3000788 Intro to Comp Molec Biol

Lecture 7: Phylogenetics and molecular evolution

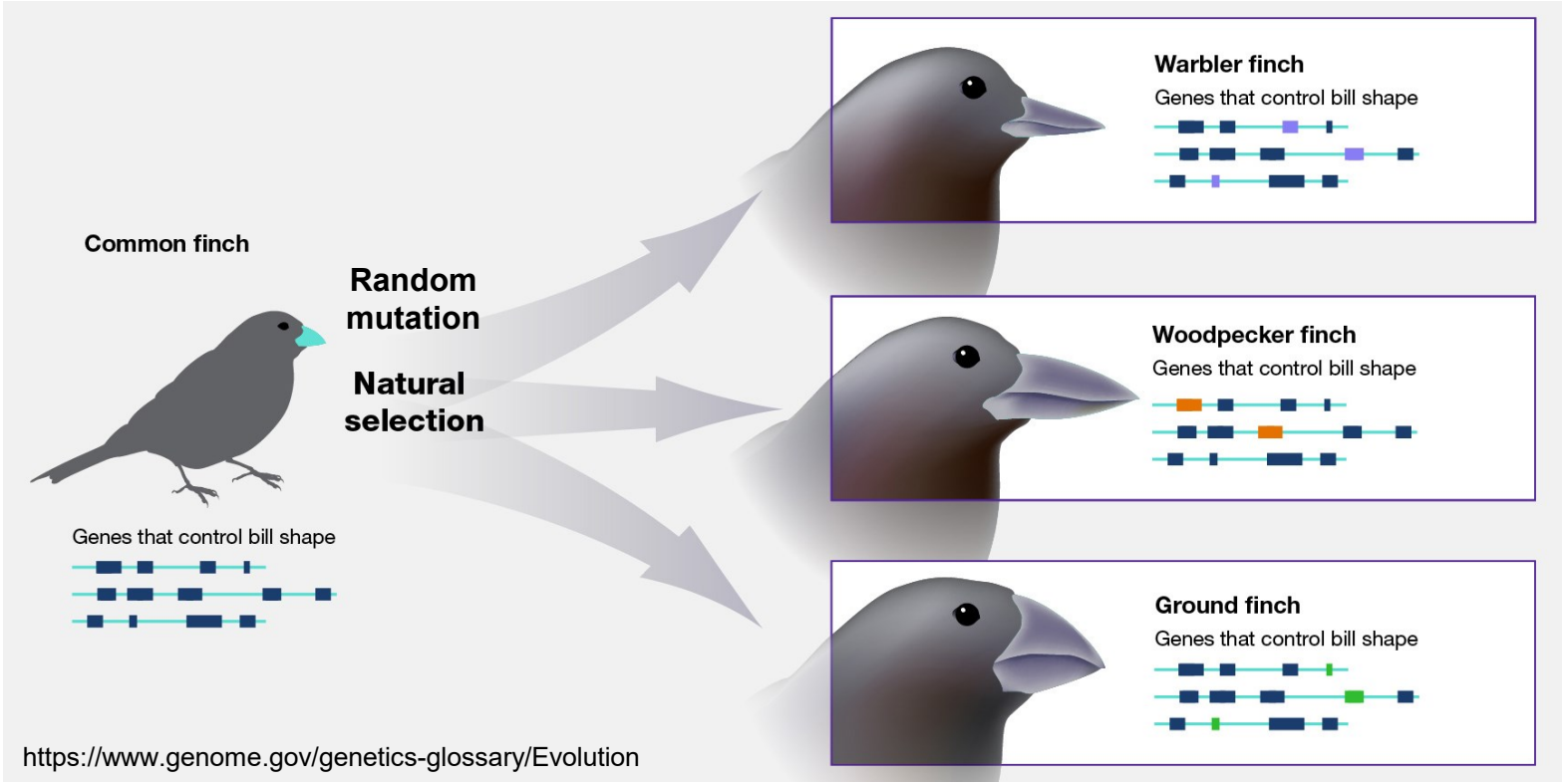
September 7, 2023

Sira Sriswasdi, PhD

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

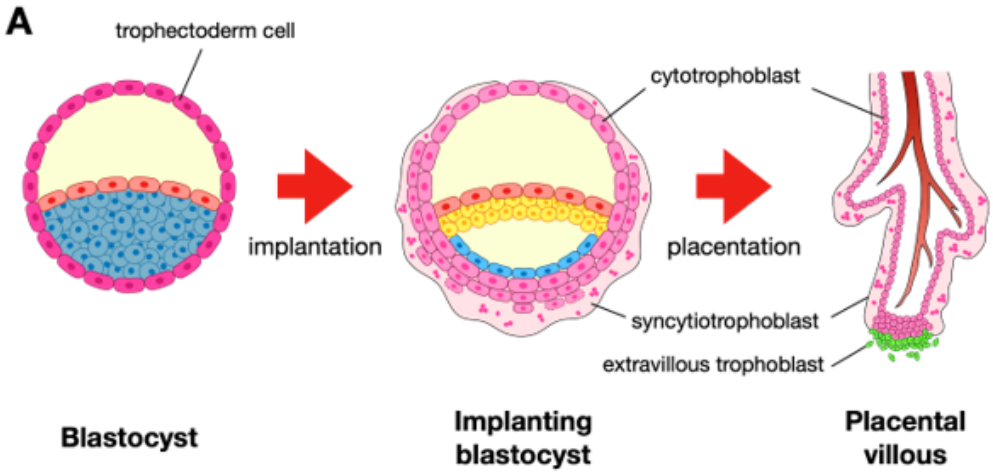
Evolution

Random mutation with natural selection



https://www.genome.gov/genetics-glossary/Evolution

Emergence of form and function



B

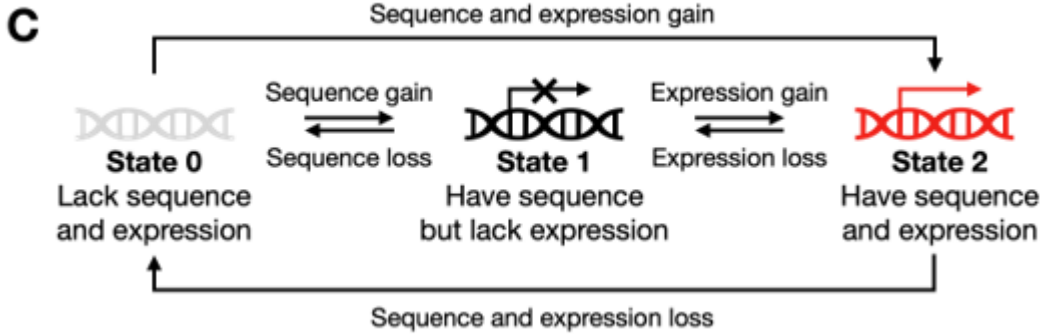
		Genome (spp.)	Transcriptome (spp.)
Lissamphibia		6	0
Sauria		106	1 (chorioallantoic membrane)
Prototheria		2	0
Metatheria		6	2 (placenta)
Eutheria		147	11 (placenta)
Total		267	14

Tetrapod
Amniota
Mammalia
Theria

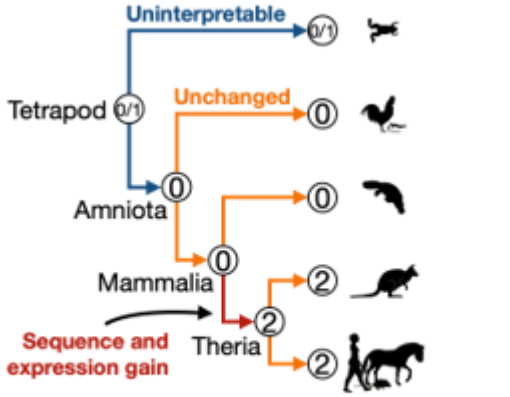
Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)

- Placenta emerged as a new organ in mammalian ancestor

The search for genes that give rise to placenta



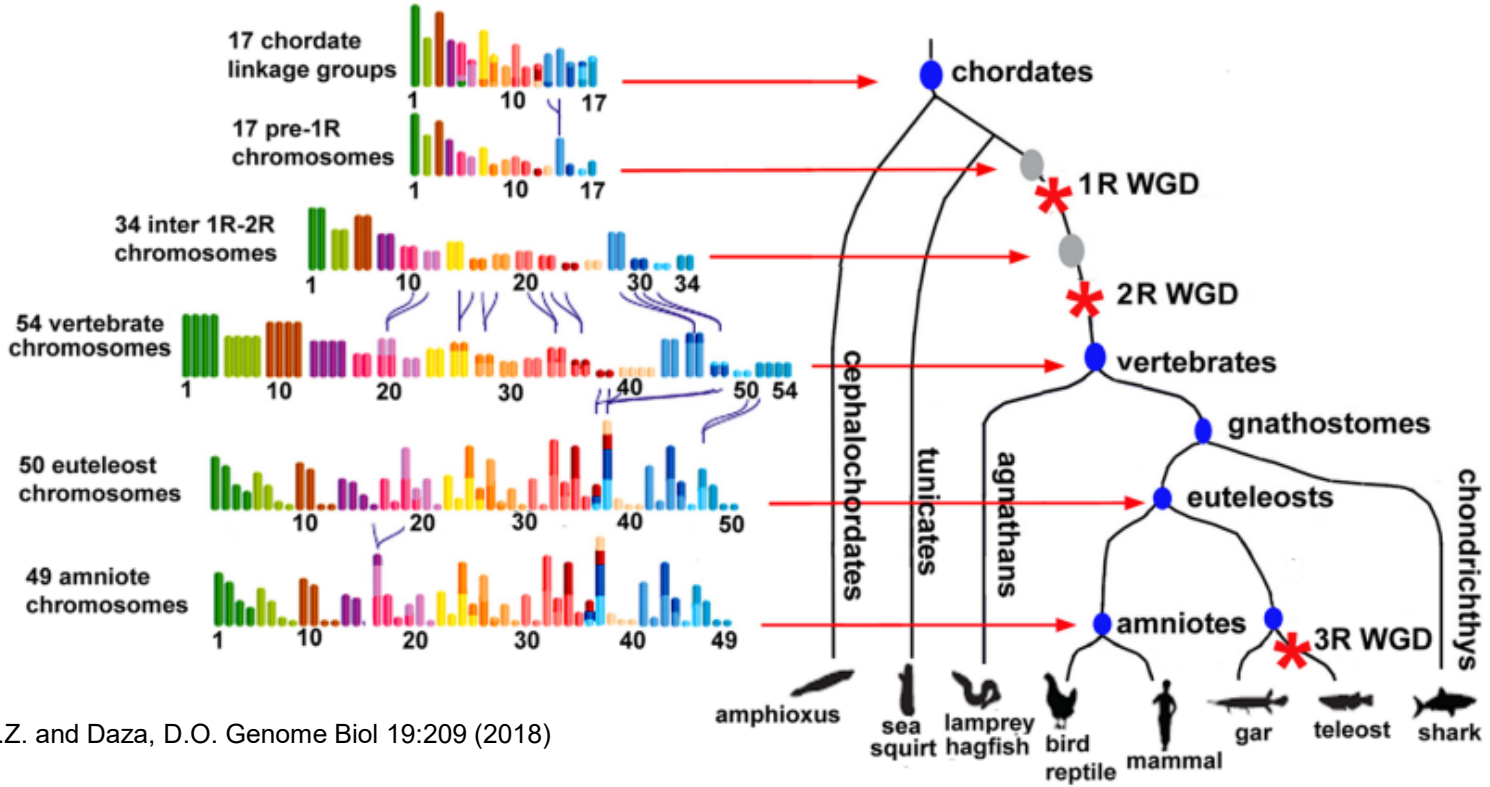
Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)



Interpreting character state changes from ancestors to descendants to transition events

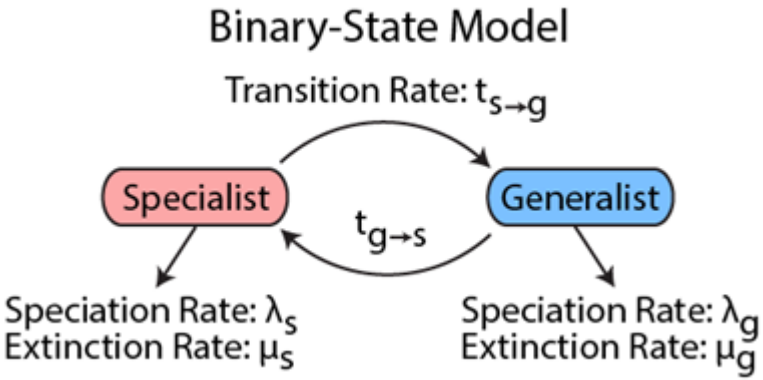
- Identify gene and expression gain events during mammalian evolution

History of whole-genome duplications in animal

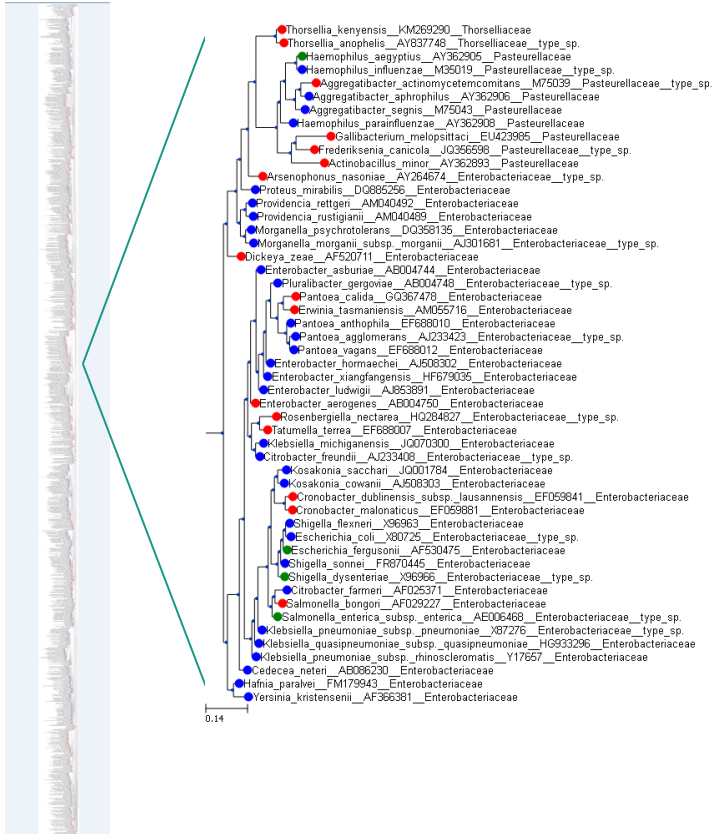


Holland, L.Z. and Daza, D.O. Genome Biol 19:209 (2018)

