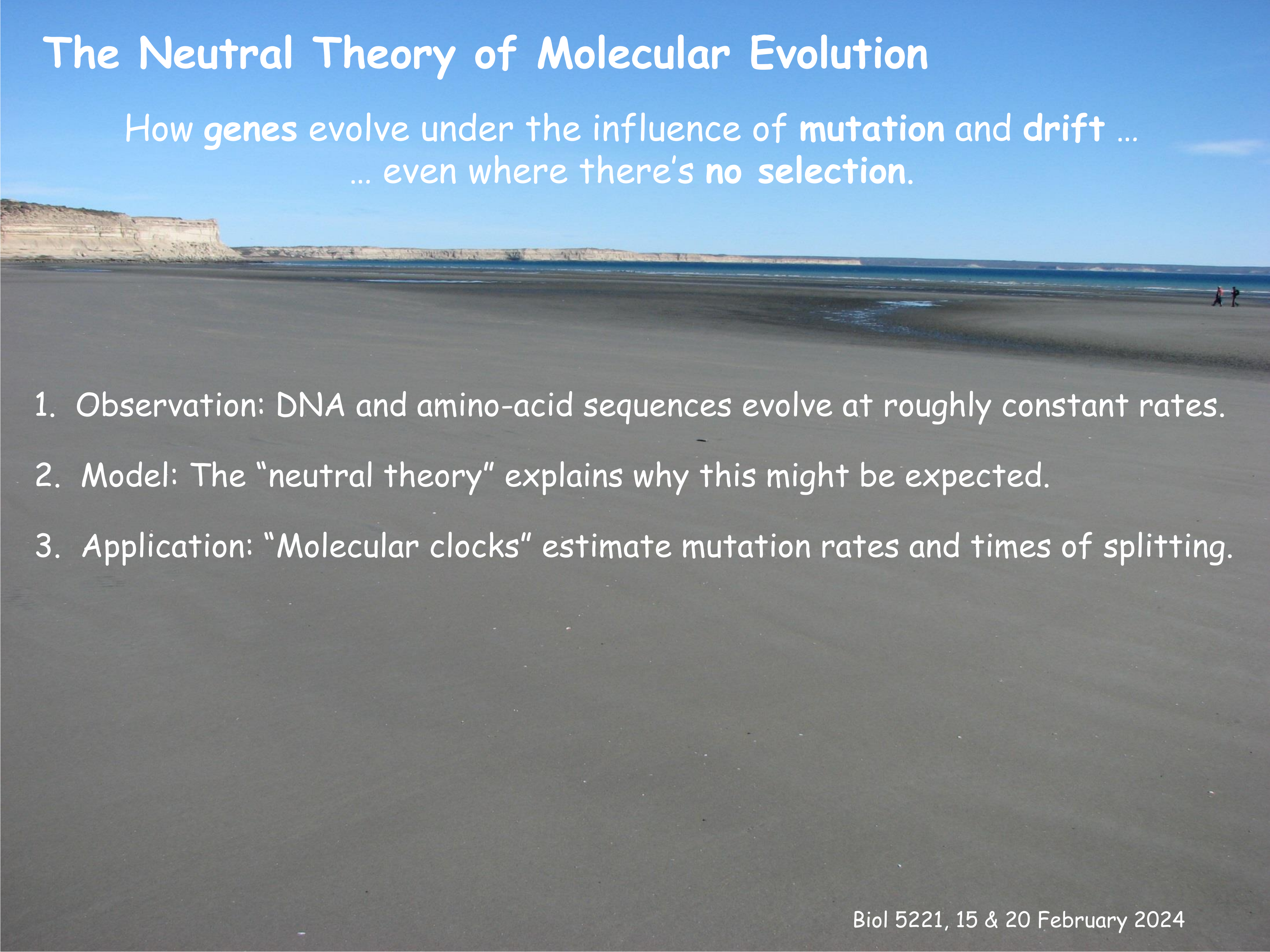
The Neutral Theory of Molecular Evolution

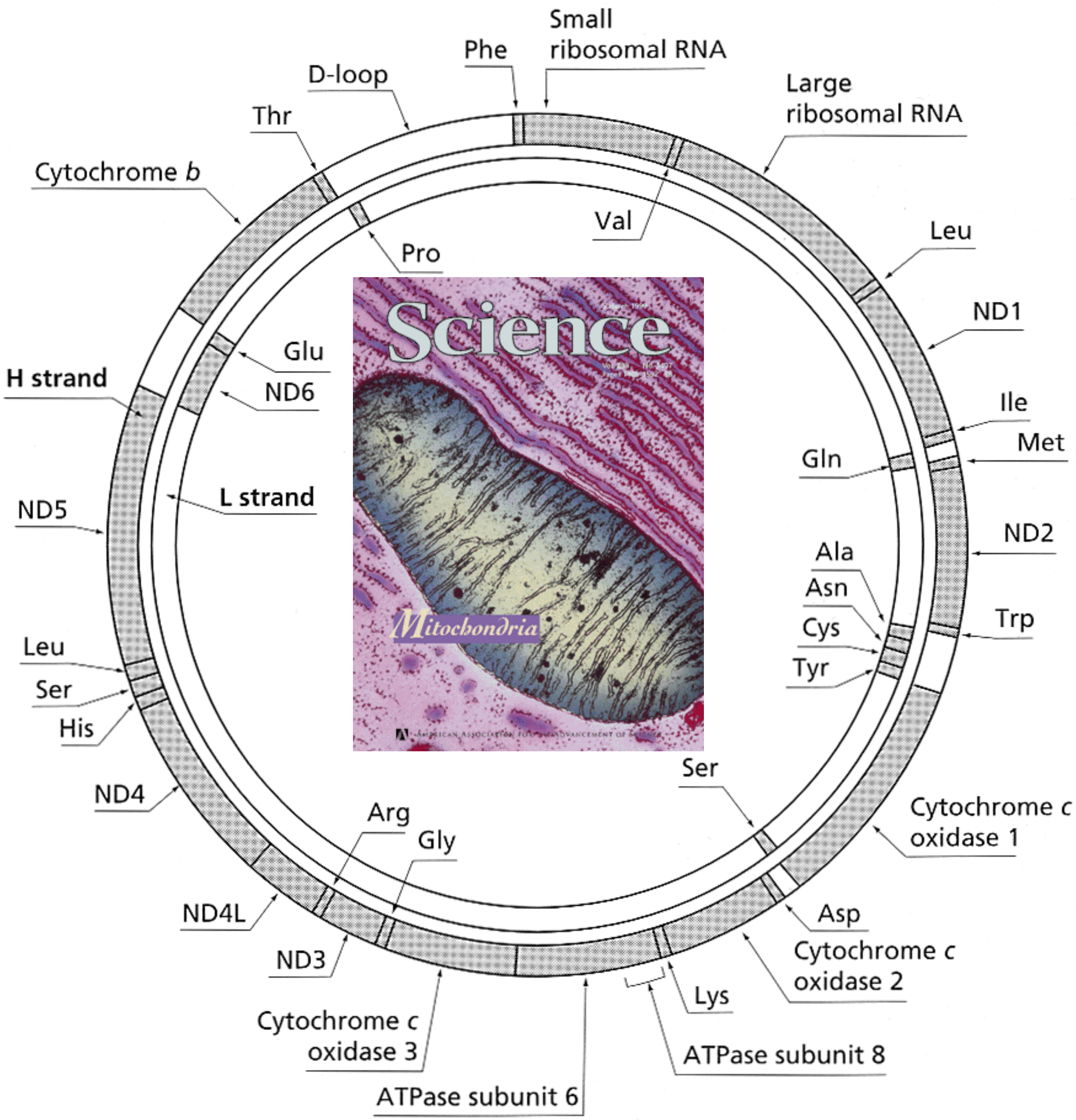
How **genes** evolve under the influence of **mutation** and **drift** …
… even where there's **no selection**.



1. Observation: DNA and amino-acid sequences evolve at roughly constant rates.
2. Model: The "neutral theory" explains why this might be expected.
3. Application: "Molecular clocks" estimate mutation rates and times of splitting.

Biol 5221, 15 & 20 February 2024

The human mitochondrial genome



Structurally identical in almost all mammals.

Tiny remnant of a formerly free-living bacterium that became an endosymbiont … then an organelle!

The human reference genome is 16,569 base pairs long.

Same genes as in all animals:

- 13 protein-coding genes
- 22 tRNA genes
- 2 ribosomal RNA genes

Most are encoded on the "heavy" (H) strand (clockwise).

ND6 and some tRNAs are encoded on the "light" (L) strand (counter-clockwise).

No introns, transposons, or "junk".

Highly A/T biased.

Mutation rate ~10x higher than that of the nuclear chromosomes.

Our mt genome can easily be *aligned* with those of other primates.

At most nucleotide ***positions*** (“sites”), ***everyone*** has the ***same*** nucleotide ***state***.

But ***some sites are variable***.

At these variable sites, ***some patterns are more common*** than others.

Here are the first ***180 bp*** of the ~16.5 kb alignment for some famous hominoids.

```
modern    GTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCA
Neander.  GTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCA
chimp     GTTTATGTAGCTTACCCCCTCAAAGCAATACACTGAAAATGTTTCGACGGGTTTACATCA
gorilla   GTTTATGTAGCTTACCTCCCCAAAGCAATACACTGAAAATGTTTCGACGGGCTCACATCA
          **************** ** ************************ ****** * ******

modern    CCCCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACAT
Neander.  CCCCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACAT
chimp     CCCCATAAACAAACAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACAT
gorilla   CCCCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAACTCTTAGTAGGATTACACAT
          ************* ************************* ********* **********

modern    GCAAGCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGC
Neander.  GCAAGCATCCCCATTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGGACAAGC
chimp     GCAAGCATCCCCGCCCC-GTGAGT-CACCCTCTAAATCGCCATGATCAAAAGGAACAAGT
gorilla   GCAAGCATCCCCGCCCCAGTGAGT-CACCCTCTAAATCACCACGATCAAAAGGAACAAGC
          ************   ** ****** ************* *** ********** *****
```

Of those 180 positions, only 16 vary among the species.

modern	T	T	A	C	C	T	G	A	G	T	T	A	T	A	A	C
Neanderthal	T	T	A	C	C	T	G	A	A	T	T	A	T	A	G	C
chimp	C	T	C	T	T	C	G	A	G	C	C	-	-	G	A	T
gorilla	T	C	C	C	C	T	A	G	G	C	C	A	-	A	A	C

164/180 (91%) ***do not*** vary, implying they have ***not evolved*** since the last common ancestor of all four hominoids.

Pairwise Differences	m	N	c	g
modern	-	2	11	7
Neanderthal	2	-	13	9
chimp	11	13	-	10
gorilla	7	9	10	-

How did these differences accumulate?

The evolutionary relationships of the four species can be inferred securely from the matrix of pairwise differences for all 16.5 kb.

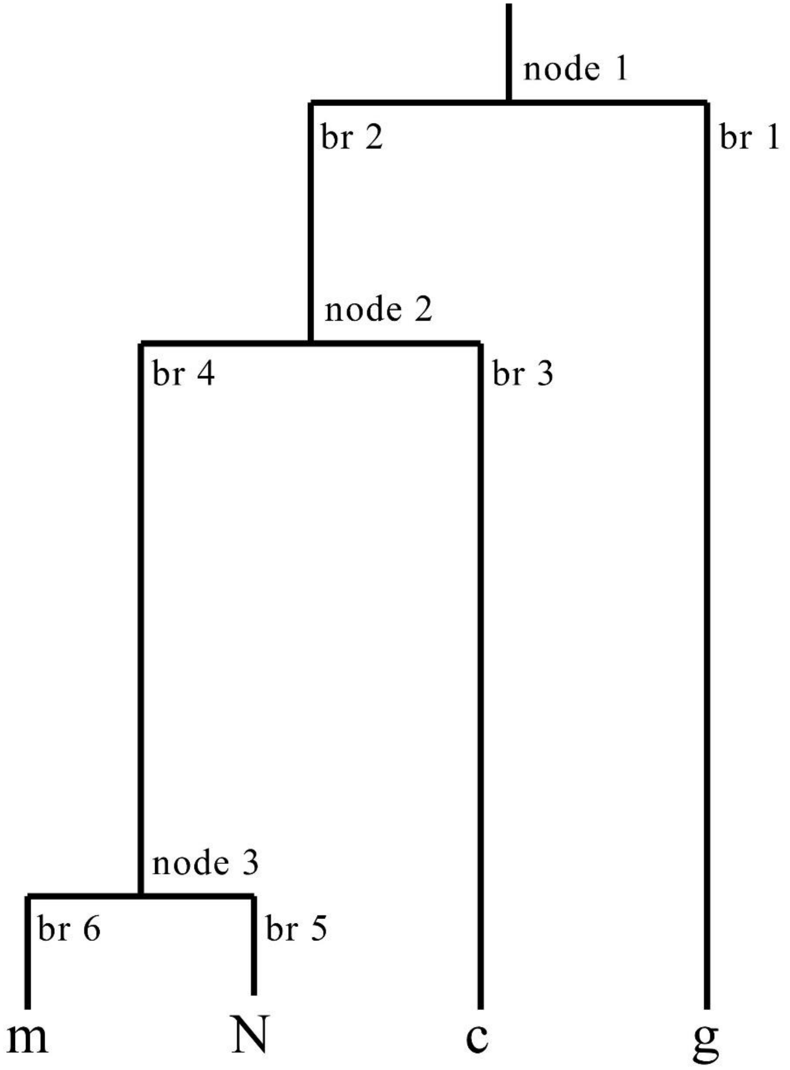
		mod	Nea	chi	gor
modern human	(m)	-	168	1305	1605
Neanderthal	(N)	168	-	1290	1597
chimpanzee	(c)	1305	1290	-	1557
gorilla	(g)	1605	1597	1557	-

And also from the distribution of site patterns

	m	N	c	g	m	N	c	g	#
p1	1	1	1	2	T	T	T	C	884
p2	1	1	2	2	A	A	C	C	589
p3	1	1	2	1	T	T	C	T	583
p4	1	2	2	2	T	A	A	A	63
p5	1	2	1	1	G	A	G	G	53
p6	1	1	2	3	T	T	C	A	40
p7	1	2	2	1	T	C	C	T	23
p8	1	2	1	2	T	C	T	C	20
p9	1	2	2	3	G	C	C	T	4
p10	1	2	3	3	T	C	A	A	2
p11	1	2	1	3	T	C	T	A	2
p12	1	2	3	2	G	C	A	C	1
									2264

1 mut

2 muts

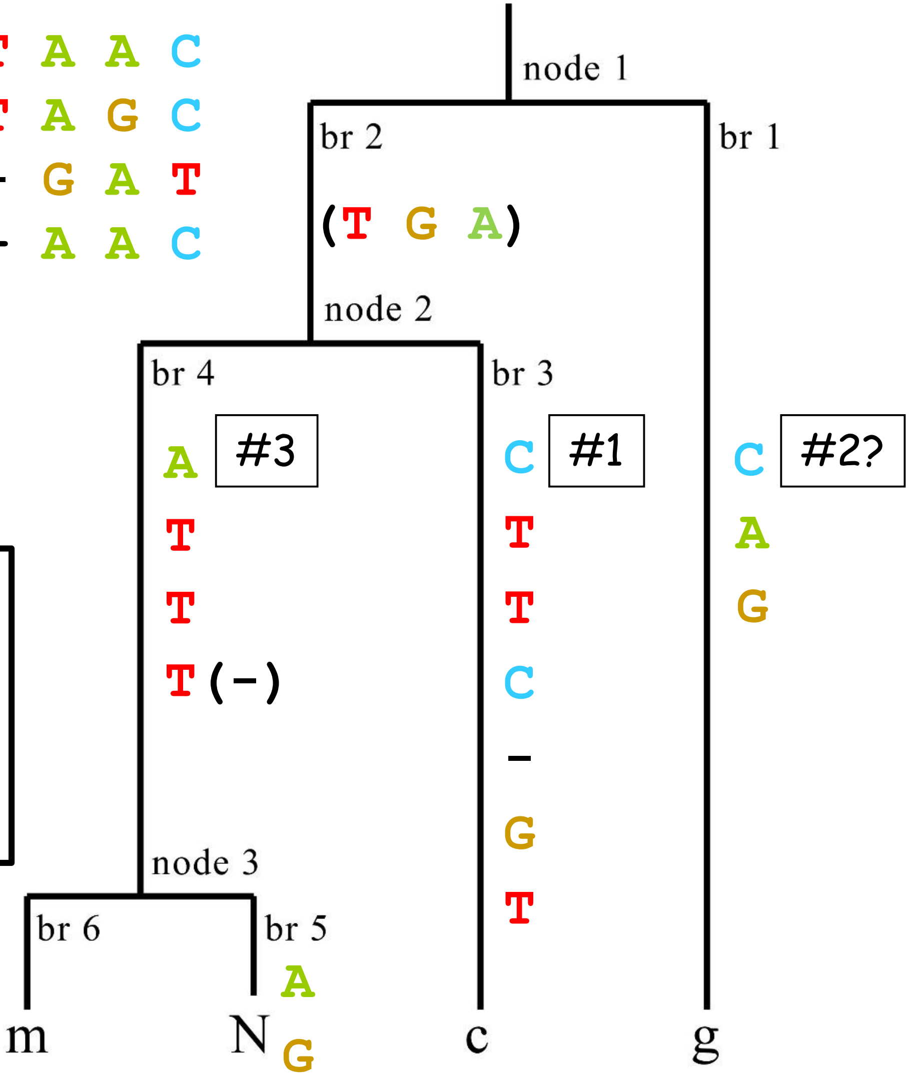


Then given the tree, we can easily "reconstruct" the mutations at the variable sites (e.g., the first 16 of them).

m	T	T	A	C	C	T	G	A	G	T	T	A	T	A	A	C
N	T	T	A	C	C	T	G	A	A	T	T	A	T	A	G	C
c	C	T	C	T	T	C	G	A	G	C	C	-	-	G	A	T
g	T	C	C	C	C	T	A	G	G	C	C	A	-	A	A	C

Pairwise Differences

	m	N	c	g
modern	-	2	11	7
Neanderthal	2	-	13	9
chimp	11	13	-	10
gorilla	7	9	10	-



But 180 bp with 16 variable sites is NOT enough sequence to correctly infer the tree!

Differences *within* species are like those *between species*, but less so

Many modern human and chimpanzee mitochondrial genome sequences have been determined and aligned.

