Using Phylogenies

Assessing Robustness and Genomic Epidemiology

Finlay Maguire

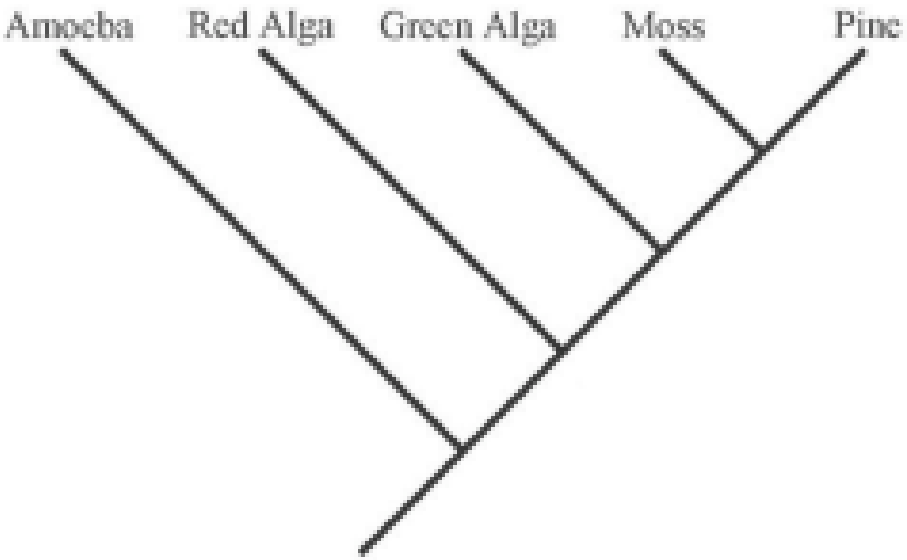
April 1, 2020

FCS, Dalhousie

Table of contents

Tree Thinking Refresher

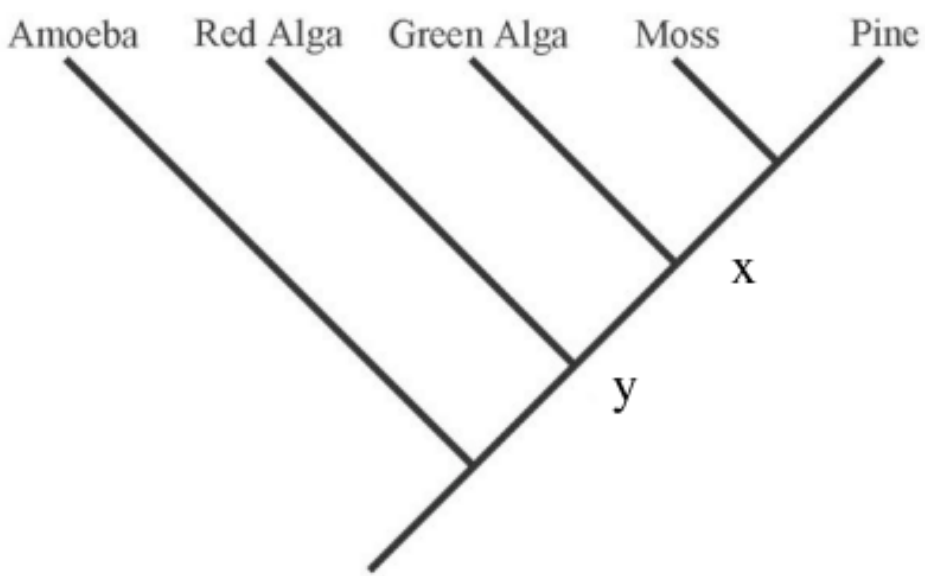
Reading a tree



which of the following is an accurate statement of relationships?

1. A green alga is more closely related to a red alga than to a moss
2. A green alga is more closely related to a moss than to a red alga
3. A green alga is equally related to a red alga and a moss
4. A green alga is related to a red alga, but is not related to a moss

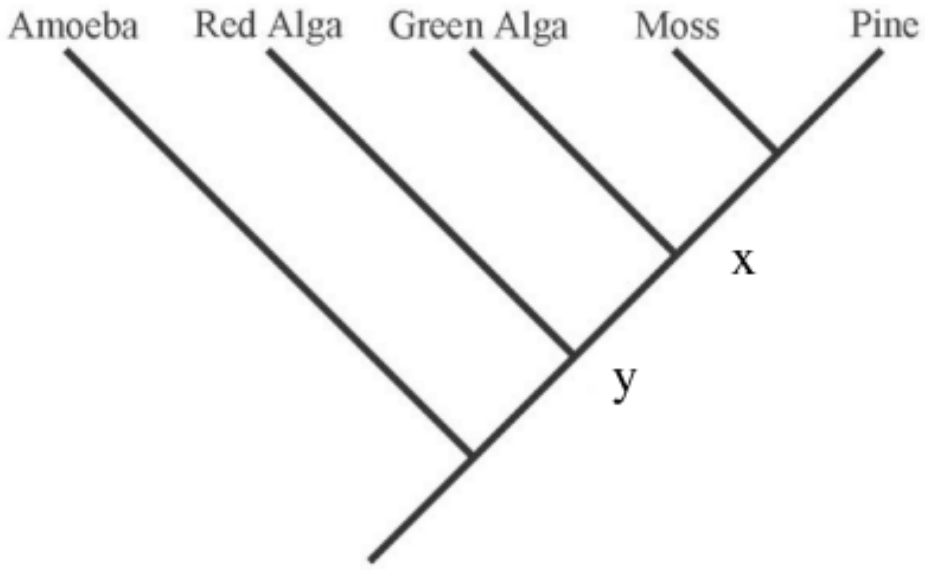
Reading a tree



which of the following is an accurate statement of relationships?

1. A green alga is more closely related to a red alga than to a moss
2. A green alga is more closely related to a moss than to a red alga
3. A green alga is equally related to a red alga and a moss
4. A green alga is related to a red alga, but is not related to a moss

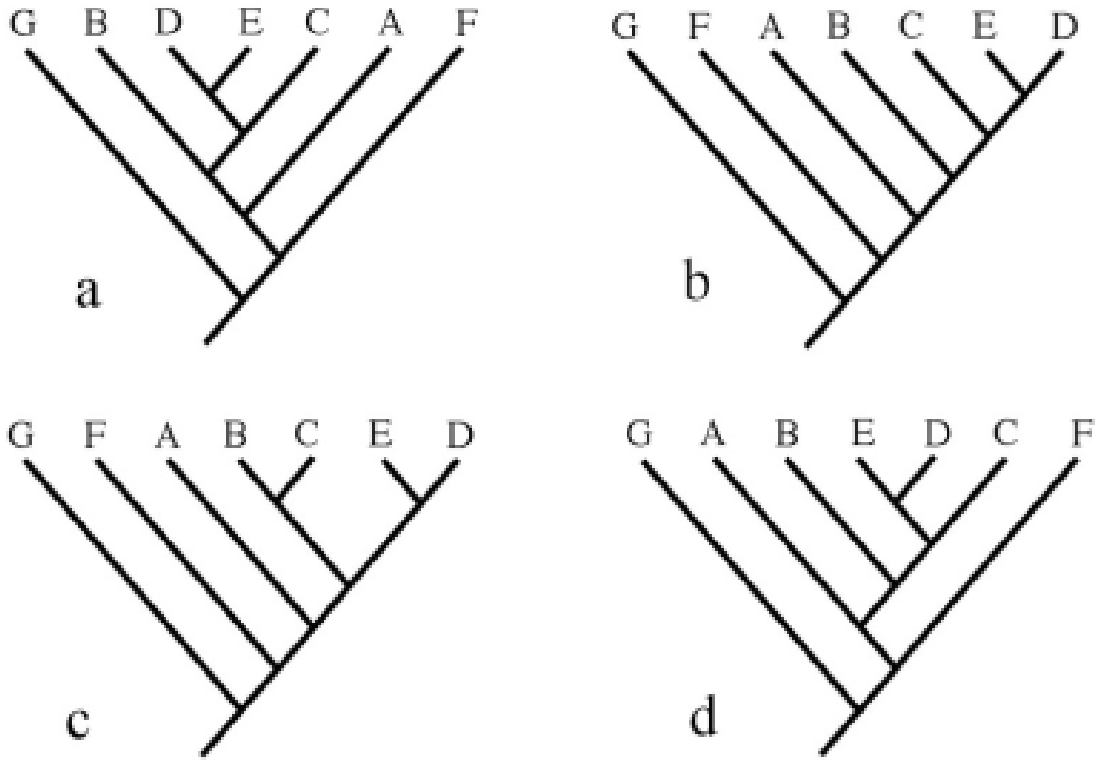
Reading a tree



which of the following is an accurate statement of relationships?

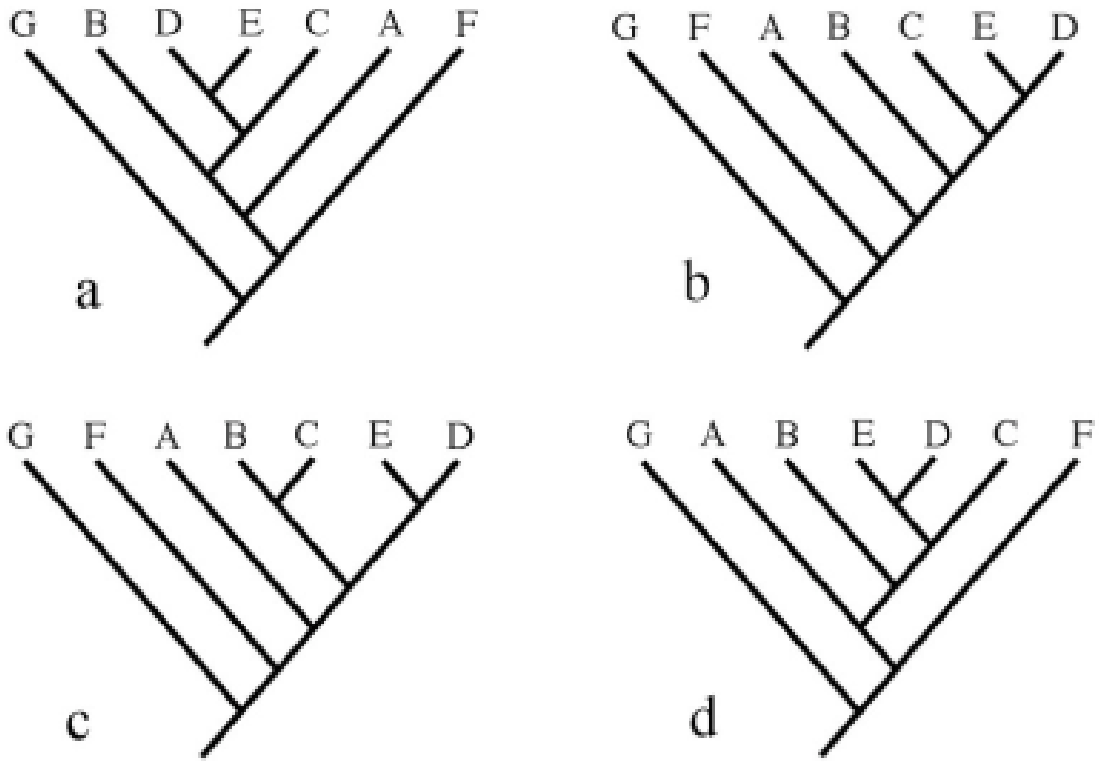
1. A green alga is more closely related to a red alga than to a moss
2. **A green alga is more closely related to a moss than to a red alga**
3. A green alga is equally related to a red alga and a moss
4. A green alga is related to a red alga, but is not related to a moss

Comparing Topologies



Which of the four trees depicts a different pattern of relationships to the others?

Comparing Topologies

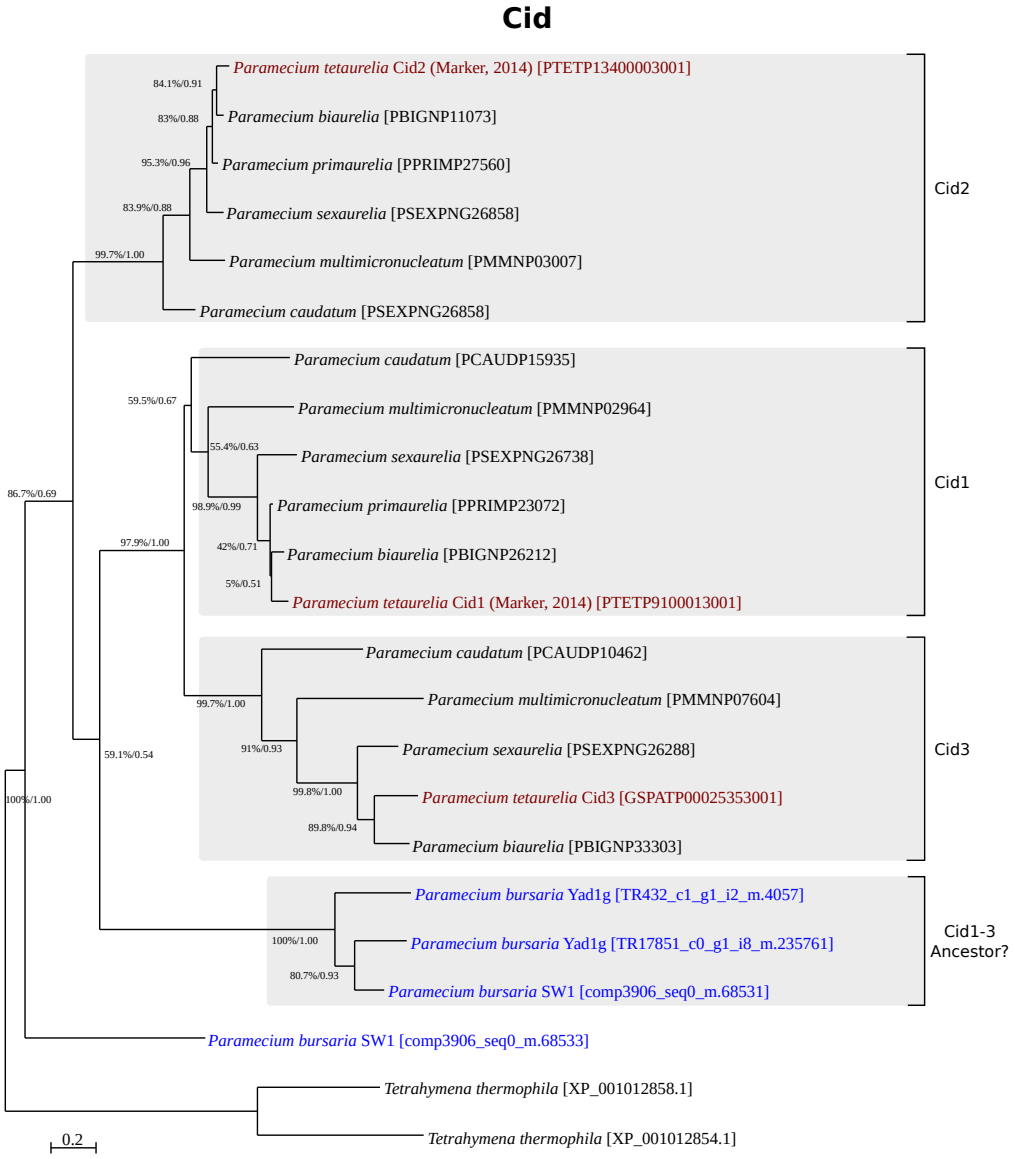


Which of the four trees depicts a different pattern of relationships to the others?

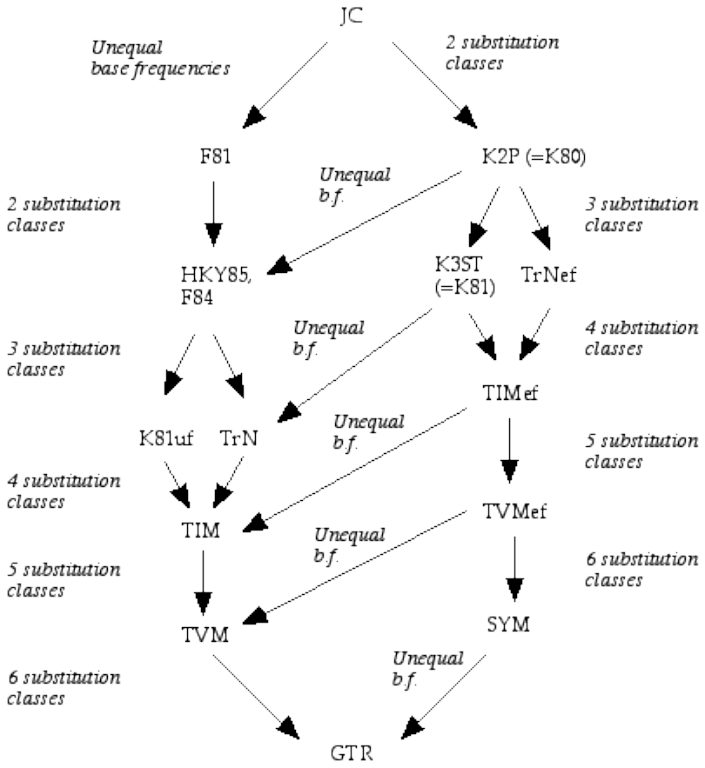
c: C is more closely related to E and D than to B in other trees.

Sequence Model Selection

Phylogenies are hypotheses

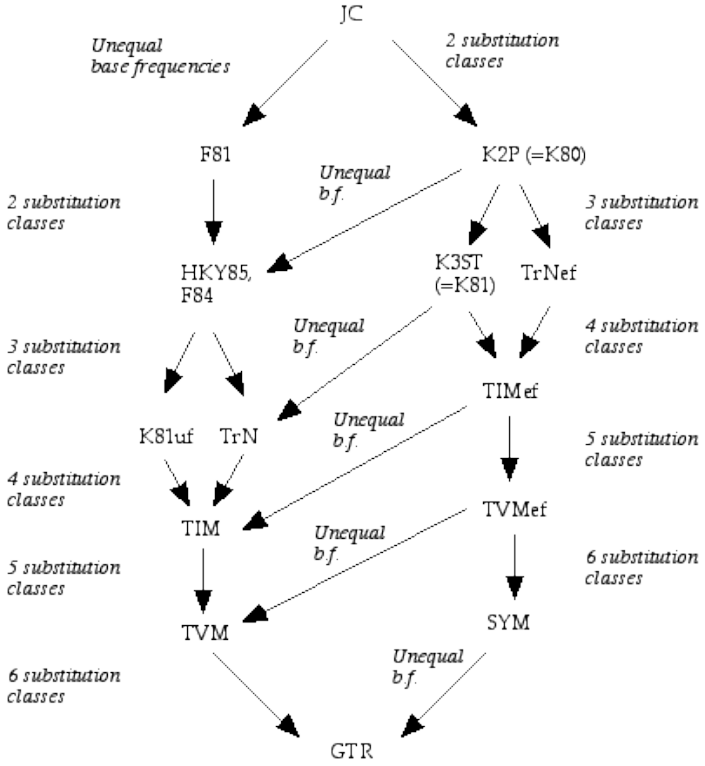


Hypothesis testing



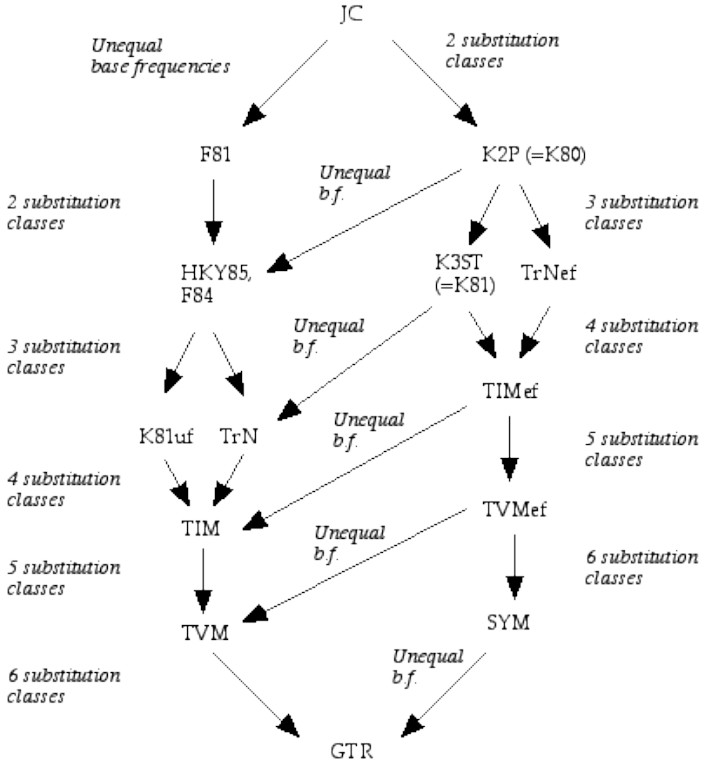
- Does another model of sequence evolution fit the data better?
- How well supported are individual branches in a tree?
- Does another tree explain the data better?

