## What about…

- violation of the simple assumptions?
  - different rates for transitions and transversions (and for different kinds of transitions and transversions)?
  - unequal base frequencies?
  - differences in base composition among lineages?
  - differences in rates among nucleotide sites? (gamma distributed rates)
  - different evolutionary rates in different organisms?
- all this and more!

comparison of 9504 bp of mtDNA sequence between *C. ciconia* and *Ciconia boyciana*

| | A | C | G | T | totals |
|---|---|---|---|---|---|
| A | 2755 | 9 | 44 | 2 | 2810 |
| C | 8 | 3214 | 1 | 80 | 3303 |
| G | 42 | 1 | 1137 | 2 | 1182 |
| T | 3 | 65 | 1 | 2140 | 2209 |
| totals | 2808 | 3289 | 1183 | 2224 | 9504 |

Proportion of sites that differ = 258/9504 = **0.0271**

Apparent transition/transversion ratio = 231/27 = 8.5556 comparison of 9502 bp of mtDNA sequence between *Falco peregrinus* and *Falco sparverius*

|  | A | C | G | T | totals |
|---|---|---|---|---|---|
| A | 2683 | 49 | 120 | 27 | 2879 |
| C | 51 | 2879 | 5 | 262 | 3197 |
| G | 104 | 3 | 1040 | 2 | 1149 |
| T | 15 | 299 | 3 | 1960 | 2277 |
| totals | 2853 | 3230 | 1168 | 2251 | 9502 |

Proportion of sites that differ = 940/9502 = **0.0989**

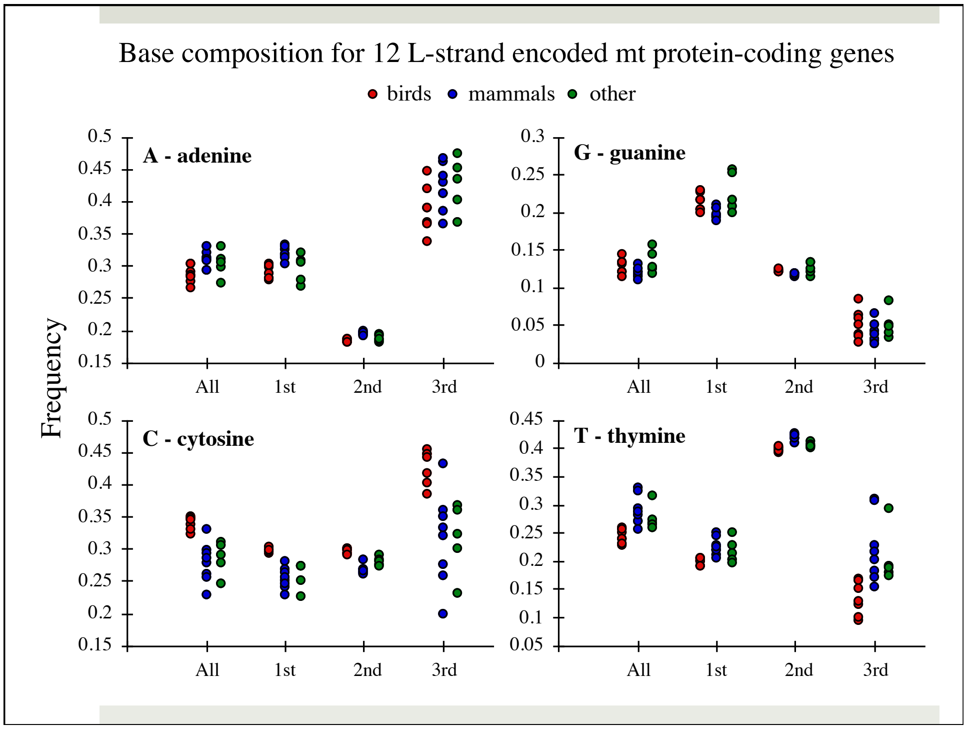
Apparent transition/transversion ratio = 785/155 = 5.0645

