Recent advances in rooted phylogenetic networks: the long road to explicit hypothesis generation

Steven Kelk, Department of Knowledge Engineering (DKE), Maastricht University (NL)

Outline of talk

1. Introduction to rooted/evolutionary phylogenetic networks
2. In how far have these models been seriously used for hypothesis generation and testing?
3. Conclusions

Part 1.

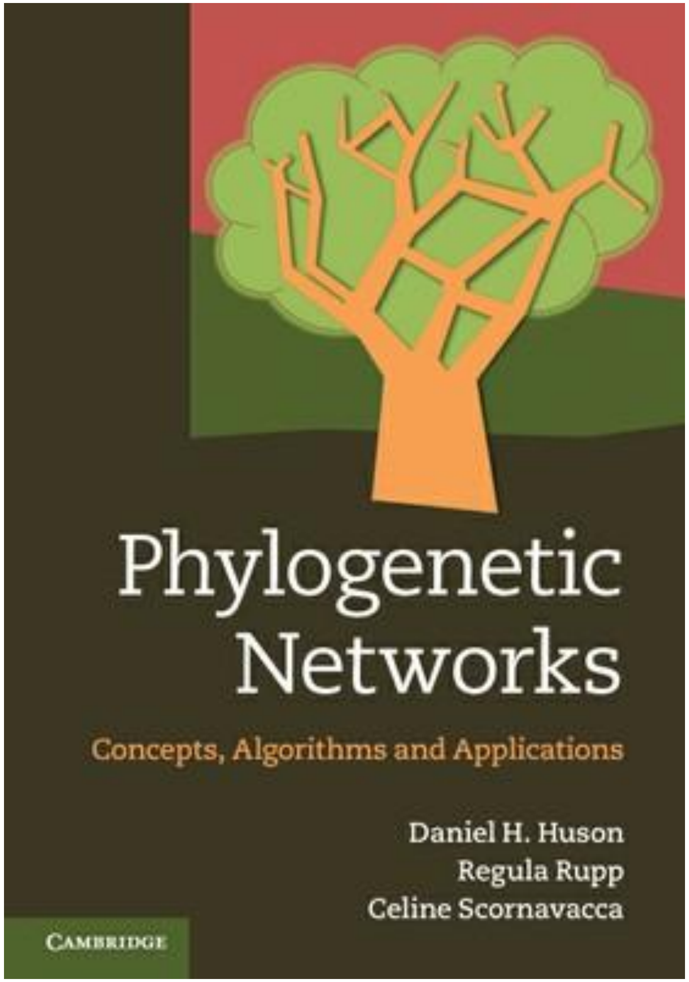
Rooted/evolutionary phylogenetic networks

“Phylogenetic networks”

- Phylogenetic networks generalise phylogenetic trees

- The description “phylogenetic network” is a source of considerable confusion…
  - It suggests that some models that are fundamentally different, are the same ☹
  - It suggests that some models that are actually very similar, are different ☹

- What unifies the models, however, is the idea that it is sometimes neither possible nor desirable to seek a single tree hypothesis to explain observed biological data

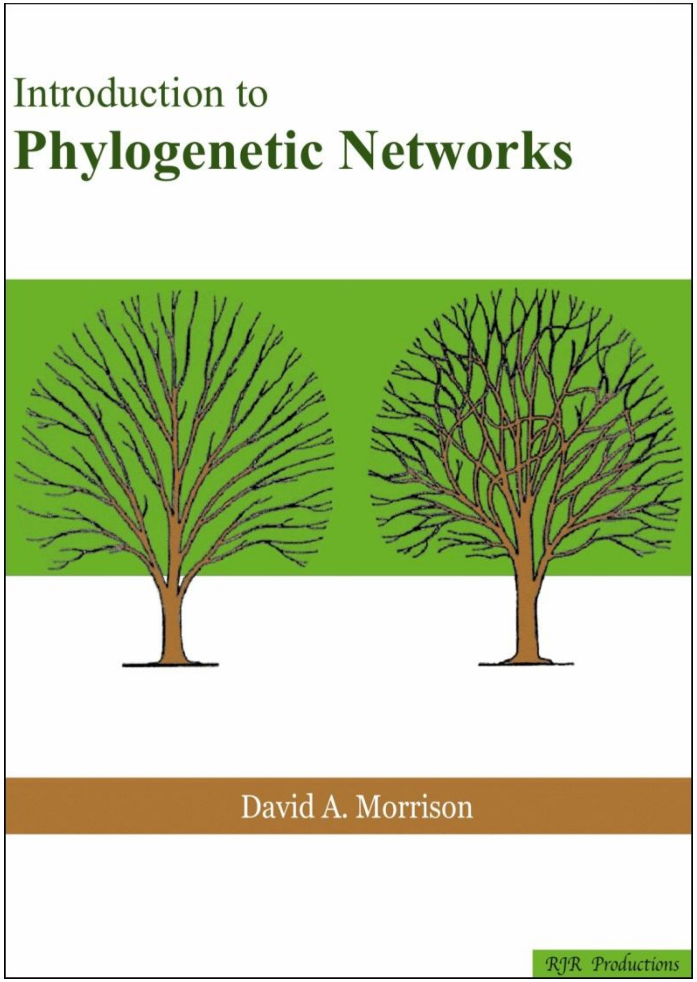


Phylogenetic Networks

Concepts, Algorithms and Applications

Daniel H. Huson
Regula Rupp
Celine Scornavacca

CAMBRIDGE

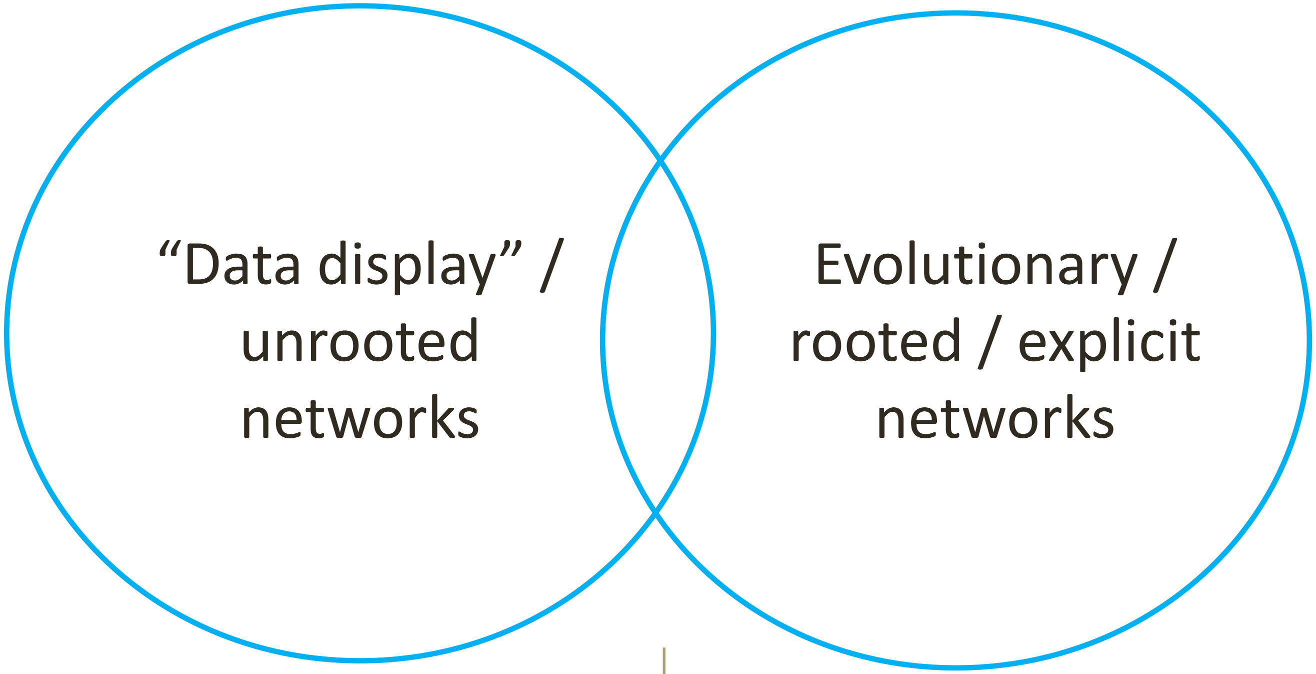


Introduction to
Phylogenetic Networks

David A. Morrison

RJR Productions

Phylogenetic networks: 2 types



"Data display" / unrooted networks

Evolutionary / rooted / explicit networks

No (explicit) model of evolution: tries to graphically represent **where** the data is non-treelike.

Does not generate a hypothesis of "what happened".

Tries to model the **events** that caused the data to be non-treelike.

Tries – in some limited way – to generate a hypothesis of "what happened".

Briefly: data-display networks

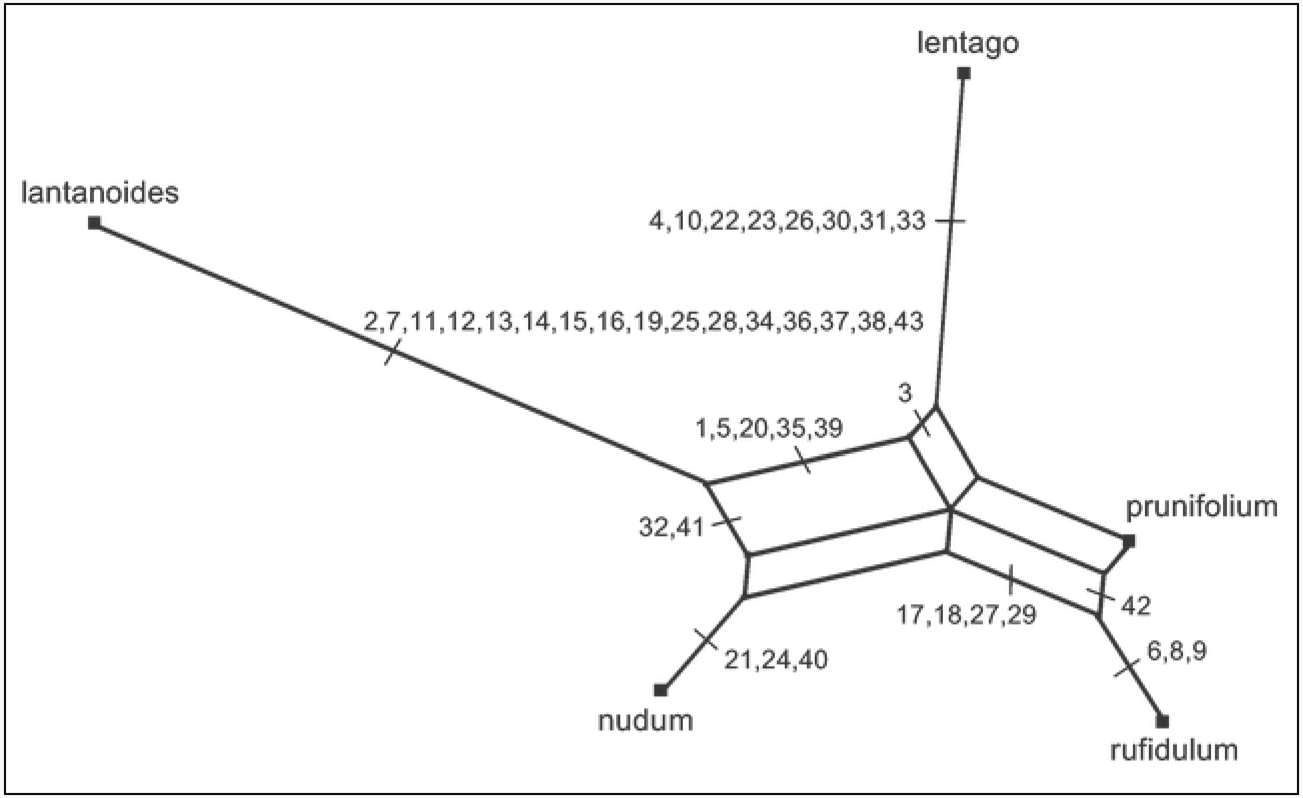


Figure 6. The Median Network for the *Viburnum* sequence, showing the edges (or sets of parallel edges) associated with each of the 43 characters.

Briefly: data-display networks

- In practice data-display phylogenetic networks are still used more than evolutionary phylogenetic networks.
- Why? Because they let the biologist *explore* the data, and to draw his/her own conclusions. They do not impose a (probably controversial) *hypothesis* on the biologist.

Evolutionary phylogenetic networks

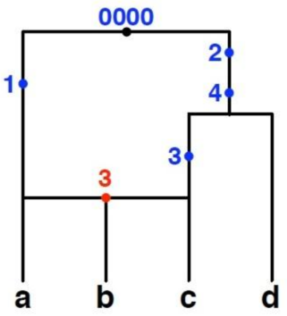
- A rooted phylogenetic tree can be viewed as a hypothesis about when and where vertical evolutionary phenomena (e.g. speciation, mutation) occurred.

- Evolutionary phylogenetic networks extend this to include horizontal ("reticulate") evolutionary phenomena, e.g.
  - Horizontal gene transfer (HGT)
  - Hybridization
  - Recombination

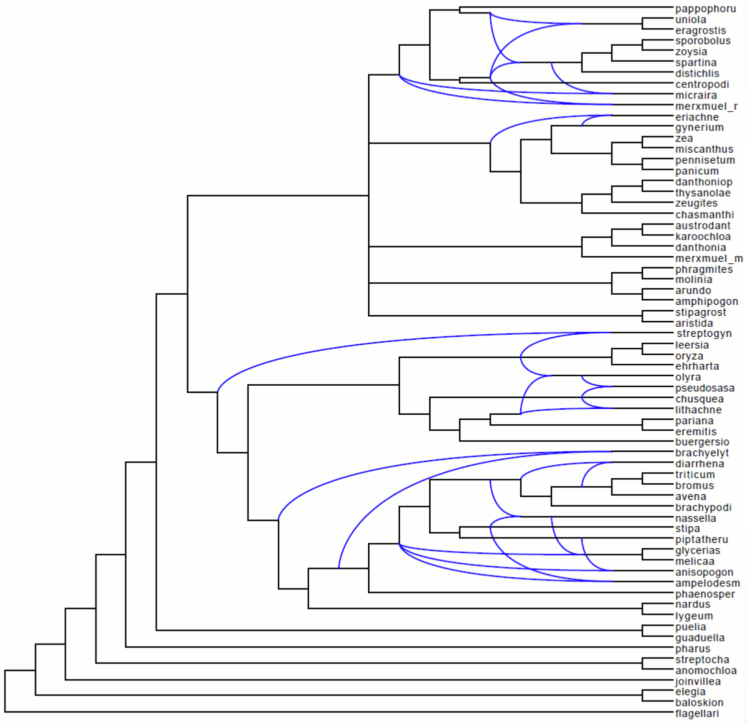
- Often modelled as rooted, directed acyclic graphs, which extend trees to also allow vertices with indegree 2 or higher: reticulations

Evolutionary phylogenetic networks: many different models and evolutionary scales…

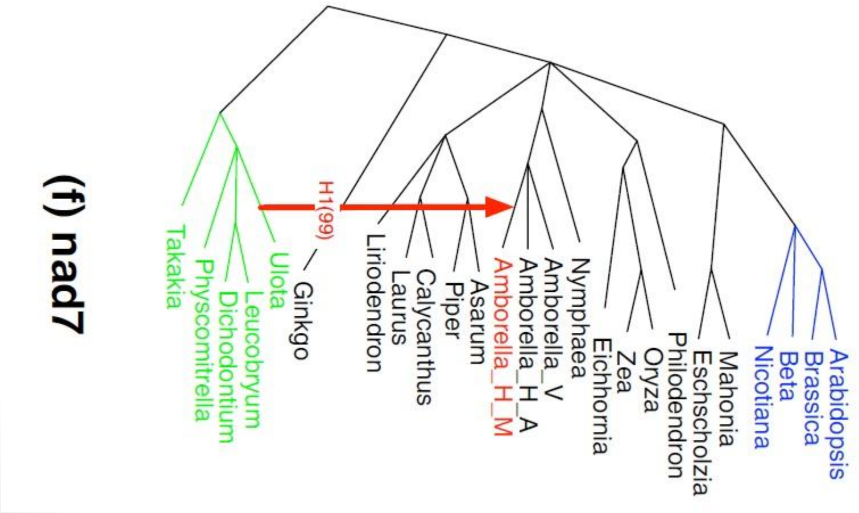
	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1



Ancestral Recombination Graph (ARG)



“Hybridization” network



Horizontal Gene Transfer (HGT)

Why? Reticulation exists!

- It is well known that reticulate events (especially HGT) are influential in the evolution of prokaryotes.

- But it is also becoming clear that hybridization (and even HGT) has a role within eukaryotic evolution.

Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee

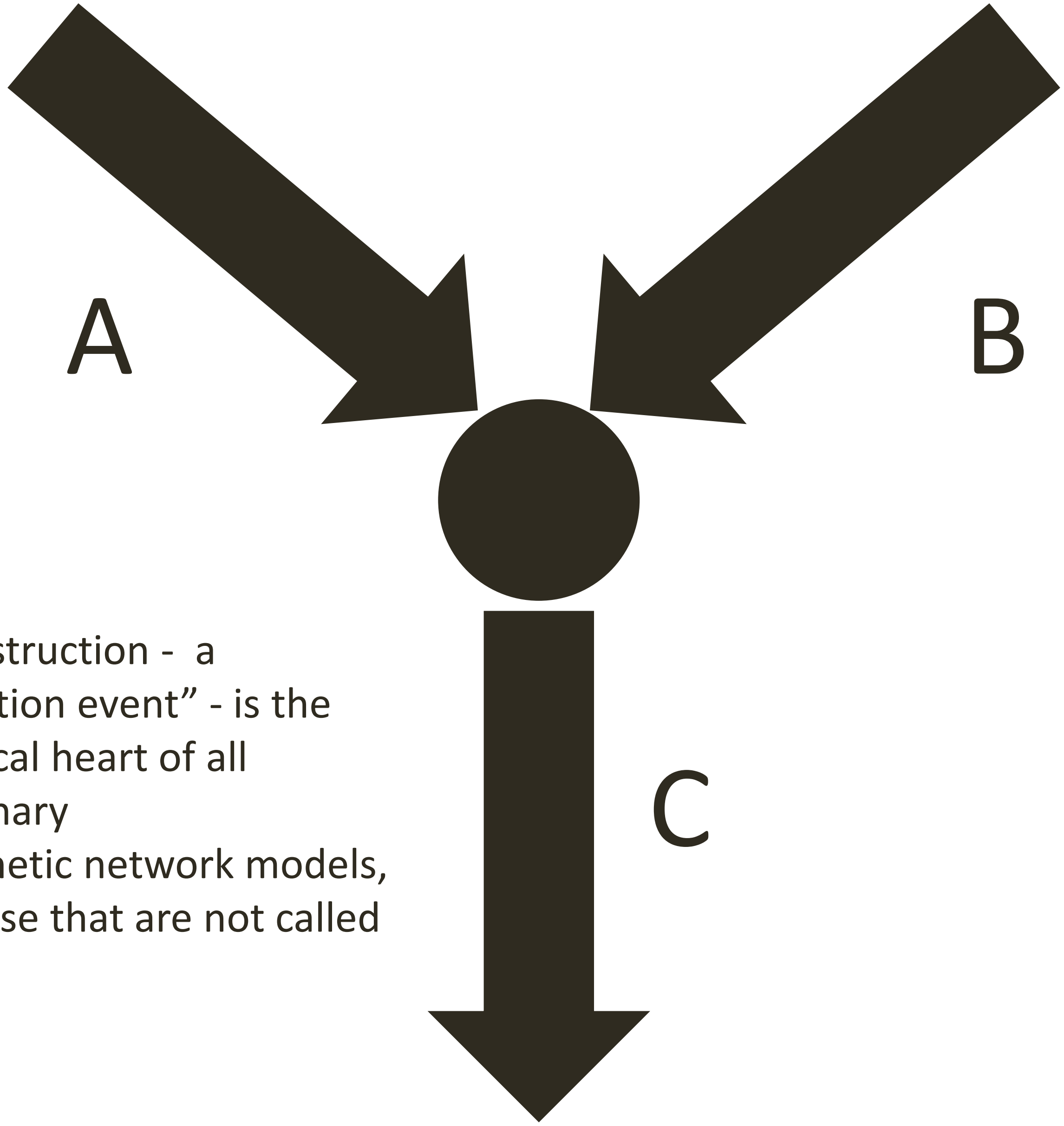
Ricardo Acuña[a,1], Beatriz E. Padilla[a,1], Claudia P. Flórez-Ramos[a], José D. Rubio[a], Juan C. Herrera[a], Pablo Benavides[a], Sang-Jik Lee[b,c], Trevor H. Yeats[b], Ashley N. Egan[b,d], Jeffrey J. Doyle[b], and Jocelyn K. C. Rose[b,2]

[a]Plant Breeding, Biotechnology, and Entomology Departments, Cenicafé, A.A. 2427 Manizales, Colombia; [b]Department of Plant Biology, Cornell University, Ithaca, NY 14853; [c]Biotechnology Institute, Nongwoo Bio Co., Ltd., Gyeonoggi 469-885, Korea; and [d]Department of Biology, East Carolina University, Greenville, NC 27858

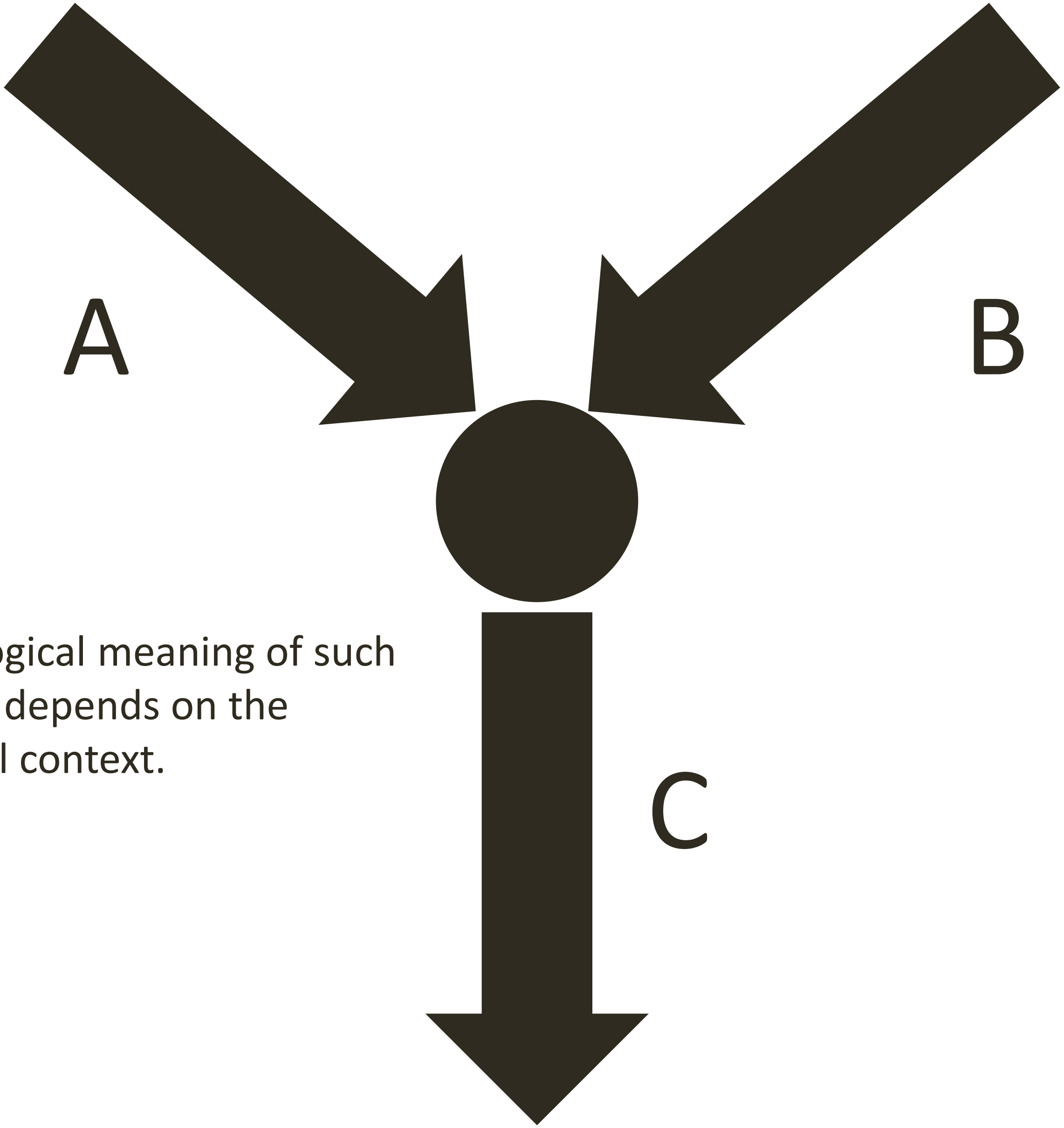
Edited by Nancy A. Moran, Yale University, West Haven, CT, and approved February 1, 2012 (received for review December 21, 2011)

Why? Reticulation exists!