Life style evolution in bacteria



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

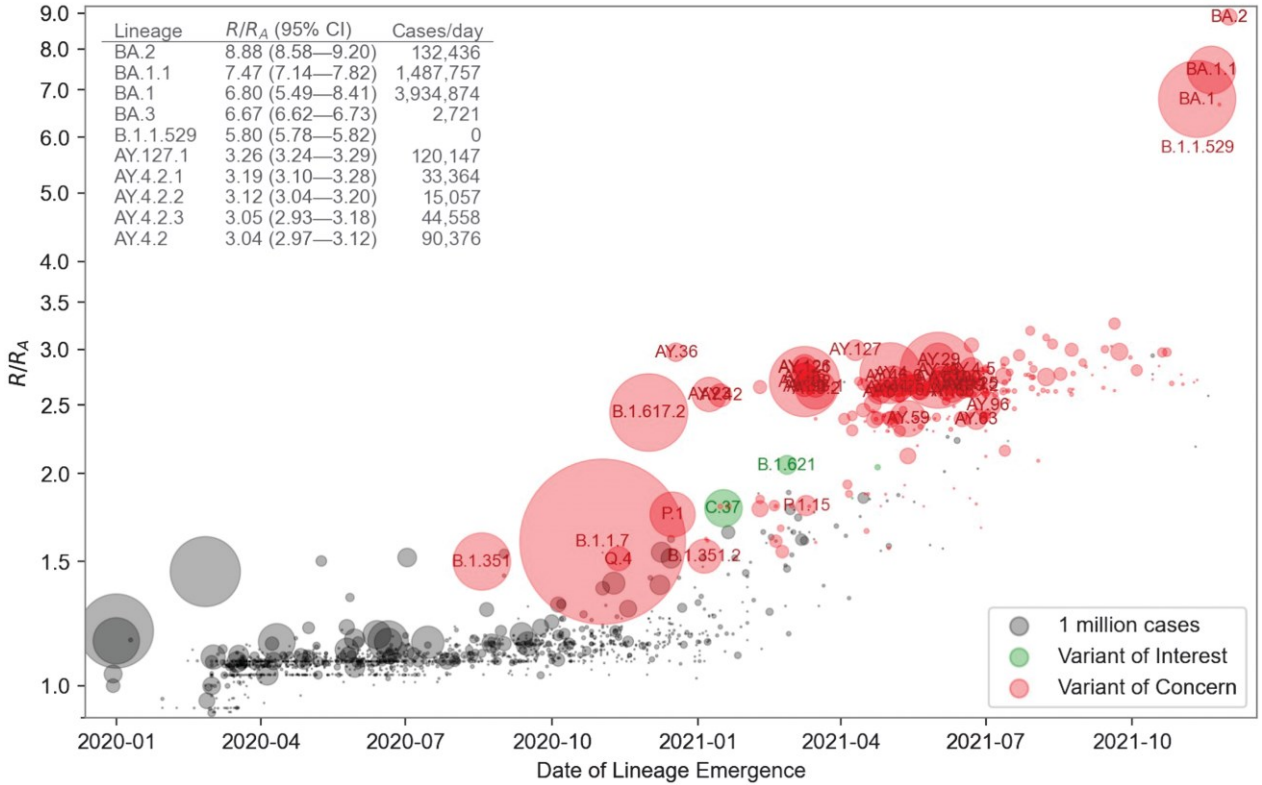


- Generalist has higher survivability
- Nature drives generalist to specialist

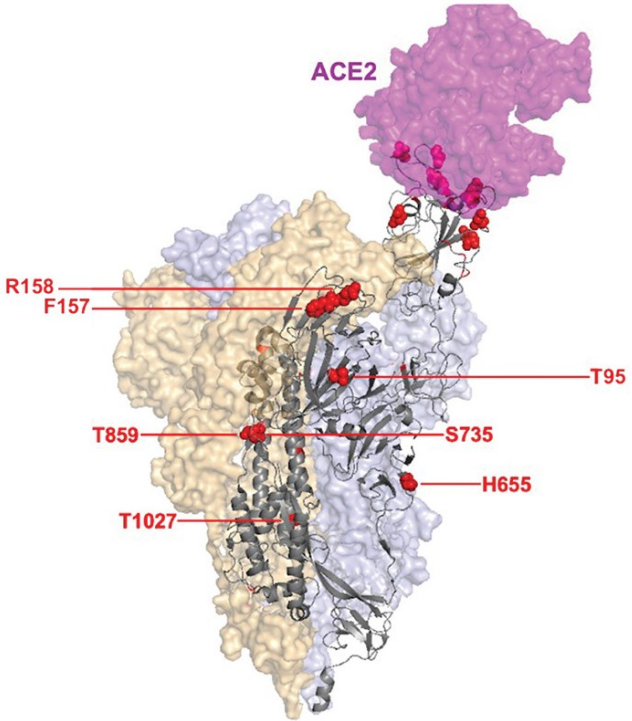
COVID-19 evolution

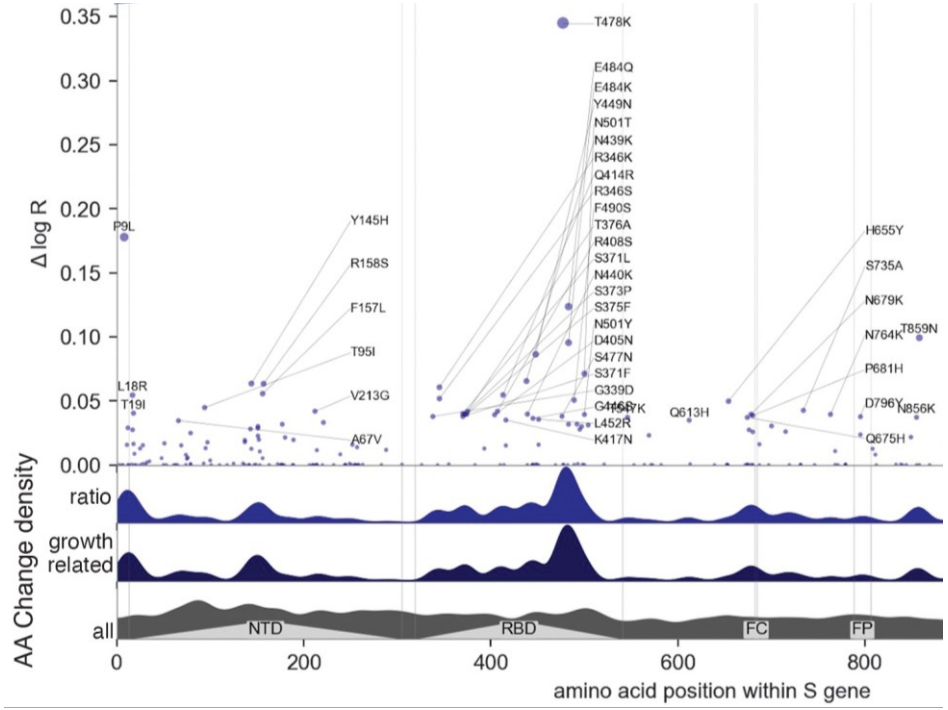
Obermeyer, F. et al. Science 376:1327-1332 (2022)

Fitness increase over the original Wuhan strain



Linking evolution to function

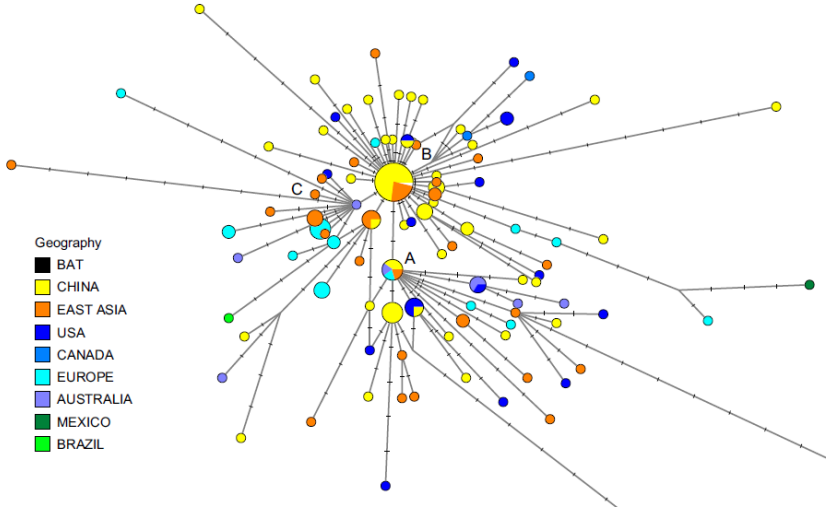




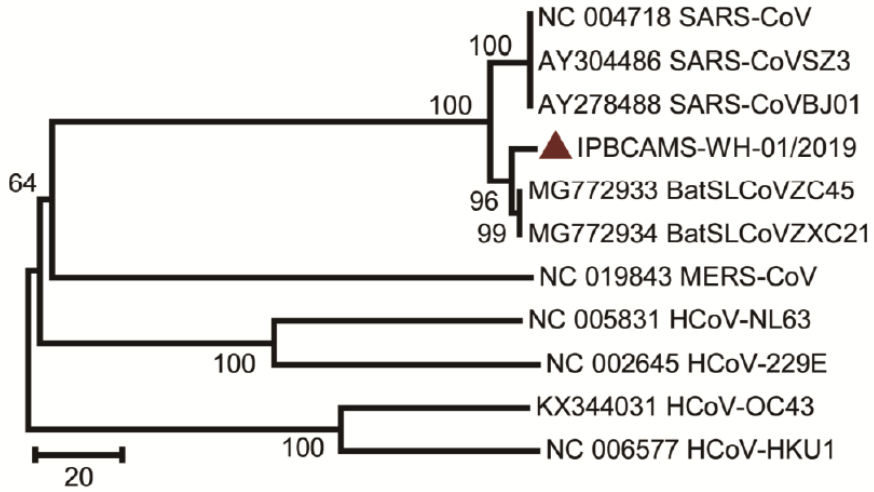
Obermeyer, F. et al. Science 376:1327-1332 (2022)

Phylogenetics

Phylogeny



Forster *et al*. PNAS (2020)



Guo *et al*. Clin Infect Dis (2020)

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

Types of phylogeny

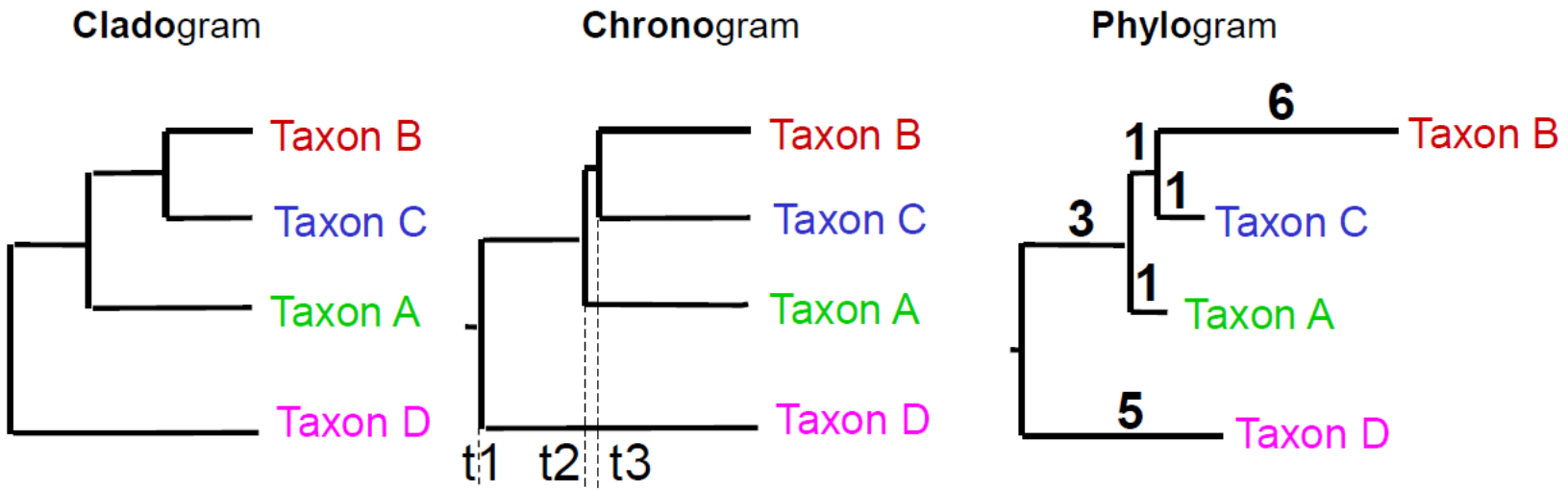


Image from Lecture 18 6.047 MIT OCW

- The choice depends on research question

Ingredients for phylogenetic reconstruction

Multiple sequence alignment

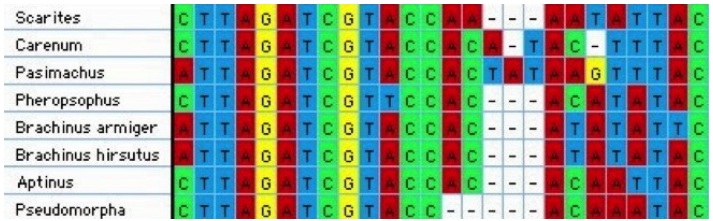
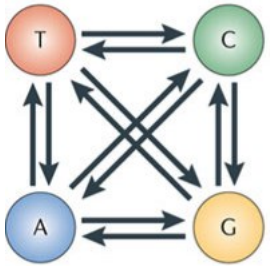


