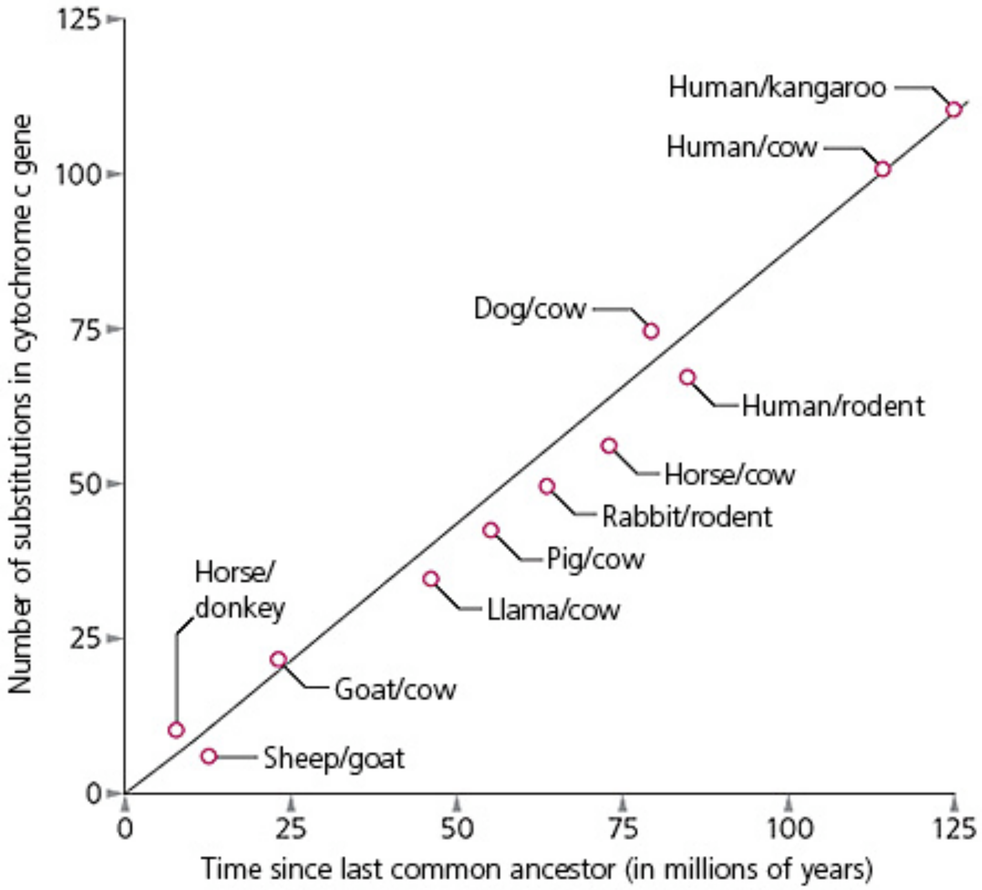
The Molecular Clock

Observation: Rate of **amino acid** substitution in many, but not all, proteins surprisingly constant over time, i.e. evolve in a clock-like manner.

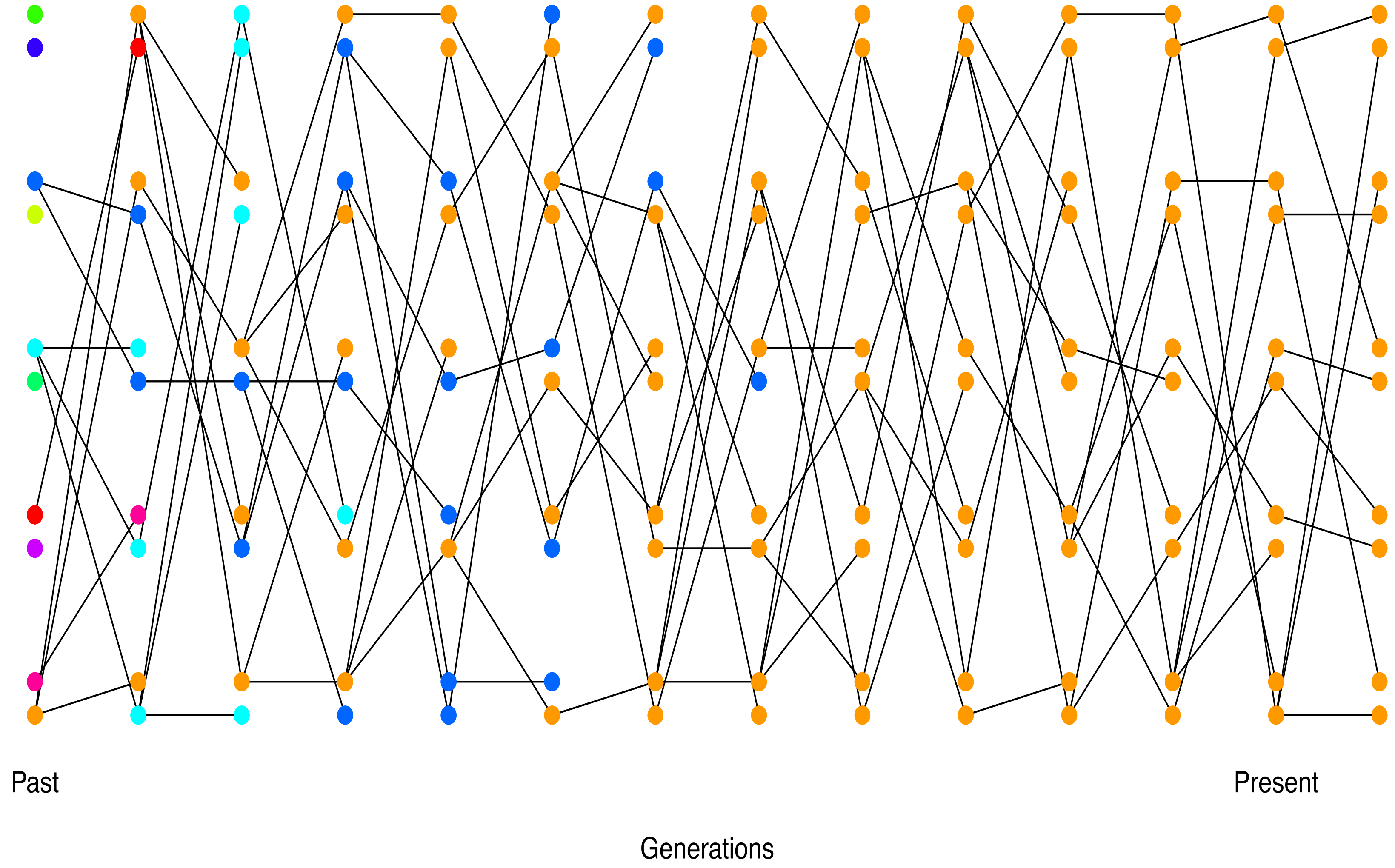
- Neutral theory Claim: Due to most amino acid replacement substitutions between species being neutral



