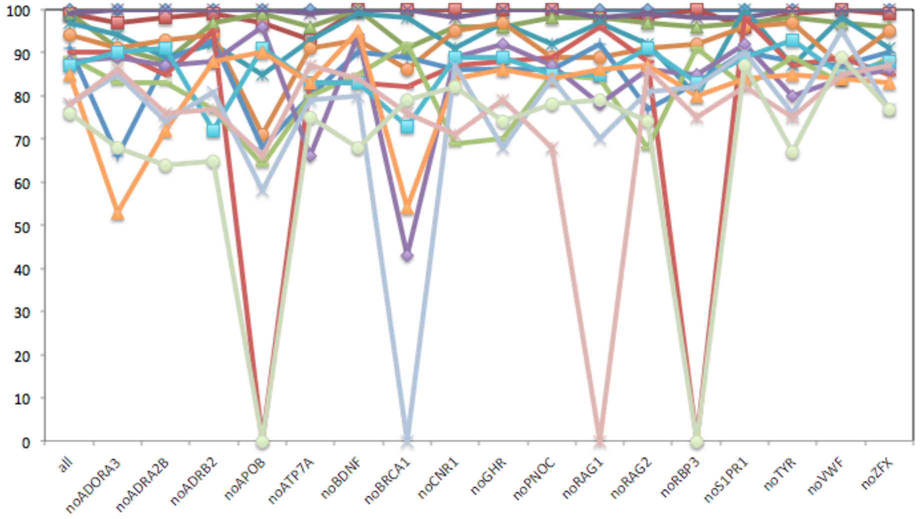
MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY

June 15-19, 2014

France: Hameau de l'Etoile

Exploring Topological Incongruence for Detecting Contaminations in Phylogenomic Data Sets

Frédéric Delsuc, Khalid Belkhir & Celine Scornavacca



isem

Institut des Sciences de l'Evolution-Montpellier

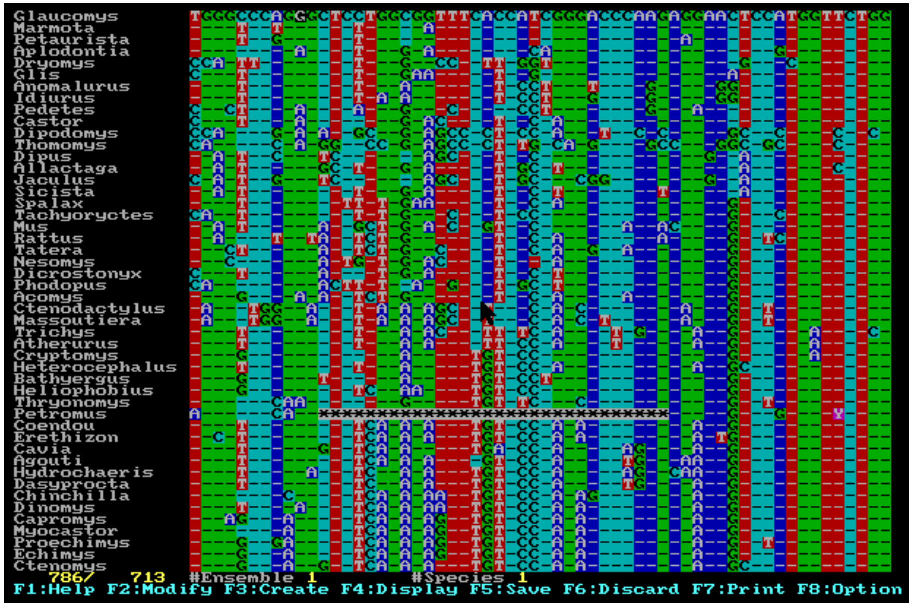
cnrs dépasser les frontières

Institut des Sciences de l'Evolution - UMR 5554 - CNRS - IRD
Université Montpellier 2 - France

um2 UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES

Phylogenetic Reconstruction in Practice

1. Organisms



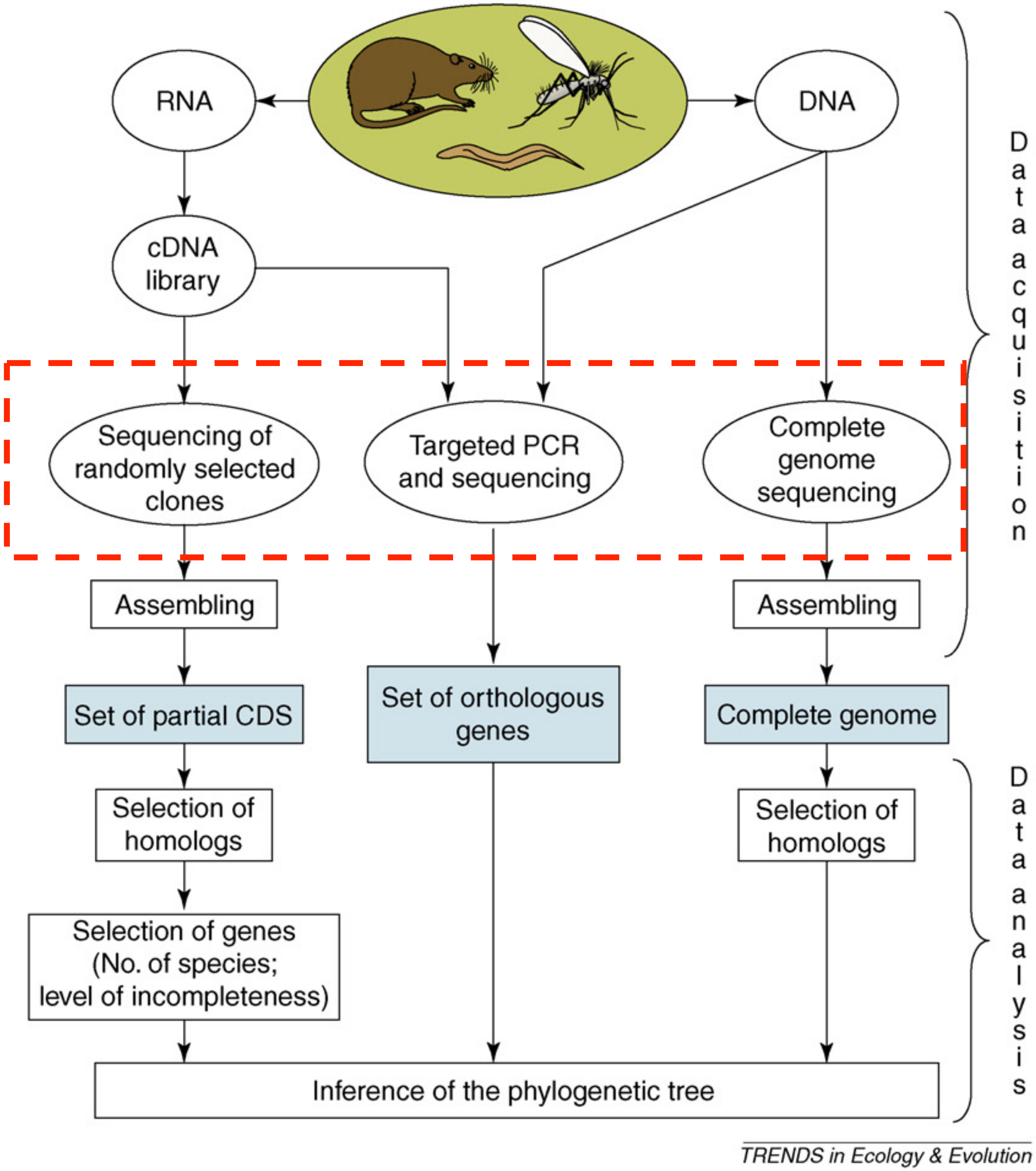
```
Glaucomys         TGGGCCCAGGGCTCCTGGCGGTTTCACCATCGGGACCCAAGAGGAACTCCATGGTTCTGG
Marmota
Petaurista
Aplodontia
Dryomys
Glis
Anomalurus
Idiurus
Pedetes
Castor
Dipodomys
Thomomys
Dipus
Allactaga
Jaculus
Sicista
Spalax
Tachyoryctes
Mus
Rattus
Tatera
Nesomys
Dicrostonyx
Phodopus
Acomys
Ctenodactylus
Massoutiera
Trichys
Atherurus
Cryptomys
Heterocephalus
Bathyergus
Heliophobius
Thryonomys
Petromus
Coendou
Erethizon
Cavia
Agouti
Hydrochaeris
Dasyprocta
Chinchilla
Dinomys
Capromys
Myocastor
Proechimys
Echimys
Ctenomys
   786/    713  #Ensemble 1          #Species  1
F1:Help F2:Modify F3:Create F4:Display F5:Save F6:Discard F7:Print F8:Option
```

2. Homologous characters

3. Reconstruction methods

Data Sources in Phylogenomics

NGS



RNA
DNA
cDNA library
Sequencing of randomly selected clones
Targeted PCR and sequencing
Complete genome sequencing
Assembling
Assembling
Set of partial CDS
Set of orthologous genes
Complete genome
Selection of homologs
Selection of homologs
Selection of genes (No. of species; level of incompleteness)
Inference of the phylogenetic tree
Data acquisition
Data analysis
TRENDS in Ecology & Evolution

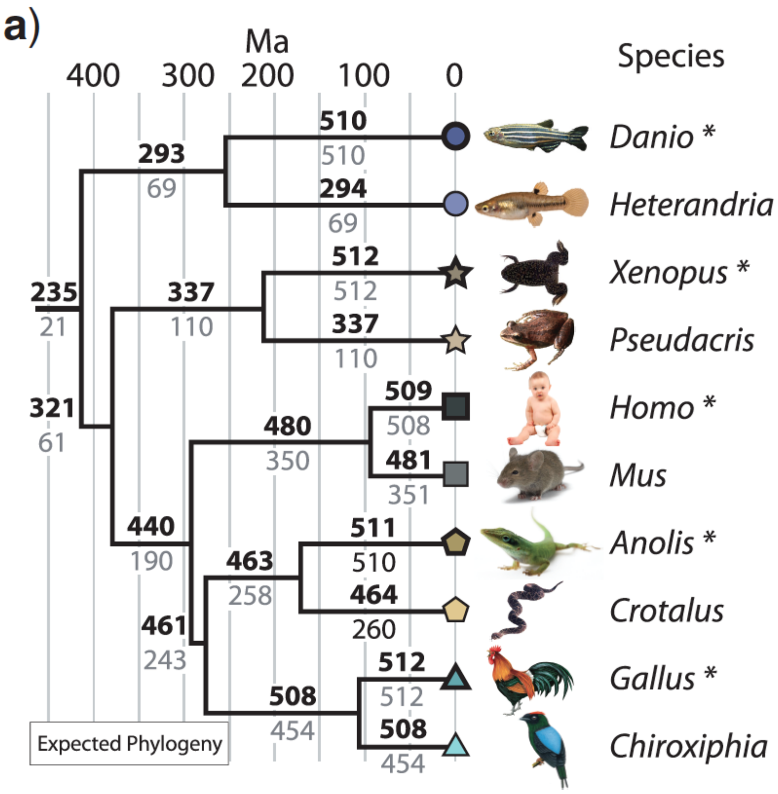
Philippe & Telford (2006) *Trends Ecol. Evol.*

Sequence Capture Methods for Phylogenomics

CENTER FOR ANCHORED PHYLOGENOMICS
ACCELERATING THE RESOLUTION OF LIFE™

http://anchoredphylogeny.com/

512 nuclear loci
for vertebrates

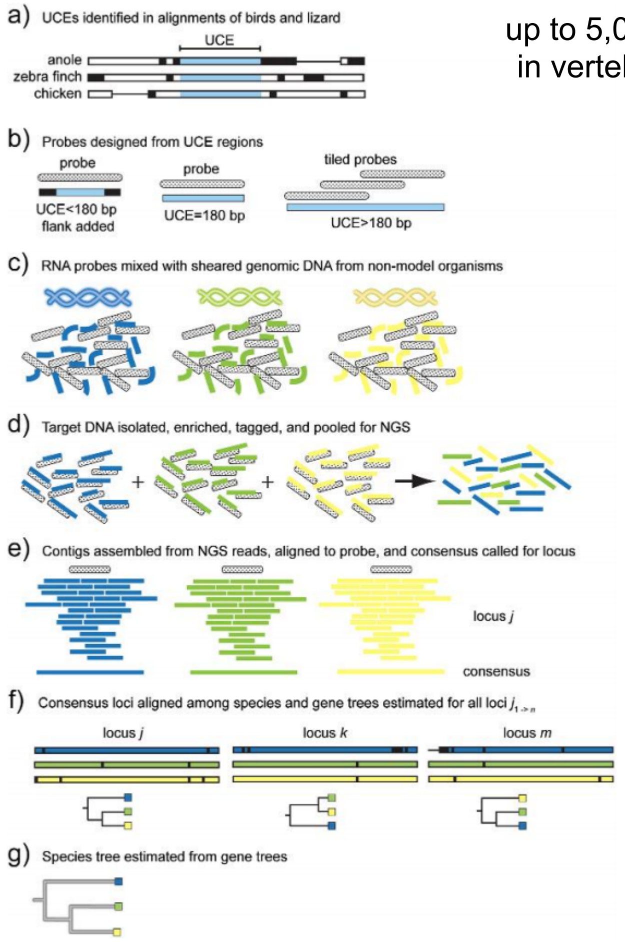


Lemmon *et al.* (2012) *Syst. Biol.*

Ultra Conserved Elements

http://ultraconserved.org/

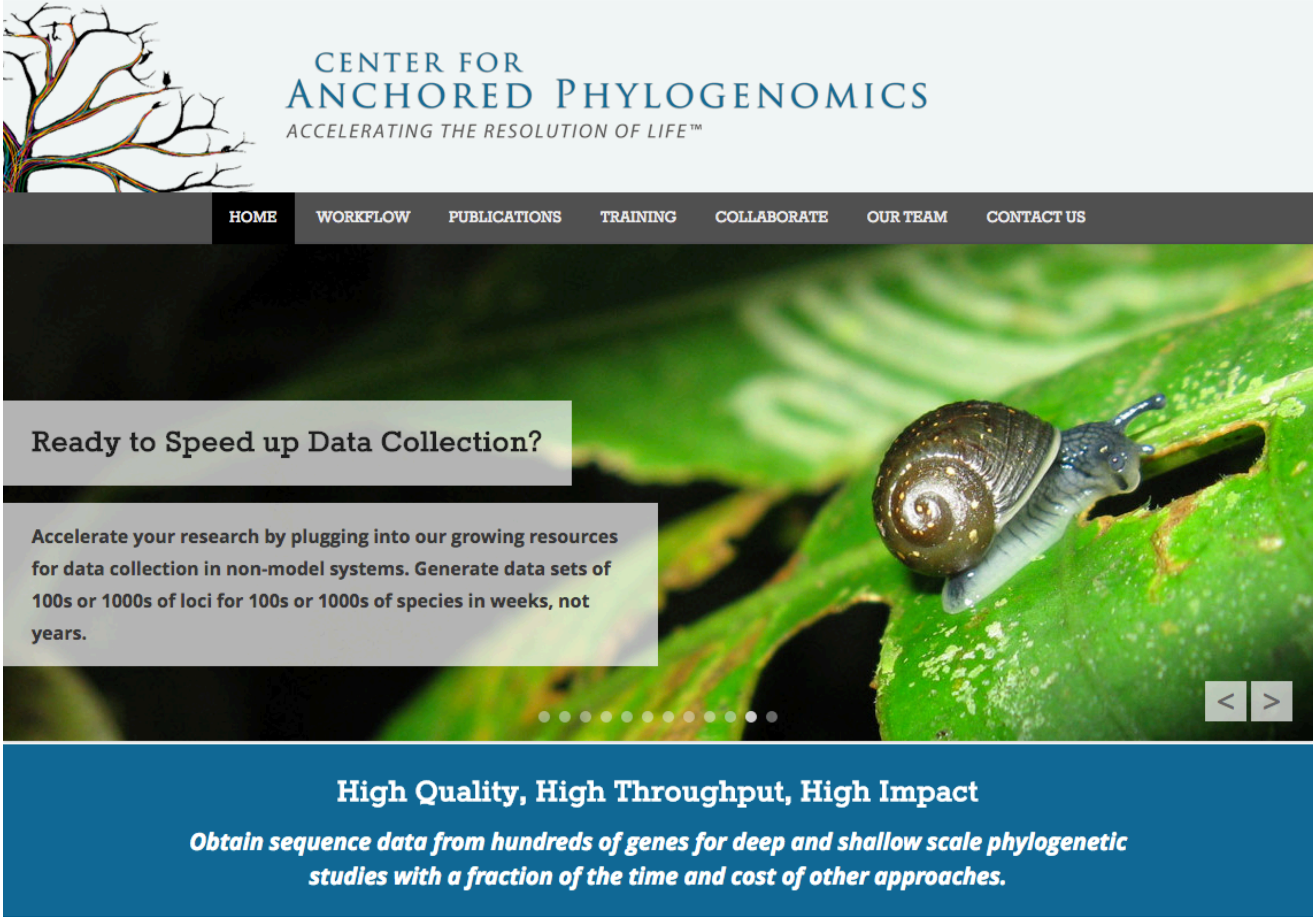
up to 5,000 loci
in vertebrates



a) UCEs identified in alignments of birds and lizard
UCE
anole
zebra finch
chicken
b) Probes designed from UCE regions
probe
UCE<180 bp
flank added
probe
UCE=180 bp
tiled probes
UCE>180 bp
c) RNA probes mixed with sheared genomic DNA from non-model organisms
d) Target DNA isolated, enriched, tagged, and pooled for NGS
e) Contigs assembled from NGS reads, aligned to probe, and consensus called for locus
locus j
consensus
f) Consensus loci aligned among species and gene trees estimated for all loci j1->n
locus j
locus k
locus m
g) Species tree estimated from gene trees

