Phylogenetics and Molecular Evolution/Filogenética e Evolução Molecular

Octávio S. Paulo
Computational Biology and Population Genomics Group (CoBiG2)

Introdução à Filogenómica

Sumário:
A análise de grandes matrizes de dados com múltiplas partições.
Evidência total versus árvore de espécie.

Phylogenomics

Opinion TRENDS in Genetics Vol.22 No.4 April 2006

Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc and Hervé Philippe

Canadian Institute for Advanced Research, Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec, Canada, H3C3J7

Phylogenomics

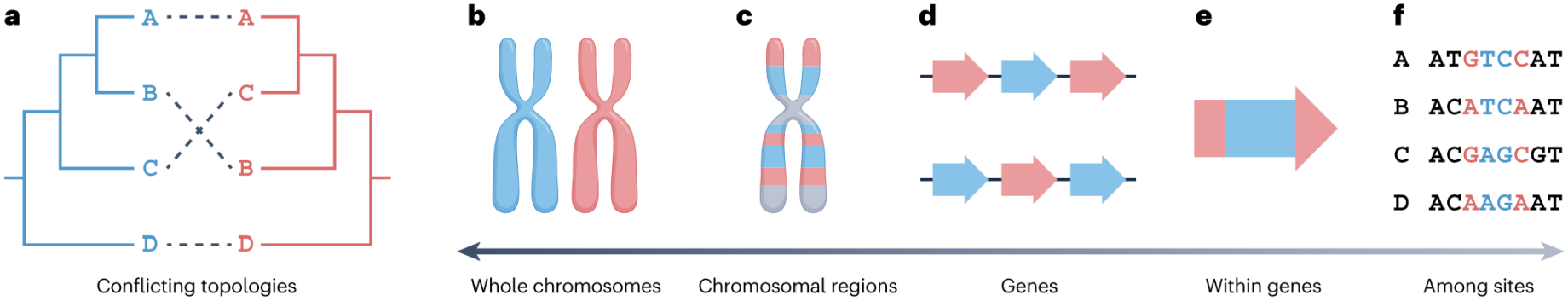
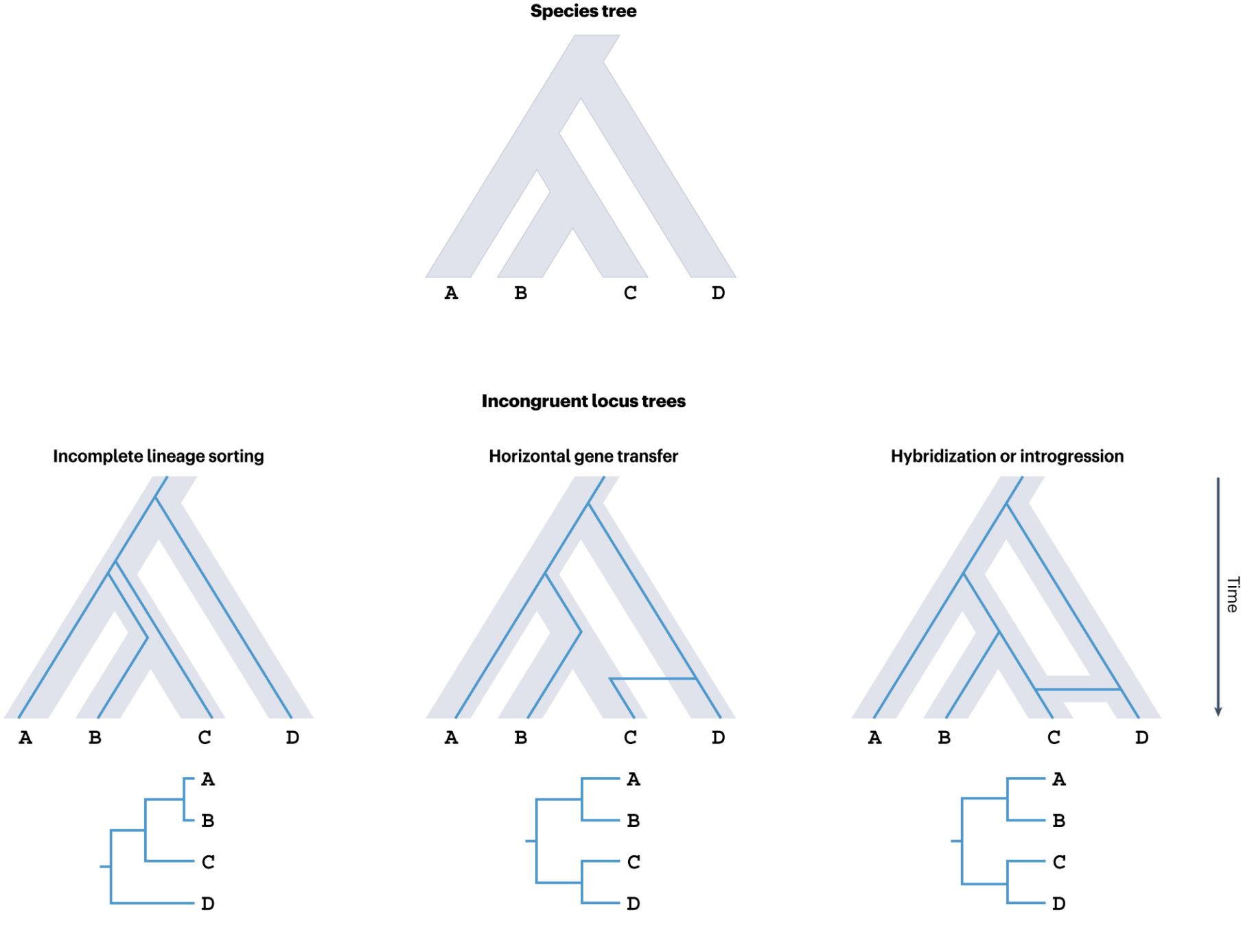
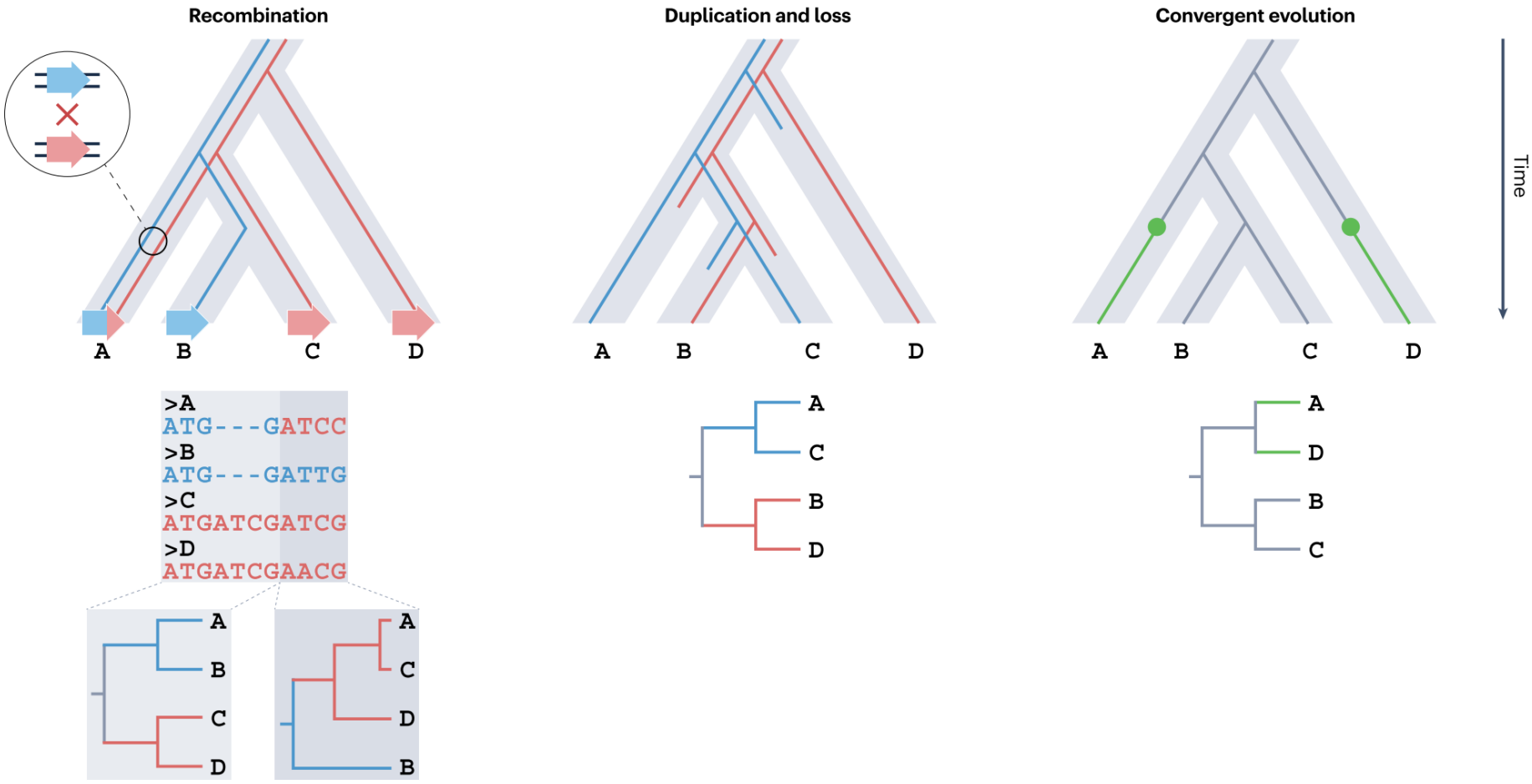


Fig. 1 | Incongruence at different levels of genomic organization. The topology shown in blue supports a sister group relationship of taxa A and B, whereas the red topology supports a sister group relationship of taxa A and C (part **a**). The inference of such conflicting topologies defines incongruence. Incongruence can occur at different levels in the genome, such as among whole chromosomes (for example, analyses of one chromosome support the blue topology but analyses of another support the red topology) (part **b**), regions of a chromosome (grey regions represent lack of homology) (part **c**), genes (or loci) (part **d**), within a gene or locus (for example, different domains support different topologies) (part **e**) and among sites in a multiple sequence alignment (part **f**). Note that incongruence is also prevalent in other types of data (for example, behavioural or morphological traits) and can occur at all evolutionary depths.


Species tree
A
B
C
D
Incongruent locus trees
Incomplete lineage sorting
Horizontal gene transfer
Hybridization or introgression
Time

Phylogenomics



Phylogenomics

Table 1 | Drivers of incongruence

Driver of incongruence	Factor
Incomplete lineage sorting	Biological
Horizontal gene transfer	Biological
Hybridization or introgression and recombination	Biological
Natural selection	Biological
Sampling (taxon and locus)	Analytical, stochastic error
Insufficient number of genes or divergent sites	Analytical, stochastic error
Erroneous orthologue detection	Analytical, systematic error
Model misspecification	Analytical, systematic error
Multiple sequence alignment errors	Analytical, treatment error
Excessive trimming	Analytical, treatment error
Inappropriate character recoding	Analytical, treatment error

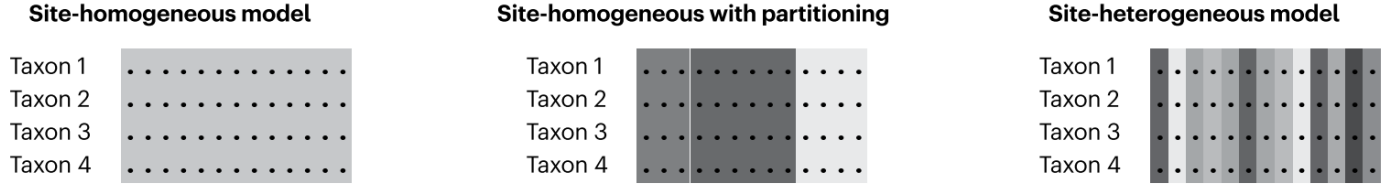
Phylogenomics

a Taxon selection

b Orthology inference

c Alignment and site trimming

d Selection of substitution model



Contributor of incongruence

- Insufficient taxon sampling
- Insufficient locus sampling
- Fast-evolving lineages
- Rogue taxa
- Outgroup choice

- Sequence length biases
- Erroneous orthologue inference (hidden paralogy and orthology)

- Misalignment
- Excessive trimming
- Inappropriate recoding

- Long-branch attraction
- Model misspecification
- Inadequate model complexity

Phylogenomics

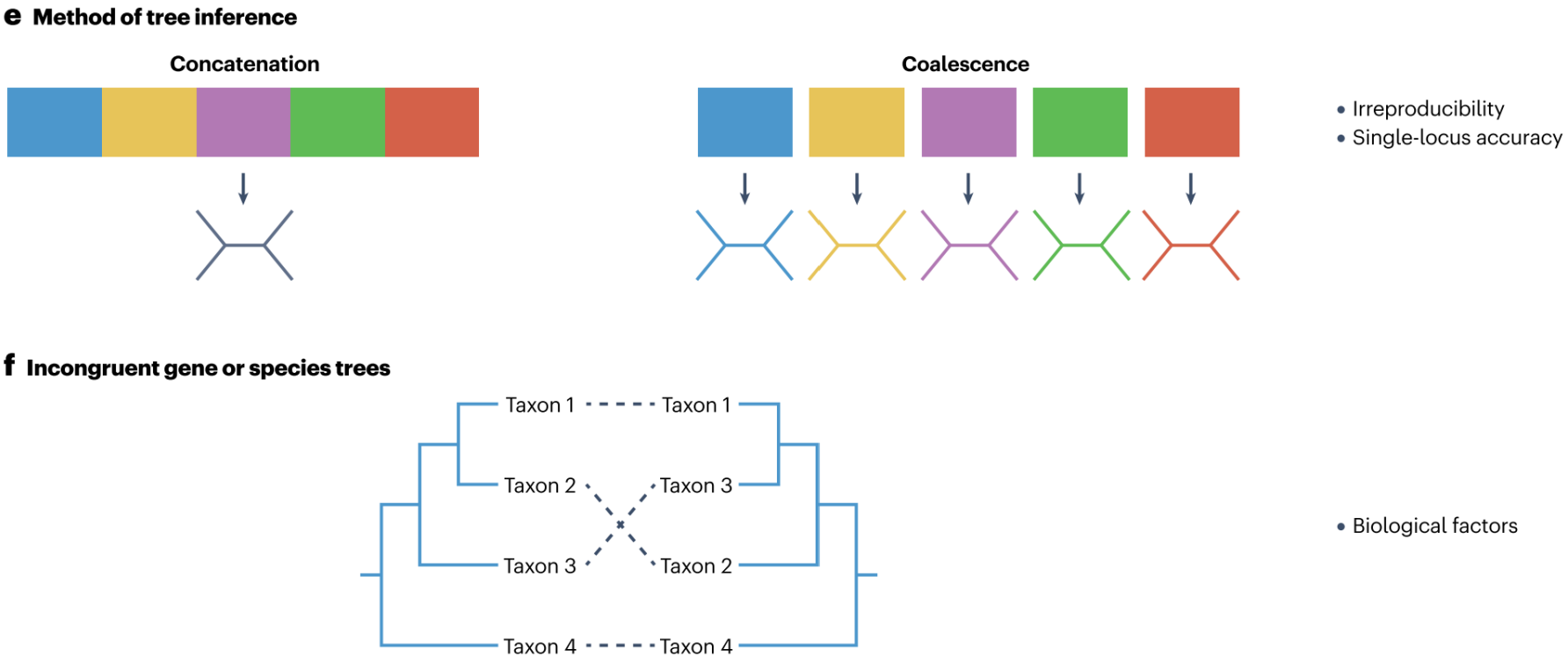


Table 2 (continued) | Tools to investigate incongruence in large genomic data sets

Software or method	Utility category	Utility details
SplitsTree	Phylogenetic network inference	Splits graph inference using multiple sequence alignments, distance matrices or sets of trees
GHOST	Substitution models	Edge-unlinked mixture model consisting of several site classes with separate sets of model parameters and edge lengths on the same tree topology
QMaker	Substitution models	Estimates general time-reversible protein matrices, which describe rates of substitutions between amino acids, from multiple sequence alignments
Asteroid	Tree inference	Supertree method for species tree inference that is robust to missing data
ASTRAL, ASTRAL-PRO and ASTER	Tree inference	Quartet-based supertree method that accounts for partial gene trees, paralogs and gene tree uncertainty
BEAST	Tree inference	Bayesian approach for phylogenetic tree inference and divergence time estimation
BPP	Tree inference	Full-likelihood implementation of the multispecies coalescent
IQ-TREE 2	Tree inference	Maximum likelihood tree inference method that uses hill-climbing and stochastic perturbation to search tree space; moreover, the Gentrius function can help identify and characterize phylogenetic terraces
MP-EST	Tree inference	Maximum pseudo-likelihood approach for species tree inference
PhyloBayes MPI	Tree inference	Bayesian tree inference method that incorporates finite and infinite mixture models to account for site variation
RAxML-NG	Tree inference	Maximum likelihood tree inference method that uses a greedy tree search algorithm to explore tree space
STAR	Tree inference	Inference of species trees using average ranks of coalescences
SpeciesRax	Tree inference	Maximum likelihood species tree inference method that explicitly accounts for incomplete lineage sorting, gene duplication, gene loss and horizontal gene transfer
SVDQuartets	Tree inference	Inference of relationships using quartets and the coalescent model

Phylogenomics

Glossary

Convergent molecular evolution
Independent evolution of similar or identical molecular changes (for example, gene deletions, nucleotide substitutions, gene order rearrangements) in organisms from different lineages that exhibit similar adaptations.

