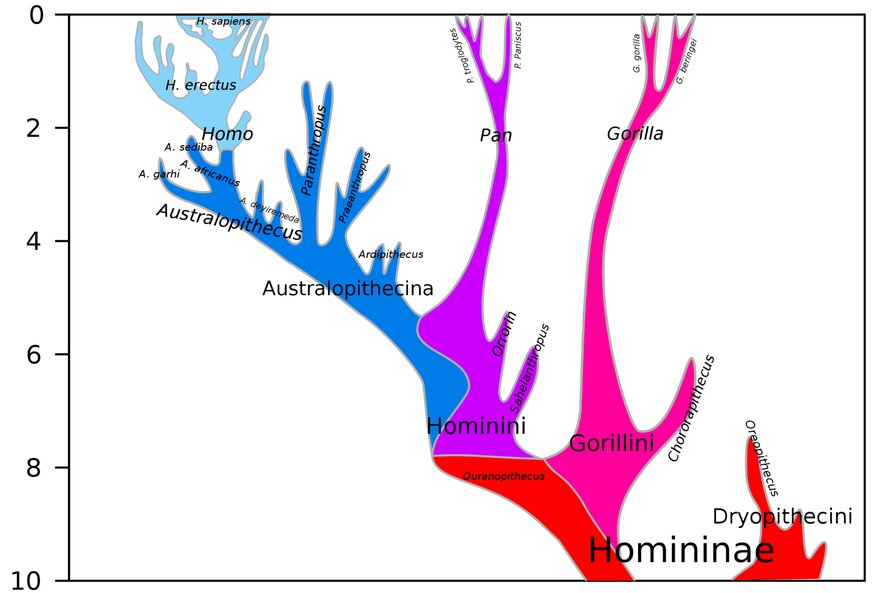
Lecture 7: Divergence and Molecular Substitution

Population genetic PCB4553/6685

Long term molecular evolution is just one substitution after another

- 30 million base pair *substitutions* between human and chimpanzees, at orthologous locations.
- Occurred since human and chimp last shared a common ancestor (7 MYA).
- Long-term evolution, from the molecular perspective, is just one substitution after another.
- Each of the substitutions must have arisen as a mutation in the population, spread through the population as a polymorphism before eventually reaching fixation.



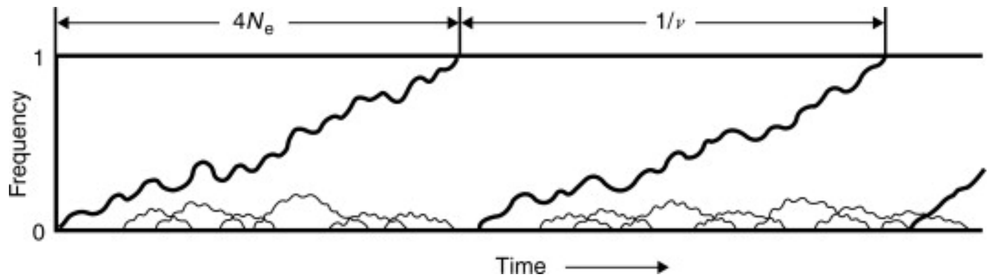
Human	accacagcatttgttagttactgccaagaagcctgtatctgtagggtaaaatcctcgctgaagtgggttg
Chimp	.....g.....c.....
Gorilla	.....cc.....
Orangutan	.....c.....c.....c.....
Gibbon	.....c.....---.....
Crab-eating macaque	g.....gg...c.....c..t.t.....

- Variable positions in a primate alignment of orthologous sequences of a 136bp region.

The time scale

- Human Ne = 10000
- Generation time 20 years?
- Mean time to MRCA 800,000
- Divergence time between human and chimp: 7,000,000
- Separation of time scale
- The existing polymorphisms are transient and will be fixed or lost to drift
- In the meantime new mutations will occur

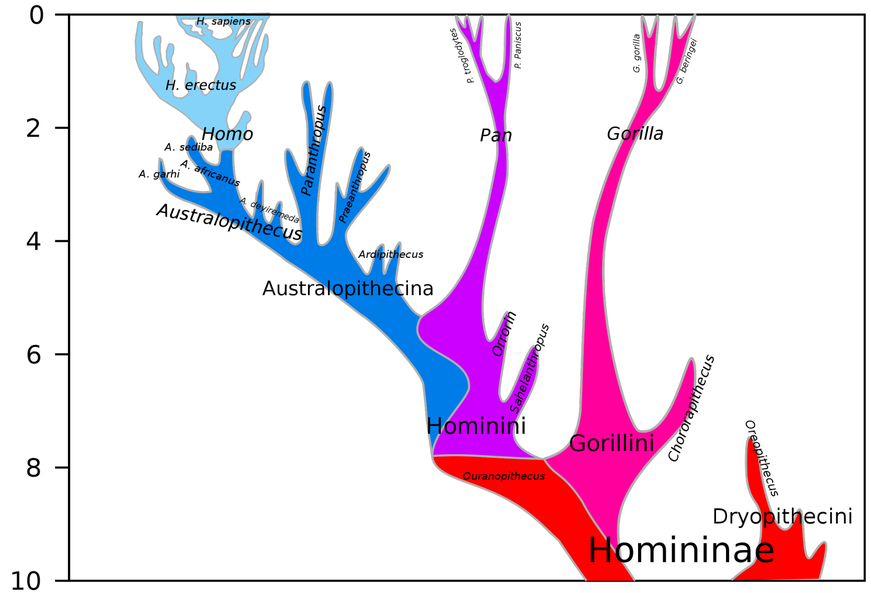
- Polymorphism is simply a phase of substitution



