nature ecology & evolution

ARTICLES

PUBLISHED: 13 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0020

Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life

Dahiana Arcila[1,2], Guillermo Ortí[1], Richard Vari[2,†], Jonathan W. Armbruster[3], Melanie L. J. Stiassny[4], Kyung D. Ko[1], Mark H. Sabaj[5], John Lundberg[5], Liam J. Revell[6] and Ricardo Betancur-R.[2,7*]

Much progress has been achieved in disentangling evolutionary relationships among species in the tree of life, but some taxonomic groups remain difficult to resolve despite increasing availability of genome-scale data sets. Here we present a practical approach to studying ancient divergences in the face of high levels of conflict, based on explicit gene genealogy interrogation (GGI). We show its efficacy in resolving the controversial relationships within the largest freshwater fish radiation (Otophysi) based on newly generated DNA sequences for 1,051 loci from 225 species. Initial results using a suite of standard methodologies revealed conflicting phylogenetic signal, which supports ten alternative evolutionary histories among early otophysan lineages. By contrast, GGI revealed that the vast majority of gene genealogies supports a single tree topology grounded on morphology that was not obtained by previous molecular studies. We also reanalysed published data sets for exemplary groups with recalcitrant resolution to assess the power of this approach. GGI supports the notion that ctenophores are the earliest-branching animal lineage, and adds insight into relationships within clades of yeasts, birds and mammals. GGI opens up a promising avenue to account for incompatible signals in large data sets and to discern between estimation error and actual biological conflict explaining gene tree discordance.

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

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

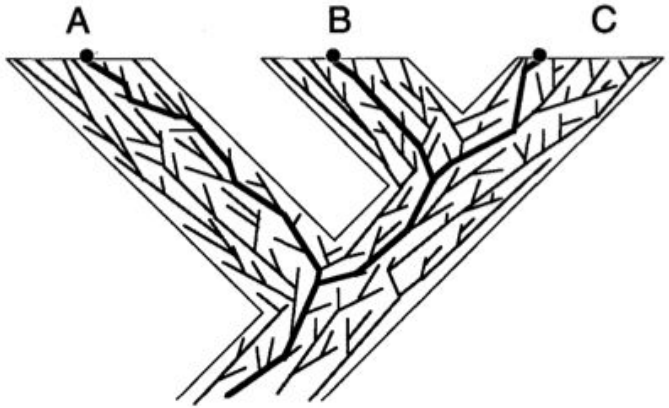


FIGURE 1. A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene tree show relationships among the sampled copies of the gene (●). Sampled copies from sister species B and C are sister copies.

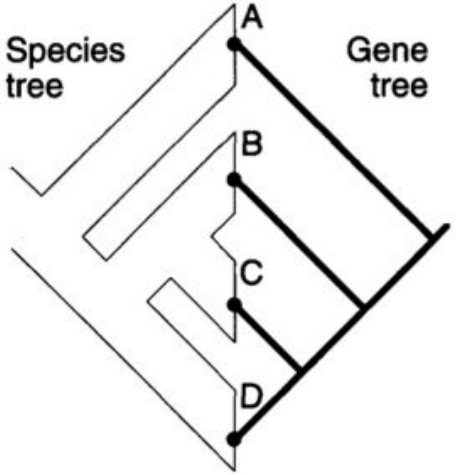


FIGURE 2. Discord between gene and species trees. At left is the species tree of four species, A, B, C, and D, and at right is the tree of a gene sampled one copy per species. Species B and C are sister species, but their gene copies are not sister copies.

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

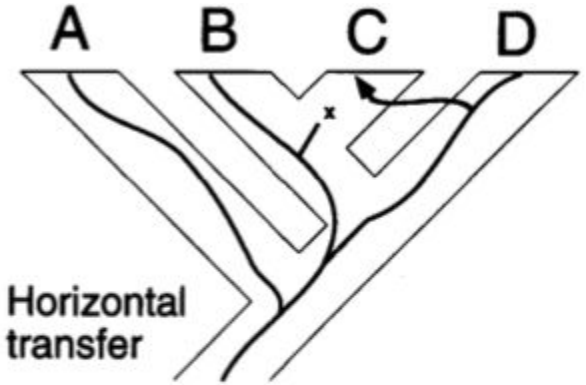


FIGURE 3. Horizontal transfer. A branch of the gene tree jumps between species lineages. If the indigenous gene copy in the receiving species lineage goes extinct or is not sampled (✖), then the gene tree will disagree with the species tree, as shown in Figure 2.

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

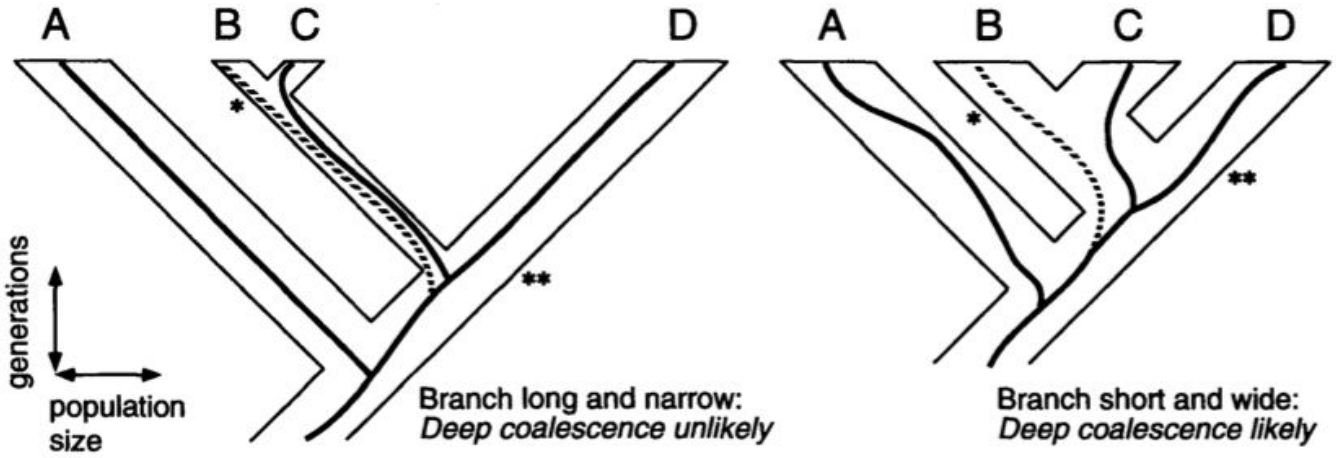


FIGURE 4. Lineage sorting (deep coalescence). Described in a time-forward sense as lineage sorting, an ancestral polymorphism at ** is retained through a lineage to the next speciation event at *, where different forms are sampled in different descendant species. Described in a time-backward sense as deep coalescence, two gene copies from species B and C meet at * but fail to coalesce until deeper than the speciation event at **, at which point the gene from C coalesces first with the gene from D. Failure to coalesce is more likely the shorter (in generations) and wider (in effective population size) the branch is between ** and *.

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

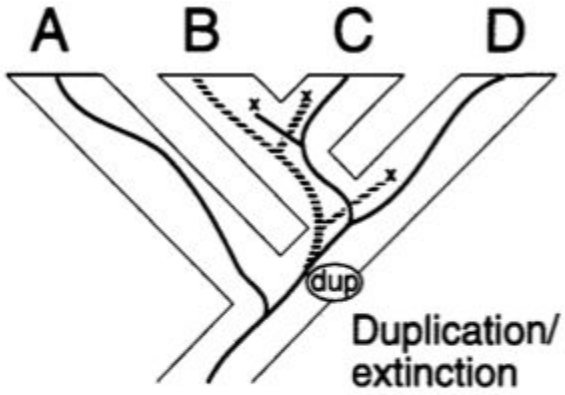


FIGURE 5. Gene duplication and extinction (or paralogous sampling). The gene is duplicated to a different locus, indicated by the dashed lines. If in descendant species one or the other locus goes extinct or is not sampled (✖), then the gene tree will disagree with the species tree, as shown in Figure 2.

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

FIGURE 9. Phylogeny as a cloud of gene histories. Phylogeny is more like a statistical distribution than a simple tree of discrete thin branches. It has a central tendency, but it also has a variance because of the diversity of gene trees. Gene trees that disagree with the central tendency are not wrong; rather, they are part of the diffuse pattern that is the genetic history.

Estimating phylogenetic trees from genome-scale data

Liang Liu,[1,2] Zhenxiang Xi,[3] Shaoyuan Wu,[4] Charles C. Davis,[3] and Scott V. Edwards[3]

Ann. N.Y. Acad. Sci. 1360 (2015) 36–53

Estimating phylogenetic trees from genome-scale data

Liang Liu,[1,2] Zhenxiang Xi,[3] Shaoyuan Wu,[4] Charles C. Davis,[3] and Scott V. Edwards[3]

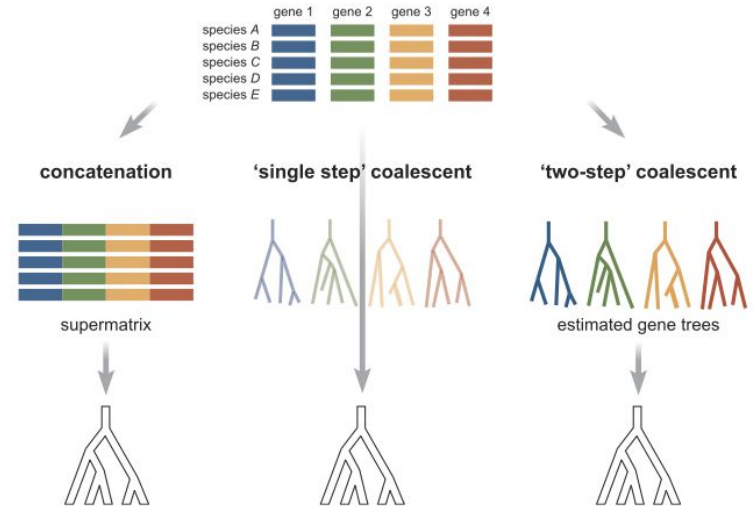


Figure 1. Schematic of the concatenation and coalescent paradigms in phylogenetics. At the top is depicted a multilocus data set consisting of five species (A–E) and four genes (1–4). On the left is indicated the classic supermatrix approach, in which all genes are concatenated to produce a single supergene, which is then subjected to phylogenetic analysis by classical or updated traditional algorithms, such as RAxML,[107] ExaBayes,[84] or MrBayes.[108] Although the resulting tree at lower left is in truth a gene tree, it is often called a species tree or phylogeny because it is the result of analysis of a complete data set. In the center is depicted a class of species tree (coalescent) methods in which both gene trees and species trees are estimated concurrently according to multilocus sequence data, priors, and a multispecies coalesent likelihood model. Examples of algorithms estimating species trees in this way include *BEAST[49] and BEST.[109] On the right are depicted so-called two-step species tree methods, in which gene trees are first estimated using classical approaches and then used as input data to estimate a species tree using algorithms such as MP-EST,[48] STAR,[47] STEM,[87] NJst,[110] STELLS,[88] or ASTRAL.[89] The methods depicted typically used sets of loci each consisting of linked DNA sites.[2] Yet other species tree methods such as SNAPP[90] and quartet inference[92] use unlinked SNPs rather than linked sites, and require different statistical models.