## Kimura 2-Parameter Model

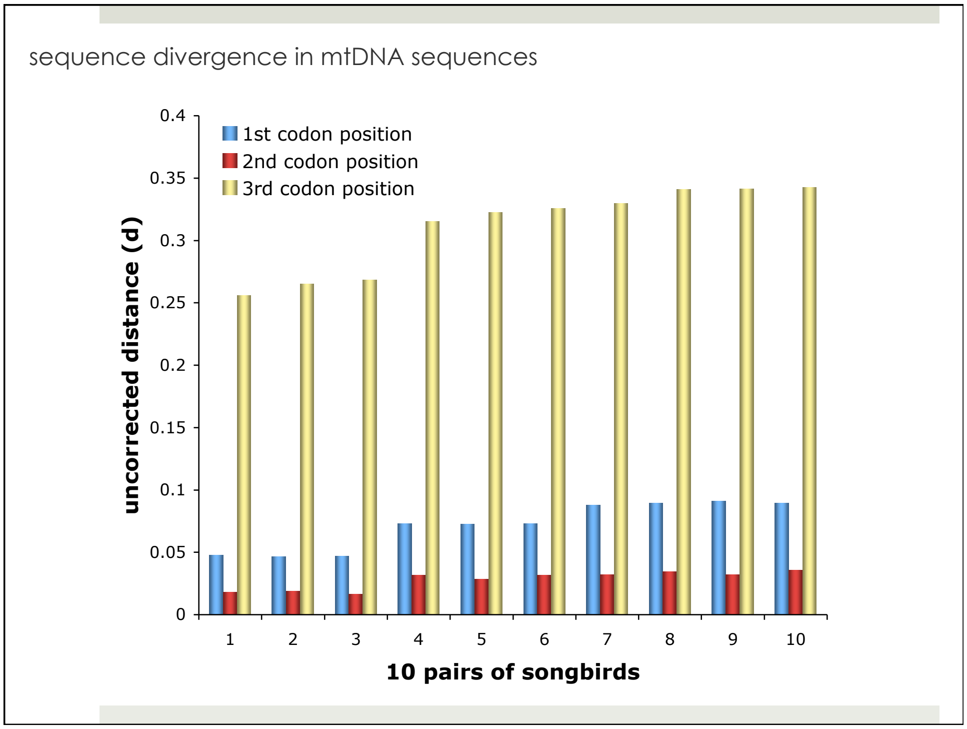
- assumes two different rates for transitions ($\alpha$) and transversions ($2\beta$)

$$d = -\frac{1}{2} \times \ln(1 - 2P - Q) - \frac{1}{4} \times \ln(1 - 2Q)$$

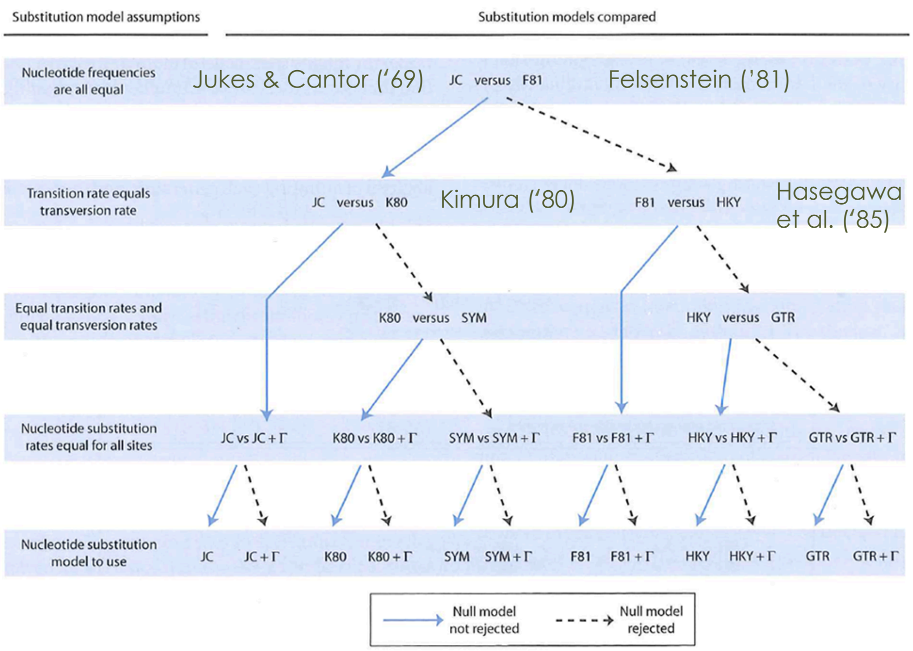
- where $P$ and $Q$ are the proportion of sites that differ by transitions and transversions, respectively
- equation is undefined for $P + Q \geq 0.75$



