david posada @ university of vigo, spain

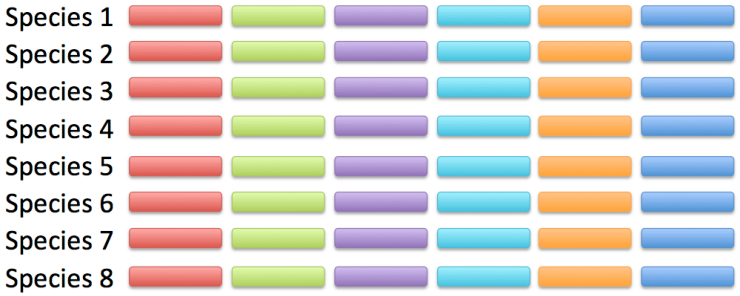
·species tree inference from multilocus datasets

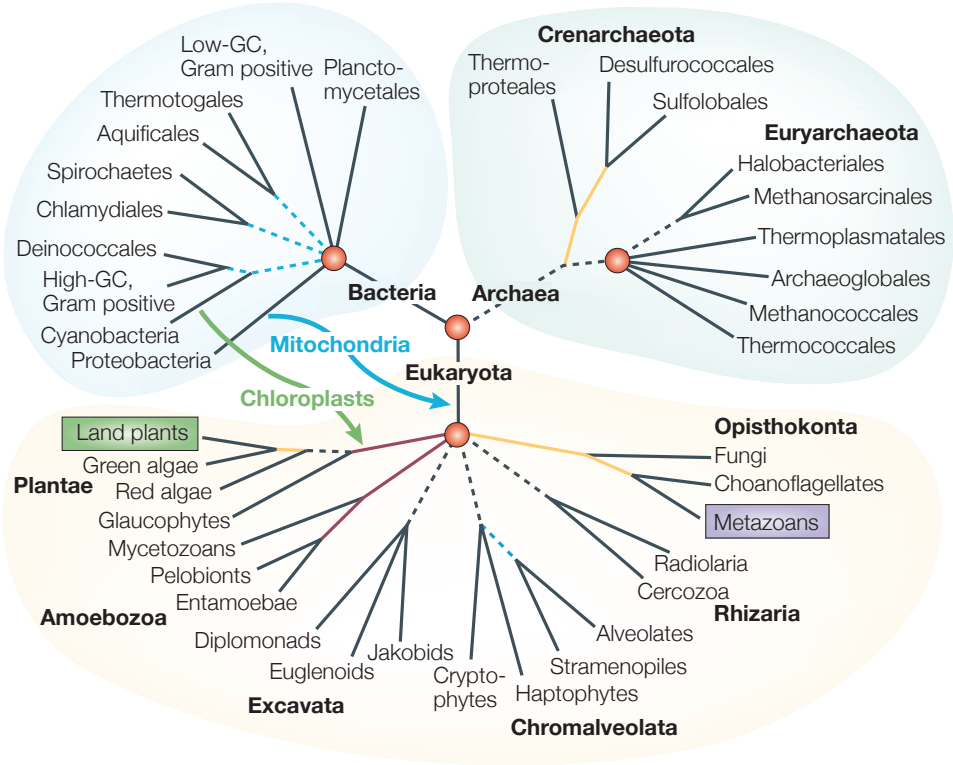
6th International Barcode of Life Conference
August 18-21, 2015
University of Guelph, Canada

phylogenomics lab
University of Vigo

.phylogenomics

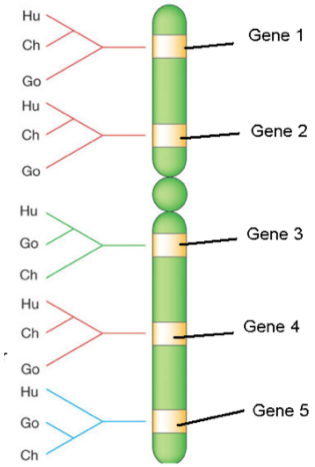
- reconstruction of phylogenies from multilocus data

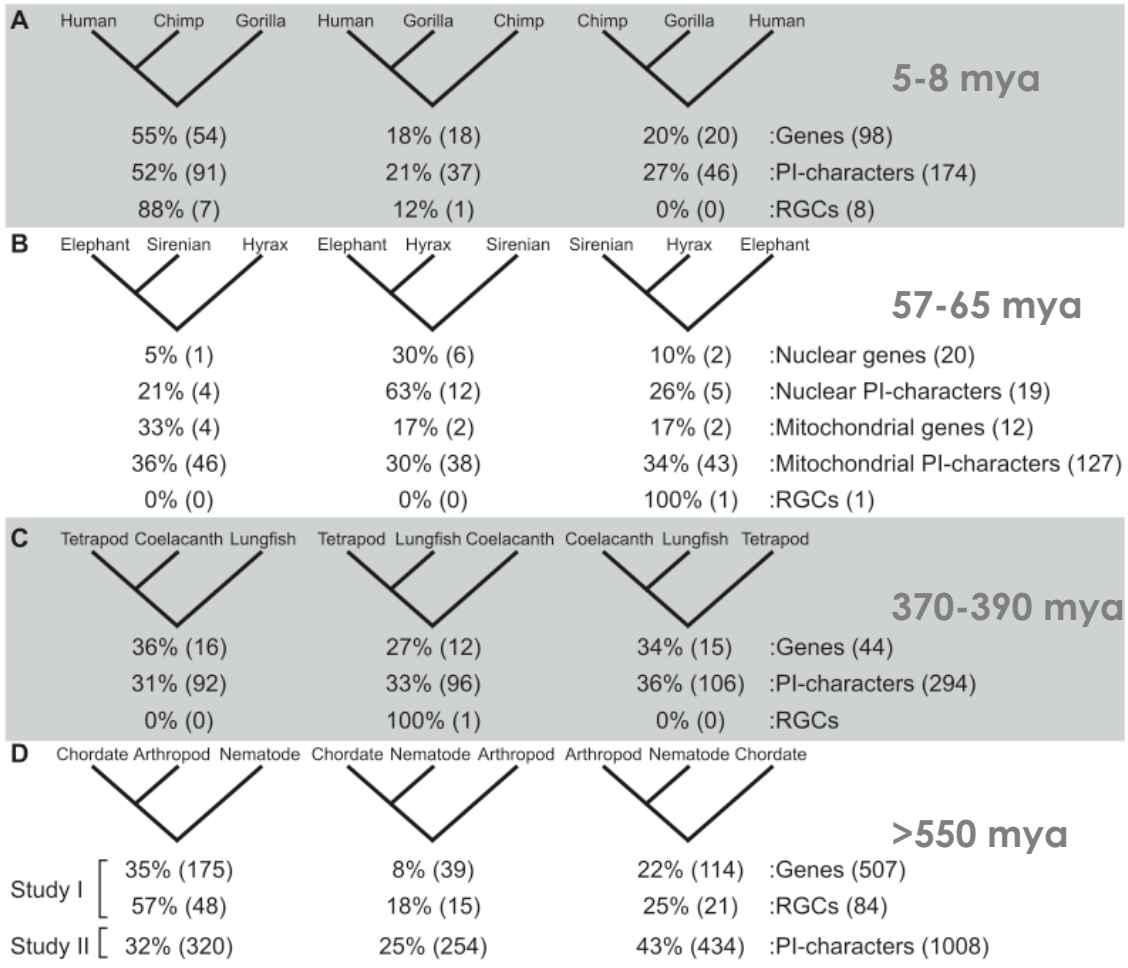




.phylogenomic incongruence

- phylogenomics has unveiled a **significant amount** of conflicting signal



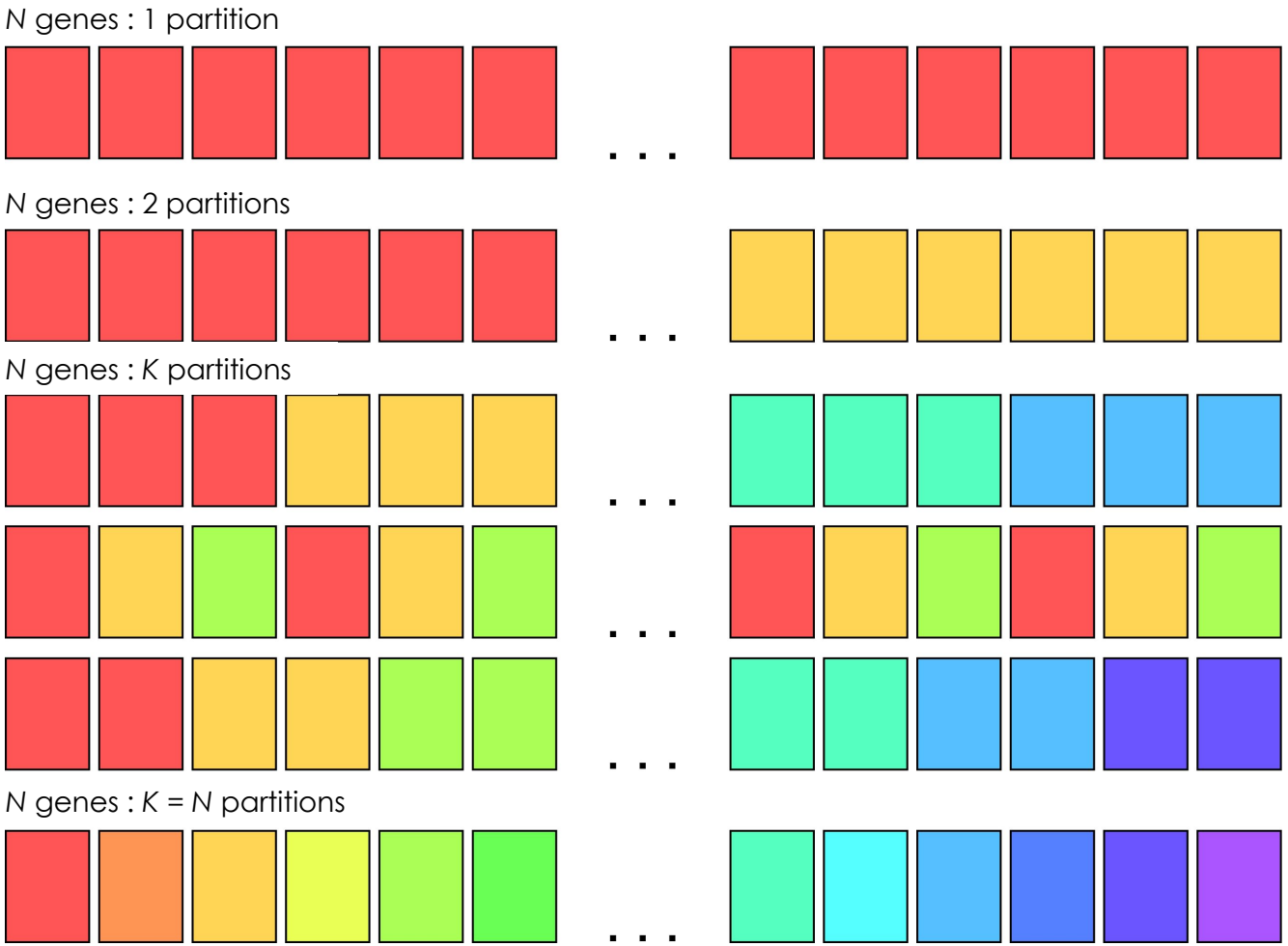


Rokas and Carroll (2006)

.why such incongruence?