Ann. N.Y. Acad. Sci. 1360 (2015) 36–53

Estimating phylogenetic trees from genome-scale data

Liang Liu,[1,2] Zhenxiang Xi,[3] Shaoyuan Wu,[4] Charles C. Davis,[3] and Scott V. Edwards[3]

Table 1. Studies evaluating the robustness of species tree phylogenetic methods to various genetic forces and sampling schemes

Topic	Conclusions/comments[a]
General violation of multispecies coalescent model[104]	• Claims the majority of multilocus sequence data sets are a poor fit to the multispecies coalescent model, although much of the violation stems from fit of substitution model or unknown sources on a minority of genes.
Gene flow[73,105]	• The coalescent method is robust to low levels of gene flow • Concatenation performs poorly relative to the coalescent methods in the presence of gene flow. • Gene flow can lead to overestimation of population sizes and underestimation of species divergence times in species trees.
Sampling/mutation[33,74,94]	• Increased sampling of individuals per species can significantly improve the estimation of shallow species trees. • Sampling more individuals does not significantly improve accuracy in estimating deep species trees. Adding more loci can improve the estimation of deep relationships. • Mutational variance is a major source of error in estimates of species trees.

Estimating phylogenetic trees from genome-scale data

Liang Liu,[1,2] Zhenxiang Xi,[3] Shaoyuan Wu,[4] Charles C. Davis,[3] and Scott V. Edwards[3]

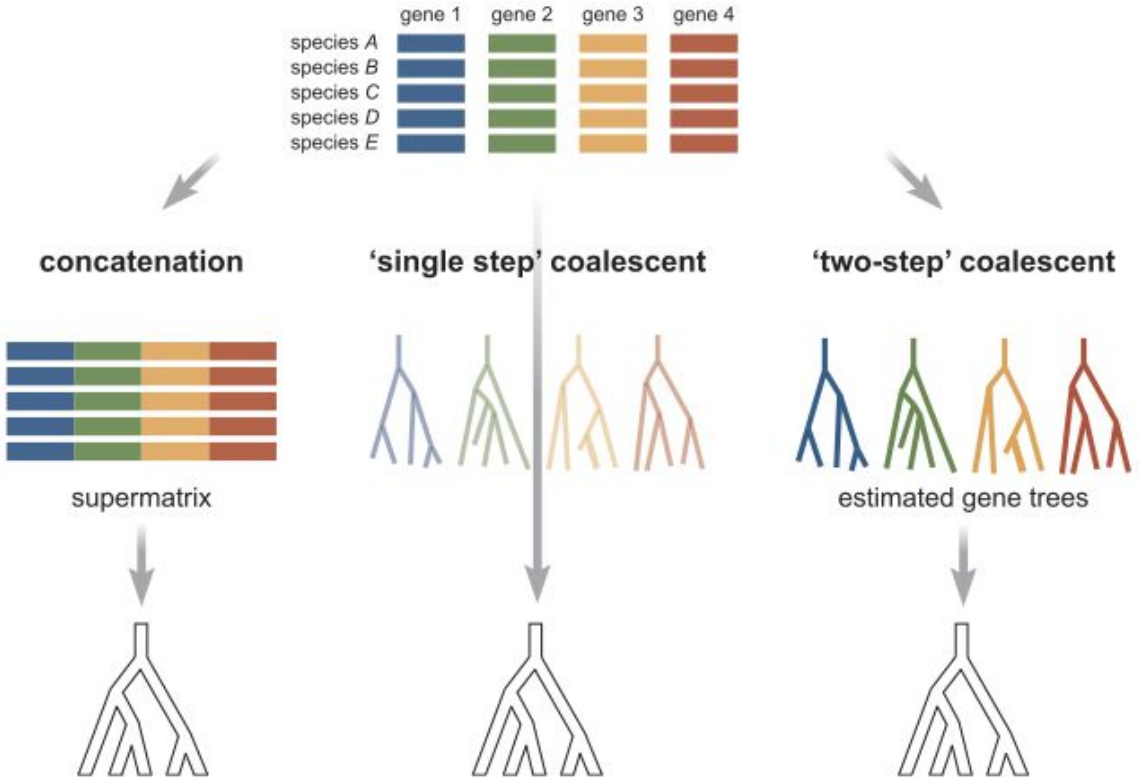
Table 1. Studies evaluating the robustness of species tree phylogenetic methods to various genetic forces and sampling schemes

Topic	Conclusions/comments[a]
Recombination[39,106]	• Recombination has minor effect on species tree estimation except on extremely short species trees. • The negative effects of recombination can be easily overcome by increased sampling of alleles
Missing data[19,63,67]	• Missing data can decrease the support of species tree estimates • Missing data can significantly affect the accuracy of species tree estimation • Species tree methods are "remarkably resilient" to missing data[68]
Taxon sampling[15]	• Compared to concatenation, coalescent methods are more robust to poor taxon sampling
Long-branch attraction[29]	• Species tree methods are more resilient to the effects of long-branch attraction than concatenation methods
Random rooting of gene trees[36,37]	• Misrooting of gene trees can mimic the coalescent process
Other[80]	• Anomalous gene trees are unlikely to pose a significant danger to empirical phylogenetic study, in part because species trees in the anomaly zone are likely to be rare.

Ann. N.Y. Acad. Sci. 1360 (2015) 36–53

Estimating phylogenetic trees from genome-scale data

Liang Liu,[1,2] Zhenxiang Xi,[3] Shaoyuan Wu,[4] Charles C. Davis,[3] and Scott V. Edwards[3]



Ann. N.Y. Acad. Sci. 1360 (2015) 36–53

Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life

Dahiana Arcila[1,2], Guillermo Ortí[1], Richard Vari[2,†], Jonathan W. Armbruster[3], Melanie L. J. Stiassny[4], Kyung D. Ko[1], Mark H. Sabaj[5], John Lundberg[5], Liam J. Revell[6] and Ricardo Betancur-R.[2,7*]

A major assumption of these 'summary'[13] or 'short-cut'[14] coalescent methods is that individual gene trees accurately depict the genealogical history of fragments of the genome that independently segregate (coalescent genes, or c-genes).

Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life

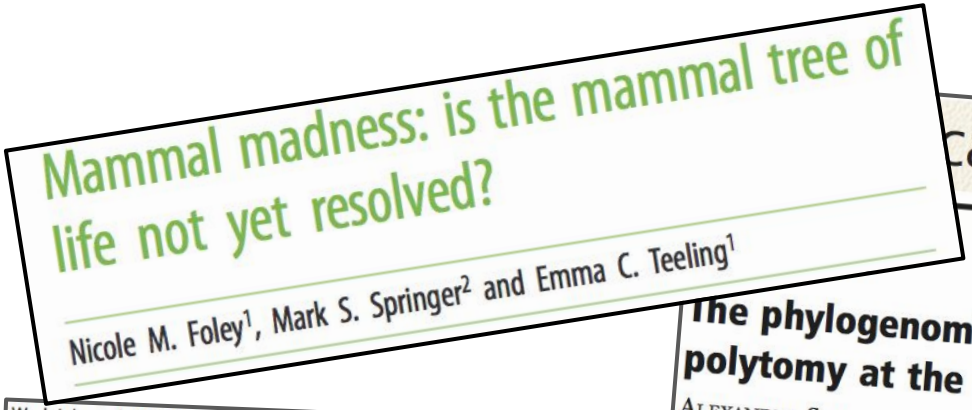
Dahiana Arcila[1,2], Guillermo Ortí[1], Richard Vari[2,†], Jonathan W. Armbruster[3], Melanie L. J. Stiassny[4], Kyung D. Ko[1], Mark H. Sabaj[5], John Lundberg[5], Liam J. Revell[6] and Ricardo Betancur-R.[2,7*]

the analysis of short, recombination-free genes (consisting of a few hundred sites) are error-prone due to limited signal-to-noise content[2,3,14–17]. On the other extreme, long genes or full-length transcripts with thousands of sites harbor more phylogenetic information, reducing (but not necessarily removing) stochastic error[2,18]. Longer genes, however, are more likely to carry past recombination events, violating the assumption of a single genealogical history[8]. Both situations lead to statistical inconsistency under the multi-species coalescent,

Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life

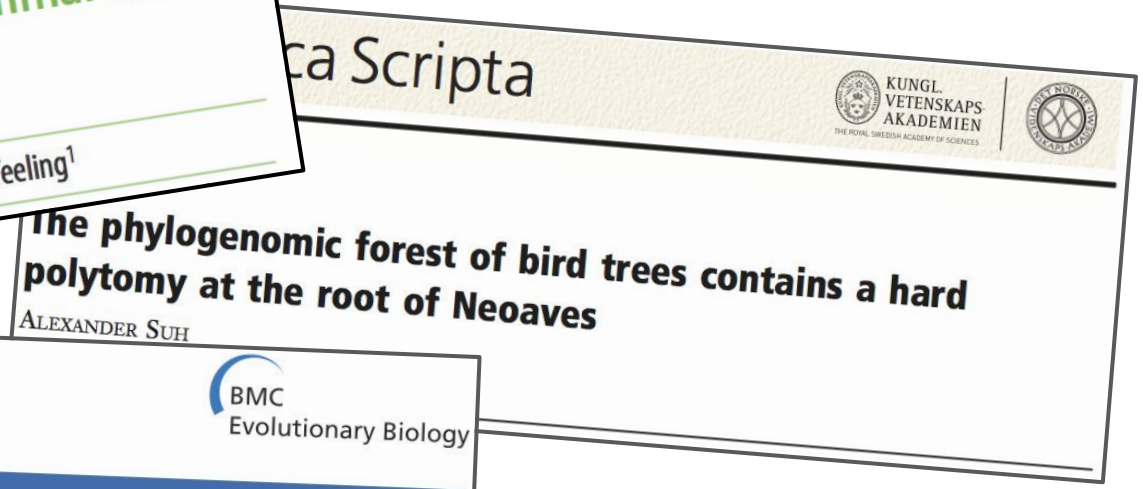
Dahiana Arcila[1,2], Guillermo Ortí[1], Richard Vari[2,†], Jonathan W. Armbruster[3], Melanie L. J. Stiassny[4], Kyung D. Ko[1], Mark H. Sabaj[5], John Lundberg[5], Liam J. Revell[6] and Ricardo Betancur-R.[2,7*]

Here, we present a phylogenomic approach that efficiently extracts the genealogical signal from short c-genes by reducing the complexity of tree space on the basis of topological constraints. This method is similar to others that place priors on gene tree topologies[25], but is unique in that priors are set to test specific hypotheses directly.



Mammal madness: is the mammal tree of life not yet resolved?

