

---

# 3000788 Intro to Comp Molec Biol

## Lecture 6: Phylogenetics and molecular evolution

September 1, 2022



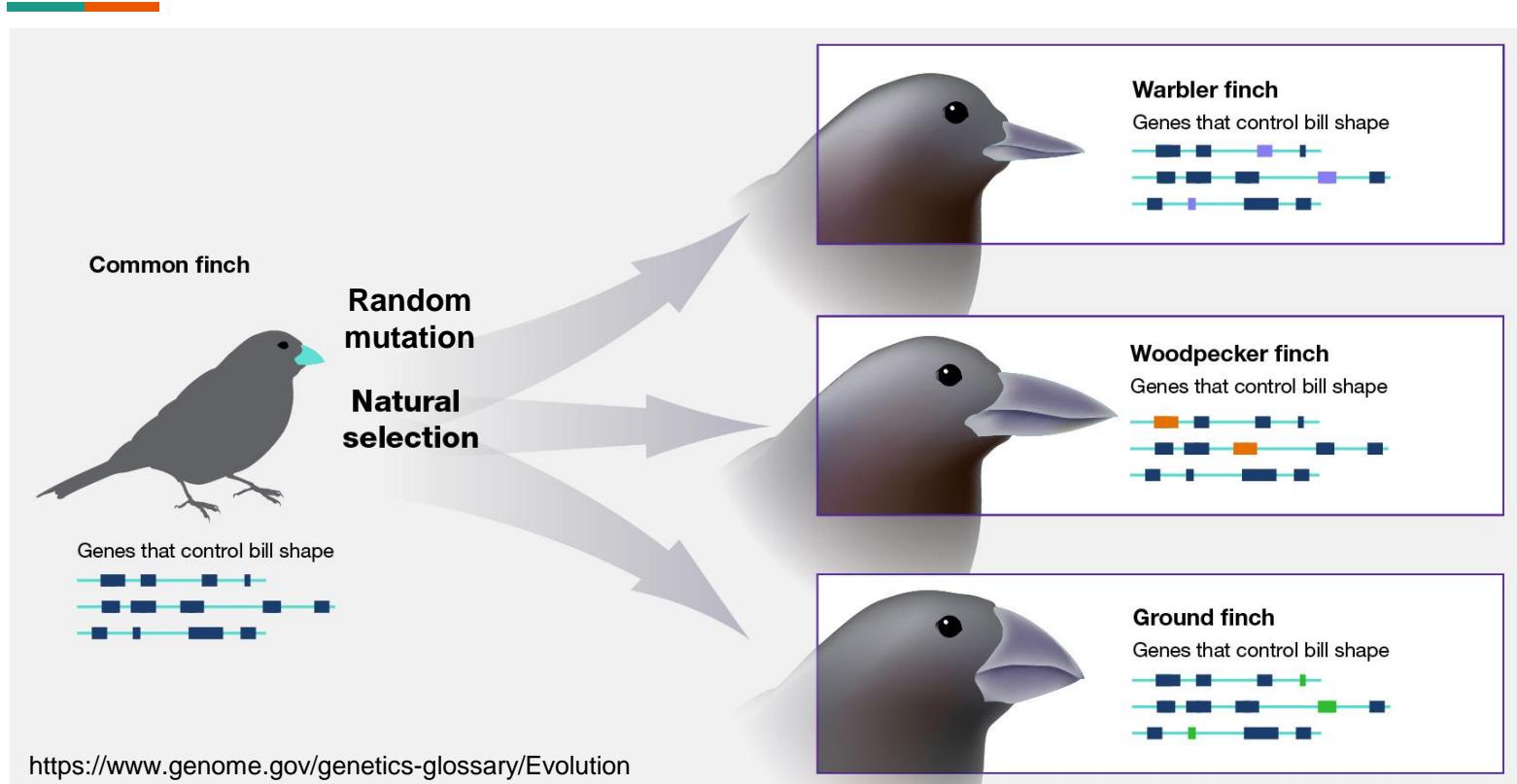
**Sira Sriswasdi, PhD**

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

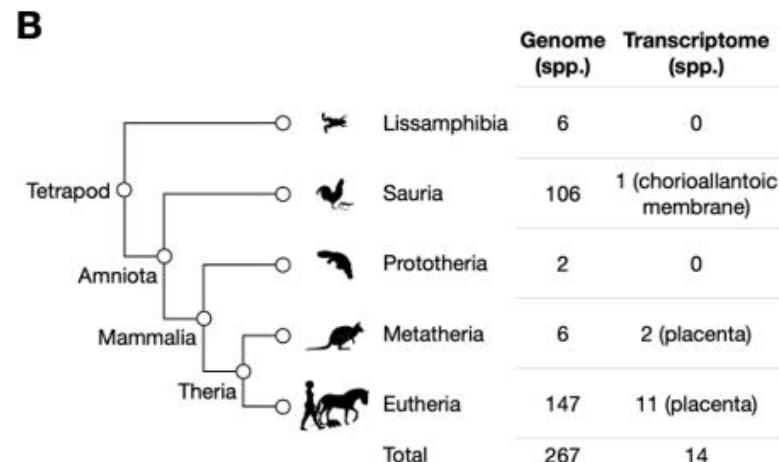
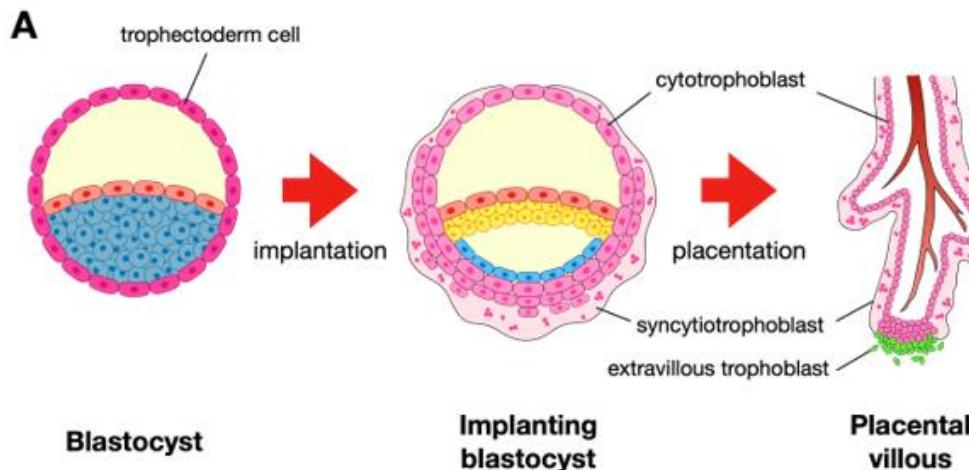


# Evolution

# Random mutation with natural selection



# Emergence of form and function

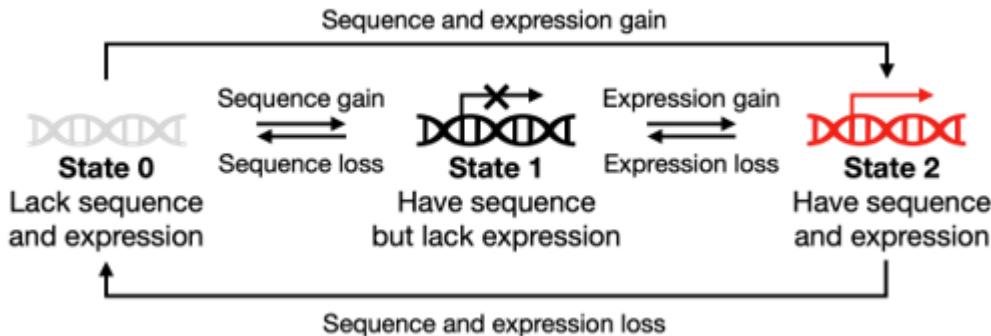


Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)

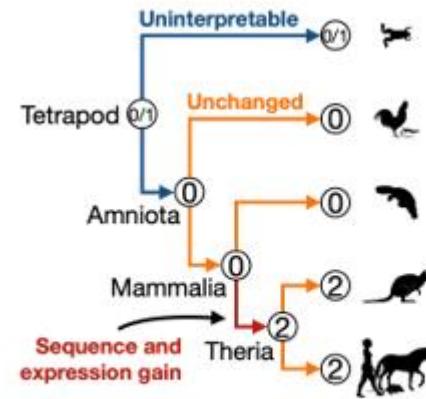
- Placenta emerged as a new organ in mammalian ancestor

# The search for genes that give rise to placenta

C



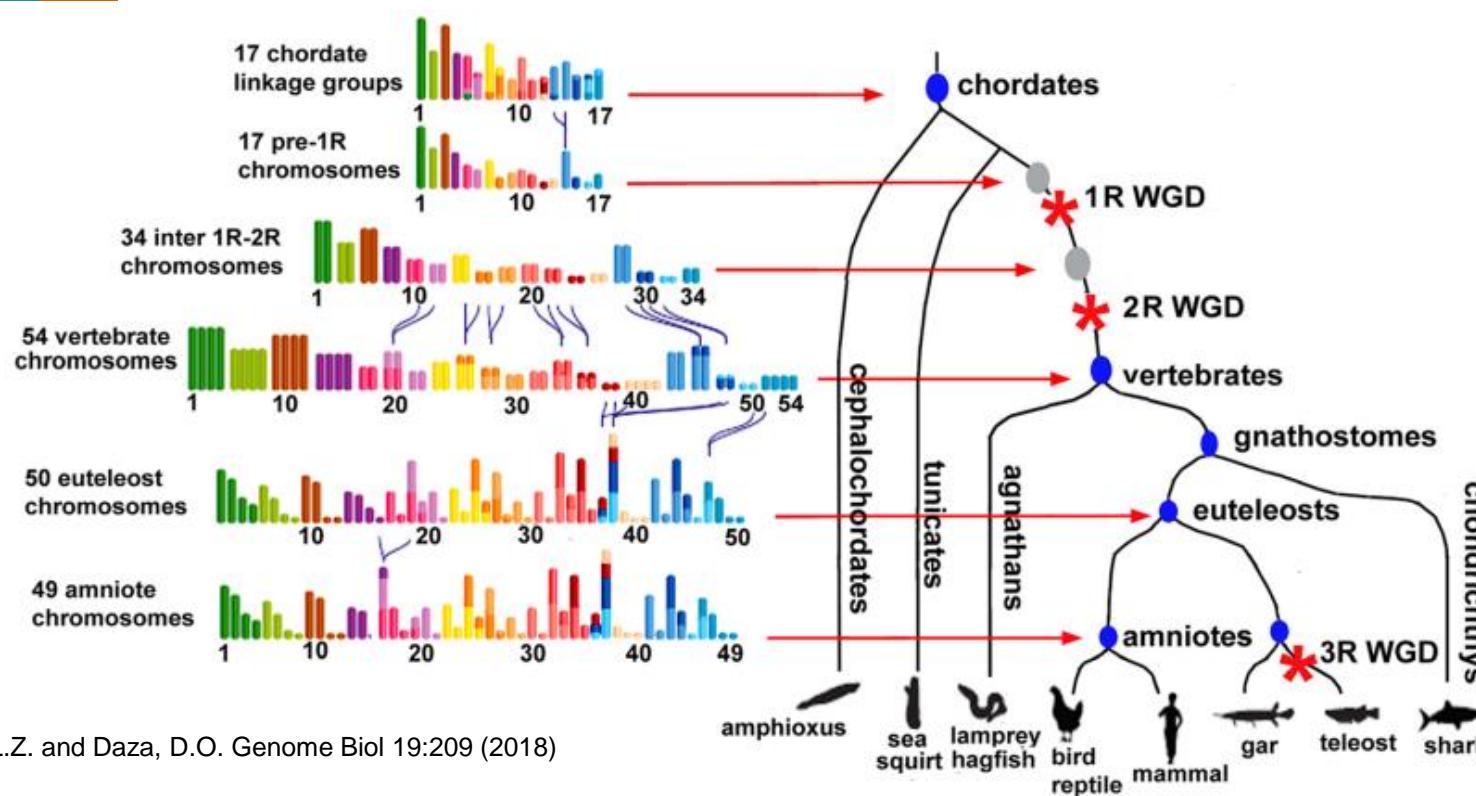
Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)



Interpreting character state changes from ancestors to descendants to transition events

- Identify gene and expression gain events during mammalian evolution

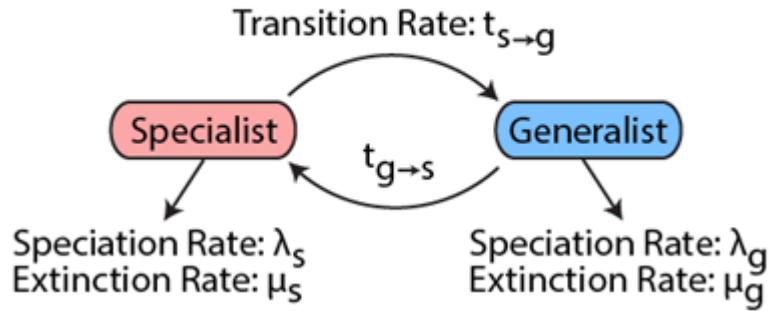
# History of whole-genome duplications in animal