Ohta 2001

Long term molecular evolution is just one substitution after another

- What forces drove the spread of these alleles through the population to become substitutions?
- Natural selection may only affect a small percentage of all substitutions
- Today we'll focus on *neutral* substitutions
- And learn to identify potential adaptive substitutions



Human	accacagcatttgttagttactgccaagaagcctgtatctgtagggtaaaatcctcgctgaagtgggttg
Chimp	.....g.....c.....
Gorilla	.....cc.....
Orangutan	.....c.....c.....c.....
Gibbon	.....c.....---.....
Crab-eating macaque	g.....gg...c.....c..t.t.....

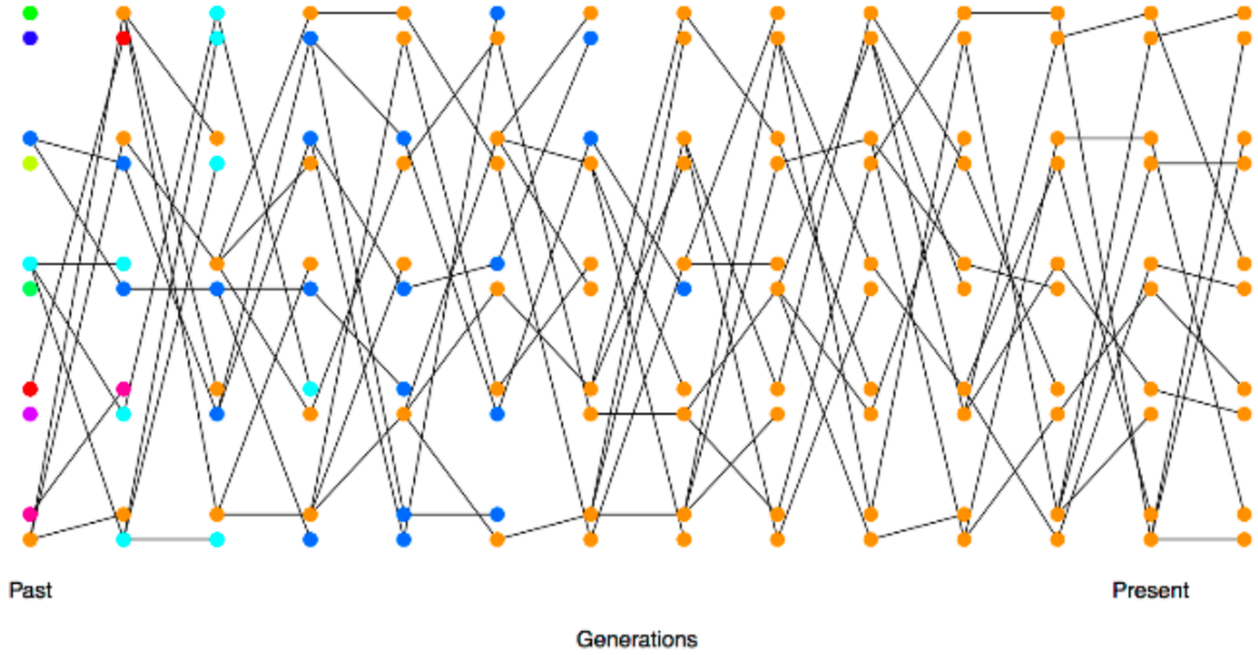
- Variable positions in a primate alignment of orthologous sequences of a 136bp region.

The Neutral Substitution process

- How do neutral substitutions occur?
  - Rare neutral allele usually becomes lost from the population and rarely drifts to fixation
  - But populations experience a large and constant influx of rare alleles due to mutation
  - Some neutral alleles will fix by chance
  - What is the probability that a neutral mutation fixes?
  - How does neutral substitutions accumulate over time?