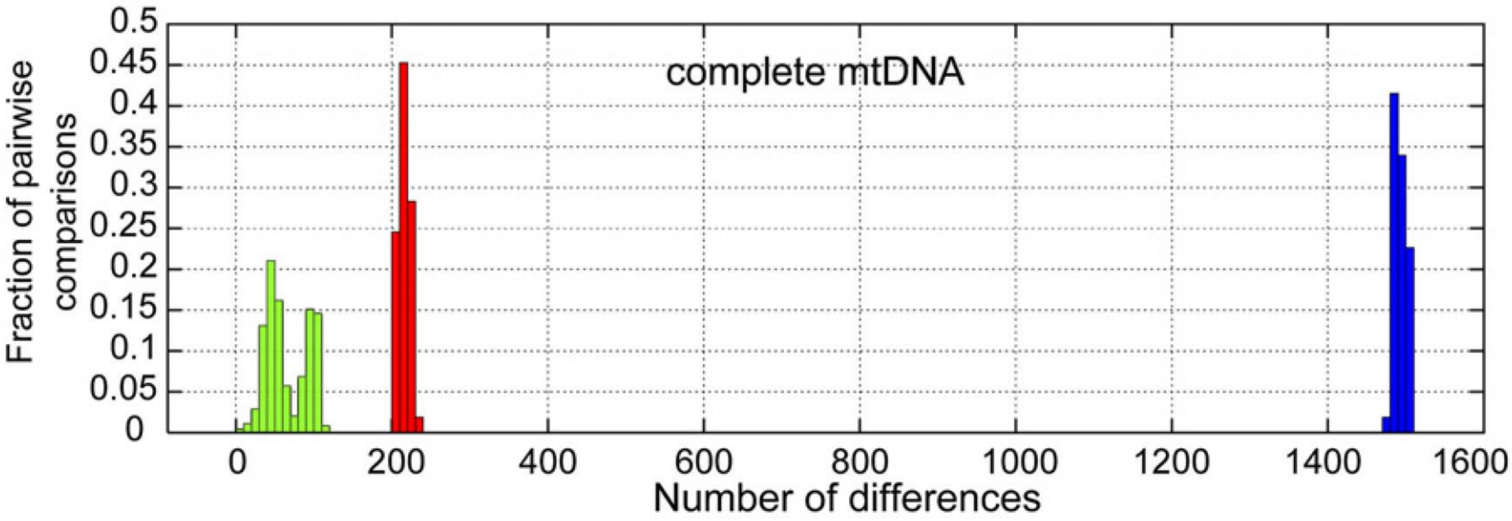
Also a few Neanderthal individuals and other pre-moderns (from fossils).

Here's the distribution of the *pairwise differences* (out of ~16.5 kb in all) for 53 modern humans, one Neanderthal and one chimpanzee.

Green histogram: distances among 53 modern humans

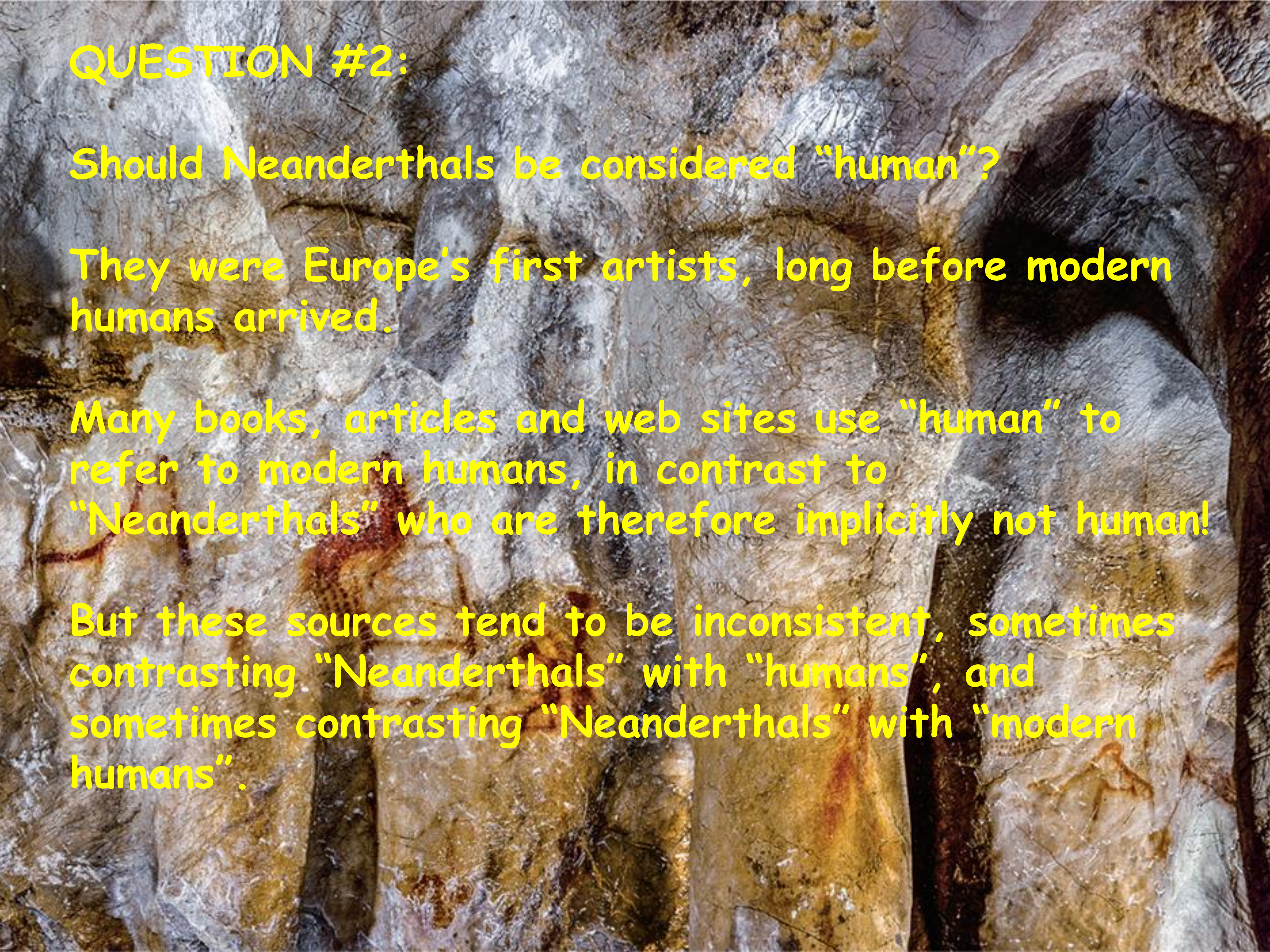
Red: distances from one Neanderthal to all 53 modern humans

Blue: distances from a typical chimp to modern and Neanderthal humans



Green *et al.* (2008) *Cell* **134**:416-426

QUESTION #1: How can the variation *among* modern humans be greater than the variation *between* those same humans and Neanderthal or chimp?



QUESTION #2:

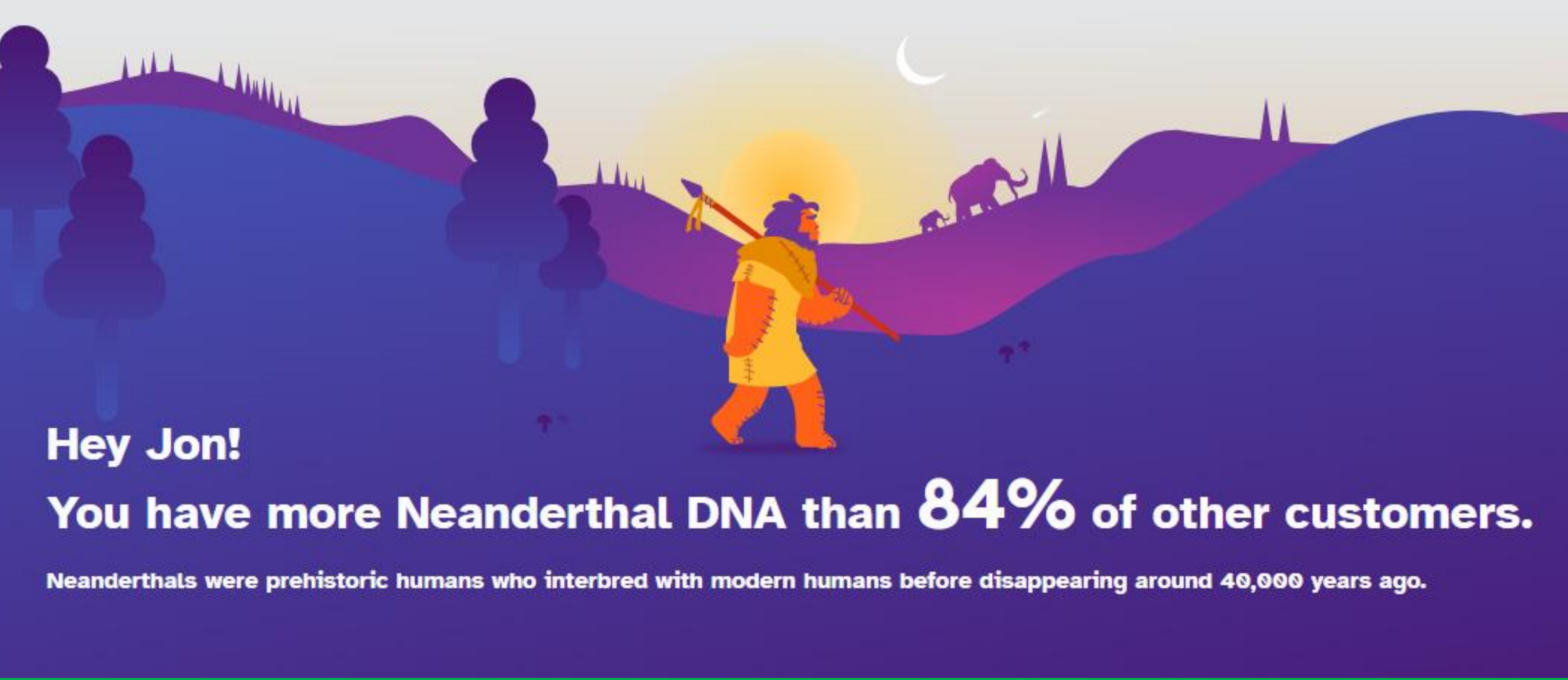
Should Neanderthals be considered “human”?

They were Europe’s first artists, long before modern humans arrived.

Many books, articles and web sites use “human” to refer to modern humans, in contrast to “Neanderthals” who are therefore implicitly not human!

But these sources tend to be inconsistent, sometimes contrasting “Neanderthals” with “humans”, and sometimes contrasting “Neanderthals” with “modern humans”.

Even the very sophisticated 23andMe!



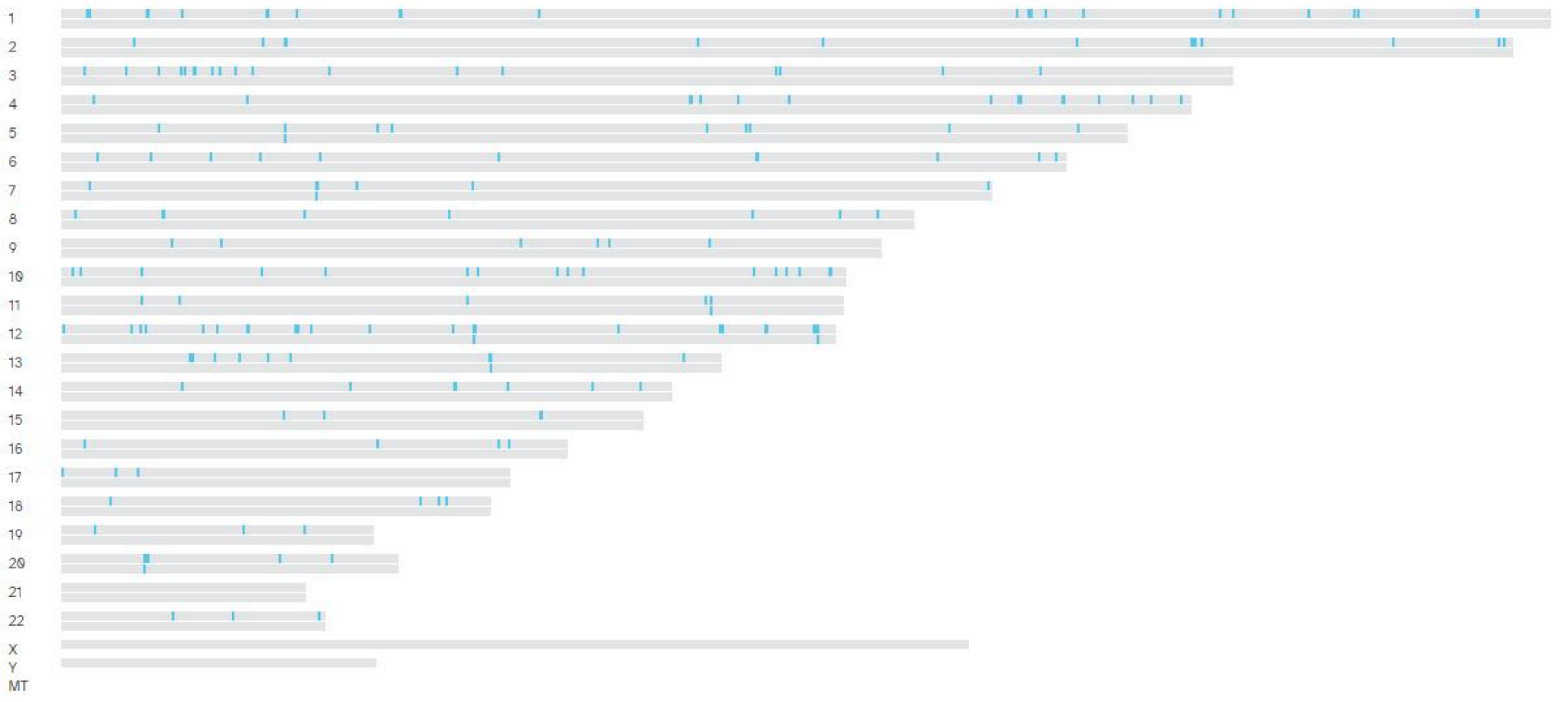
Hey Jon!

You have more Neanderthal DNA than 84% of other customers.

Neanderthals were prehistoric humans who interbred with modern humans before disappearing around 40,000 years ago.

(The total is around 2% of my genome.)

It appears as more than 250 small fragments, scattered over all the chromosomes.



My sister, and most of you, have fewer.
Am I *less human* than you?

Three observations about protein evolution stimulated development of the "neutral theory of molecular evolution" in the early 1970s.

Pattern 1: Seemingly constant rates of amino-acid evolution over many millions of years, by individual proteins (e.g. β-globin)