Faircloth *et al.* (2012) *Syst. Biol.*

The Future of Phylogenomics



CENTER FOR ANCHORED PHYLOGENOMICS

ACCELERATING THE RESOLUTION OF LIFE™

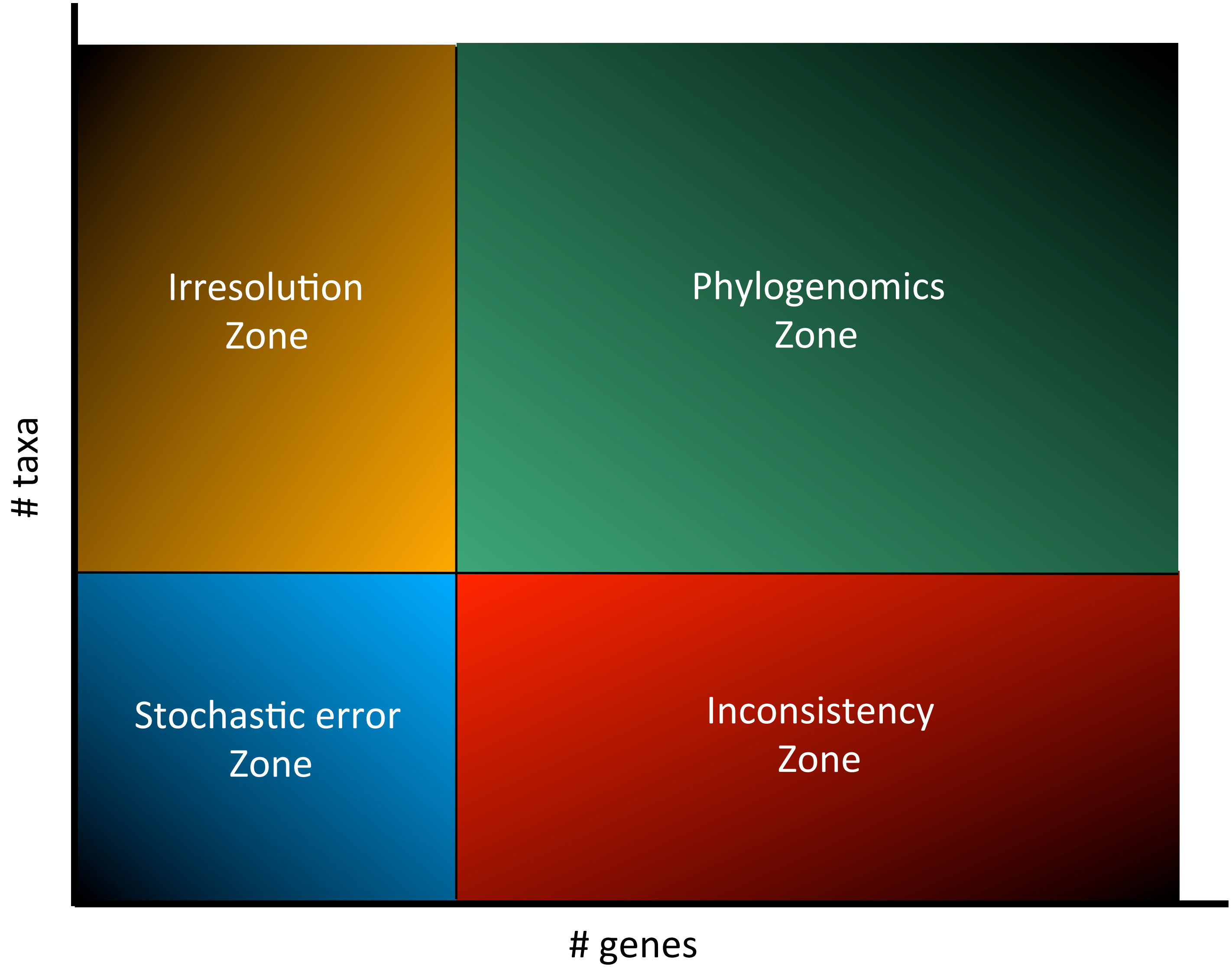
Ready to Speed up Data Collection?

Accelerate your research by plugging into our growing resources for data collection in non-model systems. Generate data sets of 100s or 1000s of loci for 100s or 1000s of species in weeks, not years.

High Quality, High Throughput, High Impact

Obtain sequence data from hundreds of genes for deep and shallow scale phylogenetic studies with a fraction of the time and cost of other approaches.

From Phylogenetics to Phylogenomics



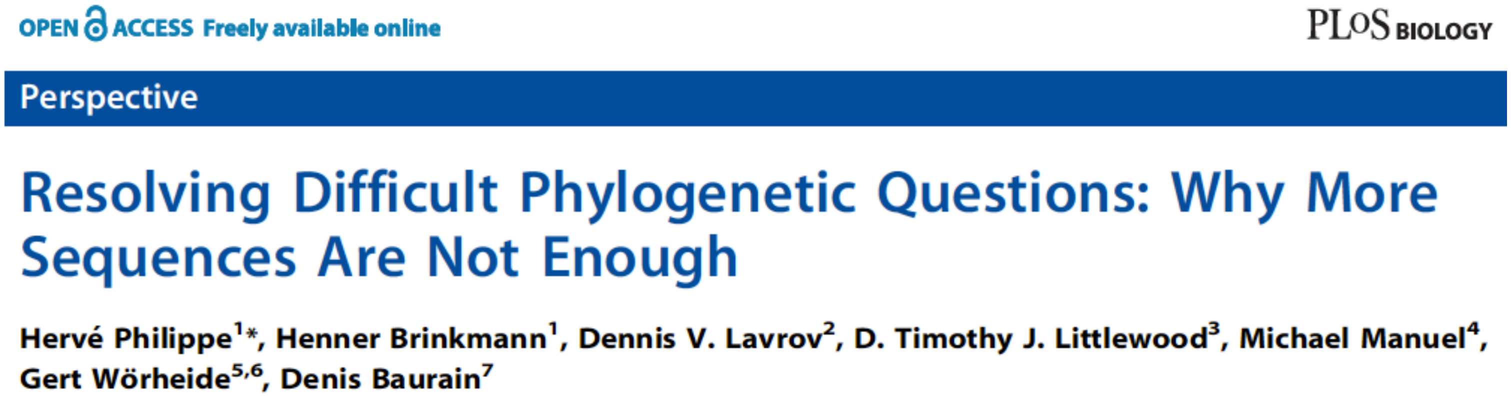
Irresolution Zone

Phylogenomics Zone

Stochastic error Zone

Inconsistency Zone

Towards a Full Resolution of the Tree of Life?



OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough

Hervé Philippe[1]*, Henner Brinkmann[1], Dennis V. Lavrov[2], D. Timothy J. Littlewood[3], Michael Manuel[4], Gert Wörheide[5,6], Denis Baurain[7]

Sources of topological incongruence in phylogenomics:

Biological reasons:

- Incomplete lineage sorting
- Horizontal gene transfer
- Hidden paralogy

Artificial reasons:

- Phylogenetic reconstruction artefacts (LBA, Compositional biases, ...)
- **Sequence misidentifications / contaminations**

=> Need for quality control methods and data exploration tools.

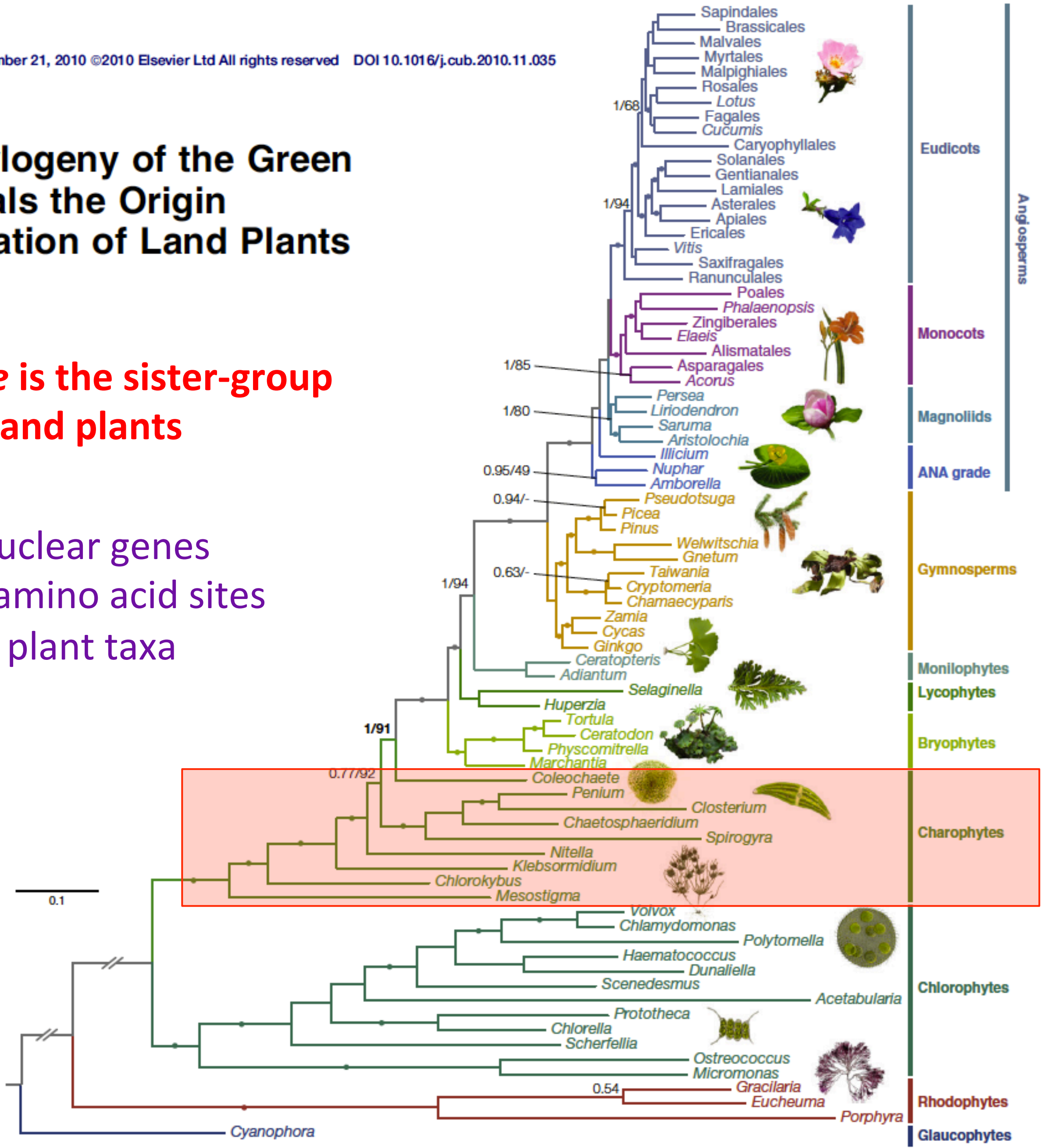
Phylogenomics and the Origin of Land Plants

Current Biology *20*, 2217–2222, December 21, 2010 ©2010 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2010.11.035

Multigene Phylogeny of the Green Lineage Reveals the Origin and Diversification of Land Plants

Coleochaete **is the sister-group of land plants**

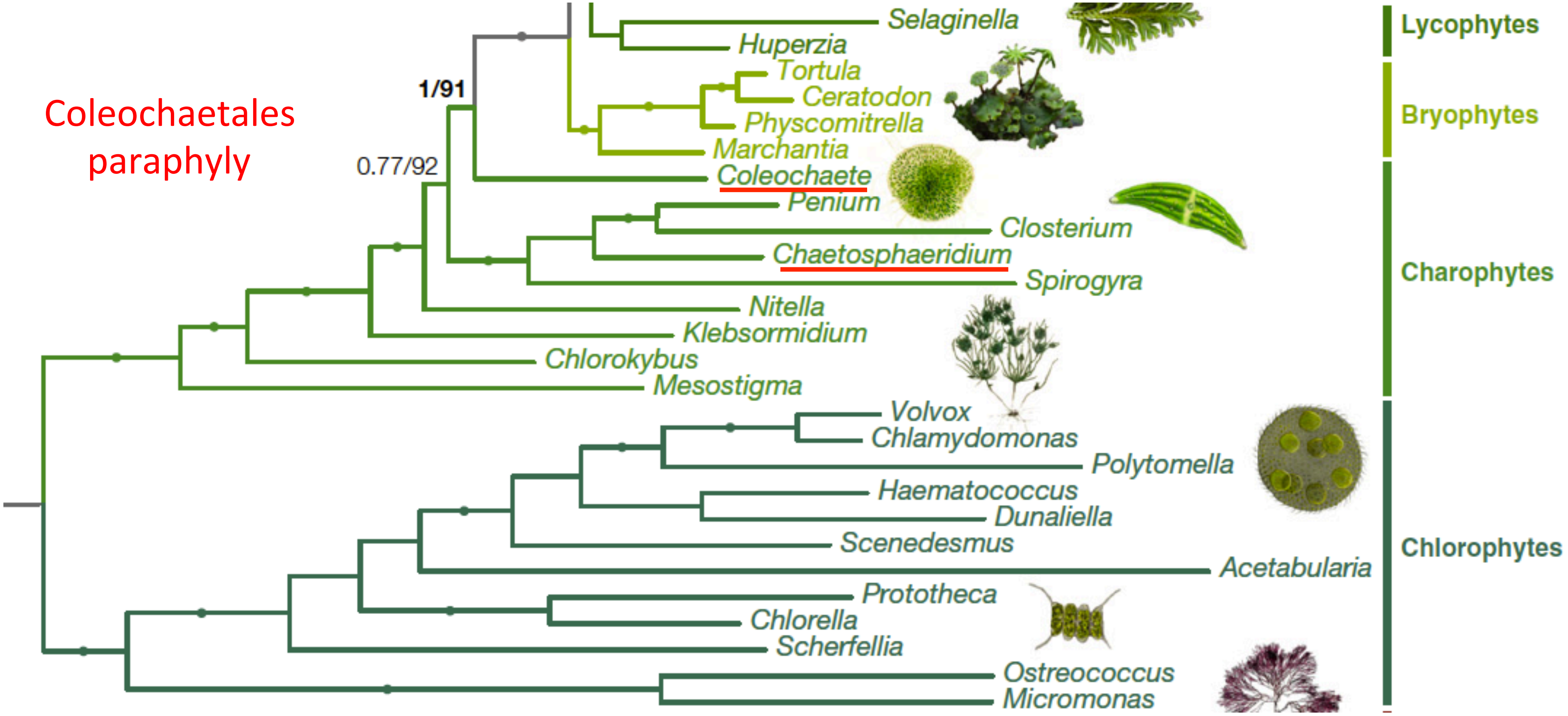
77 nuclear genes
12,149 amino acid sites
77 plant taxa