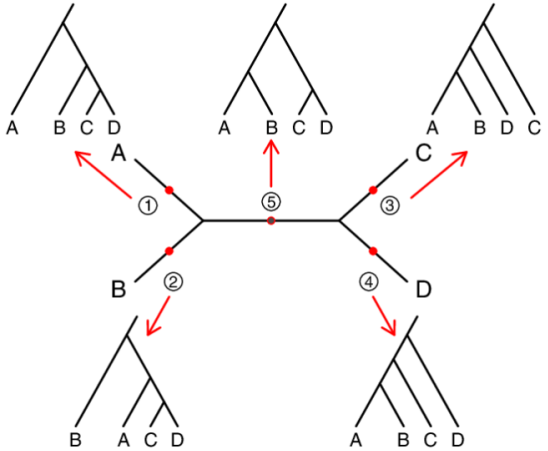
Image from www.mcqbiology.com

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

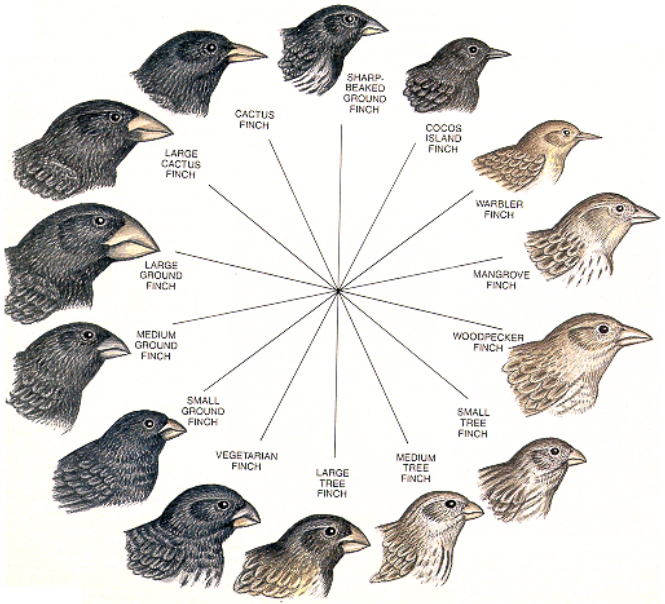
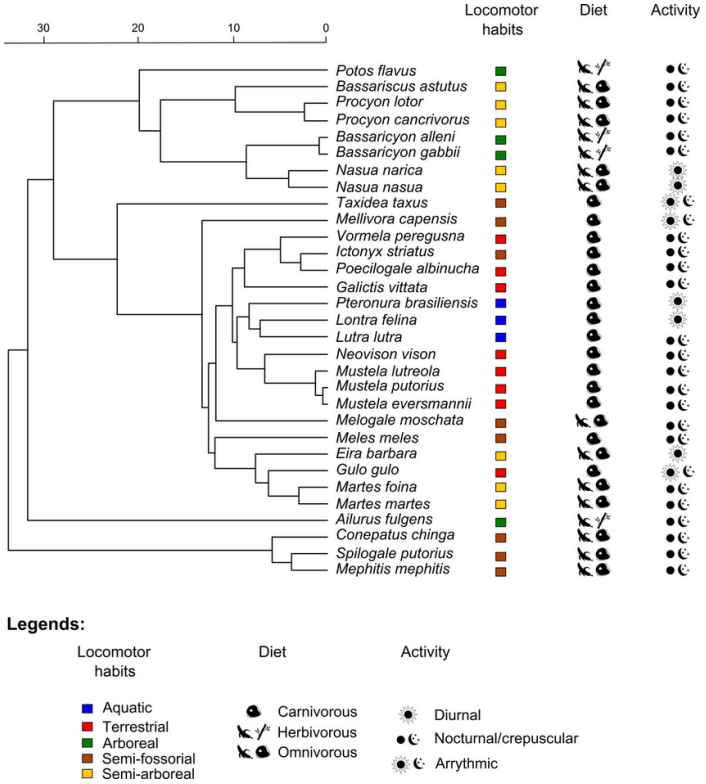


Image from pinterest

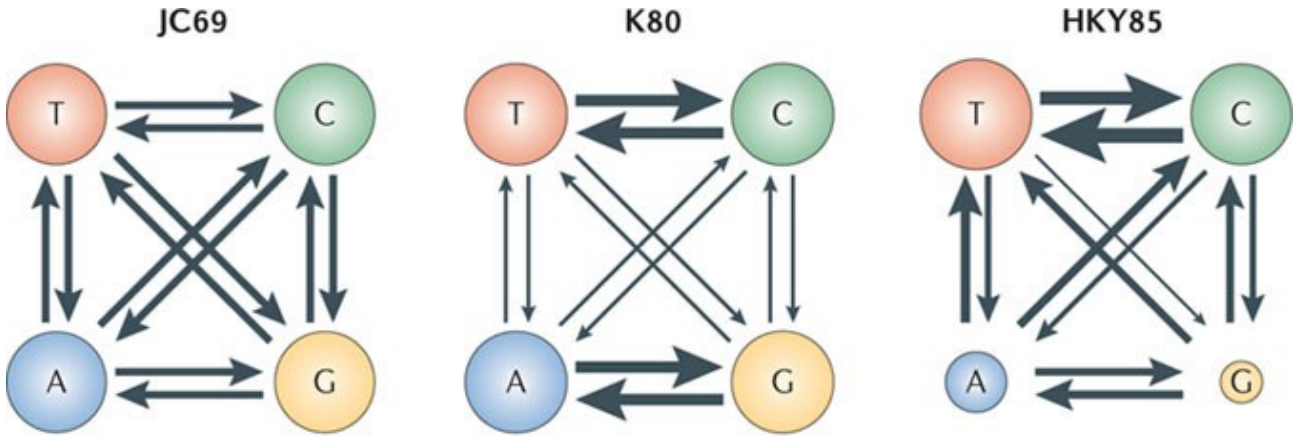


Dumont, M. et al. Biol J Linnean Soc (2015)

- Morphology, diet, or life-style

Evolutionary (substitution) model

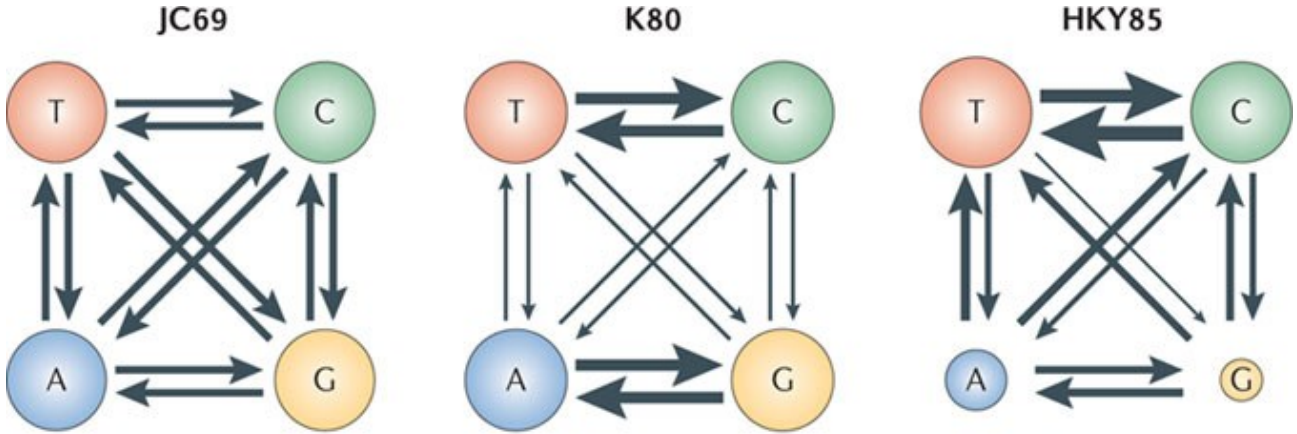
Basic nucleotide substitution models



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)

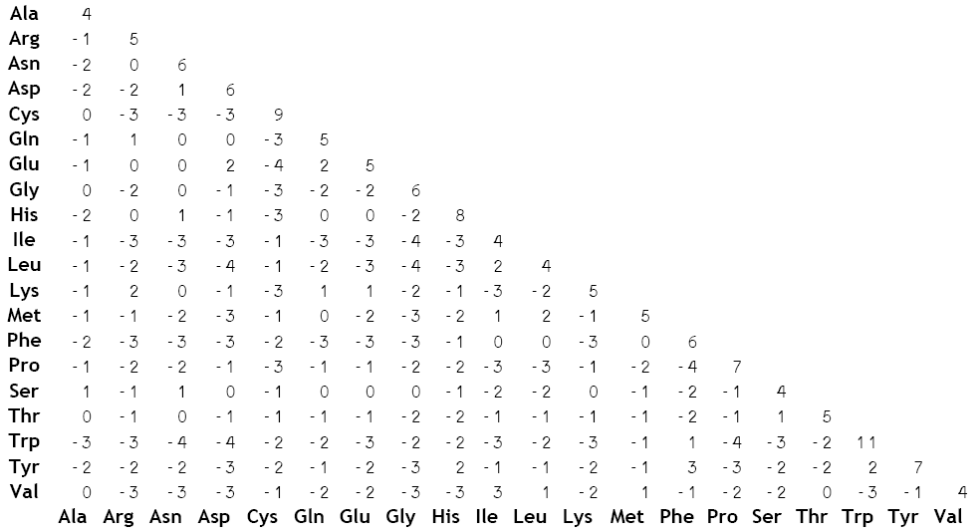
- Symmetric substitution rates: P(A→G) = P(G→A)
  - 6 parameters (4 nucleotides choose 2)
  - Time-reversible: switching ancestral and descendant taxa does not change the calculation
  - Useful in practice because we often don't know which taxon came first

- Different base frequencies
  - 3 parameters

- This is the most generalized time-reversible model possible
- Often used

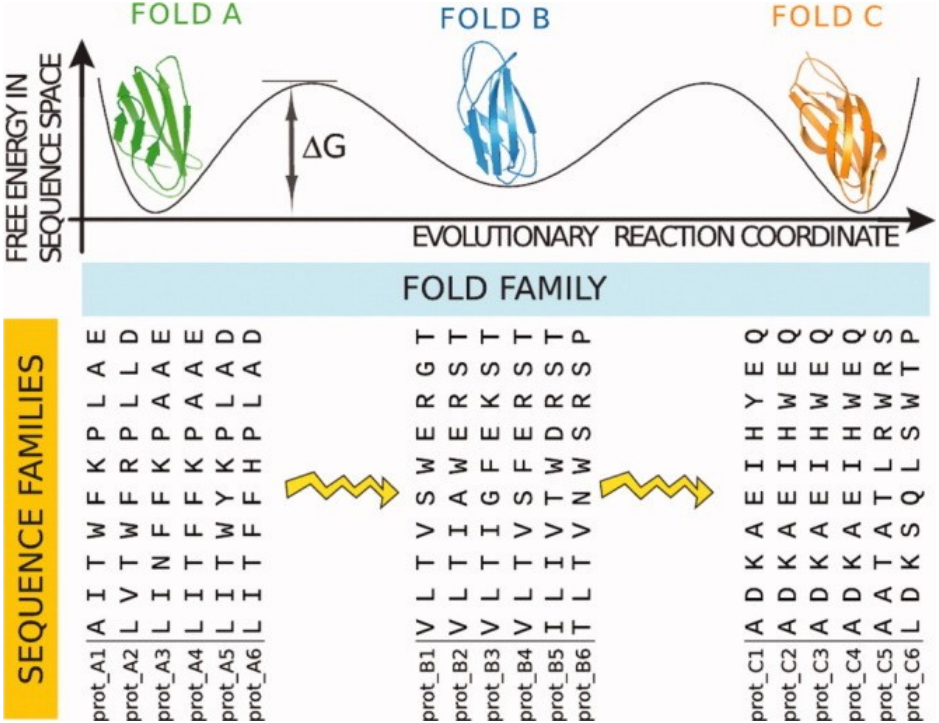
Basic amino acid substitution models

- 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		
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

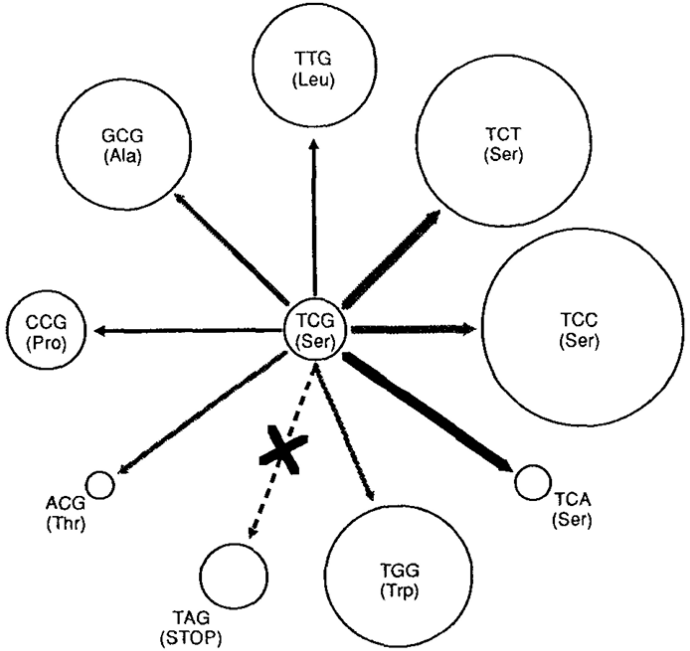


- Different positions on different protein families have different evolutionary constraints
- Use energy from interactions between residues in 3D to develop substitution matrix
- Must be tuned for each use

