Lecture 17

- Detecting sequence conservation with PhyloHMMs
- PhastCons

PhyloHMMs

- Yang 1995; Felsenstein & Churchill 1996

- Siepel A. *et al.* (2005): Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50
  - basis of PhastCons conservation scores (UCSC genome browser)

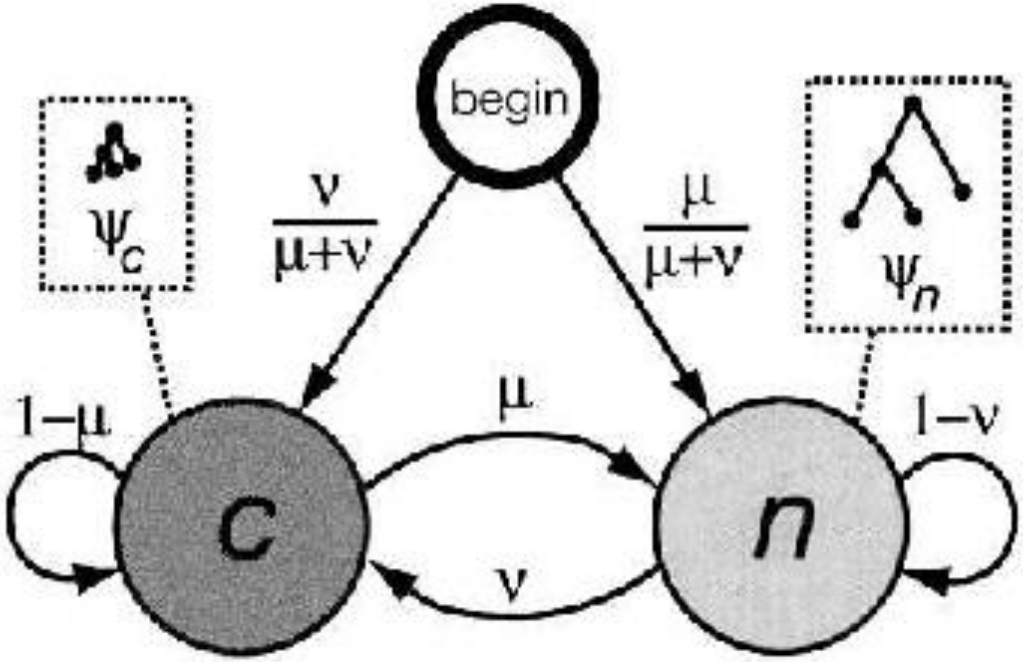
- Goal: starting from multiple genome sequence alignment, identify
  - conserved regions (regions under purifying selection),

  against background of
  - neutrally evolving regions

PhastCons PhyloHMM

- model:
  - 2-state HMM
    - c: conserved state
    - n: neutral (or nonconserved) state
  - emitted symbols are *alignment columns*
  - emission probabilities based on *phylogenetic tree* relating sequences
  - gaps in alignment treated as *missing data*

