

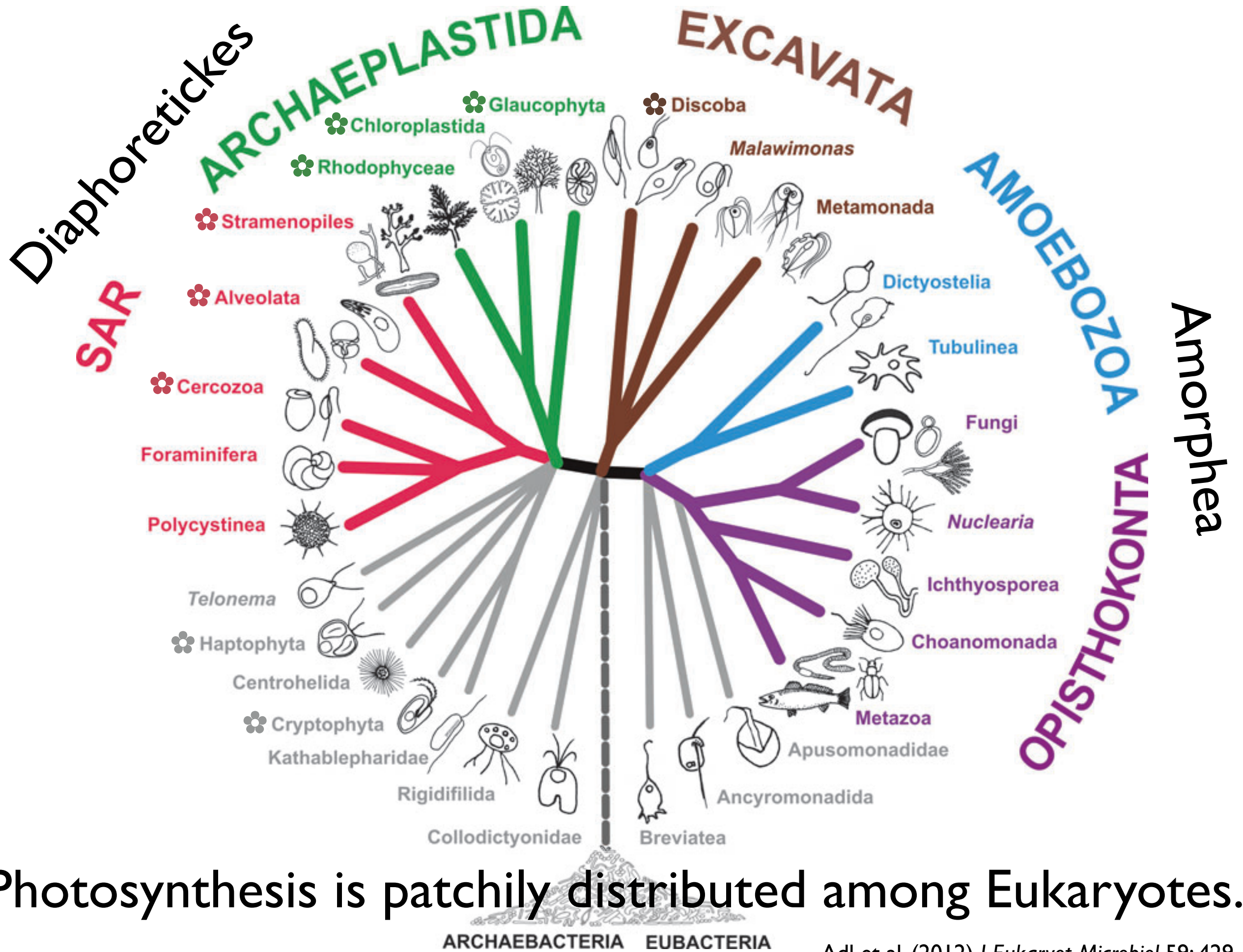
Pinpointing the origin of plastids using phylogenomics



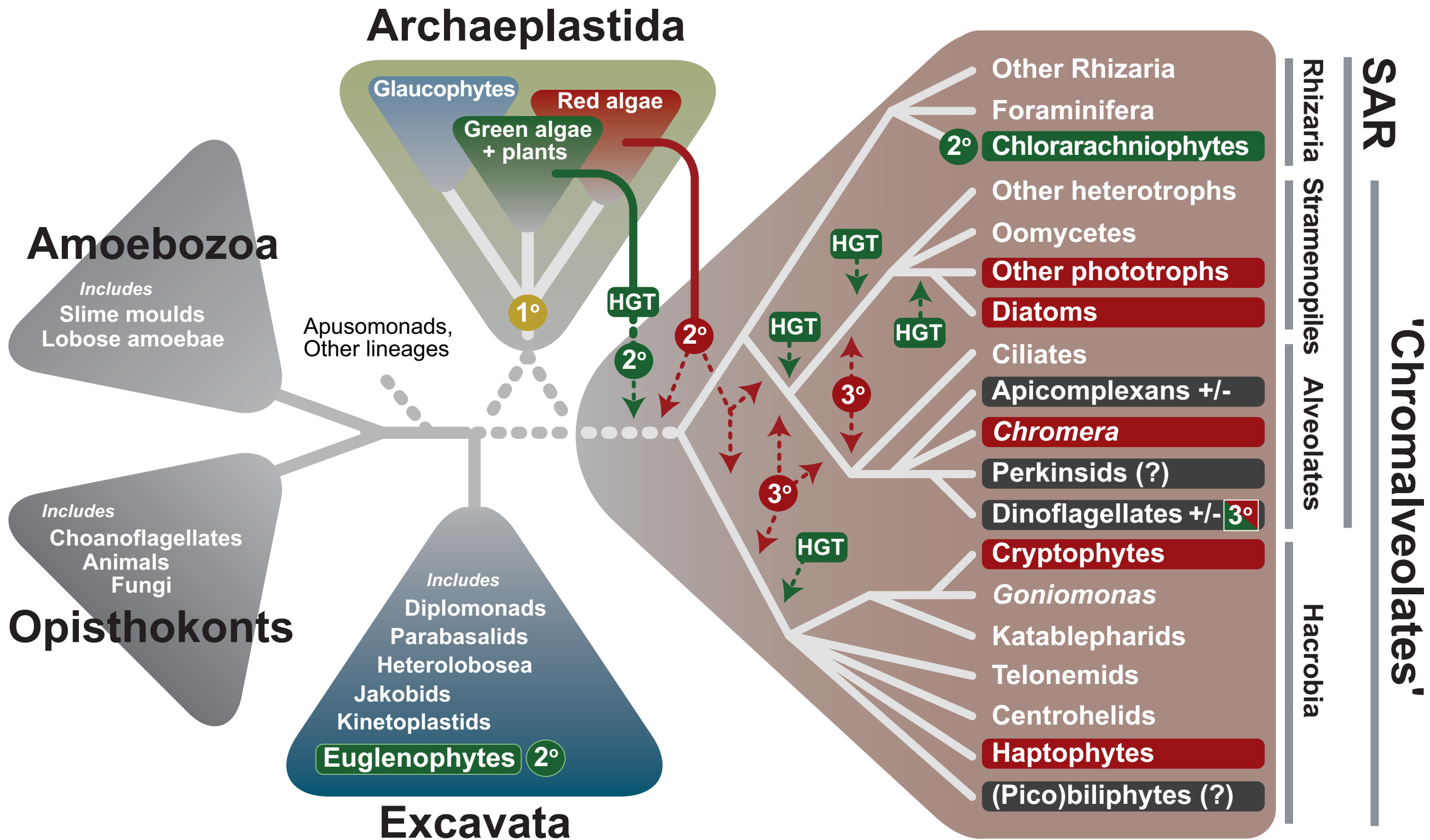
Denis BAURAIN

Eukaryotic Phylogenomics
PhytoSYSTEMS / University of Liège

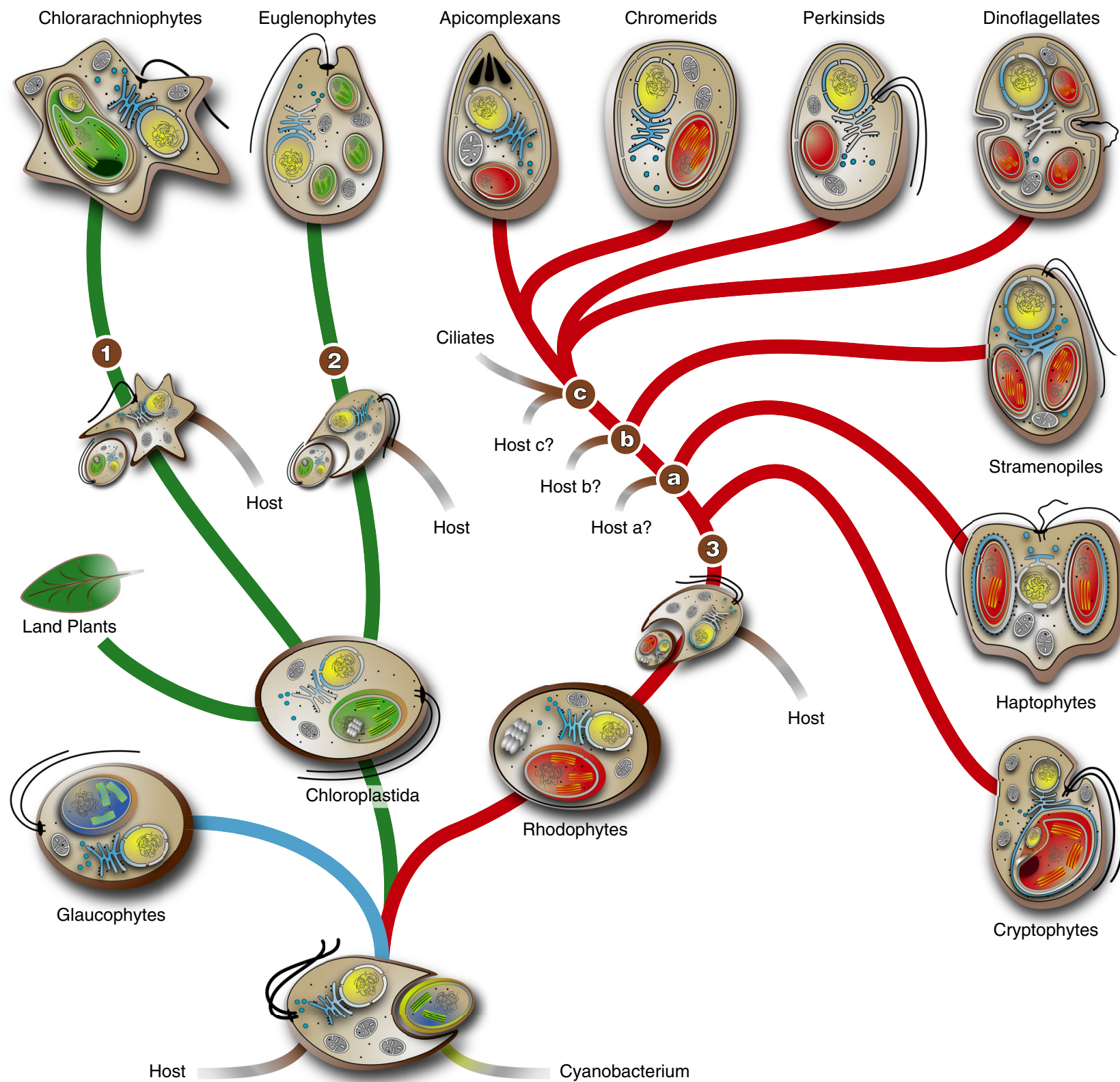
Lille, April 17th, 2015



Photosynthesis is patchily distributed among Eukaryotes.



Eukaryotes acquired their photosynthetic abilities through several endosymbiotic events.

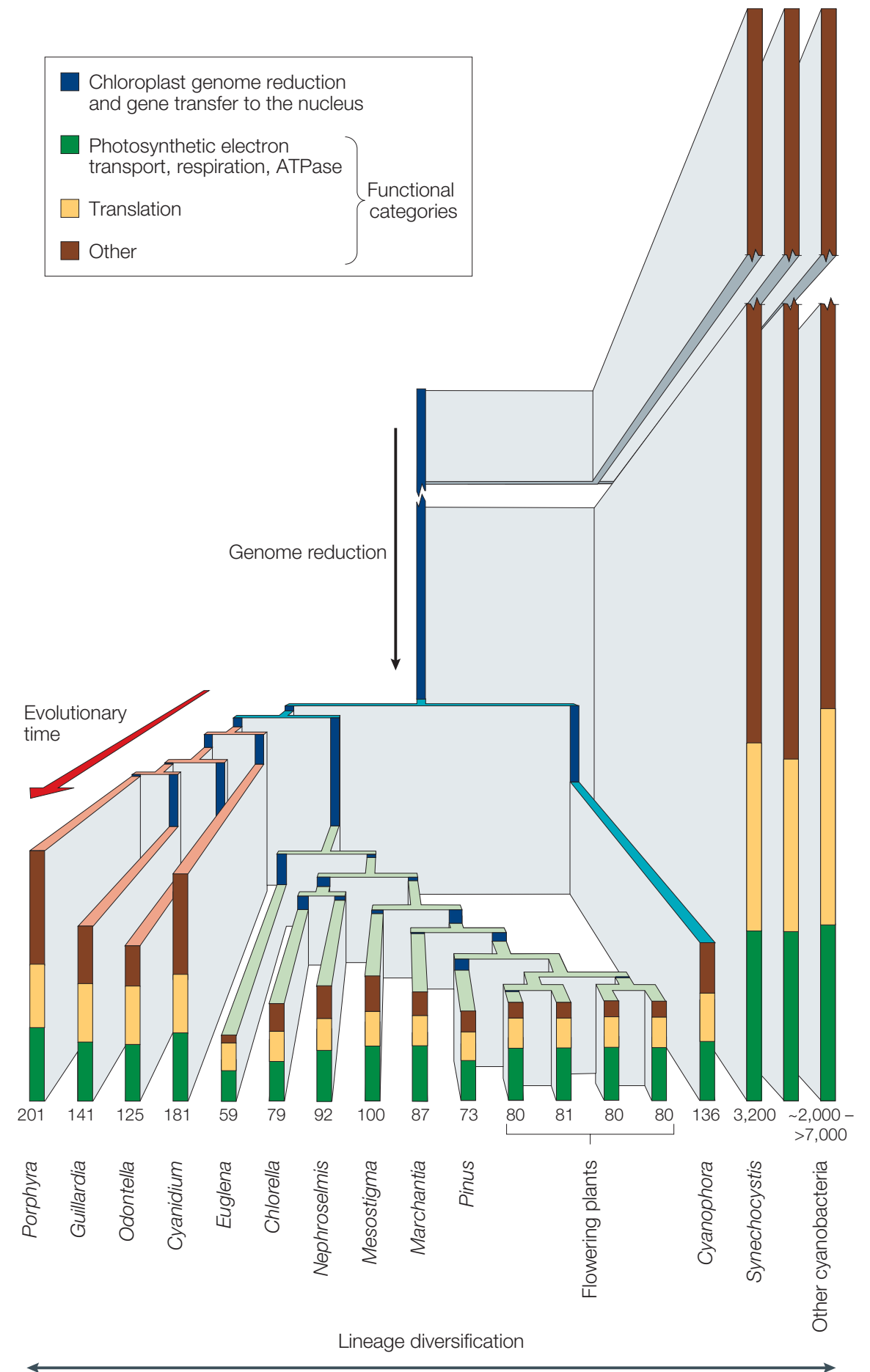


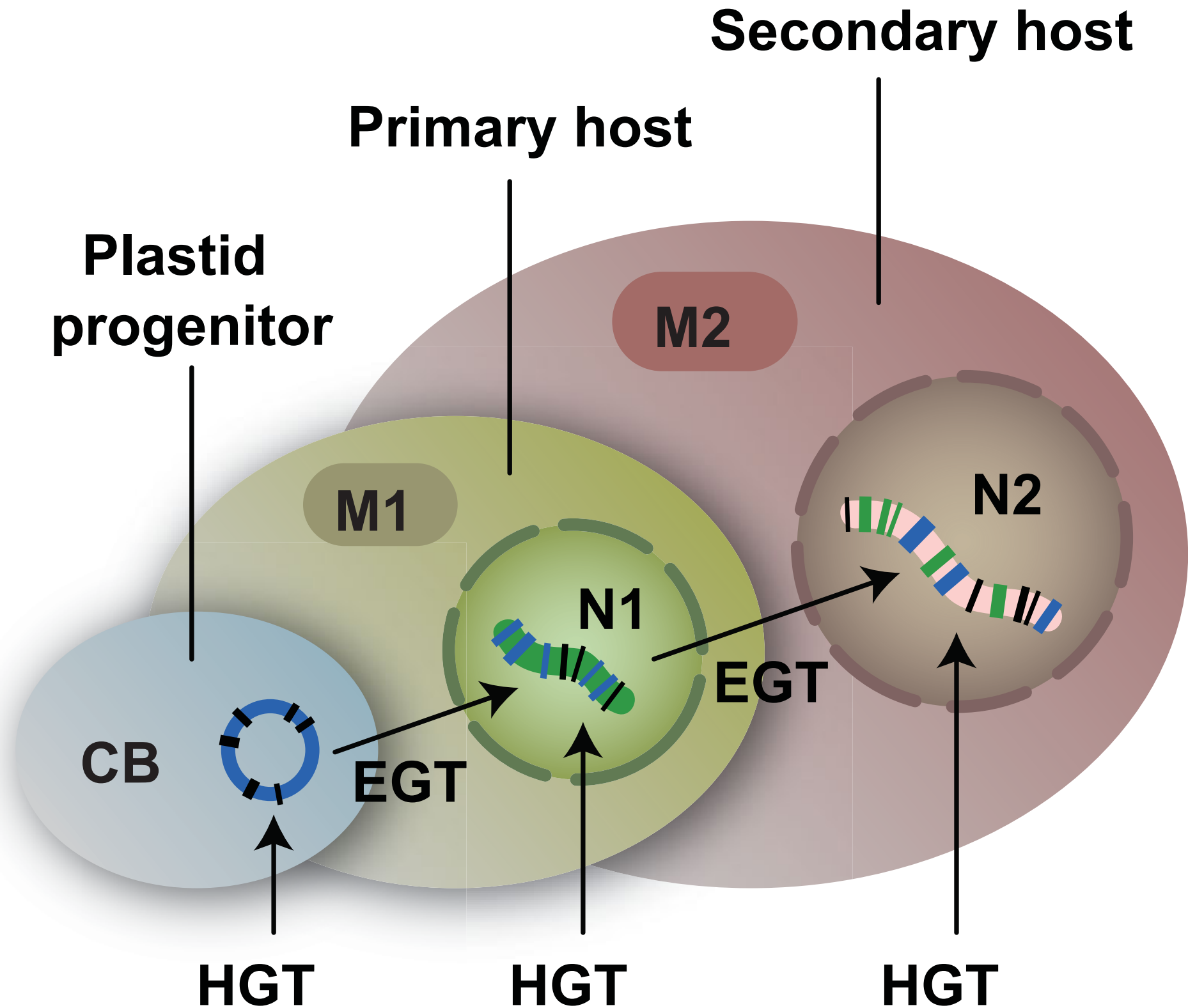
The single primary endosymbiont was a cyanobacterium, whereas subsequent endosymbioses involved algae.

Turning an endosymbiont into a plastid requires:

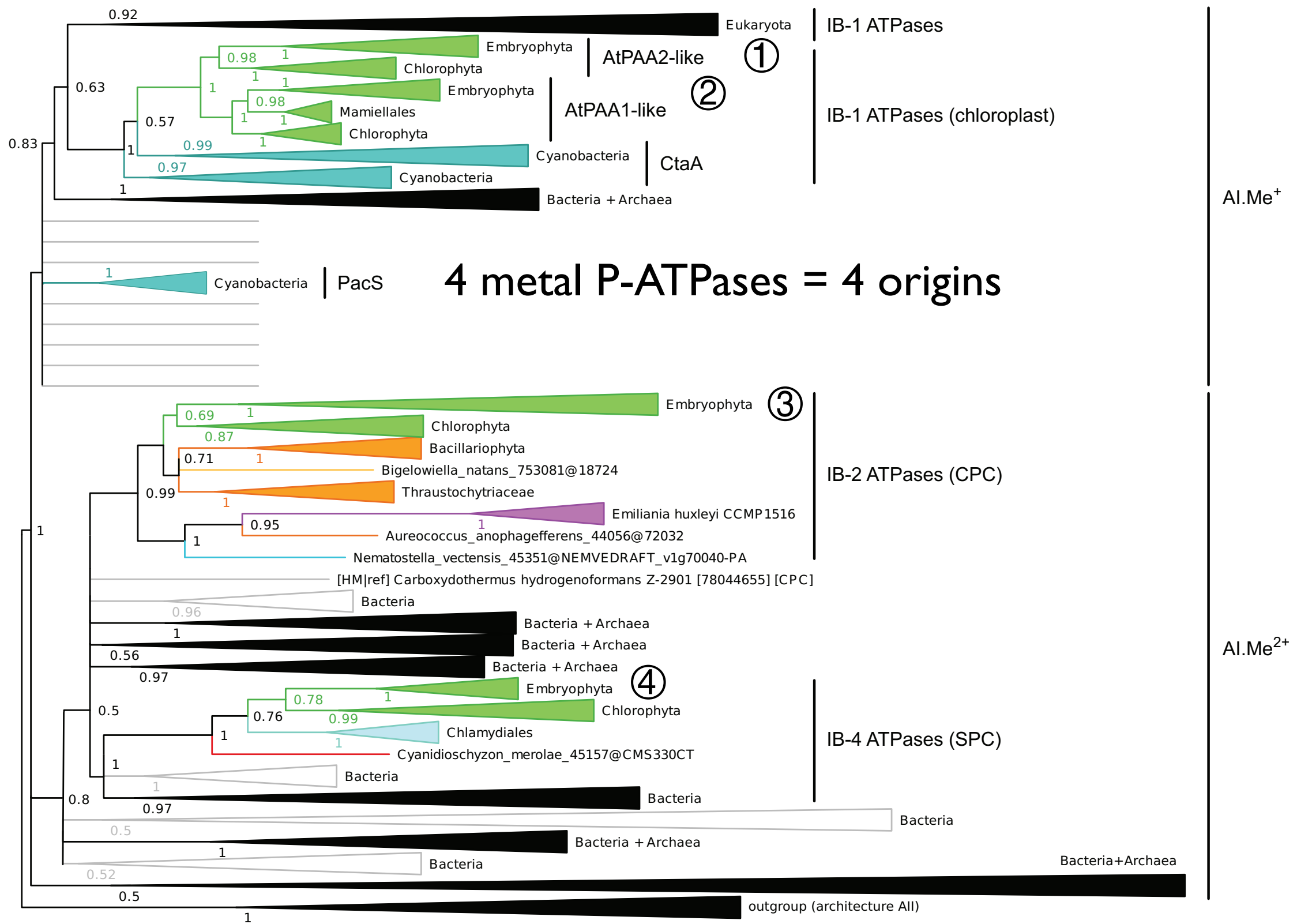
1. integration of the metabolisms
2. synchronization of the cell divisions
3. endosymbiotic gene transfer (EGT) coupled to a protein import apparatus

Hence the streamlined organellar genomes.





Algae are genetic chimeras, as most symbiont-encoded genes have been relocated to the host nuclei.



Phylogenetics is the most powerful tool we have for reconstructing gene histories and endosymbioses.

Open Questions

People in my lab work on most of them.

Plastid origins

- 1. What is the closest extant cyanobacterial lineage?*
- 2. What was the phenotype of the plastid ancestor?*
- 3. When did the primary endosymbiosis happen?*
- 4. Were there other contributors than Cyanobacteria?*
- 5. What was the incentive for the primary endosymbiosis?*

Eukaryotic evolution

- 6. What is the branching order among primary algae?*
- 7. How did primary plastids spread to other algae?*

Open Questions

But today we will focus on the first one.

Plastid origins

- 1. What is the closest extant cyanobacterial lineage?*
- 2. What was the phenotype of the plastid ancestor?*
- 3. When did the primary endosymbiosis happen?*
- 4. Were there other contributors than Cyanobacteria?*
- 5. What was the incentive for the primary endosymbiosis?*

Eukaryotic evolution

- 6. What is the branching order among primary algae?*
- 7. How did primary plastids spread to other algae?*

Phylogénomique des Eucaryotes

Luc Cornet

Contact

luc.cornet@ulg.ac.be

+32 0(4)3669040

Université de Liège

Département des Sciences de la Vie

Bld du Rectorat, 27, Bât. B22, +3

B-4000 Liège (Sart Tilman)

Belgique



Publications

[A Phylogenomic analysis of the origin of plastids](#)

Cornet, Luc; Javaux, Emmanuelle; Wilmotte, Annick; Philippe, Hervé; Baurain, Denis

Communication orale (2014, June 24)

[The unusual Gasteromycetes *Lycogalopsis solmsii* belongs to the gomphoid-phalloid group](#)

Demoulin, Vincent; Cornet, Luc; Delbruyère, Emilie; Baurain, Denis

in *Acta Mycologica* (2013), 48(1), 13-20

[A Middle Devonian Callixylon \(Archaeopteridales\) from Ronquières, Belgium](#)

Cornet, Luc; Gerrienne, Philippe; Meyer-Berthaud, Brigitte; Prestianni, Cyrille

in *Review of Palaeobotany and Palynology* (2012), 183

The content of this talk is the work of Luc Cornet.

Luc is a former paleobotanist now turned phylogeneticist.

Other contributors:


- Raphaël Léonard (Euk. Phyl.)
- Damien Sirjacobs (Euk. Phyl.)

- Emmanuelle Javaux (ULg)
- Annick Wilmotte (ULg)

- Hervé Philippe (CNRS, Moulis)

Outline

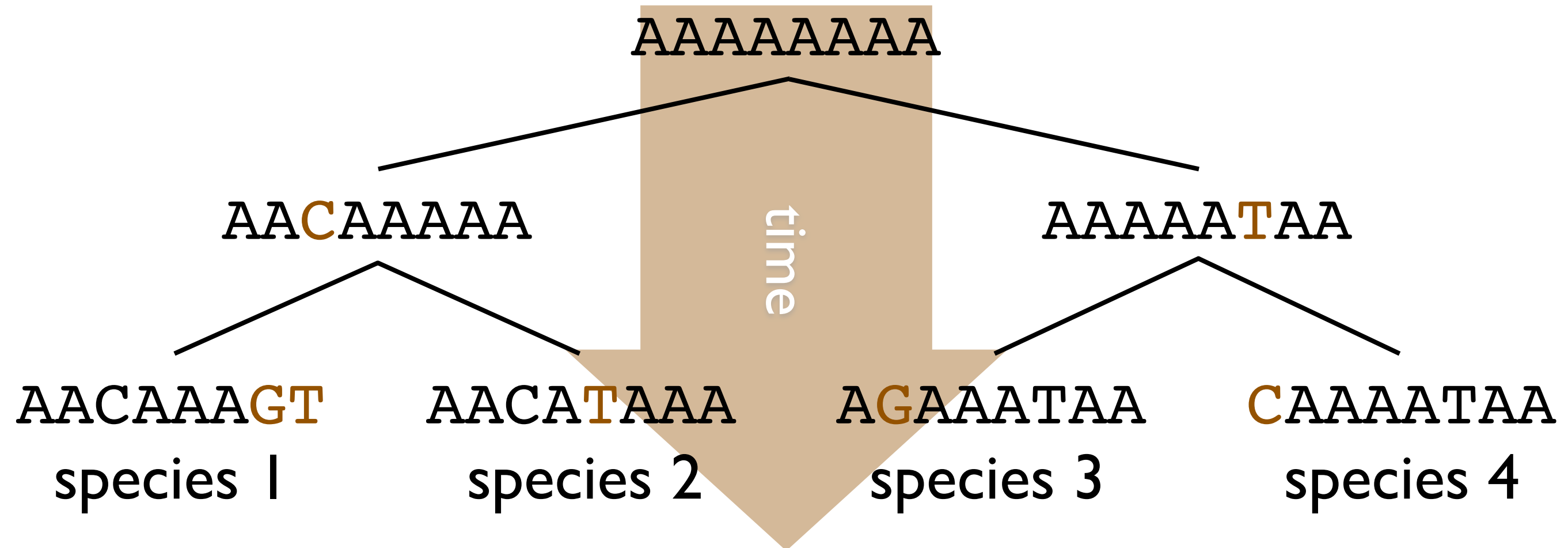
- 1. Crash Course on Molecular Phylogenetics*
- 2. Why Phylogenomics?*
- 3. Current Models for the Origin of Plastids*
- 4. Supermatrix Assembly – Pitfalls and Solutions*
- 5. Towards a Branching Point for Plastids*



*Crash Course
in Molecular Phylogenetics*

Phylogenetic Signal

Shared mutations are indicative of shared ancestry.