Finet *et al.* (2010) *Current Biol.*

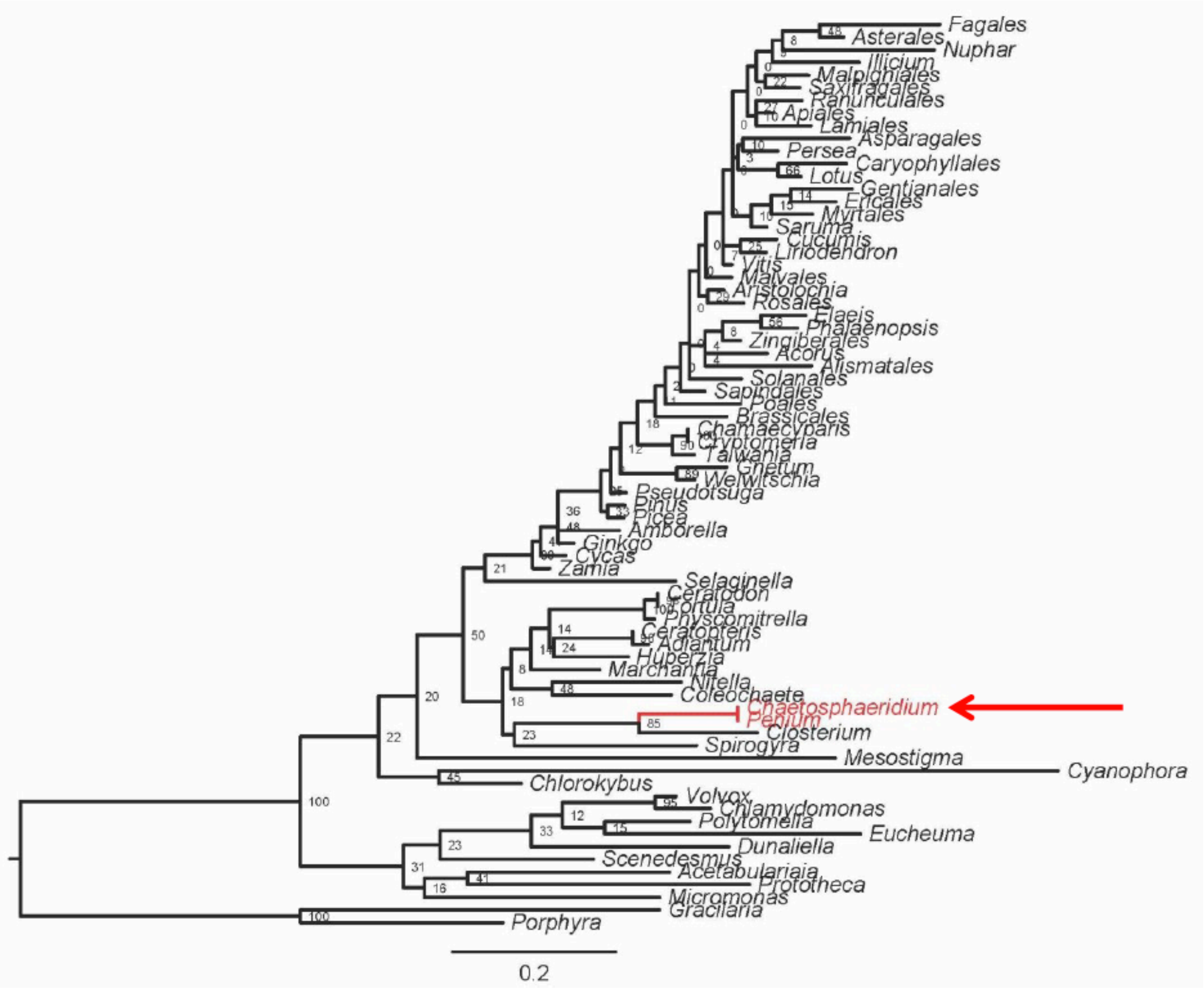
Phylogenomics and the Origin of Land Plants

Coleochaetales paraphyly



Finet *et al.* (2010) *Current Biol.*

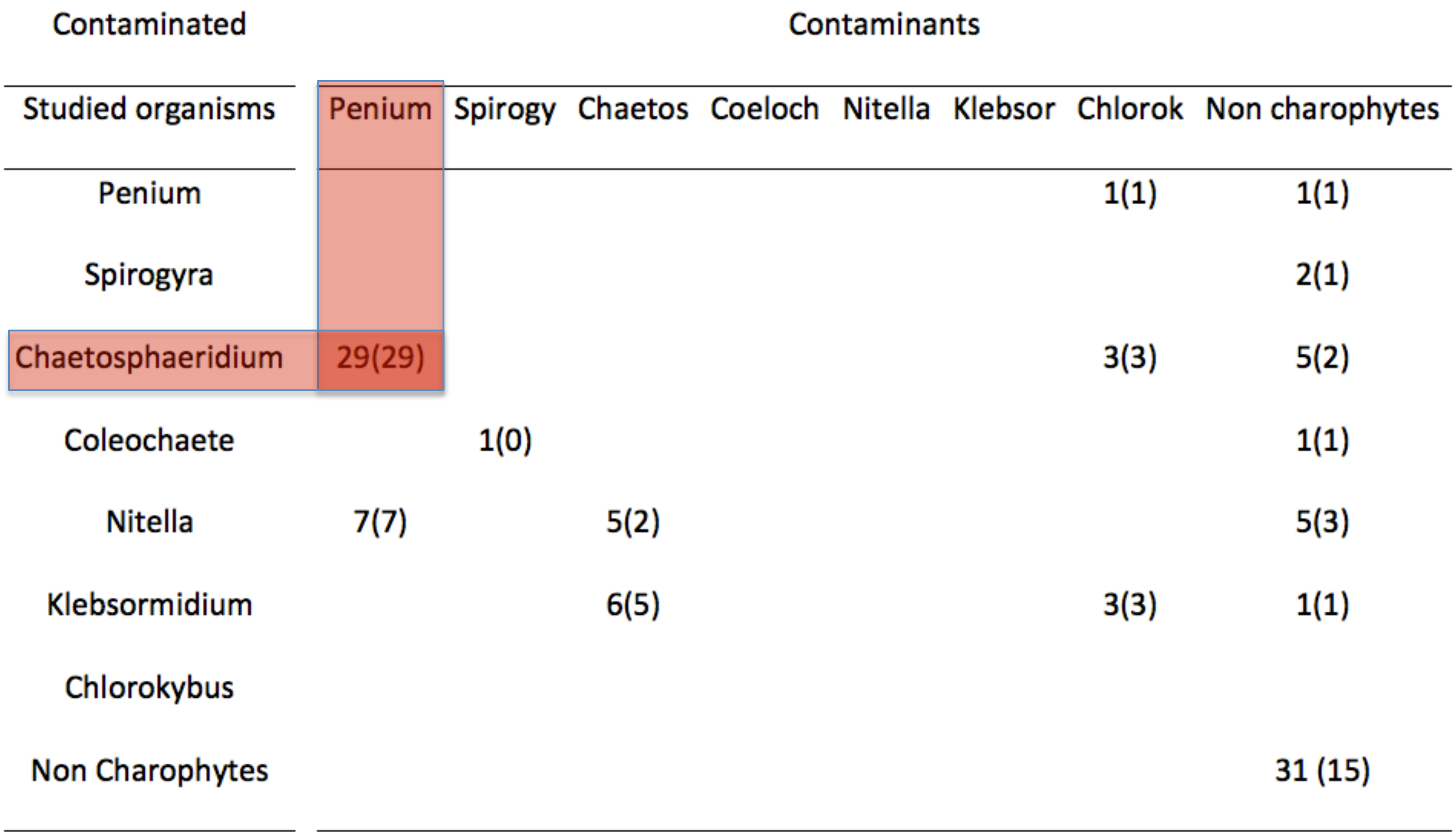
Cross-contaminations among New Transcriptomes



Laurin-Lemay, Brinkmann & Philippe (2012) *Current Biol.*

Massive Contaminations in Finet et al. Data Set

We found a total of 101 contaminated sequences, including a rotifer instead of the charalean Nitella (rpl27), or a diatom instead of the chlorophyte Volvox (rpl11b); contaminations by parasites, symbionts or commensals are not rare in transcriptomic datasets ([5] and unpublished results) and should be systematically verified and discarded. More problematically, most (55 out of 101) correspond to cross-contaminations among the seven newly sequenced charophytes [4], i.e. sequences from distantly related charophytes are virtually identical at the nucleotide level. In particular, 29 sequences from the coleochaetalean Chaetosphaeridium are from the zygnematalean Penium.

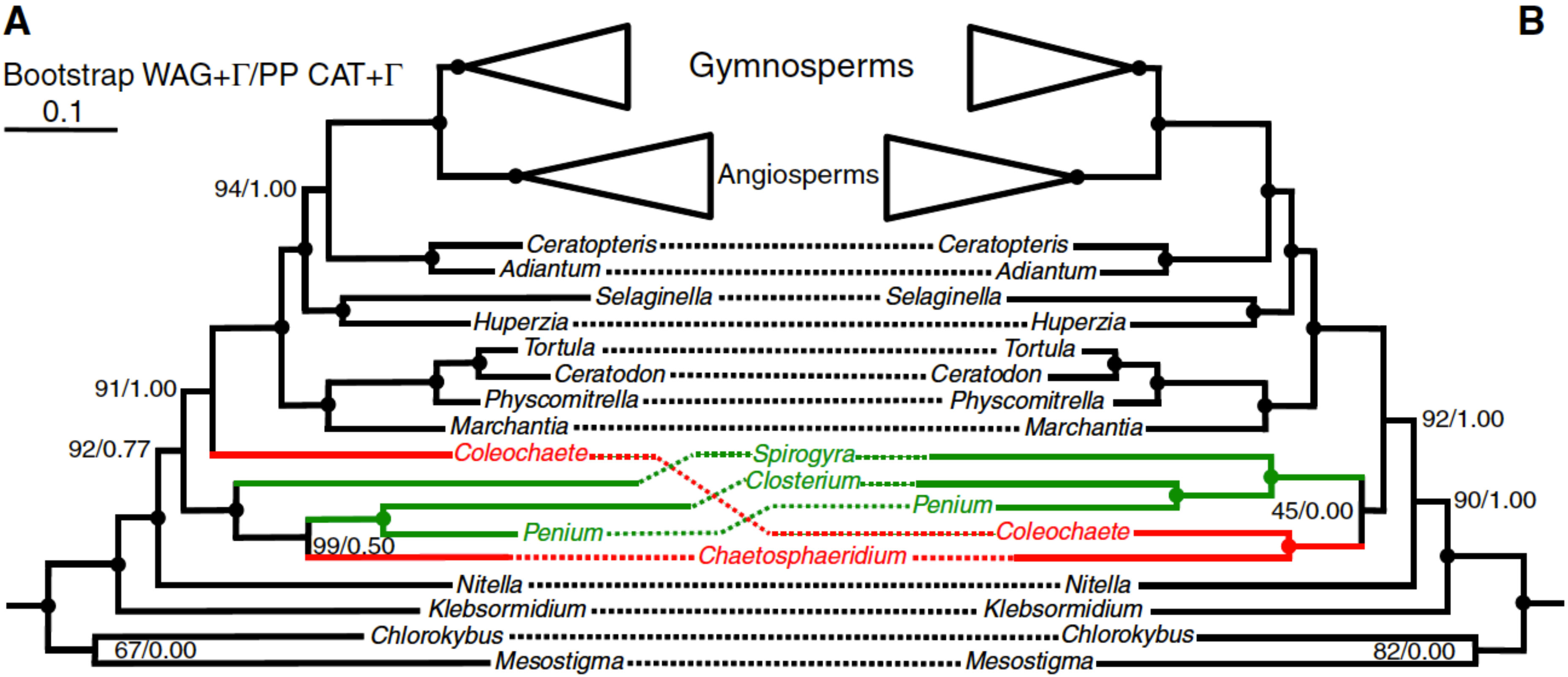


Contaminated	Contaminants							
Studied organisms	Penium	Spirogy	Chaetos	Coeloch	Nitella	Klebsor	Chlorok	Non charophytes
Penium							1(1)	1(1)
Spirogyra								2(1)
Chaetosphaeridium	29(29)						3(3)	5(2)
Coleochaete		1(0)						1(1)
Nitella	7(7)		5(2)					5(3)
Klebsormidium			6(5)				3(3)	1(1)
Chlorokybus								
Non Charophytes								31 (15)

Laurin-Lemay, Brinkmann & Philippe (2012) *Current Biol.*

Contaminations and the Origin of Land Plants

The congruence test revealed **74 contaminant sequences in the 77 ribosomal protein alignments** of Finet et al., and yielded to the **removal of 99 sequences** (because in 25 cases it was not possible to determine which is the correct sequence).

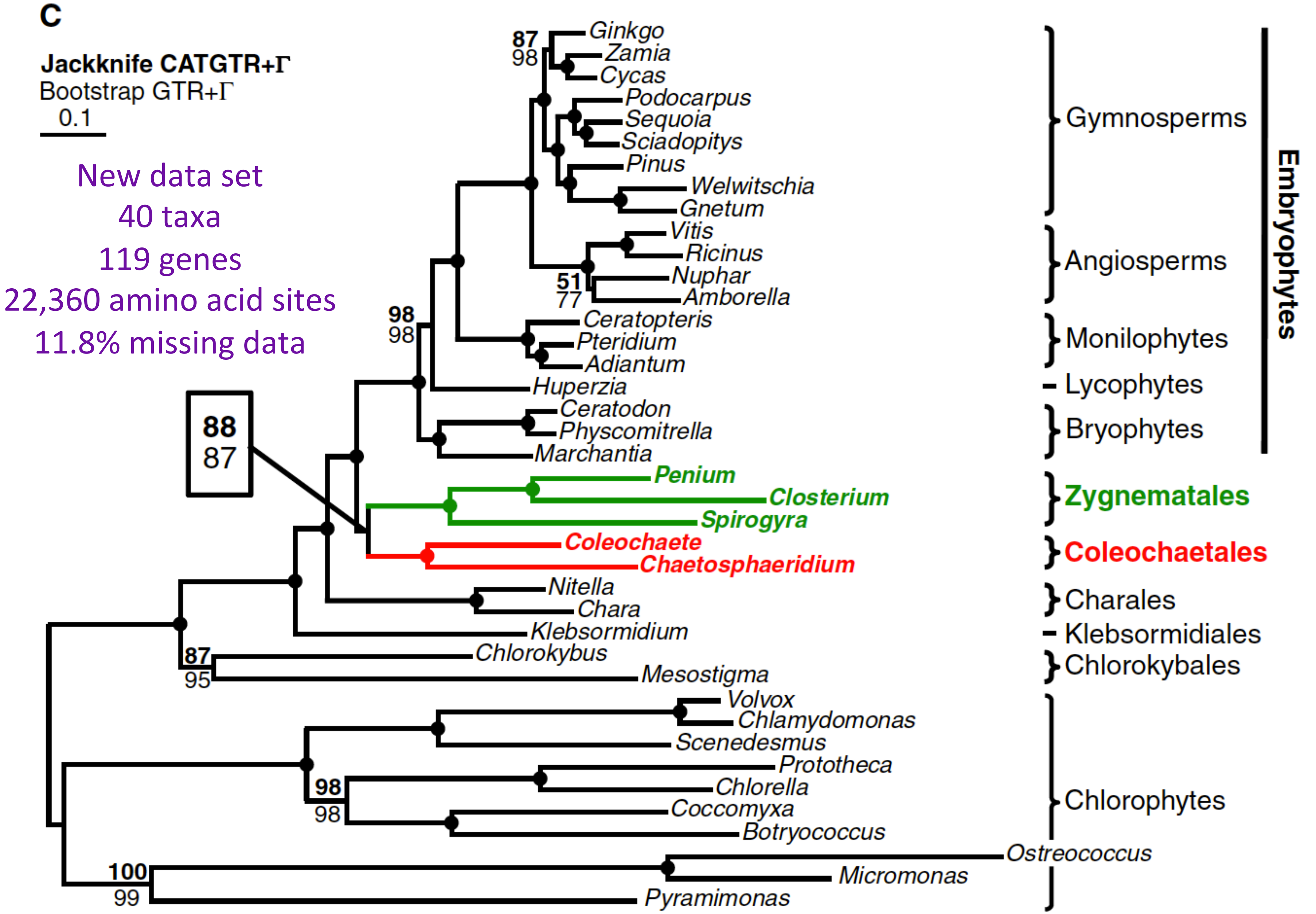


Original data set

Decontaminated data set

Laurin-Lemay *et al.* (2012) *Current Biol.*

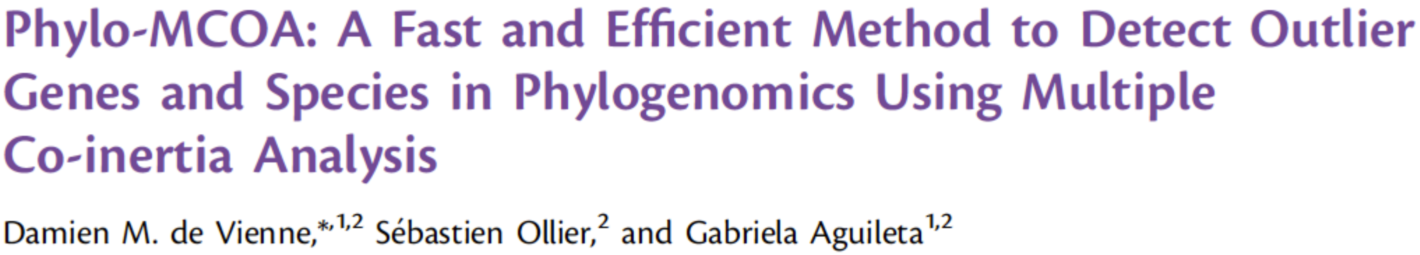
Contamination-free Phylogenomics



New data set
40 taxa
119 genes
22,360 amino acid sites
11.8% missing data

Laurin-Lemay, Brinkmann & Philippe (2012) *Current Biol.*

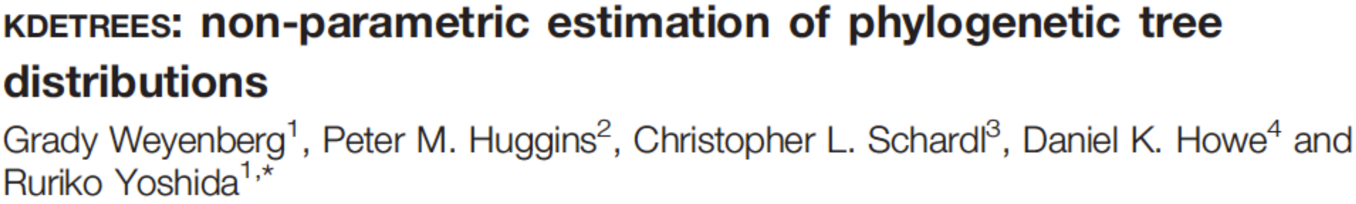
Methods for Detecting Outliers in Phylogenomic Data Sets



Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis

Damien M. de Vienne,*[1,2] Sébastien Ollier,[2] and Gabriela Aguileta[1,2]

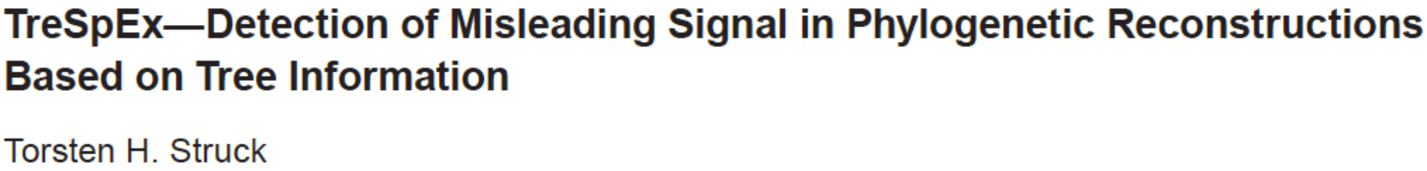
=> Multiple co-inertia analysis (MCOA) extracting the similarities and discrepancies among genes in terms of pairwise distances.