Evolutionary radiation
The occurrence of an elevated rate of speciation events in a narrow window of evolutionary time.

Heterotachy
The phenomenon of changes in the evolutionary rate of a nucleotide or amino acid sequence through time.

Hidden orthology
Undetected orthologous relationships of genes.

Hidden paralogy
Orthologous groups of genes that contain orthologues and paralogues (inparalogues and outparalogues) stemming from asymmetric patterns of duplication and loss.

Horizontal gene transfer
Also known as lateral gene transfer. The transfer of genetic material between organisms of the same or different species through non-reproductive means.

Hybridization
The interbreeding of two distinct species or lineages.

Inparalogues
Lineage-specific or species-specific paralogues wherein the duplication event occurred after divergence from a reference common ancestor.

Introgression
The interbreeding of two distinct species or lineages, followed by backcrossing with one of the parental species.

Long-branch attraction
The inaccurate inference of taxa with high evolutionary rates (giving rise to long branches in their phylogenetic trees) as closely related.

Model of sequence evolution
Also known as the substitution model. Markov models that describe rates of nucleotide or amino acid substitutions in a locus during evolution.

Partial or incomplete taxon coverage
The lack of sequences (either because they are genuinely absent or because they were not collected) from particular taxa in a group of orthologous genes.

Phylogenetic irreproducibility
Lack of reproducibility of a tree topology between two replicate tree inferences using the same software parameters (for example, same model of sequence evolution or starting seed).

Phylogenetic networks
Graphs of evolutionary relationships that, in addition to depicting the splitting of lineages, also depict the merging of lineages (due to events such as hybridization and convergent molecular evolution or due to different gene tree topologies).

Taxon sampling
Which and how many taxa are selected for a phylogenetic analysis.

Nature Reviews Genetics

Phylogenomics

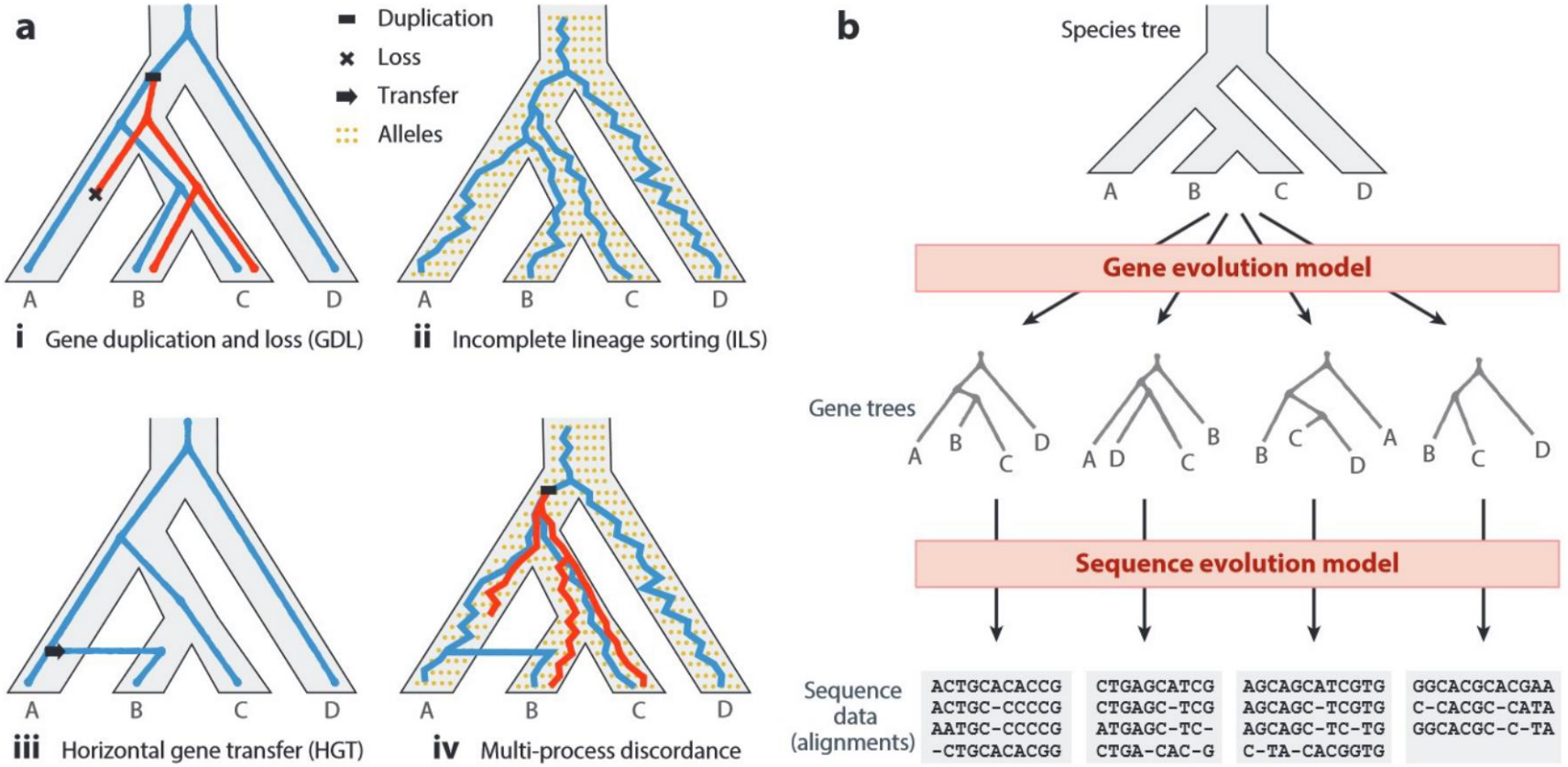


Figure 1

(*a*) A gene tree may differ from the species tree due to (*i*) gene duplication and loss, (*ii*) incomplete lineage sorting, (*iii*) horizontal gene transfer, or (*iv*) a combination of these processes. (*b*) A hierarchical model of evolution where gene trees evolve within or across the branches of a species tree according to processes illustrated in panel *a* and molecular sequences of individual loci evolve down their respective trees. A–D represent example species; two copies of a gene are shown in red and blue.

concatenation

ARTICLE

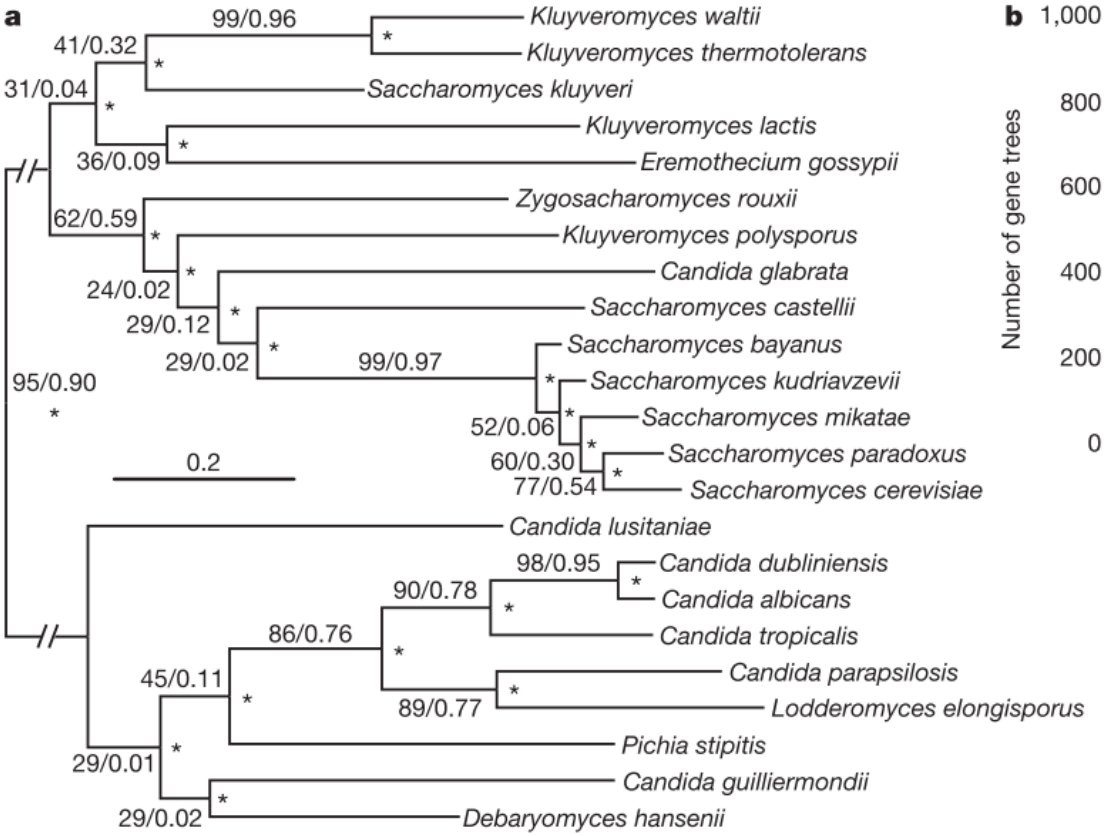
doi:10.1038/nature12130

Inferring ancient divergences requires genes with strong phylogenetic signals

Leonidas Salichos[1] & Antonis Rokas[1]

To tackle incongruence, the topological conflict between different gene trees, phylogenomic studies couple concatenation with practices such as rogue taxon removal or the use of slowly evolving genes. Phylogenomic analysis of 1,070 orthologues from 23 yeast genomes identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation. Incongruence severity increased for shorter internodes located deeper in the phylogeny. Notably, whereas most practices had little or negative impact on the yeast phylogeny, the use of genes or internodes with high average internode support significantly improved the robustness of inference. We obtained similar results in analyses of vertebrate and metazoan phylogenomic data sets. These results question the exclusive reliance on concatenation and associated practices, and argue that selecting genes with strong phylogenetic signals and demonstrating the absence of significant incongruence are essential for accurately reconstructing ancient divergences.

Salichos & Rokas 2013 Nature



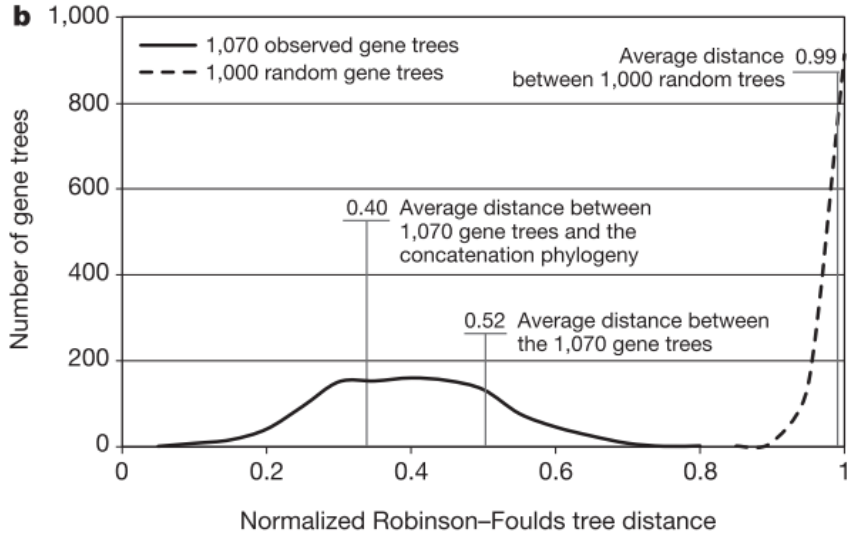


Figure 1 | The yeast species phylogeny recovered from the concatenation analysis of 1,070 genes disagrees with every gene tree, despite absolute bootstrap support. **a**, The yeast species phylogeny recovered from concatenation analysis of 1,070 genes using maximum likelihood. Asterisks denote internodes that received 100% bootstrap support by the concatenation analysis. Values near internodes correspond to gene-support frequency and internode certainty, respectively. The scale bar is in units of amino-acid substitutions per site. **b**, The distribution of the agreement between the bipartitions present in the 1,070 individual gene trees and the concatenation phylogeny, as well as the distribution of the agreement between the bipartitions present in 1,000 randomly generated trees of equal taxon number and the concatenation phylogeny, measured using the normalized Robinson–Foulds tree distance. Average distances between the 1,070 gene trees and the concatenation phylogeny, between the 1,070 gene trees themselves, and between 1,000 randomly generated gene trees that have equal taxon numbers, are also shown. The phylogeny of the 23 yeast species analysed in this study is unrooted and contains 20 non-trivial bipartitions; because the divergence of *Saccharomyces* and *Candida* lineages is well established, the mid-point rooting of the phylogeny is shown for easier visualization.

Salichos & Rokas 2013 Nature

nature ecology & evolution

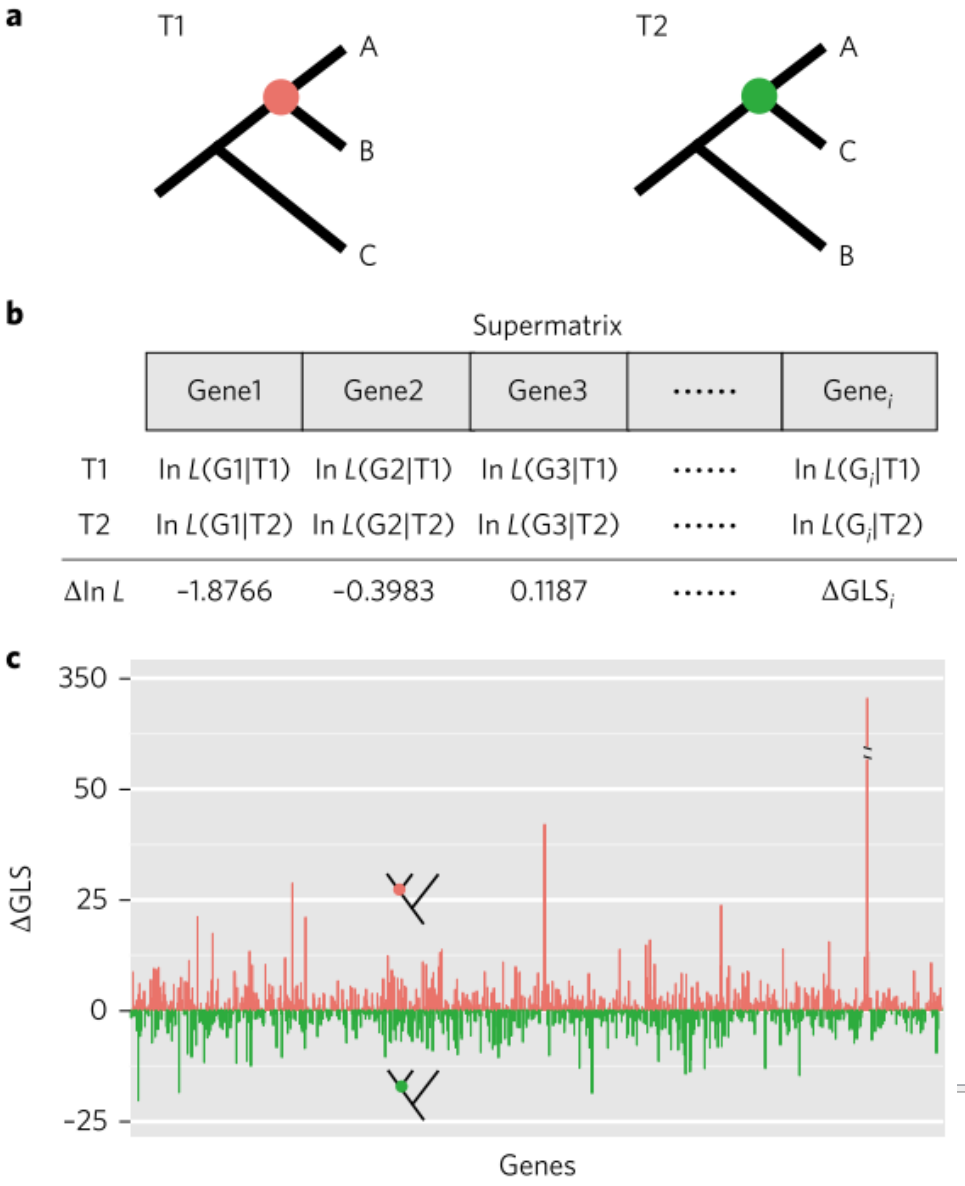
ARTICLES

PUBLISHED: 10 APRIL 2017 | VOLUME: 1 | ARTICLE NUMBER: 0126

Contentious relationships in phylogenomic studies can be driven by a handful of genes

Xing-Xing Shen[1], Chris Todd Hittinger[2] and Antonis Rokas[1*]

Phylogenomic studies have resolved countless branches of the tree of life, but remain strongly contradictory on certain, contentious relationships. Here, we use a maximum likelihood framework to quantify the distribution of phylogenetic signal among genes and sites for 17 contentious branches and 6 well-established control branches in plant, animal and fungal phylogenomic data matrices. We find that resolution in some of these 17 branches rests on a single gene or a few sites, and that removal of a single gene in concatenation analyses or a single site from every gene in coalescence-based analyses diminishes support and can alter the inferred topology. These results suggest that tiny subsets of very large data matrices drive the resolution of specific internodes, providing a dissection of the distribution of support and observed incongruence in phylogenomic analyses. We submit that quantifying the distribution of phylogenetic signal in phylogenomic data is essential for evaluating whether branches, especially contentious ones, are truly resolved. Finally, we offer one detailed example of such an evaluation for the controversy regarding the earliest-branching metazoan phylum, for which examination of the distributions of gene-wise and site-wise phylogenetic signal across eight data matrices consistently supports ctenophores as the sister group to all other metazoans.