# Life style evolution in bacteria

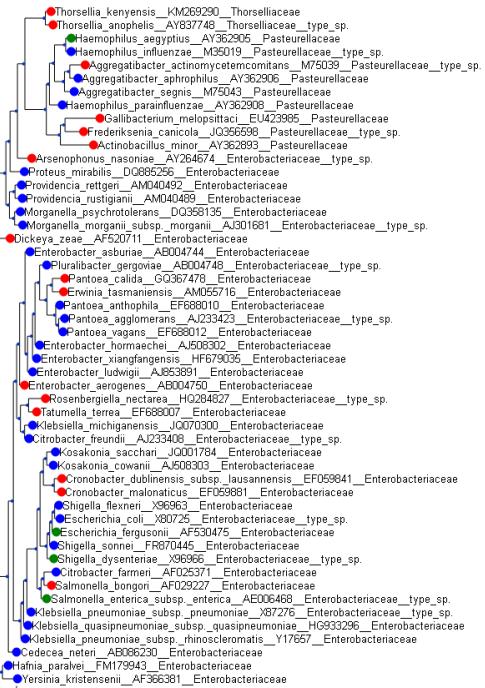


## Binary-State Model



Maddison et al., Syst Biol 2007 and FitzJohn et al. Syst Biol 2009

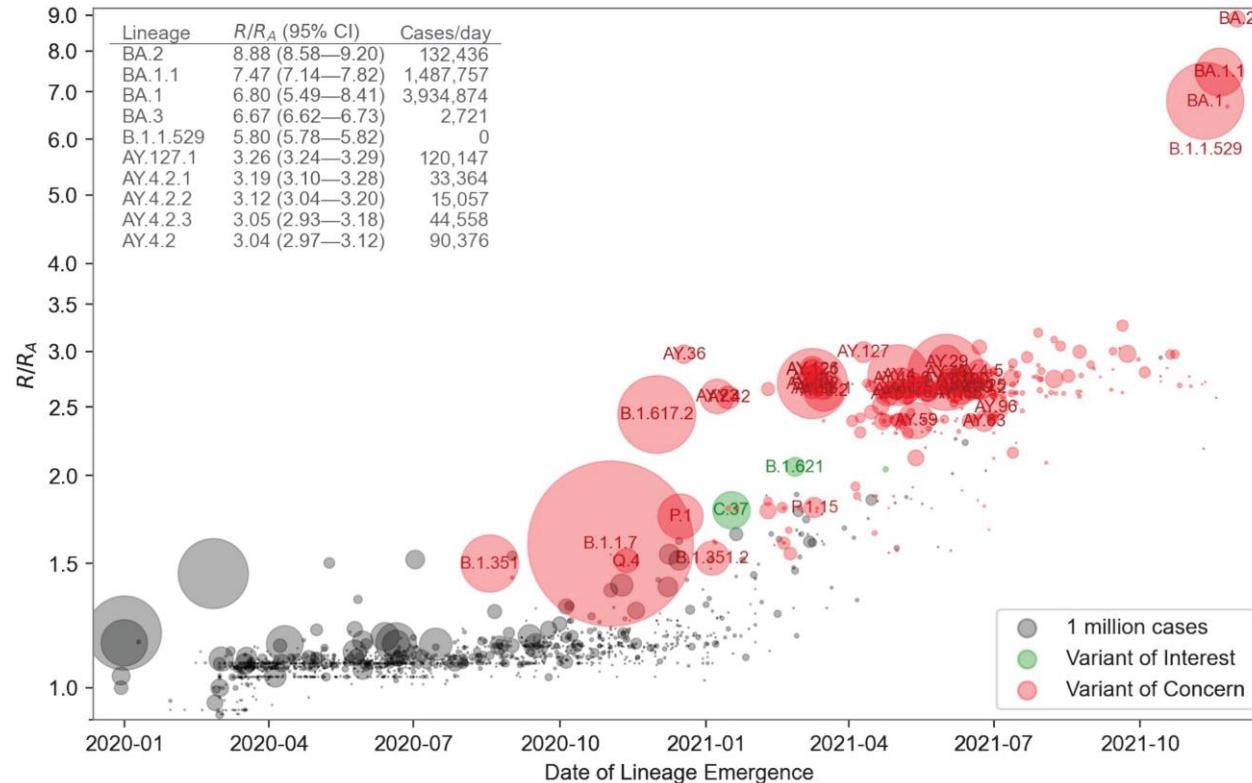
- Generalist has higher survivability
- Nature drives generalist to specialist



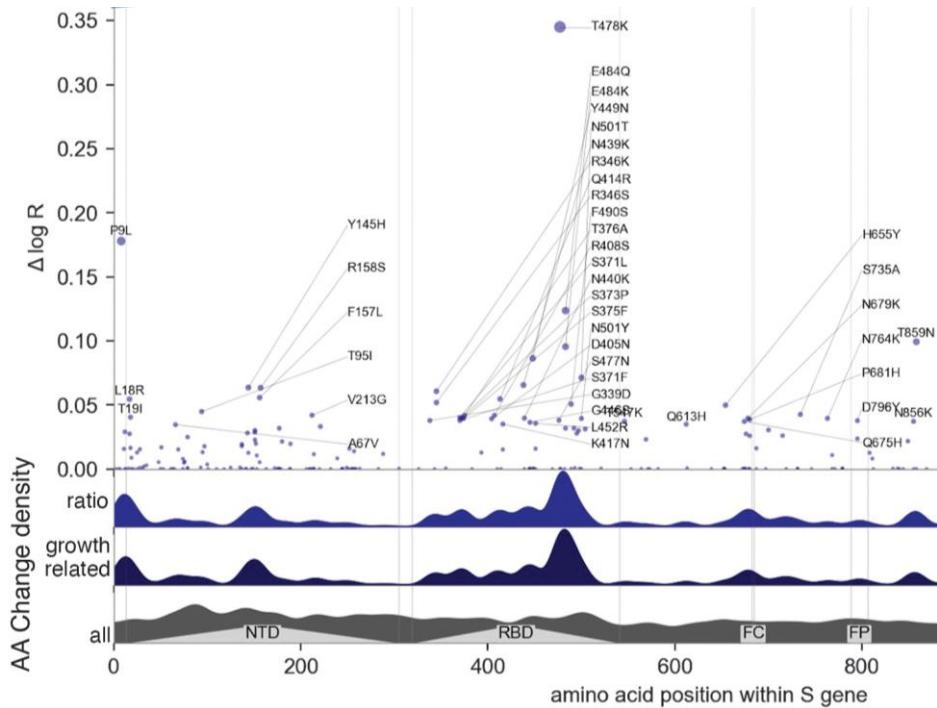
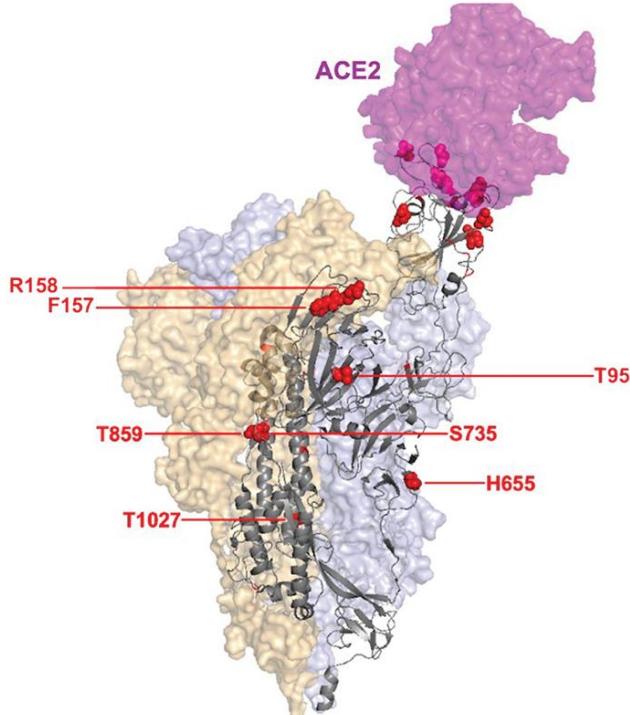
# COVID-19 evolution

Obermeyer, F. et al. Science 376:1327-1332 (2022)

Fitness increase over the original Wuhan strain



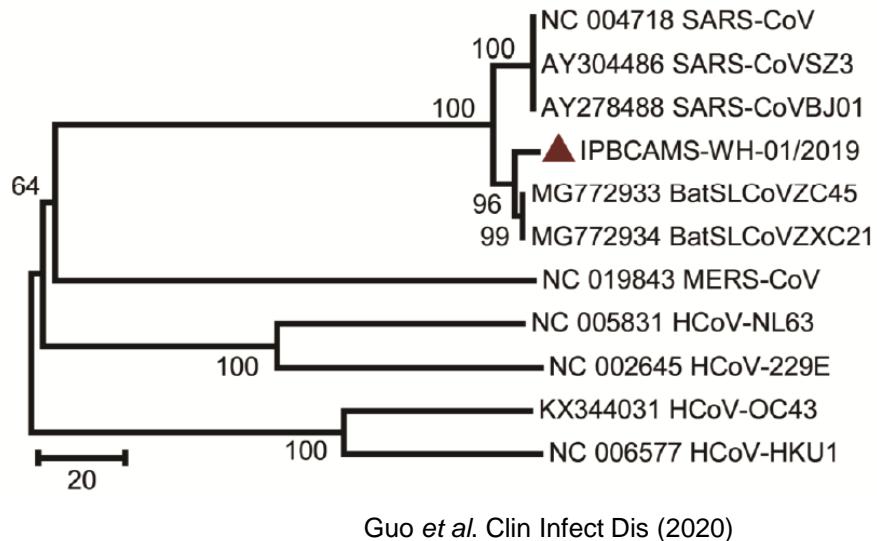
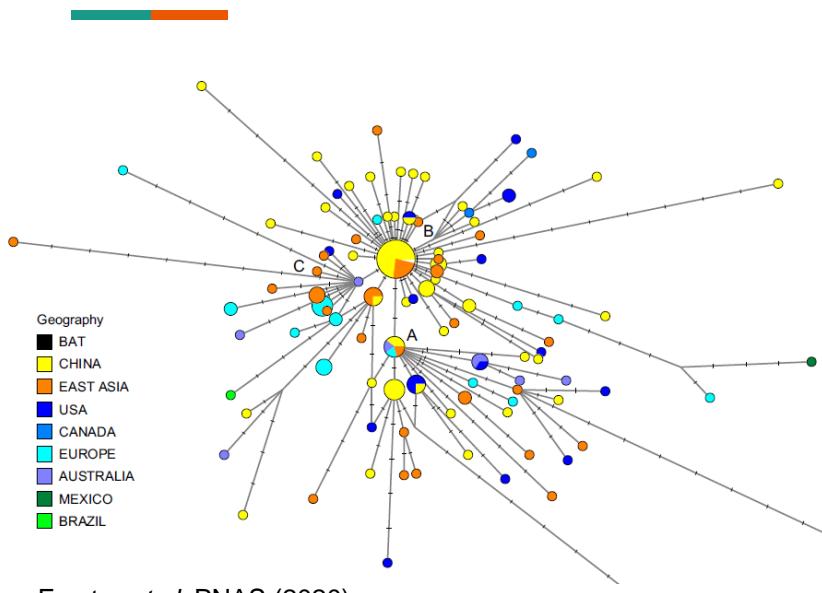
# Linking evolution to function





# Phylogenetics

# Phylogeny



- Clustering of similar taxa with respect to evolutionary distances
- Branch length reflect **time** and **mutation rate**

# Types of phylogeny

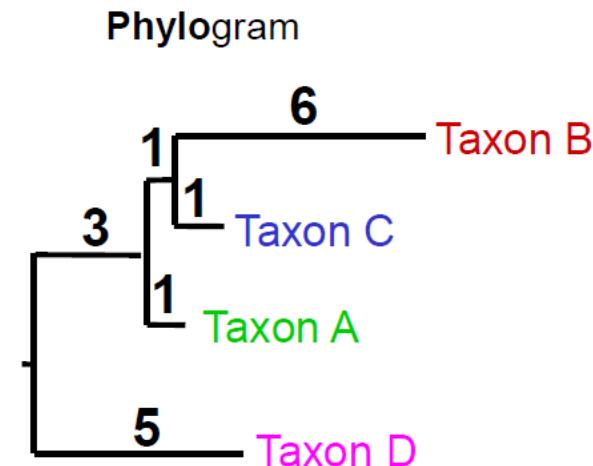
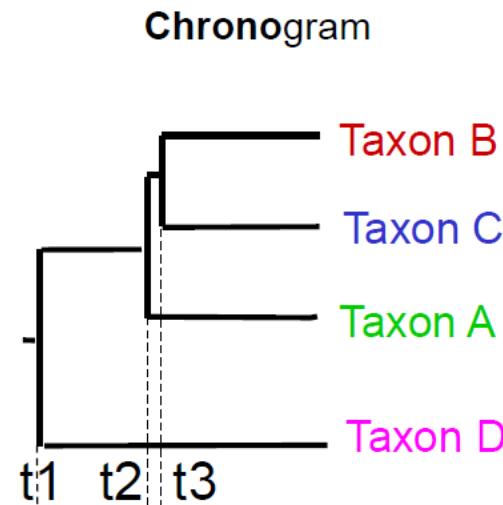
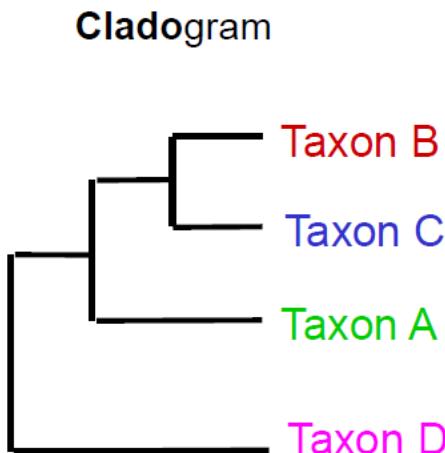