Nicole M. Foley[1], Mark S. Springer[2] and Emma C. Teeling[1]



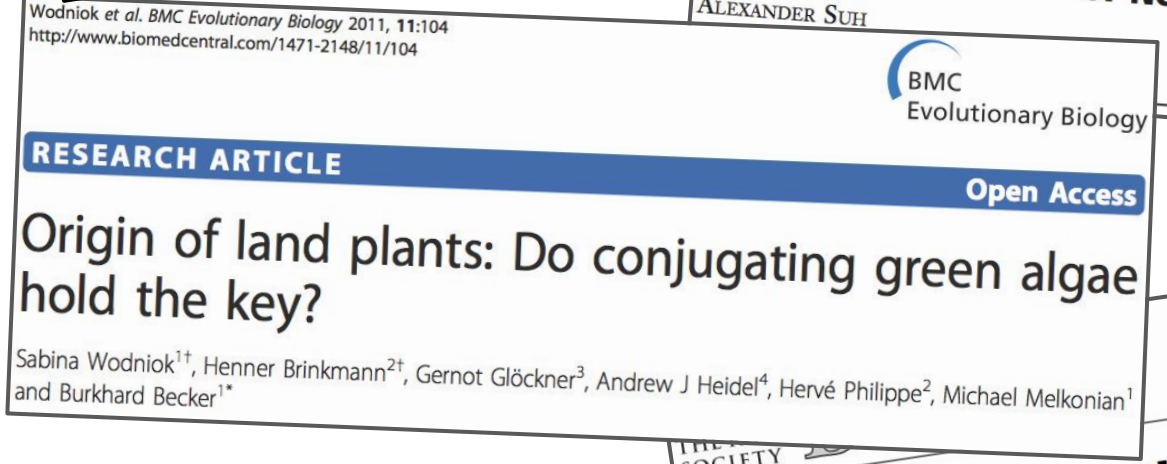
ca Scripta

KUNGL. VETENSKAPS-AKADEMIEN

THE ROYAL SWEDISH ACADEMY OF SCIENCES

The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves

ALEXANDER SUH



Wodniok *et al. BMC Evolutionary Biology* 2011, **11**:104
http://www.biomedcentral.com/1471-2148/11/104

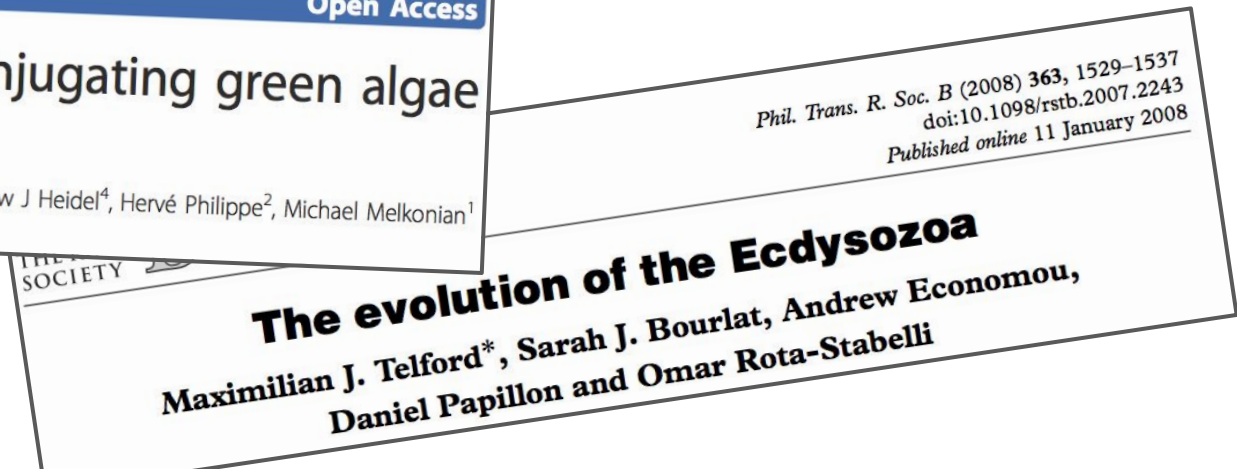
BMC Evolutionary Biology

RESEARCH ARTICLE

Open Access

Origin of land plants: Do conjugating green algae hold the key?

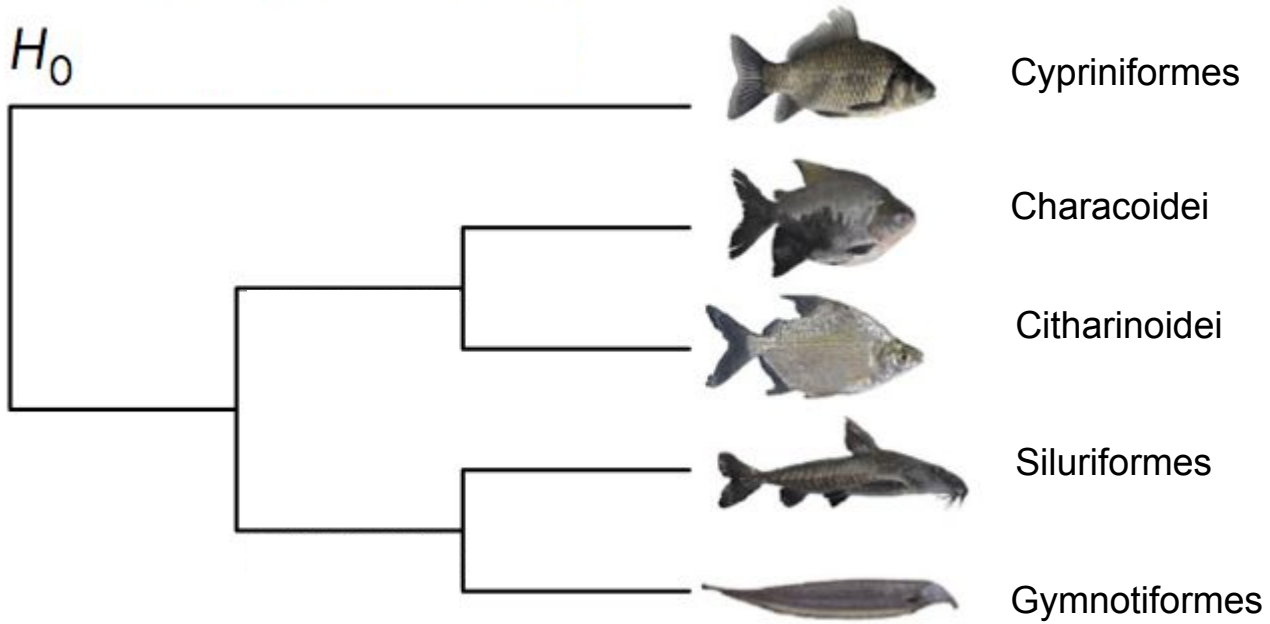
Sabina Wodniok[1†], Henner Brinkmann[2†], Gernot Glöckner[3], Andrew J Heidel[4], Hervé Philippe[2], Michael Melkonian[1] and Burkhard Becker[1*]

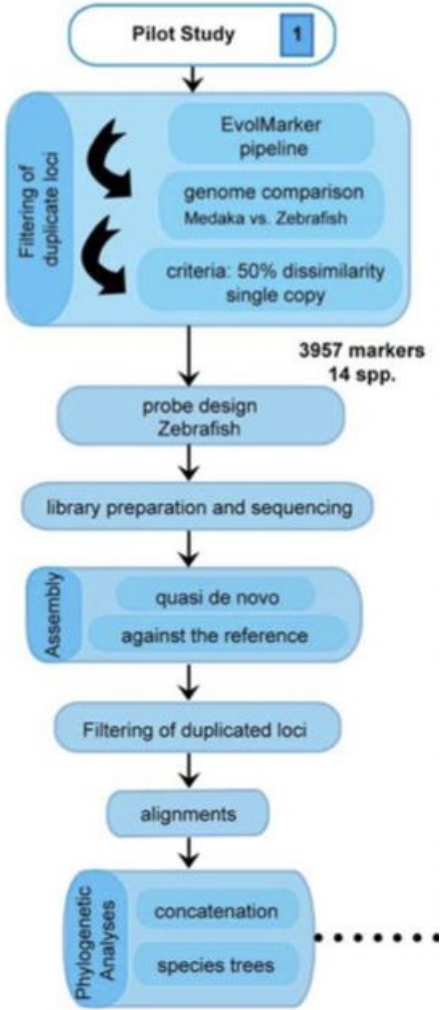


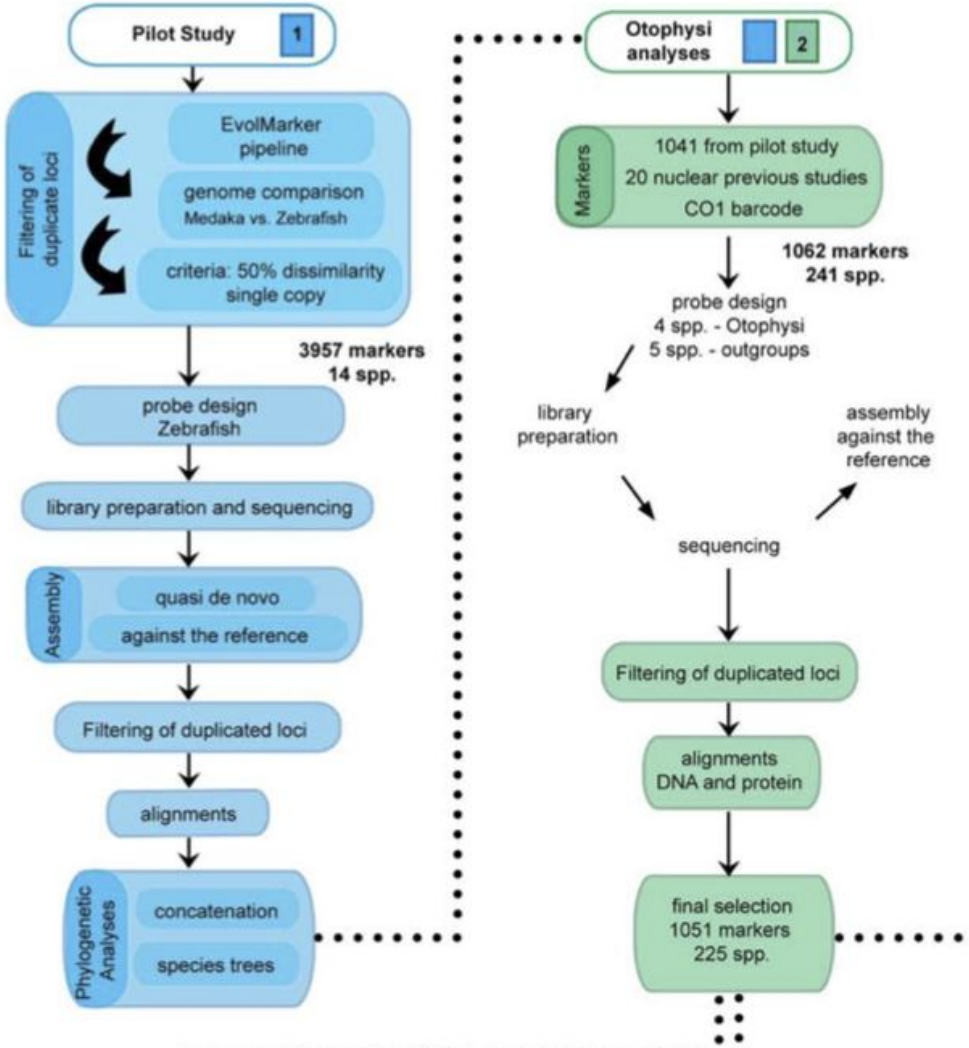
Phil. Trans. R. Soc. B (2008) **363**, 1529–1537
doi:10.1098/rstb.2007.2243
Published online 11 January 2008