From fossil record

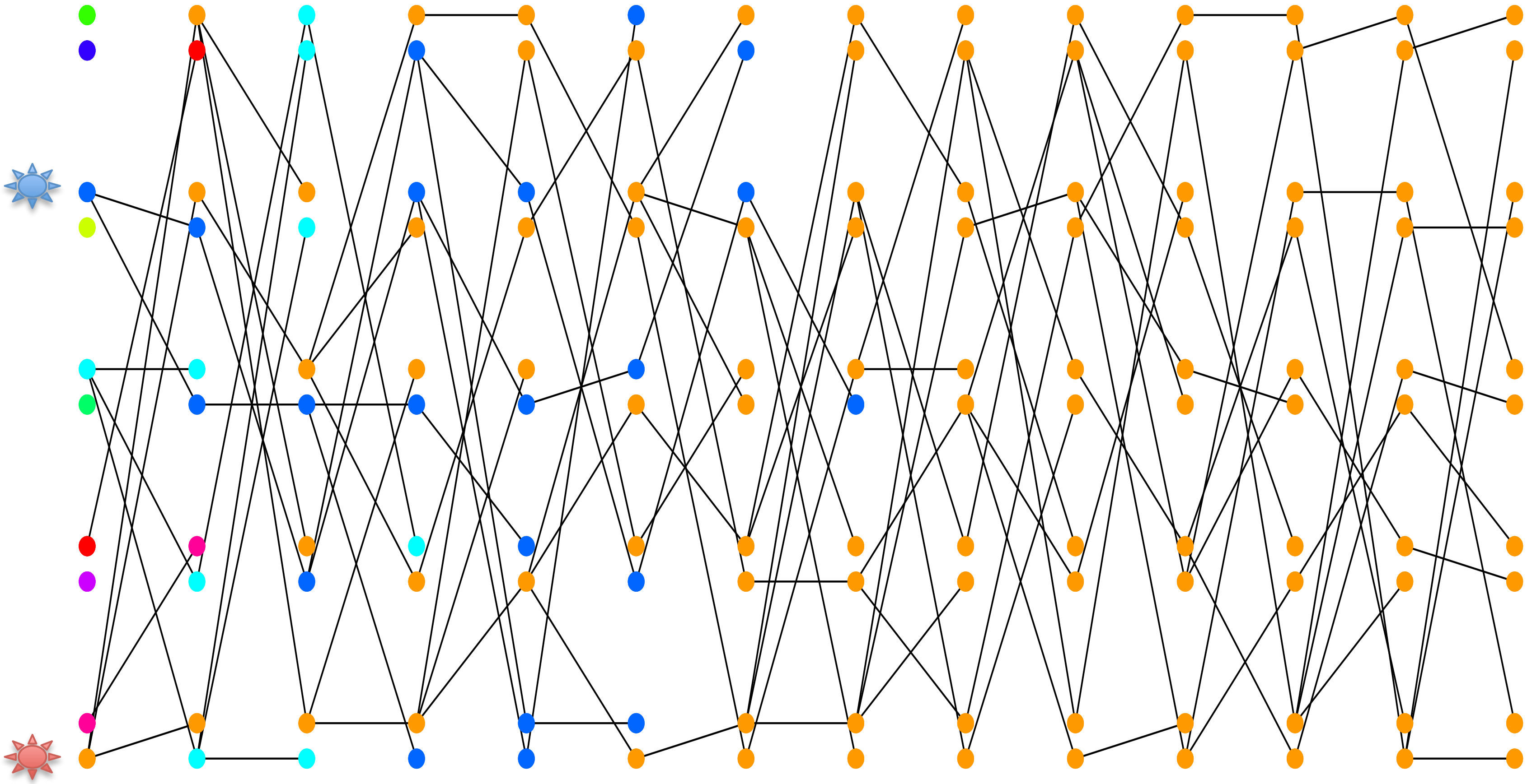
Zimmer book

Neutral evolution and a molecular clock



there are 2N alleles in our population

Neutral evolution and a molecular clock



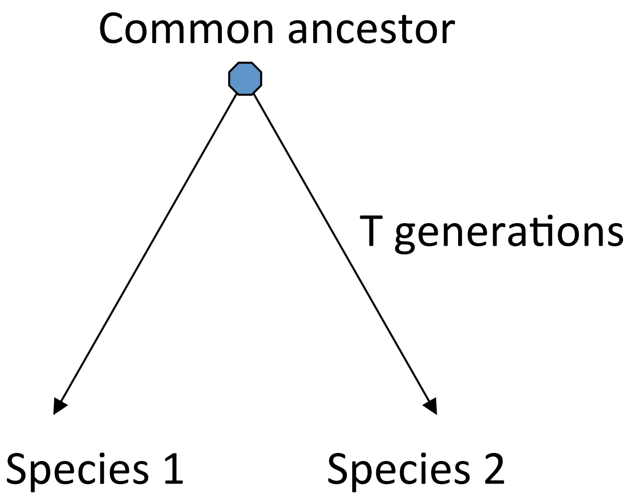
The probability that all individuals are descended from a particular neutral allele at a locus is 1/(2N)
Each generation there are 2Nμ new mutations per site

- Substitution rate per generation = 2Nμ x 1/[2N] = μ per generation

Independent of population size!

Neutral evolution and a molecular clock

- Substitution rate per generation = $2N\mu \times 1/[2N] = \mu$ per generation

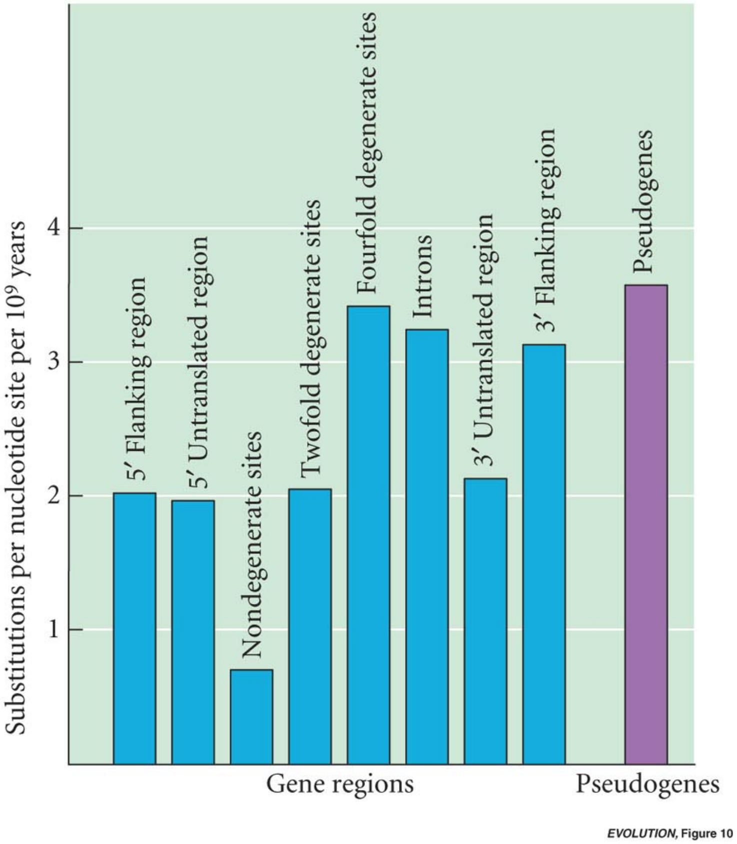


The expected number of neutral substitutions = $2T\mu$

i.e. substitutions occur at a linear rate, a molecular clock. Also gives a way to indirectly estimate mutations from divergence if T is known.

Levels of constraint

- Neutral theory claims most new mutations are deleterious and are lost immediately. Only neutral mutations contribute to substitution.