Image from Lecture 18 6.047 MIT OCW

- The choice depends on research question

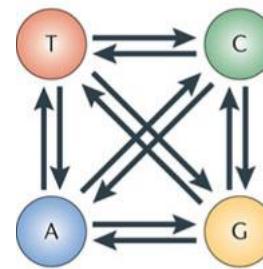
# Ingredients for phylogenetic reconstruction

## Multiple sequence alignment

Scarites	C T T A G A T C G T A C C A A -   - A T T T T A C
Carenum	C T T A G A T C G T A C C A A - T A C - T T T T A C
Pasimachus	A T T A G A T C G T A C C A C T - T - G T T T T A C
Pheropsophus	C T T A G A T C G T T C C C C - - - A C A T T A T T C
Brachinus armiger	A T T A G A T C G T A C C A C - - - A T A T T A T T C
Brachinus hirsutus	T T A G A T C G T A C C A C - - - A T A T T A T T C
Aptinus	C T T A G A T C G T A C C A C - - - A C A T T A T T C
Pseudomorpha	C T T A G A T C G T A C C A C - - - C A T T A T T C

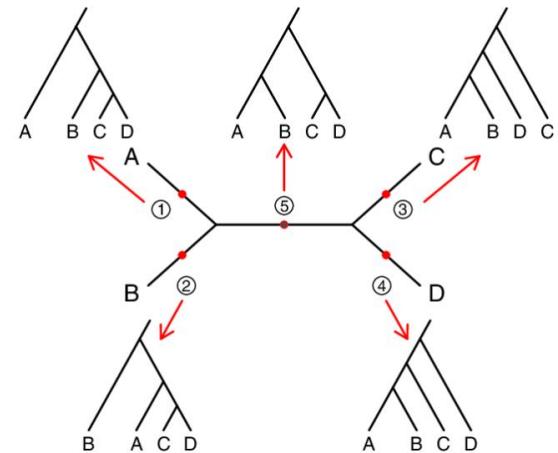
Image from [www.mcqbiology.com](http://www.mcqbiology.com)

## Evolutionary model



Yang & Rannala. Nat Rev Genetics (2012)

## Tree building algorithm



Tian, Y. and Kubatko, L.S. BMC Evol Biol 17(1) (2017)

- Sequence data + substitution model + tree topology and branch length
- Other constraints and non-molecular data can also aid the reconstruction

# Taxon features for phylogenetic reconstruction

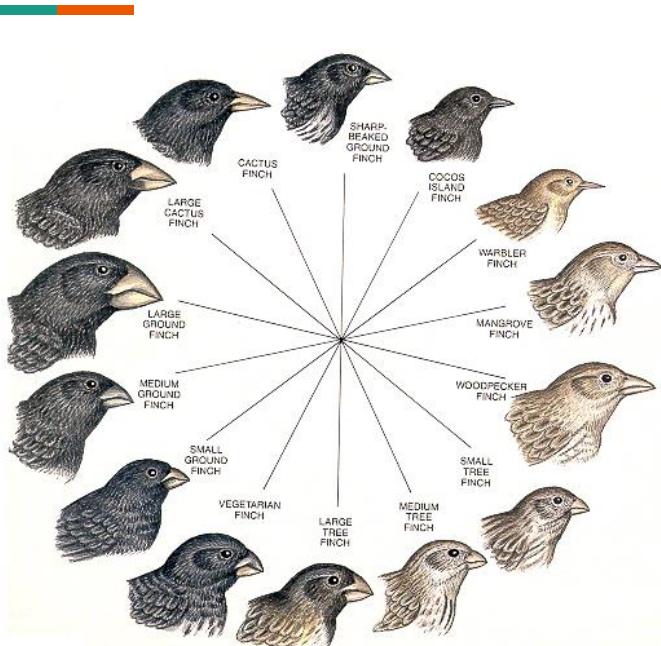
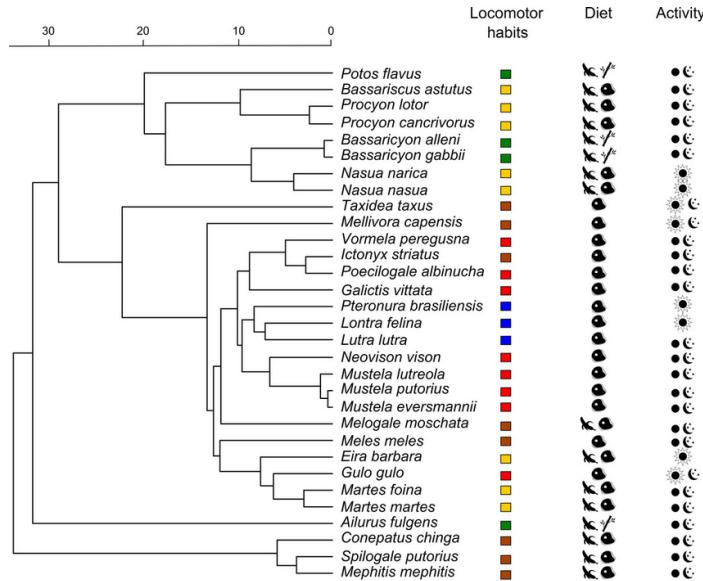


Image from pinterest

- Morphology, diet, or life-style



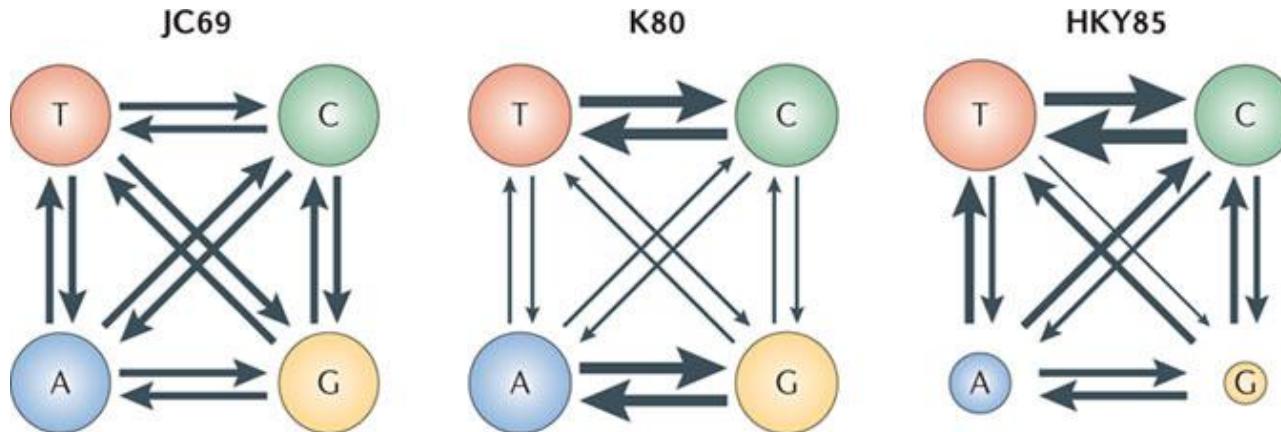
## Legends:

Locomotor habits	Diet	Activity
■ Aquatic	● Carnivorous	○ Diurnal
■ Terrestrial	● Herbivorous	● Nocturnal/crepuscular
■ Arboreal	● Omnivorous	● Arrhythmic
■ Semi-fossorial		
■ Semi-arboreal		



# Evolutionary (substitution) model

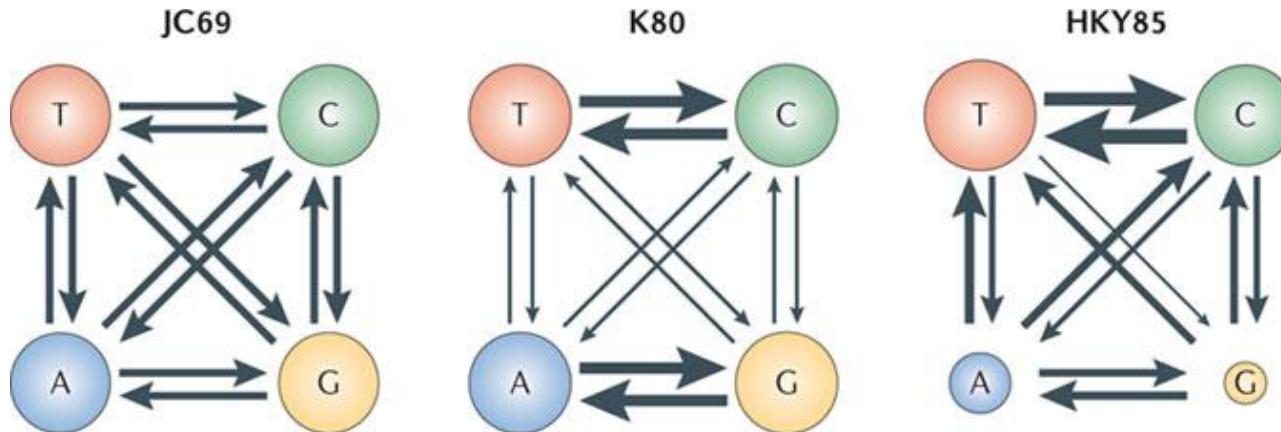
# Basic nucleotide substitution models



Yang & Rannala. Nat Rev Genetics, 13: 303-314 (2012)

- Juke-Cantor assumes **equal base frequencies** and **equal substitution rates**
- Kimura adds **transition rate** and **transversion rate**
- Hasegawa-Kishino-Yano adds **different base frequencies**

# How many parameters?



Yang & Rannala. Nat Rev Genetics, 13: 303-314 (2012)

- Juke-Cantor = 1 (substitution rate)
- Kimura = 2 (transition rate, transversion rate)
- Hasegawa-Kishino-Yano = 5 (transition rate, transversion rate, 3 base frequencies)

# General time-reversible model (GTR)



- Symmetric substitution rates:  $P(A \rightarrow G) = P(G \rightarrow A)$ 
  - 6 parameters (4 nucleotides choose 2)
  - Time-reversible: switching ancestral and descendant taxa does not change the calculation
  - Useful in practice because we often don't know which taxon came first
- Different base frequencies
  - 3 parameters
- This is the most generalized time-reversible model possible
- Often used

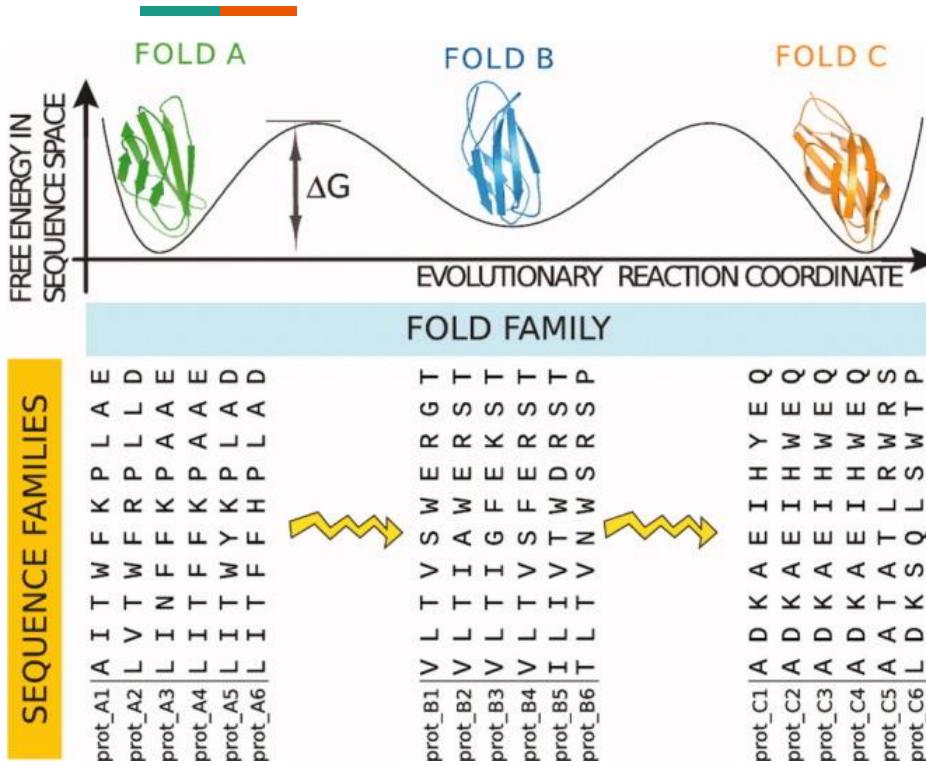
# Basic amino acid substitution models



- BLOSUM & PAM
- Dayhoff 1978
- JTT (Jones-Taylor-Thornton, 1992)
  - Essentially PAM250
- WAG (Whelan-Goldman, 2001)
  - More protein families included

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	7	
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