KDETREES: non-parametric estimation of phylogenetic tree distributions

Grady Weyenberg[1], Peter M. Huggins[2], Christopher L. Schardl[3], Daniel K. Howe[4] and Ruriko Yoshida[1,*]

=> Non-parametric method based on topological distances with the goal of identifying trees that are significantly different from the rest of distribution.



TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information

Torsten H. Struck

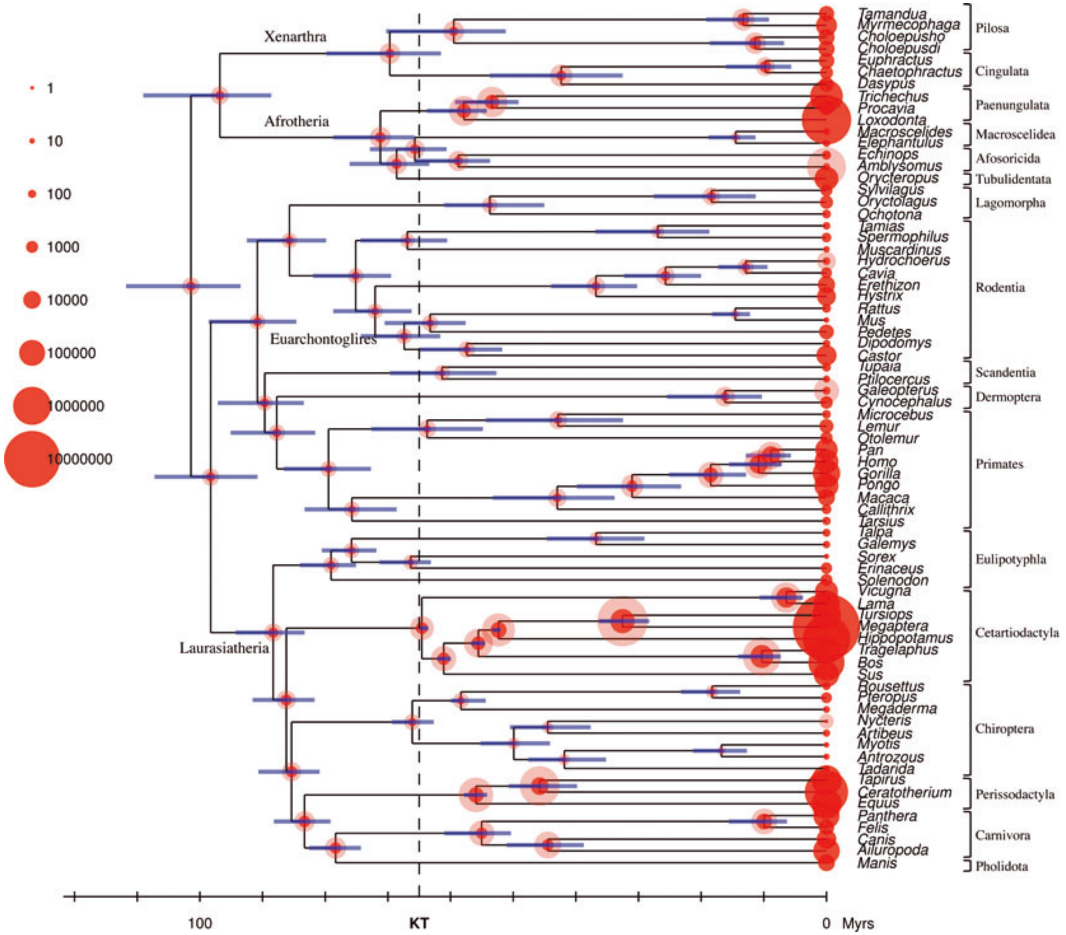
=> Combines different approaches utilizing tree-based information (nodal support or patristic distances) to identify misleading signals.

de Vienne *et al.* (2012) *Mol. Biol. Evol.*; Weyenberg et al. (2014) *Bioinformatics*; Struck (2014) *Evol. Bioinf.*

A Biological Case Study: The Sister-group of Primates

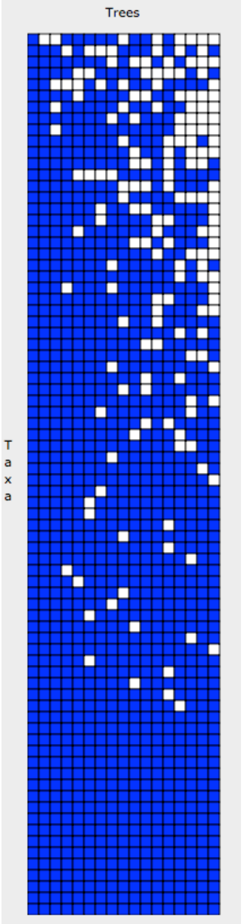
JOINT RECONSTRUCTION OF DIVERGENCE TIMES AND LIFE-HISTORY EVOLUTION IN PLACENTAL MAMMALS USING A PHYLOGENETIC COVARIANCE MODEL

Nicolas Lartillot[1] and Frédéric Delsuc[2]



Mining GenBank

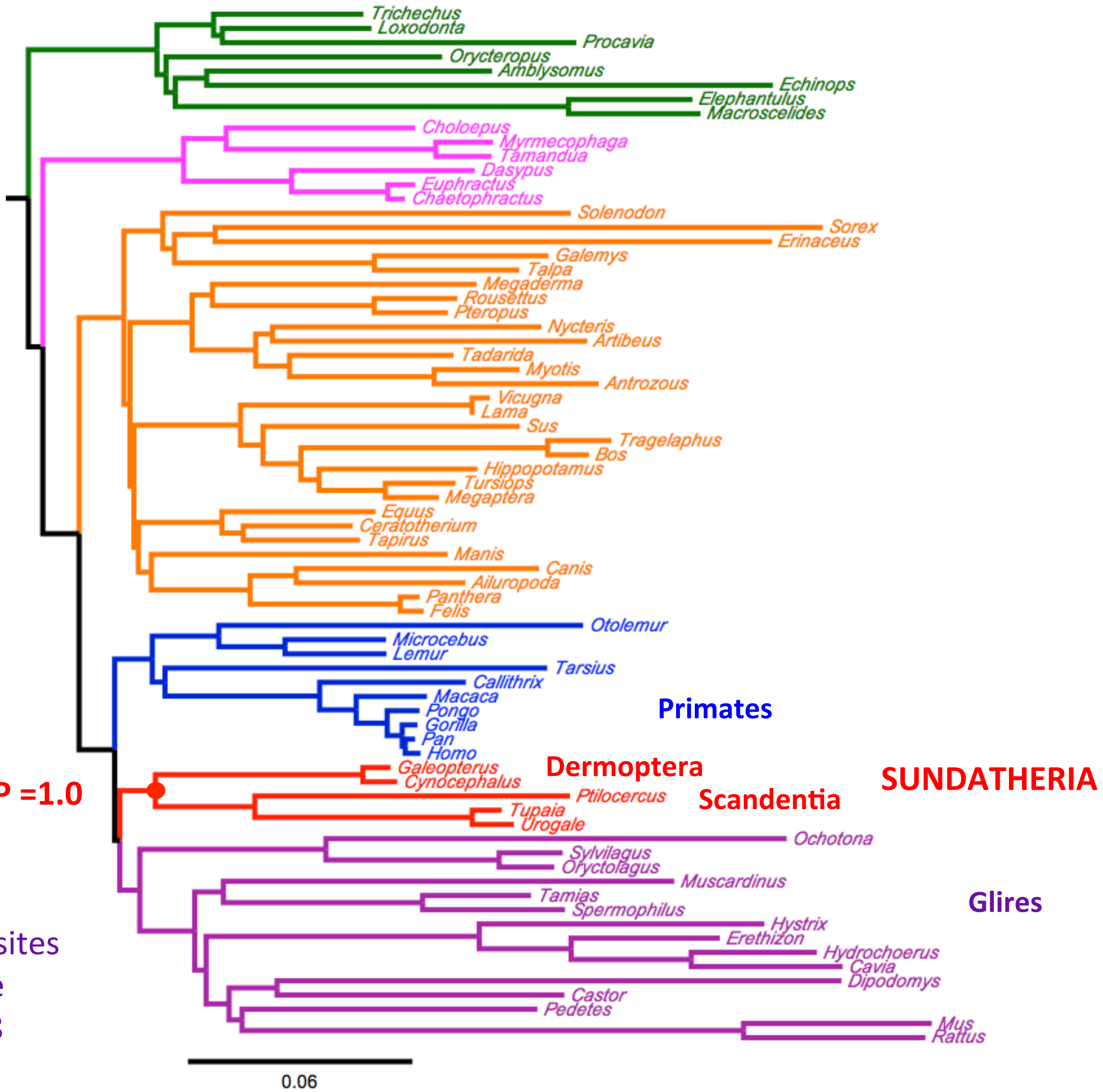
17 genes



78 species

Lartillot & Delsuc (2012) *Evolution*.

An Unexpected (yet Exciting) Result!

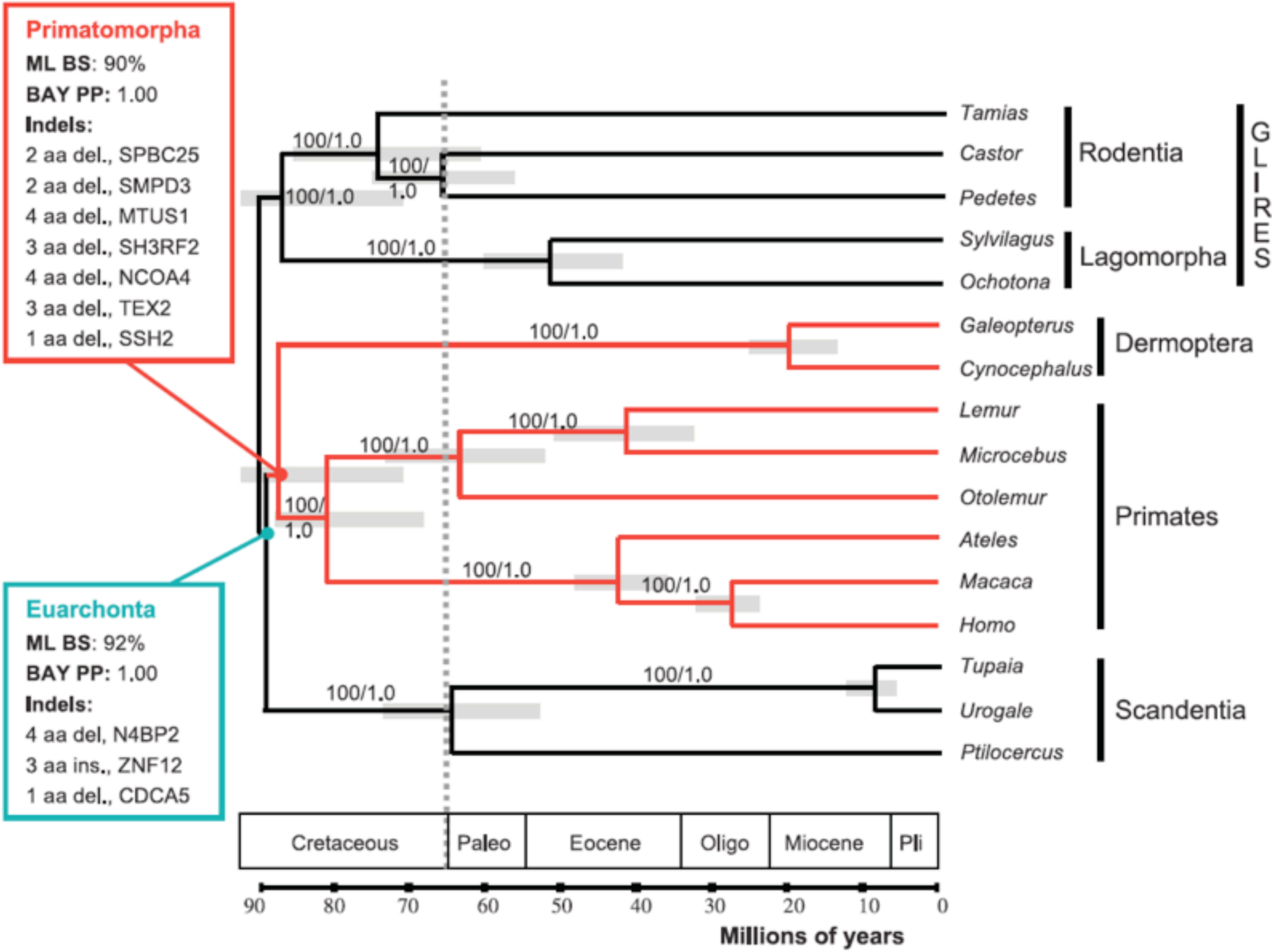


BP = 100 / PP =1.0

15,515
nucleotide sites
ML Tree
GTR+G8

+

Support for Primatomorpha



Primatomorpha

ML BS: 90%

BAY PP: 1.00

Indels:

2 aa del., SPBC25
2 aa del., SMPD3
4 aa del., MTUS1
3 aa del., SH3RF2
4 aa del., NCOA4
3 aa del., TEX2
1 aa del., SSH2