Base composition for 12 L-strand encoded mt protein-coding genes



sequence divergence in mtDNA sequences



Substitution model assumptions

Substitution models compared

Nucleotide frequencies are all equal

Jukes & Cantor ('69) JC versus F81 Felsenstein ('81)

Transition rate equals transversion rate

JC versus K80 Kimura ('80)

F81 versus HKY Hasegawa et al. ('85)

Equal transition rates and equal transversion rates

K80 versus SYM

HKY versus GTR

Nucleotide substitution rates equal for all sites

JC vs JC + Γ

K80 vs K80 + Γ

SYM vs SYM + Γ

F81 vs F81 + Γ

HKY vs HKY + Γ

GTR vs GTR + Γ

Nucleotide substitution model to use

JC JC + Γ K80 K80 + Γ SYM SYM + Γ F81 F81 + Γ HKY HKY + Γ GTR GTR + Γ

# Generalized Model of Nucleotide Substitution - the Q matrix

$$\mathbf{Q} = \begin{array}{c} \\ \end{array}\begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array}$$

$$\mathbf{Q} = \begin{pmatrix} -\mu(a\pi_{\mathrm{C}} + b\pi_{\mathrm{G}} + c\pi_{\mathrm{T}}) & \mu a\pi_{\mathrm{C}} & \mu b\pi_{\mathrm{G}} & \mu c\pi_{\mathrm{T}} \\ \mu g\pi_{\mathrm{A}} & -\mu(g\pi_{\mathrm{A}} + d\pi_{\mathrm{G}} + e\pi_{\mathrm{T}}) & \mu d\pi_{\mathrm{G}} & \mu e\pi_{\mathrm{T}} \\ \mu h\pi_{\mathrm{A}} & \mu j\pi_{\mathrm{C}} & -\mu(h\pi_{\mathrm{A}} + j\pi_{\mathrm{C}} + f\pi_{\mathrm{T}}) & \mu f\pi_{\mathrm{T}} \\ \mu i\pi_{\mathrm{A}} & \mu k\pi_{\mathrm{C}} & \mu l\pi_{\mathrm{G}} & -\mu(i\pi_{\mathrm{A}} + k\pi_{\mathrm{C}} + l\pi_{\mathrm{G}}) \end{pmatrix}$$

Where:
$\mu$ = mean instantaneous substitution rate
$\pi$ = frequency of nucleotide denoted by the subscript
$a,b,c,\dots l$ = relative rates of transformation from one base to another

Nearly all models of DNA substitution are special cases of this matrix (derived by implementing various simplifying assumptions).

## Jukes-Cantor Model

- assumes that all substitutions occur at the same rate $(a = b = c = d = e = f)$ and that the bases are in equal frequency ($\pi_A = \pi_C = \pi_G = \pi_T$)

$$\mathbf{Q} = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

## Which substitution model to use?

- p-distance (observed proportion of sites that change)
- Jukes-Cantor
- Felsenstein 1981
- Tajima-Nei
- Kimura 2-parameter
- Felsenstein 1984
- Hasegawa, Kishino, Yano (HKY) 1985
- Kimura 3-parameter
- Tamura-Nei
- General Time Reversible
- etc…