Probability of the eventual fixation of a neutral allele

- What's the probability that a mutation drift to fixation?
- Under neutrality, simply
- $1/2N$
- Argument: every allele in the pop has equal chance of fixation

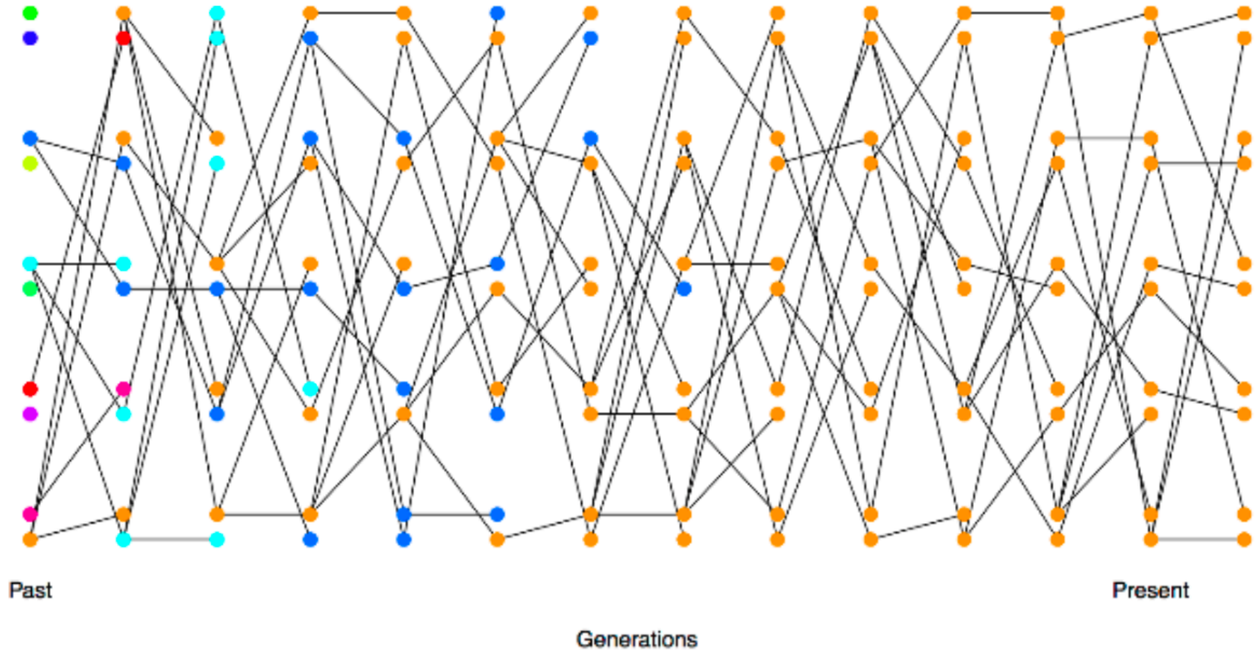


Probability of the eventual fixation of a neutral allele

- Can also be derived using the formula for fixation probability under selection when selection coefficient $s \to 0$

- $P_F(p) = \dfrac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}$

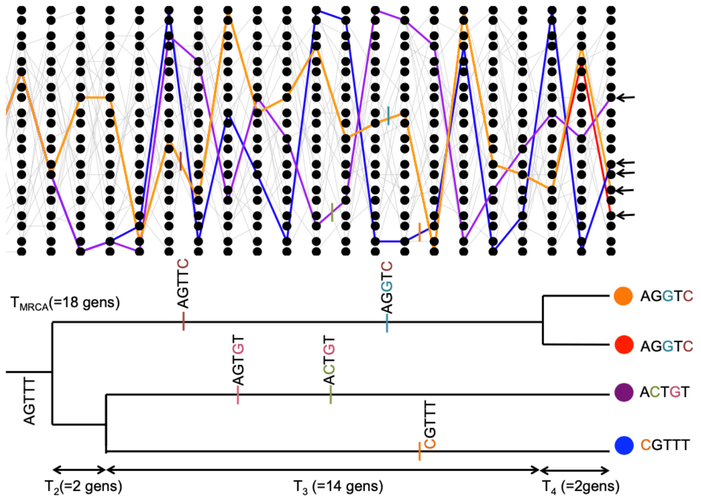
- $P_F(1/2N) = \dfrac{1 - e^{-s}}{1 - e^{-2Ns}}$



Time to eventual fixation of a neutral allele

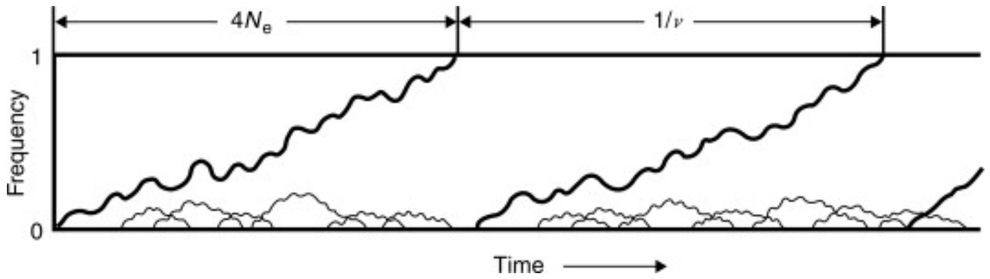
- Time to most recent common ancestor (MRCA) in a sample of n alleles:
- $$\mathbb{E}(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right)$$

- Roughly $4N$ generations for a neutral allele present in a single copy within the population to fix



Rate of substitution of neutral alleles

- What is the distribution (expectation) of the number of substitutions along a lineage?
- Assumption: a fraction C of all mutational changes are highly deleterious, and cannot possibly contribute to substitution nor polymorphism. The other $1 - C$ fraction of mutations are neutral.