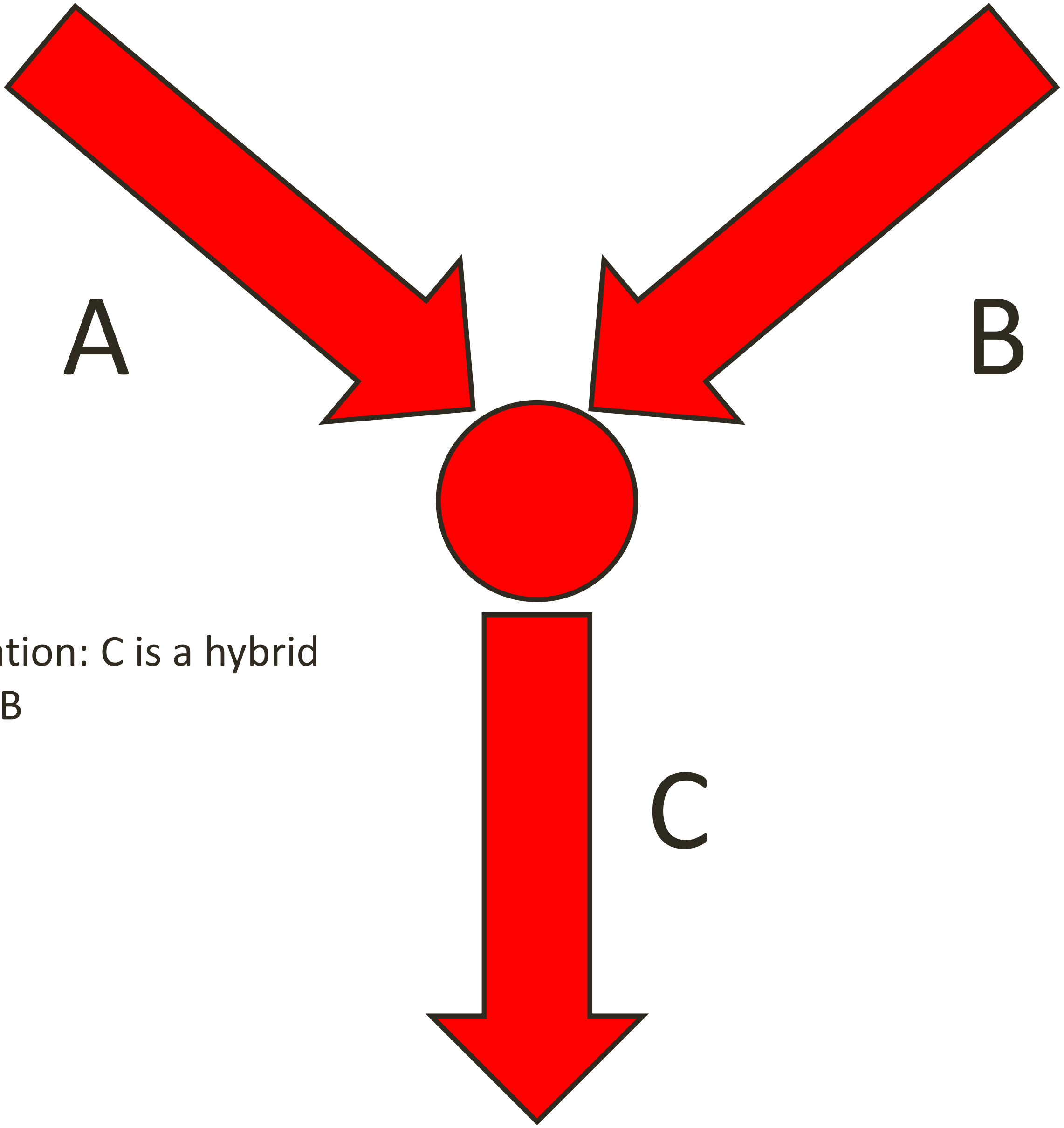
- It is well known that reticulate events (HGT) are influential in the evolution of prokaryotes.
- But it is also becoming clear that hybridization (and even HGT) has a role within eukaryotic evolution.
- The “omics” approach – collect more and more data to try and resolve controversial clades – will not work if the signal is fundamentally non-treelike. If anything, the more data we collect, the more we are confronted with reticulation.
- You can always build a tree, but if the signal is not treelike then the tree will be a potentially meaningless mathematical average of many incongruous tree signals.



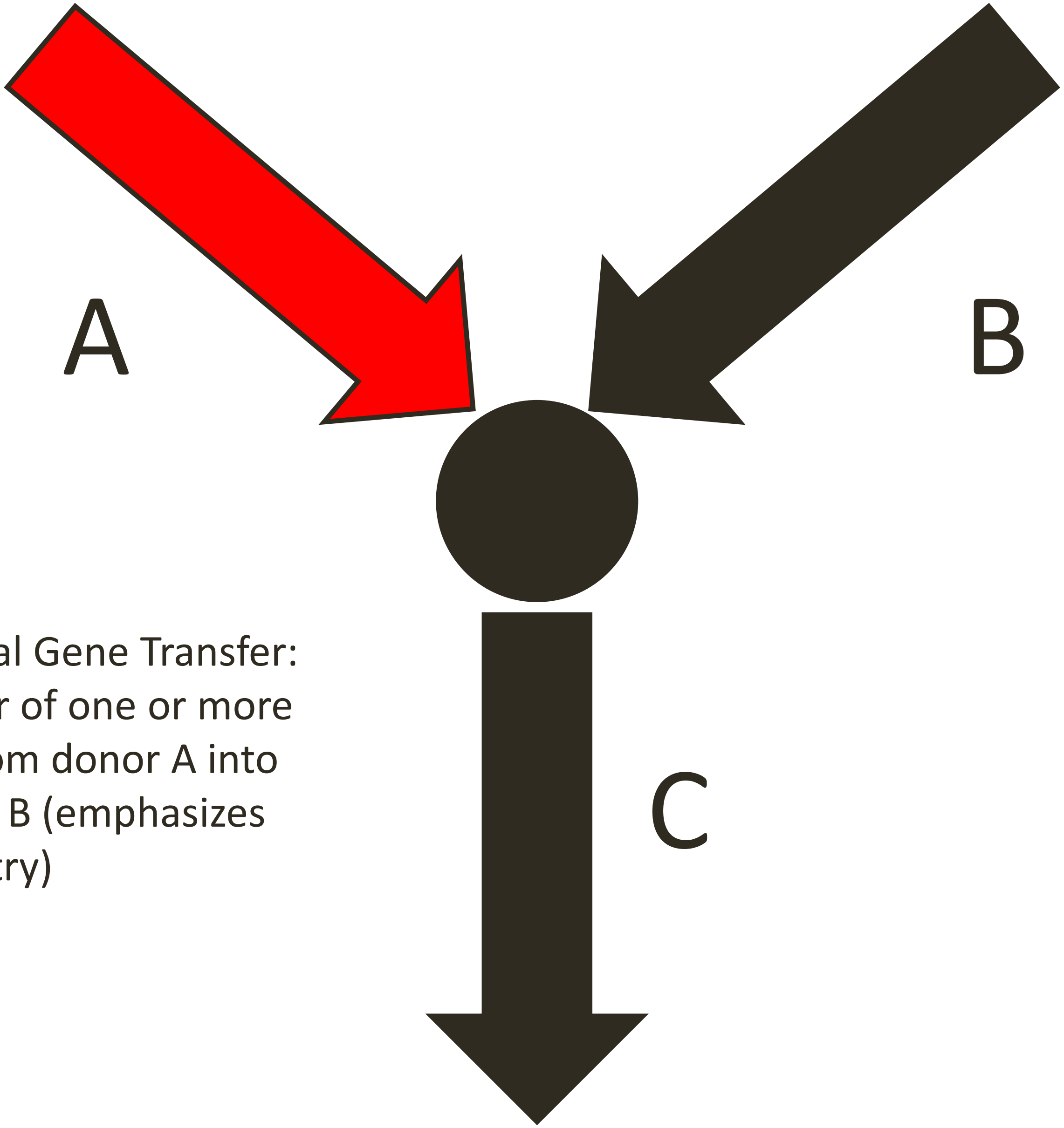
This construction - a “reticulation event” - is the topological heart of all evolutionary phylogenetic network models, even those that are not called as such…



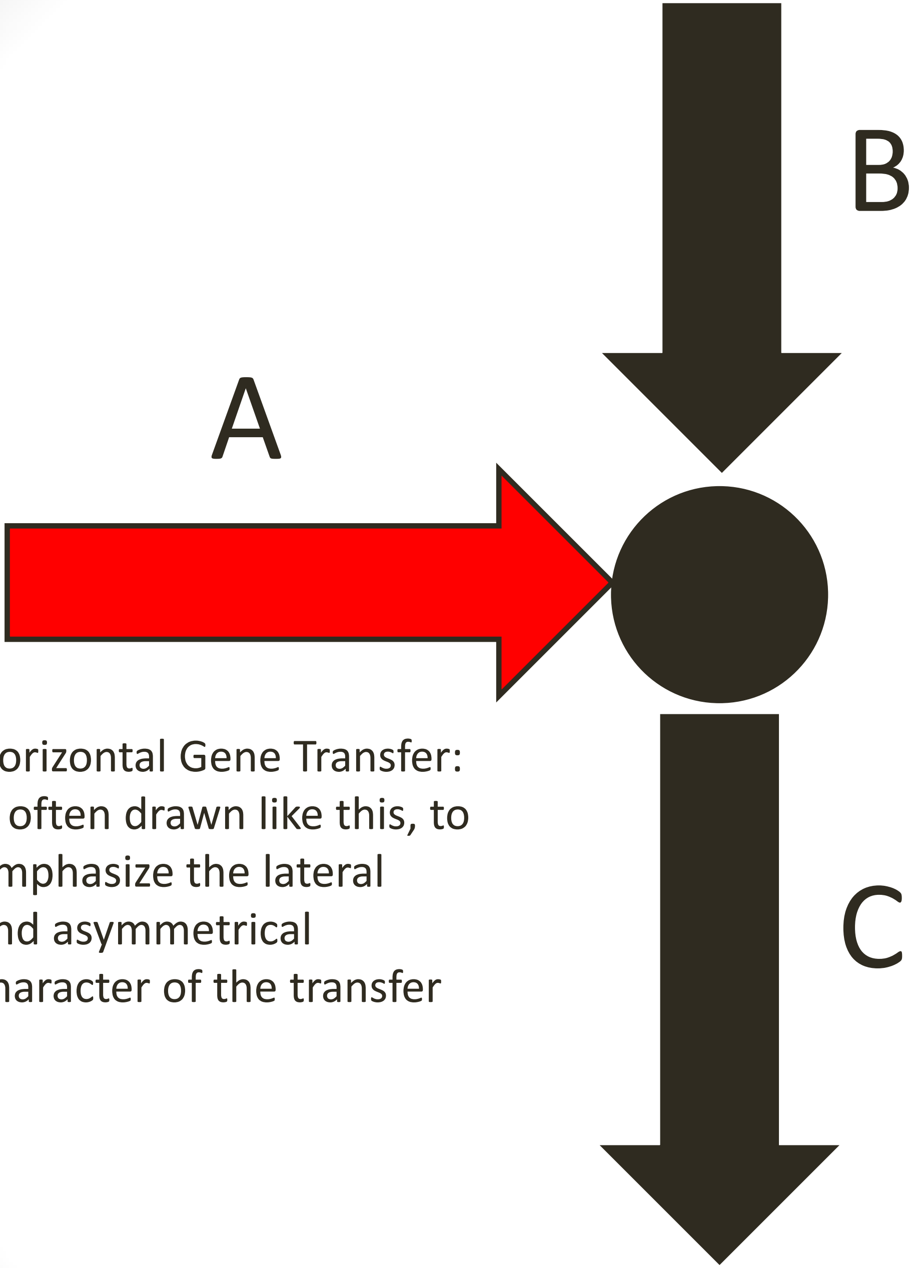
The biological meaning of such an event depends on the biological context.



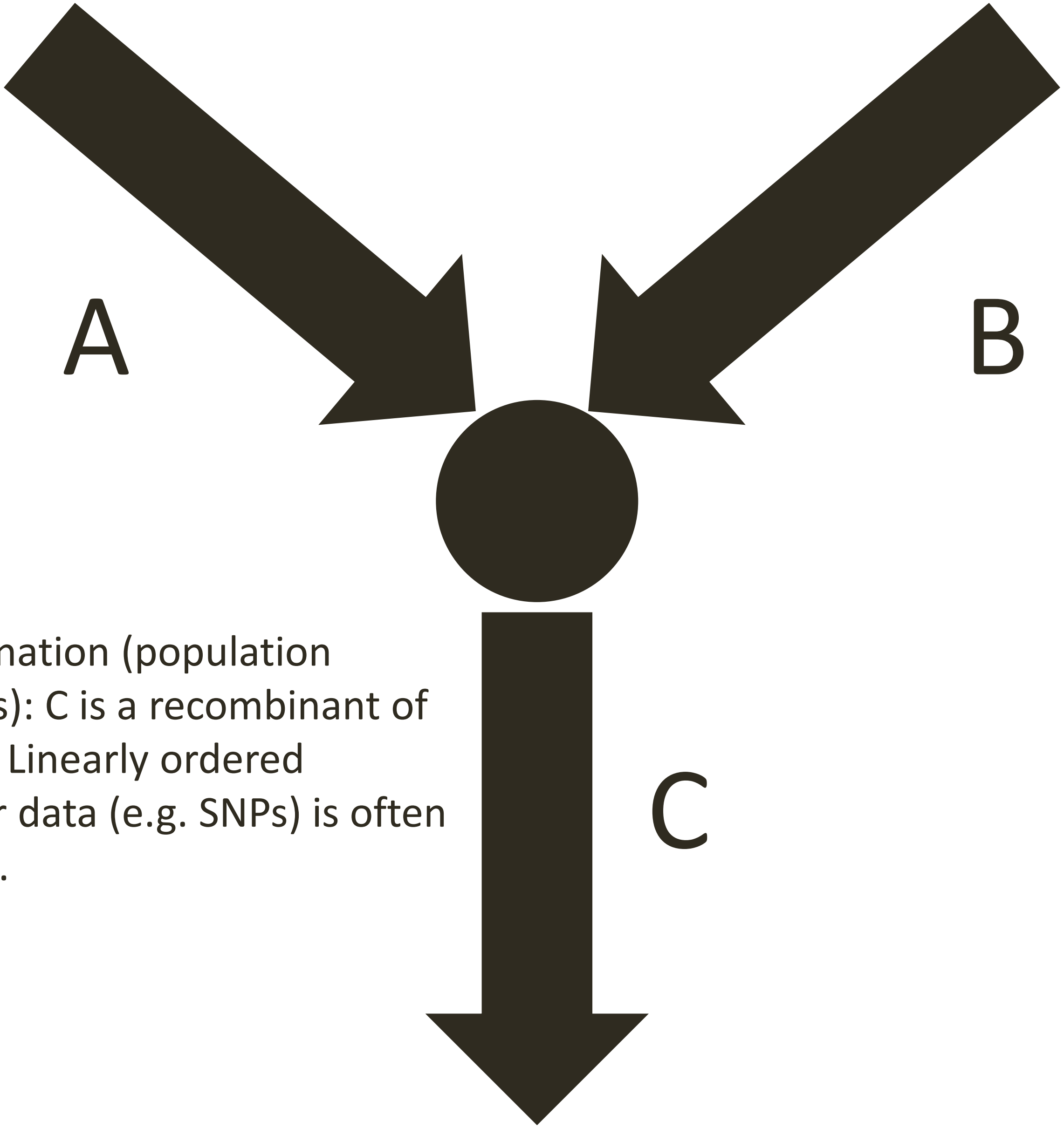
Hybridization: C is a hybrid of A and B



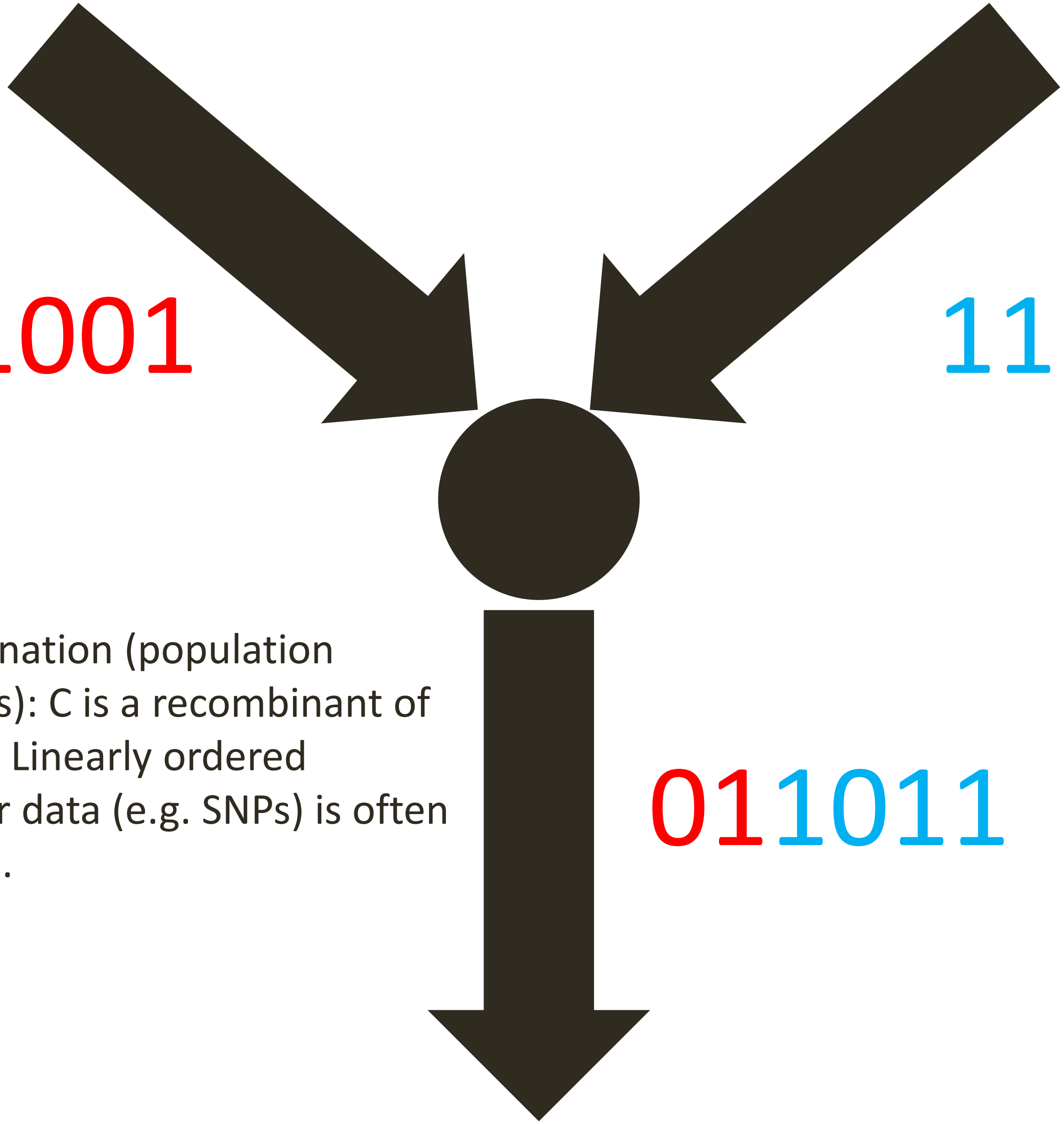
Horizontal Gene Transfer: a transfer of one or more genes from donor A into recipient B (emphasizes asymmetry)



Horizontal Gene Transfer:
is often drawn like this, to
emphasize the lateral
and asymmetrical
character of the transfer



Recombination (population genomics): C is a recombinant of A and B. Linearly ordered character data (e.g. SNPs) is often assumed.



011001

111011

011011

Recombination (population genomics): C is a recombinant of A and B. Linearly ordered character data (e.g. SNPs) is often assumed.

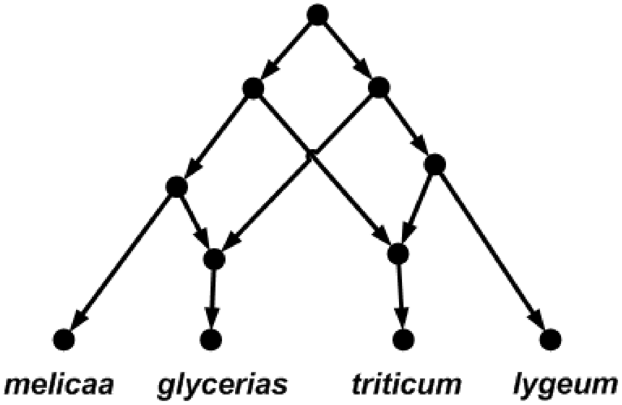
Sets of trees

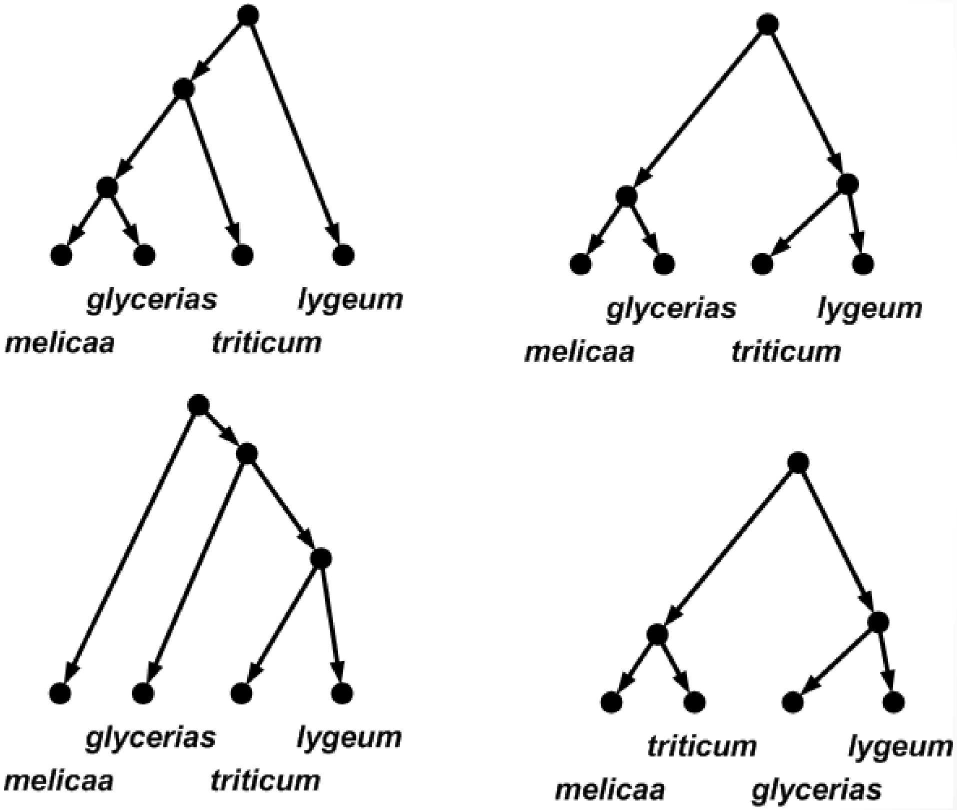
- A recurring theme - sometimes implicit - is the idea that an evolutionary phylogenetic network has many different trees topologically embedded within it.

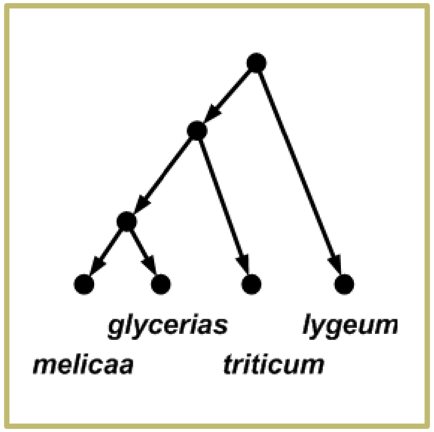
- That is: it is the simultaneous representation of the multiple distinct tree signals that can be present in a genome.