How do we select a sequence model?



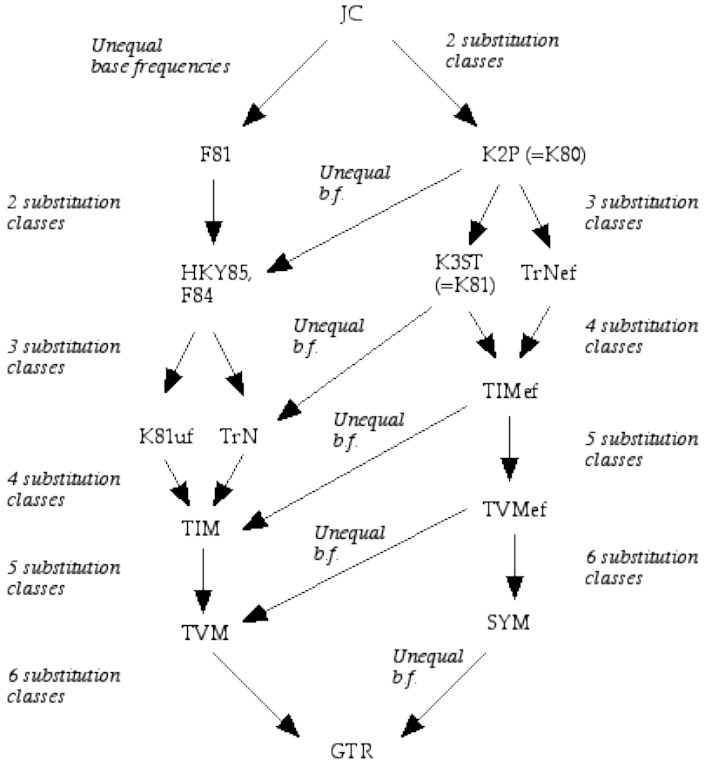
- Likelihood ratio test (LRT δ i.e. p(data| model)

How do we select a sequence model?



- Likelihood ratio test (LRT δ i.e. p(data| model)
- $\delta = 2(ln(L_1) - ln(L_0))$

How do we select a sequence model?



- Likelihood ratio test (LRT δ i.e. p(data| model)
- $\delta = 2(ln(L_1) - ln(L_0))$
- Limitations: nested models (i.e. hLRT), order matters, no regularisation

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.
- Corrected Akaike Information Criterion (**AICc**)

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.
- Corrected Akaike Information Criterion (**AICc**)
- $AICc = AIC + \frac{2K(K+1)}{n-K-1}$

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.
- Corrected Akaike Information Criterion (**AICc**)
- $AICc = AIC + \frac{2K(K+1)}{n-K-1}$
- Alternatively, there is the Bayesian Information Criterion (**BIC**):

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.
- Corrected Akaike Information Criterion (**AICc**)
- $AICc = AIC + \frac{2K(K+1)}{n-K-1}$
- Alternatively, there is the Bayesian Information Criterion (**BIC**):
- $BIC = -2ln(L) + Kln(n)$

Information Criterion

- Akaike Information Criterion (**AIC**), penalising number of parameters:
- $AIC = -2ln(L) + 2K$
- However, this penalises all high K models even if sample size is large too.
- Corrected Akaike Information Criterion (**AICc**)
- $AICc = AIC + \frac{2K(K+1)}{n-K-1}$
- Alternatively, there is the Bayesian Information Criterion (**BIC**):
- $BIC = -2ln(L) + Kln(n)$
- Decision Theory (**DT**) risk minimisation approach.

Limitations

- What if everything fits poorly?

Limitations

- What if everything fits poorly?
- Information criterion test relative goodness of fit instead of absolute

Limitations

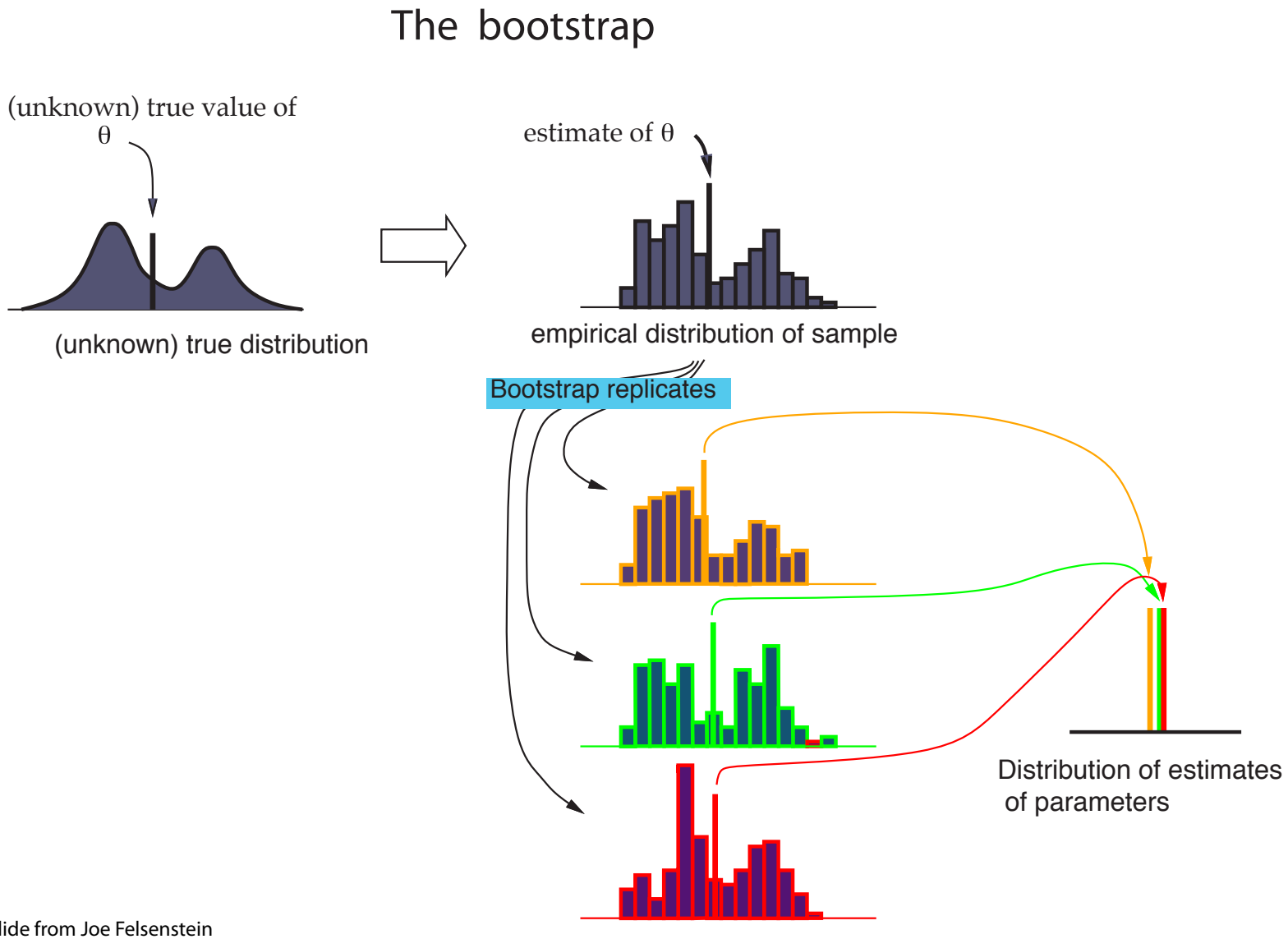
- What if everything fits poorly?
- Information criterion test relative goodness of fit instead of absolute
- Parametric Bootstrapping/Posterior Predictive Simulation

Limitations

- What if everything fits poorly?
- Information criterion test relative goodness of fit instead of absolute
- Parametric Bootstrapping/Posterior Predictive Simulation
- If the model is reasonable then data simulated under should resemble the empirical data

Branch Support Testing

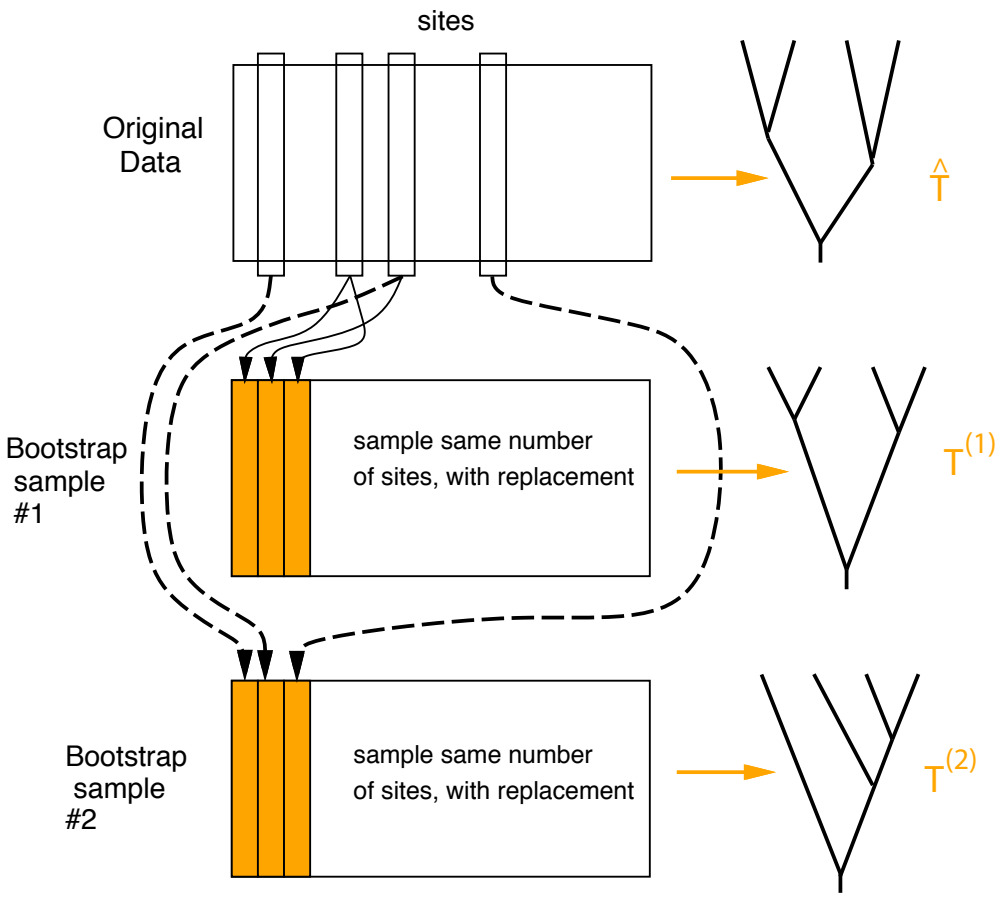
Bootstrapping in General



Slide from Joe Felsenstein

Bootstrapping Phylogenies

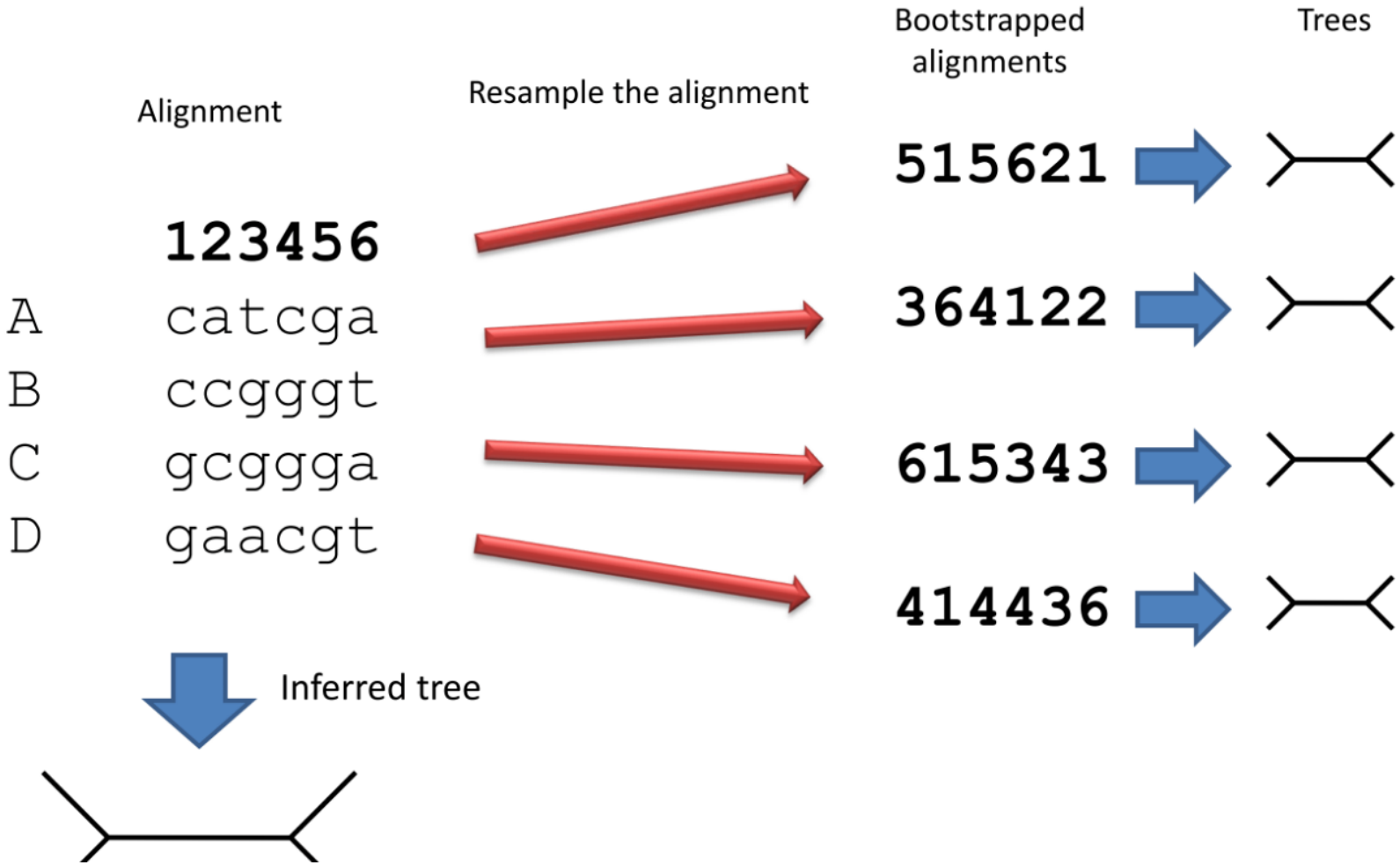
The bootstrap for phylogenies



Slide from Joe Felsenstein

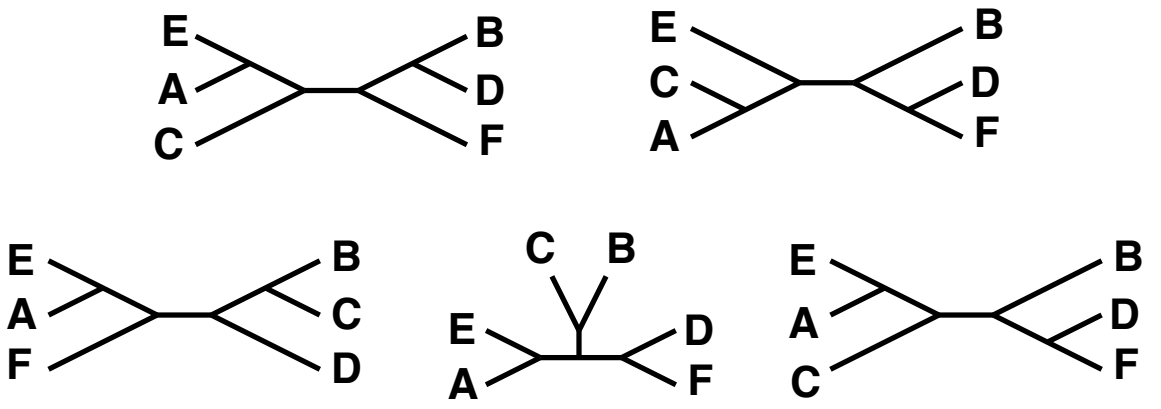
(and so on)

Bootstrapping Phylogenies

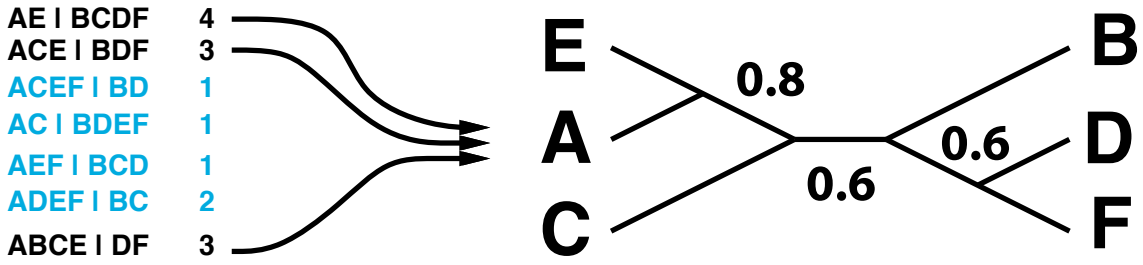


The majority-rule consensus tree

Trees:

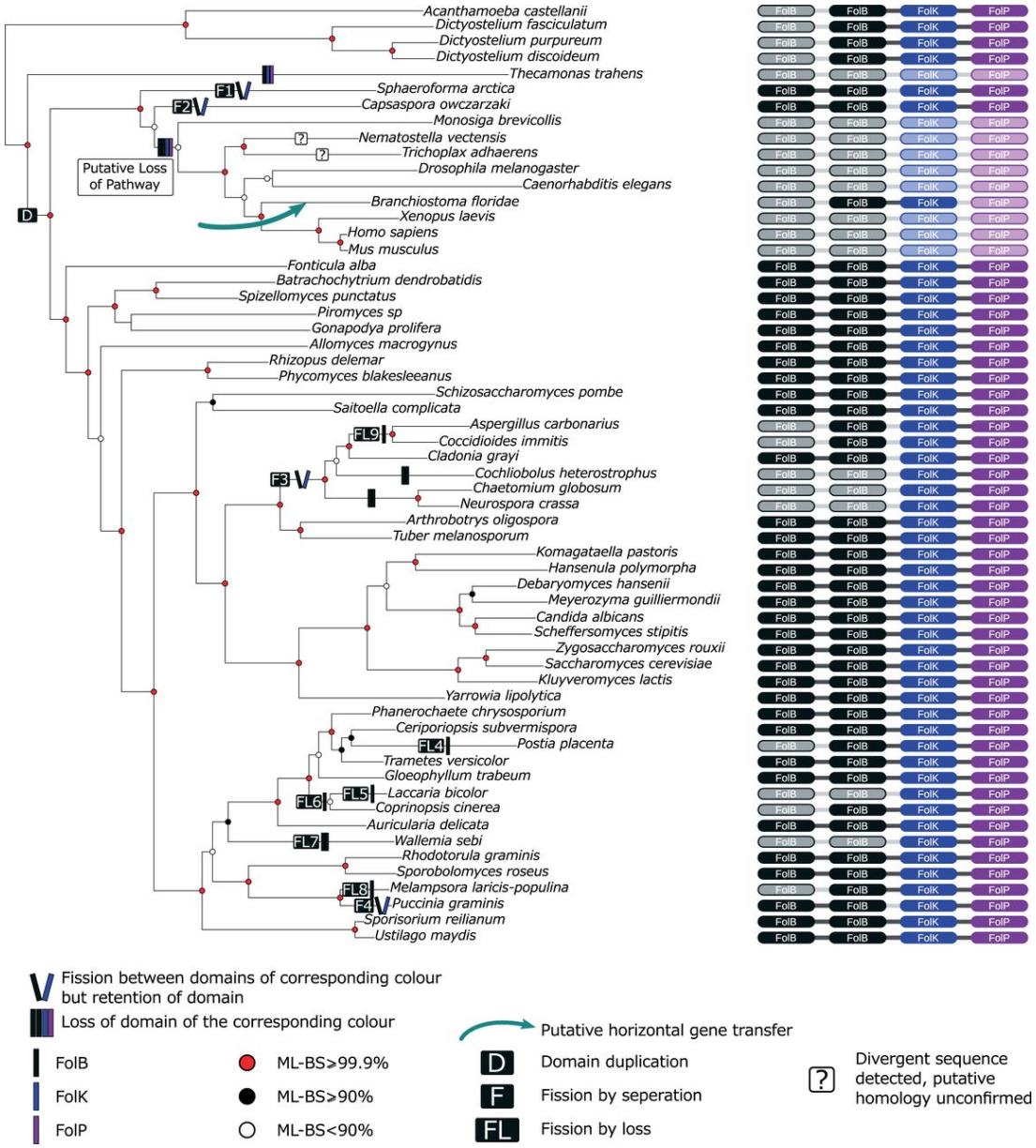


How many times each partition of species is found:



Slide from Joe Felsenstein

Combining the results



What is the bootstrap doing?