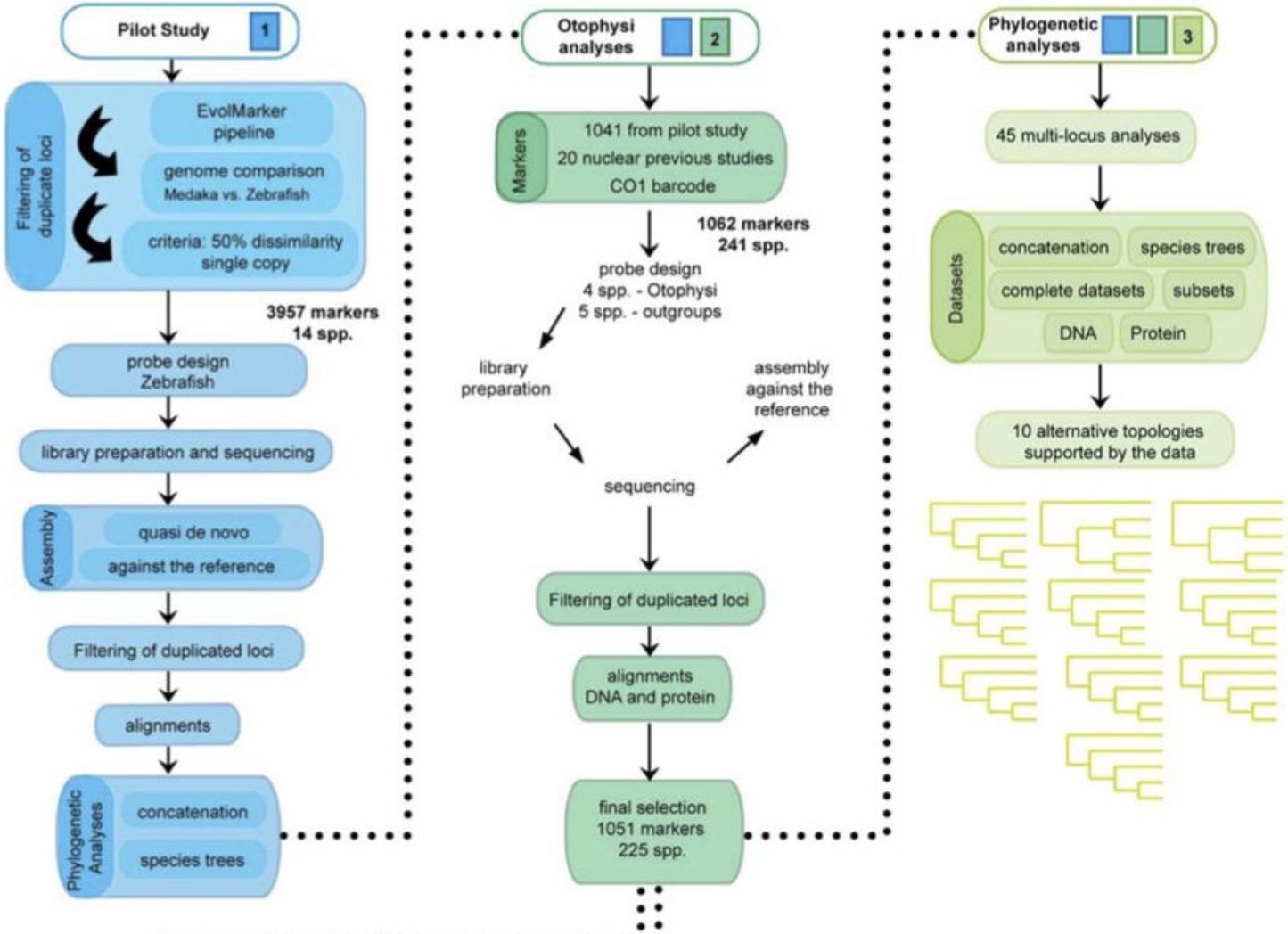
The evolution of the Ecdysozoa

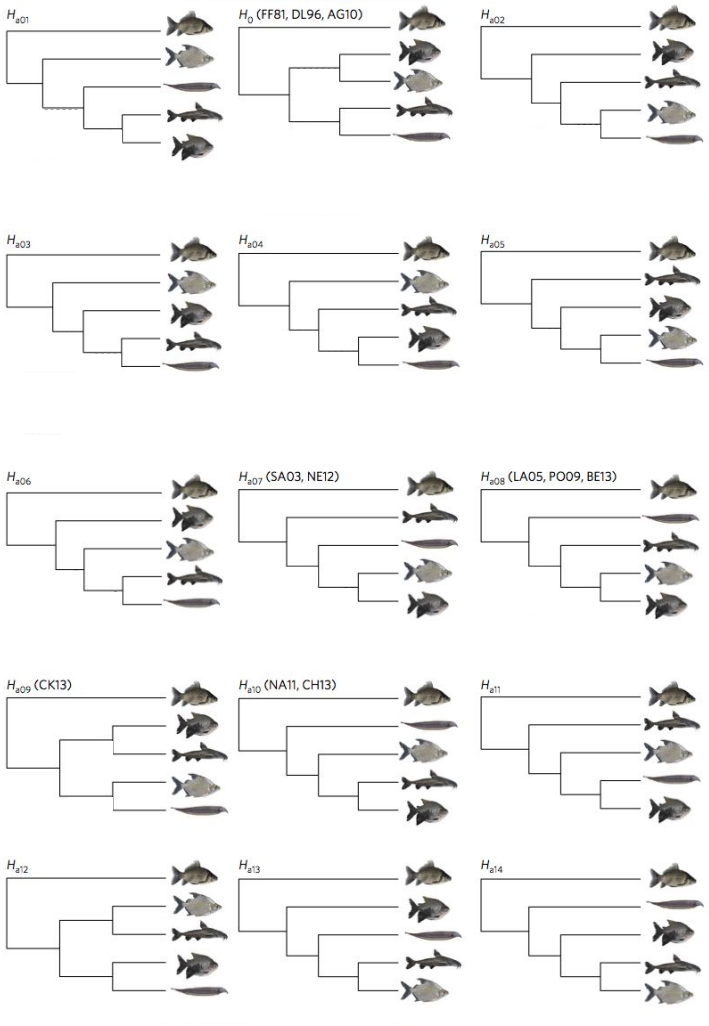
Maximilian J. Telford*, Sarah J. Bourlat, Andrew Economou, Daniel Papillon and Omar Rota-Stabelli

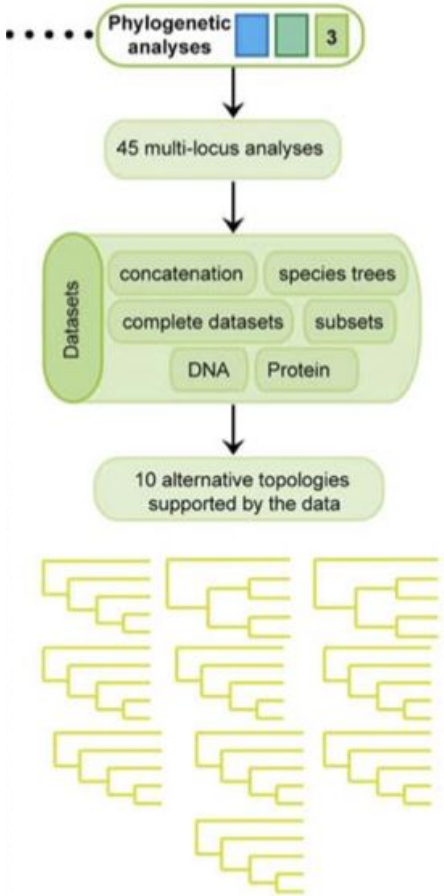

H_0
Cypriniformes
Characoidei
Citharinoidei
Siluriformes
Gymnotiformes


Pilot Study
1
Filtering of duplicate loci
EvolMarker pipeline
genome comparison
Medaka vs. Zebrafish
criteria: 50% dissimilarity
single copy
3957 markers
14 spp.
probe design
Zebrafish
library preparation and sequencing
Assembly
quasi de novo
against the reference
Filtering of duplicated loci
alignments
Phylogenetic Analyses
concatenation
species trees


Pilot Study
1
Filtering of duplicate loci
EvolMarker pipeline
genome comparison
Medaka vs. Zebrafish
criteria: 50% dissimilarity
single copy
3957 markers
14 spp.
probe design
Zebrafish
library preparation and sequencing
Assembly
quasi de novo
against the reference
Filtering of duplicated loci
alignments
Phylogenetic Analyses
concatenation
species trees
Otophysi analyses
2
Markers
1041 from pilot study
20 nuclear previous studies
CO1 barcode
1062 markers
241 spp.
probe design
4 spp. - Otophysi
5 spp. - outgroups
library preparation
assembly against the reference
sequencing
Filtering of duplicated loci
alignments
DNA and protein
final selection
1051 markers
225 spp.


Pilot Study
1
Filtering of duplicate loci
EvolMarker pipeline
genome comparison
Medaka vs. Zebrafish
criteria: 50% dissimilarity single copy
3957 markers
14 spp.
probe design Zebrafish
library preparation and sequencing
Assembly
quasi de novo
against the reference
Filtering of duplicated loci
alignments
Phylogenetic Analyses
concatenation
species trees
Otophysi analyses
2
Markers
1041 from pilot study
20 nuclear previous studies
CO1 barcode
1062 markers
241 spp.
probe design
4 spp. - Otophysi
5 spp. - outgroups
library preparation
assembly against the reference
sequencing
Filtering of duplicated loci
alignments DNA and protein
final selection
1051 markers
225 spp.
Phylogenetic analyses
3
45 multi-locus analyses
Datasets
concatenation
species trees
complete datasets
subsets
DNA
Protein
10 alternative topologies supported by the data


H_{a01}
H_0 (FF81, DL96, AG10)
H_{a02}
H_{a03}
H_{a04}
H_{a05}
H_{a06}
H_{a07} (SA03, NE12)
H_{a08} (LA05, PO09, BE13)
H_{a09} (CK13)
H_{a10} (NA11, CH13)
H_{a11}
H_{a12}
H_{a13}
H_{a14}


Phylogenetic analyses
3
45 multi-locus analyses
Datasets
concatenation
species trees
complete datasets
subsets
DNA
Protein
10 alternative topologies supported by the data