- $2N\mu(1 - C)$ mutations enters per generation
- On average $1/2N$ will be fixed
- Substitution rate $\mu(1 - C)$

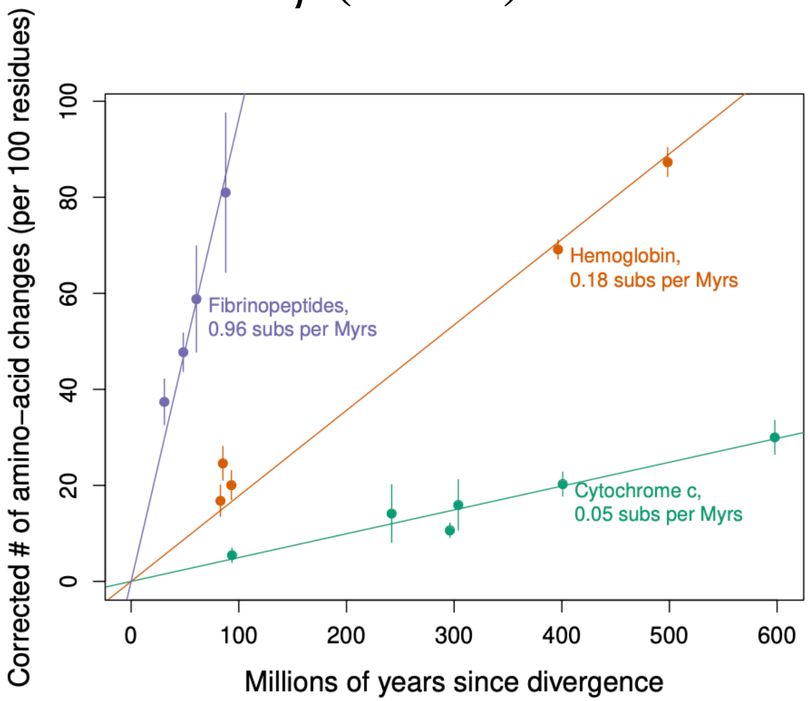
Rate of substitution of neutral alleles

- Consider a pair of species that have diverged for T generations
- i.e. orthologous sequences shared between the species last shared a common ancestor T generations ago
- $2\mu(1 - C)T$ neutral substitutions on average
- Assumption: T is a lot longer than the time it takes to fix a neutral allele, such that the total number of alleles introduced into the population that will eventually fix is the total number of substitutions.

Rate of substitution of neutral alleles

- Primary determinant of patterns of molecular evolution in a genomic region is the level of constraint (C).
- Empirically supported
  - non-coding regions often evolve more rapidly than coding regions,
  - synonymous substitutions accumulate faster than nonsynonymous
  - nonsynonymous substitutions accumulate faster in less vital proteins
- Allows one to spot putatively functional non-coding regions by looking for genomic regions that have very low levels of divergence among distantly related species.