PhastCons PhyloHMM

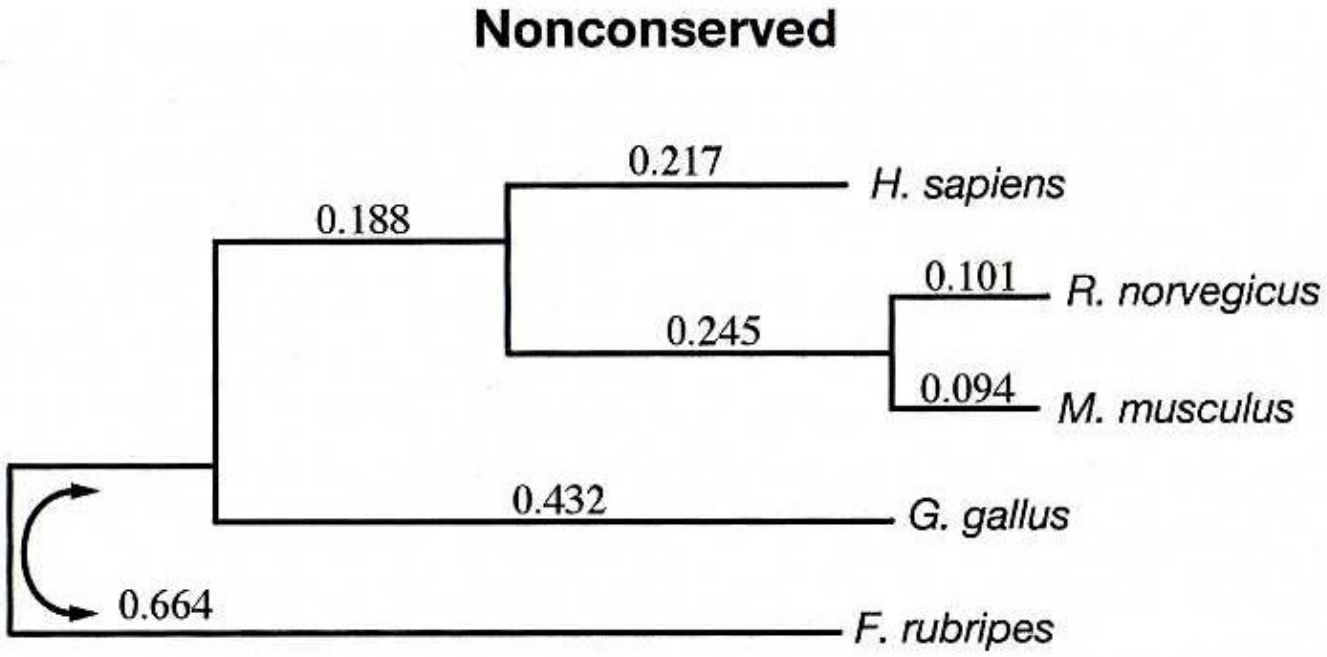


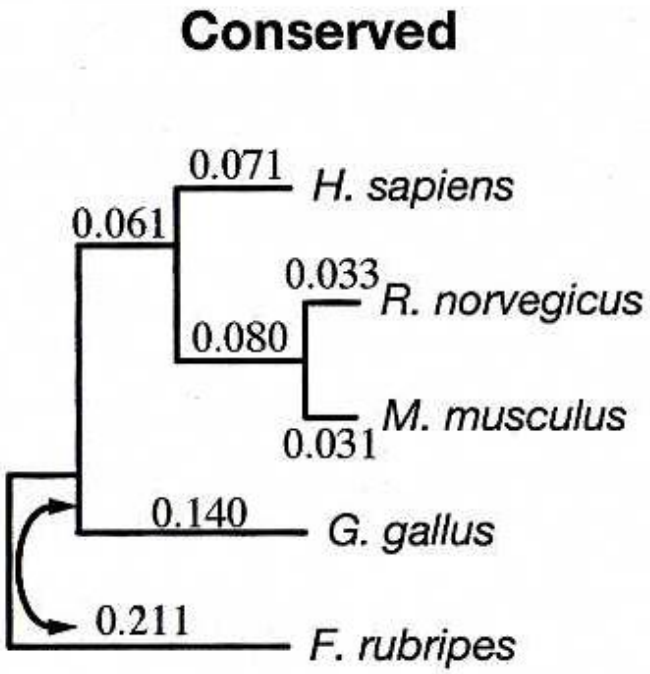
$\mu = a_{cn}$

$\nu = a_{nc}$

$$\mathbf{x} = \begin{matrix} \text{TCGCGACATATACGA}\ldots \\ \text{TTGGGGCATGTGGGT}\ldots \\ \text{AGCAGACGTCCGCAA}\ldots \end{matrix}$$