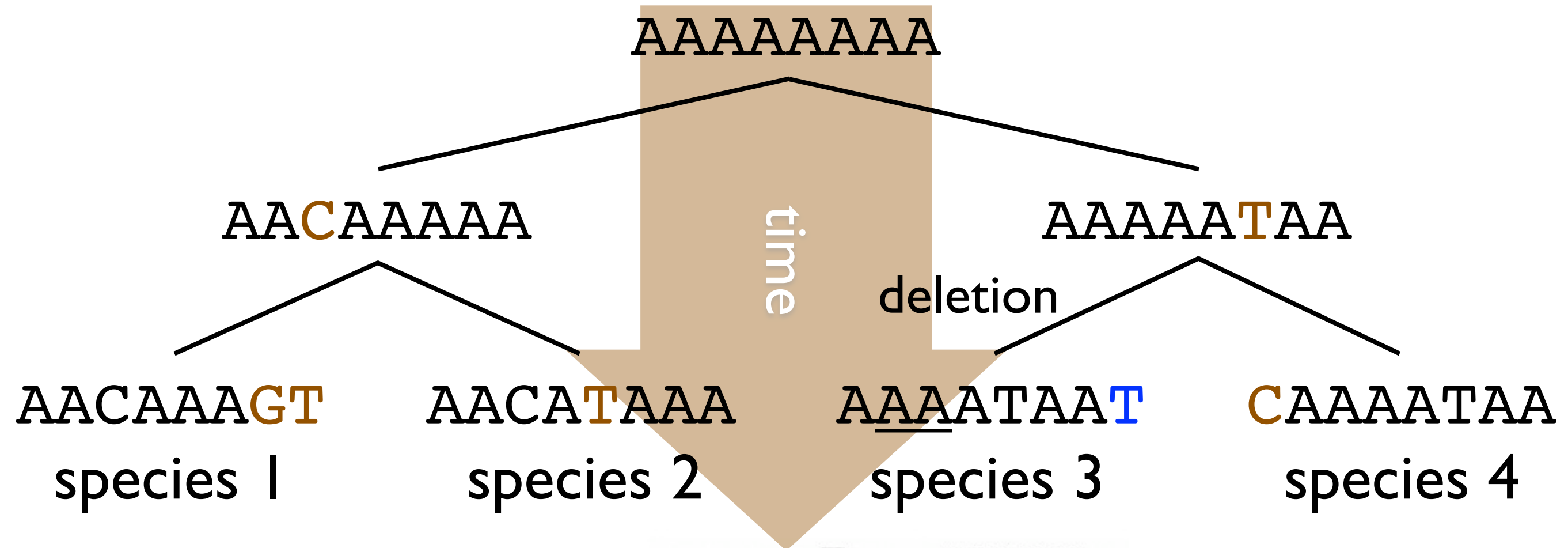
- 1 AACAAAAGT
- 2 AACATAAAA
- 3 AGAAAATAA
- 4 CAAAAATAA

New mutations at any one position are relatively **rare**.

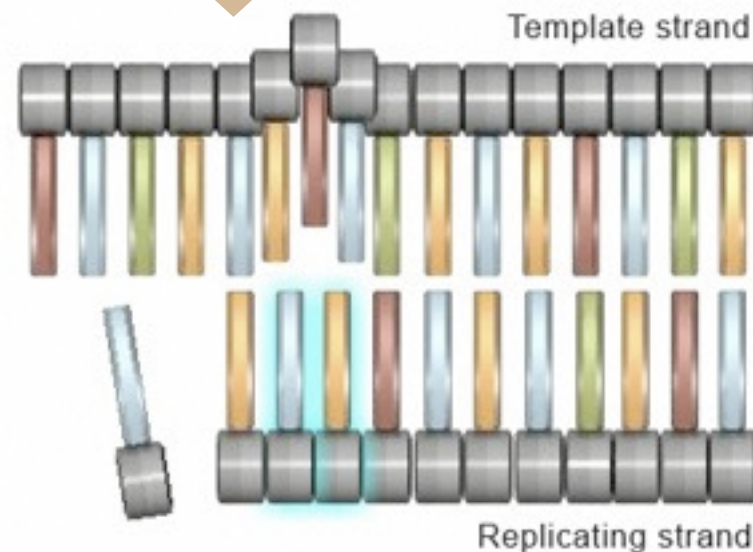
Most genetic differences between individuals are **inherited mutations**.

Phylogenetic Signal

... at least if we compare homologous positions!



1 AACAAAGT
2 AACATAAA
3 AAAATAAT
4 CAAAATAA



1 AACAAAGT
2 AACATAAA
3 AA-AAATAA
4 CAAAATAA

Sequence Alignment

example: alternative oxidase (protein seqs)

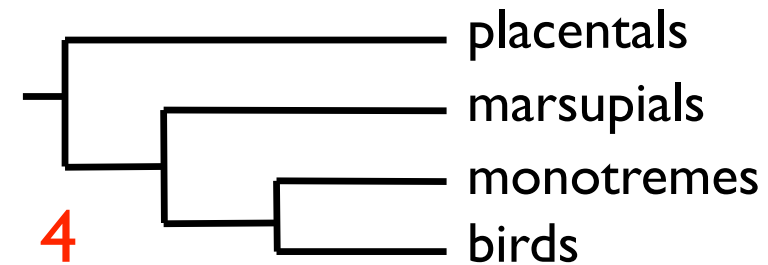
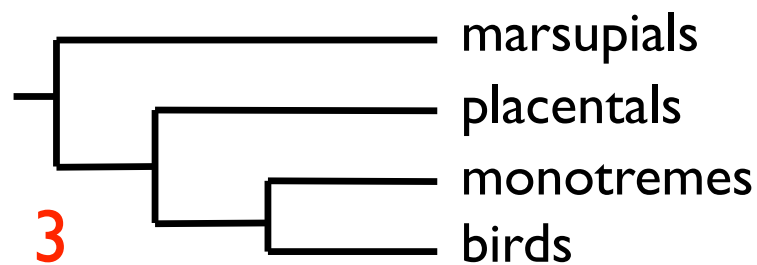
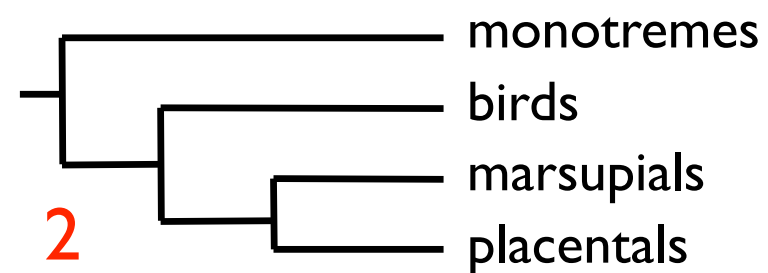
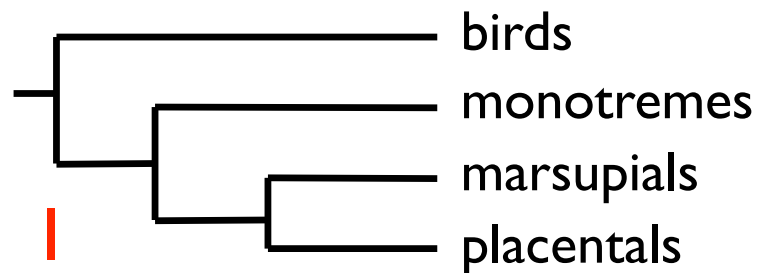
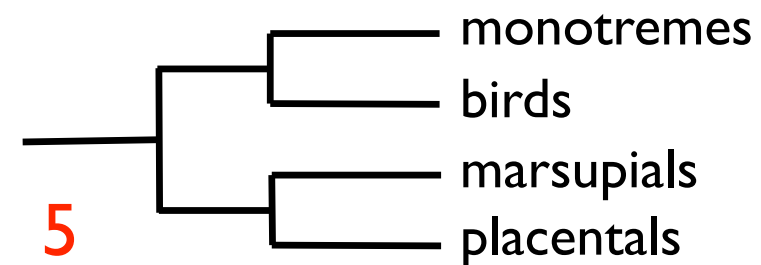
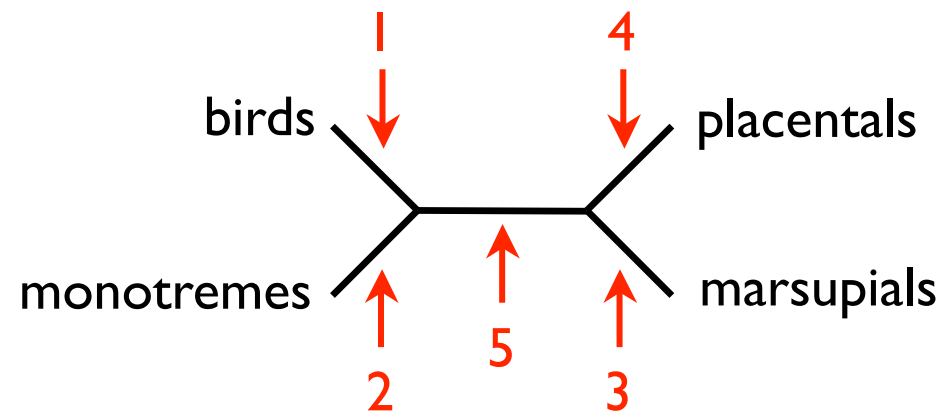
```
          100          110          120          130          140          150          160          170          180
Cr1  FAAGAVA P H P G I N P A R M A A D S A S A A A G A S G D A A L A E S Y M A H P A Y S D E Y V E S V R P T H V T P Q K L H Q H V G L R T I O V F R Y L F D K A T G Y T P T G S -
Cr2  F A T S G V A P H P G M K A P S P P T D D E V E A C W - - - - - R P V Y D T A Y L E K V K P F H I T P E R L Y Q R I G F R A I M A A R W T F D K L T G Y G P - - N -
Ac   A A T S N S N M R Y F S S T S R R W I K E F F A P P K E T D H I V E S V T T W K H P V F T E K O M K E I A I A H R E A K N W S D W V A L G T V R F L R W A T D L A T G Y R H A A - -
An   N R H Q T A G K R F I S T T P K S Q I K E F F - P P P T A P H V K E V E T A W V H P V Y T E E Q M K Q V A I A H R D A K N W A D W V A L G T V R M L R W G M D L V T G Y R H P P - -
Ca   E K P G T I P T K H K P F N I Q T E V Y N K A G I E A N D D K F L T K P T Y R H E D F T E A G V Y R V H V T H R P P R T I G D K I S C Y G T L F F R K C F D L V T G Y A V P D P -
Mg   L P R L A A S P R L F S T T S S A Q L R D F F - P V K E T E H I R O T P P T W P H H G L T E K E M V D V V P G H R K P R T L G D K F A W S L V R I S R W G M D K V S G L S S E Q Q Q
Nc   T N L S S P S P R N F S T T S V T R L K D F F - P A K E T A Y I R O T P P A W P H H G W T E E E M T S V V P E H R K P E T V G D W L A W K L V R I C R W A T D I A T G I R P E - Q Q
Pa   P S P H S K D P N S K S I F D I G T K L I V N P P P O M A D N O Y V T H P L F P H P K Y S D E D C E A V H F V H R E P K T I G D K I A D R G V K F C R A S F D F V T G Y K K P K D V
At   G G D A A G G N N K G D K G I A S Y W G V E P N K I T K E D G S E W K W N C F R P W E T Y K A D I T I D L K K H H V P T T F L D R I A Y W T V K S L R W P T D L - - - - -
Sg   D G G A E - - - - - K E A V V S Y W A V P P S K V S K E D G S E W R W T C F R P W E T Y Q A D L S I D L H K H H V P T T I L D K L A L R T V K A L R W P T D I - - - - -
          *
```

```
          190          200          210          220          230          240          250          260          270
Cr1  - - - - - M T E A Q W L R R M I F L E T V A G C P G M V A G M L R H L K S L R S M S R D R G W I H T L L E E A E N E R M H L I T F L Q L R Q P G P A F R A M V I
Cr2  - - - - - M T E A K W L Q R M I F L E T I A G V P G M V A G V L R H L K S L R S M K R D H G W I H T L L Q E A E N E R M H L L T F F E L R K P G P L F R A S I I
Ac   - - P G K Q G V E V P E Q F O M T E R K W V I R F I F L E T V A G V P G M V G G M L R H L R S L R R M K R D N G W I E T L L E E A Y N E R M H L L S F L K L A Q P G W F M R L M V L
An   - - P G R - - E H E A R F K M T E Q K W L T R F I F L E S V A G V P G M V G G M L R H L R S L R R M K R D N G W I E T L L E E A Y N E R M H L L T F L K L A E P G W F M R L M V L
Ca   - D K P D Q Y K G T - - R W E M T E E K W M T R C I F L E S I A G V P G S V A G F V R H L H S L R M L T R D K A W I E T L H D E A Y N E R M H L L T F I K I G K P S W F T R S I I Y
Mg   I N K G S P T T S I V A A K P L T E A Q W L S R F I F L E S I A A V P G M V A G M L R H L H S L R R L K R D N G W I E T L L E E A Y N E R M H L L T F L K M C E P G W L M K I L I I
Nc   V D K H H P T T A T S A D K P L T E A Q W L V R F I F L E S I A G V P G M V A G M L R H L H S L R R L K R D N G W I E T L L E E S Y N E R M H L L T F M K M C E P G L L M K T L I L
Pa   N G M L K S W E G T - - R Y E M T E E K W L T R C I F L E S V A G V P G M V A F I R H L H S L R L L K R D K A W I E T L L D E A Y N E R M H L L T F I K I G N P S W F T R F I I Y
At   - - - - - F F Q R R Y G C R A M M L E T V A A V P G M V G G M L L H C K S L R R F E Q S G G W I K A L L E E A E N E R M H L M T F M E V A K P K W Y E R A L V I
Sg   - - - - - F F Q R R Y A C R A M M L E T V A A V P G M V G G V L L H L K S L R R F E H S G G W I R A L L E E A E N E R M H L M T F M E V A Q P R W Y E R A L V L
          *
```

```
          280          290          300          310          320          330          340          350          360
Cr1  L A O G V F F N A Y F I A Y L L S P R T C H A F V G F L E E E A V K T Y T H A L V E I D A G R L - - - - W K D T P A P P V A V Q Y W G L K P G - A N M R D L I L A V R A D E A C H A
Cr2  V A O G V F W N L Y F I G Y L V S P R T C H A A V G F L E E E A V K T Y T H A L Q E I D A G R L - - - - W K G K V A P P I A C E Y W G L K P G - A S M R D L I L A V R A D E A C H A
Ac   G A O G V F F N G F F I S Y L I S P R T C H R F V G Y L E E E A V M T Y T H A I K D L E S G K L P N W - - A N O P A P D I A V A Y W Q M P E G K R T I L D L Y Y I R A D E A K H R
An   G A O G V F F N G F F L S Y L M S P R I C H R F V G Y L E E E A V I T Y T R A I K E I E A G S L P A W - - E K T E A P E I A V O Y W K M P E G Q R S M K D L L Y V R A D E A K H R
Ca   I G O G V F T N I F F L V Y L M N P R Y C H R F V G Y L E E E A V R T Y T H L I D E L D D P N K L P - D F O K L P I P N I A V O Y W P E L T P E S S F K D L I L R I R A D E A K H R
Mg   G A O G V Y F N A M F V A Y L I S P K I C H R F V G Y L E E E A V H T Y T R S I E E L E R G D L P K W S D P K F O V P E I A V S Y W G M P E G H R T M R D L L Y I R A D E A N H R
Nc   G A O G V F F N A M F L S Y L I S P K I T H R F V G Y L E E E A V H T Y T R C I R E I E E G H L P K W S D E K F E I P E M A V R Y W R M P E G K R T M K D L I H Y I R A D E A V H R
Pa   M G O G V F A N L F F L V Y L I K P R Y C H R F V G Y L E E E A V S T Y T H L I K D I D S - K R L P - K F D D V N L P E I S W L Y W T D L N E K S T F R D L I Q R I R A D E S K H R
At   T V O G V F F N A Y F L G Y L I S P K F A H R M V G Y L E E E A I H S Y T E F L K E L D K G N I - - - - E N V P A P A I A I D Y W R L P A D - A T L R D V V M V V R A D E A H H R
Sg   A V O G V F F N A Y F L G Y L L S P K F A H R V V G Y L E E E A I H S Y T E F L K D I D N G A I - - - - Q D C P A P A I A L D Y W R L P O G - S T L R D V V T V V R A D E A H H R
```

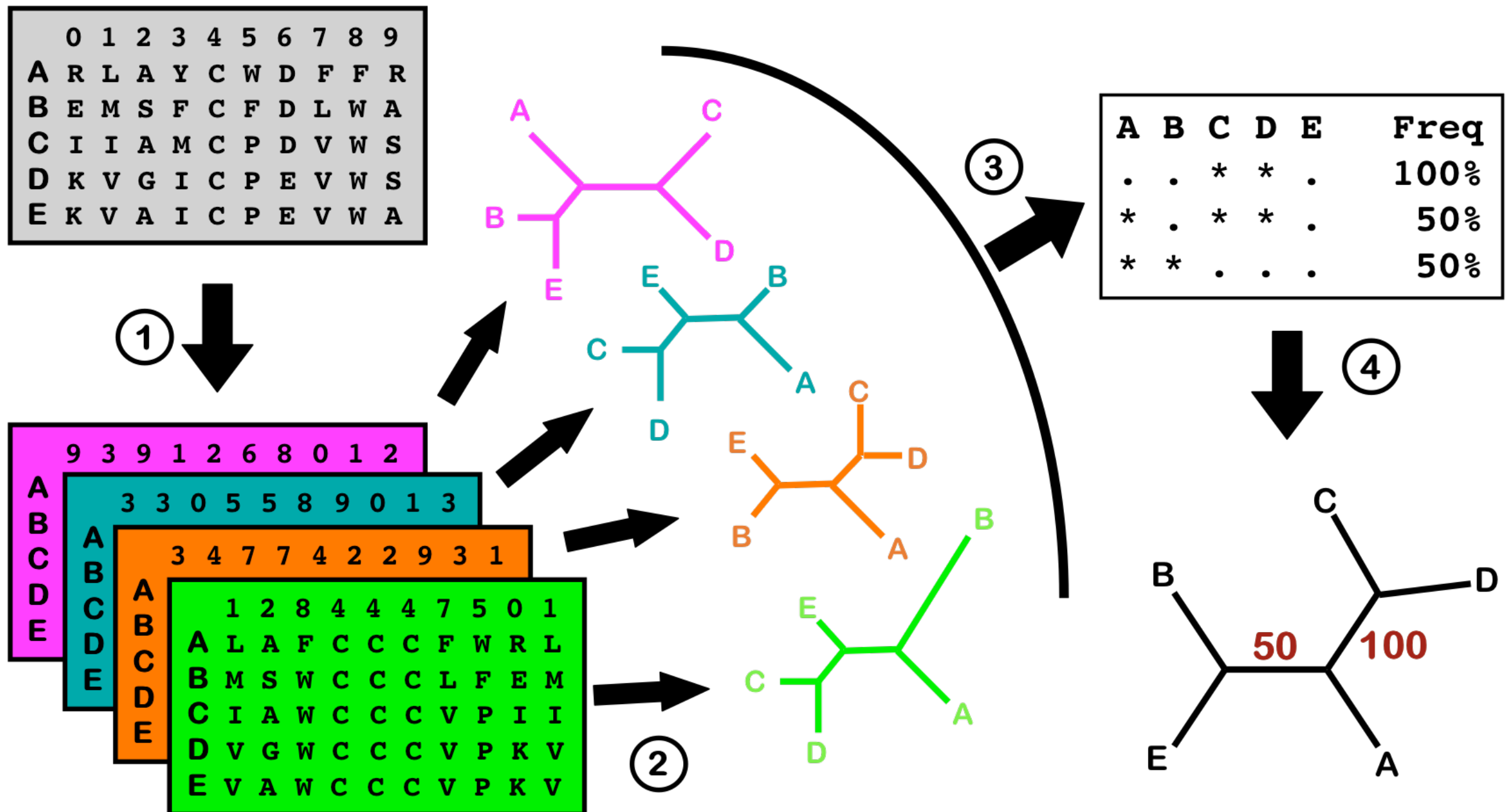
Phylogenetic Trees

unrooted vs. rooted trees / outgroup



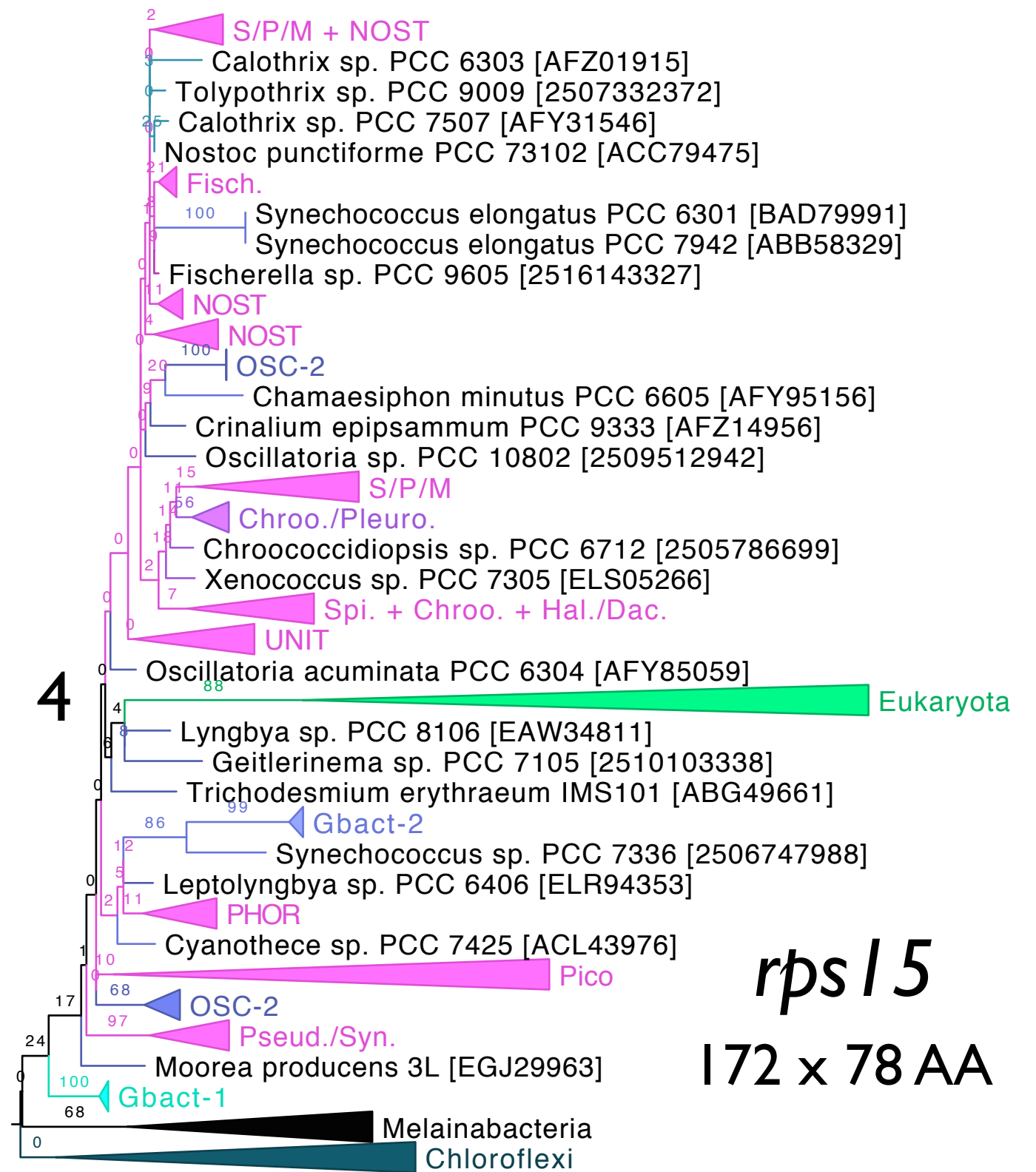
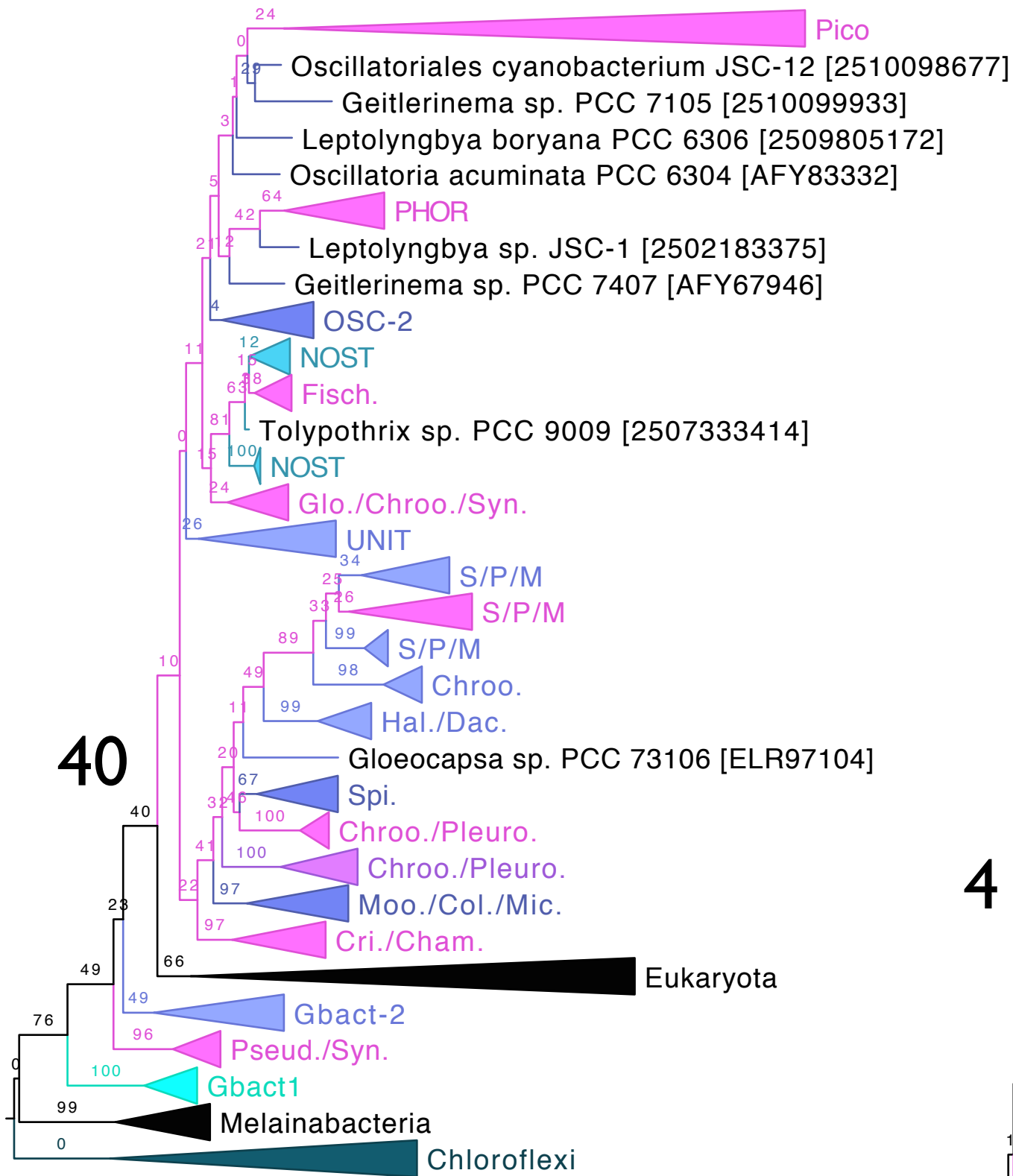
Phylogenetic Trees

statistical support / bootstrap proportions



A large, bold, brown number '2' is centered on the page. The text 'Why Phylogenomics?' is overlaid on the number.

