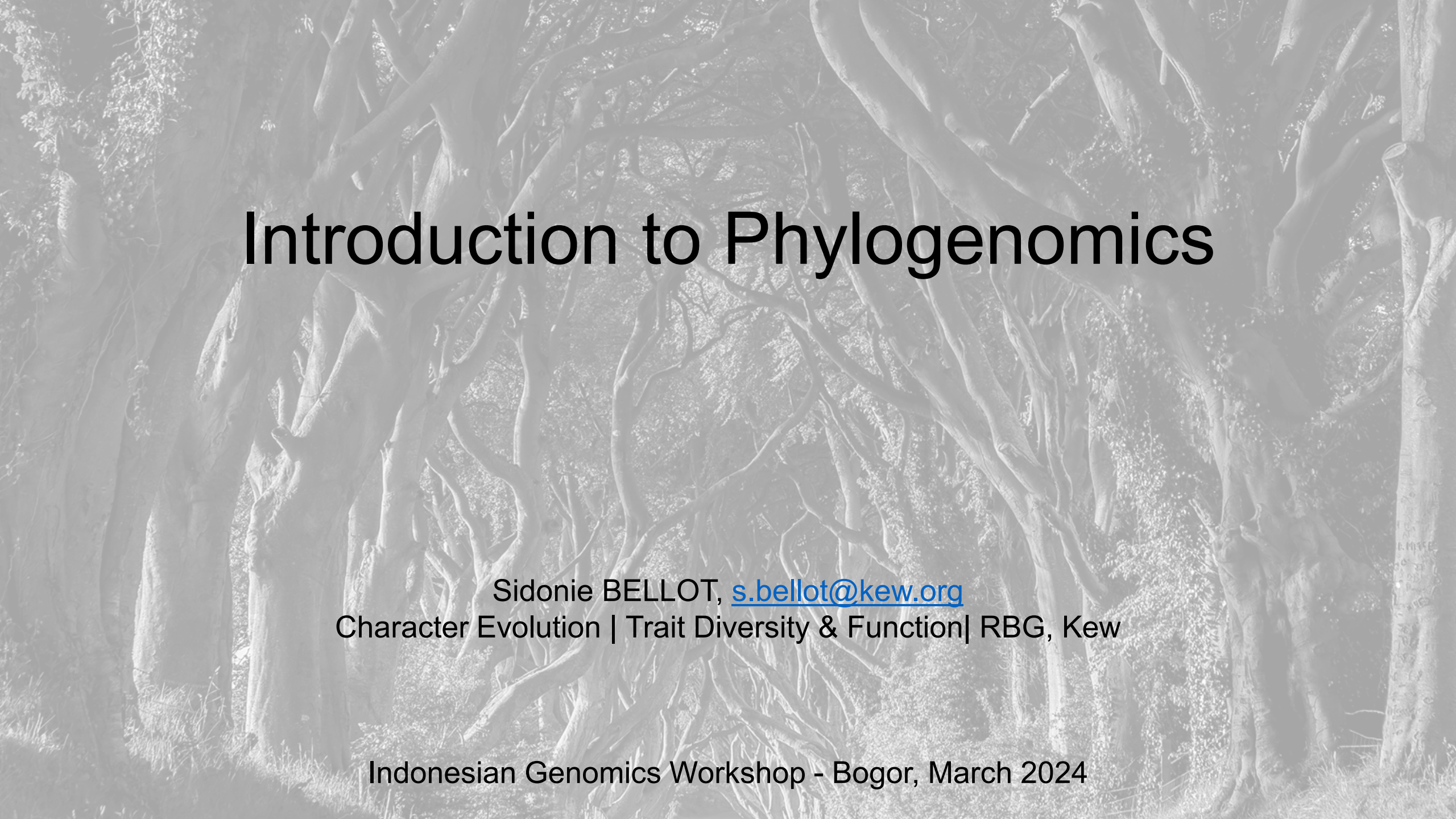


Introduction to Phylogenomics

Sidonie BELLOT, s.bellot@kew.org
Character Evolution | Trait Diversity & Function| RBG, Kew

Indonesian Genomics Workshop - Bogor, March 2024

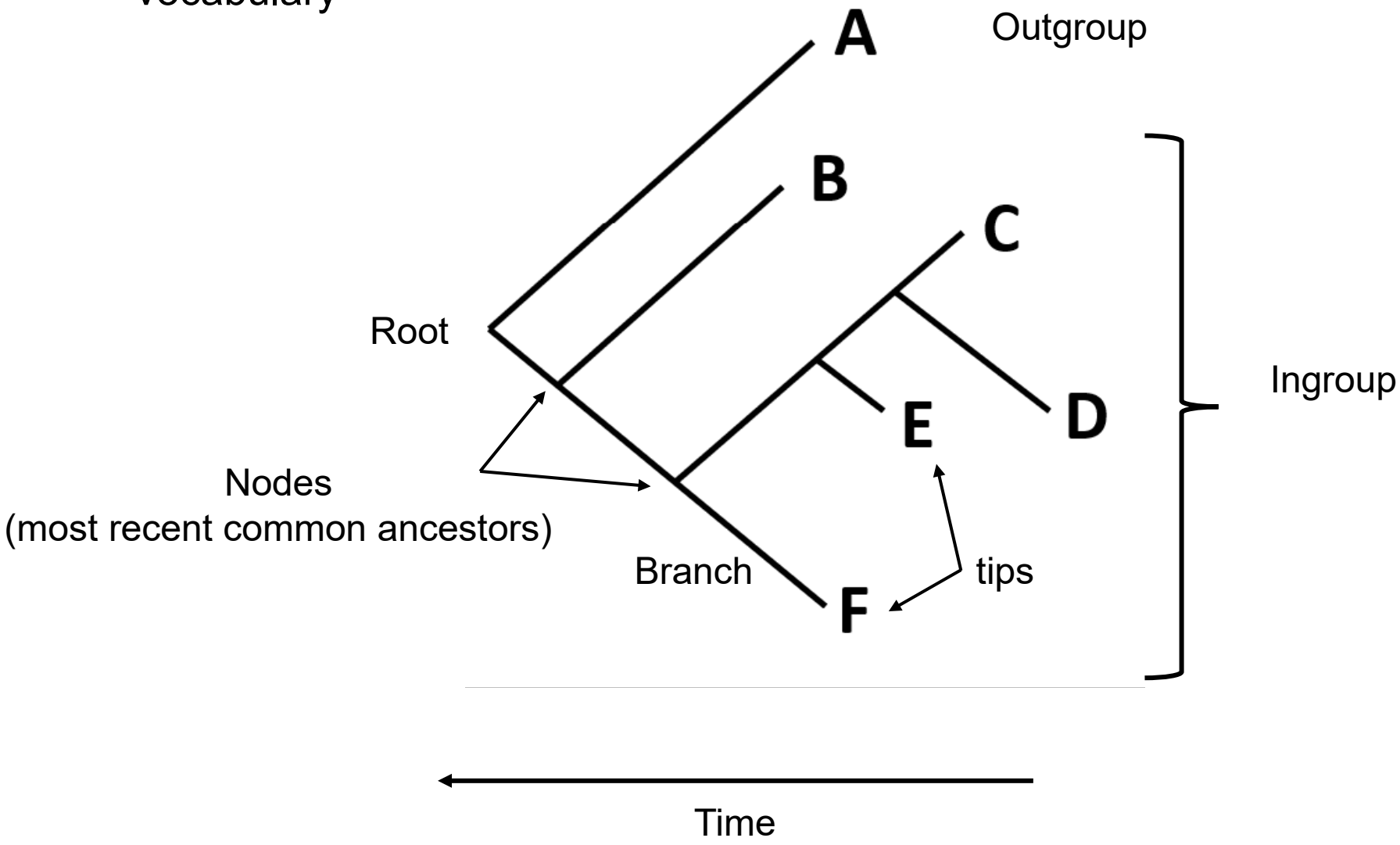
Outline

1. Phylogenetics
2. Phylogenomics
3. Approaches
4. Confidence

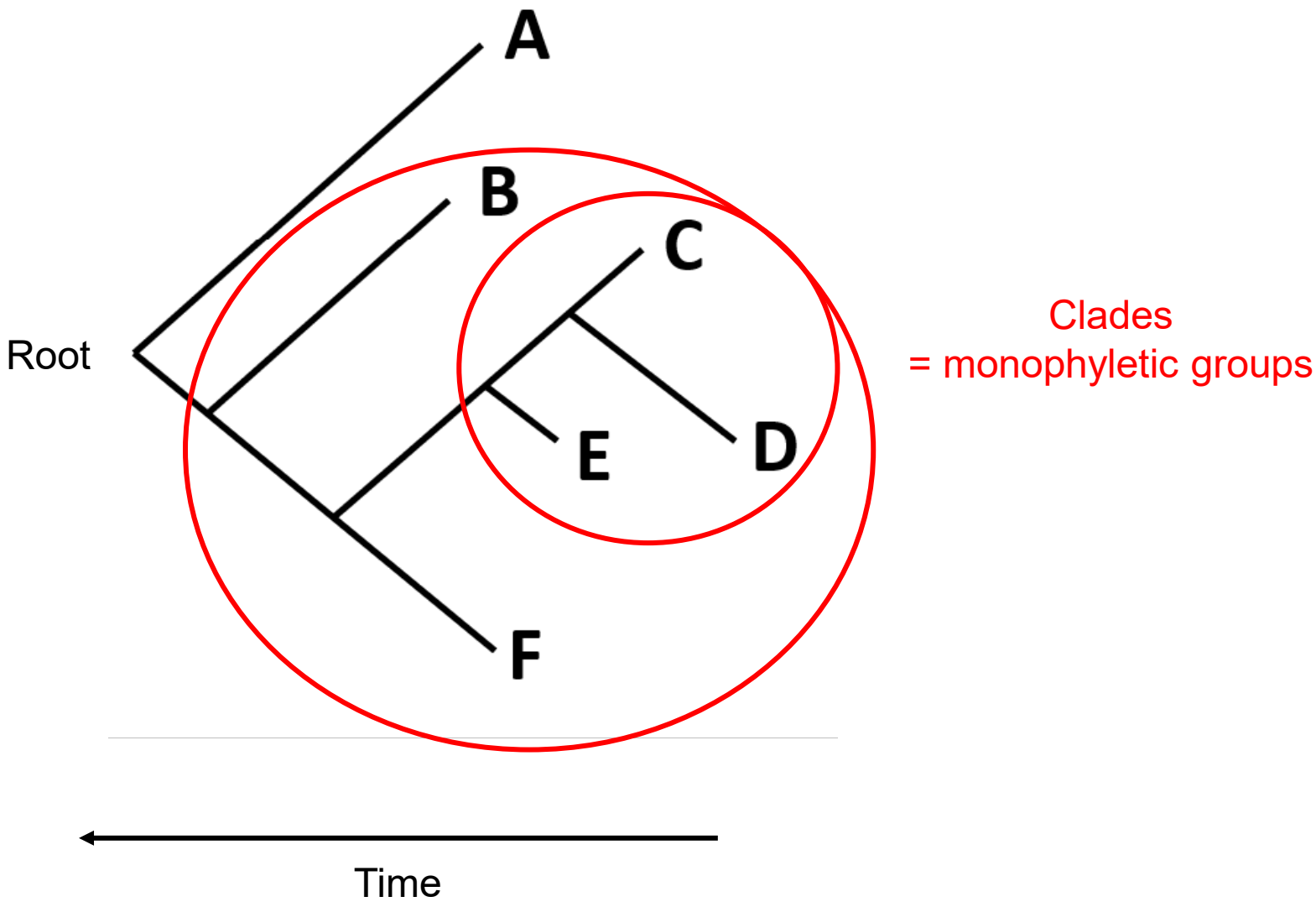
Outline

1. Phylogenetics
2. Phylogenomics
3. Approaches
4. Confidence

Phylogenetic tree vocabulary



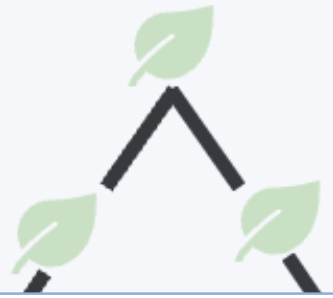
Phylogenetic tree vocabulary



Clades
= monophyletic groups

Basic principles reminder…

Features are transmitted from ancestors to descendants

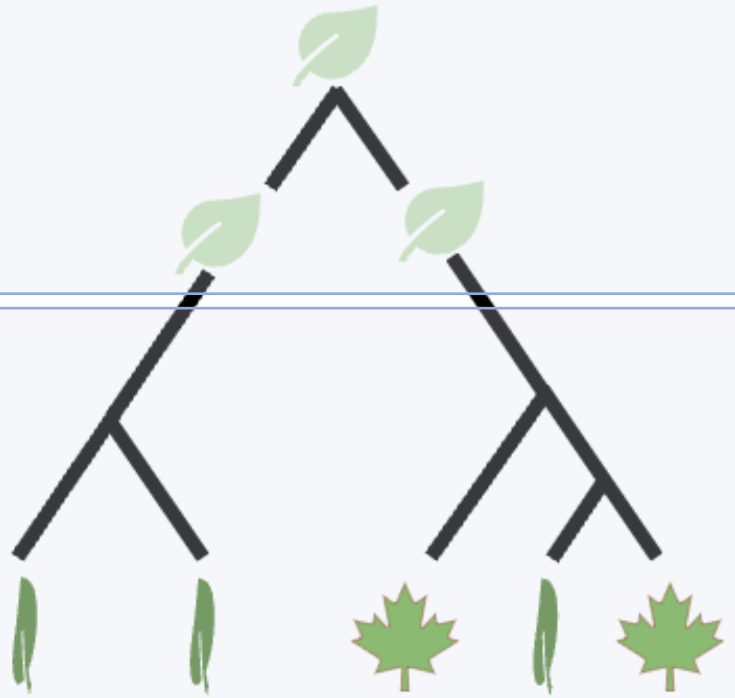


→ Homology:
similarity due to ancestry

Basic principles reminder…

Features are transmitted from ancestors to descendants

→ Homology:
similarity due to ancestry



Features change through generations (descent with modification, Darwin)

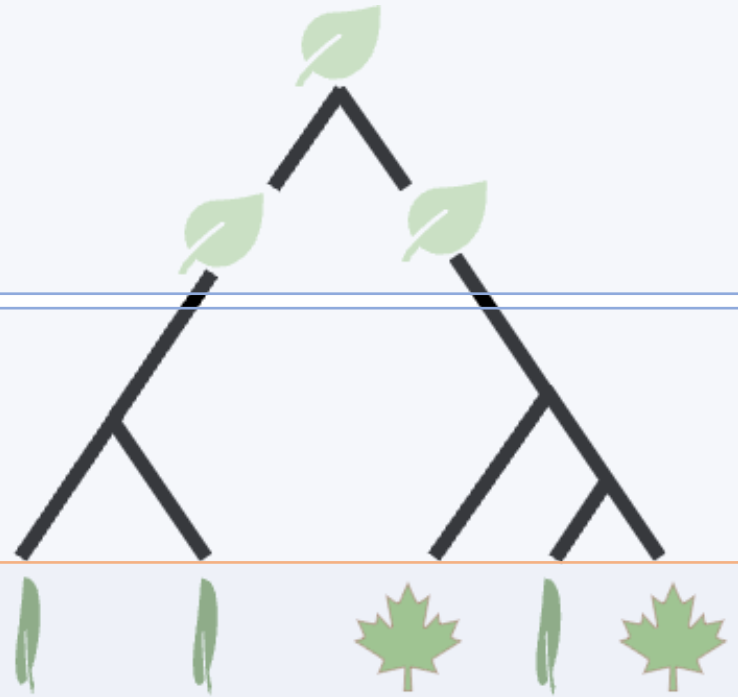
→ Closer similarity:
more recent ancestry

Phylogenetic inferences are based on homology

Basic principles reminder…

Features are transmitted from ancestors to descendants

→ Homology:
similarity due to ancestry



Features change through generations (descent with modification, Darwin)

→ Closer similarity:
more recent ancestry

Some similarities are not inherited from a common ancestor!

→ Homoplasy:
similarity not due to ancestry

Phylogenetic inferences are based on homology and can be mislead by homoplasy

Basic principles reminder…

Features are transmitted from ancestors to descendants

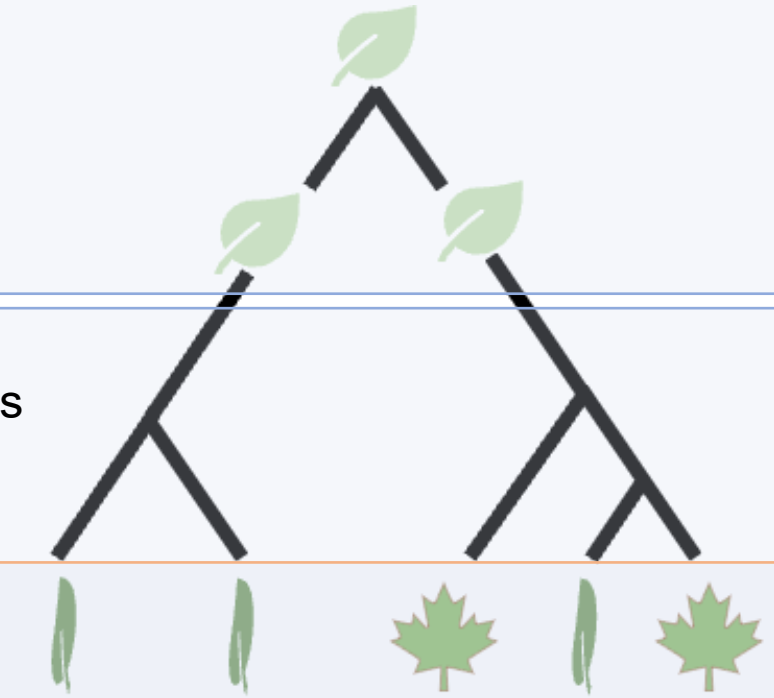
→ Homology: similarity due to ancestry

Features change through generations (descent with modification, Darwin)

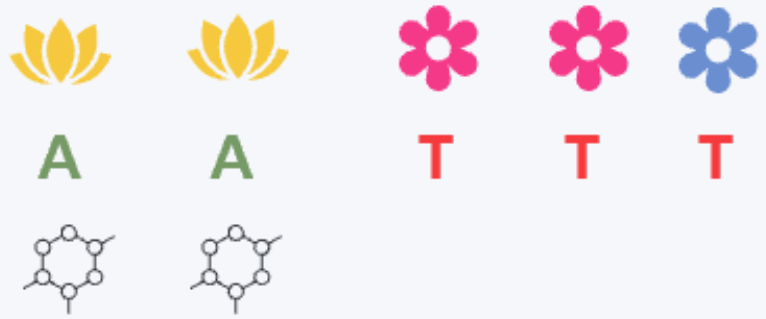
→ Closer similarity: more recent ancestry

Some similarities are not inherited from a common ancestor!

→ Homoplasy: similarity not due to ancestry



Looking at many characters helps to distinguish between homologies and homoplasies



Character	States
Leaf shape	Entire/Divided
Petal number	5/6
Site 3 of rbcL	A/T/G/C/-
Cucurbitacin E	Present/Absent

Phylogenetic inferences are based on homology and can be mislead by homoplasy

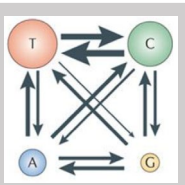
In DNA, one nucleotide site = one character → many characters!

Sequence DNA

```
                          11111111112222222222233333
          Site            12345678012345678901234567890123
          Species A       AAGTTTAAAATGTATCGGCGGCTTAAACGTGTA

                          11111111112222222222333
          Site            12345678012345678901234567890123
          Species C       CGTTTAAAATGTATCGGCGGCTAAAACGTGTA

                          111111111122222222
          Site            12345678012345678901234567
          Species B       AAGTTTAAAATGGATGGGCGGTTTAA

                          11111111112222222222233333
          Site            12345678012345678901234567890123
          Species D       GTTTTAAAATGGATGGGCGGTTAAAACGTGTAGG
```

Aligning sequences allows to find informative characters

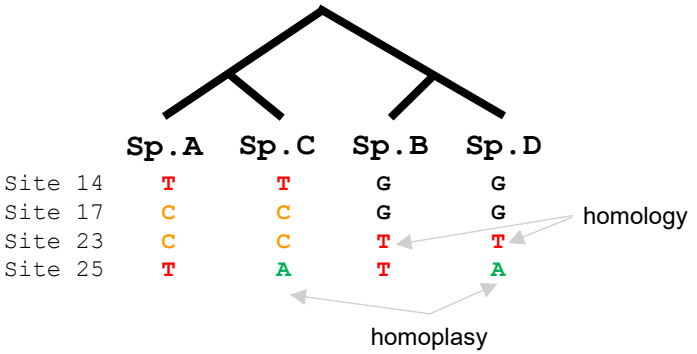
```
                    1111111111222222222233333333
Site                12345678012345678901234567890123456
Species A           AAGTTTAAAATGTATCGGCGGCTTAAACGTGTA--
Species B           AAGTTTAAAATGGATGGGCGGTTTAA---------
Species C           CG-TTTAAAATGTATCGGCGGCTAAAACGTGTA--
Species D           -GTTTTAAAATGGATGGGCGGTTAAAACGTGTAGG
```



Model of nucleotide substitution

Combining the alignment with knowledge about DNA evolution

Species phylogeny based on this gene (aka a gene tree)



How to go from similarities to relationships?