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

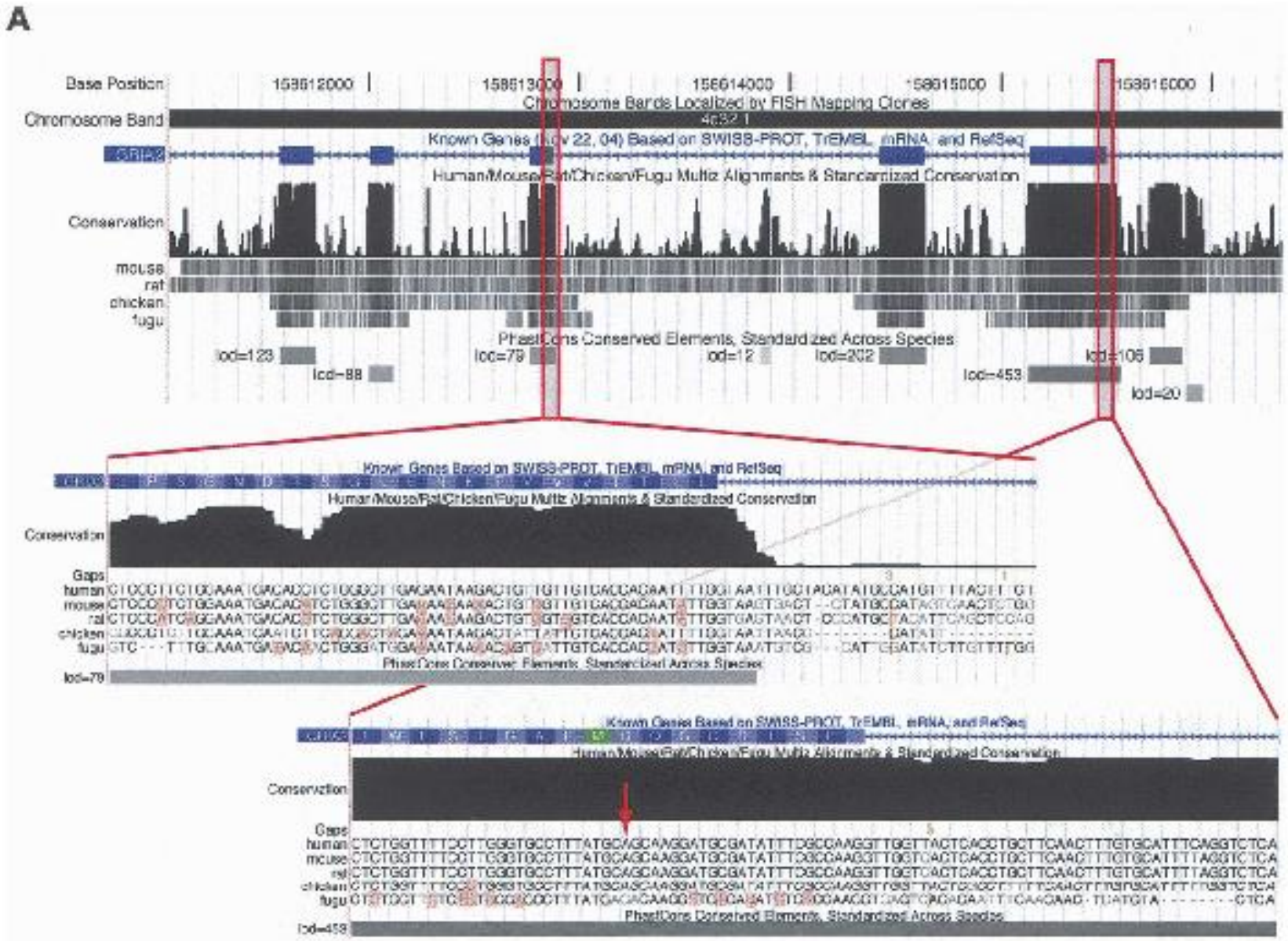

Nonconserved
0.217
H. sapiens
0.188
0.101
R. norvegicus
0.245
0.094
M. musculus
0.432
G. gallus
0.664
F. rubripes


