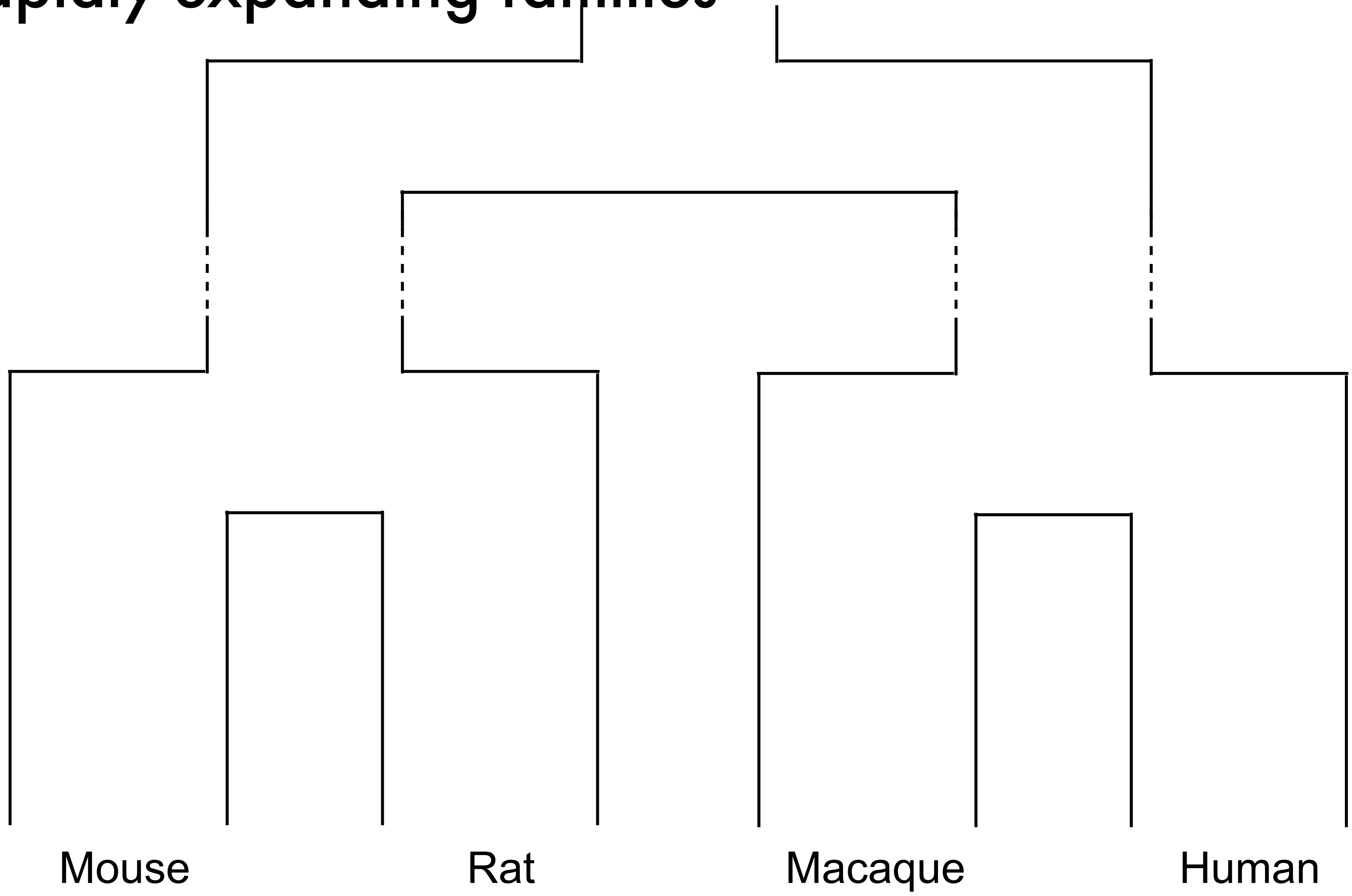
5kb upstream

5kb upstream

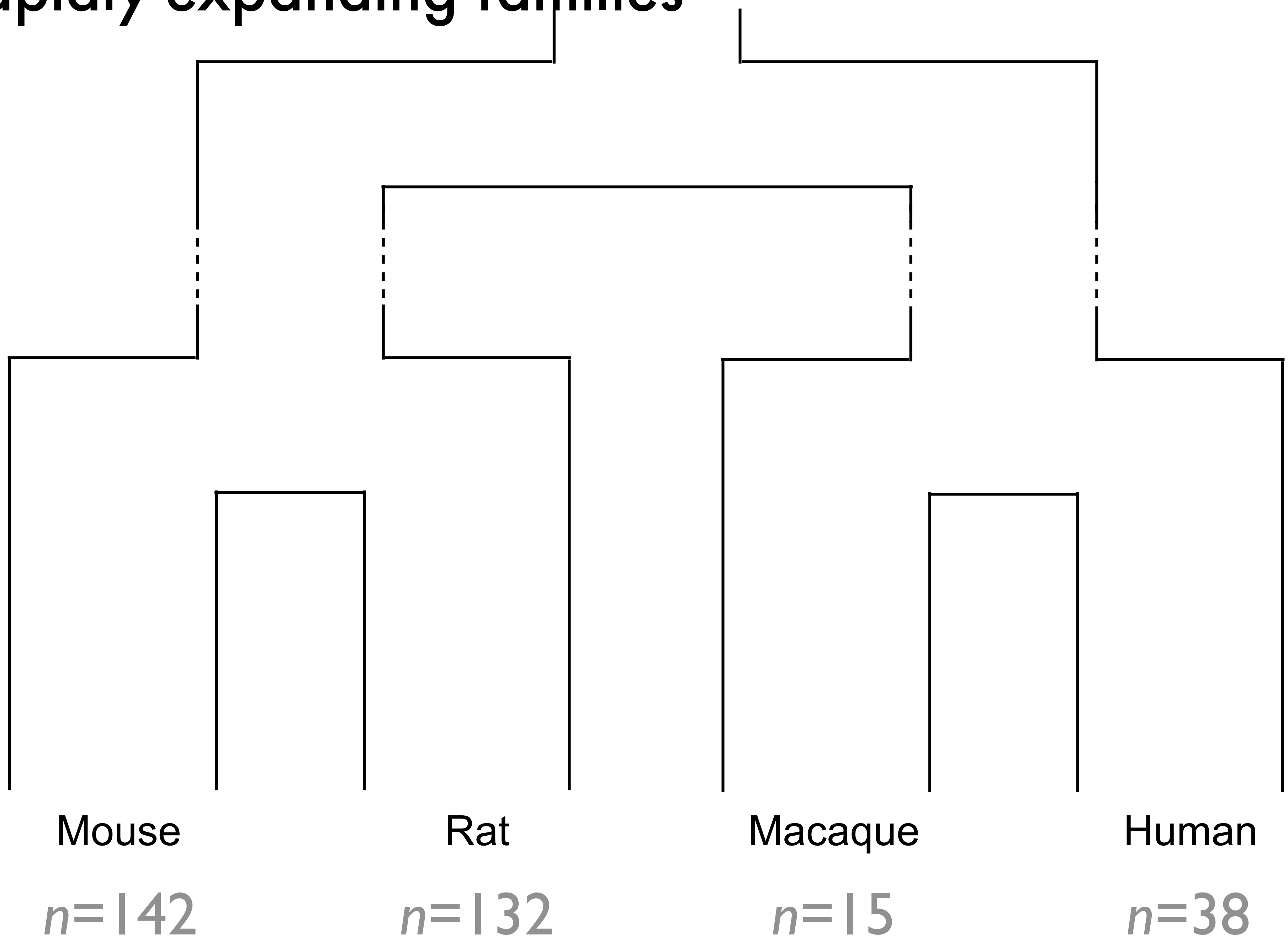


K. Pollard, UCSF

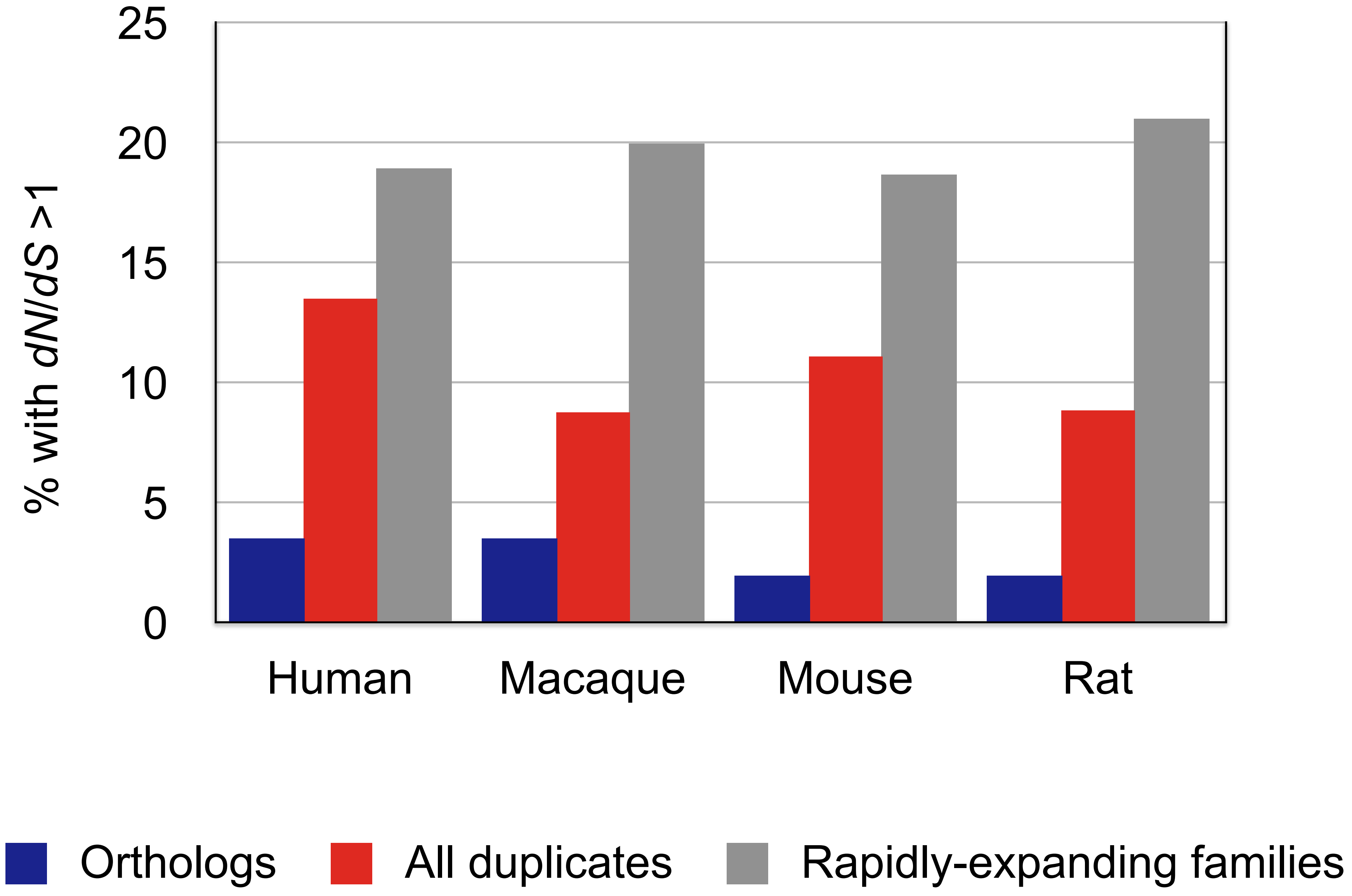
Rapidly-expanding families



Rapidly-expanding families



Rapidly-expanding families