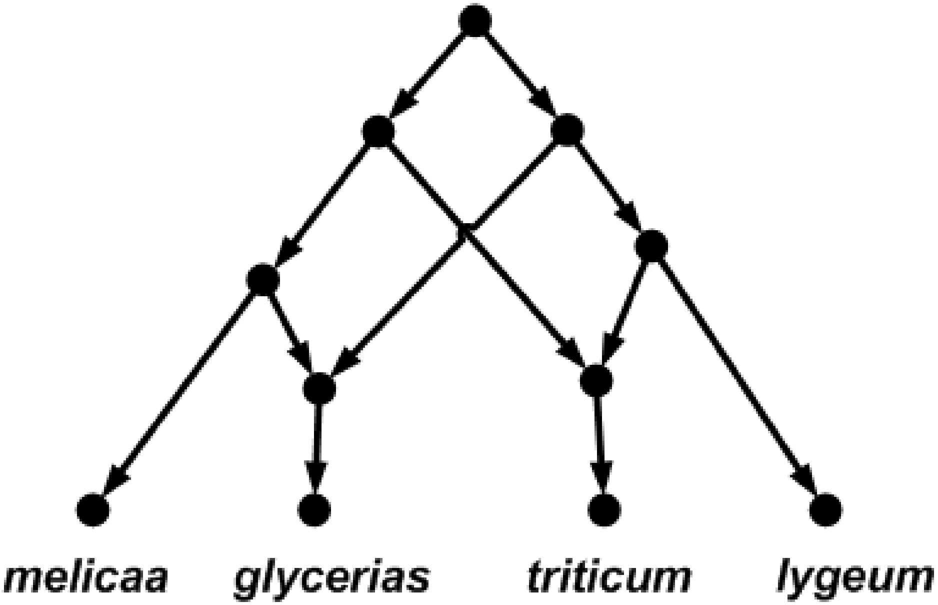
- Arguably the central question in this field is: under which circumstances can this this topological summary be a regarded as a meaningful approximation of “what actually happened”?

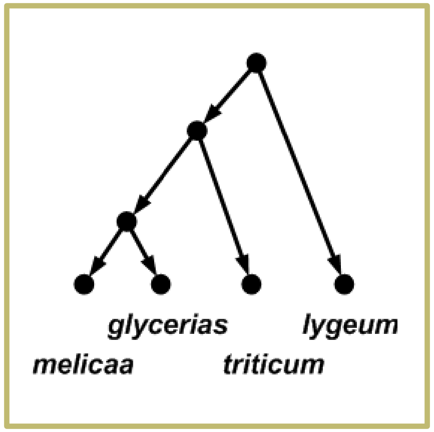
Networks contain many trees

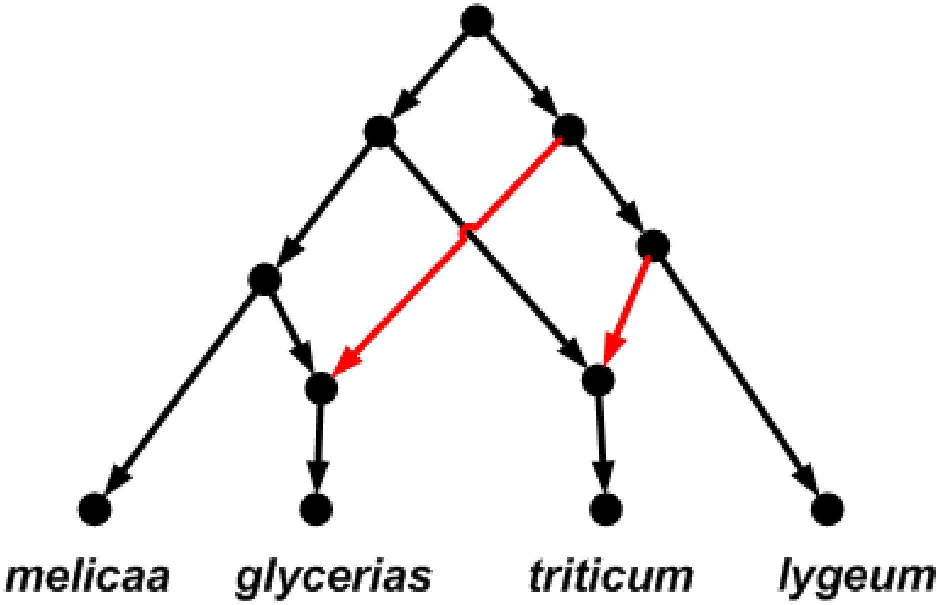


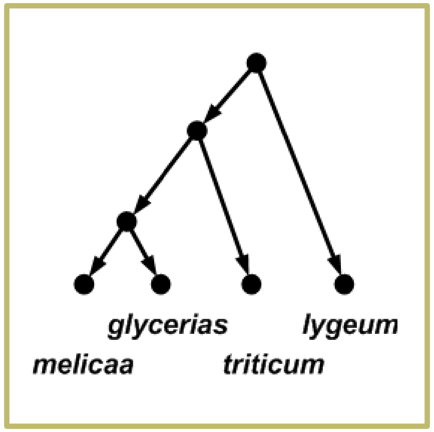


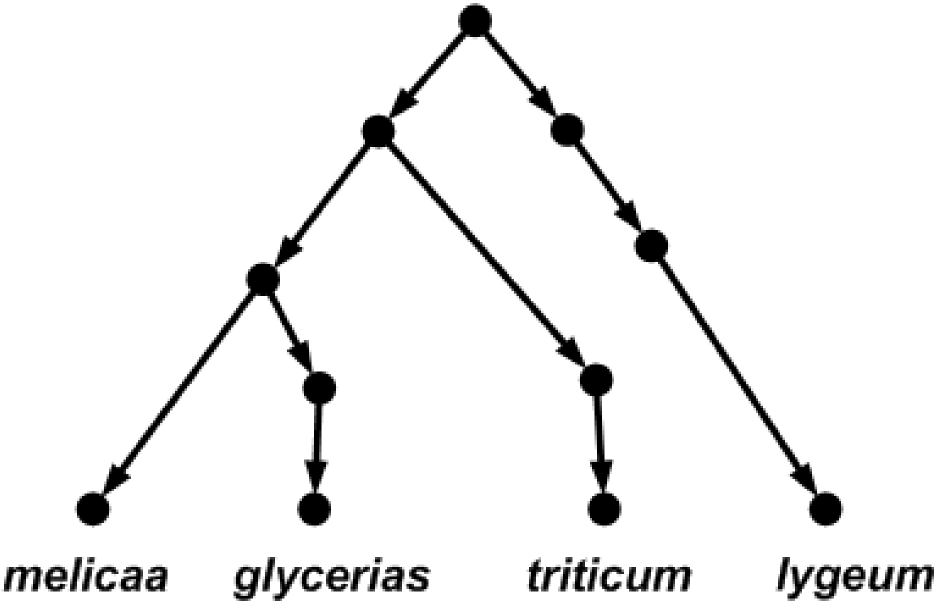

glycerias
lygeum
melicaa
triticum


melicaa
glycerias
triticum
lygeum

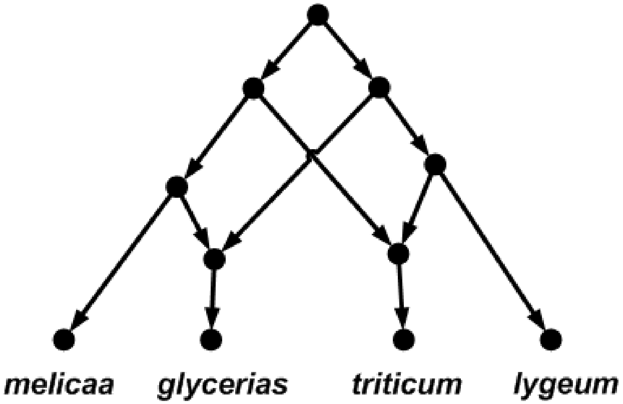

glycerias
lygeum
melicaa
triticum

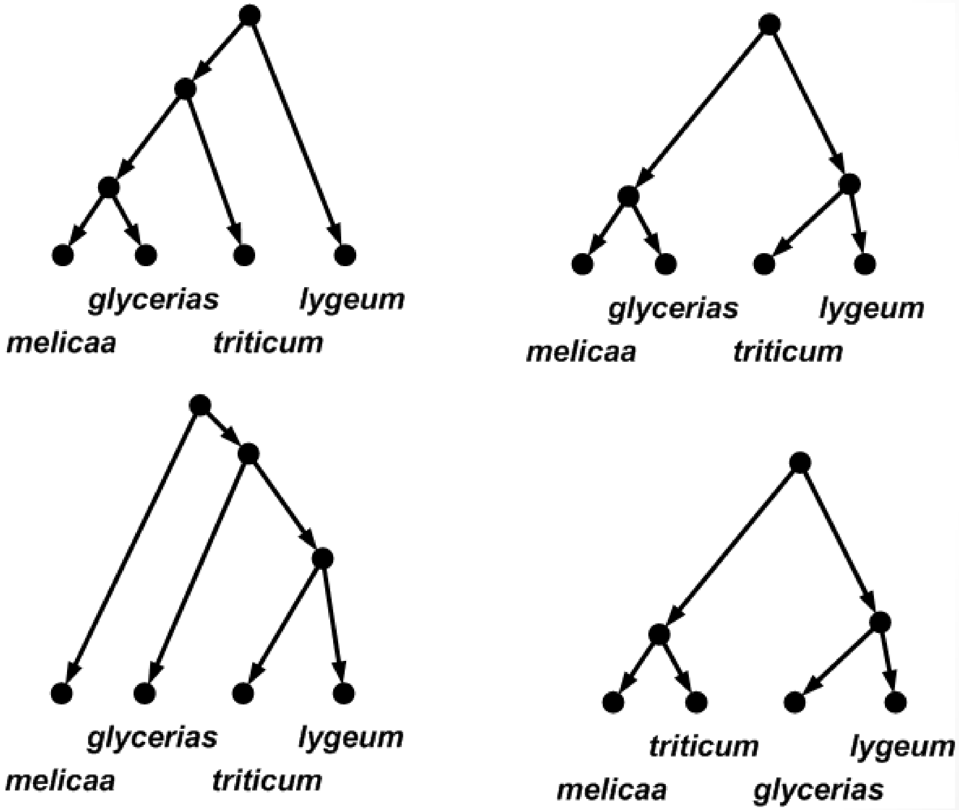

melicaa
glycerias
triticum
lygeum


glycerias
lygeum
melicaa
triticum

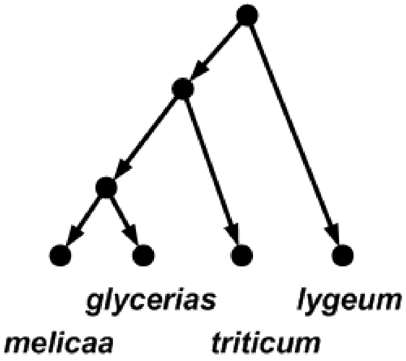

melicaa
glycerias
triticum
lygeum

Can we reconstruct the network, knowing (some of) the trees inside it?

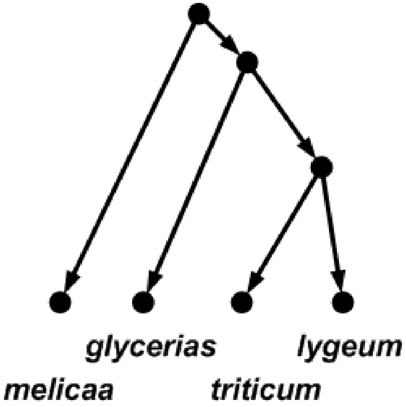




Can we reconstruct the network, knowing (some of) the trees inside it?

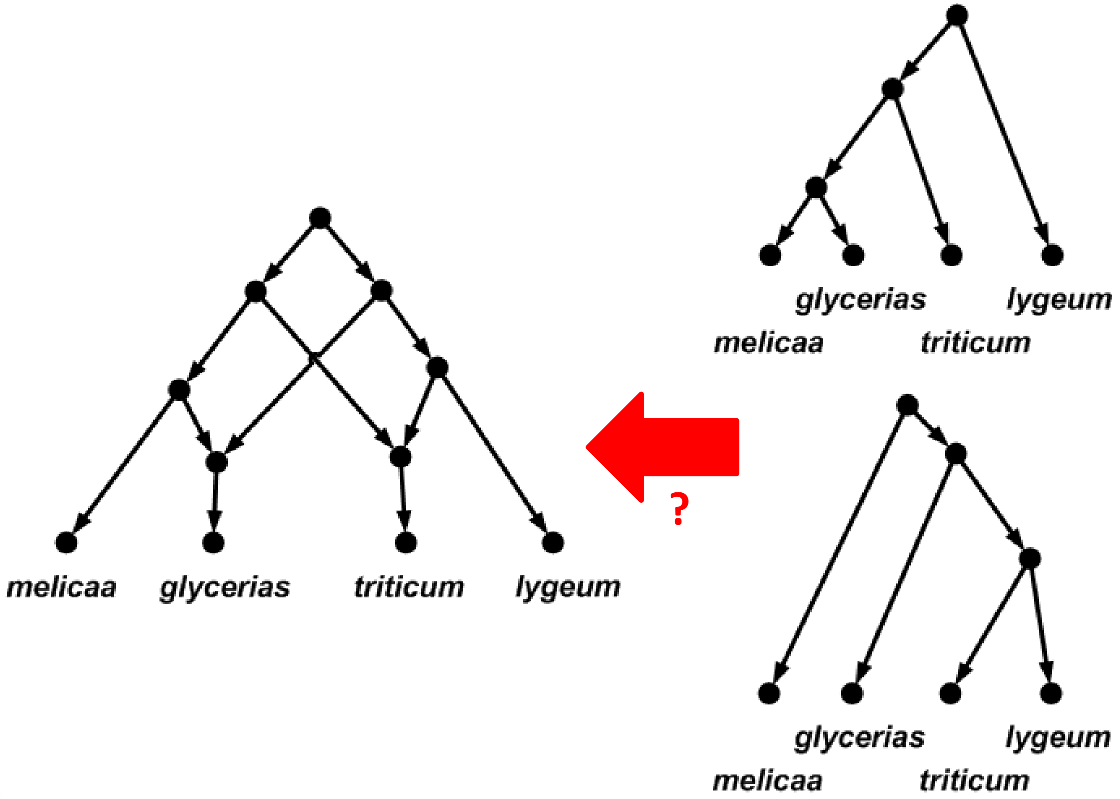


This tree
not known or
doesn’t exist



This tree
not known or
doesn’t exist

Can we reconstruct the network, knowing (some of) the trees inside it?



This tree
not known or
doesn’t exist

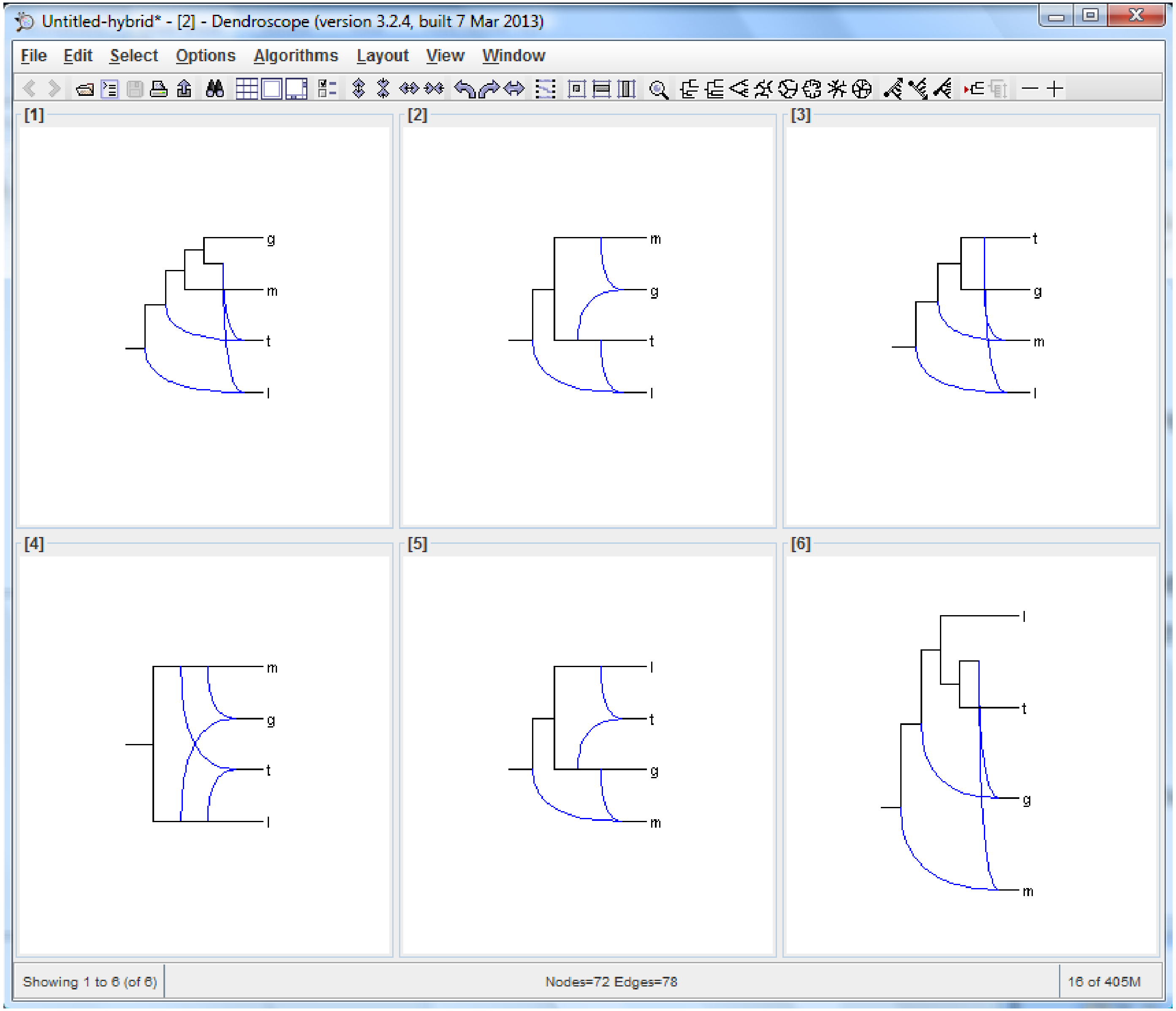
This tree
not known or
doesn’t exist

Reticulation parsimony

- **Input**: an set of incongruent rooted gene trees
- **Output:** a phylogenetic network that “contains” all the trees and which uses as few reticulation events as possible.

- If model assumptions are correct (...) and input data is well-behaved (...) then this gives a lower-bound on the number of reticulate events required to explain the incongruence.

- Does it also have predictive power? That is, can it predict topology? This is the big question for all optimization criteria, not just reticulation parsimony.



Untitled-hybrid* - [2] - Dendroscope (version 3.2.4, built 7 Mar 2013)

Showing 1 to 6 (of 6)

Nodes=72 Edges=78

16 of 405M

Incongruence ≠ reticulation

- A major complication is that incongruence between trees can be caused by many factors.
- Reticulation events can and do cause incongruence. But so can all kinds of other phenomena, both experimental and biological.
- Distinguishing between all these different phenomena is a major challenge.
- The following list is taken from *Introduction to Phylogenetic Networks.*

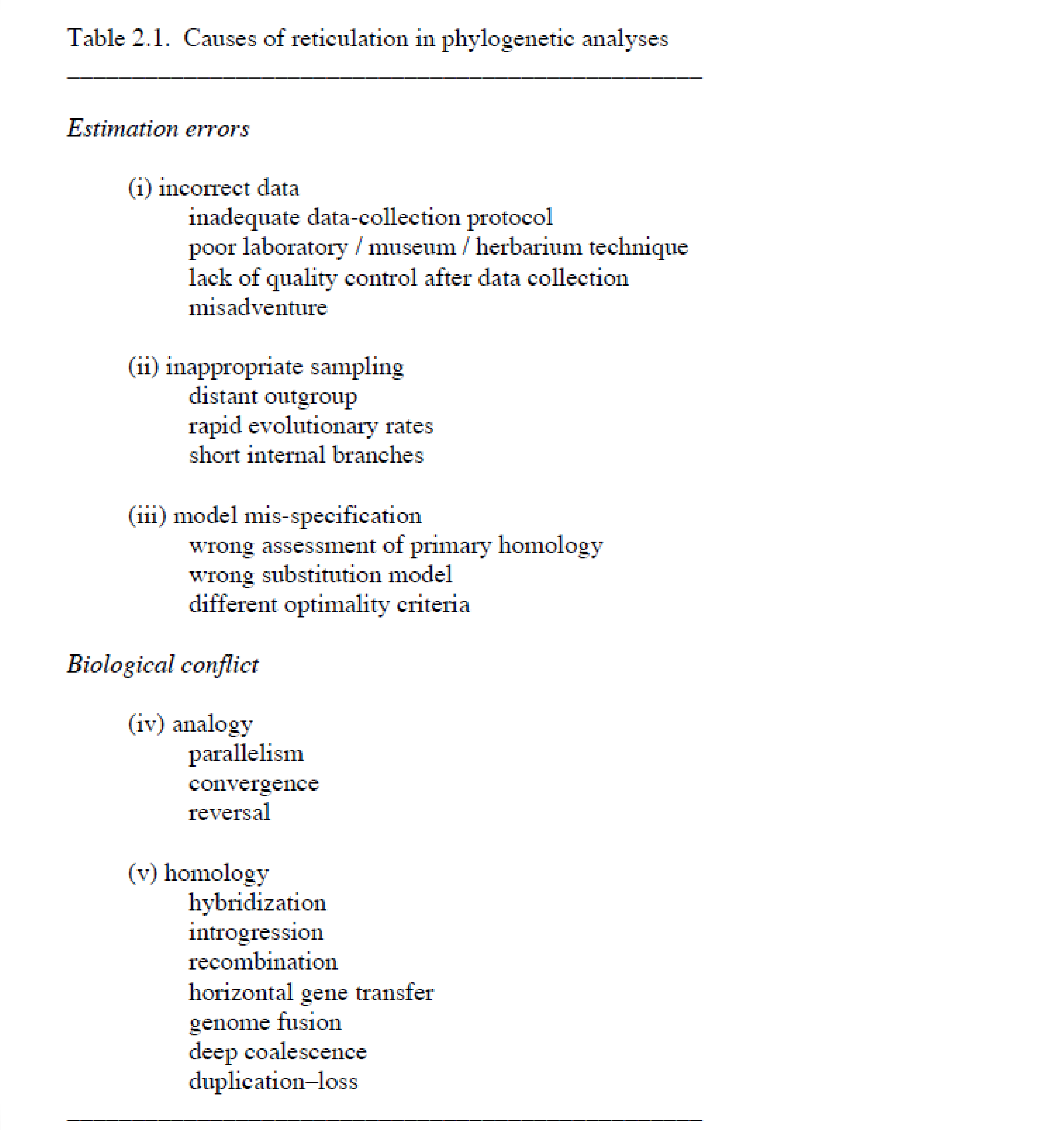


Table 2.1. Causes of reticulation in phylogenetic analyses

Estimation errors

- (i) incorrect data
  - inadequate data-collection protocol
  - poor laboratory / museum / herbarium technique
  - lack of quality control after data collection
  - misadventure
- (ii) inappropriate sampling
  - distant outgroup
  - rapid evolutionary rates
  - short internal branches
- (iii) model mis-specification
  - wrong assessment of primary homology
  - wrong substitution model
  - different optimality criteria

Biological conflict

- (iv) analogy
  - parallelism
  - convergence
  - reversal
- (v) homology
  - hybridization
  - introgression
  - recombination
  - horizontal gene transfer
  - genome fusion
  - deep coalescence
  - duplication–loss

"A rose by any other name, would smell as sweet…"

What about reconciliation?

- So far I discussed mapping trees into networks.
- What is the link with models in which the goal is to parsimoniously reconcile a species tree and a gene tree under the influence of duplication, loss and **transfer** (DLT) events?
- This is a very good question. In the DLT reconciliation literature, the terminology "phylogenetic network" is seldom used.
- There are some nontrivial mathematical differences. But there are many fundamental similarities.