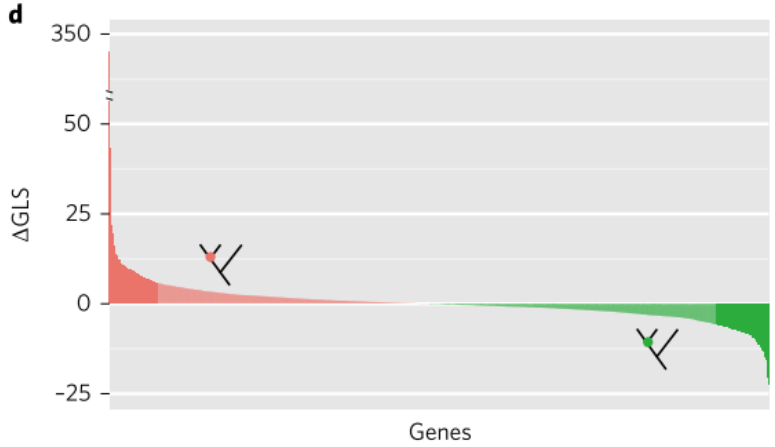


Figure 1 | A schematic representation of our approach for quantifying and visualizing phylogenetic signal in a phylogenomic data matrix. a, Two alternative phylogenetic hypotheses (T1, the unconstrained ML tree under concatenation; T2, the ML tree constrained to recover the T2 branch). **b**, Calculation of the difference in the gene-wise log-likelihood scores (ΔGLS) of T1 versus T2 for each gene in the data matrix. The difference in the site-wise log-likelihood scores, ΔSLS, of T1 versus T2 for each site in the data matrix is also calculated but is not shown here. **c**,**d**, The gene-wise phylogenetic signal (ΔGLS) for T1 versus T2 can be visualized by arranging genes either in the order of their placements in the data matrix (**c**) or in descending order of their ΔGLS values (**d**). Red bars denote genes supporting T1, whereas green bars denote genes supporting T2. The data for panels **c** and **d** are the actual values from the analysis of the Ascoideaceae branch in the fungal phylogenomic data matrix (Table 1). The schematic representation of our approach for quantifying and visualizing phylogenetic signal among three alternative phylogenetic hypotheses (T1, T2 and T3) is shown in Supplementary Fig. 1.

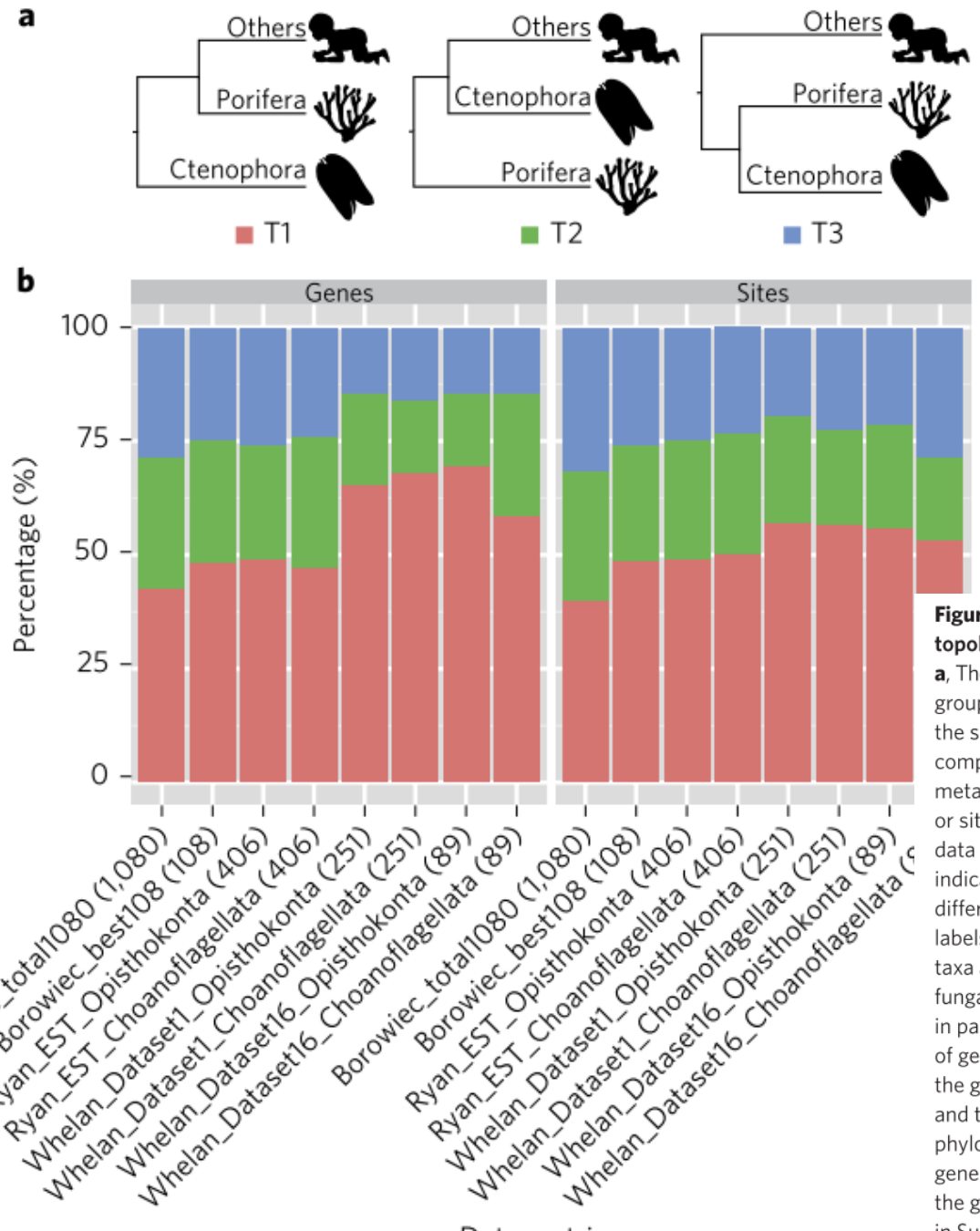


Figure 5 | The distribution of phylogenetic signal for three alternative topological hypotheses on the earliest-branching metazoan lineage. **a**, The three alternative topological hypotheses are: ctenophores as the sister group to all other metazoan phyla (Ctenophora-sister; T1), sponges as the sister group to all other metazoans (Porifera-sister; T2), or a clade composed of ctenophores and sponges as the sister group to all other metazoans (Porifera + Ctenophora-sister; T3). **b**, Proportions of genes or sites supporting each of three alternative hypotheses for each of eight data matrices from three phylogenomic studies[11,29,30] (in the matrix names indicate references: Borowiec[30]; Ryan[11]; and Whelan[29]). Note that two different non-animal outgroup sets are used in refs [11,29]: datasets whose labels include the word 'Choanoflagellata' use only choanoflagellate taxa as outgroups, whereas datasets labelled with 'Opisthokonta' use fungal, holozoan taxa, including choanoflagellates, as outgroups. Values in parentheses next to the names of data matrices indicate the number of genes present in each phylogenomic data matrix. The ΔGLS values for the genes across each data matrix are provided in Supplementary Table 9, and their distributions are shown in Supplementary Figs 62 and 63. The phylograms of all concatenation ML analyses following the removal of the gene with the highest ΔGLS value as well as those following the removal of the genes with outlier ΔGLS values in the eight data matrices can be found in Supplementary Fig. 65a–h.

Phylogenomic Subsampling

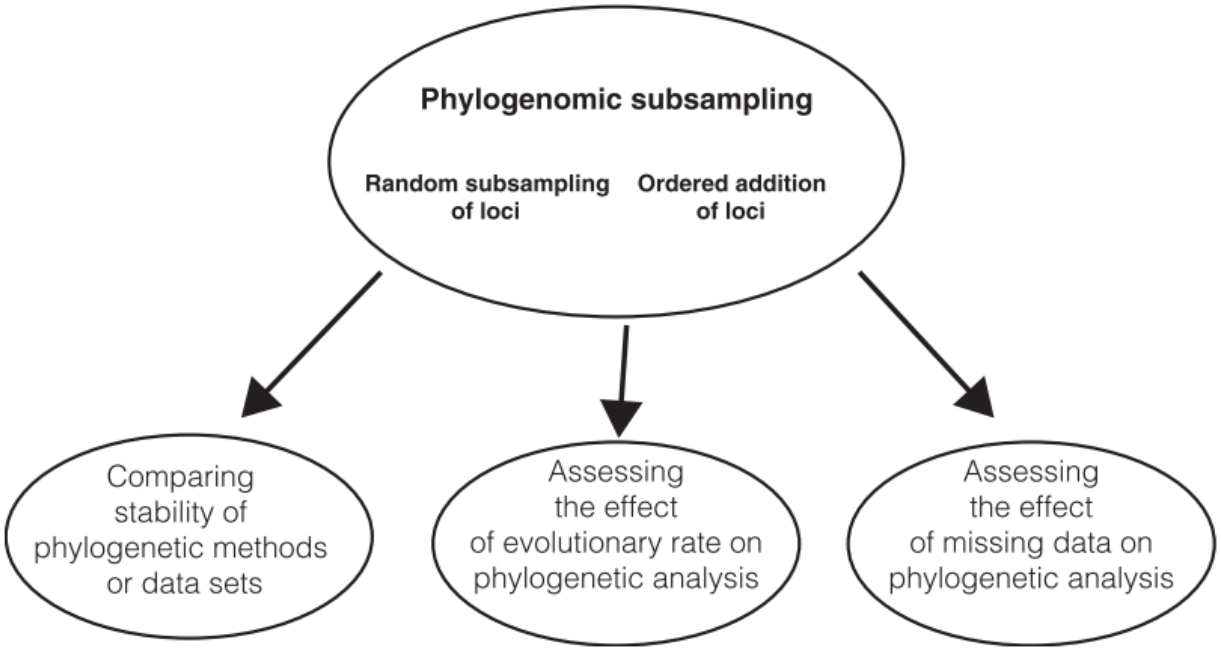


Fig. 1 Overview of uses of phylogenomic subsampling. Two types of subsampling are indicated, those which sample loci at random (the main method discussed in this paper) and those that add loci to matrices of increasing size in some ordered fashion (e.g. by increasing evolutionary rate). Phylogenomic subsampling has been used for three main purposes, as discussed in the text: to test the stability of various phylogenetic methods on matrices of different size and composition; to test the effect of differences in evolutionary rate on phylogenomic analysis; and to test the effects of missing data on phylogenomic analysis. This paper focuses on (and advocates the use of) subsampling primarily for the leftmost purpose, but acknowledges the use of ordered subsampling as well. As matrices become increasingly occupied (e.g. filled with sampled loci, as opposed to empty), the rightmost purpose will become less important.

Edwards 2016 ZooScripta

Networks

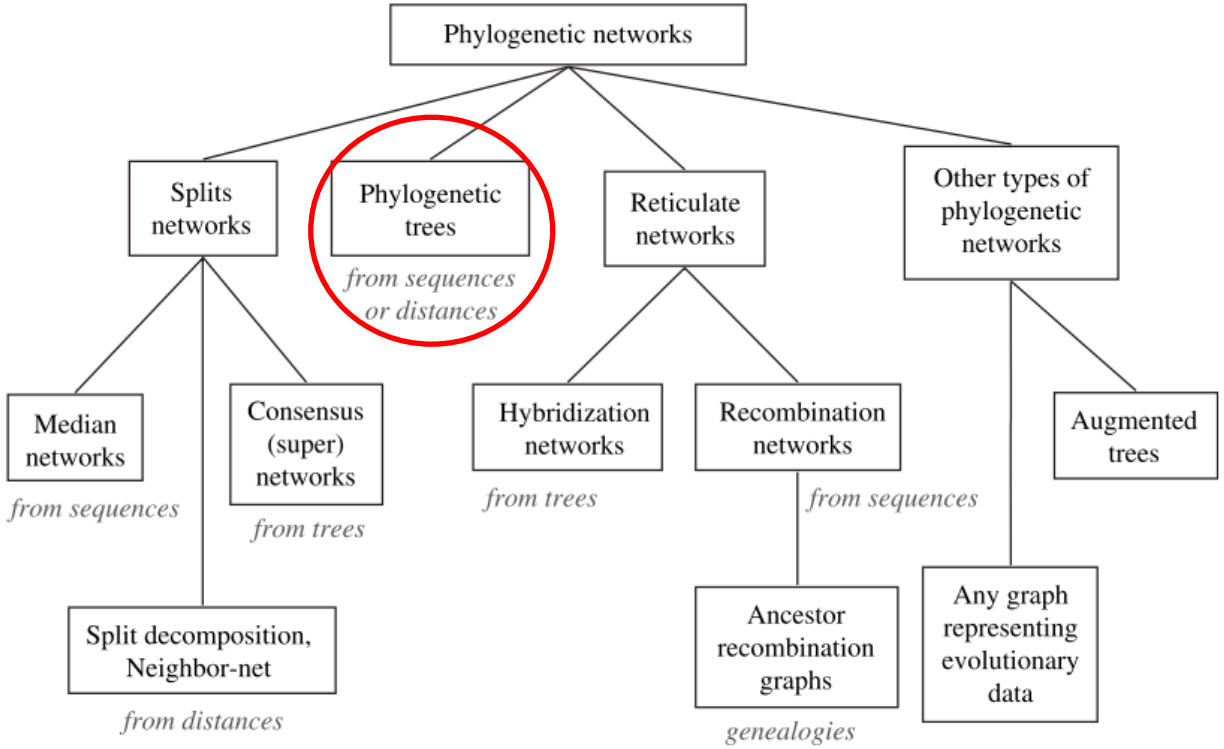


FIG. 1.—The term phylogenetic network encompasses a number of different concepts, including phylogenetic trees, split networks, reticulate networks, the latter covering both "hybridization" and "recombination" networks, and other types of networks such as "augmented trees." Recombination networks are closely related to ancestor recombination graphs used in population studies. Split networks can be obtained from character sequences, for example, as a median network, and from distances using the split decomposition or neighbor-net method or from trees as a consensus network or super-network. Augmented trees are obtained from phylogenetic trees by inserting additional edges to represent, for example, horizontal gene transfer. Other types of phylogenetic networks include host-parasite phylogenies or haplotype networks. Diagram adapted from Huson and Kloepper (2005).

Huson and Bryant 2006 MBE

Phylogenomics

Research

Open Access

Analysis of 142 genes resolves the rapid diversification of the rice genus

Xin-Hui Zou[¤*], Fu-Min Zhang[¤*], Jian-Guo Zhang[¤†], Li-Li Zang[*], Liang Tang[*], Jun Wang[†], Tao Sang[‡] and Song Ge[*§]

Addresses: [*]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China. [†]Beijing Genomics Institute, Beijing, 101300, China. [‡]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA. [§]The Graduate School, Chinese Academy of Sciences, Beijing, 100039, China.

[¤] These authors contributed equally to this work.

Correspondence: Song Ge. Email: gesong@ibcas.ac.cn

Published: 3 March 2008

Genome Biology 2008, 9:R49 (doi:10.1186/gb-2008-9-3-r49)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/3/R49

Received: 21 December 2007
Revised: 18 February 2008
Accepted: 3 March 2008

Phylogenomics

Table 1

Information on the materials used in this study

Species	Genome	Accession number*	Origin	No. of genes sequenced	No. of sites aligned
Oryza sativa†	A	93-11	China	62	52,092
O. rufipogon	A	105480	India	142	124,079
O. barthii	A	104132	Cameroon	62	52,092
O. punctata	B	103903	Tanzania	141	124,079
O. officinalis	C	104972	China	142	124,079
O. rhizomatis†	C	103410	Sri Lanka	62	52,092
O. eichingeri†	C	105415	Sri Lanka	62	52,092
O. australiensis	E	105263, 101410	Australia	135	124,079
O. brachyantha	F	105151	Sierra Leone	124	124,079
O. granulata	G	M8-15, 106469	China, Vietnam	124	124,079
Leersia tisserantii	-	105610	Cameroon	122	124,079

*All accession numbers were obtained from the International Rice Research Institute at Los Banos, Philippines, except for M8-15, which was collected by the authors. †Sixty-two genes were sequenced for these species and used only for testing the effect of dense sampling. Sequences of *O. sativa* (*93-11*) were retrieved from the BGI-RIS database.