Genome scan for positive selection on duplicates

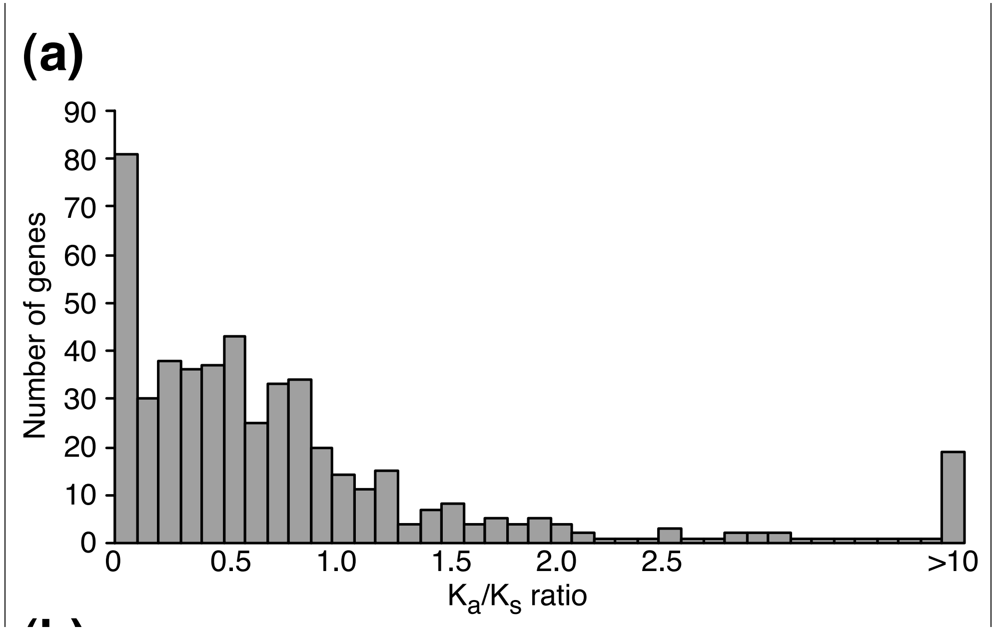
Genome scan for positive selection on duplicates

There's lots of positive selection!
(at least twice as much as on single-copy genes)

Why hasn’t anyone seen this before?