Conserved
0.071
H. sapiens
0.061
0.033
R. norvegicus
0.080
M. musculus
0.031
0.140
G. gallus
0.211
F. rubripes

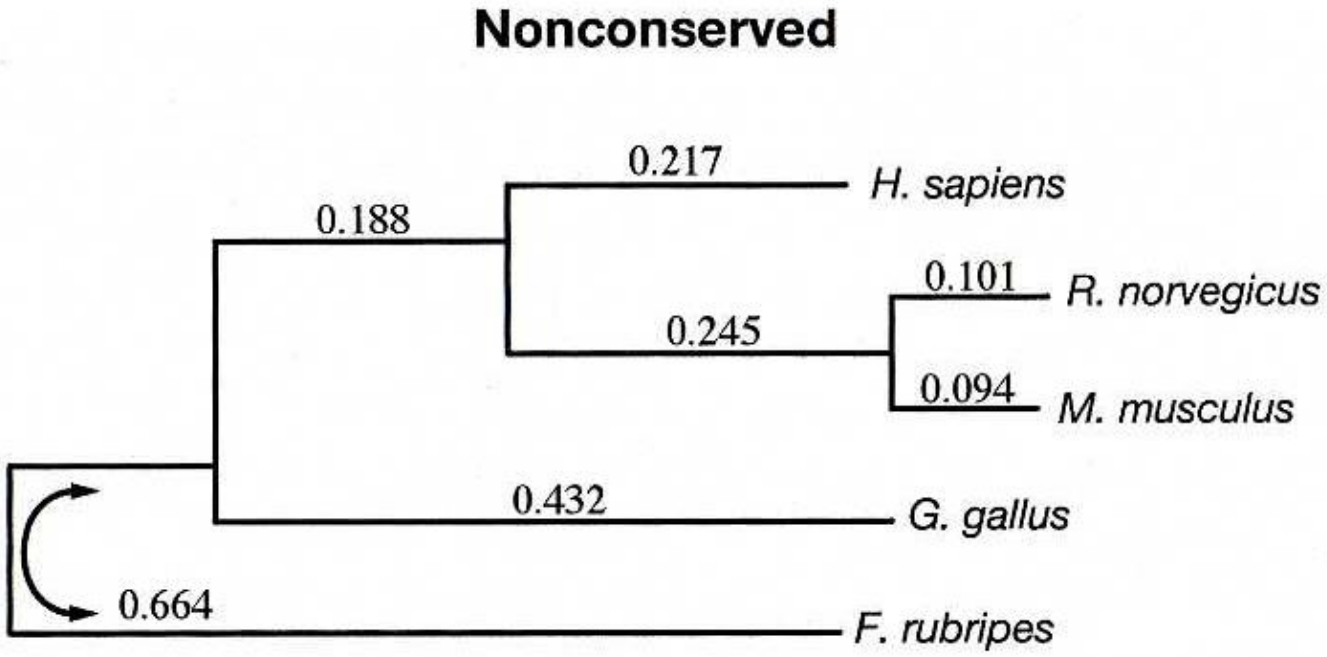
Siepel *et al* evolutionary model

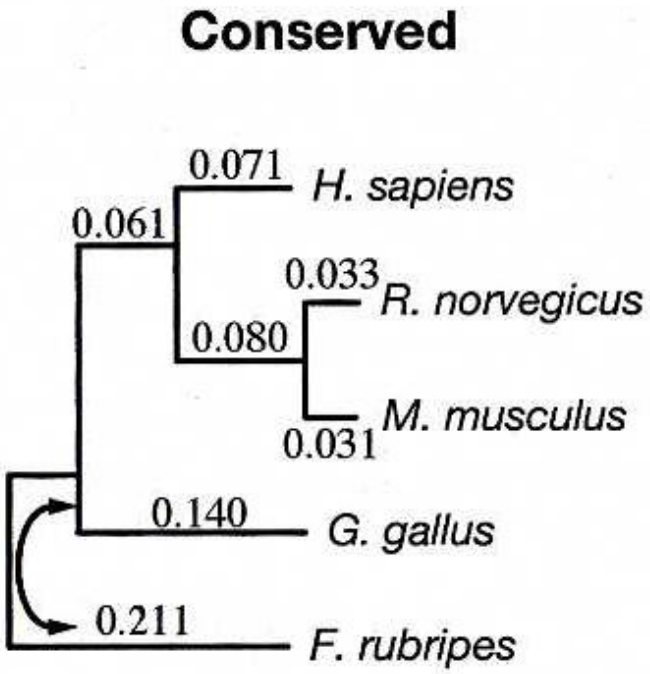
- single, reversible, infinitesimal mutation process across tree
- branches differ only in their lengths
- selection strength same across tree and sites