Methods of phylogenetic inference (1st semester + practical)

Character-based methods

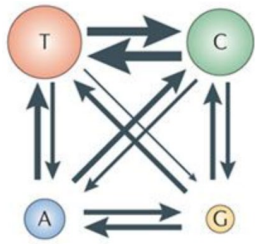
1. Optimize branch lengths of alternative topologies based on evolutionary model
2. Select the highest scoring tree based on an optimality criterion

Character matrix

	Taxon					
Character	A	B	C	D	E	F
Site 1	C	C	C	C	C	T
Site 2	A	A	C	C	C	C
Site 3	G	G	G	T	T	T
Site 4...	A	A	A	T	T	A

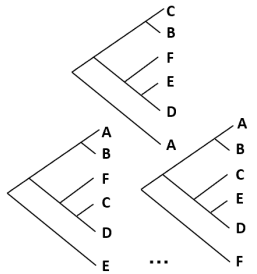
+

Evolutionary model



+

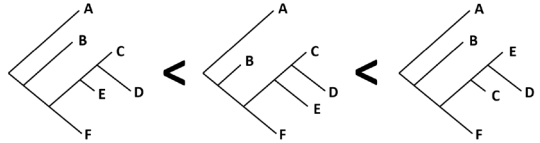
Tree space



Optimize trees following the evolutionary model

Score and rank trees according to optimality criterion

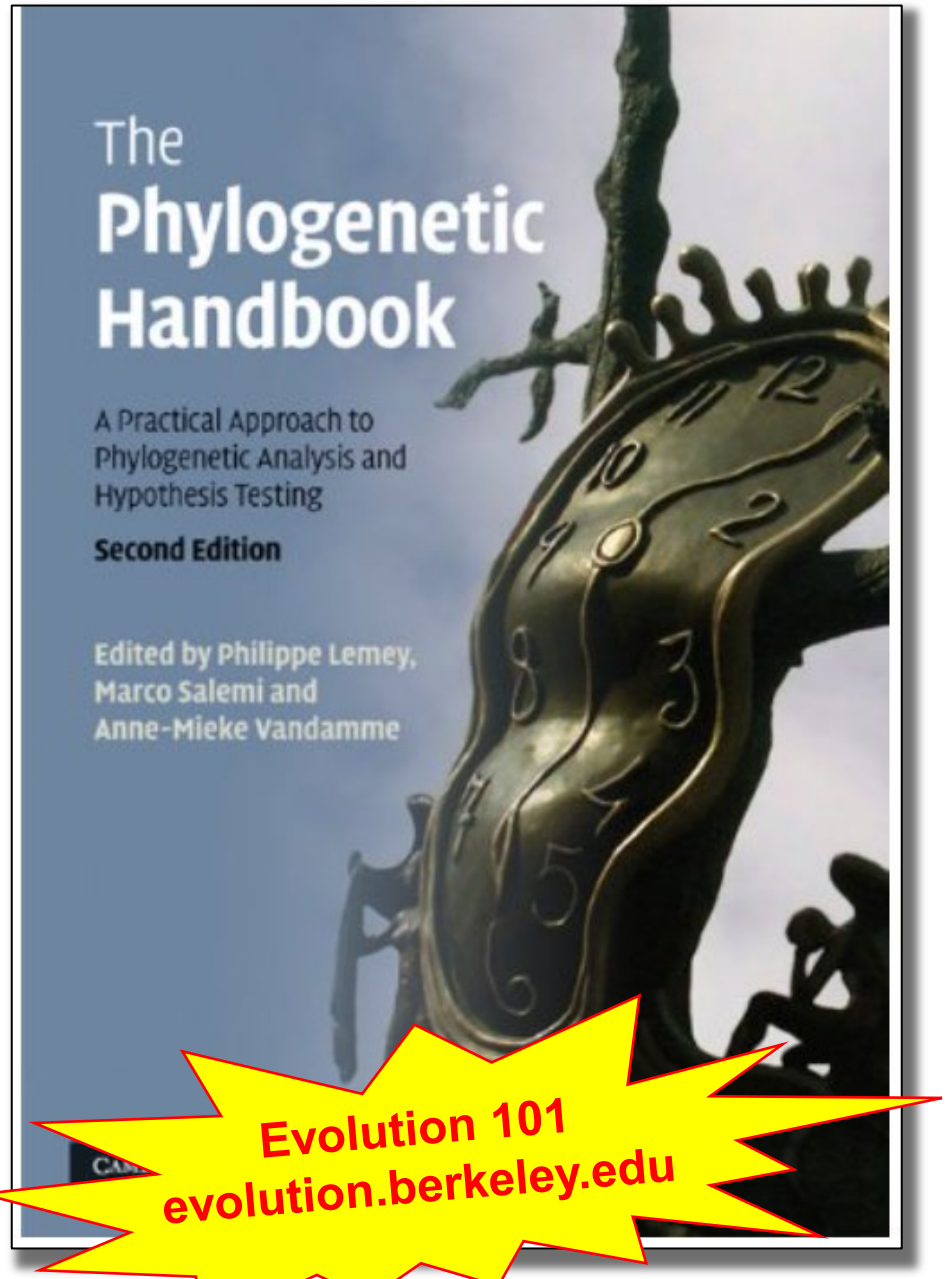
Selected tree*



Methods & optimality criteria:

- **Maximum Parsimony** – Criterion: minimum number of character changes (length of the tree)
- **Maximum Likelihood** – Criterion: maximum log-likelihood of the tree
- **Bayesian Inference** – Criterion: maximum posterior probability of the tree

*The trees shown here are random, they are not the trees that you should obtain using the example data



Evolution 101
evolution.berkeley.edu

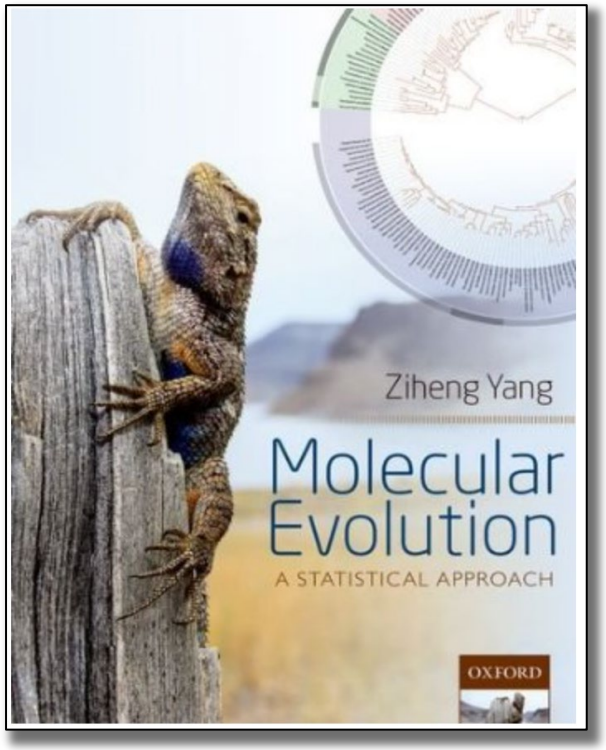
NATURE REVIEWS | GENETICS
VOLUME 13 | MAY 2012 | 303

Molecular phylogenetics: principles and practice

Ziheng Yang[1,2] and Bruce Rannala[1,3]

THE ROYAL SOCIETY

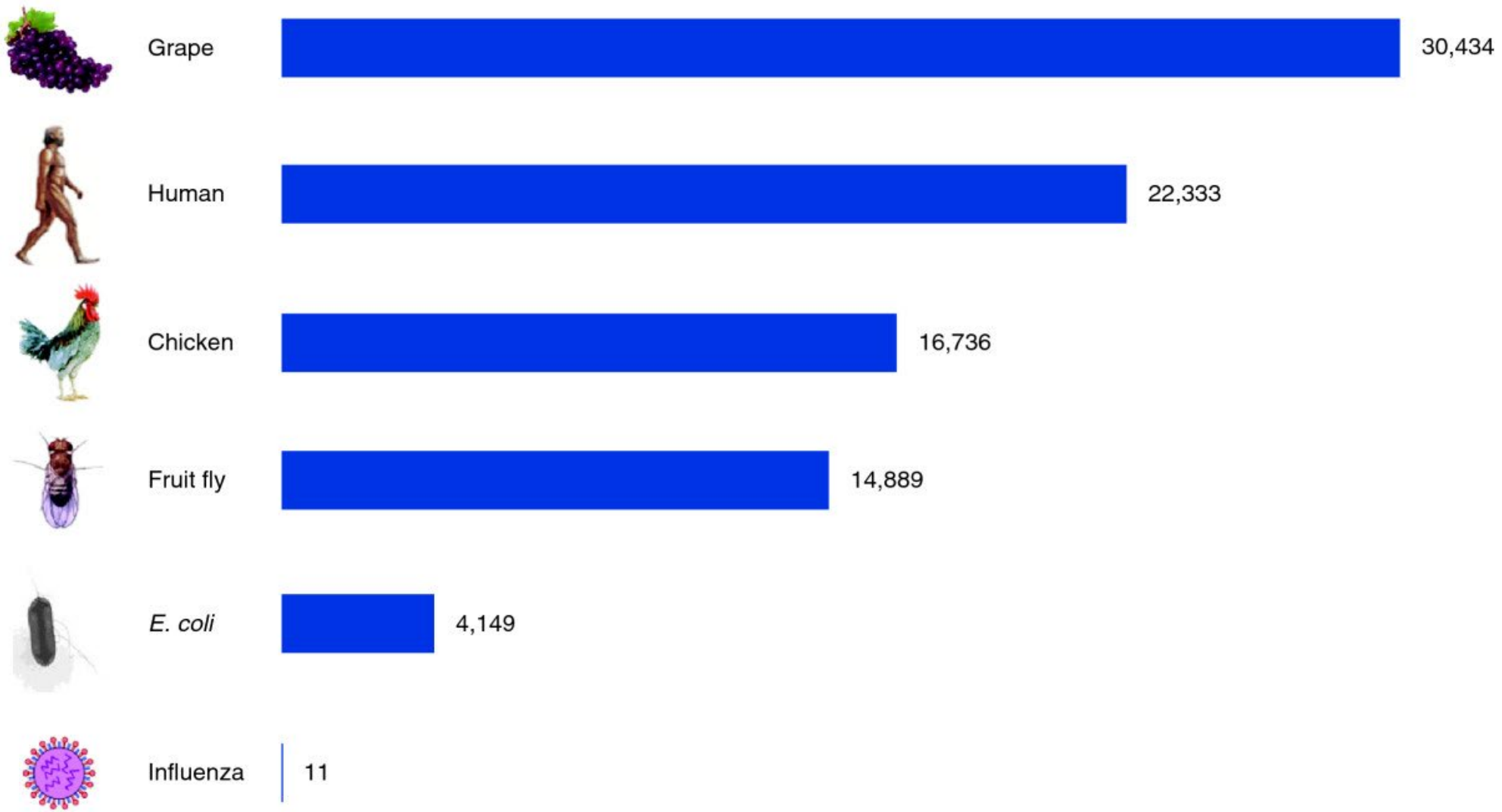
Professor Ziheng Yang FRS



Outline

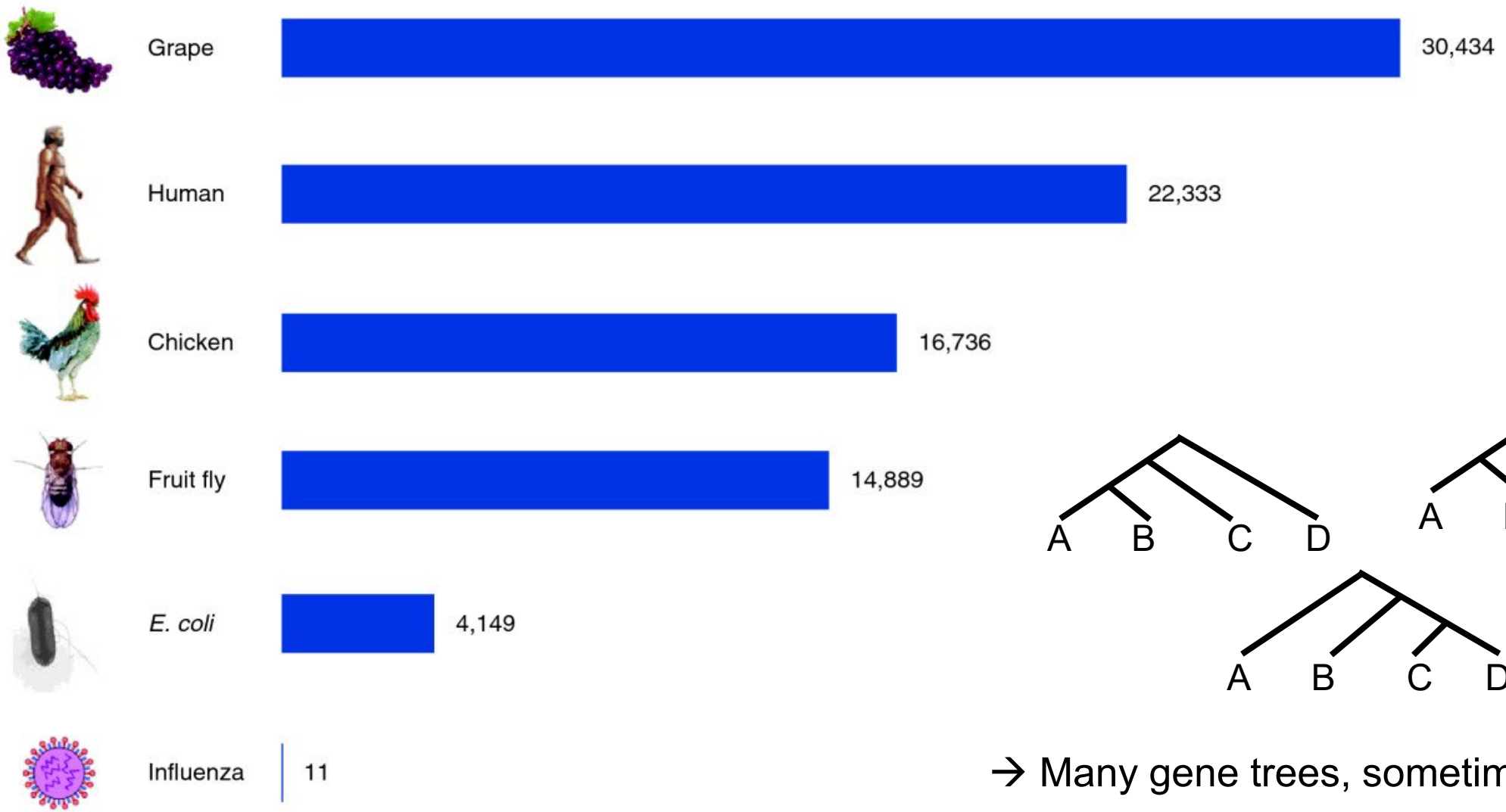
1. Phylogenetics
2. Phylogenomics
3. Approaches
4. Confidence

Number of protein-coding genes



www.clinicaleducation.org

Number of protein-coding genes



→ Many gene trees, sometimes different!

www.clinicaleducation.org

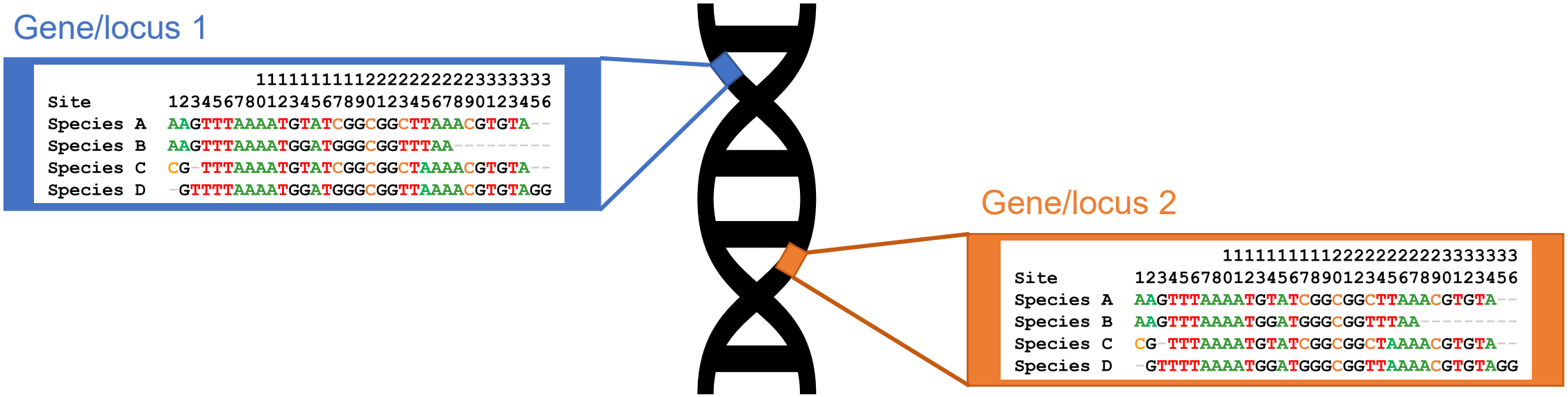

Phylogenomics:
Phylogenetics
considering
many genes

A “gene” or “locus” is here understood as:

A set of DNA sites that are close enough to each-other to not recombine

→ Sites inside a gene/locus are assumed to have a same evolutionary history, while two genes/loci may not

This is a key assumption made by many phylogenomics methods.



Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Horizontal gene transfer

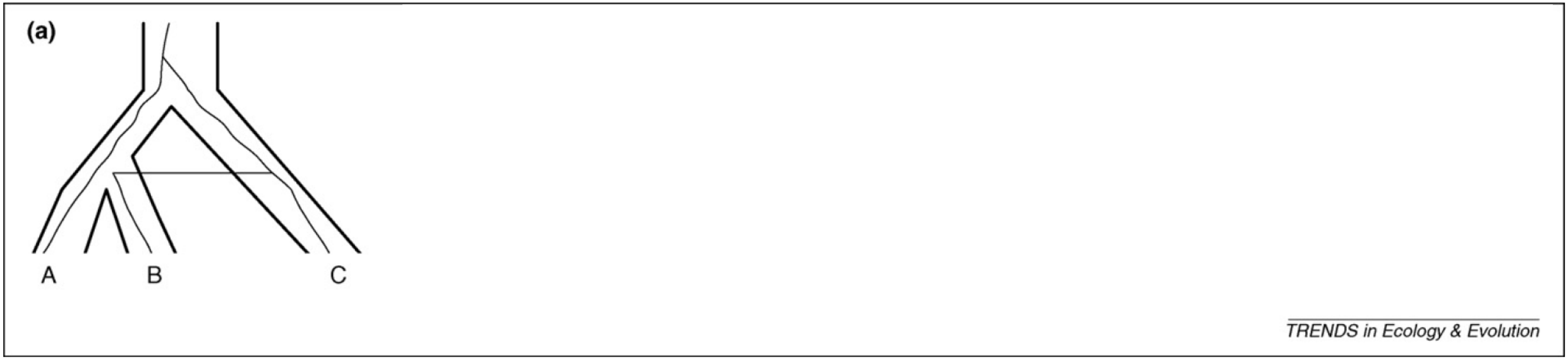
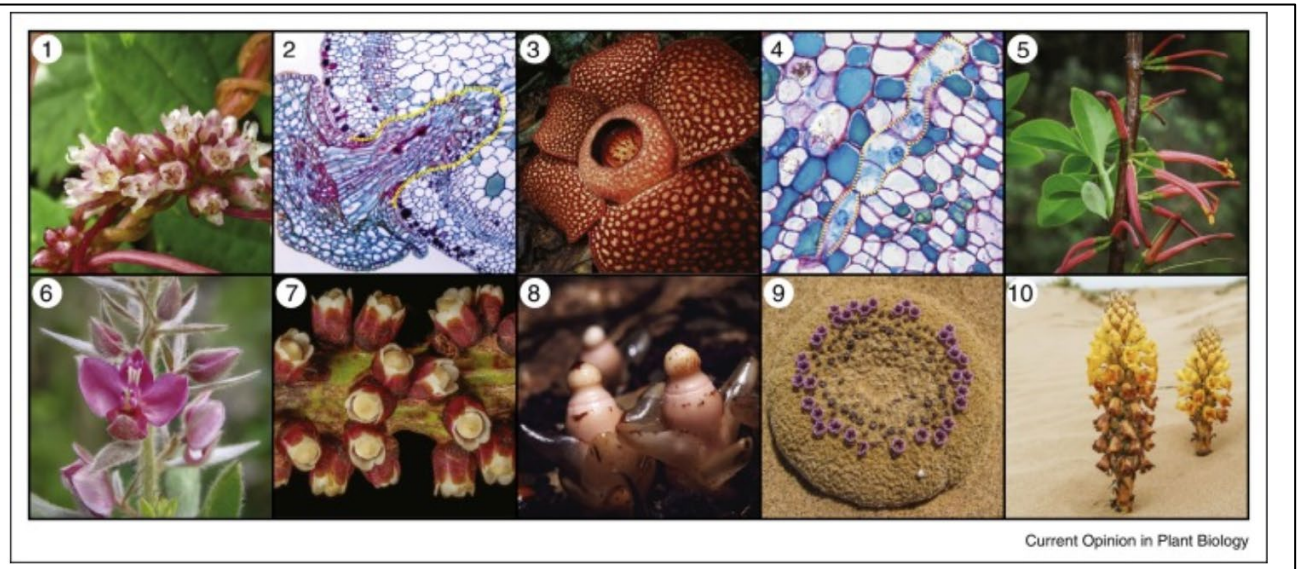


Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)).