What about reconciliation?

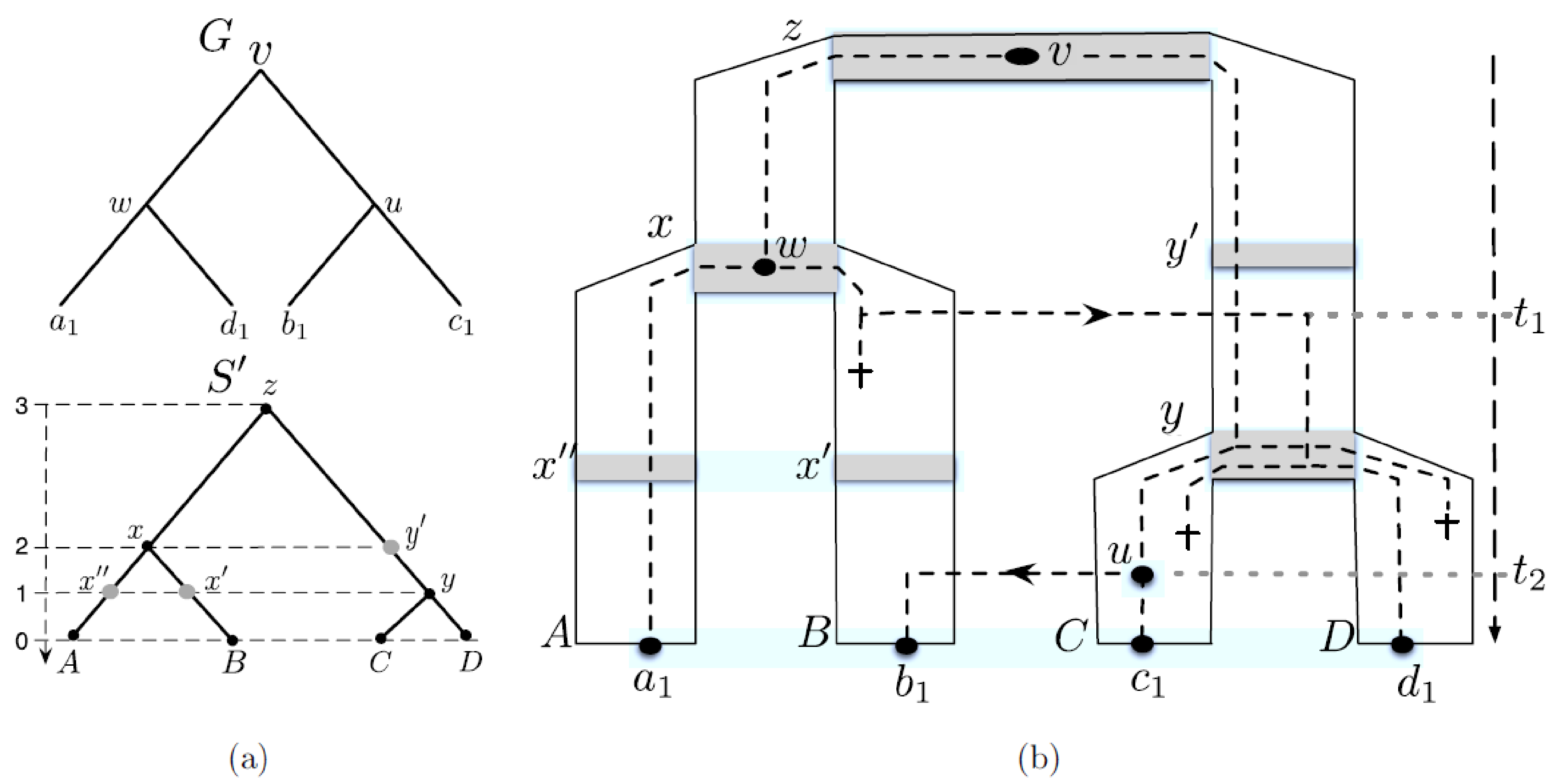


Fig. 1. (a) A gene tree G and a subdivided species tree S' (b) A reconciliation α between G and S', where α is defined as follows: $\alpha(v) = (z)$, $\alpha(u) = (y', y, C)$, $\alpha(w) = (x)$, $\alpha(a_1) = (x'', A)$, $\alpha(b_1) = (B)$, $\alpha(c_1) = (C)$ and $\alpha(d_1) = (x', y, D)$.

(FROM: REPRESENTING A SET OF RECONCILIATIONS IN A COMPACT WAY, CELINE SCORNAVACCA, WOJCIECH PAPROTNY, VINCENT BERRY and VINCENT RANWEZ, JBCB 2013)

What about reconciliation?

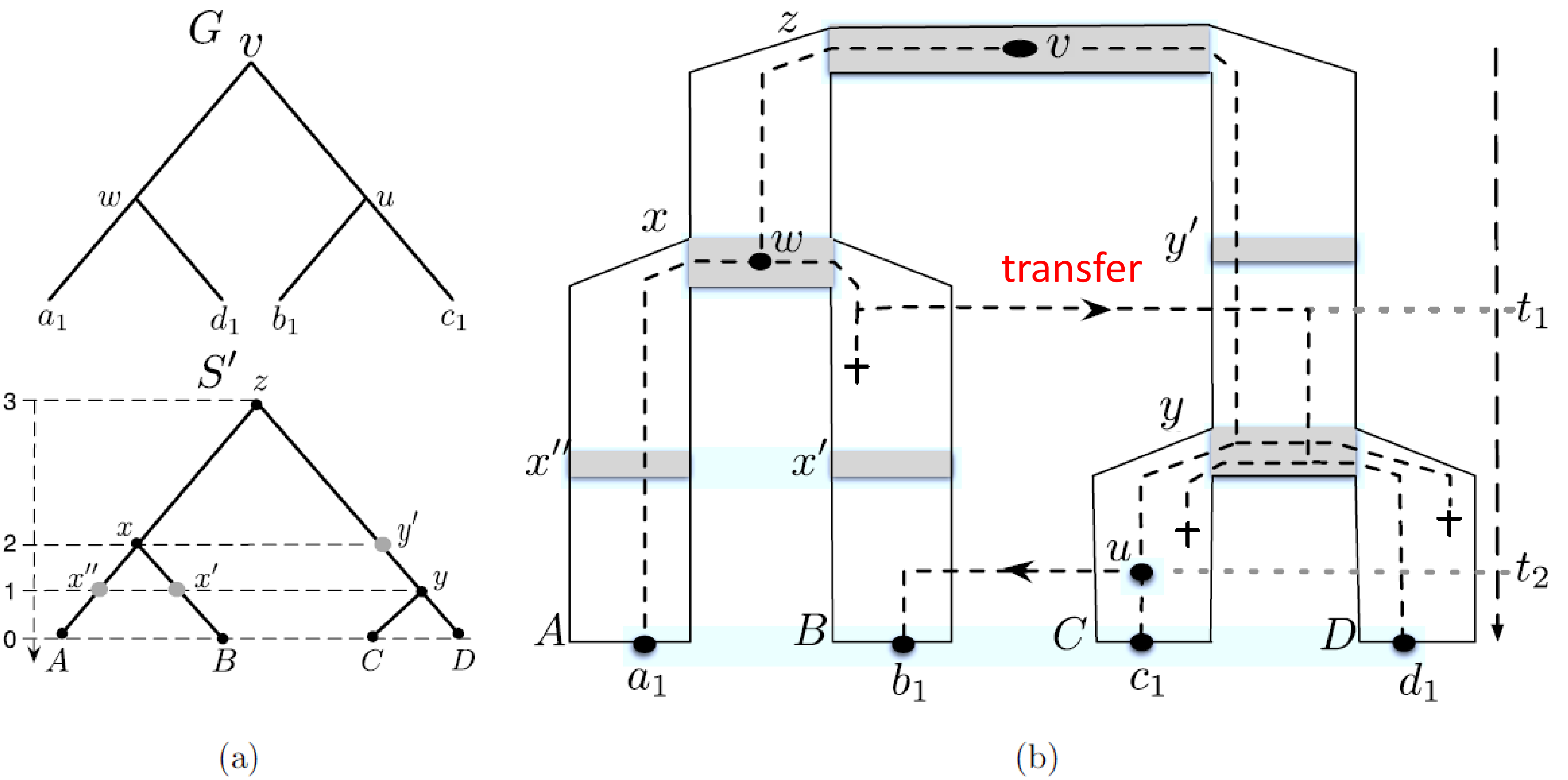


Fig. 1. (a) A gene tree G and a subdivided species tree S' (b) A reconciliation α between G and S', where α is defined as follows: $\alpha(v) = (z)$, $\alpha(u) = (y', y, C)$, $\alpha(w) = (x)$, $\alpha(a_1) = (x'', A)$, $\alpha(b_1) = (B)$, $\alpha(c_1) = (C)$ and $\alpha(d_1) = (x', y, D)$.

(FROM: REPRESENTING A SET OF RECONCILIATIONS IN A COMPACT WAY, CELINE SCORNAVACCA, WOJCIECH PAPROTNY, VINCENT BERRY and VINCENT RANWEZ, JBCB 2013)

What about reconciliation?

- At the moment the literature on reconciliation is (almost) entirely disjoint from the evolutionary phylogenetic networks literature.

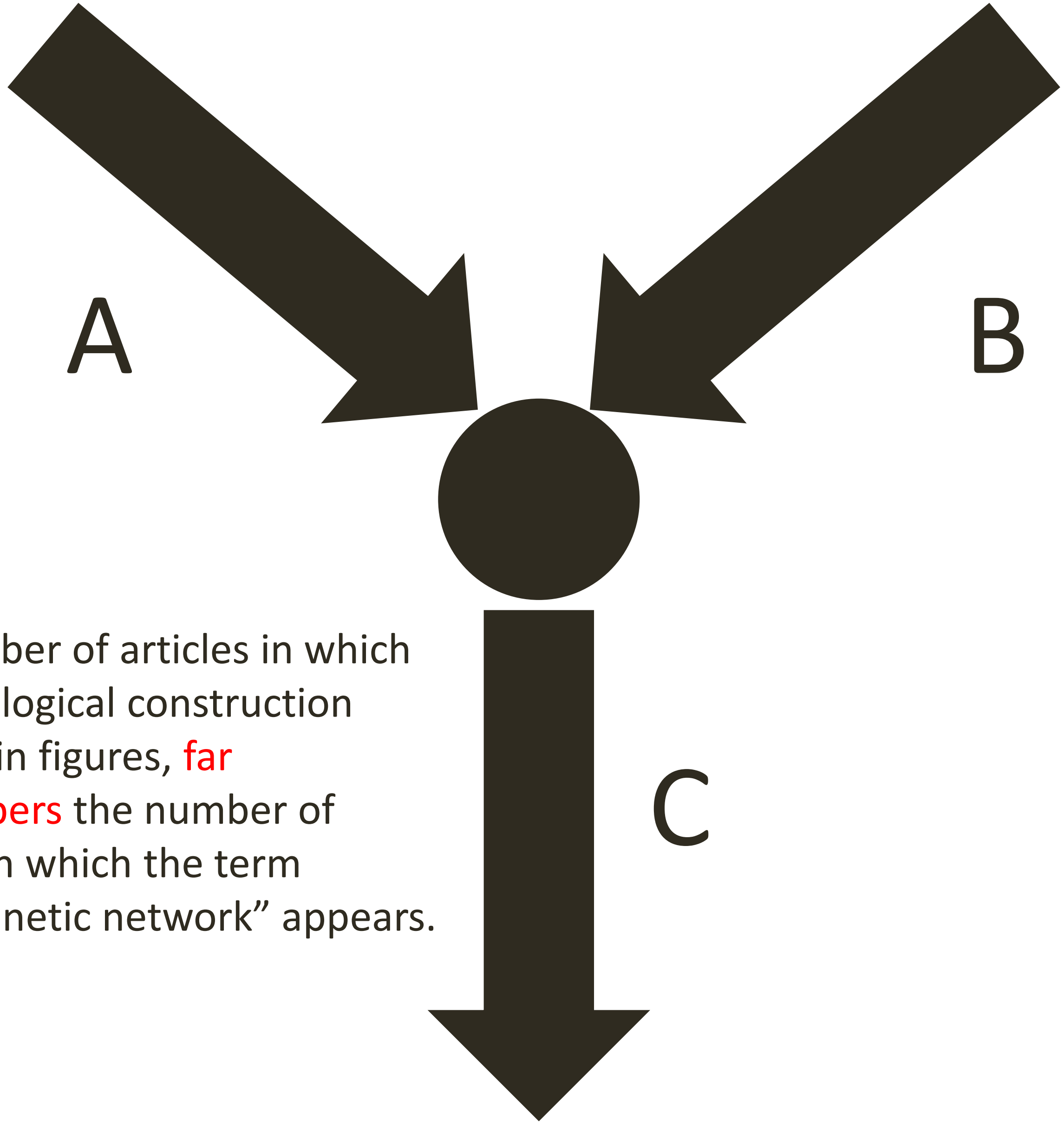
- This artificial gap needs to be closed.

- The often-heard claim *“nobody uses evolutionary phylogenetic networks in practice...”* no longer holds if we include reconciliation and other conceptually similar models in our definition of phylogenetic networks.

A rose by any other name…?

- A new claim:

- *“Many people are using evolutionary phylogenetic networks in practice – they just don’t use that terminology. Maybe they call it ad-hoc experimental determination of HGT events, or DLT-reconciliation, or Ancestral Recombination Graphs (ARGs), etc., but they are all phylogenetic networks...”*



The number of articles in which this topological construction appears in figures, far outnumbers the number of articles in which the term “phylogenetic network” appears.

A rose by any other name…?

- *"Many people are using evolutionary phylogenetic networks in practice – they just don't use that terminology. Maybe they call it ad-hoc experimental determination of HGT events, or DLT-reconciliation, or Ancestral Recombination Graphs (ARGs), etc., but they are all phylogenetic networks…"*

- *"…Very few people, however, are using <u>integrated, (semi-)automated methods</u> for constructing phylogenetic networks. That is, very few people use and/or trust existing software that takes raw biological data as input and generates a network hypothesis."*

What can we currently do?

- Here’s a summary of what we can currently do. I am using my expanded definition of phylogenetic network here.
- Hence, not all methods are automated or even semi-automated. Complicated experimental pipelines are common.
- In the second part of the talk I will investigate how far the (semi-)automated techniques have been seriously used for hypothesis generation and testing.

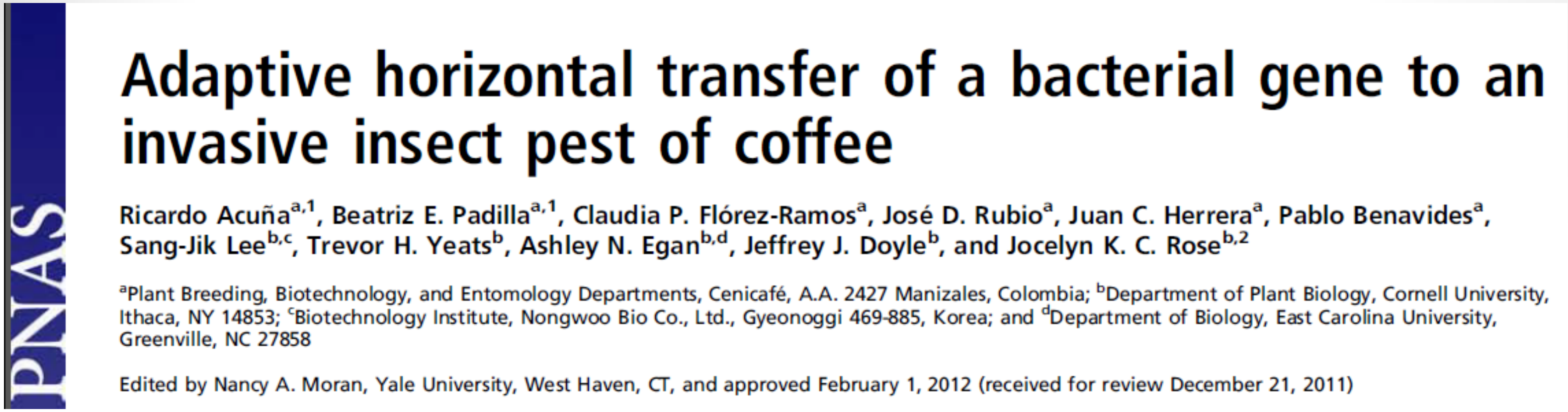
What can we currently do? (1)

- **Ad-hoc / pipeline experimental analysis**

- Context: in Google Scholar there are currently approximately 60,000 articles that refer to HGT/LGT.

- Many of these articles are concerned with quantifying and locating HGT events within various different groups of organisms.

- In the absence of standard(ized) computational tools for quantifying/locating HGT, and in the spirit of experimental computational biology, many of the articles use a huge array of (more conventional) software packages and phylogenetic techniques to gather evidence for conclusions.



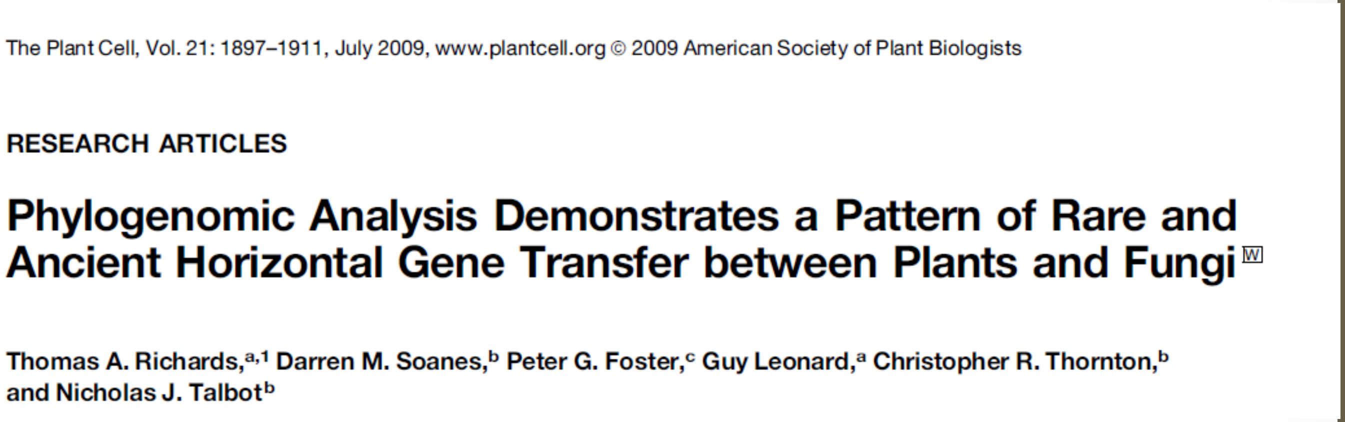
PNAS

Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee

Ricardo Acuña[a,1], Beatriz E. Padilla[a,1], Claudia P. Flórez-Ramos[a], José D. Rubio[a], Juan C. Herrera[a], Pablo Benavides[a], Sang-Jik Lee[b,c], Trevor H. Yeats[b], Ashley N. Egan[b,d], Jeffrey J. Doyle[b], and Jocelyn K. C. Rose[b,2]

[a]Plant Breeding, Biotechnology, and Entomology Departments, Cenicafé, A.A. 2427 Manizales, Colombia; [b]Department of Plant Biology, Cornell University, Ithaca, NY 14853; [c]Biotechnology Institute, Nongwoo Bio Co., Ltd., Gyeonoggi 469-885, Korea; and [d]Department of Biology, East Carolina University, Greenville, NC 27858

Edited by Nancy A. Moran, Yale University, West Haven, CT, and approved February 1, 2012 (received for review December 21, 2011)

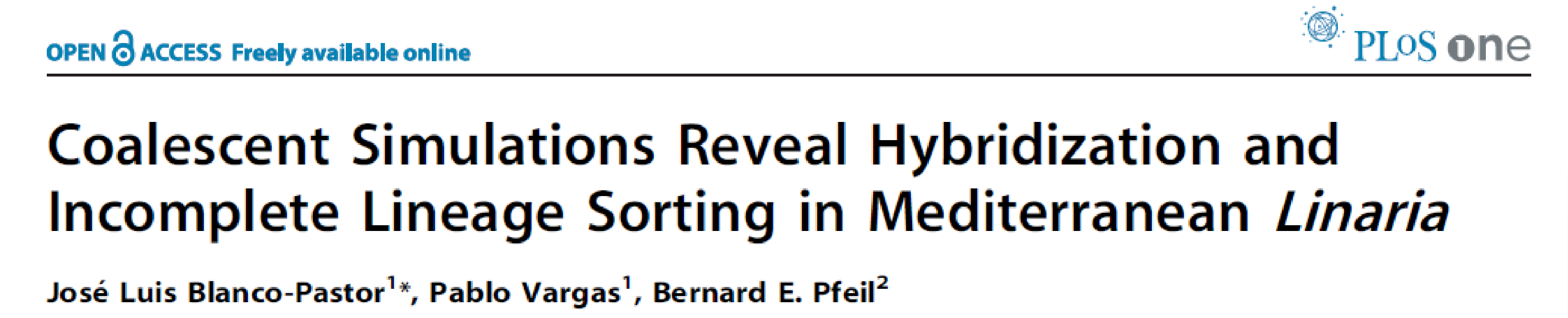


The Plant Cell, Vol. 21: 1897–1911, July 2009, www.plantcell.org © 2009 American Society of Plant Biologists

RESEARCH ARTICLES

Phylogenomic Analysis Demonstrates a Pattern of Rare and Ancient Horizontal Gene Transfer between Plants and Fungi [W]

Thomas A. Richards,[a,1] Darren M. Soanes,[b] Peter G. Foster,[c] Guy Leonard,[a] Christopher R. Thornton,[b] and Nicholas J. Talbot[b]



OPEN ACCESS Freely available online

PLoS one

Coalescent Simulations Reveal Hybridization and Incomplete Lineage Sorting in Mediterranean *Linaria*

José Luis Blanco-Pastor[1]*, Pablo Vargas[1], Bernard E. Pfeil[2]

(2012)

What can we currently do? (2)

- **Many flavours of parsimony**

- “Reticulation parsimony” : combinatorial algorithms to assemble trees or fragments of trees (triplets, clusters, SNPs) into networks with few reticulation events.

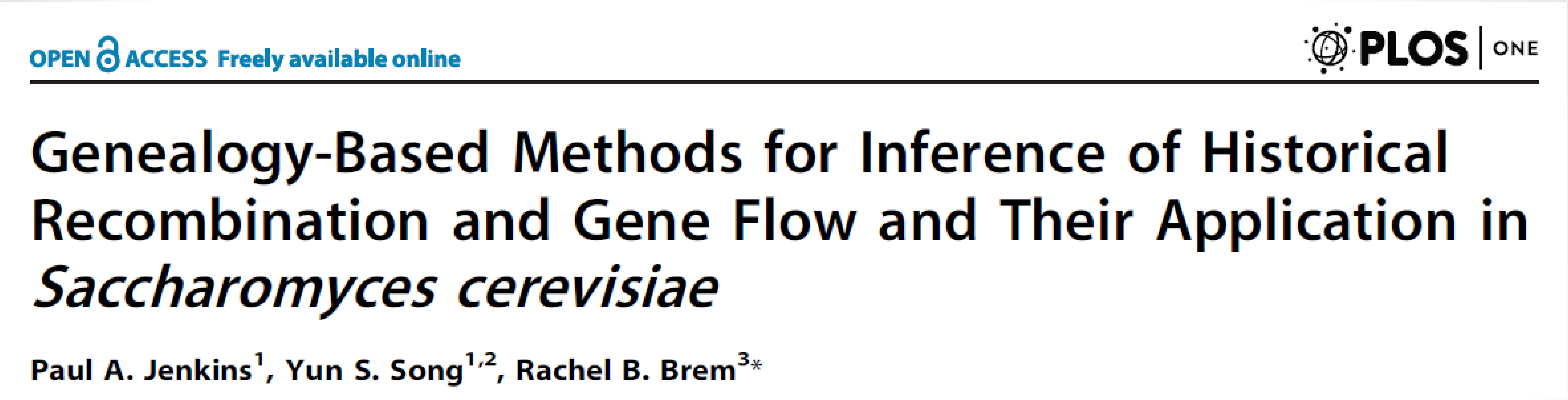
- Parsiminous gene-species tree DLT reconciliation : fit a gene tree into a species tree whilst minimizing weighted cost of speciation, gene duplication, gene loss, horizontal gene transfer events.