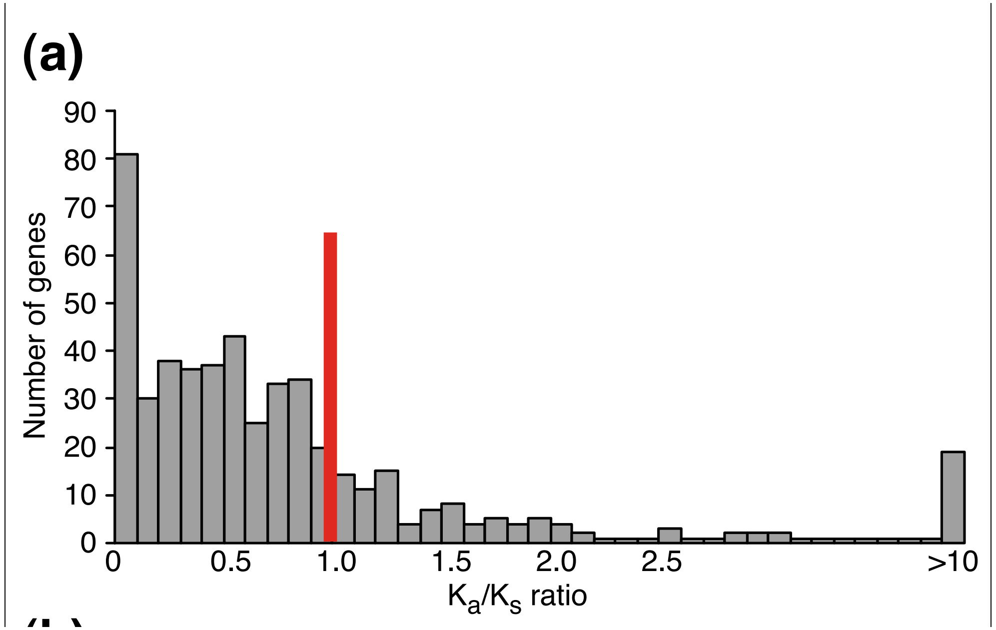
Why hasn't anyone seen this before?

Zhang, Gu, and Li 2003



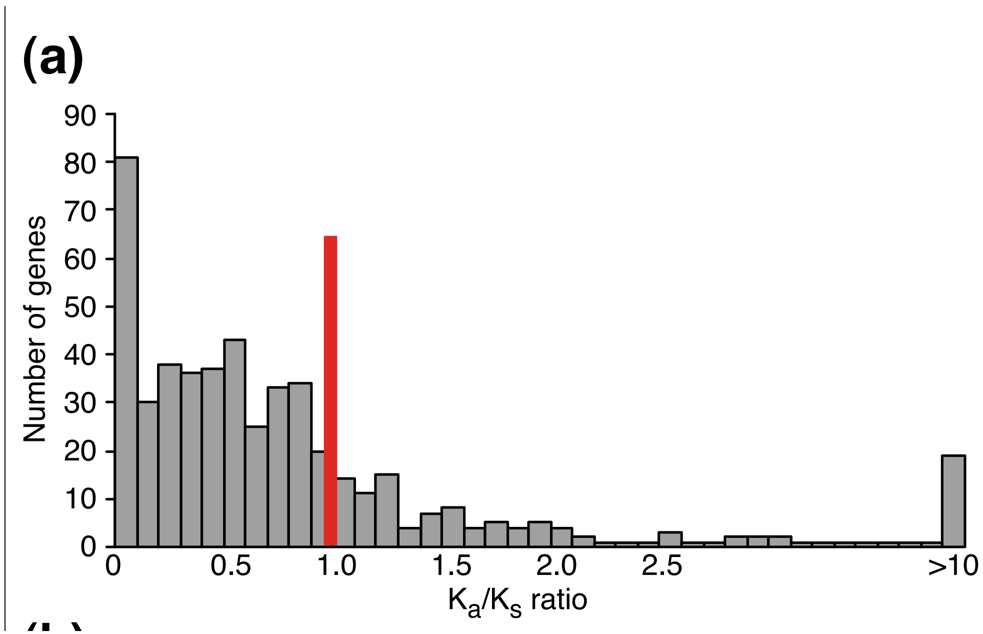
Why hasn't anyone seen this before?

Zhang, Gu, and Li 2003

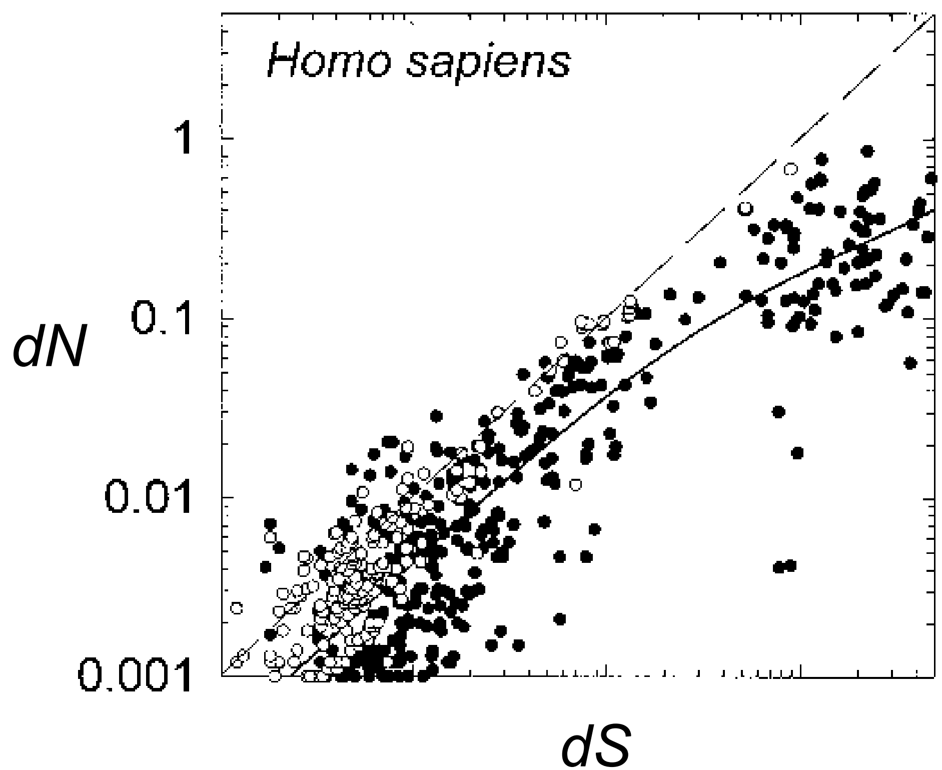


Why hasn't anyone seen this before?