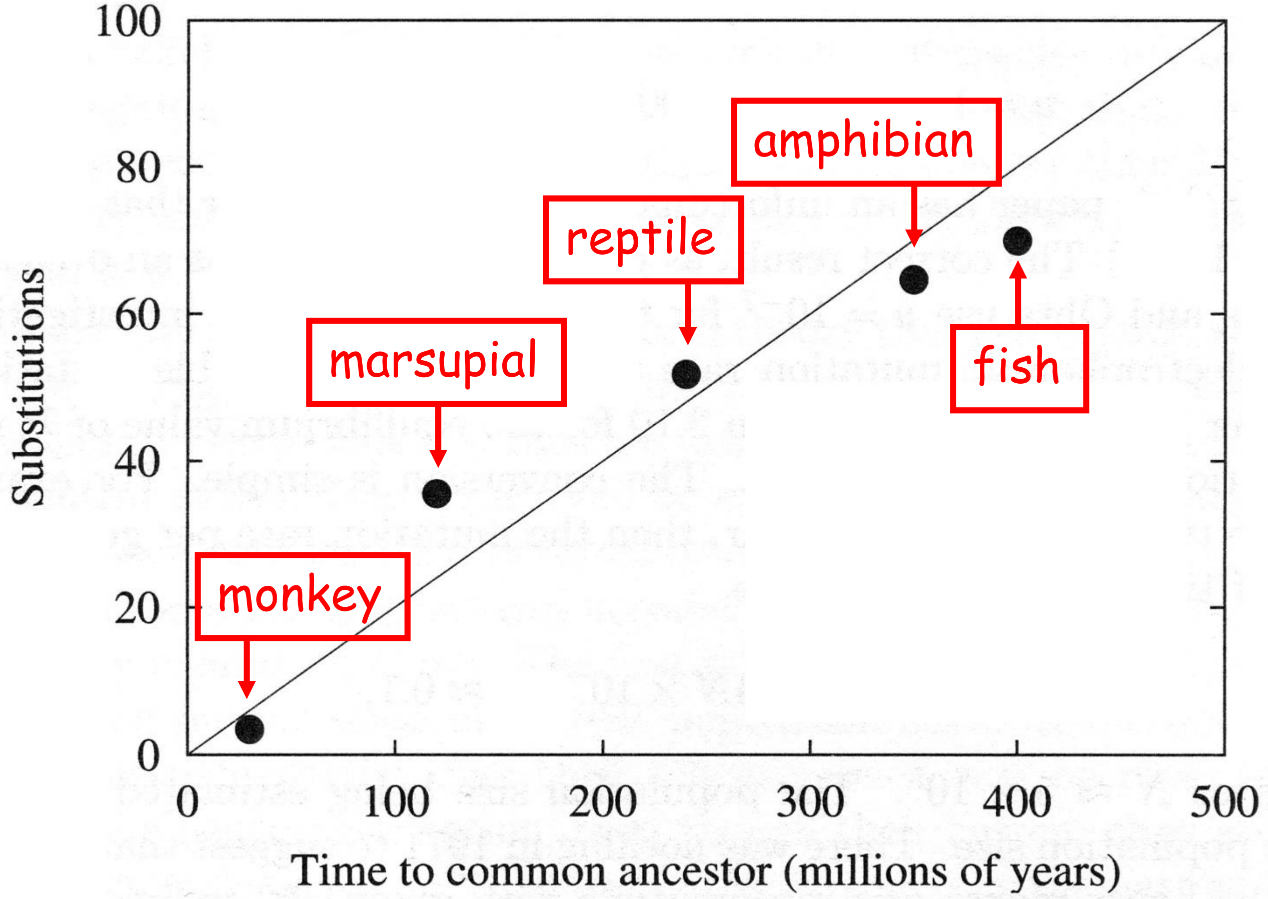


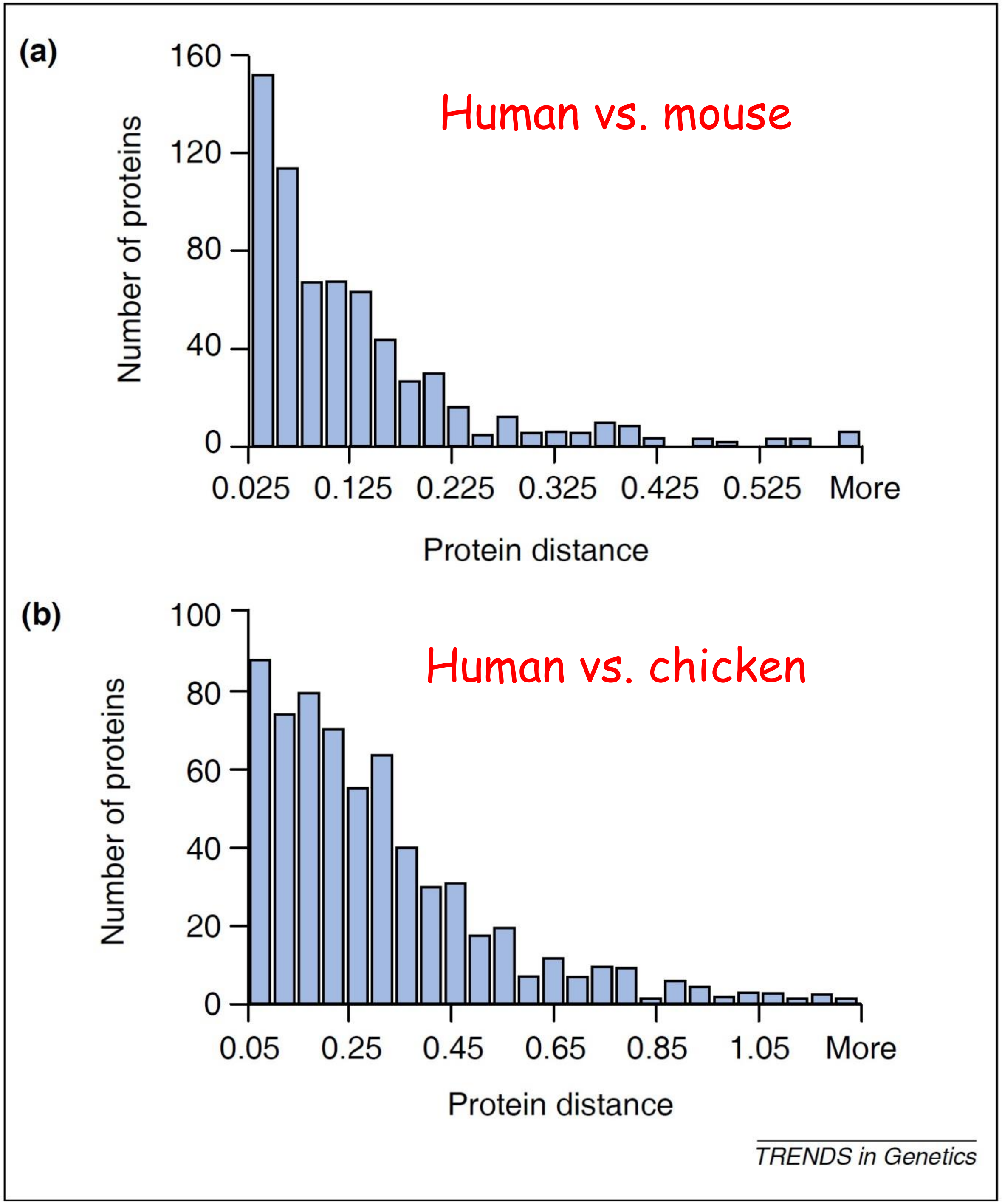
Figure 2.6: The number of amino acid substitutions in beta globin that occurred in the lineages leading to humans and various species as a function of the time back to their common ancestors.

Pattern 2: Different proteins evolving at characteristically very different rates.

This recent analysis uses the genome sequences of human, mouse and chicken, comparing the accumulated differences of 647 proteins.

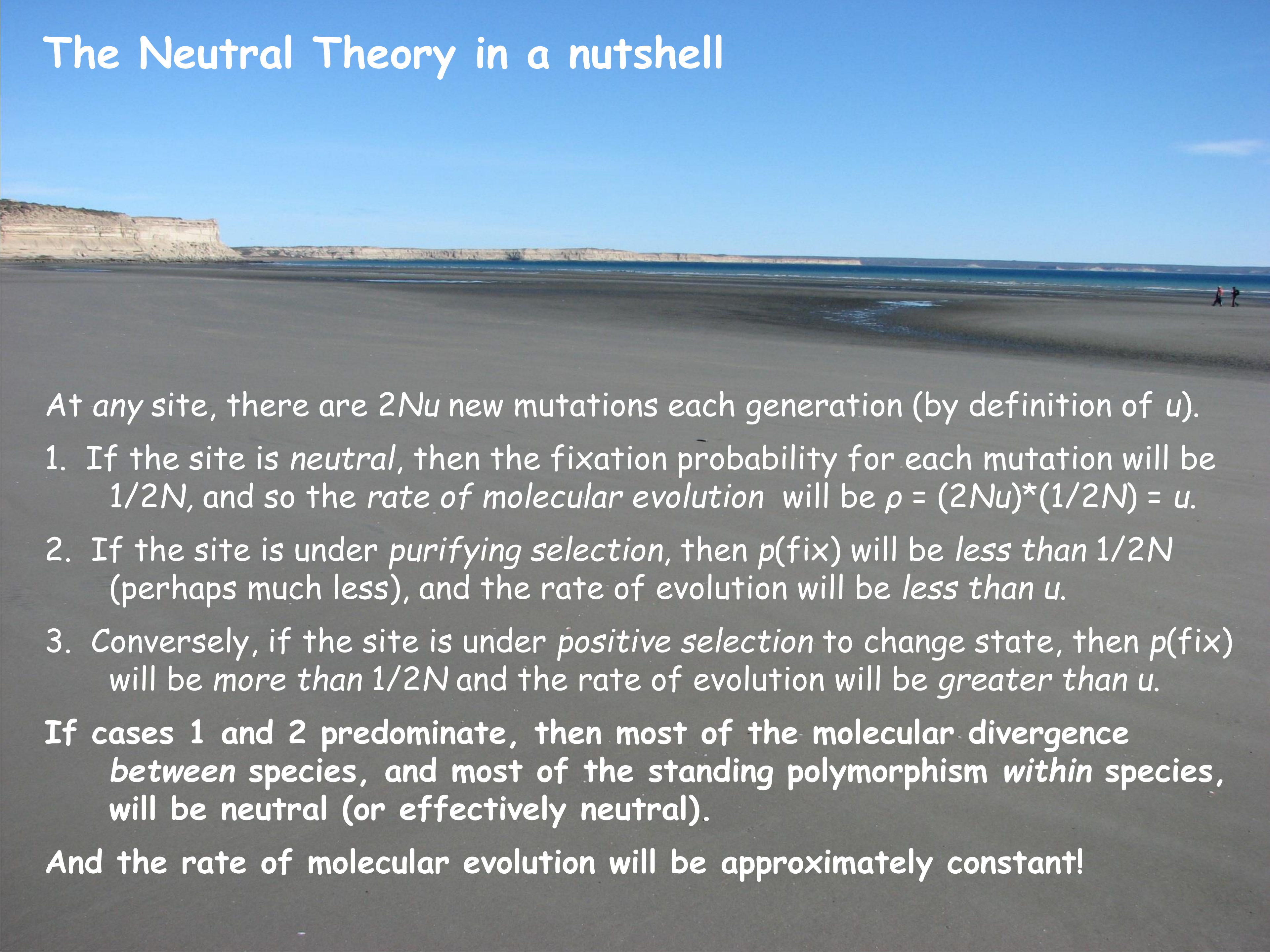
Pattern 3: Different *parts of the same* protein evolving at very different rates.

(And later, different rates at synonymous and nonsynonymous sites in coding DNA sequences.)



TRENDS in Genetics

The Neutral Theory in a nutshell



At *any* site, there are 2*Nu* new mutations each *generation* (by definition of *u*).

1. If the site is *neutral*, then the fixation probability for each mutation will be 1/2*N*, and so the *rate of molecular evolution* will be $\rho = (2Nu)*(1/2N) = u$.
2. If the site is under *purifying selection*, then *p*(fix) will be *less than* 1/2*N* (perhaps much less), and the rate of evolution will be *less than u*.
3. Conversely, if the site is under *positive selection* to change state, then *p*(fix) will be *more than* 1/2*N* and the rate of evolution will be *greater than u*.

If cases 1 and 2 predominate, then most of the molecular divergence *between* species, and most of the standing polymorphism *within* species, will be neutral (or effectively neutral).

And the rate of molecular evolution will be approximately constant!

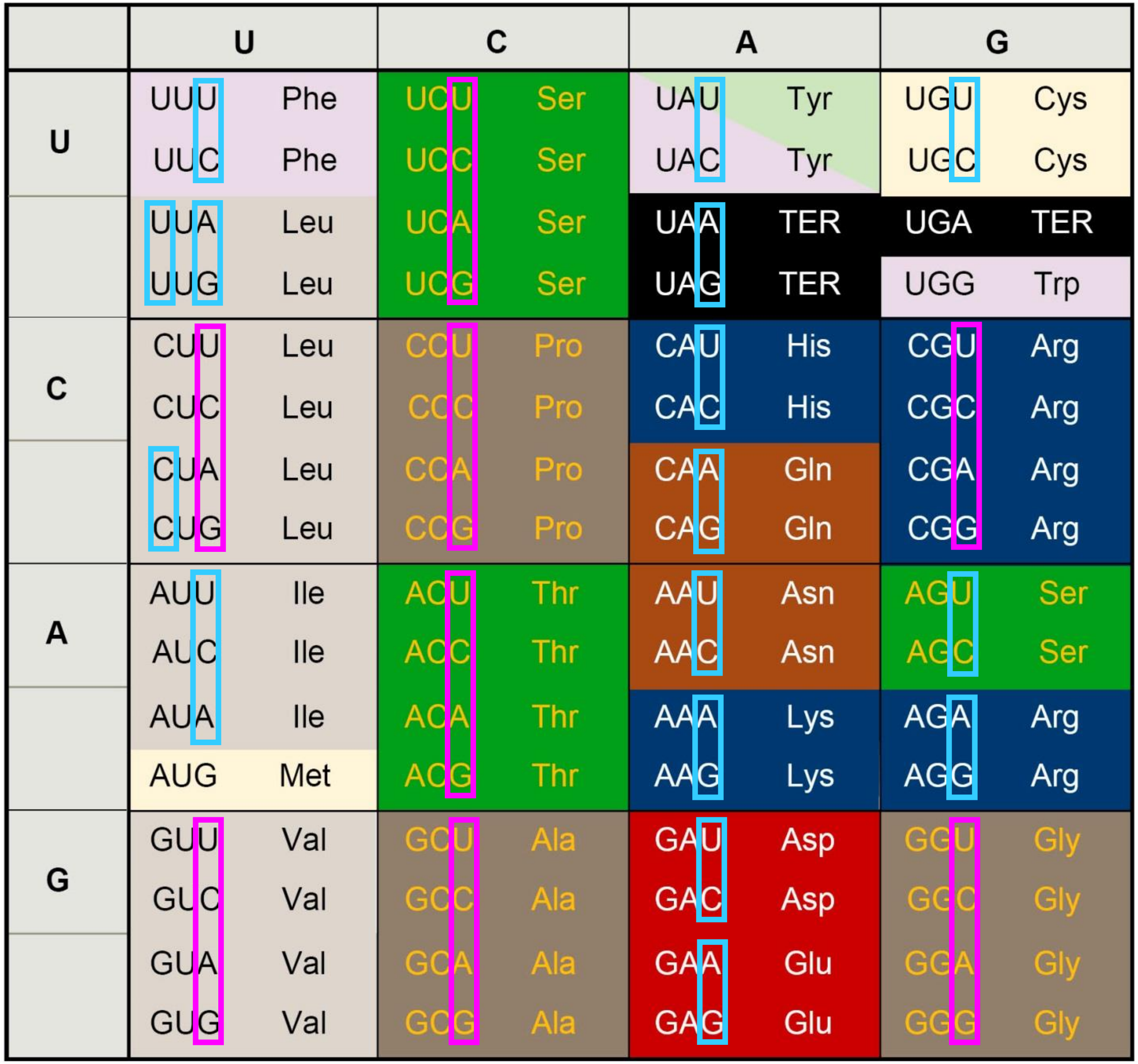
Most sites in coding sequences are under purifying selection, so they evolve slowly and show little variation within species.

But "synonymous" sites can mutate without changing the amino-acid sequence of the protein.

4-fold synonymous or "degenerate" sites can mutate to any of the other three bases.

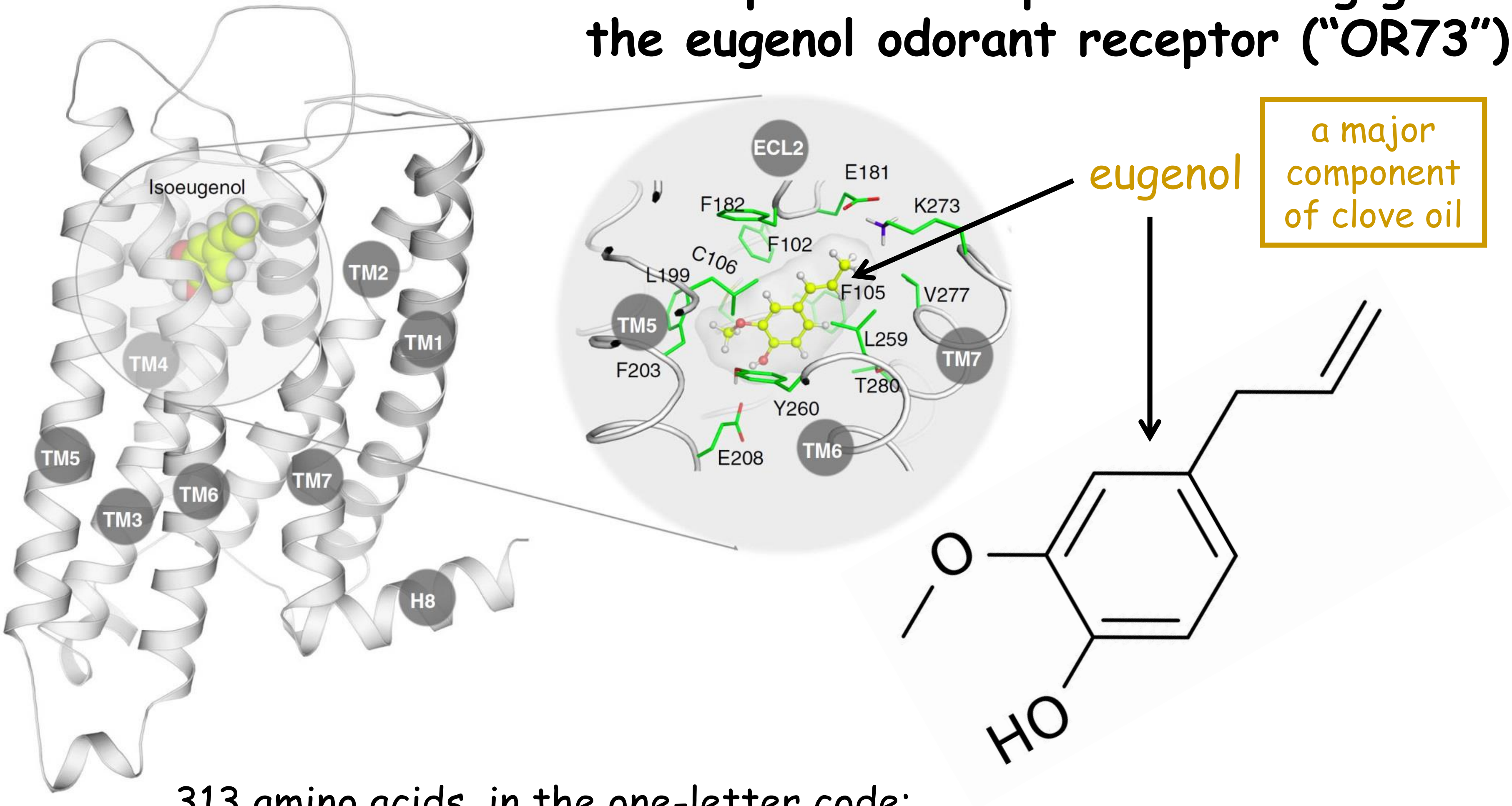
2-fold degenerate sites can mutate to the other purine (A ↔ G) or pyrimidine (C ↔ T=U).

Overall, *roughly* 25% of random nucleotide substitutions in a typical coding sequence will be synonymous, and 75% will be non-synonymous.