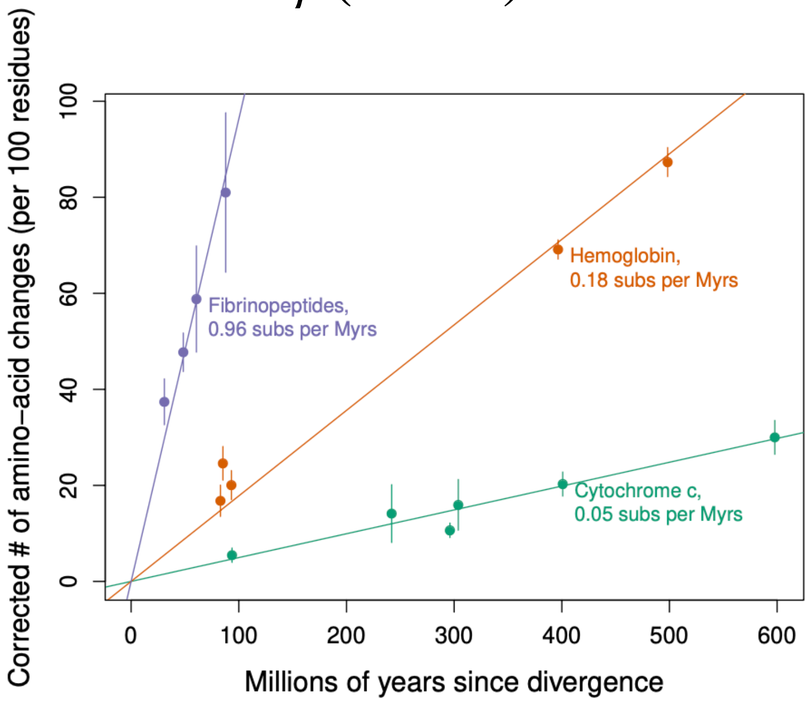
$$2\mu(1-C)T$$



Molecular clock

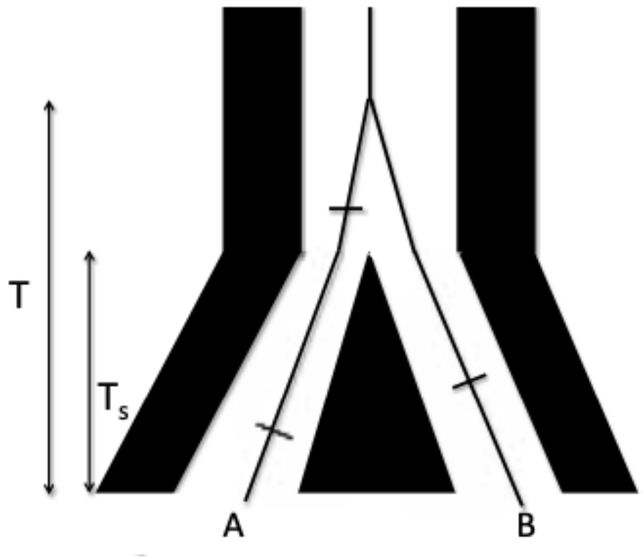
- Nucleotides accrue substitutions at a constant rate
- Number of substitutions scale linearly with time
- Empirical evidence
  - divergence time between species (e.g. from fossil record)
- Useful for dating divergence times

$$2\mu(1 - C)T$$



The contribution of ancestral polymorphism to divergence

- T_s: time since the two species split from the common ancestor
- $T = T_s + 2N_A$
- If recent split, then ancestral polymorphisms contribute to divergence



Tests of molecular evolution - dN/dS

- Neutral model offers null hypothesis for us to test for possible selection
- dN/dS
- ratio of the rates of non-synonymous (dN) to synonymous substitutions (dS) in ***a given gene***
- We can calculate **dN** by counting up the non-synonymous changes divided by the total number of positions in the gene where a nonsynonymous point mutation could occur and then divide by time.
- **dS** can be calculated in a similar way

Tests of molecular evolution - dN/dS

- $dN = (1 - C)\mu$
- $dS = \mu$
- $dN/dS = (1 - C)$
- Assuming non-synonymous mutations can only be strongly deleterious or neutral, i.e. only negative selection
- $C = 1 - dN/dS$
- C: fraction of non-synonymous mutations that are quickly weeded out of the population by selection, and so do not contribute to divergence among species.

Tests of molecular evolution - dN/dS

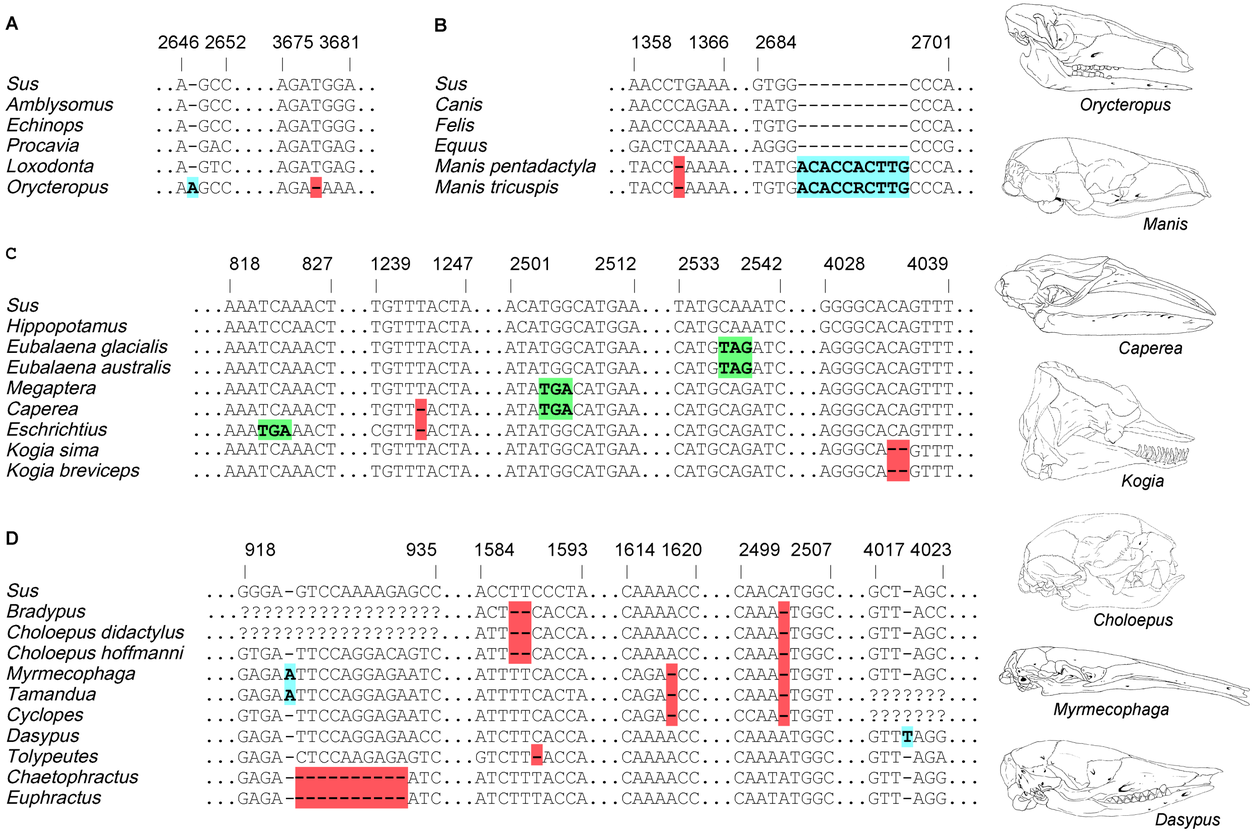
- Can use dN/dS to test if a gene evolves in a constrained way (i.e. under purifying selection)