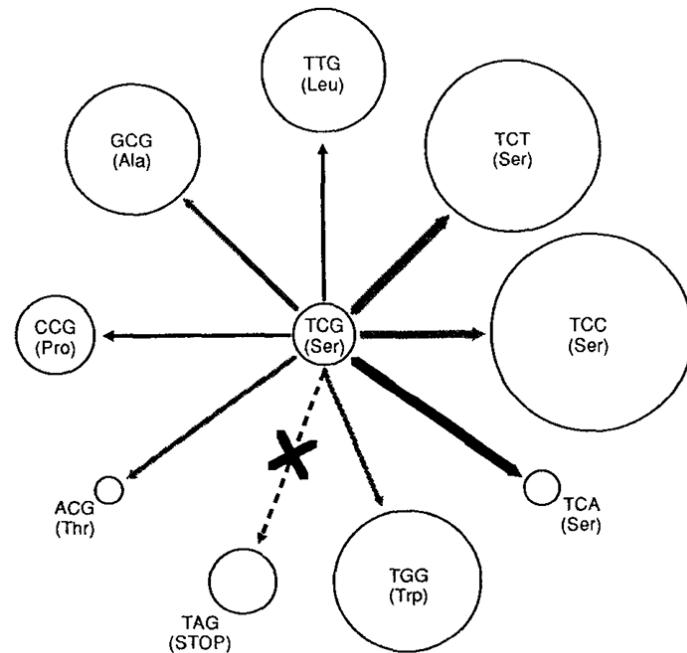
# Structure-based substitution model



- Different positions on different protein families have different evolutionary constraints
- Use energy from interactions between residues in 3D to develop substitution matrix
- Must be tuned for each use

# Codon substitution model

- GY94 (Goldman-Yang)
- Codon frequencies
  - Nucleotide and amino acid frequencies
- Codon neighborhood
  - Substitution rate depends on similarity between coded amino acids
- Natural selection
  - Non-synonymous / synonymous rate



# Codon alignment: amino acid → nucleotide

A. DNA alignment

Q9FPK4	ATGGGTGTTTCAGCTACGAGATGAGGCCACC	TCCGTTATCCTCCGGCTA	GGCTGTTCAAGTCC	TTTGTCTTAGATGCCGAC	AAACTCATT
Q9FPK3	ATGGGTGTTTCAGCTACGAGATGAGGCCACC	TCCGTTATCCTCCGGCTA	GGCTGTTCAAGTCC	TTTGTCTTAGATGCCGAC	AAACTCATT
Q945E7	ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACCC	TCCCGAACATTGCTCCAGCCA	GGCTTTCAAGCT	TTTGTCTTGAGGCTGCC	AAGATTGG
Q6XC94	ATGGGTGTTGGCAGTTATGAGTTTGAGGTAACCC	TCCCGAACATTGCTCCAGCCA	GGCTTTCAAGCT	TTTGTCTTGAGGCTGCC	AAGATTGG
Q6Q4B5	ATGGGTGTTGTGAGTTATGACTTGAGGTAACCC	CTCCCCAATTGCTCCAGCCA	GGCTTTCAAGCT	TTTGTCTTGAGGCTGCC	AAGATTGG
Q43549	ATGGGTGTTTCATTAGAAAACGTGAGTTTACCC	TCCGTCATTCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AAACTCATC
Q4VPJ1	ATGGGTGTTTCACATACGAACTCTGAGTCCACC	TCCGTCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AAACTCATC
Q84LA7	ATGGGTGTCCTCACATACGAACTCCGAATTGACCC	TCTATCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AAACTCATC
Q4VPI3	ATGGGTGTTTCACATACGAACTCCGAATTGACCC	TCTATCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AAACTCATC

B. Back-translation from protein alignment

Q9FPK4	ATCGGTGTTTCAGCTACCACGATGAGGCCACC	TCCGTTATCCTCCGGCTA	AGGCTGTT	--AAGTCC	TTTGTCTTAGATGCCGACAACCTCATT
Q9FPK3	ATCGGTGTTTCAGCTACCACGATGAGGCCACC	TCCGTTATCCTCCGGCTA	AGGCTGTT	--AAGTCC	TTTGTCTTAGATGCCGACAACCTCATT
Q945E7	ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACCC	CTCCCCAATTGCTCCAGGCC	AGGCTT	--AAGGCT	TTTGTCTTGAGGCTGCCAAGATTGG
Q6XC94	ATGGGTGTTGGCAGTTATGACTTGAGGTAACCC	CTCCCCAATTGCTCCAGGCC	AGGCTT	--AAGGCT	TTTGTCTTGAGGCTGCCAAGATTGG
Q6Q4B5	ATGGGTGTTGTGAGTTATGACTTGAGGTAACCC	CTCCCCAATTGCTCCAGGCC	AGGCTT	--TCAAGGCT	TTTGTCTTGACCTGCCAAGATTGG
Q43549	ATGGGTGTTTCATTAGAAAACGTGAGTTTACCC	CTCCCCCTGCTA	AGGCTT	--AATGCC	TTTGTCTTGATGCTGACAAACCTCATC
Q4VPJ1	ATGGGTGTTTCACATACGAACTCTGAGTCCACC	CTCCCCCTGCTA	AGGCTT	--AATGCC	TTTGTCTTGATGCTGACAAACCTCATC
Q84LA7	ATGGGTGTCCTCACATACGAACTCCGAATTGACCC	CTCCCCCTGCTA	AGGCTT	--AATGCC	TTTGTCTTGATGCTGACAAACCTCATC
Q4VPI3	ATGGGTGTTTCACATACGAACTCCGAATTGACCC	CTCCCCCTGCTA	AGGCTT	--AATGCC	TTTGTCTTGATGCTGACAAACCTCATC



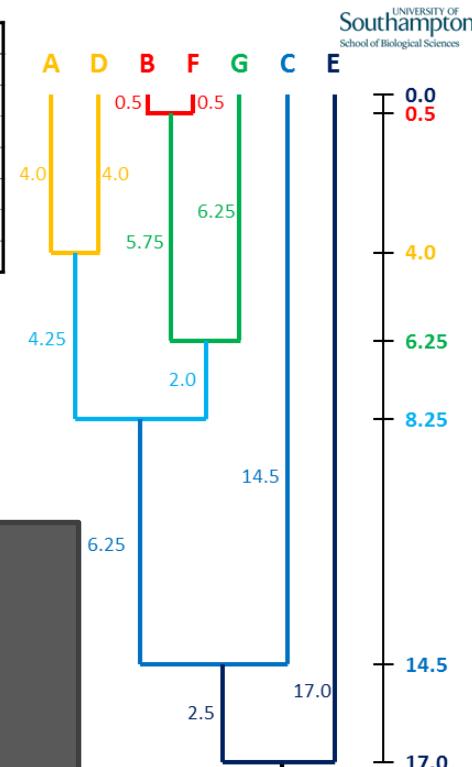


# Tree building algorithms

# UPGMA: Unweighted pair-group method arithmetic mean

1

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

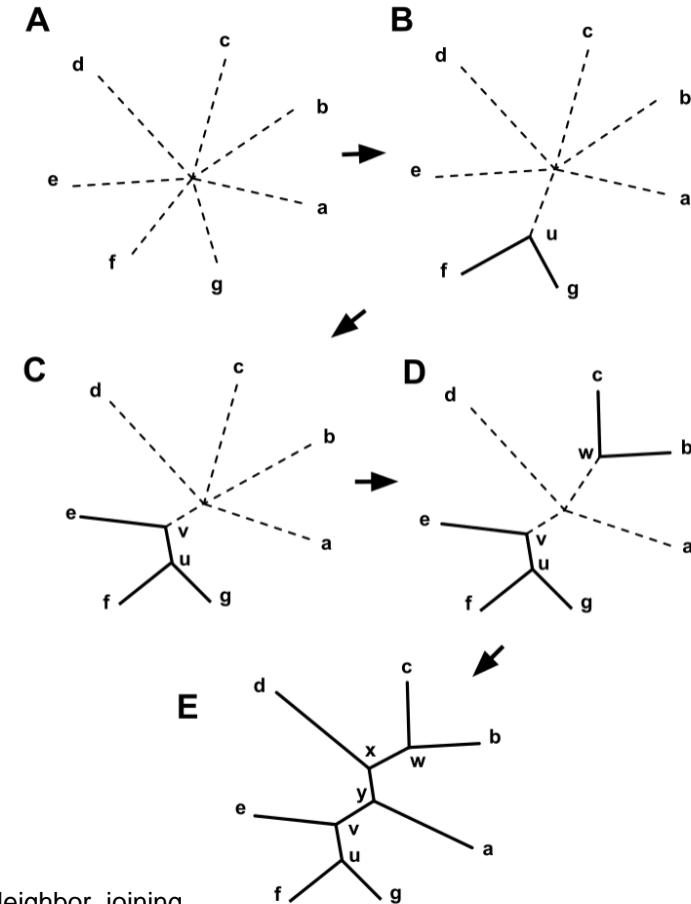


# Neighbor joining

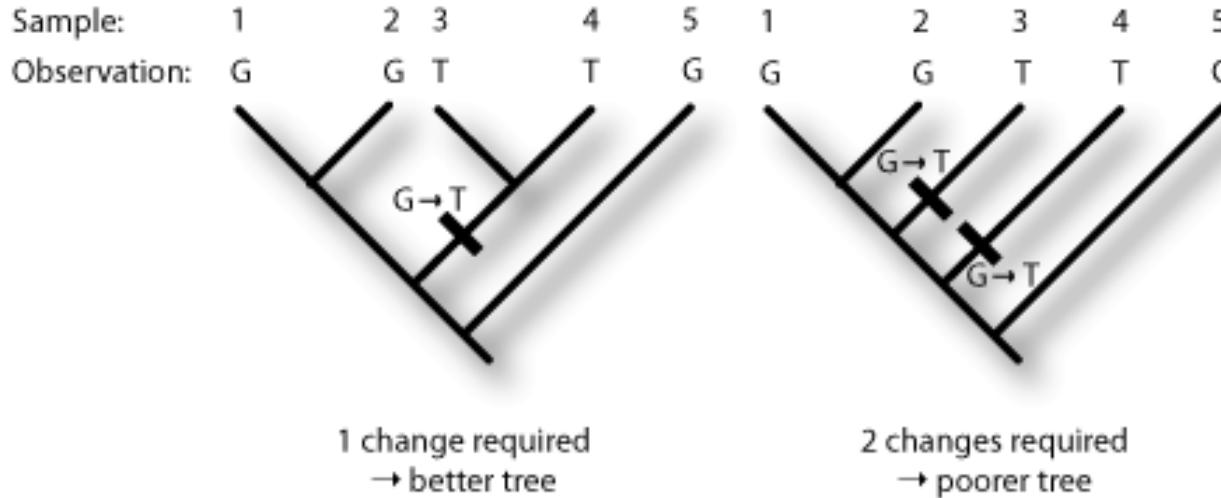
- Start with a star tree
- Join taxa  $i, j$  with smallest Q score

$$Q(i,j) = (n - 2)d(i,j) - \sum_{k=1}^n [d(i,k) + d(j,k)]$$

- Join taxa that are **similar to each other** but **dissimilar from other taxa**



# Maximum parsimony / Minimum evolution

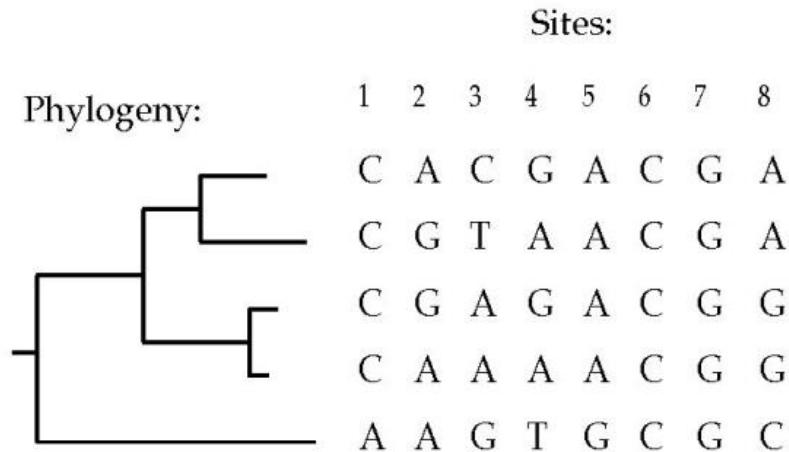


<https://biology.stackexchange.com/questions/60161/which-nucleotide-as-start-point-for-maximum-parsimony>

- Explanation requiring minimum number of changes is the most likely
- Cannot handle distant evolution: A → T → A

# Maximum likelihood approach

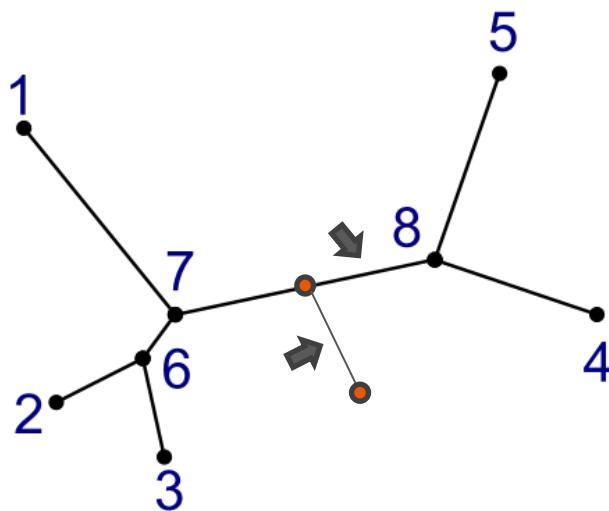
---



[https://homes.cs.washington.edu/~ruzzo/courses/gs559/09wi/lectures/8A\\_likelihood.pdf](https://homes.cs.washington.edu/~ruzzo/courses/gs559/09wi/lectures/8A_likelihood.pdf)