Degnan & Rosenberg, Trends in Ecology and Evolution, 24

Examples of horizontal gene transfer in plants



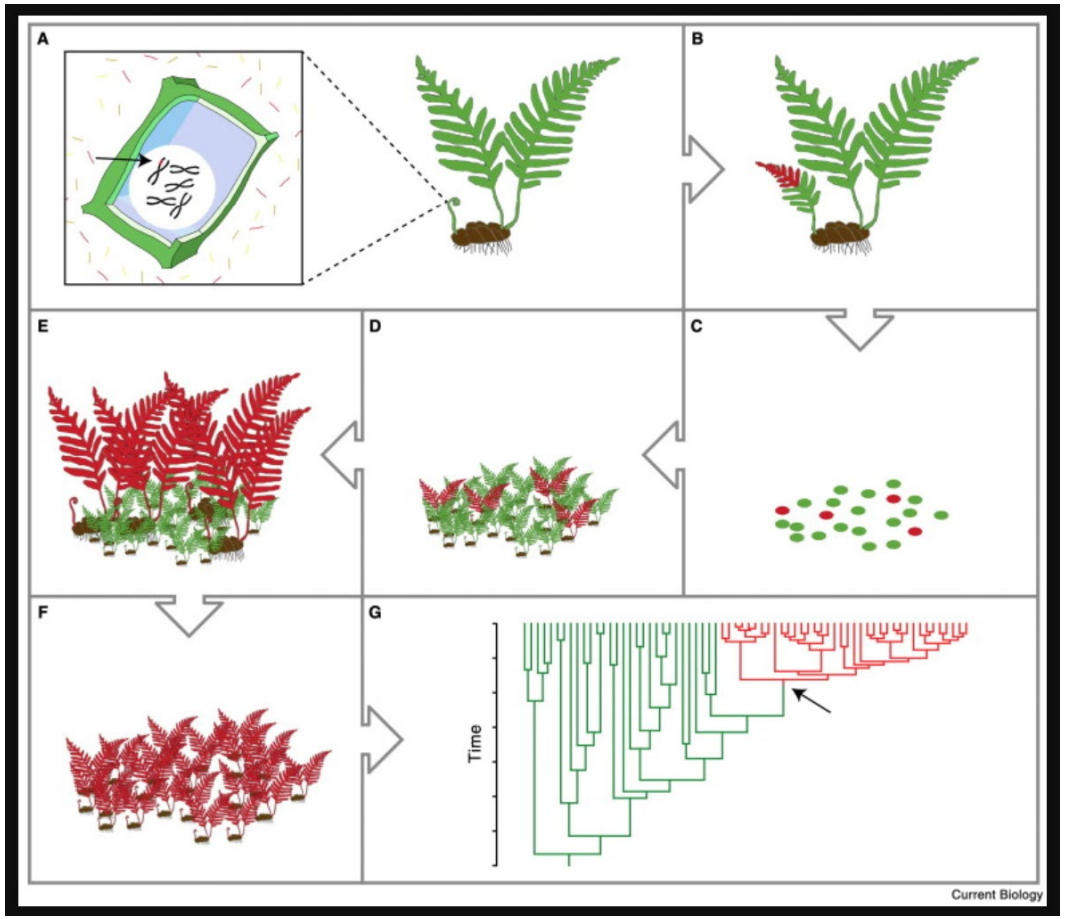
Download : Download high-res image (1MB) Download : Download full-size image

Figure 1. Parasitic plant diversity. (1) Holoparasitic *Cuscuta europaea* (Convolvulaceae); (2) *Cuscuta campestris* (Convolvulaceae) penetrating host tissue via a haustorium (highlighted with dotted yellow lines); (3) Holoparasitic *Rafflesia arnoldii* (Rafflesiaceae), which produces the world's largest flowers; (4) Holoparasitic *Rhizanthes lowii* (Rafflesiaceae) showing its very reduced endophyte (marked with dotted yellow lines) in the phloem of the host — the only remnants of a vegetative body; (5) Hemiparasitic *Taxillus delavayi* (Loranthaceae, Santalales); (6) Hemiparasitic *Krameria argenteaI* (Krameriaceae); (7) Holoparasitic *Pilostyles hamiltonii* (Apodanthaceae); (8) Holoparasitic *Mitrastema yamamotoi* (Mitrastemonaceae); (9) Holoparasitic *Pholisma sonorae* (Lennoaceae); and (10) Holoparasitic *Cistanche phelyphaea* (Orobanchaceae). Images copyright Dave, M. Costea, J. Holden, L. Nikolov, J. Lundberg, J. Medeiros, K. Thiele, C. Tada, J. Bartel, and P. Precey, respectively.

PNAS

Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns

Fay-Wei Li, Juan Carlos Villarreal, Steven Kelly, Carl J. Rothfels, Michael Melkonian, Eftychios Frangedakis, Markus Ruhsam, Erin M. Sigel, Joshua P. Der, Jarmila Pittermann, Dylan O. Burge, Lisa Pokorny, Anders Larsson, Tao Chen, Stina Weststrand, Philip Thomas, Eric Carpenter, Yong Zhang, Zhijian Tian, Li Chen, Zhixiang Yan, Ying Zhu, Xiao Sun, Jun Wang, Dennis W. Stevenson, Barbara J. Crandall-Stotler, A. Jonathan Shaw, Michael K. Deyholos, Douglas E. Soltis, Sean W. Graham, Michael D. Windham, Jane A. Langdale, Gane Ka-Shu Wong, Sarah Mathews, and Kathleen M. Pryer



Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Horizontal gene transfer

Gene duplication & loss

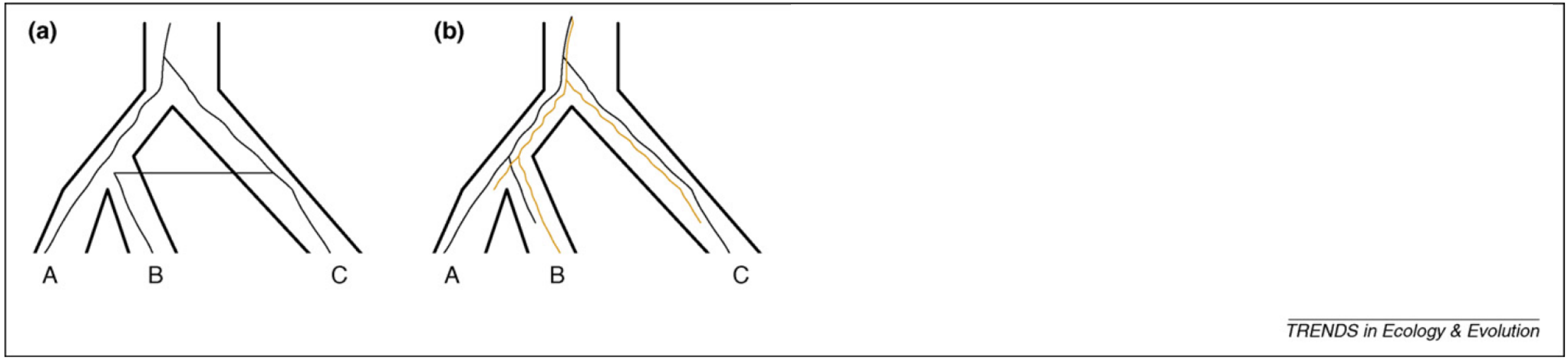
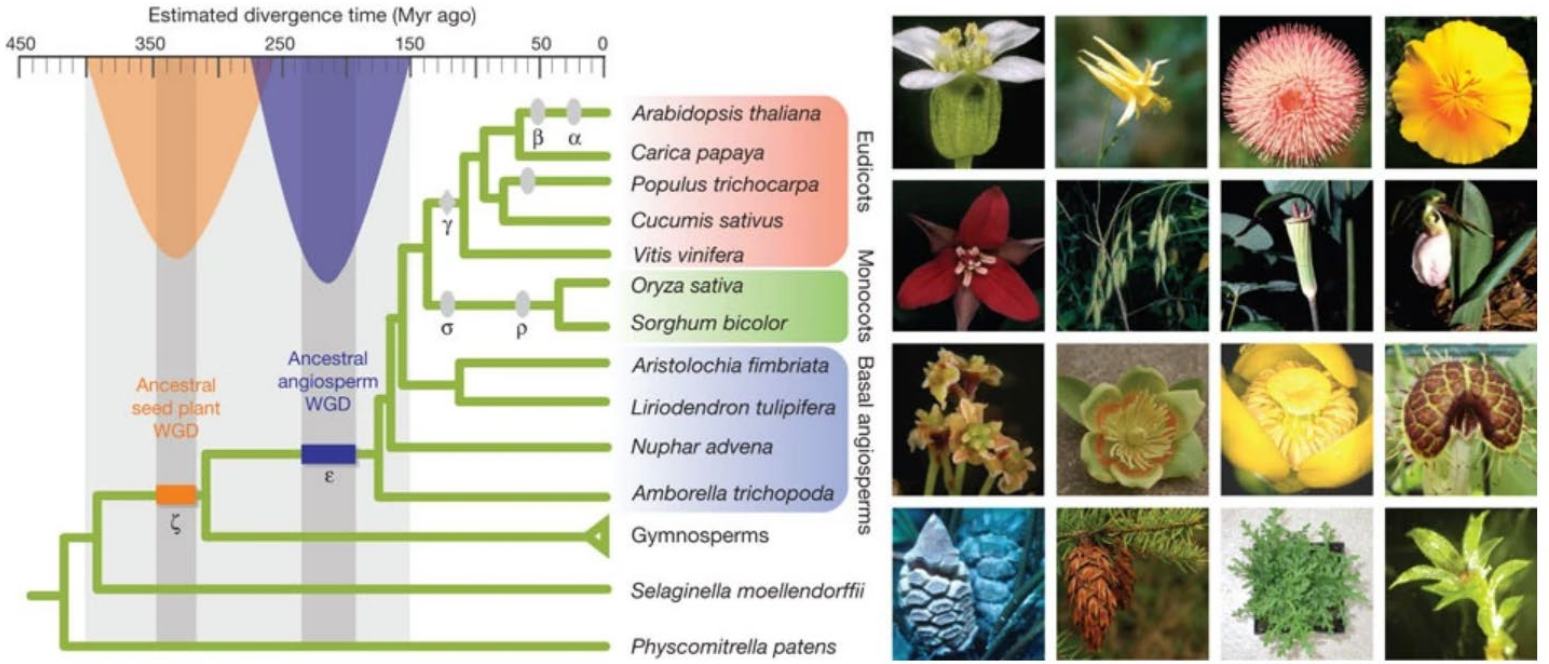


Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)). **(b)** Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause lineages from A and C to appear more closely related to each other than either is to B.

Degnan & Rosenberg, Trends in Ecology and Evolution, 24

From: Ancestral polyploidy in seed plants and angiosperms



Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see text). The diamond refers to the triplication event probably shared by all core eudicots. Horizontal bars denote confidence regions for ancestral seed plant WGD and ancestral angiosperm WGD, and are drawn to reflect upper and lower bounds of mean estimates from Fig. 2 (more orthogroups) and Supplementary Fig. 5 (more taxa). The photographs provide examples of the reproductive diversity of eudicots (top row, left to right: *Arabidopsis thaliana*, *Aquilegia chrysantha*, *Cirsium pumilum*, *Eschscholzia californica*), monocots (second row, left to right: *Trillium erectum*, *Bromus kalmii*, *Arisaema triphyllum*, *Cypripedium acaule*), basal angiosperms (third row, left to right: *Amborella trichopoda*, *Liriodendron tulipifera*, *Nuphar advena*, *Aristolochia fimbriata*), gymnosperms (fourth row, first and second from left: *Zamia vazquezii*, *Pseudotsuga menziesii*) and the outgroups *Selaginella moellendorfii* (vegetative; fourth row, third from left) and *Physcomitrella patens* (fourth row, right). See Supplementary Table 4 for photo credits.

Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Horizontal gene transfer Gene duplication & loss Hybridization

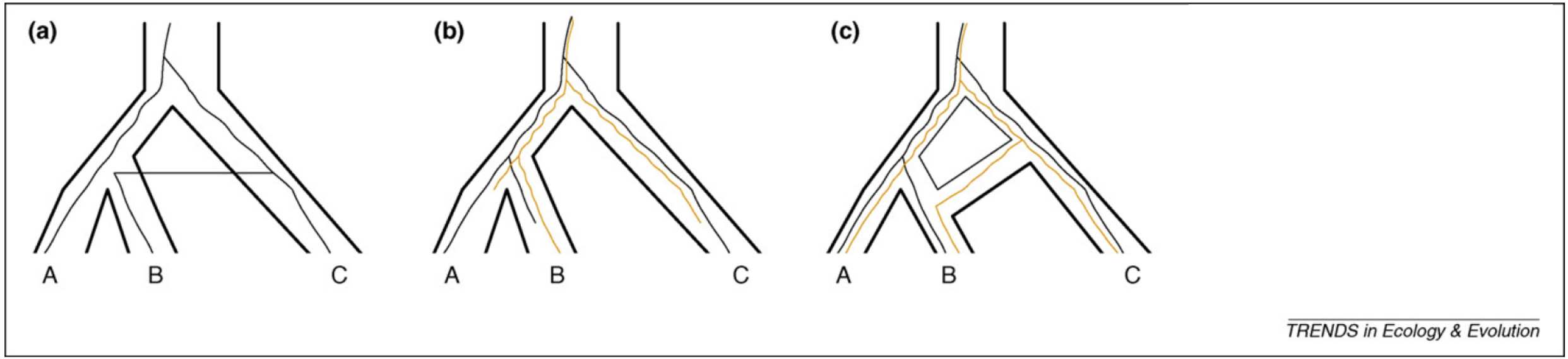


Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)). **(b)** Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause lineages from A and C to appear more closely related to each other than either is to B. **(c)** Hybridization causes some genes sampled from species B to descend from the population ancestral to A and B, whereas others descend from the population ancestral to B and C. The two gene trees depicted in (c) are ((AB)C) (black) and (A(BC)) (orange). Hybridization affects whole genomes, whereas HGT typically affects only small DNA segments.

Degnan & Rosenberg, Trends in Ecology and Evolution, 24

Fig. 3. Putative natural hybrids in Disa. A. Inflorescences of a hybrid (center) between D. atricapilla (left) and D. bivalvata (right). B. Inflorescence of a hybrid (center) between D. fragrans (left) and D. sankeyi (right). C. Inflorescences of a hybrid (center) between D. albomagentea (left) and D. obtusa subsp. picta (right). D. Inflorescence of a hybrid (center) between D. graminifolia (left) and D. ferruginea (right). E. Inflorescence of a hybrid (center) between D. sabulosa (left) and D. pygmaea (right). F. Inflorescence of a hybrid (center) between D. cephalotes subsp. frigida (left) and D. cephalotes subsp. cephalotes (right). Inflorescence of a hybrid (right) between D. versicolor (left) and either D. polygonoides or D. woodii. H. Inflorescence of a hybrid between D. versicolor (see panel G) and D. sankeyi (see panel C). I. Inflorescence of a hybrid (center) between Disa rhodantha (left) and Disa versicolor (right). J. Flowers of a hybrid (center) between Disa amoena (left) and Disa vigilans (right). Scale bars: A-J = 10 mm. Image credits: A, B, D (right), F (left), I (right) by S.D. Johnson; C (left and right), D (left and center), E by W. Liltved; C (center), F (right), G, H by H. Stärker; F (center) by J. Jersakova, I (left and center) by K. Wodrich, J by D. Bellstedt.

Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Horizontal gene transfer | Gene duplication & loss | Hybridization | Recombination

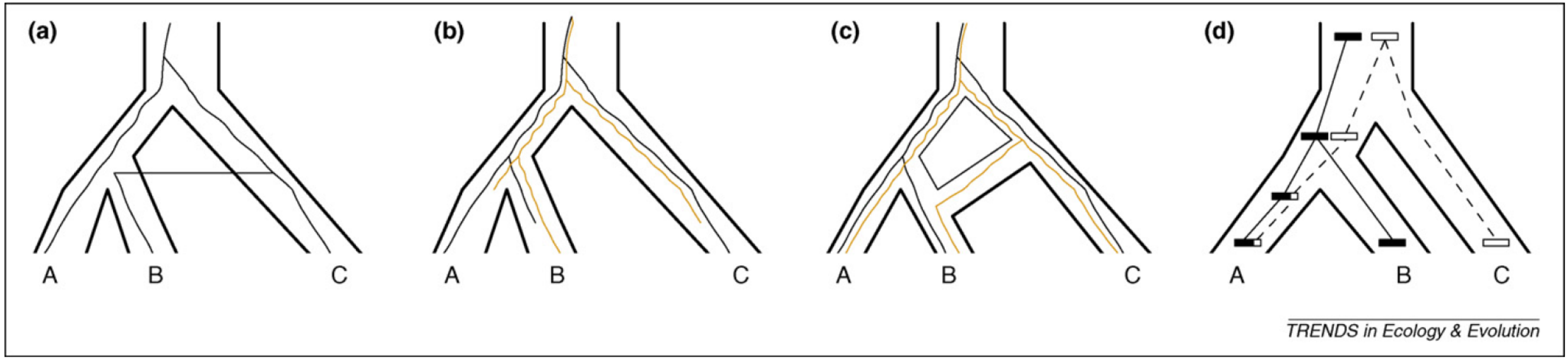
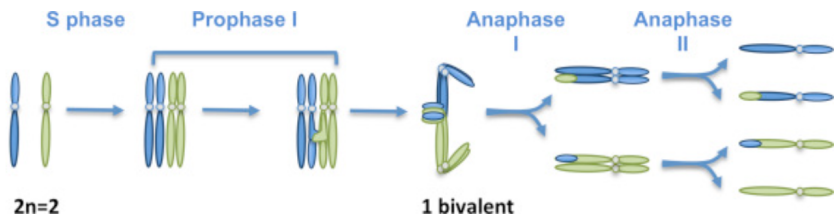


Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)). **(b)** Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause lineages from A and C to appear more closely related to each other than either is to B. **(c)** Hybridization causes some genes sampled from species B to descend from the population ancestral to A and B, whereas others descend from the population ancestral to B and C. The two gene trees depicted in (c) are ((AB)C) (black) and (A(BC)) (orange). Hybridization affects whole genomes, whereas HGT typically affects only small DNA segments. **(d)** Recombination can lead to different histories for neighboring segments within a gene. For the DNA segment depicted in black, the gene tree is ((AB)C), but for the segment in white, the gene tree is ((AC)B).