	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	TER	UGA	TER
	UUG	Leu	UCG	Ser	UAG	TER	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

A simple nuclear protein-coding gene: the eugenol odorant receptor ("OR73")



a major component of clove oil

313 amino acids, in the one-letter code:

```
MTLSDGNHSGAVFTLLGFSDYPELTIPLFLIFLTIYSITVVGNIGMIVIIRINPKLHIPMYFF
LSHLSFVDFCYSSIVAPKMLVNLVTMNRGISFVGCLVQFFFFCTFVVTESFLLGVMAYDRFVA
IRNPLLYTVAMSQRLCAMLVLGSYAWGVVCSLILTCSALNLSFYGFNMINHFFCEFSSLLSLS
RSDTSVSQLLLFVFATFNEISTLLIILLSYVLIVVTILKMKSASGRRKAFSTCASHLTAITIF
HGTILFLYCVPNSKNSRHTVKVASVFYTVVIPMLNPLIYSLRNKDVKDTVKKIIGTKVYSS
```

Translated human and mouse OR73 (“eugenol receptor”) coding sequences

Anth/Biol 5221, 18 February 2020

```
314 codons (313 amino acids), 942 base pairs
  44 first-position differences   (14.0%)
  30 second-position differences  ( 9.6%)
 113 third-position differences   (36.0%)
 187 total nucleotide differences (19.9%)
  56 amino-acid differences       (17.9%)
```

First- and second-position differences, and amino-acid differences, are much less common than third-position differences!

```
human      M  L  L  T  D  R  N  T  S  G  T  T  F  T  L  L  G  F  S  D  Y  P  E  L  Q  V  P  L  F  L    30
          atgctgctgacagatagaaatacaagtgggaccacgttcaccctcttgggcttctcagattacccagaactgcaagtcccactcttcctg   90
mouse      .  T  .  S  .  G  .  H  .  .  A  V  .  .  .  .  .  .  .  .  .  .  .  .  T  I  .  .  .  .    30
          ...act...t.....g.....cac......g.tgt...........................t......t..ac.a....t.....tt..   90

human      V  F  L  A  I  Y  N  V  T  V  L  G  N  I  G  L  I  V  I  I  K  I  N  P  K  L  H  T  P  M    60
          gtttttctggccatctacaatgtcactgtgctagggaatattgggttgattgtgatcatcaaaatcaaccccaaactgcatacccccatg  180
mouse      I  .  .  T  .  .  S  I  .  .  V  .  .  .  .  M  .  .  .  .  R  .  .  .  .  .  .  I  .  .    60
          a.a.....ca.........gca........g....a........ca.......c..a....g...t..t...........c.t.......  180

human      Y  F  F  L  S  Q  L  S  F  V  D  F  C  Y  S  S  I  I  A  P  K  M  L  V  N  L  V  V  K  D    90
          tacttttcctcagccaactctcctttgtggatttctgctattcctccatcattgctcccaagatgttggtgaaccttgttgtcaaagac  270
mouse      .  .  .  .  .  H  .  .  .  .  .  .  .  .  .  .  .  V  .  .  .  .  .  .  .  .  .  T  M  N    90
          .....c..t........c.....t...........t..t...........tg..............c....a..t..a...aca.tga..  270

human      R  T  I  S  F  L  G  C  V  V  Q  F  F  F  F  C  T  F  V  V  T  E  S  F  L  L  A  V  M  A   120
          agaaccatttcattttaggatgcgtagtacaattctttttcttctgtacctttgtggtcactgaatcctttttattagctgtgatggcc  360
mouse      .  G  .  .  .  V  .  .  L  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  G  .  .  .   120
          ...gg...a......g........t...g...........t........t..c.....a........t...c......ga........t  360

human      Y  D  R  F  V  A  I  C  N  P  L  L  Y  T  V  D  M  S  Q  K  L  C  V  L  L  V  V  G  S  Y   150
          tatgaccgcttcgtggccatttgcaaccctctgctctacacagttgacatgtcccagaaactctgcgtgctgctggttgtgggatcctat  450
mouse      .  .  .  .  .  .  .  R  .  .  .  .  .  .  .  A  .  .  .  R  .  .  A  M  .  .  L  .  .  .   150
          ......a.g..t........cc..........a...........g.c...........gg.....t.cca.......at...........  450
```

OR “I7” orthologs in rat and mouse

```
327 codons

  8 first-position differences
  8 second-position differences
 32 third-position differences
 48 total differences

 15 amino-acid differences
```

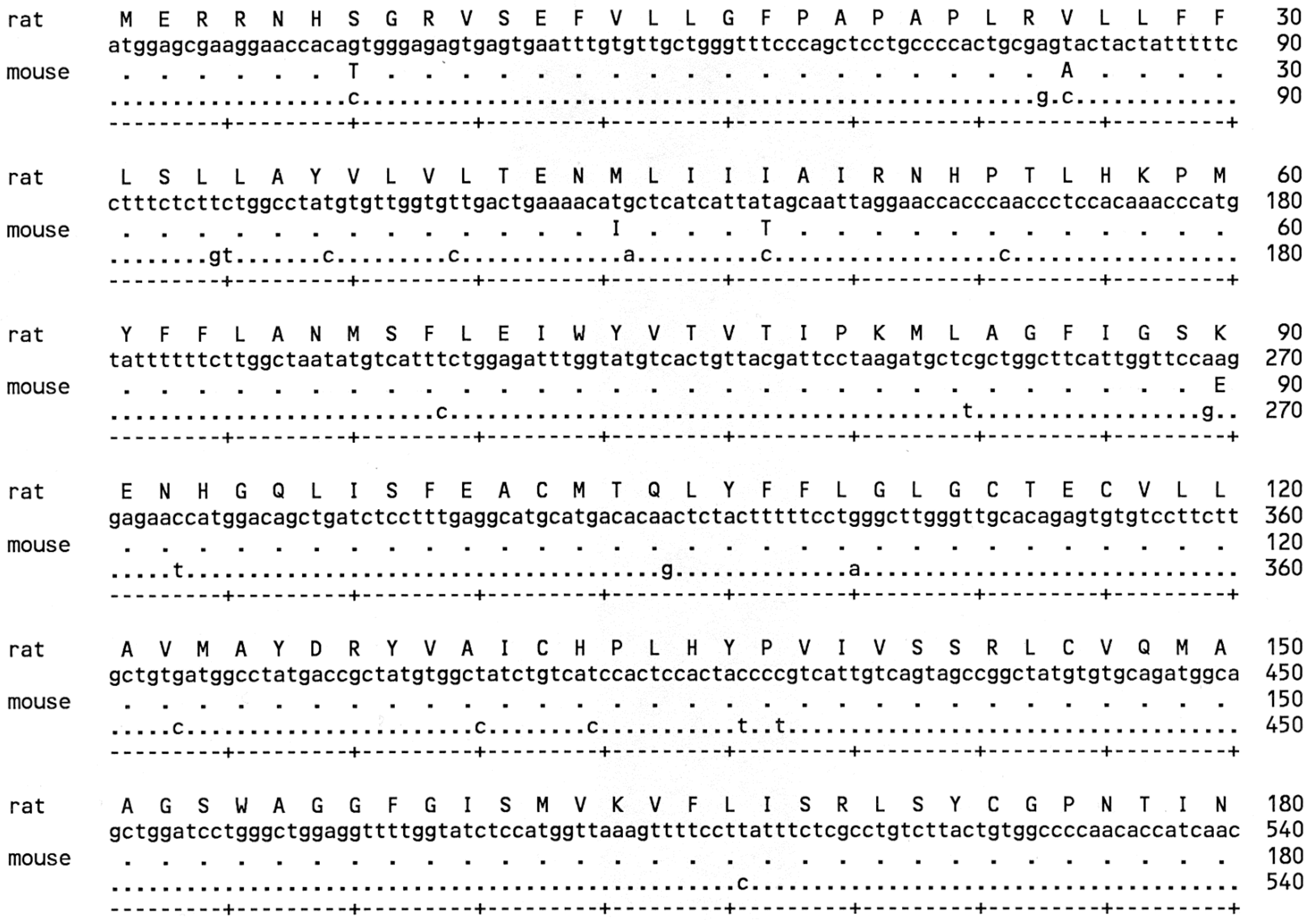
```
Ks = 0.125    Ka = 0.024      Ka/Ks = 0.193    (Ks/Ka =    5.2)
```

Ks: synonymous substitutions per synonymous site

Ka: non-synonymous substitutions per non-synonymous site

In this type of alignment, both the DNA and amino-acid sequences are shown.

For ease of comprehension, sequences after the first one (here rat) are shown as *differences* from the first one. (A dot means “same as in the first sequence”.)



```
rat        M  E  R  R  N  H  S  G  R  V  S  E  F  V  L  L  G  F  P  A  P  A  P  L  R  V  L  L  F  F   30
          atggagcgaaggaaccacagtgggagagtgagtgaatttgtgttgctgggtttcccagctcctgccccactgcgagtactactattttc   90
mouse      .  .  .  .  .  .  T  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  A  .  .  .  .   30
          ...................c.......................................................g.c.............   90
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+

rat        L  S  L  L  A  Y  V  L  V  L  T  E  N  M  L  I  I  I  A  I  R  N  H  P  T  L  H  K  P  M   60
          ctttctcttctggcctatgtgttggtgttgactgaaaacatgctcatcattatagcaattaggaaccacccaaccctccacaaacccatg  180
mouse      .  .  .  .  .  .  .  .  .  .  .  .  .  I  .  .  .  T  .  .  .  .  .  .  .  .  .  .  .  .   60
          ........gt.......c.........c.............a..........c..................c..................  180
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+

rat        Y  F  F  L  A  N  M  S  F  L  E  I  W  Y  V  T  V  T  I  P  K  M  L  A  G  F  I  G  S  K   90
          tatttttcttggctaatatgtcatttctggagatttggtatgtcactgttacgattcctaagatgctcgctggcttcattggttccaag  270
mouse      .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  E   90
          ..........................c.........................................t..................g..  270
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+

rat        E  N  H  G  Q  L  I  S  F  E  A  C  M  T  Q  L  Y  F  F  L  G  L  G  C  T  E  C  V  L  L  120
          gagaaccatggacagctgatctcctttgaggcatgcatgacacaactctacttttttcctgggcttgggttgcacagagtgtgtccttctt  360
mouse      .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  120
          .....t......................................g..............a..............................  360
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+

rat        A  V  M  A  Y  D  R  Y  V  A  I  C  H  P  L  H  Y  P  V  I  V  S  S  R  L  C  V  Q  M  A  150
          gctgtgatggcctatgaccgctatgtggctatctgtcatccactccactaccccgtcattgtcagtagccggctatgtgtgcagatggca  450
mouse      .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  150
          .....c.......................c........c...........t..t....................................  450
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+

rat        A  G  S  W  A  G  G  F  G  I  S  M  V  K  V  F  L  I  S  R  L  S  Y  C  G  P  N  T  I  N  180
          gctggatcctgggctggaggttttggtatctccatggttaaagttttccttatttctcgcctgtcttactgtggccccaacaccatcaac  540
mouse      .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  180
          ..................................................c.......................................  540
          ---------+---------+---------+---------+---------+---------+---------+---------+---------+
```

A central prediction of the Neutral Theory:

The overall rate of molecular evolution should be roughly proportional to the mutation rate, other things being equal.

Here are the five bands in the *human-chimp* I7 alignment where the nucleotide differences (just 7 of them) occur.

In this nuclear gene:

7 nt diffs in 981 bp =
0.71 %

Mitochondrial genome:

1305 nt diffs in 15.5kb =
8.44%

I7 orthologs in human and chimpanzee

```
327 codons
  2 first-position differences
  1 second-position differences
  4 third-position differences
  7 total differences

Human       M  E  W  R  N  H  S  G  R  V  S  E  F  V  L  L  G  F  P  A    20
           atggagtggcggaaccatagtgggagagtgagtgagtttgtgttgctgggcttccctgct   60
chimp       .  .  .  .  .  .  .  .  I  .  .  .  .  .  .  .  .  .  .  .    20
           .........................t..................................   60
           ---------+---------+---------+---------+---------+---------+

Human       Y  F  F  L  A  N  M  S  F  L  E  I  W  Y  V  T  V  T  I  P    80
           tactttttctagctaatatgtcctttctggagatctggtatgtcactgtcactattccc  240
chimp       .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    80
           .....c......................................................  240
           ---------+---------+---------+---------+---------+---------+

Human       G  C  M  T  Q  L  Y  F  F  L  G  L  G  C  T  E  C  V  L  L   120
           ggatgcatgacacagctctacttttttccttggcttgggctgcactgagtgtgtccttctc  360
chimp       .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   120
           .................g..........................................  360
           ---------+---------+---------+---------+---------+---------+

Human       S  M  V  K  V  F  L  I  S  G  L  S  Y  C  G  P  N  I  I  N   180
           tccatggtcaaagtttttcttatttctggcctctcttactgtggccccaacatcatcaac  540
chimp       .  .  .  .  .  .  .  .  .  R  .  .  .  .  .  .  .  .  .  .   180
           ...........................c................................  540
           ---------+---------+---------+---------+---------+---------+

Human       K  A  F  S  T  C  A  S  H  L  T  V  V  I  I  F  Y  A  A  S   260
           aaggccttttccacctgtgcctctcatctcactgttgtgataatcttctatgcagccagt  780
chimp       .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  S  .  .   260
           ..................................................ct........  780
           ---------+---------+---------+---------+---------+---------+

Human       E  V  K  R  A  L  C  C  T  L  H  L  Y  Q  H  Q  D  P  D  P   320
           gaggtcaagagagccctatgctgtactctgcacctgtaccagcaccaggatcctgacccc  960
chimp       .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   320
           .................c..........................................  960
           ---------+---------+---------+---------+---------+---------+b
```

“Molecular clocks” keep time (not precisely, but remarkably well)

Rat and mouse last had a common ancestor around 15 million years ago (mya).

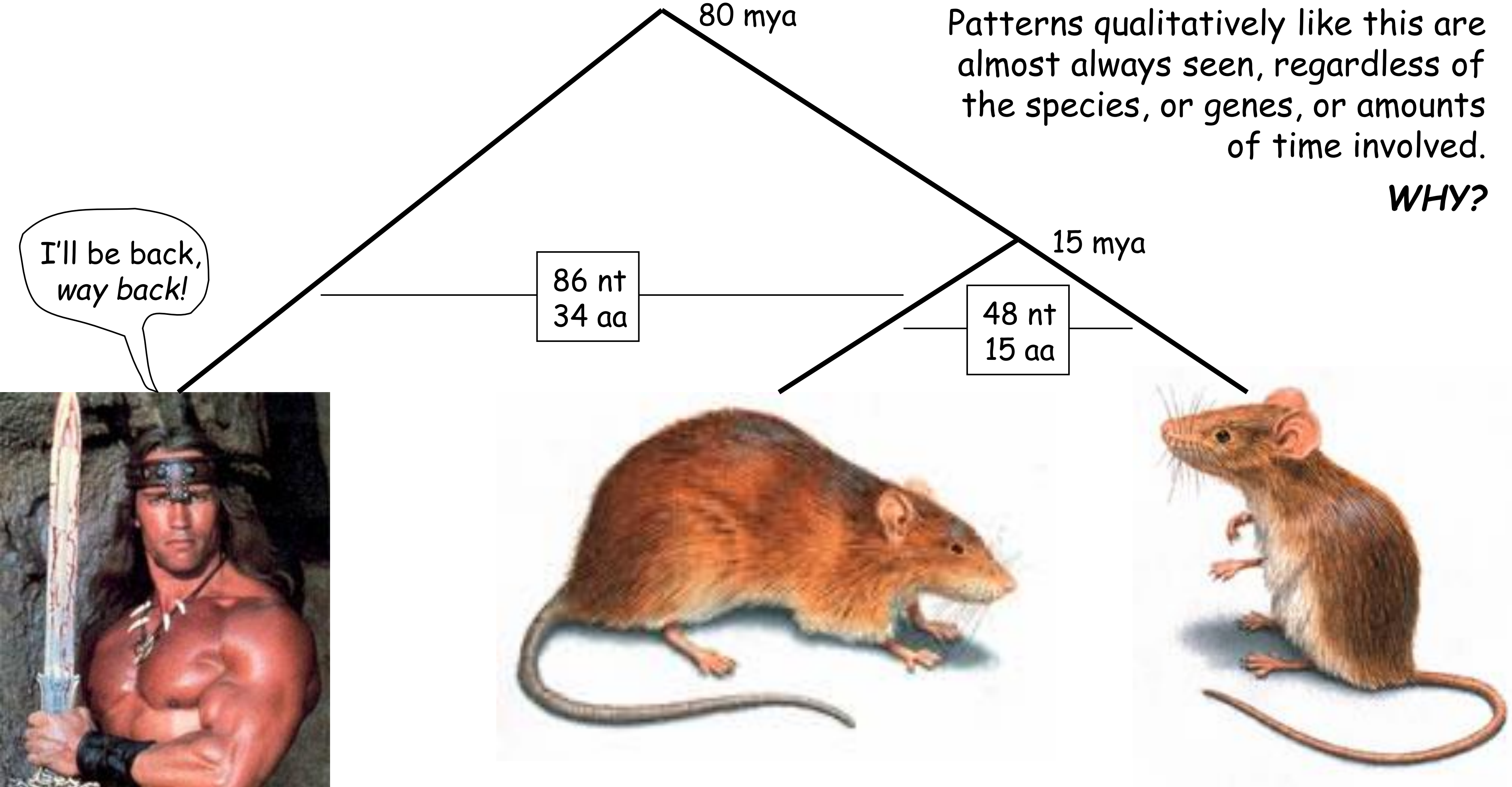
Their I7 genes differ at 48/981 nucleotide positions, and the I7 proteins encoded by those genes differ at 15/327 amino-acid positions.

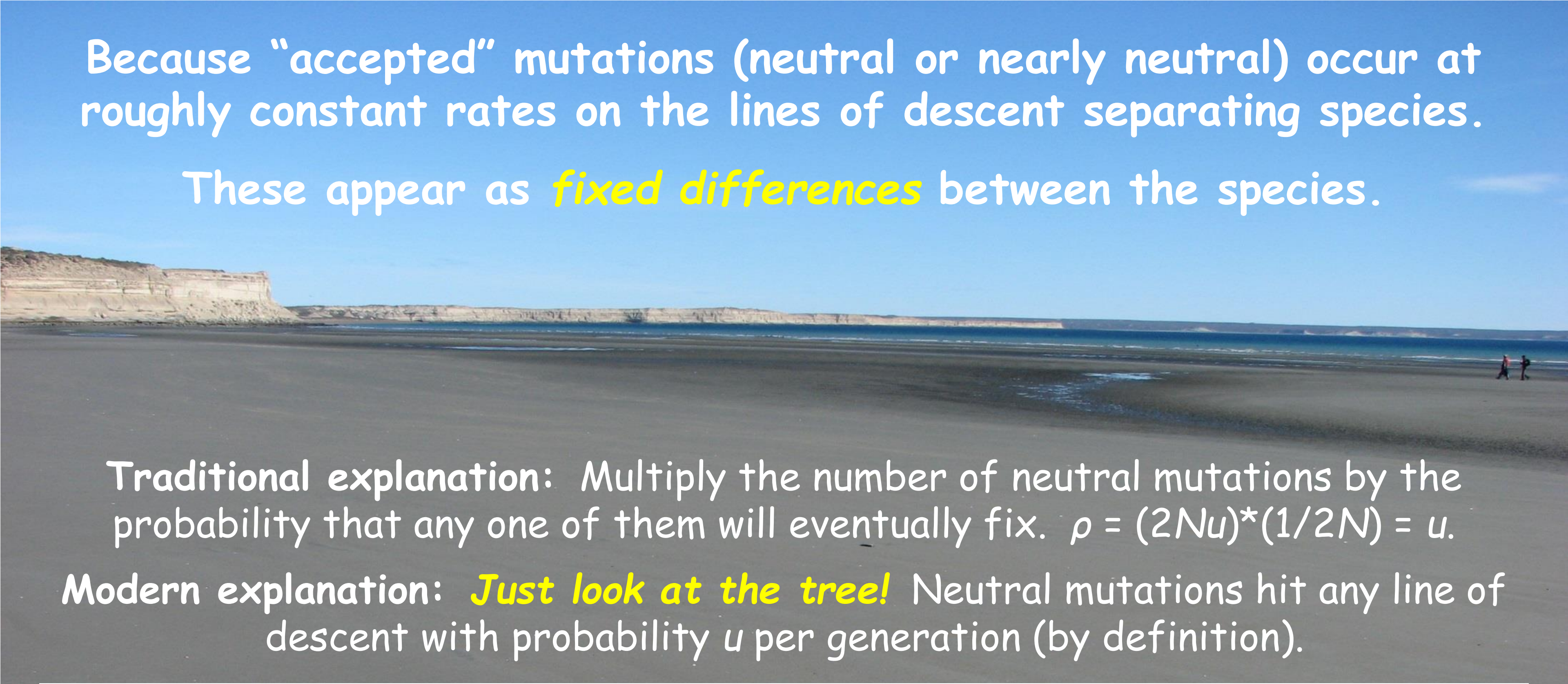
Humans and rodents last had a common ancestor around 80 mya.