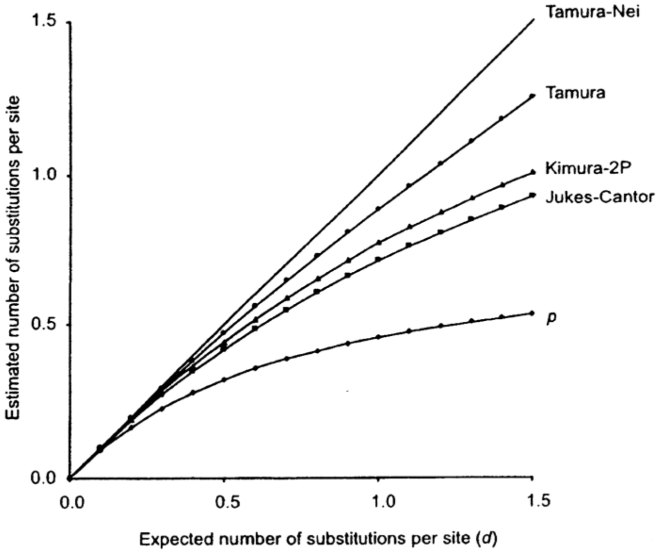
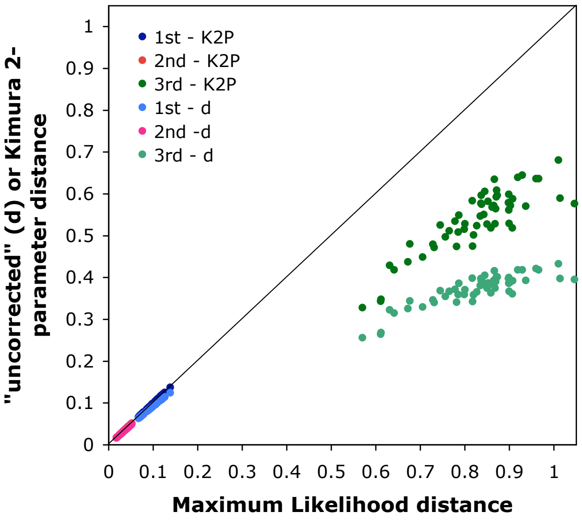


FIGURE 3.1. Estimates of the number of nucleotide substitutions obtained by different distance measures when actual nucleotide substitution follows the Tamura-Nei model. The nucleotide frequencies assumed are $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$; and the two transition/transversion rate ratios assumed are $\alpha_1/\beta = 4$ and $\alpha_2/\beta = 8$.

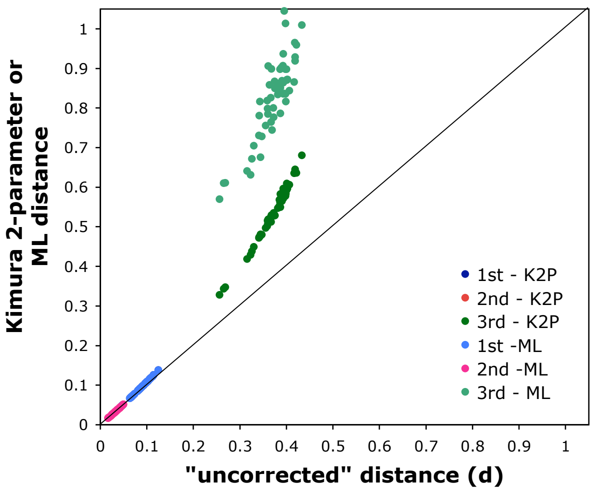
increasing complexity of model

genetic divergence in mtDNA protein-coding genes among 11 songbirds (Passeriformes)



*K2P*

*p (=d)*



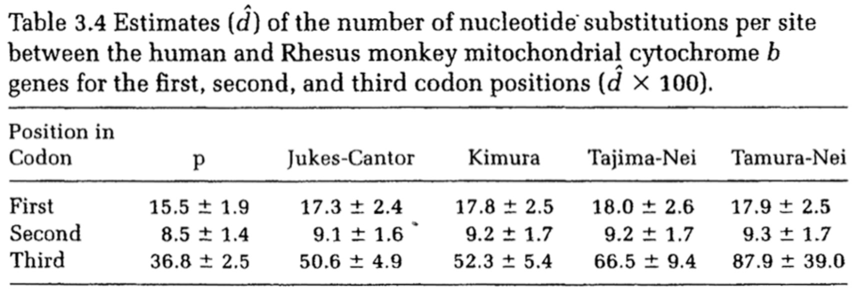


Table 3.4 Estimates ($\hat{d}$) of the number of nucleotide substitutions per site between the human and Rhesus monkey mitochondrial cytochrome *b* genes for the first, second, and third codon positions ($\hat{d} \times 100$).