Degnan & Rosenberg, Trends in Ecology and Evolution, 24

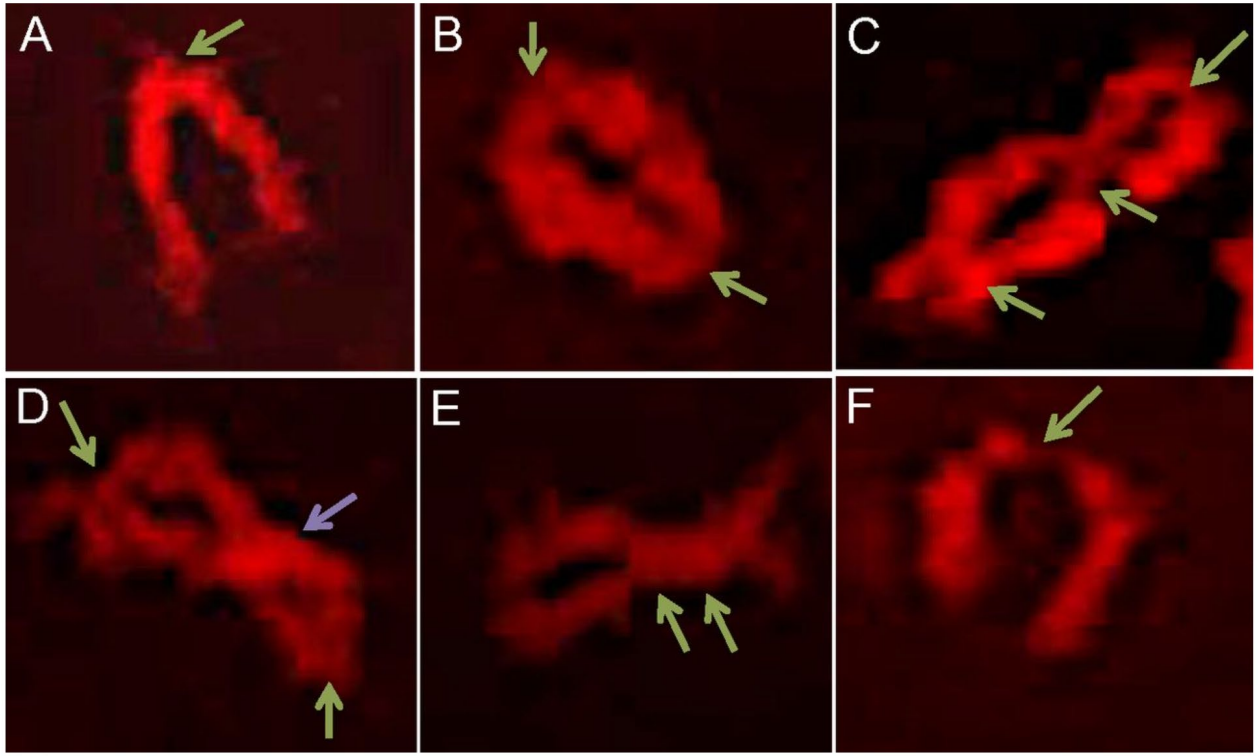
Meiotic recombination (simplified)



Download : Download full-size image

Fig. 1. Schematic representation of a meiotic division. In meiocytes, as in any sporophytic cell, a set of chromosomes of maternal origin (blue) coexists with a set of chromosomes of paternal origin (green). They correspond to pairs of homologous chromosomes (or homologs). Here, a hypothetical organism with a diploid number of chromosomes of 2 has been chosen. Replication (S phase) duplicates each chromosome into two sister chromatids that are kept together by the action of cohesins (not shown). Meiosis consists in the succession of two rounds of chromosome segregation (Anaphases I and II) after a single S phase. During prophase I, homologous chromosomes recombine and associate into bivalents. Meiosis I separates the homologous chromosomes, while meiosis II separates the sister chromatids.

Meiotic recombination in maize



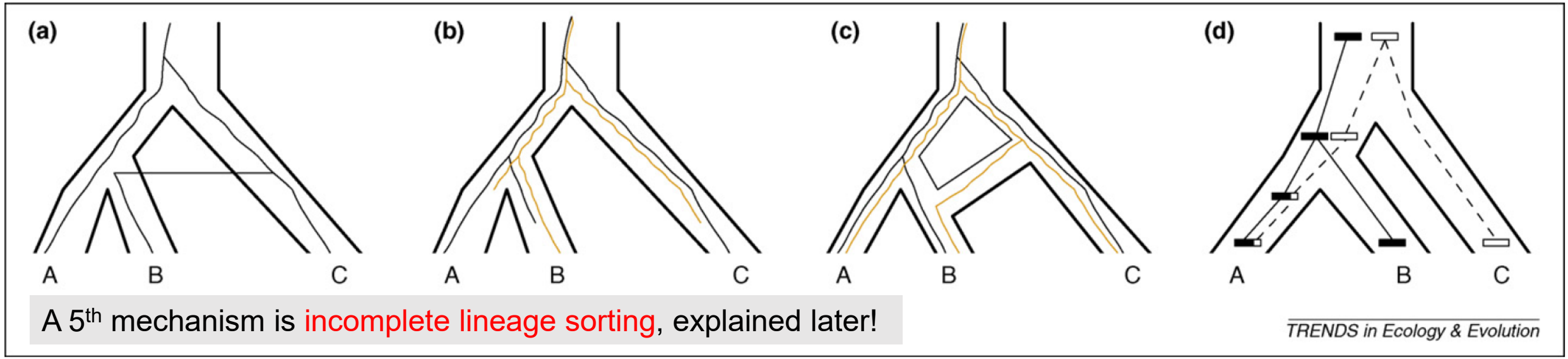
Quantification of chiasmata in maize. Chiasmata were identified using a combination of chromosome spreading and 3D image reconstruction. (*A*) Typical morphology of a rod bivalent with a single chiasma at a chromosome end. (*B*) A ring bivalent with two chiasmata, one at each end. (*C*) A ring bivalent with three chiasmata, two terminal and one interstitial. (*D*) A bivalent with two terminal chiasmata and a chromosome twist. (*E*) A bivalent with two chiasmata located close to each other. (*F*) A bivalent with a poorly visible terminal chiasma that could make the bivalent be confused for two univalents. Chiasmata are indicated by green arrows. A chromosome twist is indicated by a purple arrow.

Different genes may have different histories!

Not all genes reflect the species relationships, or they may reflect different aspects of them.

Possible mechanisms for this are:

Horizontal gene transfer | Gene duplication & loss | Hybridization | Recombination



A 5th mechanism is incomplete lineage sorting, explained later!

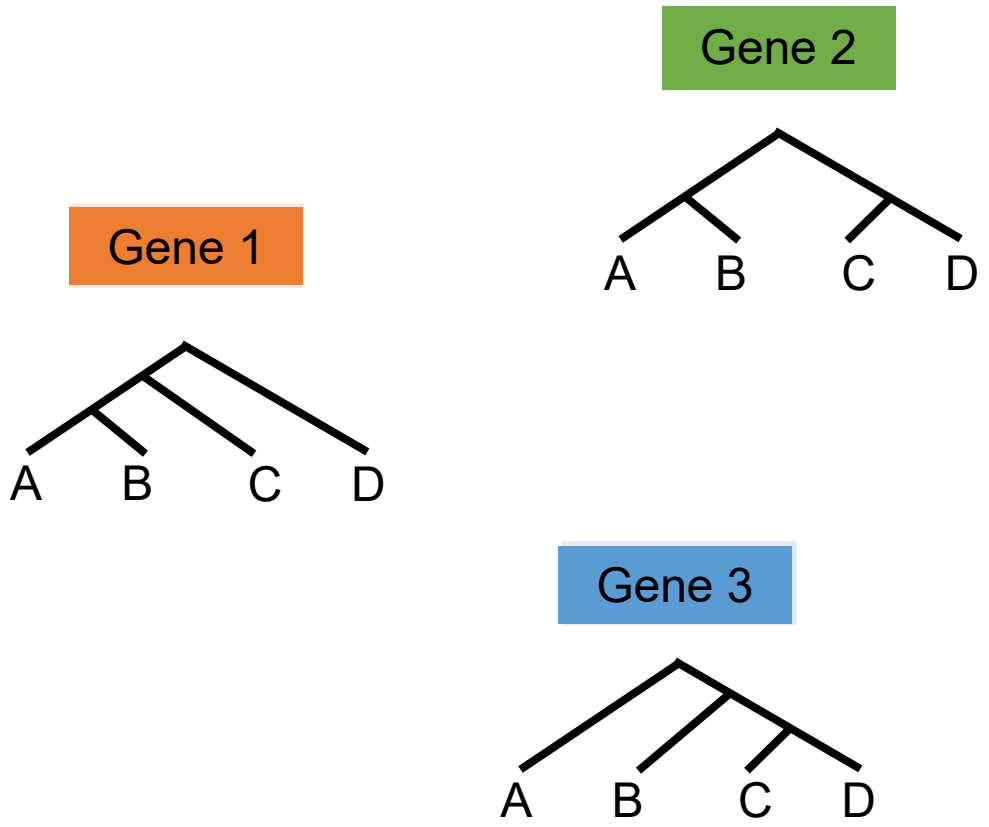
Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)). **(b)** Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause lineages from A and C to appear more closely related to each other than either is to B. **(c)** Hybridization causes some genes sampled from species B to descend from the population ancestral to A and B, whereas others descend from the population ancestral to B and C. The two gene trees depicted in (c) are ((AB)C) (black) and (A(BC)) (orange). Hybridization affects whole genomes, whereas HGT typically affects only small DNA segments. **(d)** Recombination can lead to different histories for neighboring segments within a gene. For the DNA segment depicted in black, the gene tree is ((AB)C), but for the segment in white, the gene tree is ((AC)B).

Degnan & Rosenberg, Trends in Ecology and Evolution, 24

Phylogenomics allow to study many DNA characters

→ They can provide new resolution, but…

→ They also reveal conflicts between gene trees



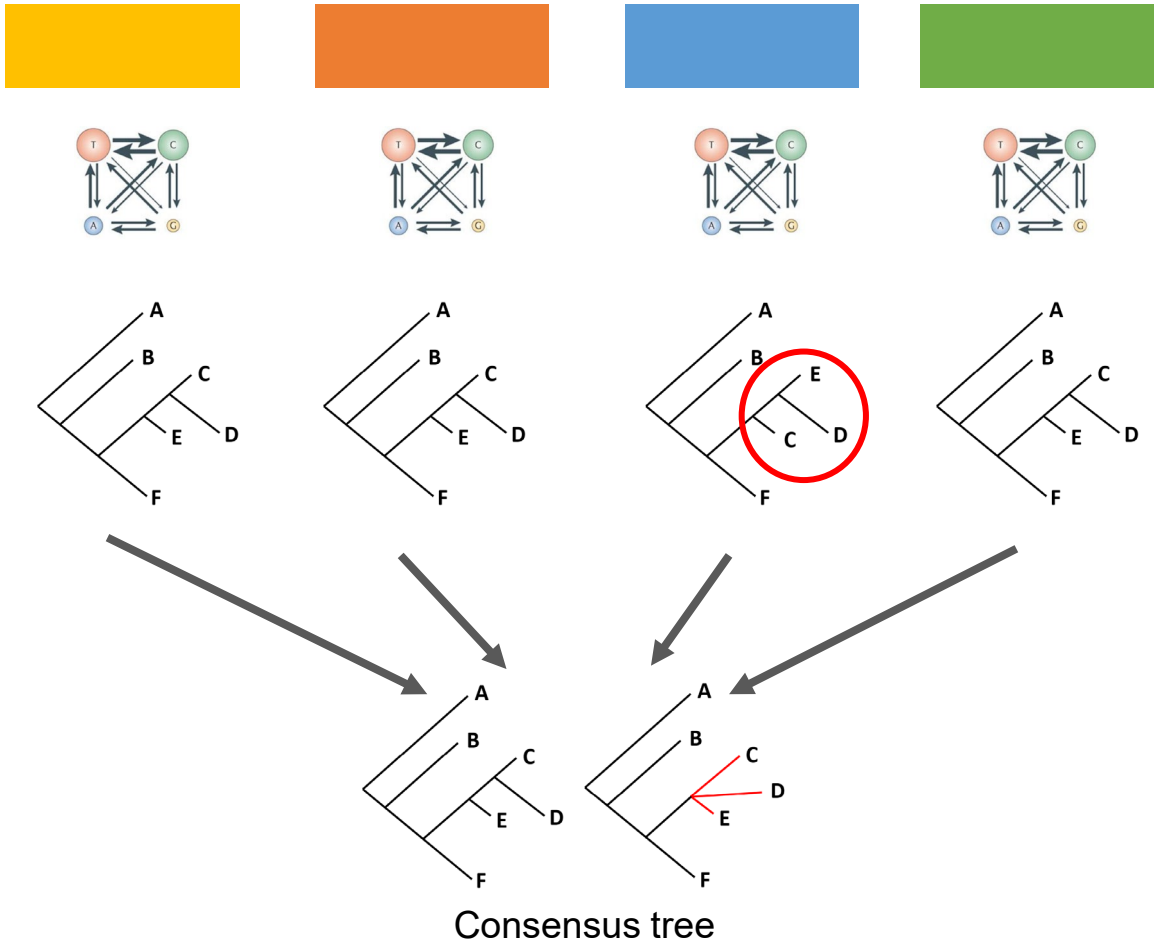
How to obtain a species tree from many gene trees?

Outline

1. Phylogenetics
2. Phylogenomics
3. Approaches
4. Confidence

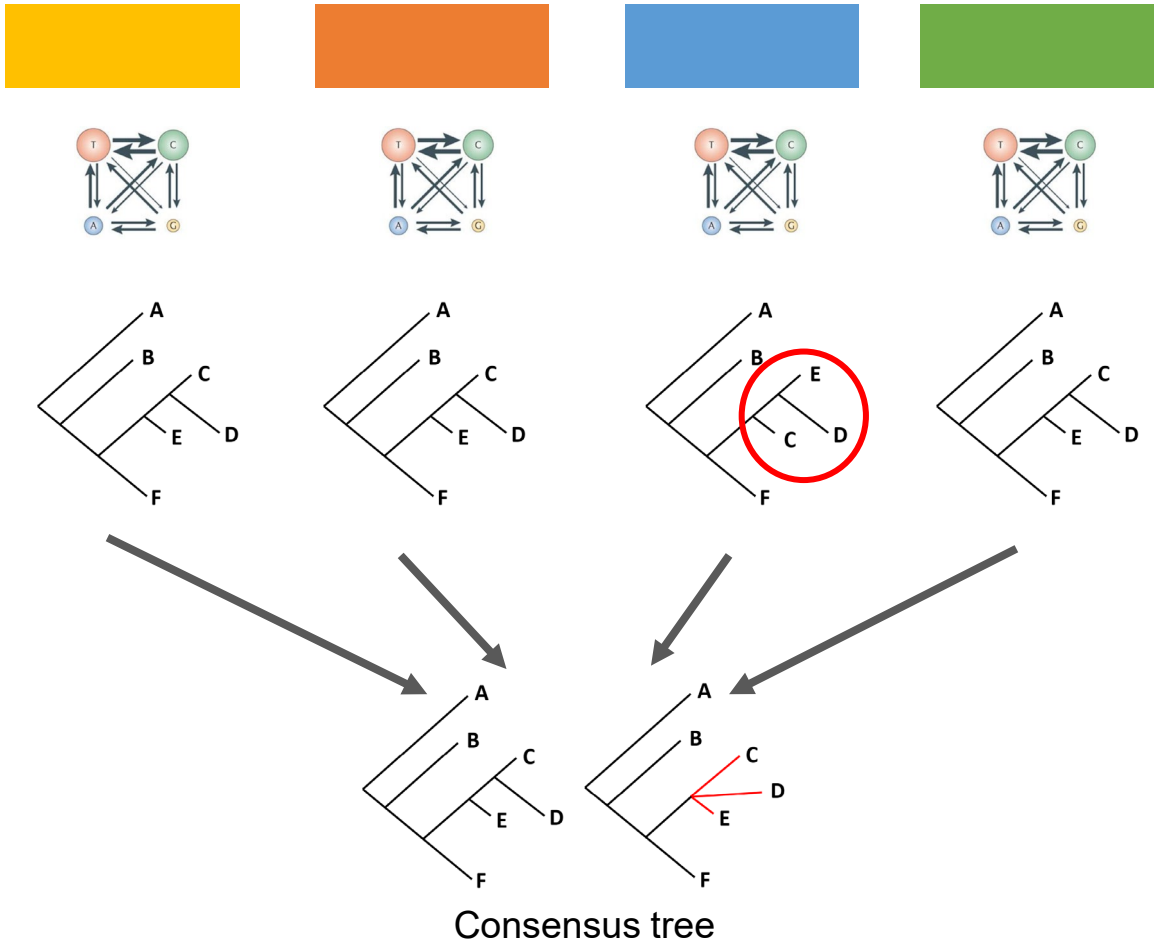
How to obtain a species tree from many gene trees?

Supertree approach

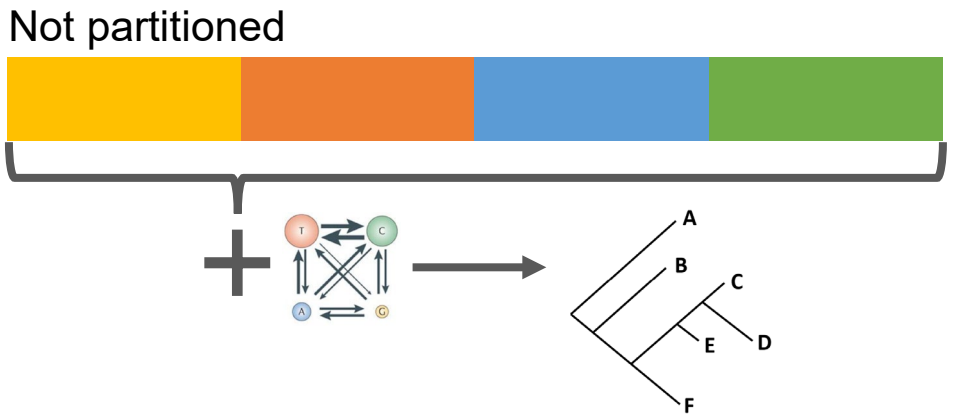


How to obtain a species tree from many gene trees?