Their I7 genes differ by around 86 nucleotides and 34 amino acids, on average.

Patterns qualitatively like this are almost always seen, regardless of the species, or genes, or amounts of time involved.

WHY?





Because "accepted" mutations (neutral or nearly neutral) occur at roughly constant rates on the lines of descent separating species.

These appear as *fixed differences* between the species.

Traditional explanation: Multiply the number of neutral mutations by the probability that any one of them will eventually fix. $\rho = (2Nu)*(1/2N) = u$.

Modern explanation: ***Just look at the tree!*** Neutral mutations hit any line of descent with probability u per generation (by definition).

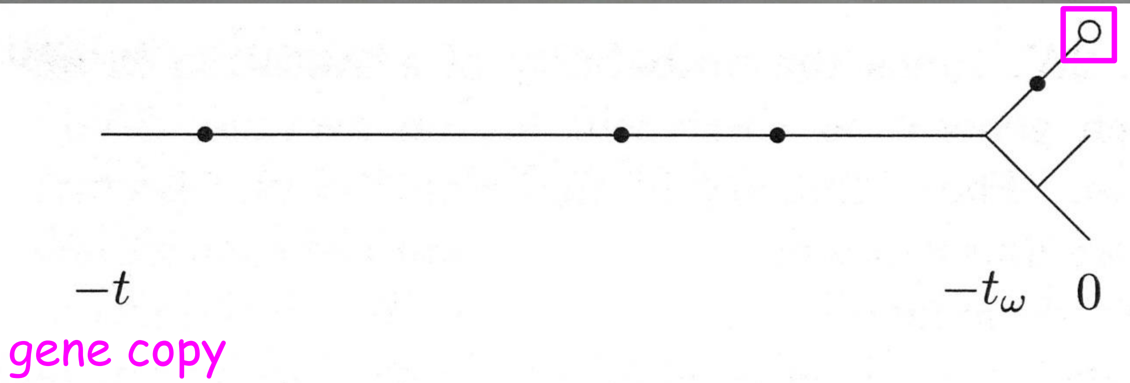
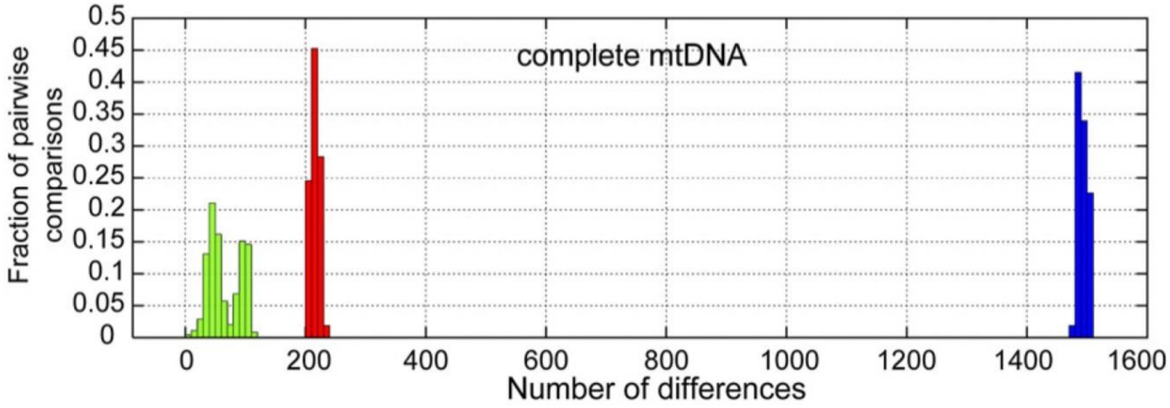


Figure 2.4: The allele picked at random from the population at time zero is indicated by the open circle. The closed circles represent mutations on the lineage. The first three mutations are substitutions; the fourth mutation is polymorphic.

Back to Question #1:

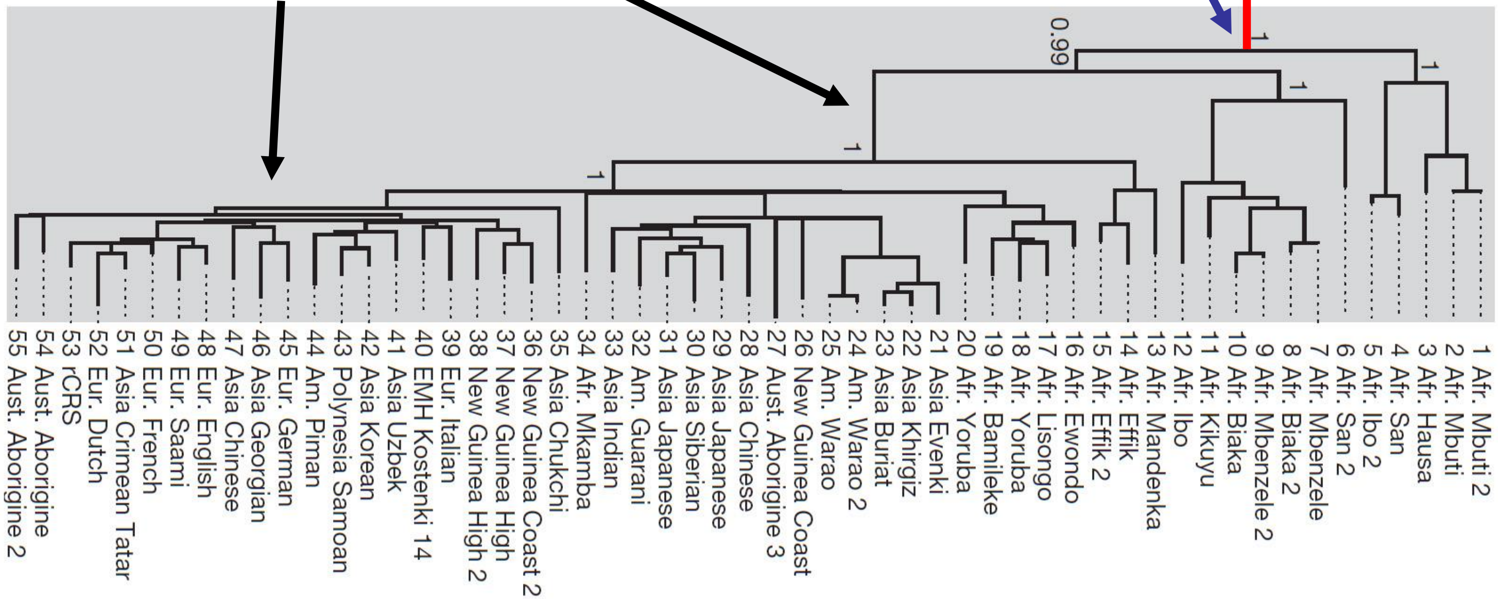
How can the variation among modern humans be greater than the variation between those same humans and a Neanderthal or a chimp?



But TIMES of separation VARY greatly for pairs of modern mitochondria.

TIME from here to tips, and E(# of diffs), is *also* the SAME in every case. So the N-m variation is **purely mutational**.

~180/210 differences are *all the same* (fixed) between N and moderns.



Krause *et al.* (2010)
Nature **464**:894-897

Neanderthal

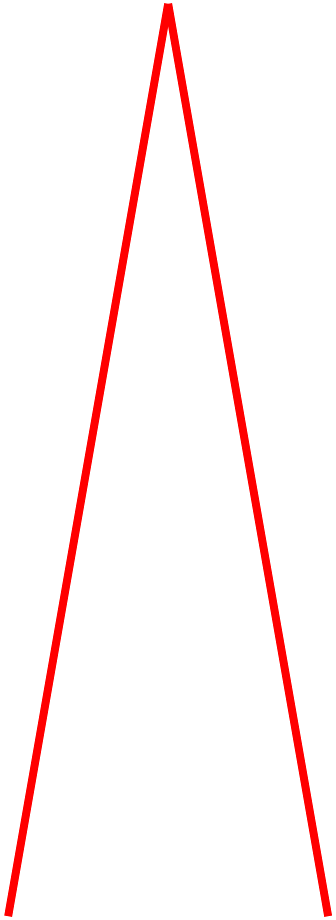
Distribution of k (muts)

2 :	1
3 :	3
4 :	26
5 :	43
6 :	63
7 :	96
8 :	119
9 :	111
10 :	125
11 :	106
12 :	106
13 :	68
14 :	52
15 :	38
16 :	17
17 :	14
18 :	3
19 :	3
20 :	1
21 :	3
22 :	1
25 :	1

mean = 9.9

var = 10.2

1000 trees
all L=10



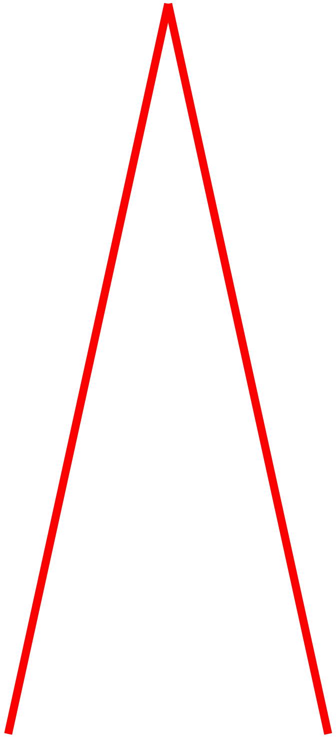
Distribution of k (muts)

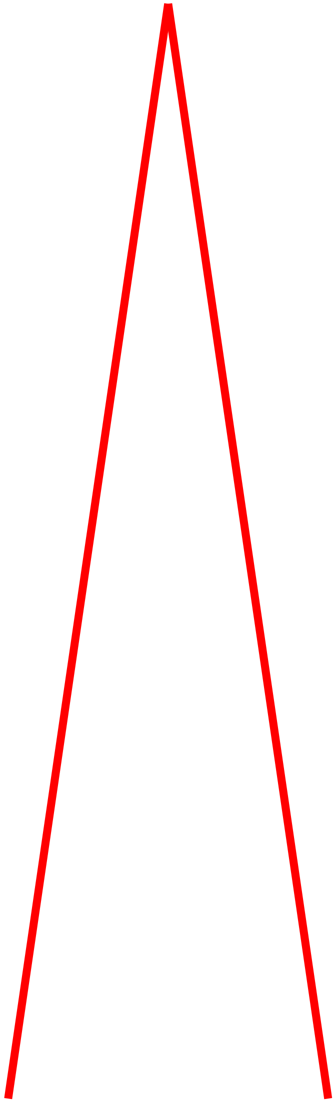
1 :	1
2 :	4
3 :	22
4 :	32
5 :	48
6 :	62
7 :	93
8 :	96
9 :	111
10 :	108
11 :	105
12 :	72
13 :	62
14 :	61
15 :	33
16 :	39
17 :	15
18 :	21
19 :	7
20 :	5
21 :	1
22 :	1
25 :	1

mean = 10.0

var = 13.9

1000 trees,
half L=8 half L=12



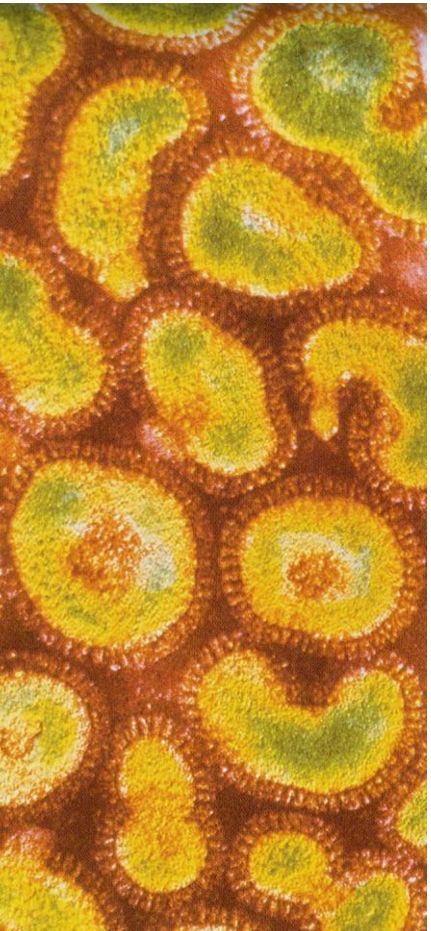


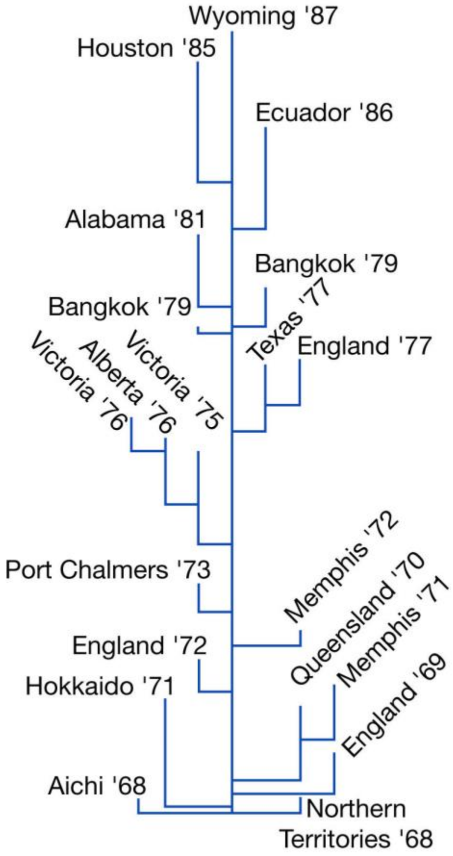
The variance of 8 and 12 is 4!

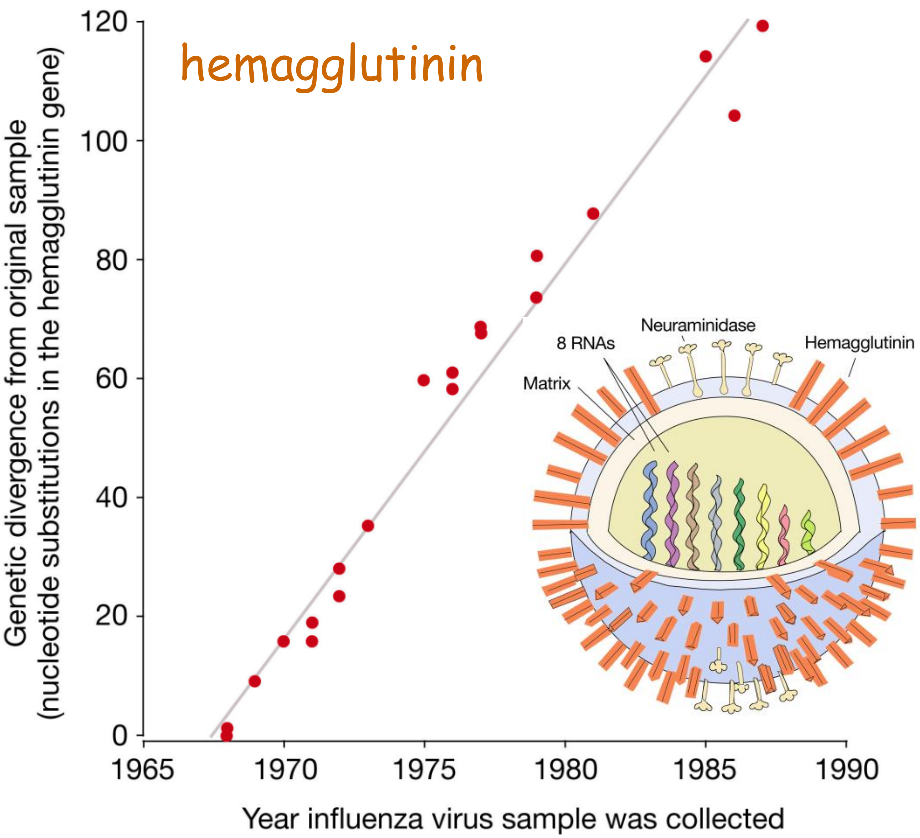
How can we calibrate molecular clocks?

The flu-virus clock has been calibrated directly, by analyzing viruses sampled at many times during the last several decades.

These data for the virus's hemagglutinin gene show a steady accumulation of nucleotide substitutions over a period of more than 20 years.