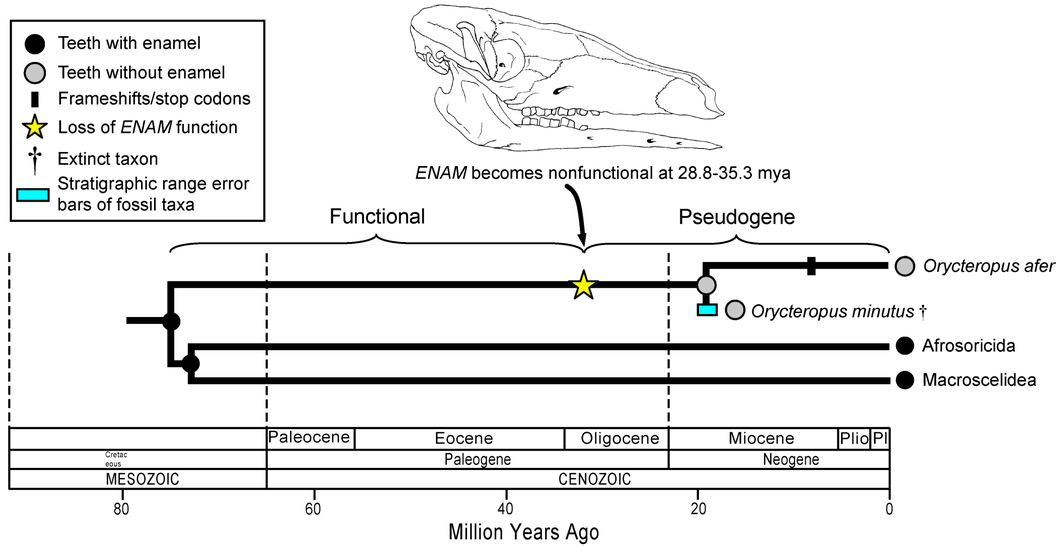
- By testing dN/dS this is significantly less that 1

- Can provide evolutionary evidence that a stretch of DNA proposed to be protein-coding is subject to selective constraint, and so likely does encode for a functional protein.

- Can also be applied to specific branches of a phylogeny for a gene, to test on which branches the gene is subject to constraint, or to test for changes in the level of constraint across the phylogeny.

Loss of function of enamel gene result in dN/dS close to 1

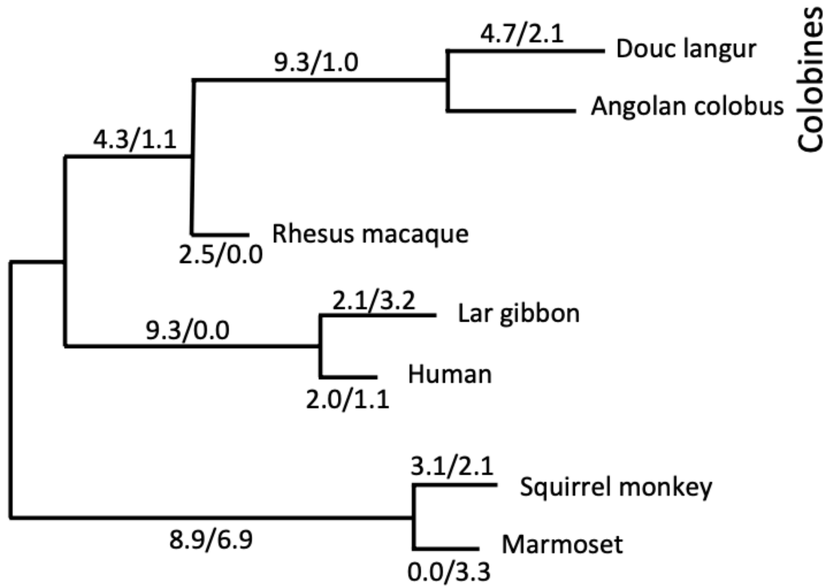




- branches of the enamlin phylogeny with a functional enamlin gene had an estimated dN/dS = 0.51
- the branches with a pseudogenized Enamlin had dN/dS = 1.02