Liberles, D.A. Protein Sci. 21: 769-785 (2012)

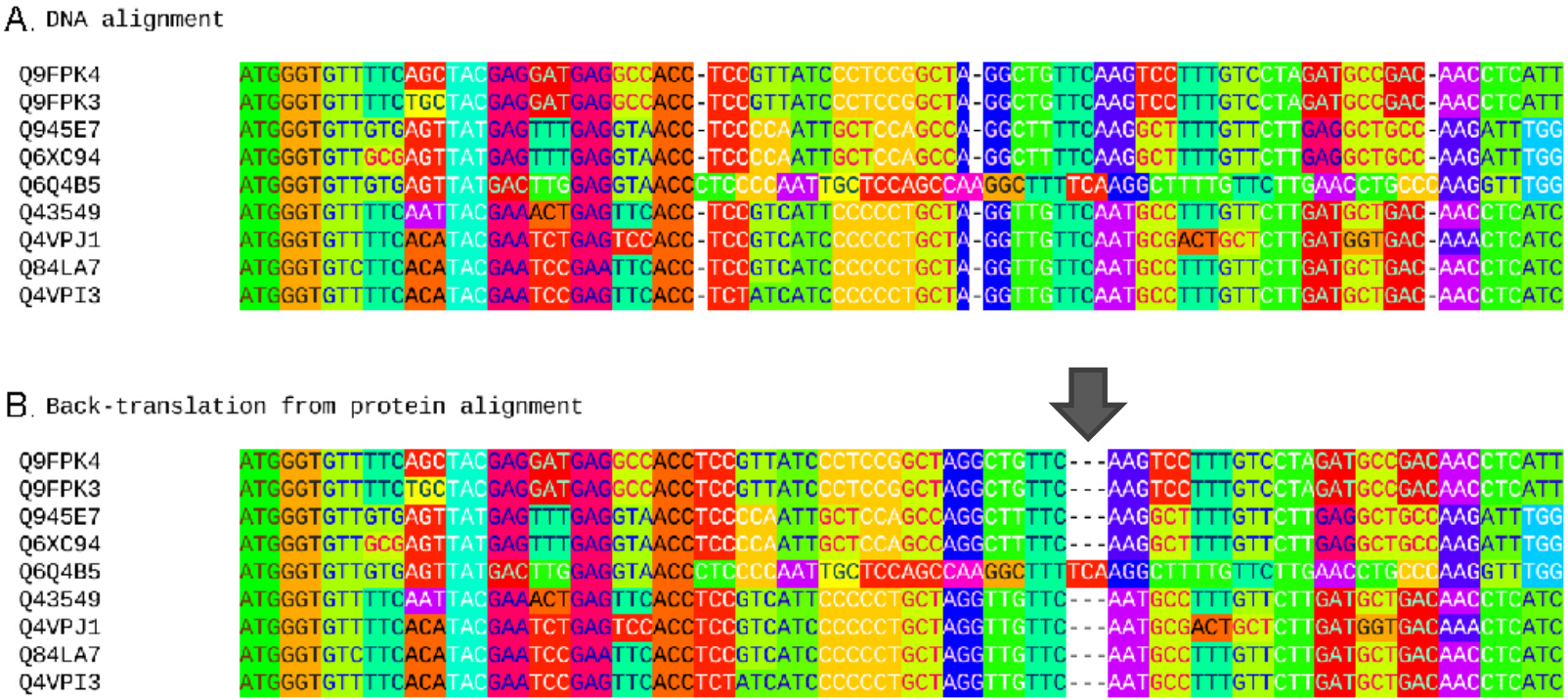
Codon substitution model

- 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



Goldman, N. and Yang, Z. Mol Biol Evol 11:725-736 (1994)

Codon alignment: amino acid → nucleotide



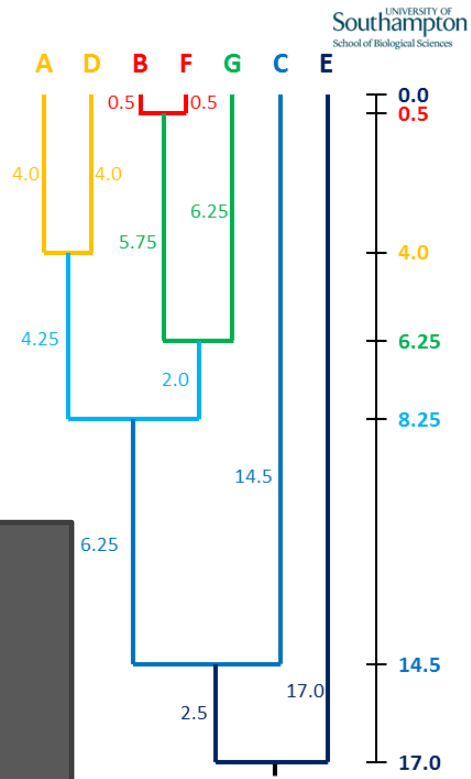
Scherer, N. and Basso, D.M. Genet. Mol. Res. 7:853-860 (2008)

Tree building algorithms

UPGMA: Unweighted pair-group method arithmatic 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	



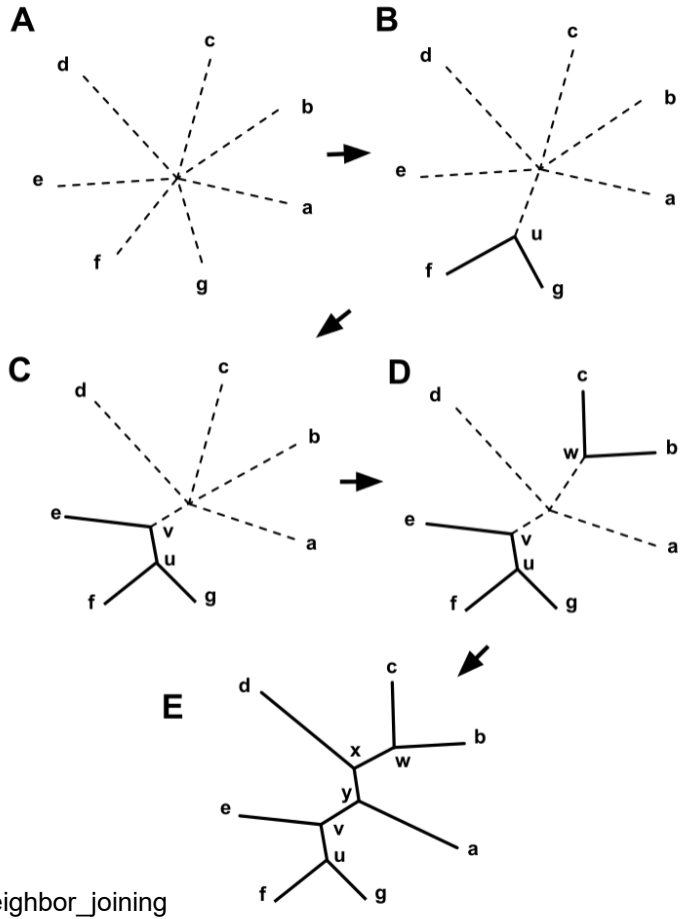
http://www.slimsuite.unsw.edu.au/teaching/upgma/

Neighbor joining

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

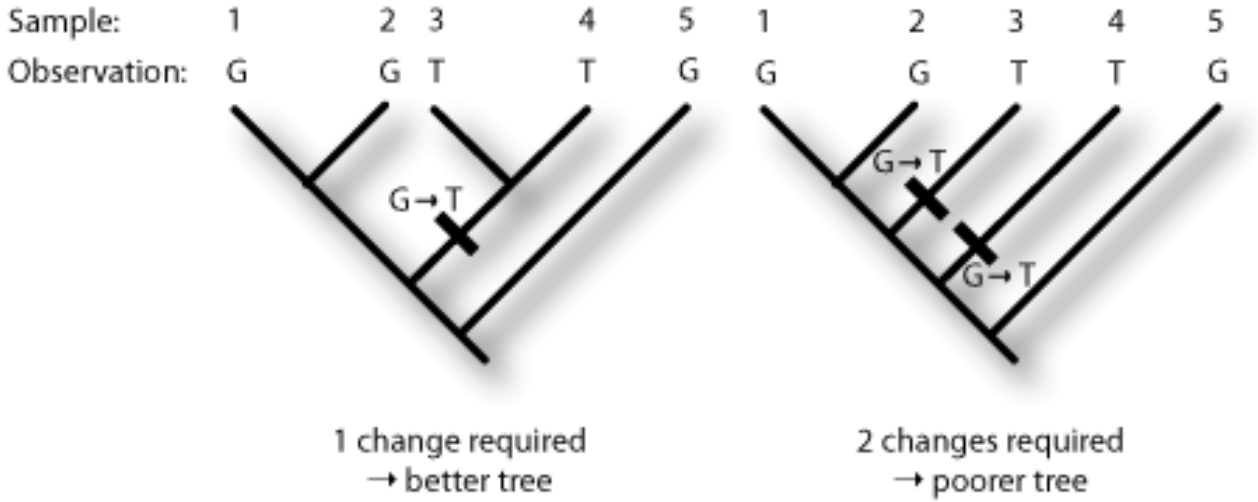
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} [d(i,k) + d(j,k)]$$

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



https://en.wikipedia.org/wiki/Neighbor_joining

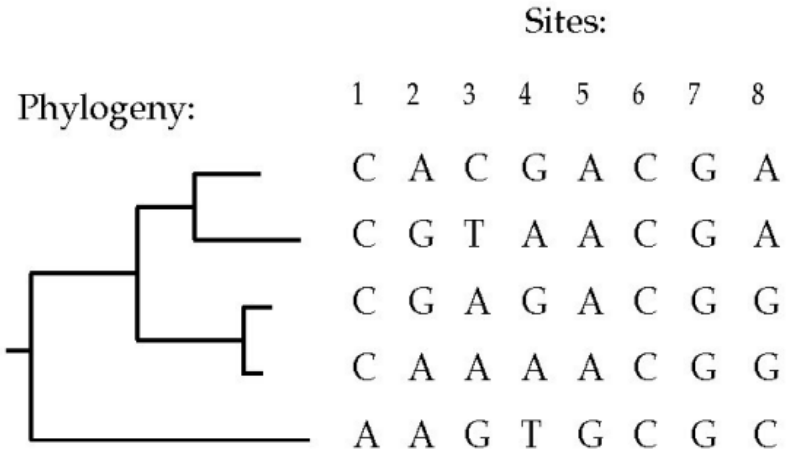
Maximum parsimony / Minimum evolution



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

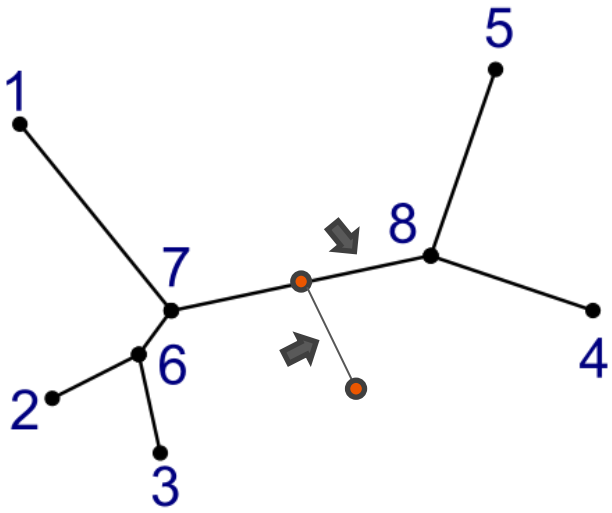
Maximum likelihood approach



https://homes.cs.washington.edu/~ruzzo/courses/gs559/09wi/lectures/8A_likelihood.pdf

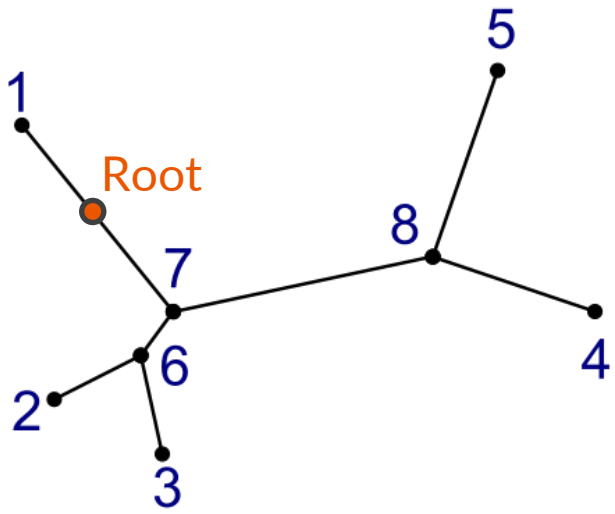
- Likelihood = P(sequence data | substitution model, tree topology, branch lengths)
- Given a phylogenetic tree topology, we can alter branch lengths (one branch at a time) to search for the best answer
- The problem is how to find the best tree topology