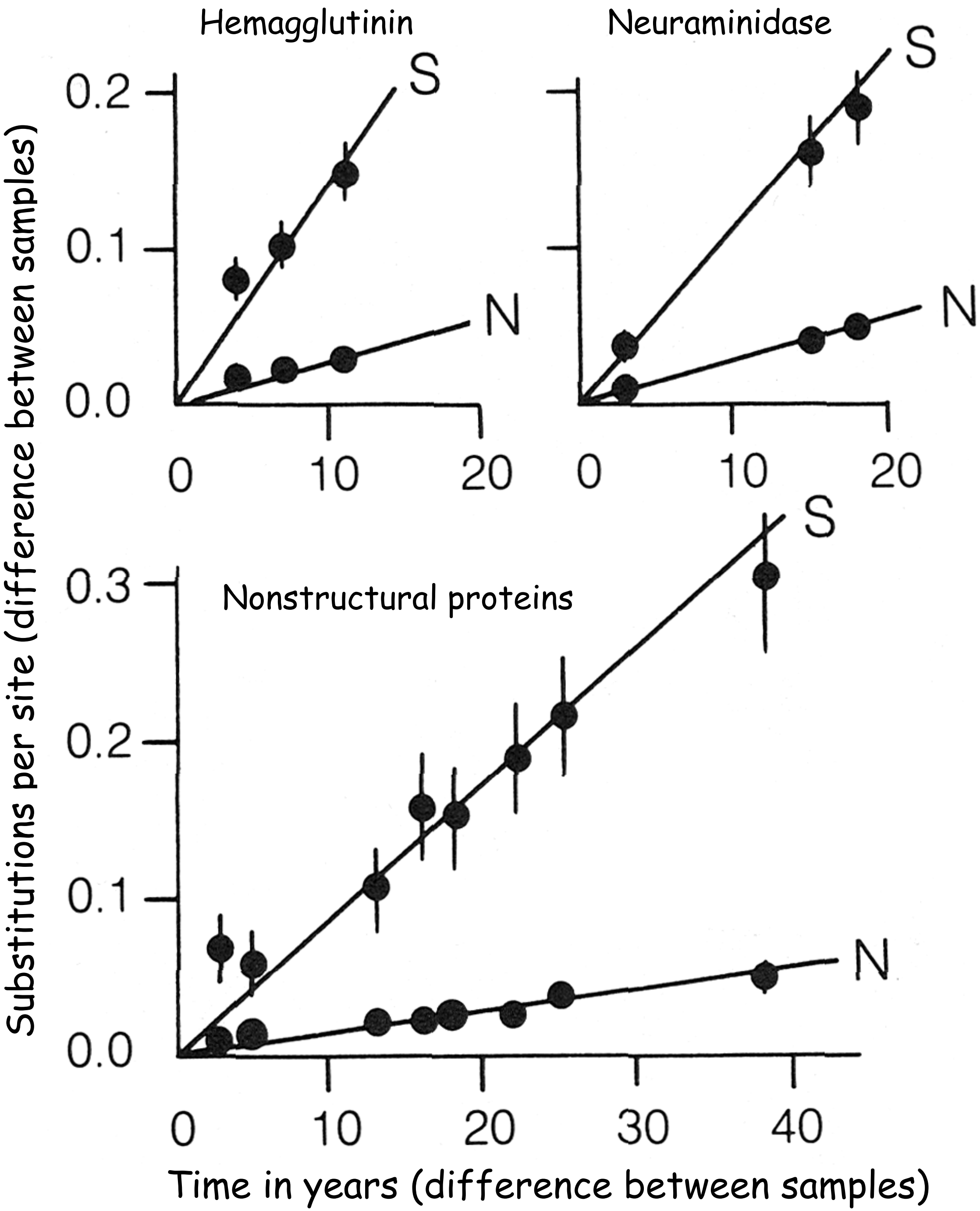
These data for several genes show higher rates for the surface-expressed hemagglutinin and neuraminidase genes than for nonstructural proteins, and higher rates for synonymous (S) than for nonsynonymous (N) substitutions.

The apparent rates of synonymous substitution per synonymous site per year are 0.014, 0.011, 0.009.

The rates of nonsynonymous substitution per nonsynonymous site per year are are 0.0029, 0.0028, and 0.0015.

Thus the synonymous sites evolve around five times as fast as the nonsynonymous sites.

But ***either*** kind of site could be used as a molecular clock, as could any of the genes.



Calibrating the molecular clock "retrospectively"

If substitutions occur at a more or less constant rate, then the total molecular ***divergence*** is simply the ***product*** of the elapsed ***time*** and the ***rate of substitution***.

It follows that if we know any ***two*** of these quantities, we can infer the ***other one!***

The divergence (***K***) is our primary observation, from alignments of present-day sequences.

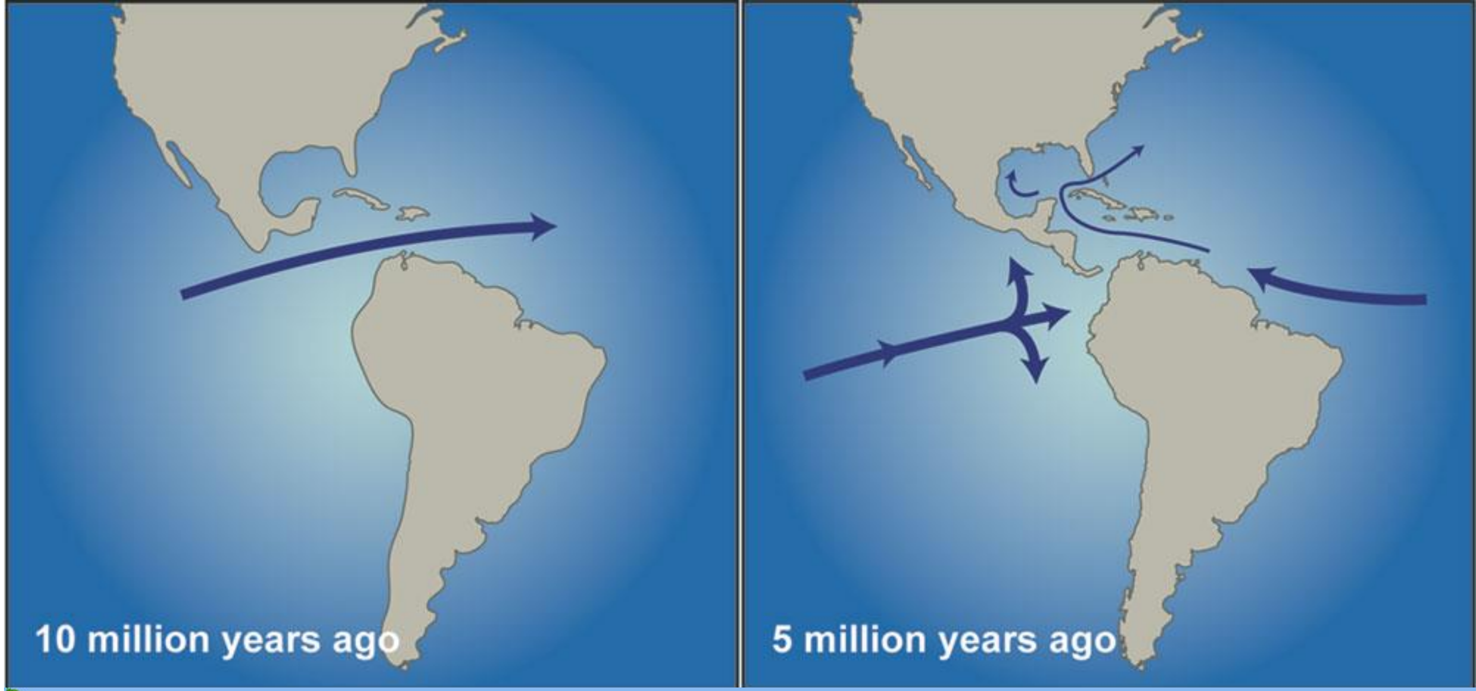
Sometimes we can also know the time (***T***), from fossils or other geological events.

Then we can ***estimate*** the rate of substitution (***μ***).

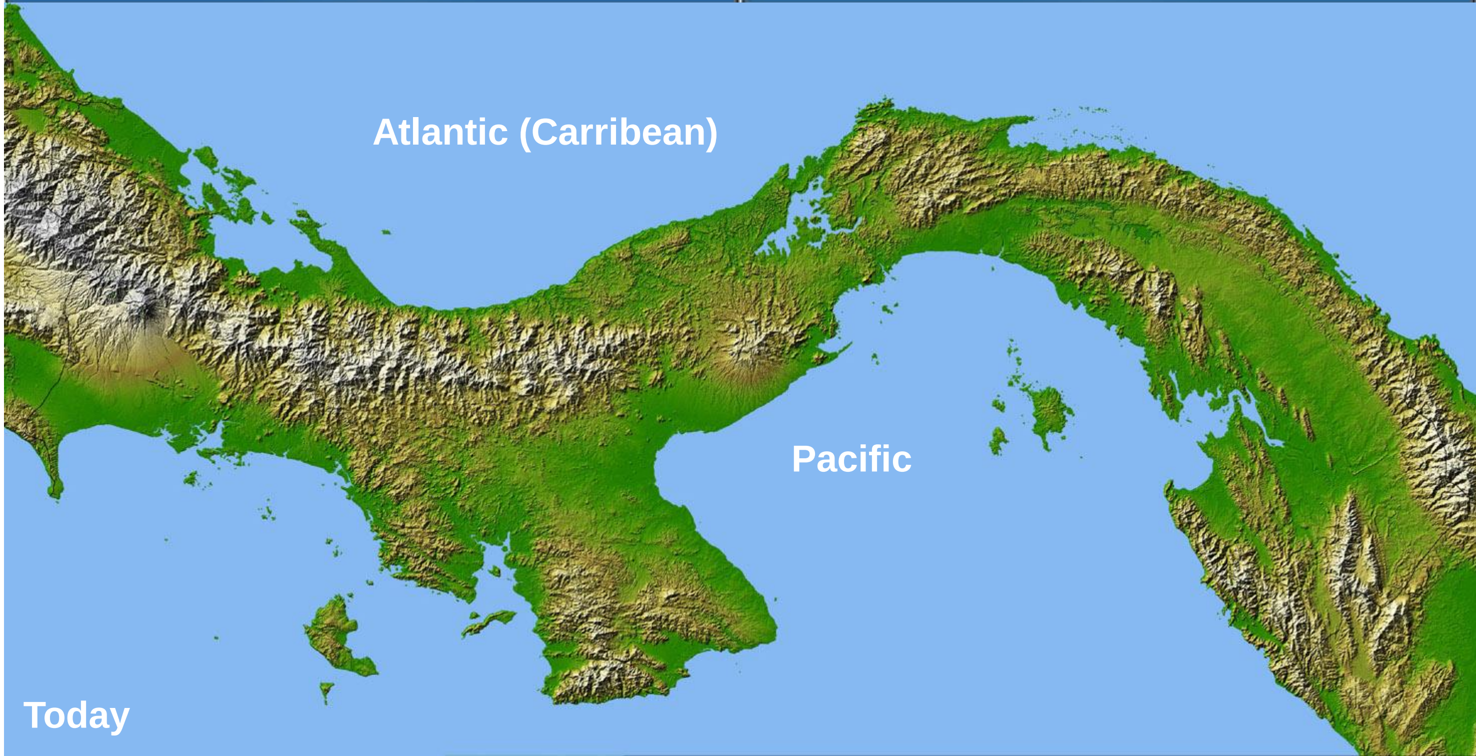
A snapping shrimp (*Alpheus*)

NICARAGUA
COSTA RICA
PANAMA
COLOMBIA
PACIFIC OCEAN
Golfo de Panamá
Golfo del Darién
Golfo de los Mosquitos
Lago de Nicaragua
Managua
Granada
Bluefields
SAN JOSE
Limón
Colón
Barranquilla
Cartagena
Santa Marta

The Isthmus of Panama emerged as a wrinkle in the earth's crust during the Miocene, as the South American Plate pushed into the North American Plate.



Epoch	Age Ma
Holocene	
Pleistocene	1.8
Pliocene	5.2
Miocene	23.8
Oligocene	33.5
Eocene	55.6
Paleocene	65

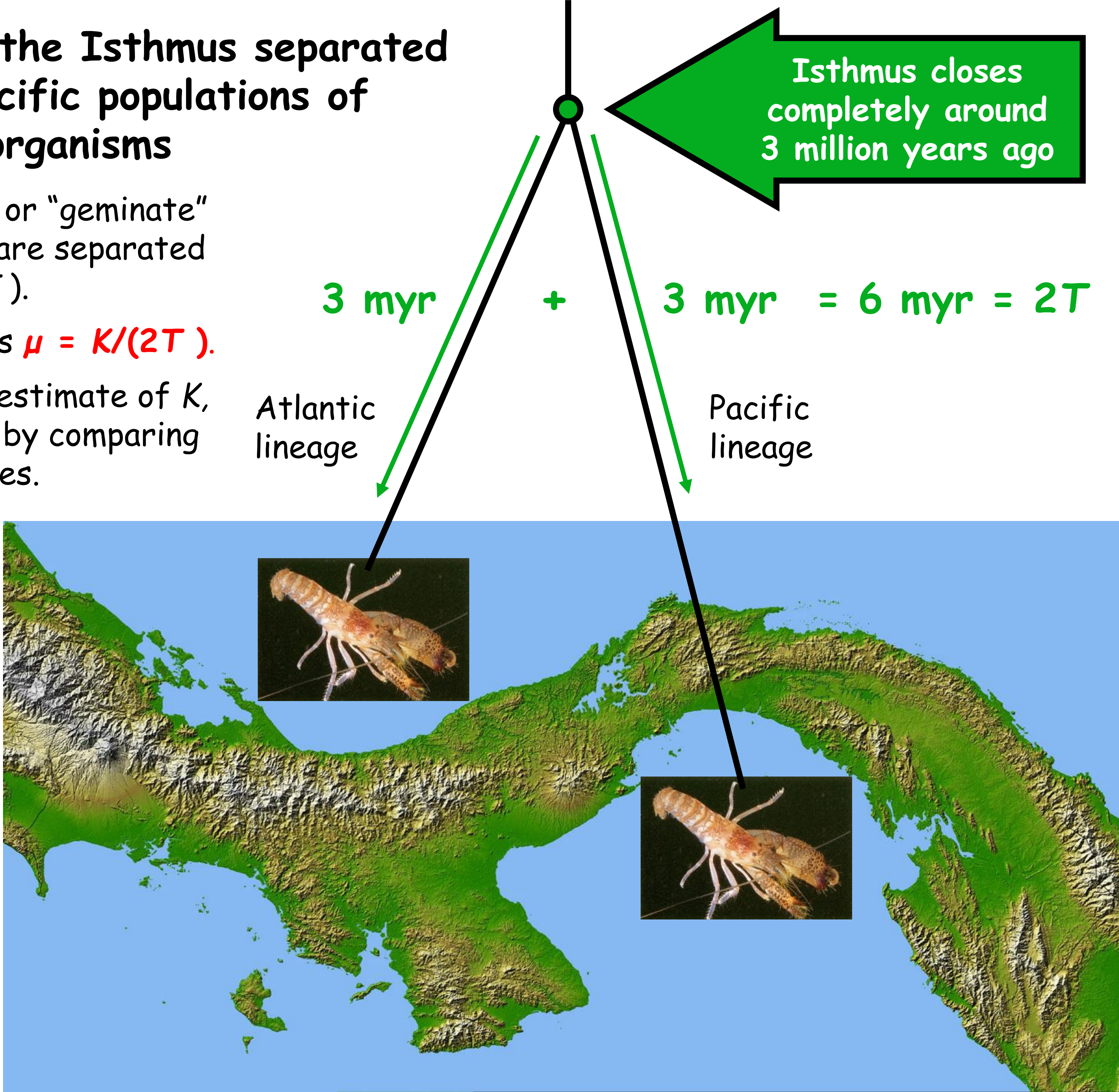


The closure of the Isthmus separated Atlantic and Pacific populations of shallow-water organisms

Today, these sibling or "geminate" (twin) species pairs are separated by 3 million years (T).

$K = 2T\mu$, which means $\mu = K/(2T)$.

So all we need is an estimate of K, which we can obtain by comparing orthologous sequences.



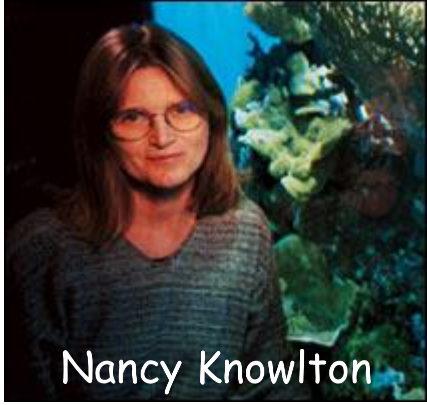
Isthmus closes completely around 3 million years ago

3 myr + 3 myr = 6 myr = 2T

Atlantic lineage

Pacific lineage

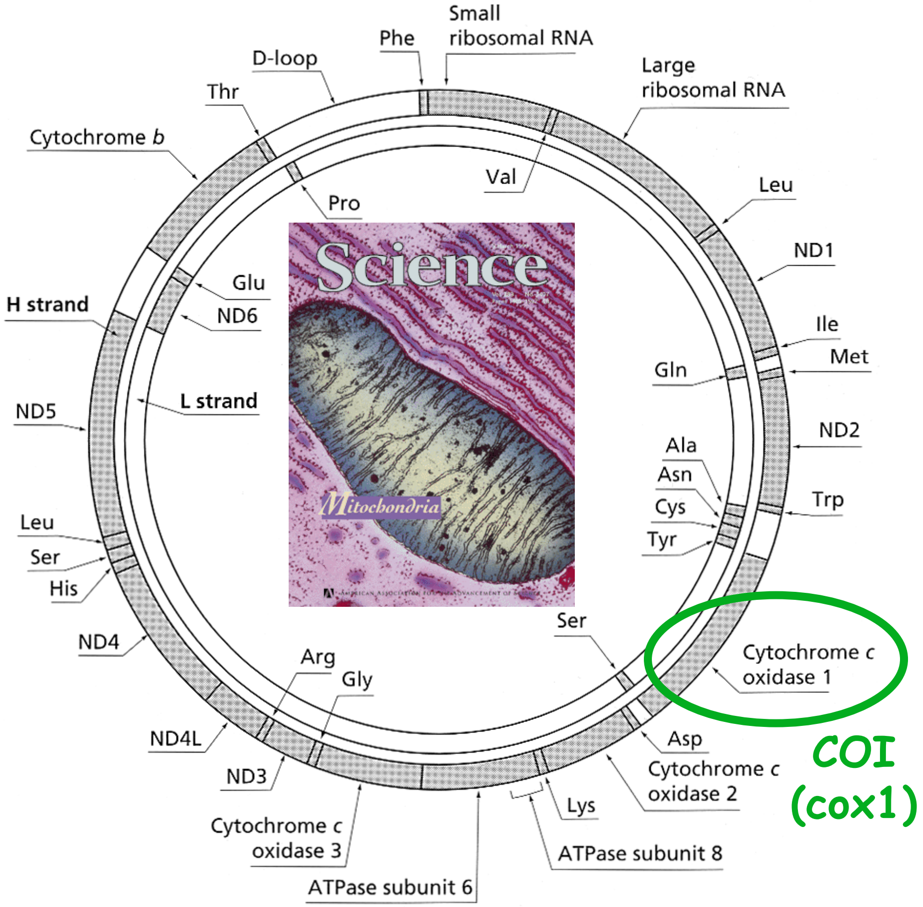
But *which* orthologous sequences, in *which* species?