Supertree approach

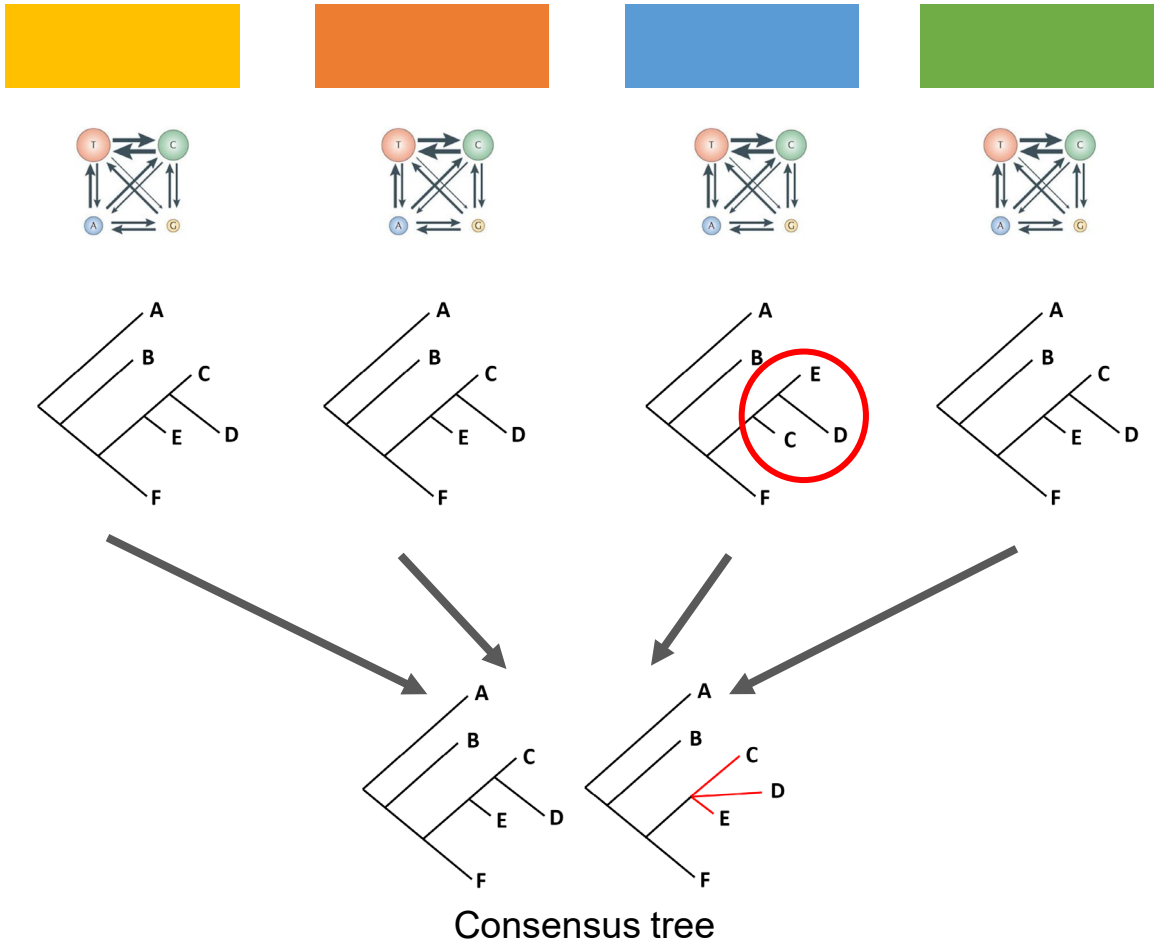


Supermatrix = concatenation approach

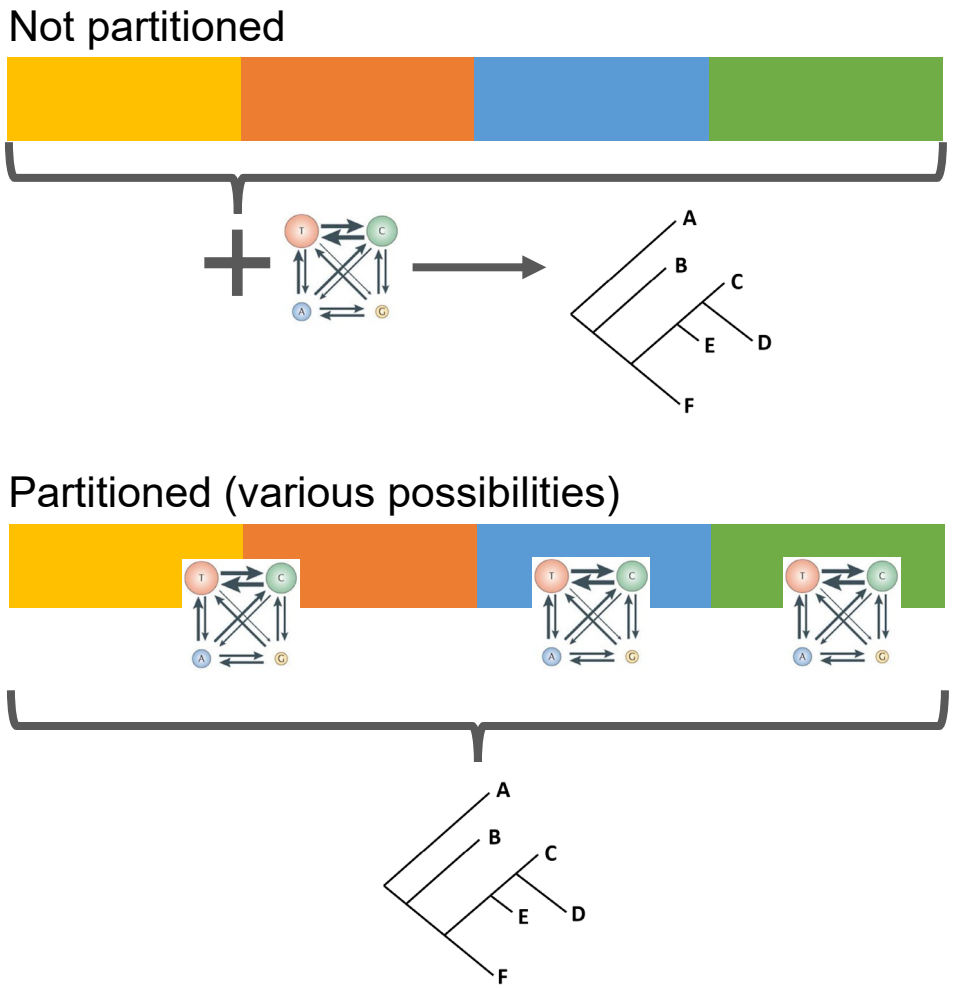


How to obtain a species tree from many gene trees?

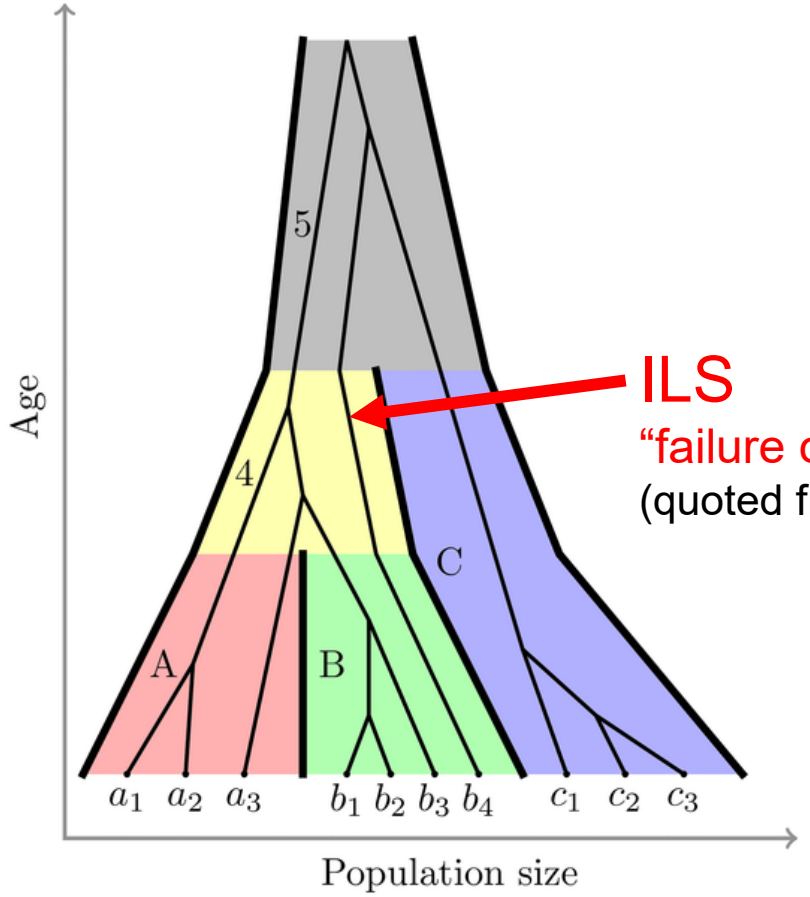
Supertree approach



Supermatrix = concatenation approach



The issues with concatenation: incomplete lineage sorting, the anomaly zone and more



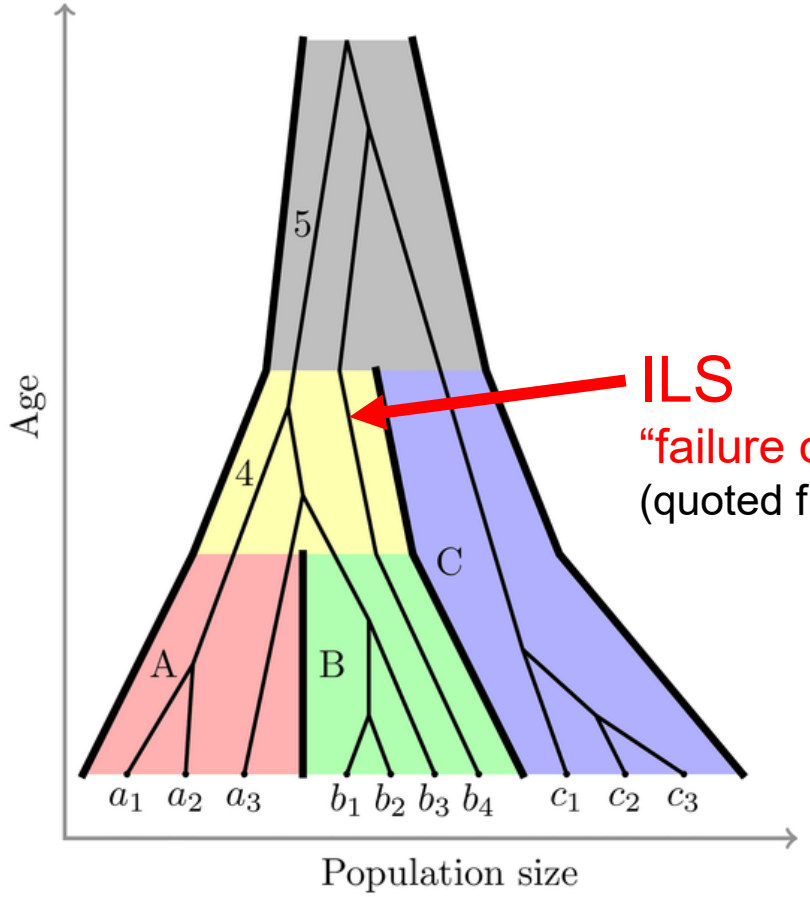
ILS

"failure of gene lineages in a population to coalesce"
(quoted from Degnan&Rosenberg, 2009, Trends Ecol. Evol. 24)

Fig 4. The multispecies coalescent (MSC) model with three species and a single gene tree. A separate coalescent process applies to each of the five branches in the tree; the branches for the extant species A (red), B (green) and C (blue), the ancestral branch of A and B (yellow), and the root branch (grey). Several individuals have been sampled per species. In this example the ancestral lineage of individual b_4 does not coalesce in species B or ancestral species 4. In ancestral species 5, it coalesces with the ancestral lineage of species C. This leads to incomplete lineage sorting and enables gene tree discordance—in this example b_4 is a sister taxon to individuals from species C, rather than to individuals from its own species, or sister species A. If b_4 was the representative individual for its species, then this gene would exhibit gene tree discordance. Other individuals which show concordance at this locus are expected to show discordance at other unlinked loci when populations are large or speciation times are recent.

Bouckaert et al. 2019. PLOS Computational Biology 15(4): e1006650

The issues with concatenation: incomplete lineage sorting, the anomaly zone and more



The probability of Incomplete Lineage Sorting (ILS) increases with shorter branch lengths and larger population size (N_e)

ILS

"failure of gene lineages in a population to coalesce"
(quoted from Degnan&Rosenberg, 2009, Trends Ecol. Evol. 24)

Fig 4. The multispecies coalescent (MSC) model with three species and a single gene tree. A separate coalescent process applies to each of the five branches in the tree; the branches for the extant species A (red), B (green) and C (blue), the ancestral branch of A and B (yellow), and the root branch (grey). Several individuals have been sampled per species. In this example the ancestral lineage of individual b_4 does not coalesce in species B or ancestral species 4. In ancestral species 5, it coalesces with the ancestral lineage of species C. This leads to incomplete lineage sorting and enables gene tree discordance—in this example b_4 is a sister taxon to individuals from species C, rather than to individuals from its own species, or sister species A. If b_4 was the representative individual for its species, then this gene would exhibit gene tree discordance. Other individuals which show concordance at this locus are expected to show discordance at other unlinked loci when populations are large or speciation times are recent.

Bouckaert et al. 2019. PLOS Computational Biology 15(4): e1006650

The issues with concatenation: incomplete lineage sorting, the anomaly zone and more

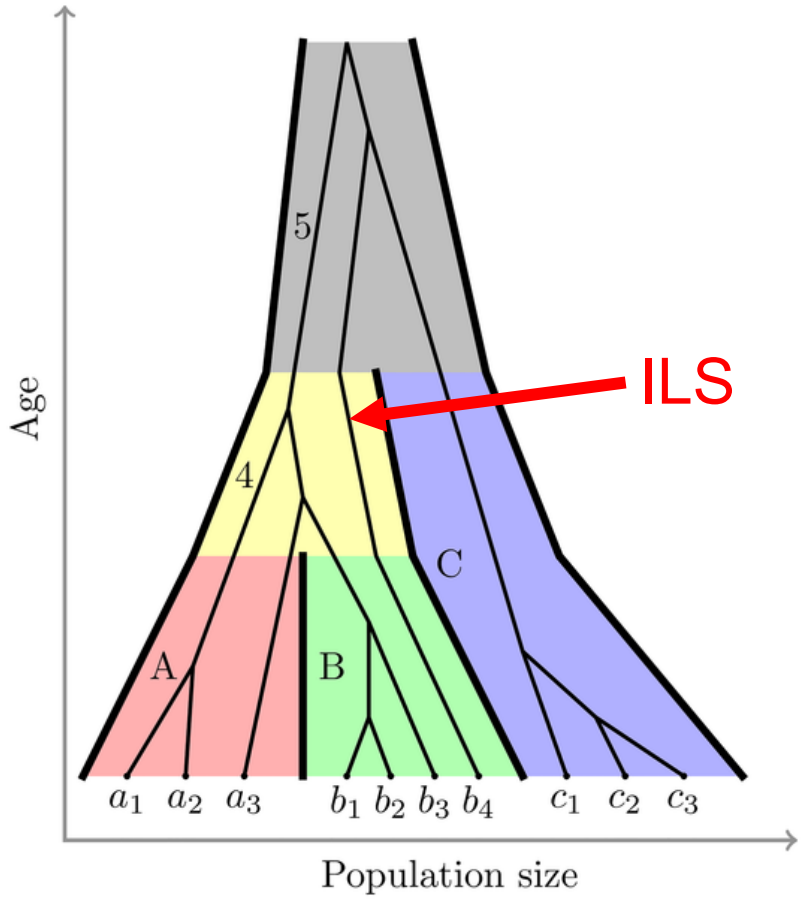
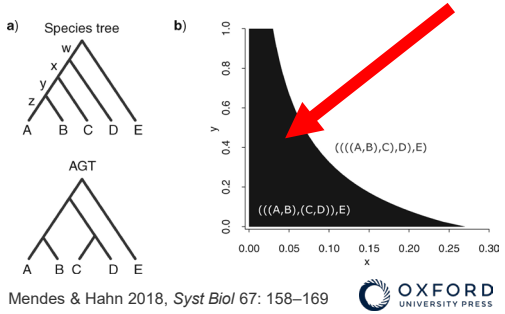


Fig 4. The multispecies coalescent (MSC) model with three species and a single gene tree. A separate coalescent process applies to each of the five branches in the tree; the branches for the extant species A (red), B (green) and C (blue), the ancestral branch of A and B (yellow), and the root branch (grey). Several individuals have been sampled per species. In this example the ancestral lineage of individual b_4 does not coalesce in species B or ancestral species 4. In ancestral species 5, it coalesces with the ancestral lineage of species C. This leads to incomplete lineage sorting and enables gene tree discordance—in this example b_4 is a sister taxon to individuals from species C, rather than to individuals from its own species, or sister species A. If b_4 was the representative individual for its species, then this gene would exhibit gene tree discordance. Other individuals which show concordance at this locus are expected to show discordance at other unlinked loci when populations are large or speciation times are recent.

Bouckaert et al. 2019. PLOS Computational Biology 15(4): e1006650

The probability of Incomplete Lineage Sorting (ILS) increases with shorter branch lengths and larger population size (N_e)

→ For a tree with at least 4 ingroup species,
there is a zone of the parameter space (branch length space) where most gene trees have a topology different from the species tree: the "Anomaly Zone".



Mendes & Hahn 2018, *Syst Biol* 67: 158–169

"When two or more consecutive internal branches on a species tree are sufficiently short, gene trees incongruent with the species tree can be more common than congruent gene trees"

Figure 1 a) Top tree: smallest species tree for which an anomaly zone can be defined, where z is the length of terminal branches A and B, and w, x, and y are the lengths of the three internal branches (oldest to youngest), respectively. Bottom tree: the most common gene tree (an anomalous gene tree, AGT) when species tree ((((A,B),C),D),E) (top tree) is inside the anomaly zone. Branch lengths are arbitrary and were not drawn in proportion to theoretical or simulated averages. b) Phylogenetic tree space for species tree ((((A,B),C),D),E), where x and y correspond to the lengths of the oldest and youngest ingroup internal branches, respectively, (as shown in [a]; x and y are measured in coalescent units, i.e., N$_{\rm e}$ generations). The region shaded in black corresponds to the anomaly zone (Degnan and Rosenberg 2006), in which the most common gene tree is AGT (((A,B),(C,D)),E).}

The issues with concatenation: incomplete lineage sorting, the anomaly zone and more

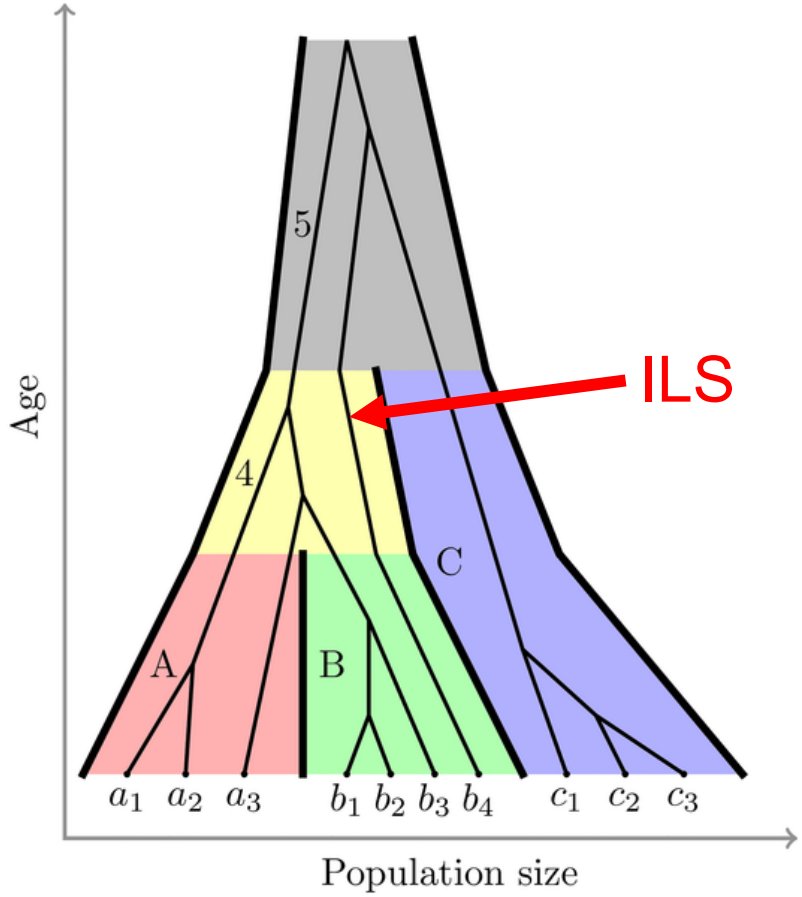
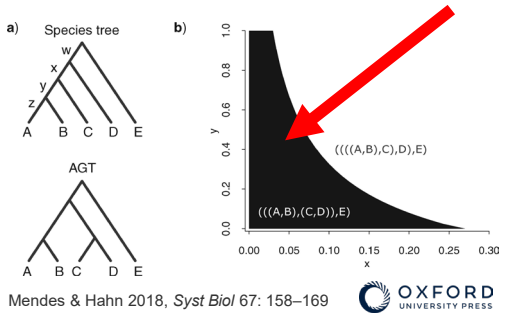


Fig 4. The multispecies coalescent (MSC) model with three species and a single gene tree. A separate coalescent process applies to each of the five branches in the tree; the branches for the extant species A (red), B (green) and C (blue), the ancestral branch of A and B (yellow), and the root branch (grey). Several individuals have been sampled per species. In this example the ancestral lineage of individual b_4 does not coalesce in species B or ancestral species 4. In ancestral species 5, it coalesces with the ancestral lineage of species C. This leads to incomplete lineage sorting and enables gene tree discordance—in this example b_4 is a sister taxon to individuals from species C, rather than to individuals from its own species, or sister species A. If b_4 was the representative individual for its species, then this gene would exhibit gene tree discordance. Other individuals which show concordance at this locus are expected to show discordance at other unlinked loci when populations are large or speciation times are recent.

Bouckaert et al. 2019. PLOS Computational Biology 15(4): e1006650

The probability of Incomplete Lineage Sorting (ILS) increases with shorter branch lengths and larger population size (N_e)

→ For a tree with at least 4 ingroup species, there is a zone of the parameter space (branch length space) where most gene trees have a topology different from the species tree: the "Anomaly Zone".