- reconstruction **artifacts**
  - systematic and random error
    - substitution model misspecification
  - short branches and bushes
- different **gene trees** do exist within a **species tree**
  - lineage sorting
  - gene duplication and loss
  - horizontal gene transfer (hybridization, recombination)

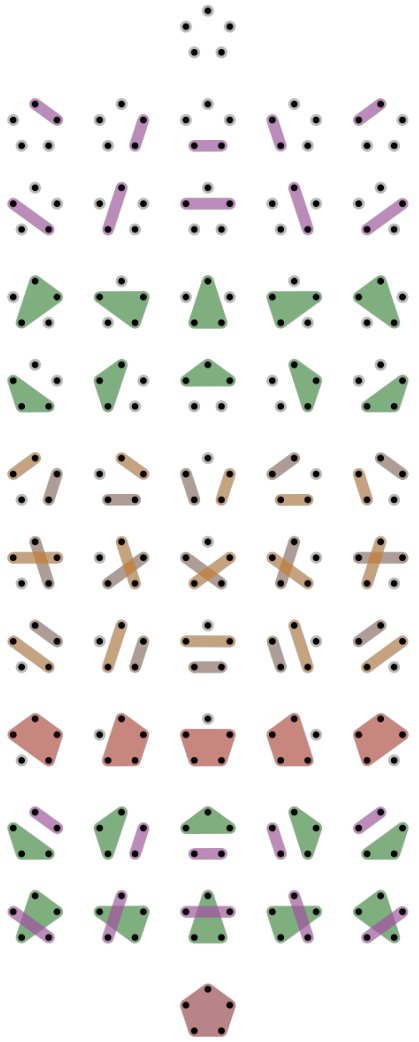
.substitution model partitioning



.many solutions

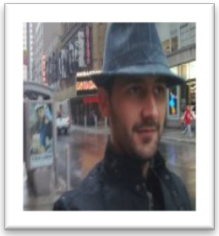
.for 5 genes there are 52 partitioning schemes, for 12, 4 million, for 20, 51×10^{12}.

n k	1	2	3	4	5	6	7	8	9	10	B(n)
1	1										1
2	1	1									2
3	1	3	1								5
4	1	7	6	1							15
5	1	15	25	10	1						52
6	1	31	90	65	15	1					203
7	1	63	301	350	140	21	1				877
8	1	127	966	1701	1050	266	28	1			4140
9	1	255	3025	7770	6951	2646	462	36	1		21147
10	1	511	9330	34105	42525	22827	5880	750	45	1	115975



.partitioning scheme identification

	partitiontest (ours)		partitionfinder	
	hcluster	greedy	hcluster	greedy
PPR	0.20	**0.30**	0.01	0.25
RI	**0.97**	0.93	0.85	0.95
ARI	**0.78**	0.70	0.03	0.77
Kdiff	2.01	**-1.71**	13.68	-1.77
runtime	**01:20:25**	05:25:50	01:59:00	14:31:20



Diego Darriba

Table 6. Simulation summary. Parameter va

	Sim1
N, number of genes	U(10,50)
K, number of partitions	U(1,N)
Gene length	U(500,1500)
Number of taxa	U(6,40)
Topology	Fixed
Number of replicates	4,000
Tree length	U(0.5,15)

.phylogenetic accuracy

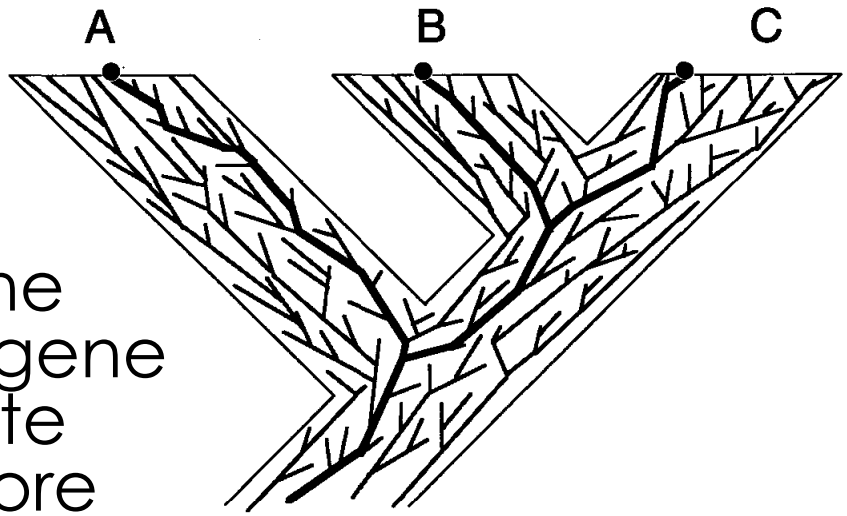
	a priori partitions			partitiontest (ours)		partitionfinder	
	K=1 (GTR+G)	K=true (GTR+G)	K=N (GTR+G)	hcluster	greedy	hcluster	greedy
% true topology	0.787	0.890	0.890	**0.892**	0.842	0.885	0.820
RF	0.018	**0.007**	**0.007**	**0.007**	0.012	**0.007**	0.013
runtime	--	--	--	**01:20:25**	05:25:50	01:59:00	14:31:20

Table 6. Simulation summary. Parameter va

	Sim1
N, number of genes	U(10,50)
K, number of partitions	U(1,N)
Gene length	U(500,1500)
Number of taxa	U(6,40)
Topology	Fixed
Number of replicates	4,000
Tree length	U(0.5,15)

.species trees and gene trees

- a **species tree** represents the pattern of branching of species lineages via the process of speciation.



- a **gene tree** represents the pattern of branching of gene copies after they replicate and are passed on to more than one offspring.