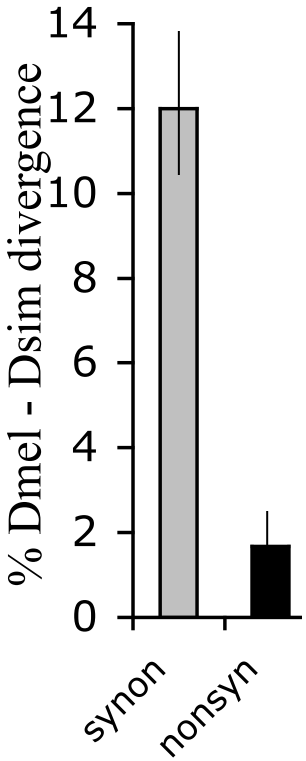


Fut. 10.14

Consistent with neutral theory slower rate of substitution at more constrained sites.

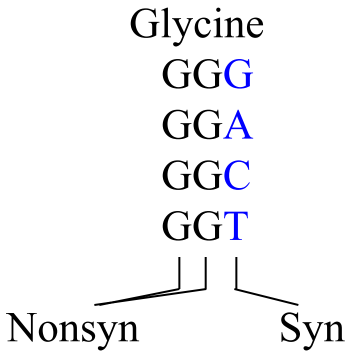
Variation in divergence (substitution rate) across classes of nucleotide sites (human vs. rodent)

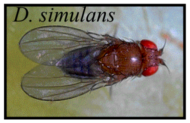
Selectively constrained (functional) sites evolve more slowly than less functional sites. (*C*=constraint)



Andolfatto Nature 2005

e.g. amino-acid coding sites versus synonymous sites

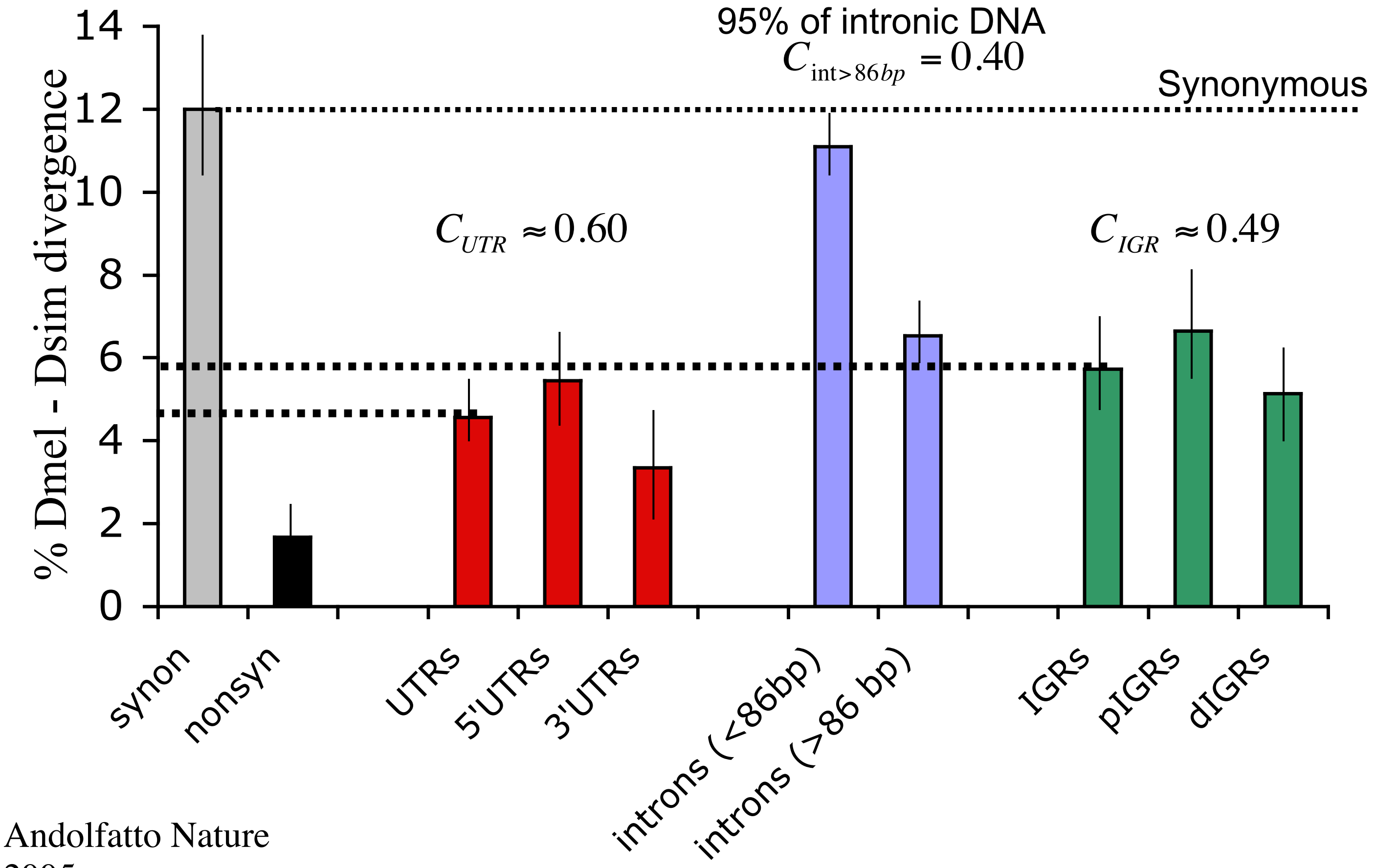

Glycine
GGG
GGA
GGC
GGT
Nonsyn
Syn

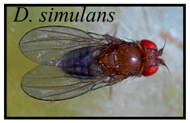


$$C_{NonSyn} = 1 - \frac{D_{obs}}{D_{\exp}} = 1 - \frac{D_{nonsyn}}{D_{syn}} = 0.865$$

Assuming all synonymous substitutions are neutral we estimate that 86.5% of mutations at nonsynonymous sites are deleterious enough to have been removed by selection

Most non-coding DNA evolves slower than synonymous sites in the *D.* melanogaster group