Mendes & Hahn 2018, *Syst Biol* 67: 158–169

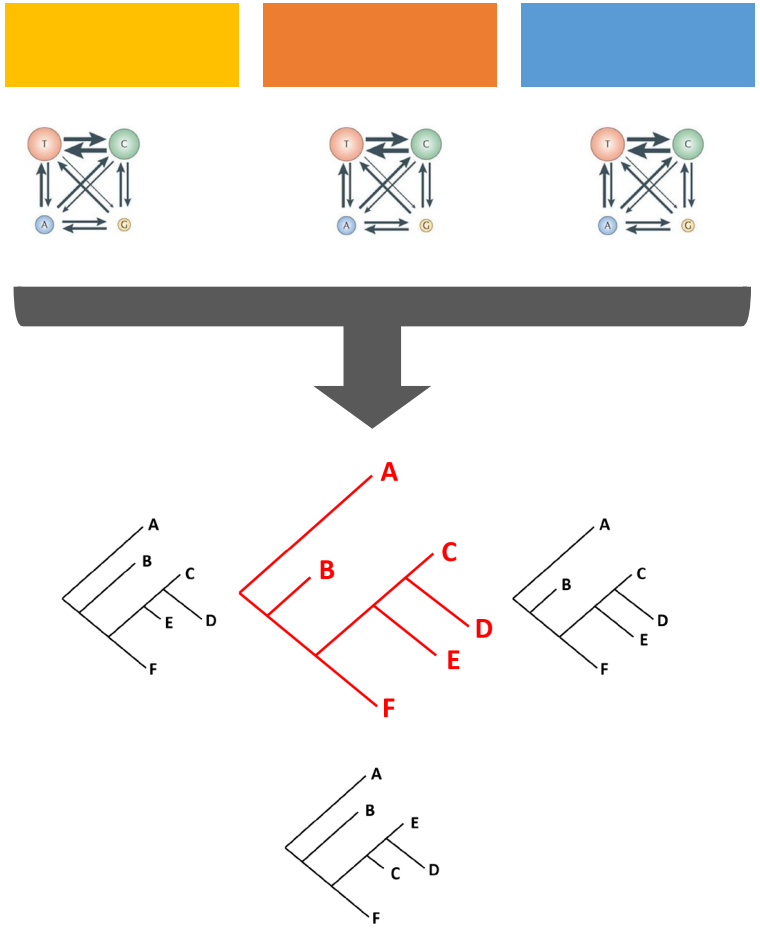
"When two or more consecutive internal branches on a species tree are sufficiently short, gene trees incongruent with the species tree can be more common than congruent gene trees"

Figure 1 a) Top tree: smallest species tree for which an anomaly zone can be defined, where z is the length of terminal branches A and B, and w, x, and y are the lengths of the three internal branches (oldest to youngest), respectively. Bottom tree: the most common gene tree (an anomalous gene tree, AGT) when species tree ((((A,B),C),D),E) (top tree) is inside the anomaly zone. Branch lengths are arbitrary and were not drawn in proportion to theoretical or simulated averages. b) Phylogenetic tree space for species tree ((((A,B),C),D),E), where x and y correspond to the lengths of the oldest and youngest ingroup internal branches, respectively, (as shown in [a]; x and y are measured in coalescent units, i.e., N$_{\rm e}$ generations). The region shaded in black corresponds to the anomaly zone (Degnan and Rosenberg 2006), in which the most common gene tree is AGT (((A,B),(C,D)),E).}

→ If you concatenate just select the majority tree among the gene trees, you may get the wrong species tree
(but see Mendes & Hahn: sometimes that can happen outside of the AZ, and sometimes one can still get the right tree even if it is in the AZ)

→ Solution: the multispecies coalescent model, which accounts for gene tree discordance due to ILS
(More details on the MSC in Degnan & Rosenberg, 2009. Trends in Ecology and Evolution, 24)

Multispecies coalescent (MSC) approaches
Co-estimation methods

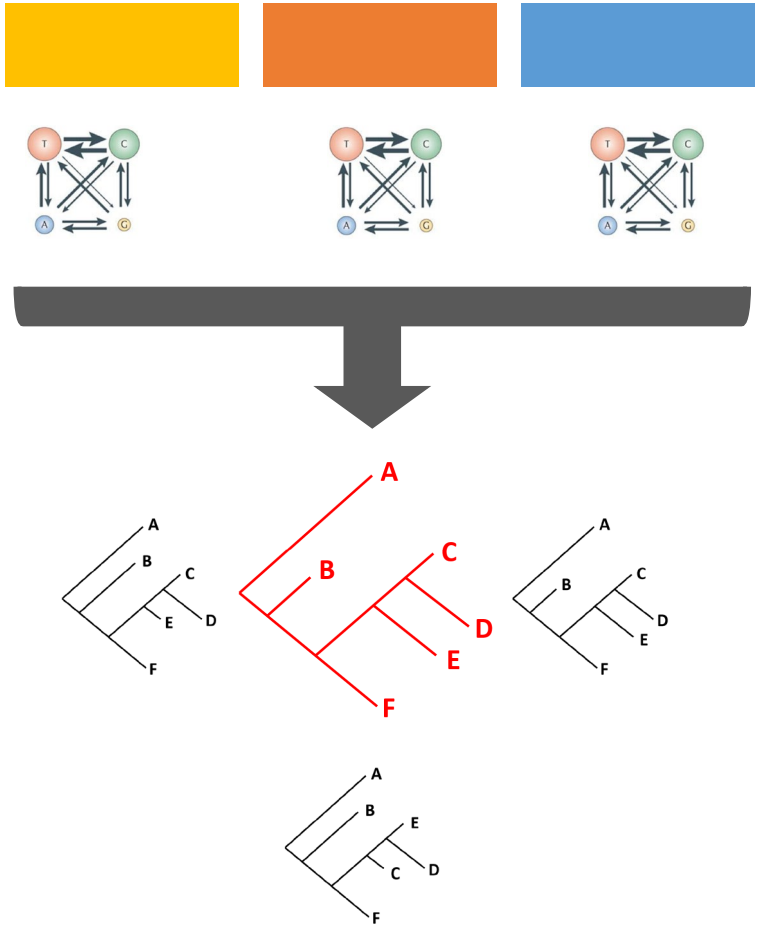


Sequence alignments
for each gene
+
Models of evolution

Phylogenetic inference using the MSC

Co-estimate gene trees
and species tree

Multispecies coalescent (MSC) approaches
Co-estimation methods



Sequence alignments
for each gene
+
Models of evolution

Phylogenetic inference using the MSC

Co-estimate gene trees
and species tree

***BEAST**

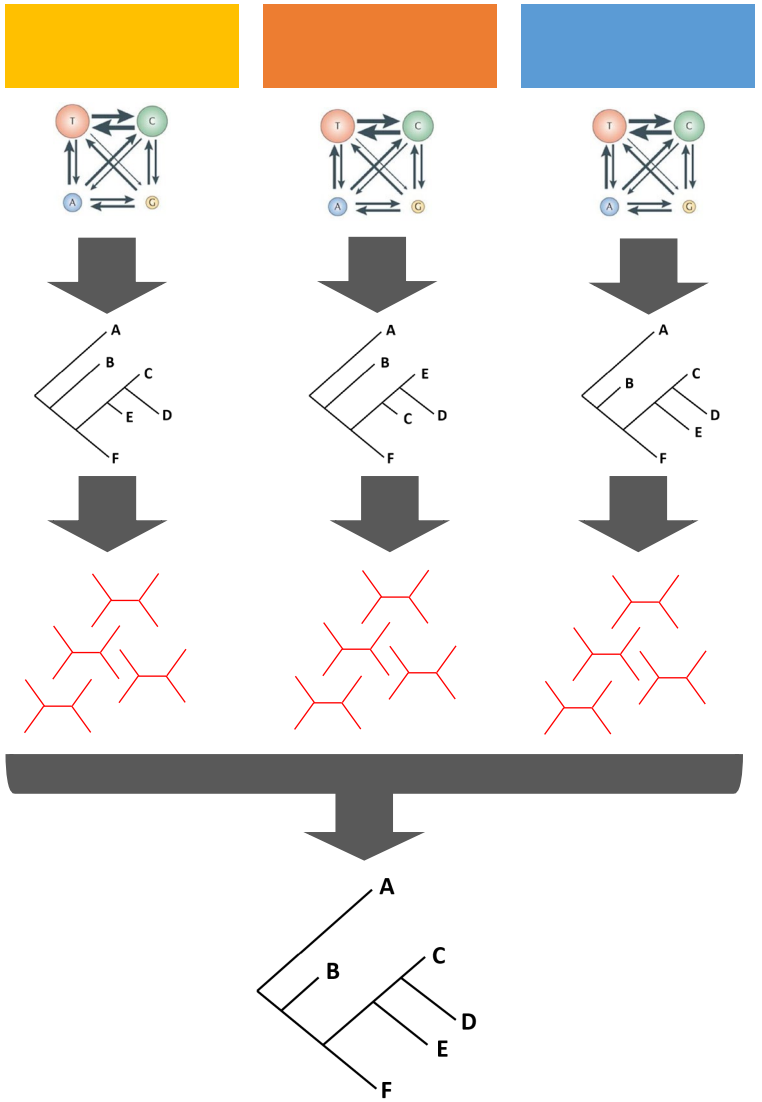
Bayesian Inference of Species Trees from Multilocus Data

Joseph Heled, Alexei J. Drummond Author Notes

Molecular Biology and Evolution, Volume 27, Issue 3, March 2010, Pages 570–

Very slow!
Impossible for hundreds of tips

Multispecies coalescent (MSC) approaches
Quartet-based (summary) methods



Sequence alignments

+

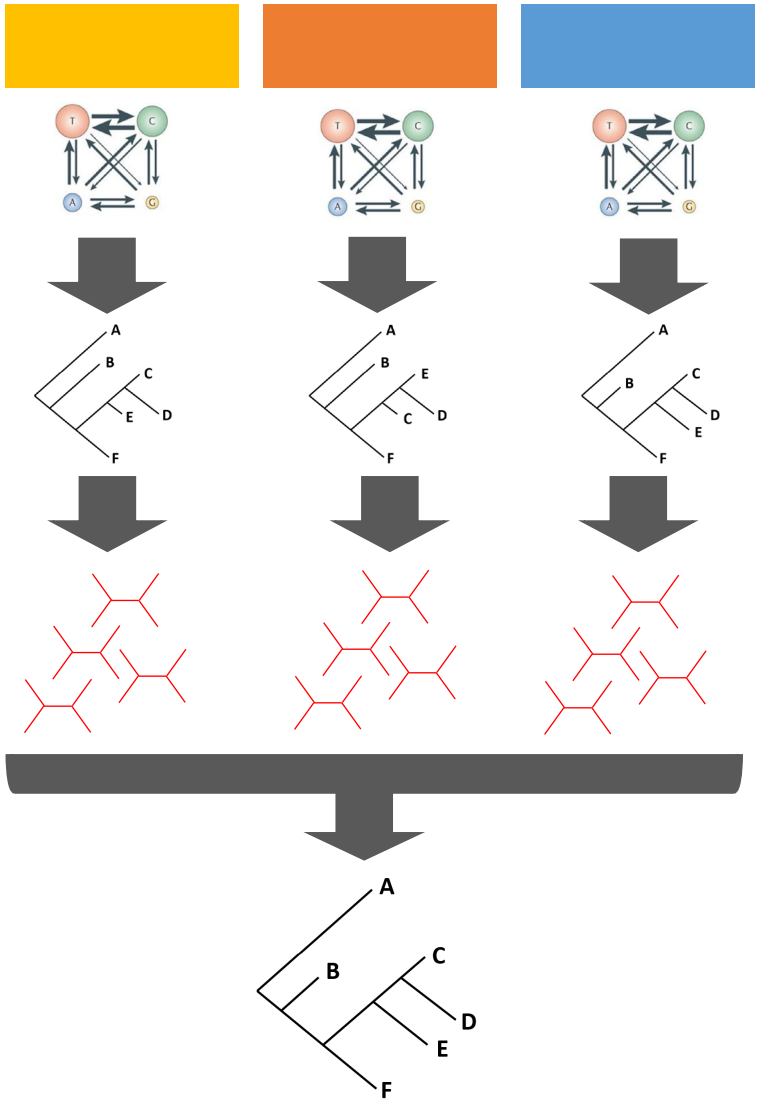
Models of evolution

Phylogenetic inference (ML, BI…)

Gene trees

Quartets extracted from each gene tree

Multispecies coalescent (MSC) approaches
Quartet-based (summary) methods

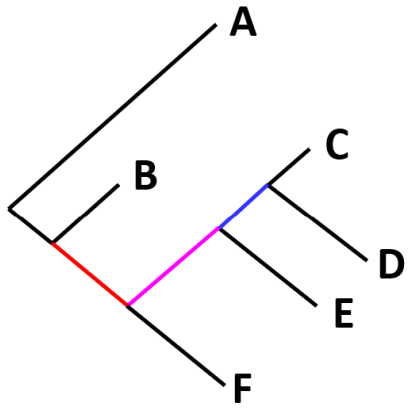


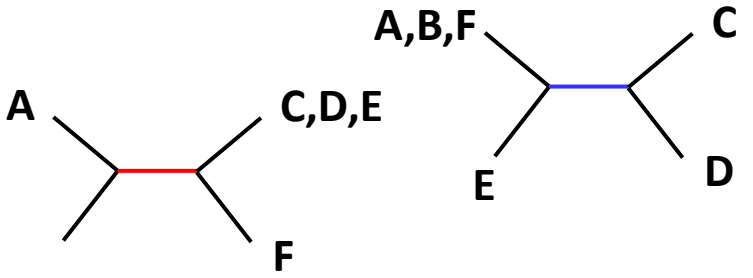
Sequence alignments
+
Models of evolution

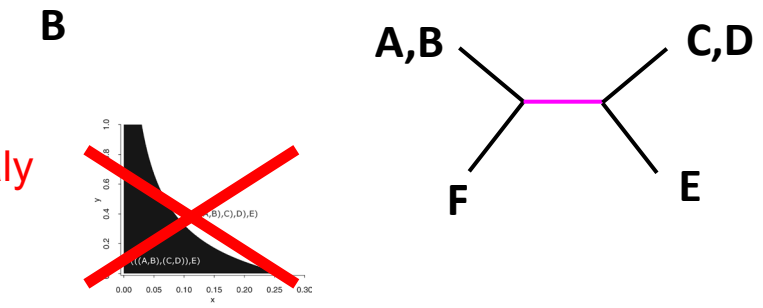
Phylogenetic inference (ML, BI…)

Gene trees

Quartets extracted from each gene tree

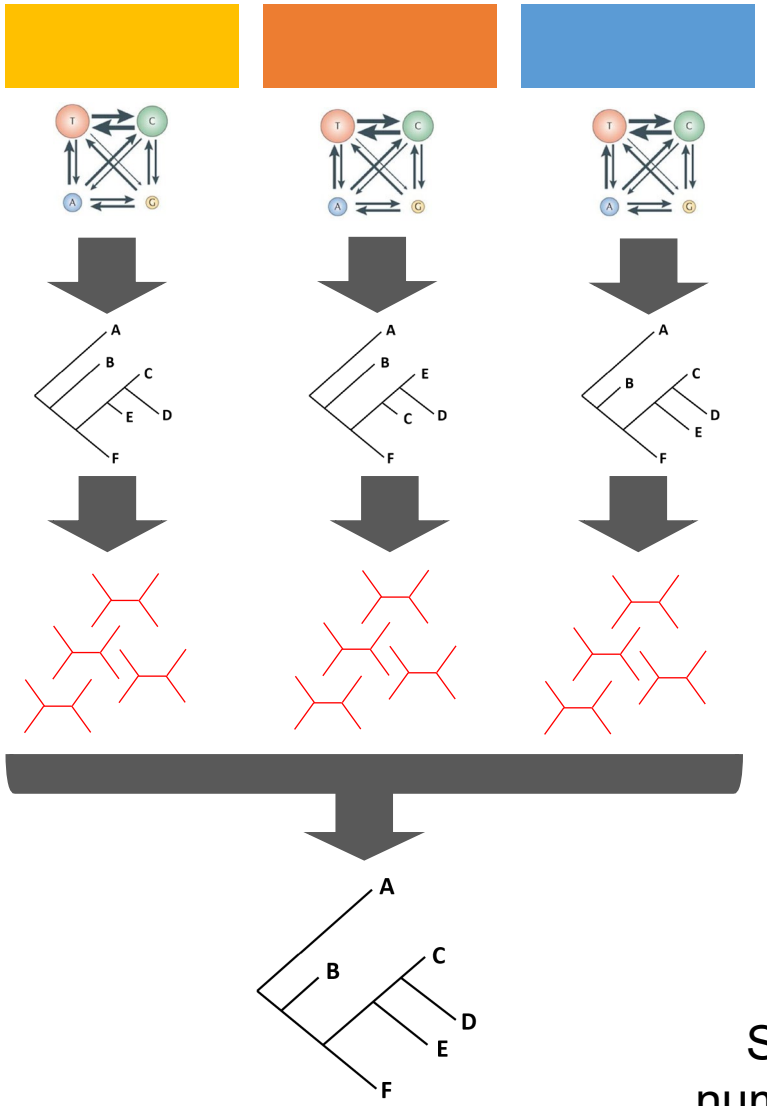






There is no anomaly zone for unrooted quartets! ☺

Multispecies coalescent (MSC) approaches
Quartet-based (summary) methods



Sequence alignments
+
Models of evolution

Phylogenetic inference (ML, BI…)

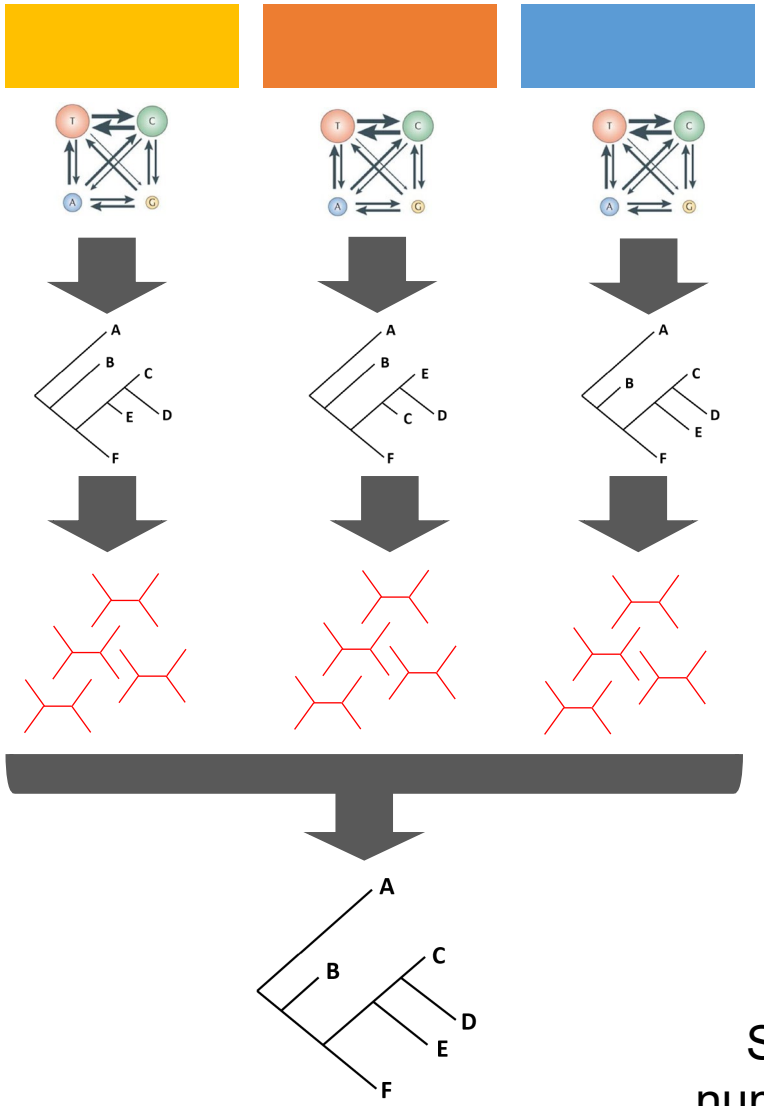
Gene trees

Quartets extracted from each gene tree

Phylogenetic inference statistically consistent with the MSC

Species tree supported by the highest number of quartets found in the gene trees

Multispecies coalescent (MSC) approaches
Quartet-based (summary) methods



Sequence alignments
+
Models of evolution

Phylogenetic inference (ML, BI…)

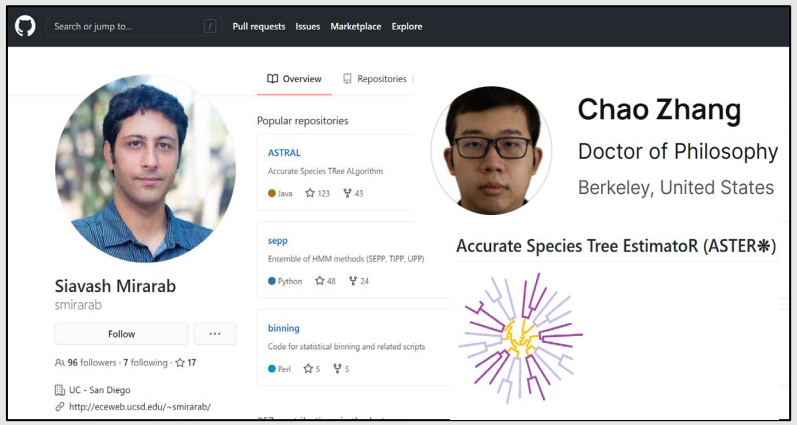
Gene trees

Quartets extracted from each gene tree

Phylogenetic inference statistically consistent with the MSC

Species tree supported by the highest number of quartets found in the gene trees

ASTRAL / ASTER



ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees

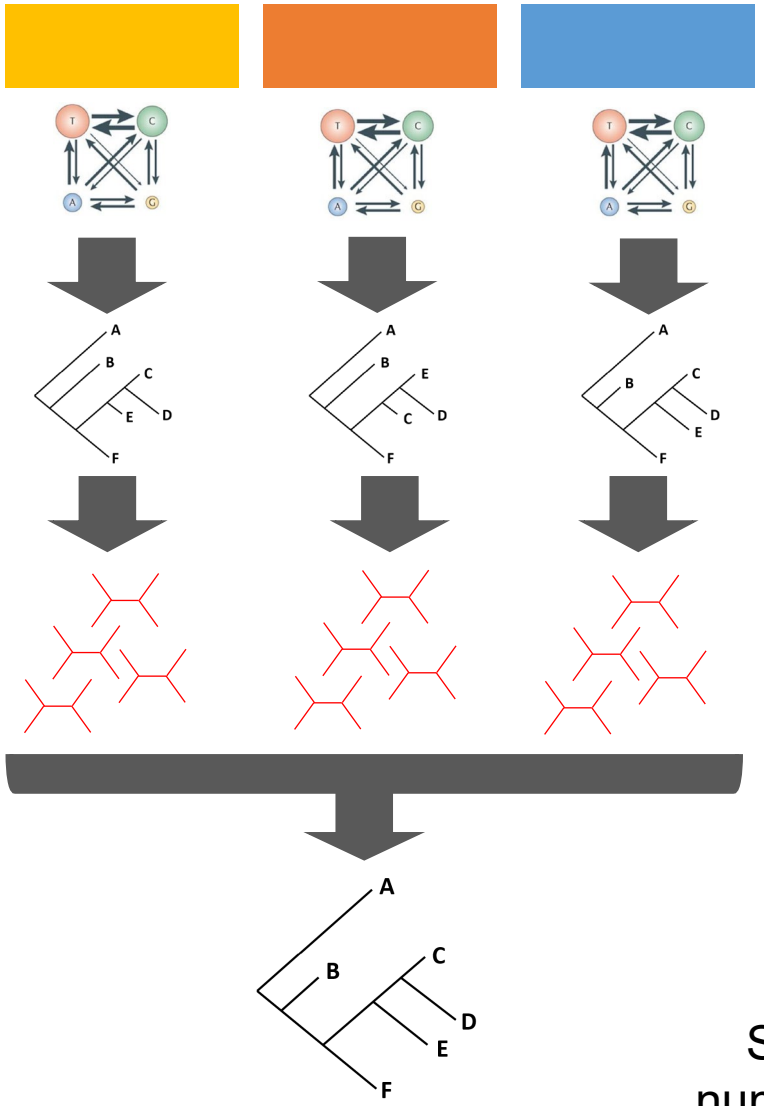
Chao Zhang, Maryam Rabiee, Erfan Sayyari & Siavash Mirarab

BMC Bioinformatics **19**, Article number: 153 (2018) | Cite this article

Commonly used
Quick, accurate
Works with hundreds of tips

Sensitive to gene tree error!