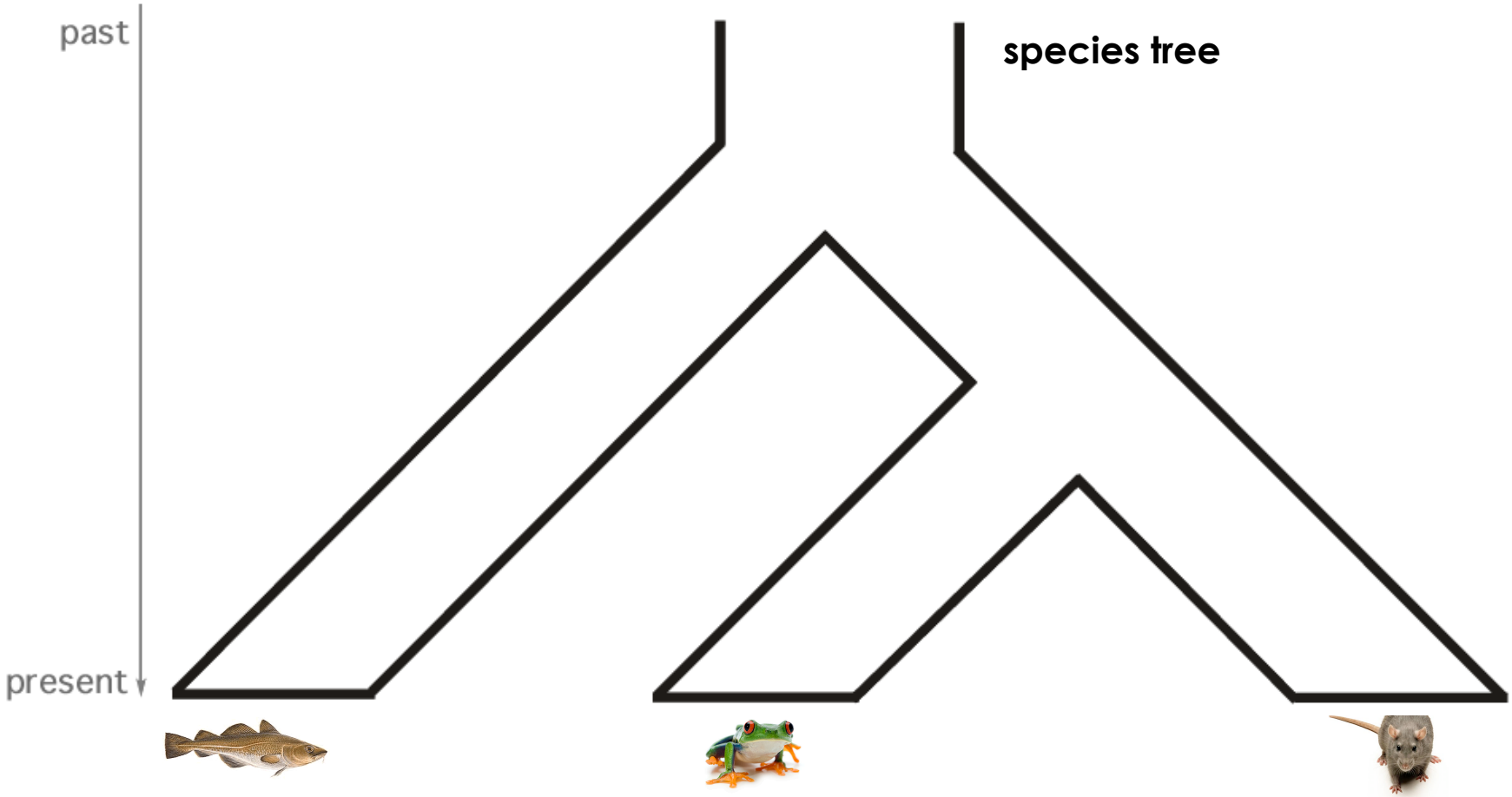
Syst. Biol. 46(3):523–536, 1997

GENE TREES IN SPECIES TREES

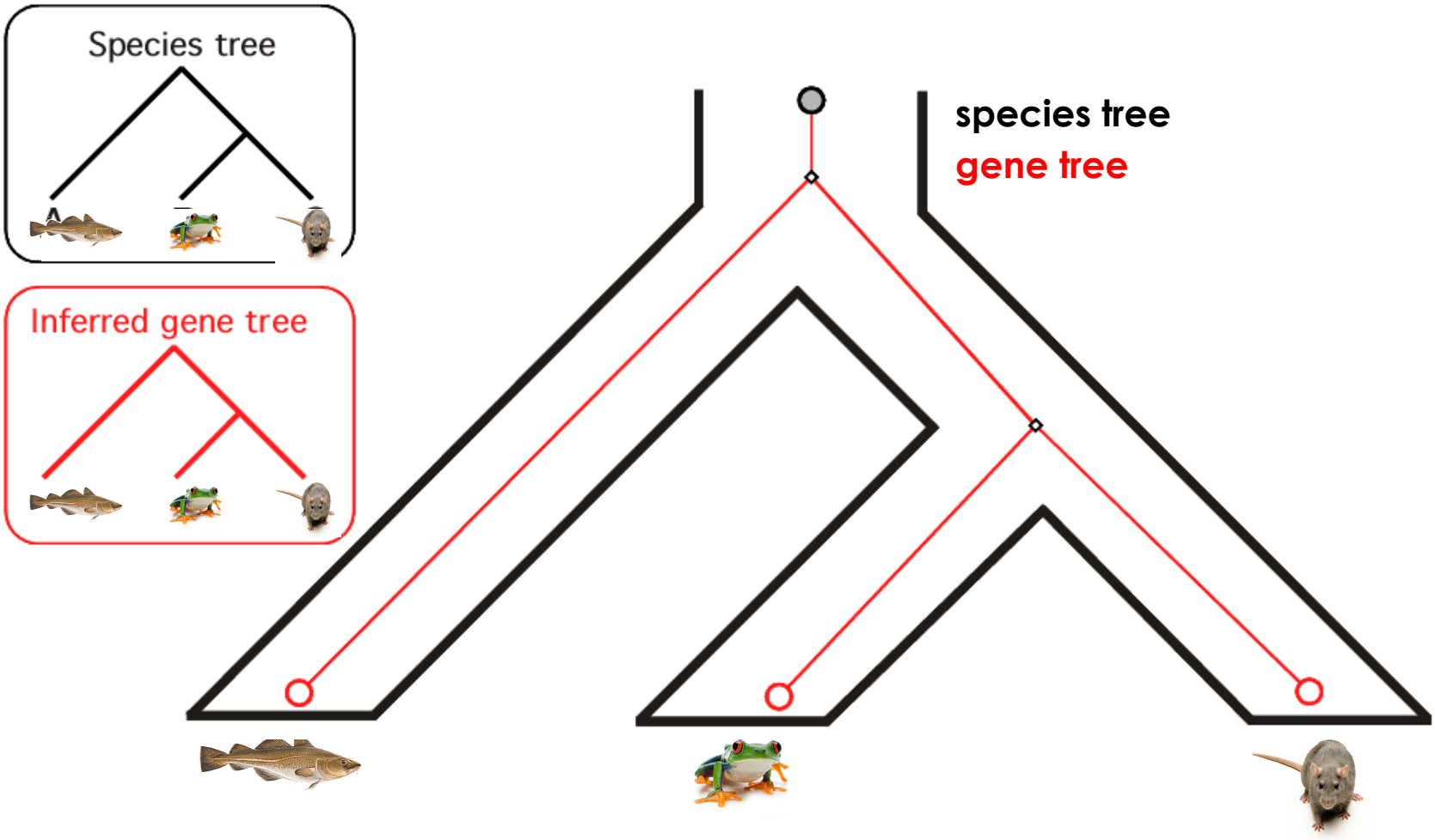
WAYNE P. MADDISON

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

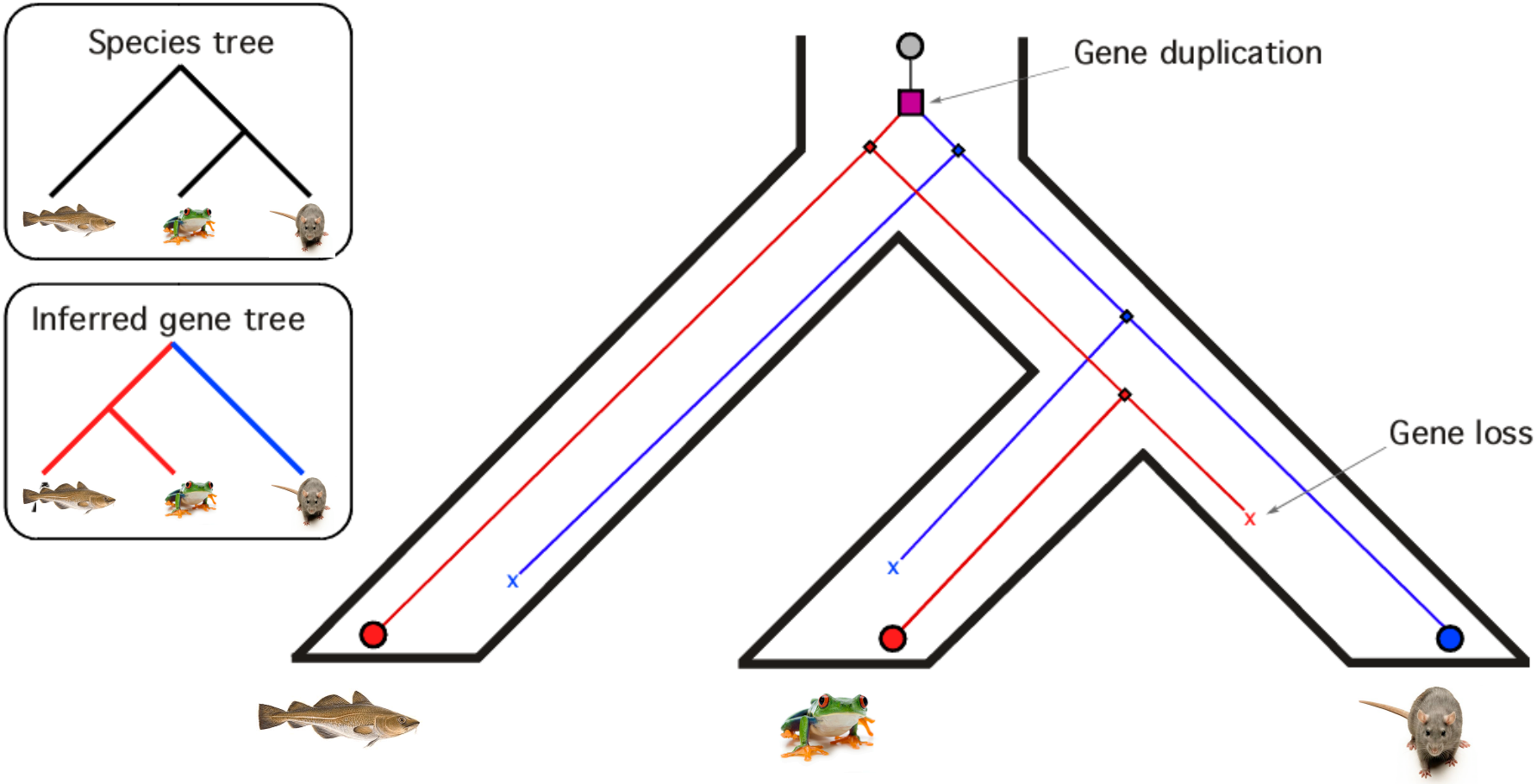
.species tree



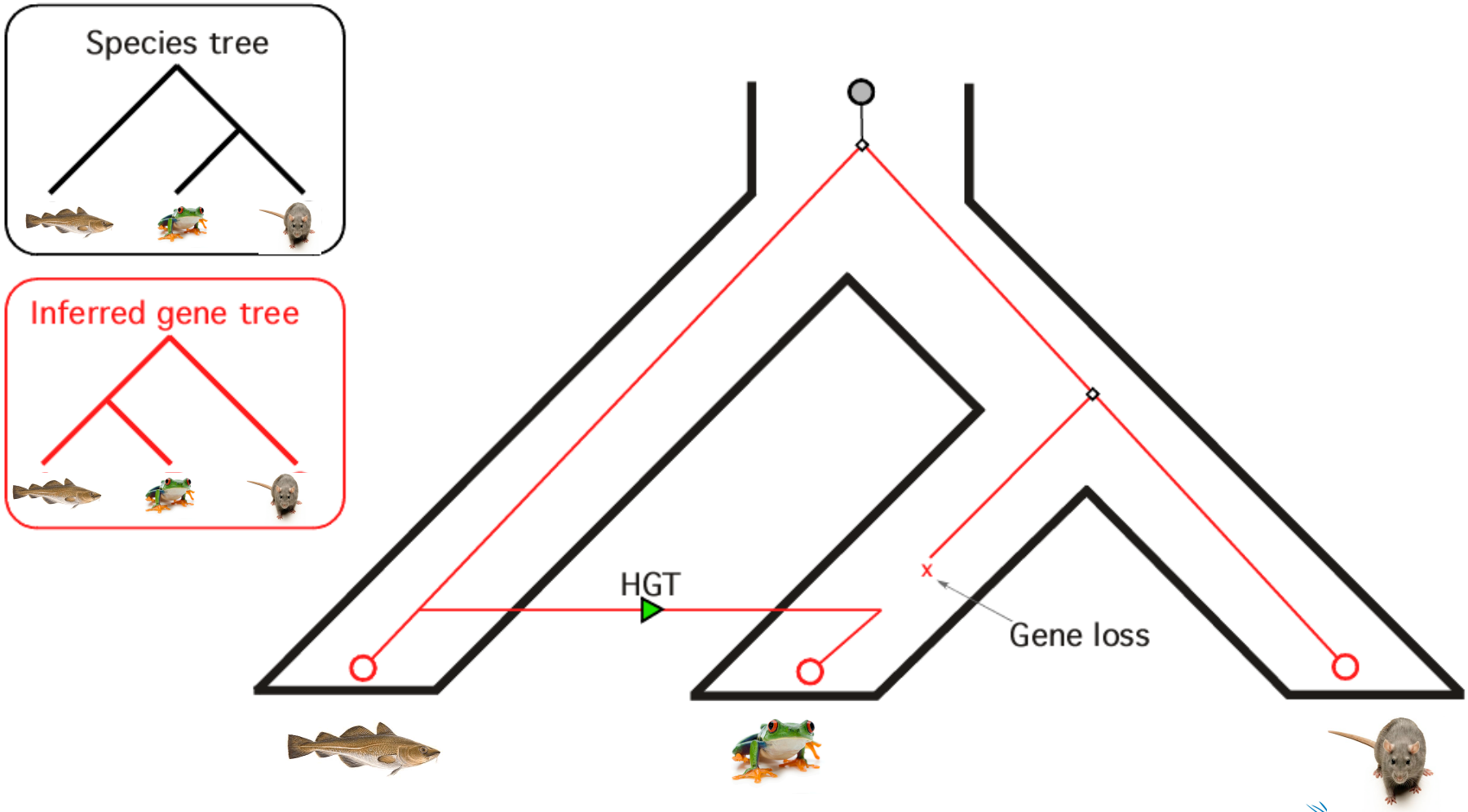
.dominant paradigm



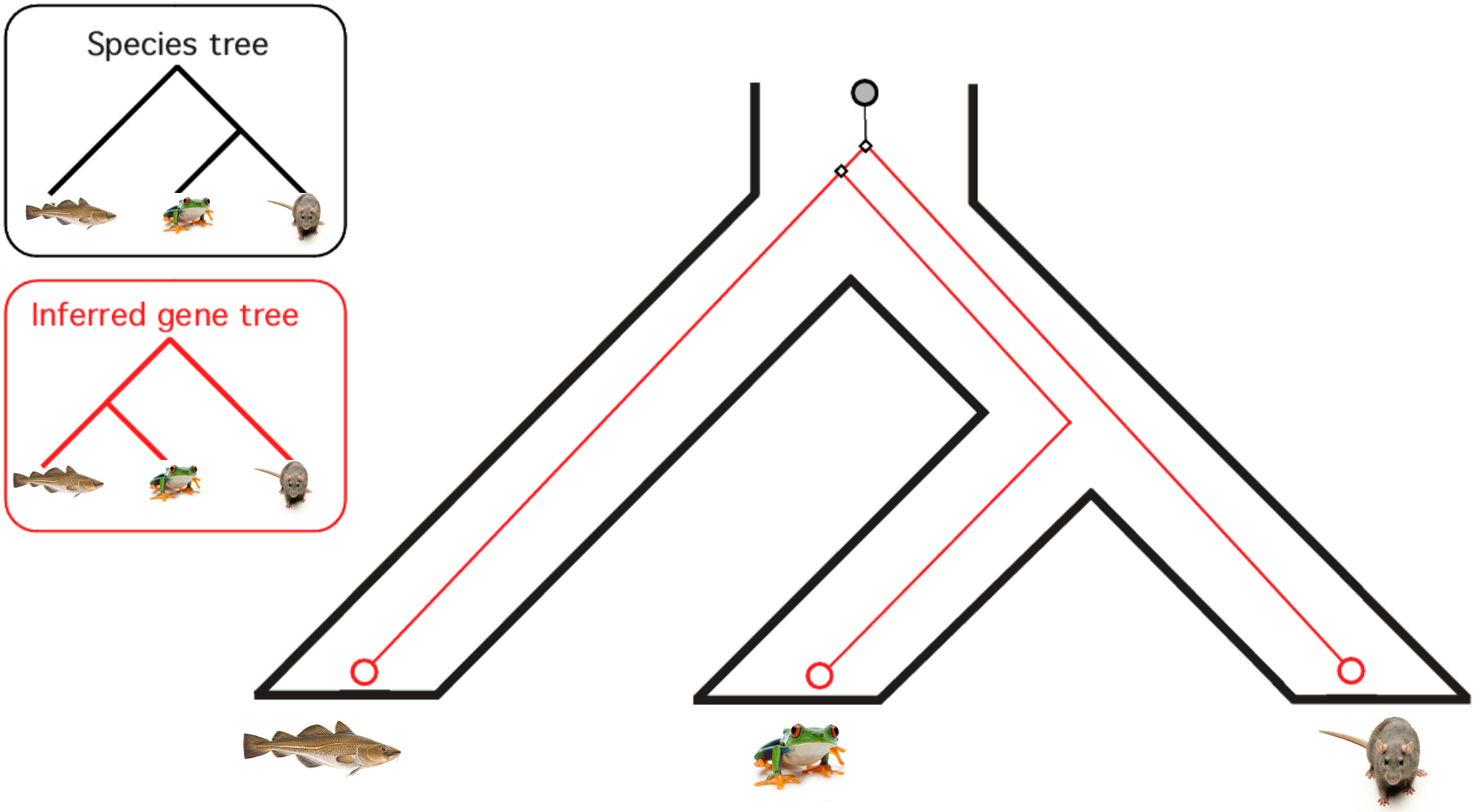
.gene duplication and loss



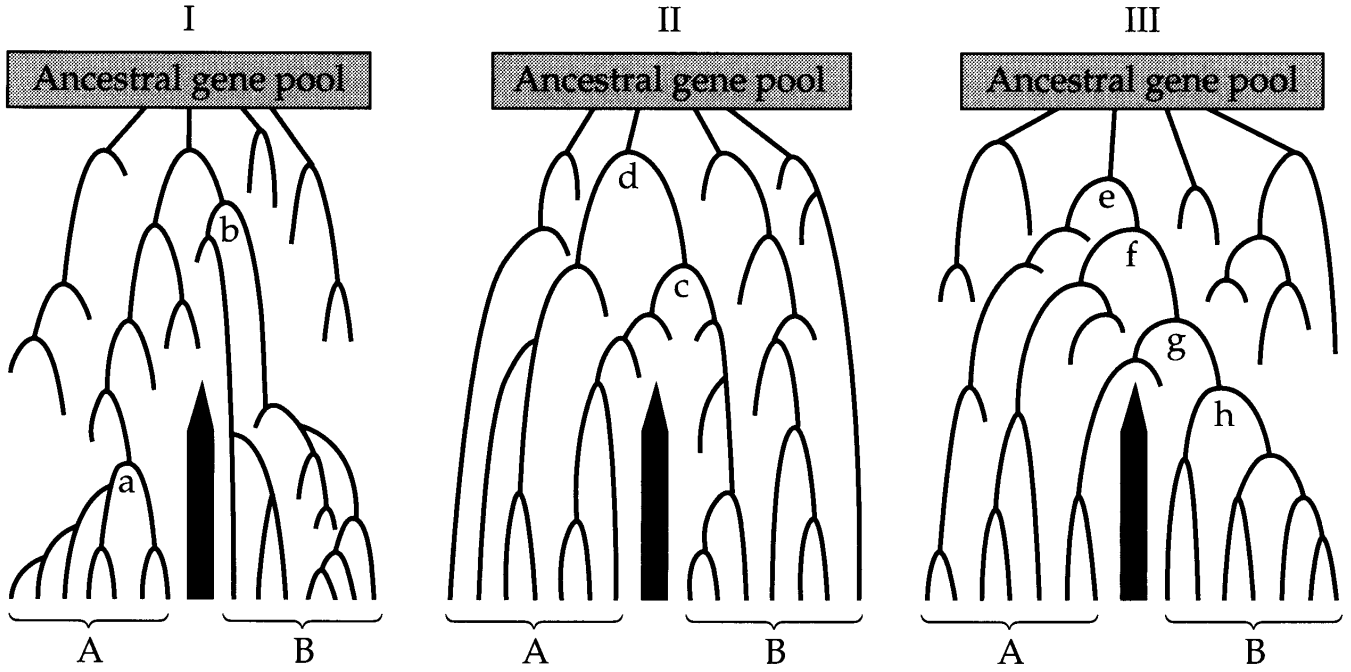
.horizontal gene transfer



.incomplete lineage sorting



.incomplete lineage sorting

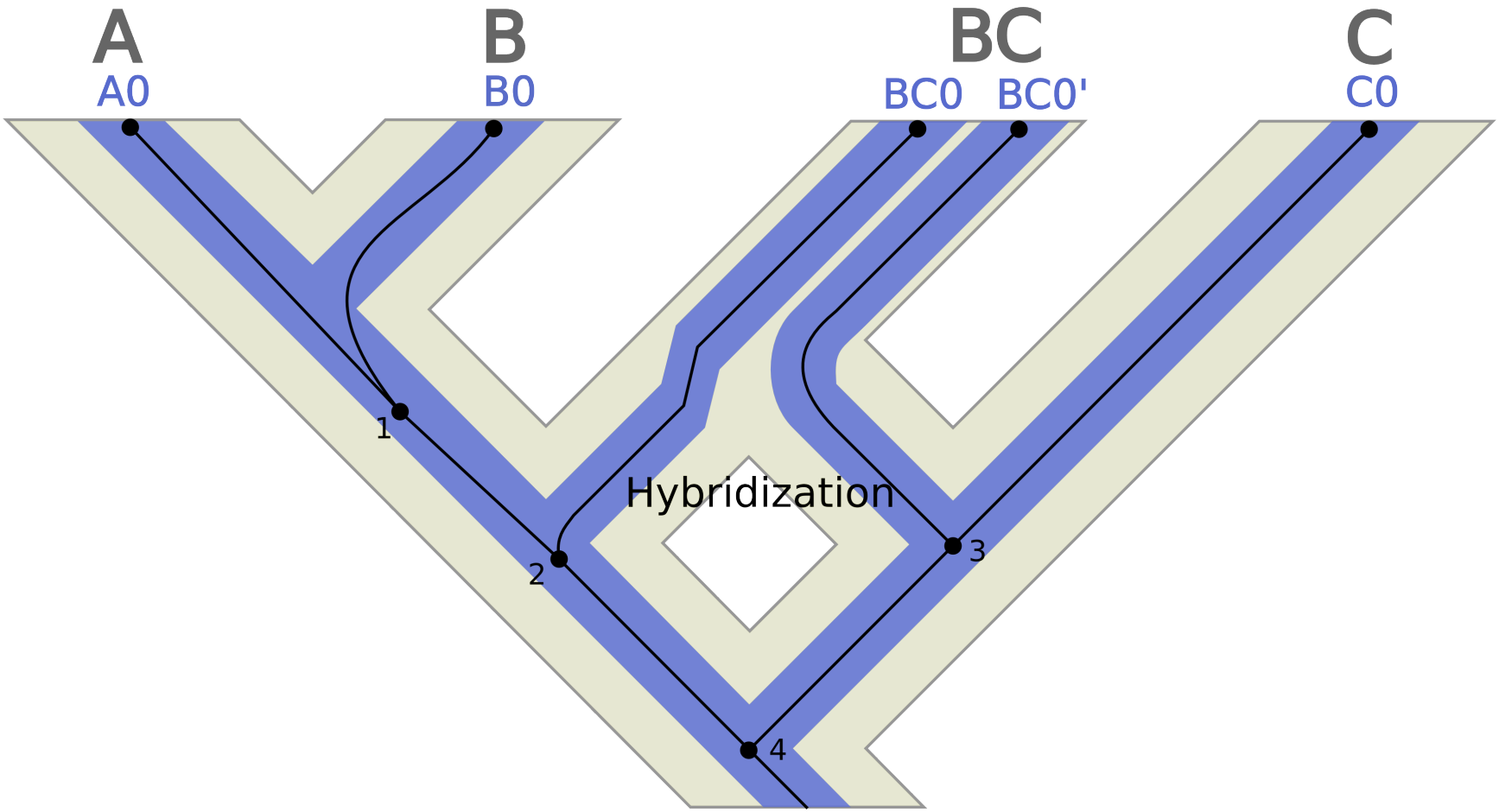


~20% animal mtDNA

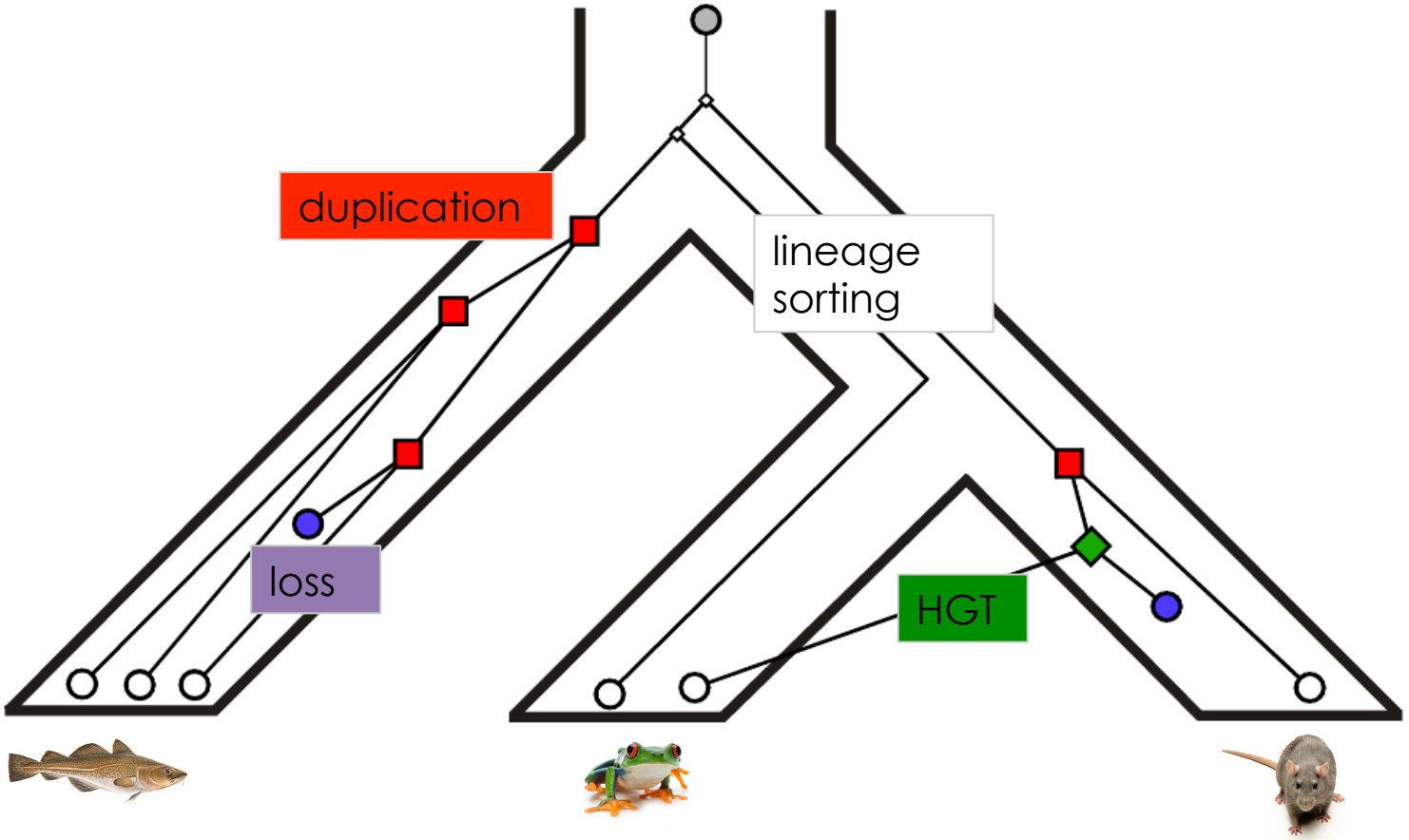
Figure 4.11 **Three categories of phylogenetic relationships** between two sister taxa (A and B) are possible with respect to an allelic genealogy. Lowercase letters point out important ancestral nodes to which extant alleles or haplotypes trace. Solid dark bars indicate barriers to reproduction (extrinsic or intrinsic). The phylogenetic categories in the gene tree are as follows: I, reciprocal monophyly; II, polyphyly; III, paraphyly of A with respect to B. (After Avise et al. 1983.)