- Randomly reweighing the sites in an alignments

What is the bootstrap doing?

- Randomly reweighing the sites in an alignments
- Probability of a site being excluded $1 - \frac{1}{n} n$

What is the bootstrap doing?

- Randomly reweighing the sites in an alignments
- Probability of a site being excluded $1 - \frac{1}{n}n$
- Asymptotically approximately 0.36

What is the bootstrap doing?

- Randomly reweighing the sites in an alignments
- Probability of a site being excluded $1 - \frac{1}{n}n$
- Asymptotically approximately 0.36
- Goal to simulate an infinite population (number of alignment columns)

Limitations

- Typically underestimates the true probabilities

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding
- Assumes independence of sites

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding
- Assumes independence of sites
- Relies on good input data

Limitations

- Typically underestimates the true probabilities
- i.e biased but conservative
- Computationally demanding
- Assumes independence of sites
- Relies on good input data
- Only answers to what extent does input data support a given part of the tree

Parametric Bootstraps

- Simulate data sets of this size assuming the estimate of the tree is the truth
- Key for many more sophisticated tests.
- Can be used to generate p-values, but non-trivial

Alternative Approaches

- Resampling estimated log-likelihoods (**RELL**)

Alternative Approaches

- Resampling estimated log-likelihoods (**RELL**)
- Instead of re-doing the full ML inference just re-sample the site $ln(L)$ values and sum

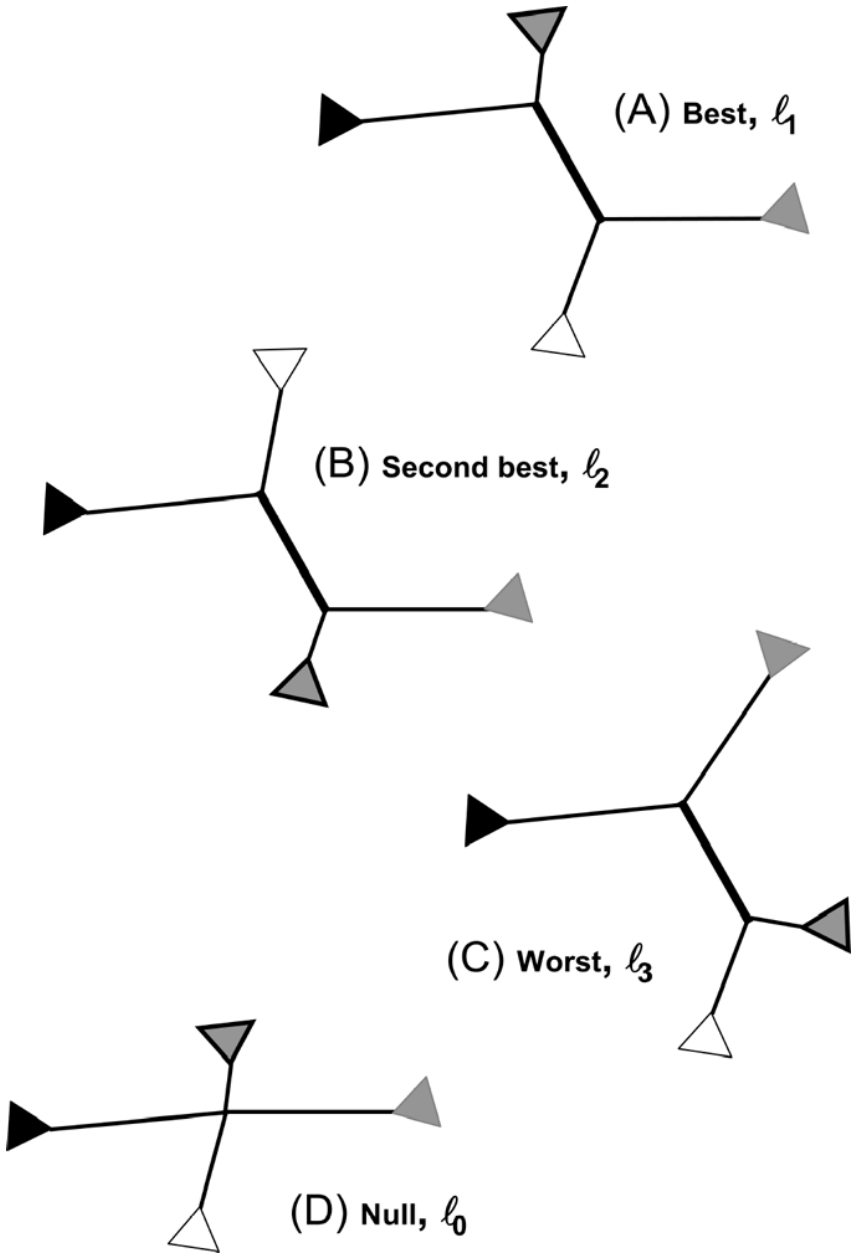
Alternative Approaches

- Resampling estimated log-likelihoods (**RELL**)
- Instead of re-doing the full ML inference just re-sample the site $ln(L)$ values and sum
- Rapid Bootstraps (**RBS**)

Alternative Approaches

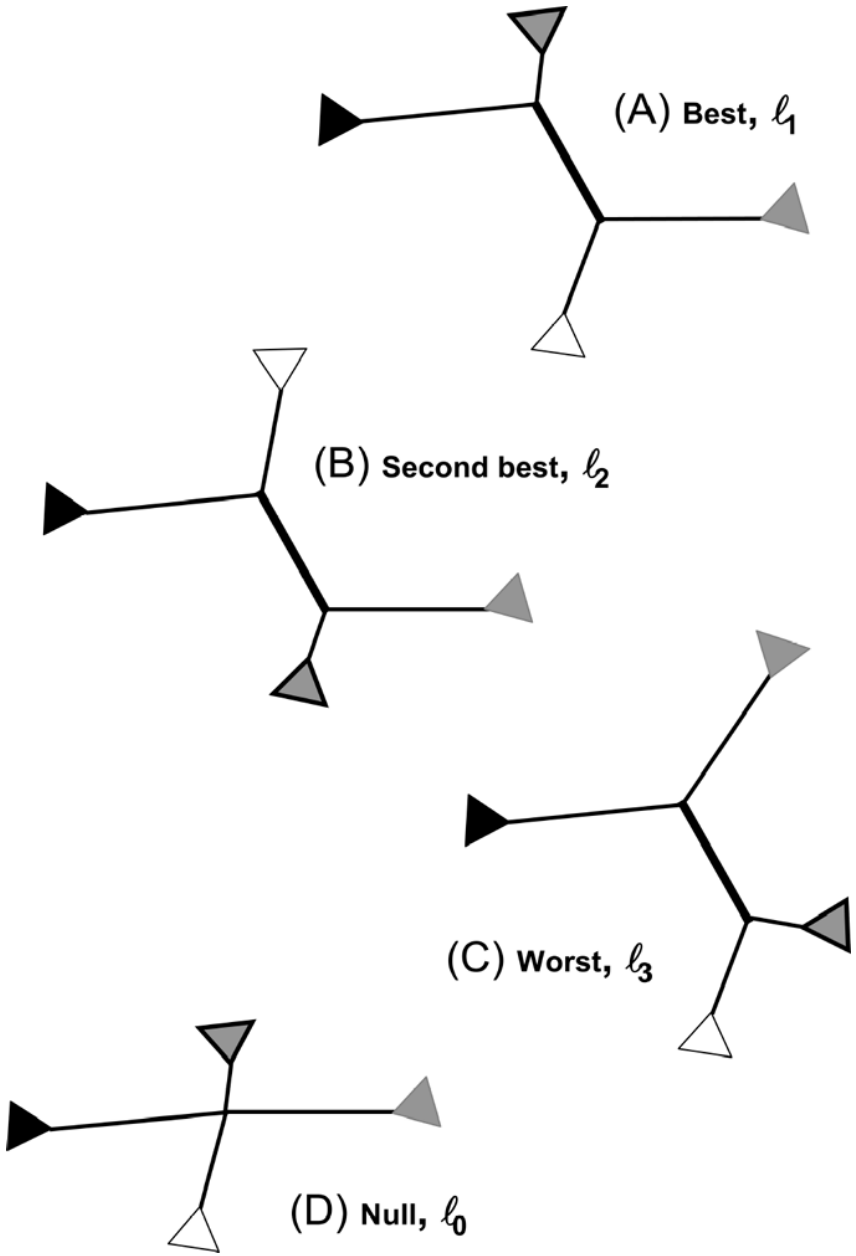
- Resampling estimated log-likelihoods (**RELL**)
- Instead of re-doing the full ML inference just re-sample the site $ln(L)$ values and sum
- Rapid Bootstraps (**RBS**)
- Ultrafast Bootstraps (**UFBoot**)

Likelihood Tests



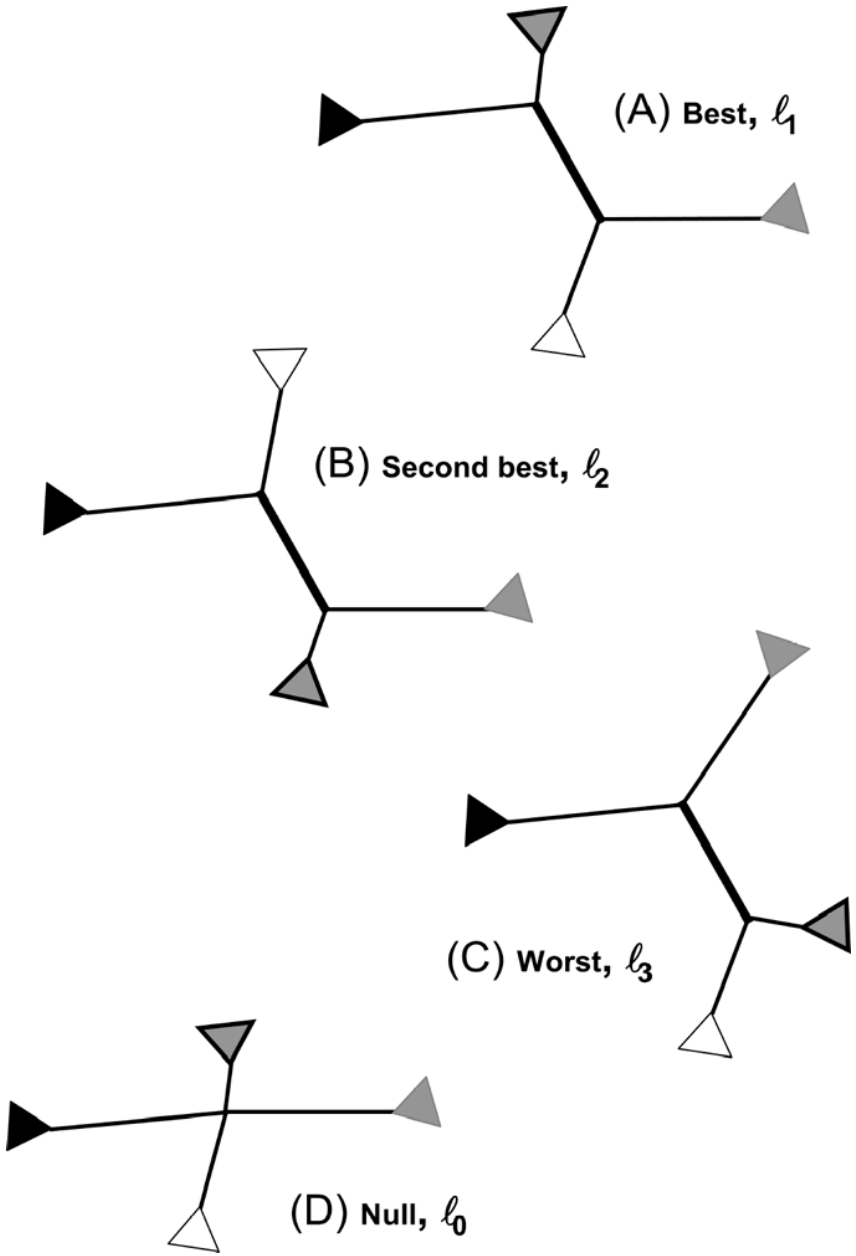
- Comparing the 3 nearest NNIs to a given branch:

Likelihood Tests



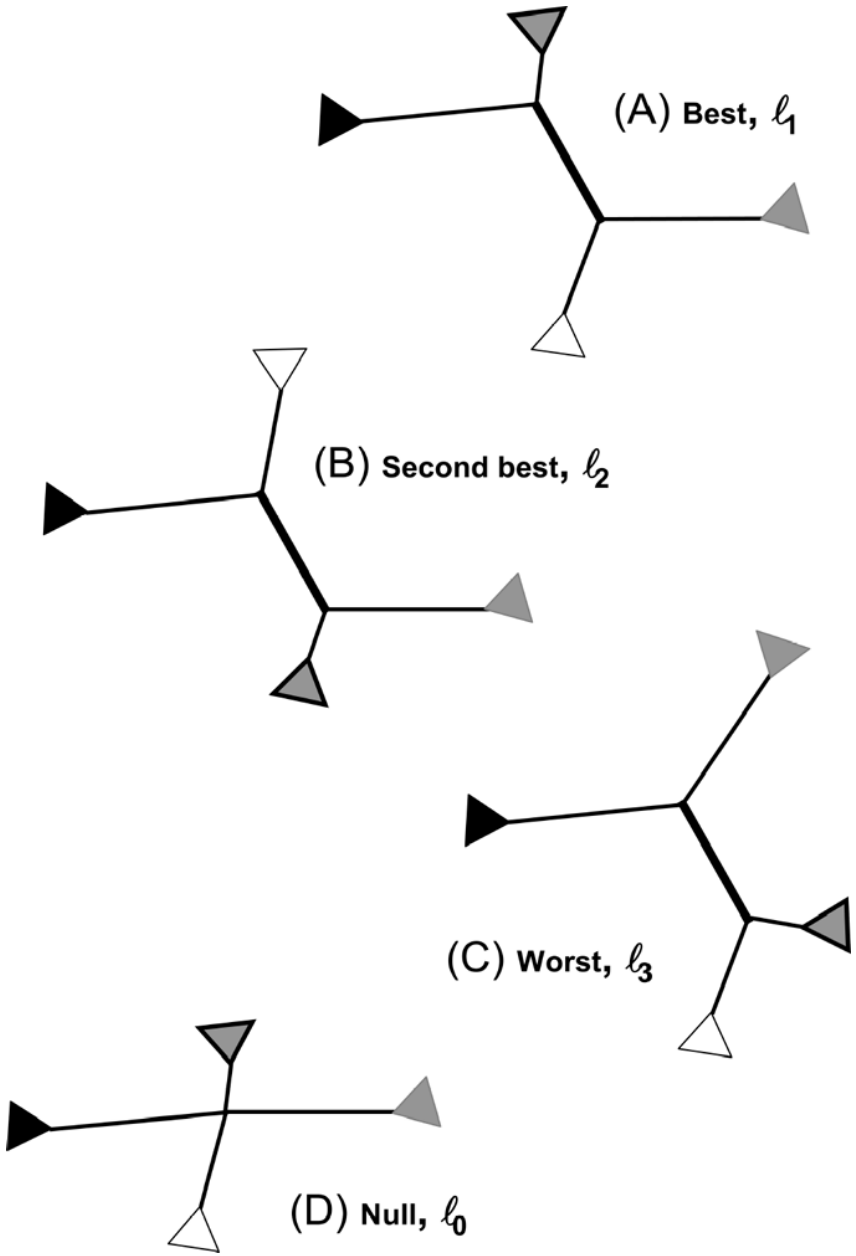
- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs

Likelihood Tests



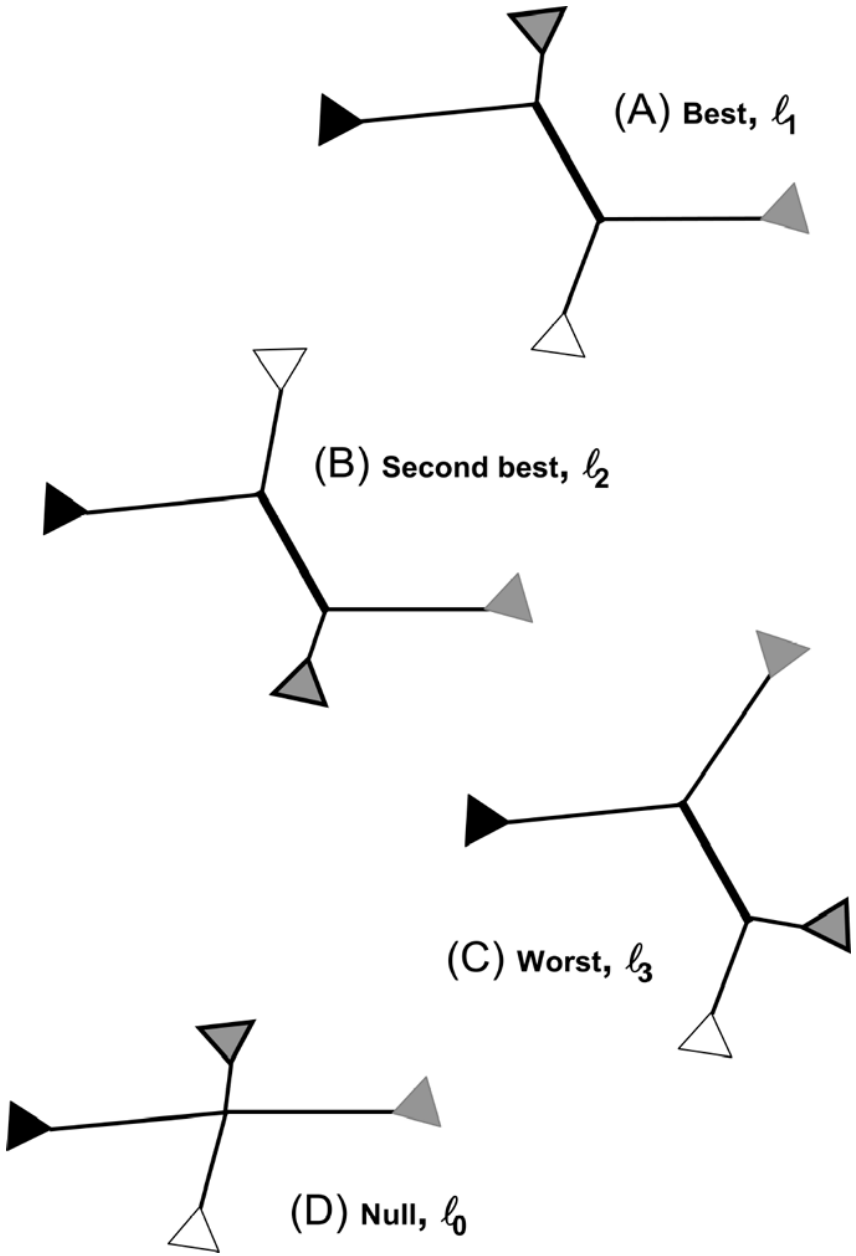
- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs
- Non-parametric **SH-aLRT** based on RELL

Likelihood Tests



- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs
- Non-parametric **SH-aLRT** based on RELL
- **aBayes**:

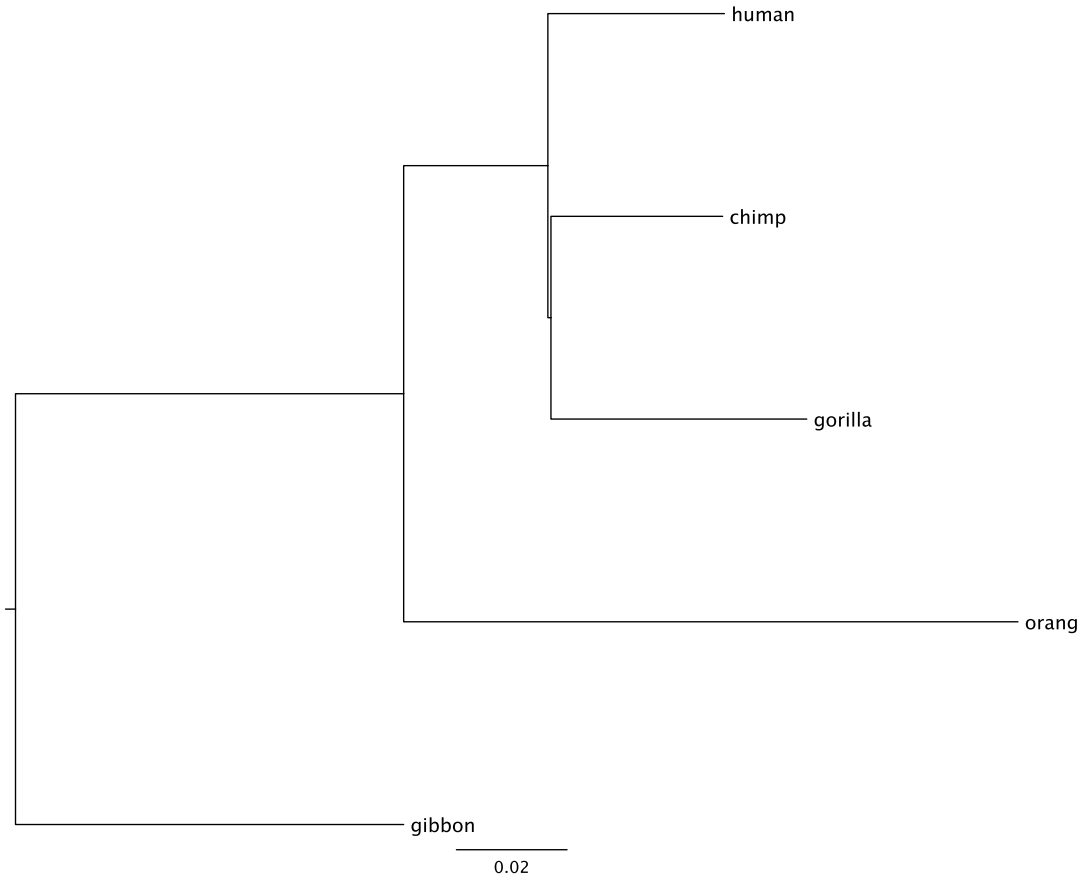
Likelihood Tests



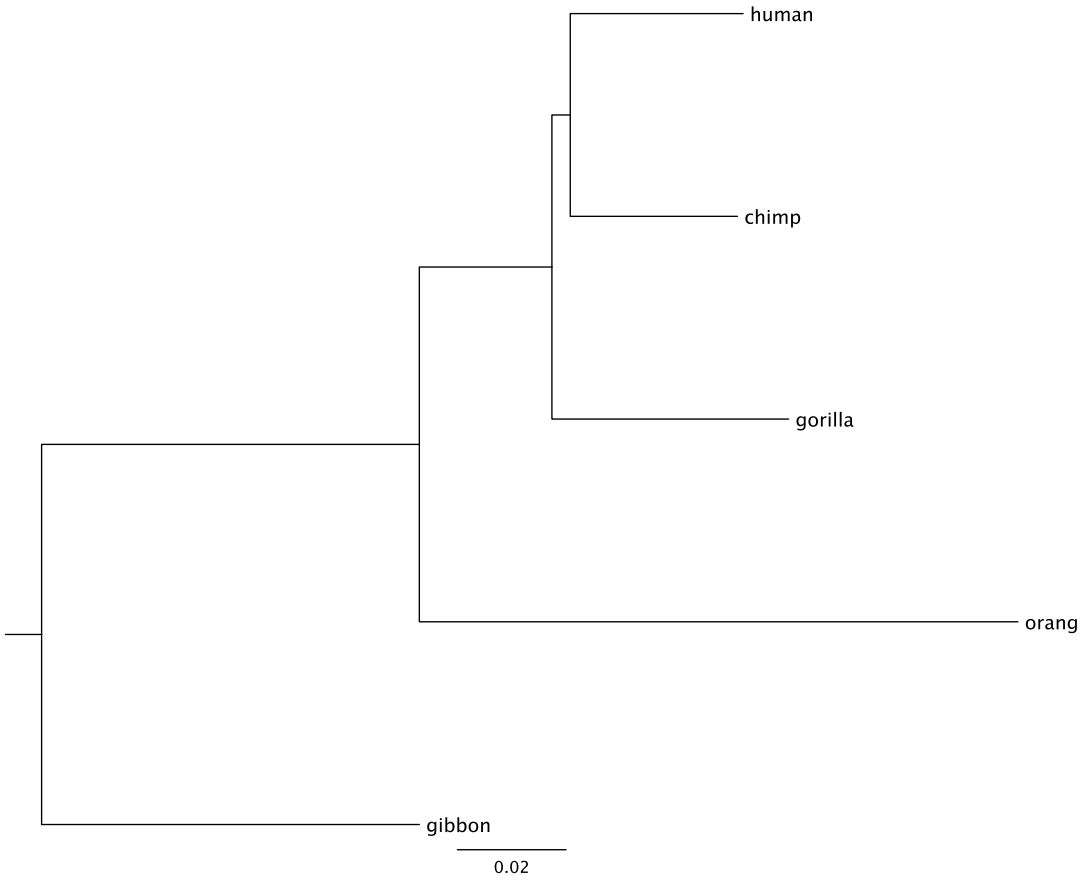
- Comparing the 3 nearest NNIs to a given branch:
- Parametric **aLRT**: χ^2 of δ for branch vs. closest NNIs
- Non-parametric **SH-aLRT** based on RELL
- **aBayes**:
- $P(T_c \mid X) = \frac{P(X|T_c)P(T_c)}{\sum_i^2 = 0 P(X||T_i)P(T_i)}$ with flat prior $P(T_0) = P(T_1) = P(T_2)$

Comparing Trees

How to compare competing hypotheses?

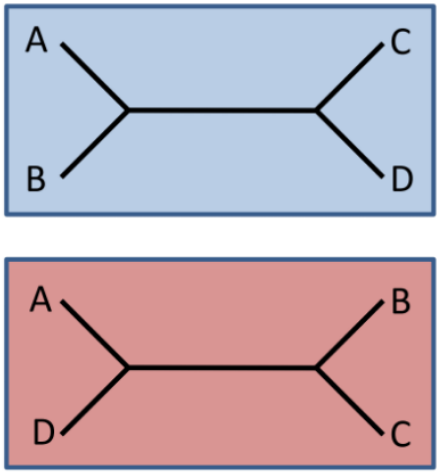


How to compare competing hypotheses?



Simplistic Comparison

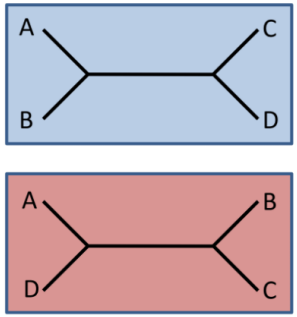
	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2



19

Qualitative Comparison

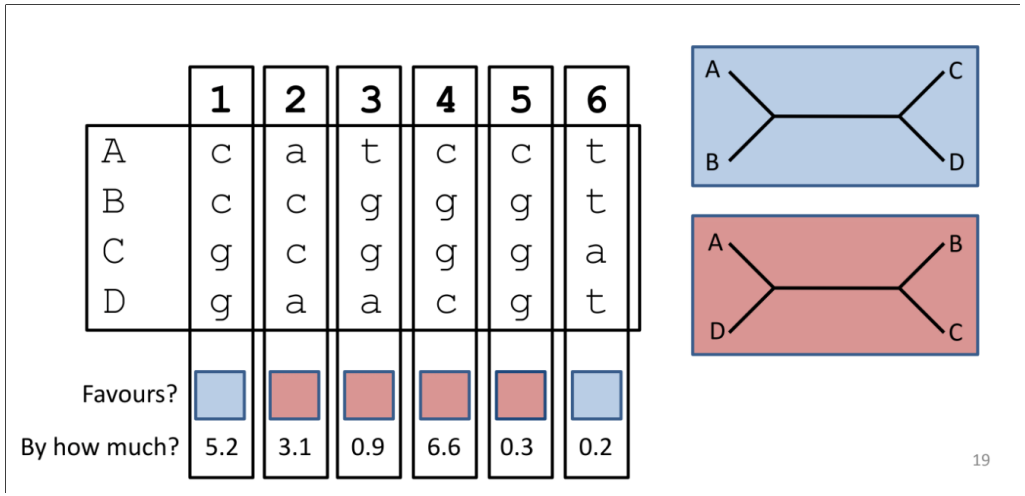
	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2



19

- 4 sites favour the red tree, 2 favour the blue

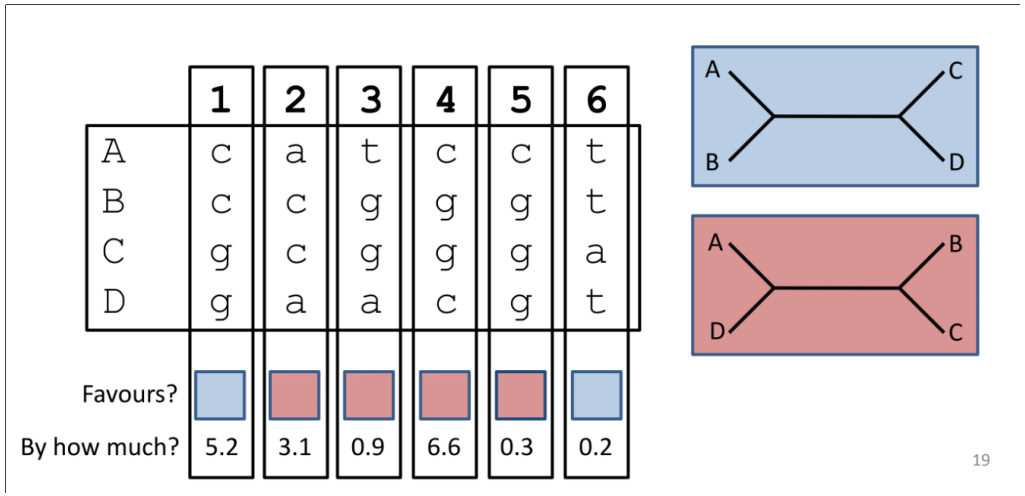
Qualitative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$

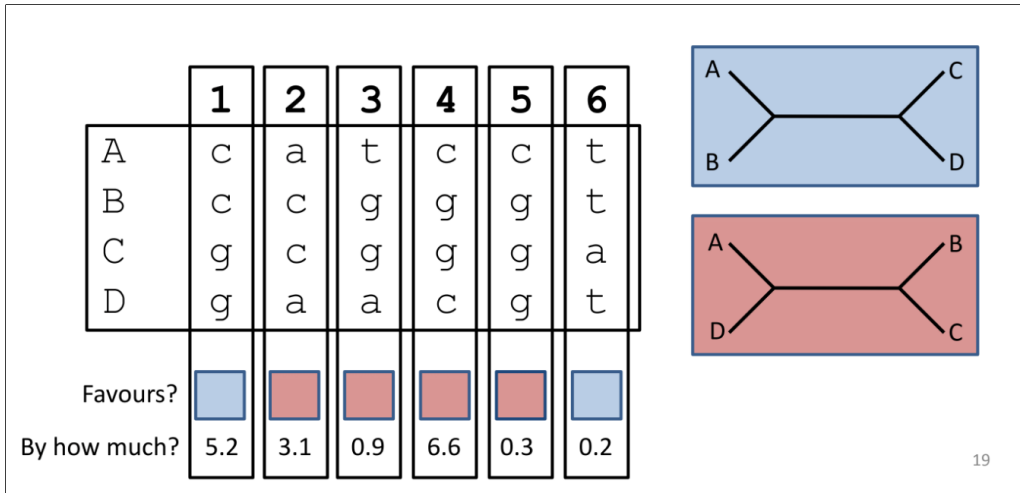
Qualitative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$

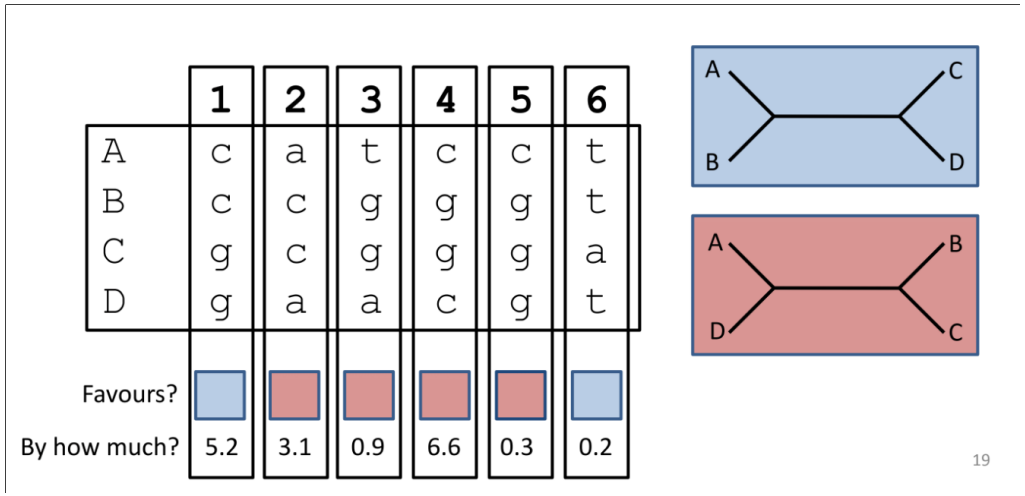
Qualitative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$
- 40 out of 60 $p = 0.0124$

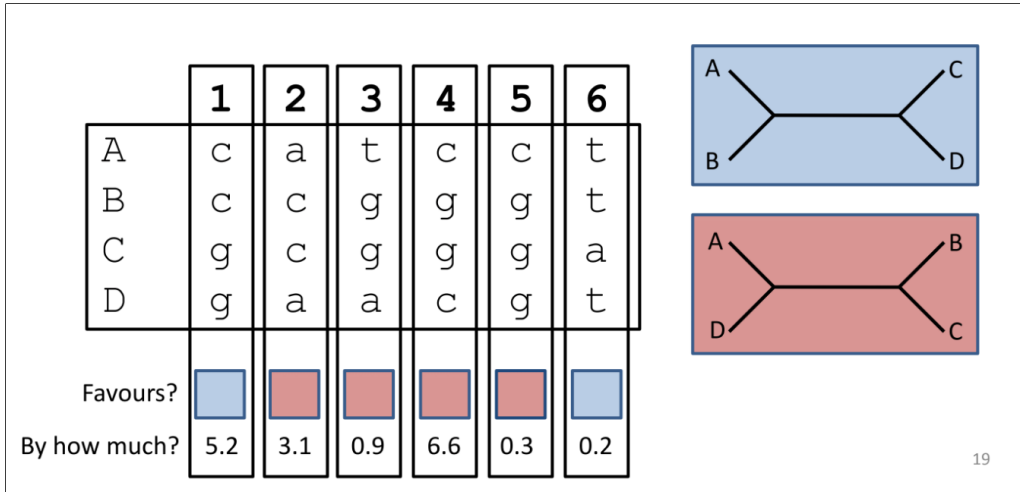
Qualitative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- 4 sites favour the red tree, 2 favour the blue
- $\binom{n}{k} p^k (1-p)^{n-k}$
- 4 out of 6 $p = 0.6875$
- 40 out of 60 $p = 0.0124$
- 400 out of 600 $p = 2.3 * 10^{-16}$

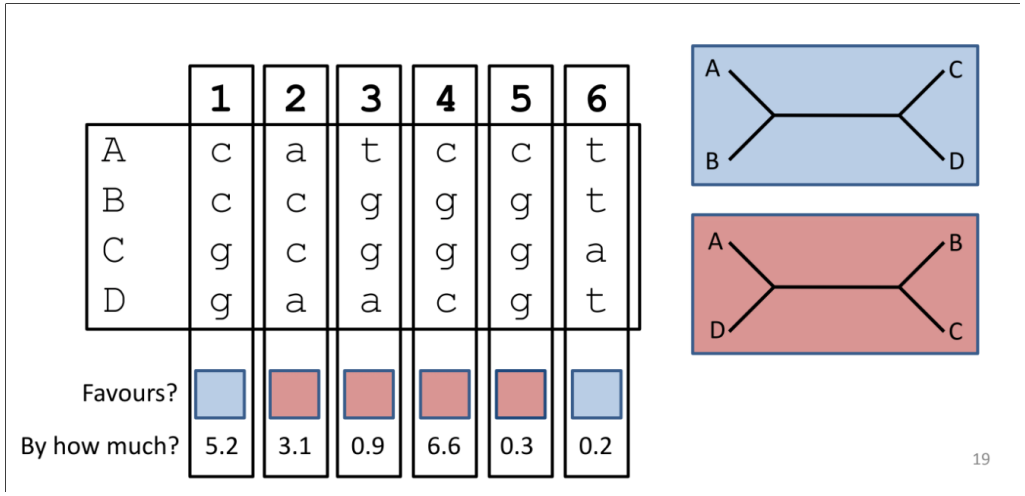
Quantiative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- $\mu = (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 - 0.2)/6 = 0.916$

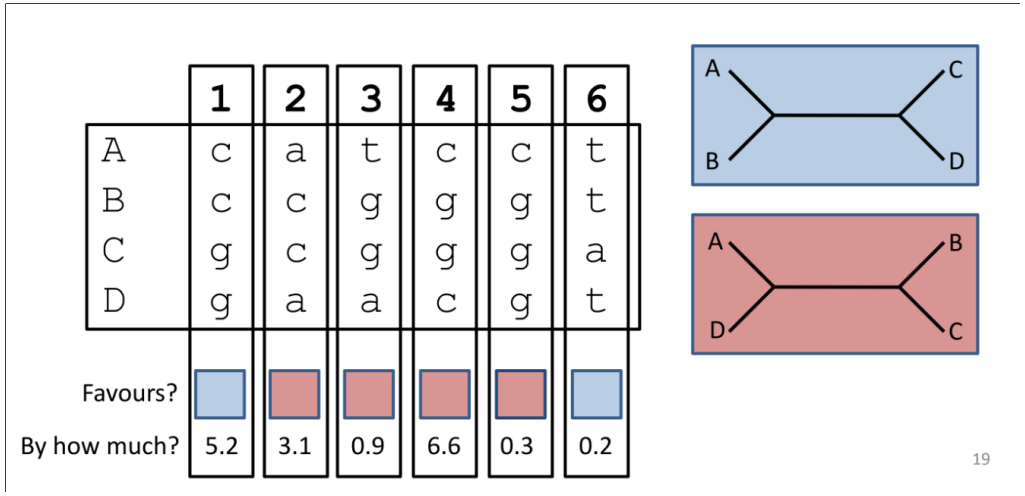
Quantiative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- $\mu = (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 - 0.2)/6 = 0.916$
- $\sigma^2 = 15.22$

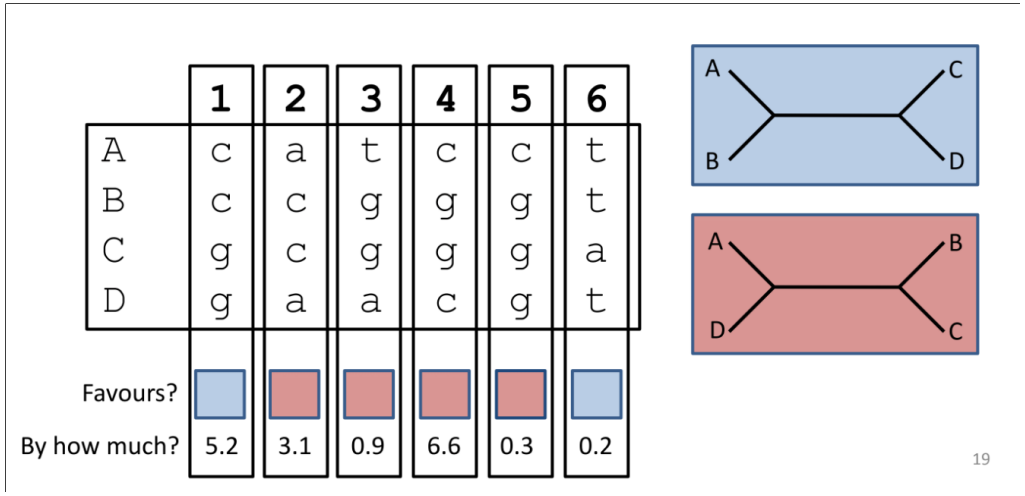
Quantiative Comparison



	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

- $\mu = (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 - 0.2)/6 = 0.916$
- $\sigma^2 = 15.22$
- $t = \frac{\mu}{\sigma^2} * \sqrt{N} = 0.148$

Quantiative Comparison



- $\mu = (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 - 0.2)/6 = 0.916$
- $\sigma^2 = 15.22$
- $t = \frac{\mu}{\sigma^2} * \sqrt{N} = 0.148$
- therefore: $p = 0.888$ under $5 d.f.$

More robust approaches

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.