Number of possible trees with distinct leaves



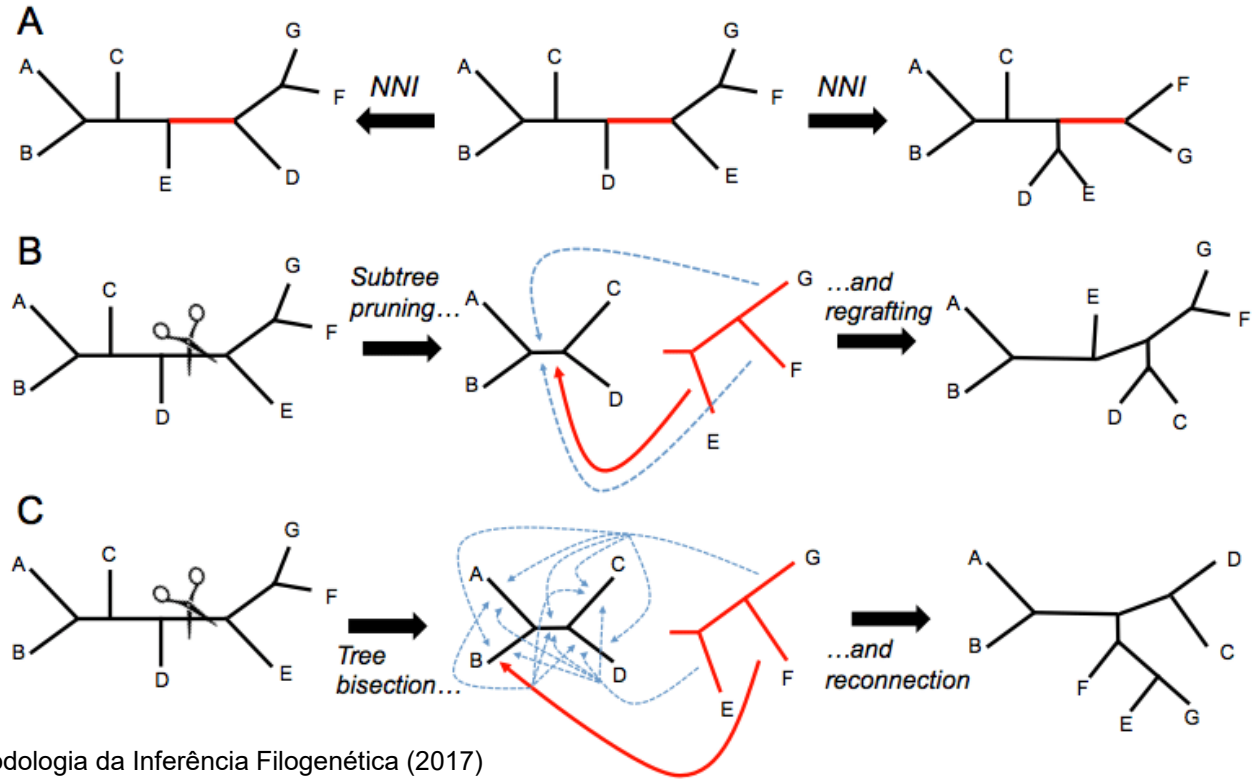
- Unrooted tree with n leaves has $2n-2$ nodes
  - Adding a leaf will also create one internal node
- There are $2n-3$ branches
  - Adding a leaf will create two branches
  - One new branch and split an existing branch
- From each unrooted tree with n leaves, there are $2n-3$ locations to attach a new leaf
- Number of unrooted trees with n leaves, for n>2, is $(2n-5)x(2n-7)x(2n-9)x \ldots x1$

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 $(2n-3)x(2n-5)x(2n-7)x(2n-9)x \ldots x1$
- Number of rooted trees with 10 leaves = 17x15x13x11x9x7x5x3 = 34,459,425

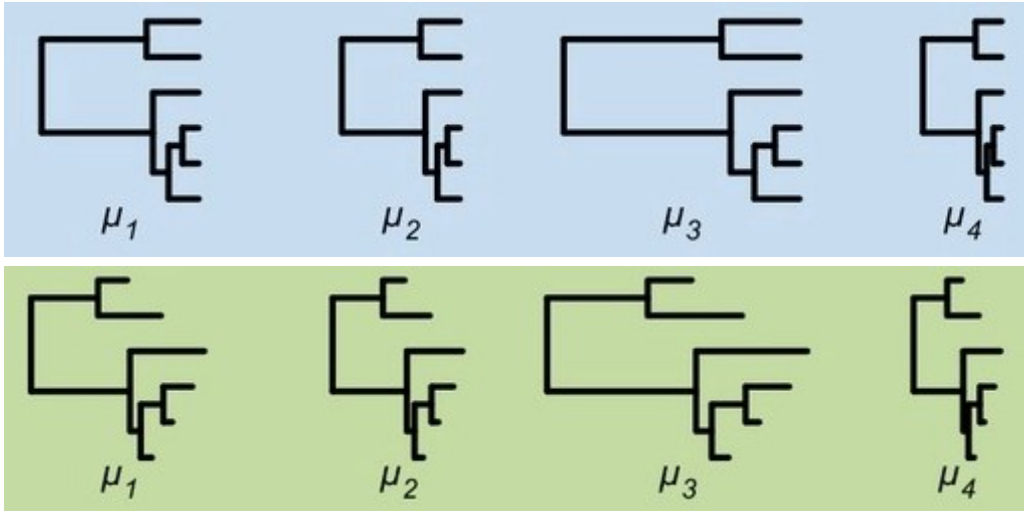
Heuristic tree search algorithms



Gainett *et al*. Metodologia da Inferência Filogenética (2017)

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

Invariant sites

Gamma-distributed mutation rates

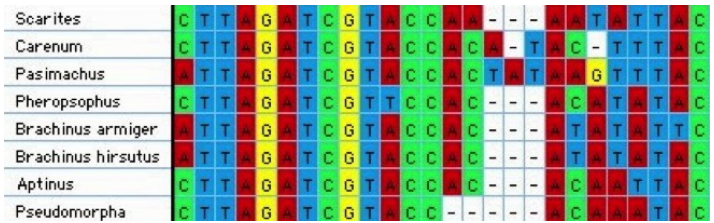
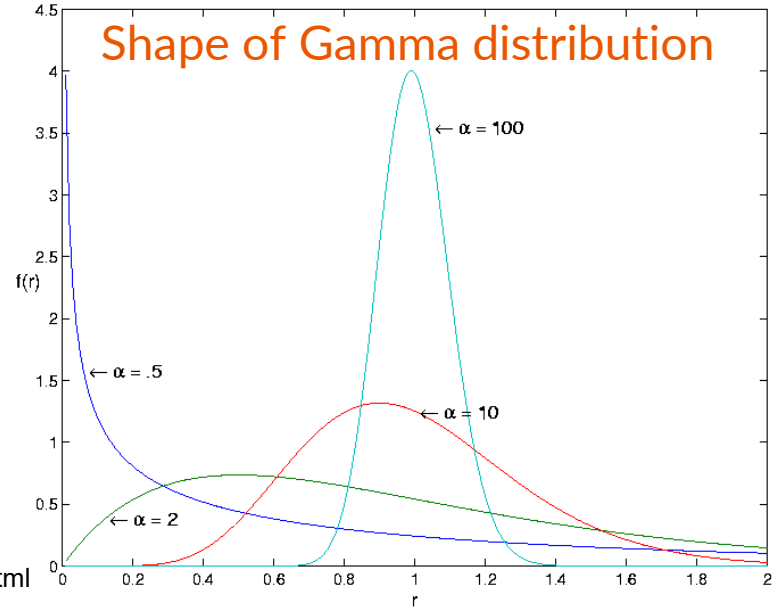


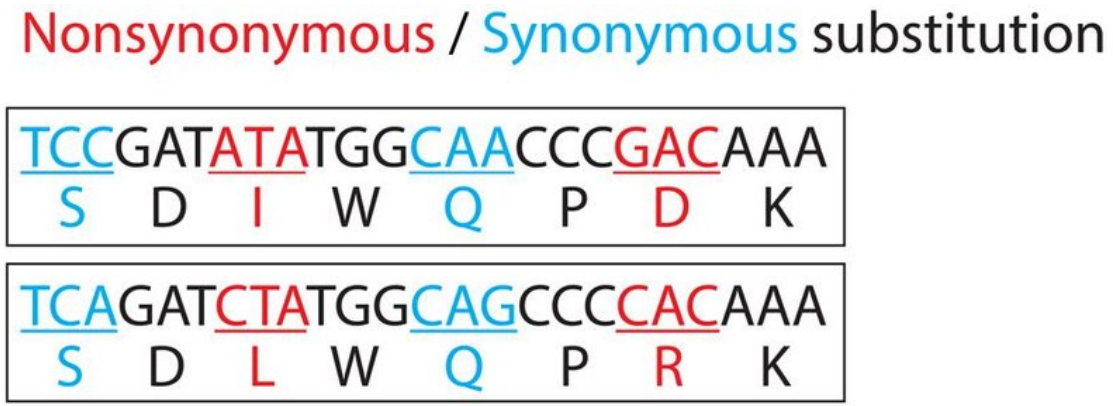
Image from www.mcqbiology.com



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

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

Natural selection on codon model



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

Model	-lnL
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

models	diff. DF = q	X^2	P	
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

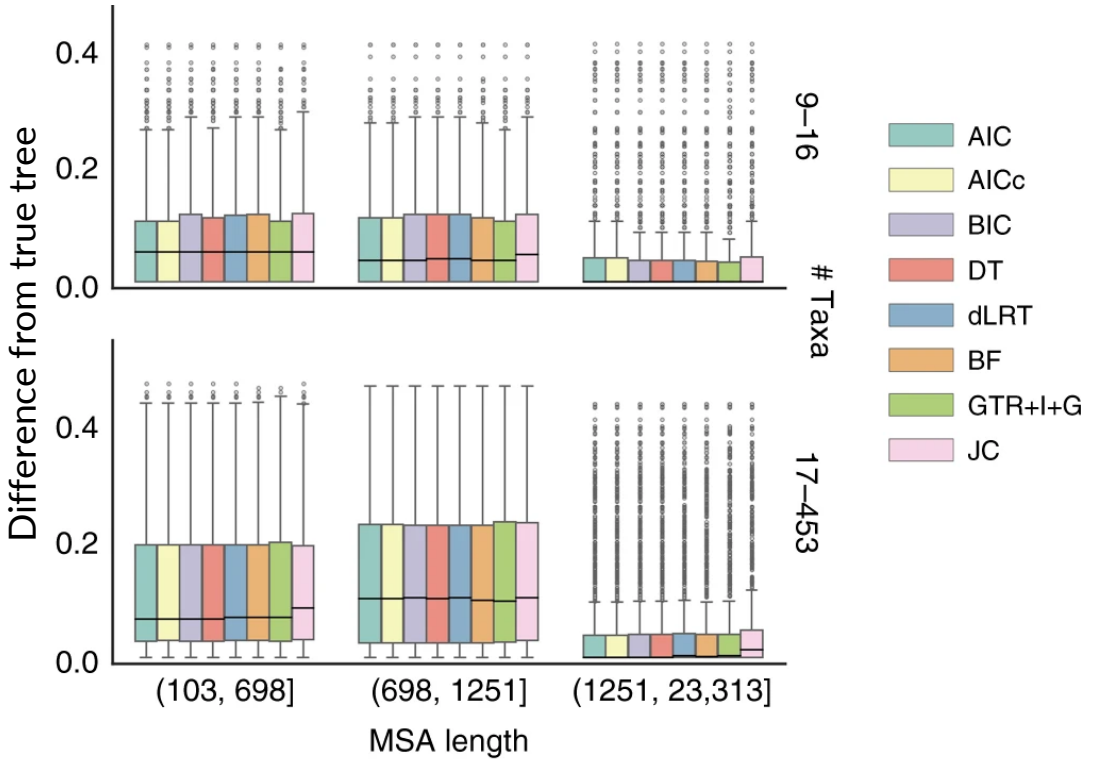
Model	-lnL
HKY85	3233.34395
HKY85 +G	3145.29031
HKY85 +I+G	3142.36439

models	diff. DF = q	X^2	P	
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

https://www.ccg.unam.mx/~vinuesa/Model_fitting_in_phylogenetics.html

Minimal impact from model selection

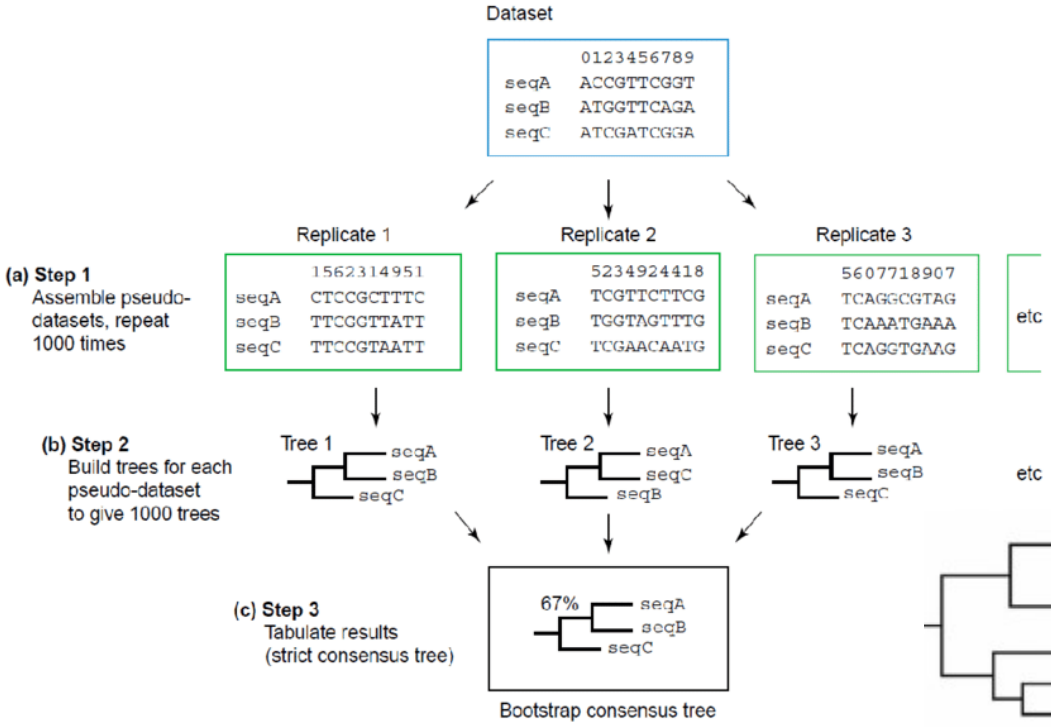
- Evaluate tree topology and ancestral sequence reconstruction
- GTR+I+G (the most complex model) yields similar results as JC (the simplest model)
- Consistent results when using various criteria to select the best nucleotide model (AIC, BIC, etc.)