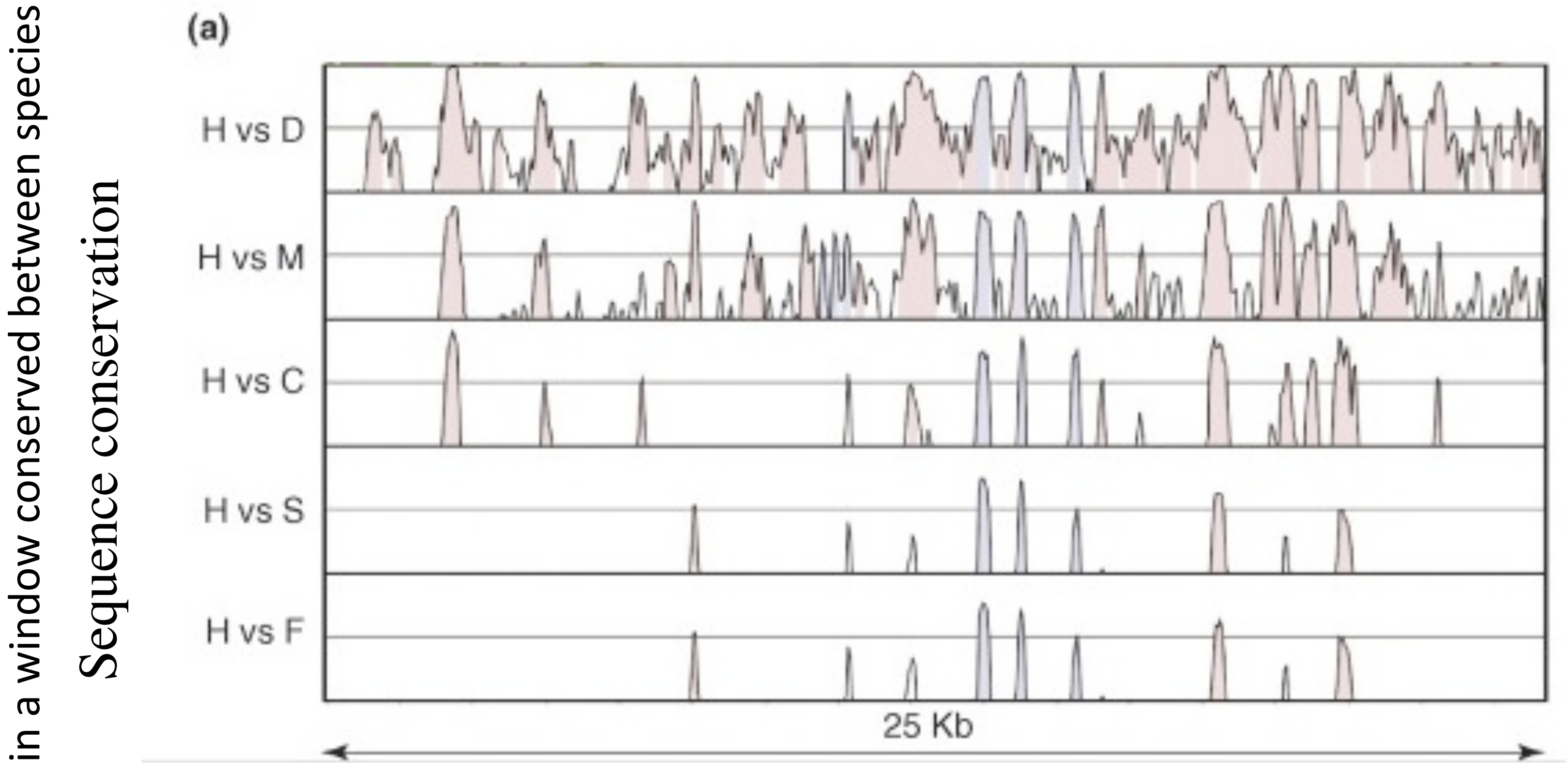


Andolfatto Nature
2005

Implication: 40-70% of non-coding sites are constrained by selection

Levels of constraint

Functional elements in genomes are often identified by their conservation over species.



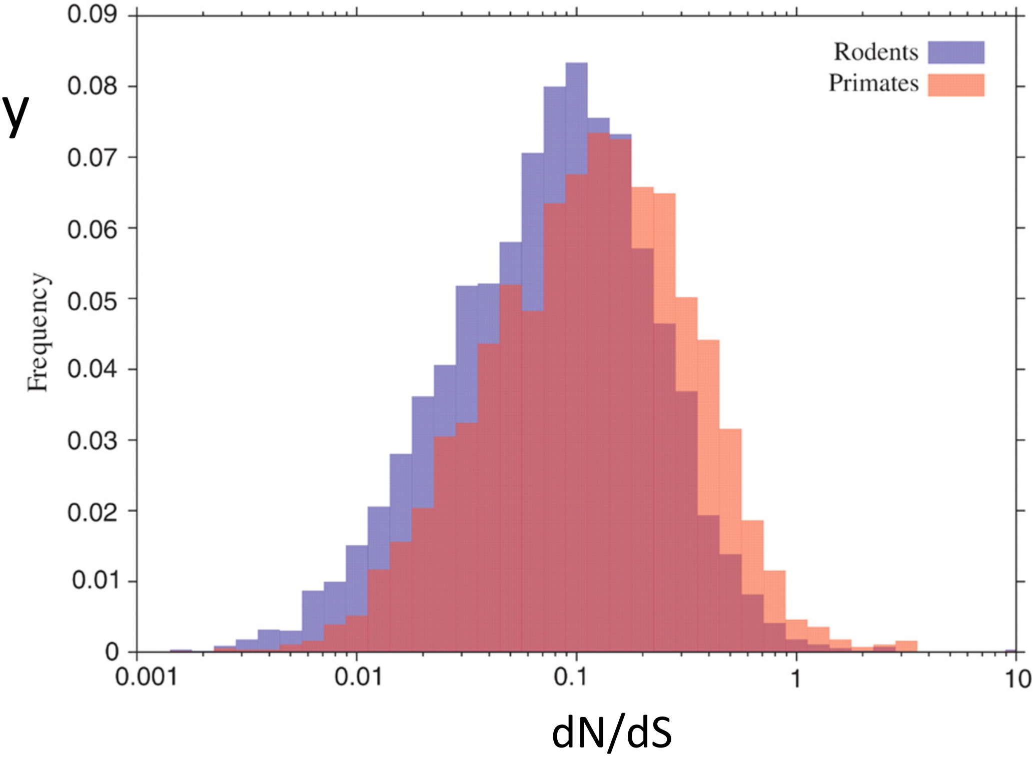
5’ region of Pax7 in humans vs dog, mouse, chicken, stickleback, fugu

Levels of constraint

- Neutral theory claims most new mutations are deleterious and are lost immediately. Only neutral mutations contribute to substitution.

- dN = nonsynonymous subst./site
- dS = synonymous subst./site
- Expected dN/dS for a "typical" gene. < 1

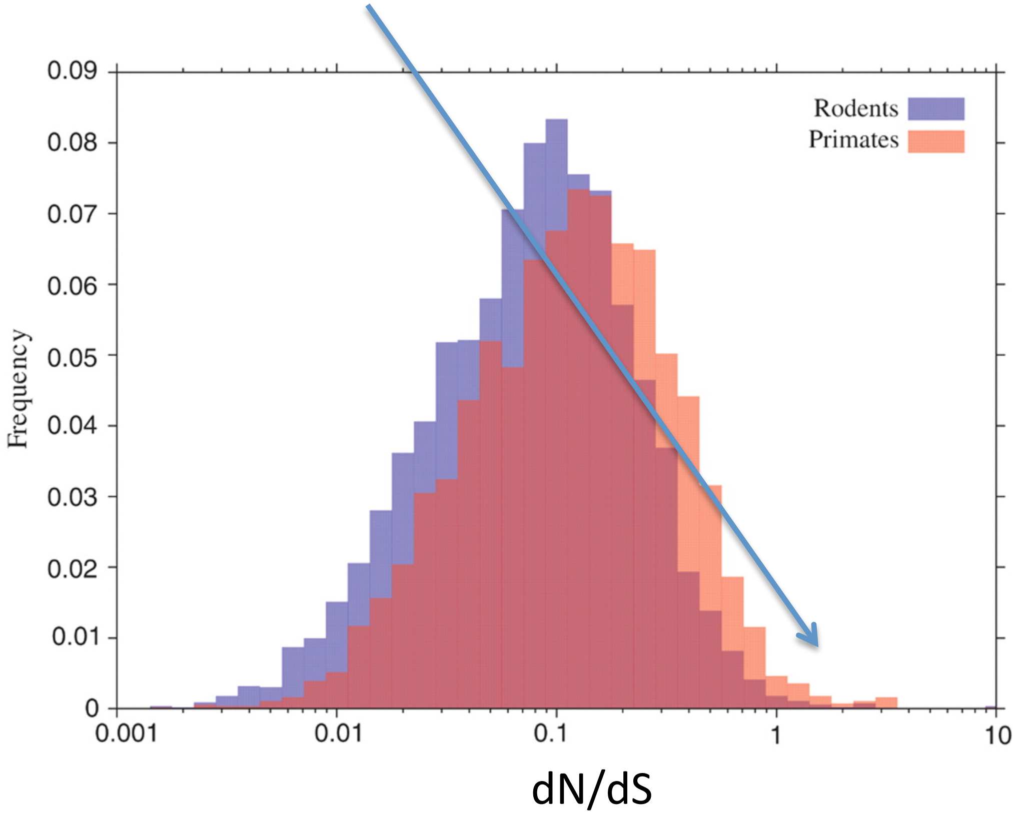
Consistent with neutral theory slower rate of substitution at more constrained sites.