Zhang, Gu, and Li 2003

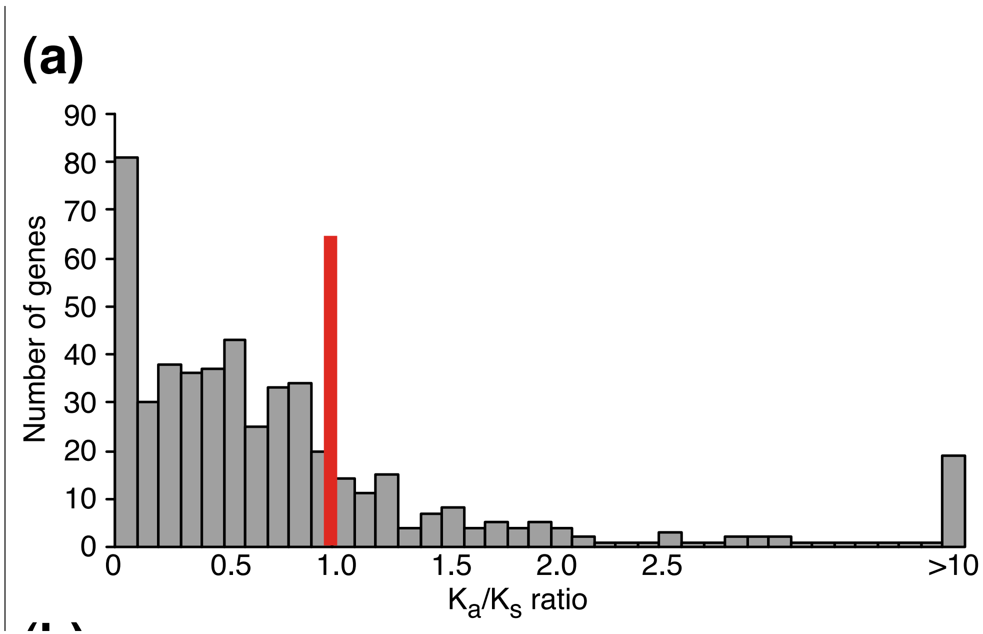


Lynch and Conery 2000

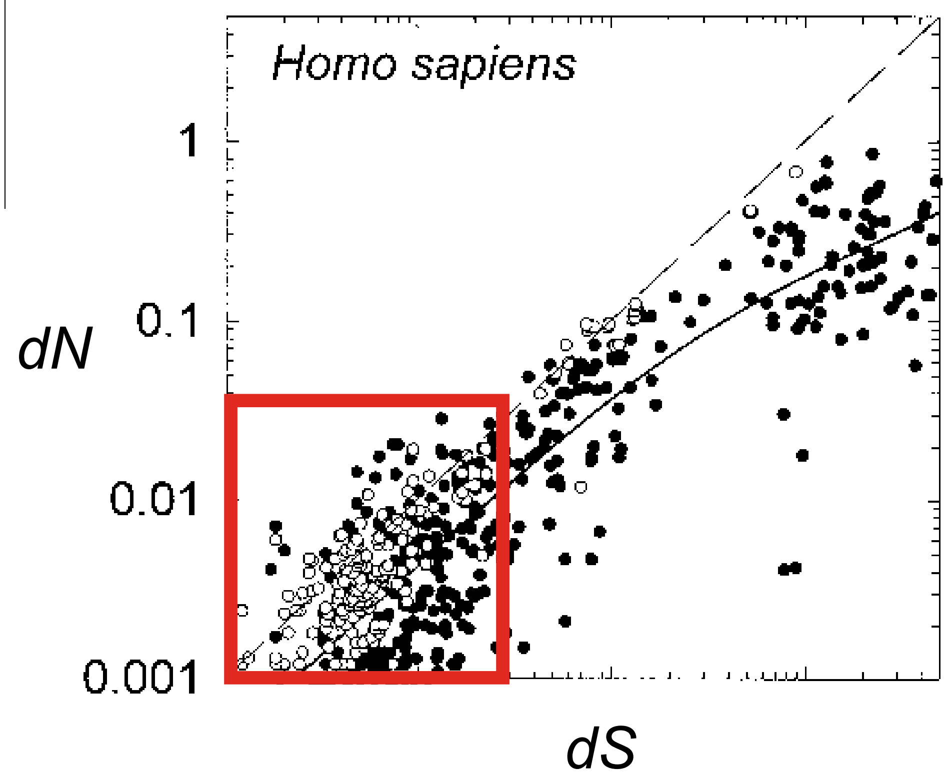


Why hasn't anyone seen this before?