| Position in Codon | p | Jukes-Cantor | Kimura | Tajima-Nei | Tamura-Nei |
|---|---|---|---|---|---|
| First | 15.5 ± 1.9 | 17.3 ± 2.4 | 17.8 ± 2.5 | 18.0 ± 2.6 | 17.9 ± 2.5 |
| Second | 8.5 ± 1.4 | 9.1 ± 1.6 | 9.2 ± 1.7 | 9.2 ± 1.7 | 9.3 ± 1.7 |
| Third | 36.8 ± 2.5 | 50.6 ± 4.9 | 52.3 ± 5.4 | 66.5 ± 9.4 | 87.9 ± 39.0 |

## Generalizations…

- estimated genetic distance generally increases with the complexity of the underlying substitution model
  - but even the most "aggressively corrected" distances may still be underestimates
- increased model complexity generally leads to greater accuracy but also reduced precision (i.e., greater variance)
- when dates of evolutionary events are based on estimates of sequence divergence, interpret cautiously!

## Neutral Expectations…

- ...with constant population size and mutation

**nucleotide diversity**

$$E(\Pi)=\theta=4N\mu$$

**# segregating sites**

$$E(S)=\theta\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots+\frac{1}{k-1}\right)$$

**homozygosity**

$$\hat{F}=\frac{1}{1+4N\mu}$$

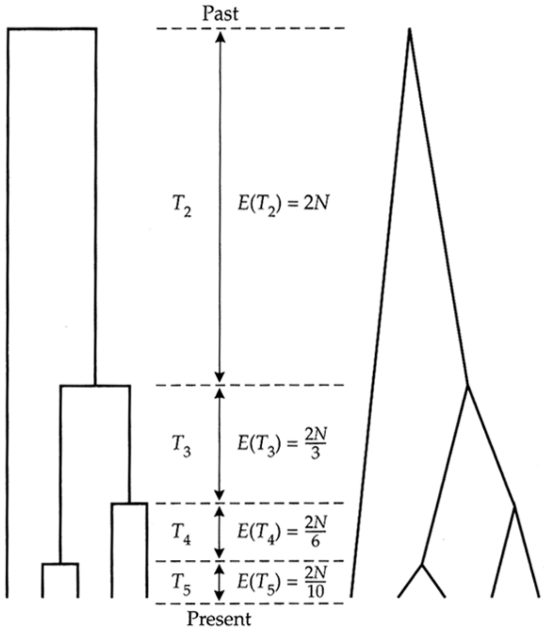
**# unique alleles**

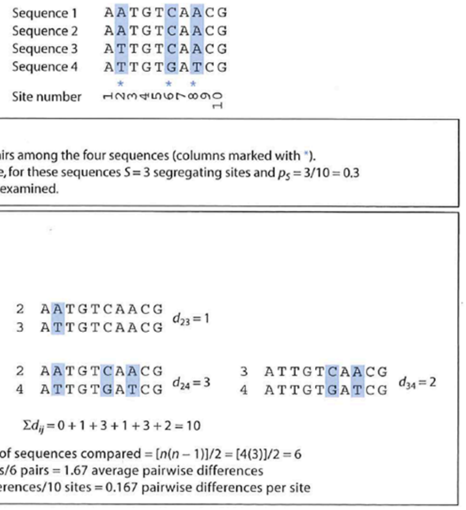
$$E(k)=1+\frac{\theta}{\theta+1}+\frac{\theta}{\theta+2}+\ldots+\frac{\theta}{\theta+n-1}$$

**allele frequency distribution**

$$\Pr\{a_1,\ldots,a_n\}=\frac{n!}{\theta(\theta+1)\ \ldots\ (\theta+n-1)}\prod_{j=1}^{n}\frac{\theta^{a_j}}{j^{a_j}a_j!}$$

**FIGURE 3.15** Two completely equivalent ways of illustrating the coalescences in a gene tree. On the left, the coalescent events are represented as horizontal lines, on the left they are represented as nodes. In any each generation, if there are $k$ alleles present, the expected time back to the next coalescence is given by $4N/[k(1-k)]$. For example, starting with five alleles, the expected time back to the first coalescence is $4N/[(5)(4)] = 2N/10$. Note that the successive times get longer. When there are only two alleles, the time back to the final coalescence is $2N$ generations.





```
Sequence 1   AATGTCAACG
Sequence 2   AATGTCAACG
Sequence 3   ATTGTCAACG
Sequence 4   ATTGTGATCG
              *   * *
Site number  12345678910
```

Segregating sites ($S$ and $p_S$):

Sites 2, 6, and 8 have variable base pairs among the four sequences (columns marked with *). These are segretating sites. Therefore, for these sequences $S = 3$ segregating sites and $p_S = 3/10 = 0.3$ segregating sites per nucleotide site examined.