More robust approaches

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(X \mid T_1, T_2) = 2\,[\ln L(X \mid T_1) - \ln L(X \mid T_2)]$

More robust approaches

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(X \mid T_1, T_2) = 2\left[\ln L(X \mid T_1) - \ln L(X \mid T_2)\right]$
- Expectation under null $\mathbb{E}\left[\delta(X \mid T_1, T_2)\right] = 0$

More robust approaches

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(X \mid T_1, T_2) = 2\left[\ln L(X \mid T_1) - \ln L(X \mid T_2)\right]$
- Expectation under null $\mathbb{E}\left[\delta(X \mid T_1, T_2)\right] = 0$
- Why can't we just use χ^2 to get a critical value for δ?

More robust approaches

- Null: if no sampling error (infinite data) T_1 and T_2 would explain the data equally well.
- $\delta(X \mid T_1, T_2) = 2\left[\ln L(X \mid T_1) - \ln L(X \mid T_2)\right]$
- Expectation under null $\mathbb{E}\left[\delta(X \mid T_1, T_2)\right] = 0$
- Why can't we just use χ^2 to get a critical value for δ?
- Tree space is difficult.

Estimating variance of the null

- Many avenues:
- Non-parametric bootstrapping
- Parametric bootstrapping
- Related approaches.

Alternative tests

- Shimodaira-Hasegawa Test

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**
- Achieves weighted by varying bootstrap size for each tree.

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**
- Achieves weighted by varying bootstrap size for each tree.
- Better for larger comparisons, can have issues with P-space curvature.

Alternative tests

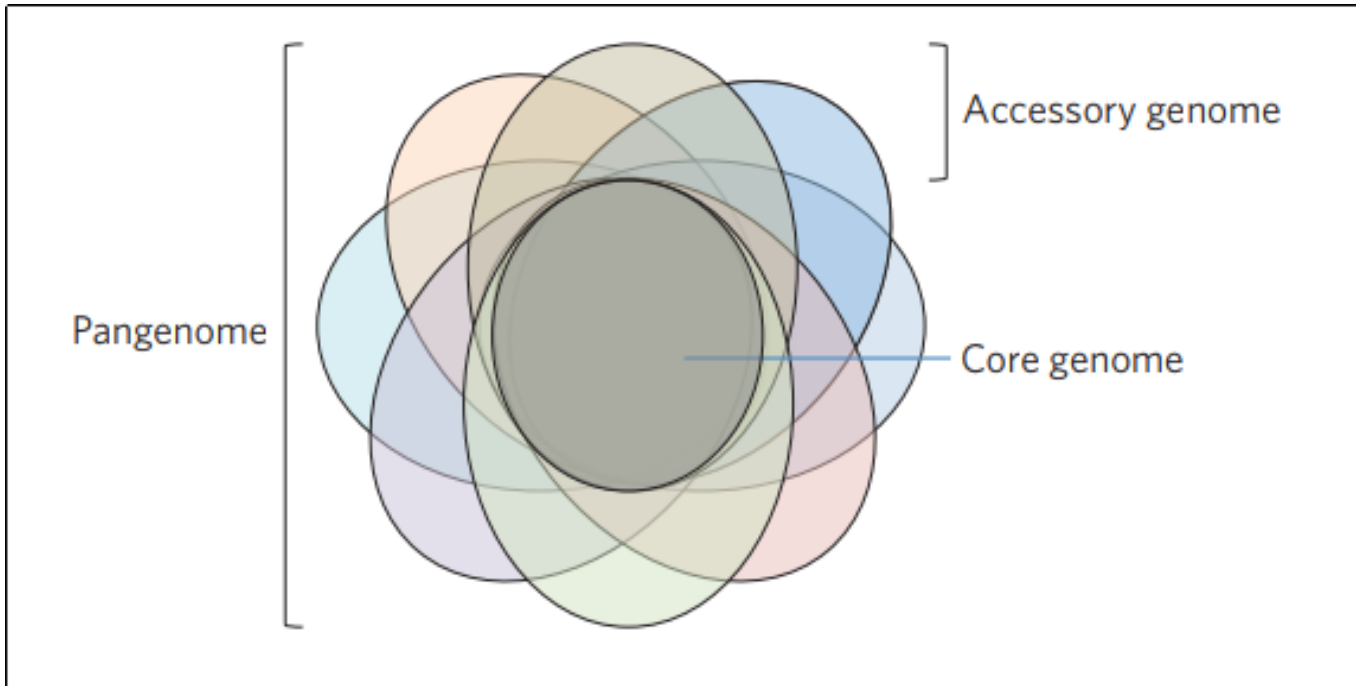
- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**
- Achieves weighted by varying bootstrap size for each tree.
- Better for larger comparisons, can have issues with P-space curvature.
- **Swofford–Olsen–Waddell–Hillis** same idea but uses parametric bootstraps instead.

Alternative tests

- **Shimodaira-Hasegawa Test**
- Compares candidate tree sets
- $H_0 =$ all topologies equally good
- Very conservative when the number of candidate trees is large
- Can be corrected with weighted SH-test overcomes.
- **Approximately Unbiased Test**
- Achieves weighted by varying bootstrap size for each tree.
- Better for larger comparisons, can have issues with P-space curvature.
- **Swofford–Olsen–Waddell–Hillis** same idea but uses parametric bootstraps instead.
- Sensitive to model misspecification.

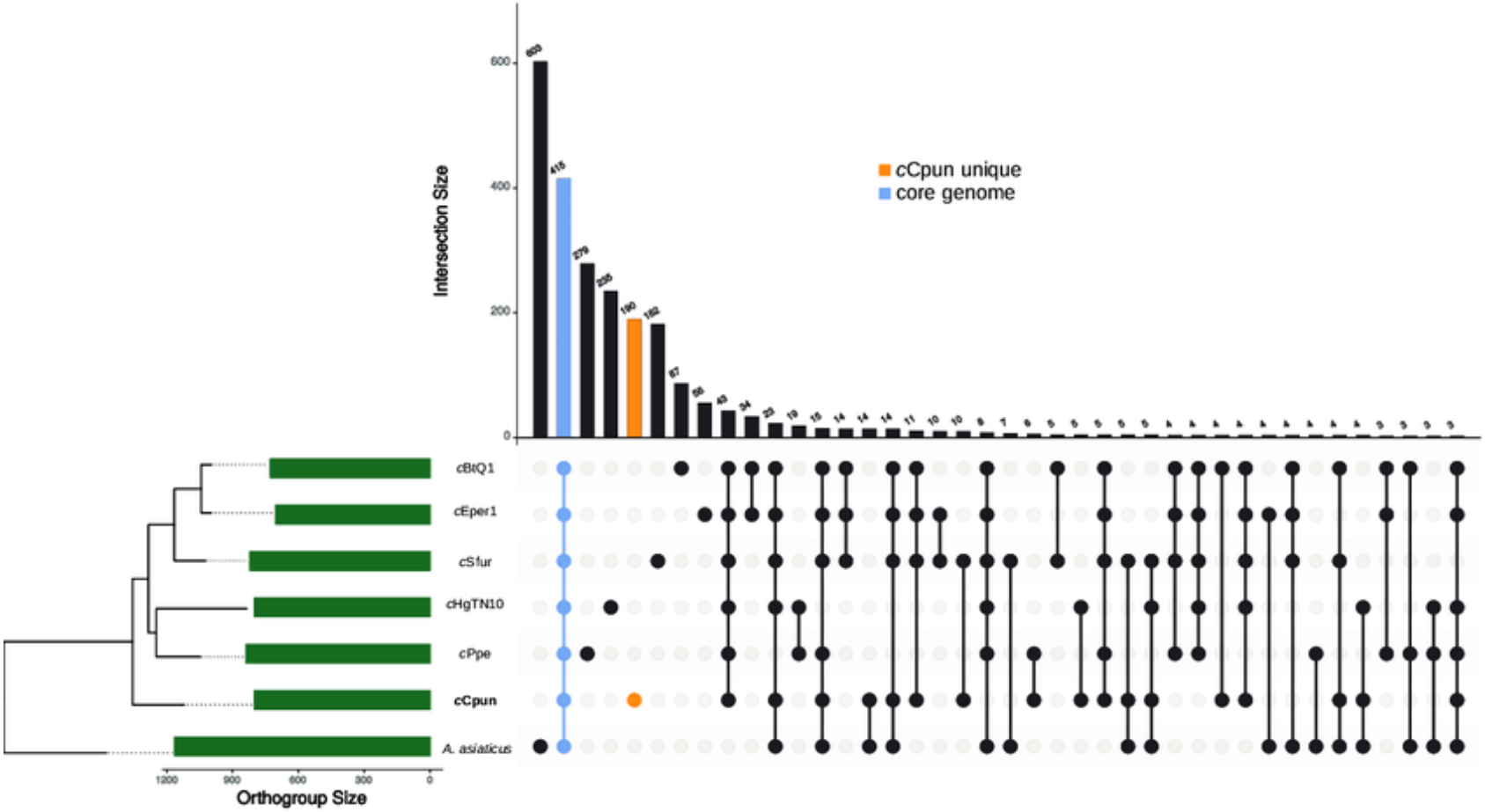
From A Single Gene to Many Genes

Core vs Pan-Genome

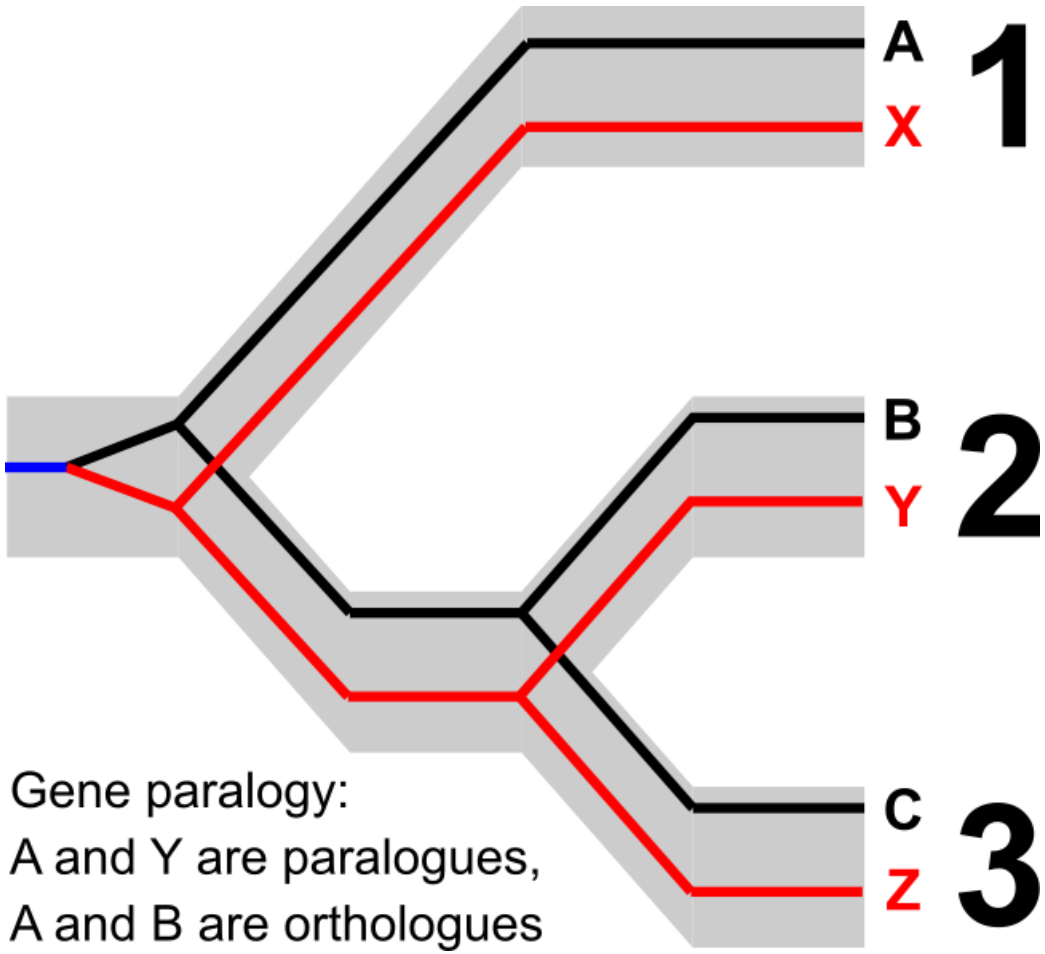


- How do we go from a bunch of individual genes to a species phylogeny?

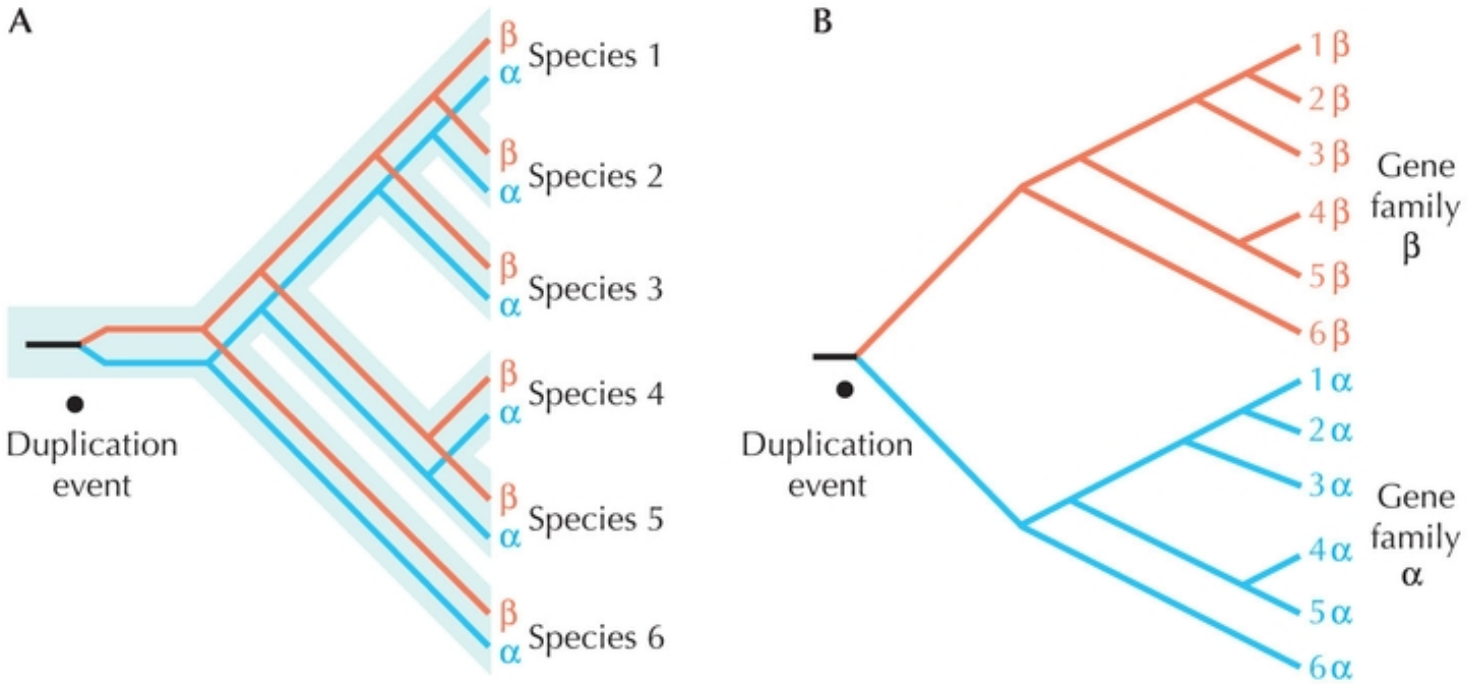
Venn/Euler plots should be avoided



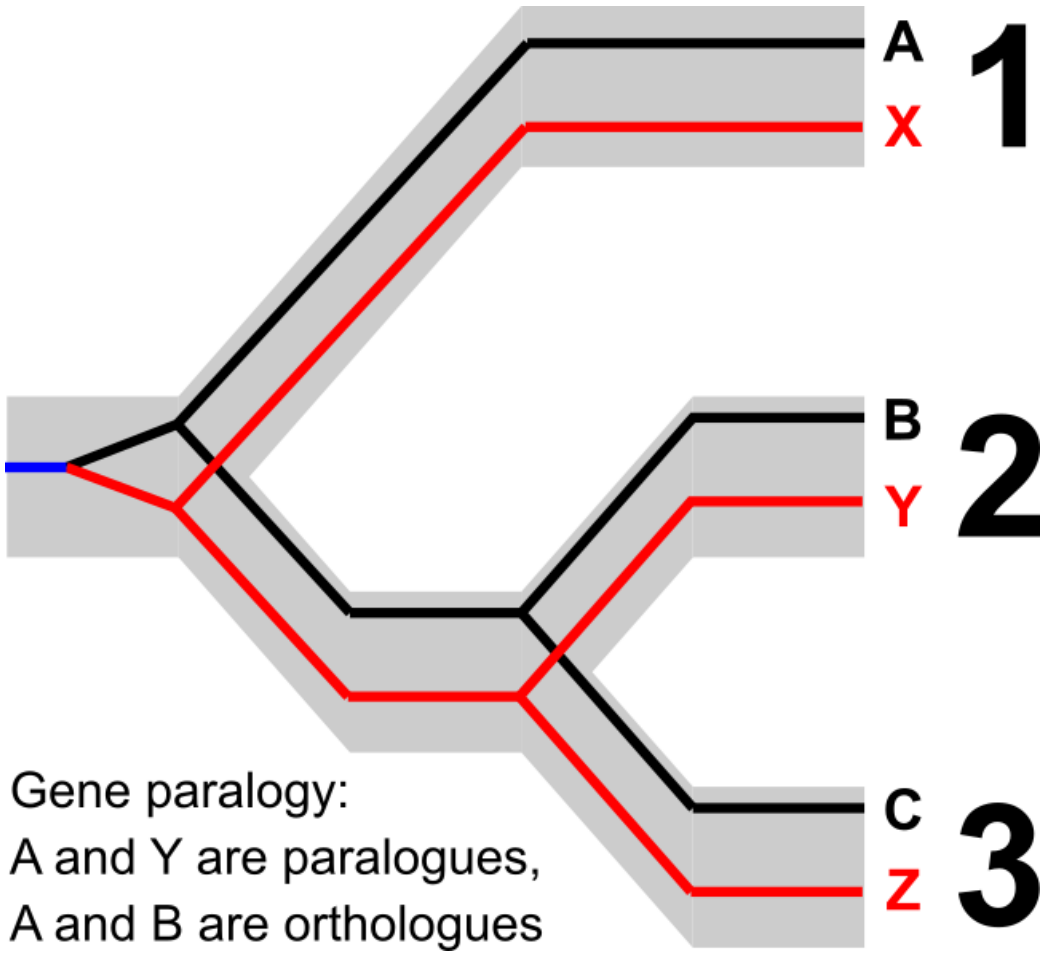
Why can't just use a single gene tree?