Zhang, Gu, and Li 2003



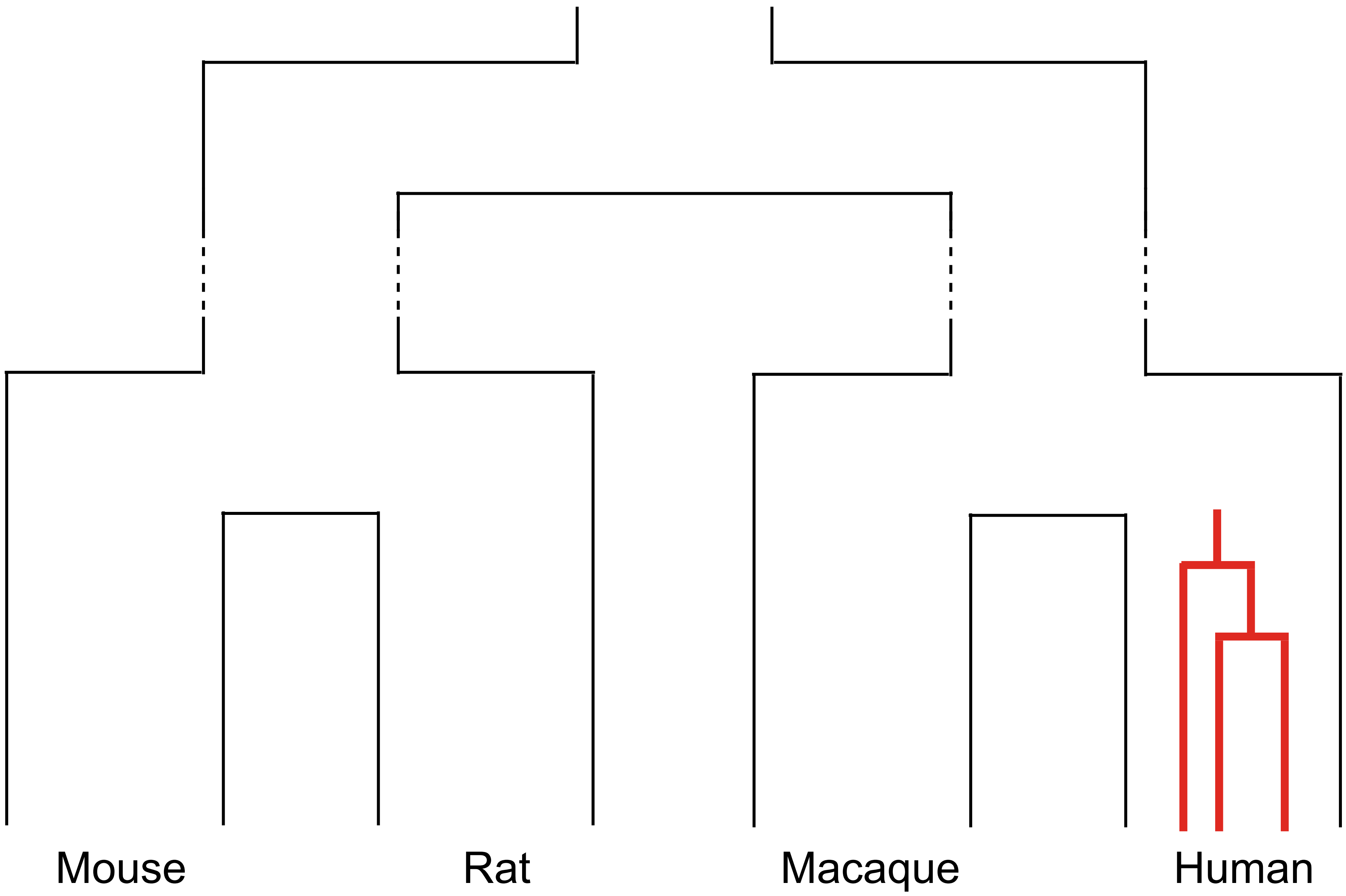
Lynch and Conery 2000

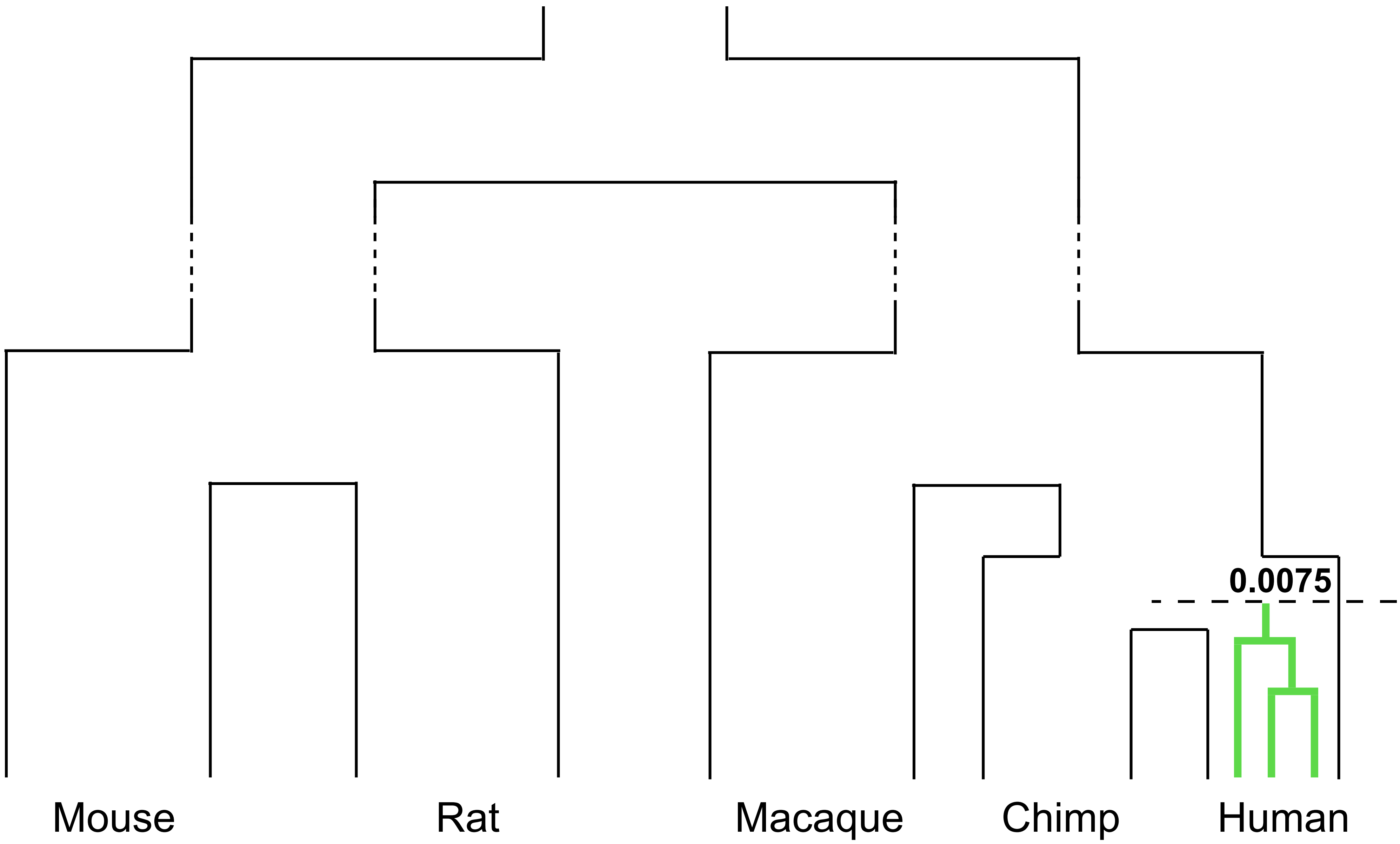


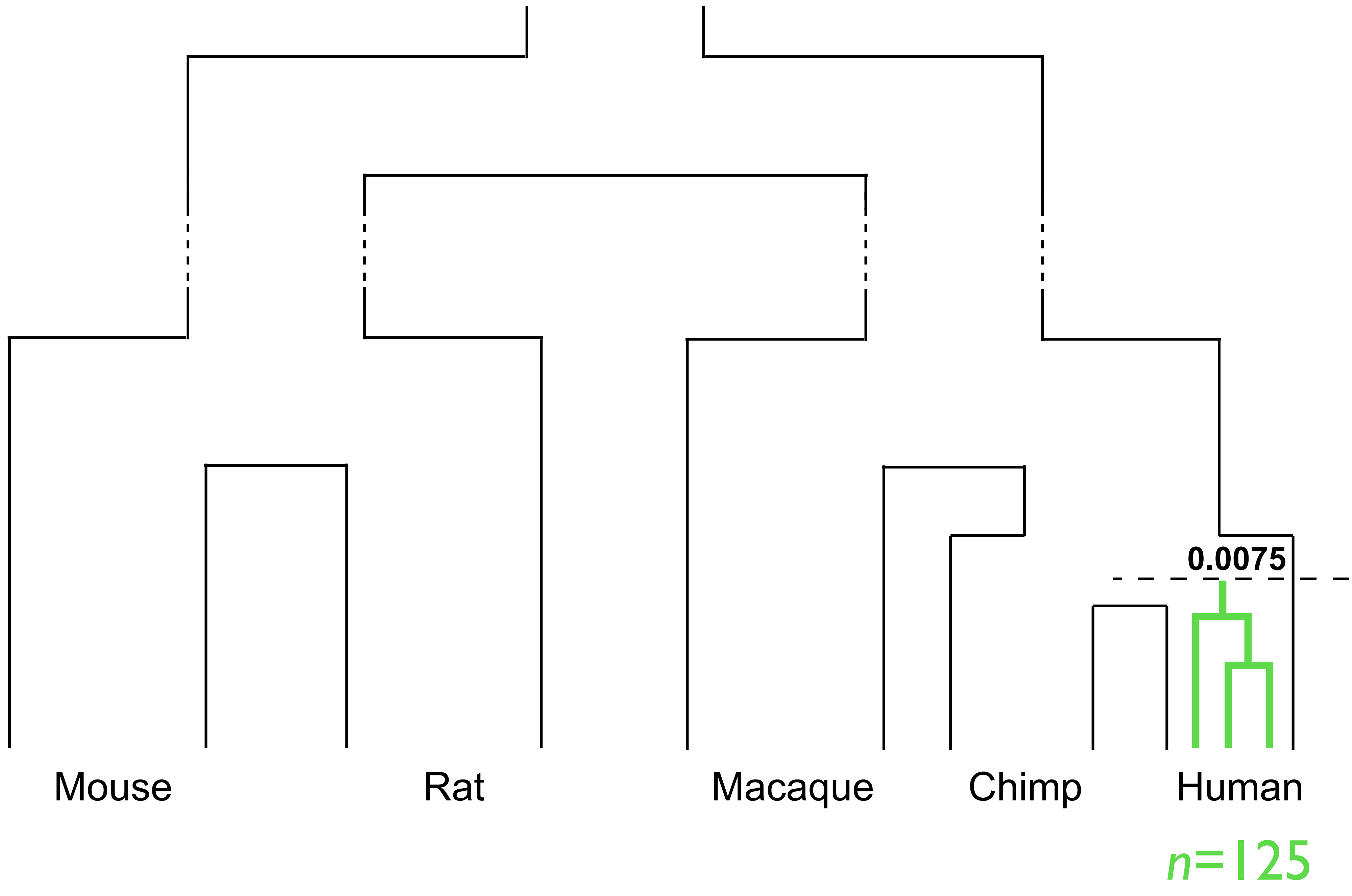
Genome scans for positive selection

Positive selection in humans:

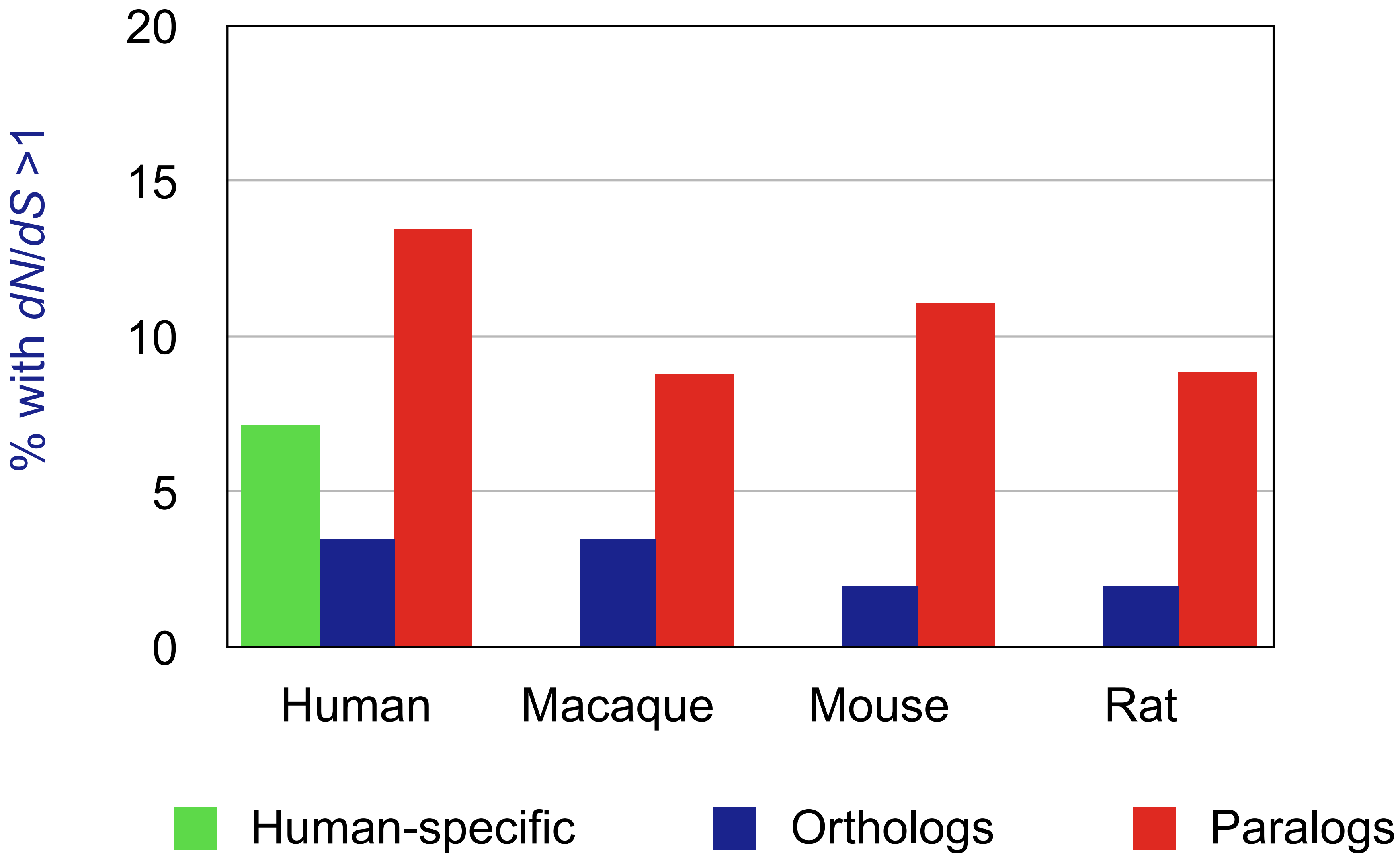
	P<0.05	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0
Bakewell et al. 2007 Gibbs et al. 2007	154/13,888 (1.1%)	2


Mouse
Rat
Macaque
Human


0.0075
Mouse
Rat
Macaque
Chimp
Human


0.0075
Mouse
Rat
Macaque
Chimp
Human
n=125

All duplicates + human-specific duplicates

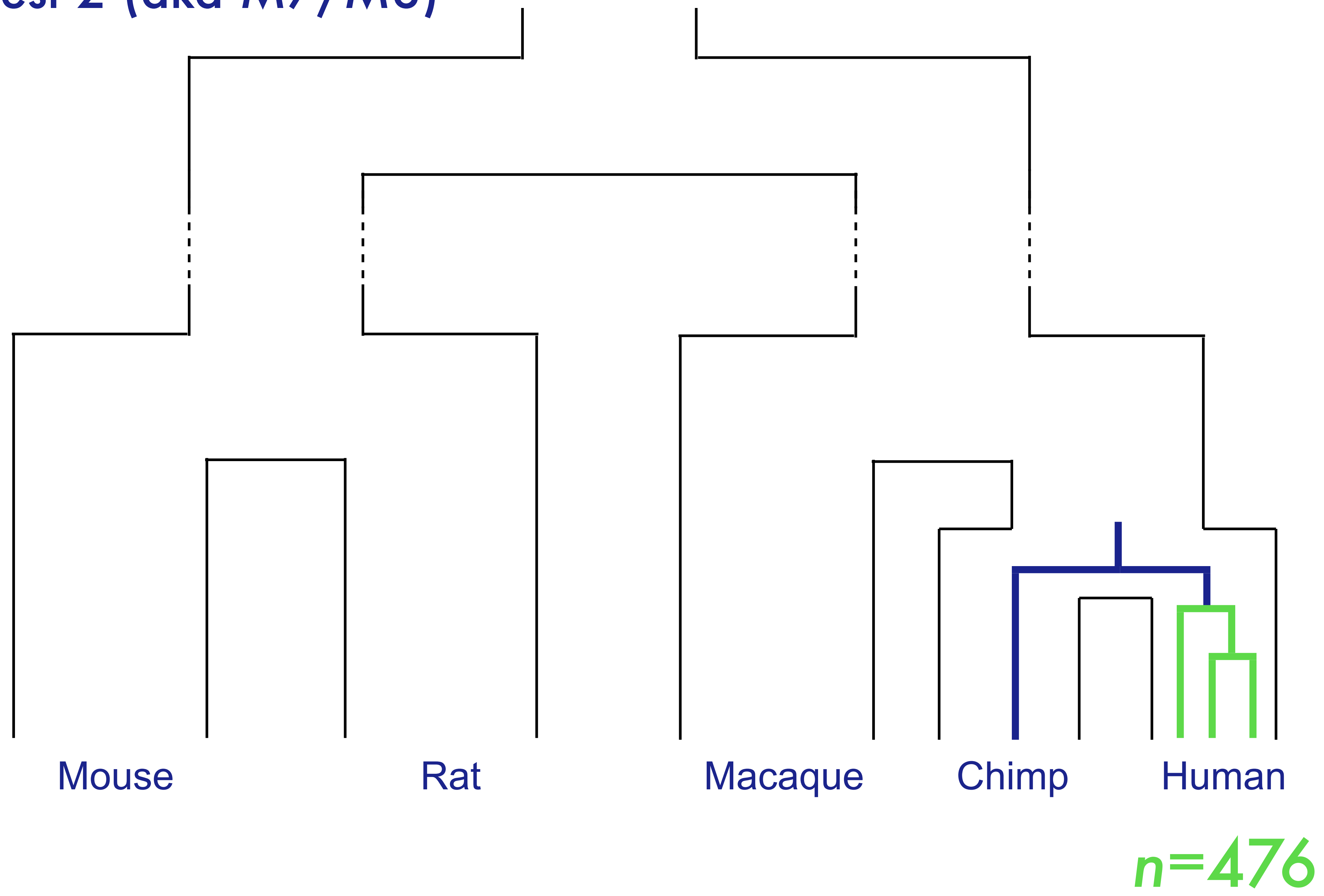


Genome scans for positive selection

Positive selection in humans:

	P<0.05	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0
Bakewell et al. 2007 Gibbs et al. 2007	154/13,888 (1.1%)	2
our data	9/125 (7.2%)	3

Test 2 (aka M7/M8)



Genome scans for positive selection

Positive selection in humans:

	P<0.05	FDR<0.05
Nielsen et al. 2005	35/13,653 (0.2%)	>0
Bakewell et al. 2007 Gibbs et al. 2007	154/13,888 (1.1%)	2
our data	17/476 (3.6%)	3

Interesting human genes

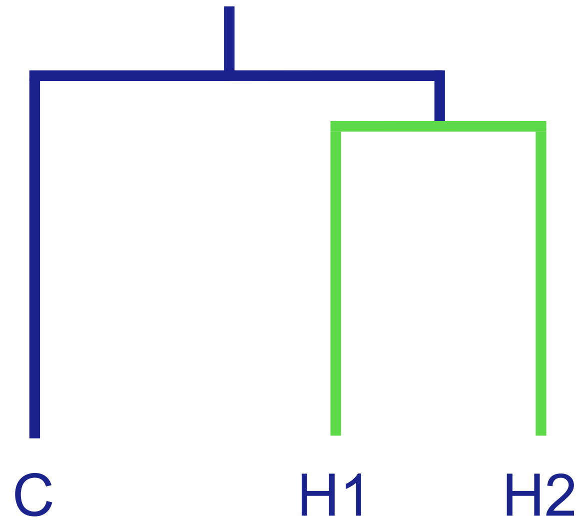
BZRP2

Interesting human genes

BZRP2

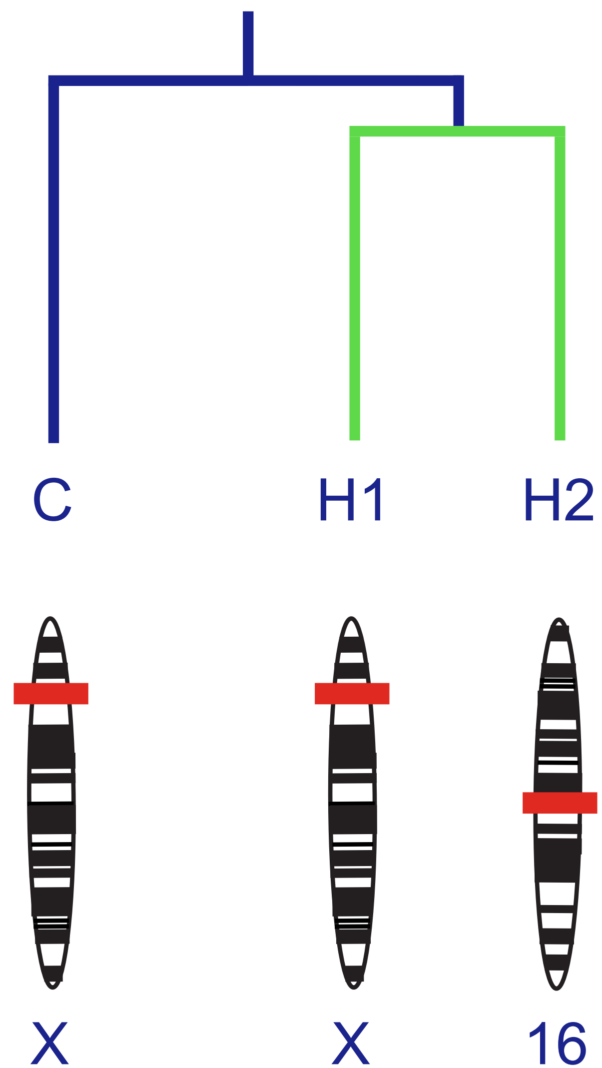
(Benzodiazepine receptor protein)