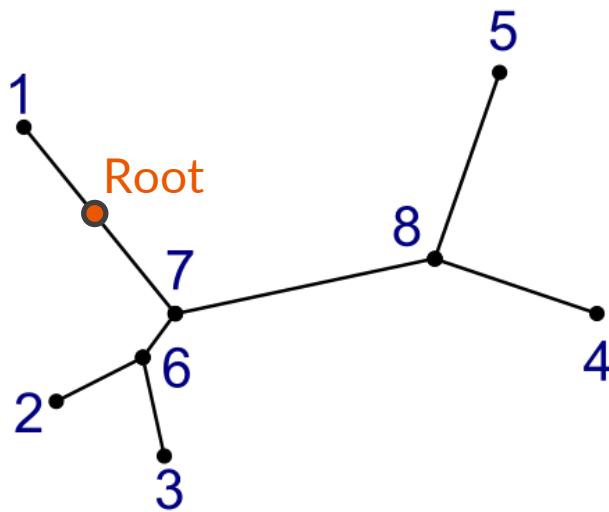
- Likelihood =  $P(\text{sequence data} \mid \text{substitution model, tree topology, branch lengths})$
- Given a phylogenetic tree topology, we can alter branch lengths (one branch at a time) to search for the best answer
- The problem is how to find the best tree topology

# Number of possible trees with distinct leaves



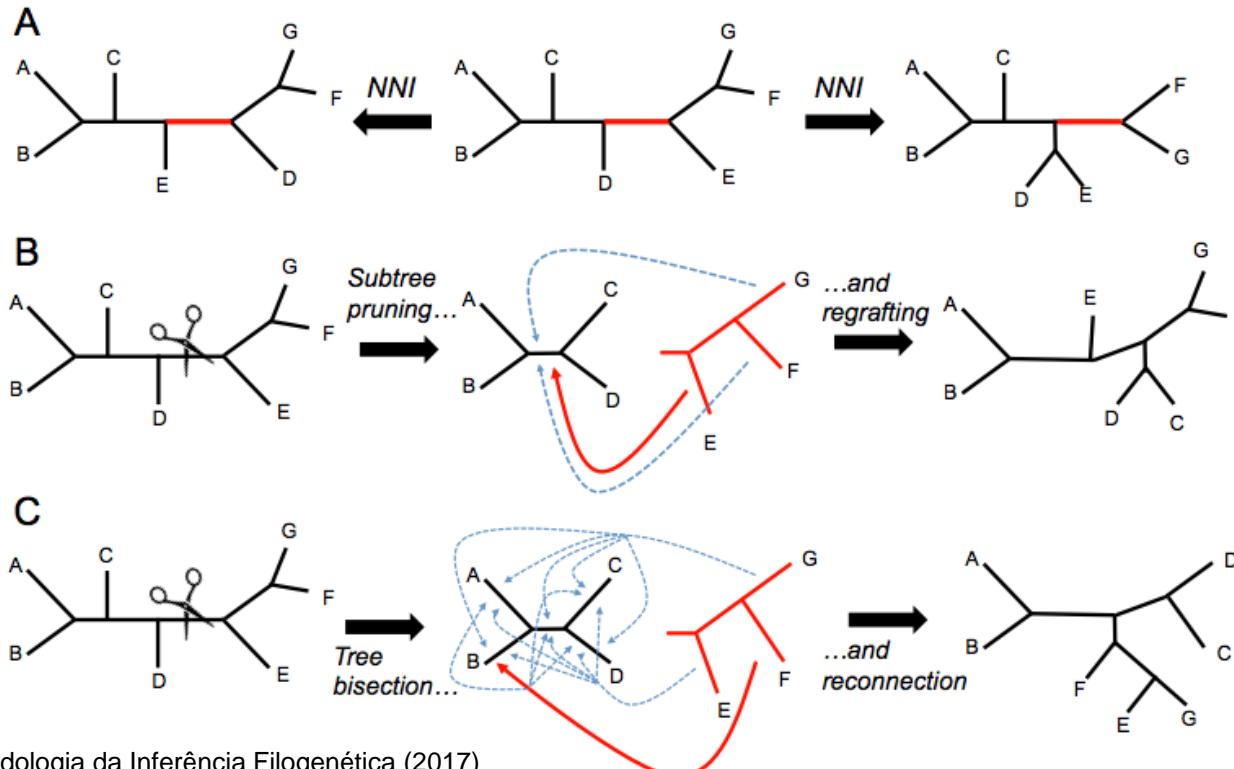
- Unrooted tree with  $n$  leaves has  $2n-2$  nodes
  - Adding a leaf will also create one internal node
- There are  $2n-3$  branches
  - Adding a leaf will create two branches
  - One new branch and split an existing branch
- From each unrooted tree with  $n$  leaves, there are  $2n-3$  locations to attach a new leaf
- Number of unrooted trees with  $n$  leaves, for  $n > 2$ , is  $(2n-5) \times (2n-7) \times (2n-9) \times \dots \times 1$

# Number of possible trees with distinct leaves



- For each unrooted tree with  $n$  leaves, there are  $2n-3$  locations to designate as the root (common ancestor)
- Number of rooted trees with  $n$  leaves, for  $n > 2$ , is  $(2n-3) \times (2n-5) \times (2n-7) \times (2n-9) \times \dots \times 1$
- Number of rooted trees with 10 leaves =  $17 \times 15 \times 13 \times 11 \times 9 \times 7 \times 5 \times 3 = 34,459,425$

# Heuristic tree search algorithms

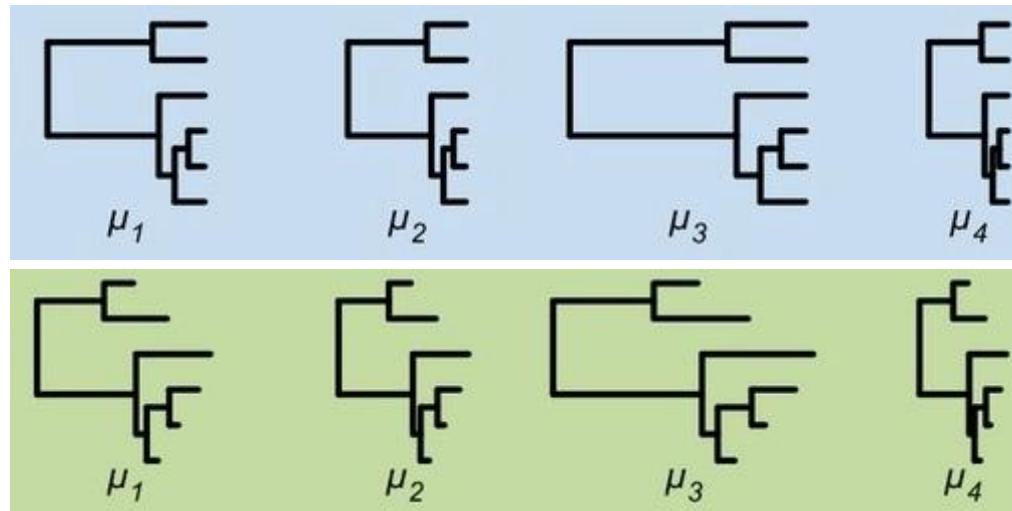




# Additional evolutionary parameters

# Molecular clock assumption

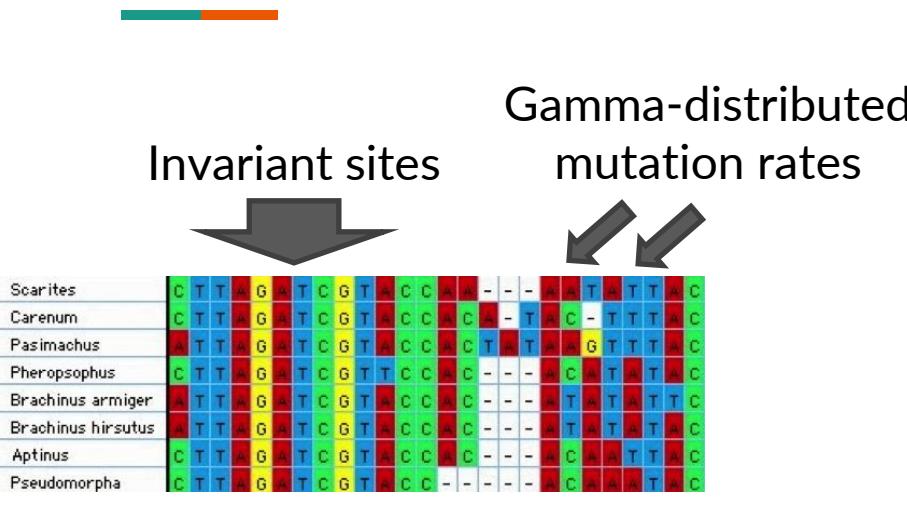
---



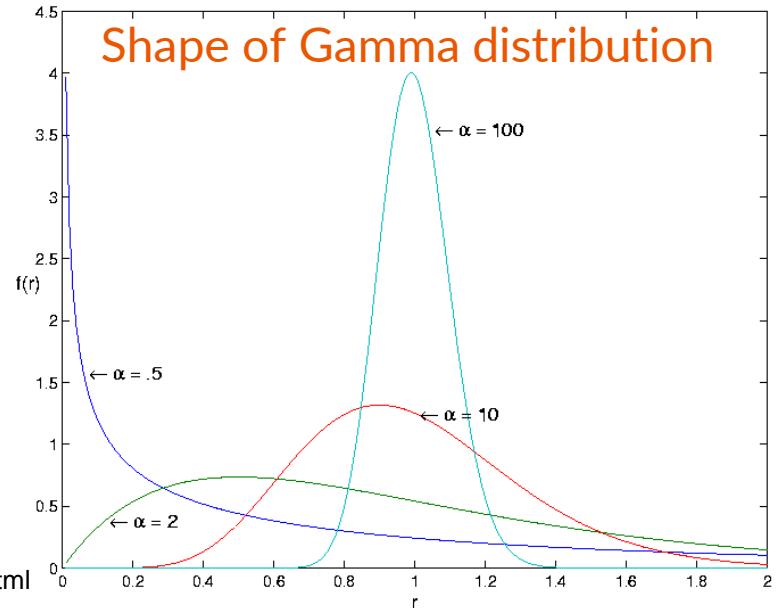
Ho, S.Y.W. and Duchene, S. Molecular Ecology 23:5947-65 (2014)

- Molecular clock assumes constant evolutionary rate throughout the tree
- Same root-to-tip distance → allow dating of evolutionary events

# Site-specific evolutionary models



<http://www.bioinf.man.ac.uk/resources/phase/manual/node81.html>



- Gamma distribution  $\Gamma(\alpha, \alpha)$  can mimic diverse shapes with mean = 1
- High  $\alpha$  results in bell shape while low  $\alpha$  yields higher variance =  $1/\alpha$

# Natural selection on codon model

---

Nonsynonymous / Synonymous substitution

TCC	GAT	<u>ATA</u>	TGG	<u>CAA</u>	CCC	<u>GAC</u>	AAA
S	D	I	W	Q	P	D	K
TCA	GAT	<u>CTA</u>	TGG	<u>CAG</u>	CCC	<u>CAC</u>	AAA
S	D	L	W	Q	P	R	K

Luo, H. Frontiers in Microbiology 6:191 (2015)

- Null hypothesis: synonymous and non-synonymous occurs proportional to the number of corresponding codon positions ( $dN/dS = 1$ )
- Alternative hypothesis:  $dN/dS$  can differ from 1

# Nested relationship between nucleotide models

---

- GTR → HKY85 if we set all transitions with the same rate and all transversions with the same rate
- HKY85 → K80 if we set all base frequencies to be 1/4
- K80 → JC69 if we set transition rate and transversion rate to be the same
- Model with invariant site → without invariant site
- Model with site-specific Gamma rate → without site-specific Gamma rate
- Model without molecular clock → with molecular clock
- Model with any dN/dS → with  $dN/dS = 1$
- Nested model testing

# Nested model testing (chi-squared & likelihood)

Model complexity ↓

<u>Model</u>	<u>-lnL</u>
JC69	3585.54820
F81	3508.04085
<b>HKY85</b>	<b>3233.34395</b>
TrN93	3232.29439

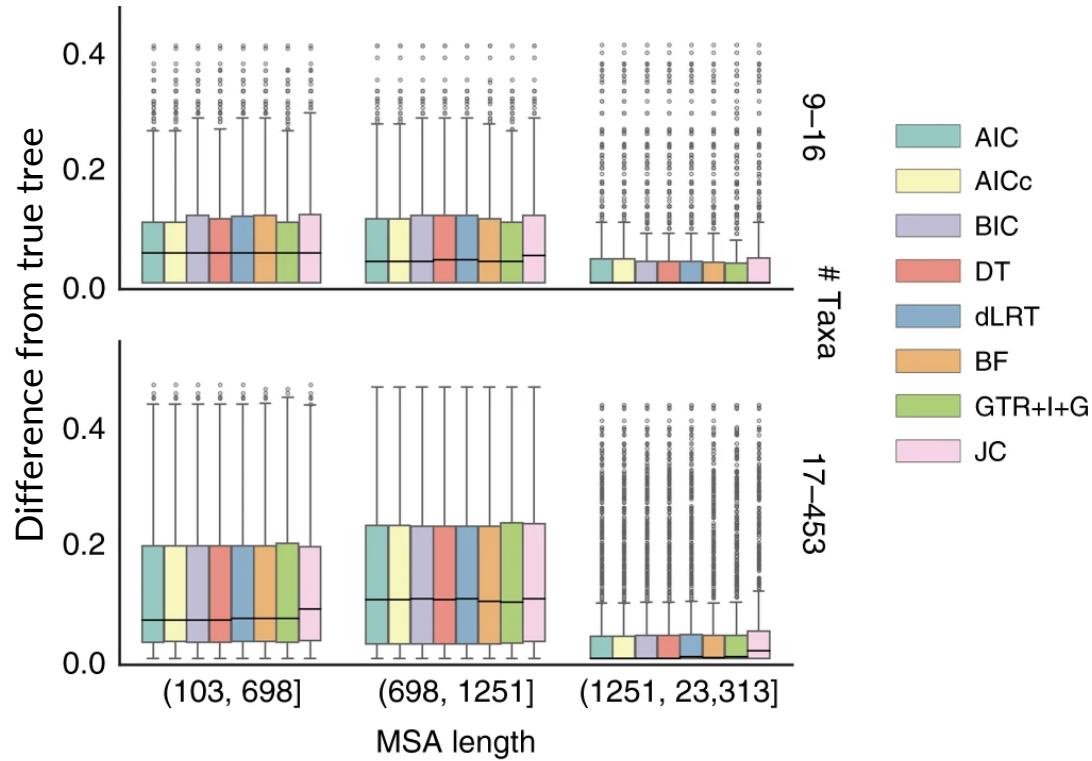
<u>models</u>	<u>diff. DF = q</u>	<u>X<sup>2</sup></u>	<u>P</u>	
JC-F81	3 - 0 = 3	155	0	F81 model fits the data significantly better than JC
F81-HKY85	4 - 3 = 1	549.4	0	HKY85 model fits the data significantly better than F81
KHY-TrN	5 - 4 = 1	2.1	0.15	TrN model does not fit the data significantly better than HKY85