- Extension of classical Maximum Parsimony (MP) tree-building technique to networks: network is viewed as a set of trees, the network is as good as its best tree.

What can we currently do?

- Reticulation parsimony attracts a lot of attention from mathematicians but is not used much in practice.

- Recent exception:



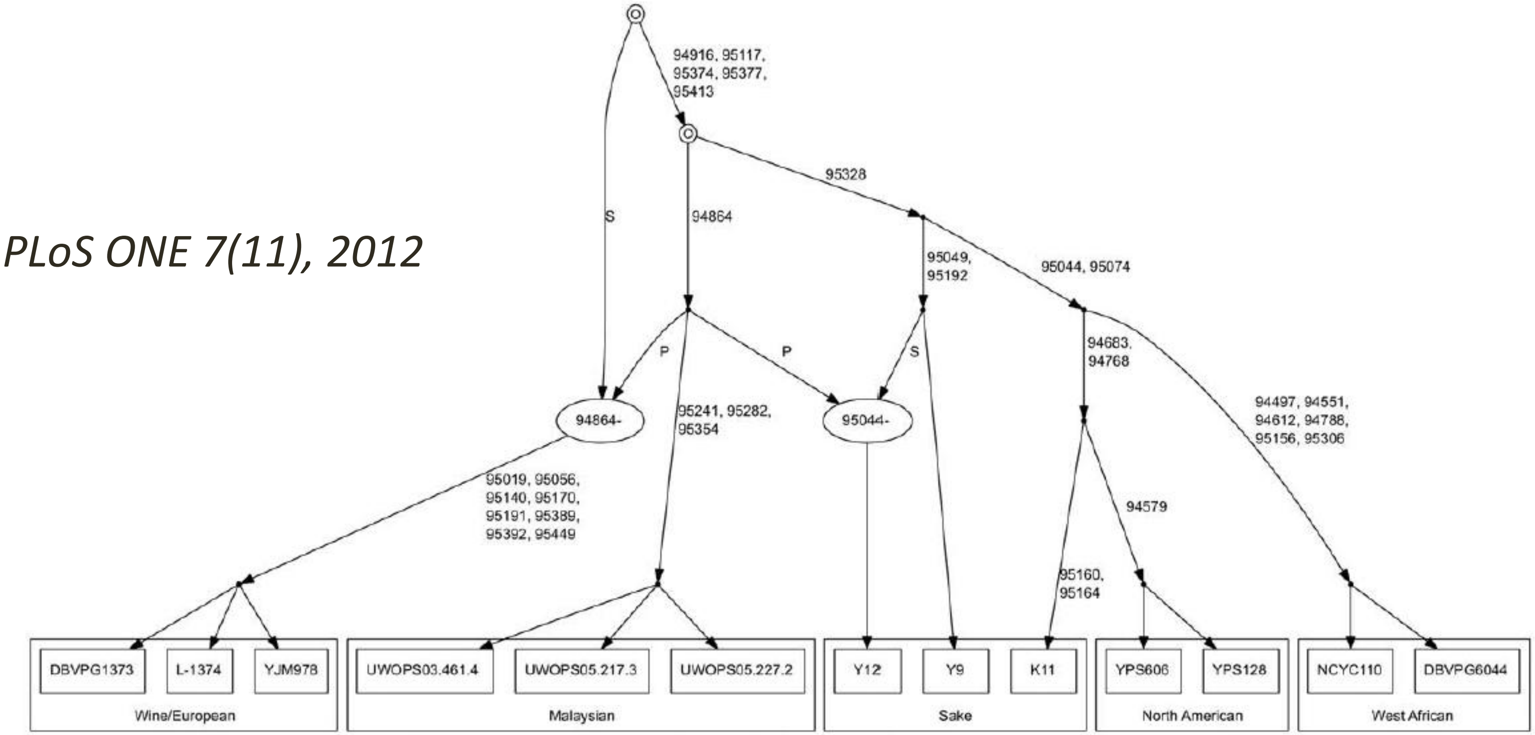
OPEN ACCESS Freely available online

PLOS | ONE

Genealogy-Based Methods for Inference of Historical Recombination and Gene Flow and Their Application in *Saccharomyces cerevisiae*

Paul A. Jenkins[1], Yun S. Song[1,2], Rachel B. Brem[3]*

PLoS ONE 7(11), 2012



What can we currently do? (3)

- **Other methods based on topological dissimilarity**

- Combinatorial methods that attempt to quantify and model HGT using gene-species tree incongruence measures (various rearrangement distances e.g. rSPR; quartet decomposition, bipartition analysis etc.)

- Statistical likelihood-based tests to determine whether incongruence between a given species tree and a gene tree is statistically significant (AU, SH, KH, ILD...)

What can we currently do? (4)

- **Statistical methods**

- Coalescent with recombination / stochastic analysis of Ancestral Recombination Graphs (ARGs)
- Bayesian
- Methods to understand when/if a species tree signal can be recovered in the corrupting presence of HGT
- Statistical methods for distinguishing between hybridization and incomplete lineage sorting
- Statistical reconciliation models
- Extension of ML to networks (similar idea to MP on networks)

What can we currently do? (5)

- **Combinations of different techniques and models**
- Due to the strengths and weaknesses of different individual techniques, there is a growing tendency towards combining multiple techniques.
- Example: using statistical methods to discriminate between network topologies generated by a low-resolution parsimony-based method.
- There is also a tendency towards computational models that incorporate multiple incongruence-causing events, especially: hybridization vs. incomplete lineage sorting

Computational headaches

- The space of phylogenetic networks is vast; far larger than the space of trees. We still don’t understand how to deal with this.
- This heavily constrains all methods, parsimony-based or statistical, whose mathematical core is based on enumerating or integrating over this space.
- NP-hardness (or worse) is everywhere.
- Many multiple optima.
- Sensitivity of topology-based methods to noise.
- Dealing with multiple sources of incongruence.

Part 2.

In how far have these models been seriously used for hypothesis generation and testing?

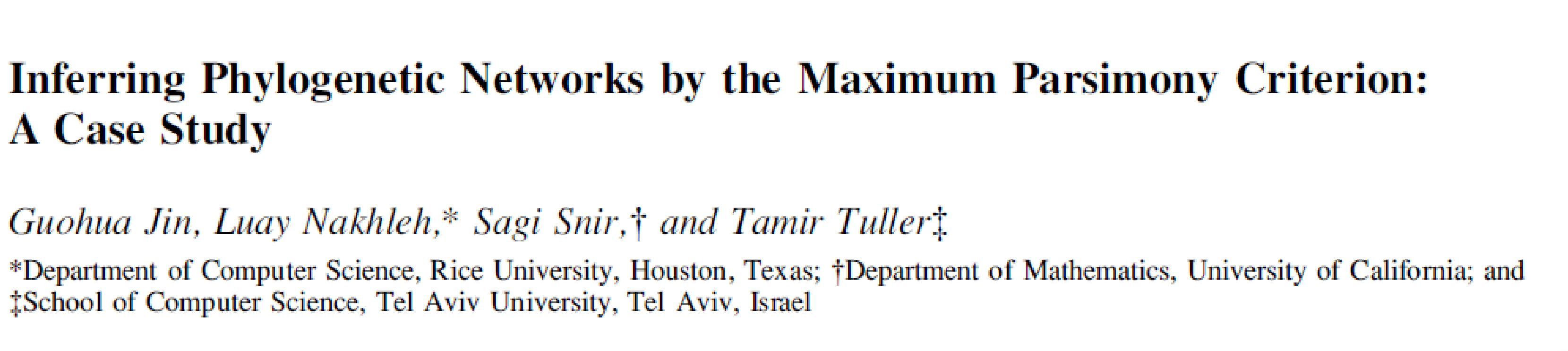
Ground rules

- I will now look at 4 case-studies, based on published articles, to try and answer this question.
- In all cases existing datasets are re-analysed.
- No simulated data. No *Poacea* dataset!

Preliminary reflections

- The bad news is that I only have time today to look at a few case studies ☹
- The good news is that I could have included many more ☺
- Someone should write a book about this – the time is right.

Case study 1: MP on networks



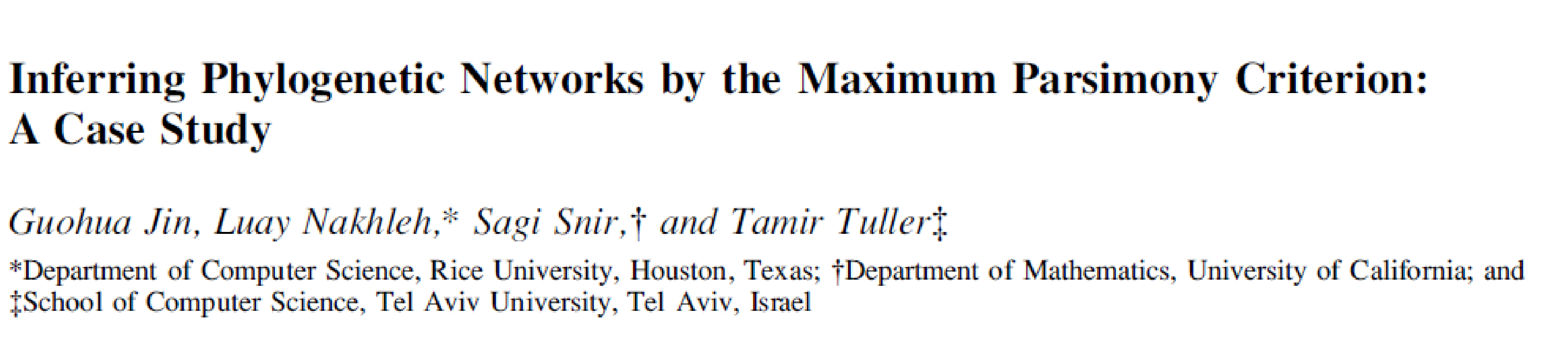
Inferring Phylogenetic Networks by the Maximum Parsimony Criterion: A Case Study

Guohua Jin, Luay Nakhleh,[*] Sagi Snir,[†] and Tamir Tuller[‡]

[*] Department of Computer Science, Rice University, Houston, Texas; [†] Department of Mathematics, University of California; and [‡] School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Mol Biol Evol (2007) 24 (1): 324-337.

Case study 1: MP on networks



Inferring Phylogenetic Networks by the Maximum Parsimony Criterion: A Case Study

Guohua Jin, Luay Nakhleh, Sagi Snir,† and Tamir Tuller‡*

*Department of Computer Science, Rice University, Houston, Texas; †Department of Mathematics, University of California; and ‡School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Mol Biol Evol (2007) 24 (1): 324-337.

"In this article, we investigate the performance and robustness of the MP criterion for phylogenetic networks on real biological data sets. In particular, we study the performance of the MP criterion with respect to detecting the actual number and location of HGT events, the robustness of the criterion with respect to incomplete taxon sampling and different site substitution matrices, and the applicability of the criterion to detecting HGT in chimeric genes."

Inferring phylogenetic networks by the maximum parsimony criterion: a case study

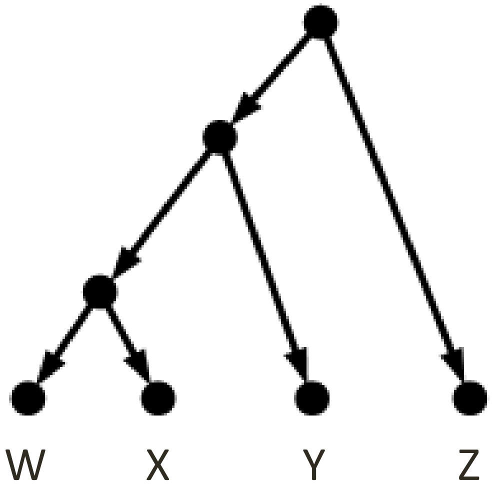
- Re-analyses 4 biological datasets:
  - The rubisco gene rbcL of a group of 46 plastids, cyanobacteria, and proteobacteria, which was analyzed by Delwiche and Palmer (1996).
  - The ribosomal protein rpl12e of a group of 14 archaeal organisms, which was analyzed by Matte-Tailliez et al. (2002).
  - The ribosomal protein gene rps11 of a group of 47 flowering plants, which was analyzed by Bergthorsson et al. (2003).
  - The mitochondrial gene cox2 of a group of 25 seed and nonseed plants, which was analyzed by Bergthorsson et al. (2004).

Inferring phylogenetic networks by the maximum parsimony criterion: a case study

- In each case, the starting point is a reliable/plausible species tree. This is obtained either from earlier literature or from separate analysis (…)

- The gene is left as sequence data (so no intermediate tree-building step for the gene)

- The goal is to fit HGT events onto the species tree, creating a phylogenetic network, to improve the fit of the sequence data

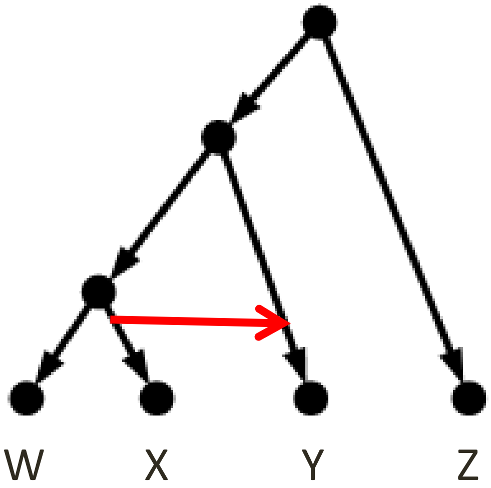


Species tree

W	GACATATC--A--A
X	ATCGCTA-C--AAT
Y	TG-TAAA---C-T-A
Z	GTAACACATCAT-

Gene alignment

Add HGT edges to improve the (parsimony) fit of the alignment

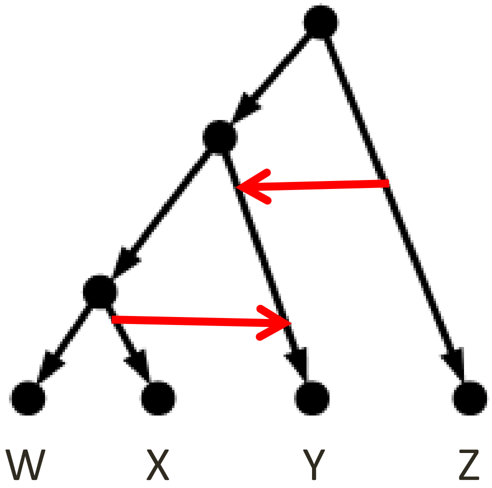


Species tree

W	GACATATC--A--A
X	ATCGCTA-C--AAT
Y	TG-TAAA---C-T-A
Z	GTAACACATCAT-

Gene alignment

Add HGT edges to improve the (parsimony) fit of the alignment

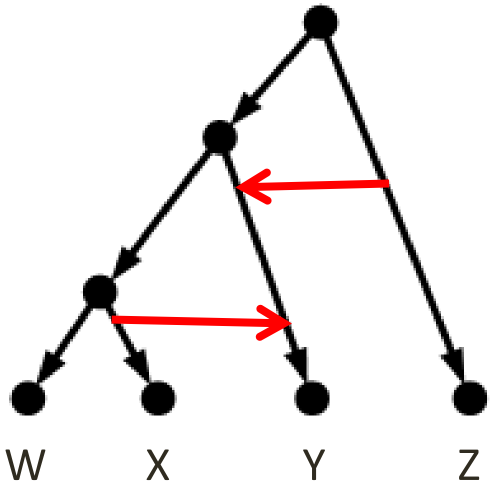


Species tree

W	GACATATC--A--A
X	ATCGCTA-C--AAT
Y	TG-TAAA---C-T-A
Z	GTAACACATCAT-

Gene alignment

Add HGT edges to improve the (parsimony) fit of the alignment



Species tree

W	GACATATC--A--A
X	ATCGCTA-C--AAT
Y	TG-TAAA---C-T-A
Z	GTAACACATCAT-

Gene alignment

Stop adding HGT edges when the improvement is no longer significant.

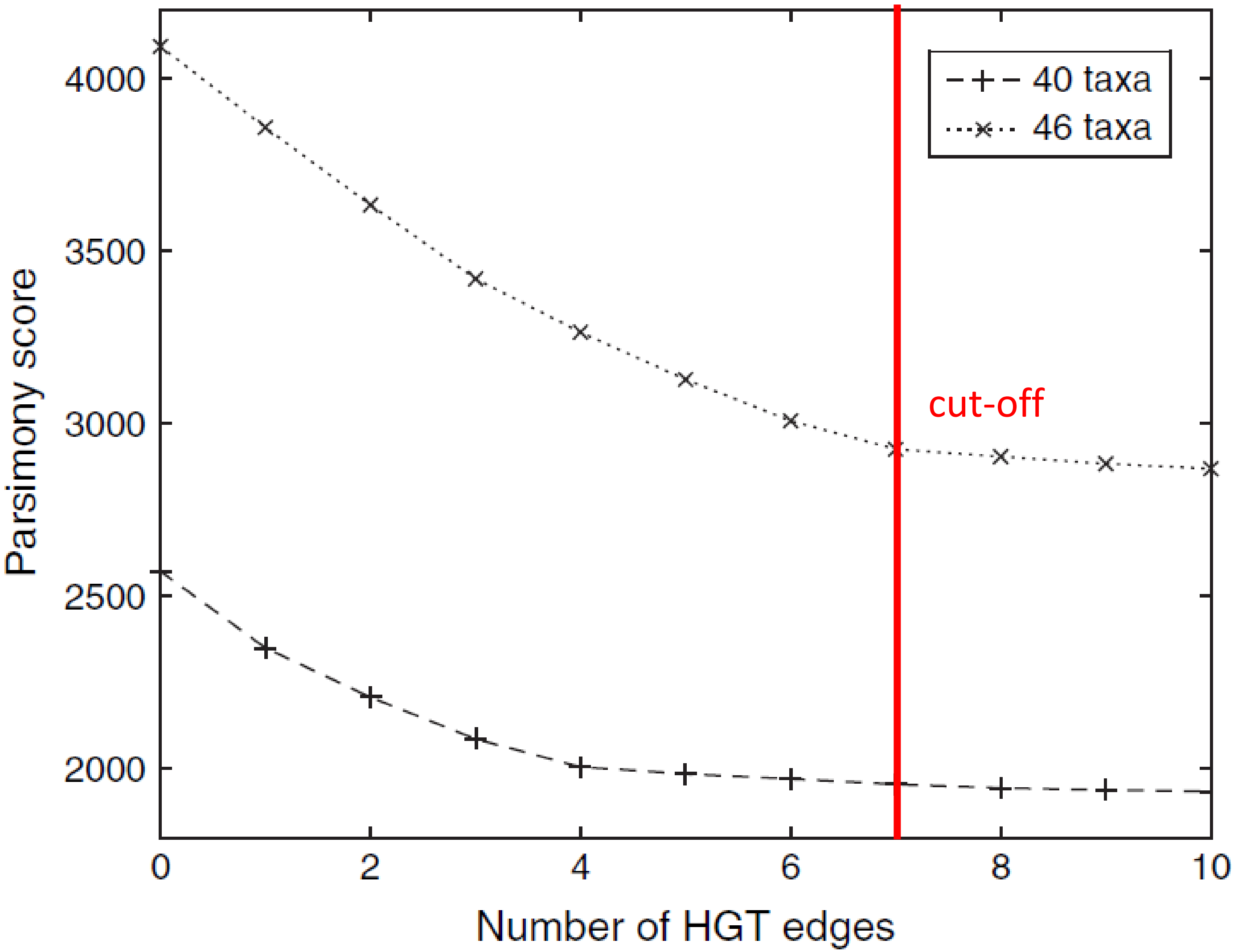
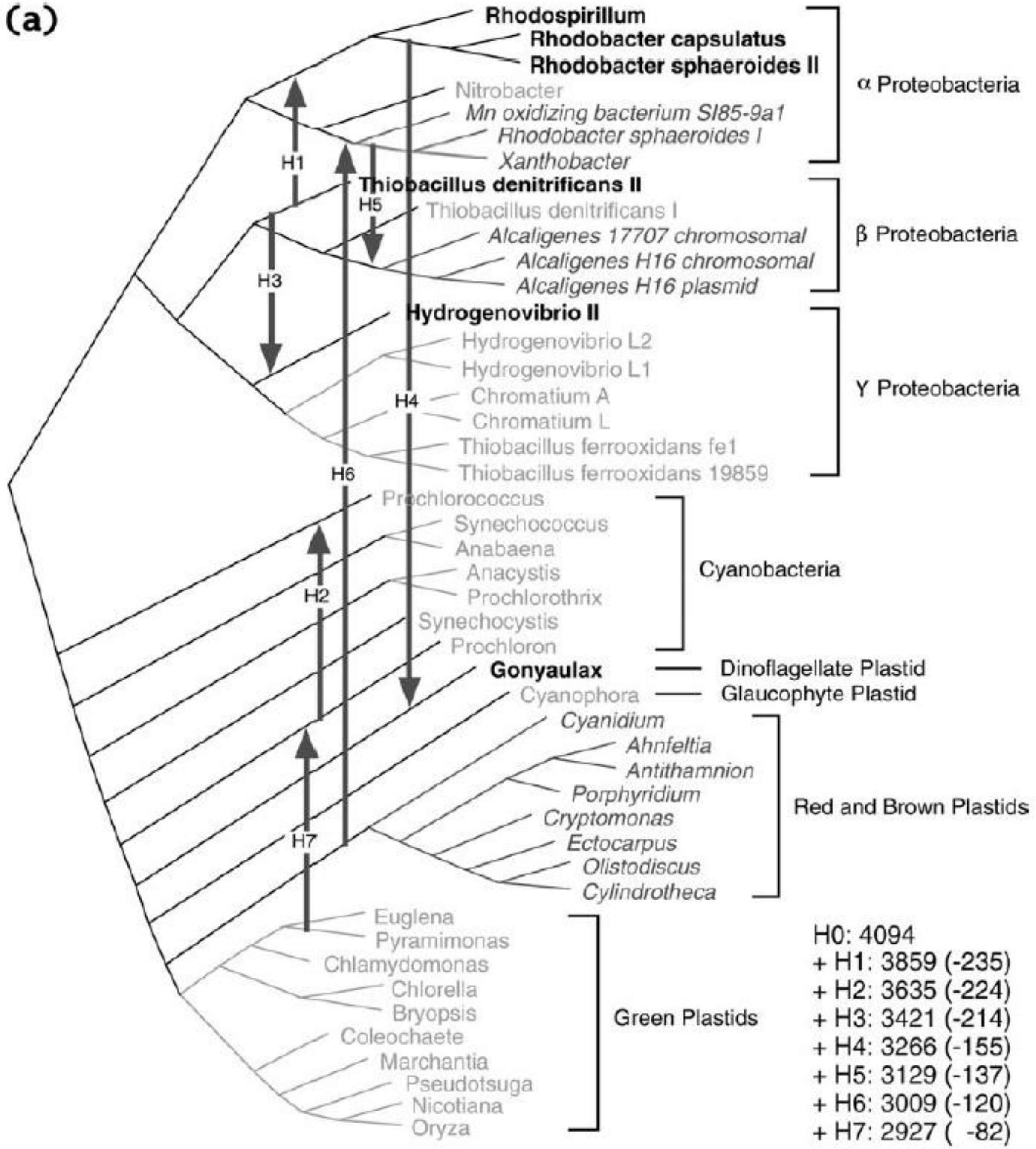


FIG. 4.—Optimal improvement in the parsimony score as extra edges are added to the species tree to obtain a phylogenetic network on the *rbcL* data set. The most significant improvements are obtained by adding the first 7 HGT edges in the case of the 46-taxon data set and the first 4 HGT edges in the case of the 40-taxon data set.