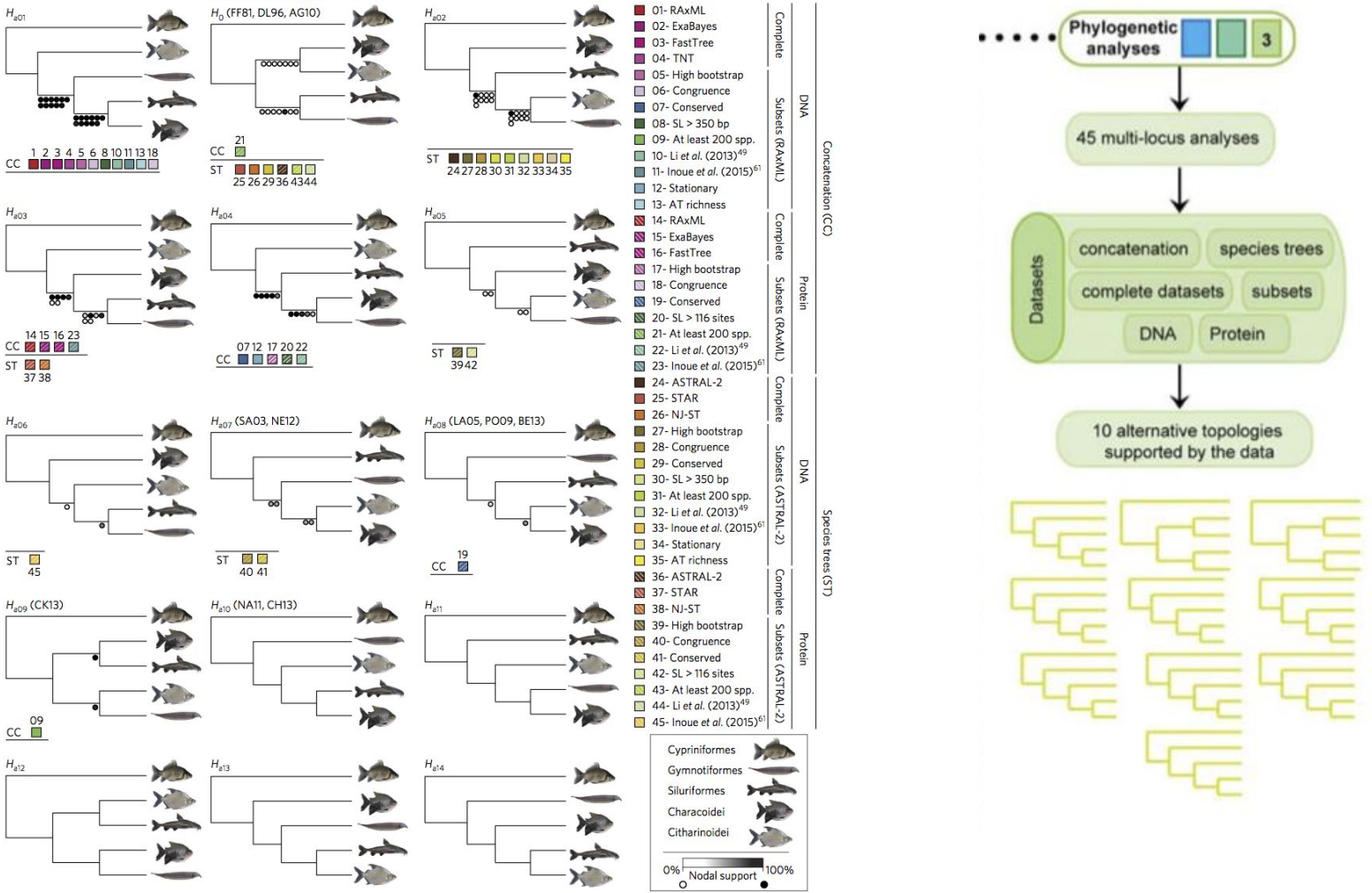
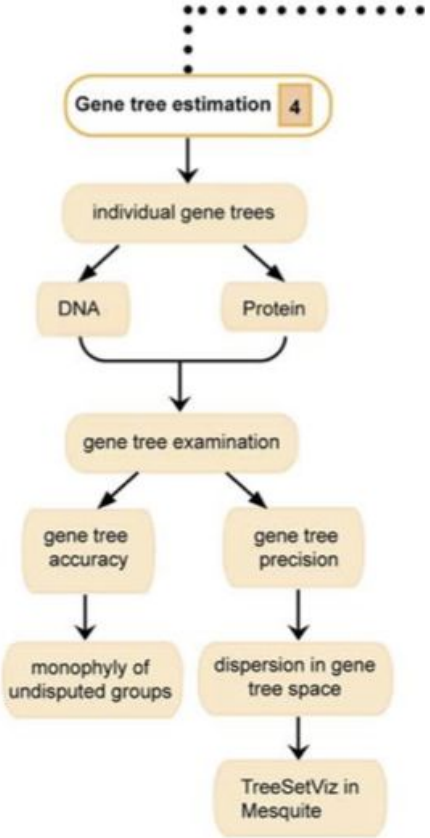


Figure 1 | Null morphological hypothesis (H_0) and all 14 possible alternative trees for the five major lineages in Otophysi. Previous studies supporting each