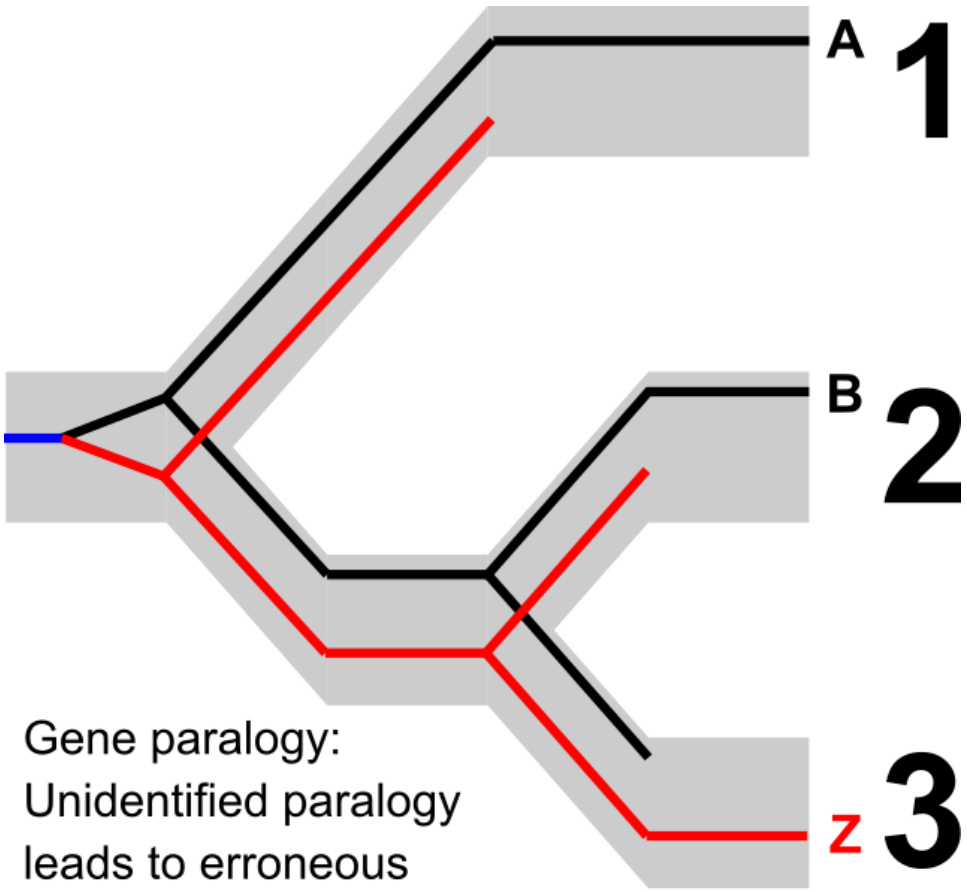
Paralogy


A
β
α
Species 1
β
α
Species 2
β
α
Species 3
β
α
Species 4
β
α
Species 5
β
α
Species 6
Duplication
event
B
1 β
2 β
3 β
4 β
5 β
6 β
Gene
family
β
1 α
2 α
3 α
4 α
5 α
6 α
Gene
family
α
Duplication
event

Paralogy

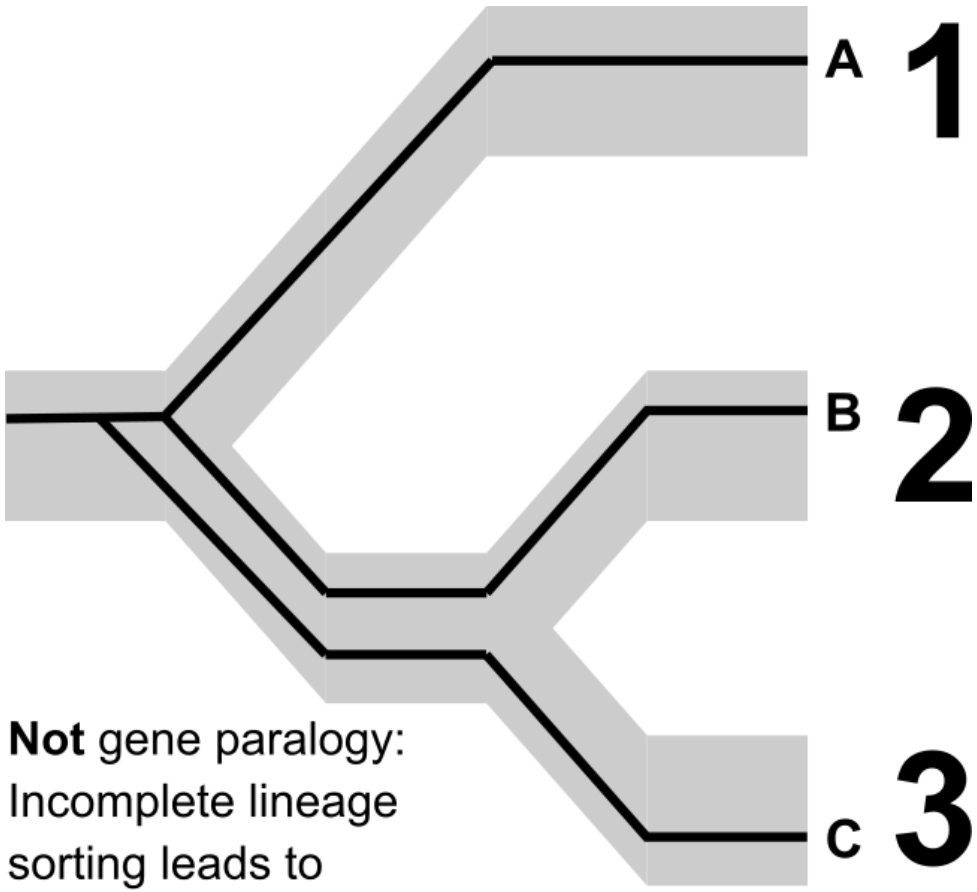

A
X
1
B
Y
2
C
Z
3
Gene paralogy:
A and Y are paralogues,
A and B are orthologues

Hidden Paralogy



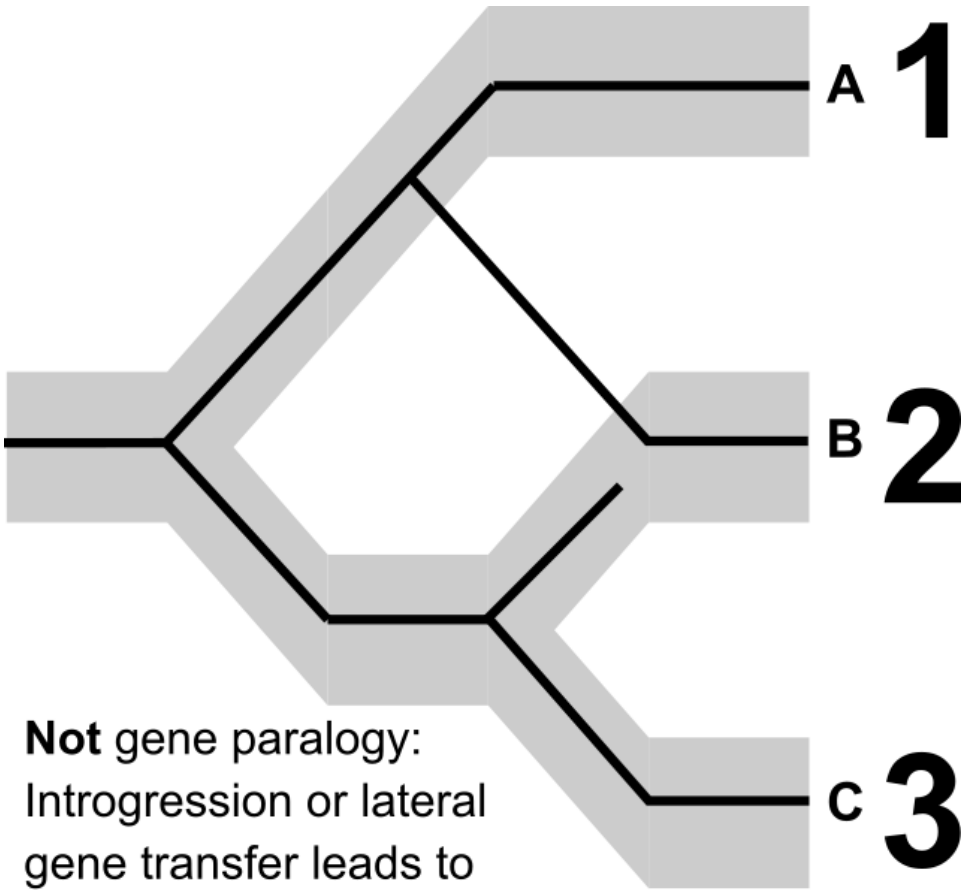
Gene paralogy:
Unidentified paralogy
leads to erroneous
inference of ((1,2),3)

Incomplete Lineage Sorting



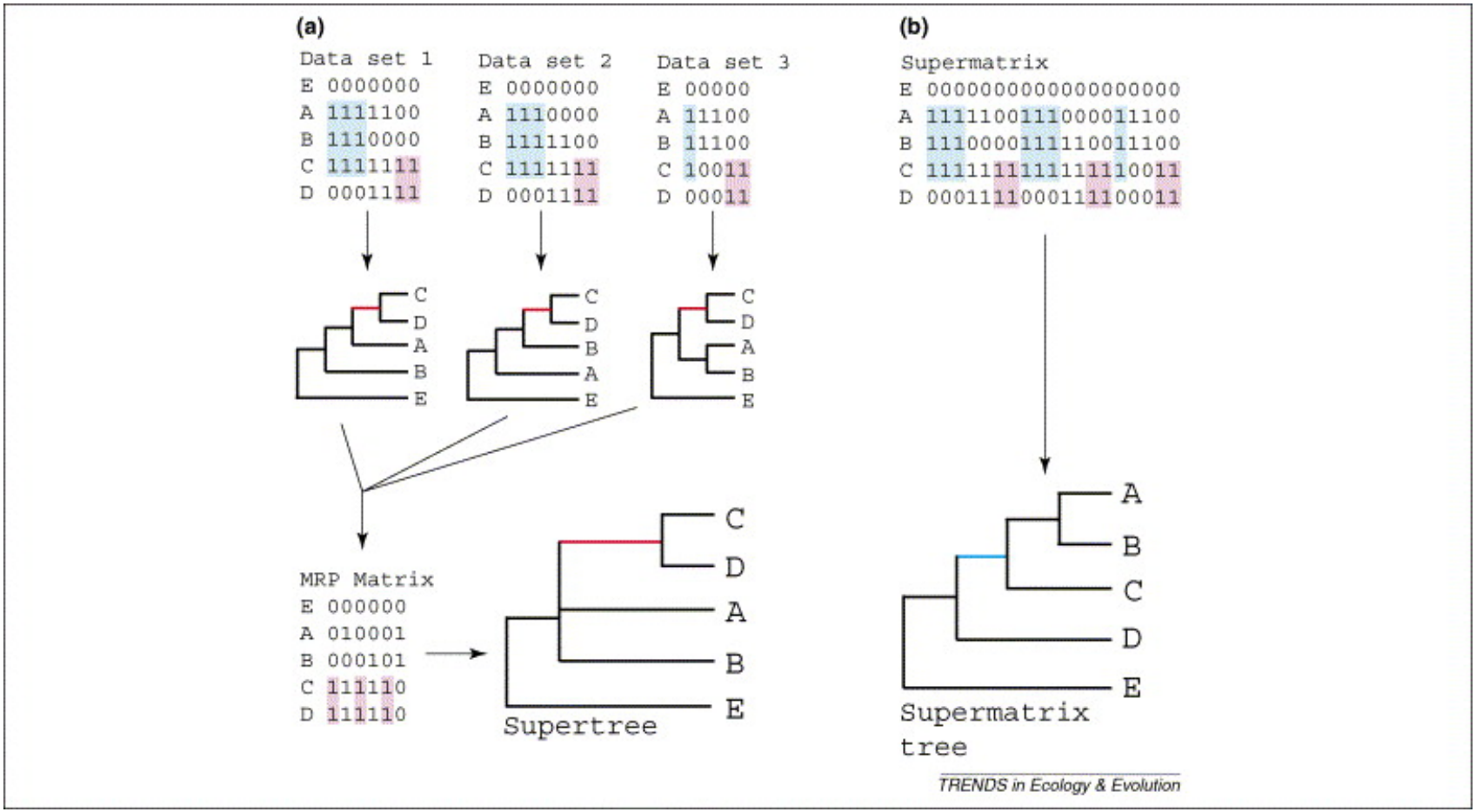
Not gene paralogy:
Incomplete lineage
sorting leads to
erroneous inference of ((1,2),3)

Lateral/Horizontal Gene Transfer

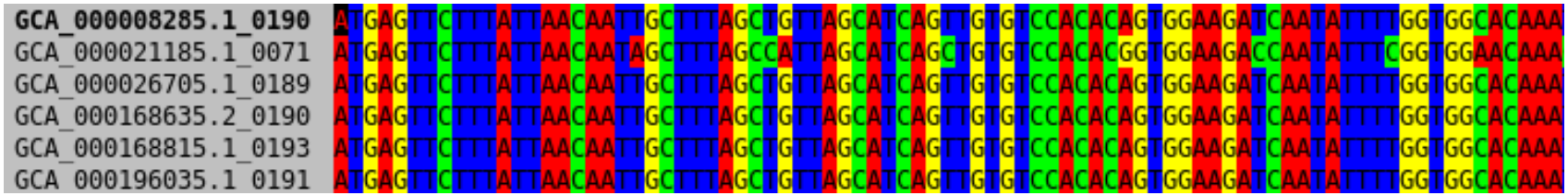


Not gene paralogy: Introgression or lateral gene transfer leads to erroneous inference of ((1,2),3)

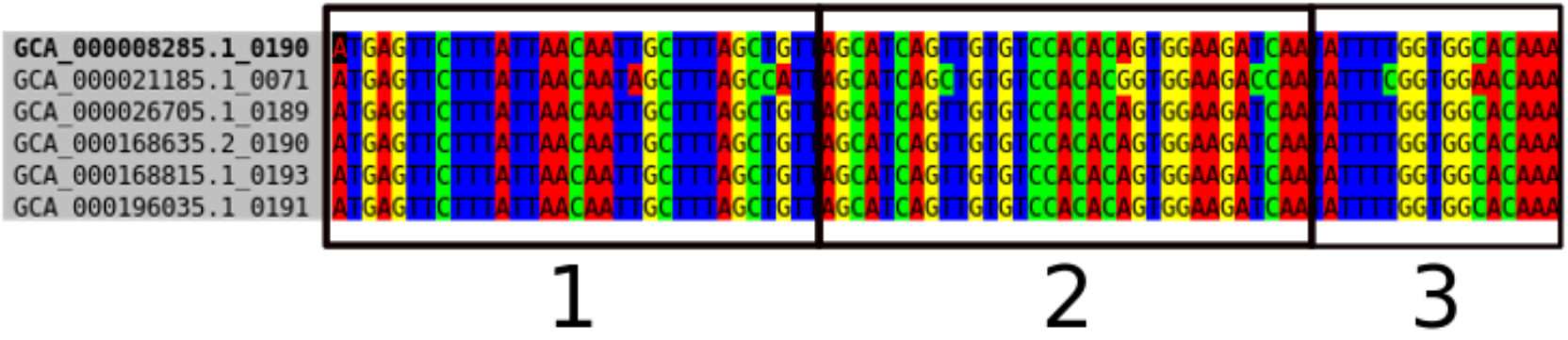
Supermatrix and SuperTrees



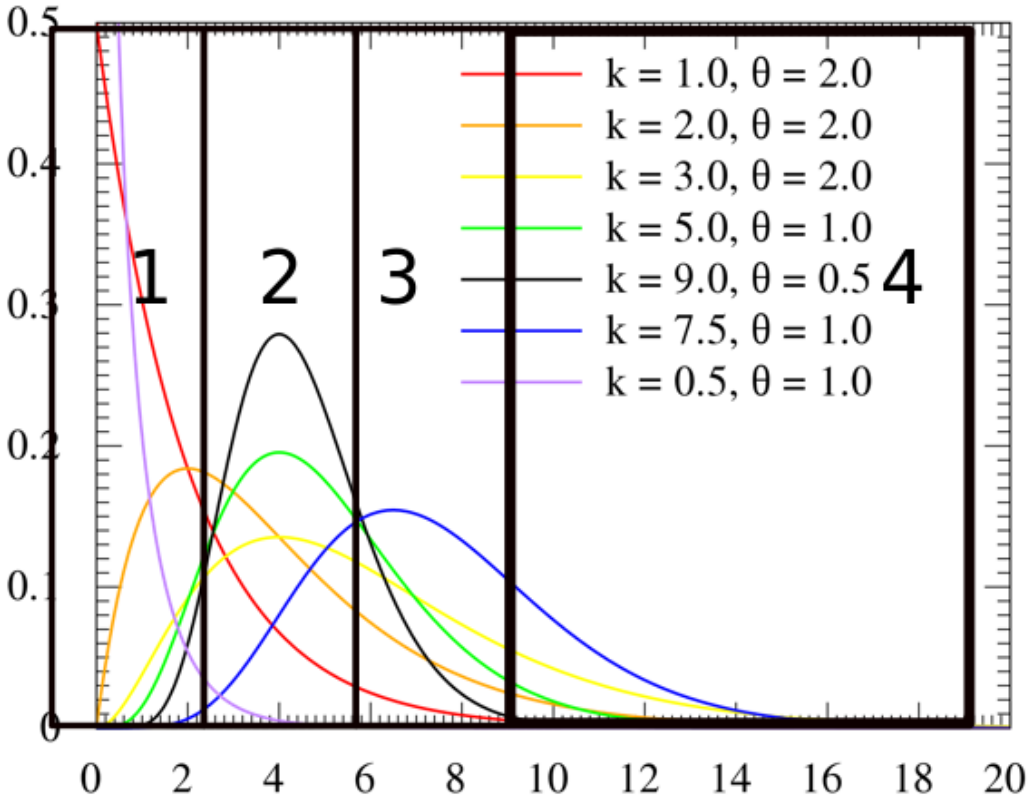
Supermatrix Evolution Models

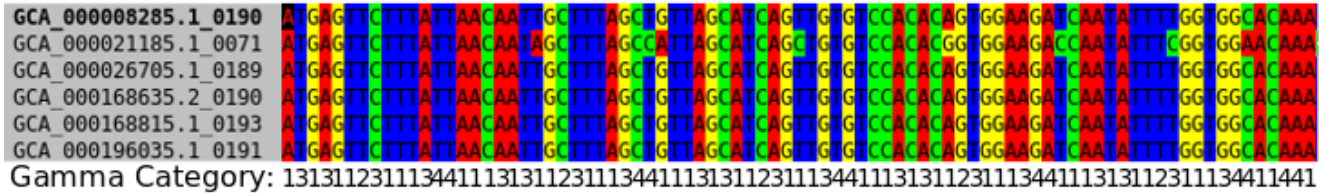


Gene Partitions

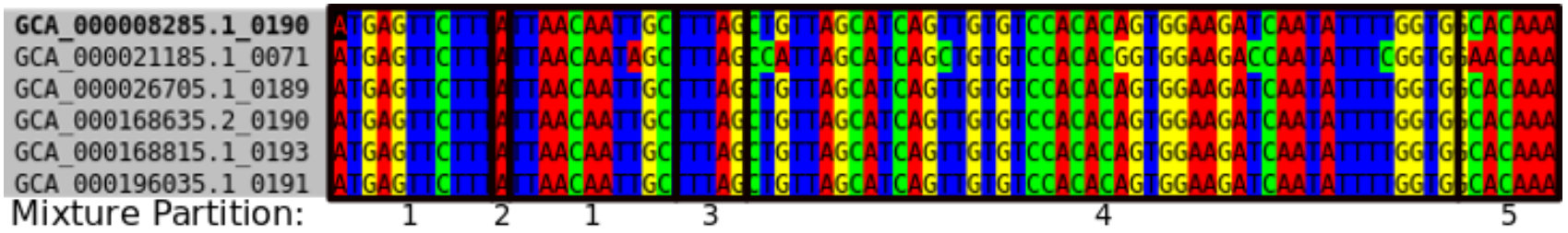


Gamma Rates


k = 1.0, θ = 2.0
k = 2.0, θ = 2.0
k = 3.0, θ = 2.0
k = 5.0, θ = 1.0
k = 9.0, θ = 0.5
k = 7.5, θ = 1.0
k = 0.5, θ = 1.0
1
2
3
4
0.5
0.4
0.3
0.2
0.1
0
0 2 4 6 8 10 12 14 16 18 20


GCA_000008285.1_0190
GCA_000021185.1_0071
GCA_000026705.1_0189
GCA_000168635.2_0190
GCA_000168815.1_0193
GCA_000196035.1_0191
Gamma Category: 1313112311134411131311231113441113131123111344111313112311134411131311231113441441