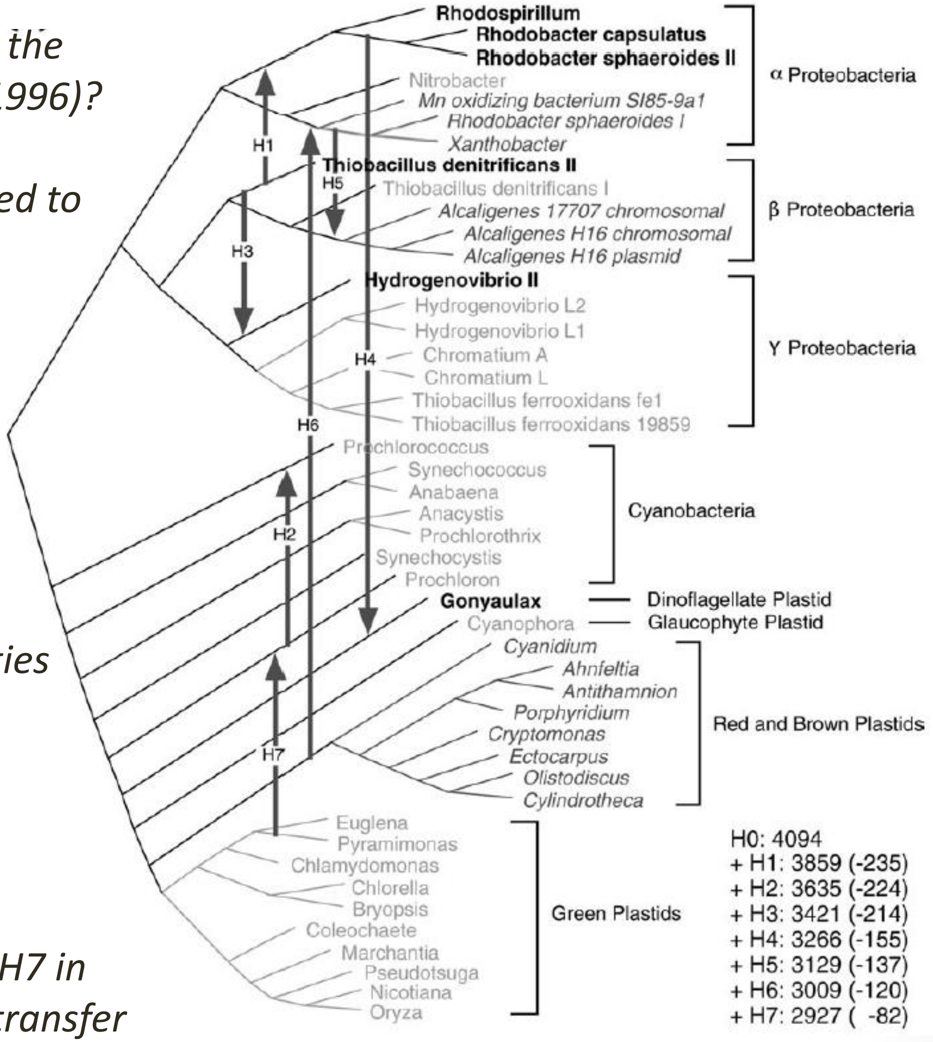


H0: 4094
+ H1: 3859 (-235)
+ H2: 3635 (-224)
+ H3: 3421 (-214)
+ H4: 3266 (-155)
+ H5: 3129 (-137)
+ H6: 3009 (-120)
+ H7: 2927 (-82)

"How do these findings compare with the hypotheses of Delwiche and Palmer (1996)? The authors postulated that at least 4 independent HGTs were required to explain the division of plastids and proteobacteria into the greenlike and redlike groups…

…Furthermore, they postulated 3 more HGTs to account for incongruities in the rbcL phylogeny….

Finally, our analysis gave rise to edge H7 in figure 5a, which gives indication of a transfer that was not postulated by the authors, but among all 7 edges found in our analysis, this edge led to the smallest improvement in the parsimony score…."



"How do these findings compare with the hypotheses of Delwiche and Palmer (1996)? The authors postulated that at least 4 independent HGTs were required to explain the division of plastids and proteobacteria into the greenlike and redlike groups…

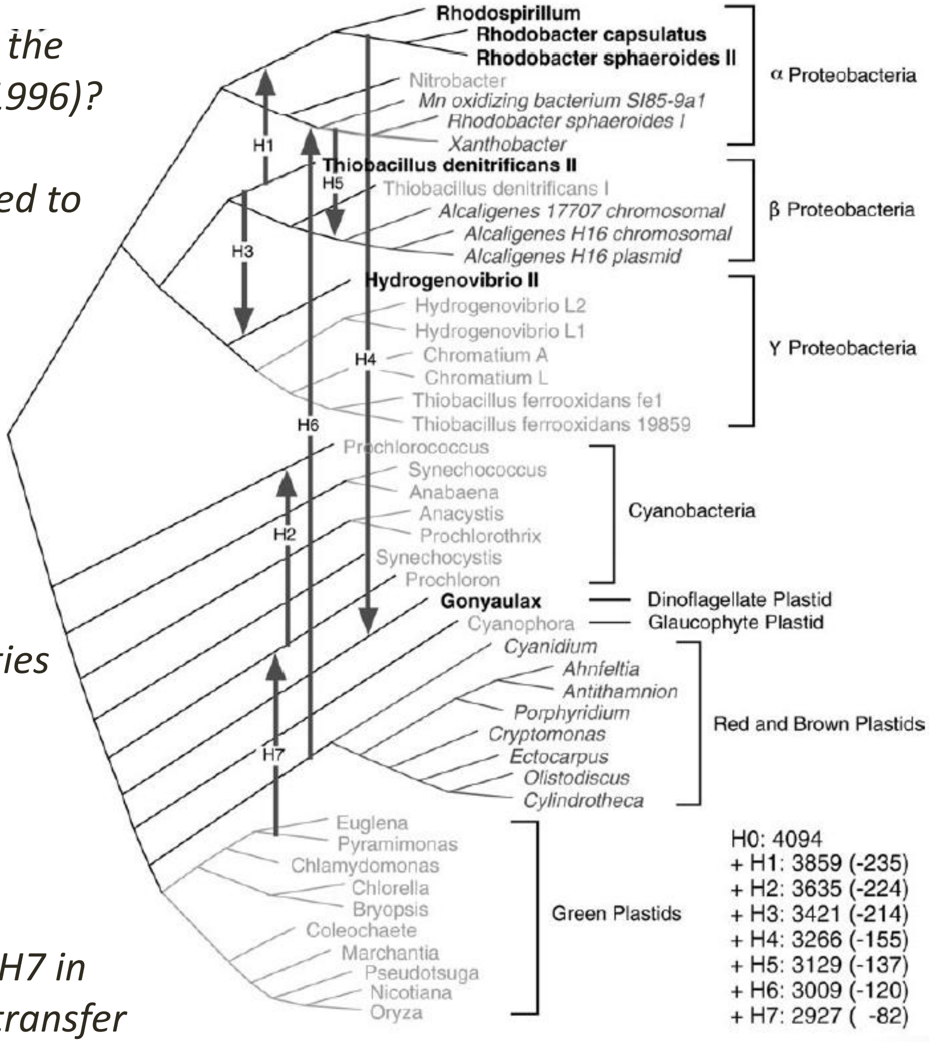
Functional equivalence:
H6, H1, H3

…Furthermore, they postulated 3 more HGTs to account for incongruities in the rbcL phylogeny….

Functional equivalence:
H4, H2, H5

Finally, our analysis gave rise to edge H7 in figure 5a, which gives indication of a transfer that was not postulated by the authors, but among all 7 edges found in our analysis, this edge led to the smallest improvement in the parsimony score…."

"Weak false positive" : H7



“How do these findings compare with the hypotheses of Delwiche and Palmer (1996)? The authors postulated that at least 4 independent HGTs were required to explain the division of plastids and proteobacteria into the greenlike and redlike groups…

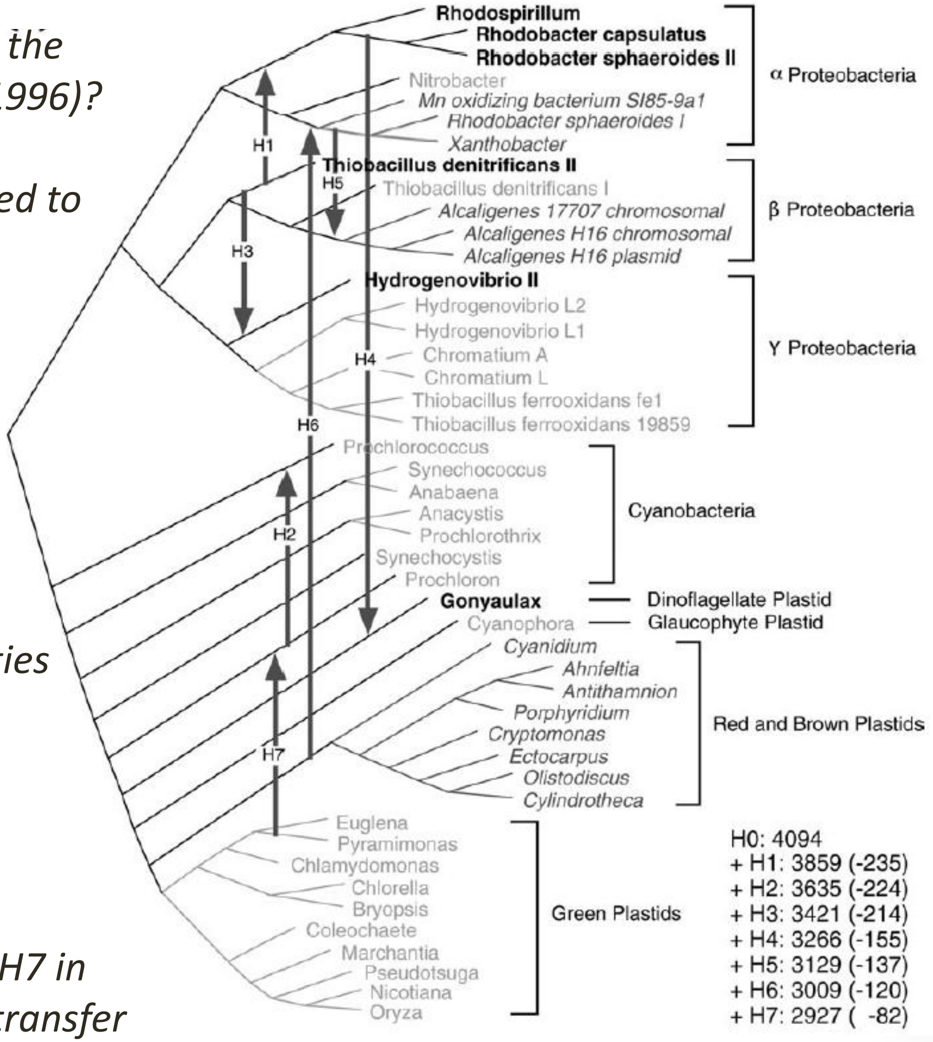
Functional equivalence:
H6, H1, H3

…Furthermore, they postulated 3 more HGTs to account for incongruities in the rbcL phylogeny….

Functional equivalence:
H4, H2, H5

Finally, our analysis gave rise to edge H7 in figure 5a, which gives indication of a transfer that was not postulated by the authors, but among all 7 edges found in our analysis, this edge led to the smallest improvement in the parsimony score….”

“Weak false positive” : H7

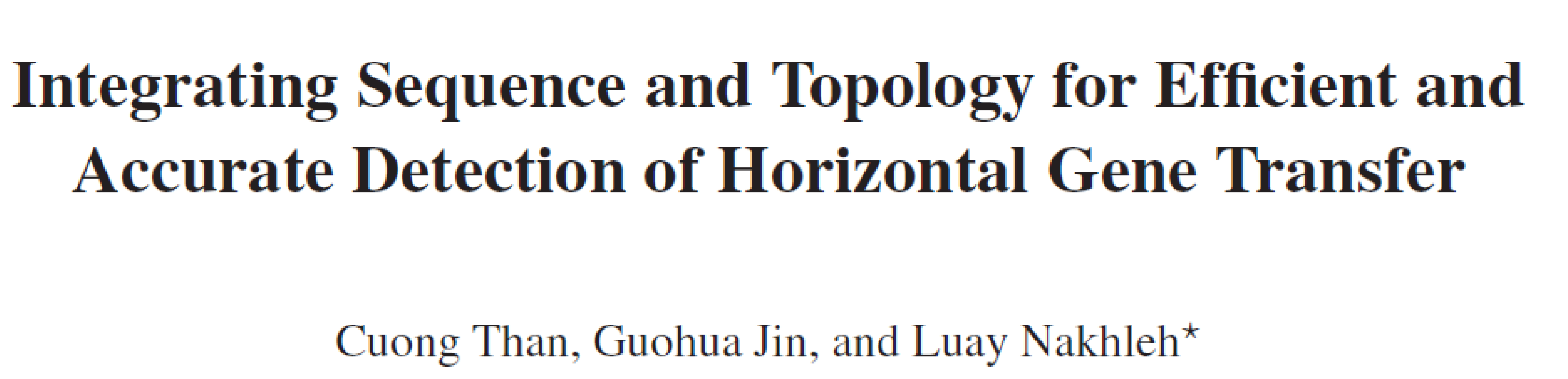


Conclusion: output of MP algorithm mostly consistent with predictions of Delwich and Palmer (1996)

Inferring phylogenetic networks by the maximum parsimony criterion: a case study

- **Similar message for other 3 datasets:**
  - High-confidence HGT events hypothesized by earlier articles are recovered (at least in a functional sense)
  - Other HGT events seem to reflect species-gene tree incongruence observed in earlier articles (this is not obvious: MP is a topology vs. sequence method, not topology vs. topology)
  - Some “false positives” but not too many

Case study 2: MP follow-up



Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

Cuong Than, Guohua Jin, and Luay Nakhleh*

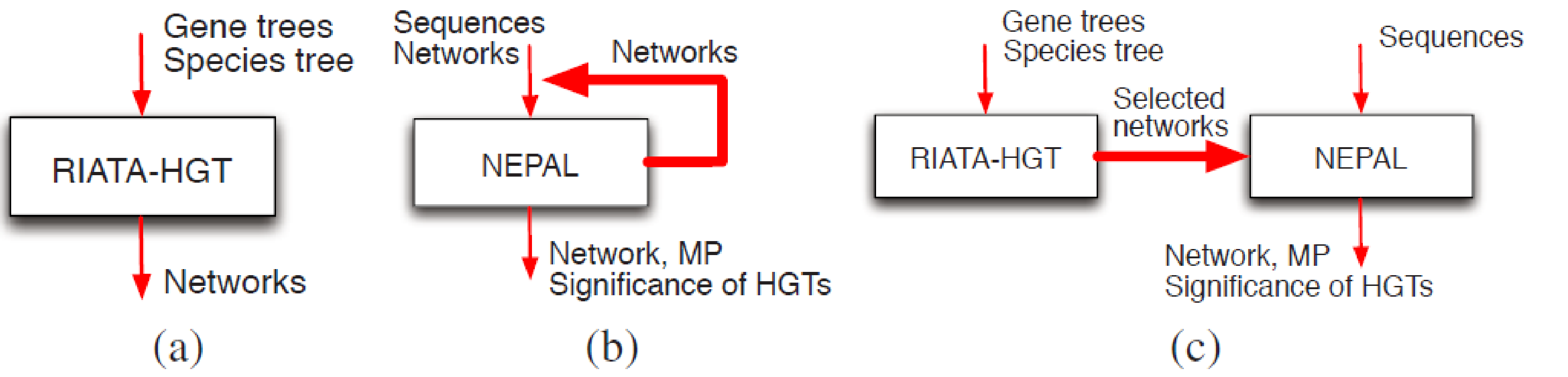
RECOMB-CG 2008

Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

- The previous MP algorithm was topology vs. sequence
- But generating HGT events this way is computationally devastatingly hard (=too slow)

- So generate the putative HGT events with a classical reconciliation-style approach (i.e. species tree topology vs. gene tree topology) and then use MP to filter them.
- Also: quality of the HGT events postulated by topology vs. topology analysis can be assessed by considering bootstraps of input trees

Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer



Conclusion: approach (c) seems to combine speed of (a) with accuracy of (b)

Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

- Re-analysis of *“a data set of 20 genes that exhibited massive HGT in the basal angiosperm Amborella according to Bergthorsson et. al. (2004)”*

- Species tree obtained from NCBI (http://www.ncbi.nih.gov)

- *[4] Bergthorsson, U., Richardson, A., Young, G.J., Goertzen, L., Palmer, J.D.: Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. Proc. Nat’l Acad. Sci., USA 101, 17747–17752 (2004)*

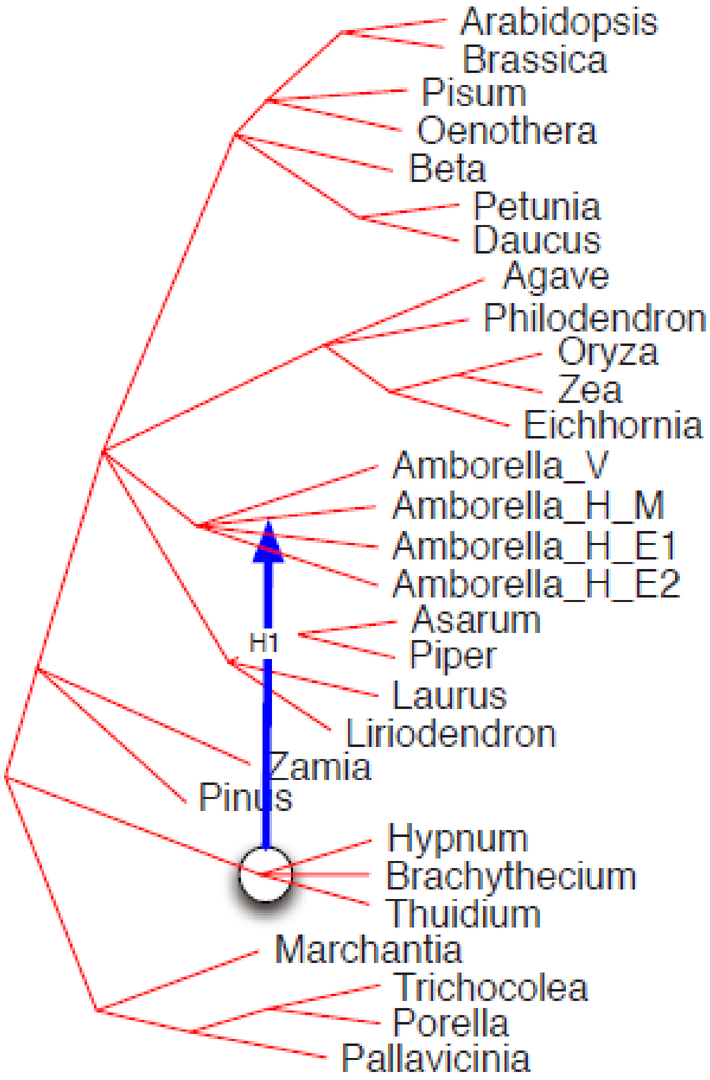
Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

- Bergthorsson et. al. applied the likelihood-based SH test to 25 putative HGT events:
  - 13 were supported, 9 unsupported, and 3 (the 3 intron data sets) had no reported support.

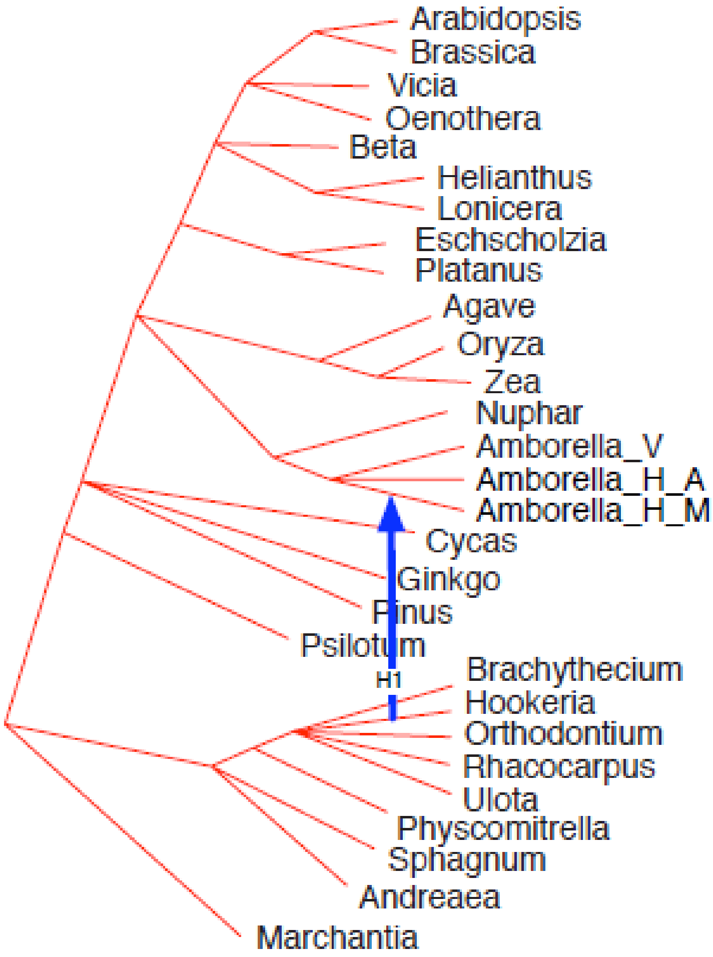
- In all cases the different topology/sequence combinations tried could recover most (11-12) of the 13. Conclusion: few false negatives with respect to SH test…?

- Of the 9 unsupported, 5 were not recovered, but 4 were. Conclusion: some false positives with respect to SH test…?

- 8 events not identified by Bergthorsson et al. Conclusion: events missed by original “into Amborella” analysis? Other problems?

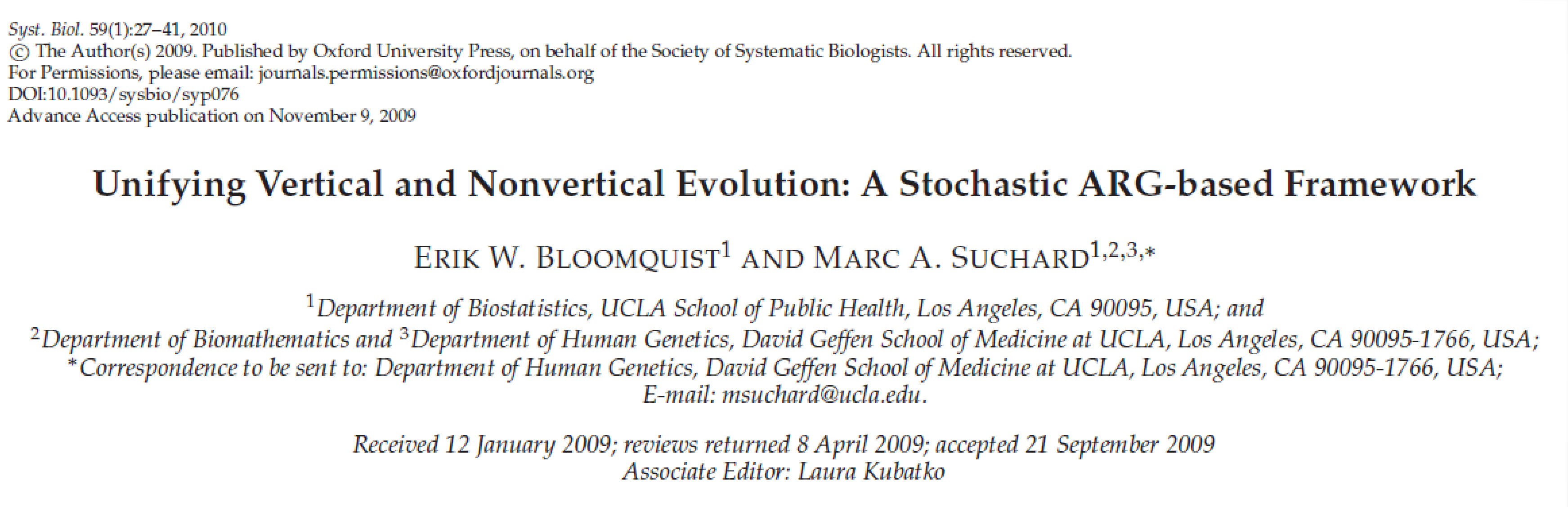


The *cox2* gene data set



The *nad5* gene data set

Case study 3: statistical methods



Syst. Biol. 59(1):27–41, 2010

DOI:10.1093/sysbio/syp076
Advance Access publication on November 9, 2009

Unifying Vertical and Nonvertical Evolution: A Stochastic ARG-based Framework

ERIK W. BLOOMQUIST[1] AND MARC A. SUCHARD[1,2,3,*]

[1]*Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095, USA; and*
[2]*Department of Biomathematics and* [3]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-1766, USA;*
[*]*Correspondence to be sent to: Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-1766, USA;*
E-mail: msuchard@ucla.edu.

Received 12 January 2009; reviews returned 8 April 2009; accepted 21 September 2009
Associate Editor: Laura Kubatko

Case study 3: statistical methods

- Idea: most statistical methods only provide indirect information about reticulation events. ARGs (i.e. evolutionary phylogenetic networks) provide a way to make this information explicit.