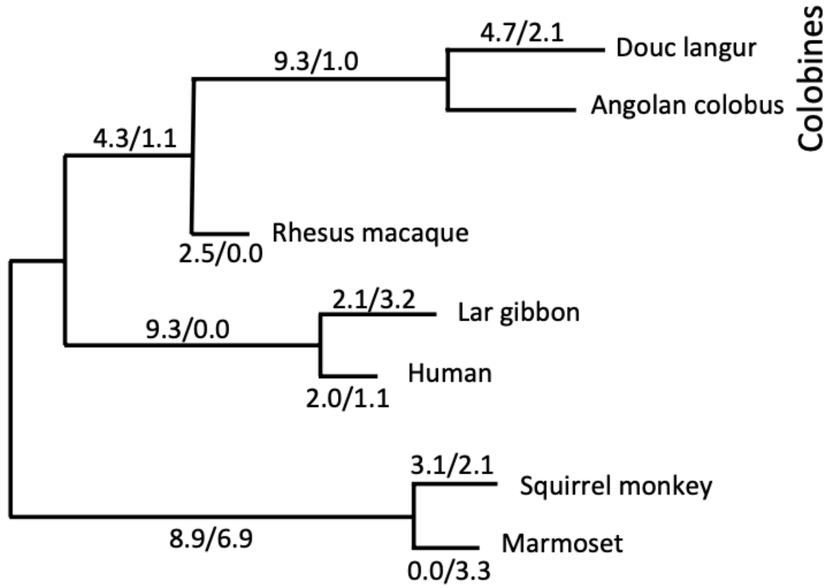
Adaptive evolution and dN/dS

- A gene can also be under positive selection: beneficial mutations must also arise and fix from from time to time
- B: fraction of non-synonymous mutations that are beneficial
- $2N\mu B$ beneficial mutations arise per generation
- With fixation probability $f_B > 1/2N$
- $dN = (1 - C - B)\mu + (2N\mu B) \times f_B$.
- $dN/dS = (1 - C - B) + 2NBf_B$
- Can be positive if non-synonymous mutations occur at faster rate!



Adaptive evolution and dN/dS

- Evolution of the lysozyme gene in primates (Messier and Stewart, 1997; Yang, 1998)
- Lysozyme: key for the breaking down bacterial walls.
- Lysozyme gene shows very fast protein evolution, notably on the lineages leading to apes and Colobines
- Colobine lysozyme protein has convergently evolved very similar amino-acid changes at 5 key residuals in cows and Hoatzins

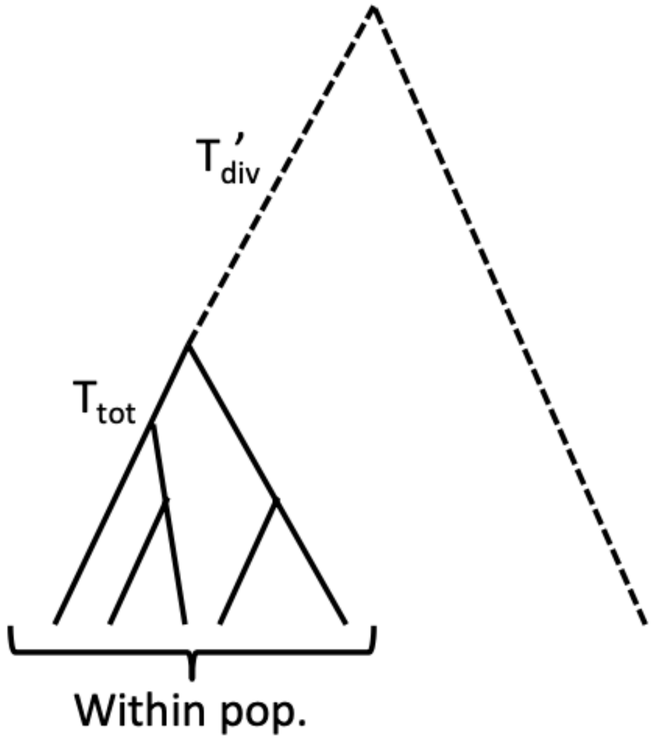


Tests of molecular evolution: McDonald-Kreitman

- dN/dS to very conservative for detecting selection
- For a more powerful test of rapid divergence, we need to adjust for the level of constraint a gene experiences at non-synonymous sites.
- Idea: use polymorphism data as an internal control