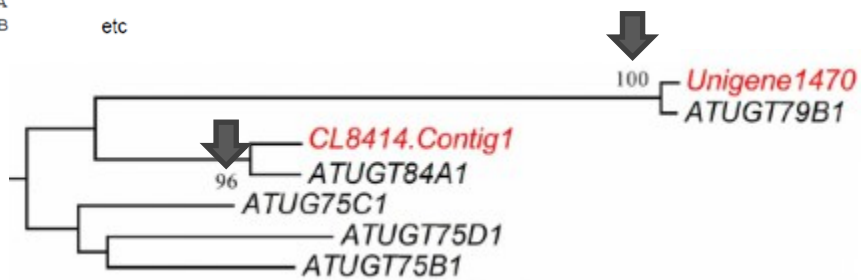
Abadi, S. Nat. Comm. 10:934 (2019)

Reliability test with bootstrapping

Bootstrap support for taxa group



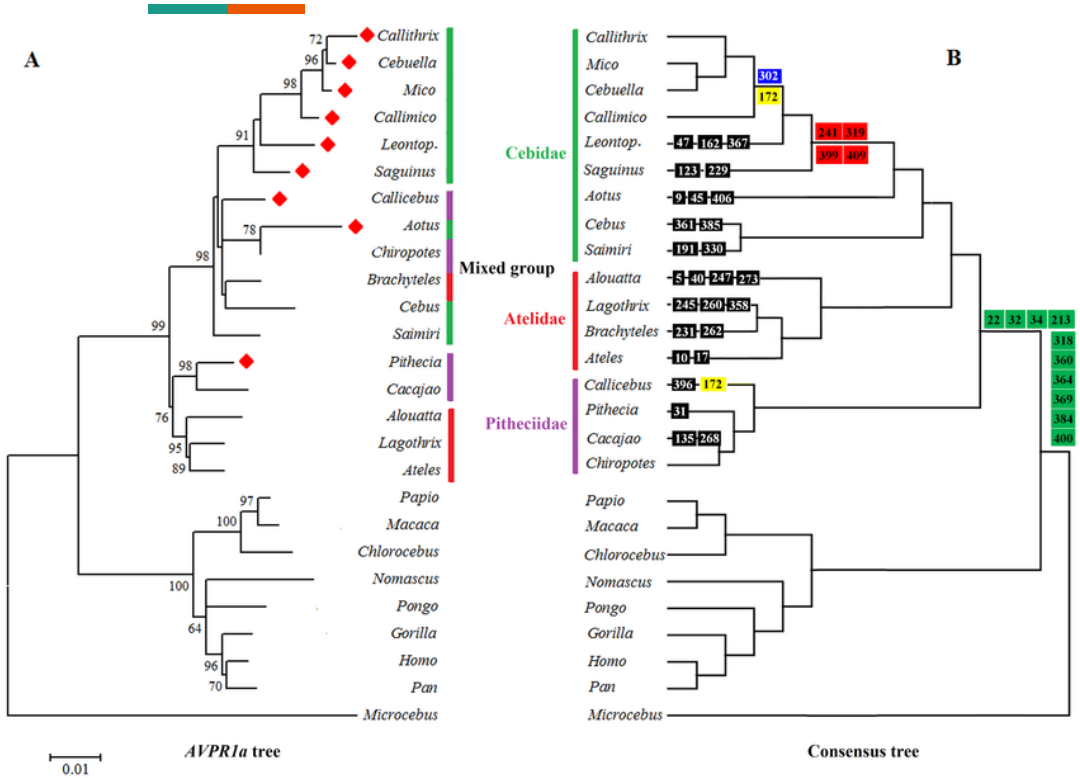
- Bootstrapping = sampling with repeats
- Allow multiple trees to be constructed from a single sequence alignment



Wyant *et al*. Archives of Virology (2011)

Wei et al. PLoS ONE 2013

Gene tree vs concatenated tree

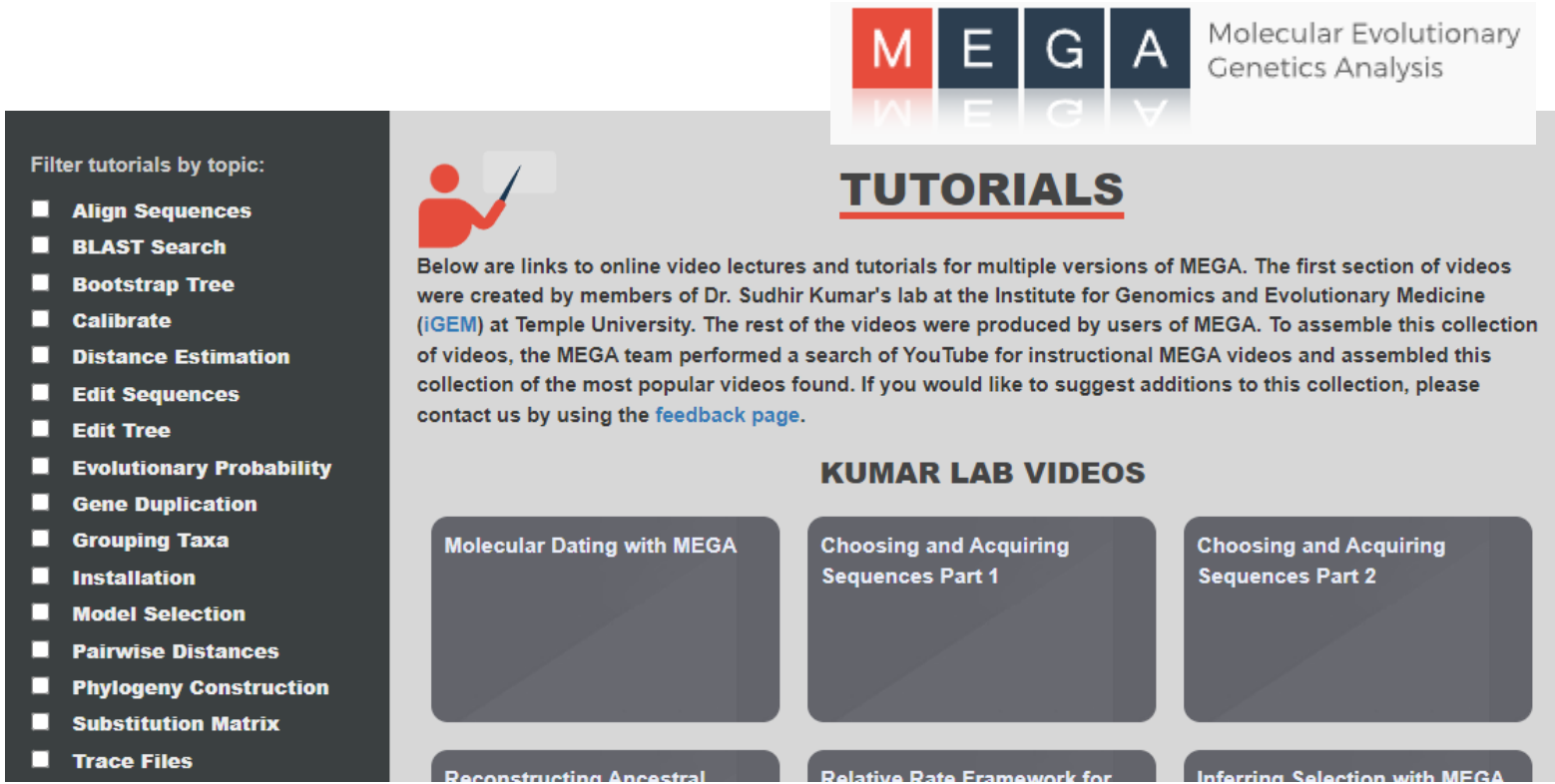


Ren, D. et al. PLoS ONE 9:e222638 (2014)

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

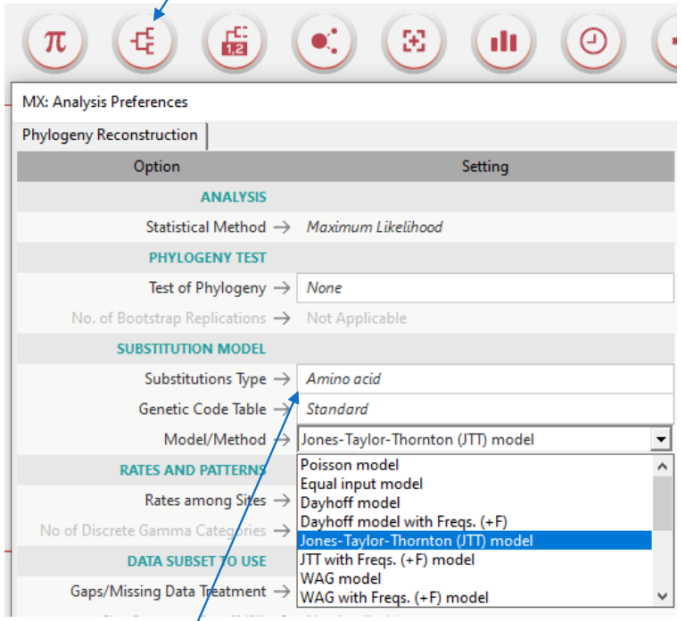
Example of how to setup phylogenetic reconstruction

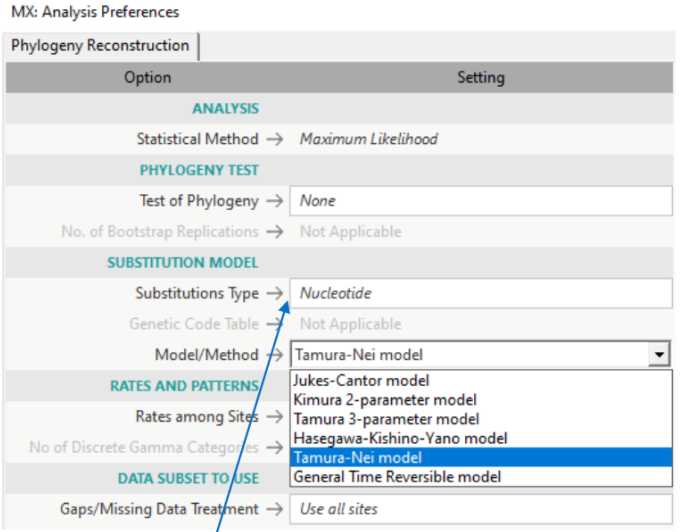
MEGA: A Windows GUI tool for phylogenetic analysis