(from Avise, 2004, Molecular Markers, Natural History and Evolution, 2nd Ed.)

.hybridization



.'full' model



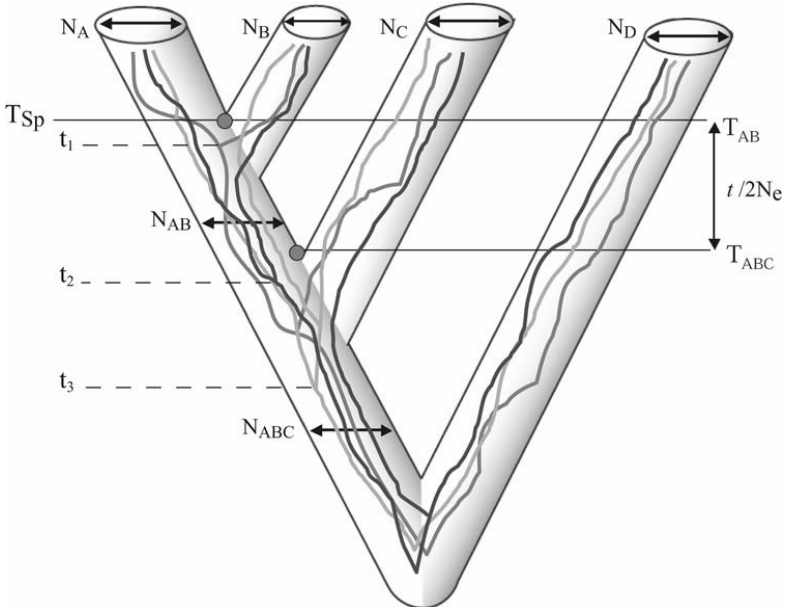
$$P(S \mid D_1 \ldots D_k) \Leftarrow P(S, G_1 \ldots G_k, \lambda, \mu, \varphi, N, t \mid D_1 \ldots D_k) = \frac{P(D_1 \ldots D_k, G_1 \ldots G_k \mid S, \lambda, \mu, \varphi, N, t)}{P(D_1 \ldots D_k)}$$

.species tree methods

- **supermatrix** = concatenation
  - phylogenetic "trend"
  - binning
  - orthologous genes
- **supertree**
  - supertrees *sensu est*. : (e.g., MRP, RFst, MLst, GTP, ASTRAL,…)
  - consensus (e.g., BUCKY,…)
  - parametric (e.g., STEM, STAR, MP-EST,…)
- **full probabilistic** (e.g., *BEAST, PhylDog,…)

.multispecies coalescent

- 'censored' coalescence of a gene tree *g* within a species tree *S*.
  - standard coalescent within species
  - after species splits, lineages from descendant groups can coalesce
- 'species' = **any diverging group of individuals** or lineage



$$P(g \mid S) = \prod_{b \in S} P(L_b(g) \mid N_b(t))$$

$$P(L_b(g) \mid N_b(t)) = \prod_{i=0}^{k-1} \frac{1}{N_b(t_{i+1})} \prod_{i=0}^{k} \exp\left(-\int_{t_i}^{t_{i+1}} \frac{\binom{l-i}{2}}{N_b(t)} dt \right)$$