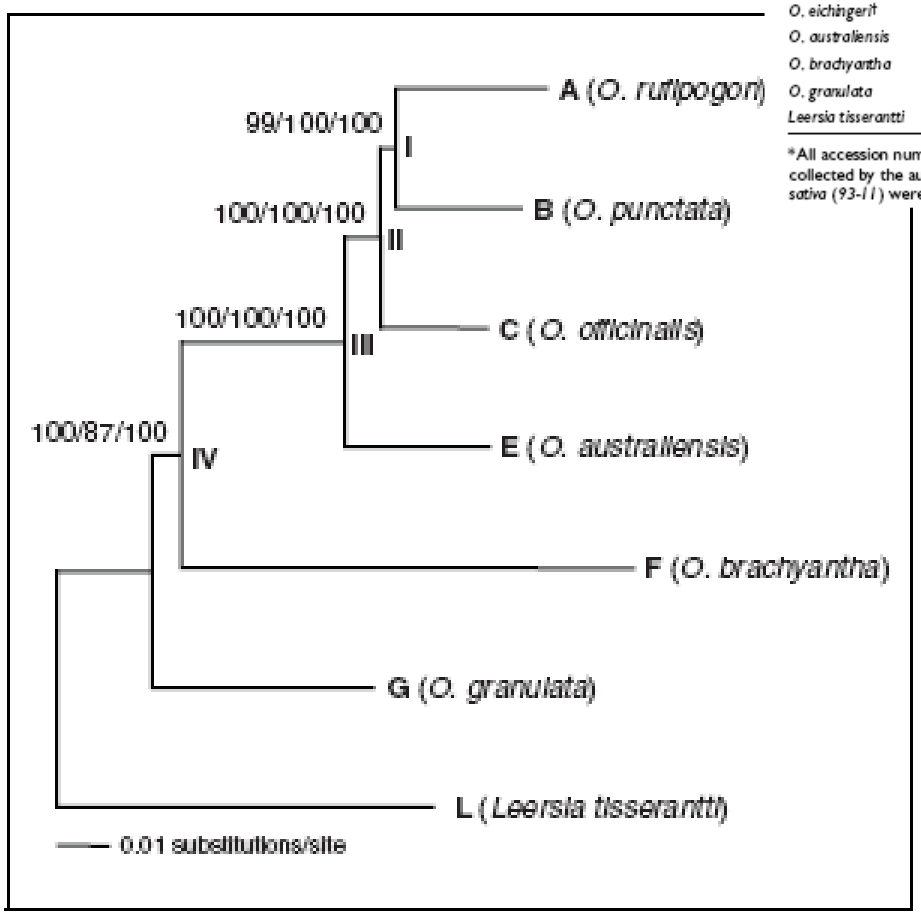


Figure 1
ML tree inferred from the concatenated sequences of 142 genes using the GTR+Γ model. The same topology was obtained from MP and BI. The letters A, B, C, E, F, and G represent all recognized diploid genome types of *Oryza*, and L represents the outgroup. The names of the species that represent the genome types and outgroup are in parentheses. Numbers above branches indicate bootstrap support of ML and MP, and posterior probability of BI, respectively. Four internal branches of *Oryza* genome types are indicated with I, II, III, and IV. Branch length is proportional to the number of substitutions measured by the scale bar.

Phylogenomics

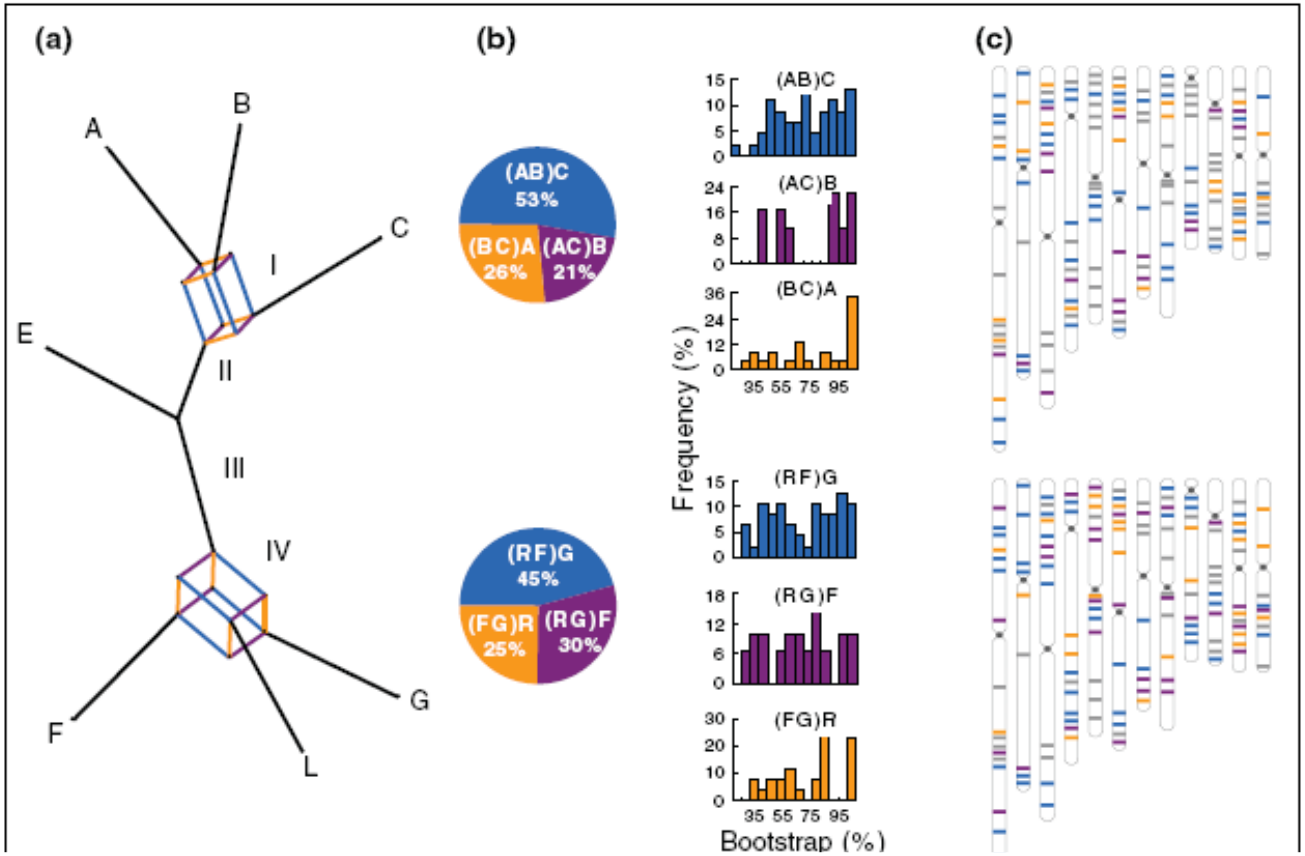


Figure 3

Genome-wide incongruence. A, B, C, E, F, and G represent *Oryza* genome types and L represents the outgroup, *Leersia*. (a) Consensus network constructed from ML trees at a threshold of 0.15. The two boxes indicate the relatively high levels of incongruence among gene trees associated with internal branches I and IV. Branch length is proportional to the frequency of occurrence of a particular split of all gene trees. R represents the rest of the genome types, including A-, B-, C-, and E-genomes. Color schemes: for the box associated with branch I, blue, orange, and purple illustrate splits supporting alternative topologies, (AB)C, (BC)A, and (AC)B, respectively; for the box associated with branch IV, blue, orange, and purple illustrate splits supporting alternative topologies, (RF)G, (FG)R, and (RG)F, respectively. (b) Pie graphs indicate the proportions of gene trees that support alternative splits in the corresponding boxes at the left. Histograms at the right illustrate the distribution of ML bootstrap support for the corresponding split (in the corresponding colors). (c) Illustration of the relative physical locations of the 142 sampled genes on the 12 rice chromosomes based on rice genome sequences. The colors indicate genes supporting a split or topology coded in the same color in the corresponding boxes on the consensus network. Genes coded in gray are those that had no input in the topology illustrated in the pie graphs and those not included for the construction of the consensus network because of missing data.

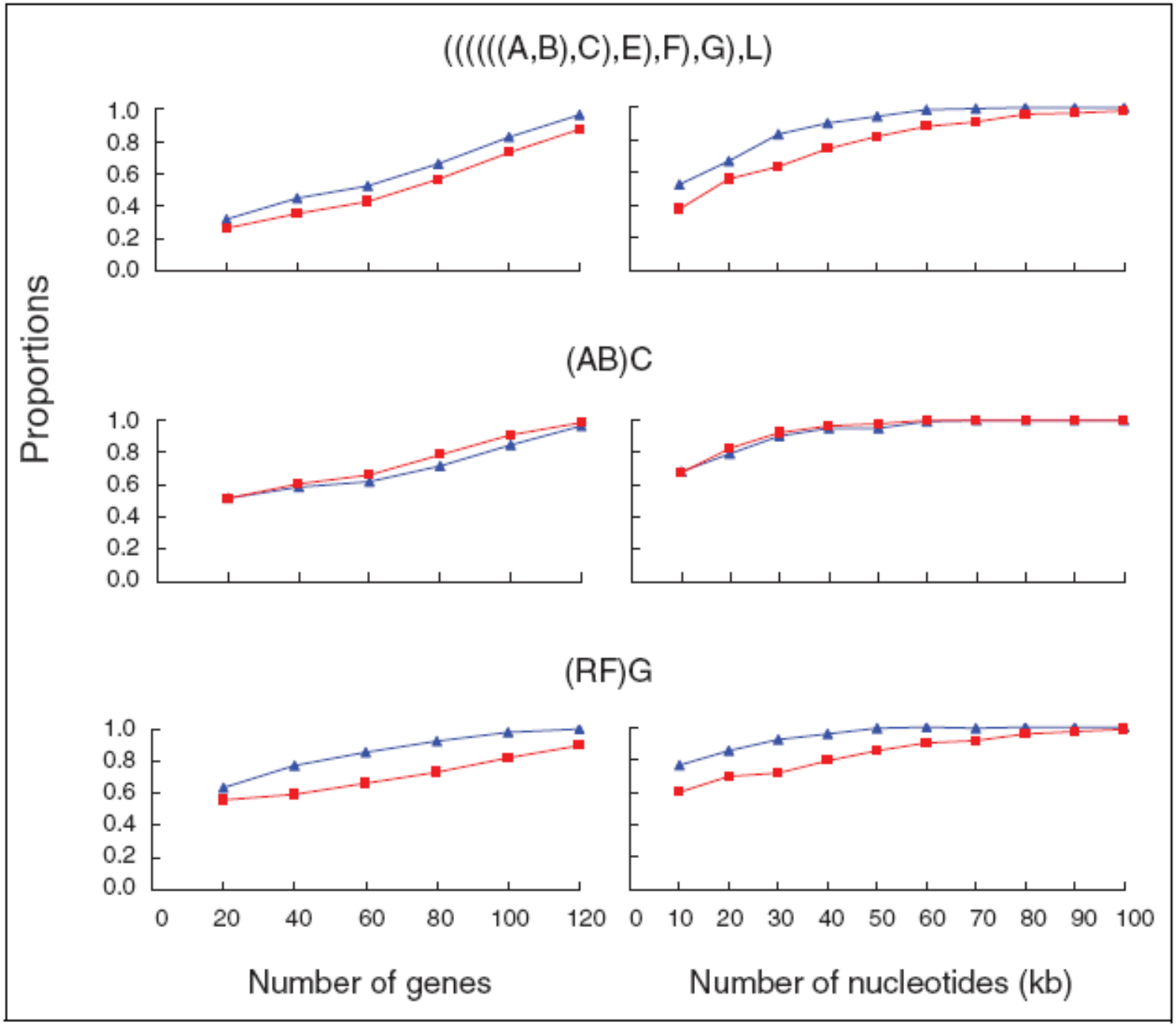
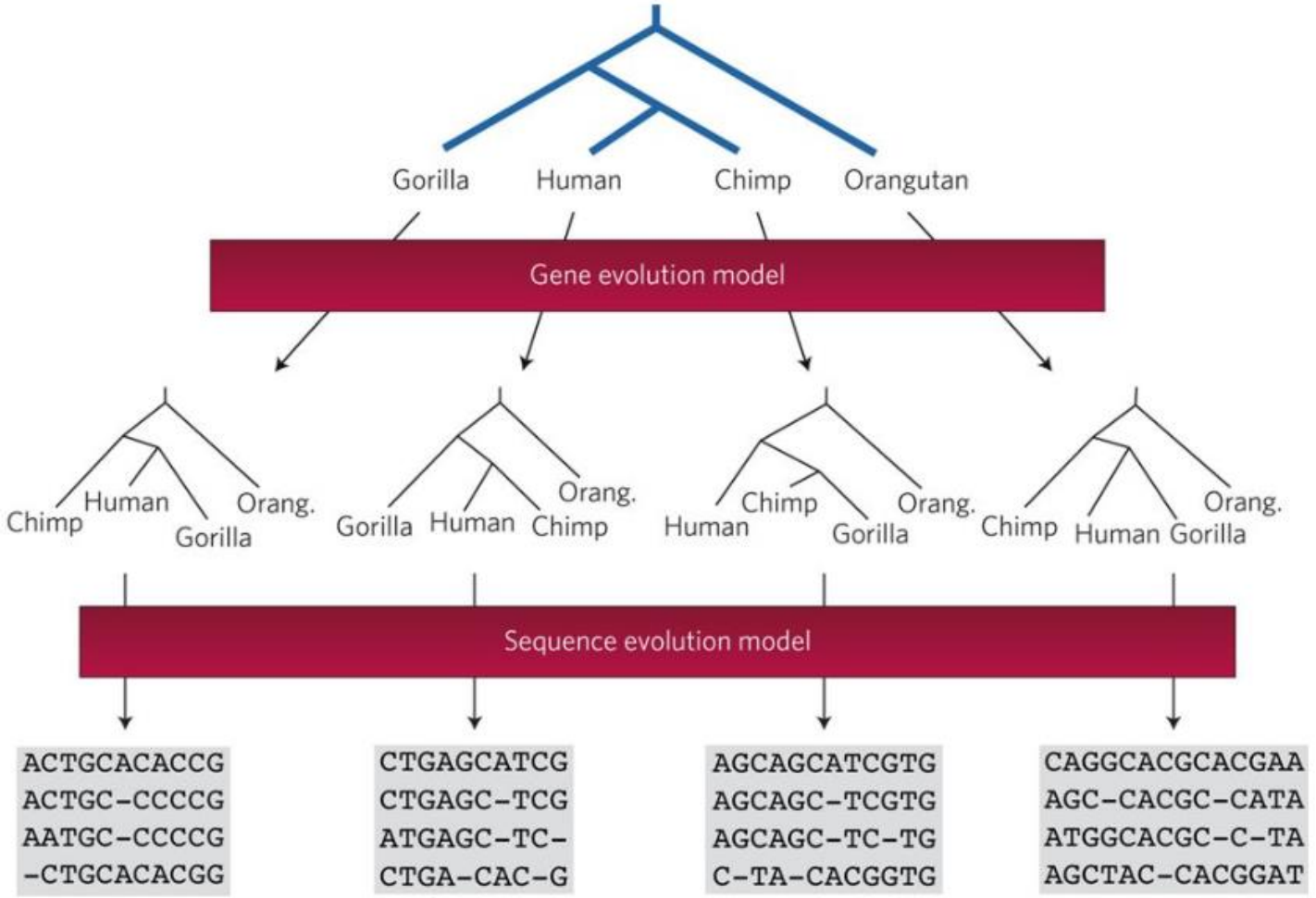


Figure 4
The proportions of topologies (or clades) that are identical to those shown in Figure 1 based on resampling of 142 gene sequences at various scales. Results of ML and MP analyses are indicated by blue and red, respectively. Genome types are represented with the same capital letters as in Figure 3.

Conclusions

For soft polytomies, an obviously interesting question is how many DNA sequences would be needed to resolve rapid speciation considering that DNA sequences have been, and will remain, major sources of biological data. The mosaic genome or different evolutionary histories of genes under rapid speciation, in conjunction with other factors associated with species divergence (for example, selection and high homoplasy of ancient speciation, brings about difficulties in resolving speciation events when using a small number of regions/genes or limited characters. This study shows that as many as 120 genes with an average length of 874 bp or 50 kb of randomly sampled nucleotides from 142 genes are needed to resolve clades I and IV simultaneously with over 95% confidence (Figure 4). Clearly, blocks of contiguous nucleotide sites were less powerful in phylogenetic resolution than samples consisting of sites drawn randomly from the genome because nucleotides within genes do not evolve independently. This implies that for the same amount of sequence data, a larger number of unlinked shorter DNA fragments are preferred over a smaller number of larger fragments for resolving short branches.

Genes trees- species tree



Schematics of a generative model adopted by the community. The species tree (top) generates gene trees (middle) using a model of gene tree evolution (for example, multi-species coalescence[9] or birth–death models of gene birth), and then each gene separately generates sequence data (bottom) using models of sequence evolution. Inference of the species tree starts from the data and follows the opposite directions of the generative model, either in two stages (summary methods), all at once (co-estimation), or skipping the middle layer (site-based methods).