Multispecies coalescent (MSC) approaches
Quartet-based (summary) methods



Sequence alignments
+
Models of evolution

Phylogenetic inference (ML, BI…)

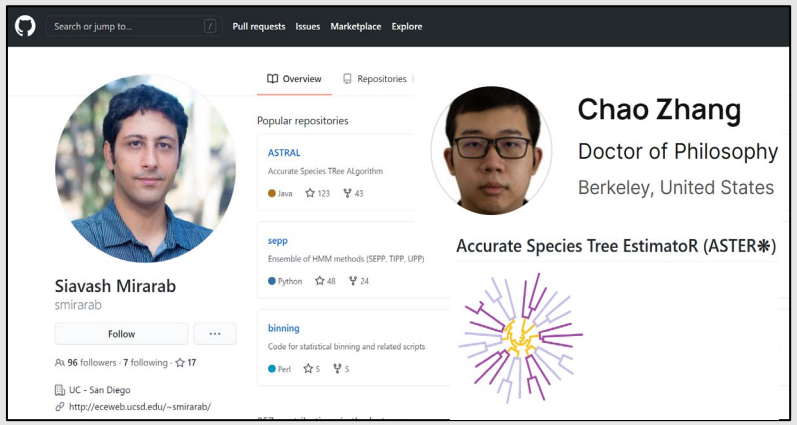
Gene trees

Quartets extracted from each gene tree

Phylogenetic inference statistically consistent with the MSC

Species tree supported by the highest number of quartets found in the gene trees

ASTRAL / ASTER



ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees

Chao Zhang, Maryam Rabiee, Erfan Sayyari & Siavash Mirarab

BMC Bioinformatics **19**, Article number: 153 (2018) | Cite this article

Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees

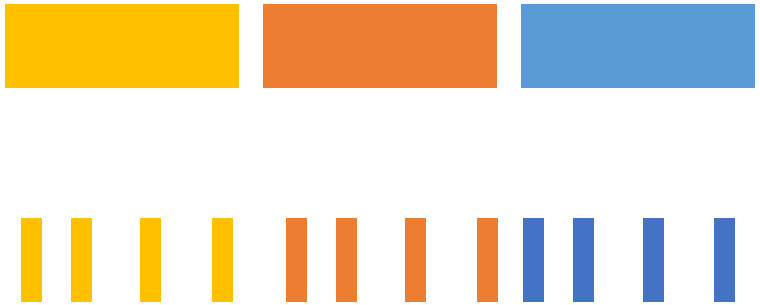
Chao Zhang, Siavash Mirarab

Molecular Biology and Evolution, Volume 39, Issue 12, December 2022, msac215,

Are all gene trees equally trustable?

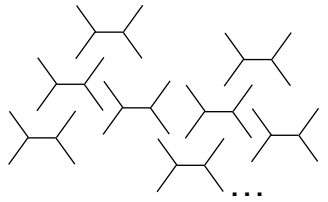
How do we estimate confidence in a tree? In a clade?

Multispecies coalescent approaches – Site-based methods: no reliance on gene tree accuracy

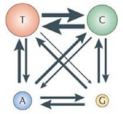


Sequence alignments for each gene

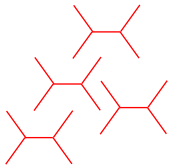
Informative sites only



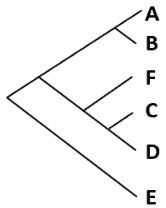
Compute all possible quartets for the taxa



Chose most plausible topology for each quartet (minimize SVD score)



Reconstruct full species tree from selected quartets



Julia Chifman
Assistant Professor
Mathematics & Statistics
Contact
chifman@american.edu
(202) 885-3686
CAS - Math and Statistics
Don Myers Building - 208D
Office Hours: M 4:00 - 5:30; W 11:00 -

SVD quartets promising, studies needed

Laura S. Kubatko
Professor
Departments of Statistics and Evolution, Ecology, and Organismal Biology
Co-Director, Mathematical Biosciences Institute

Quartet Inference from SNP Data Under the Coalescent Model FREE

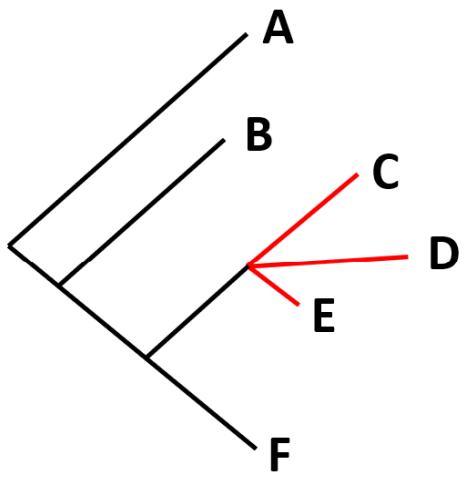
Julia Chifman, Laura Kubatko ✉ Author Notes

Bioinformatics, Volume 30, Issue 23, 1 December 2014, Pages 3317–3324,

Outline

1. Phylogenetics
2. Phylogenomics
3. Approaches
4. Confidence

How much information underlies a gene tree?
Do we need more data to solve a node?
Can we even hope to resolve a node?



Measuring gene(s)' phylogenetic informativeness
can help answering these questions

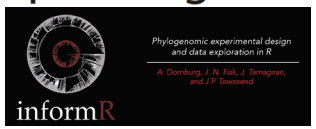
López-Giráldez and Townsend *BMC Evolutionary Biology* 2011, **11**:152
http://www.biomedcentral.com/1471-2148/11/152

BMC
Evolutionary Biology

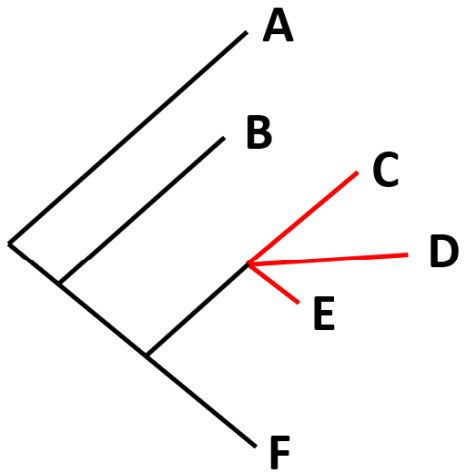
SOFTWARE Open Access

PhyDesign: an online application for profiling phylogenetic informativeness

Francesc López-Giráldez* and Jeffrey P Townsend*



How much information underlies a gene tree?
Do we need more data to solve a node?
Can we even hope to resolve a node?



Measuring gene(s)' phylogenetic informativeness
can help answering these questions

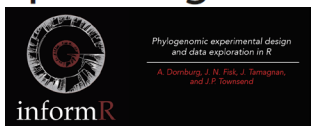
López-Giráldez and Townsend *BMC Evolutionary Biology* 2011, **11**:152
http://www.biomedcentral.com/1471-2148/11/152

BMC Evolutionary Biology

SOFTWARE Open Access

PhyDesign: an online application for profiling phylogenetic informativeness

Francesc López-Giráldez* and Jeffrey P Townsend*



THE ROYAL SOCIETY

Phylogenetic information and experimental design in molecular systematics

Nick Goldman

Syst. Biol. 61(5):835–849, 2012

DOI:10.1093/sysbio/sys036
Advance Access publication on March 3, 2012

Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny

JEFFREY P. TOWNSEND[1,2,*], ZHUO SU[1], AND YONAS I. TEKLE[1,3]

Syst. Biol. 61(5):811–821, 2012

DOI:10.1093/sysbio/sys033
Advance Access publication on February 15, 2012

The Probability of Correctly Resolving a Split as an Experimental Design Criterion in Phylogenetics

EDWARD SUSKO[1,*] AND ANDREW J. ROGER[2]

Points of View

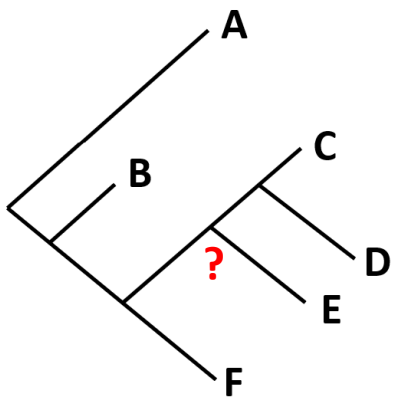
Syst. Biol. 68(1):145–156, 2019

DOI:10.1093/sysbio/syy047
Advance Access publication June 25, 2018

Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets

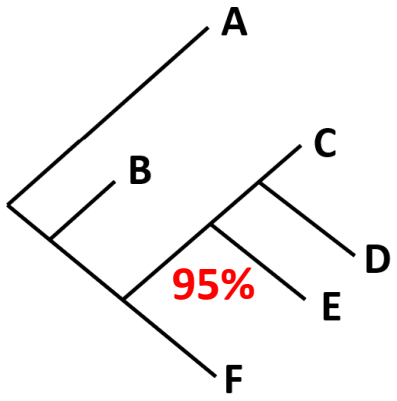
ALEX DORNBURG[1,*], ZHUO SU[2], AND JEFFREY P. TOWNSEND[2,3,4]

How do we measure our confidence in phylogenetic results?



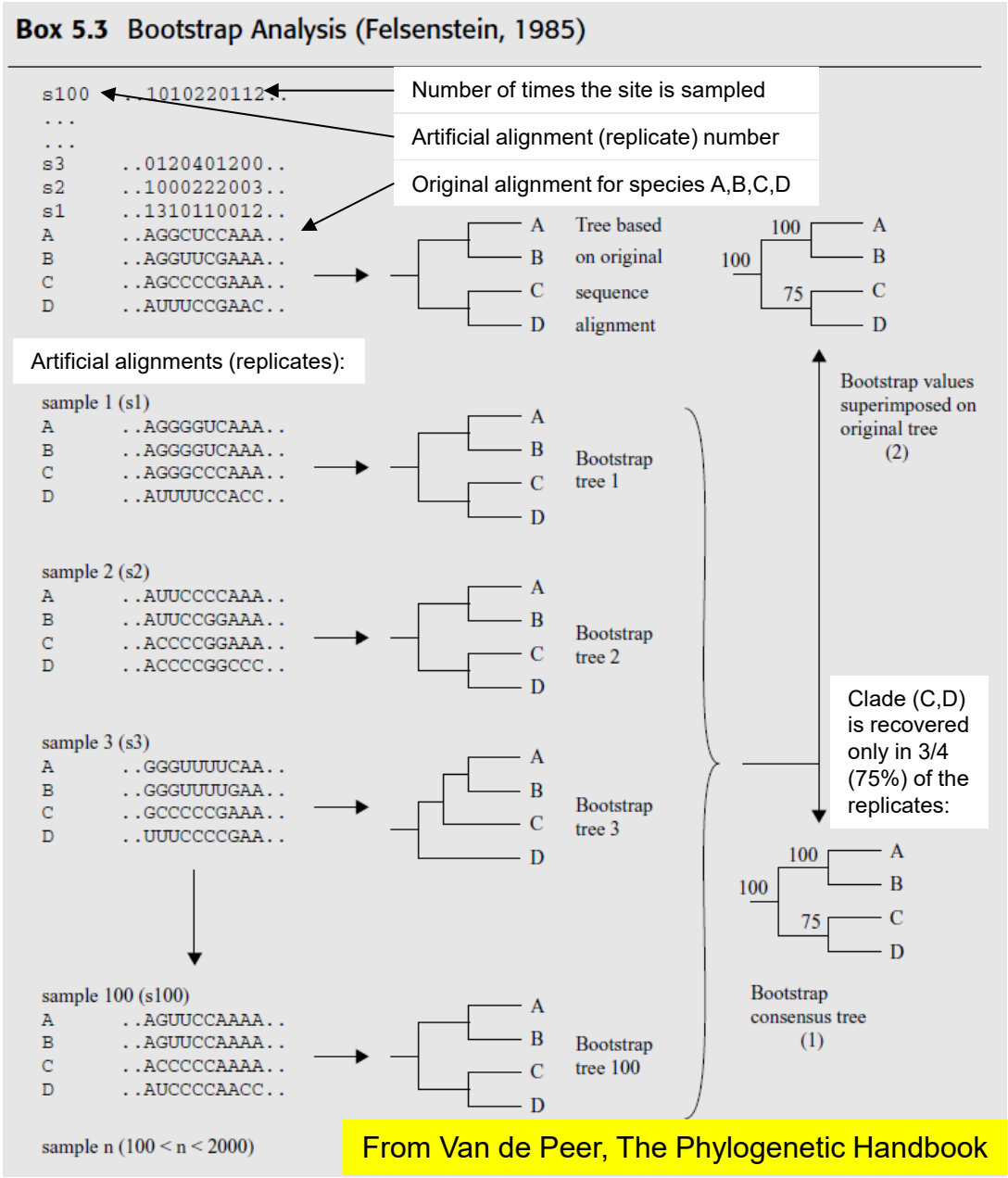
- Bootstrap support
- Posterior probability
- Concordance gene trees vs species trees

How do we measure our confidence in phylogenetic results?



- Bootstrap support

(most often for Maximum parsimony or Maximum Likelihood)



Bootstrap analysis

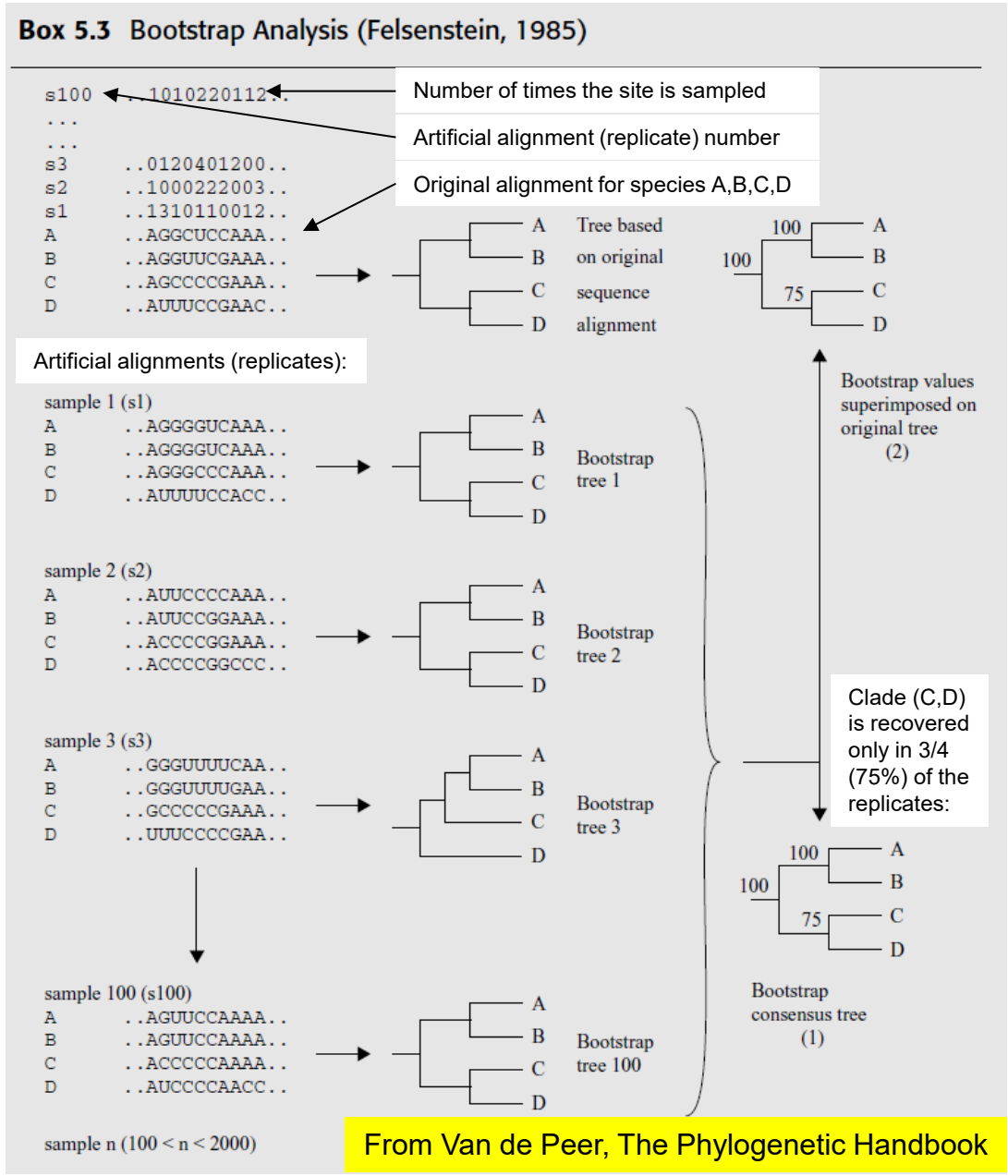
- **Generate many artificial alignments** of the same length as the original by sampling sites from the original alignment with repetition allowed.

- **Make a tree for each artificial alignment** using the same method and models as the ones used to make the tree based on the original alignment.

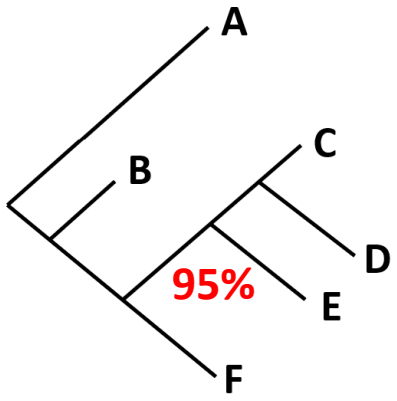
- **Count how many times a clade in the "real" tree occurs in the bootstrap trees.**

OR

Make a consensus of the bootstrap trees and **count how many times a clade in the consensus tree occurs in the bootstrap trees.**



How do we measure our confidence in phylogenetic results?