.problems

- fully probabilistic models, although more realistic, tend to be **slow**
- usually **limited** to particular sources of gene tree / species tree disagreement
  - the multispecies coalescent assumes that all genes from the same species are orthologous
  - duplication and loss models assume that sequences mapped to one species are necessarily the product of a duplication

Systematic Biology Advance Access published November 4, 2014

Syst. Biol. 0(0):1–20, 2014

DOI:10.1093/sysbio/syu082

A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction

LEONARDO DE OLIVEIRA MARTINS*, DIEGO MALLO, AND DAVID POSADA

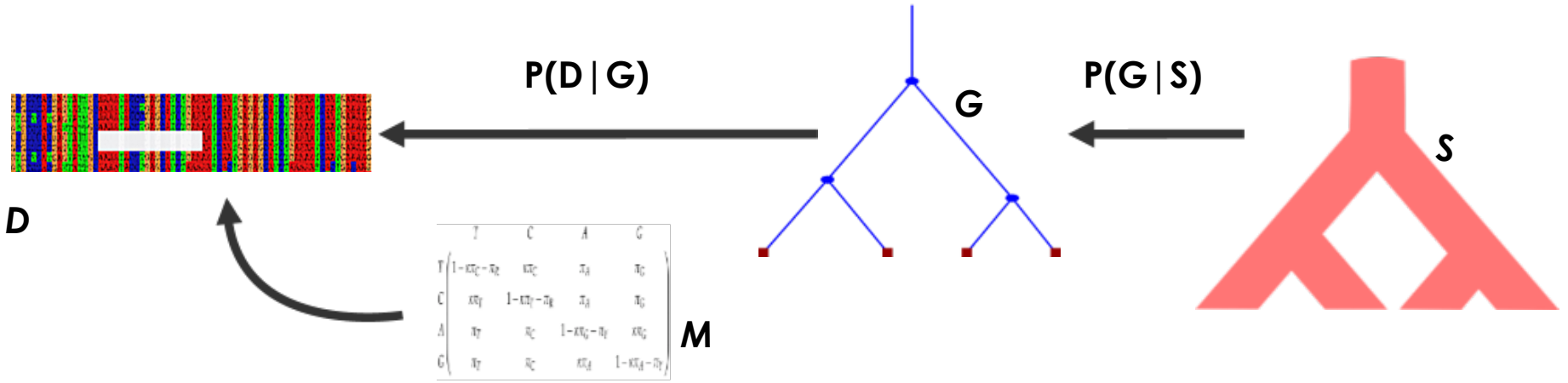
Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, 36310, Spain
**Correspondence to be sent to: Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, 36310, Spain; E-mail: leomrtns@uvigo.es.*

Received 5 Februay 2014; reviews returned 4 June 2014; accepted 30 September 2014
Associate Editor: Laura Kubatko

Leo Martins

https://bitbucket.org/leomrtns/guenomu/

.P(S | D)