Mirarab 2017 NatEcolEvo

Species trees

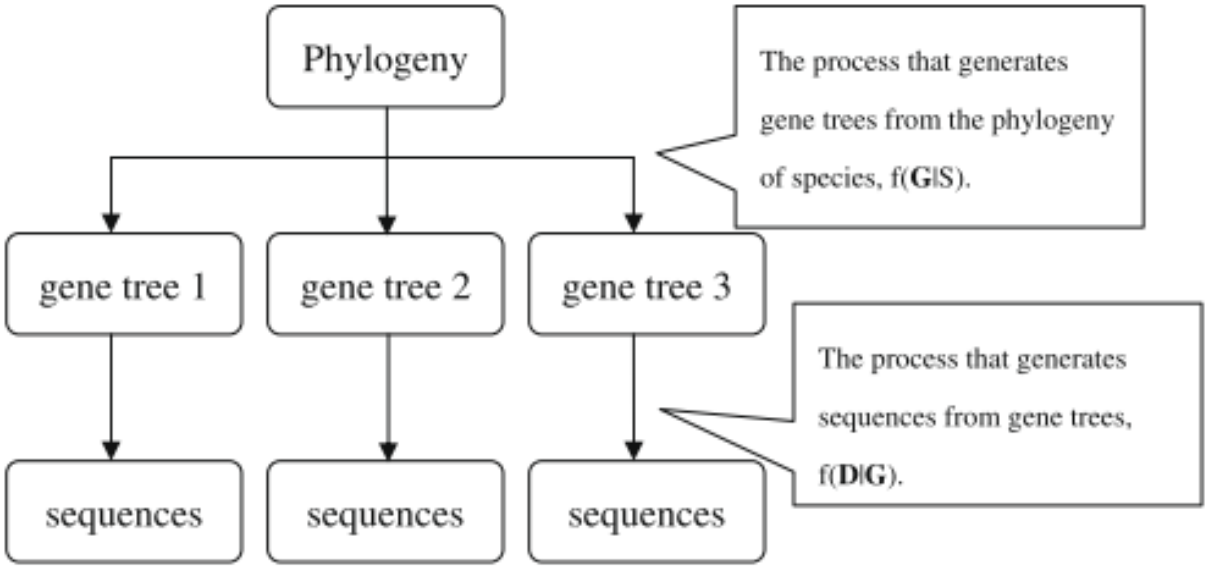


Fig. 1. Flow chart for demonstrating the statistical model for multilocus sequences generated from the phylogeny of species. The chart illustrates the independence of the two major stochastic processes in generating molecular data from species trees: the generation of gene trees from the species tree and the generation of DNA sequences from the constituent gene trees.

Liu *et al.* 2009 MPE

Phylogenomics

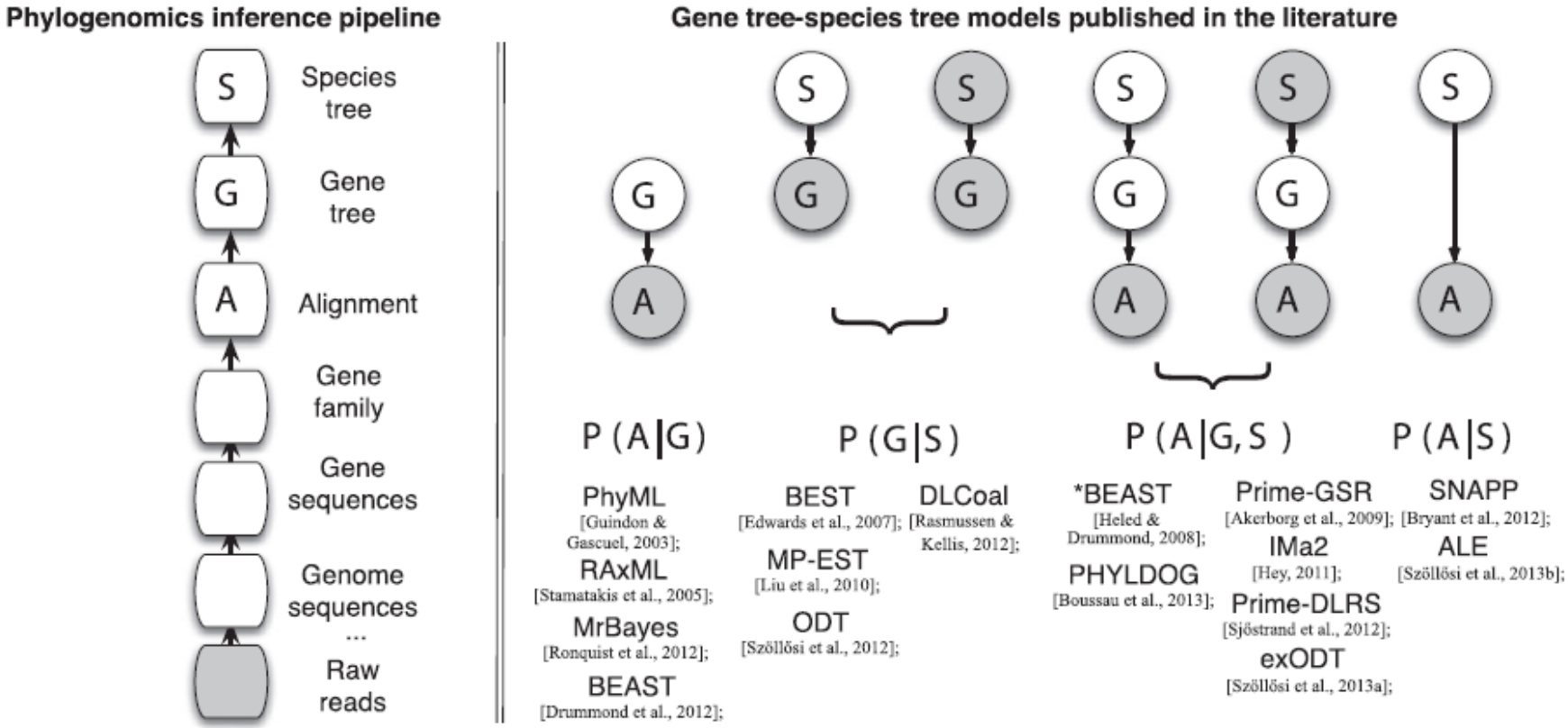


FIGURE 5. Gene tree–species tree models in the context of the phylogenomics inference pipeline. Left: the inference pipeline (some steps are not represented, such as sequencing error correction). Right: graphical representation of the inferential problem for a selection of the models and associated phylogenetic software discussed in the main text. The sequence of steps in the graphical model representations correspond to the hierarchical sequence of evolutionary process generating genomic sequences (cf. Fig. 1). The likelihood that must be computed is also shown. Graphical model conventions are observed: stochastic nodes, nodes corresponding to data considered as known are gray, and nodes whose states are inferred are in white. The models have been simplified, and parameters others than the gene tree and the species tree have not been represented.

Szllozi *et al.* 2015 SysBiol

Phylogenomics

- Summary methods, which operate by first estimating gene trees (one for each locus) and then using the information in these gene trees to estimate the species trees. The most well-known such method is ASTRAL (Mirarab et al. 2014b).
- Site-based methods, which calculate small trees (typically unrooted quartet trees or rooted triplet trees) from the site patterns and then combine these small trees into a tree on the full data set. The most well-known such method is SVDquartets (Chifman & Kubatko 2014), which is available through PAUP* (Swofford 2002).
- Coestimation methods, which coestimate the species tree and the set of gene trees. The most well-known such methods are StarBEAST (Heled & Drummond 2010) and its improved version, StarBEAST2 (Ogilvie et al. 2017).

multispecies coalescent model MSC

The multispecies coalescent model is preferred to the 'super-matrix' method for phylogenetic inference when population sizes are large relative to the ages of the species being considered, because considerable differences are expected between individual gene trees and the species tree they evolve within

Heled *et al.* 2013 BMC EvolBio

The multispecies coalescent (MSC) model is a relatively new and arguably successful approach to phylogenomics in which individual gene trees are estimated simultaneously or separately with a species tree as a means of estimating phylogenetic relationships.

Edwards *et al.* 2016 MPE

Total evidence vs species trees

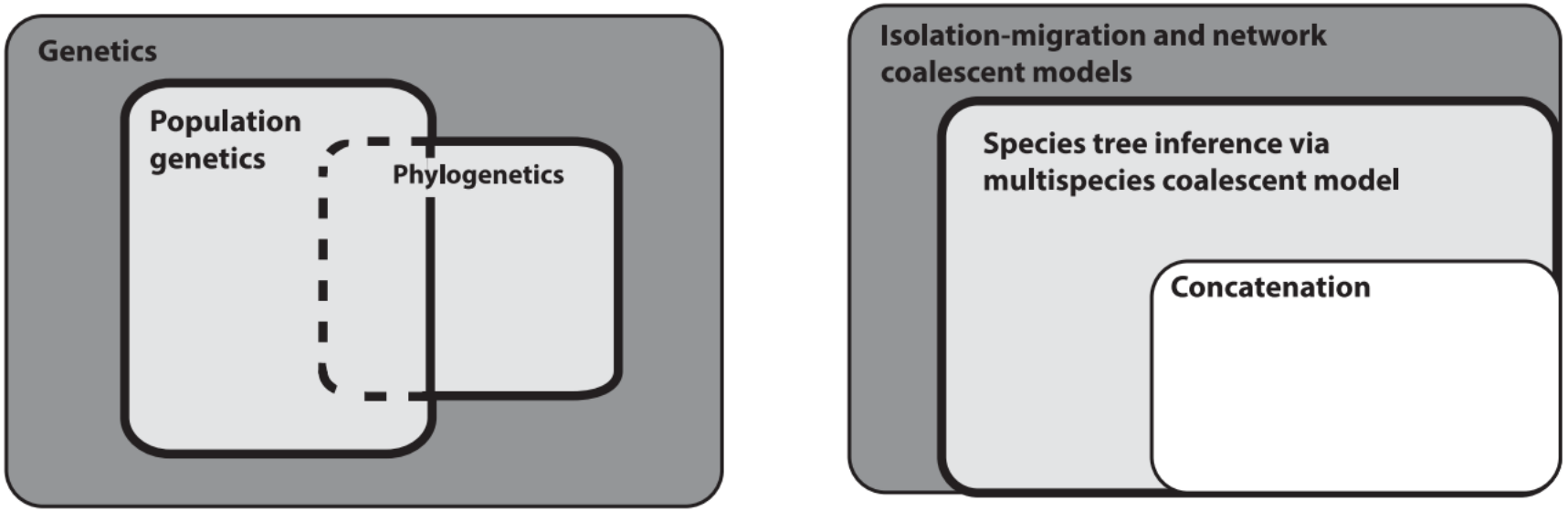
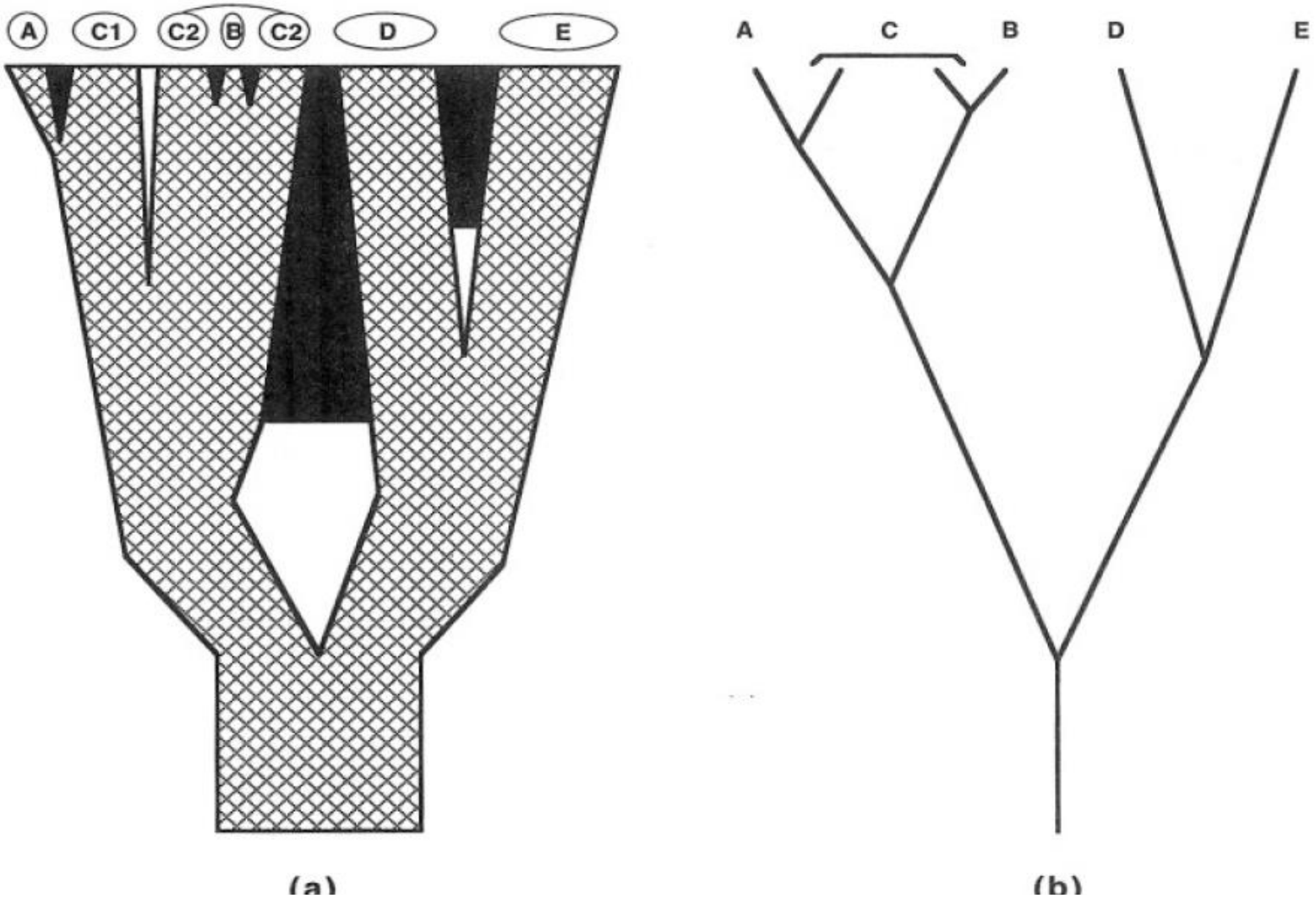


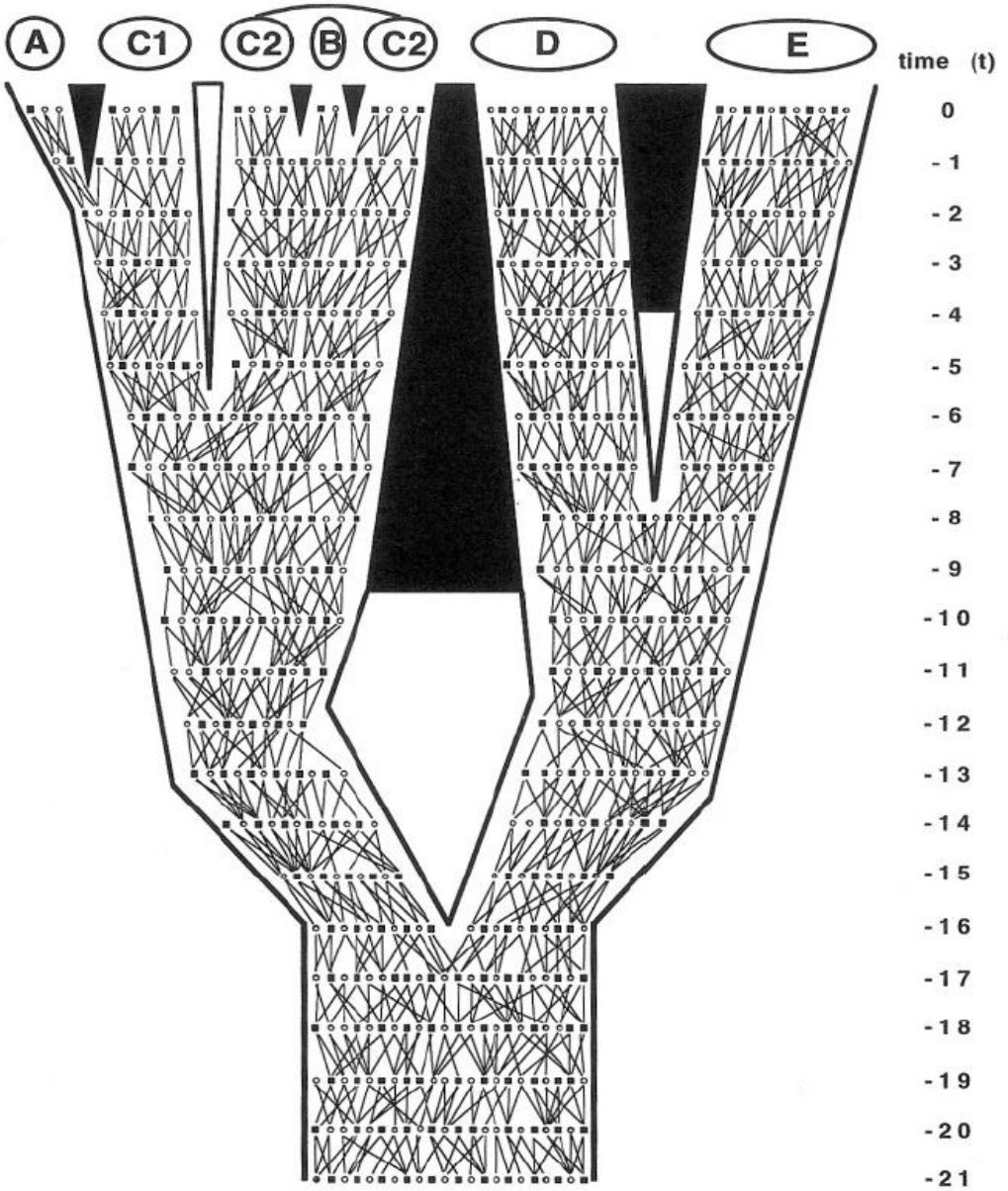
Fig. 1. Hierarchy of domains and models in the fields of genetics and phylogenetics. Phylogenetics is viewed as a particular instance of the broader field of genetics in which phylogenetics is nested. (Top) A view of the relationship among the domains of genetics, phylogenetics and population genetics encourages approaches to phylogenetics that are consistent with the major tenets of genetics, such as the chromosomal structure of genomes, random assortment of alleles during meiosis, independent transmission of alleles at unlinked loci and other mainstays of genetics. Most sampling schemes for phylogenetics (sampling multiple alleles per species, or single alleles from multiple loci from multiple species) demand consideration of population genetic principles. (Bottom) Relationship among models in phylogenetics, including the MSC and concatenation models. Concatenation is best viewed as a particular case of the broader model inherent in the MSC, which is itself a particular case of models that incorporate gene flow and other reticulations, such as recently proposed MSC network models.