Adaptive protein divergence and the dN/dS ratio

- dN = nonsynonymous subst./site
- dS = synonymous subst./site
- Expected dN/dS for a “typical” gene. < 1

- Expected dN/dS for a gene coding for an unconstrained protein = 1
- Recurrent directional selection, dN/dS > 1 *



* This is a very conservative test.

An example

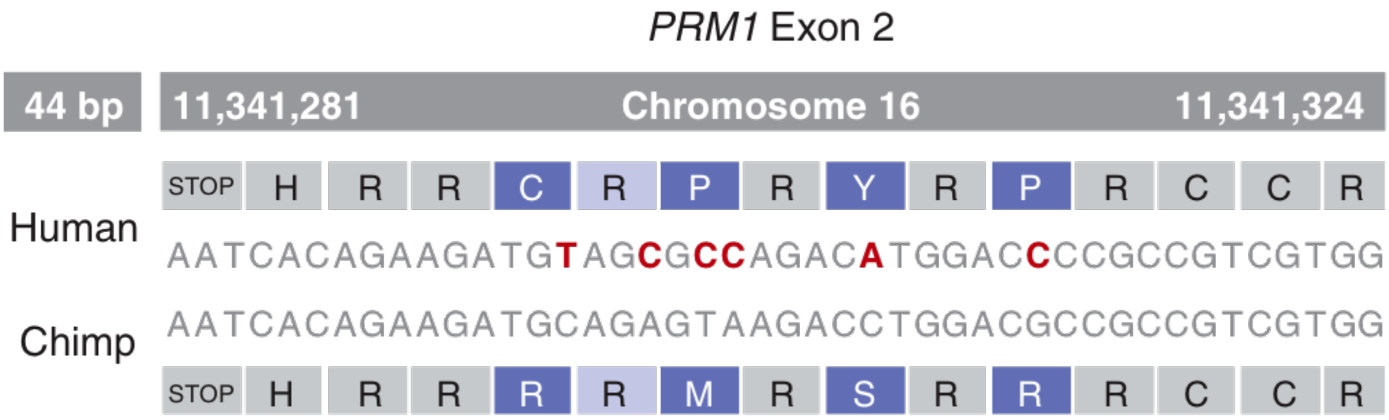
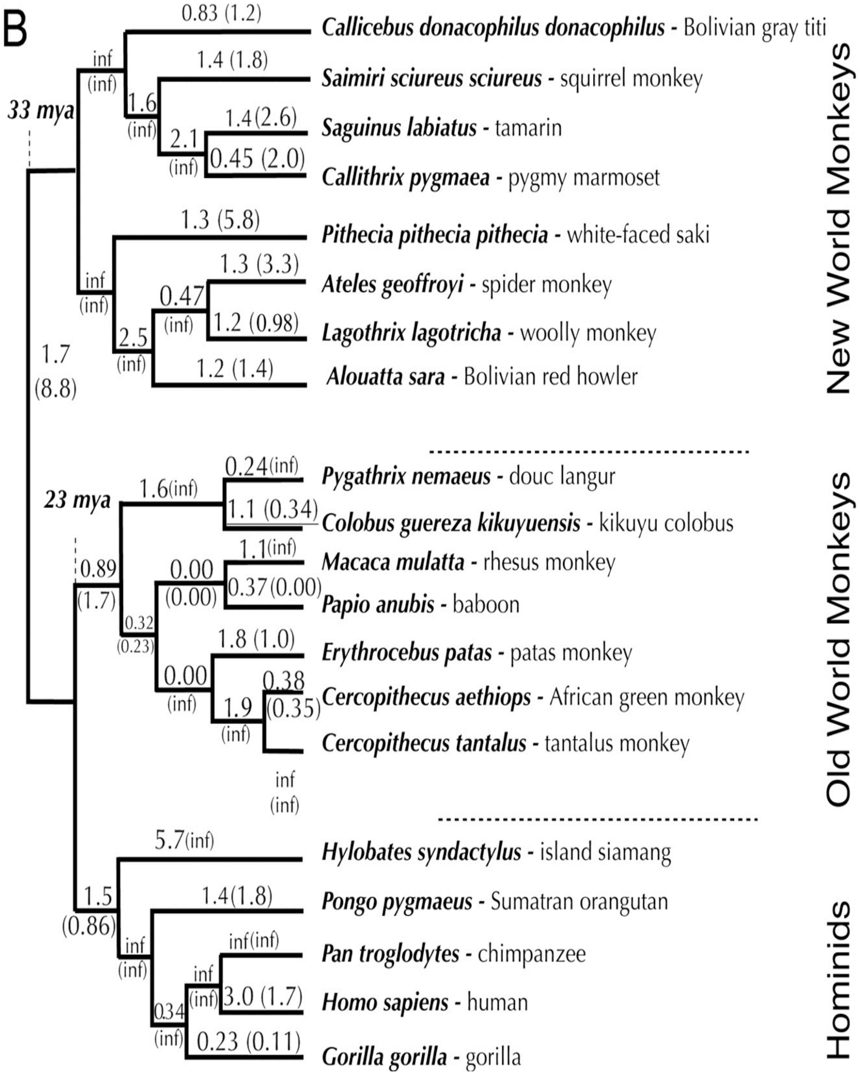


Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

Sabeti et al 2007

TRIM5α gene involved in primate retrovirus defense.
dN/dS> 1 for the majority of branches in phylogeny