Why Phylogenomics?

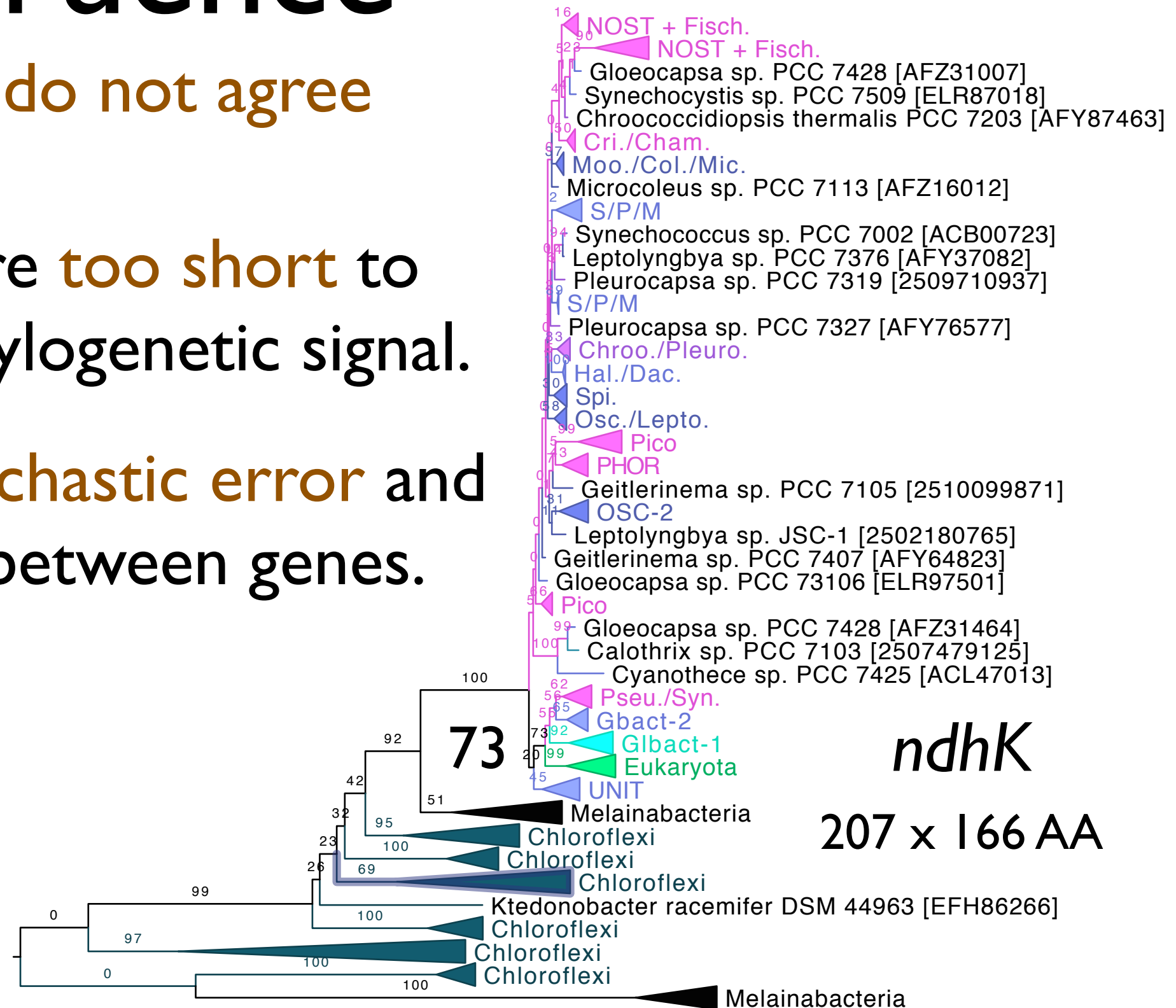


Incongruence

when genes do not agree

Single genes are **too short** to record much phylogenetic signal.

This leads to **stochastic error** and incongruence between genes.



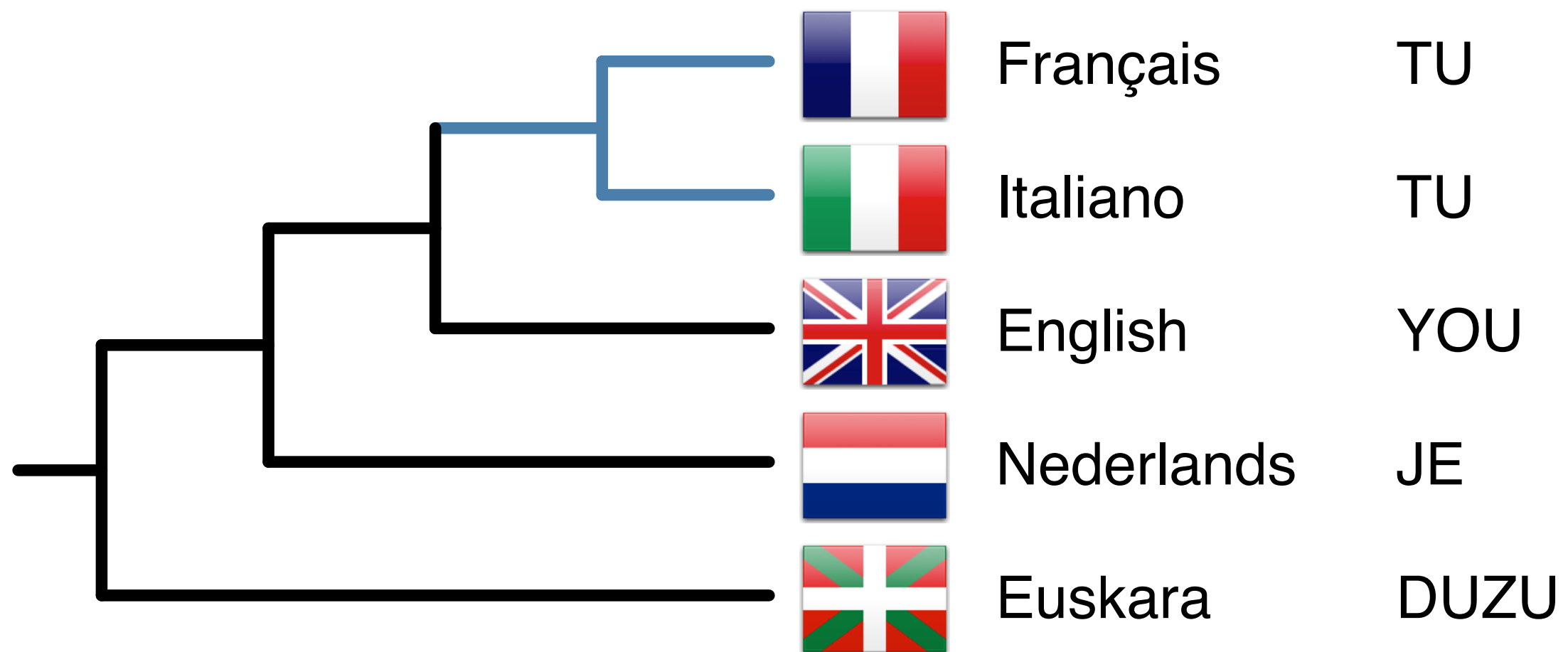
ndhK
207 x 166 AA

Incongruence

a linguistic analogy

Potential relationships among European languages

based on the naive analysis of 1 word

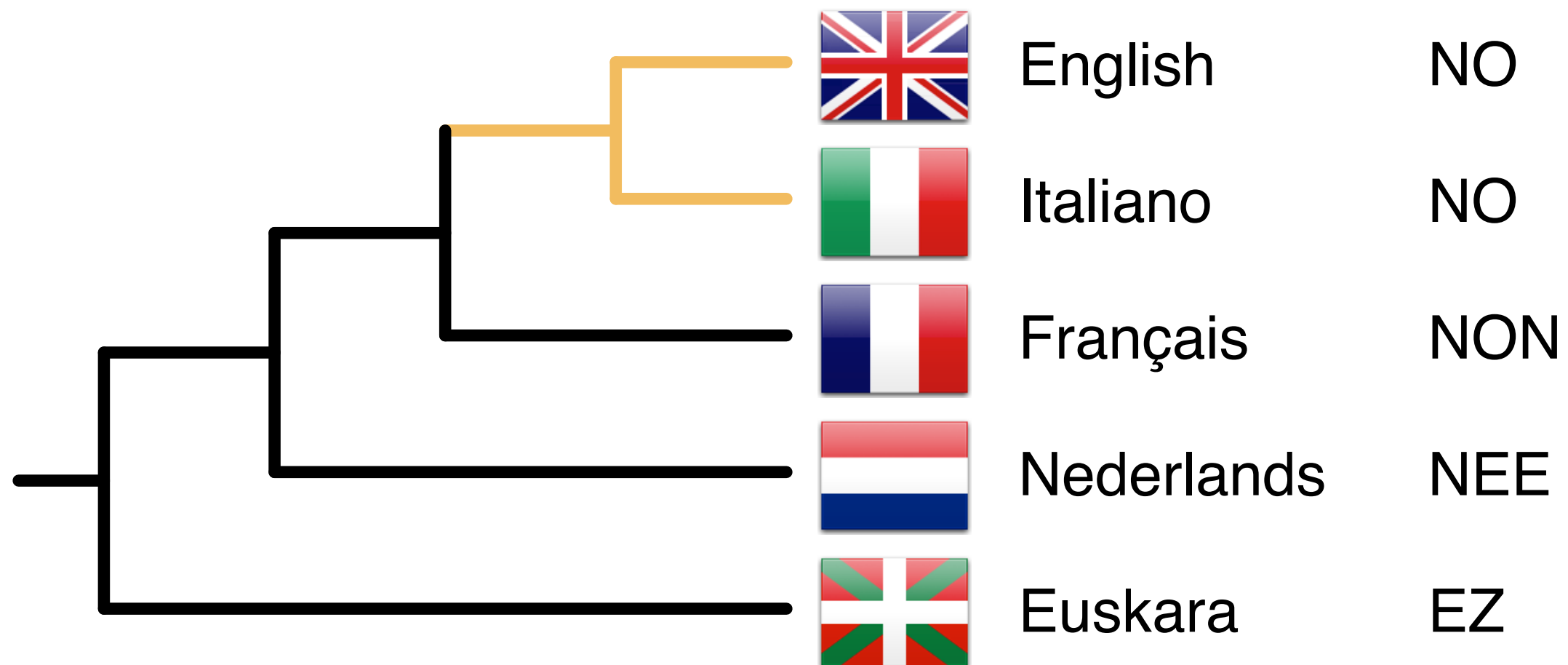


Incongruence

a linguistic analogy

Potential relationships among European languages

based on the naive analysis of 1 word



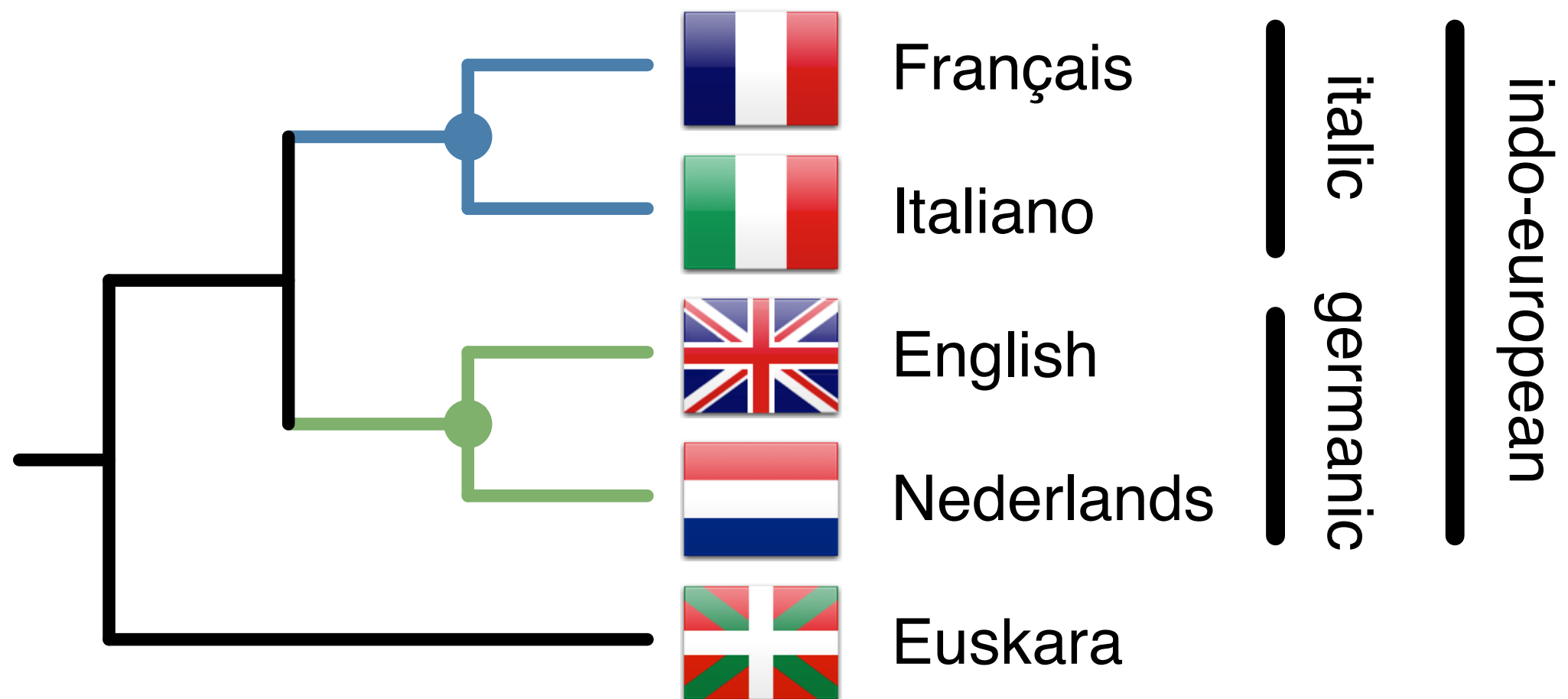
	français	italiano	english	nederlands	euskara
1	un	uno	one	een	bat
2	deux	due	two	twee	bi
3	trois	tre	three	drie	hiru
4	je	io	I	ik	ni
5	tu	tu	you	je	duzu
6	qui ?	chi?	who?	wie?	nor?
7	oui	si	yes	ja	bai
8	non	no	no	nee	ez
9	mère	madre	mother	moeder	ama
10	père	padre	father	vader	aita
11	dent	dente	tooth	tand	hortz
12	coeur	cuore	heart	hart	bihotza
13	pied	piede	foot	voet	oinez
14	souris	topolino	mouse	muis	saguaren

Incongruence

a linguistic analogy

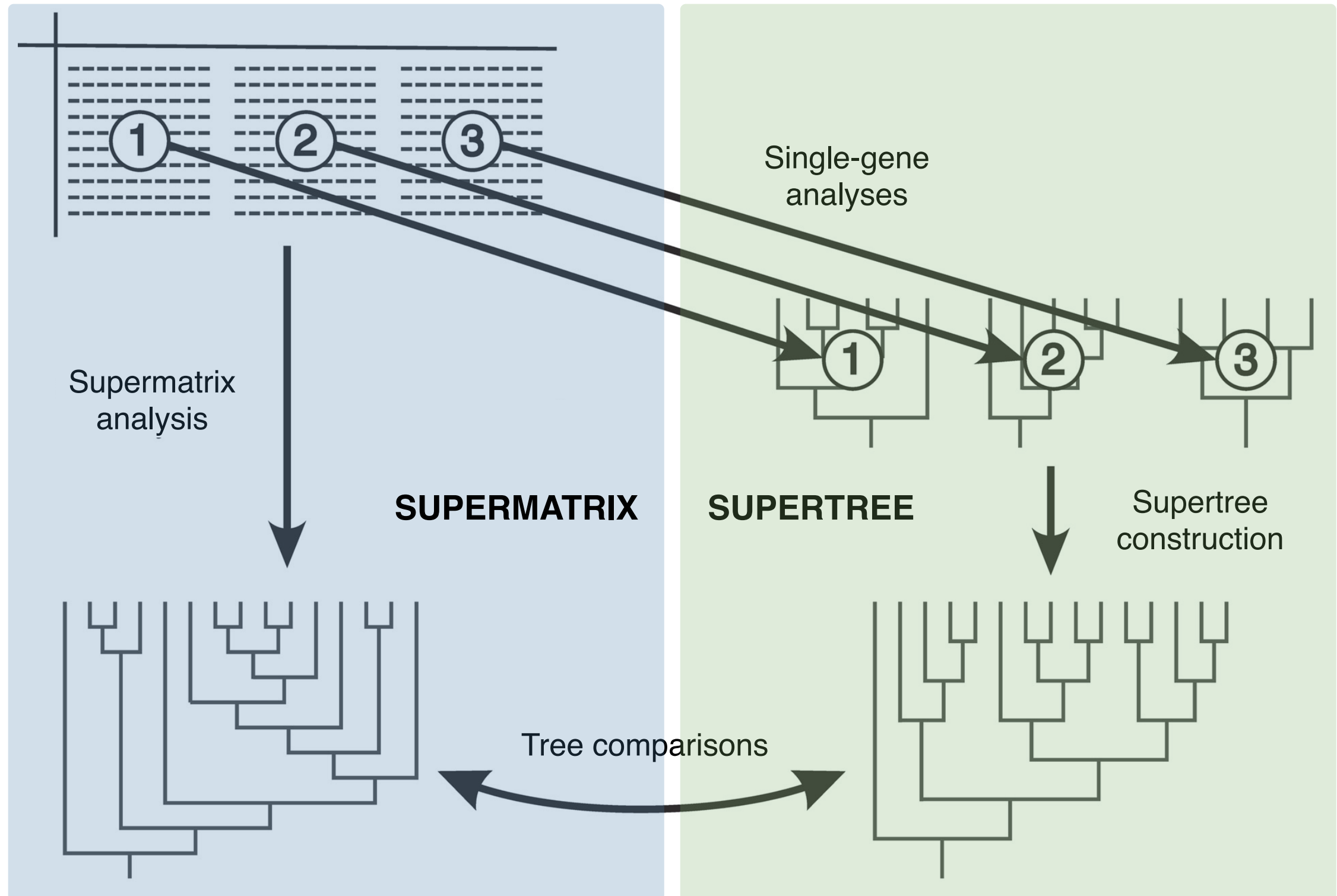
Known relationships among European languages

strongly supported by the naive analysis of 14 words



Phylogenomics

use of genome-scale data for inferring phylogenies



Phylogenomics

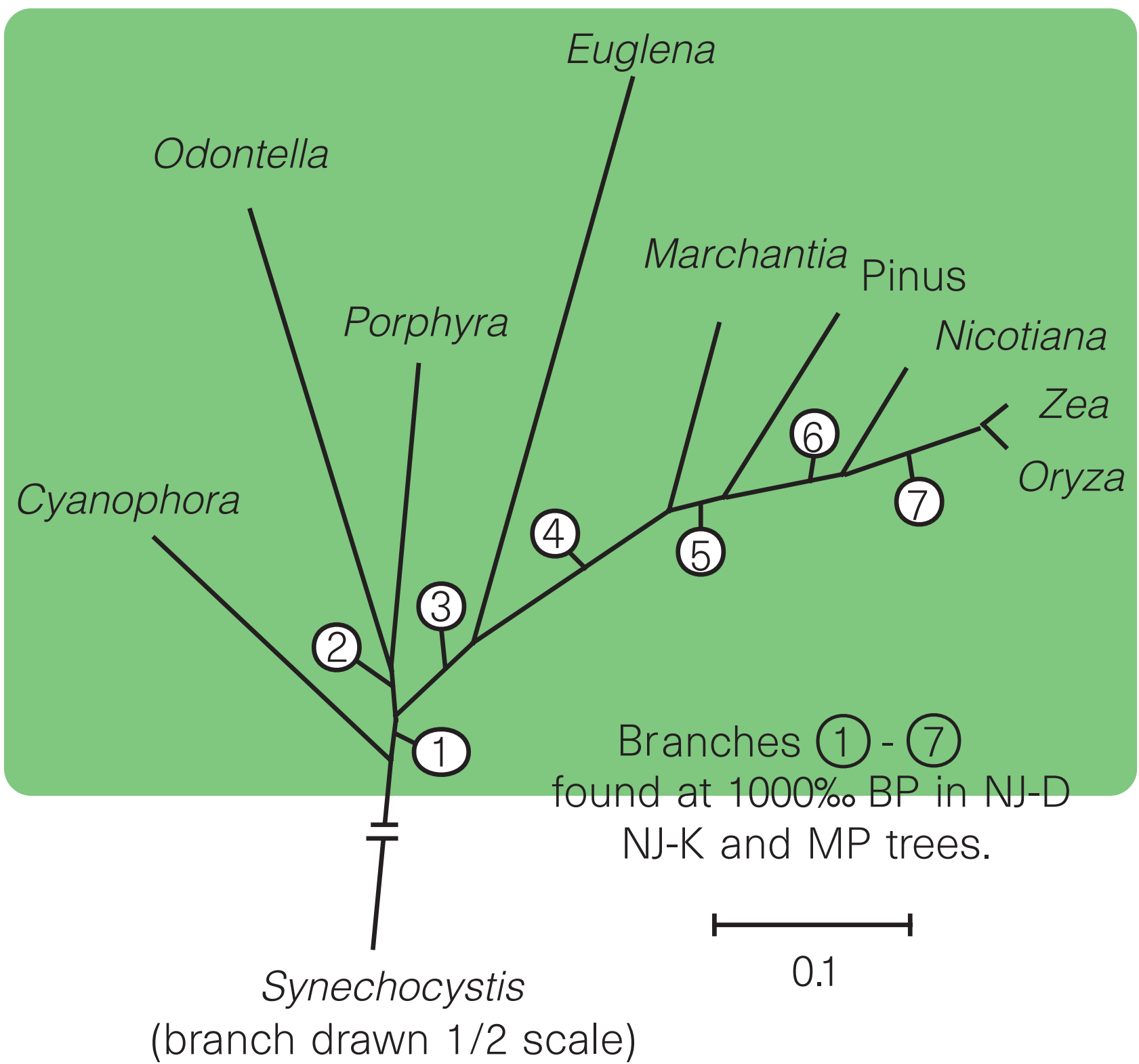
concatenation of alignments into supermatrices

	S^1		S^2		S^3
S_1	A C G T C A A G		T G G - - T		C G G A C T A C G T
S_2	A C - T C C A G		A G C T C C		C C C T - - - - G G
S_3	A C - T C G A C		A G C T C G		C G T T C G A C G T

	S^1	S^2	S^3
S_1	A C G T C A A G	T G G - - T	C G G A C T A C G T
S_2	A C - T C C A G
S_3	A C - T C G A C	A G C T C C
S_4	A G C T C G	C C C T - - - - G G
S_5	C G T T C G A C G T



The few missing sequences make the gene boundaries apparent in this supermatrix containing more than 25,000 AA.

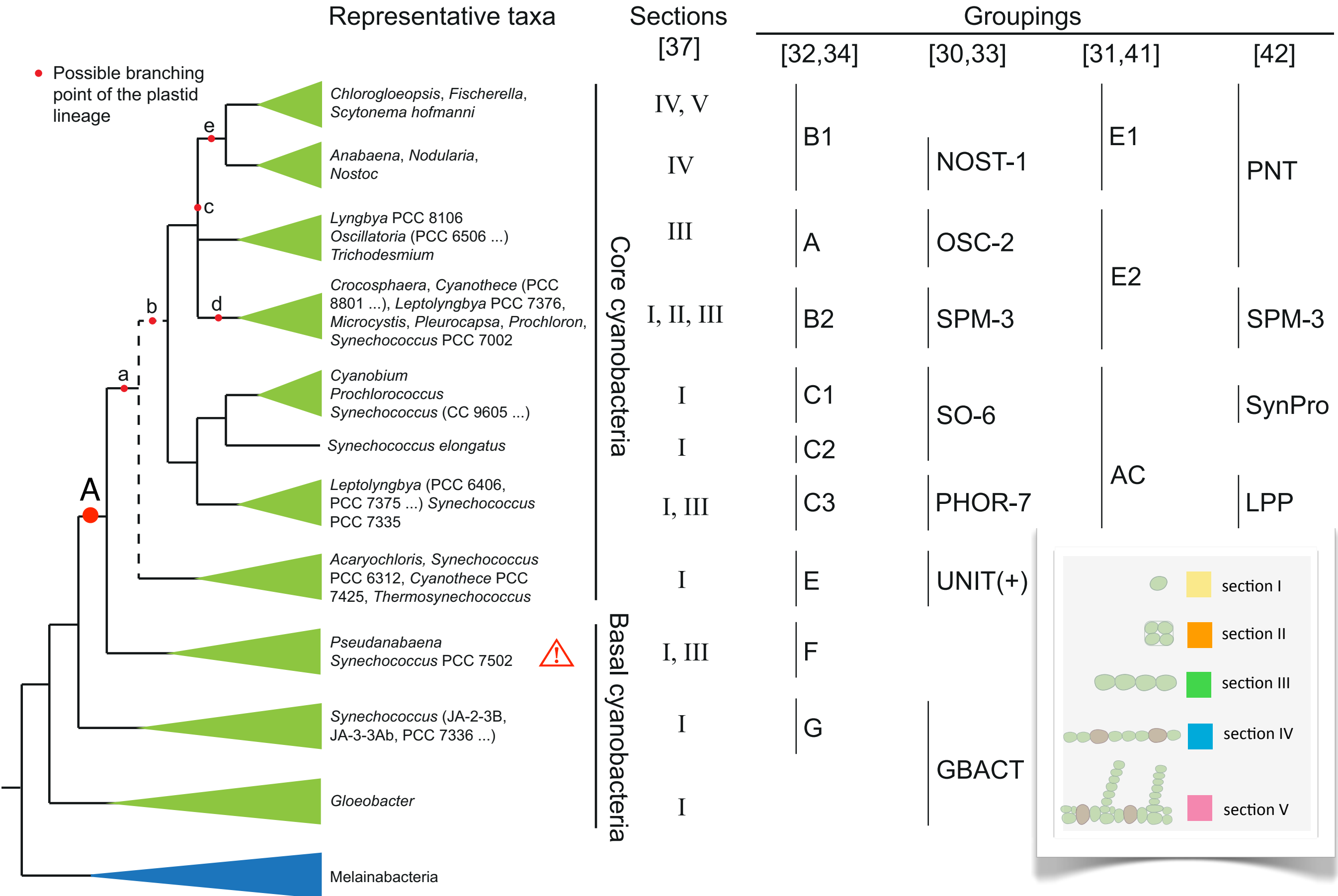


compartment	plastid
# genes	45
# cyanobacteria	1
# plastids	9
supermatrix	10 x 11,039 AA
model	Dayhoff (NJ)
source	[52] Fig. 1a
branching point	n.a.