- Likelihood-based Coalescent with Recombination is computationally far too intensive

- Uses Bayesian/MCMC methods to move through ARG-space

- “To demonstrate our model, we analyze 2 empirical examples. The first examines a *Leptospira interrogans* data set in order to gain more information on the evolutionary history (Stevenson et al. 2007), and the second explores a *Saccharomyces* data set taken from Rokas et al. (2003).”

Case study 3: statistical methods

- *“Stevenson et al. (2007) suggest that the lenF gene in several serovars (lineages) of L. interrogans is actually the product of a nonvertical transmission event and fusion between an ancestral lenC lineage and lenF lineage using the gene-tree methodology of Suchard et al. (2005). Specifically, Stevenson et al. (2007) use a Bayes’s factor test to determine whether the lenF lineage forms a monophyletic clade.”*

- They re-analyse this dataset with their statistical method to compute “the most probable ARG”

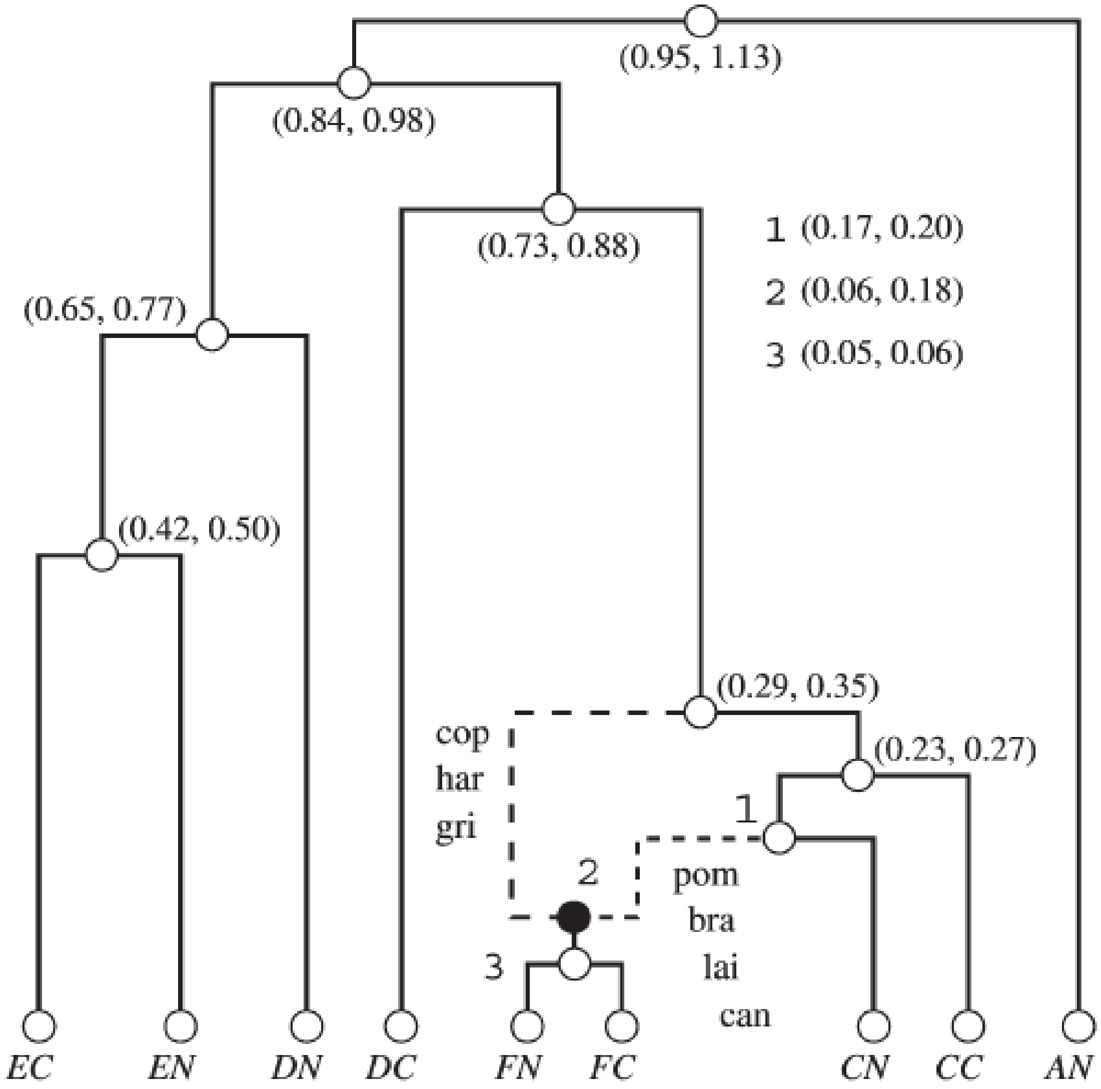


FIGURE 1. Nonvertical evolution confirmation and event dating. The figure shows most probable ARG that represents the evolutionary history of 9 members of the *len* family in *Leptospira interrogans*. For

Case study 3: statistical methods

- *“Although SMARTIE recovers a single nonvertical event and essentially confirms the results of Stevenson et al. (2007), SMARTIE provides numerous advantages over the previous analysis. Importantly, we gain substantially more information on the evolutionary history….”*

- More information = lots more statistical support information

- And then *Saccharomyces* again… (Rokas 2003):

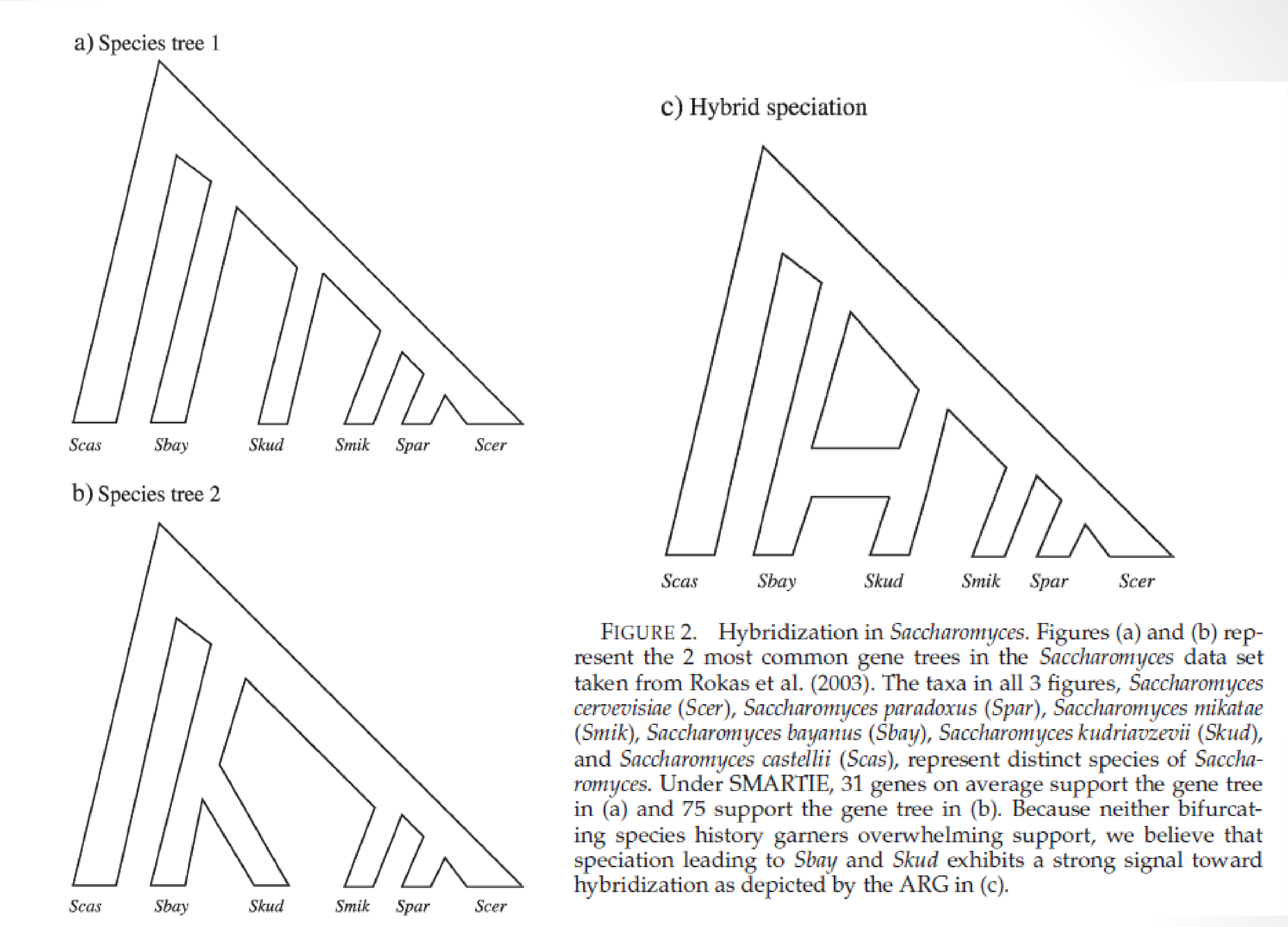


FIGURE 2. Hybridization in *Saccharomyces*. Figures (a) and (b) represent the 2 most common gene trees in the *Saccharomyces* data set taken from Rokas et al. (2003). The taxa in all 3 figures, *Saccharomyces cervevisiae* (*Scer*), *Saccharomyces paradoxus* (*Spar*), *Saccharomyces mikatae* (*Smik*), *Saccharomyces bayanus* (*Sbay*), *Saccharomyces kudriavzevii* (*Skud*), and *Saccharomyces castellii* (*Scas*), represent distinct species of *Saccharomyces*. Under SMARTIE, 31 genes on average support the gene tree in (a) and 75 support the gene tree in (b). Because neither bifurcating species history garners overwhelming support, we believe that speciation leading to *Sbay* and *Skud* exhibits a strong signal toward hybridization as depicted by the ARG in (c).

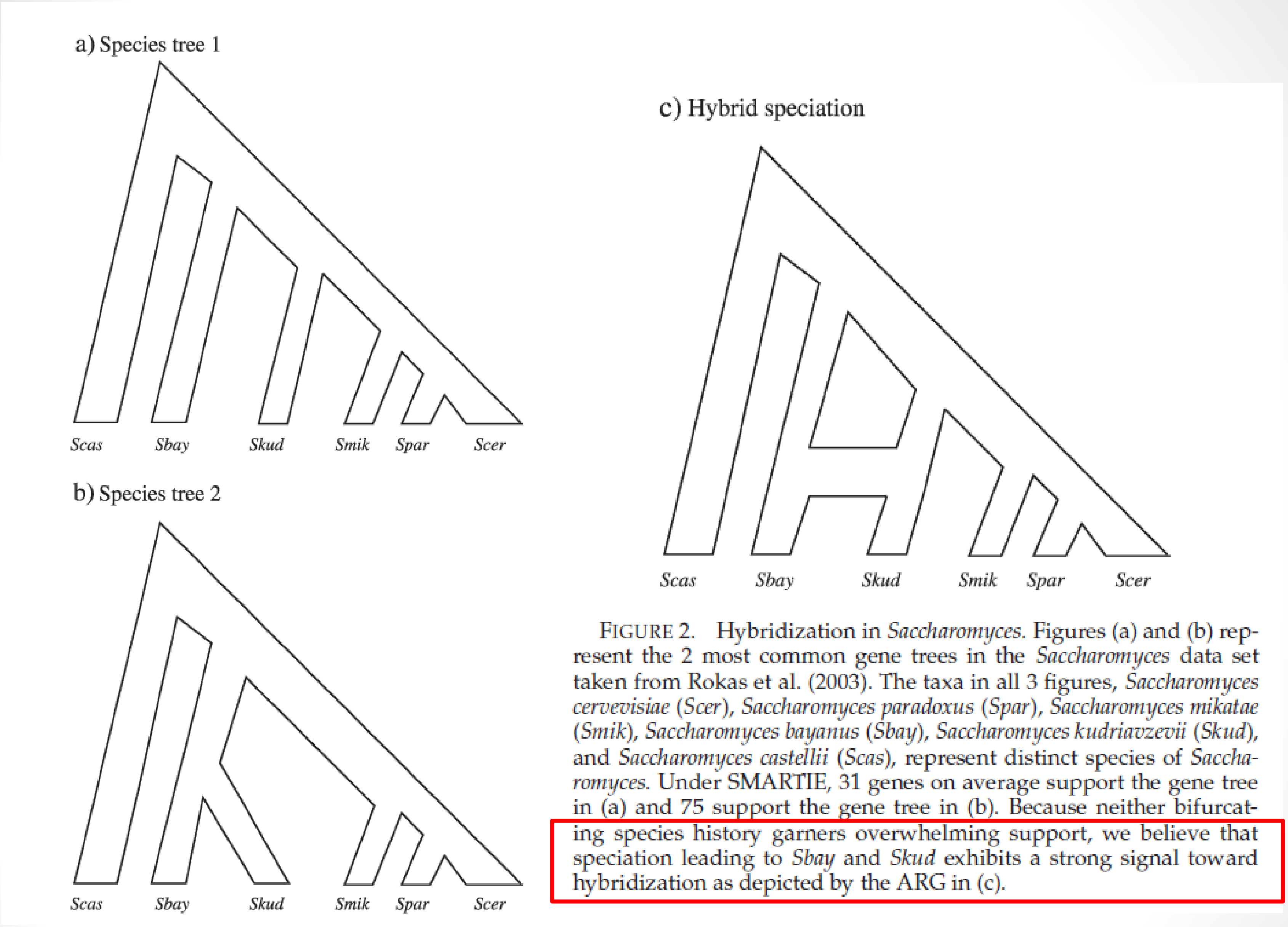


FIGURE 2. Hybridization in *Saccharomyces*. Figures (a) and (b) represent the 2 most common gene trees in the *Saccharomyces* data set taken from Rokas et al. (2003). The taxa in all 3 figures, *Saccharomyces cervevisiae* (*Scer*), *Saccharomyces paradoxus* (*Spar*), *Saccharomyces mikatae* (*Smik*), *Saccharomyces bayanus* (*Sbay*), *Saccharomyces kudriavzevii* (*Skud*), and *Saccharomyces castellii* (*Scas*), represent distinct species of *Saccharomyces*. Under SMARTIE, 31 genes on average support the gene tree in (a) and 75 support the gene tree in (b). Because neither bifurcating species history garners overwhelming support, we believe that speciation leading to *Sbay* and *Skud* exhibits a strong signal toward hybridization as depicted by the ARG in (c).

Unclear how they obtained (c)…

Case study 3: statistical methods

- Conclusion later supported and refined in *"The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection",* Yu Y, Degnan JH, Nakhleh L, PLoS Gen. 8(4), 2012

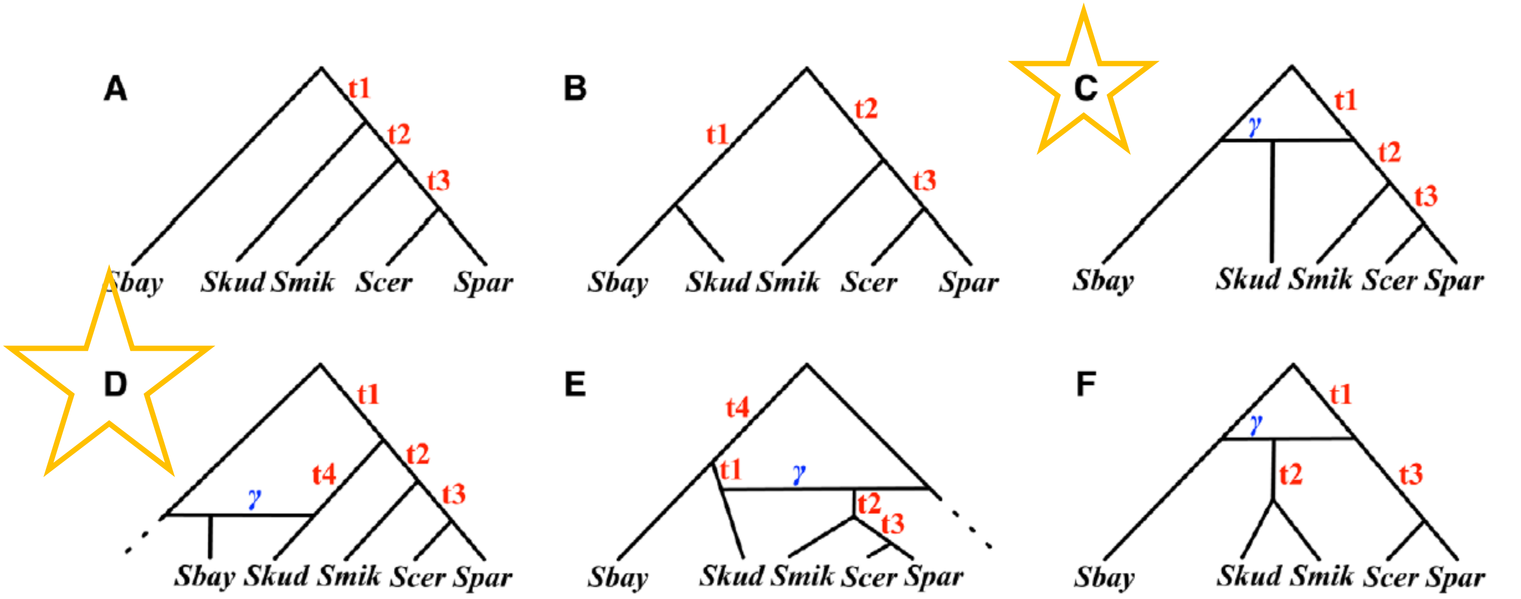


Figure 2. Various hypotheses for the evolutionary history of a yeast data set. (A) The species tree for the five species *Sbay*, *Skud*, *Smik*, *Scer*, and *Spar*, as proposed in [35], and inferred using a Bayesian approach [39] and a parsimony approach [36]. (B) A slightly suboptimal tree for the five species, as identified in [36,39]. (C–E) The three phylogenetic networks that reconcile both trees in (A) and (B), and which we reported as equally optimal evolutionary histories under a parsimony criterion in [30]. (F) A phylogenetic network that postulates *Smik* and *Skud* as two sister taxa whose divergence followed a hybridization event.

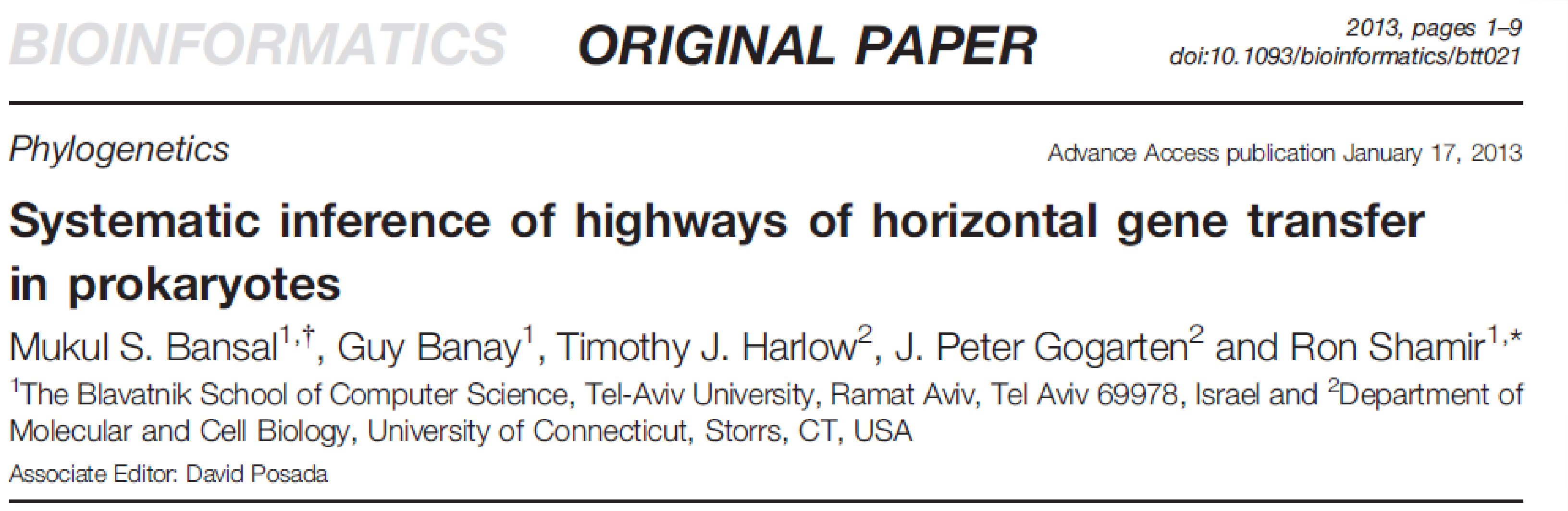
Case study 3: statistical methods

- Conclusion later supported and refined in *“The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection”,* Yu Y, Degnan JH, Nakhleh L, PLoS Gen. 8(4), 2012

- Yu et al: *“In summary, our analysis gives higher support for the hypothesis of extensive hybridization, a low degree of deep coalescence, and long branch lengths than to the hypothesis of a species tree with short branches and extensive deep coalescence.”*

- (Deep Coalescence = Incomplete Lineage Sorting)

Case study 4: “highways”



BIOINFORMATICS ORIGINAL PAPER

2013, pages 1–9
doi:10.1093/bioinformatics/btt021

Phylogenetics Advance Access publication January 17, 2013

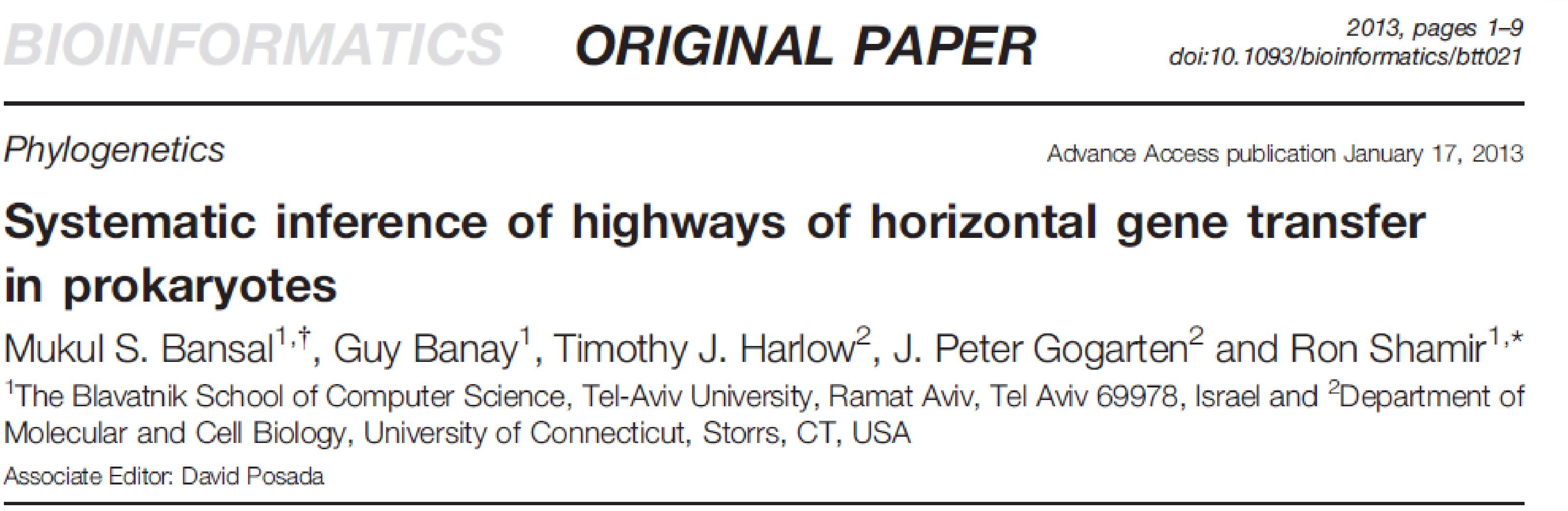
Systematic inference of highways of horizontal gene transfer in prokaryotes

Mukul S. Bansal[1,†], Guy Banay[1], Timothy J. Harlow[2], J. Peter Gogarten[2] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv, Tel Aviv 69978, Israel and [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Associate Editor: David Posada

Case study 4: “highways”



BIOINFORMATICS **ORIGINAL PAPER**

2013, pages 1–9
doi:10.1093/bioinformatics/btt021

Phylogenetics

Advance Access publication January 17, 2013

Systematic inference of highways of horizontal gene transfer in prokaryotes

Mukul S. Bansal[1,†], Guy Banay[1], Timothy J. Harlow[2], J. Peter Gogarten[2] and Ron Shamir[1,*]

[1]The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv, Tel Aviv 69978, Israel and [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Associate Editor: David Posada

“Highways” are simply HGT events that many different genes seem to support (when mapped individually onto the same species tree).