Sawyer et al 2005

Examples of genes showing dN/dS > 1

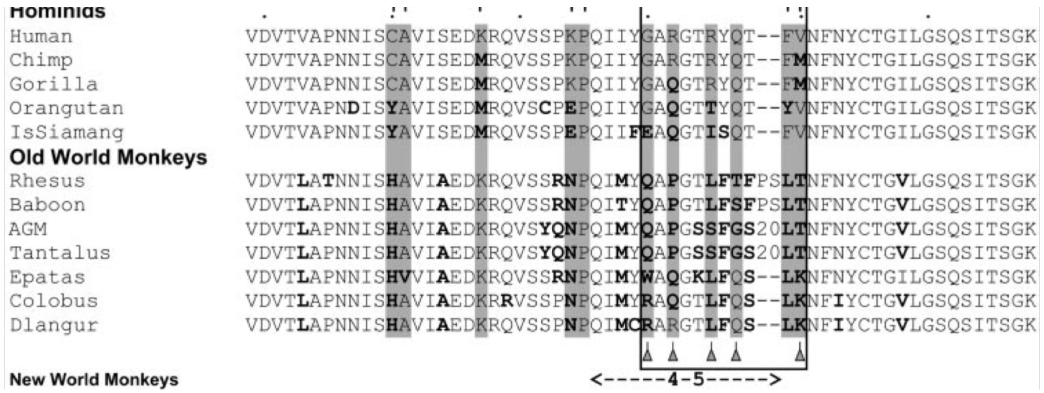
--Vertebrate immune system

--Viral coat proteins in viruses

--Reproduction related genes

Invertebrate gamete recognition proteins

Vertebrate gametogenesis



```
Hominids
Human       VDVTVAPNNISCAVISEDKRQVSSPKPQIIYGARGTRYQT--FVNFNYCTGILGSQSITSGK
Chimp       VDVTVAPNNISCAVISEDMRQVSSPKPQIIYGARGTRYQT--FMNFNYCTGILGSQSITSGK
Gorilla     VDVTVAPNNISCAVISEDMRQVSSPKPQIIYGAQGTRYQT--FMNFNYCTGILGSQSITSGK
Orangutan   VDVTVAPNDISYAVISEDMRQVSCPEPQIIYGAQGTTYQT--YVNFNYCTGILGSQSITSGK
IsSiamang   VDVTVAPNNISYAVISEDMRQVSSPEPQIIFEAQGTISQT--FVNFNYCTGILGSQSITSGK
Old World Monkeys
Rhesus      VDVTLATNNISHAVIAEDKRQVSSRNPQIMYQAPGTLFTFPSLTNFNYCTGVLGSQSITSGK
Baboon      VDVTLAPNNISHAVIAEDKRQVSSRNPQITYQAPGTLFSFPSLTNFNYCTGVLGSQSITSGK
AGM         VDVTLAPNNISHAVIAEDKRQVSYQNPQIMYQAPGSSFGS20LTNFNYCTGVLGSQSITSGK
Tantalus    VDVTLAPNNISHAVIAEDKRQVSYQNPQIMYQAPGSSFGS20LTNFNYCTGVLGSQSITSGK
Epatas      VDVTLAPNNISHVVIAEDKRQVSSRNPQIMYWAQGKLFQS--LKNFNYCTGILGSQSITSGK
Colobus     VDVTLAPNNISHAVIAEDKRRVSSPNPQIMYRAQGTLFQS--LKNFIYCTGVLGSQSITSGK
Dlangur     VDVTLAPNNISHAVIAEDKRQVSSPNPQIMCRARGTLFQS--LKNFIYCTGVLGSQSITSGK

New World Monkeys
                                      <-----4-5----->
```

Logic of the MK test

	Syn.	Replace.	R/S ratio
Fixed	a	b	b/a
Polymorphic	c	d	d/c

Expectation of polymorphic and fixed variants under neutrality
Proportional to

	Syn.	Replace.	R/S Ratio
Fixed	u_S	u_R	u_R/u_S
Polymorphic	$4Nu_S$	$4Nu_R$	u_R/u_S

Thus under neutrality: $b/a = d/c$
Can test as a 2 x 2 table
Under a model where a proportion of replacements experience:
directional selection or are mildly deleterious?

MK test

An example of a dataset compatible with neutral protein variation

	Syn.	Replace.	R/S Ratio
Fixed	10	5	0.5
Polymorphic	20	10	0.5

An example of a dataset incompatible with neutral protein variation

Polymorphism and divergence in *a Nuclearporin gene* (*Nup96*) in
D. melanogaster and *D. simulans*

	Syn.	Replace.	R/S Ratio
Fixed	34	27	0.794
Polymorphic	108	27	0.250

P-value by G-test
=0.0006

Presgraves and Stephan MBE

What fraction of protein divergence results from directional selection?

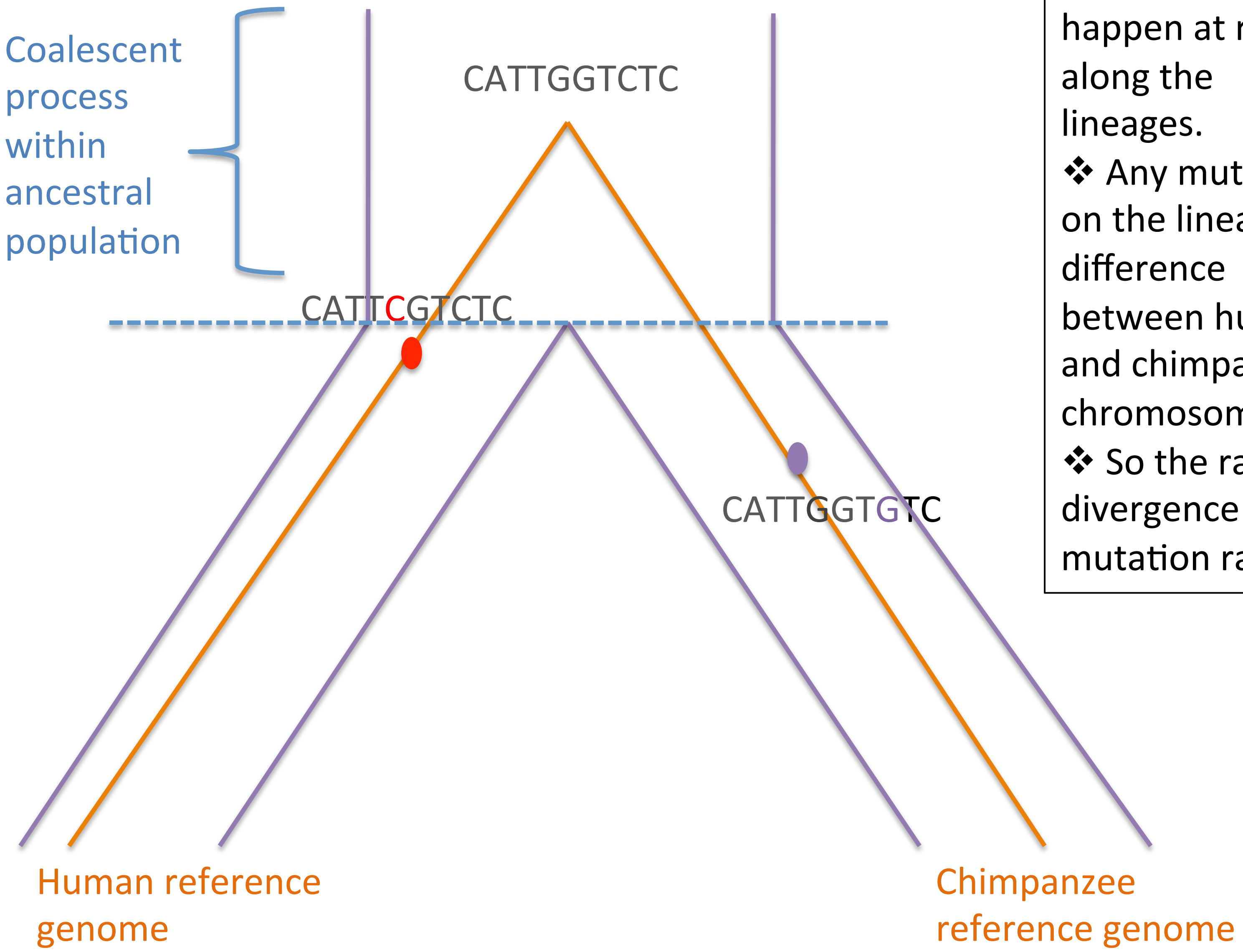
- From *D. simulans* population genomics (Begun *et al*. 2007)
    - 20% of genes show evidence of adaptive protein evolution by MK test
    - 30-50% of amino acid fixations due to directional selection

- These approaches applied to non-coding DNA suggest adaptive evolution of non-coding DNA is also common in flies

- ~50 % of amino-acid substitutions due to directional selection in between mouse subspecies (Halligan et al. 2010)

- ~10 % of amino-acid substitutions due to directional selection in the human lineage (Boyko et al 2009)