Substitution model choices

PHYLOGENY mode

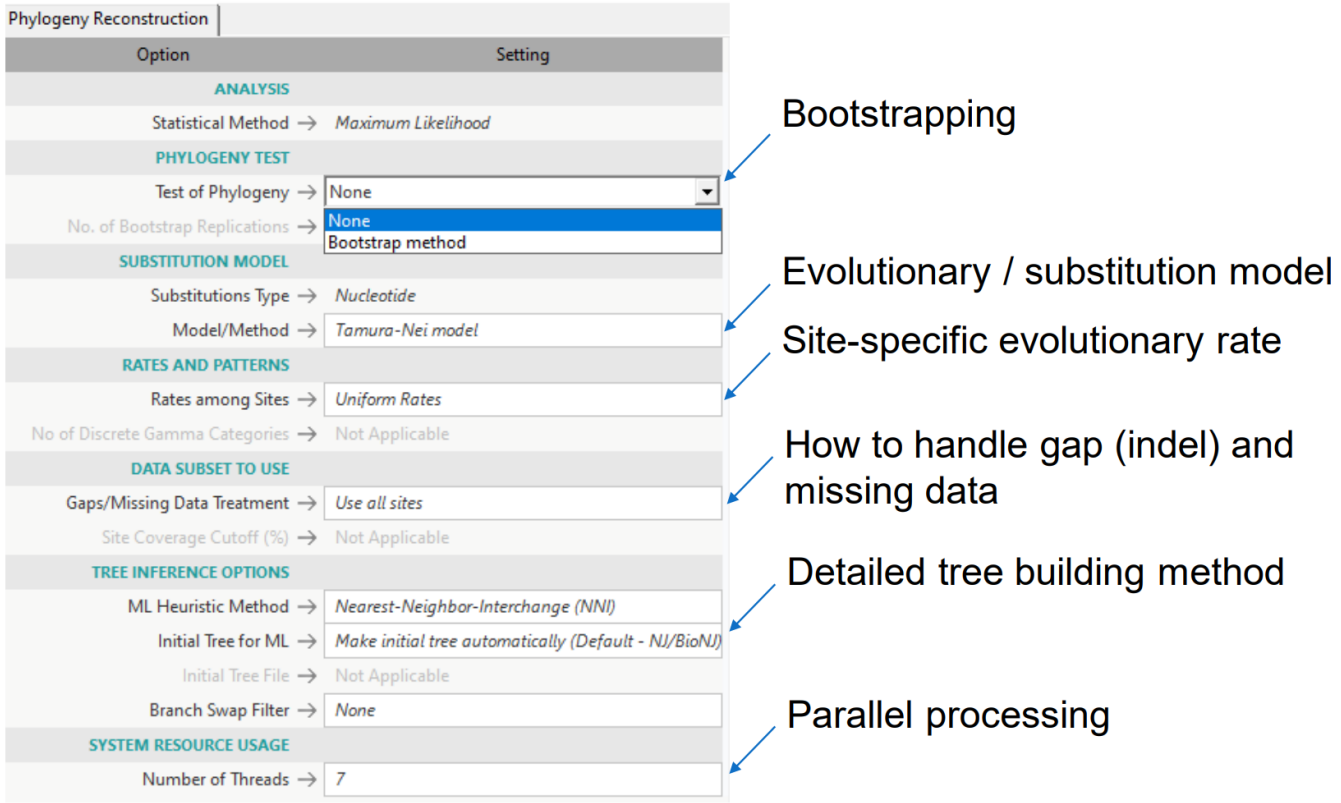




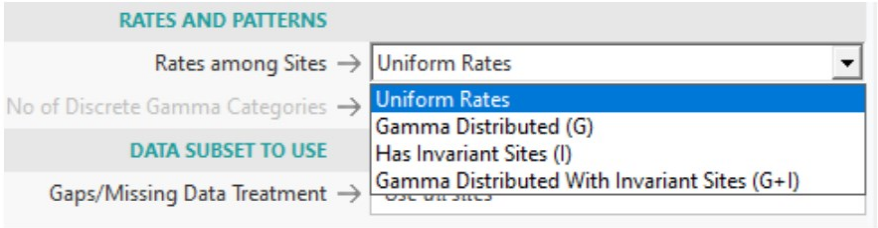
Amino acid-based models

Nucleotide-based models

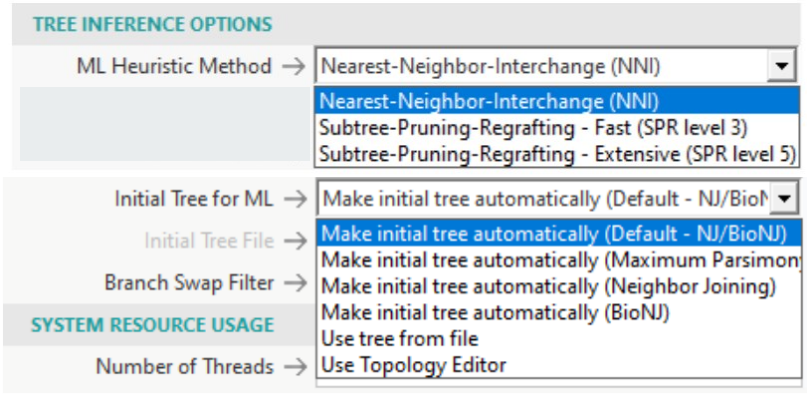
Maximum likelihood parameters



Maximum likelihood parameters



Invariant site and Gamma-distributed site-specific substitution rates



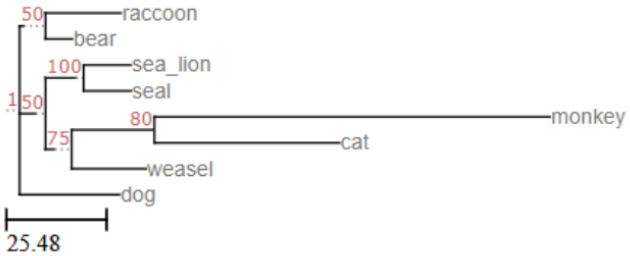
Tree search method

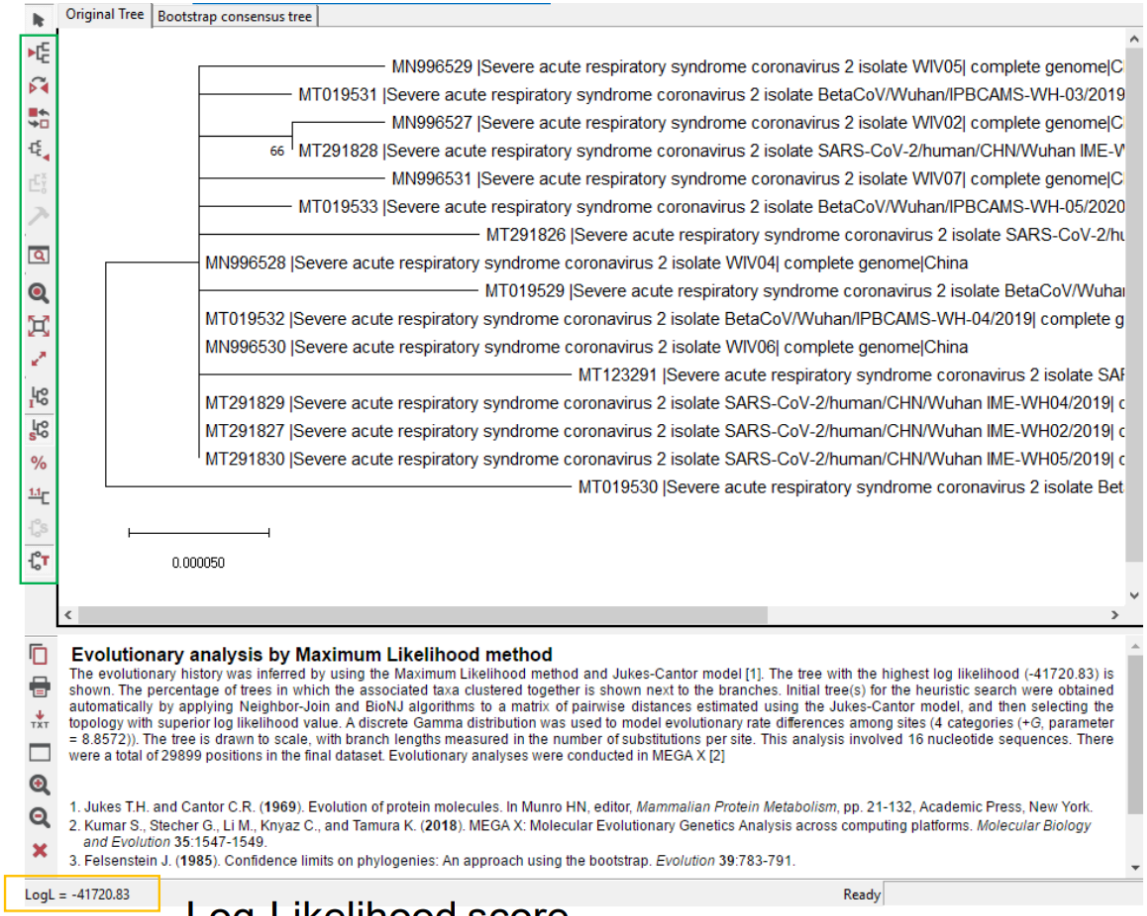
Initial tree can be built using quick and simple method like neighbor joining

Output

NEWICK format

((raccoon:19.19959,bear:6.80041)50:0.84600,((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);





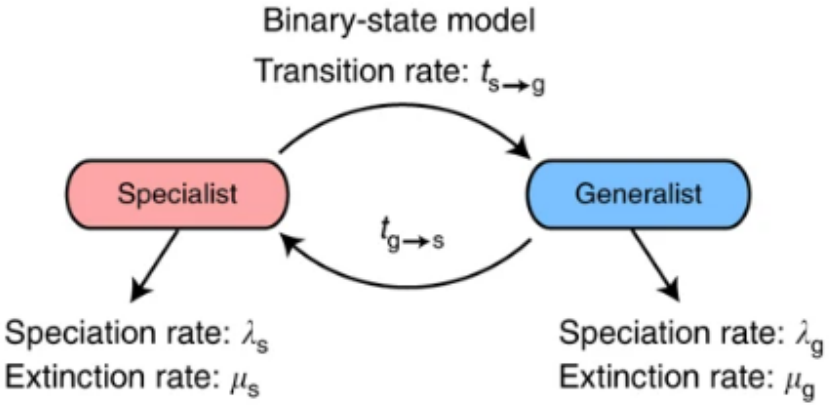
Log-Likelihood score

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(tree topology, branch lengths | substitution model, sequence data)
- More: https://evolution.genetics.washington.edu/phylip/software.html#methods

Character state evolution

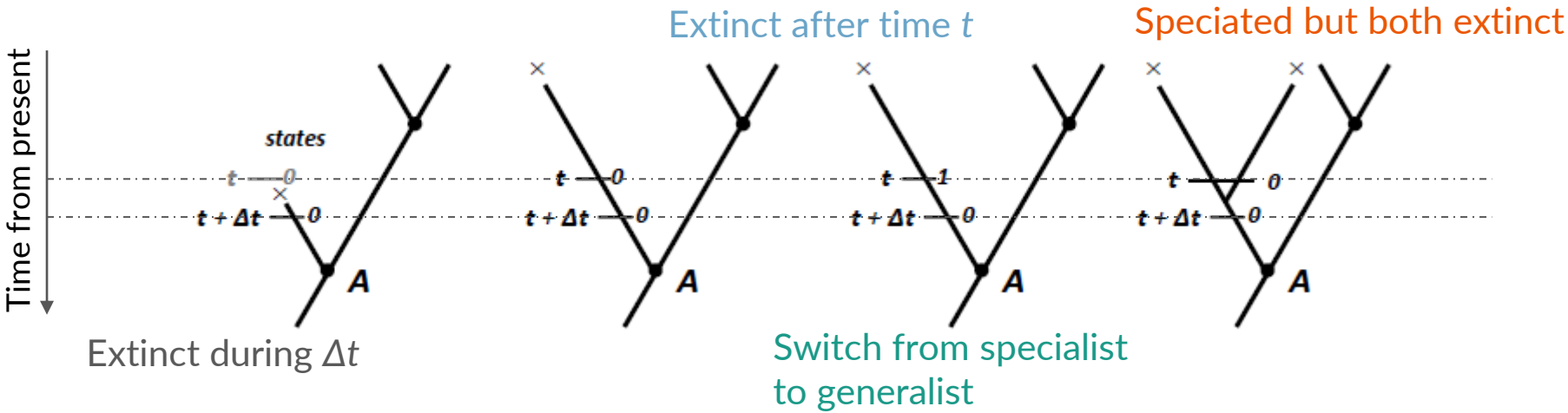
Impact of character state on evolutionary process



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

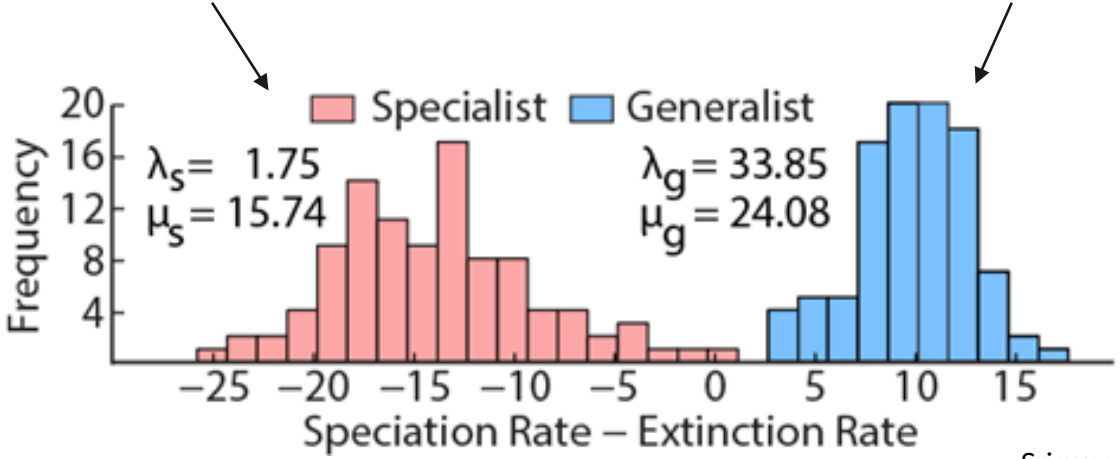


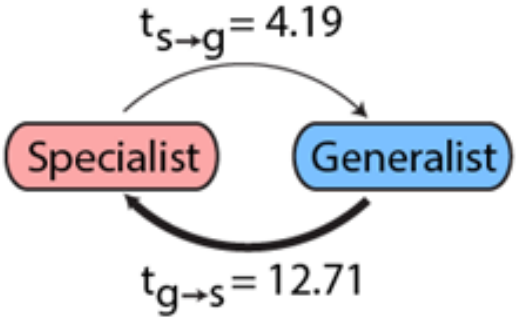
$$Extinct_s(t + \Delta t) = \mu_s \Delta t$$
$$+ (1 - \mu_s \Delta t)(1 - \tau_{s \to g} \Delta t)(1 - \lambda_s \Delta t) Extinct_s(t)$$
$$+ (1 - \mu_s \Delta t)(\tau_{s \to g} \Delta t)(1 - \lambda_s \Delta t) Extinct_g(t)$$
$$+ (1 - \mu_s \Delta t)(1 - \tau_{s \to g} \Delta t)(\lambda_s \Delta t) Extinct_s(t)^2$$