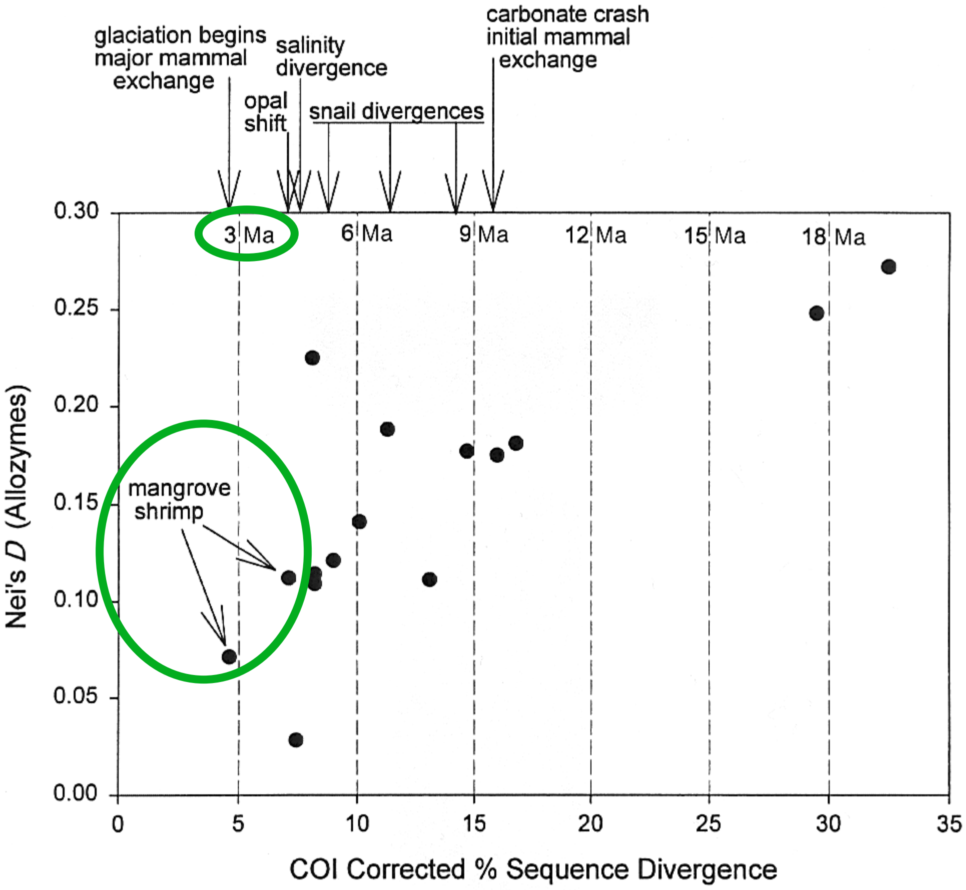


Nancy Knowlton

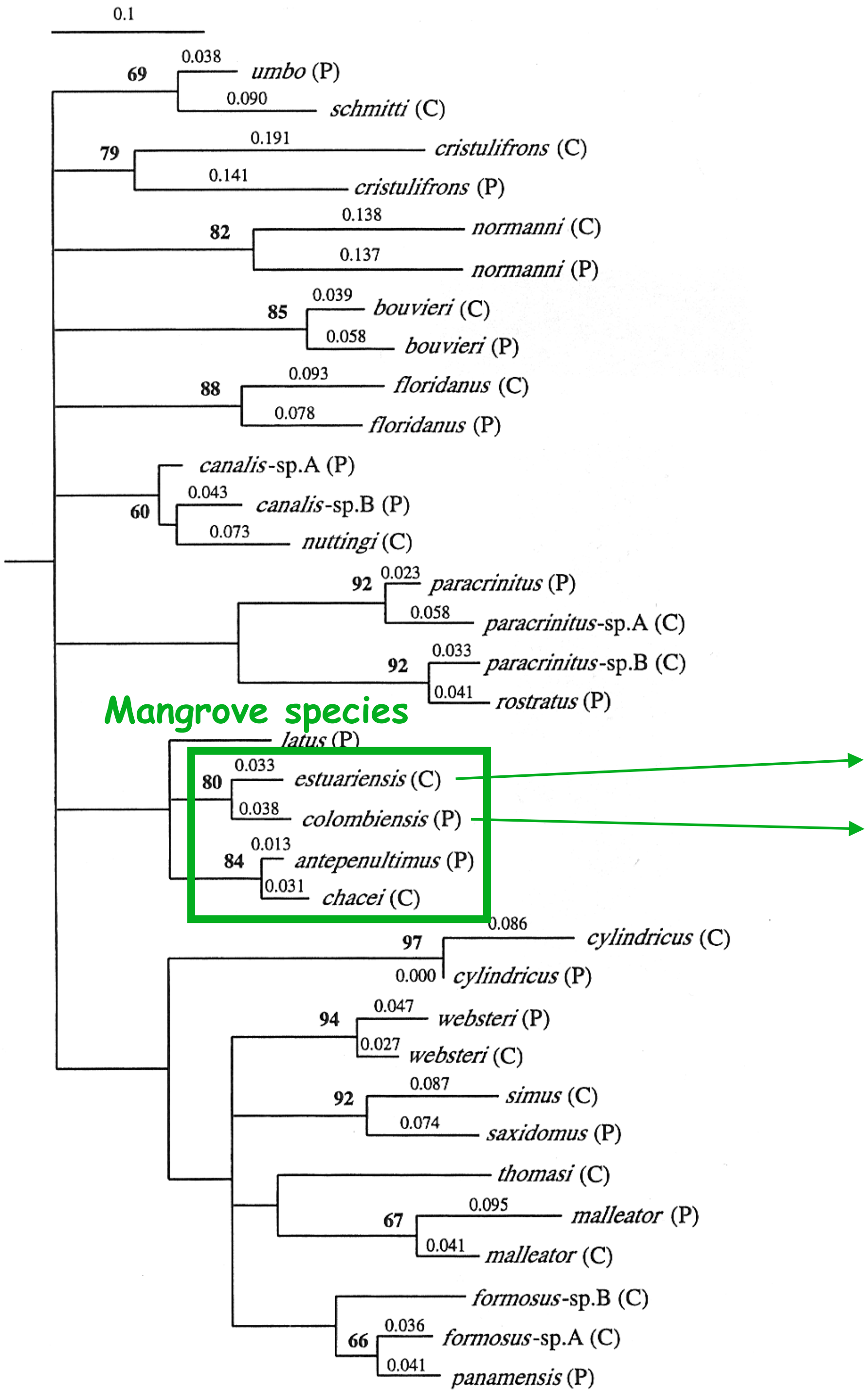
Nancy Knowlton and her colleagues collected many species of snapping shrimp (genus *Alpheus*) from both sides of the Isthmus, and sequenced part of their COI (cox1) genes.

They found much variation in levels of divergence between trans-isthmian sibling species. Those living at greater depths were more diverged than those from shallow, inshore habitats.





From RDM Page and EC Holmes, *Molecular Evolution: A Phylogenetic Approach* (Blackwell, 1998)



Mangrove species

Sibling species pair #1:

A. colombiensis (Pacific)

A. estuariensis (Atlantic/Carribean)

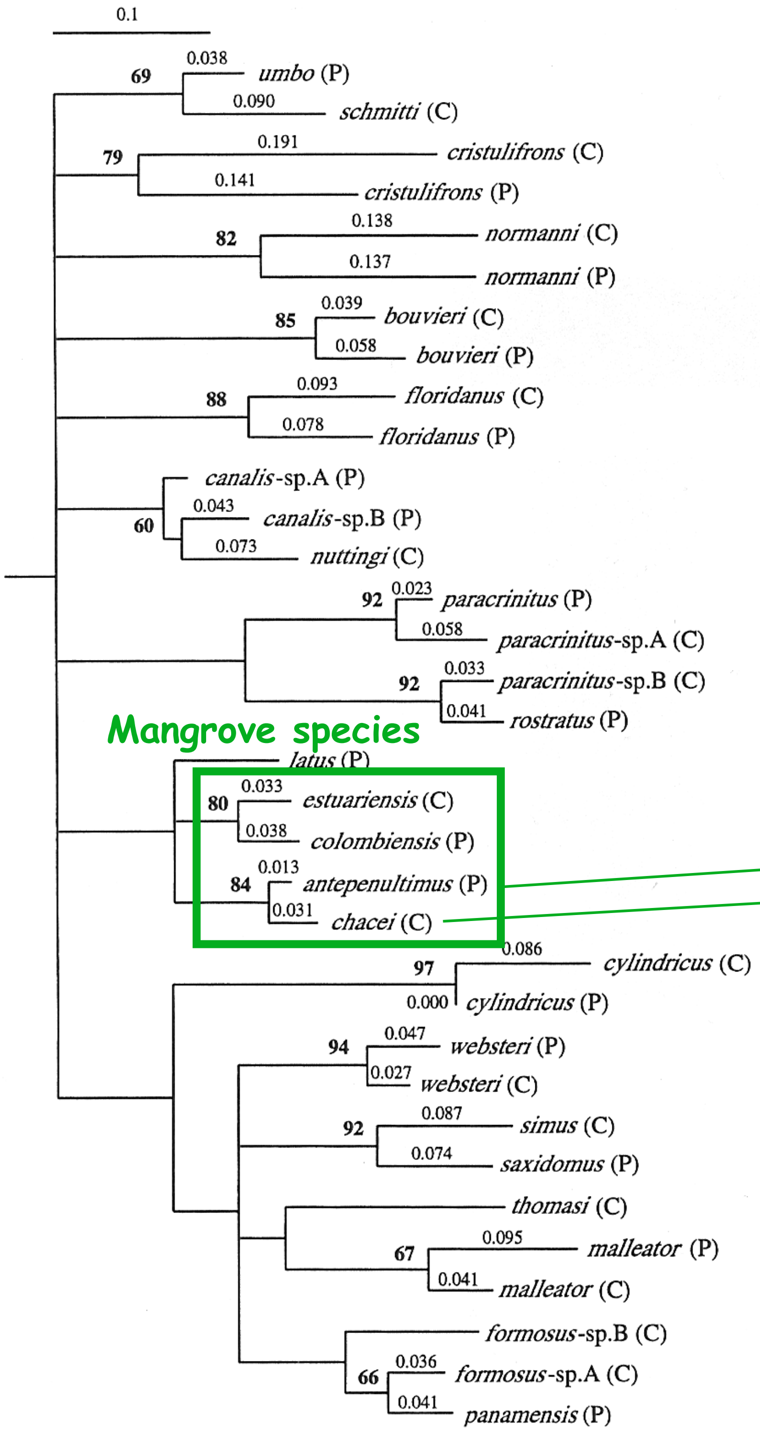
```
 1 first-position differences
 0 second-position differences
32 third-position differences
33 total differences

 0 amino-acid differences

Ks = 0.234    (sd = 0.0443)
Ka = 0.000    (sd = 0.0000)
```

33 synonymous differences

```
Acolo(P)     H  P  E  V  Y  I  L  I  L  P  A  F  G  M  I  S  H  I  I  N    20
            cacccagaagtttatattctaattctaccagctttcggtataatctcccacatcatcaat   60
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    20
            ..................................................t.....t..c   60
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     Q  E  S  G  K  K  E  A  F  G  T  L  G  M  I  Y  x  M  A  A    40
            caagaatcaggaaaaaaagaagcattcggaacattaggaataatctac.n.atggctgca  120
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    40
            .....g...................................g...........a..a...  120
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     I  G  I  L  G  F  V  V  W  A  H  H  M  F  T  V  G  M  D  V    60
            attgggatccttggatttgtagtatgggcccaccacatgttcactgttggaatagatgta  180
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    60
            .....a.....c..........................a..t..............c...  180
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     D  T  R  A  Y  F  T  S  A  T  M  I  I  A  V  P  T  G  I  K    80
            gatacacgagcatacttcacatctgcaactataattattgccgttcccactggaattaaa  240
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    80
            ..............t..t.....c.....a..............................  240
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     I  F  S  W  L  G  T  L  H  G  S  Q  F  T  Y  S  P  S  L  L   100
            attttcagatgattaggaacacttcacggaagacaattcacttacagaccctcactactt  300
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   100
            .........................................c..................  300
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     W  A  L  G  F  V  F  L  F  T  M  G  x  L  T  G  V  V  L  A   120
            tgagcactagggtttgtattttattcacaatagga.n.ctgacaggggtggtcctagct  360
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   120
            .....c...................................a.....a..c.........  360
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     N  S  S  I  D  I  I  L  H  D  T  Y  Y  V  V  A  H  F  H  Y   140
            aactcctcaatcgatatcatcttacacgacacttactacgtagtggcccacttccactac  420
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   140
            ...........t........t.......................c...............  420
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     V  L  S  M  G  A  V  F  G  I  F  A  G  I  A  H  W  F  P  L   160
            gtcctatctataggggcagtattcggaatcttcgcaggtattgcccactgattcccccta  480
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   160
            ........g.....a.....g..........................t.........t..  480
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     F  T  G  L  S  L  N  P  Q  W  L  K  M  H  F  F  T  M  F  I   180
            ttcacaggcctatccctaaacccccaatgacttaaaatacacttttttaccatatttatt  540
Aestu(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   180
            ..t...........t...................................t.........  540
            ---------+---------+---------+---------+---------+---------+

Acolo(P)     G  V  N  I  x  F  F  P    188
            ggagtgaacatc.n.ttcttcccc   564
Aestu(A)     .  .  .  .  .  .  .  .    188
            .....a..................   564
            ---------+---------+----
```

Sibling species pair #2:
A. antepenultimus (Pacific)
A. chacei (Atlantic/Carribean)

22 synonymous differences
1 nonsynonymous difference



Mangrove species

```
 3 first-position differences
 0 second-position differences
20 third-position differences
23 total differences

 1 amino-acid difference

Ks = 0.132    (sd = 0.0301)
Ka = 0.002    (sd = 0.0022)

Aante(P)     H  P  E  V  Y  I  L  I  L  P  A  F  G  M  I  S  H  I  I  N    20
            cacccagaagtttatattctcattctcccagcctttggtataatctcccatattattaac   60
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    20
            ....................t..............c........................   60
            ---------+---------+---------+---------+---------+---------+

Aante(P)     Q  E  S  G  K  K  E  A  x  G  T  L  x  M  I  Y  A  M  A  A    40
            caagagtcaggaaaaaaagaagca.n.ggaaccta.n.ataatctacgctatagccgca  120
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    40
            .....a...........................t.............t...........  120
            ---------+---------+---------+---------+---------+---------+

Aante(P)     I  G  I  L  G  F  V  V  W  A  x  x  M  F  T  V  G  M  D  V    60
            atcggaatcctaggatttgtagtatgagca.n..n.atattcaccgttggaatagacgta  180
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    60
            .........t..................................t...............  180
            ---------+---------+---------+---------+---------+---------+

Aante(P)     D  T  R  A  Y  F  T  S  A  T  M  I  I  A  V  P  T  G  I  K    80
            gatacacgagcatacttcacatcagcaaccataattattgctgttcctaccggaattaaa  240
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    80
            ..c..g.................g...........c.................g......  240
            ---------+---------+---------+---------+---------+---------+

Aante(P)     I  F  S  W  L  G  T  L  H  G  S  Q  F  T  Y  S  P  S  L  L   100
            attttcagatgattaggaacacttcacggaagacaatttacatatagaccctcattactt  300
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   100
            ...........................................................c  300
            ---------+---------+---------+---------+---------+---------+

Aante(P)     W  A  L  G  F  V  F  L  F  T  M  G  G  L  T  G  V  V  L  A   120
            tgggccctaggatttgtgttcctatttacaataggaggtctaacaggagtagtcctagcc  360
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   120
            ...........................................................t  360
            ---------+---------+---------+---------+---------+---------+

Aante(P)     N  S  S  I  D  I  I  L  H  D  T  Y  Y  V  V  A  H  F  H  Y   140
            aactcatcaatcgacattattttacacgatacttattacgtggtagcccacttccactac  420
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   140
            .....c.....t.....c..........................................  420
            ---------+---------+---------+---------+---------+---------+

Aante(P)     V  L  S  M  G  A  V  F  G  I  F  A  G  I  A  H  W  F  P  L   160
            gtcctatctataggagcagtatttggaatcttcgcaggtattgcccactgattcccccta  480
Achac(A)     .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   160
            ...................................g........................  480
            ---------+---------+---------+---------+---------+---------+

Aante(P)     F  T  G  L  S  L  N  P  Q  W  L  K  M  H  F  F  T  M  F  I   180
            ttcacaggactatctttaaaccccaatgacttaaaatacacttctttactatatttatc  540
Achac(A)     .  .  .  V  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .   180
            .........g...................g..c...........................  540
            ---------+---------+---------+---------+---------+---------+

Aante(P)     G  V  N  I  T  F  F  P    188
            ggagtaaatatcacatttttcccc   564
Achac(A)     .  .  .  .  .  .  .  .    188
            ........c..............t   564
            ---------+---------+----
```

The synonymous nucleotide substitutions

	A. antepenultimus	*A. colombiensis*	
	A. chacei	*A. estuariensis*	
A / G	6	10	(Ts, purines)
A / C	1	1	(Transversions)
A / T		2	(Transversions)
G / C		2	(Transversions)
G / T		1	(Transversions)
C / T	15	17	(Ts, pyrimidines)
Totals	22	33	

(plus 1 non-syn transversion between A.ante/A.chac)

Three ways to calibrate the Alpheus COI clock

(1) Use all sites and substitutions, don't distinguish fast and slow sites, don't correct for multiple hits.

The two pairs of sequences differ by 23 and 33 of 564 base pairs (bp).

That's 28/564 = 0.05 substitutions per site (5%) *on average.*

Dividing by 3 MYr, we get a raw divergence of 1.7% per million years.

Along each branch: $\mu = P/2T$ = (0.05 subs/site)/(6 MYr) = 0.0083 subs/site/MYr.

(2) Use synonymous sites and substitutions only.

There are roughly $\frac{1}{4}$(564) = 141 effectively synonymous sites.

The sequences differ by 27.5 *synonymous* substitutions, on average.

Thus P = 27.5/141 = 0.195 subs/site (for *synonymous* substitutions).