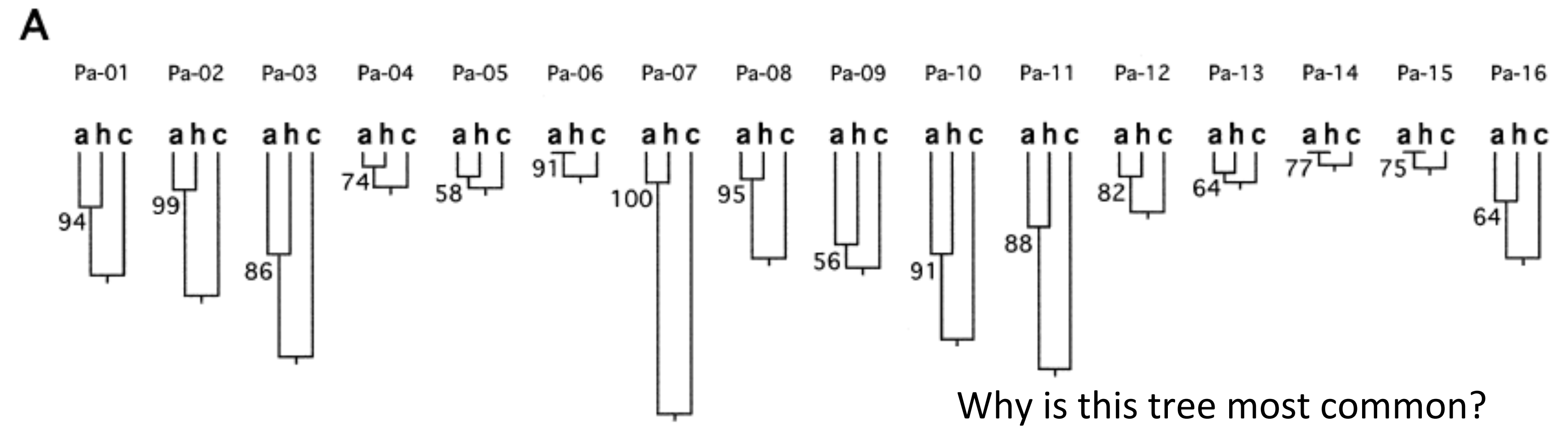
At neutral sites, the divergence rate = the mutation rate



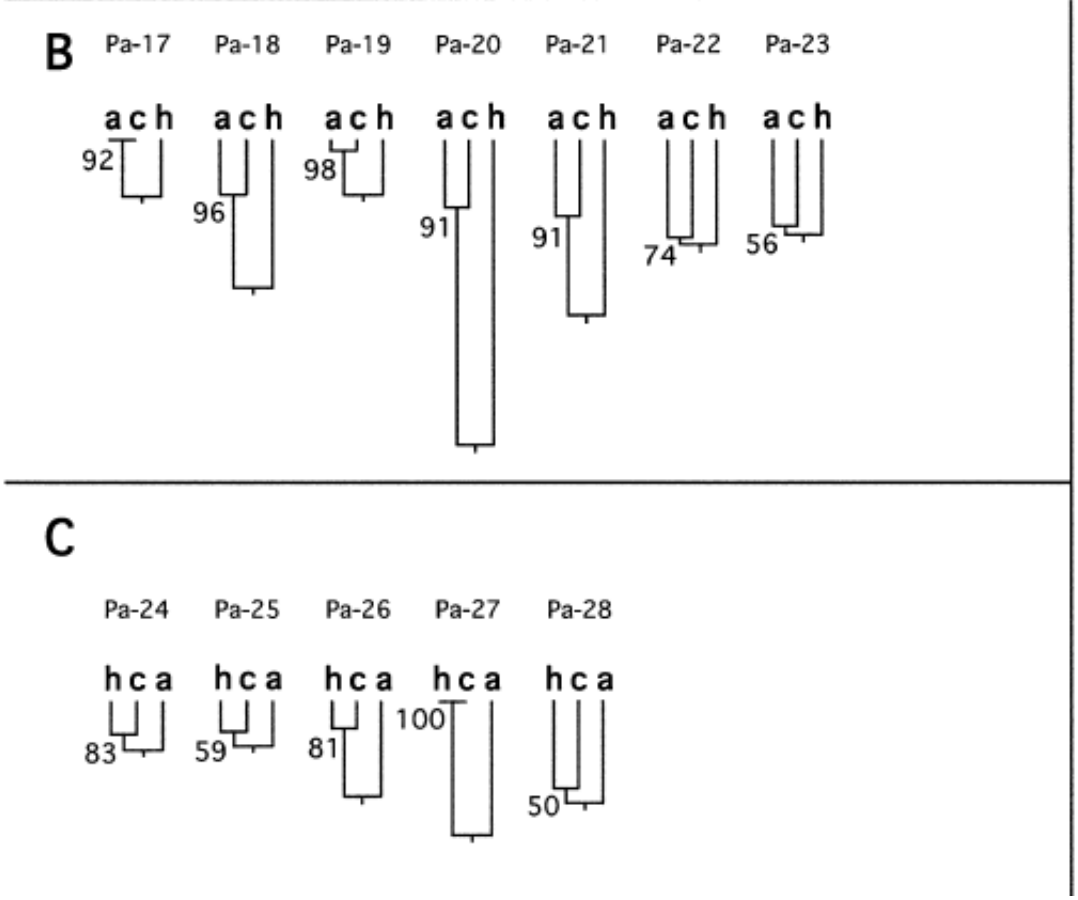
❖ Mutations happen at rate μ along the lineages.

❖ Any mutation on the lineage is a difference between human and chimpanzee chromosomes.

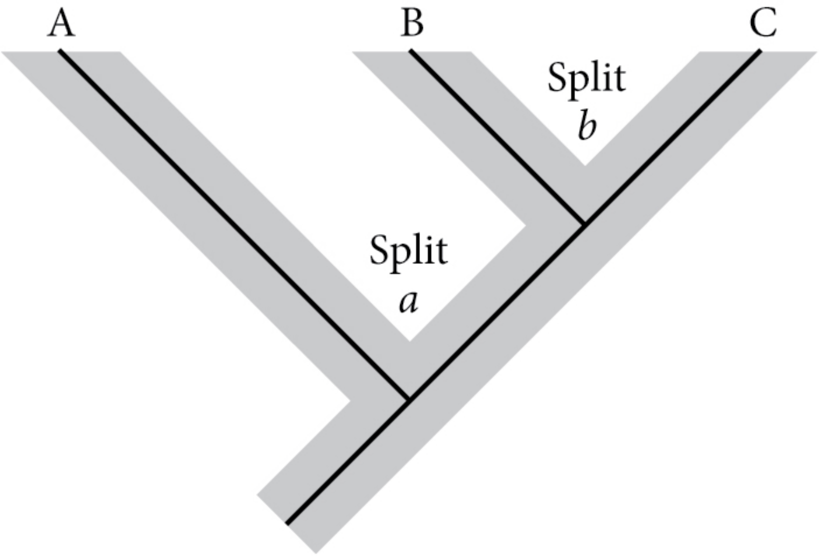
❖ So the rate of divergence = the mutation rate μ.