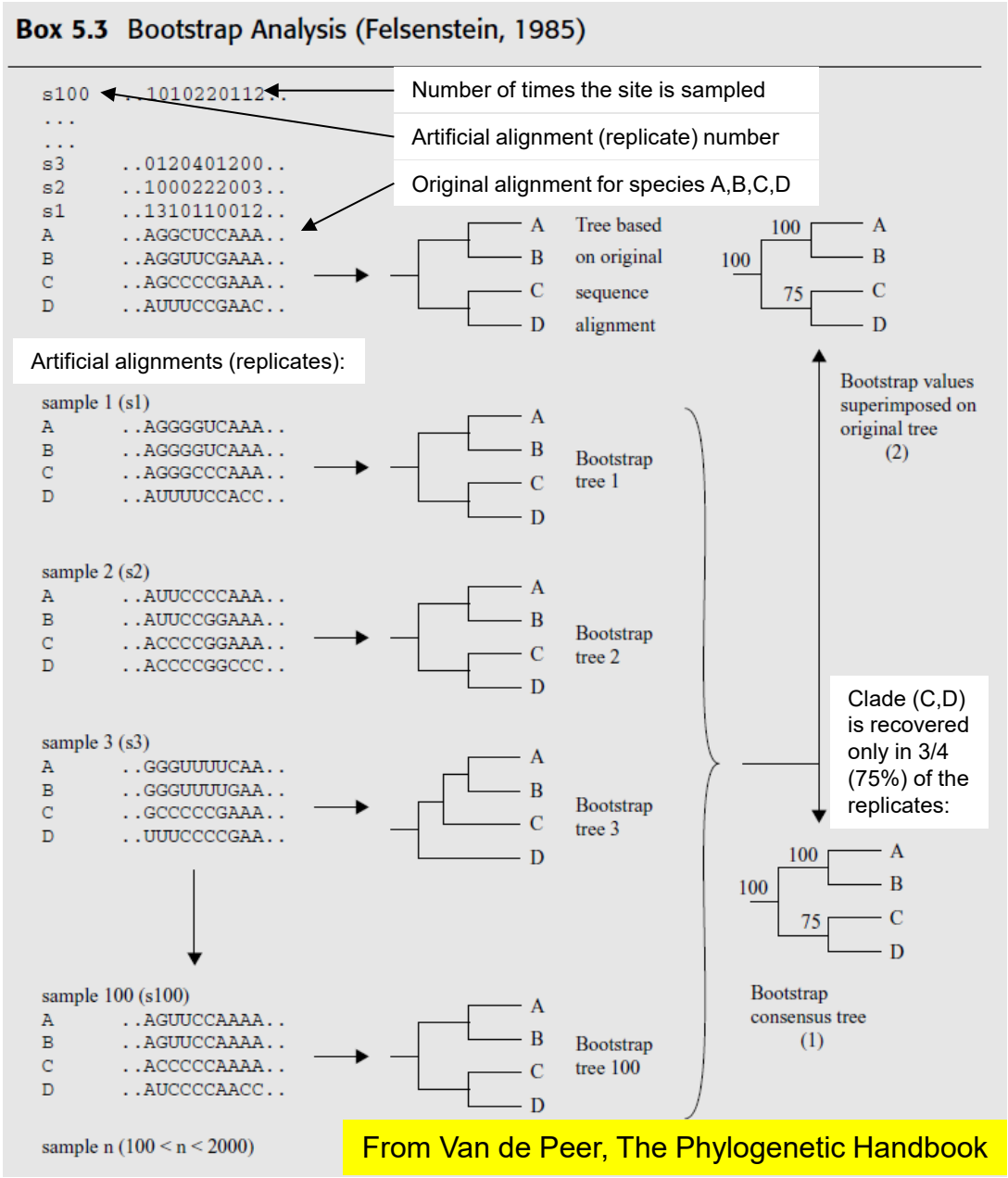


- Bootstrap support

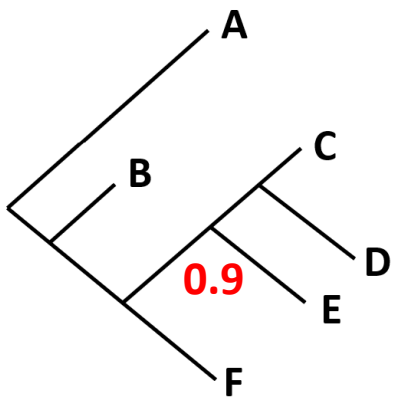
(most often for Maximum parsimony or Maximum Likelihood)

Bootstrap values are often given as percentages
Usually, values under 70% should be taken with caution

BUT even high values can be misleading!
If bias in sequences, bias in bootstrap values!
(E.g. Long Branch Attraction or base frequencies biases)



How do we measure our confidence in phylogenetic results?



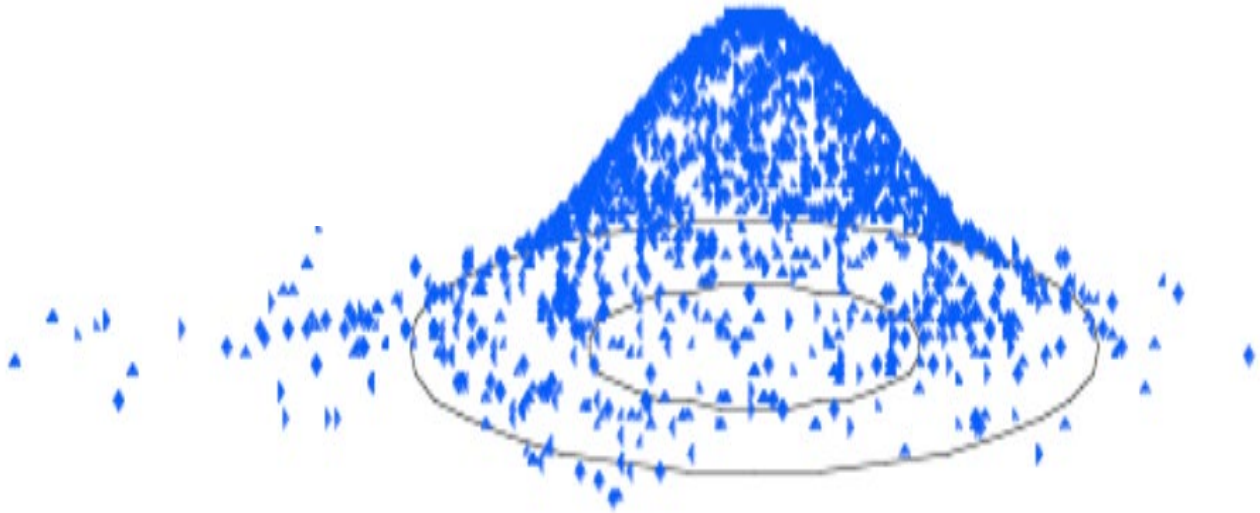
- Posterior probability

(Bayesian inference)

Probability that the clade is correct given the data and model

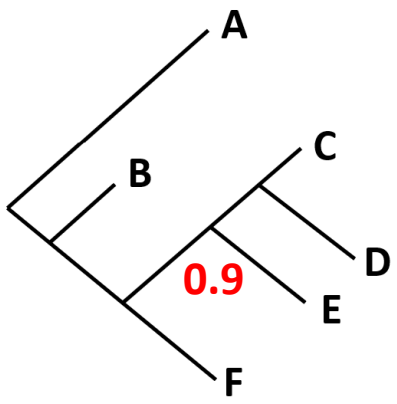
Estimated from the proportion of the clade in the sample of trees selected by the Bayesian Inference (MCMC) algorithm

Although they are proportions, posterior probabilities are estimates of probabilities, so they are provided as a number between 0 and 1.



Adapted from http://carrot.mcb.uconn.edu/~olgazh/bioinf2010/class31.html

How do we measure our confidence in phylogenetic results?



- Posterior probability

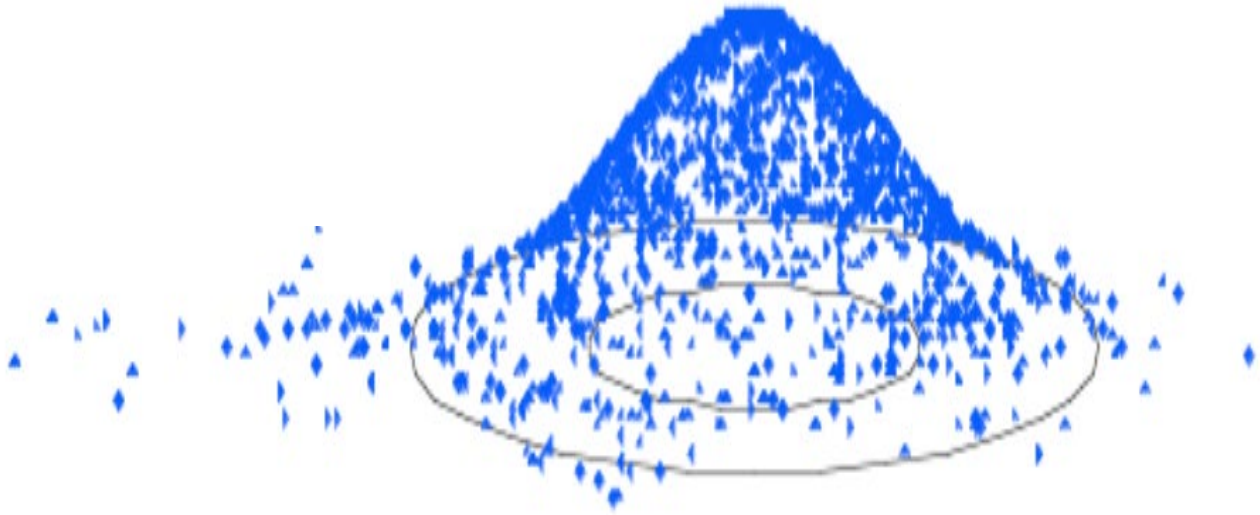
(Bayesian inference)

Posterior probabilities tend to be inflated, so PP < 0.95 should be taken with caution!

Probability that the clade is correct given the data and model

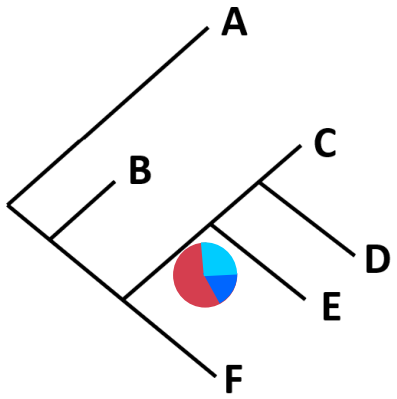
Estimated from the proportion of the clade in the sample of trees selected by the Bayesian Inference (MCMC) algorithm

Although they are proportions, posterior probabilities are estimates of probabilities, so they are provided as a number between 0 and 1.

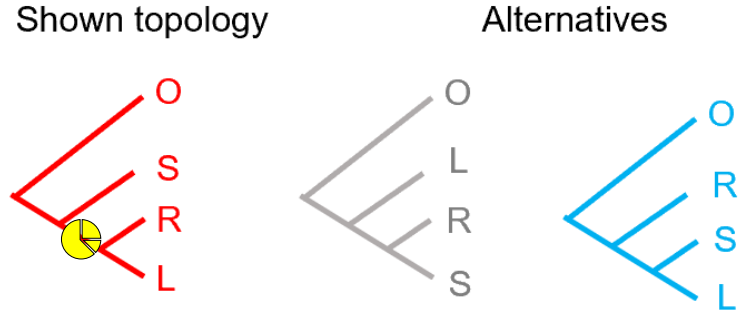


Adapted from http://carrot.mcb.uconn.edu/~olgazh/bioinf2010/class31.html

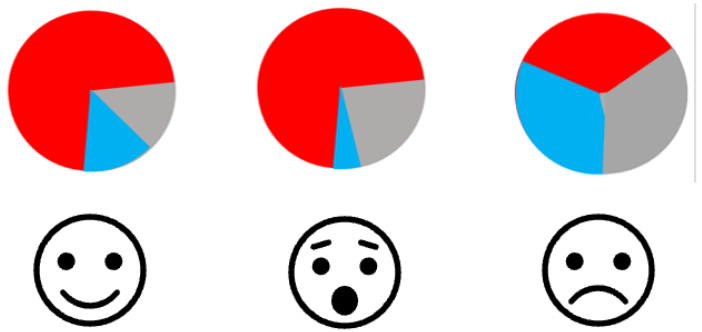
How do we measure our confidence in phylogenetic results?



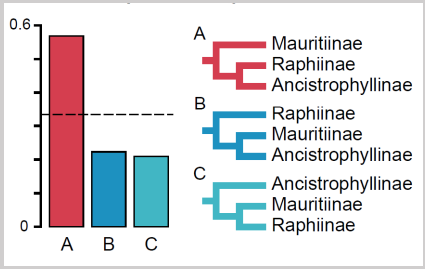
- Concordance gene trees vs species trees



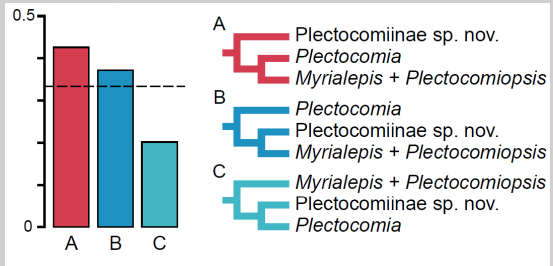
Examples of support:



Analysis of gene tree frequencies
Species tree topology vs alternatives:



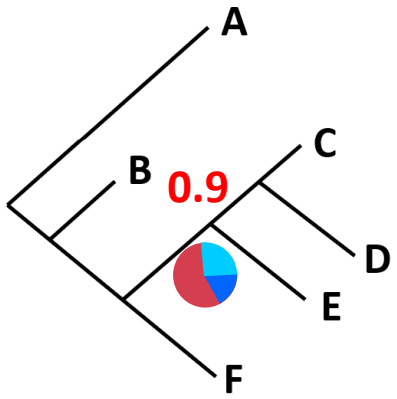
Similar frequencies of the alternative topologies
→ ILS?



One alternative topology more represented
→ Hybridisation?

Example taken from Kuhnhaeuser et al., 2020, MPE

How do we measure our confidence in phylogenetic results?



- Concordance gene trees vs species trees

Topology frequencies (quartets, bipartitions)
Gene concordance factor
Site concordance factor
Local posterior probabilities (ASTRAL)
…

ELSEVIER
Molecular Phylogenetics and Evolution
Volume 122, May 2018, Pages 110-115

Short Communication

DiscoVista: Interpretable visualizations of gene tree discordance

Erfan Sayyari [a], James B. Whitfield [b], Siavash Mirarab [a]

Stephen Smith phyparts

New Methods to Calculate Concordance Factors for Phylogenomic Datasets

Bui Quang Minh, Matthew W Hahn, Robert Lanfear

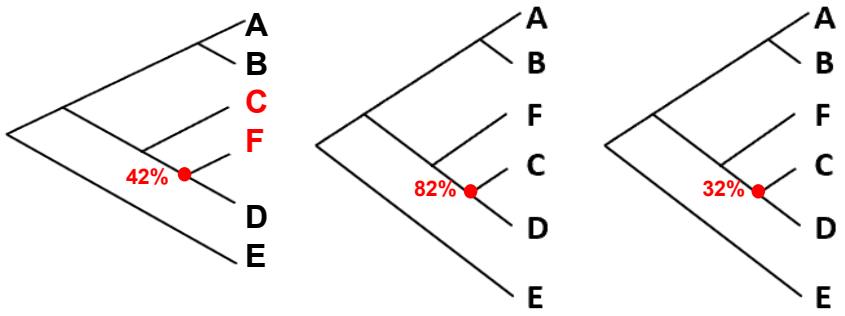
Molecular Biology and Evolution, Volume 37, Issue 9, September 2020, Pages

Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies

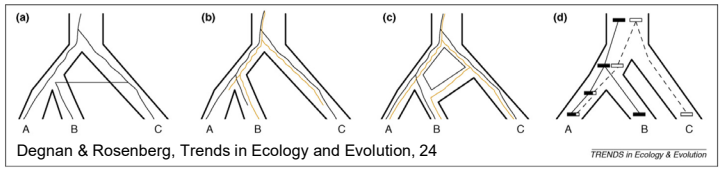
Erfan Sayyari, Siavash Mirarab Author Notes

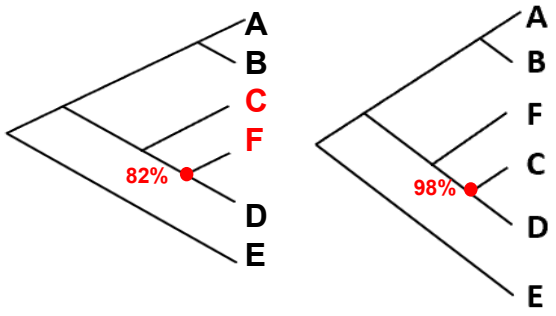
Molecular Biology and Evolution, Volume 33, Issue 7, July 2016, Pages 1654–

- Conflicting gene trees with low support may only indicate a lack of phylogenetic signal



- Conflicting gene trees with high support may indicate biological causes
  (e.g. hybridization, ILS, horizontal transfer)





→ Looking at support values and the phylogenetic informativeness of genes before and after doing the species tree can help focusing on informative genes and identifying biologically meaningful conflicts

To learn more…

Trends in Ecology & Evolution
Volume 24, Issue 6, June 2009, Pages 332-340

Review

Gene tree discordance, phylogenetic inference and the multispecies coalescent

James H. Degnan [1, 2] ✉, Noah A. Rosenberg [1, 3, 4] ✉

PeerJ Life & Environment

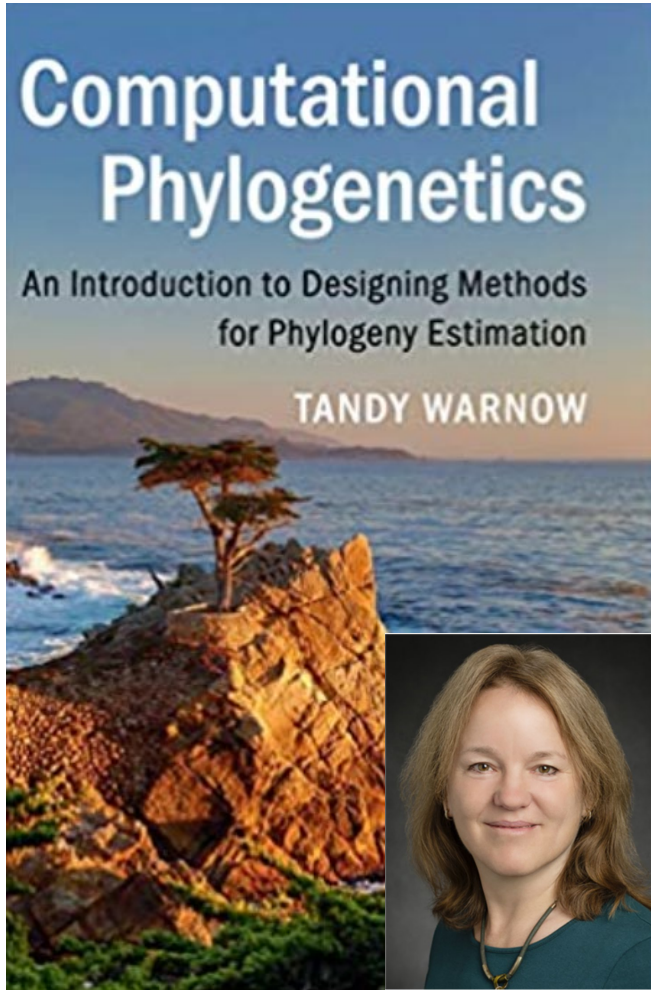
‹ BIOINFORMATICS AND GENOMICS

Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics

Literature review | Biodiversity | Computational Biology | Evolutionary Studies | Genomics

Gustavo A. Bravo [1], Alexandre Antonelli [1, 2, 3, 4], Christine D. Bacon [2, 3], Krzysztof Bartoszek [5], Mozes P. K. Blom [6], Stella Huynh [7], Graham Jones [3], L. Lacey Knowles [8], Sangeet Lamichhaney [1], Thomas Marcussen [9], Hélène Morlon [10], Luay K. Nakhleh [11], Bengt Oxelman [2, 3], Bernard Pfeil [3], Alexander Schliep [12], Niklas Wahlberg [13], Fernanda P. Werneck [14], John Wiedenhoeft [12, 15], Sandi Willows-Munro [16], Scott V. Edwards [1, 17]

Published February 14, 2019



More genes = More problems + More insights

- Phylogenomics allow to study many DNA characters, solving polytomies
- Phylogenomics also reveal conflicts between gene trees
- Conflicts can result from lack of phylogenetic signal
- Conflicts can also result from biological phenomena (e.g. ILS, HGT, hybridisation, duplications…)
- Informative conflicts can be studied to uncover new aspects of species history
- A network or multiple trees may represent species history better than a single tree!