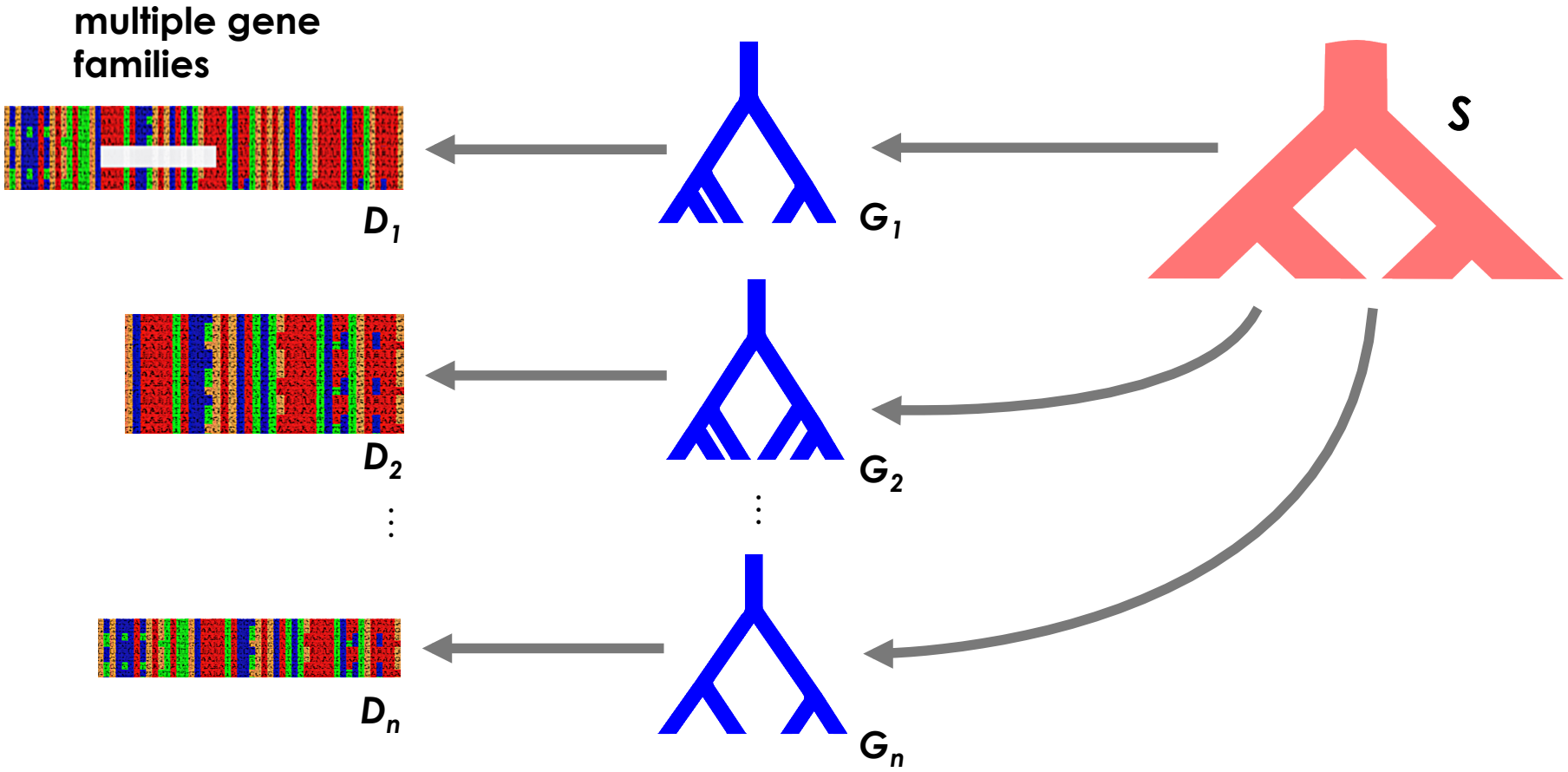
$$P(S|D) = P(D|G,M)P(M)P(G|S)P(S)$$

posterior probability species tree

prob. alignment

prob. gene tree

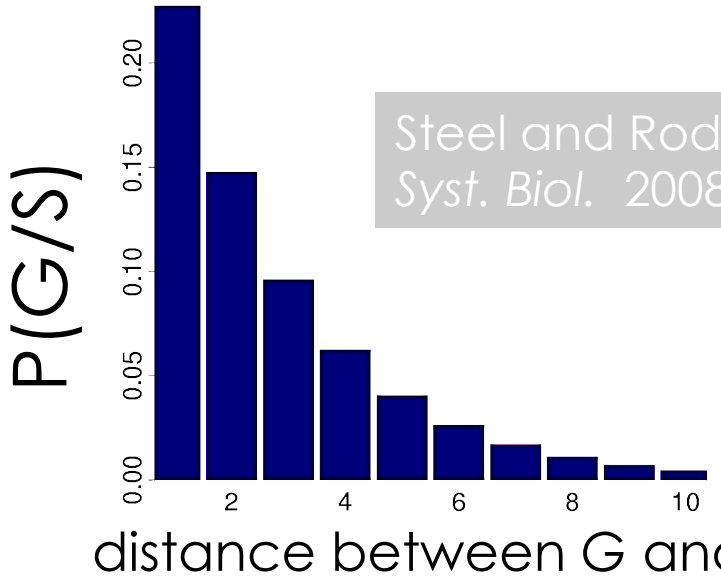
.P(S | D_1,D_2,…,D_n)



$$P\left(S\middle|D_1,D_2,\ldots,D_n\right)=\prod_{i=1}^{n}P\left(D_i\middle|G_i\right)P\left(G_i\middle|S\right)P\left(S\right)$$

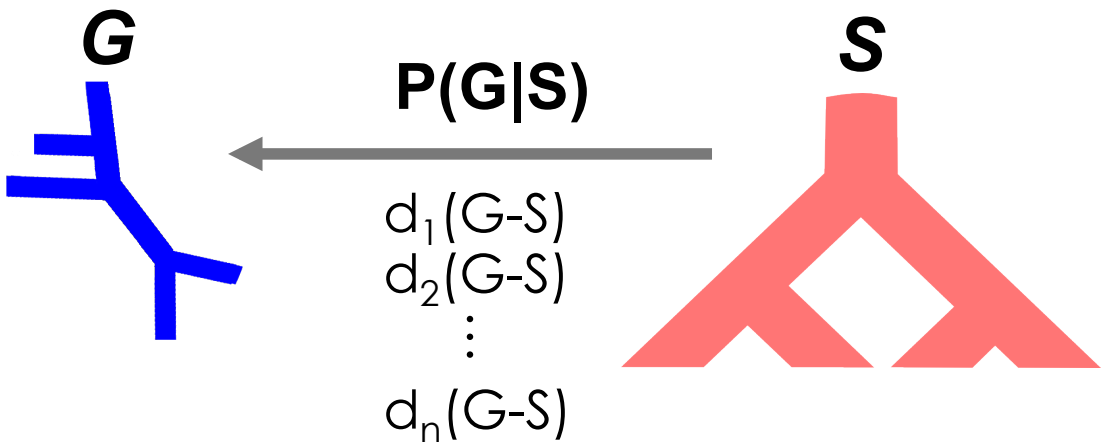
.approximating P(G|S) – ML supertrees

simplest explanation for P(G|S):



Steel and Rodrigo. *Syst. Biol.* 2008

.our approach to P(G|S)

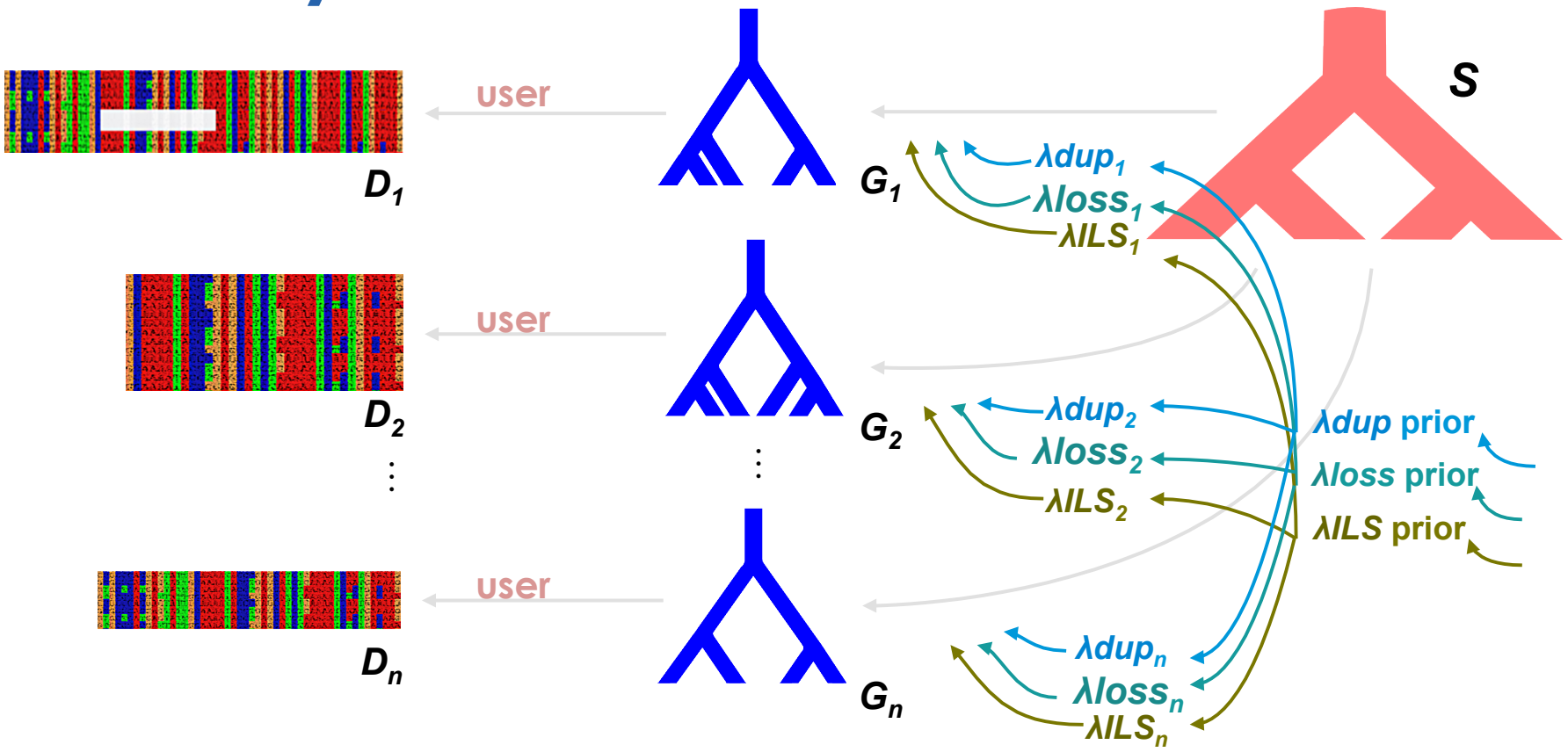


- work with **unrooted gene trees**
- **multiple individuals** per species
- **multiple** measures of disagreement between G and S

.measures of disagreement

- **reconciliation** between rooted *S* and unrooted *G*
  - duplication and loss
  - incomplete lineage sorting
  - optimal* G root location
- **non-parametric** distances
  - do not model biological phenomena
  - mulRF: multilabeled gene trees
- each distance can contribute distinctly through **different penalty parameters**
- we **ignore branch lengths** … for now

.bayesian hierarchical model

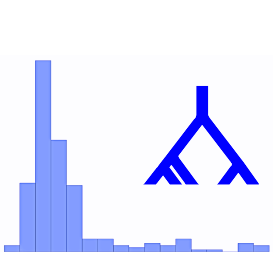


$$P(S,\Theta|\boldsymbol{D}) \propto \underbrace{P(\boldsymbol{\lambda}_0)P(S)}_{\text{genome-wide}} \times \prod_{i=1}^{N} \underbrace{P(D_i,\boldsymbol{\theta}_i|G_i)P(G_i|\lambda_{i\cdot},S)P(\lambda_{i\cdot}|\boldsymbol{\lambda}_0)}_{\text{gene-family-specific}}$$