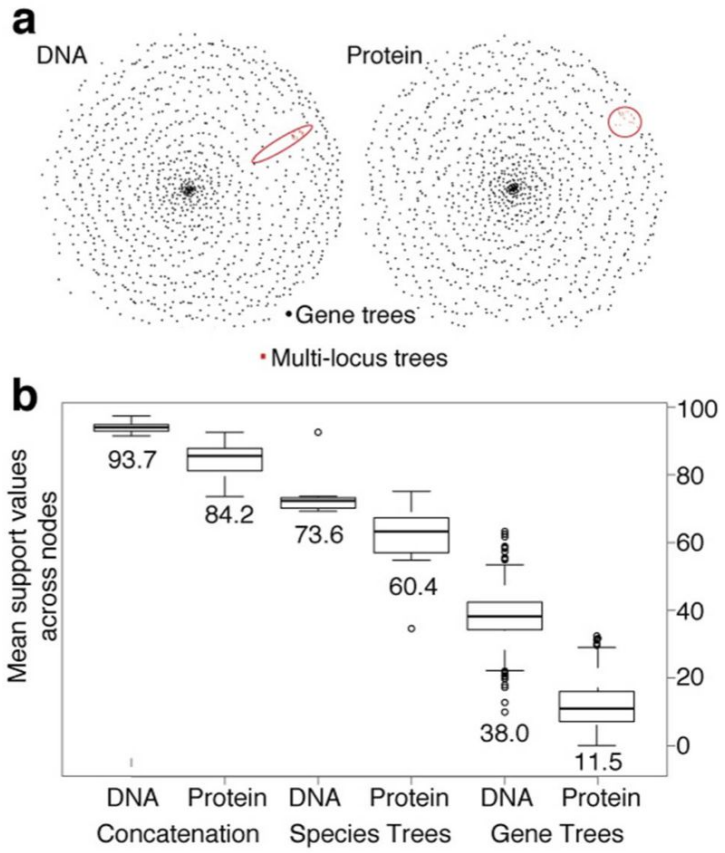
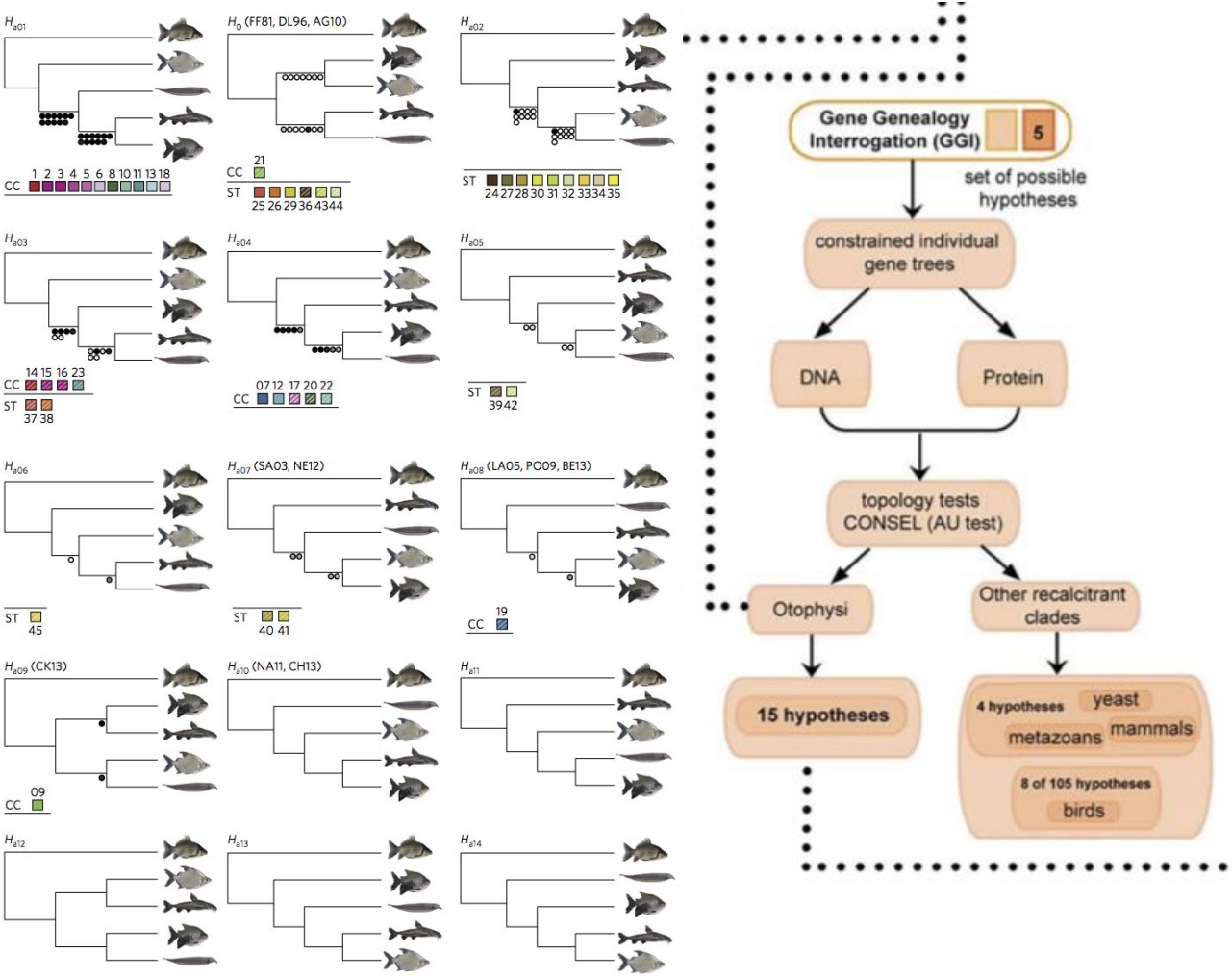
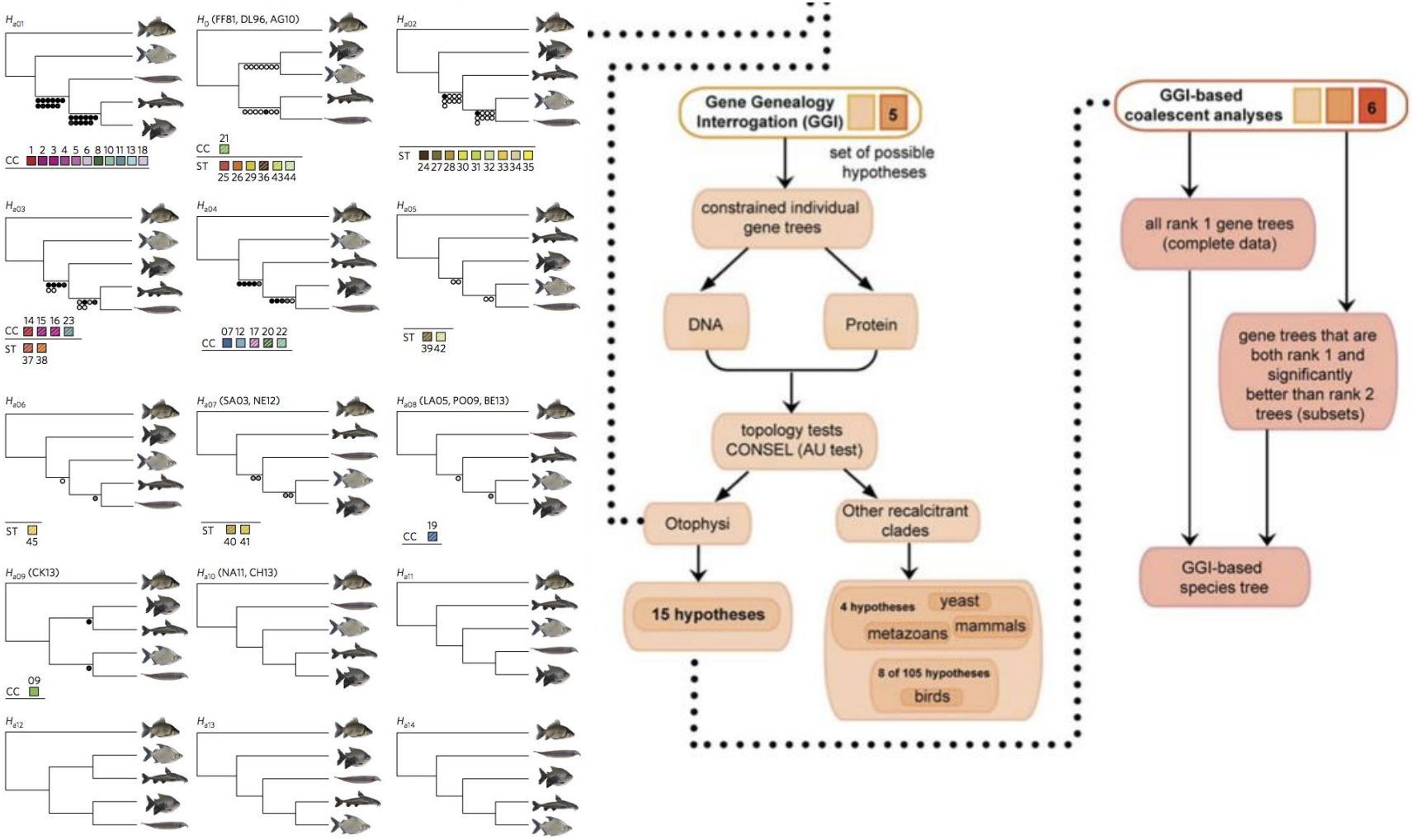
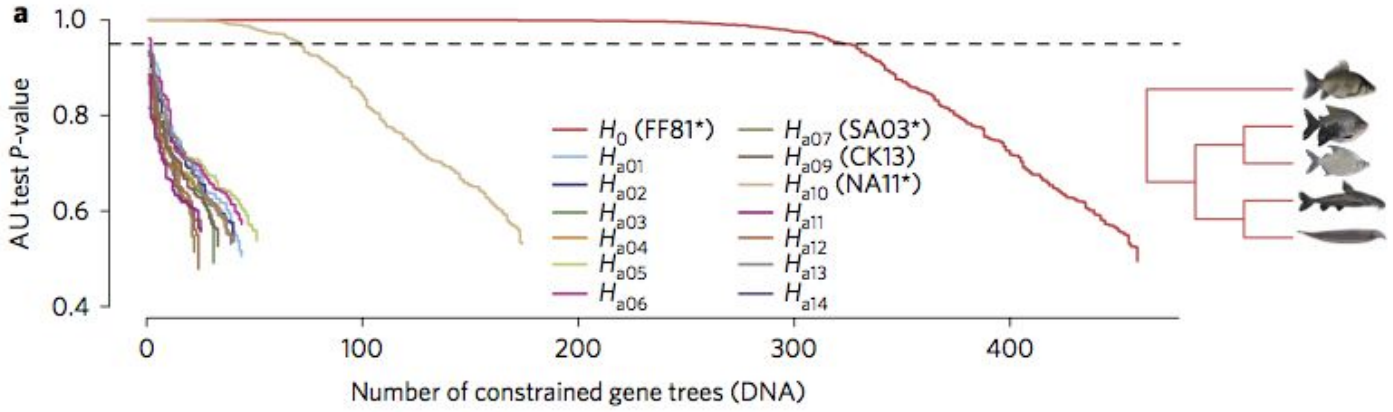
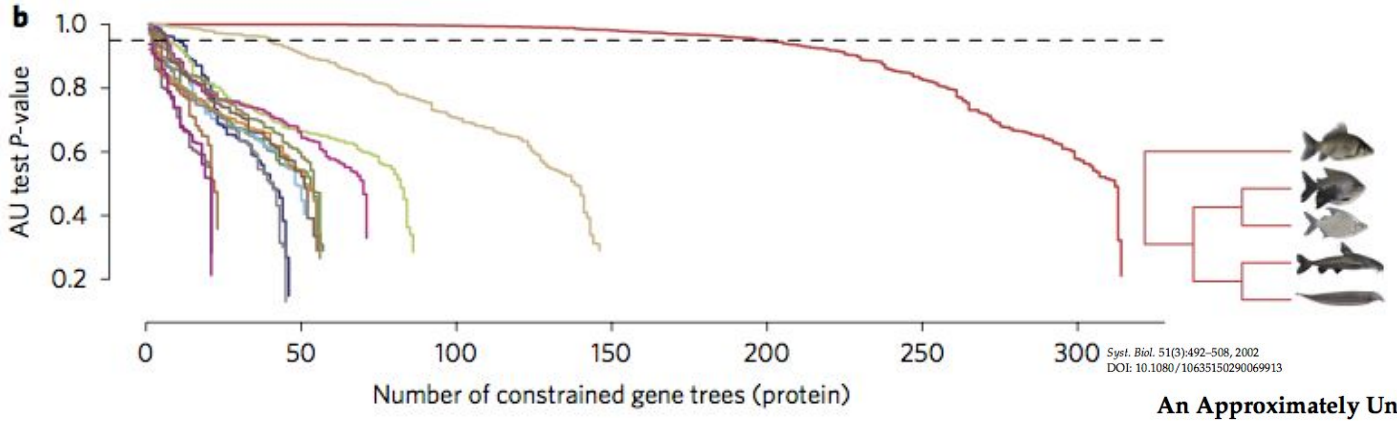


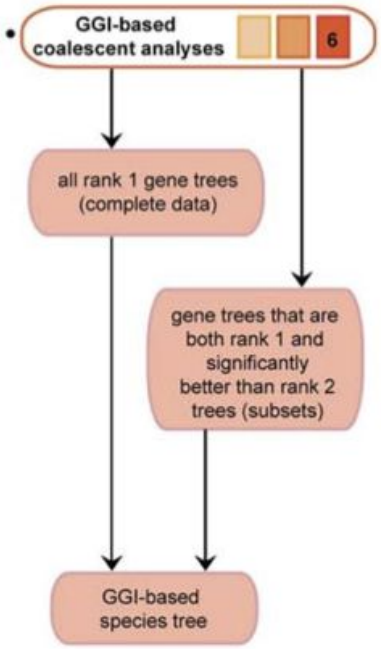
Figure S2. Assessments of phylogenetic precision for multi-locus trees and individual gene trees | a, dispersion in multidimensional tree space based on un-weighted Robinson-Foulds distances; **b,** plots of mean support values across all nodes for each tree obtained with different methods (ExaBayes: posterior probabilities; all other analyses: bootstrap values).


H_{a01}
CC 1 2 3 4 5 6 8 10 11 13 18
H_{0} (FF81, DL96, AG10)
CC 21
ST 25 26 29 36 43 44
H_{a02}
ST 24 27 28 30 31 32 33 34 35
H_{a03}
CC 14 15 16 23
ST 37 38
H_{a04}
CC 07 12 17 20 22
H_{a05}
ST 39 42
H_{a06}
ST 45
H_{a07} (SA03, NE12)
ST 40 41
H_{a08} (LA05, PO09, BE13)
CC 19
H_{a09} (CK13)
CC 09
H_{a10} (NA11, CH13)
H_{a11}
H_{a12}
H_{a13}
H_{a14}
Gene Genealogy Interrogation (GGI)
5
set of possible hypotheses
constrained individual gene trees
DNA
Protein
topology tests CONSEL (AU test)
Otophysi
Other recalcitrant clades
15 hypotheses
4 hypotheses
yeast
metazoans
mammals
8 of 105 hypotheses
birds


Gene Genealogy Interrogation (GGI)
5
set of possible hypotheses
constrained individual gene trees
DNA
Protein
topology tests CONSEL (AU test)
Otophysi
Other recalcitrant clades
15 hypotheses
4 hypotheses
yeast
metazoans
mammals
8 of 105 hypotheses
birds
GGI-based coalescent analyses
6
all rank 1 gene trees (complete data)
gene trees that are both rank 1 and significantly better than rank 2 trees (subsets)
GGI-based species tree