- branch lengths:
  - Expected # substitutions/site over corresponding evolutionary time period
  - for neutral state, should reflect underlying mutation rate
  - for conserved state: mutation rate $\times$ scaling factor ρ
    - ρ = frac of mutations that escape purifying selection
    - $\rho \approx .33$ (for vertebrates)

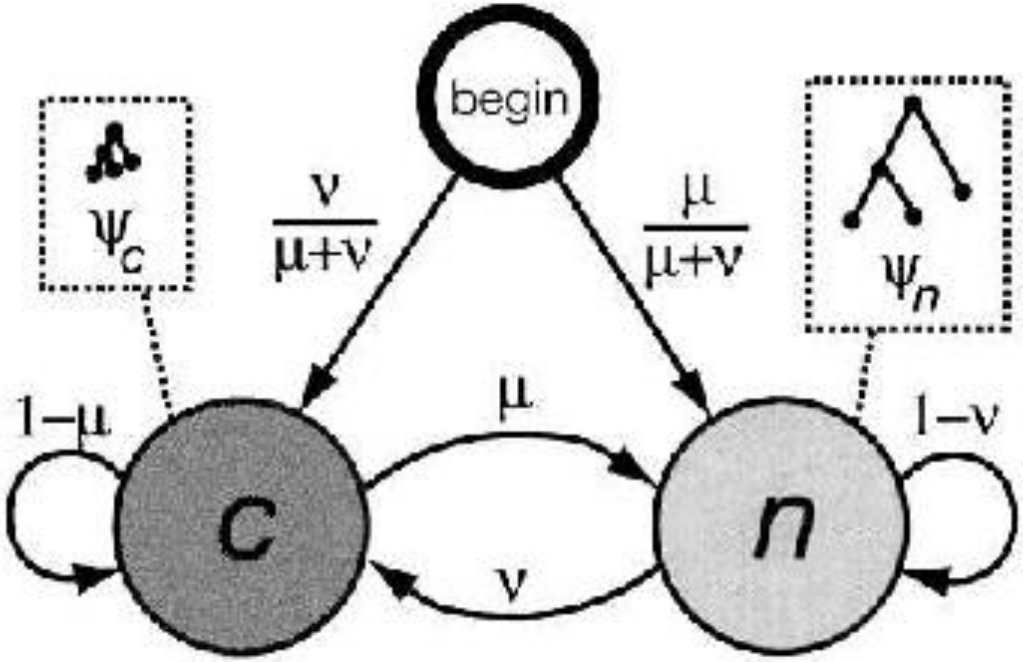


from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.


Nonconserved
0.217
H. sapiens
0.188
0.101
R. norvegicus
0.245
0.094
M. musculus
0.432
G. gallus
0.664
F. rubripes


Conserved
0.071
H. sapiens
0.061
0.033
R. norvegicus
0.080
M. musculus
0.031
0.140
G. gallus
0.211
F. rubripes

PhastCons PhyloHMM



$\mu = a_{cn}$

$\nu = a_{nc}$

$$\mathbf{x} = \begin{matrix} \text{TCGCGACATATACGA}\ldots \\ \text{TTGGGGCATGTGGGT}\ldots \\ \text{AGCAGACGTCCGCAA}\ldots \end{matrix}$$