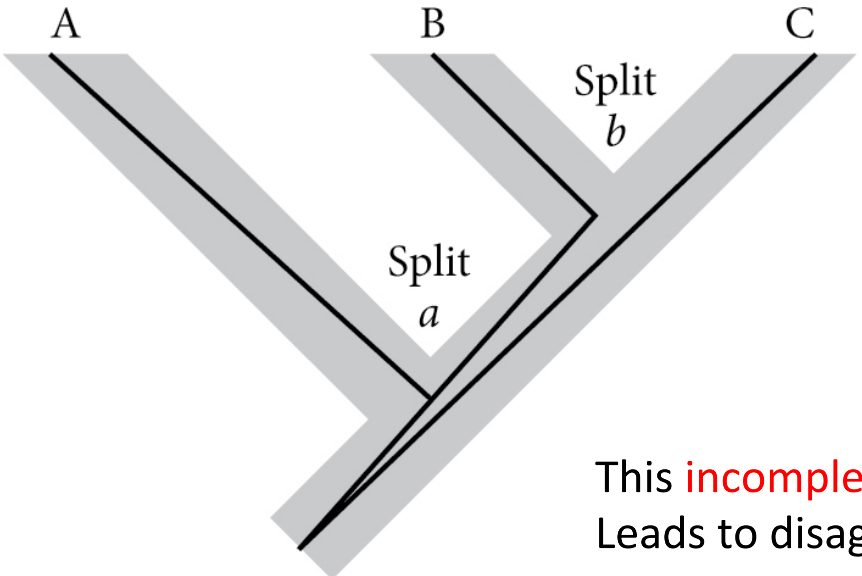


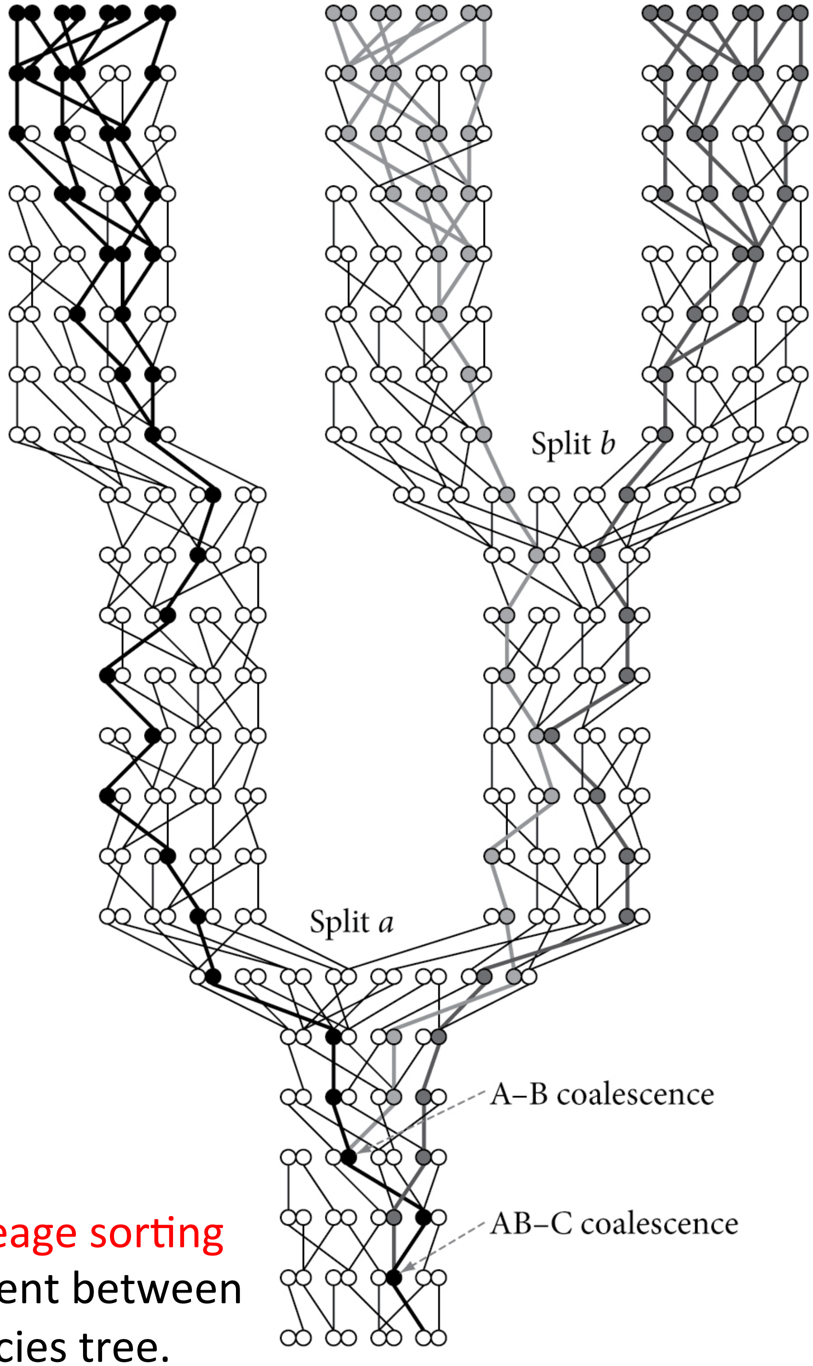
Why is this tree most common?



28 gene trees for 3
AUSTRALIAN
GRASS FINCHES (POEPHILA)
JENNINGS & EDWARDS 2005

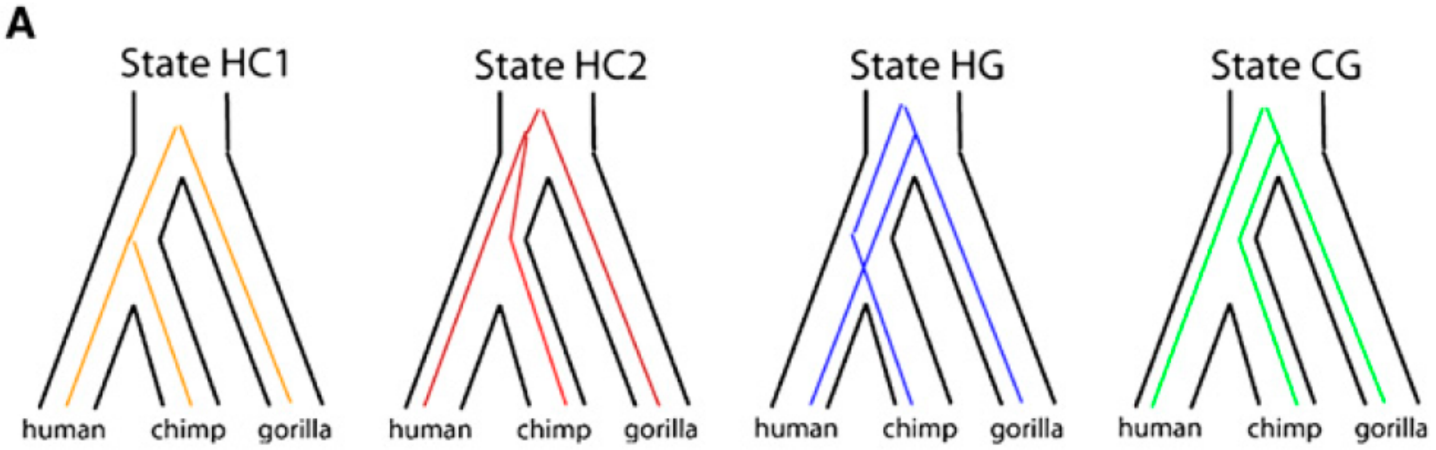






This incomplete lineage sorting
Leads to disagreement between
Gene trees and species tree.

Human-Chimp-Gorilla incomplete lineage sorting



Species tree:
Uninformative

Agrees

Disagrees