Arguably highways are the point at which species-gene tree reconciliation merges with the idea of a species network

“These highways point towards major events in evolutionary history; well corroborated examples are the uptake of endosymbionts into the eukaryotic host, and the many genes transferred from the symbiont to the host’s nuclear genome (Gary, 1993). Recent proposals for evolutionary events that may be reflected in highways are the role of Chlamydiae in establishing the primary plastid in the Archaeplastida (red and green algae, plants and glaucocystophytes) (Becker et al., 2008; Huang and Gogarten, 2007; Moustafa et al., 2008), the evolution of double membrane bacteria through an endosymbiosis between clostridia and actinobacteria (Lake, 2009) and the high rate of transfer between marine Synechococcus and Prochlorococcus (Zhaxybayeva et al., 2006, 2009a).”

biological justification…

Case study 4: “highways”

- Bansal et al (2013) detect highways by decomposing gene trees into quartets (trees on 4 taxa) and analysing the conflict between these quartets.

- *“We also applied the method to a dataset of 144 taxa and 22,430 gene trees from Beiko et al. (2005). Our results are largely consistent with previous analyses of this dataset, and the entire computational analysis of this large dataset took < 2 days (using a single CPU). Our new method thus makes it possible to easily, quickly and accurately infer highways even for large datasets as well as on datasets with high rates of HGT.”*

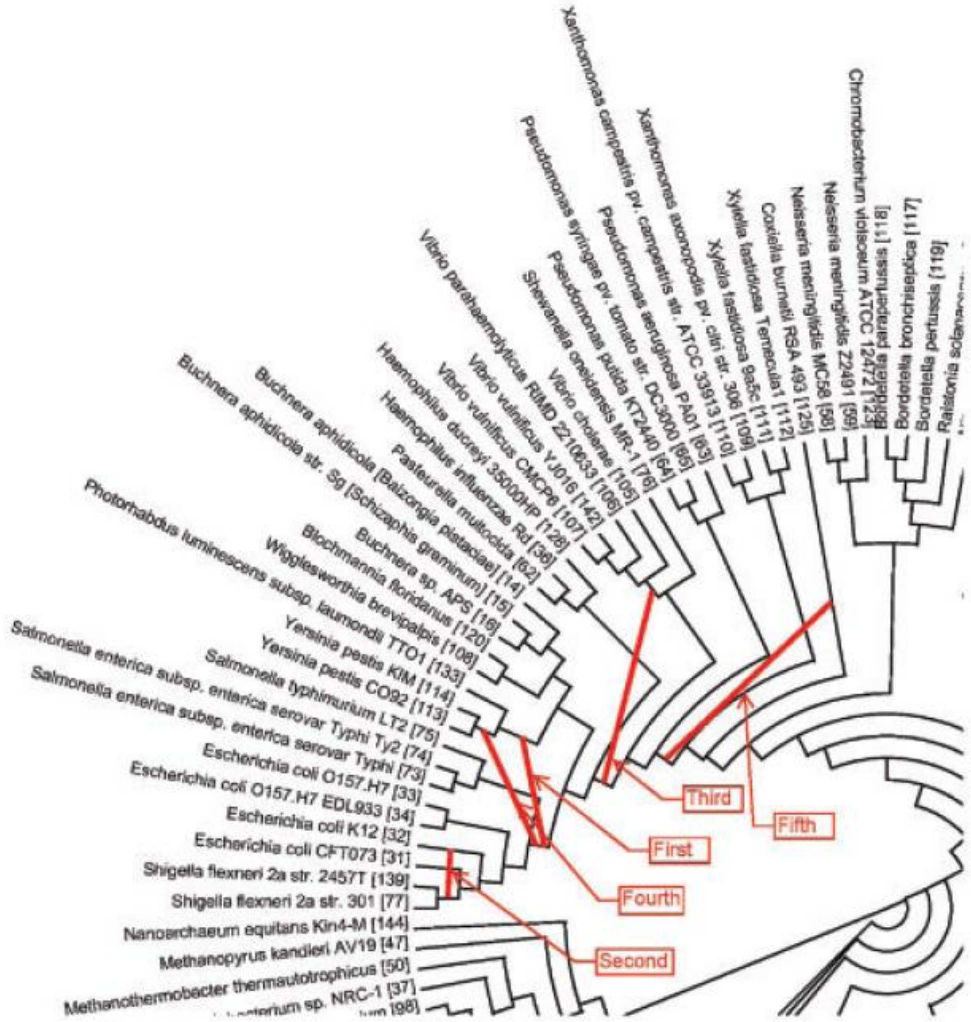

Methanothermobacter thermautotrophicus [50]
Methanopyrus kandleri AV19 [47]
Nanoarchaeum equitans Kin4-M [144]
Shigella flexneri 2a str. 301 [77]
Shigella flexneri 2a str. 2457T [139]
Escherichia coli CFT073 [31]
Escherichia coli K12 [32]
Escherichia coli O157.H7 EDL933 [34]
Escherichia coli O157.H7 [33]
Salmonella enterica subsp. enterica serovar Typhi [73]
Salmonella enterica subsp. entarica serovar Typhi Ty2 [74]
Salmonella typhimurium LT2 [75]
Yersinia pestis CO92 [113]
Yersinia pestis KIM [114]
Photorhabdus luminescens subsp. laumondii TTO1 [133]
Wigglesworthia brevipalpis [108]
Blochmannia floridanus [120]
Buchnera sp. APS [16]
Buchnera aphidicola str. Sg [Schizaphis graminum] [15]
Buchnera aphidicola [Baizongia pistaciae] [14]
Pasteurella multocida [62]
Haemophilus influenzae Rd [36]
Haemophilus ducreyi 35000HP [128]
Vibrio vulnificus CMCP6 [107]
Vibrio vulnificus YJ016 [142]
Vibrio parahaemolyticus RIMD 2210633 [106]
Vibrio cholerae [105]
Shewanella oneidensis MR-1 [76]
Pseudomonas putida KT2440 [64]
Pseudomonas syringae pv. tomato str. DC3000 [65]
Pseudomonas aeruginosa PA01 [63]
Xanthomonas campestris pv. campestris str. ATCC 33913 [110]
Xanthomonas axonopodis pv. citri str. 306 [109]
Xylella fastidiosa 9a5c [111]
Xylella fastidiosa Temecula1 [112]
Coxiella burnetii RSA 493 [125]
Neisseria meningitidis MC58 [58]
Neisseria meningitidis Z2491 [59]
Chromobacterium violaceum ATCC 12472 [123]
Bordetella parapertussis [118]
Bordetella bronchiseptica [117]
Bordetella pertussis [119]
Ralstonia solana
Third
First
Fourth
Second
Fifth
bacterium sp. NRC-1 [37]
um [98]

Fig. 4. Results on the dataset of Beiko *et al.* (2005) The top five highways, along with their ranks, computed by the method are marked in red (bold edges). The reported scores for these top five highways are 83.3, 52.5, 42.9, 35.1 and 24.3. Since the top five highways are each within the gamma proteobacteria, the figure focuses on only that portion of the phylogeny (we refer the reader to Supplementary Figure S9 for a figure showing the full phylogeny). The tree was drawn using Dendroscope (Huson *et al.*, 2007)

- The 22,430 genes generate approximately 5,000 HGT events

- There is a large gap between the number of times the top 5-6 HGT events are used, and the rest.

- Conclusion: there are approximately 5 highways, shown here

- Similar conclusion to the analysis of Beiko (2005)

Case study 4: “highways”

- Earlier, in Bansal (2011), an algorithmically slightly less advanced technique was applied to a different dataset:

- *“We applied our method to a dataset of 1128 genes from 11 cyanobacterial species, taken from Zhaxybayeva et al. 2006. The existence of a highway on this set of species was postulated in Zhaxybayeva et al. 2006, 2009 and thus this dataset serves for method validation.”*

- They identify up to 3 highways, arguing robustly that (Zhaxybayeva et al. 2006, 2009) offers biological justification for the first, and that the other two are plausible:

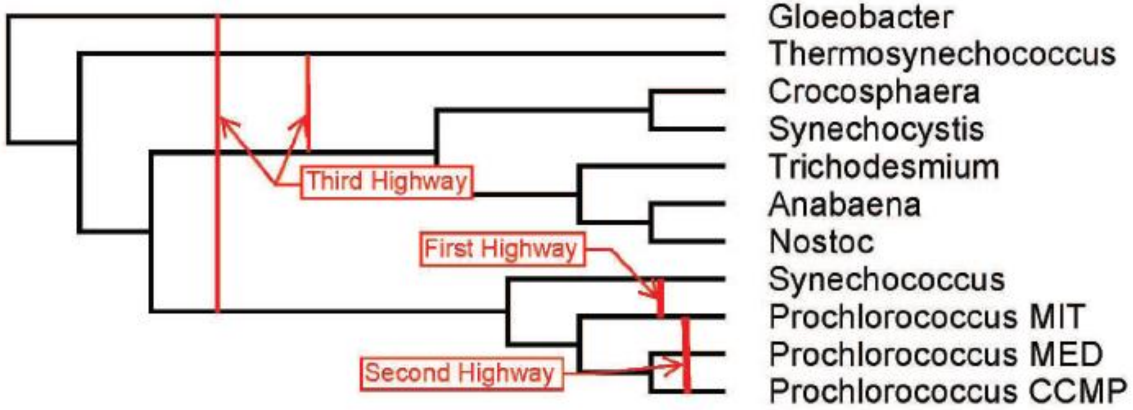


Figure 8: The 16SrRNA tree on the 11 cyanobacterial species, with detected highways marked.

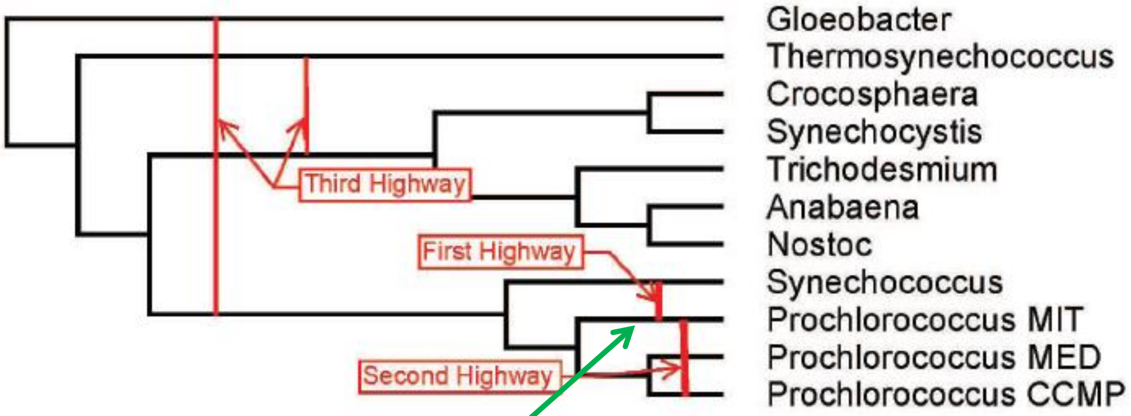
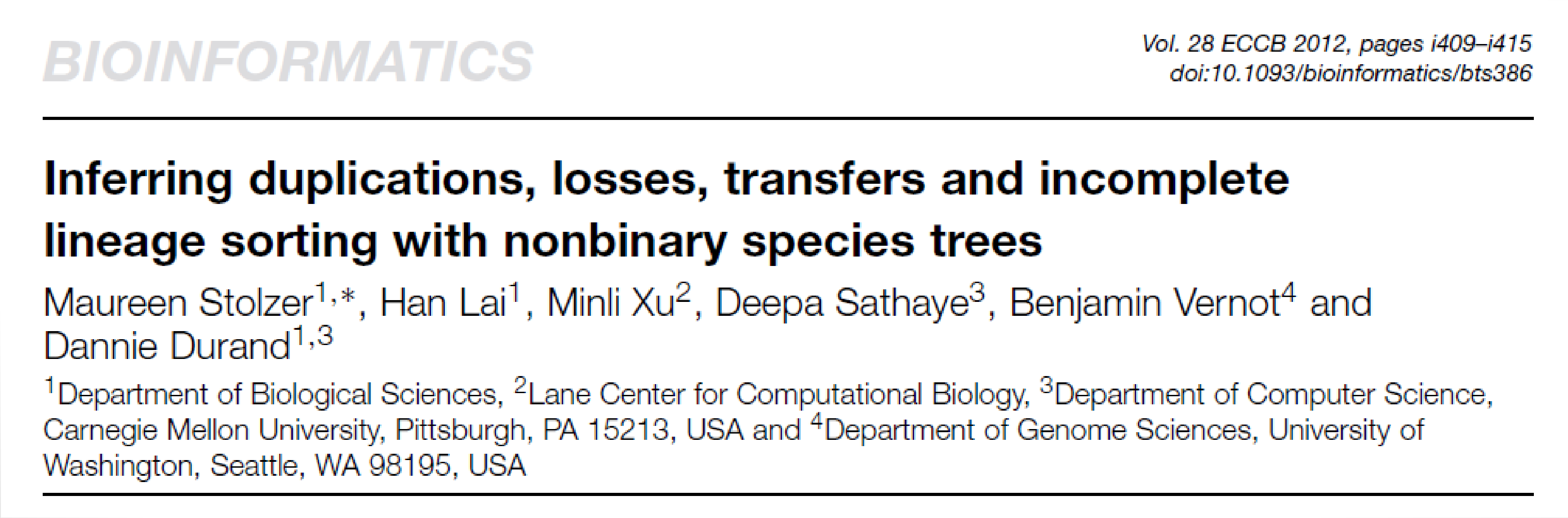


Figure 8: The 16SrRNA tree on the 11 cyanobacterial species, with detected highways marked.

This is the one they are most confident about

Case study 4: “highways”

- In the article below from 2012 the DLT-reconciliation model (deletion, loss, transfer) is extended to include Incomplete Lineage Sorting, yielding DLTI-reconciliation.
- They re-analyse the dataset discussed in Bansal (2011)…



BIOINFORMATICS

Vol. 28 ECCB 2012, pages i409–i415
doi:10.1093/bioinformatics/bts386

Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees

Maureen Stolzer[1,*], Han Lai[1], Minli Xu[2], Deepa Sathaye[3], Benjamin Vernot[4] and Dannie Durand[1,3]

[1]Department of Biological Sciences, [2]Lane Center for Computational Biology, [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

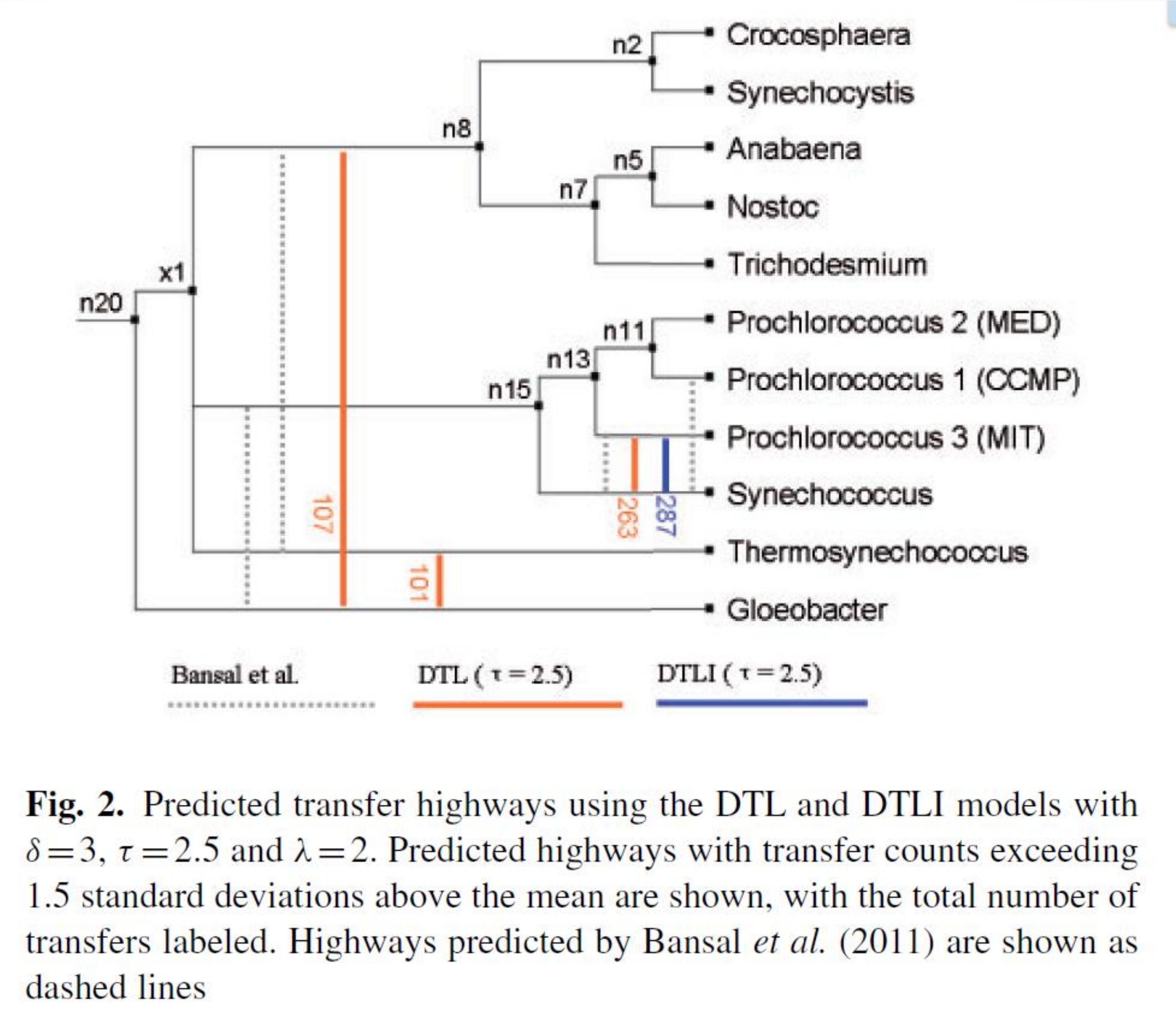


Fig. 2. Predicted transfer highways using the DTL and DTLI models with $\delta = 3$, $\tau = 2.5$ and $\lambda = 2$. Predicted highways with transfer counts exceeding 1.5 standard deviations above the mean are shown, with the total number of transfers labeled. Highways predicted by Bansal *et al.* (2011) are shown as dashed lines

- **In both the DLT and DLTI models they obtain support for the most well-supported highway identified by Bansal et al (blue 287 / orange 263 line). ☺**

- **But in the DLTI model the remaining highways vanish: they are (apparently) better explained by incomplete lineage sorting…**

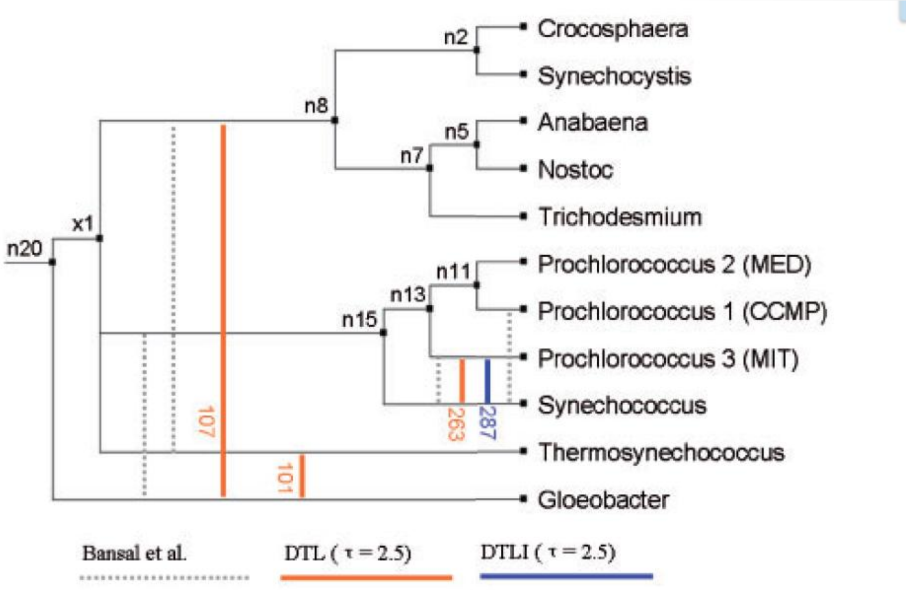


Fig. 2. Predicted transfer highways using the DTL and DTLI models with $\delta = 3$, $\tau = 2.5$ and $\lambda = 2$. Predicted highways with transfer counts exceeding 1.5 standard deviations above the mean are shown, with the total number of transfers labeled. Highways predicted by Bansal *et al.* (2011) are shown as dashed lines

- In both the DLT and DLTI models they obtain support for the most well-supported highway identified by Bansal et al (blue 287 / orange 263 line). ☺

- But in the DLTI model the remaining highways vanish: they are (apparently) better explained by incomplete lineage sorting…

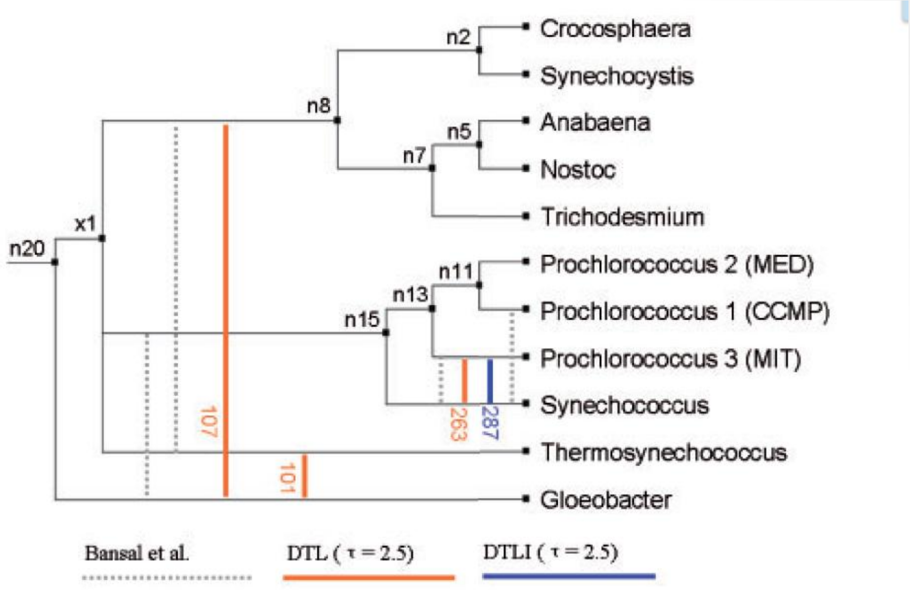


Fig. 2. Predicted transfer highways using the DTL and DTLI models with $\delta = 3$, $\tau = 2.5$ and $\lambda = 2$. Predicted highways with transfer counts exceeding 1.5 standard deviations above the mean are shown, with the total number of transfers labeled. Highways predicted by Bansal *et al.* (2011) are shown as dashed lines

"…it is possible that apparent HGT highways could be, at least in part, mis-interpretations of deep coalescence." ☹

Part 3:

Conclusions

Validation is improving

- Evolutionary phylogenetic networks – in all their different flavours – are actually being used more and more to credibly (re-)analyse "real" datasets.

- Developers of network software are getting better at leveraging the biological literature to validate the output of their software.

- The anchoring of such (re-)analysis in experimental/theoretical biology needs to be strengthened, however.

  - In some cases the biological anchor might be missing entirely ("I got the same answers as the previous group of mathematicians")

  - In some cases the biological anchor might itself be an artefact of software (circular inference)

- But overall the situation is encouraging: much better than I thought

Trends

- Pragmatic combinations of parsimony-based and statistical methods: comparative speed + resolution
- Constructive statistical methods
- Multi-event models (D-T-L-H-ILS….)
- Robustness/stability analysis (noise, uncertainty, multiple optima)
- Getting the huge size of the network search space under control (…)
- Further exploration of the interface between phylogenetics and population genetics
- Interdisciplinary research consortia

Thank you for listening!