<u>Model</u>	<u>-lnL</u>
HKY85	3233.34395
<b>HKY85 +G</b>	<b>3145.29031</b>
HKY85 +I+G	3142.36439

<u>models</u>	<u>diff. DF = q</u>	<u>X<sup>2</sup></u>	<u>P</u>	
HKY85-vs. +G	1	176	0	Adding site-specific rate fits the data significantly better
HKY85+G vs. I+G	1	5.85	0.015	Adding invariant sites does not fit the data significantly better

# Minimal impact from model selection

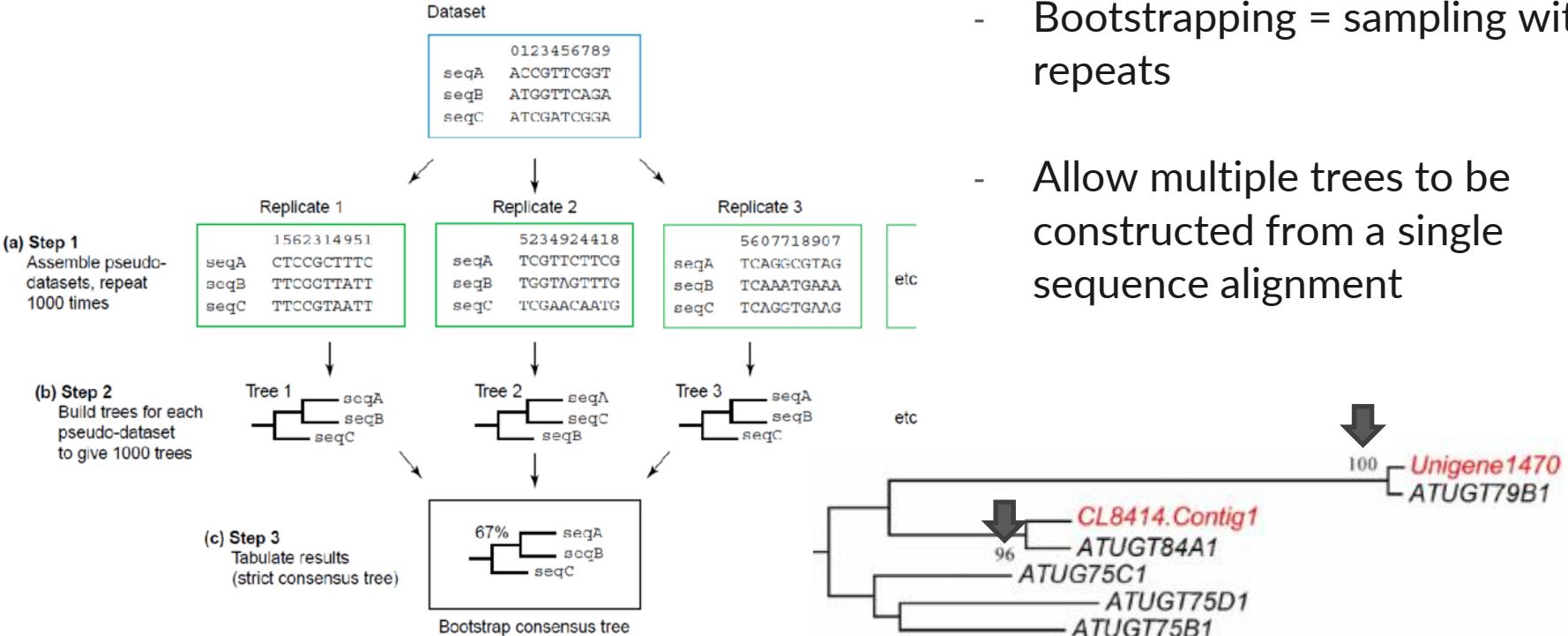
- Evaluate tree topology and ancestral sequence reconstruction
- GTR+I+G (the most complex model) yields similar results as JC (the simplest model)
- Consistent results when using various criteria to select the best nucleotide model (AIC, BIC, etc.)



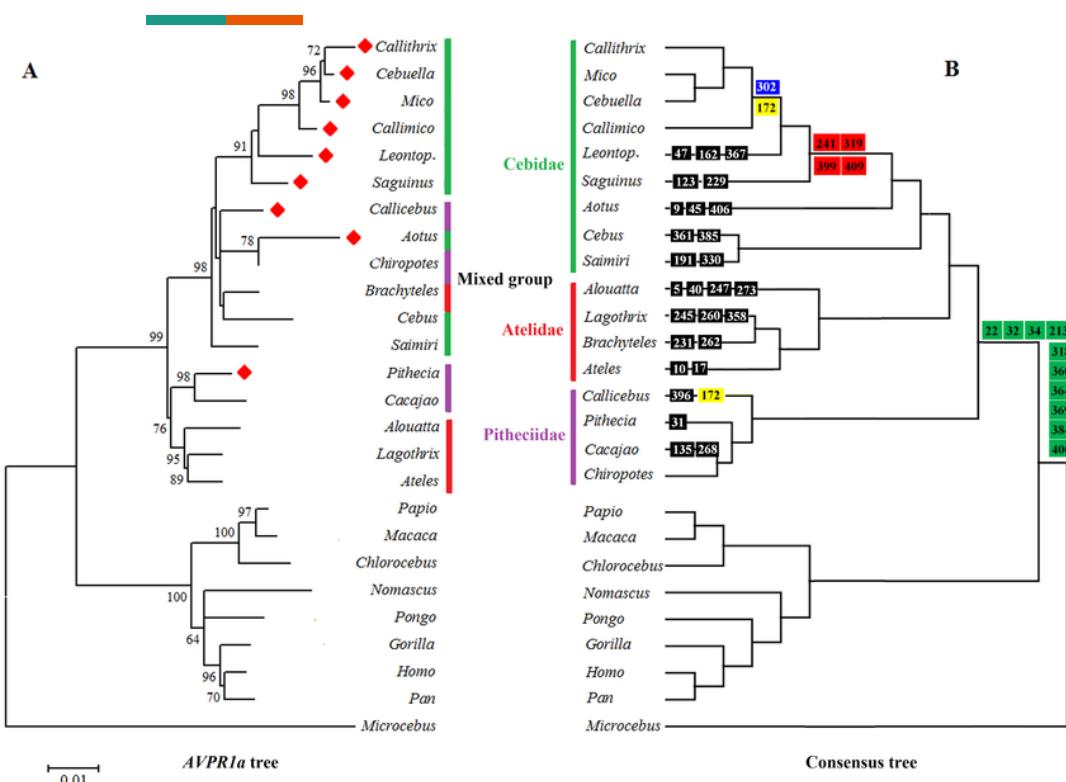


# Reliability test with bootstrapping

# Bootstrap support for taxa group



# Gene tree vs concatenated tree



- Different gene may have different evolutionary history
- Housekeeping genes vs taxaspecific genes
- Can provide insights into the evolutionary process



# **Example of how to setup phylogenetic reconstruction**

# MEGA: A Windows GUI tool for phylogenetic analysis

Filter tutorials by topic:

- Align Sequences
- BLAST Search
- Bootstrap Tree
- Calibrate
- Distance Estimation
- Edit Sequences
- Edit Tree
- Evolutionary Probability
- Gene Duplication
- Grouping Taxa
- Installation
- Model Selection
- Pairwise Distances
- Phylogeny Construction
- Substitution Matrix
- Trace Files



Molecular Evolutionary  
Genetics Analysis

## TUTORIALS

Below are links to online video lectures and tutorials for multiple versions of MEGA. The first section of videos were created by members of Dr. Sudhir Kumar's lab at the Institute for Genomics and Evolutionary Medicine ([iGEM](#)) at Temple University. The rest of the videos were produced by users of MEGA. To assemble this collection of videos, the MEGA team performed a search of YouTube for instructional MEGA videos and assembled this collection of the most popular videos found. If you would like to suggest additions to this collection, please contact us by using the [feedback page](#).

### KUMAR LAB VIDEOS

Molecular Dating with MEGA

Choosing and Acquiring  
Sequences Part 1

Choosing and Acquiring  
Sequences Part 2

Reconstructing Ancestral

Relative Rate Framework for

Inferring Selection with MEGA

# Substitution model choices

## PHYLOGENY mode

MX: Analysis Preferences

Phylogeny Reconstruction |

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Amino acid
Genetic Code Table	→ Standard
Model/Method	→ Jones-Taylor-Thornton (JTT) model
RATES AND PATTERNS	
Rates among Sites	→ Poisson model
No of Discrete Gamma Categories	→ JTT with Freqs. (+F)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ WAG model

Amino acid-based models

MX: Analysis Preferences

Phylogeny Reconstruction |

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Genetic Code Table	→ Not Applicable
Model/Method	→ Tamura-Nei model
RATES AND PATTERNS	
Rates among Sites	→ Jukes-Cantor model
No of Discrete Gamma Categories	→ Kimura 2-parameter model
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Hasegawa-Kishino-Yano model

Nucleotide-based models

# Maximum likelihood parameters

Phylogeny Reconstruction	
Option	Setting
<b>ANALYSIS</b>	
Statistical Method →	Maximum Likelihood
<b>PHYLOGENY TEST</b>	
Test of Phylogeny →	None
No. of Bootstrap Replications →	None Bootstrap method
<b>SUBSTITUTION MODEL</b>	
Substitutions Type →	Nucleotide
Model/Method →	Tamura-Nei model
<b>RATES AND PATTERNS</b>	
Rates among Sites →	Uniform Rates
No of Discrete Gamma Categories →	Not Applicable
<b>DATA SUBSET TO USE</b>	
Gaps/Missing Data Treatment →	Use all sites
Site Coverage Cutoff (%) →	Not Applicable
<b>TREE INFERENCE OPTIONS</b>	
ML Heuristic Method →	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML →	Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File →	Not Applicable
Branch Swap Filter →	None
<b>SYSTEM RESOURCE USAGE</b>	
Number of Threads →	7