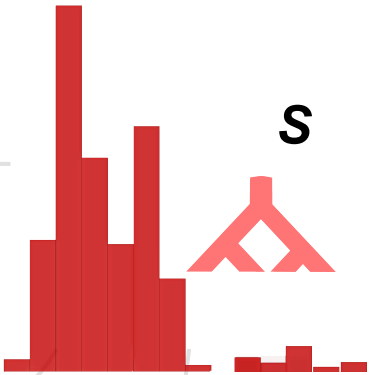
.implementation: guenomu

.2-stage MCMC approach

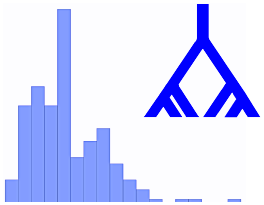
simulated annealing to find modes



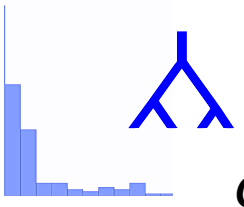
G_1 distribution



S



G_2 distribution



G_n distribution

INPUT
gene tree posterior distributions
(MrBayes, PhyloBayes,…)

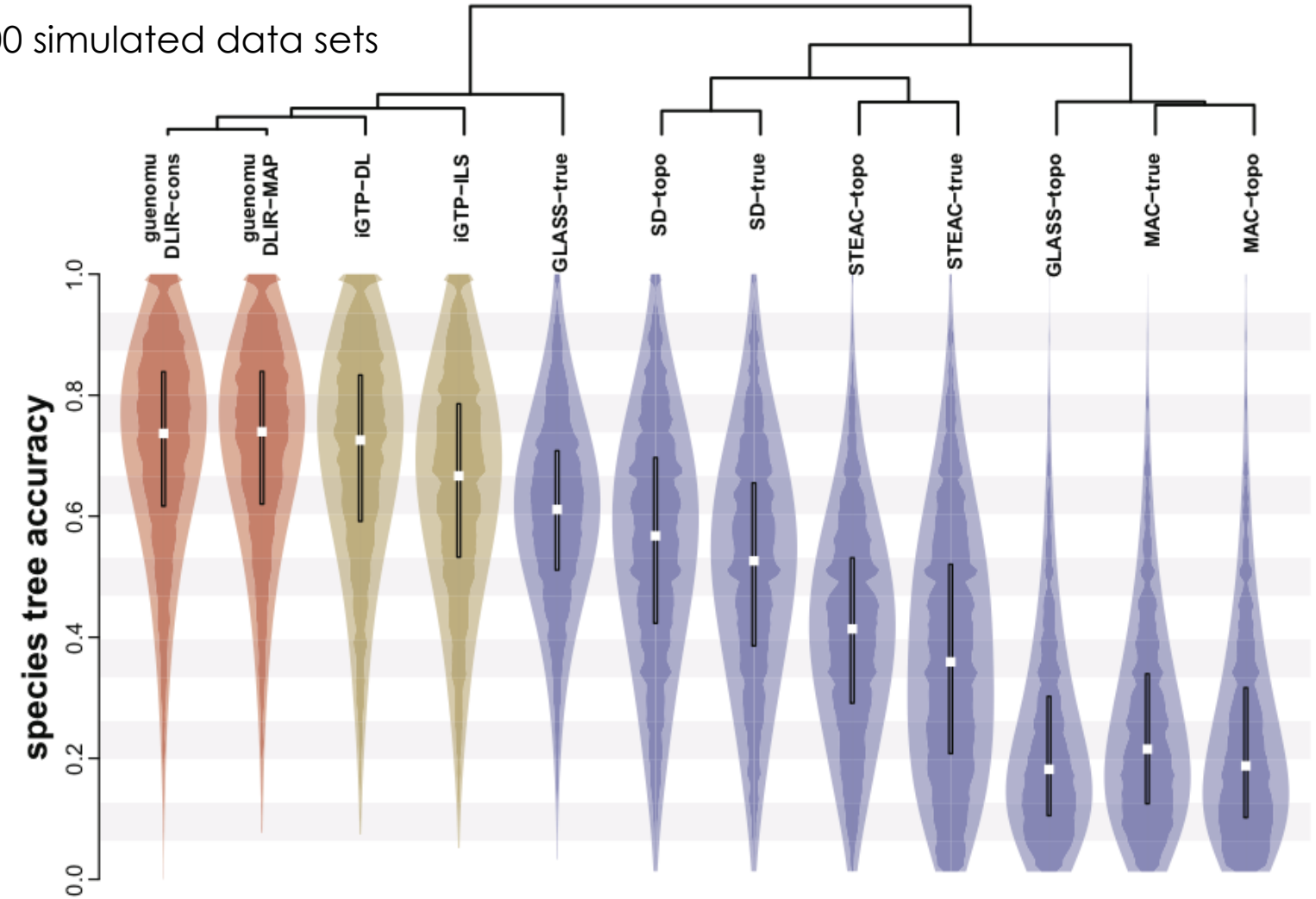
OUTPUT
species tree posterior distribution

Importance sampling

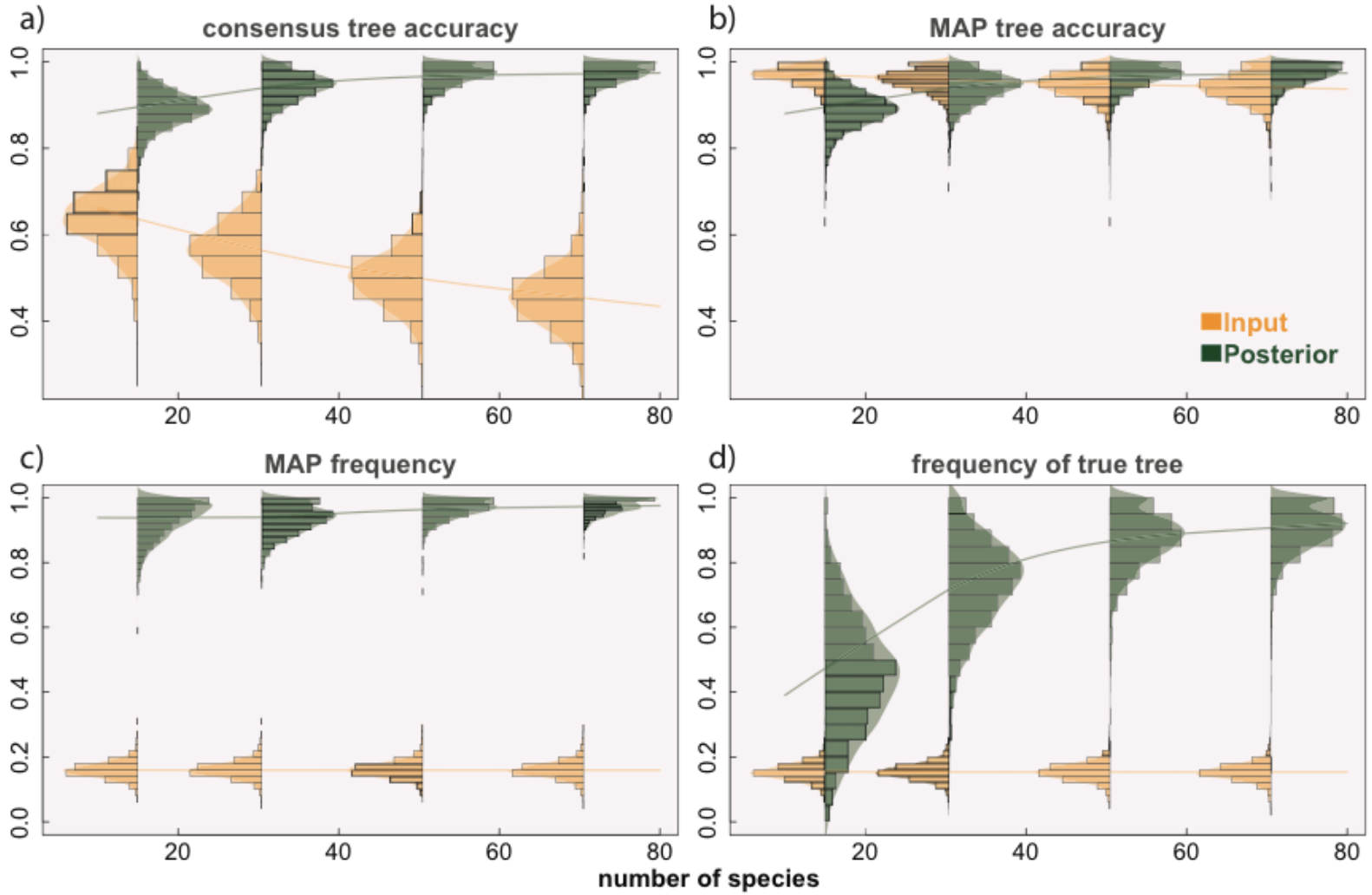
$$P(S, \Theta \mid D) \propto P(\lambda_0)P(S) \times \prod_{i=1}^{N} P(D_i, \theta_i \mid G_i) P(G_i \mid \lambda_i, S) P(\lambda_i \mid \lambda_0)$$