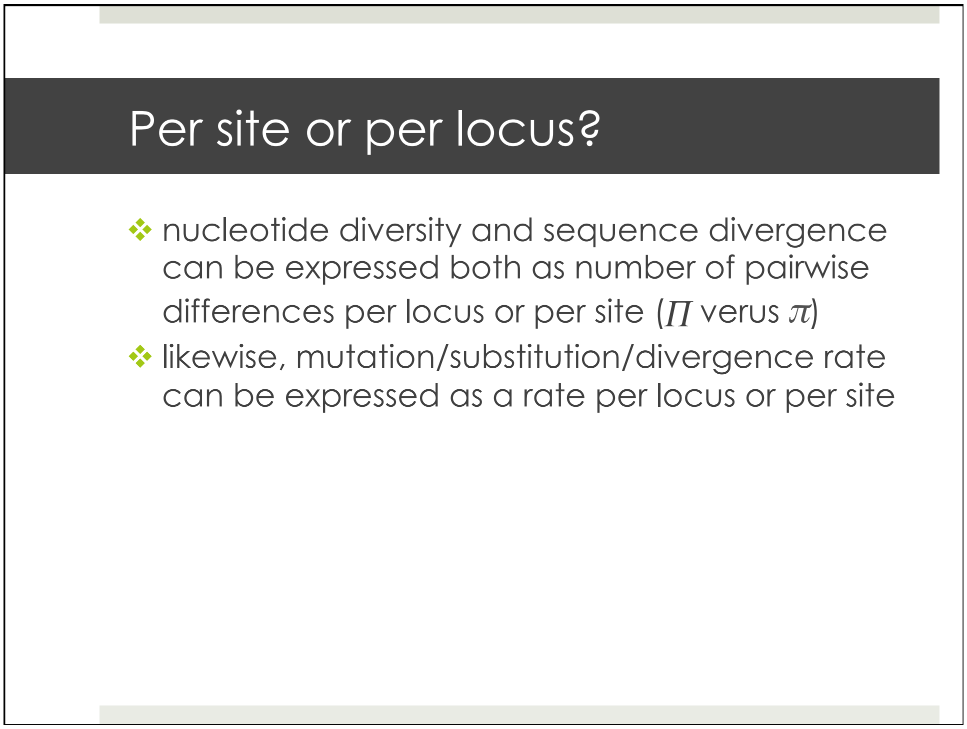
Nucleotide diversity ($\pi$):

```
1  AATGTCAACG  d12 = 0
2  AATGTCAACG

1  AATGTCAACG  d13 = 1    2  AATGTCAACG  d23 = 1
3  ATTGTCAACG             3  ATTGTCAACG

1  AATGTCAACG  d14 = 3    2  AATGTCAACG  d24 = 3    3  ATTGTCAACG  d34 = 2
4  ATTGTGATCG             4  ATTGTGATCG             4  ATTGTGATCG
```

$\Sigma d_{ij} = 0 + 1 + 3 + 1 + 3 + 2 = 10$

Number of pairs of sequences compared $= [n(n-1)]/2 = [4(3)]/2 = 6$
$\hat{\pi} = 10$ differences/6 pairs $= 1.67$ average pairwise differences
$\hat{\pi} = 1.67$ avg. differences/10 sites $= 0.167$ pairwise differences per site

**Figure 8.11** A hypothetical sample of four DNA sequences that are each 10 nucleotides long. There a total of three segregating sites ($S = 3$) or three-tenths of a segregating site per nucleotide ($p_S = 0.3$). The nucleotide diversity is calculated by summing the nucleotide sites that differ between each unique pair of DNA sequences. In this example there are 1.67 average pairwise nucleotide differences or 0.167 average pairwise nucleotide differences per nucleotide site.



# Per site or per locus?

- nucleotide diversity and sequence divergence can be expressed both as number of pairwise differences per locus or per site ($\Pi$ verus $\pi$)
- likewise, mutation/substitution/divergence rate can be expressed as a rate per locus or per site