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Some general issues in applying probability models, in the PhyloHMM context

- Is the model computable?
- Is the model ‘reasonable’?
  - 2 states enough?
    - variability of mutation, selection within genome
    - changes in selected sites over time
    - but simplicity has its advantages!
      - interpretability
      - overfitting & parameter estimation less problematic
  - Markov condition on transition probabilities
  - treatment of gaps

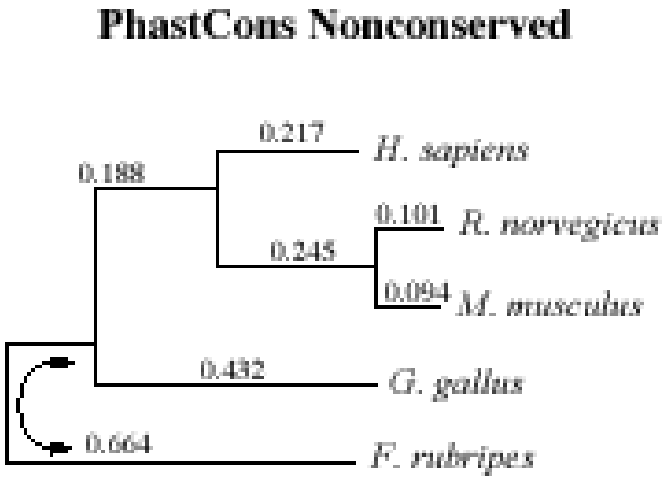
- How good is the input data?
  - alignability of neutral sequence
  - accuracy of genome sequence alignments

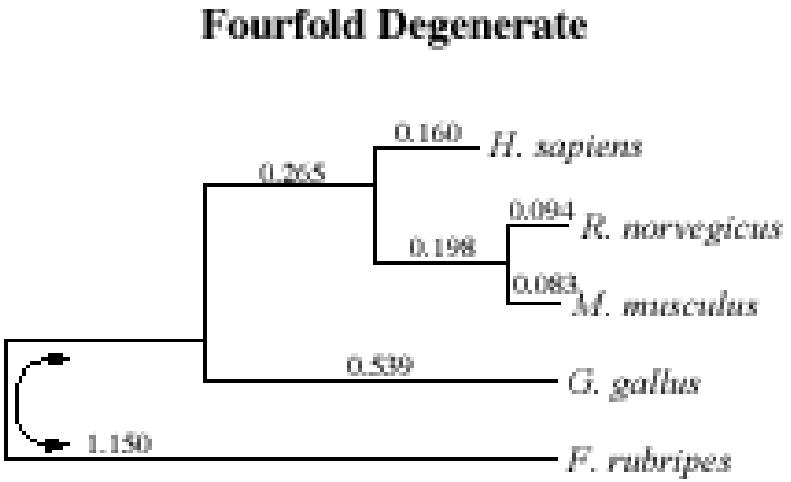
- Are results reliable?
  - no true ‘test set’ – instead, putative false positive rate, and ‘biological plausibility’ of findings

Alignment issues

- Multiz: progressive pairwise alignments
- accurate multiple genome alignment *not* a solved problem!
  - statistical assessment: Prakash & Tompa (2005, 2007, 2009)
  - ENCODE region alignment analyses: Margulies EH *et al.* 2007
  - major issues:
    - accurate gap placement (even for close species!!)
    - discrimination among paralogous sequences (e.g. repeats, duplications)
    - short 'junk' alignment segments
  - *in principle*, more sequences should give more accurate alignments
- inaccurate alignments can cause
  - neutral rate to be *overestimated*
  - conserved segments to be *overidentified*
    - because more slowly mutating (or better aligned) neutral segments may be called conserved

- for distantly related species, neutrally evolving regions no longer alignable
  - analyze 4D sites in coding sequences to estimate neutral rates
    - CDS alignments much more reliable, but
    - synonymous sites somewhat atypical (some selection; composition & mutation patterns)





The Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G