Euarchonta

ML BS: 92%

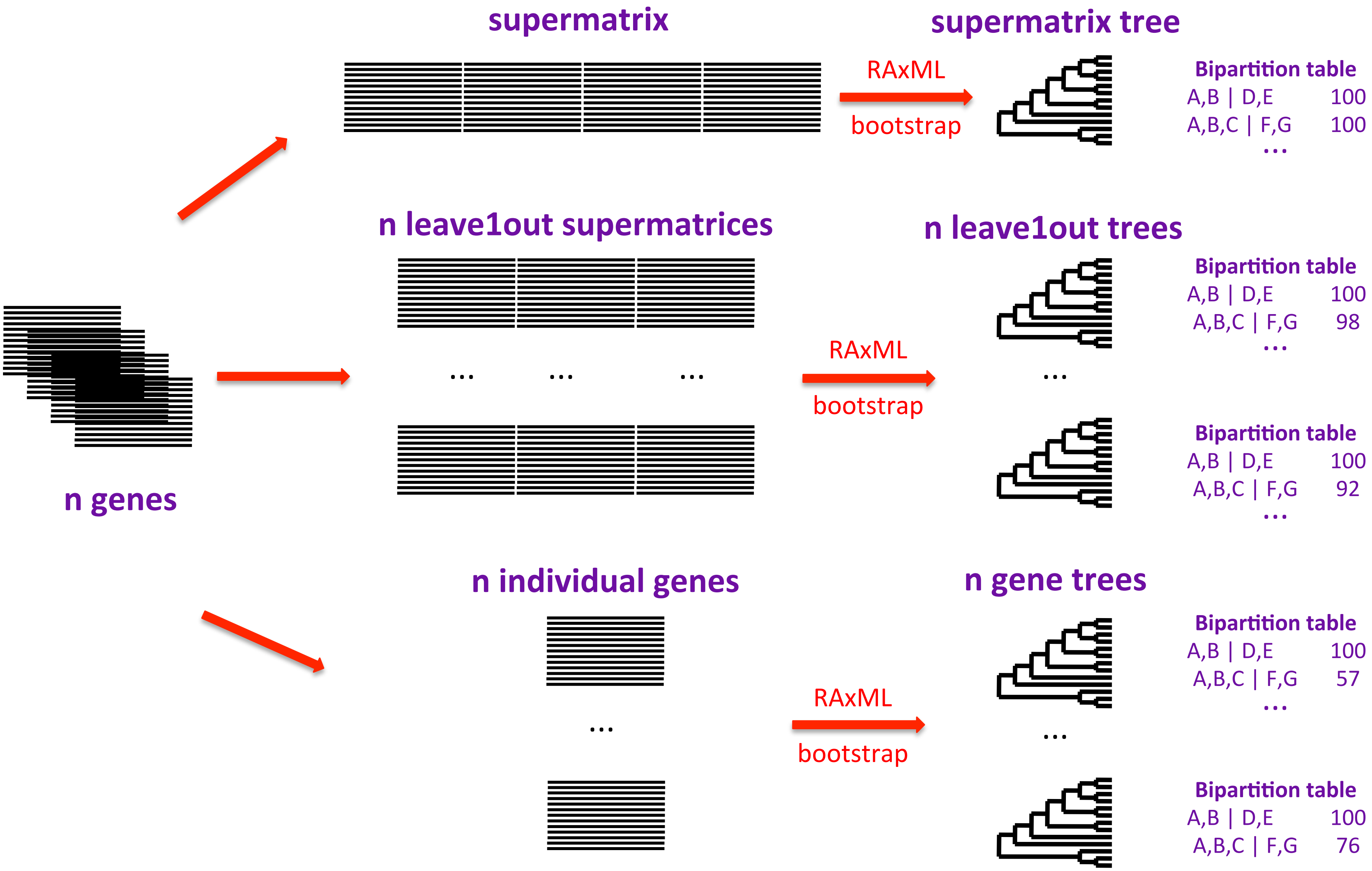
BAY PP: 1.00

Indels:

4 aa del, N4BP2
3 aa ins., ZNF12
1 aa del., CDCA5

Janecka *et al.* (2007) *Science*

A Simple Pipeline for Exploring Topological Incongruence



n genes

supermatrix

RAxML

bootstrap

supermatrix tree

Bipartition table

A,B | D,E	100
A,B,C | F,G	100
…	

n leave1out supermatrices

… … …

RAxML

bootstrap

n leave1out trees

Bipartition table

A,B | D,E	100
A,B,C | F,G	98
…	

…

Bipartition table

A,B | D,E	100
A,B,C | F,G	92
…	

n individual genes

…

RAxML

bootstrap

n gene trees

Bipartition table

A,B | D,E	100
A,B,C | F,G	57
…	

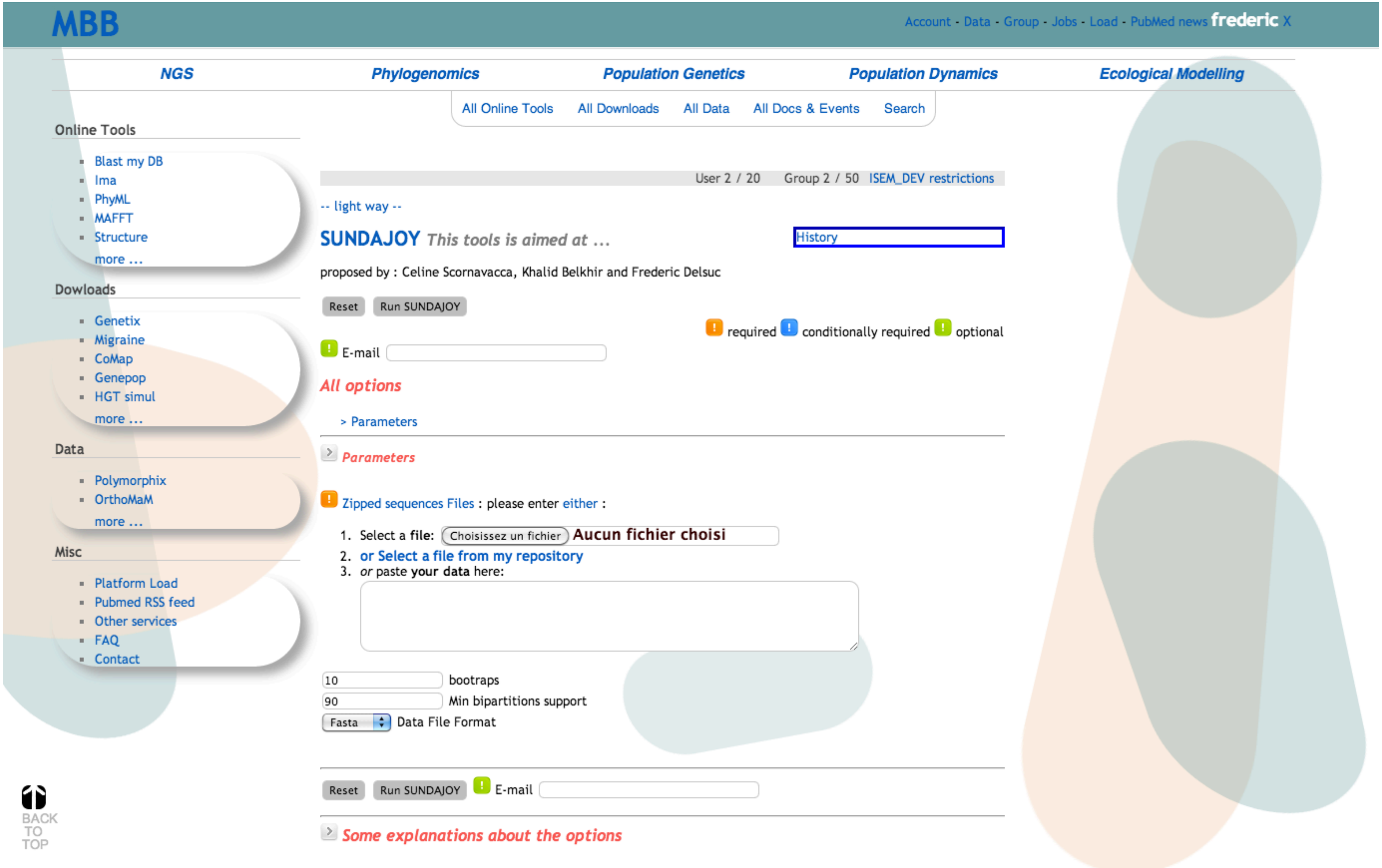
…

Bipartition table

A,B | D,E	100
A,B,C | F,G	76
…	

Montpellier Bioinformatics Biodiversity Platform (MBB)

http://mbb.univ-montp2.fr/MBB/



User 2 / 20   Group 2 / 50   ISEM_DEV restrictions

-- light way --

SUNDAJOY *This tools is aimed at ...*

proposed by : Celine Scornavacca, Khalid Belkhir and Frederic Delsuc

! required ! conditionally required ! optional

! E-mail

All options

> Parameters

Parameters

! Zipped sequences Files : please enter either :

1. Select a file: Choisissez un fichier **Aucun fichier choisi**
2. or Select a file from my repository
3. *or* paste **your data** here:

10 bootraps

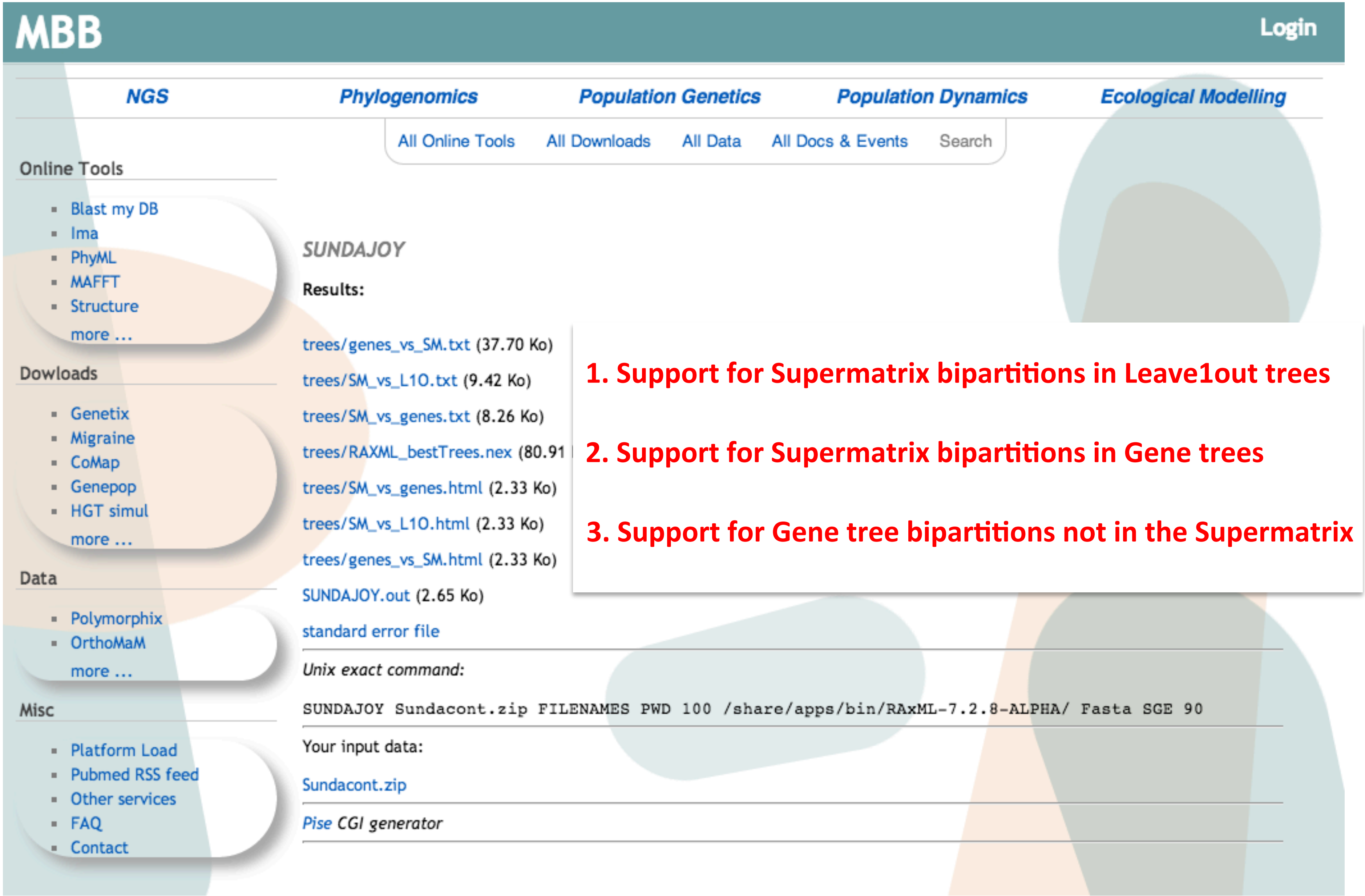
90 Min bipartitions support

Fasta Data File Format

Reset Run SUNDAJOY ! E-mail

Some explanations about the options

Exploratory Plots for Detecting Incongruent Bipartitions



SUNDAJOY

Results:

trees/genes_vs_SM.txt (37.70 Ko)

trees/SM_vs_L10.txt (9.42 Ko)

trees/SM_vs_genes.txt (8.26 Ko)

trees/RAXML_bestTrees.nex (80.91

trees/SM_vs_genes.html (2.33 Ko)

trees/SM_vs_L10.html (2.33 Ko)

trees/genes_vs_SM.html (2.33 Ko)

SUNDAJOY.out (2.65 Ko)

standard error file

Unix exact command:

```
SUNDAJOY Sundacont.zip FILENAMES PWD 100 /share/apps/bin/RAxML-7.2.8-ALPHA/ Fasta SGE 90
```

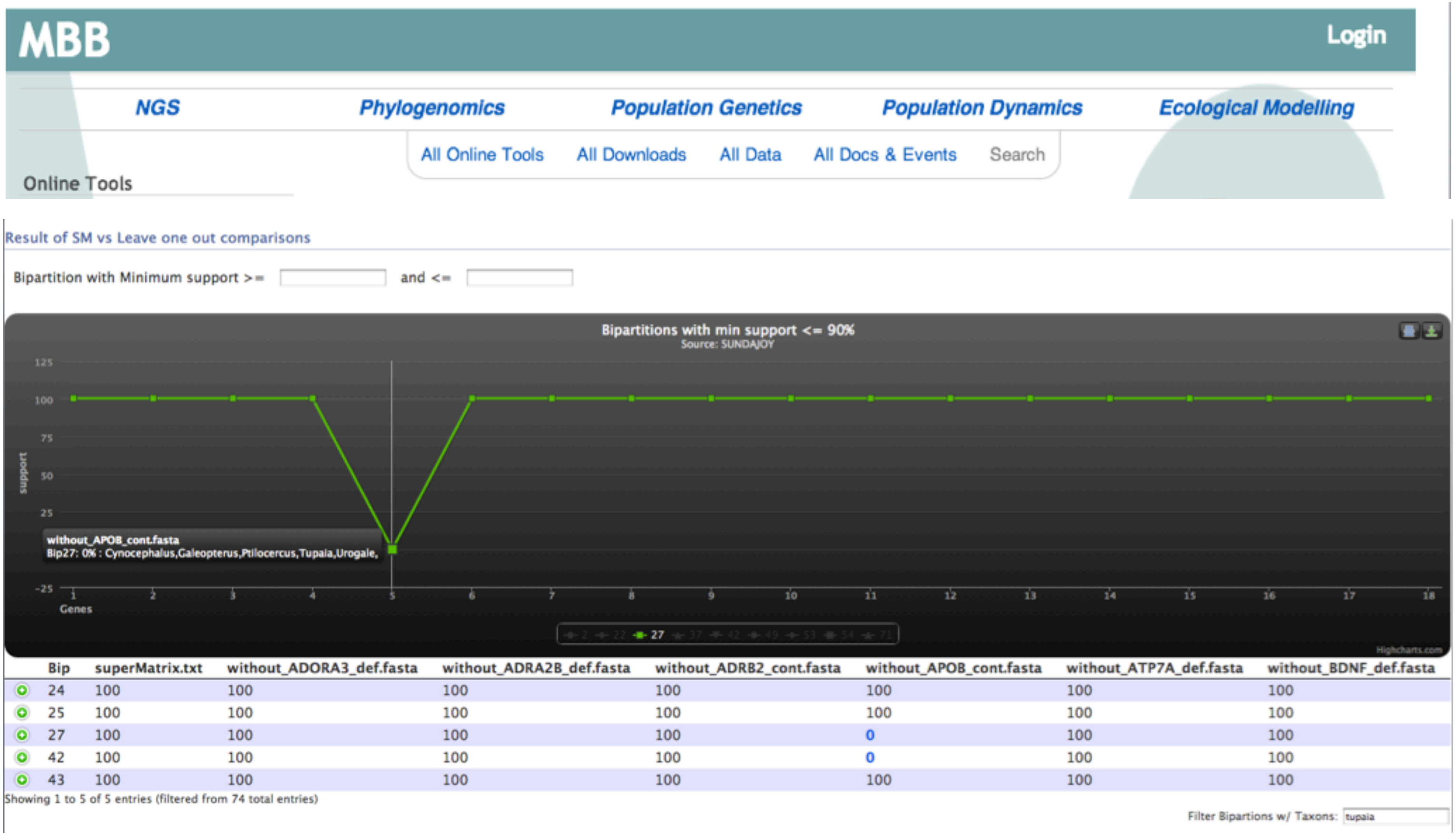
Your input data:

Sundacont.zip

Pise CGI generator

1. **Support for Supermatrix bipartitions in Leave1out trees**
2. **Support for Supermatrix bipartitions in Gene trees**
3. **Support for Gene tree bipartitions not in the Supermatrix**