.better species trees

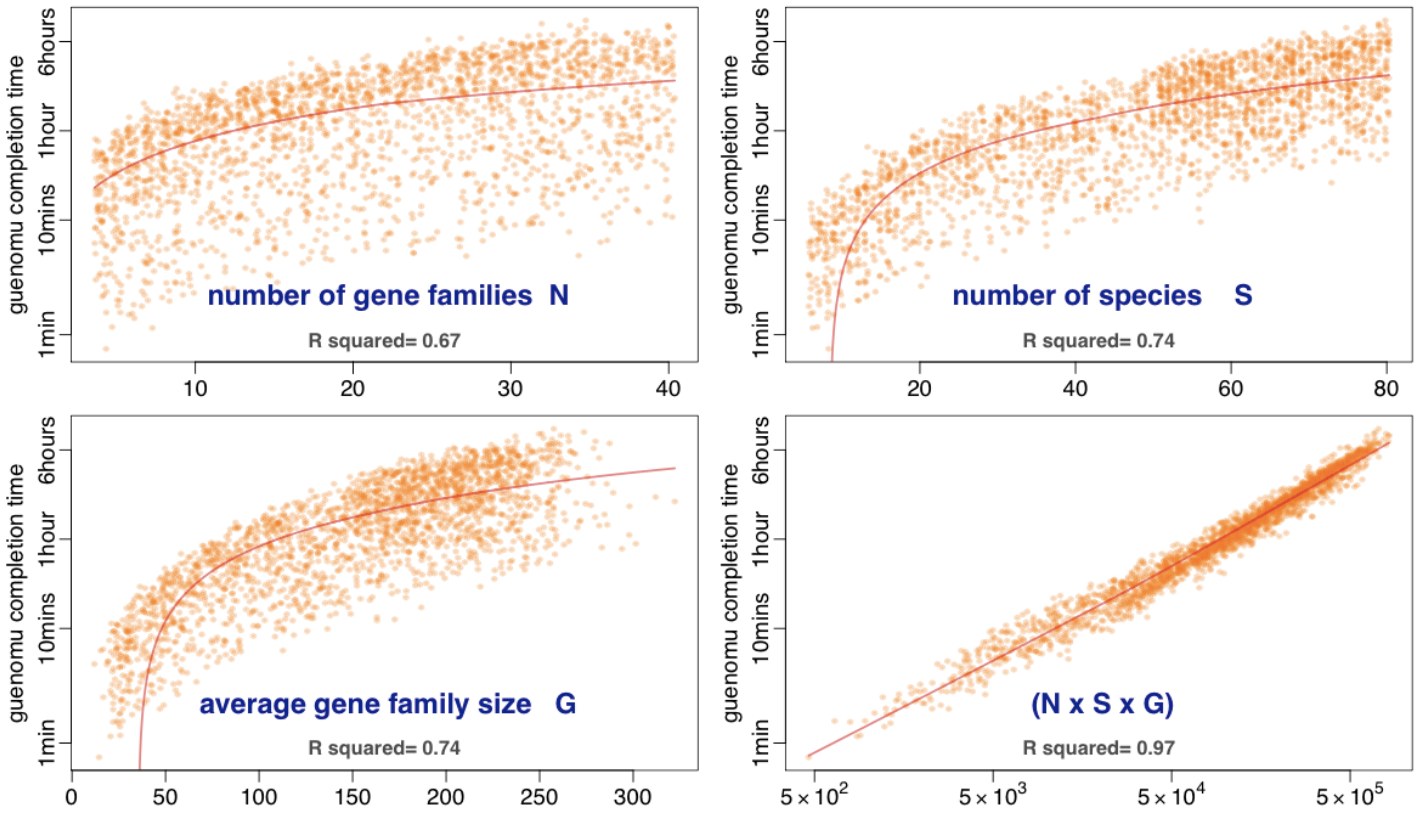
12,000 simulated data sets



.better gene trees



.very good speed

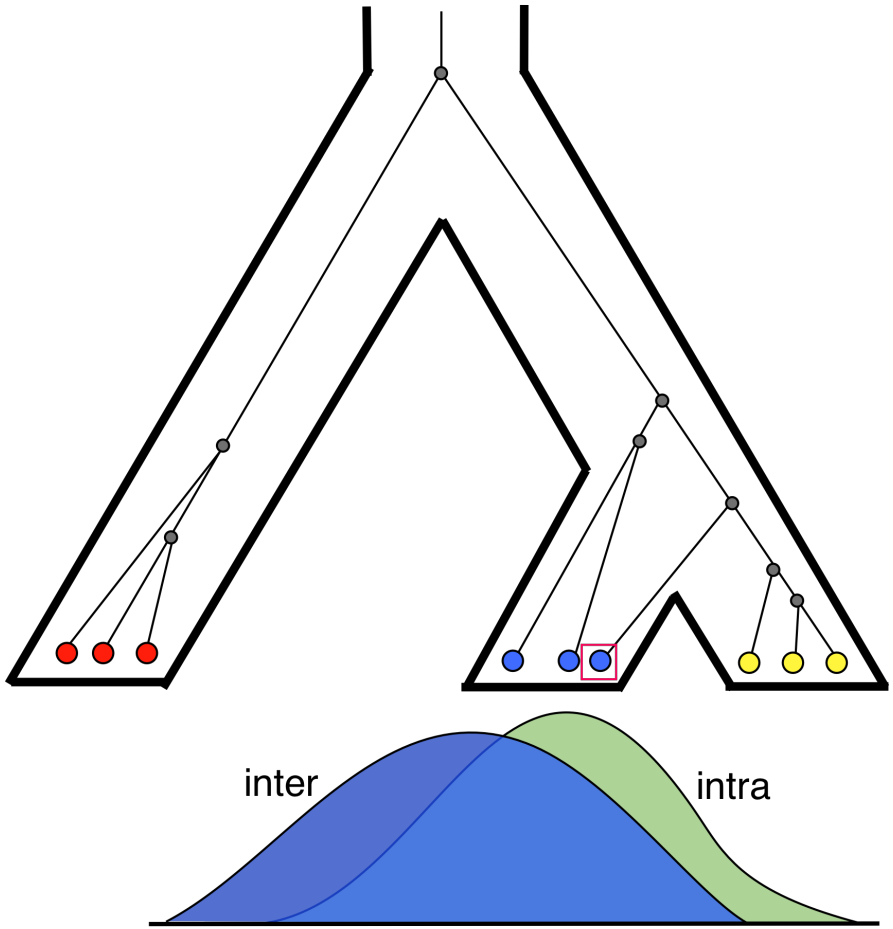


Simulations: **average run of 1.5 hours on one processor** (longest took 9 hours)

37 mammals, 447 gene families: **less than 6 hours in one processor**

.but what for barcoding?

- ILS increases the barcode overlap
- ILS can directly mislead barcoding
- in multigene datasets chances for ILS are higher



.multispecies coalescent for barcoding

Syst. Biol. 63(4):639–644, 2014

DOI:10.1093/sysbio/syu028
Advance Access publication March 28, 2014

A Preliminary Framework for DNA Barcoding, Incorporating the Multispecies Coalescent

MARK DOWTON[1,*], KELLY MEIKLEJOHN[2], STEPHEN L. CAMERON[3], AND JAMES WALLMAN[2]

[1]*Centre for Medical and Molecular Bioscience;* [2]*Institute for Conservation Biology and Environmental Management, School of Biological Sciences, University of Wollongong, NSW 2522; and* [3]*School of Earth, Environmental and Biological Sciences, Queensland University of Technology, QLD 4001, Australia*
**Correspondence to be sent to: Centre for Medical and Molecular Bioscience, School of Biological Sciences, University of Wollongong, NSW 2522, Australia; E-mail: mdowton@uow.edu.au.*

Received 27 February 2014; reviews returned 8 March 2014; accepted 18 March 2014
Associate Editor: Tanja Stadler

Syst. Biol. 63(6):1005–1009, 2014

DOI:10.1093/sysbio/syu060
Advance Access publication August 12, 2014

Known Knowns, Known Unknowns, Unknown Unknowns and Unknown Knowns in DNA Barcoding: A Comment on Dowton et al.

RUPERT A. COLLINS[1*] AND ROBERT H. CRUICKSHANK[2]

[1]*Laboratório de Evolução e Genética Animal, Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil and* [2]*Department of Ecology, Faculty of Agriculture and Life Sciences, Lincoln University, Lincoln 7647, Canterbury, New Zealand*
**Correspondence to be sent to: Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil; E-mail: rupertcollins@gmail.com*

Received 5 June 2014; reviews returned 1 August 2014; accepted 4 August 2014
Associate Editor: Tanya Stadler

.quick MSC barcoding?

- **faster** options
  - guenomu + BPP
  - guenomu with query as species X

- **good species trees** are important
  - for MSC delimitation (BPP)
  - to calibrate the barcode gap for particular groups
  - better reference libraries

.take home

- gene trees *are not* species trees
- best-fit partitioning schemes for multilocus data (**partitiontest**)
  - do not assume the same model for every partition
  - GTR+G per partition works well
- **guenomu** can offer sensible estimates of species trees (and gene trees) from multilocus data
  - gene tree **uncertainty**
  - **multiple individuals** from the same species
  - **non-overlapping species** across gene families
  - **orthologs AND paralogs**
  - **rooted** species trees without outgroup

.open questions

- is ILS relevant for barcoding?
- are species tree relevant for barcoding?
- but how much relevant?

.open questions

- is ILS relevant for barcoding? **… i think so**
- are species trees relevant for barcoding?
- but how much relevant?

.open questions

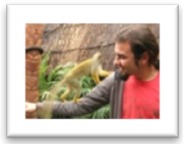
- is ILS relevant for barcoding? **… i think so**
- are species trees relevant for barcoding? **… probably in some cases**
- but how much relevant?

.open questions

- is ILS relevant for barcoding? **… i think so**
- are species trees relevant for barcoding? **… probably in some cases**
- but how much relevant? **… i don't know**

.acknowledgements

- **people**
  - leo martins (> Imperial)
  - diego mallo (> ASU)
  - diego darriba (> Heidelberg)


- **funding**

erc

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

UniversidadeVigo

- **hosts**

6th International Barcode of Life Conference August 18-21, 2015

UNIVERSITY of GUELPH

.thanks

Phylogenomics Lab

David Posada's lab at the University of Vigo

Home Publications Software People News Contact

Welcome to our lab at the University of Vigo, Spain. We have traditionally focused on the development of practical methods for phylogenetic analysis, but more recently we are also pursuing further interests in the deployment of NGS technologies and their application in the phylogeography and population genomics of marine invertebrates, in particular mollusks. Our current projects are described below.

Cancer phylogeography

phylocancer erc

The evolution of cancer tumors in a body can be likened with the evolution of populations in more or less fragmented habitats. During tumor progression, this population of cells is subject to distinct somatic evolutionary processes like mutation, drift, selection or migration, which can act at different points in time and geographical space. So far evolutionary inferences drawn from cancer genomes have been mostly qualitative. We aim to construct a robust theoretical and methodological evolutionary framework that can contribute to a better understanding of the process of somatic evolution and shed light into the biology of cancer.

Phylogenomic estimation of species trees

The estimation of species trees from genomic data is an open problem that goes beyond concatenating many genes and estimating a single tree, and distinct phenomena can explain the disagreement between gene trees and species history. We are currently working on phylogenetic models of genome evolution able to consider lineage sorting, gene duplication and loss and horizontal gene transfer. At the same time, we are developing a practical computational approach for selecting the best partition for multi-gene data sets (i.e., considering genomic heterogeneity) and comparing distinct phylogenomic strategies.

NGS phylogeography of closely related genomes

Our understanding of the mechanisms of evolution at the genomic level is being transformed by the current explosion of massive sequencing of non-model organisms. Together with Rafael Zardoya and colleagues we are using RNA-seq to obtain a large number of homologous loci from a set of marine snail species, a recent radiation of the genus *Trovaoconus* endemic from the Cape Verde islands. With Jesus Troncoso at the University of Vigo we are studying different aspects of transcriptomic evolution and using these data to decipher the role of incomplete lineage sorting and gene duplication on the rapid evolution of related genomes. Currently we are focusing in the genus *Elysia* and *Hypselodoris*.

Mussel genome

We are currently working on the *de novo* genome and transcriptome sequencing of the marine bivalve *Mytilus galloprovincialis*using NGS technologies. Mussel is a very common and commercially important

Search

Search

Recent posts

- ModelTest paper among the top 100 in history! October 30, 2014
- Postdoctoral position in NGS cancer evolution October 27, 2014
- Unsorted Homology within Locus and Species Trees October 14, 2014
- A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction October 6, 2014
- Science paper on HIV-1 hidden history October 3, 2014

calpost

December 2014

M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

« Oct

Phylogenomics papers

- The evolution of tenascins and fibronectin.

Species trees

- Species delimitation in the lichenized fungal genus Vulpicida (Parmeliaceae, Ascomycota) using gene concatenation and coalescent-based species tree approaches.

tagcloud

cape verde conus deep coalescence duplication and loss gene duplication and loss gene family gene tree HIV-1 homology categories horizontal gene transfer jobs lab locus tree models modeltest NGS phylocancer phylogeography postdoc postdocs reconciliation species tree supertree

http://darwin.uvigo.es