Edwards *et al.* 2016 MPE

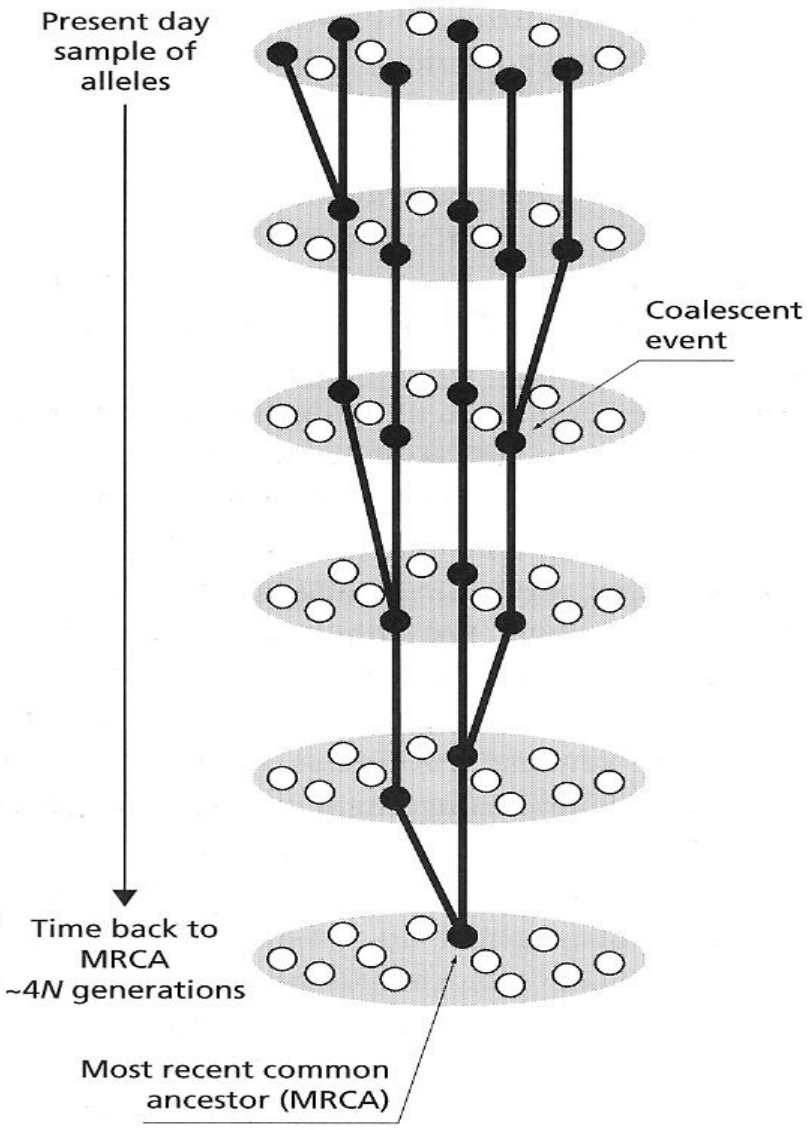
Trees and lienages



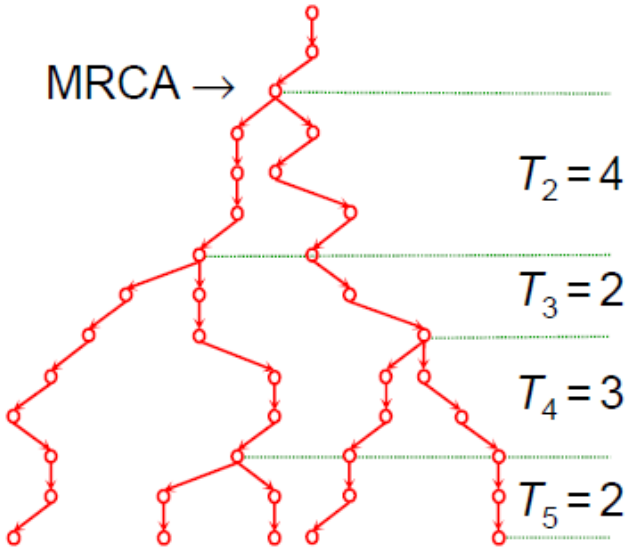
Lineages



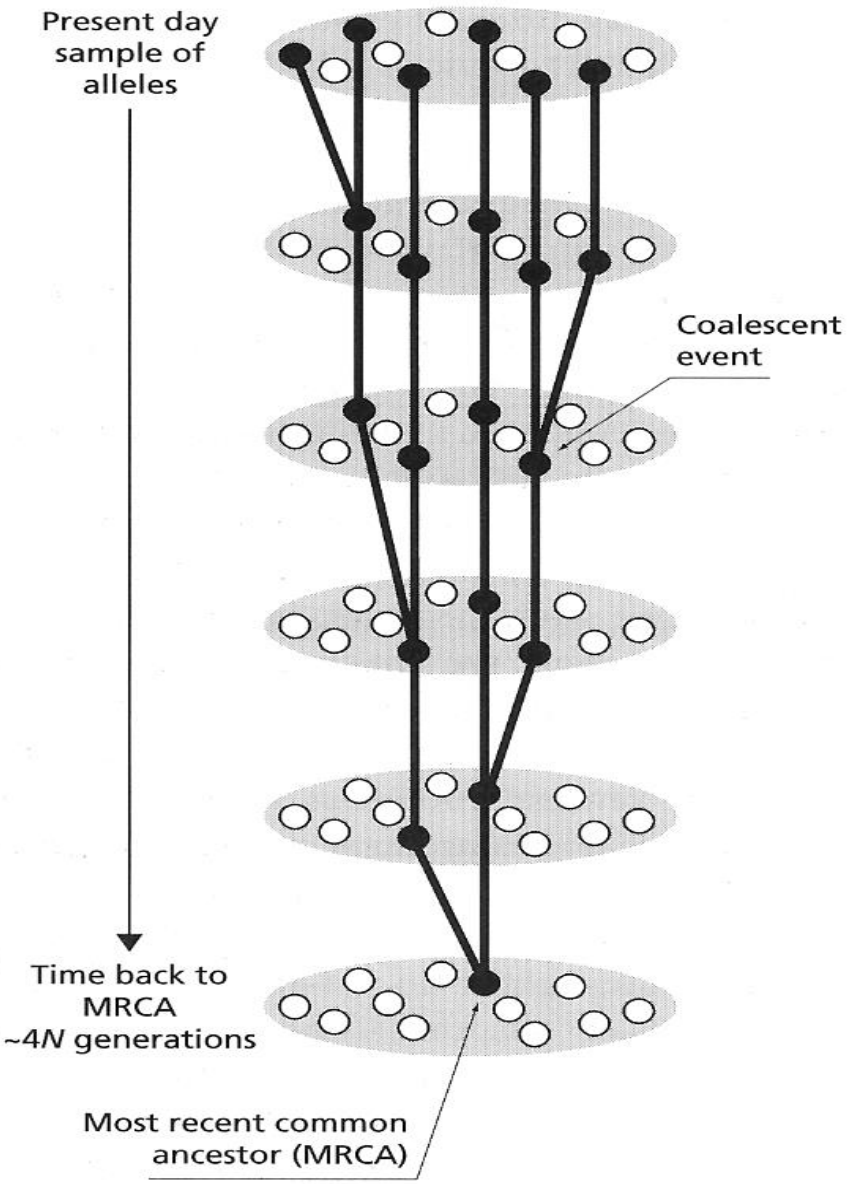
Coalescence



(**b**) Coalescent process
The process of lineage joining when one traces the genealogical history of the sample backwards in time.



Coalescence



The probability that two alleles share and not share an ancestor 1 generations ago is:

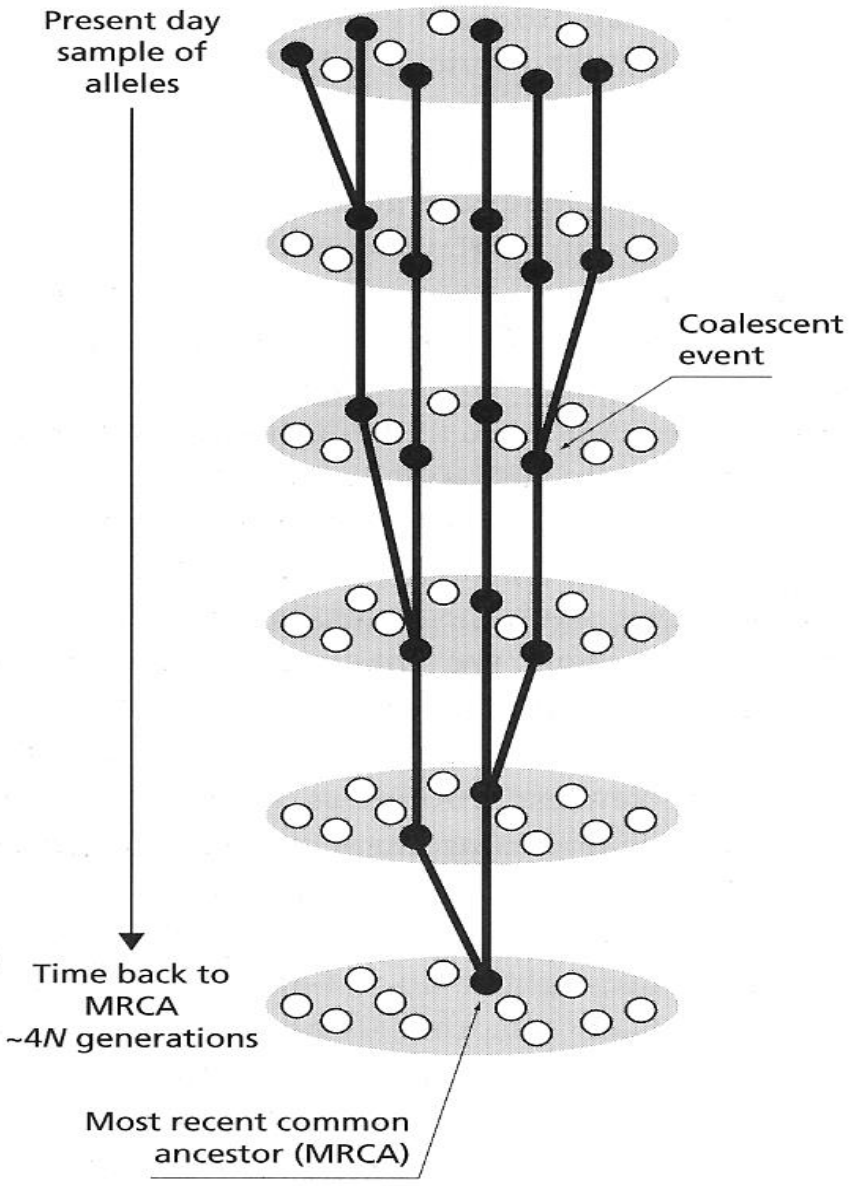
$$1/N$$

$$1-1/N$$

The probability that two alleles share and not share an ancestor 2 generations ago is:

$$1/N*(1-1/N)$$

Coalescence



The probability that two alleles share an ancestor G generations ago is:

$$f(G) = (1/2N)(1-1/2N)^{G-1}$$

$$\cong$$

$$f(G) = (1/2N)e^{-(G-1)/2N}$$

The coalescent: 2 genes

The probability that two genes find a common ancestor in the first generation back is $1/(2N)$. The probability that two genes find a common ancestor j generations back is

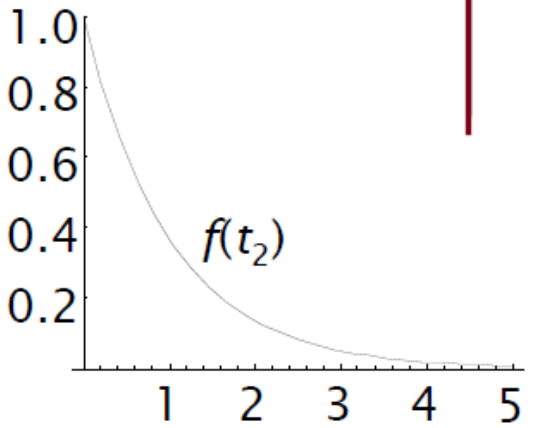
$$\Pr\{T_2 = j\} = \left(1 - \frac{1}{2N}\right)^{j-1} \times \frac{1}{2N}$$

$t_2 = T_2/(2N)$

- It takes on average $2N$ generations for two genes to coalesce.

Let $t_2 = T_2/(2N)$ so that one time unit is $2N$ generations,

$$f(t_2) = e^{-t_2}$$



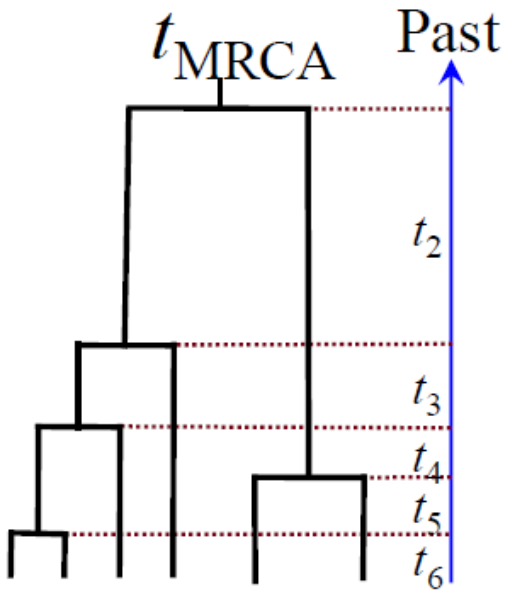
Yang 2014

The coalescent: n sequences

The waiting times t_n, t_{n-1}, …, t_2 are independent exponential variables, with