Exploratory Plots for Detecting Incongruent Bipartitions



Result of SM vs Leave one out comparisons

Bipartition with Minimum support >= and <=

Bipartitions with min support <= 90%

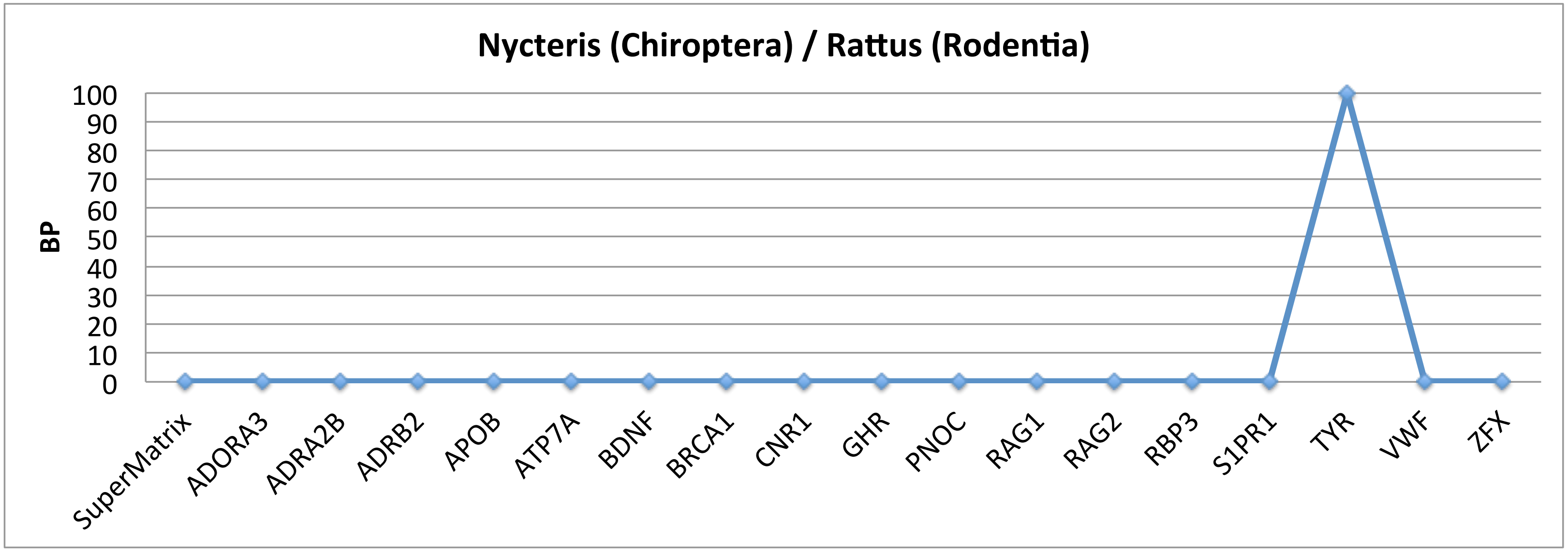
Source: SUNDAJOY

	Bip	superMatrix.txt	without_ADORA3_def.fasta	without_ADRA2B_def.fasta	without_ADRB2_cont.fasta	without_APOB_cont.fasta	without_ATP7A_def.fasta	without_BDNF_def.fasta
⊕	24	100	100	100	100	100	100	100
⊕	25	100	100	100	100	100	100	100
⊕	27	100	100	100	100	0	100	100
⊕	42	100	100	100	100	0	100	100
⊕	43	100	100	100	100	100	100	100

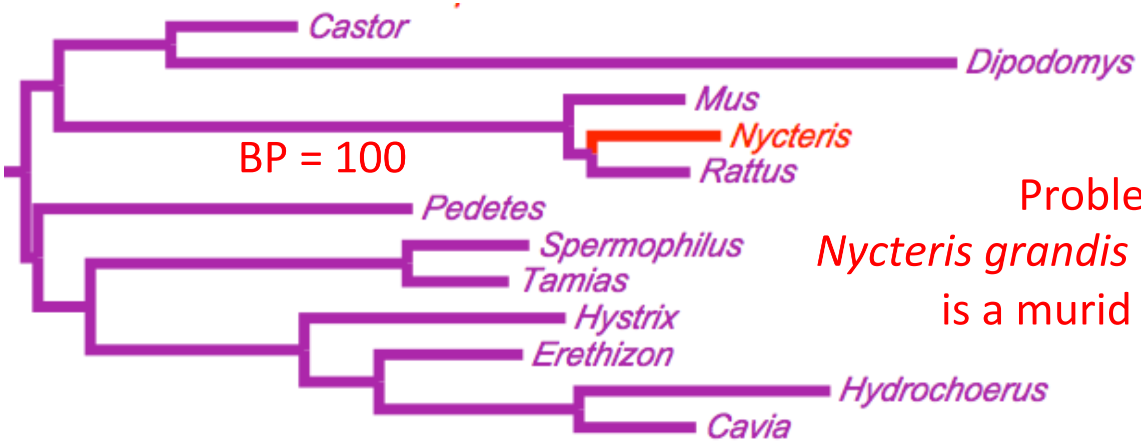
Showing 1 to 5 of 5 entries (filtered from 74 total entries)

Filter Bipartions w/ Taxons: tupaia

Contaminations / Misindentifications in GenBank

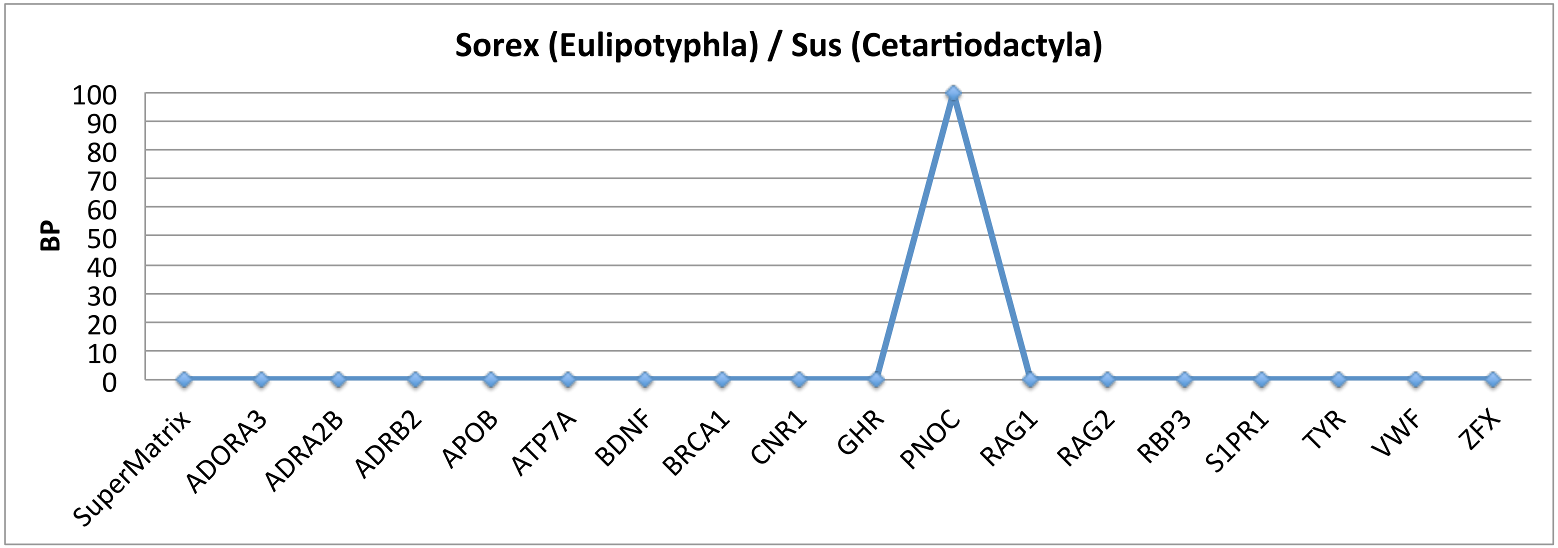


TYR (Tyrosinase)

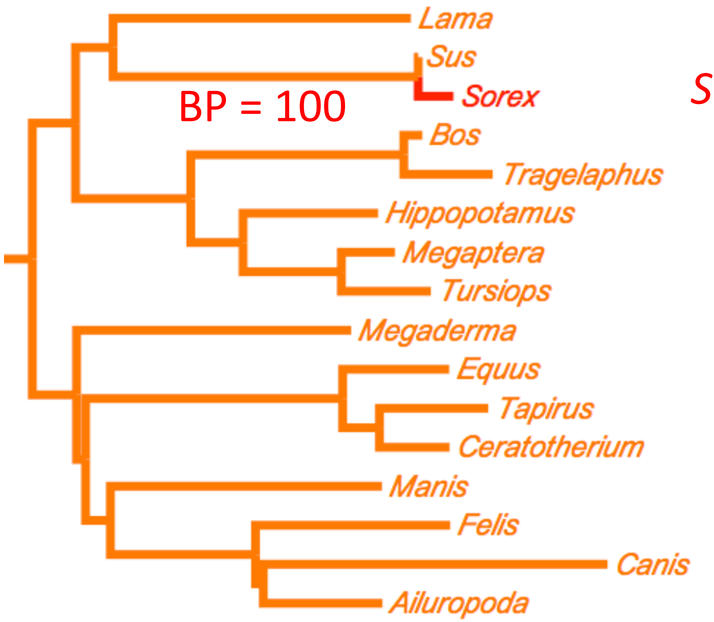


Problem:
Nycteris grandis (AY834610.1)
is a murid rodent

Contaminations / Misindentifications in GenBank



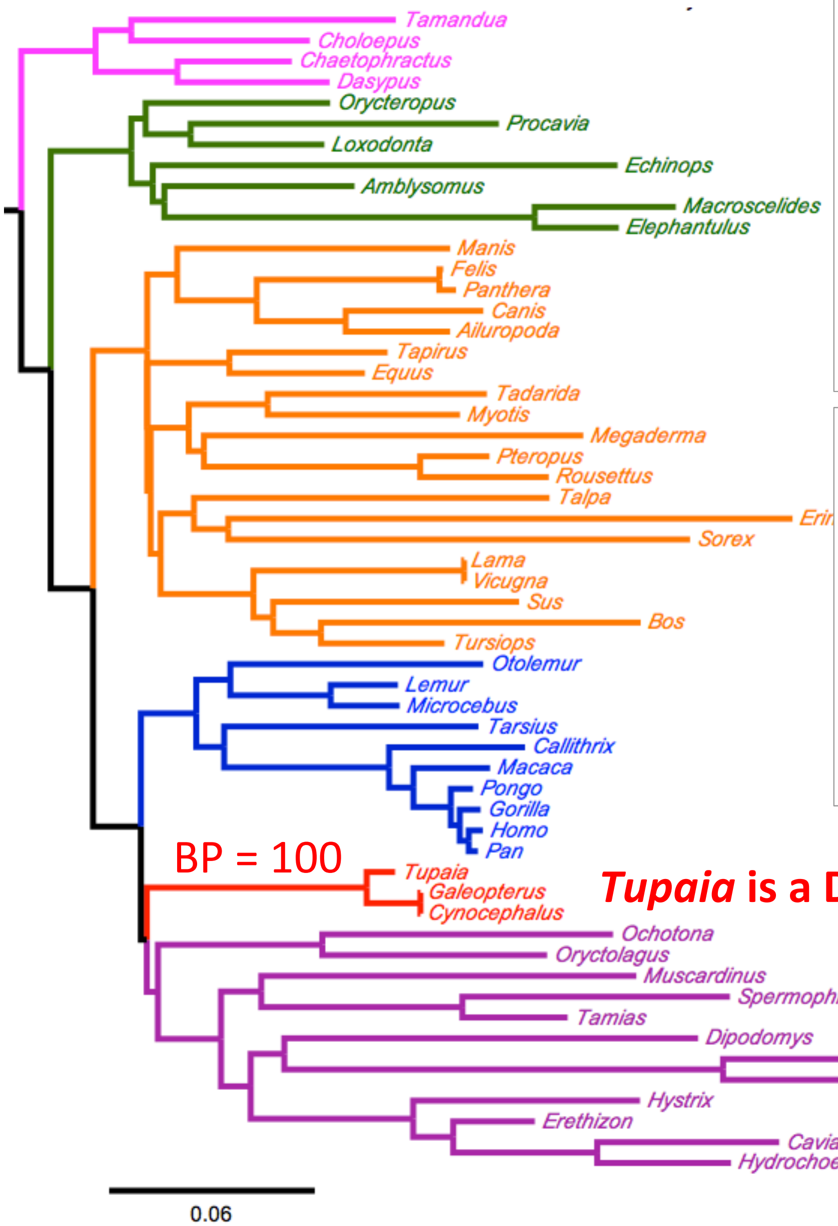
PNOC (Prepronociceptin)

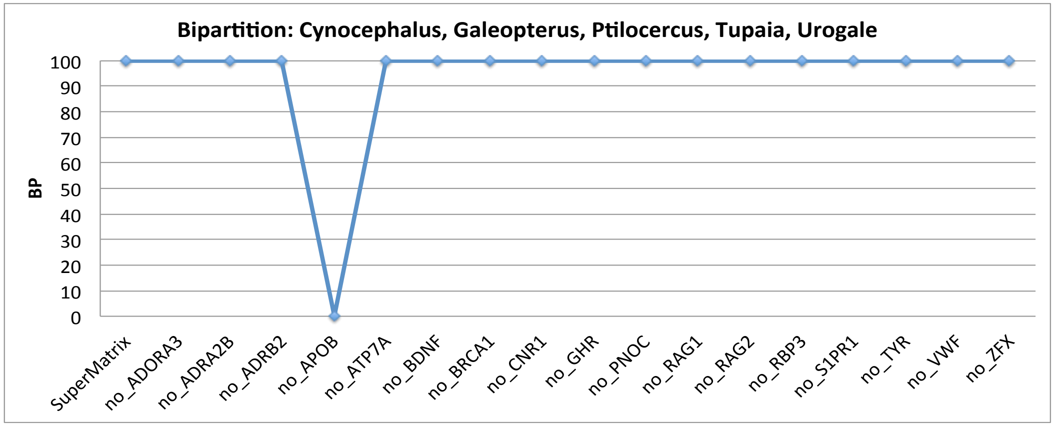


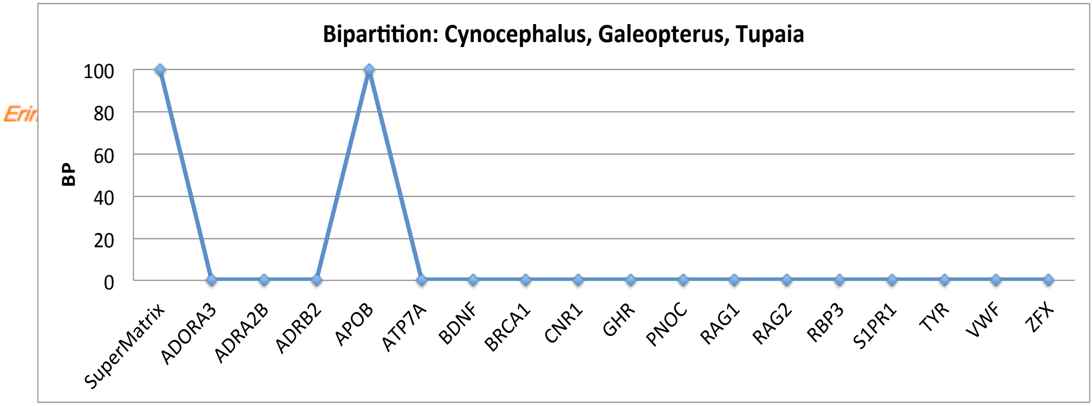
Problem:
Sorex araneus (AY011813.1)
is a Suidae

Contaminations / Misindentifications in GenBank

APOB (Apolipoprotein B)