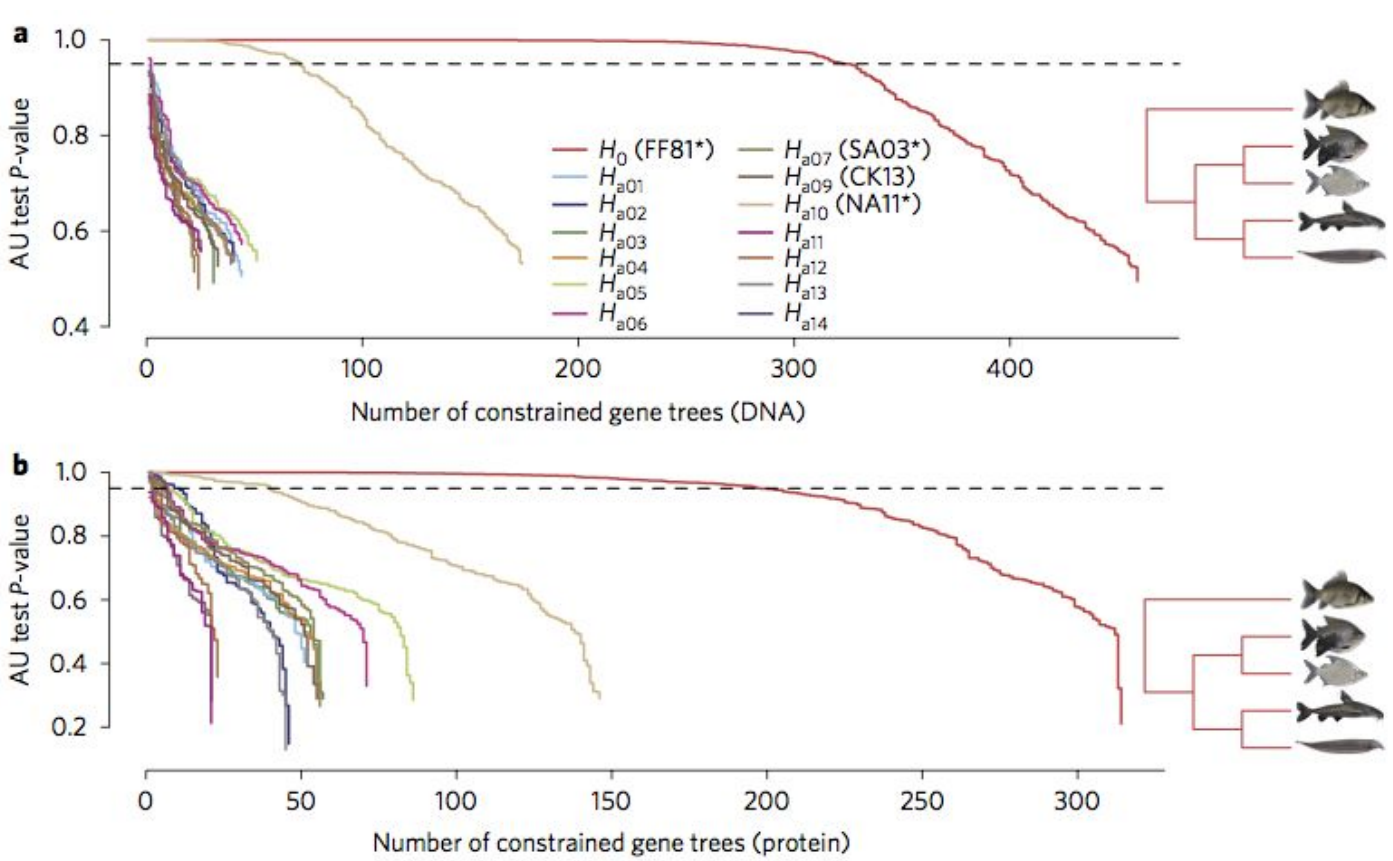


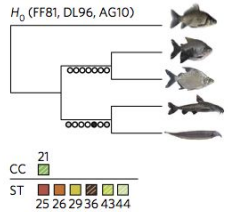
Syst. Biol. 51(3):492–508, 2002
DOI: 10.1080/10635150290069913

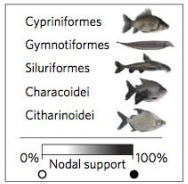
An Approximately Unbiased Test of Phylogenetic Tree Selection

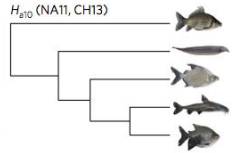
HIDETOSHI SHIMODAIRA

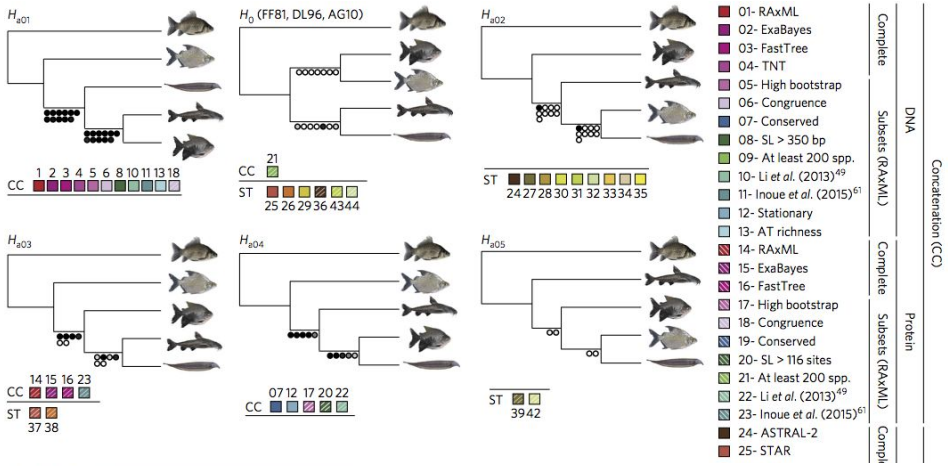
Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minatoku, Tokyo 106–8569, Japan;
E-mail: shimo@ism.ac.jp

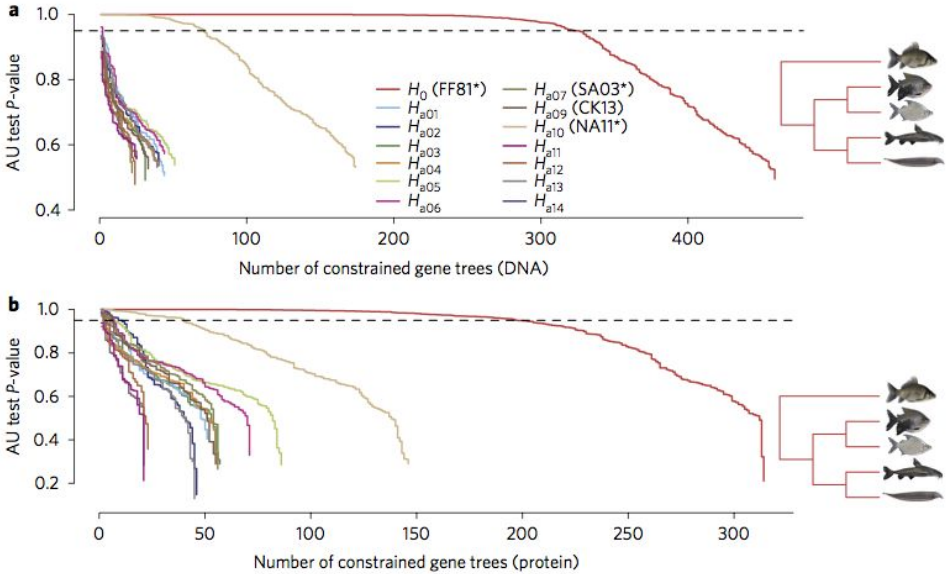

a
AU test P-value
1.0
0.8
0.6
0.4
0
100
200
300
400
Number of constrained gene trees (DNA)
H_0 (FF81*)
H_{a01}
H_{a02}
H_{a03}
H_{a04}
H_{a05}
H_{a06}
H_{a07} (SA03*)
H_{a09} (CK13)
H_{a10} (NA11*)
H_{a11}
H_{a12}
H_{a13}
H_{a14}
b
AU test P-value
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Number of constrained gene trees (protein)

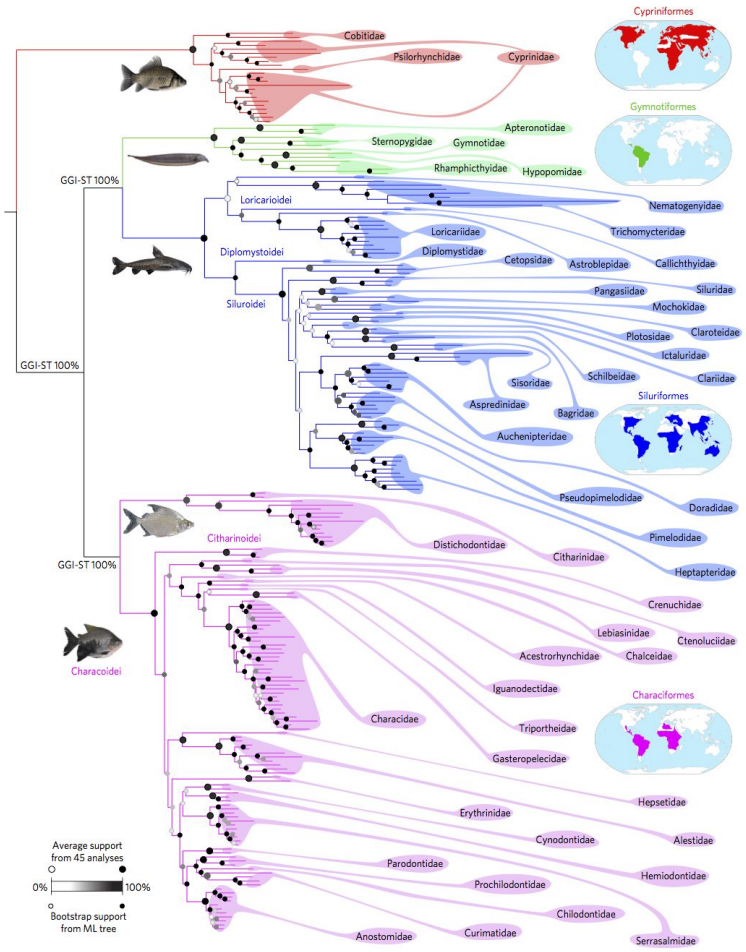

H_0 (FF81, DL96, AG10)
CC 21
ST 25 26 29 36 43 44


Cypriniformes
Gymnotiformes
Siluriformes
Characoidei
Citharinoidei
0%
Nodal support
100%


H_{a10} (NA11, CH13)







ure 3 | Otophysan phylogeny based on the concatenation analysis of 231 protein genes with minimal missing information. This preferred topology id branch lengths) is based on RAxML analysis 21, which is largely congruent with the GGI-based species tree topology obtained with STAR and TRAL-2 (Supplementary Fig. 3). Support values are based on the average obtained from 45 alternative analyses (large circles), the bootstrap support :ained from the ML tree (small circles), or the bootstrap values of the GGI-based species tree analyses using ASTRAL-2 (denoted as GGI-ST; key nodes ted only). Marine otophysan lineages are not depicted in the maps.

ILS is not the main problem

Coalescent theory predicts that phylogenetic histories of lineages evolving under a combination of short internal branches and large effective population sizes are prone to high incidence of ILS[8]. It has been demonstrated that for five or more lineages such conditions can generate gene trees with topologies that differ from the underlying species phylogeny with highest probability[36,37]. When the evolutionary history of a clade falls within this so-called anomaly zone[8], simply adopting the most frequent gene tree as a surrogate for the species phylogeny (the democratic vote procedure) is positively misleading.

To account for this possibility, as the genuine backbone tree of inter-ordinal relationships must be 1 of 15 possibilities (enumeration of all possibilities for an unrooted tree of five taxa), we used the GGI trees selected by the topology tests (the preferred constrained gene trees optimized by ML) as input for summary coalescent analyses. For this test we employed both DNA- and protein-based trees in combination with two different species-tree methods. We also applied two alternative approaches for sampling GGI trees, one using all rank 1 trees (complete data with 1,051 genes) and another using only the set of rank 1 trees that are significantly better than the alternatives ($P < 0.05$; a subset of 397 DNA trees and 275 protein trees; Supplementary Table 3). Of the eight species-tree analyses conducted, all converged on the H_0 tree, with each backbone node receiving 100% bootstrap support. Finally, an adapted version of the GGI-based coalescent method that uses constrained topologies in combination with unconstrained gene trees also supports the H_0 tree (Supplementary Information).