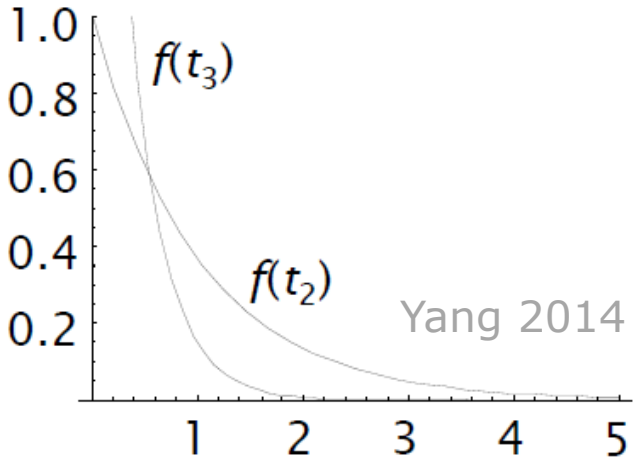
$$E(t_j) = \frac{1}{j(j-1)/2}.$$

$$E(t_{\text{MRCA}}) = 2(1 - 1/n).$$

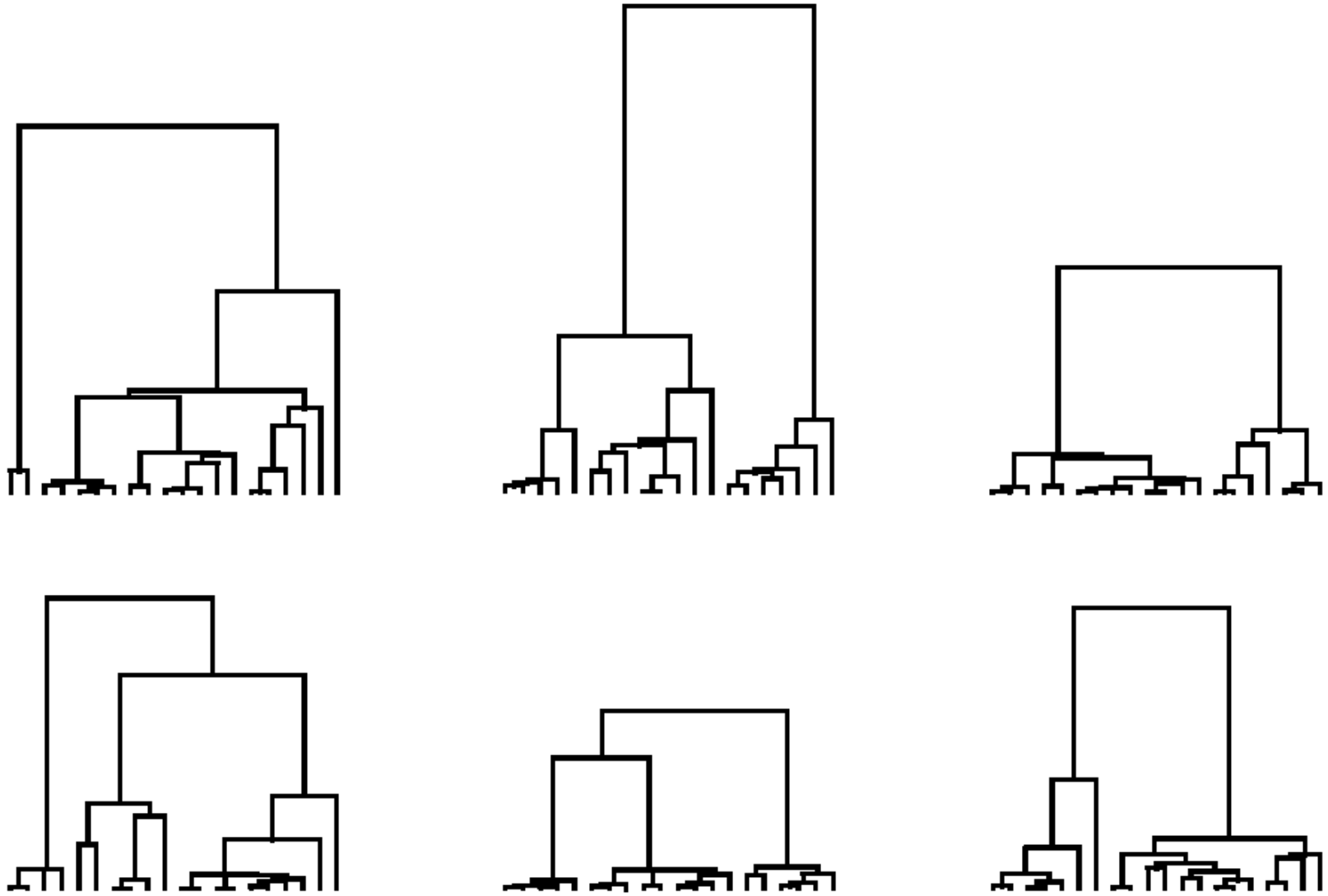


It takes on average $\sim 4N$ ($\pm 2.15N$) generations for the whole sample to coalesce.

It takes on average $2N$ generations for the last two lineages to coalesce.



The coalescent n=20

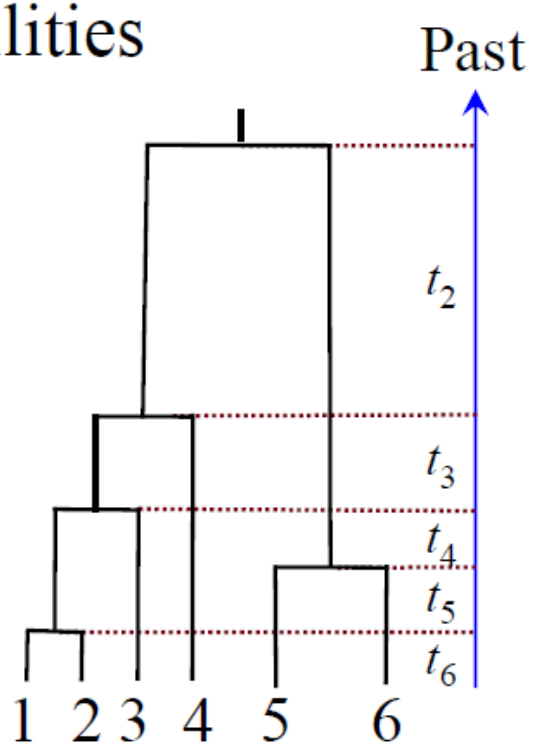


Yang 2014

Coalescence

- The coalescent model provides a probabilistic description of the genealogical tree and the coalescent times for the sample.
- The mutation model specifies the probabilities of changes over time.

```
AF310299    TCCATTCAAG AGTCTATTAT CAGTTTTTTC …
AF310267    TCCATTCAAG AGTCTATTAT CAGTTTTTTC …
AF310294    TCCATTCAAG AGTTTATTAT CAGTTTTTTC …
AF310277    TCCATTCAAG AGTCTATTAT CAGTTTTTTC …
AF310266    TCCATTCAAG AGTCTATTAT CAGTTTTTTC …
AF310322    TCCATTCAAG AGTCTATTAT CAATTTTTTC …
```



Theta

$\theta = 4N\mu$ measures the genetic variation in a population, where μ is the mutation rate.

θ is the expected number of differences per site between two randomly-drawn sequences.

$\theta_H = 0.0006$ in humans: two sequences taken at random from the human population are different at 0.06% of sites. This translates to $N \sim 10{,}000$ (using $g = 15$y, $\mu = 10^{-9}$/site/year).

Average coalescent time is $2N$ generations.

Average sequence distance is $\theta = 2N \times \mu \times 2$.

Yang 2014

Data and Model

Data: $X = \{X_i\}$, X_i is alignment of n_i sequences at locus i.
Parameter: $\theta = 4N\mu$.
G_i: gene tree topology at locus i, unobserved.
$\boldsymbol{t}_i$: $(n_i - 1)$ coalescent times in the tree, unobserved.

$f(G_i, \boldsymbol{t}_i|\theta)$ is given by the coalescent model.
$f(X_i|G_i, \boldsymbol{t}_i)$ is Felsenstein's phylogenetic likelihood.

Yang 2014

Estimation of the parameter $\theta=4N\mu$

Maximum likelihood.

Each locus i has an unobserved gene tree G_i with coalescent times $\boldsymbol{t}_i$. The likelihood averages over G_i and $\boldsymbol{t}_i$ at each locus.

$$\ell(\theta) = \sum_i \log f(X_i \mid \theta) = \sum_i \log\left\{\sum_{G_i} \int f(G_i, \mathbf{t}_i \mid \theta) f(X_i \mid G_i, \mathbf{t}_i)\, \mathrm{d}\mathbf{t}_i\right\}$$

Bayesian.

The averaging over G_i and $\boldsymbol{t}_i$ is through Monte Carlo Markov chain (MCMC).

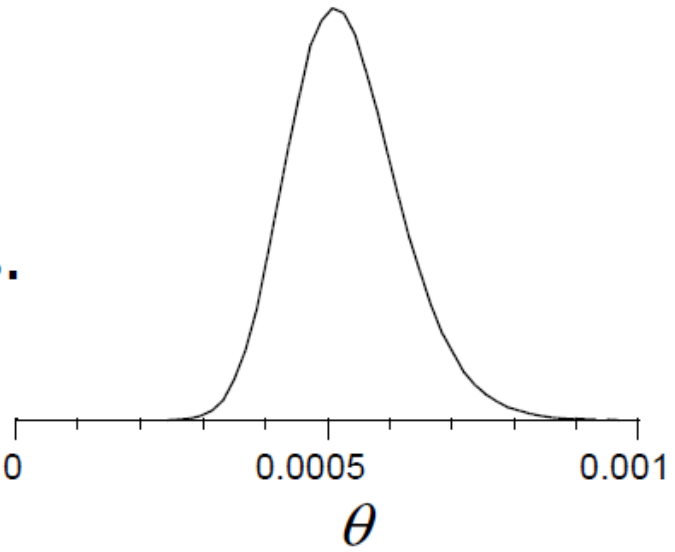
$$f(\theta, \{G_i, \mathbf{t}_i\} \mid X) \propto f(\theta) \prod_i f(G_i, \mathbf{t}_i \mid \theta) f(X_i \mid G_i, \mathbf{t}_i)$$

Yang 2014

MCMC samples from the posterior

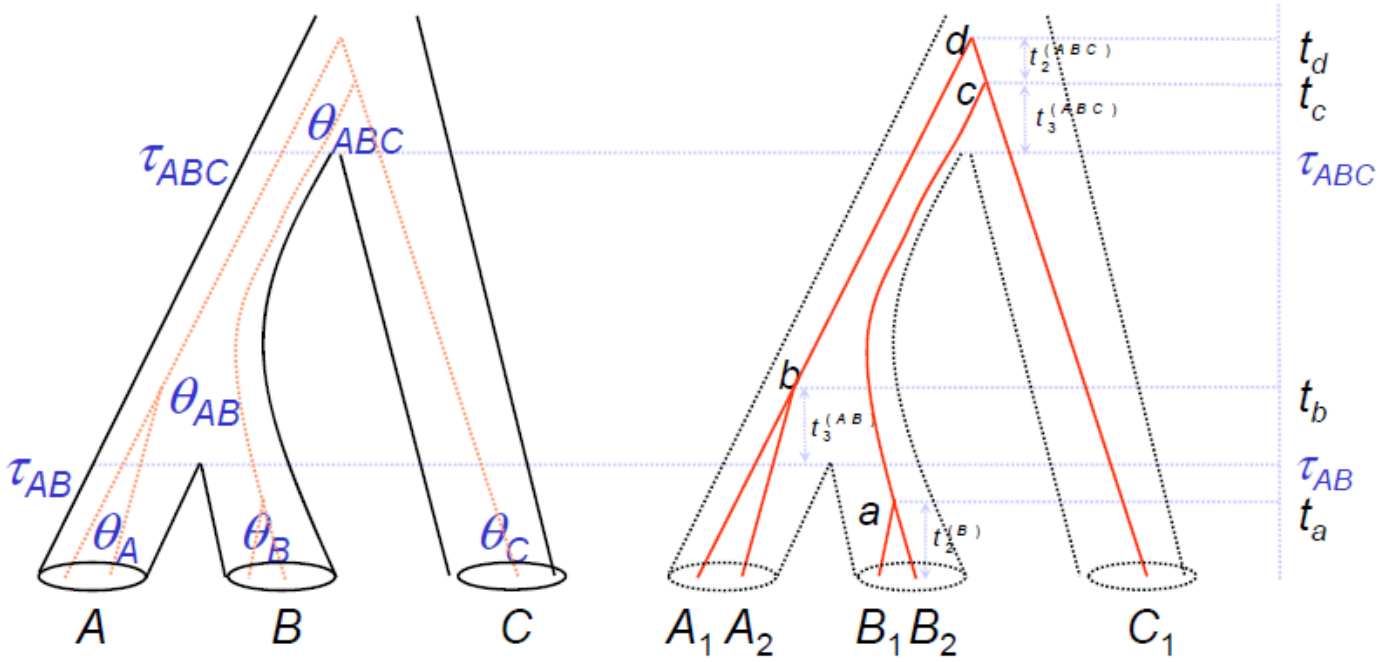
$f(\theta, \{G_i, \boldsymbol{t}_i\}|X)$

1. Initialize θ. Generate G_i and $\boldsymbol{t}_i$ from the coalescent.
2. Loop
   - Change parameter θ.
   - Change gene tree topology $\{G_i\}$.
   - Change coalescent times $\{\boldsymbol{t}_i\}$.
   - Take a sample every k iterations.



Multispecies coalescent model

- There are two sets of parameters: τs, θs.
- Lineages join independently in different populations.
- Coalescent rate is reset when lineages enter a new species.
- Genes split before species.



Rannala & Yang (2003 *Genetics* 164:1645-1656)

Multispecies coalescent model – thre species

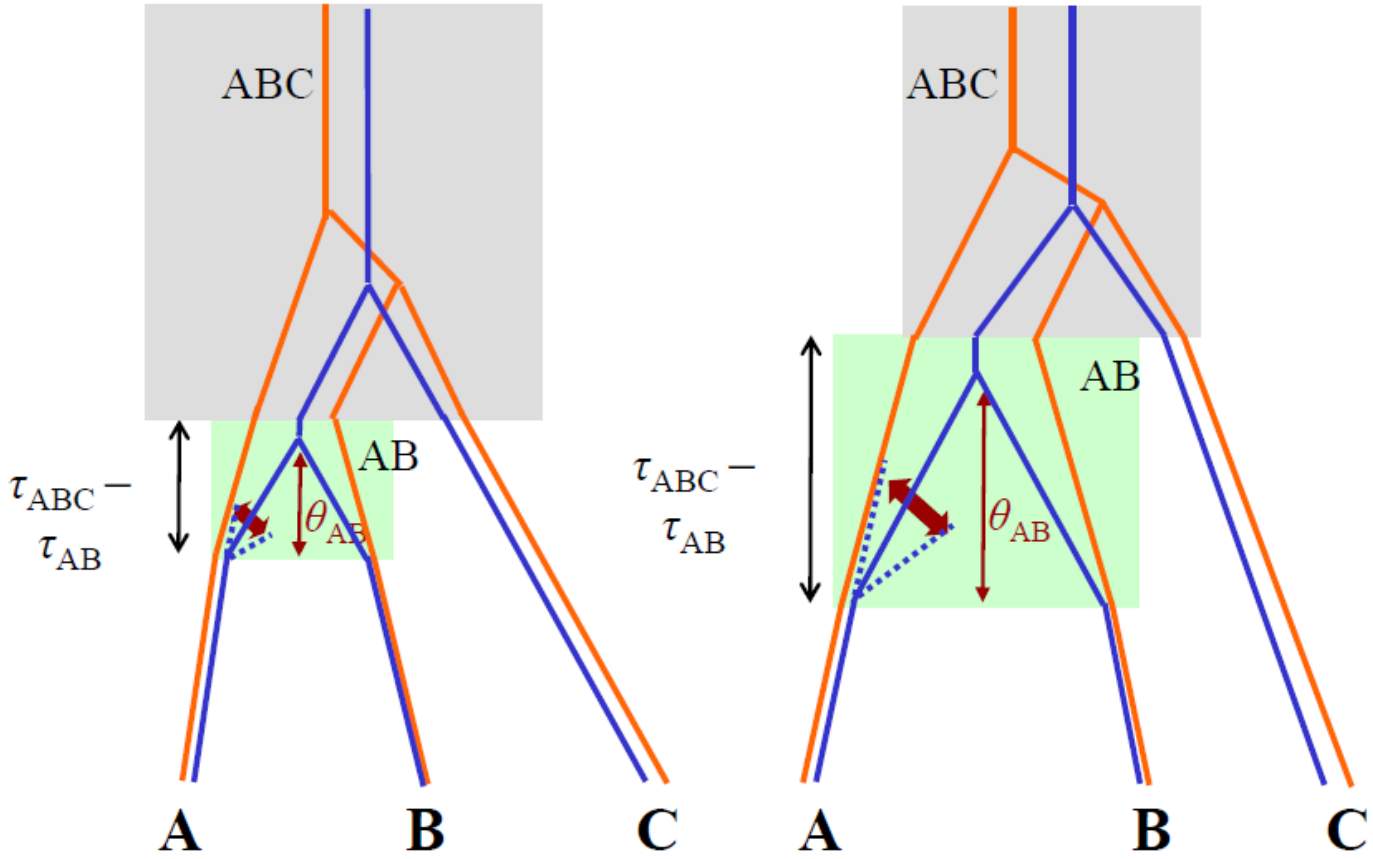
Parameters:

Speciation times:

τ_{AB}, τ_{ABC}

Population sizes:

θ_{AB}, θ_{ABC}

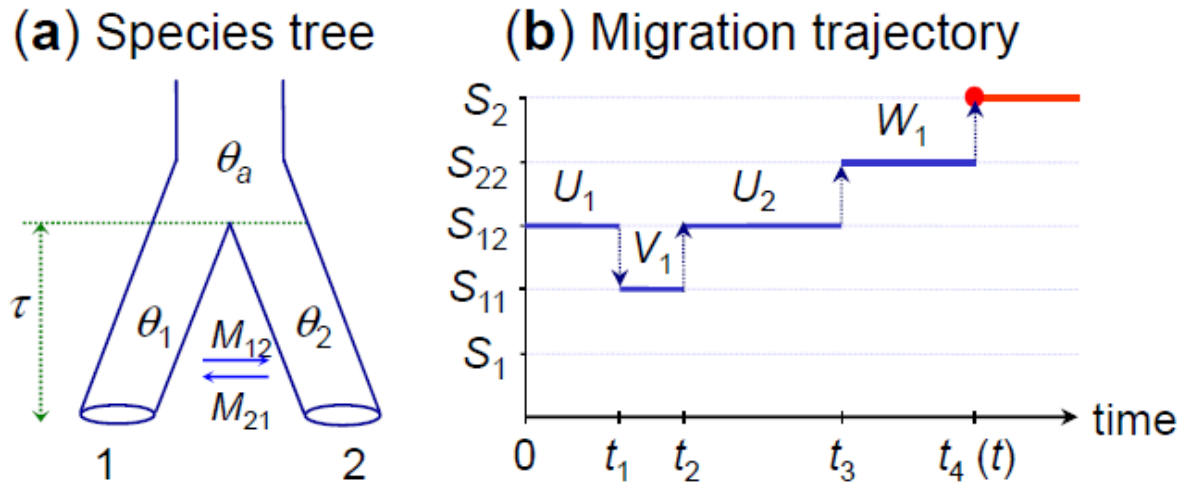


Species tree-gene tree mismatch probability:

$$P_{\text{mismatch}} = \frac{2}{3} e^{-2(\tau_{ABC} - \tau_{AB})/\theta_{AB}}$$

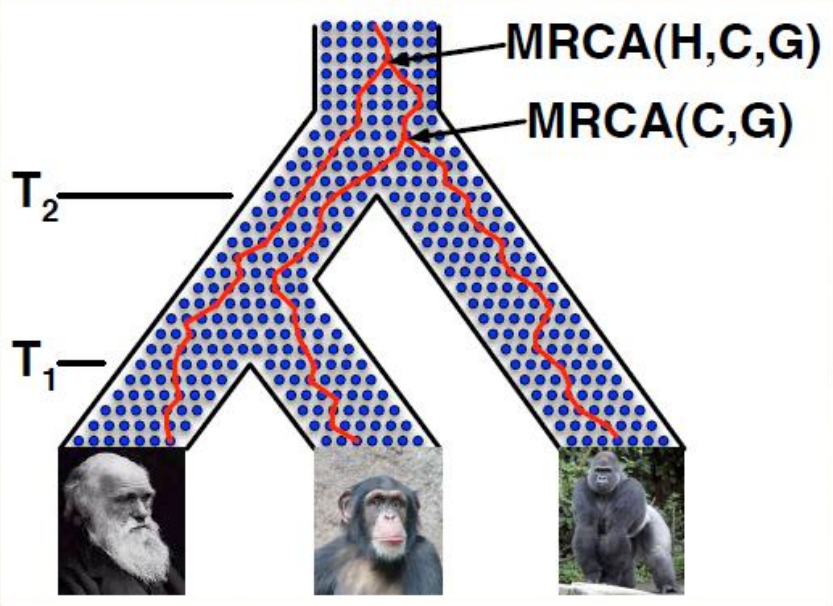
Applications o multispecies coalescent model

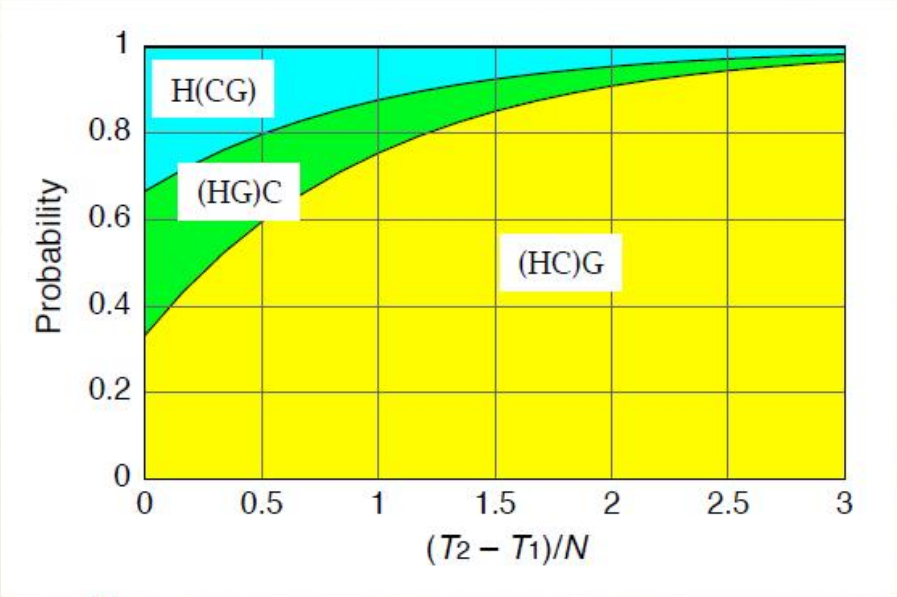
- Species tree estimation (BEST, *BEAST, STEM, etc.)
- Inference of population demographic process
- Estimation of migration patterns and rates (IMa)
- Species delimitation
- …



Yang 2014

multispecies coalescent model ILS





$$P[((HC)G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

$$P[((HG)C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

$$P[((CG)H)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

multispecies network coalescent (MSNC)

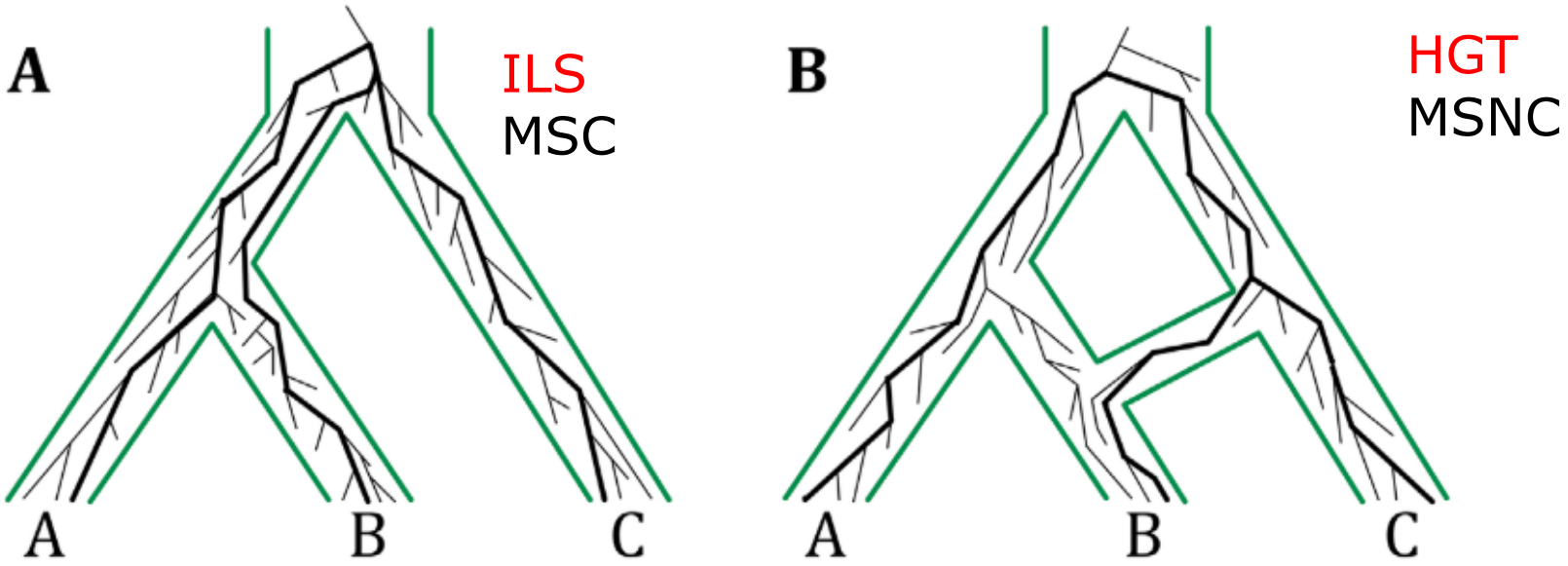
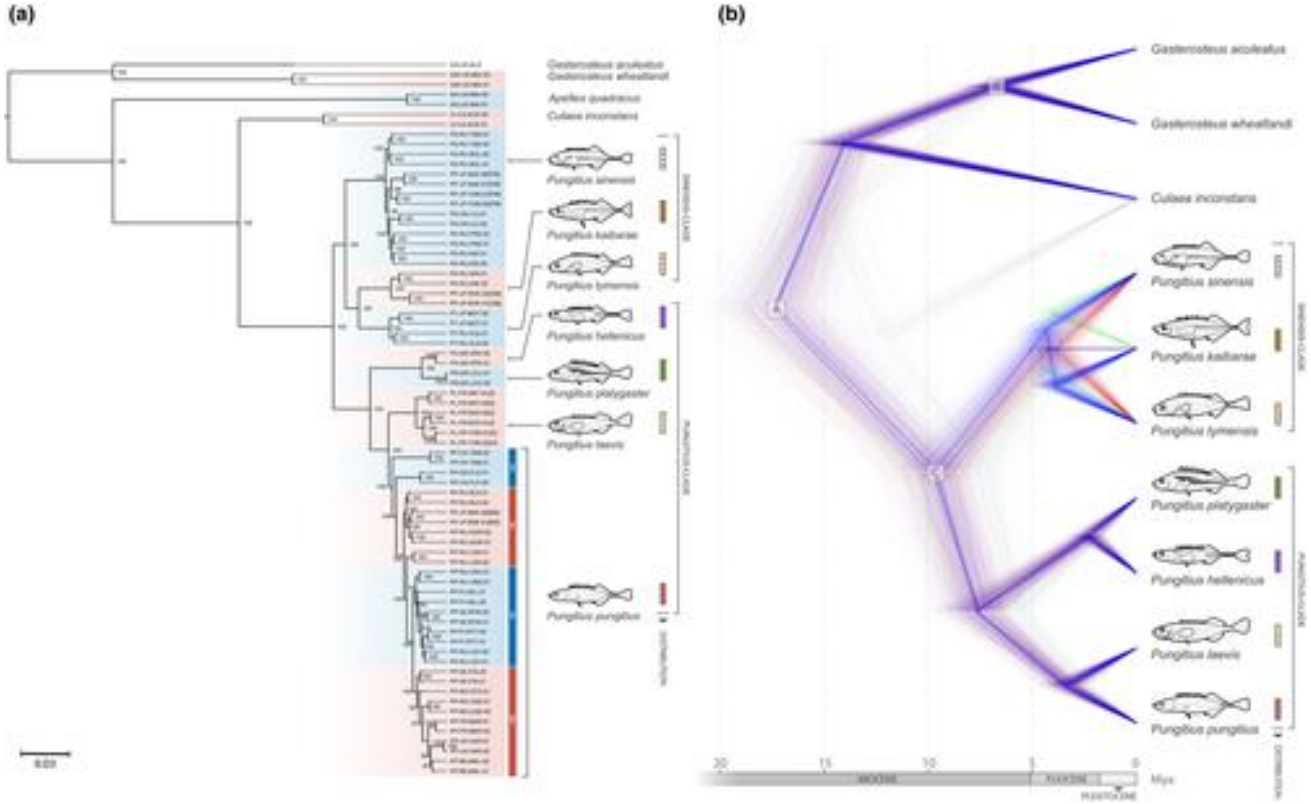


Fig 1. The multispecies coalescent on trees and networks. (**A**) The multispecies coalescent (MSC) links populations by a tree structure and allows for modeling gene genealogies within the branches of a species tree. The gene genealogy indicated by thick lines inside the species tree is incongruent with the species tree due to incomplete lineage sorting (ILS). (**B**) The multispecies network coalescent (MSNC) links populations by a network structure, thus allowing for reticulations events among populations. The gene genealogy indicated by thick lines inside the species network is involved in reticulation, e.g., hybridization. The gene genealogies in both panels have the same topologies, but have different probabilities under the MSC and MSNC models.

Wen *et al.* 2016 PlosGenetics

A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus Pungitius



Molecular Ecology, Volume: 28, Issue: 17, Pages: 4046-4064, First published: 07 August 2019, DOI: (10.1111/mec.15204)

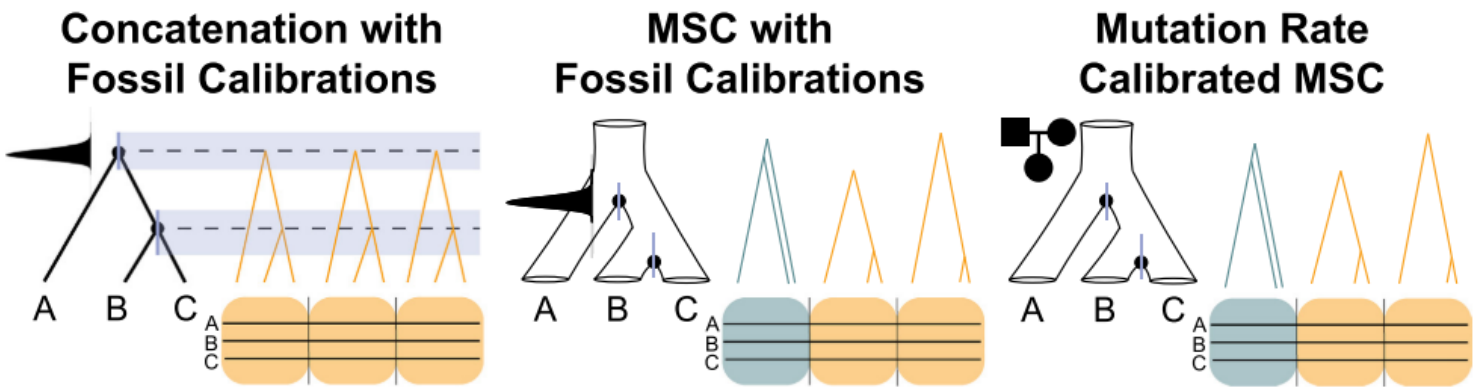
Guo *et al.* 2019 MolEco

(a) Individual-level ML phylogeny, and (b) time-calibrated species-level phylogeny. Three calibration points (A, B and C) defined in the text were used. Divergence times are given in million years ago (Mya). The maximum-clade-credibility summary tree of (b) is provided in Figure S4. The blue colour indicates the most common topology; the red colour indicates the second most common topology; the pale green colour indicates the third most common topology; and the dark green colour indicates all other trees.

Concatenation vs coalescence

Concatenation versus coalescence. Phylogenomic data matrices can be analysed as a single supermatrix (an approach known as concatenation) or each gene alignment can be analysed separately under the multispecies coalescent framework (an approach known as coalescence). The two approaches sometimes yield different tree topologies, contributing to incongruence[68,87]. Determining which approach is more appropriate for a phylogenomic data set is difficult. For example, using simulated multi-locus data, concatenation slightly outperformed a fully coalescent-based approach (wherein gene trees and species trees are coestimated), whereas using coalescent independent sites, both approaches performed comparably[156]. However, an extensive evaluation of coalescent-based and concatenation-based approaches when different biological and analytical factors are at play is lacking, hindering our knowledge of best practices. Moreover, there can be differences in the performance of fully and summary coalescent-based methods (wherein gene trees are first estimated and then the species tree is estimated by summarizing the collection of gene trees). Summary coalescent-based methods are more vulnerable to errors in gene tree inference than fully coalescent-based methods but newer implementations of summary coalescent-based methods take gene tree uncertainty into account[30]. Analyses with both fully and summary coalescent-based methods can be improved through targeted data filtering such as removing loci with low phylogenetic informativeness[157]. Loci that are inconsistent between concatenation-based and coalescence-based methods can also be pruned from data matrices[158].

Concatenation vs coalescence



Strengths

Concatenation with Fossil Calibrations	MSC with Fossil Calibrations	Mutation Rate Calibrated MSC
Computationally efficient for large numbers of tips and loci	Considers discordance between gene trees and species trees	Does not require calibrations on nodes from external information such as fossils

Weaknesses

Concatenation with Fossil Calibrations	MSC with Fossil Calibrations	Mutation Rate Calibrated MSC
May produce biased estimates when ILS is high or when gene sequence divergence is far from species divergence	Increased computational complexity from averaging over gene trees to estimate species tree parameters	Requires external mutation rate estimates from sequenced pedigrees and potentially not appropriate for distant taxa

Common Programs

Concatenation with Fossil Calibrations	MSC with Fossil Calibrations	Mutation Rate Calibrated MSC
BEAST2 [96] MCMCTREE [97] MrBayes [98] PhyloBayes [99]	BPP [5,71] StarBEAST2 [6]	BPP [5,71] StarBEAST2 [6]

Trends in Genetics

Figure 5. Differences between Bayesian Methods for Divergence Time Estimation and Programs for Implementing Them. A number of methods that estimate divergence times with concatenated data [96–99] or the MSC [5,6,71] are available with some variations in prior distributions and relaxed-clock models. The choice of concatenation or MSC methods, and whether divergence times are calibrated with fossils or mutation rates, is dependent on the data set size, prevalence of ILS among species, and appropriateness of a single germline mutation rate.