Polarizing duplicates



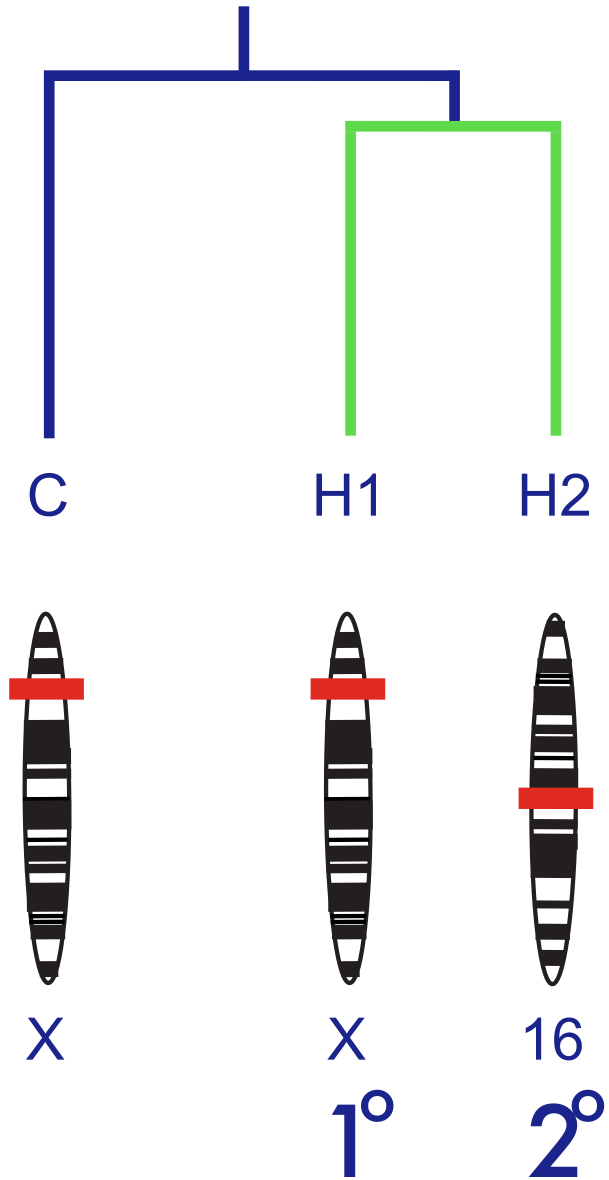
Han and Hahn (*in press*) *PSB*

Polarizing duplicates



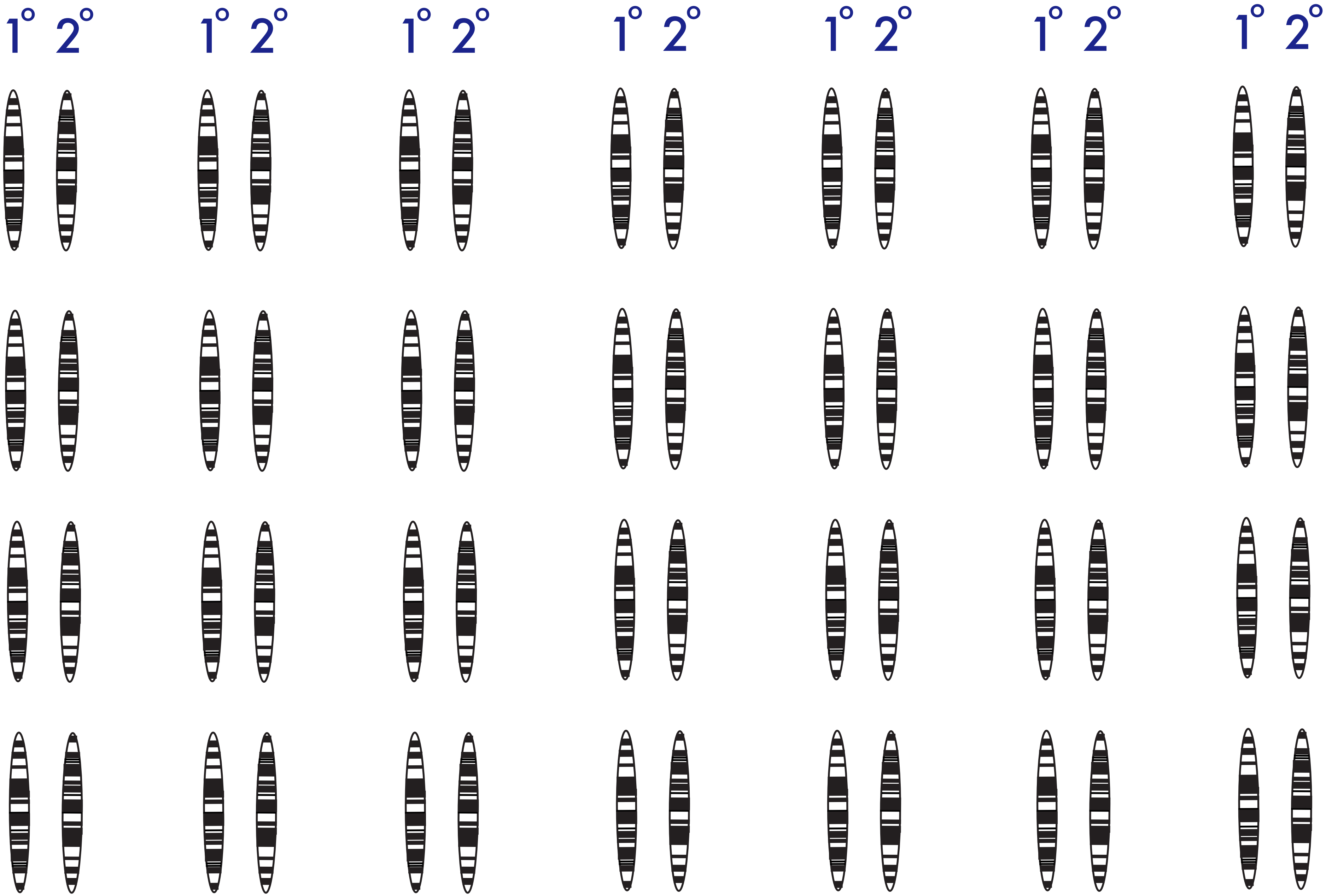
Han and Hahn (*in press*) *PSB*

Polarizing duplicates

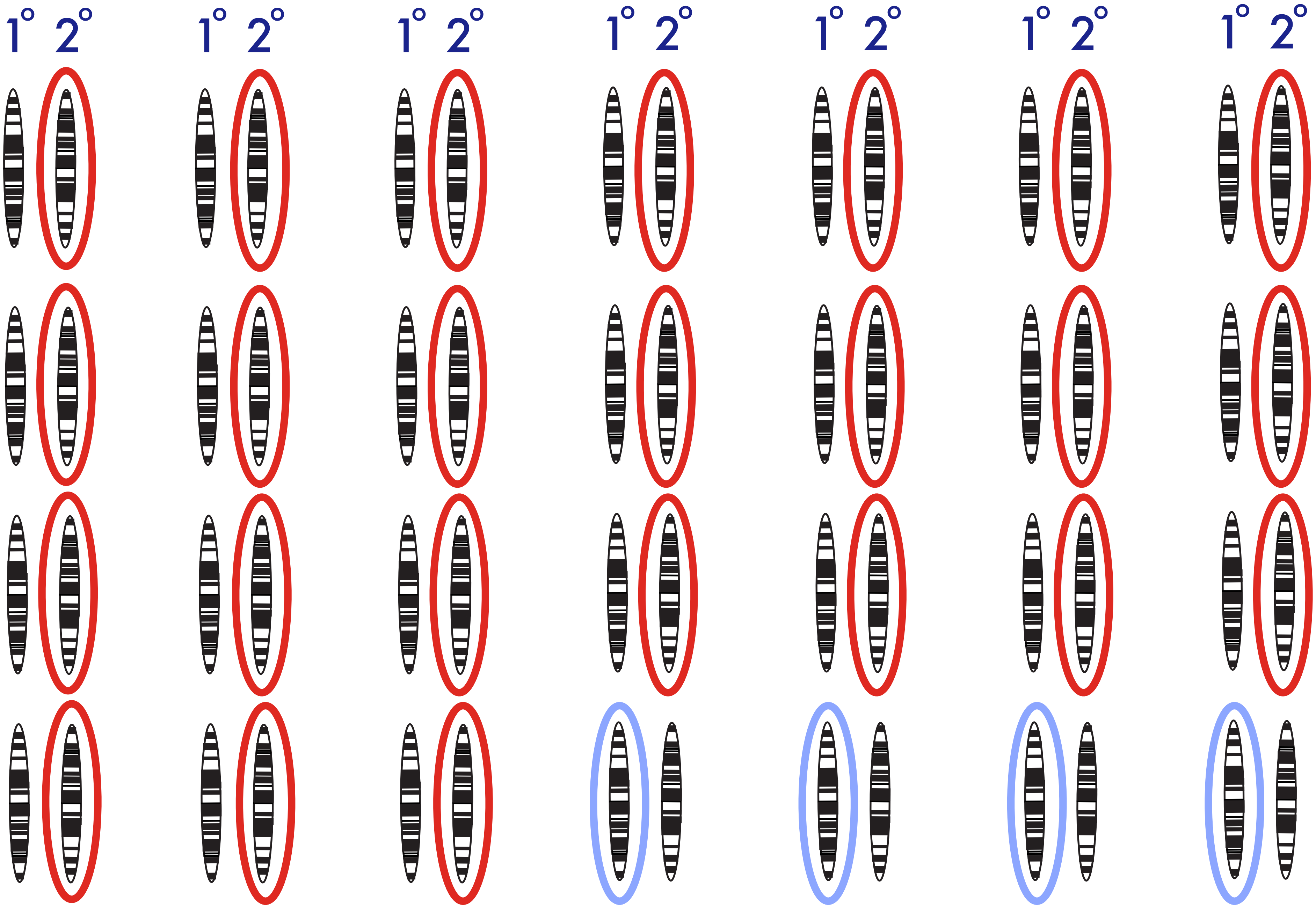


Han and Hahn (*in press*) *PSB*

Polarizing duplicates



Polarizing duplicates



Conclusions I

Conclusions I

Orthology, schmorthology

Conclusions II

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

Thanks

Jeff Demuth

Mira Han

Claudio Casola
Casey McGrath
Dan Schrider

Cross your fingers!