Along each branch: $\mu = P/2T$ = (0.195 subs/site)/(6 MYr) = 0.0325 subs/site/MYr.

Or in scientific notation, $\mu = 3.25\text{x}10^{-8}$ subs/site/yr.

This is **four** times as great as the simple estimate (1) that ignored codon structure.

Note that this is an estimate of *Ks* (synonymous substitutions per synonymous site)

(3) Use the Jukes-Cantor correction for multiple hits (to account for failure of the infinite-sites model)

Method (2) shows that the synonymous site divergence is around 20% -- large enough that we expect *multiple hits* at some sites.

The number of mutations along a branch (or branches) will follow a *Poisson* distribution.

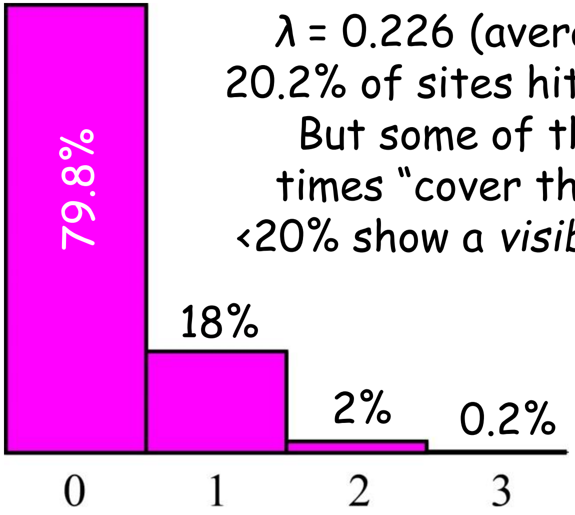
The actual or expected number (*K*) can be anything, but the proportion or probability of different states (*P*) can't exceed 0.75.

The Jukes-Cantor correction extrapolates from the observed pairwise difference (*P*) to the expected total number of substitutions (*K*) : $\mathbf{K = -\frac{3}{4} ln\,(1 - 4P/3)}$

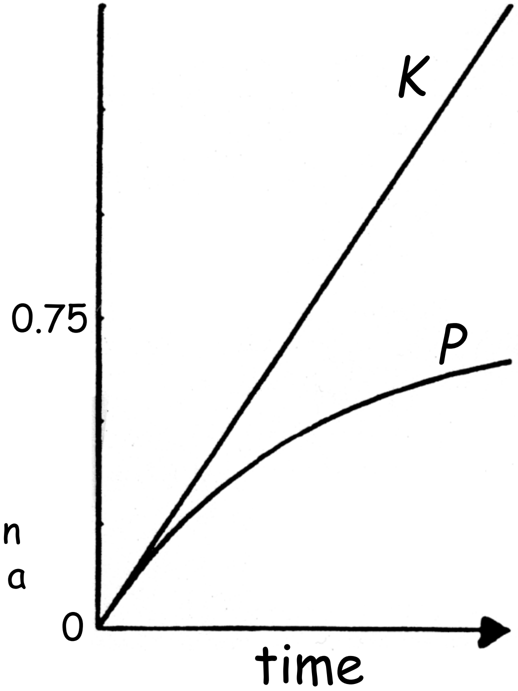
For the snapping-shrimp synonymous sites:
$K = -\frac{3}{4} ln\,(1 - 4*\mathbf{0.195}/3) = \mathbf{0.226}$ subs/site.

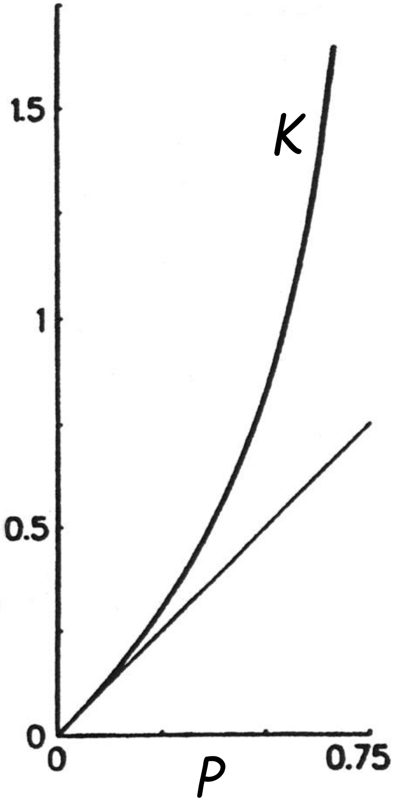
Our estimate of μ therefore increases from 3.25 to 3.8x10^{-8} subs/syn-site/yr.

Caveat: Even this model is simpler than those used in real research, but it makes the ideas clear and does a good job, under "easy" cirumstances like these.



λ = 0.226 (average # of hits).
20.2% of sites hit at least once.
But some of those hit 2 or 3 times "cover their tracks", so <20% show a *visible difference*.





Fully degenerate sites, introns and pseudogenes evolve at neutral rate

(at least in typical mammals)

These average rates for *different kinds of nucleotide positions* in the **nuclear genome** were estimated from alignments of about 50 human and rodent orthologs, assuming a last common ancestor 80 MYA.

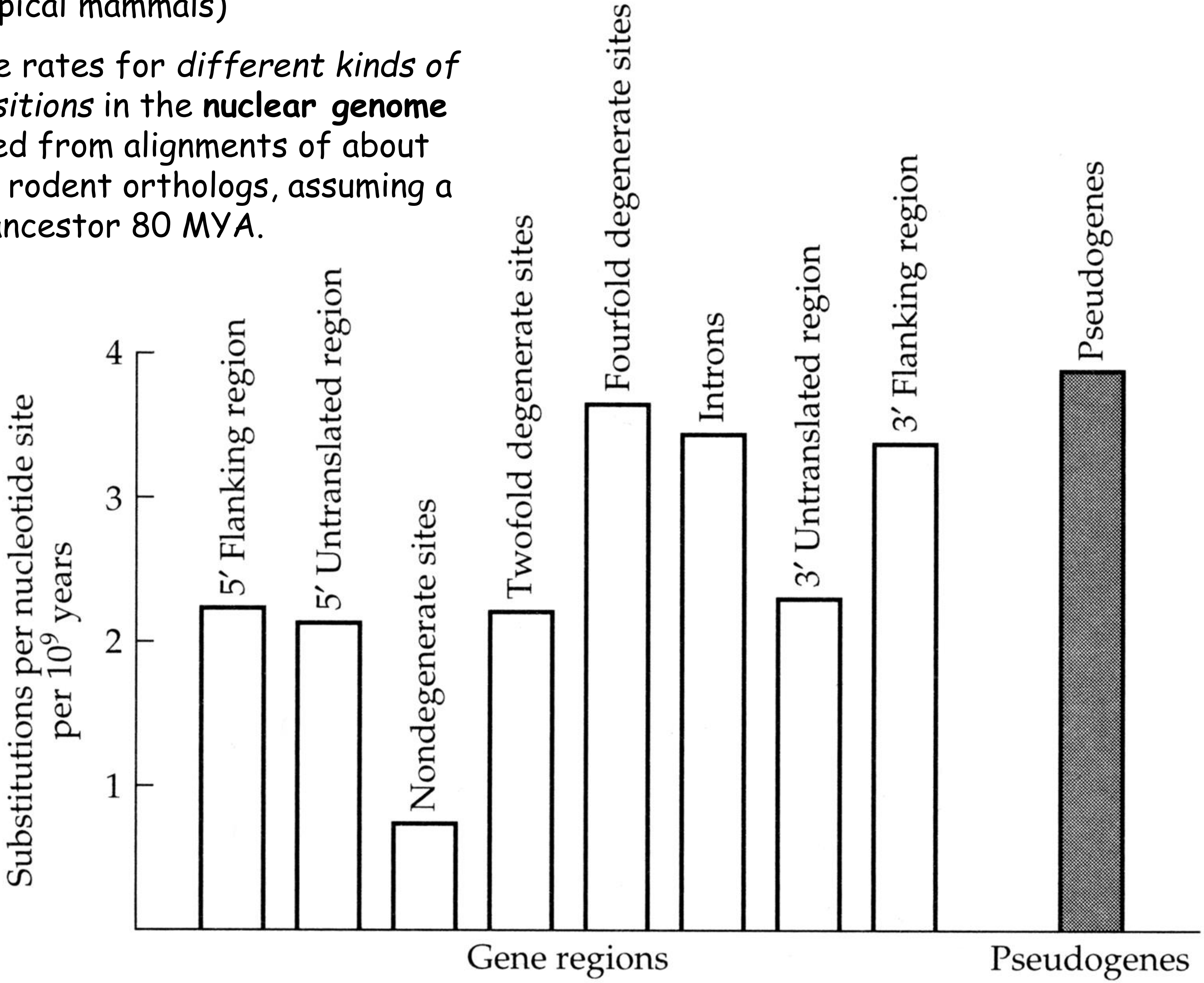


Figure from Grauer & Li, *Fundamentals of Molecular Evolution*, 2nd edn (Sinauer 2000)

What about humans and chimpanzees?

We differ by around 35,000,000 nucleotide substitutions.

Given $3\text{x}10^9$ base pairs per haploid genome, that's roughly 1/86 base pairs, or $K \approx 0.012$ per site.

Fossils suggest a last shared ancestor around $T \approx 6\text{x}10^6$ yr.

Remember, $K = 2T\mu$.

So $\mu = K/2T = 1.2\text{x}10^{-2}/2\text{x}6\text{x}10^6$

$= 1\text{x}10^{-9}/\text{yr}$.

That's a bit lower than the rates estimated for typical mammals.

But we (hominids) have had longer generation times!

Suppose 10-20 years.

Then $\mu \approx 1\text{-}2\text{x}10^{-8}$ hits/site/gen.

1 September 2005 | www.nature.com/nature | $10

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

nature

STAR FORMATION
A massive protostar unveiled

CANCER IMMUNOLOGY
How tumours dupe T cells

AIR POLLUTION
China's NO_2 build-up seen from space

NATUREJOBS
Membrane proteomics

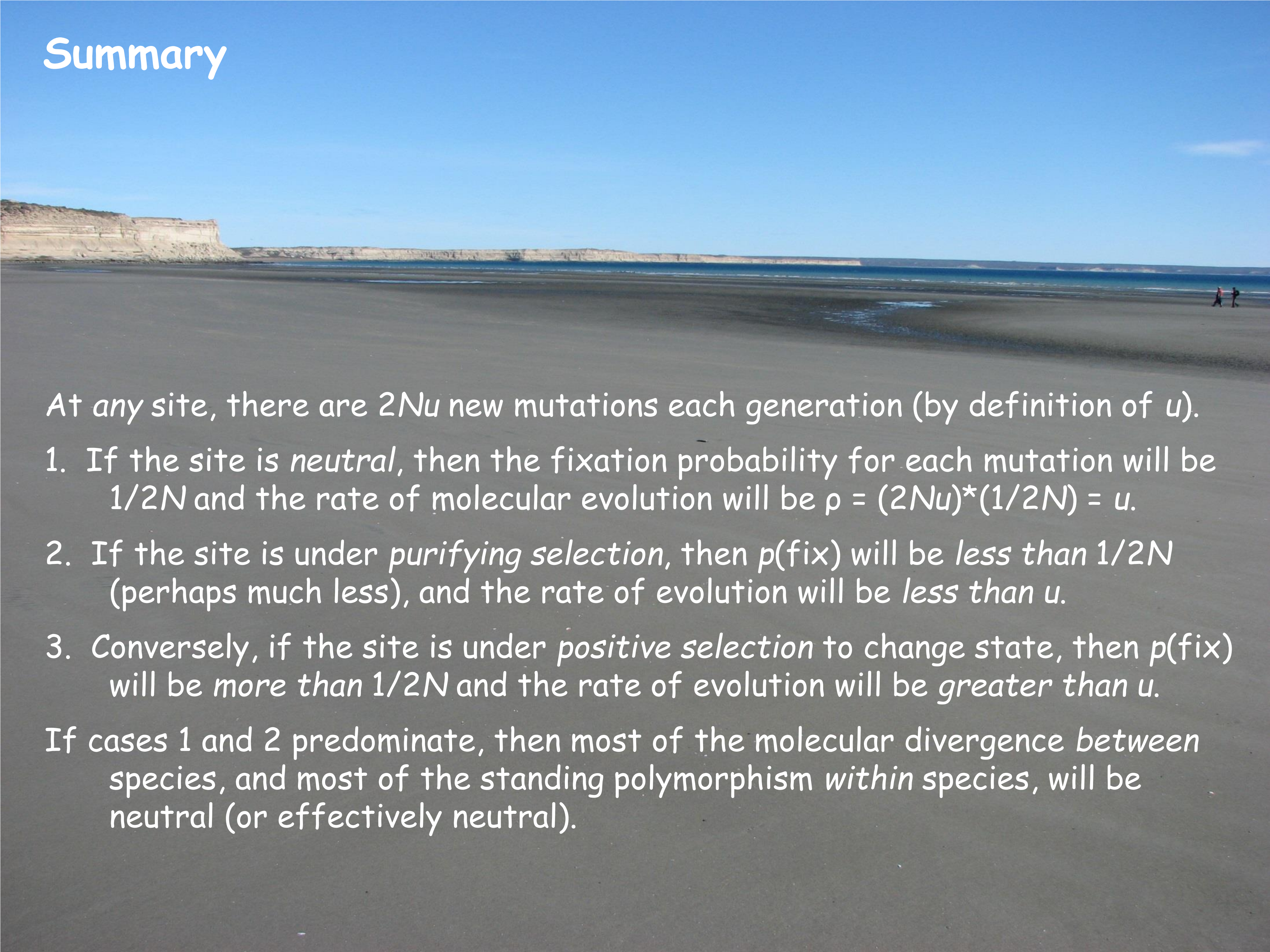
THE CHIMPANZEE GENOME

$10.00US $12.99CAN

35>

0 71486 03070 6

Summary



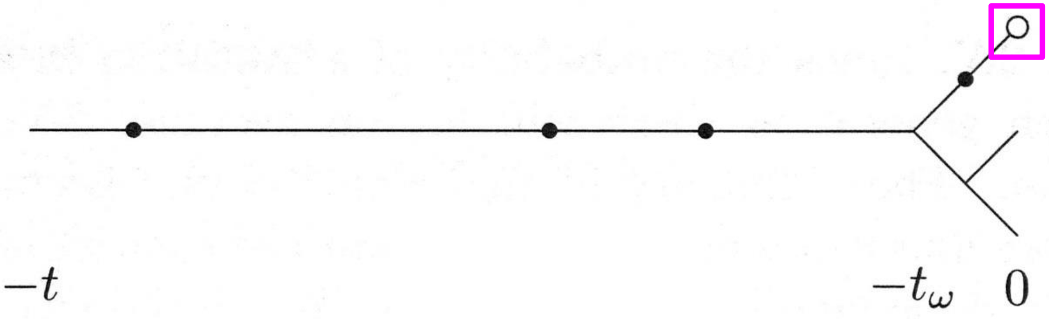
At *any* site, there are $2Nu$ new mutations each *generation* (by definition of u).

1. If the site is *neutral*, then the fixation probability for each mutation will be $1/2N$ and the rate of molecular evolution will be $\rho = (2Nu)*(1/2N) = u$.
2. If the site is under *purifying selection*, then p(fix) will be *less than* $1/2N$ (perhaps much less), and the rate of evolution will be *less than u*.
3. Conversely, if the site is under *positive selection* to change state, then p(fix) will be *more than* $1/2N$ and the rate of evolution will be *greater than u*.

If cases 1 and 2 predominate, then most of the molecular divergence *between* species, and most of the standing polymorphism *within* species, will be neutral (or effectively neutral).

Summary II

Amazingly, selection at neighboring sites does *not* affect the rate of evolution at neutral sites! (That's because the *neutral* mutations had no effect on the survival probabilities of the surviving lineage.)

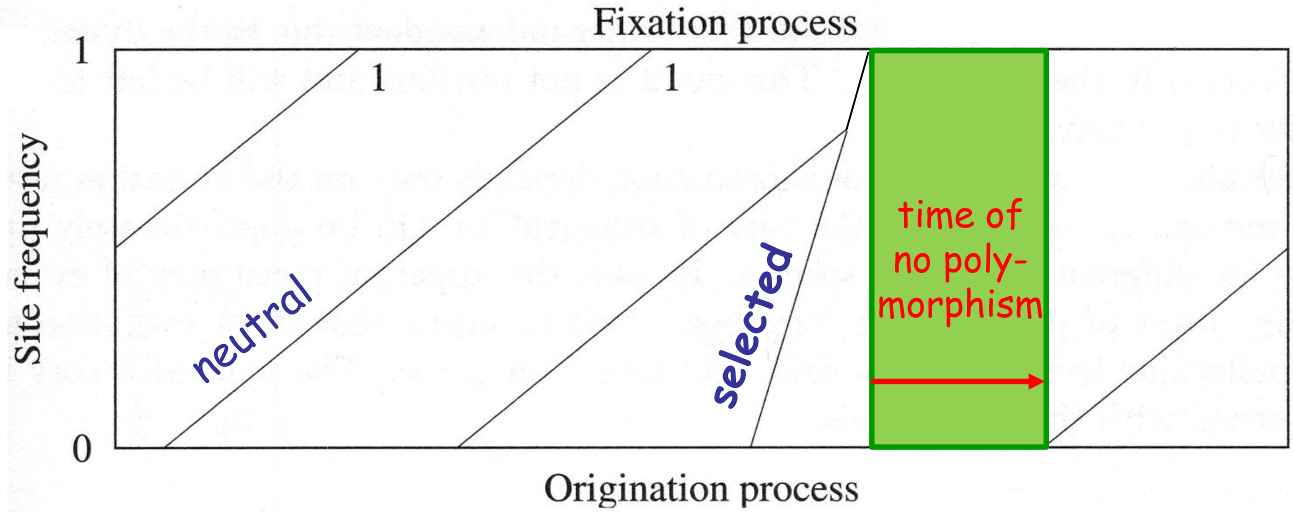


gene copy

Figure 2.4: The allele picked at random from the population at time zero is indicated by the open circle. The closed circles represent mutations on the lineage. The first three mutations are substitutions; the fourth mutation is polymorphic.

Summary III

However, selection at neighboring sites may *greatly* affect the amount of neutral *polymorphism*, and its "shape" (e.g., the site frequency spectrum).



Mutations at this locus occur at a constant rate.

Figure 2.5: A stylized view of molecular evolution in the infinite-sites, no-recombination model.