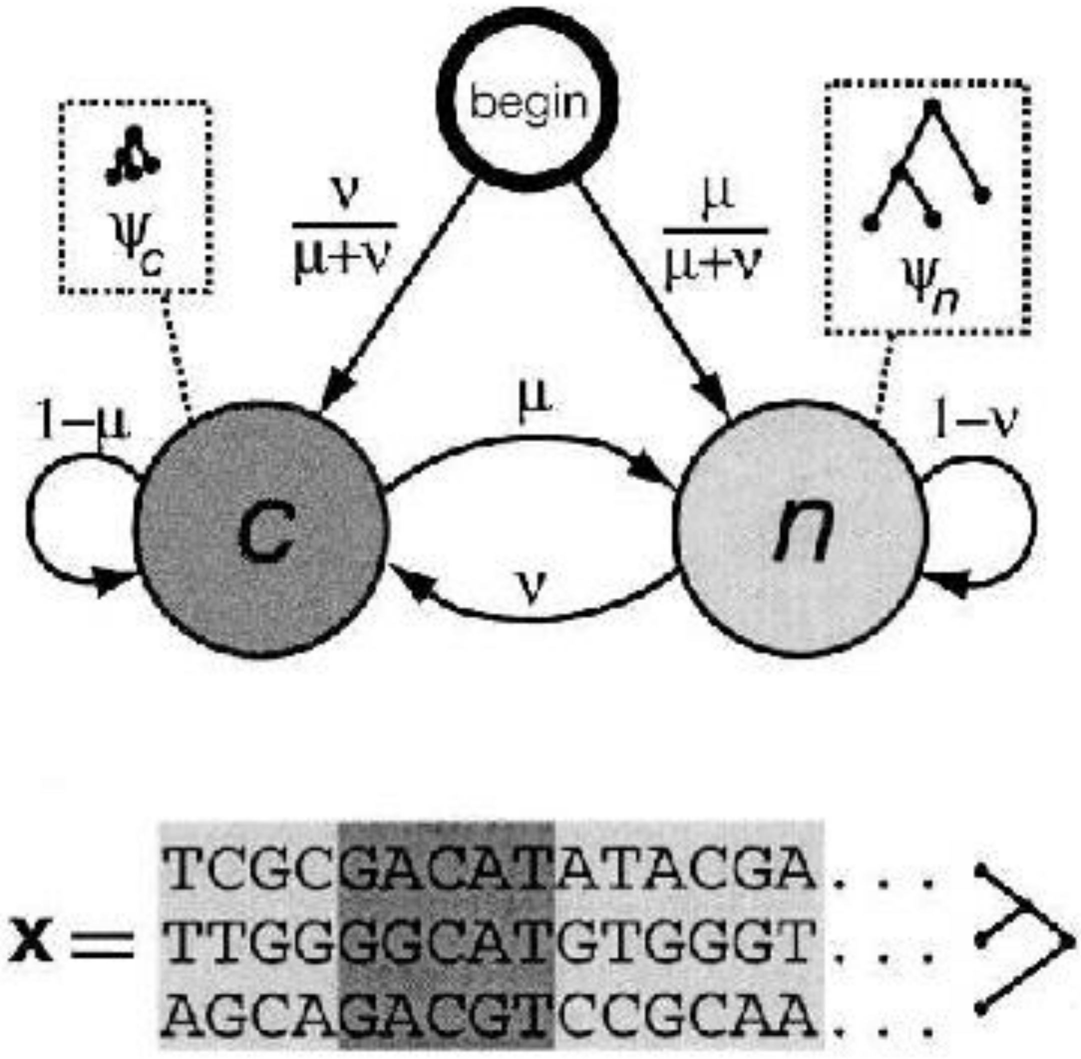
Today’s Lecture

- PhastCons

PhastCons PhyloHMM

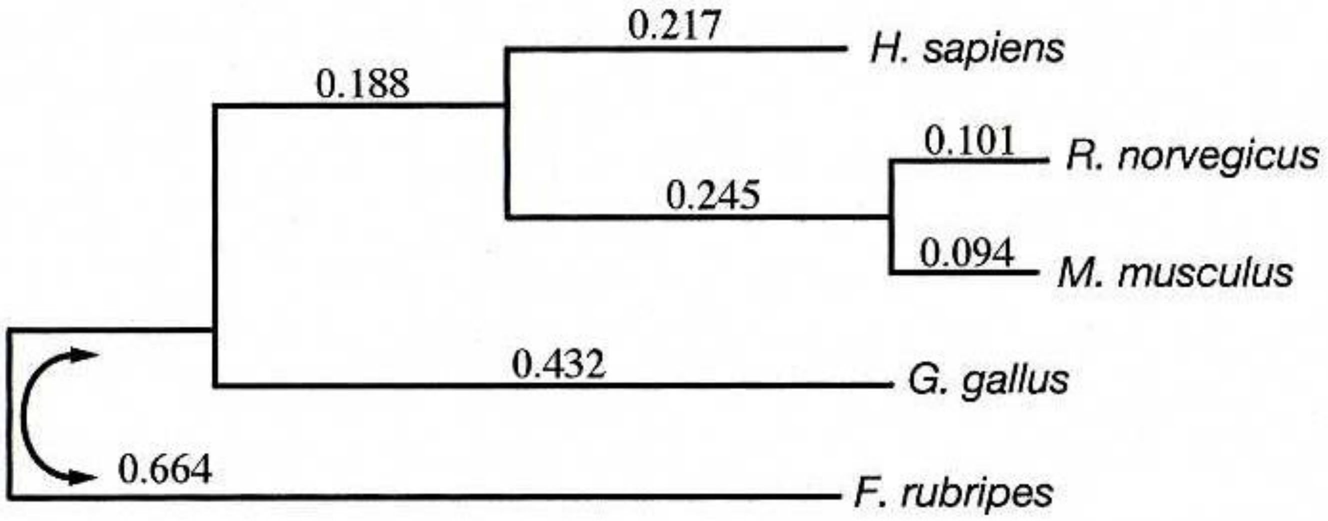


$\mu = a_{cn}$

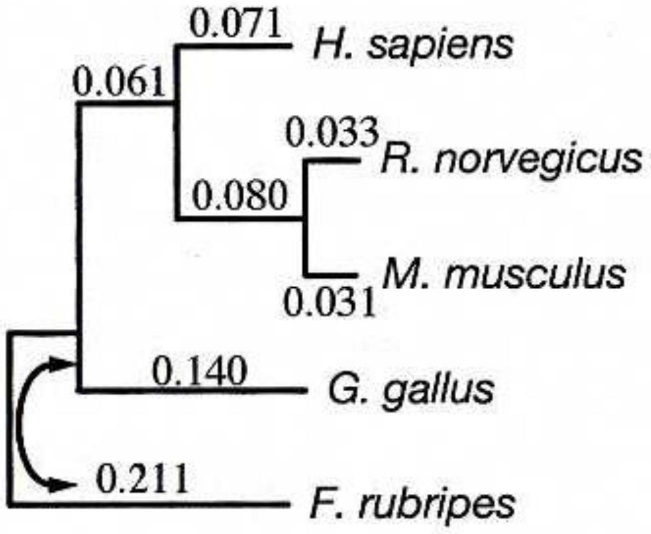
$\nu = a_{nc}$

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

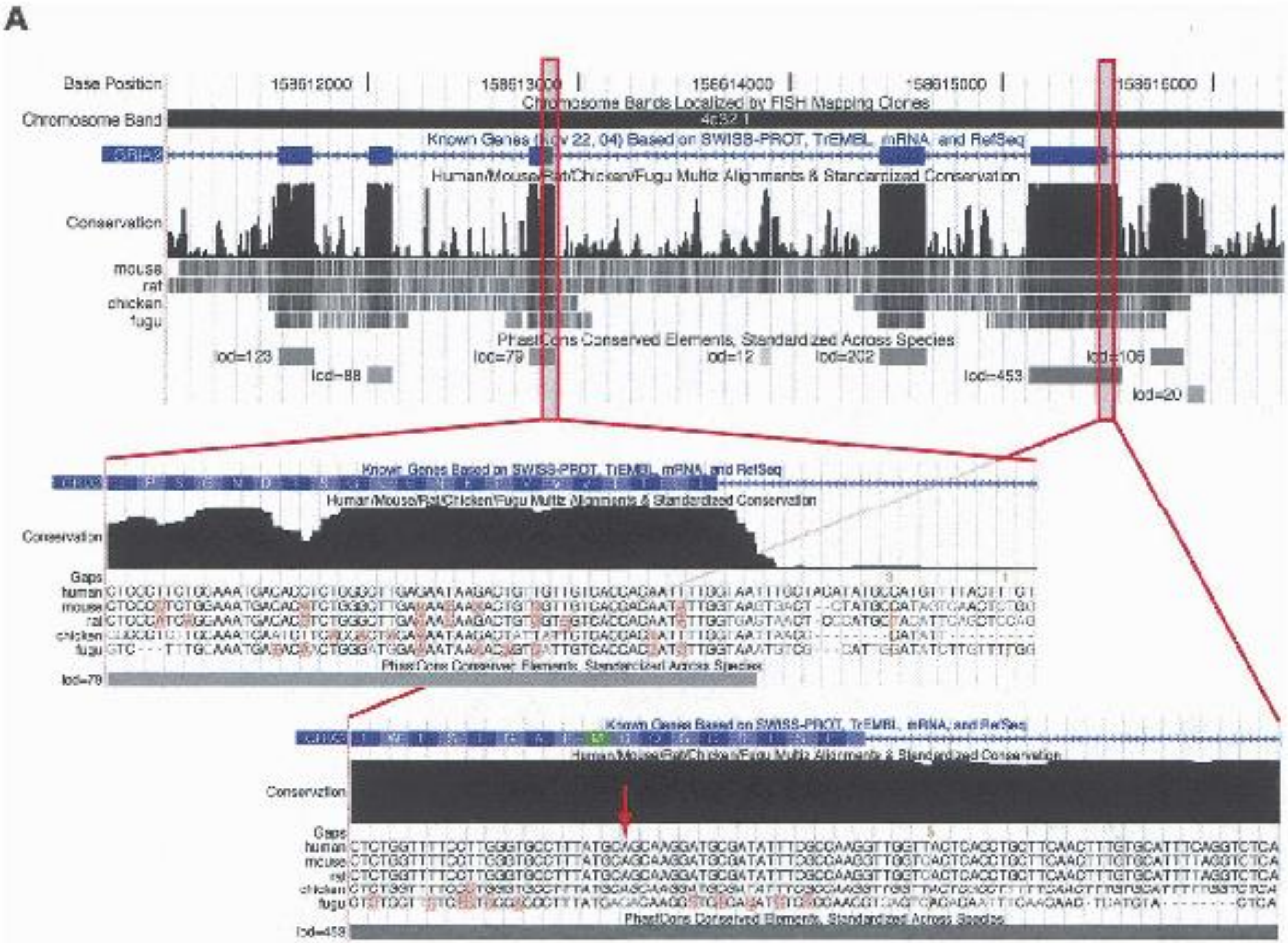
Nonconserved



Conserved



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


A
Base Position
158612000
158613000
158614000
158615000
158616000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
4q32.1
GRIA2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=123
lod=88
lod=79
lod=12
lod=202
lod=453
lod=106
lod=20
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=79
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=453

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?
  - Markov condition on transition probabilities
- 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)
- inaccurate alignments cause
  - neutral rate to be *overestimated*
  - conserved segments to be *overidentified*
    - because more slowly mutating (or better aligned) neutral segments may be called conserved