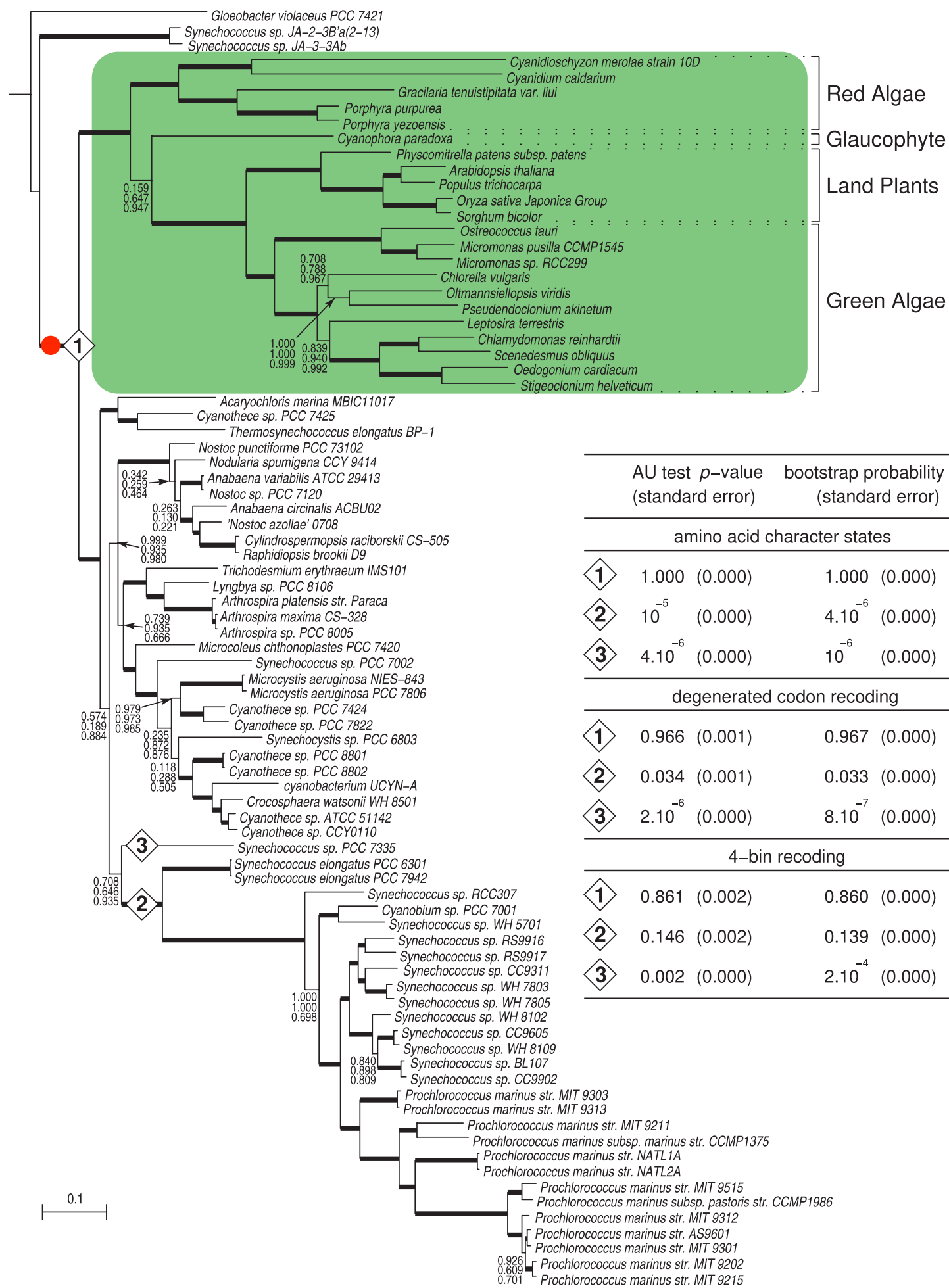
Phylogenomics increases statistical support by reducing stochastic error, which should end incongruence.



*Current Models for
the Origin of Plastids*



Early or late branching of plastids?

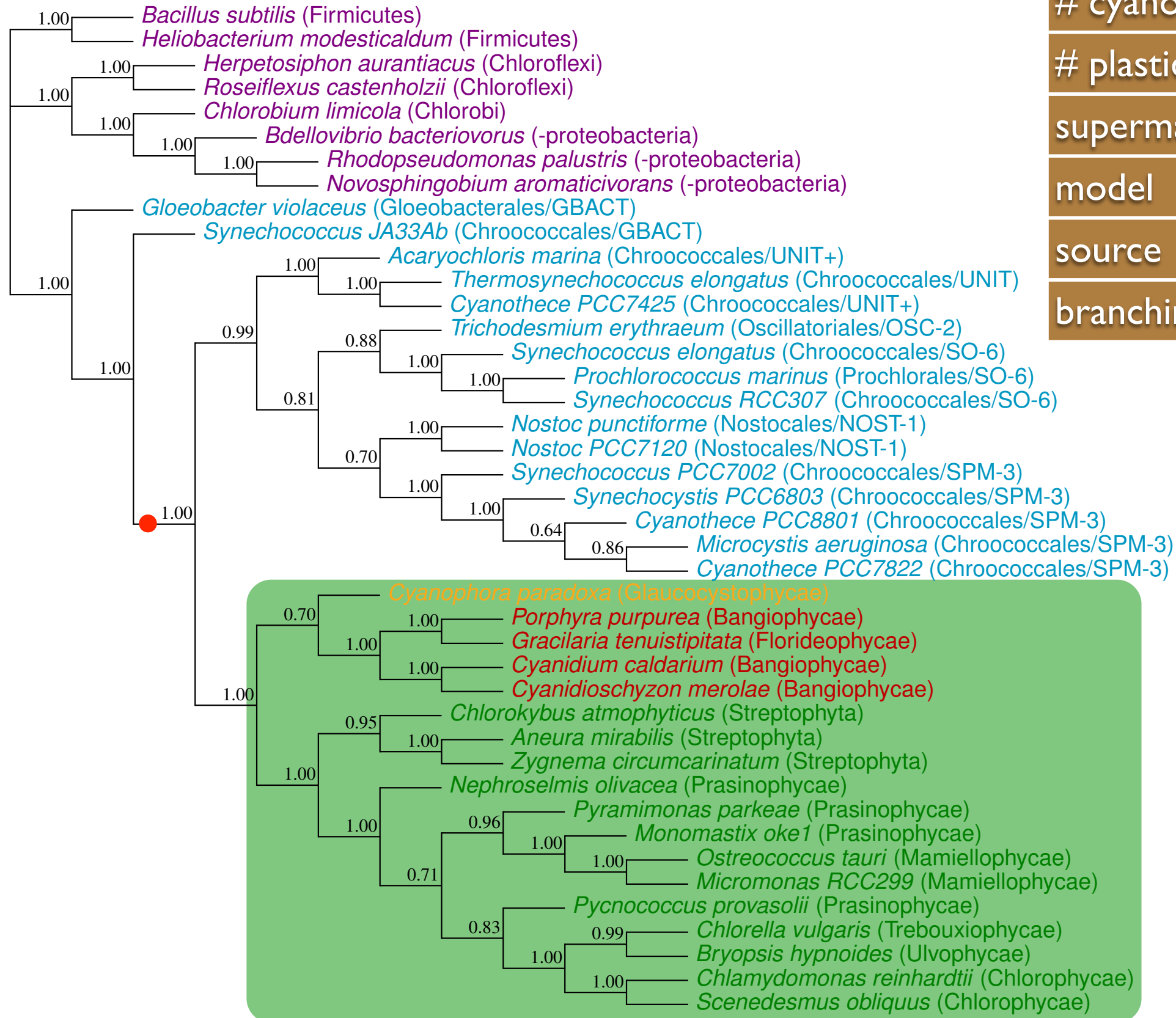


	AU test <i>p</i> -value (standard error)	bootstrap probability (standard error)
amino acid character states		
1	1.000 (0.000)	1.000 (0.000)
2	10 ⁻⁵ (0.000)	4.10 ⁻⁶ (0.000)
3	4.10 ⁻⁶ (0.000)	10 ⁻⁶ (0.000)
degenerated codon recoding		
1	0.966 (0.001)	0.967 (0.000)
2	0.034 (0.001)	0.033 (0.000)
3	2.10 ⁻⁶ (0.000)	8.10 ⁻⁷ (0.000)
4-bin recoding		
1	0.861 (0.002)	0.860 (0.000)
2	0.146 (0.002)	0.139 (0.000)
3	0.002 (0.000)	2.10 ⁻⁴ (0.000)

compartment	plastid & nucleus
# genes	191
# cyanobacteria	61
# plastids	22
supermatrix	83 x 30,149 AA
model	LG / GTR
source	[30] Fig. 2
branching point	a (A)

sophisticated methods
early branching point

sophisticated methods early branching point



compartment

plastid & nucleus

genes

75

cyanobacteria

16

plastids

18

supermatrix

32 x 38,898? AA

model

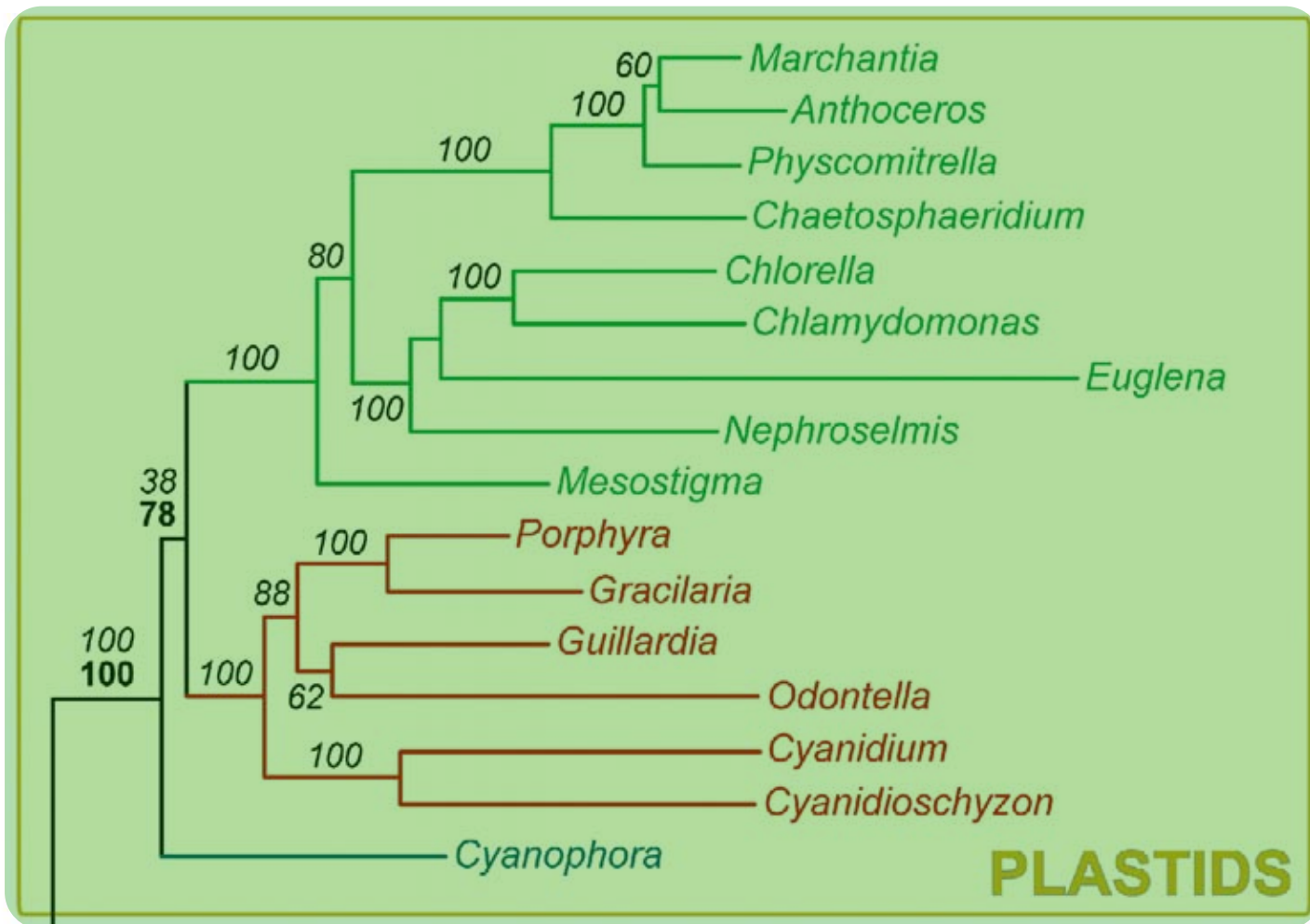
CPREV

source

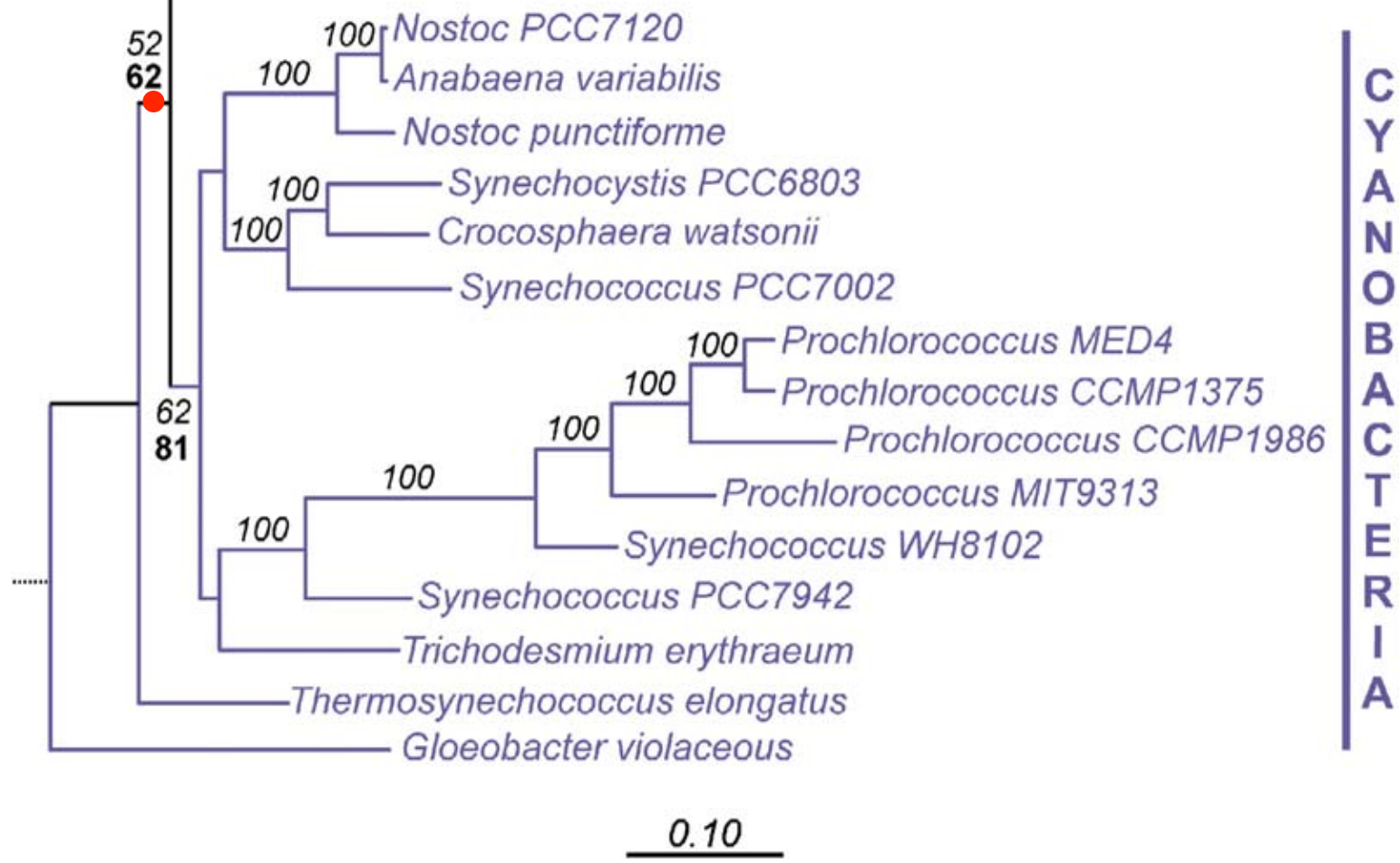
[33] Fig. 2

branching point

a (A)

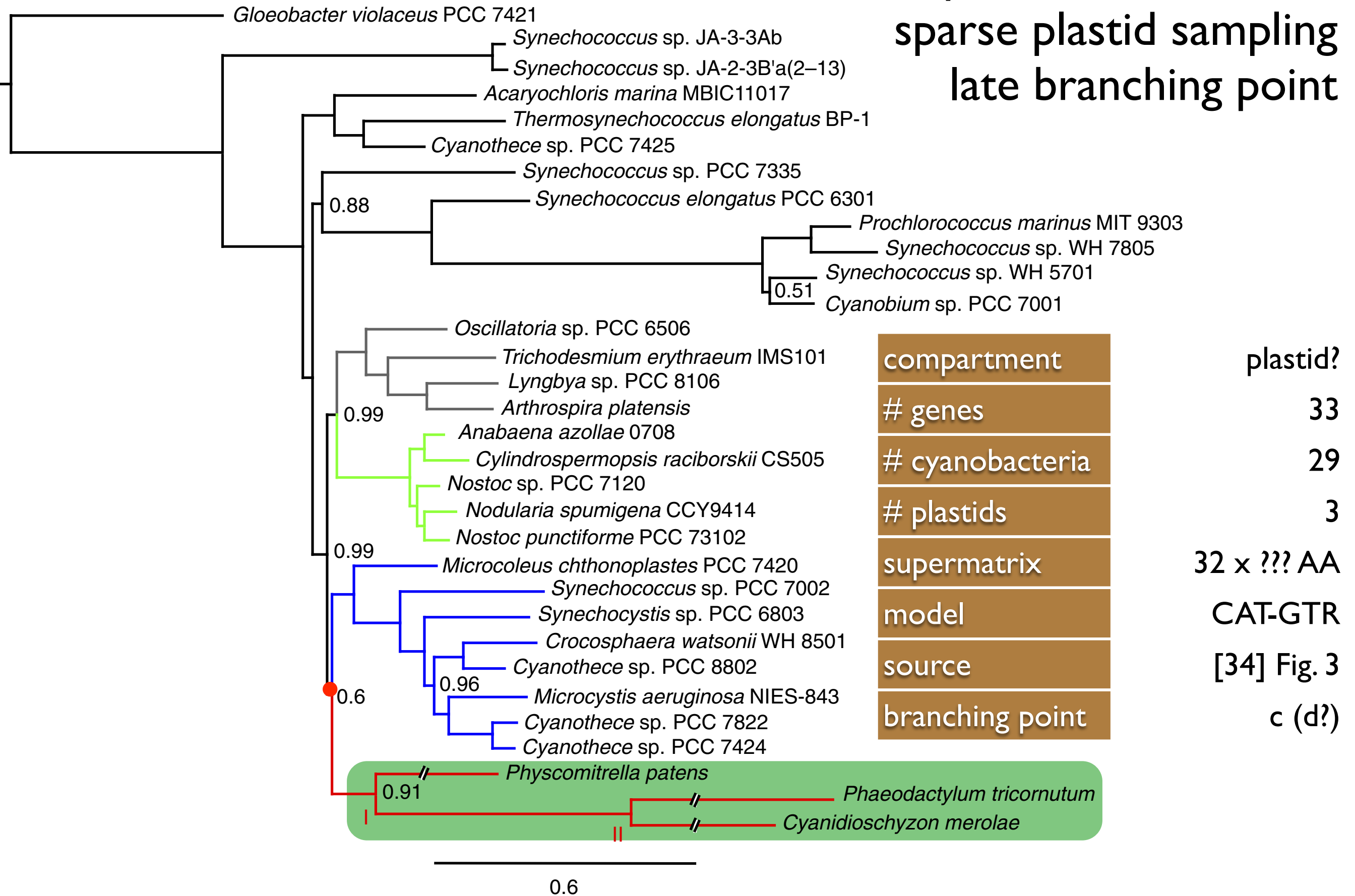


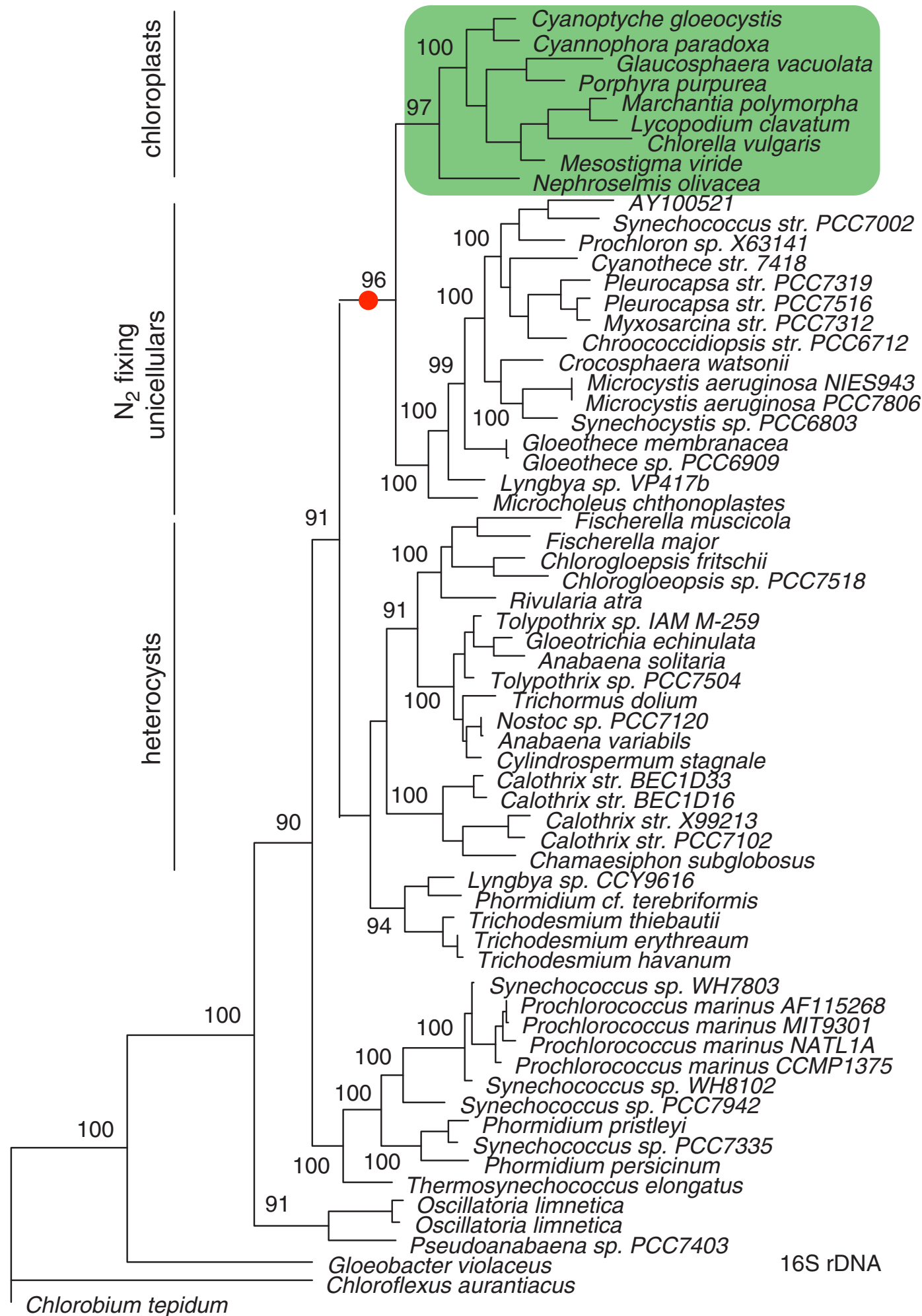
compartment	plastid
# genes	50
# cyanobacteria	15
# plastids	16
supermatrix	31 x 10,334 AA
model	WAG
source	[29] Fig. 2
branching point	b



good methods
later branching point

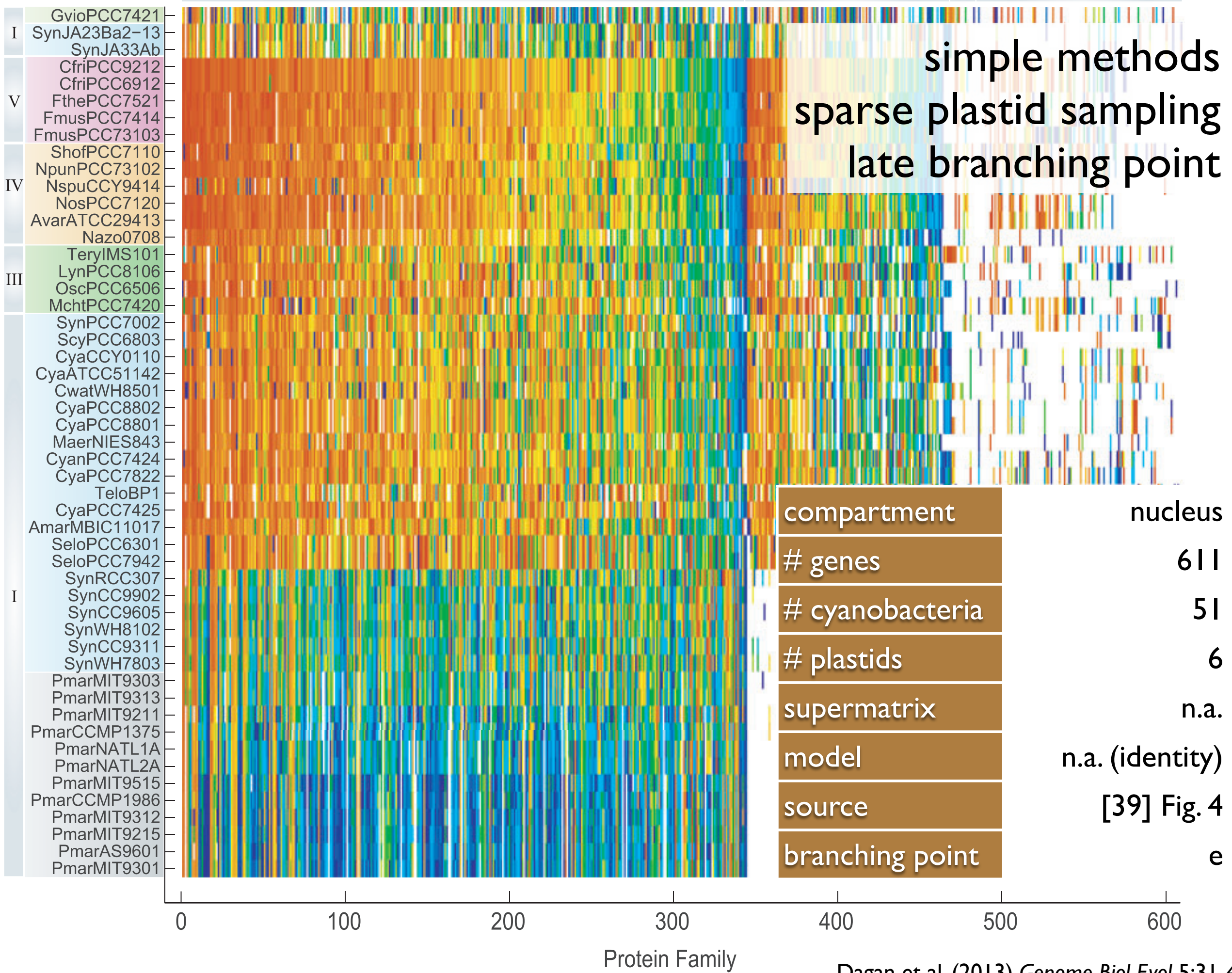
sophisticated methods sparse plastid sampling late branching point