Our results suggest that the evolutionary history of major otophysan lineages is not trapped in the anomaly zone. In fact, these analyses identify only a minor proportion of gene trees that are significantly discordant with the inferred species phylogeny (17.7–28.4%, most supporting H_{a10}), suggesting that other sources of error rather than ILS are likely the main cause of incongruence. Gene tree estimation error may be biasing summary coalescent approaches, but the causes for discrepancy between coalescent and concatenation results are unclear. For two hypotheses (H_0 and H_{a03}), some concatenation and species tree methods converge, but more often they seem to produce non-overlapping sets of results (Fig. 1). We were unable to isolate any single factor as the principal explanation for discordance in multi-locus analysis. Possibilities include the combination of slight model misspecifications interacting in analyses of large data sets and amplifying systematic biases, or processes such as horizontal gene transfer or duplication/extinction affecting some of the sampled genes[38]. What is perhaps most surprising is the observation that the most common topology from concatenation is incongruent with our GGI tree, even in the absence of evidence for substantial ILS. An investigation of factors that could account for this pattern would be a fruitful subject of future theoretical and analytical studies. In summary, the coalescent analyses using GGI trees resolve with high confidence the branching order of major otophysan groups (Supplementary Fig. 3), a result that is fully congruent with the morphological hypothesis (H_0)[27], thereby reconciling a long history of molecular and morphological conflict.

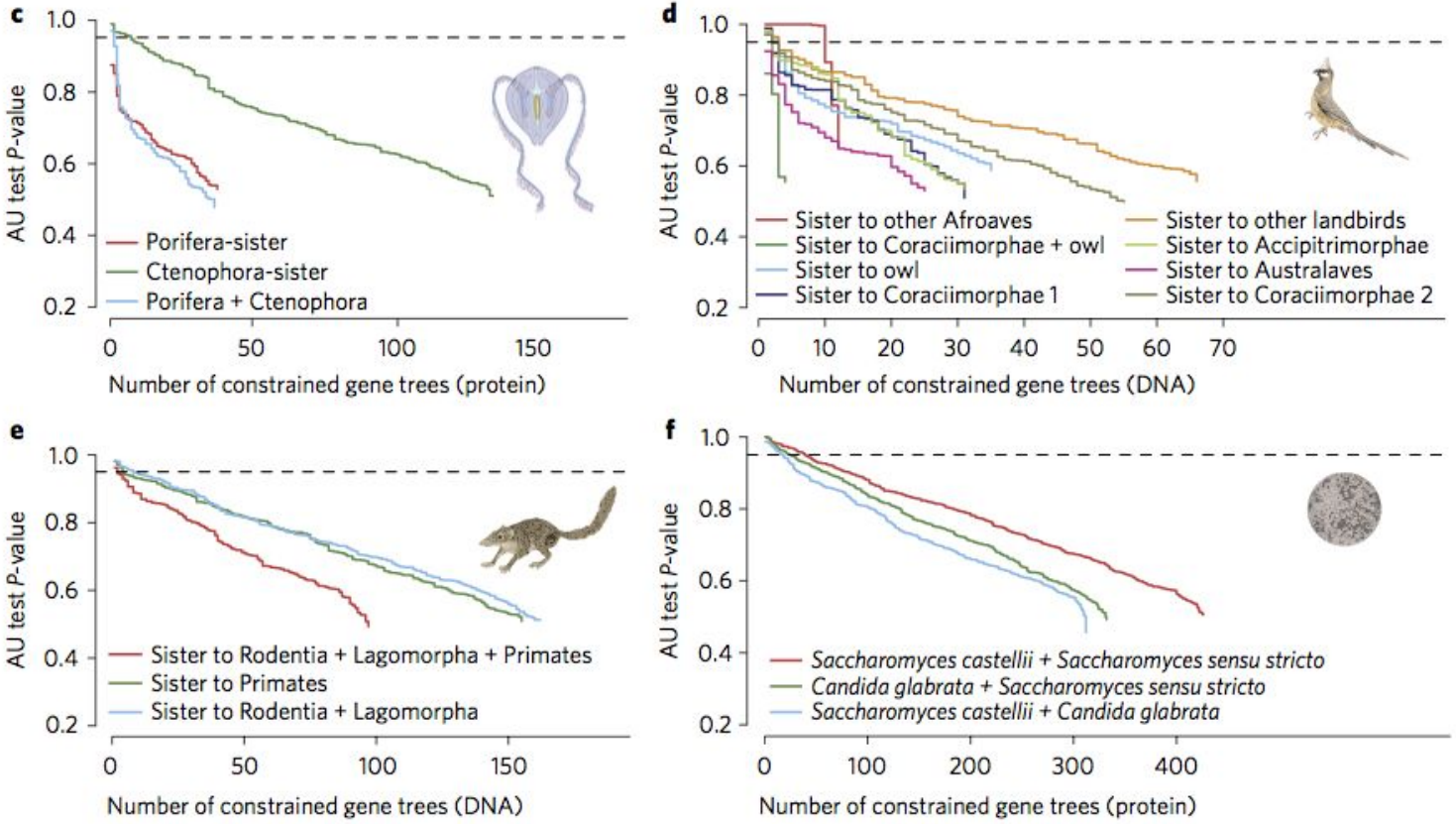


Figure 2 | Gene genealogy interrogation (GGI) applied to phylogenomic data sets to test alternative hypotheses. Lines represent the cumulative number of genes (*x* axes) supporting each hypothesis with highest probability (rank 1) and their associated *P*-values (*y* axes) according to the approximately unbiased (AU) topology test. Values above the dashed line indicate all rank 1 hypotheses that are significantly better than the alternatives ($P < 0.05$), whereas those below the dashed line are also rank 1 but without statistical significance. **a**,**b**, Otophysi (*see Supplementary Table 1). **c**, Metazoans. **d**, Neoaves (mousebird). **e**, Eutherian mammals (tree shrew). **f**, Yeast phylogeny.

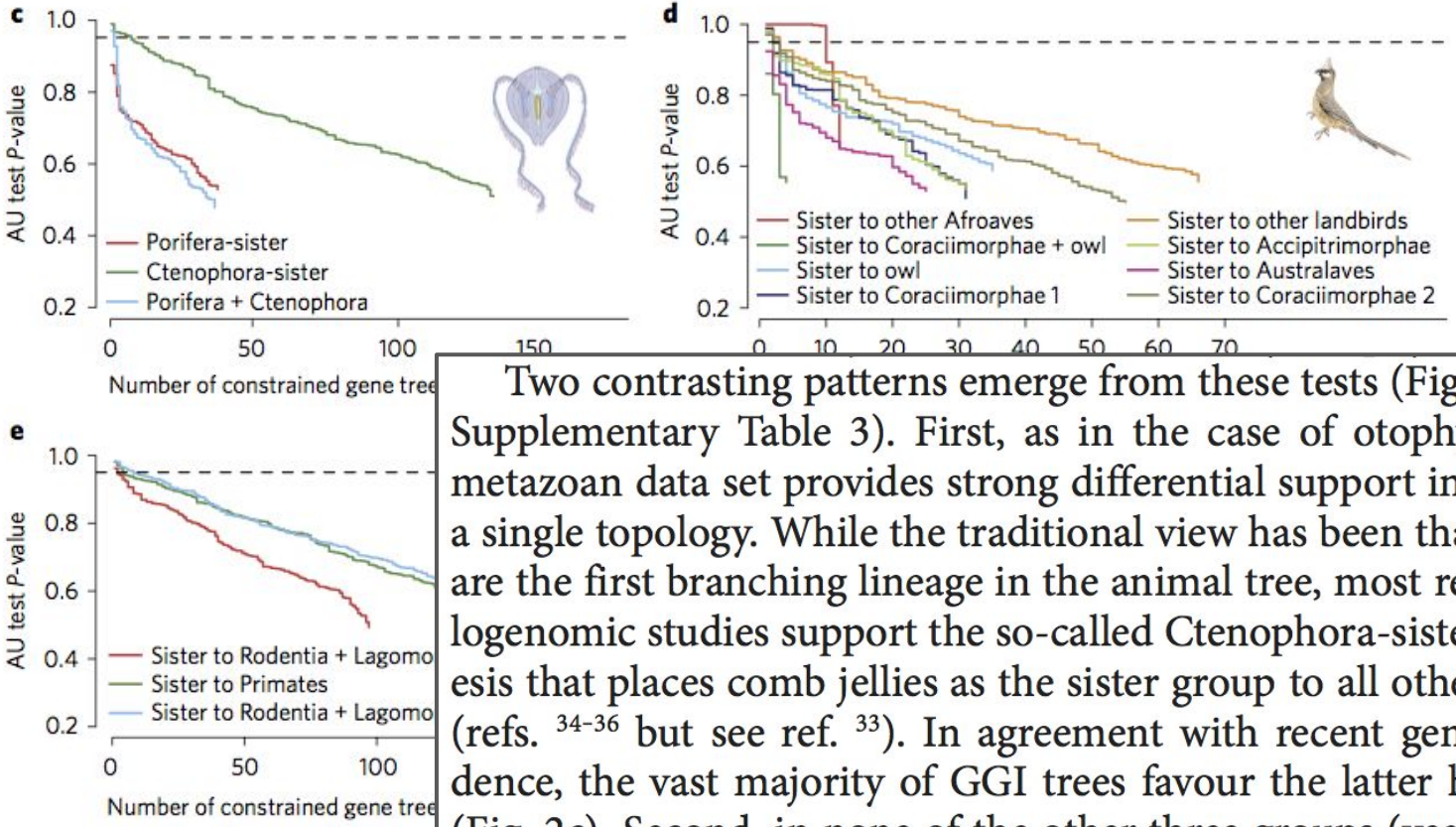


Figure 2 | Gene genealogy interrogation (GGI) appl
of genes (*x* axes) supporting each hypothesis with h
unbiased (AU) topology test. Values above the dash
whereas those below the dashed line are also rank 1
d, Neoaves (mousebird). **e**, Eutherian mammals (tree shrew). **f**, Yeast phylogeny.

Two contrasting patterns emerge from these tests (Fig. 2c–f and Supplementary Table 3). First, as in the case of otophysans, the metazoan data set provides strong differential support in favour of a single topology. While the traditional view has been that sponges are the first branching lineage in the animal tree, most recent phylogenomic studies support the so-called Ctenophora-sister hypothesis that places comb jellies as the sister group to all other animals (refs. [34–36] but see ref. [33]). In agreement with recent genomic evidence, the vast majority of GGI trees favour the latter hypothesis (Fig. 2c). Second, in none of the other three groups (yeasts, mammals and Neoaves) does GGI select a particular tree topology over another with overwhelming support (Fig. 2d–f), indicating that genealogical conflict in these groups is substantial.


gettyimages
lishenjun
165608670