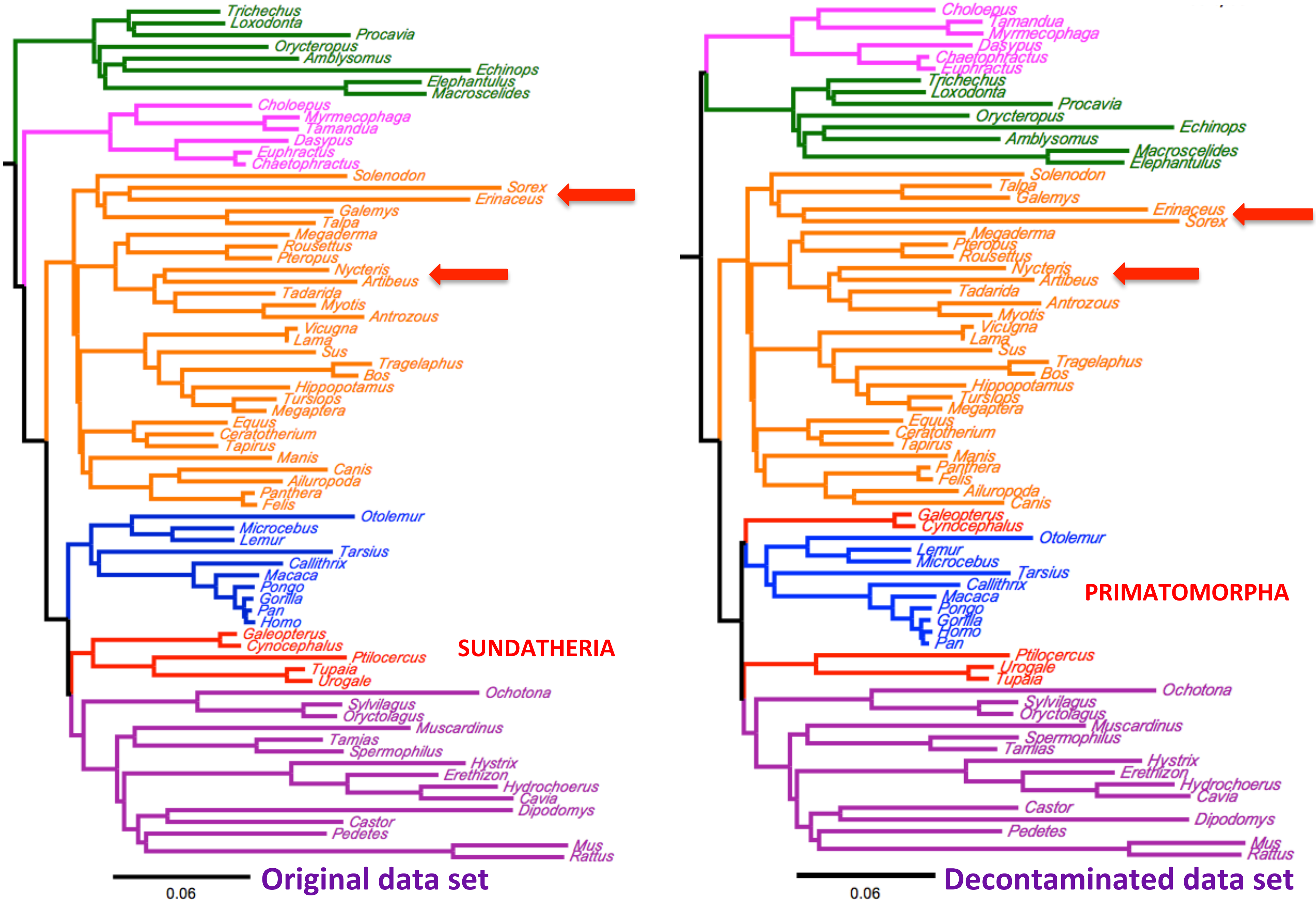




***Tupaia* is a Dermopteran (97% similarity)**

>gi|256549376|gb|FJ648363.1| *Tupaia glis* apolipoprotein B-like (APOB) gene
Submitted 29 October 2008
Ali F, Pons J, Shekelle M, Goodman M and Meier R
A sparse supermatrix recovers a well-supported primate phylogeny with dates (Unpublished)

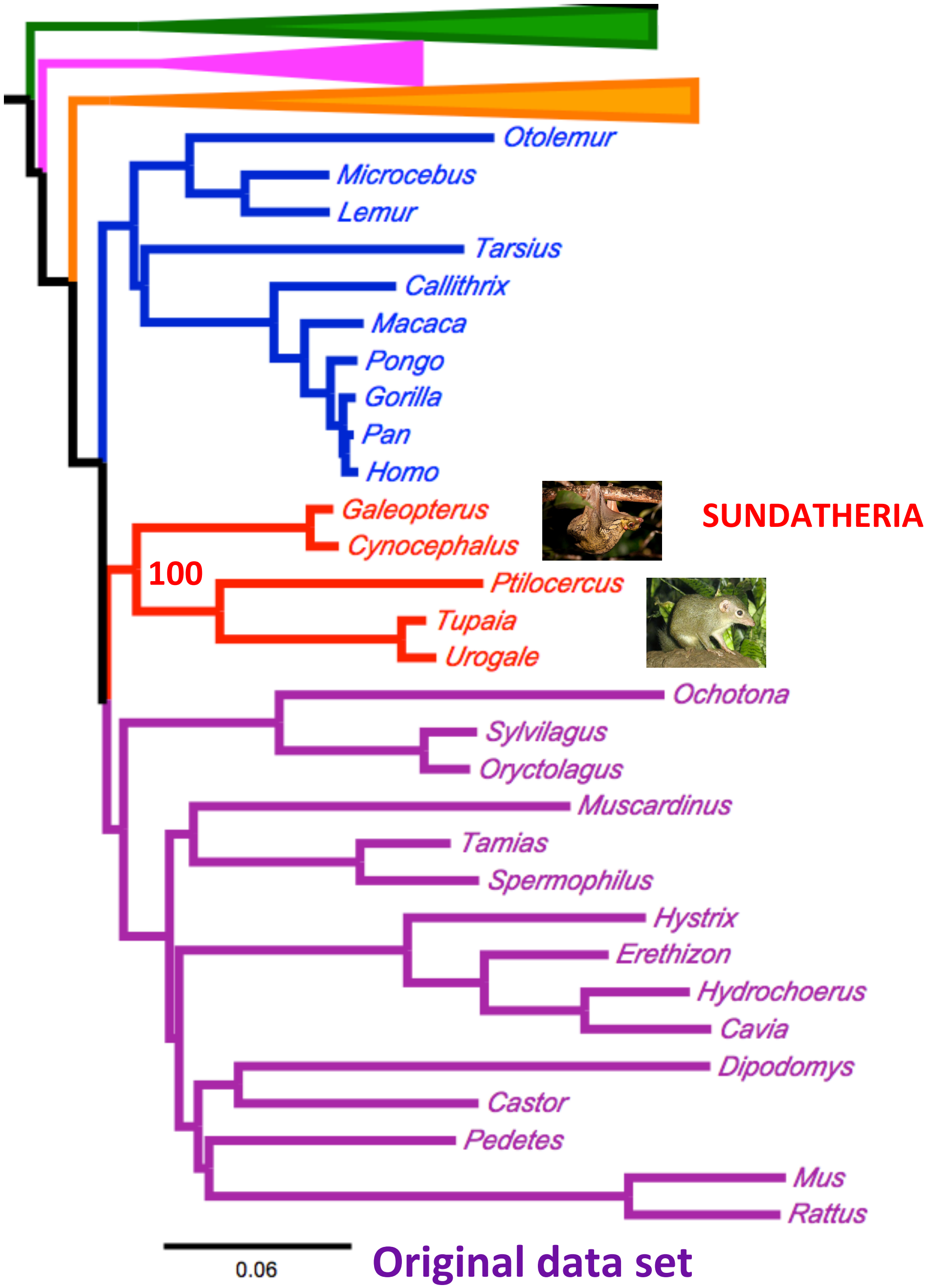
Effects of Contaminations on Phylogenomic Inference



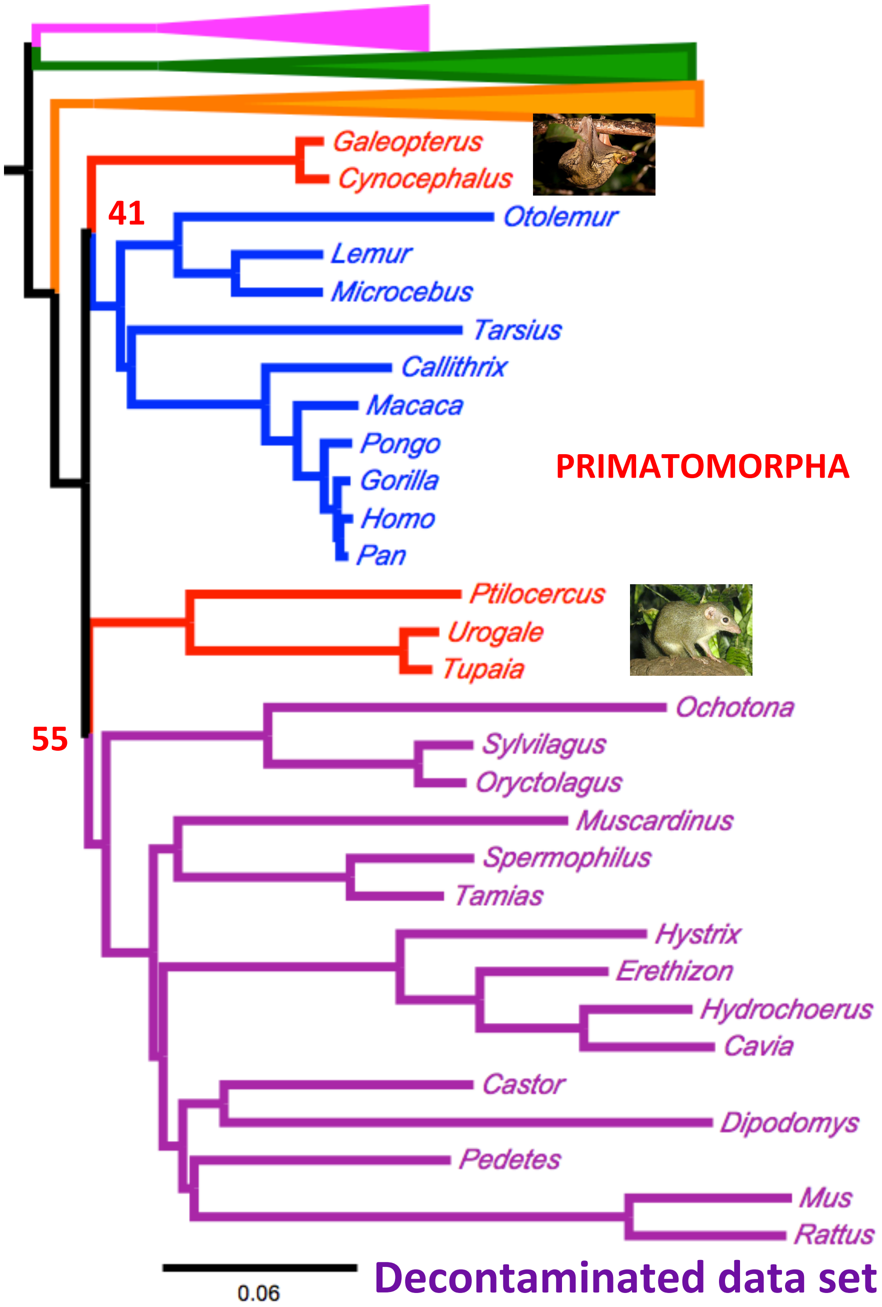
Original data set

Decontaminated data set

Effects of Contaminations on Phylogenomic Inference



Original data set



Decontaminated data set

Conclusions

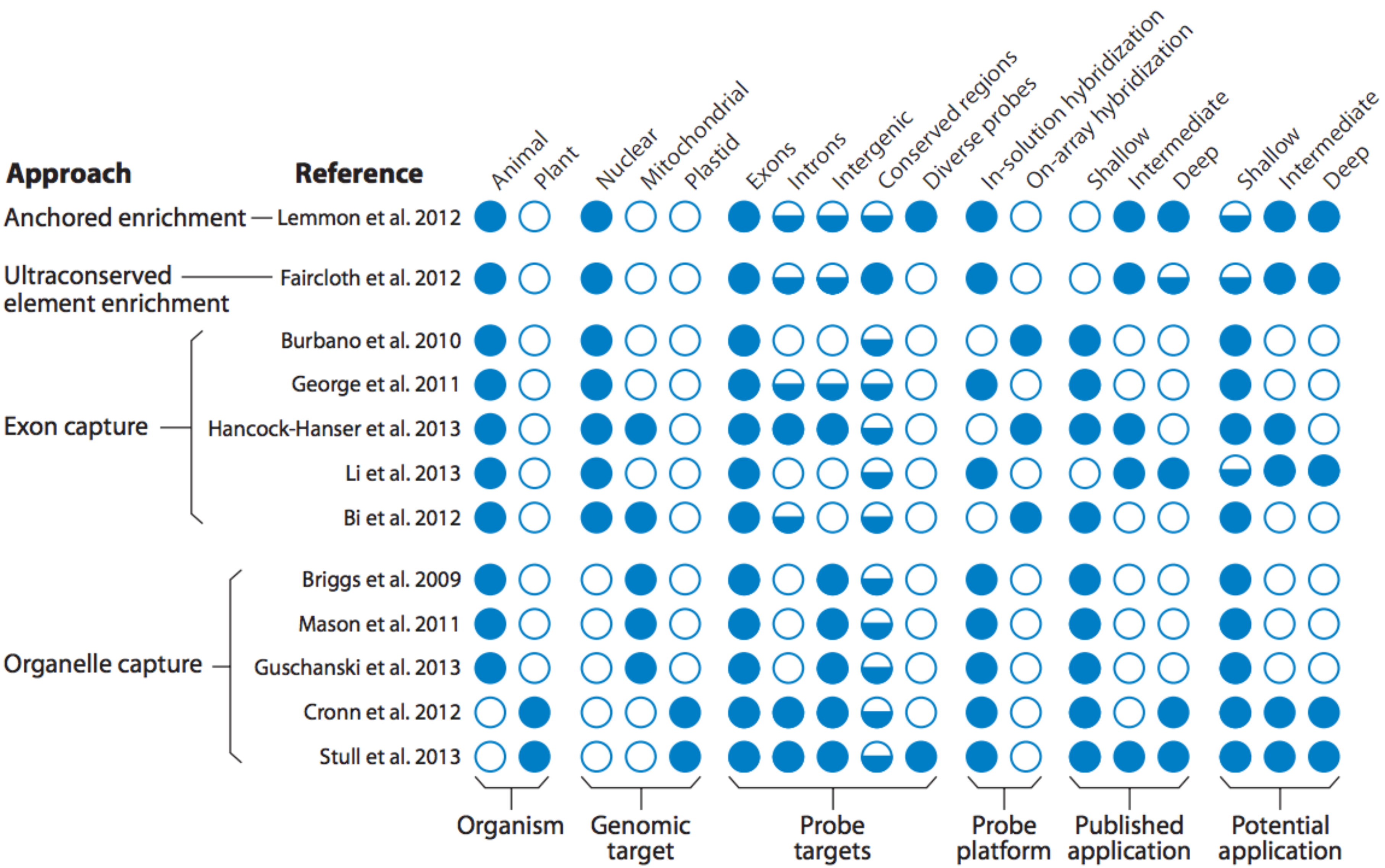
=> **Contaminations / misidentifications are frequent** in phylogenomic data sets and public databases

=> Simple **data exploration tools** based on bipartition support allow detecting incongruent signals due to misidentified / cotanminated sequences

=> Even a few number of contaminations / misidentifications **can strongly impact phylogenetic inference** when phylogenetic signal is scarce

Thanks for your attention!