simple methods
not phylogenomics
late branching point

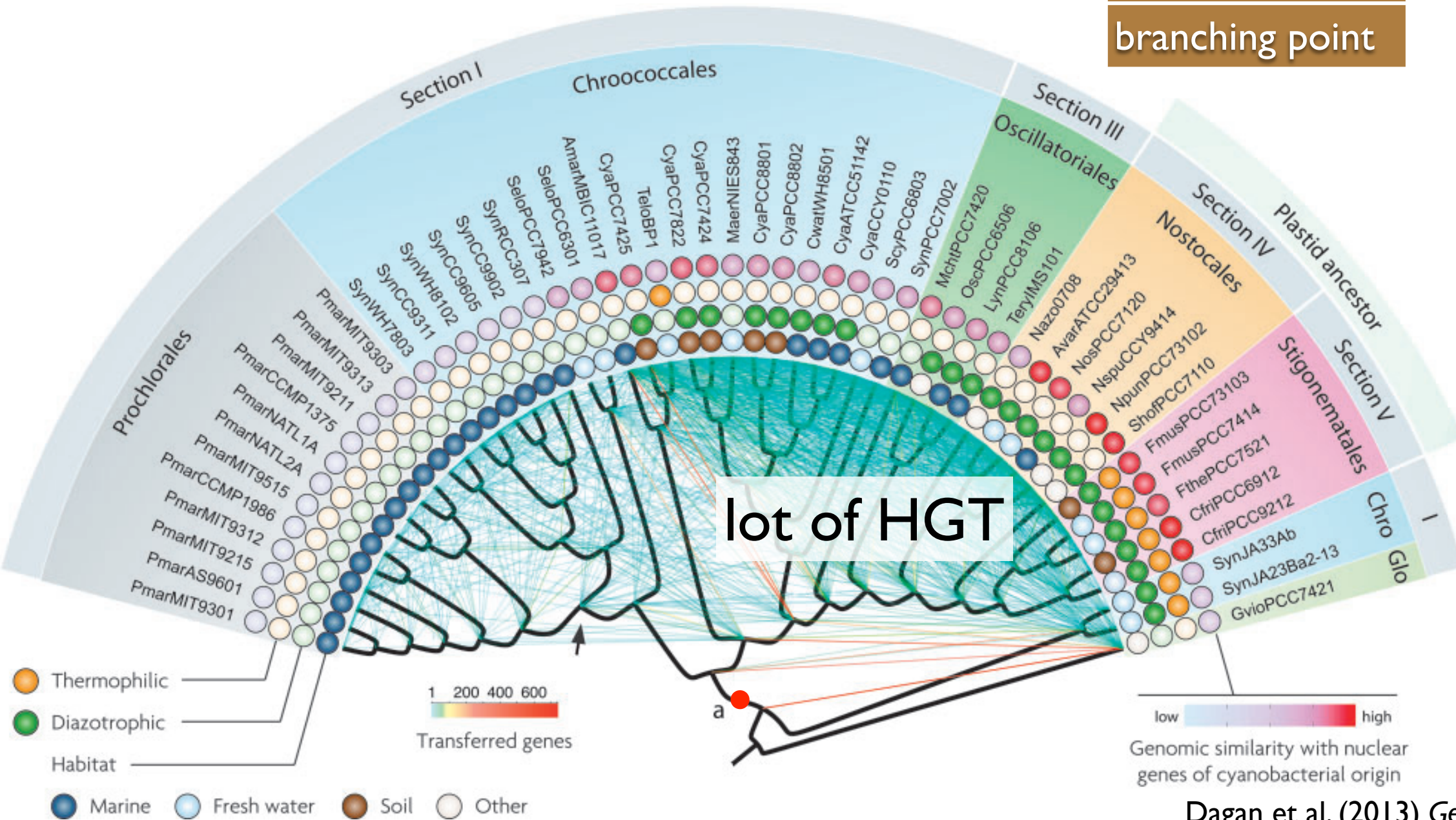
compartment	plastid
# genes	1
# cyanobacteria	56
# plastids	9
supermatrix	n.a.
model	best-fit (?)
source	[38] Fig. 1c
branching point	d



simple methods
 sparse plastid sampling
 early branching point



compartment	nucleus
# genes	23
# cyanobacteria	51
# plastids	6
supermatrix	57 x ? AA
model	JTT? (NJ?)
source	[39] Figs 2 / S6
branching point	a (A)



*Most of these phylogenomic
analyses are incongruent!*

How is it possible?

Non-phylogenetic Signal

Incongruence in phylogenomics has multiple causes.

◎ Supermatrix assembly

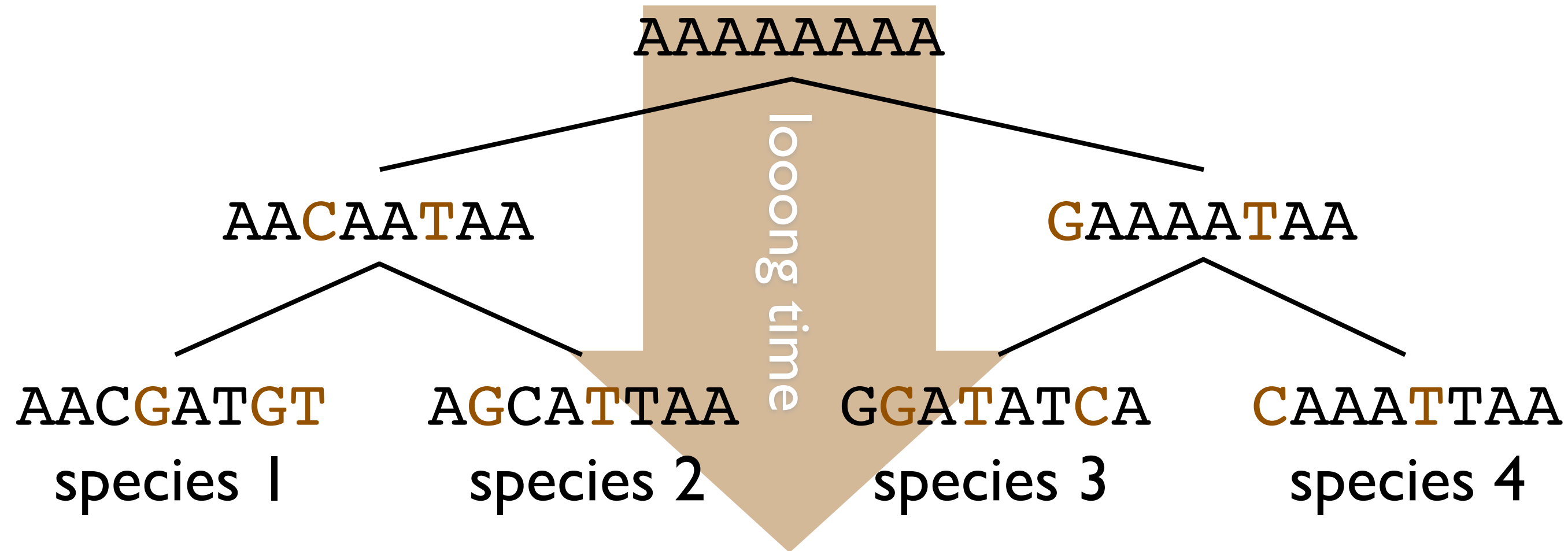
- Conflicting gene histories (transfers)
- Orthology issues (undetected paralogy)
- Contaminations and taxonomic errors
- Sequence errors and dubious alignment

◎ Phylogenetic inference

- Model violations leading to artefacts

Non-phylogenetic Signal

multiple substitutions / saturation / homoplasy

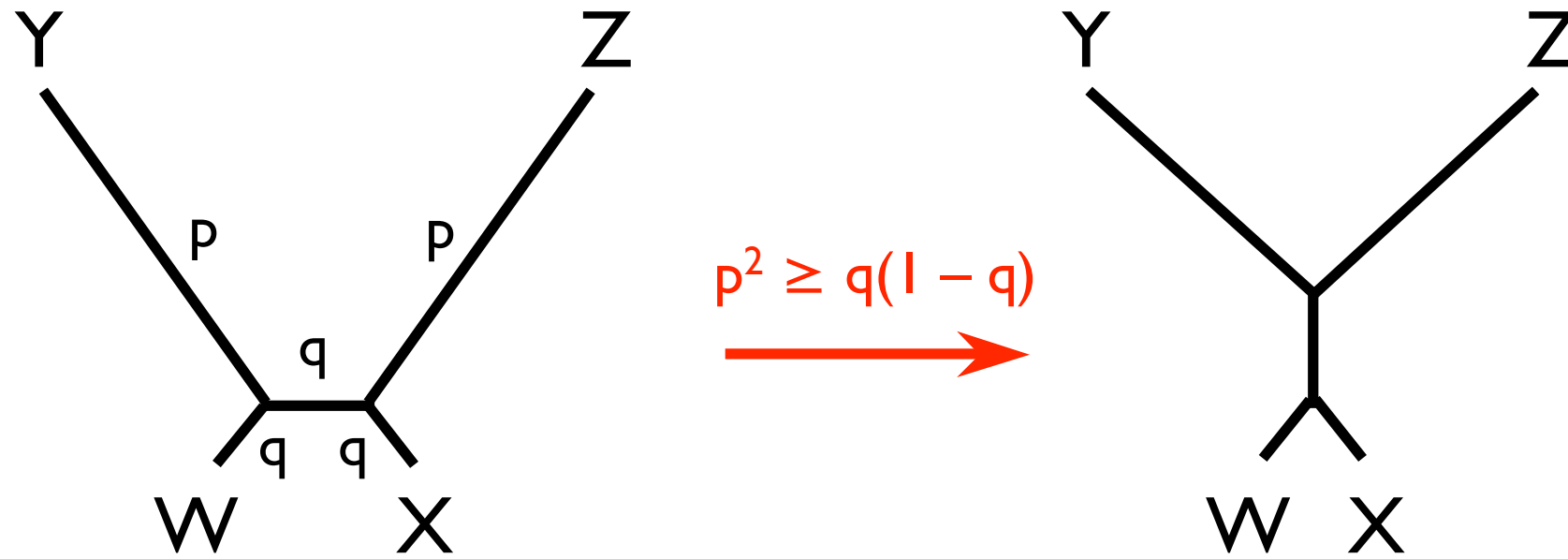


- 1 AACGATGT
- 2 AGCATTAA
- 3 GGATATCA
- 4 CAAATTAA

Due to long geological **time** (and/or to fast evolutionary **rate**), the phylogenetic signal is erased and replaced by a **misleading signal**.

Non-phylogenetic Signal

long-branch attraction artefact



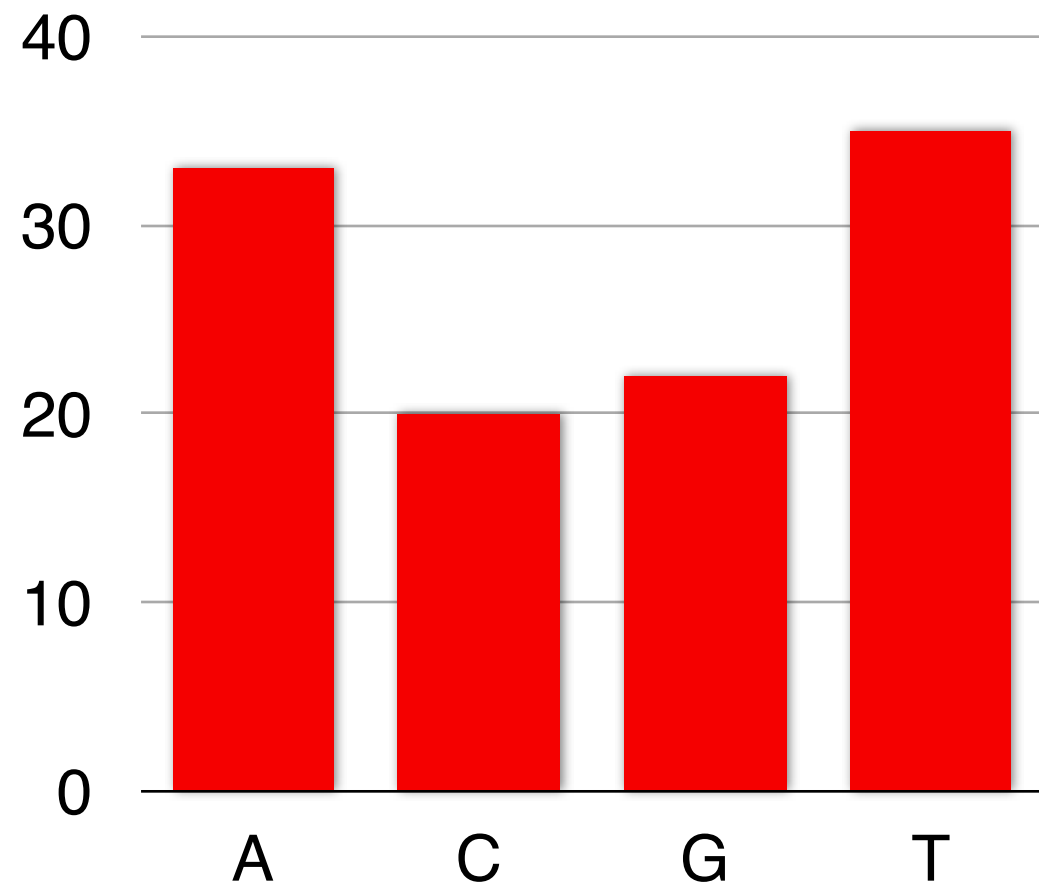
if Y is a phylogenetically distant outgroup



When two branches are **longer** than others, they tend to be **grouped together** irrespective of their true relationships.

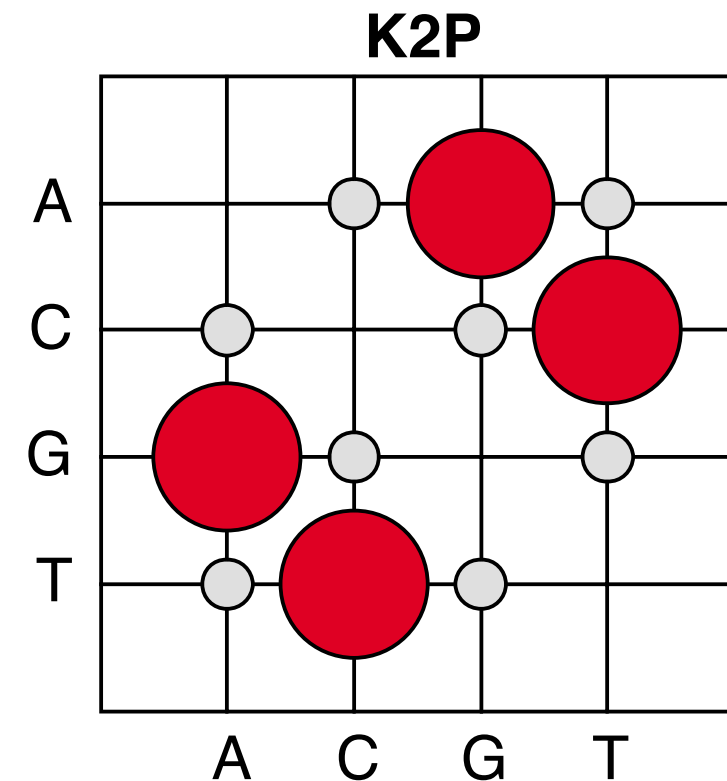
Non-phylogenetic Signal

Models are designed to detect the multiple substitutions.



composition

unequal base frequencies



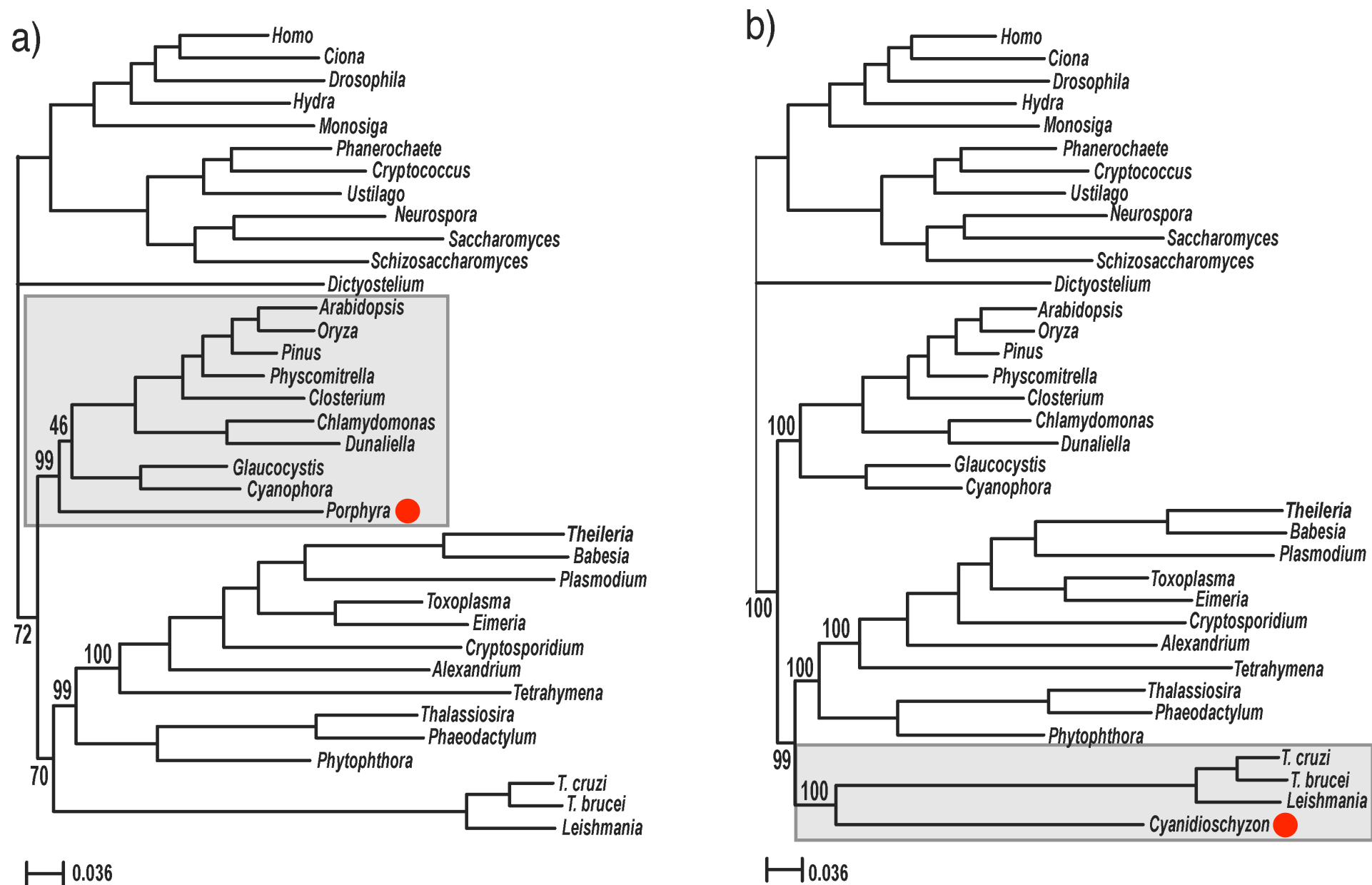
process

unequal substitution rates

Protein models are similar in spirit but larger (20 amino acids).

Non-phylogenetic Signal

Model violations (wrong assumptions) lead to artefacts.



These are due to **heterogeneities** in the evolutionary process.

Non-phylogenetic Signal

How to reduce the phylogenetic artefacts?

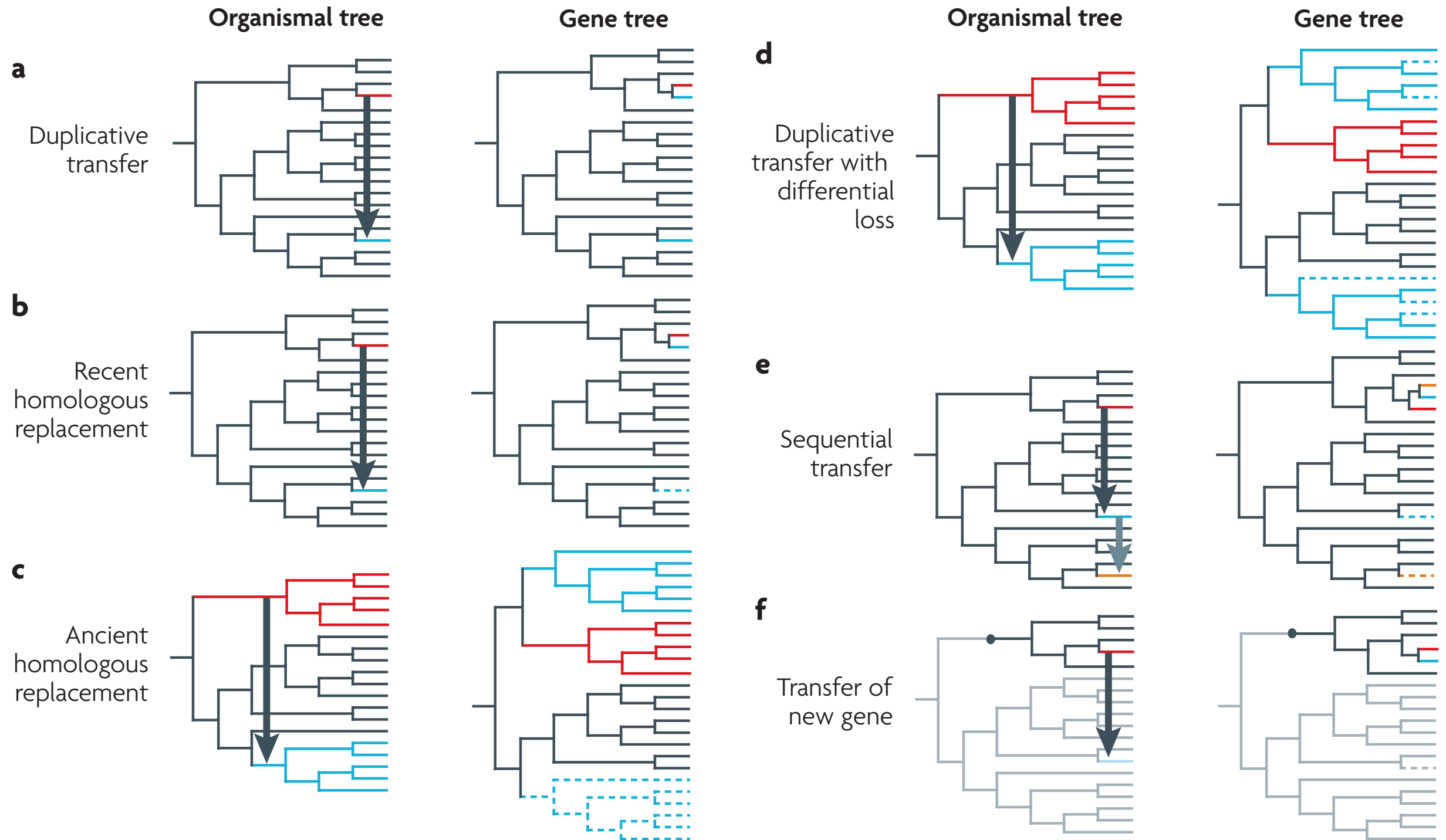
- Combination of 3 approaches
 1. Improvement of taxonomic sampling
 - a. Use of slowly evolving species*
 - b. Use of a dense taxon sampling of each group*
 2. Removal of fast evolving sites from supermatrices
 3. Use of sophisticated evolutionary models



*Supermatrix Assembly
Pitfalls and Solutions*

Gene Transfer

Gene evolution IS NOT organismal evolution.