Estimating state-specific parameters

Net reduction in # of species

Net increase in # of species



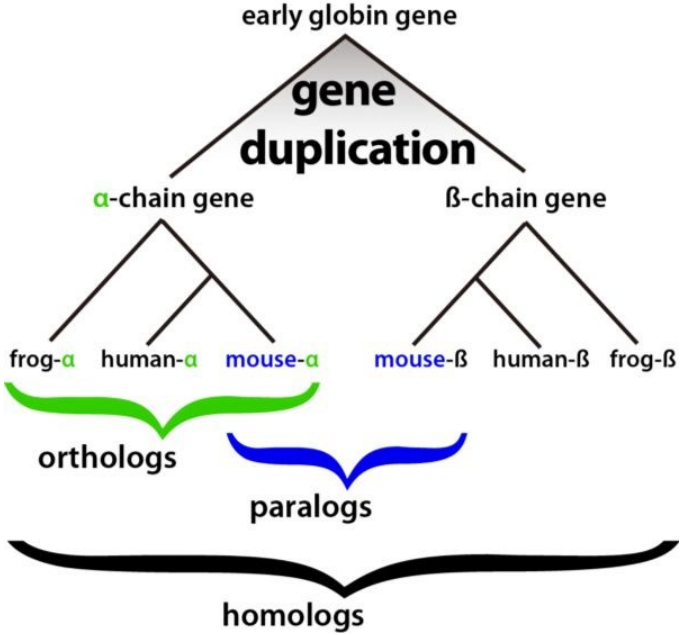


Sriswasdi *et al.*, Nature Comm 2017

Ortholog and paralog mapping

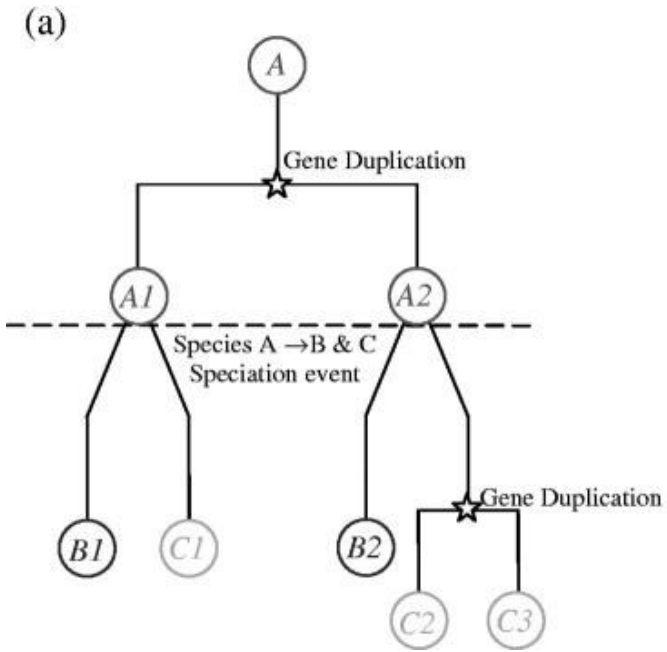
Homolog

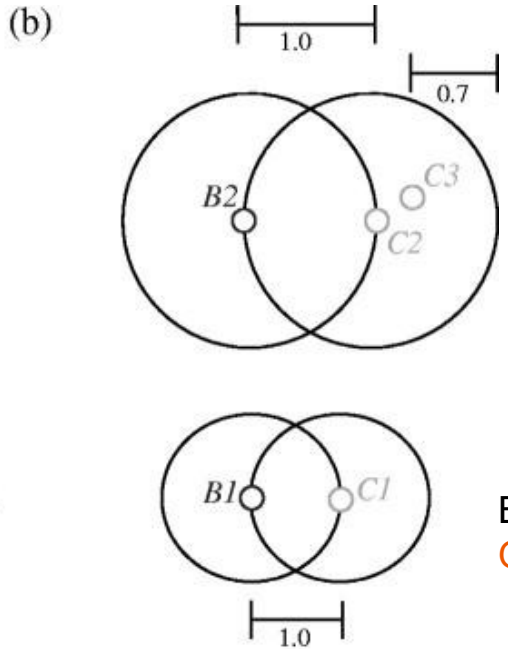
- Ortholog = genes across different species that descended from a common ancestor
- Paralog = genes within the same species that descended from duplicates of a common ancestor
- For phylogenetic reconstruction, we usually want to include all homologs together
  - Study gene family expansion



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

All-vs-all BLAST to find ortholog and paralog





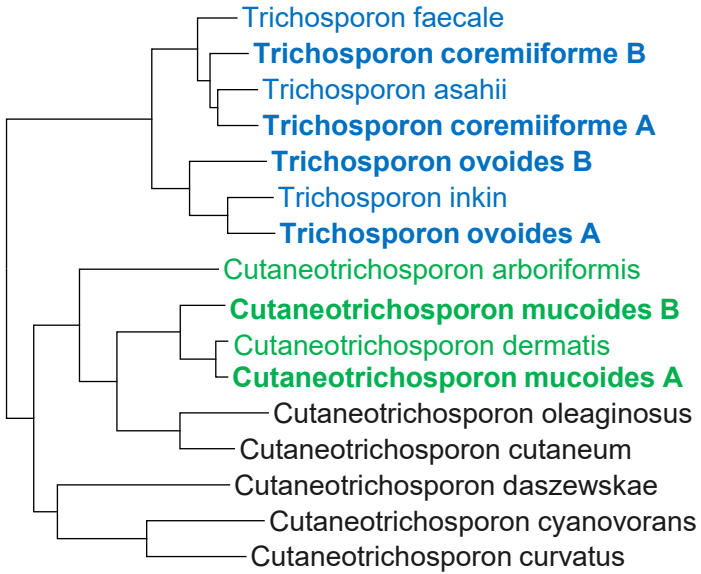
B2-C2 are reciprocal best hits
Ortholog

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

B1-C1 are reciprocal best hits
Ortholog

O'Brien, K.P. *et al.*, Nucleic Acids Res. 1:D476-80 (2005)

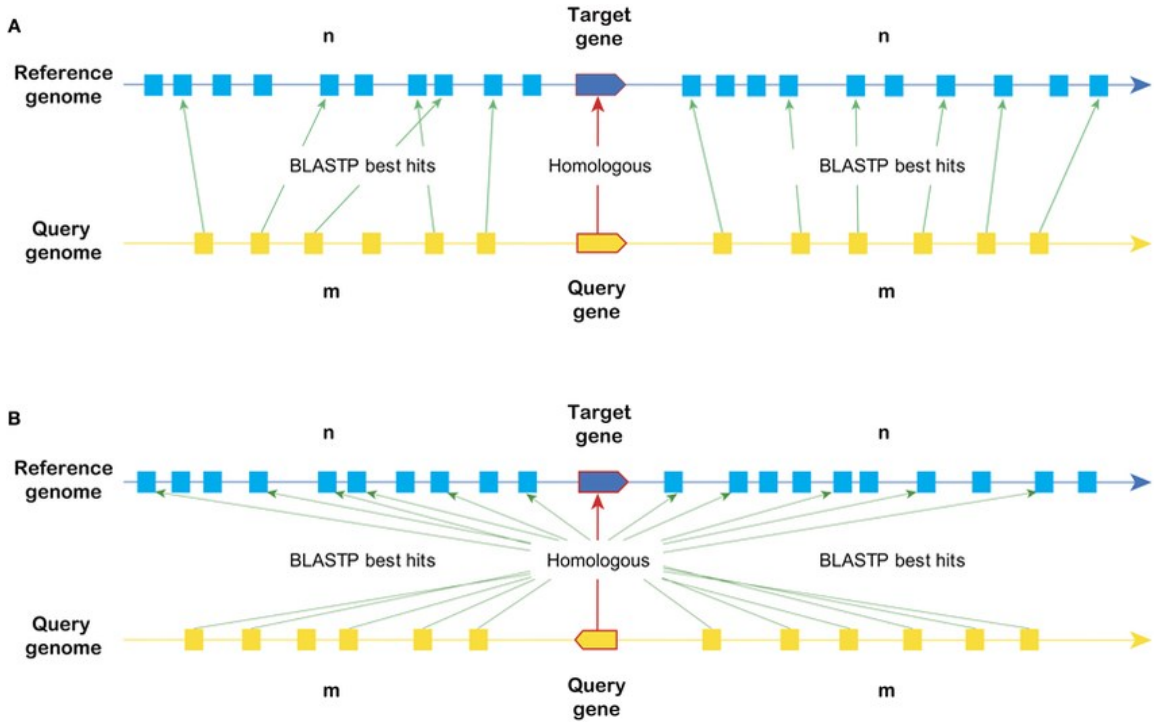
Phylogenetic reconstruction for hybrid genomes



Takashima, M. et al. Yeast 2017

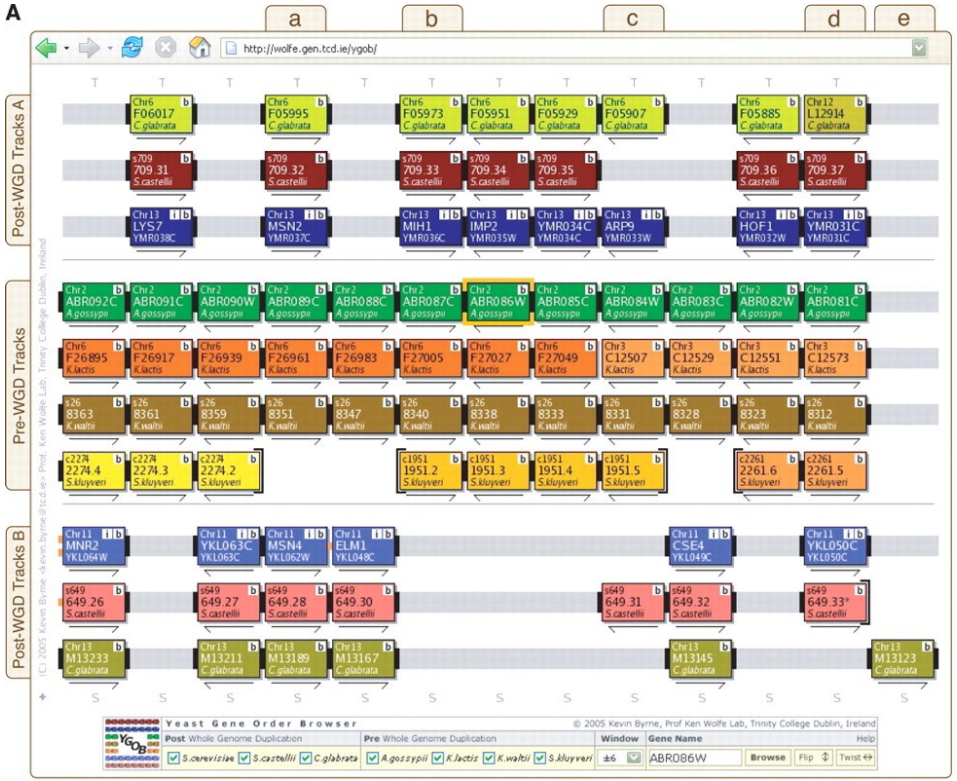
- Some species are hybrid (merging of two different genomes)
- Use all-vs-all BLAST to map orthologs and paralogs
- But which paralogs should be grouped together into *A* and *B* tracks
  - Sequence similarity to anchor taxa
  - Gene synteny structure

Synteny: conservation of gene order on chromosome



Cheng, F. et al. Frontiers in Plant Science 3:198 (2012)

Yeast Gene Order Browser

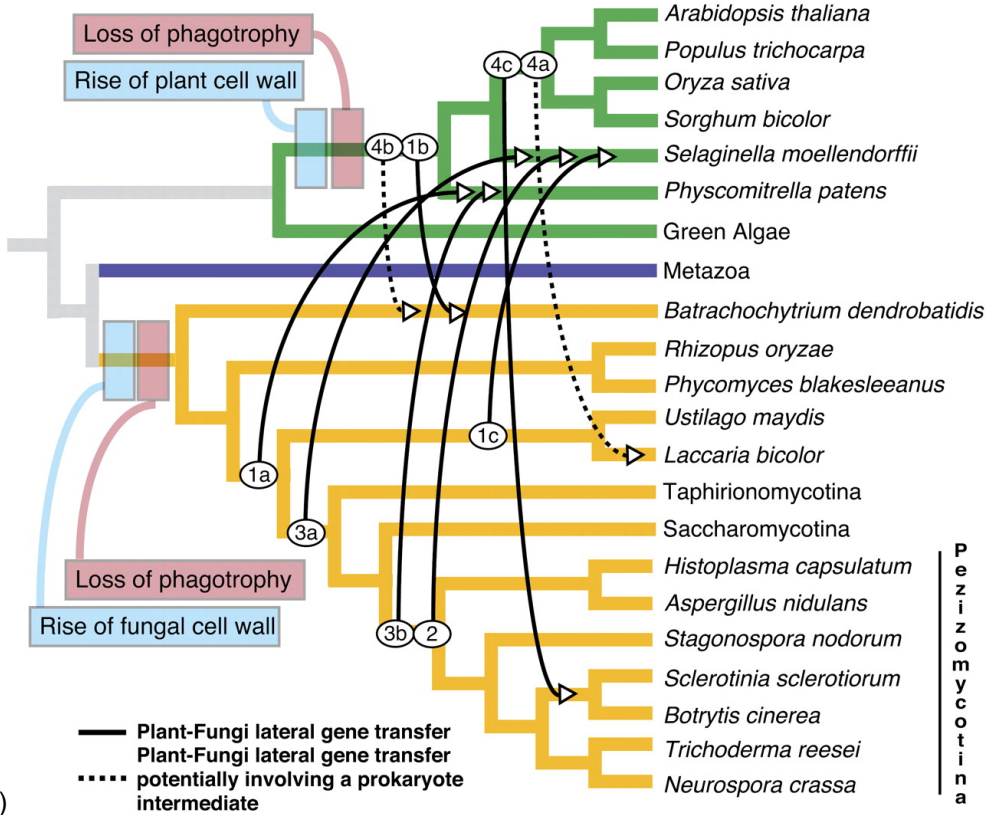


Byrne, K.P. and Wolfe, K.H. Genome Research 15:1456-61 (2005)

Horizontal gene transfer

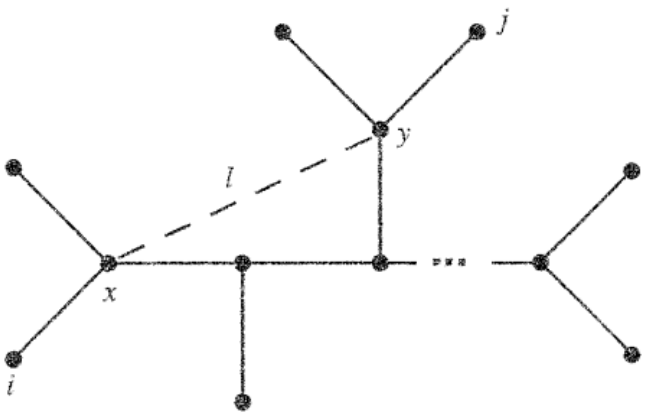
Horizontal gene transfer (HGT)

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



Richards *et al*. The Plant Cell (2009)

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
  - $\frac{\sum_{x\neq y}\left(d_{tree}(x,y)-d_{sequence}(x,y)\right)^2}{\frac{n(n-1)}{2}-N}$, where *n* is the number of taxa and *N* is the number of branches

Any question?

- See you on September 11