Bootstrapping

Evolutionary / substitution model

Site-specific evolutionary rate

How to handle gap (indel) and missing data

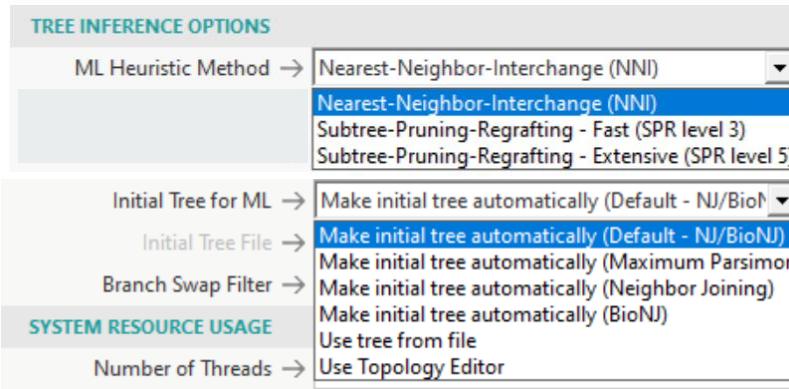
Detailed tree building method

Parallel processing

# Maximum likelihood parameters



Invariant site and Gamma-distributed site-specific substitution rates



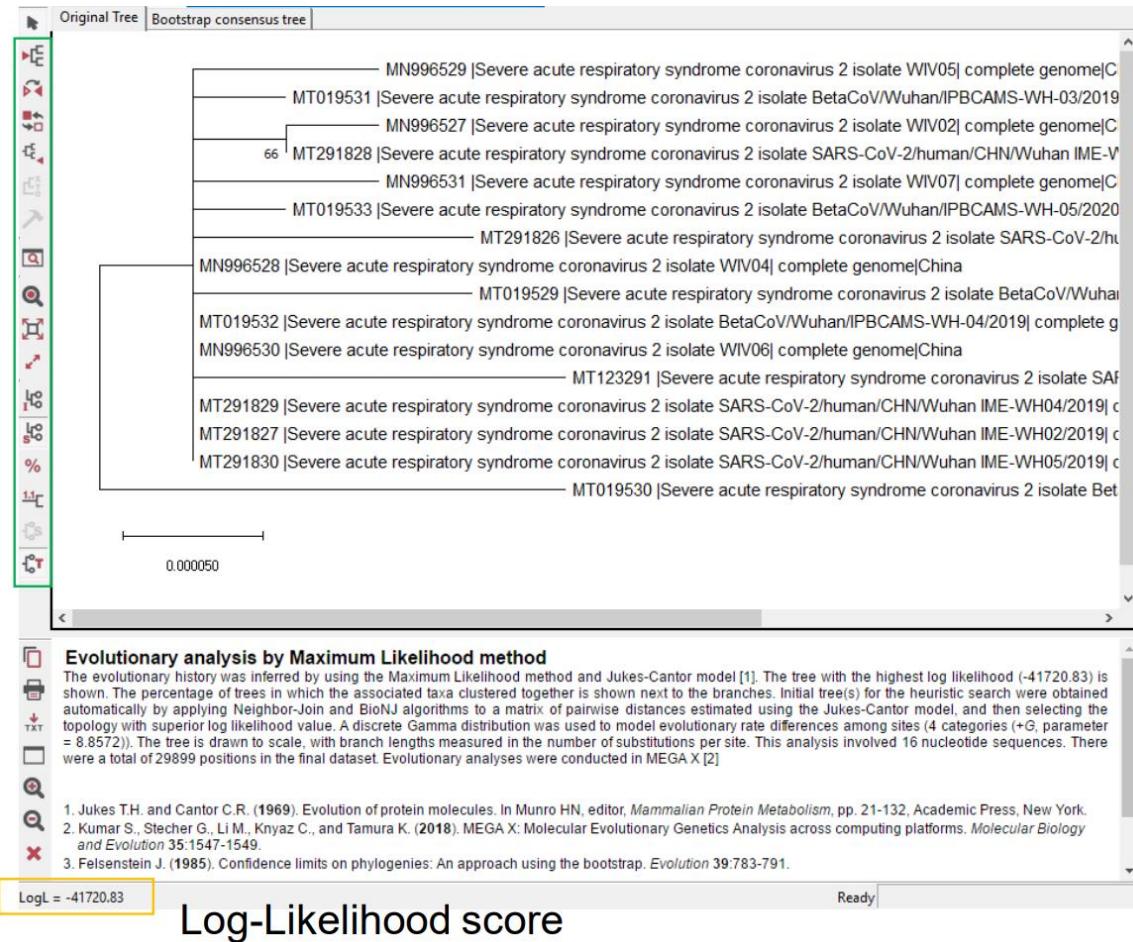
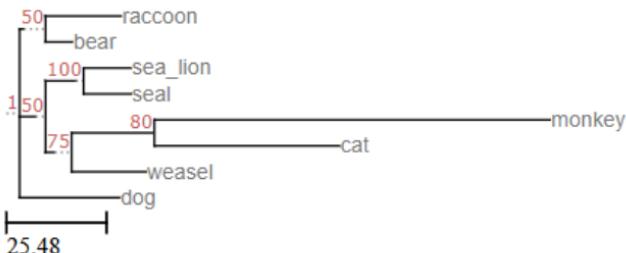
Tree search method

Initial tree can be built using quick and simple method like neighbor joining

# Output

## NEWICK format

```
((raccoon:19.19959,bear:6.80041)50:0.84  
600,((sea_lion:11.99700,seal:12.00300)  
100:7.52973,((monkey:100.85930,cat:47.  
14069)80:20.59201,weasel:18.87953)75:  
2.09460)50:3.87382,dog:25.46154);
```



## Other useful tools

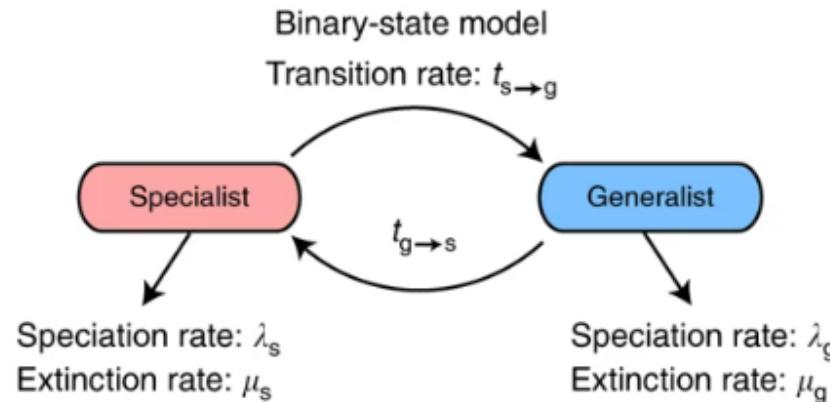
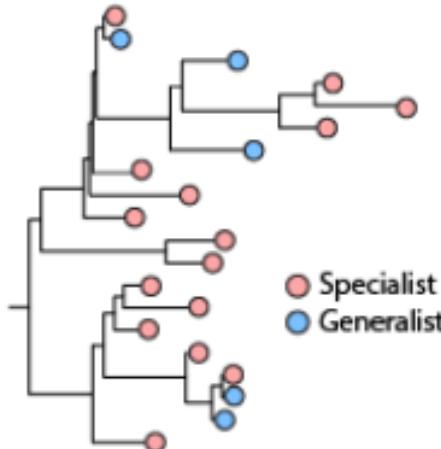
---

- RAxML, PAUP are standard phylogenetic reconstruction tools
- PAML specializes in natural selection analysis (dN/dS)
- FastTree specializes in speed for large dataset
- MrBayes, BEAST use Bayesian approach
  - $P(\text{tree topology}, \text{branch lengths} \mid \text{substitution model}, \text{sequence data})$
- More: <https://evolution.genetics.washington.edu/phylip/software.html#methods>



# Character state evolution

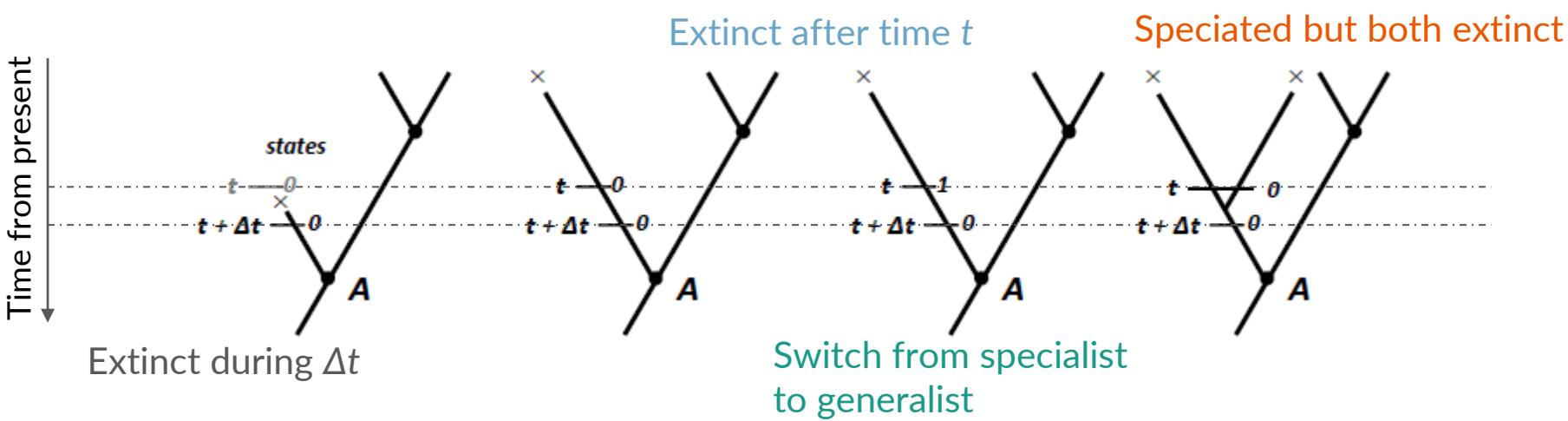
# Impact of character state on evolutionary process



Sriswasdi et al. Nat Comm (2017)

- Assume that being a generalist or a specialist would impact the speciation rate and extinction rate
- The two state can switch during evolution

# Estimating state-specific parameters

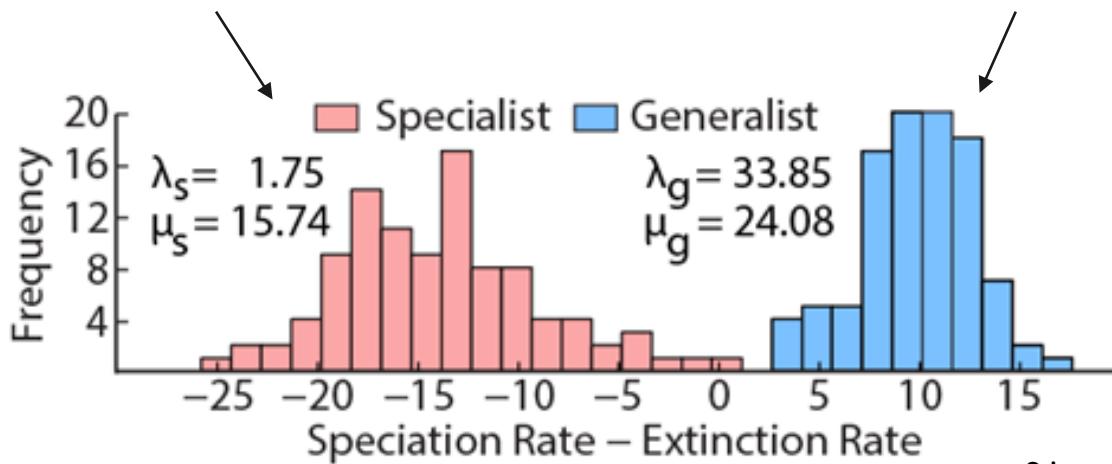


$$\begin{aligned} Extinct_s(t + \Delta t) = & \mu_s \Delta t \\ & + (1 - \mu_s \Delta t)(1 - \tau_{s \rightarrow g} \Delta t)(1 - \lambda_s \Delta t) Extinct_s(t) \\ & + (1 - \mu_s \Delta t)(\tau_{s \rightarrow g} \Delta t)(1 - \lambda_s \Delta t) Extinct_g(t) \\ & + (1 - \mu_s \Delta t)(1 - \tau_{s \rightarrow g} \Delta t)(\lambda_s \Delta t) Extinct_s(t)^2 \end{aligned}$$

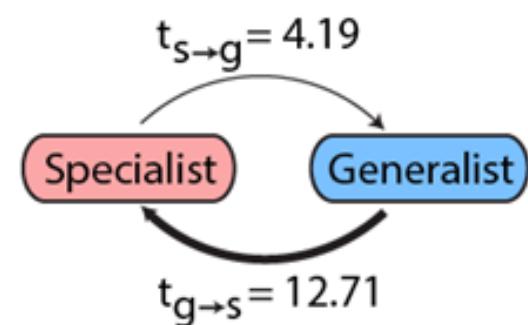
# Estimating state-specific parameters

—

Net reduction in # of species



Net increase in # of species



Sriswasdi et al., Nature Comm 2017

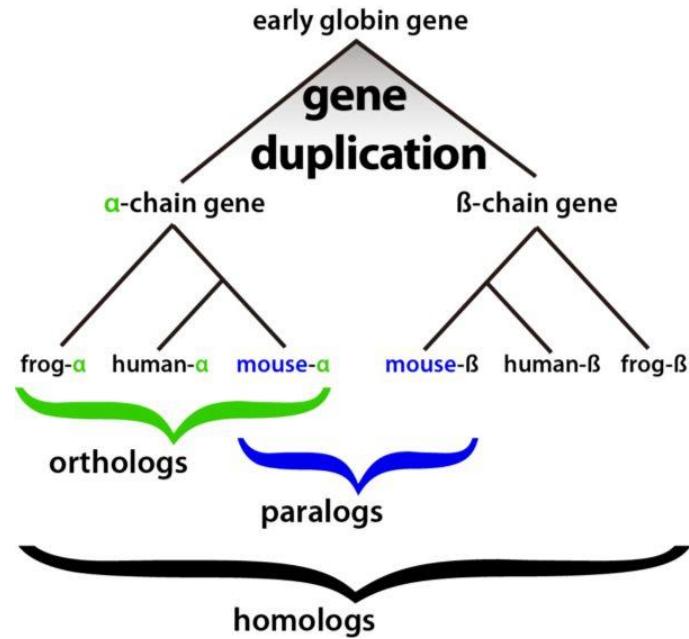


# Ortholog and paralog mapping

# Homolog

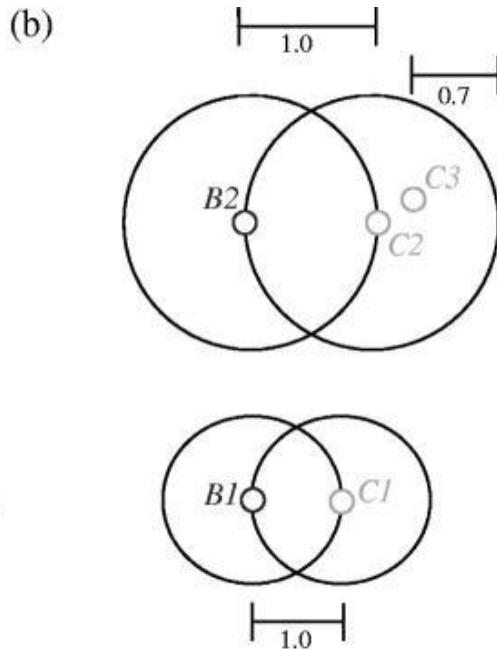
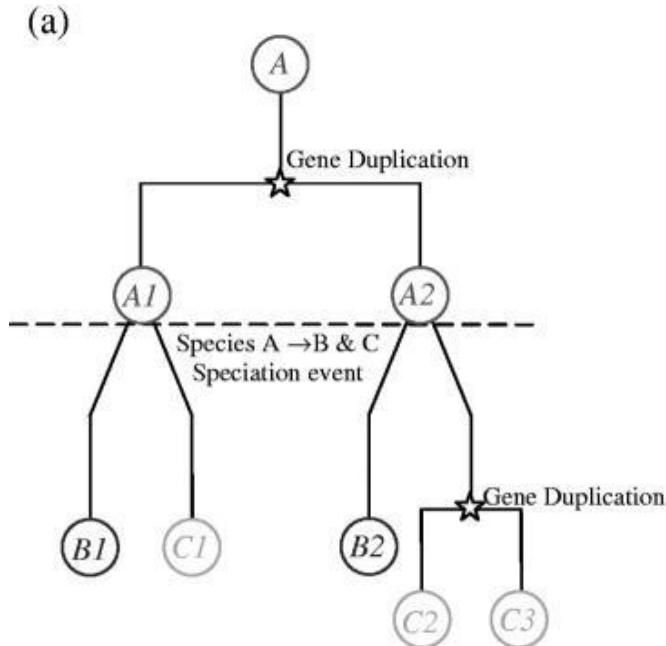
---