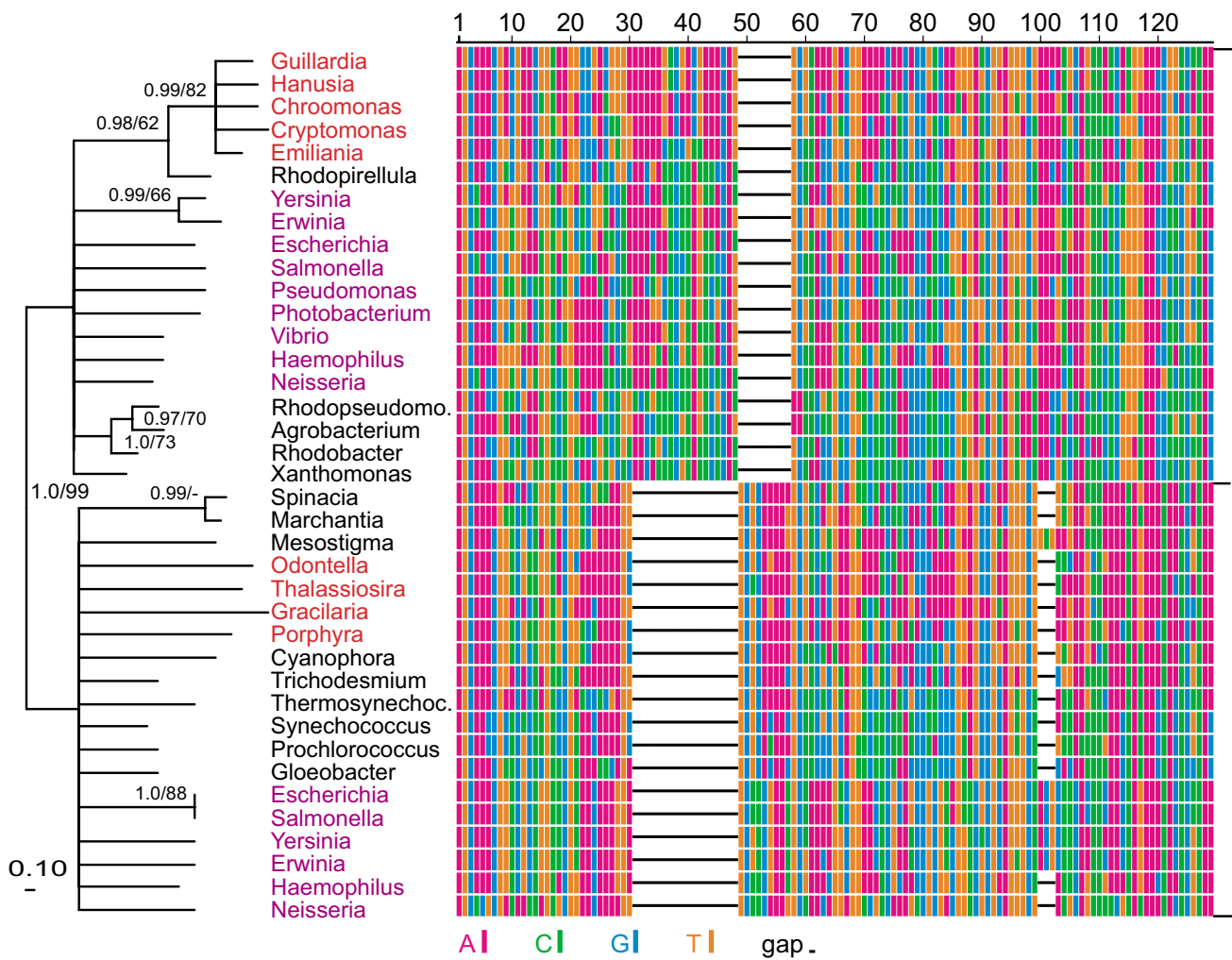


Gene Transfer

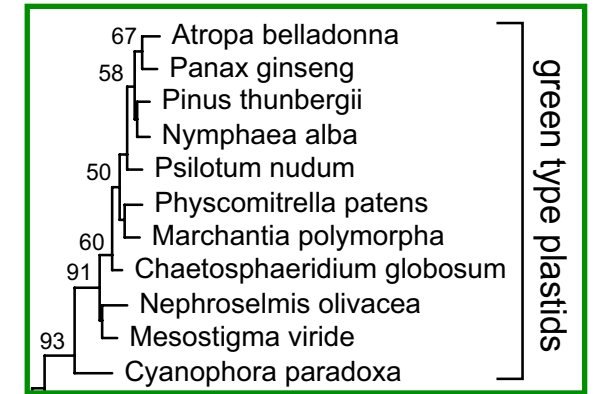
known cases of transfers in plastids

crypto- and haptophytes
(proteobacteria)

(cyanobacteria)

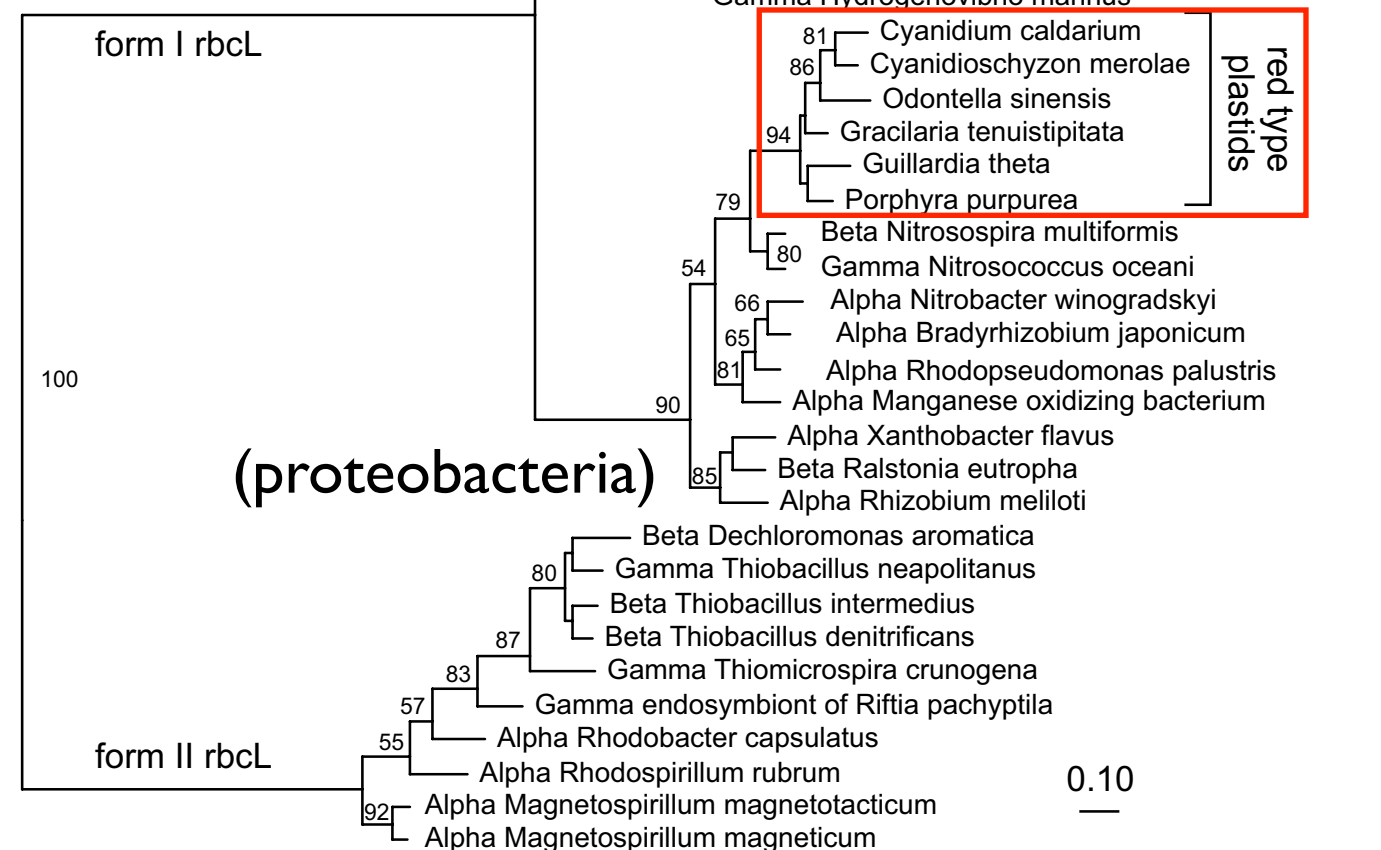


rps36

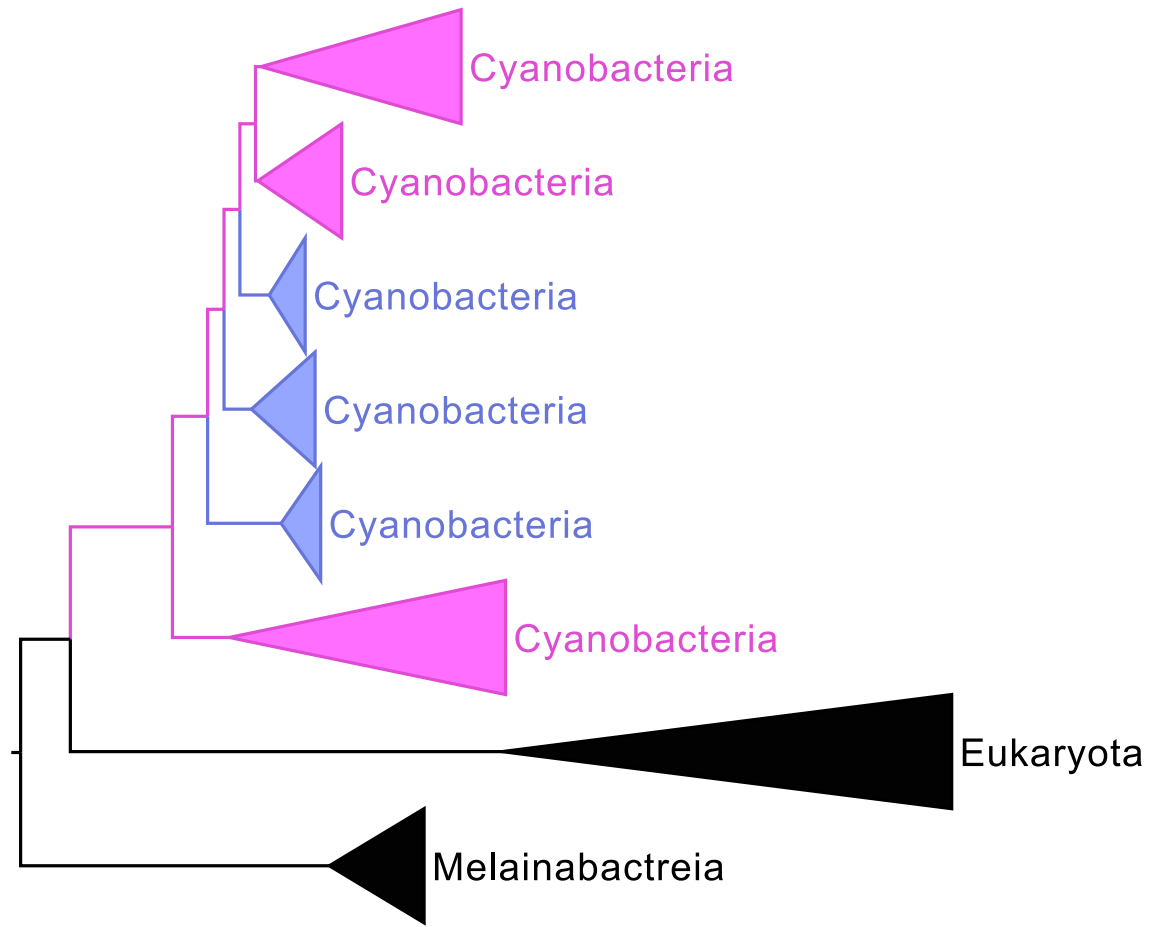


green type plastids

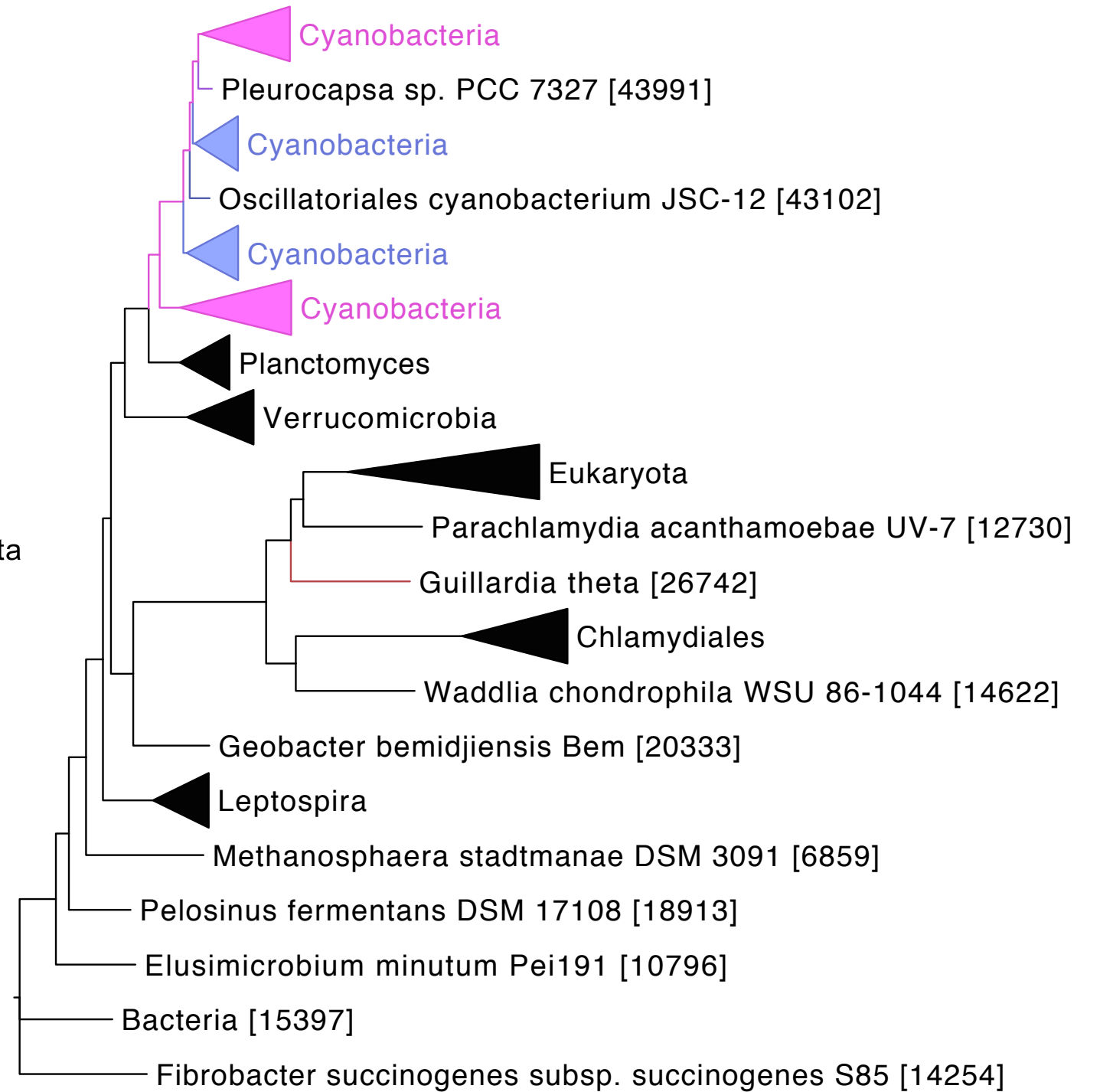
rbcl



other algae and plants
(cyanobacteria)

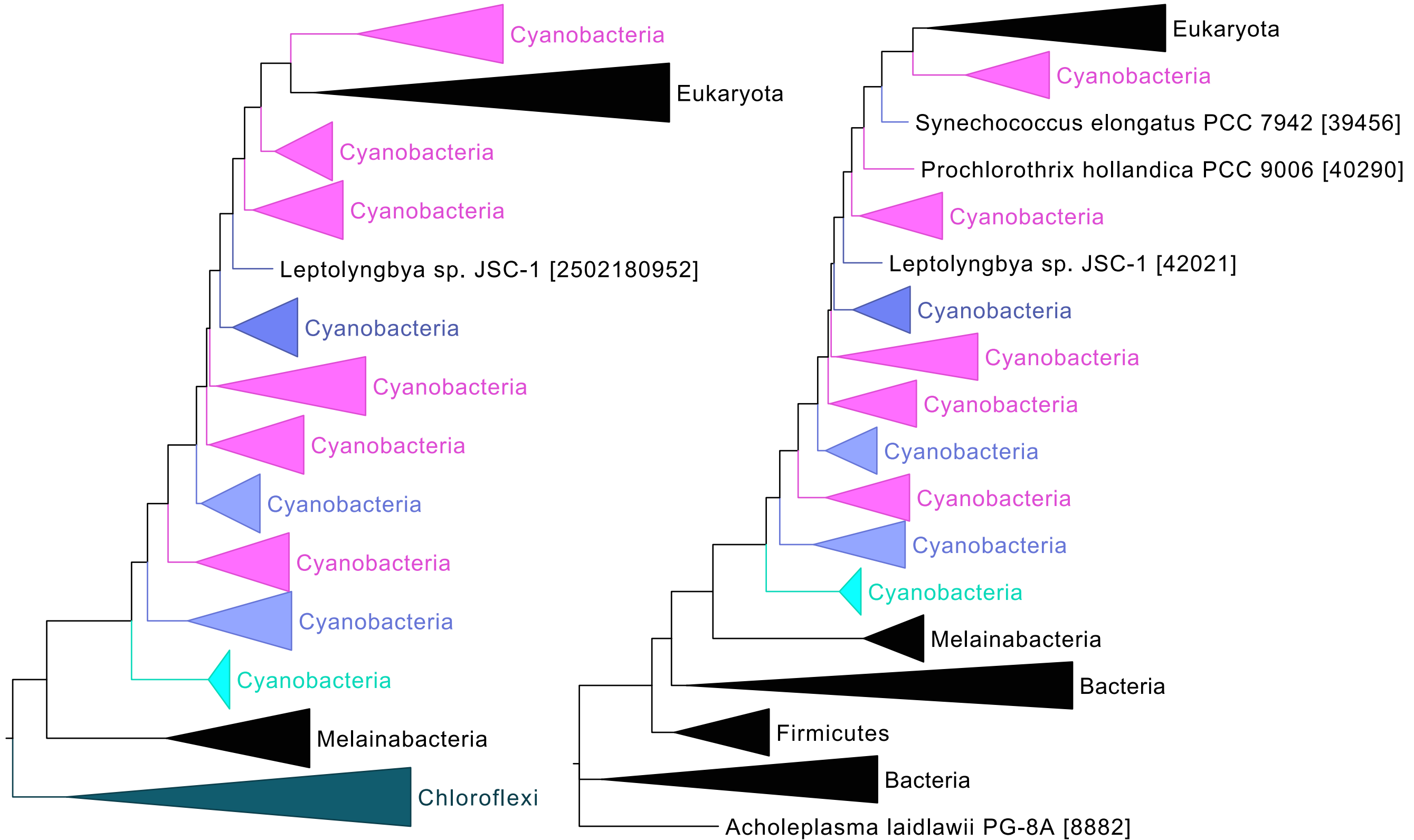


CYANO10923



Adding bacterial sequences for this nuclear gene shows that it is actually of Chlamydiales origin...

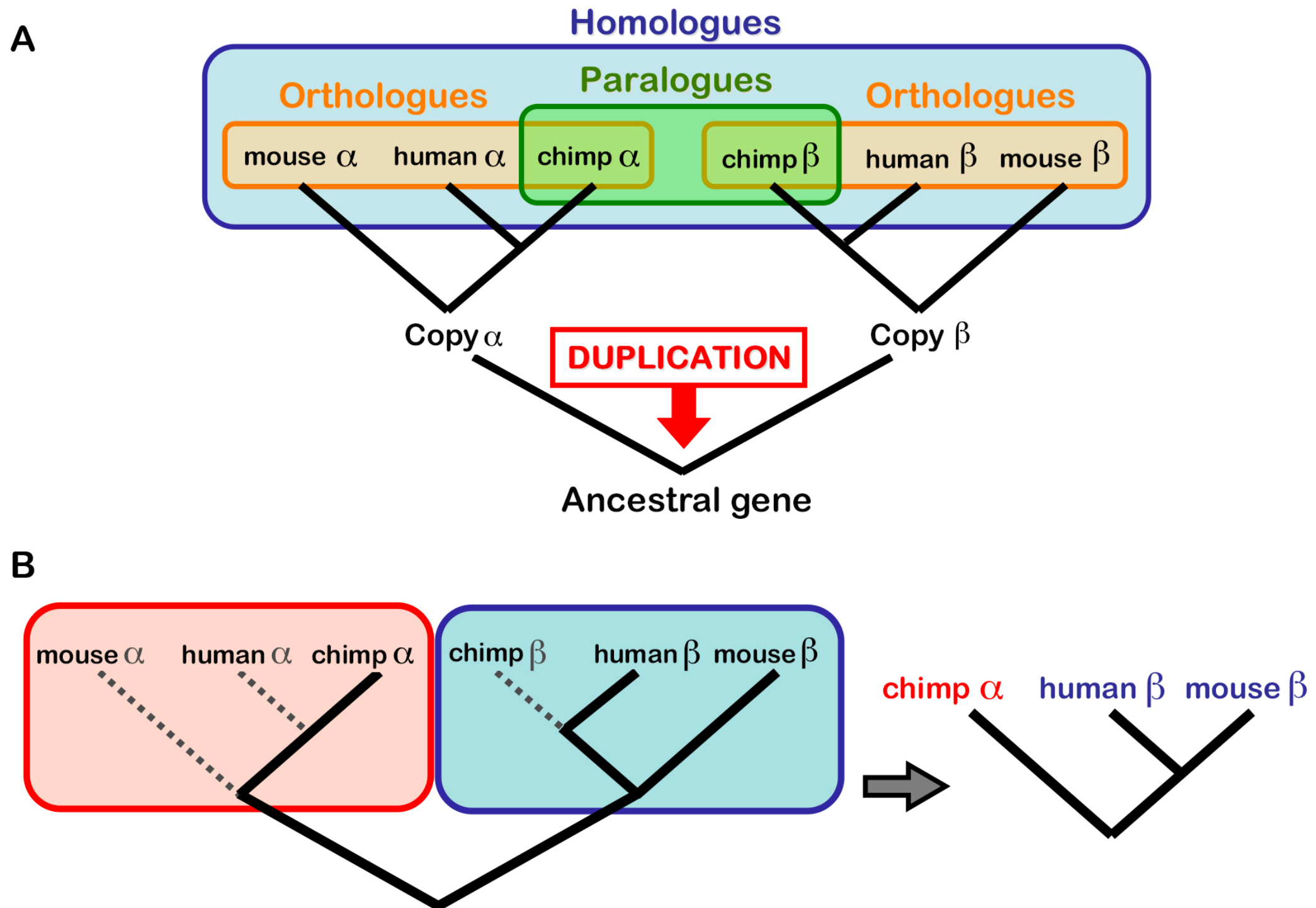
CYANO10318



... whereas this one really comes from Cyanobacteria!

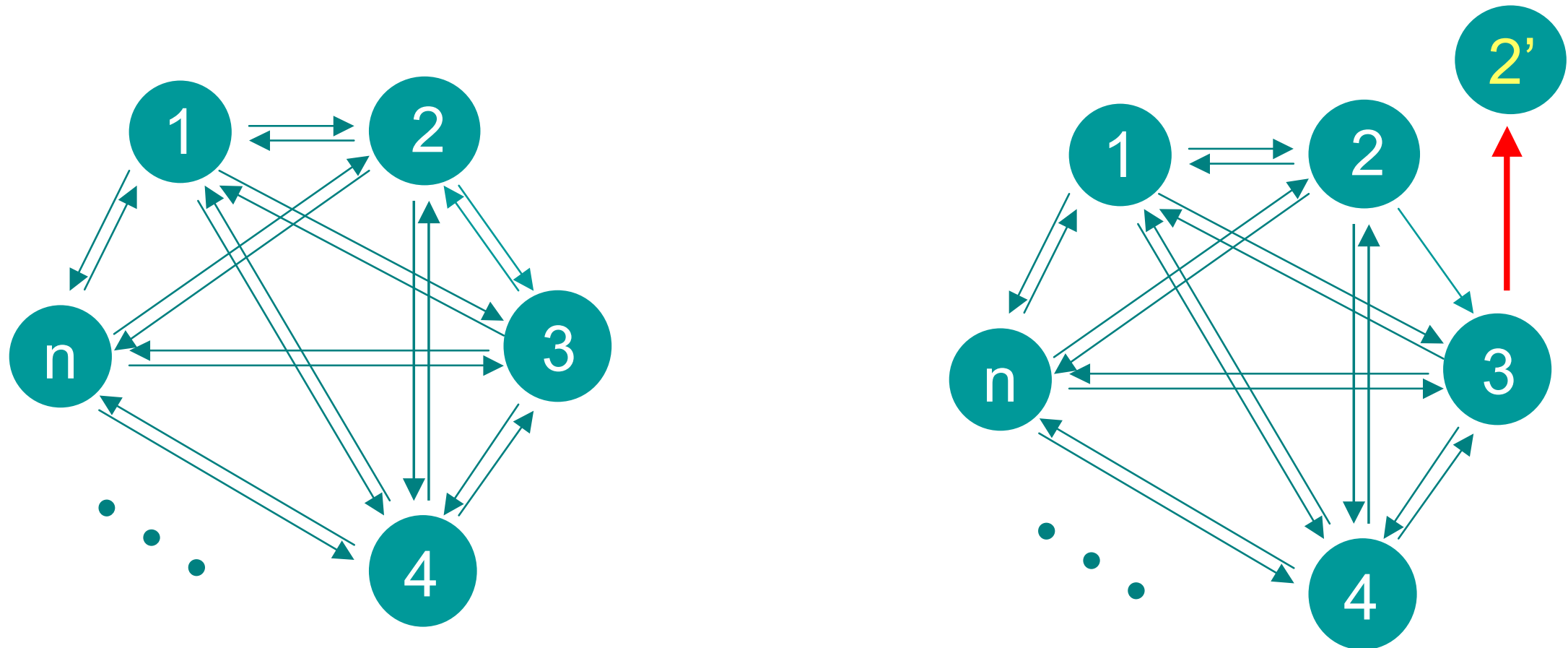
Orthology Issues

the perils of undetected paralogy



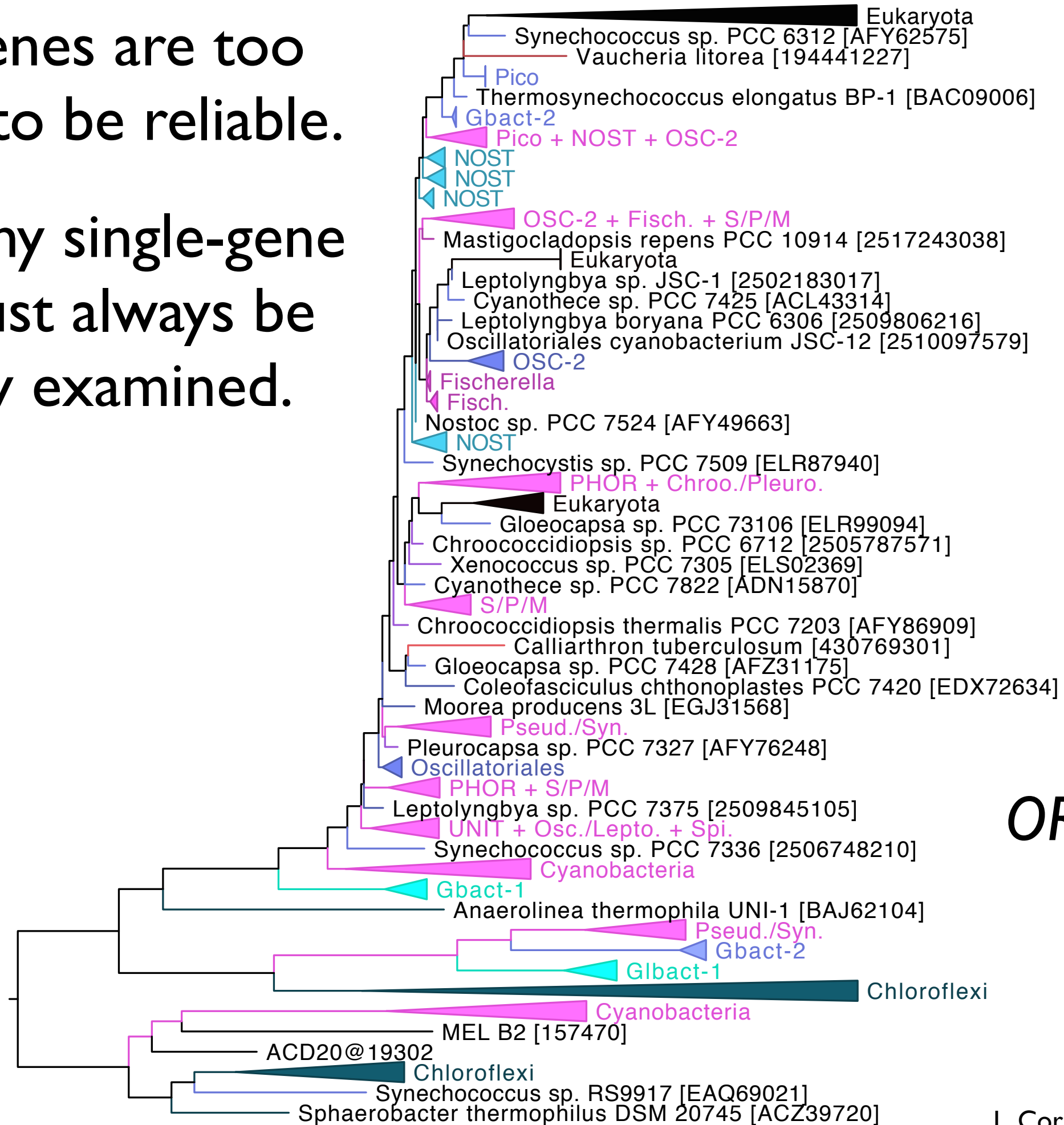
Orthology Issues

the BRH (Best Reciprocal Hits) criterion



A pair of ORFs is said to be a BRH if the two sequences are each other's best match between two genomes (using alignment-based genetic distances).

Some genes are too complex to be reliable. That is why single-gene trees must always be carefully examined.



ORF199

Contaminations

Most NGS transcriptomes datasets are affected.