Tests of molecular evolution: McDonald-Kreitman

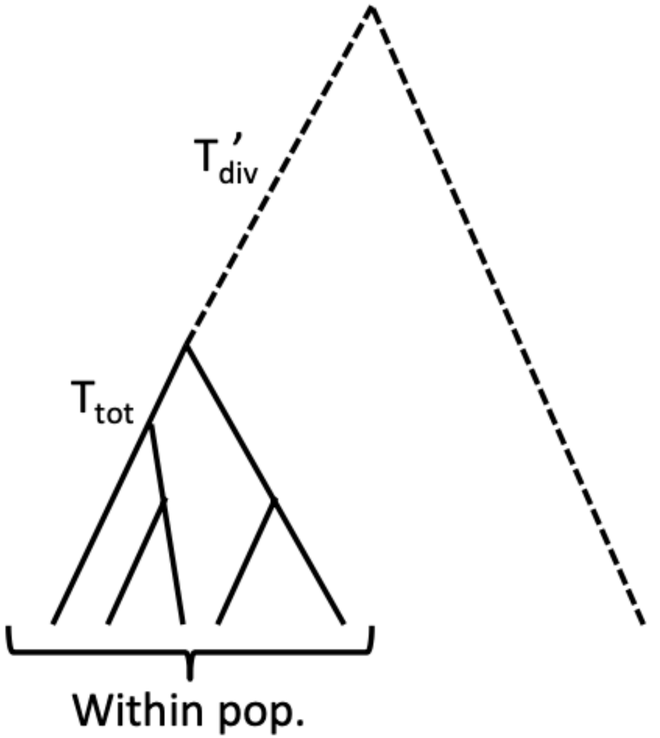
- MK test: neutral theory of molecular evolution at a gene
- Data: polymorphism at a gene for one species and divergence to a closely related species.



	Poly.	Fixed
Non-Syn.	P_N	D_N
Syn.	P_S	D_S
Ratio	P_N/P_S	D_N/D_S

Tests of molecular evolution: McDonald-Kreitman

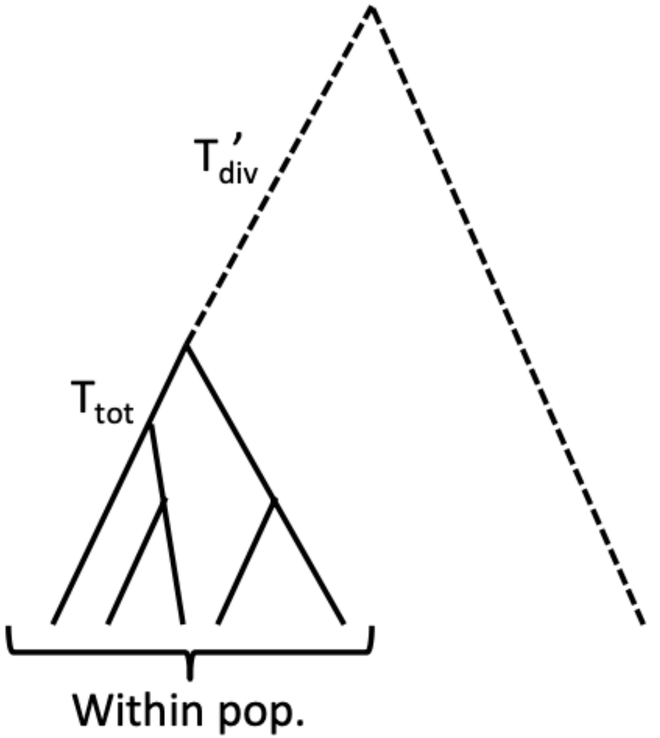
- L_s : sites where synonymous could arise
- L_N : sites where non-synonymous could arise
- Null expectation:
  - $D_N/D_S < 1$
  - $P_N/P_S < 1$
- Ratio significantly higher for divergence than polymorphism: evidence that non-synonymous substitutions are accumulating more rapidly expected given constraint alone.



	Poly.	Fixed
Non-Syn.	$\mu L_N(1-C)T_{tot}$	$\mu L_N(1-C)T'_{div}$
Syn.	$\mu L_S T_{tot}$	$\mu L_S T'_{div}$
Ratio	$L_N(1-C)/(L_S)$	$L_N(1-C)/(L_S)$

Tests of molecular evolution: McDonald-Kreitman

- Intuition:
- Strongly beneficial mutations quickly become fixed in the population so do not contribute to polymorphism
- Per site rate of mutations that contribute to polymorphism
  - $\mu(1 - C - B + B \times p)$
  - p is probability of this allele fixes before present time
  - p is small if selection is strong
- Per site rate of mutations that contribute to substitutions
  - $\mu[(1 - C - B) + B \times 2Nf_B]$



	Poly.	Fixed
Non-Syn.	$\mu L_N(1-C)T_{tot}$	$\mu L_N(1-C)T'_{div}$
Syn.	$\mu L_S T_{tot}$	$\mu L_S T'_{div}$
Ratio	$L_N(1-C)/(L_S)$	$L_N(1-C)/(L_S)$