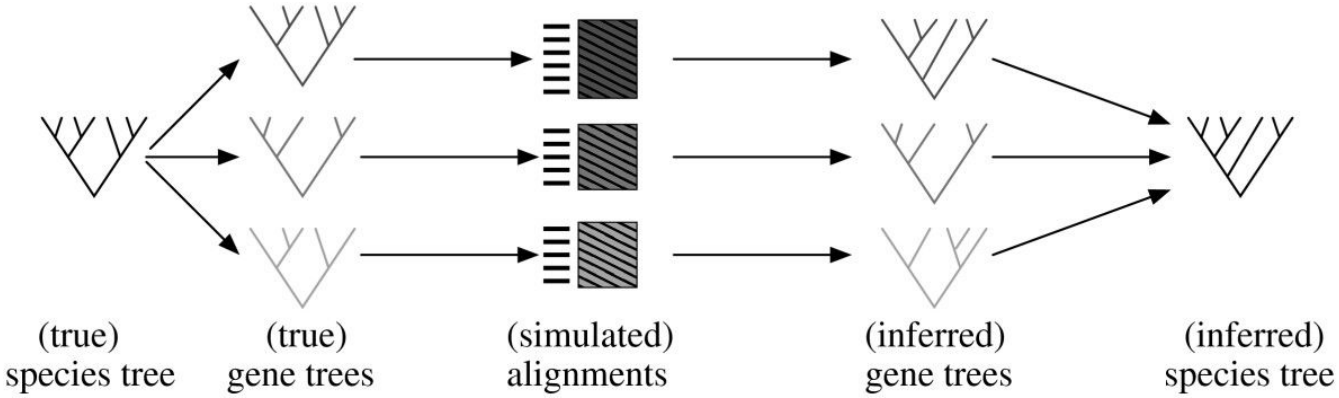
Mixture Models



Supermatrix

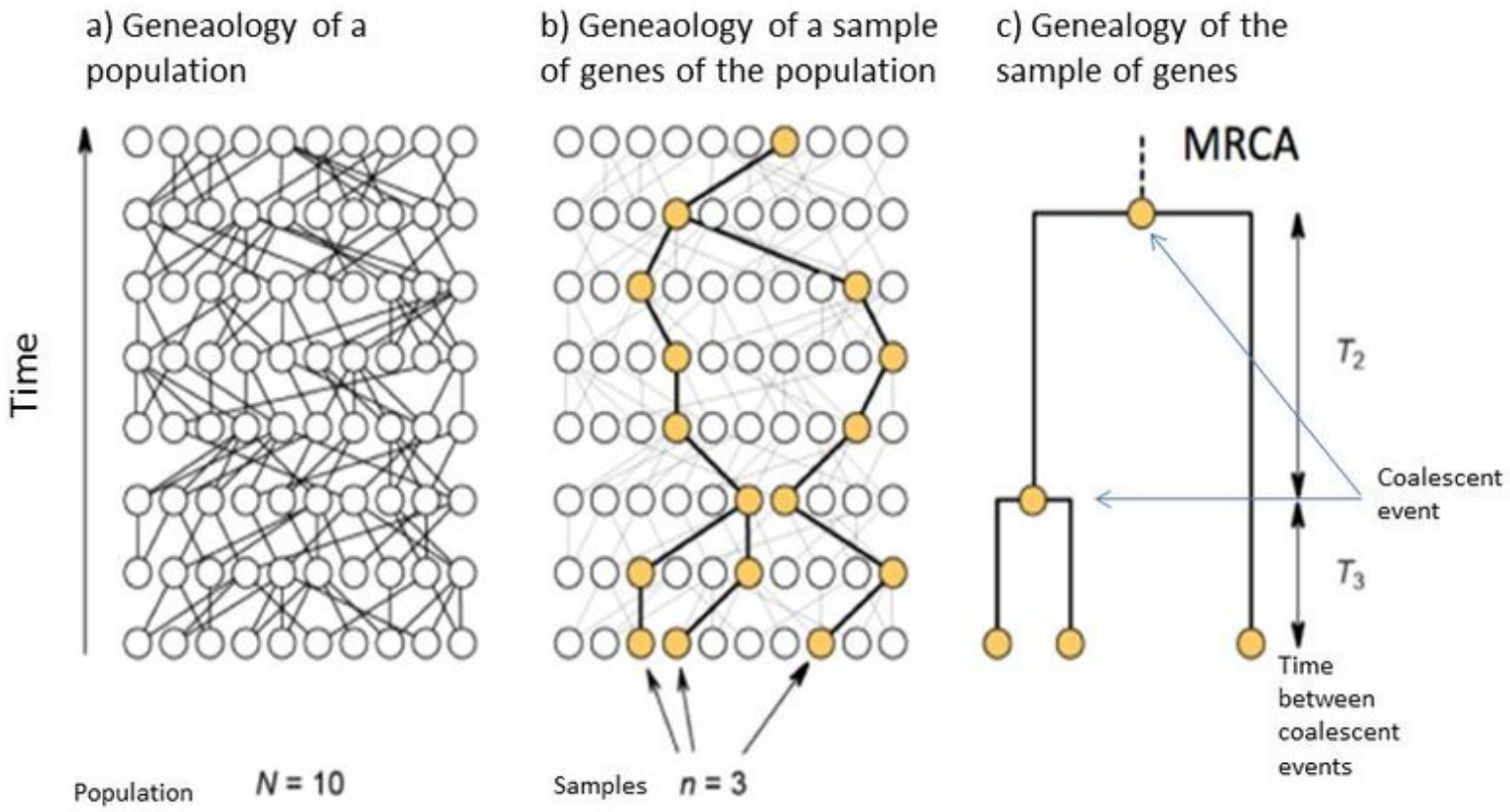
- Can be very computationally demanding
- Variable evolutionary rates/models combined (one-size fits all)
- CAT model does offer a solution but can overfit.
- More robust to random error when phylogenetic signal is consistent.

SuperTree



- Allows reconciliation of partial overlaps (i.e. not just core genome)
- Faster/more tractable
- Observed to have lower accuracy generally but more robust to incongruent signal (i.e. frequent HGT).

Coalescent Theory



Genomic Epidemiology
Phylogenetics

Identifying HGT with phylogenies

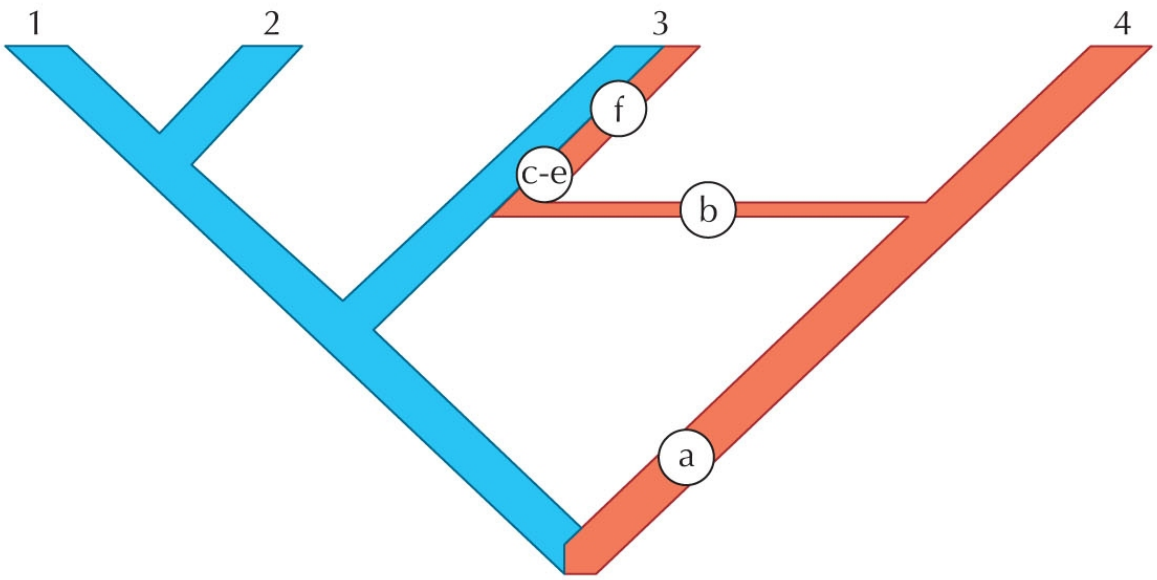
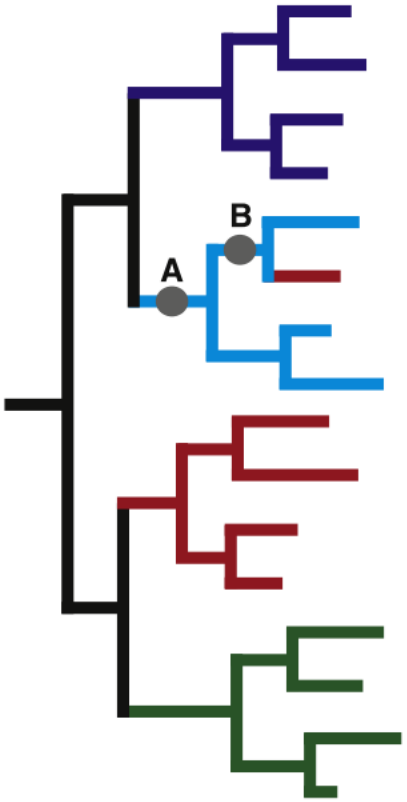


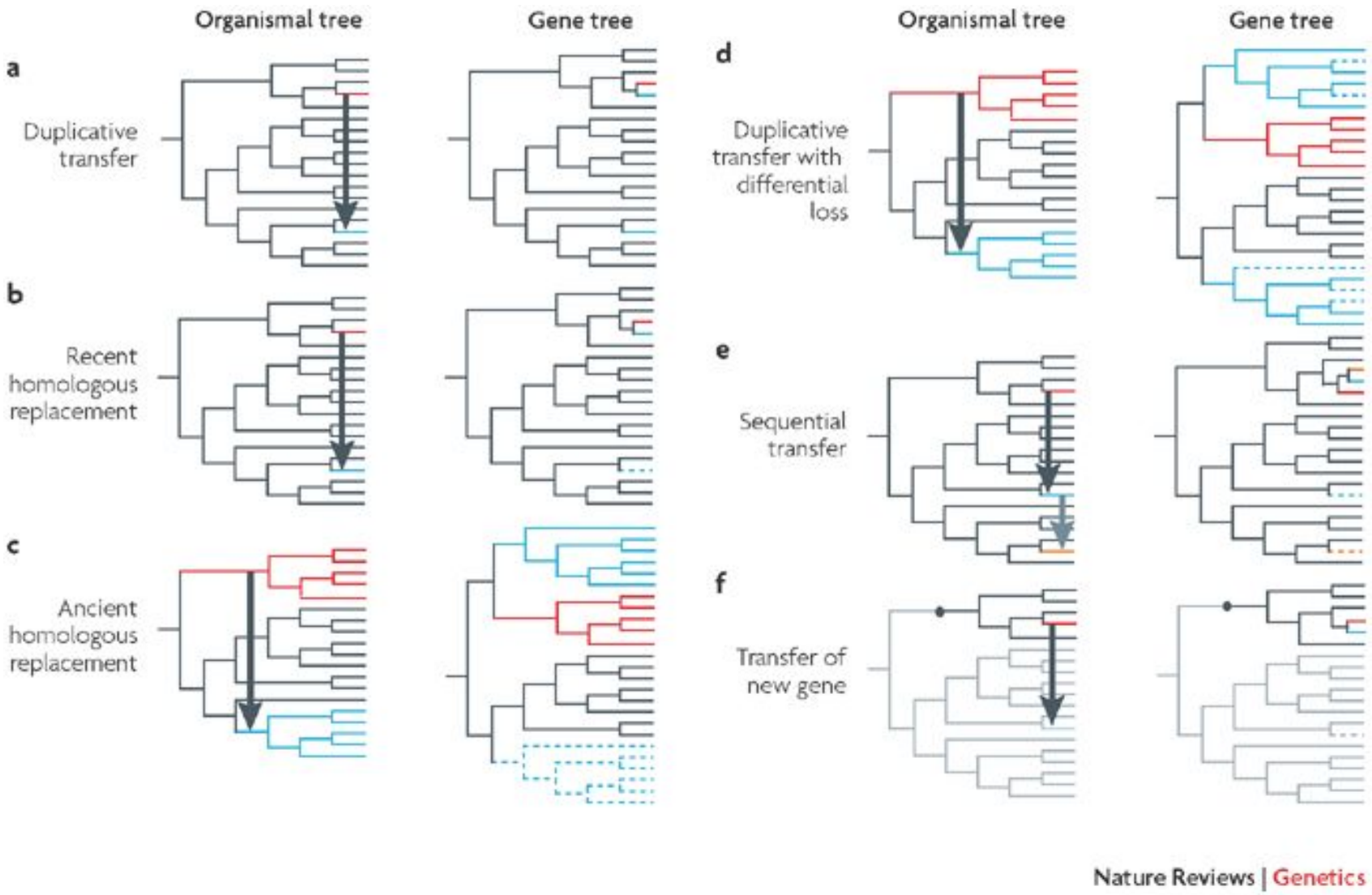
FIGURE 27.30. Stages in lateral gene transfer (LGT). The evolution of four species and one example of LGT are shown. Some key steps in LGT are labeled: (a) Divergence of genomes of different lineages; (b) movement of DNA from one lineage to another; (c) maintenance and replication of the foreign DNA; (d) possible positive selection for the foreign DNA; (e) spread into the new species' population; (f) amelioration. (Modified from Penny D. and Poole A. *Curr. Opin. Genet. Dev.* **9:** 672–677, © 1999 Elsevier.)

Schematic HGT Tree

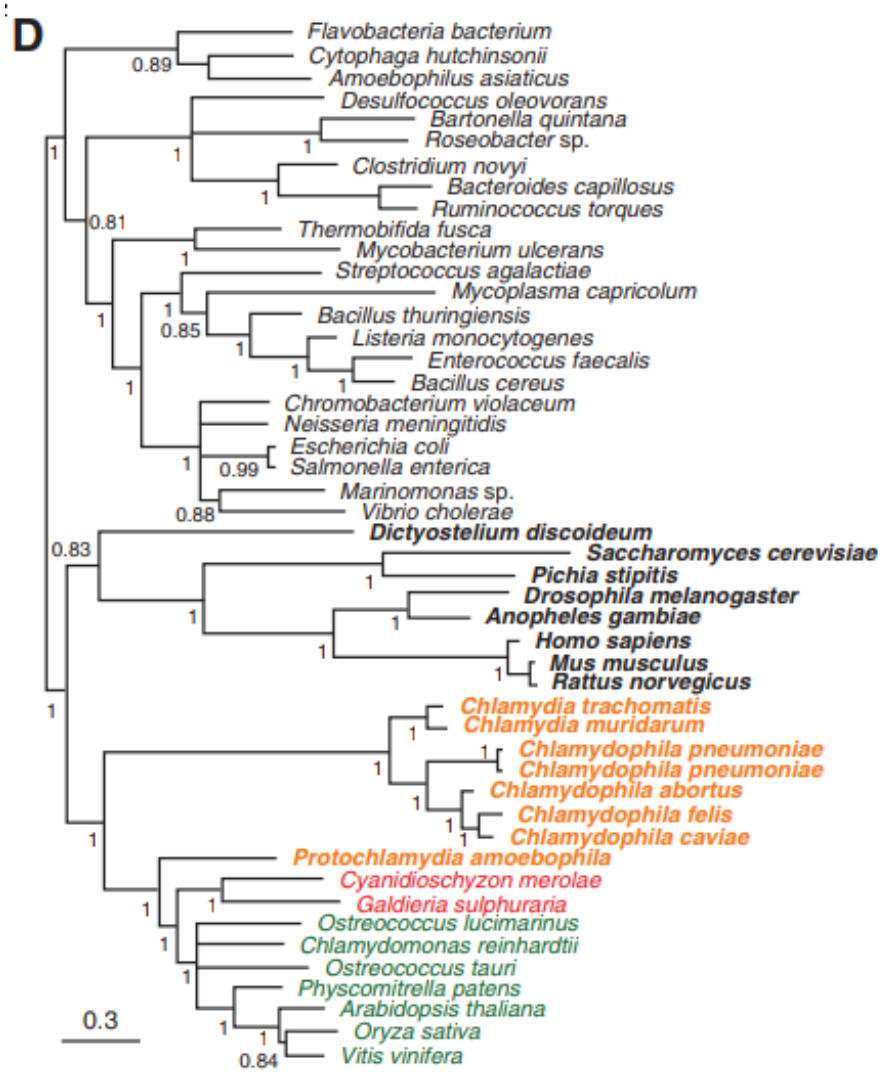


- (A) Need strong branch support for recipient branching with donor lineages
- (B) Need strong support for recipient branching within donor lineages

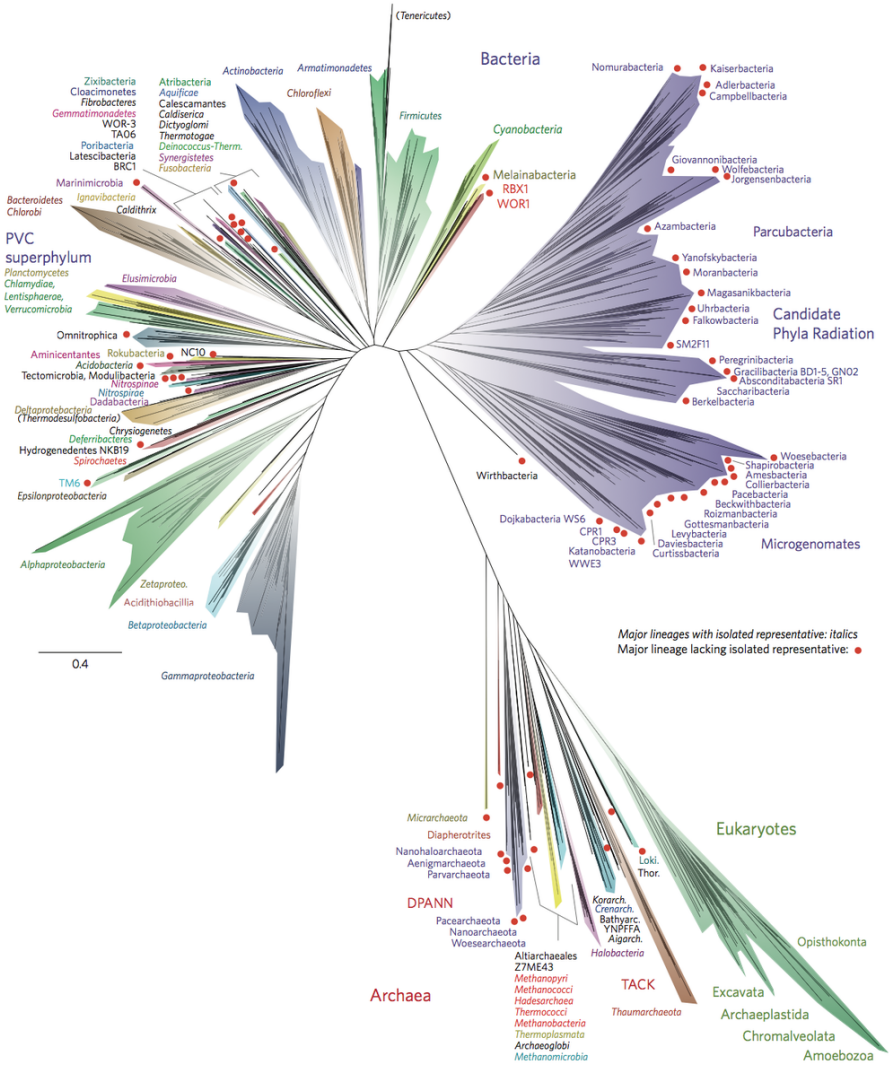
Different Types result in different topologies