Current Biology 22, 1456–1457, August 7, 2012 ©2012 Elsevier Ltd All rights reserved

Erratum

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

Cédric Finet,* Ruth E. Timme, Charles F. Delwiche, and Ferdinand Marlétaz

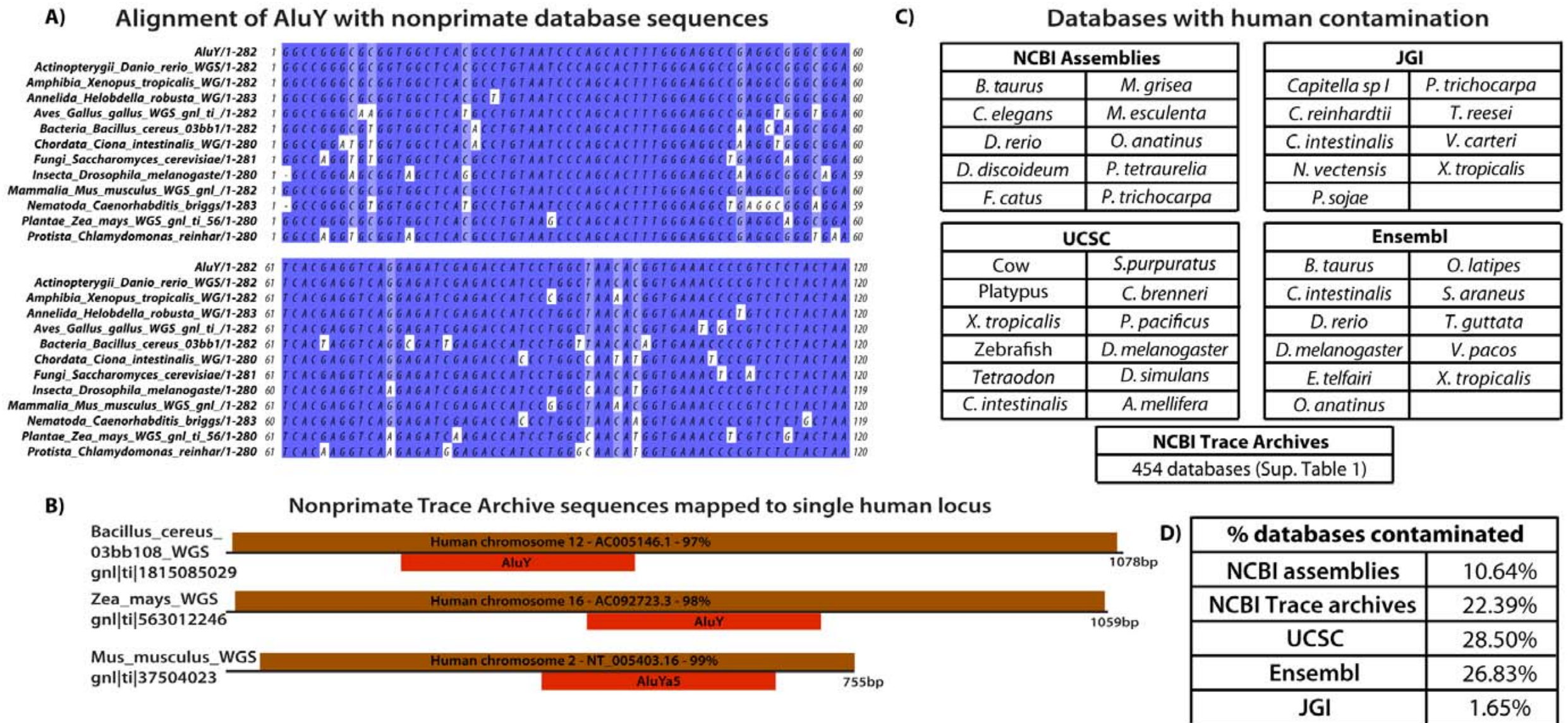
(Current Biology 20, 2217–2222; December 21, 2010)

Following the publication of our multigene phylogeny of the green lineage, some studies have proposed a distinct branching pattern among algal relatives of land plants [1, 2]. Most notably, these differences relate to the position of the taxon *Coleochaete* that we identified as a land plant sister group and the failure to recover the monophyly of Coleochaetales in our study.

Recent exchanges of views with Hervé Philippe and colleagues have prompted us to investigate the possible cause of these inconsistencies. In particular, we searched for contaminations that had been identified by Philippe and colleagues ([3], this issue of *Current Biology*). We identified cross-contaminations in two libraries sequenced for our study, corresponding to the species *Chaetosphaeridium globosum* and to a lesser extent *Nitella hyalina*. The contamination likely took place during library preparation or sequencing, because the samples were obtained from pure in-house algal cultures, and there is a correlation between time of sample processing at the sequence facility and levels of cross-contamination. It is also noteworthy that the most highly contaminated data set, with as much as 5% exogenous sequence, was from *C. globosum*, which was the one data set that was amplified prior to the sequencing reactions. *C. globosum* and *N. hyalina* had what appear to be erroneous sequences for 28 out of 77 genes, representing 47 out of 5,929 sequences in the data matrix. Unfortunately, the phylogenetic proximity of these taxa and the close similarity of these sequences caused the contaminant sequences to escape our validation protocol.

Contaminations

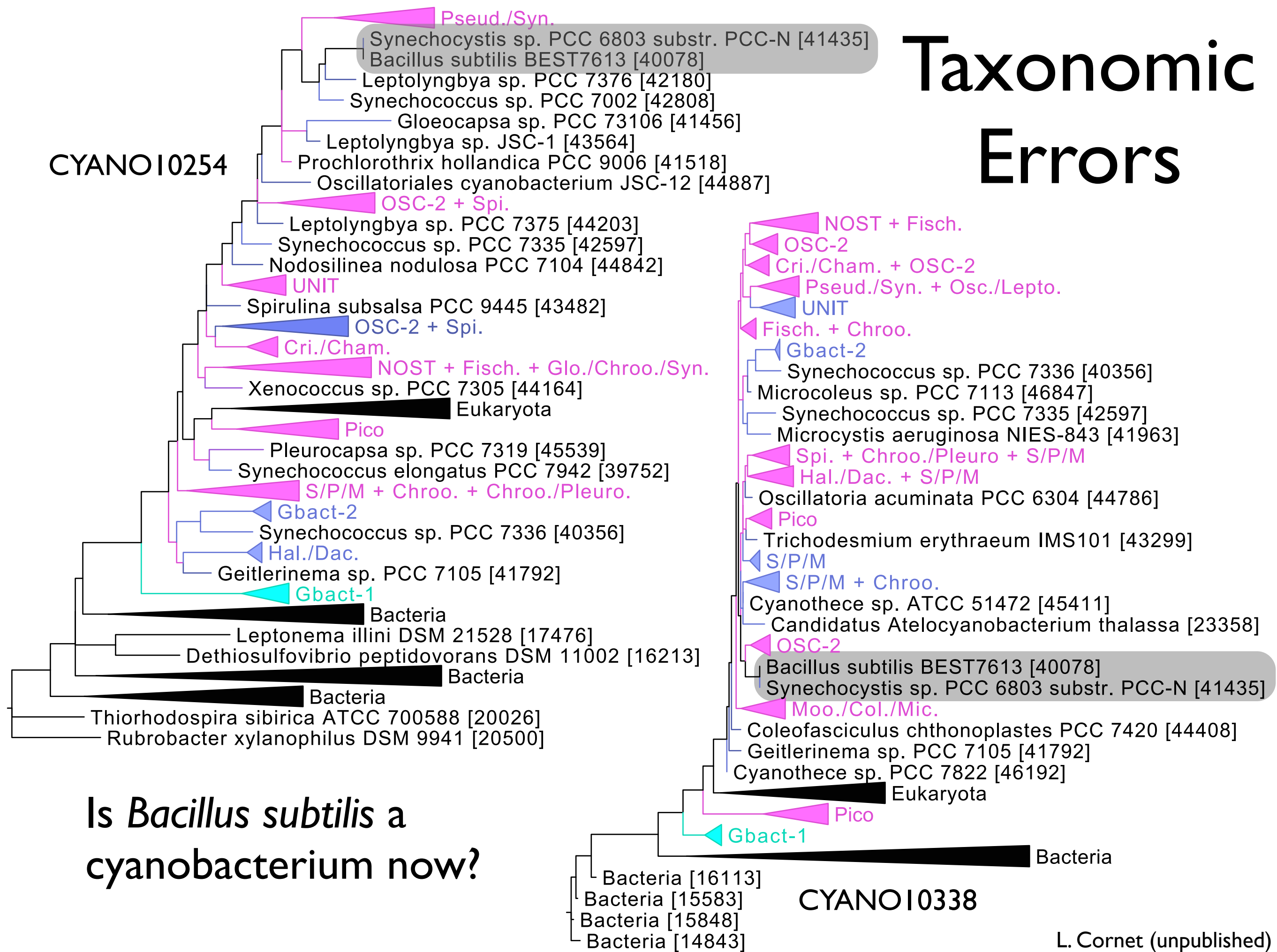
Genomes are also plagued by this problem.



492 out of 2749 non-primate public databases (NCBI, Ensembl, UCSC, JGI) were contaminated with human sequences!

Taxonomic Errors

CYANOI0254



Is *Bacillus subtilis* a cyanobacterium now?

CYANOI0338

Envoyer clear

Example:

cellular organisms - Bacteria - Firmicutes - Bacilli - Bacillales - Bacillaceae - Bacillus - Bacillus subtilis group - Bacillus subtilis - Bacillus subtilis BEST7003

Taxon	Taxonomy ID	Rank	Common Name
- Bacillus subtilis BEST7003 ↑	1204342	no rank	

Sabotage!

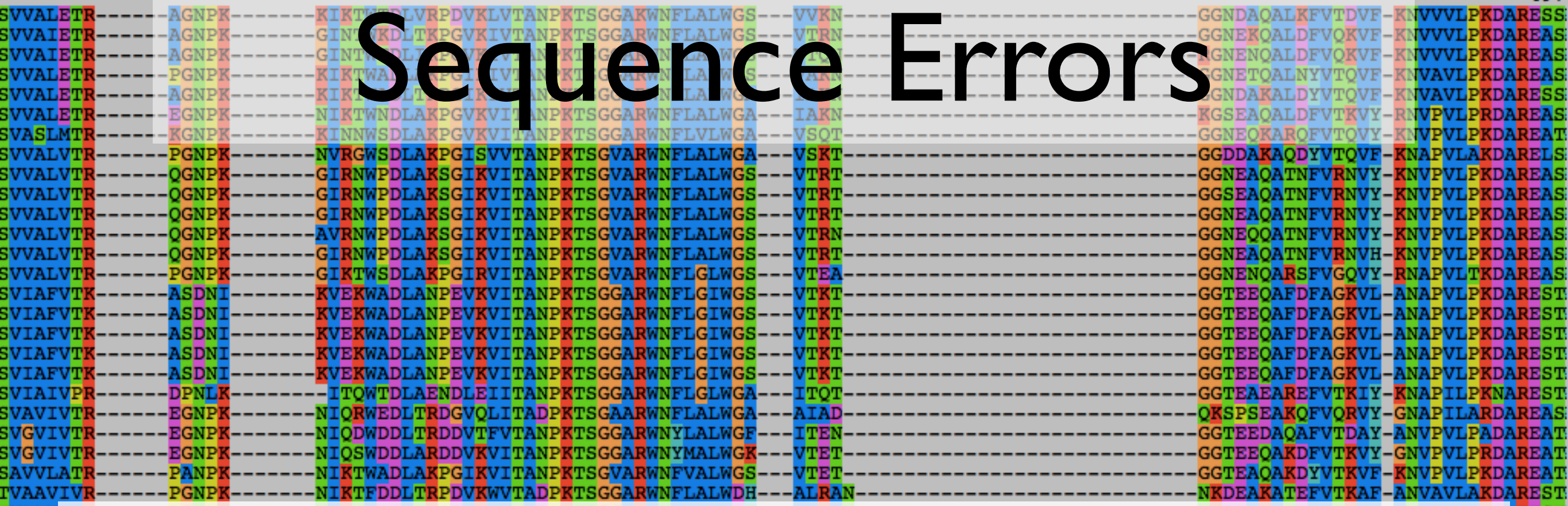
Complete Sequence of the First Chimera Genome Constructed by Cloning the Whole Genome of *Synechocystis* Strain PCC6803 into the *Bacillus subtilis* 168 Genome

Satoru Watanabe,^a Yuh Shiwa,^b Mitsuhiro Itaya,^c and Hirofumi Yoshikawa^{a,c}

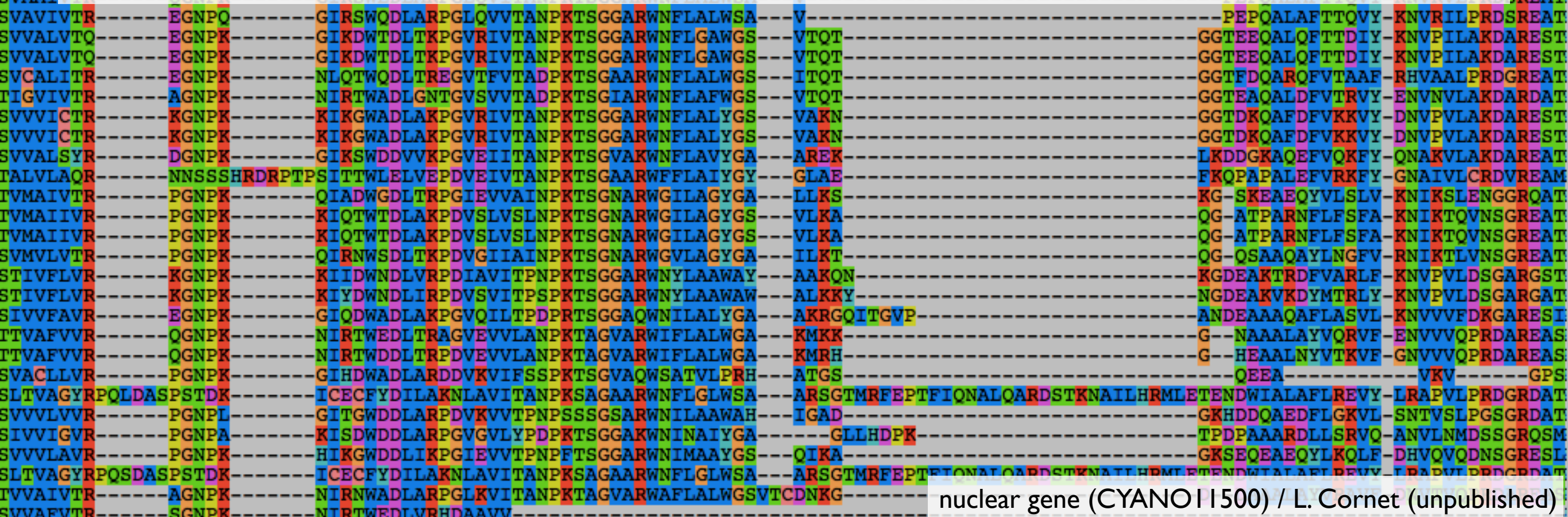
Department of Bioscience^a and Genome Research Center,^b Tokyo University of Agriculture, Tokyo, Japan, and Laboratory of Genome Designing Biology, Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan^c

Genome synthesis of existing or designed genomes is made feasible by the first successful cloning of a cyanobacterium, *Synechocystis* PCC6803, in Gram-positive, endospore-forming *Bacillus subtilis*. Whole-genome sequence analysis of the isolate and parental *B. subtilis* strains provides clues for identifying single nucleotide polymorphisms (SNPs) in the 2 complete bacterial genomes in one cell.

Sequence Errors



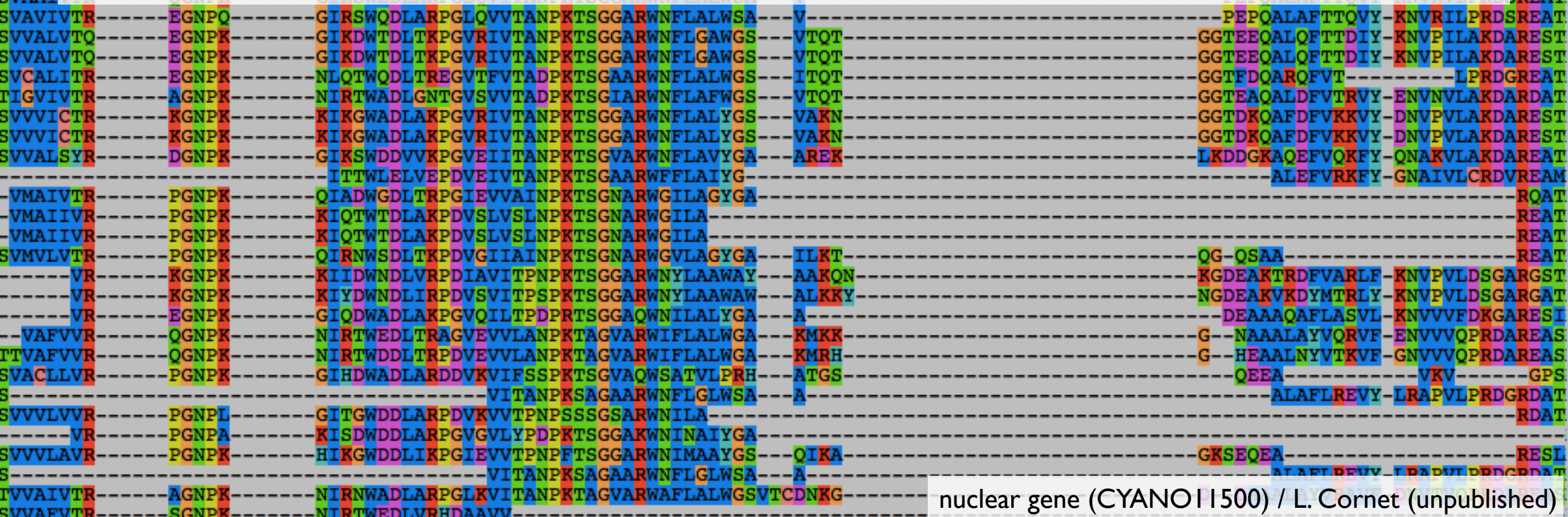
Alignments are replete with dubiously homologous regions. These have to be cleaned before phylogenetic inference.



Sequence Errors



Step 1: Removal of frameshifts and non-homologous stretches with *HMMCleanest* by Poujol, Di Franco et al. (unpublished)



Dubious Alignment

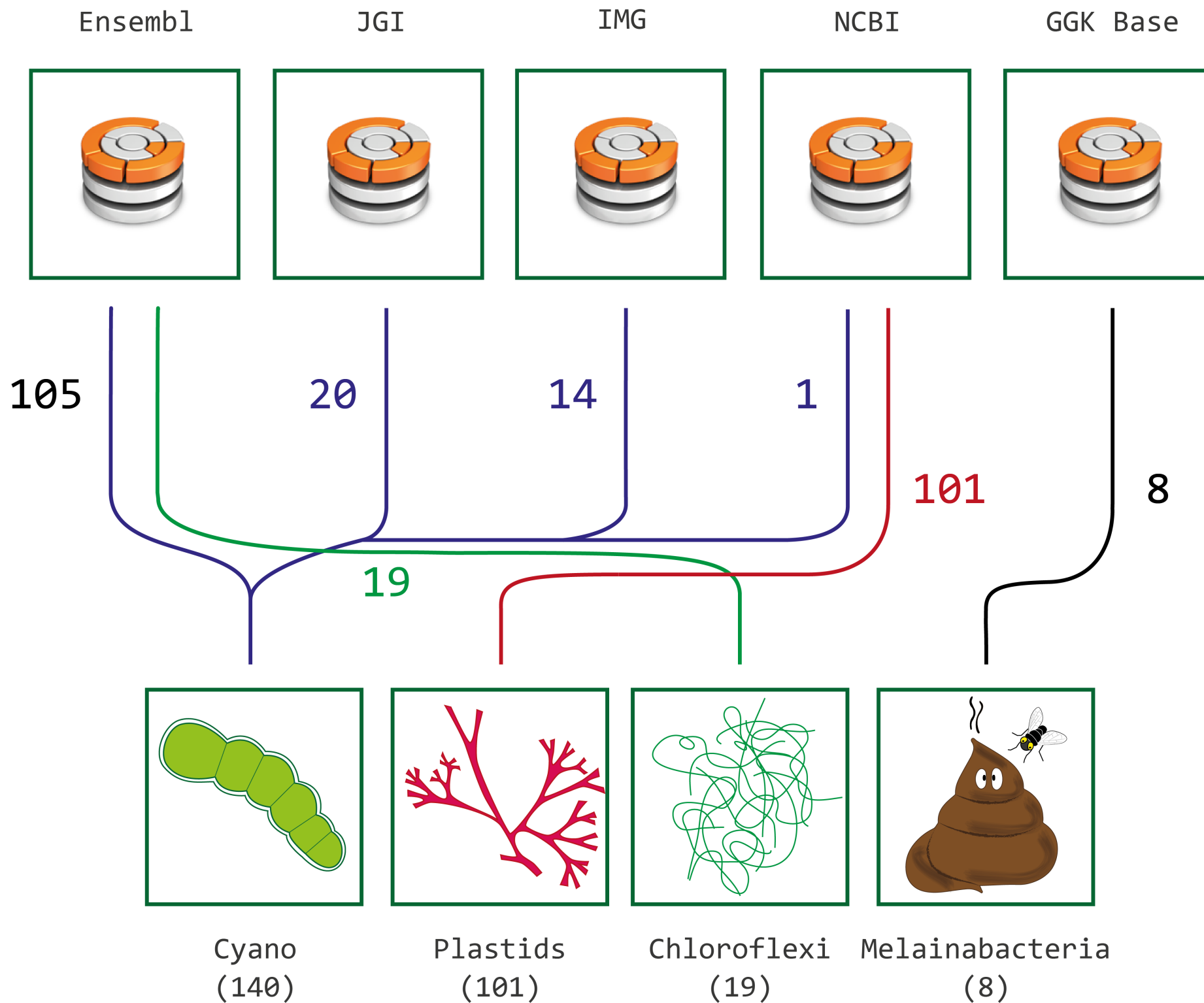
Step 2: Removal of ambiguously aligned columns
with Gblocks by Castresana (2000) *Mol Biol Evol* 17:540-552

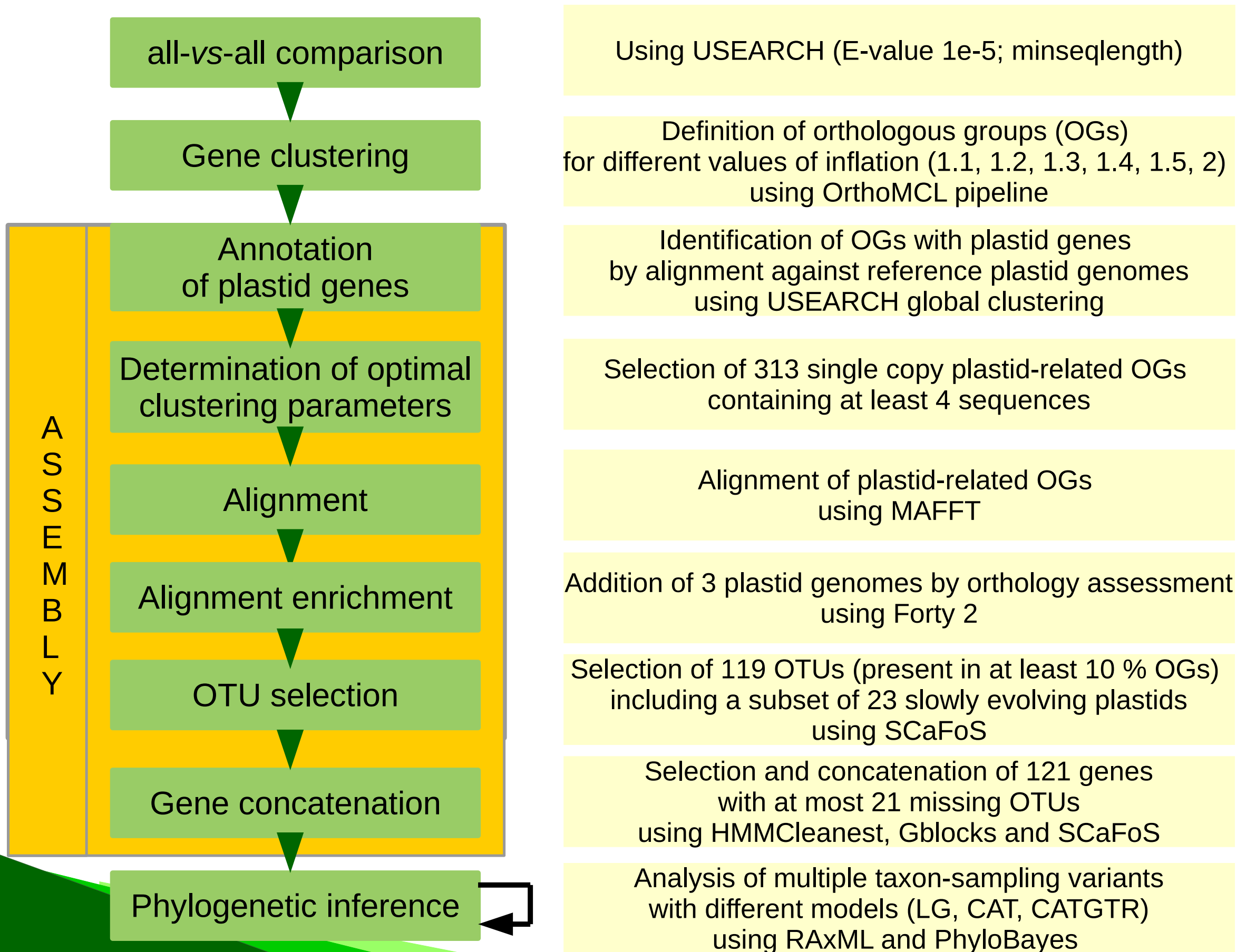
A large, bold, brown number '5' is centered on the page. The text 'Towards a Branching Point for Plastids' is overlaid on the number.