- Ortholog = genes across different species that descended from a common ancestor
- Paralog = genes within the same species that descended from duplicates of a common ancestor
- For phylogenetic reconstruction, we usually want to include all homologs together
  - Study gene family expansion



<https://sites.google.com/site/jkim339n/part2a>

# All-vs-all BLAST to find ortholog and paralog

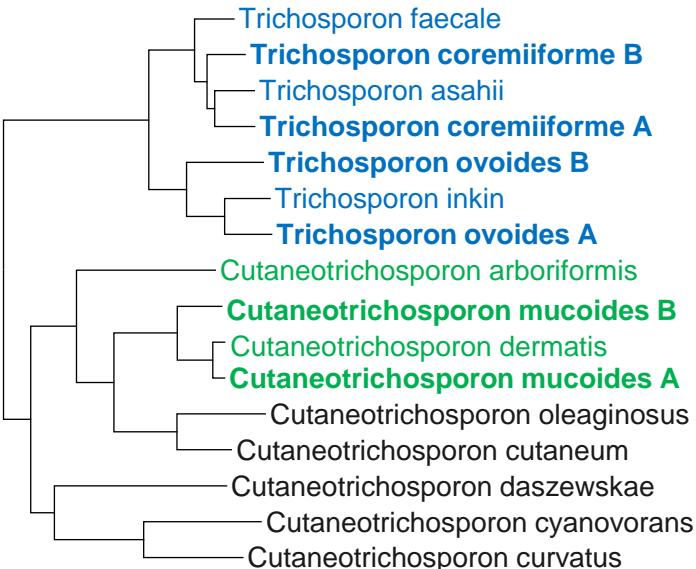


B2-C2 are reciprocal best hits  
Ortholog

C3 is similar to C2 (high BLAST score) but not as similar to B2  
Paralog

B1-C1 are reciprocal best hits  
Ortholog

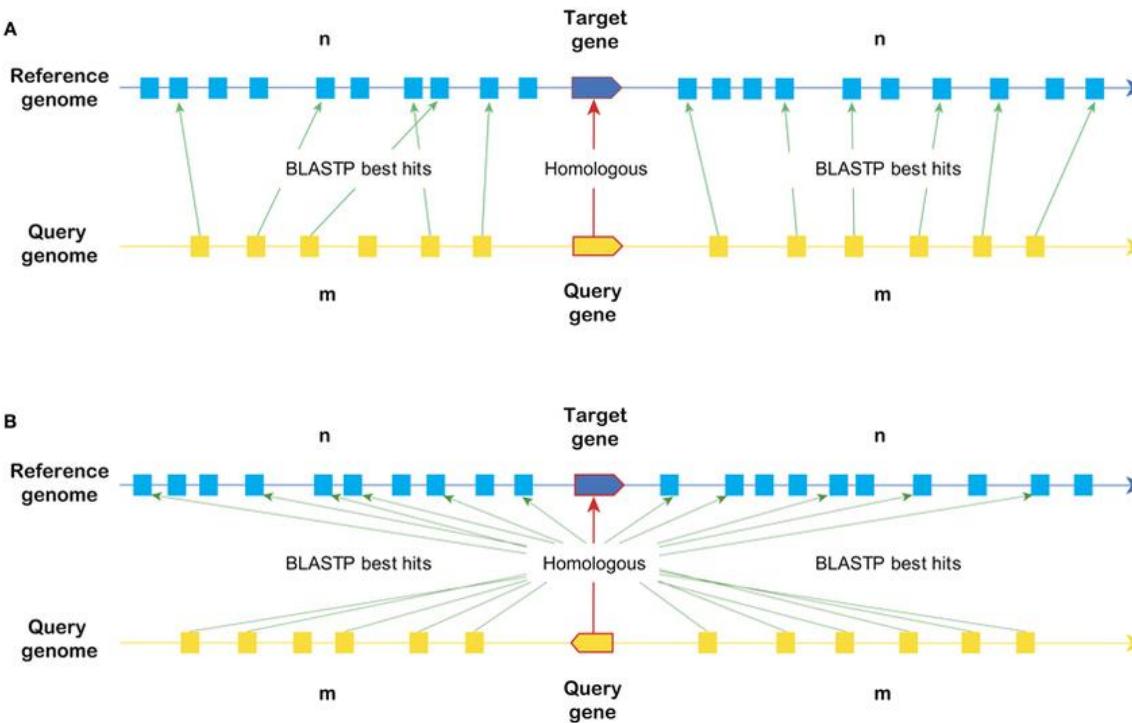
# Phylogenetic reconstruction for hybrid genomes



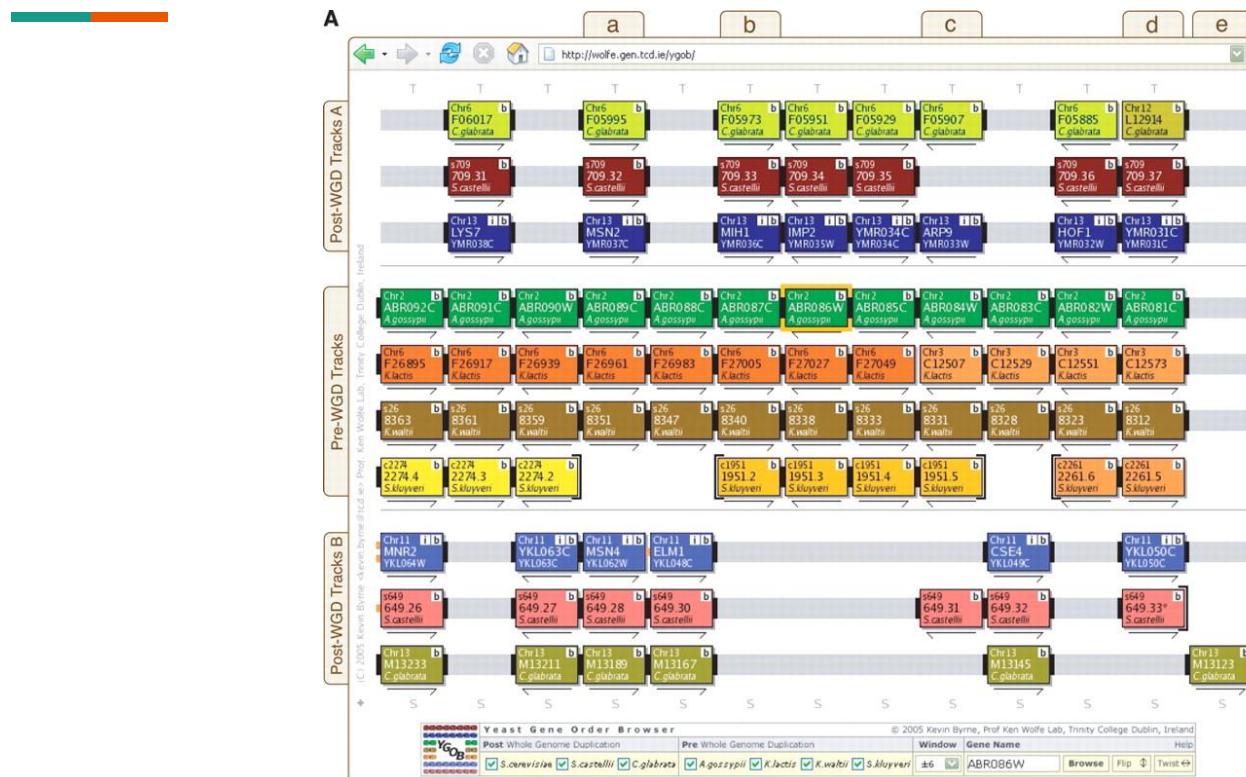
Takashima, M. et al. Yeast 2017

- Some species are hybrid (merging of two different genomes)
- Use all-vs-all BLAST to map orthologs and paralogs
- But which paralogs should be grouped together into A and B tracks
  - Sequence similarity to anchor taxa
  - Gene synteny structure

# Synteny: conservation of gene order on chromosome



# **Yeast Gene Order Browser**

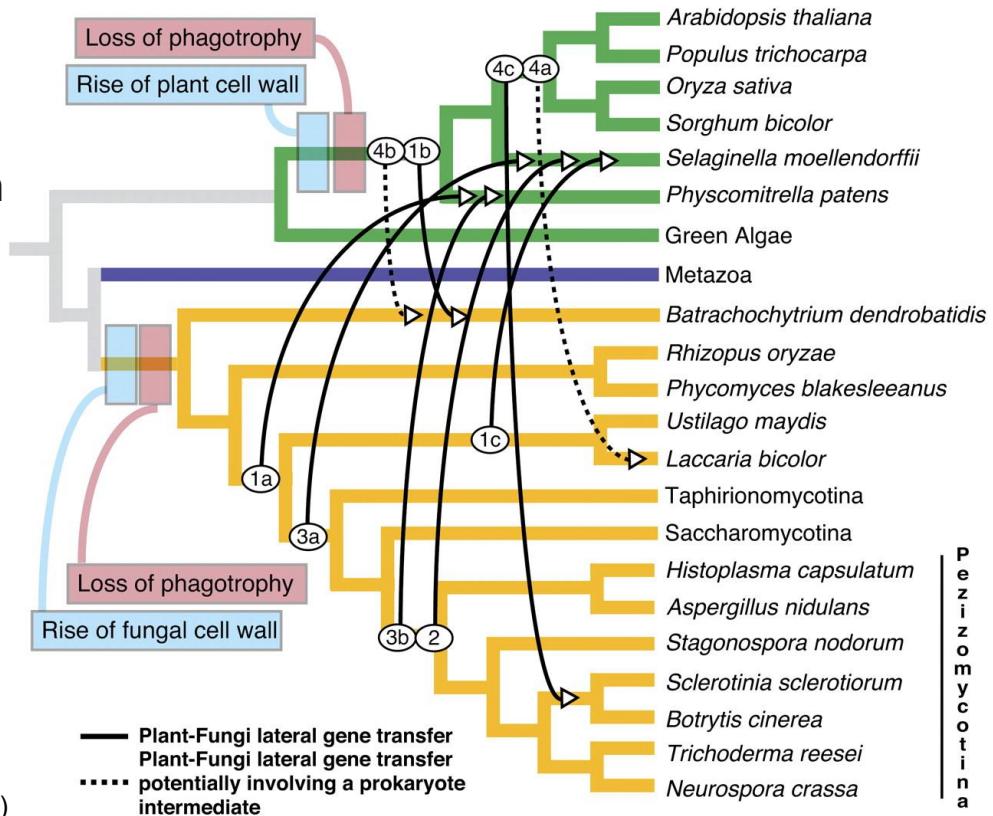




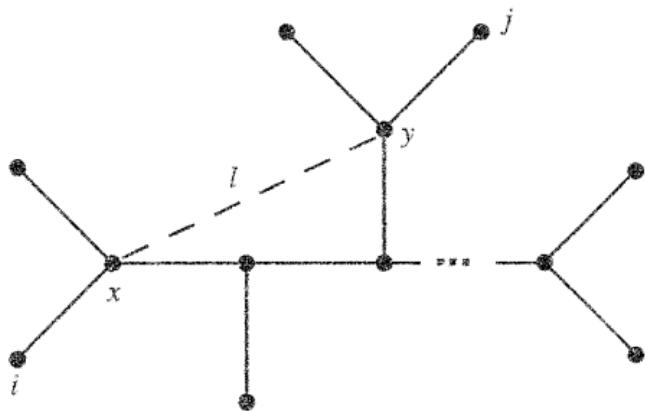
# Horizontal gene transfer

# Horizontal gene transfer (HGT)

- Evolutionary models assume vertical transmission of DNA from ancestor to offspring
- HGT creates similarity in sequence over long evolutionary distance



# Reticulogram = phylogeny with short-cut links



Legendre, P. and Makarenkov, V.  
Syst. Biol. 51:199-216 (2002)

- Adding a short-cut link between nodes  $x$  and  $y$  that share similar sequences
  - $d_{tree}(x, y)$  would better reflect  $d_{sequence}(x, y)$
- Keep adding short-cut links as long as the improvement is better than average
  - $\frac{\sum_{x \neq y} (d_{tree}(x, y) - d_{sequence}(x, y))^2}{\frac{n(n-1)}{2} - N}$ , where  $n$  is the number of taxa and  $N$  is the number of branches

# Any question?

---

- See you next week on September 6<sup>th</sup> 9-10:30am