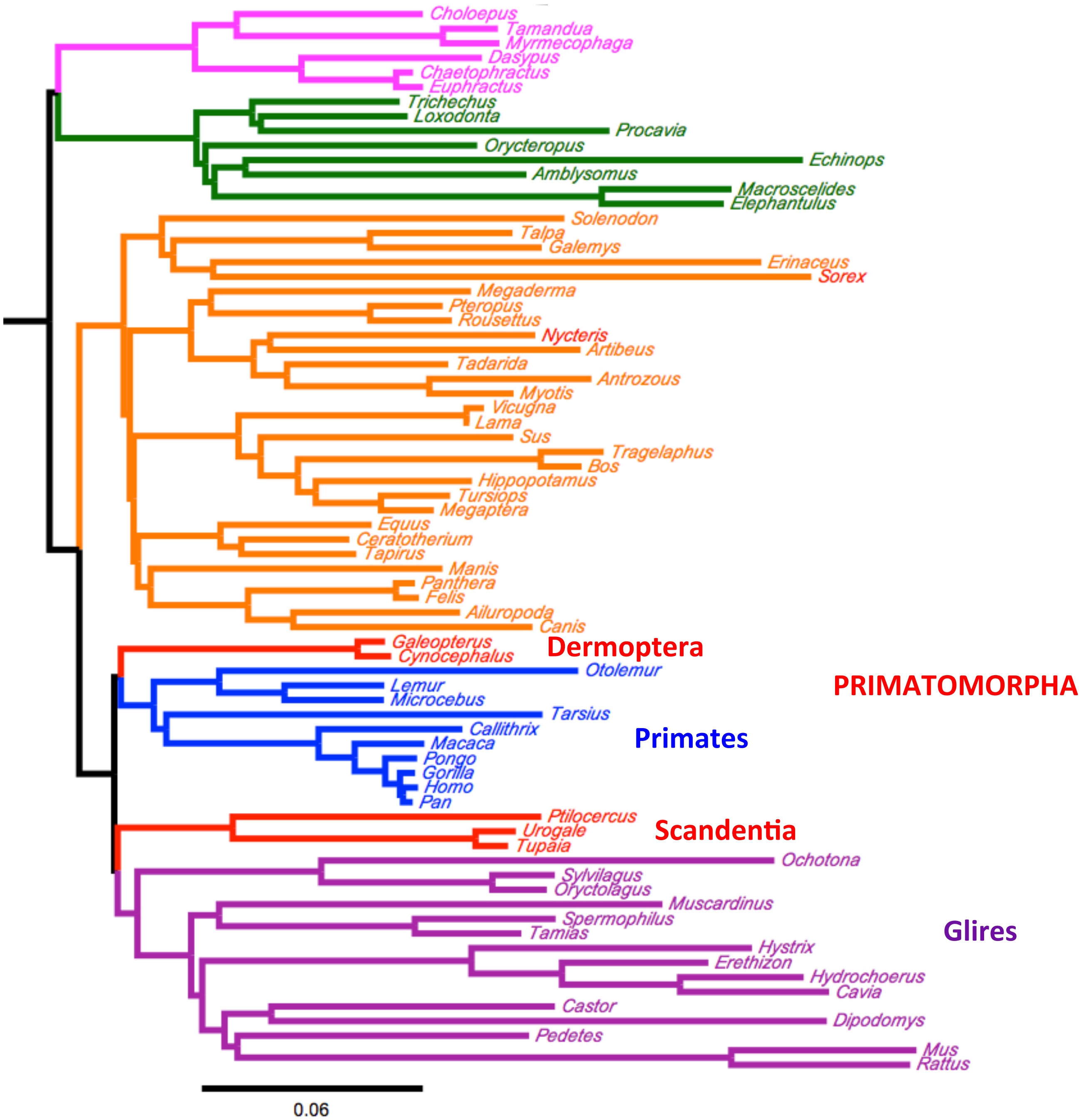
Applications of Sequence Capture Methods



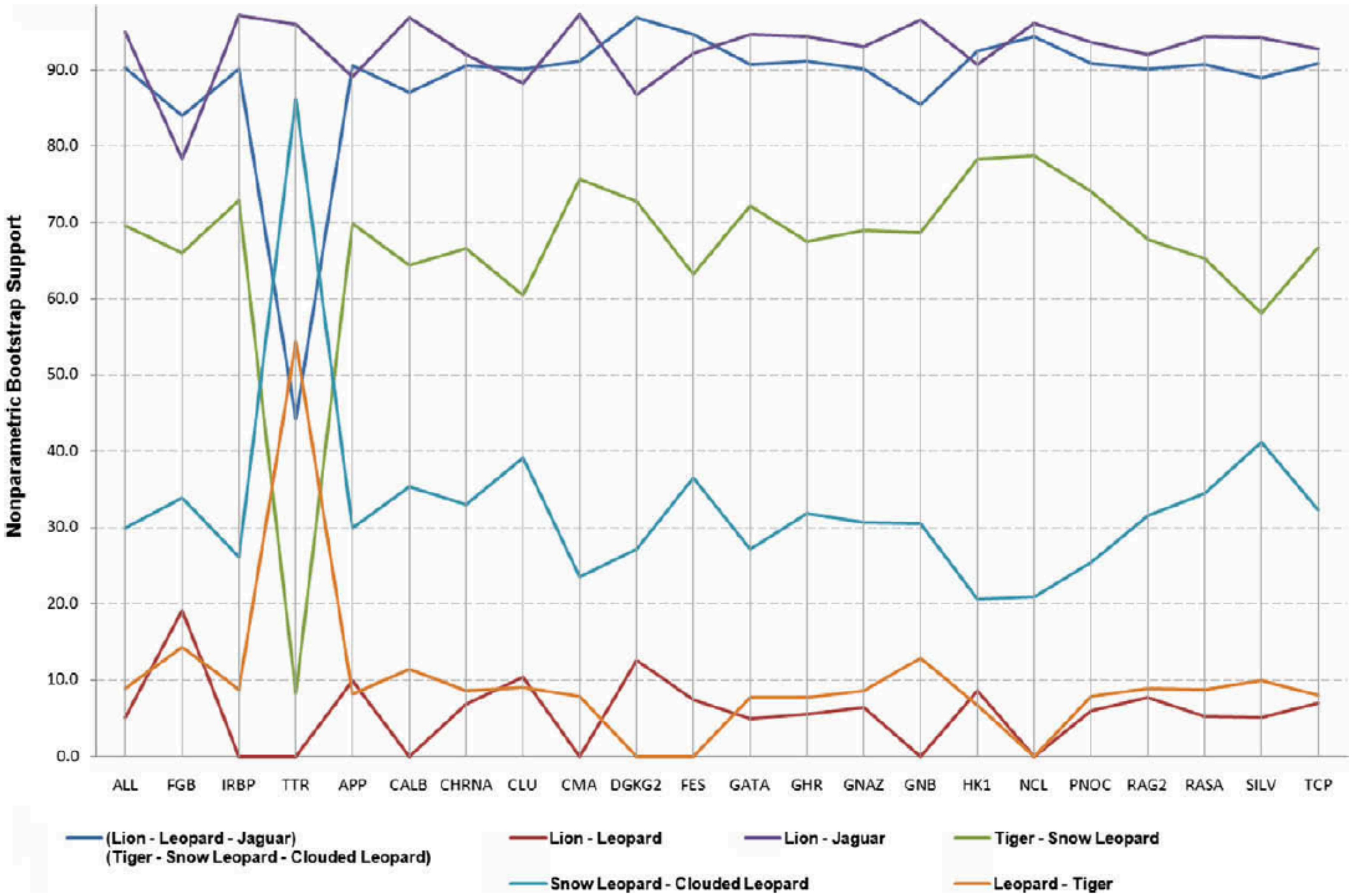
		Organism		Genomic target			Probe targets					Probe platform		Published application			Potential application		
Approach	**Reference**	Animal	Plant	Nuclear	Mitochondrial	Plastid	Exons	Introns	Intergenic	Conserved regions	Diverse probes	In-solution hybridization	On-array hybridization	Shallow	Intermediate	Deep	Shallow	Intermediate	Deep
Anchored enrichment	Lemmon et al. 2012	●	○	●	○	○	●	◒	◒	◒	●	●	○	○	●	●	◒	●	●
Ultraconserved element enrichment	Faircloth et al. 2012	●	○	●	○	○	●	◒	◒	●	○	●	○	○	●	◒	◒	●	●
Exon capture	Burbano et al. 2010	●	○	●	○	○	●	○	○	◒	○	○	●	●	○	○	●	○	○
	George et al. 2011	●	○	●	○	○	●	◒	◒	◒	○	●	○	●	○	○	●	○	○
	Hancock-Hanser et al. 2013	●	○	●	●	○	●	●	●	◒	○	○	●	●	●	○	●	●	○
	Li et al. 2013	●	○	●	○	○	●	○	○	◒	○	●	○	○	●	●	◒	●	●
	Bi et al. 2012	●	○	●	●	○	●	◒	○	◒	○	○	●	●	○	○	●	○	○
Organelle capture	Briggs et al. 2009	●	○	○	●	○	●	○	●	◒	○	●	○	●	○	○	●	○	○
	Mason et al. 2011	●	○	○	●	○	●	○	●	◒	○	●	○	●	○	○	●	○	○
	Guschanski et al. 2013	●	○	○	●	○	●	○	●	◒	○	●	○	●	○	○	●	○	○
	Cronn et al. 2012	○	●	○	○	●	●	●	●	◒	○	●	○	●	○	●	●	●	●
	Stull et al. 2013	○	●	○	○	●	●	●	●	◒	●	●	○	●	●	●	●	●	●

Lemmon & Lemmon (2013) *Annu. Rev. Ecol. Evol. Syst.*

SUNDACLEAN

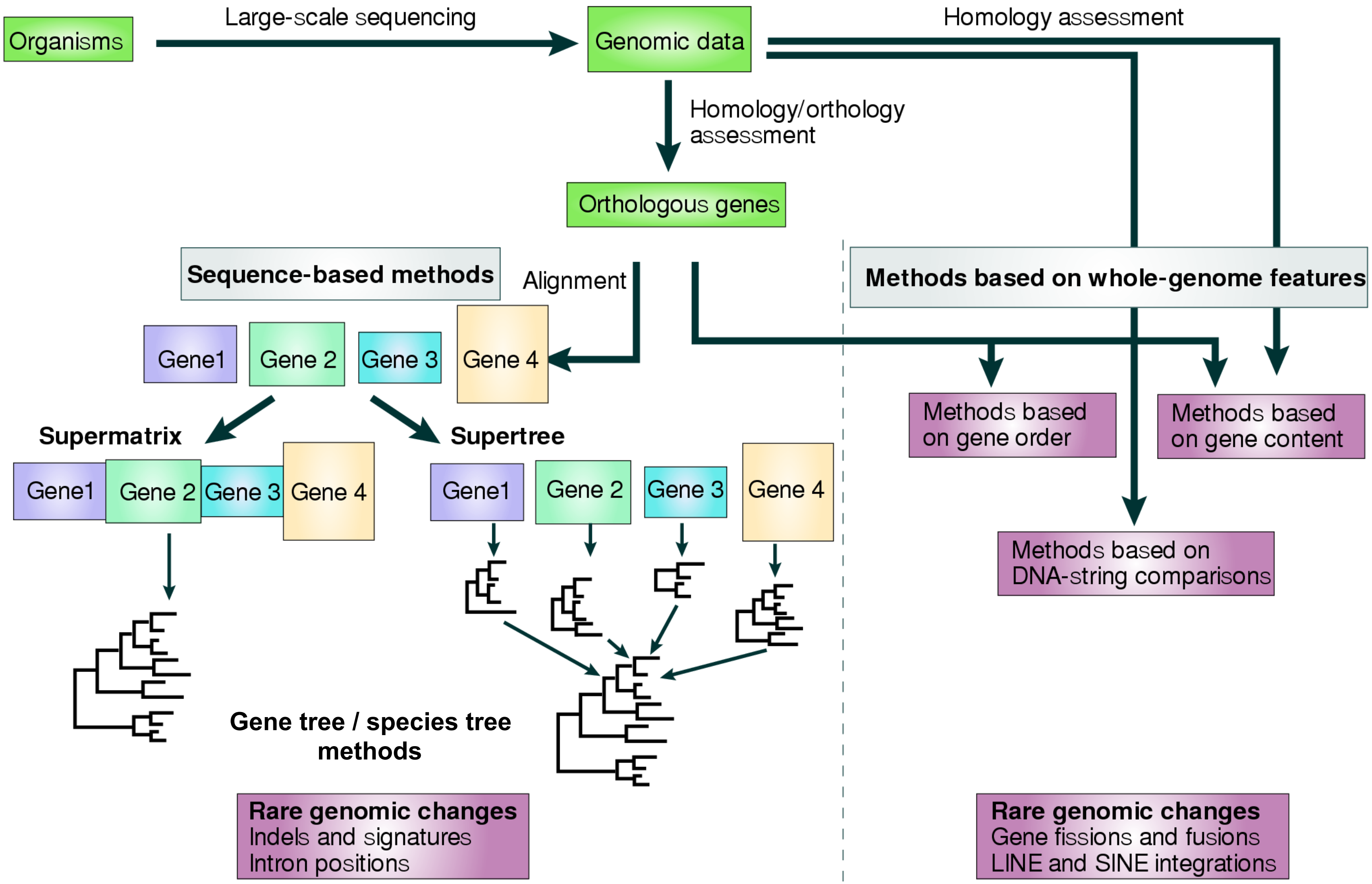

Choloepus
Tamandua
Myrmecophaga
Dasypus
Chaetophractus
Euphractus
Trichechus
Loxodonta
Procavia
Orycteropus
Echinops
Amblysomus
Macroscelides
Elephantulus
Solenodon
Talpa
Galemys
Erinaceus
Sorex
Megaderma
Pteropus
Rousettus
Nycteris
Artibeus
Tadarida
Antrozous
Myotis
Vicugna
Lama
Sus
Tragelaphus
Bos
Hippopotamus
Tursiops
Megaptera
Equus
Ceratotherium
Tapirus
Manis
Panthera
Felis
Ailuropoda
Canis
Galeopterus
Cynocephalus
Dermoptera
Otolemur
PRIMATOMORPHA
Lemur
Microcebus
Tarsius
Callithrix
Primates
Macaca
Pongo
Gorilla
Homo
Pan
Ptilocercus
Urogale
Tupaia
Scandentia
Ochotona
Sylvilagus
Oryctolagus
Muscardinus
Spermophilus
Tamias
Glires
Hystrix
Erethizon
Hydrochoerus
Cavia
Castor
Dipodomys
Pedetes
Mus
Rattus
0.06

A Jackknife Approach Based on the Supermatrix



Davis *et al.* (2010) *Mol. Phylogenet. Evol.*

Methods of Phylogenomic Inference



Organisms

Large-scale sequencing

Genomic data

Homology assessment

Homology/orthology assessment

Orthologous genes

Sequence-based methods

Alignment

Gene1

Gene 2

Gene 3

Gene 4

Supermatrix

Gene1

Gene 2

Gene 3

Gene 4

Supertree

Gene1

Gene 2

Gene 3

Gene 4

Gene tree / species tree methods

Rare genomic changes
Indels and signatures
Intron positions

Methods based on whole-genome features

Methods based on gene order

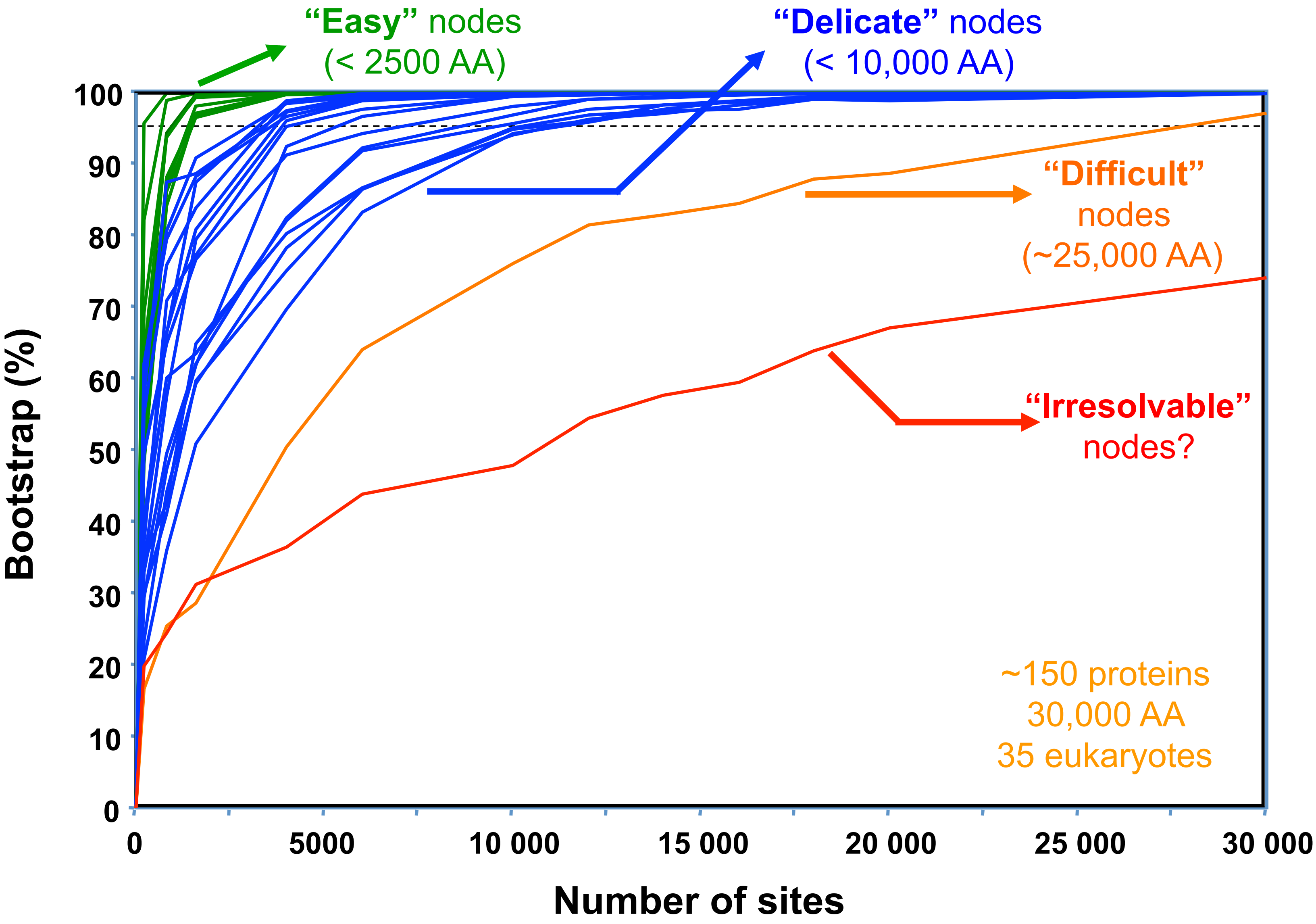
Methods based on gene content

Methods based on DNA-string comparisons

Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations

Delsuc, Brinkmann & Philippe (2005). *Nat. Rev. Genet.* 6: 361-375.

Phylogenomics Increases the Resolving Power



Delsuc, Brinkmann & Philippe (2005) *Nat. Rev. Genet.*