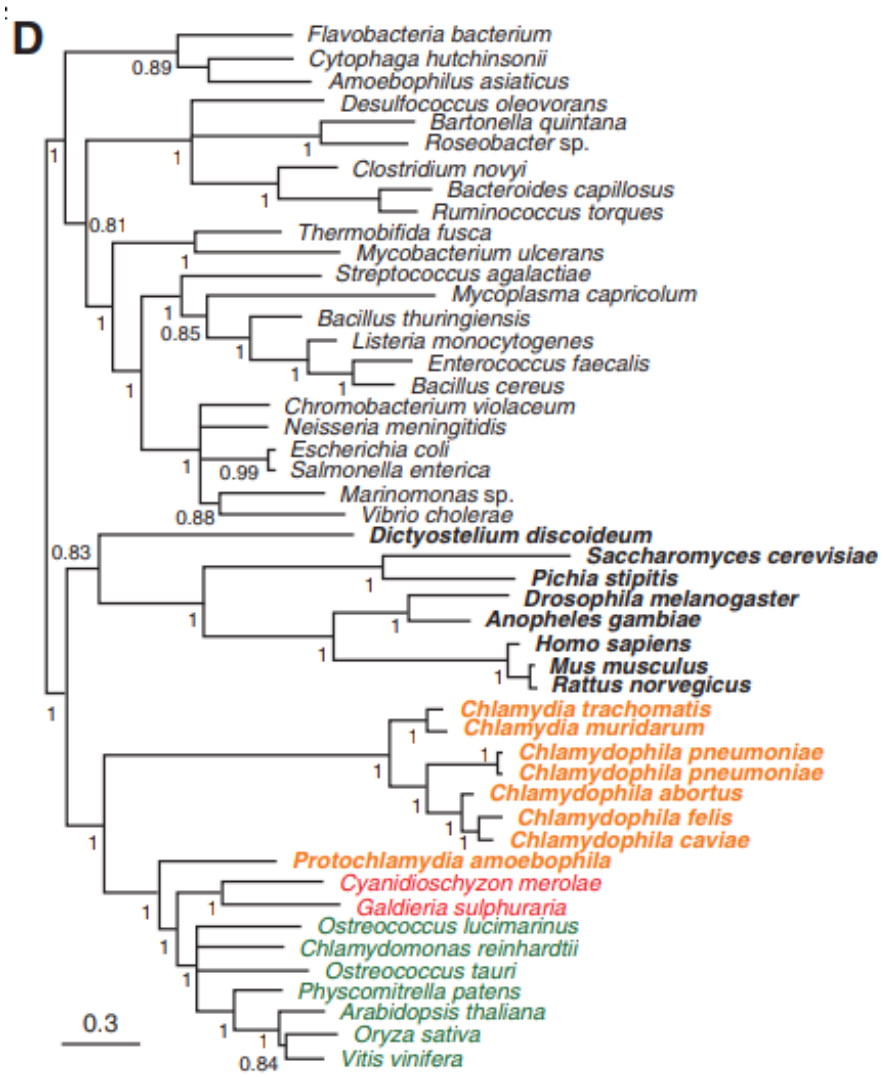
Real HGT


D
Flavobacteria bacterium
Cytophaga hutchinsonii
Amoebophilus asiaticus
Desulfococcus oleovorans
Bartonella quintana
Roseobacter sp.
Clostridium novyi
Bacteroides capillosus
Ruminococcus torques
Thermobifida fusca
Mycobacterium ulcerans
Streptococcus agalactiae
Mycoplasma capricolum
Bacillus thuringiensis
Listeria monocytogenes
Enterococcus faecalis
Bacillus cereus
Chromobacterium violaceum
Neisseria meningitidis
Escherichia coli
Salmonella enterica
Marinomonas sp.
Vibrio cholerae
Dictyostelium discoideum
Saccharomyces cerevisiae
Pichia stipitis
Drosophila melanogaster
Anopheles gambiae
Homo sapiens
Mus musculus
Rattus norvegicus
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila pneumoniae
Chlamydophila pneumoniae
Chlamydophila abortus
Chlamydophila felis
Chlamydophila caviae
Protochlamydia amoebophila
Cyanidioschyzon merolae
Galdieria sulphuraria
Ostreococcus lucimarinus
Chlamydomonas reinhardtii
Ostreococcus tauri
Physcomitrella patens
Arabidopsis thaliana
Oryza sativa
Vitis vinifera
0.3

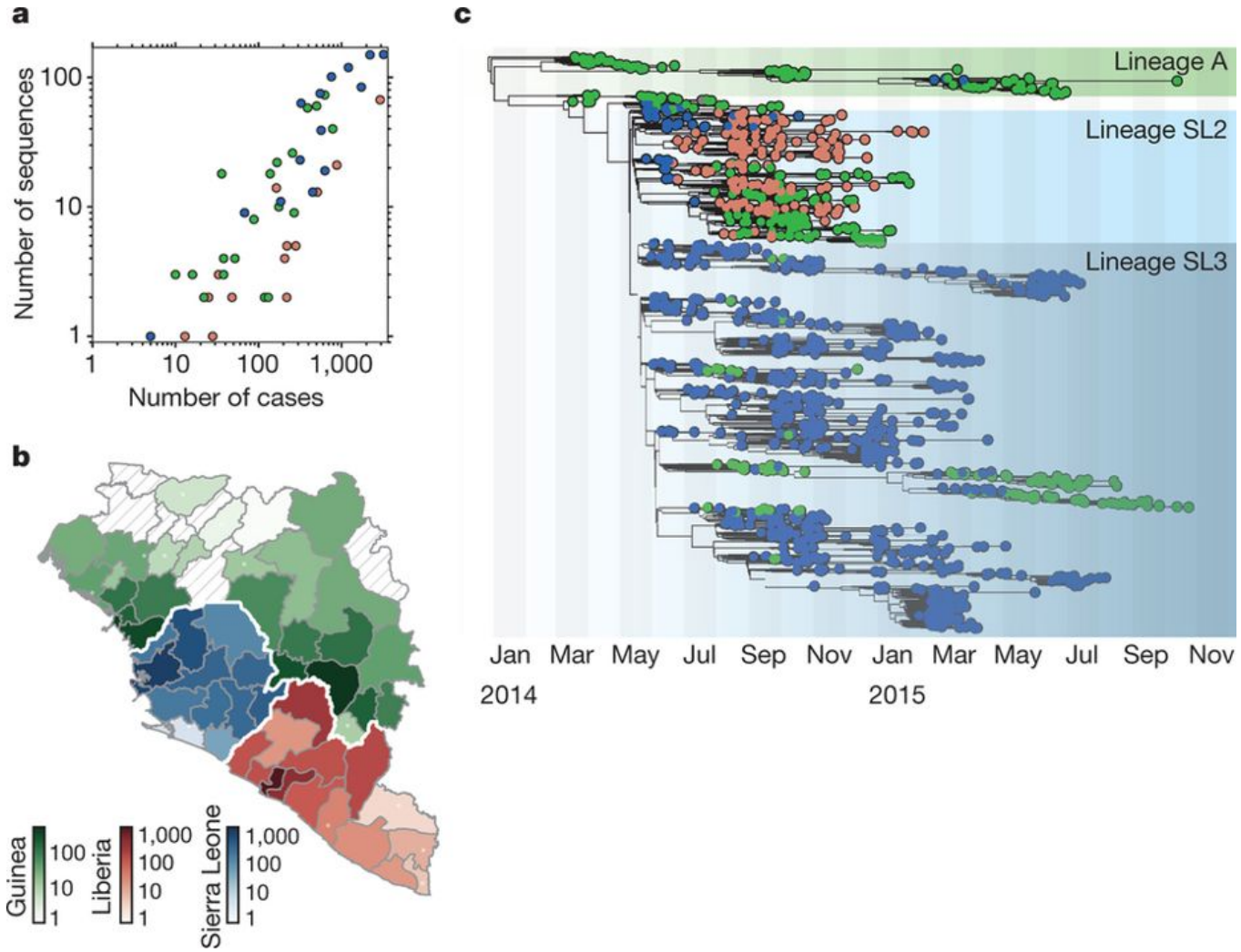
Real HGT


(Tenericutes)
Bacteria
Actinobacteria
Armatimonadetes
Chloroflexi
Firmicutes
Cyanobacteria
Melainabacteria
RBX1
WOR1
Nomurabacteria
Kaiserbacteria
Adlerbacteria
Campbellbacteria
Giovannonibacteria
Wolfebacteria
Jorgensenbacteria
Azambacteria
Parcubacteria
Yanofskybacteria
Moranbacteria
Magasanikbacteria
Uhrbacteria
Falkowbacteria
Candidate
Phyla Radiation
SM2F11
Peregrinibacteria
Gracilibacteria BD1-5, GN02
Absconditabacteria SR1
Saccharibacteria
Berkelbacteria
Zixibacteria
Cloacimonetes
Fibrobacteres
Gemmatimonadetes
WOR-3
TA06
Poribacteria
Latescibacteria
BRC1
Atribacteria
Aquificae
Calescamantes
Caldiserica
Dictyoglomi
Thermotogae
Deinococcus-Therm.
Synergistetes
Fusobacteria
Marinimicrobia
Ignavibacteria
Calditrix
Bacteroidetes
Chlorobi
PVC
superphylum
Planctomycetes
Chlamydiae,
Lentisphaerae,
Verrucomicrobia
Elusimicrobia
Omnitrophica
Aminicentantes
Rokubacteria
NC10
Acidobacteria
Tectomicrobia, Modulibacteria
Nitrospinae
Nitrospirae
Dadabacteria
Deltaprotebacteria
(Thermodesulfobacteria)
Chrysiogenetes
Deferribacteres
Hydrogenedentes NKB19
Spirochaetes
TM6
Epsilonproteobacteria
Alphaproteobacteria
Zetaproteo.
Acidithiohacillia
Betaproteobacteria
Gammaproteobacteria
0.4
Wirthbacteria
Dojkabacteria WS6
CPR1
CPR3
Katanobacteria
WWE3
Woesebacteria
Shapirobacteria
Amesbacteria
Collierbacteria
Pacebacteria
Beckwithbacteria
Roizmanbacteria
Gottesmanbacteria
Levybacteria
Daviesbacteria
Curtissbacteria
Microgenomates
Major lineages with isolated representative: italics
Major lineage lacking isolated representative:
Micrarchaeota
Diapherotrites
Nanohaloarchaeota
Aenigmarchaeota
Parvarchaeota
DPANN
Pacearchaeota
Nanoarchaeota
Woesearchaeota
Altiarchaeales
Z7ME43
Methanopyri
Methanococci
Hadesarchaea
Thermococci
Methanobacteria
Thermoplasmata
Archaeoglobi
Methanomicrobia
Archaea
Korarch.
Crenarch.
Bathyarc.
YNPFFA
Aigarch.
Halobacteria
TACK
Thaumarchaeota
Loki.
Thor.
Eukaryotes
Opisthokonta
Excavata
Archaeplastida
Chromalveolata
Amoebozoa

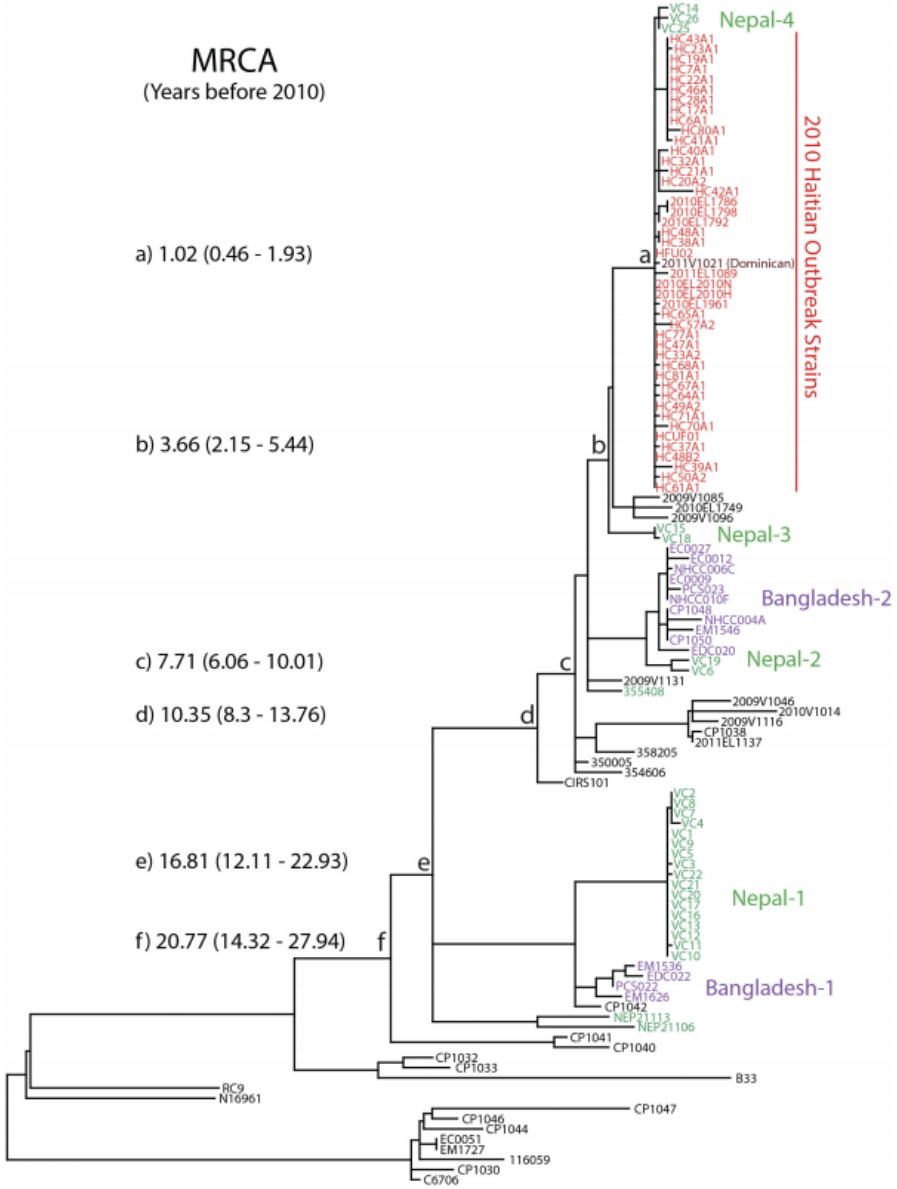
Real HGT


D
Flavobacteria bacterium
Cytophaga hutchinsonii
Amoebophilus asiaticus
Desulfococcus oleovorans
Bartonella quintana
Roseobacter sp.
Clostridium novyi
Bacteroides capillosus
Ruminococcus torques
Thermobifida fusca
Mycobacterium ulcerans
Streptococcus agalactiae
Mycoplasma capricolum
Bacillus thuringiensis
Listeria monocytogenes
Enterococcus faecalis
Bacillus cereus
Chromobacterium violaceum
Neisseria meningitidis
Escherichia coli
Salmonella enterica
Marinomonas sp.
Vibrio cholerae
Dictyostelium discoideum
Saccharomyces cerevisiae
Pichia stipitis
Drosophila melanogaster
Anopheles gambiae
Homo sapiens
Mus musculus
Rattus norvegicus
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila pneumoniae
Chlamydophila pneumoniae
Chlamydophila abortus
Chlamydophila felis
Chlamydophila caviae
Protochlamydia amoebophila
Cyanidioschyzon merolae
Galdieria sulphuraria
Ostreococcus lucimarinus
Chlamydomonas reinhardtii
Ostreococcus tauri
Physcomitrella patens
Arabidopsis thaliana
Oryza sativa
Vitis vinifera
0.3

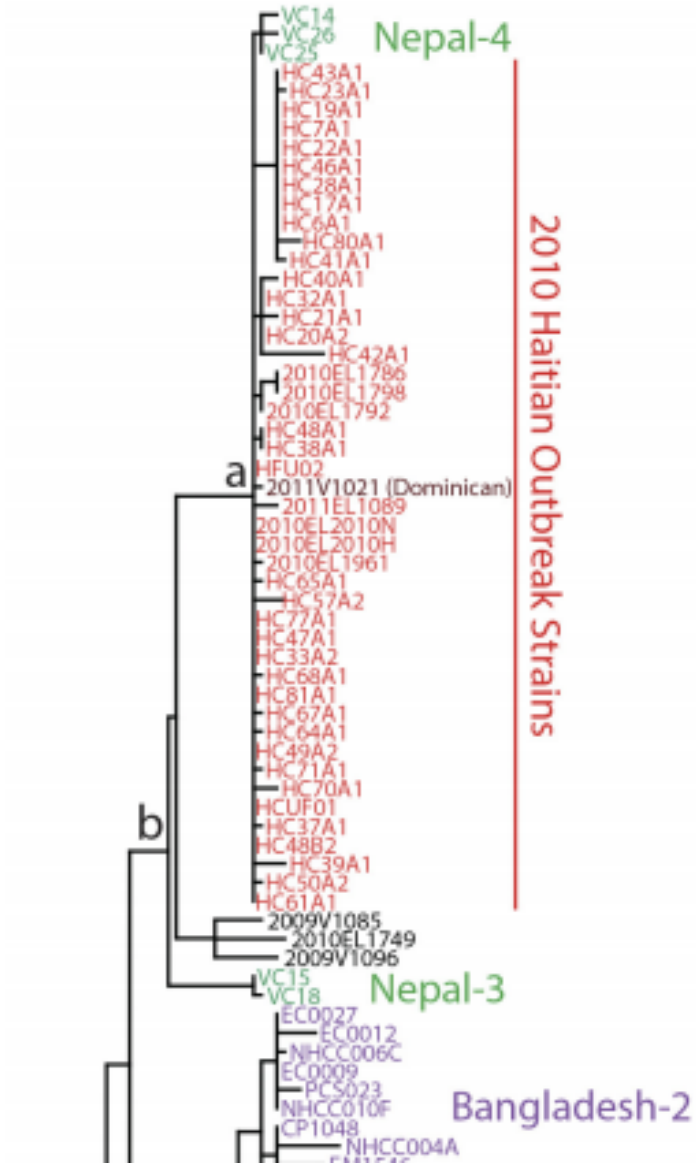
West African Ebola Epidemic



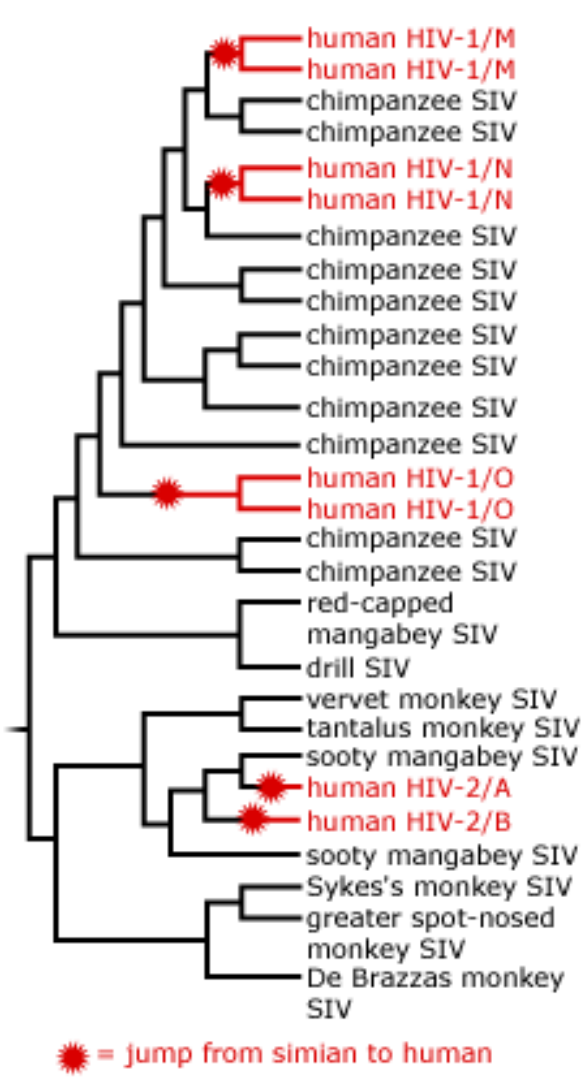
Haitian Cholera Outbreak: Identifying the origin



Haitian Cholera Outbreak: Identifying the origin



Finding zoonoses



Conclusion

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.
- Bootstrapping is a slow, biased but conservative way to estimate the support for a given branch in your tree.

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.
- Bootstrapping is a slow, biased but conservative way to estimate the support for a given branch in your tree.
- Comparing trees directly is non-trivial due to tree-space.

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.
- Bootstrapping is a slow, biased but conservative way to estimate the support for a given branch in your tree.
- Comparing trees directly is non-trivial due to tree-space.
- Supermatrix and supertree approaches allow reconciliation of data from multiple genes.

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.
- Bootstrapping is a slow, biased but conservative way to estimate the support for a given branch in your tree.
- Comparing trees directly is non-trivial due to tree-space.
- Supermatrix and supertree approaches allow reconciliation of data from multiple genes.
- Incongruence between the species tree and the gene tree might be evidence for HGT.

Summary

- Phylogenies are hypothesis and their inference includes assumptions that need testing.
- Bootstrapping is a slow, biased but conservative way to estimate the support for a given branch in your tree.
- Comparing trees directly is non-trivial due to tree-space.
- Supermatrix and supertree approaches allow reconciliation of data from multiple genes.
- Incongruence between the species tree and the gene tree might be evidence for HGT.
- Phylogenies can be used to trace outbreak origins and parameters.

Questions?