*Towards a Branching
Point for Plastids*

Plastid Supermatrix

starting from complete genomes





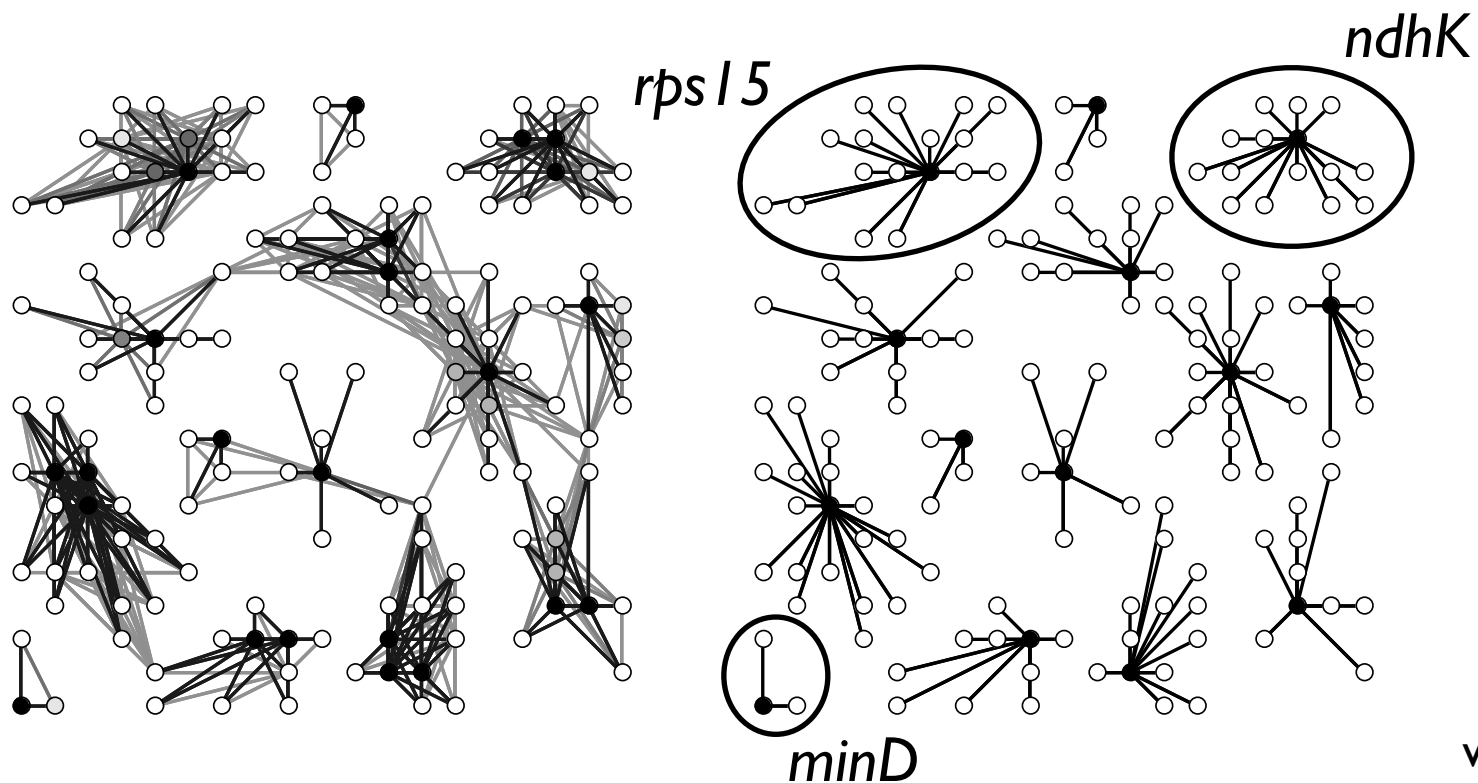
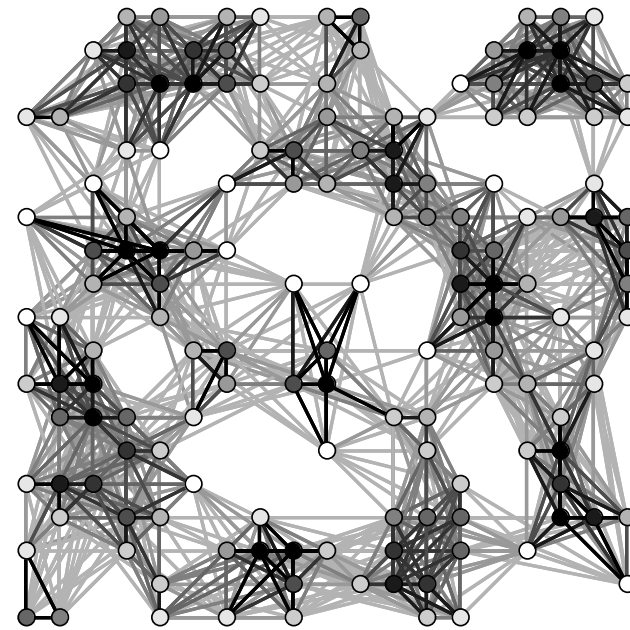
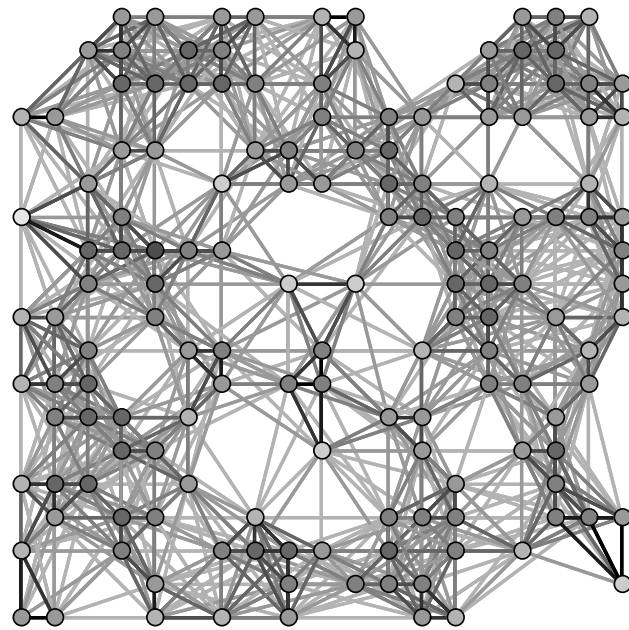
Gene Clustering

all-vs-all similarity matrix

916'		41							45					100
915'													100	
914'					89					67		100		
...														
3'					65					100				
2'		85							100					
1'	83					48		100						
1048							100							
1047	52					100								
1046					100									
...														
3			100											
2		100												
1	100													
	1	2	3	...	1046	1047	1048	1'	2'	3'	...	914'	915'	916'

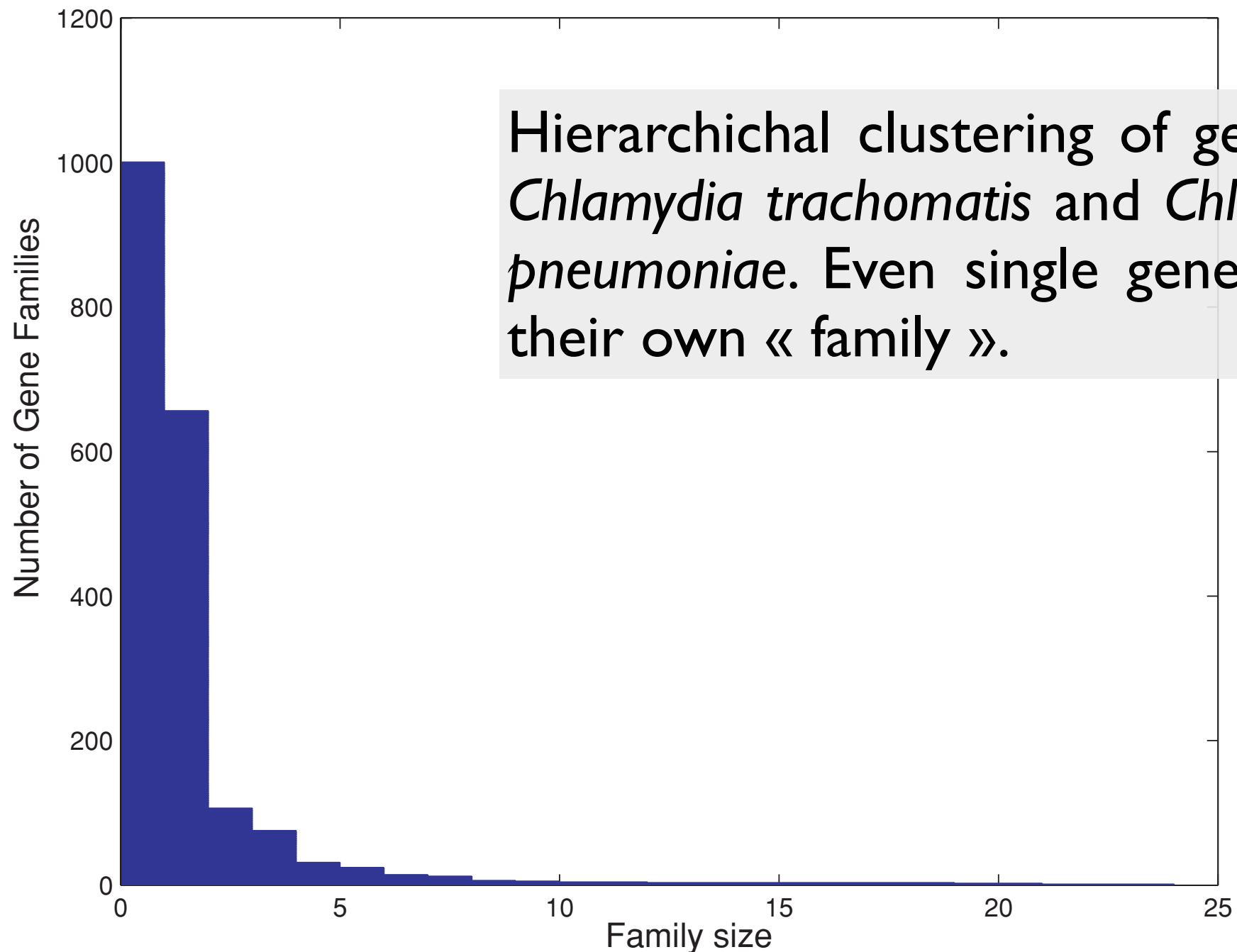
Gene Clustering

Markov Clustering / *OrthoMCL*

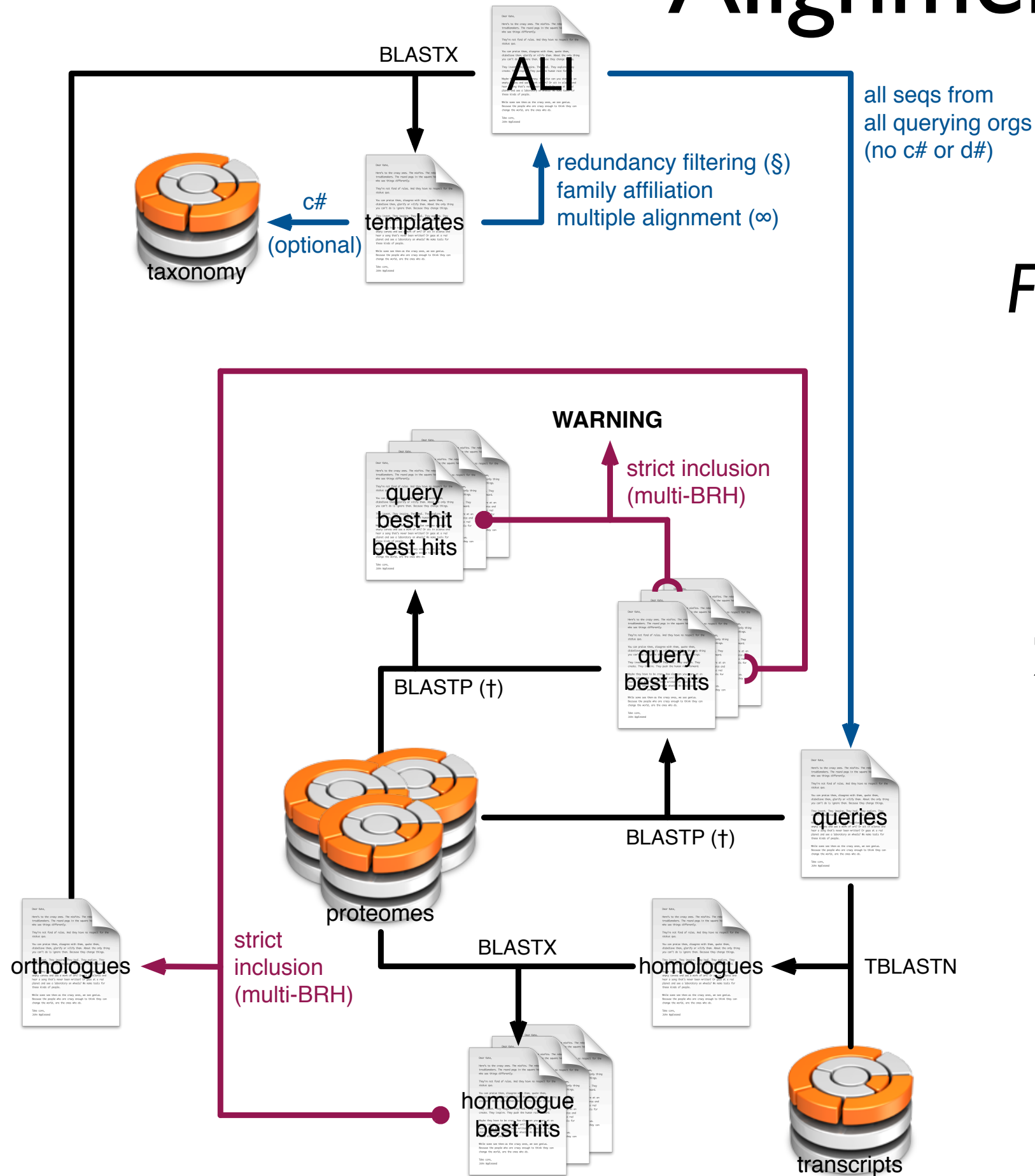


Gene Clustering

many small gene families and few large families



Alignment Enrichment

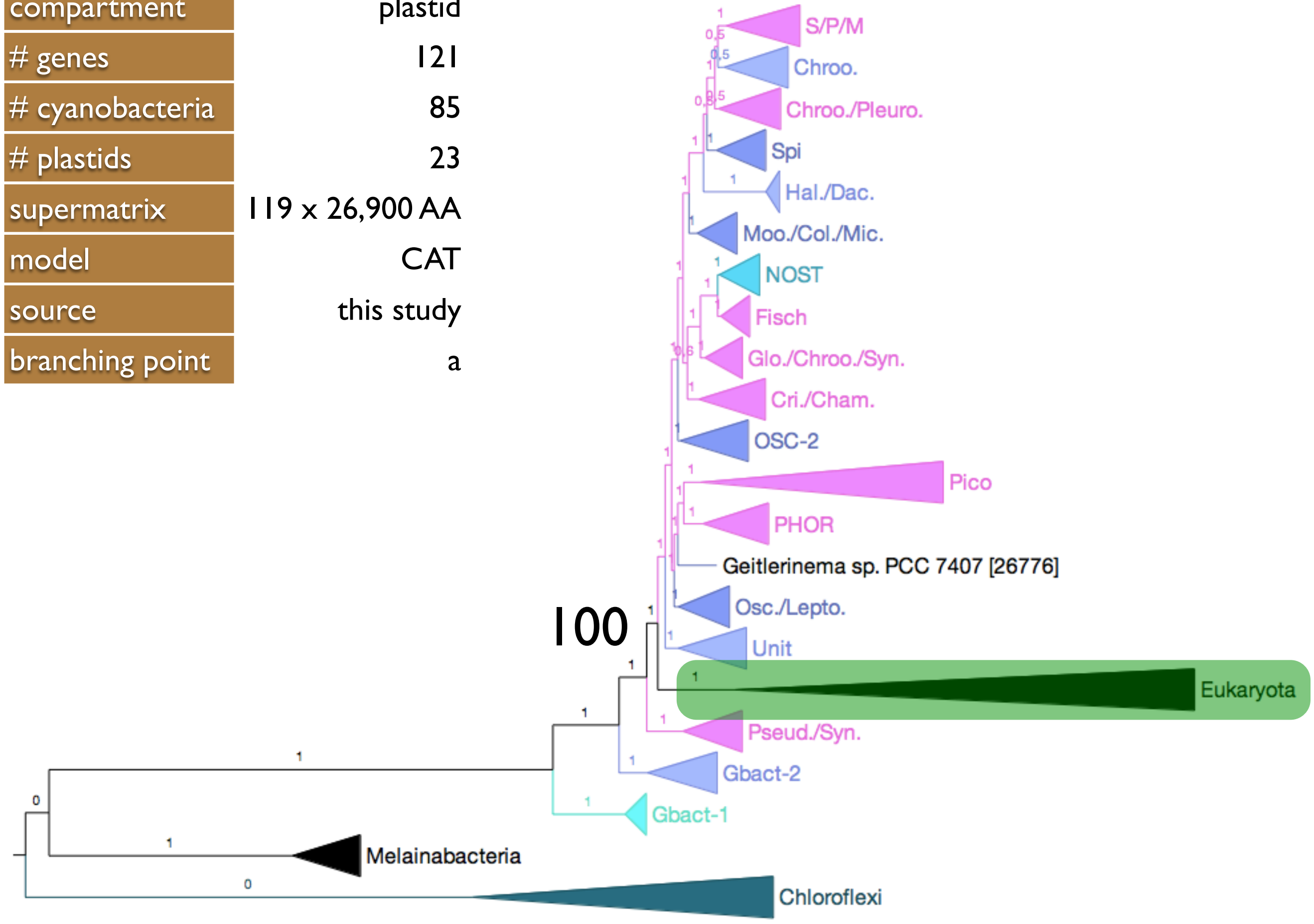


Forty 2

1. identifies **orthologs** in genomes or transcriptomes
2. controls for NGS **contaminations**
3. **aligns** the new sequences

compartment
genes
cyanobacteria
plastids
supermatrix
model
source
branching point

plastid
 121
 85
 23
 119 x 26,900 AA
 CAT
 this study
 a

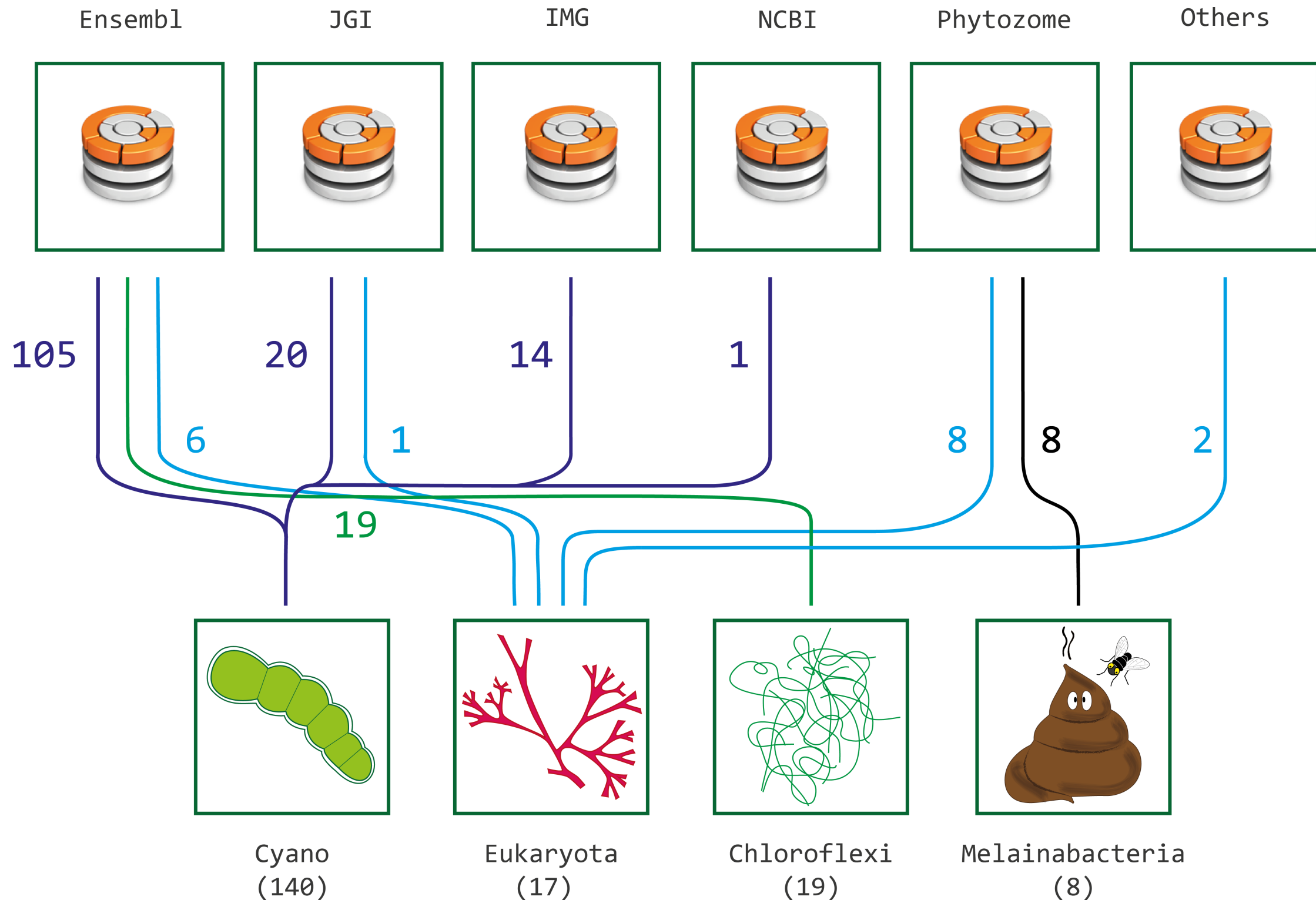


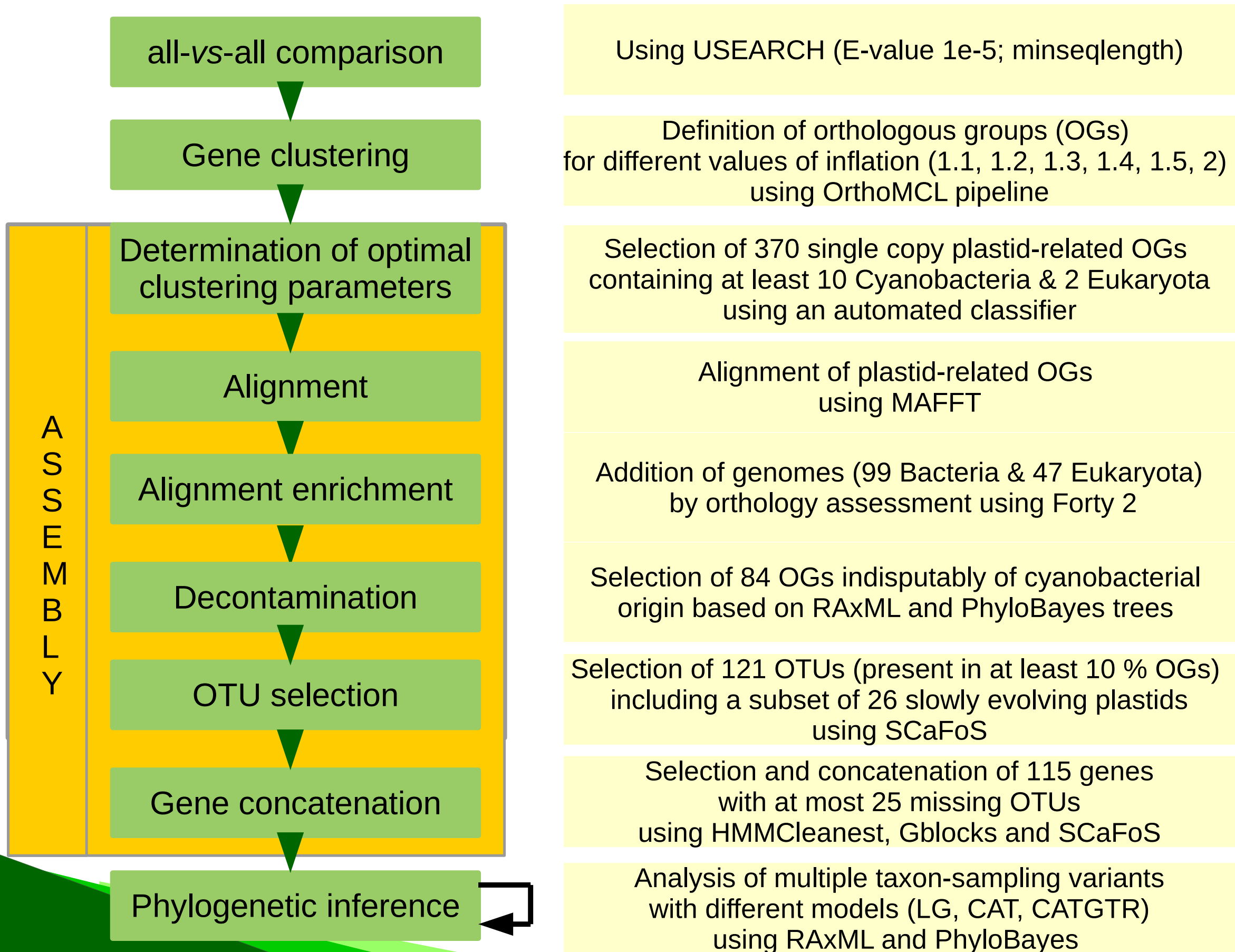
Plastid-encoded genes are difficult to analyze due to a highly biased evolutionary process.

How could we confirm this result?

Nuclear Supermatrix

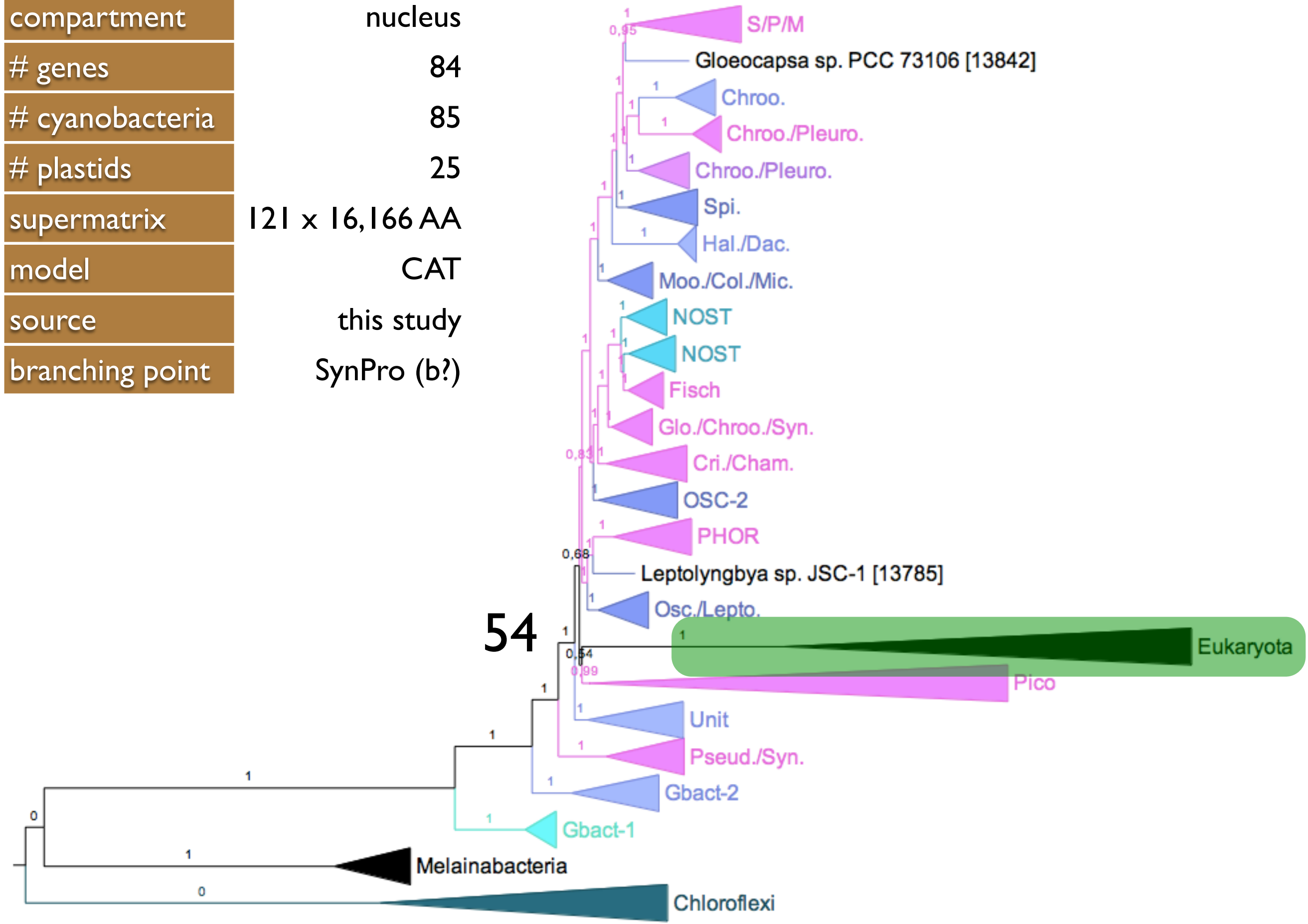
We assemble a completely disjoint dataset.





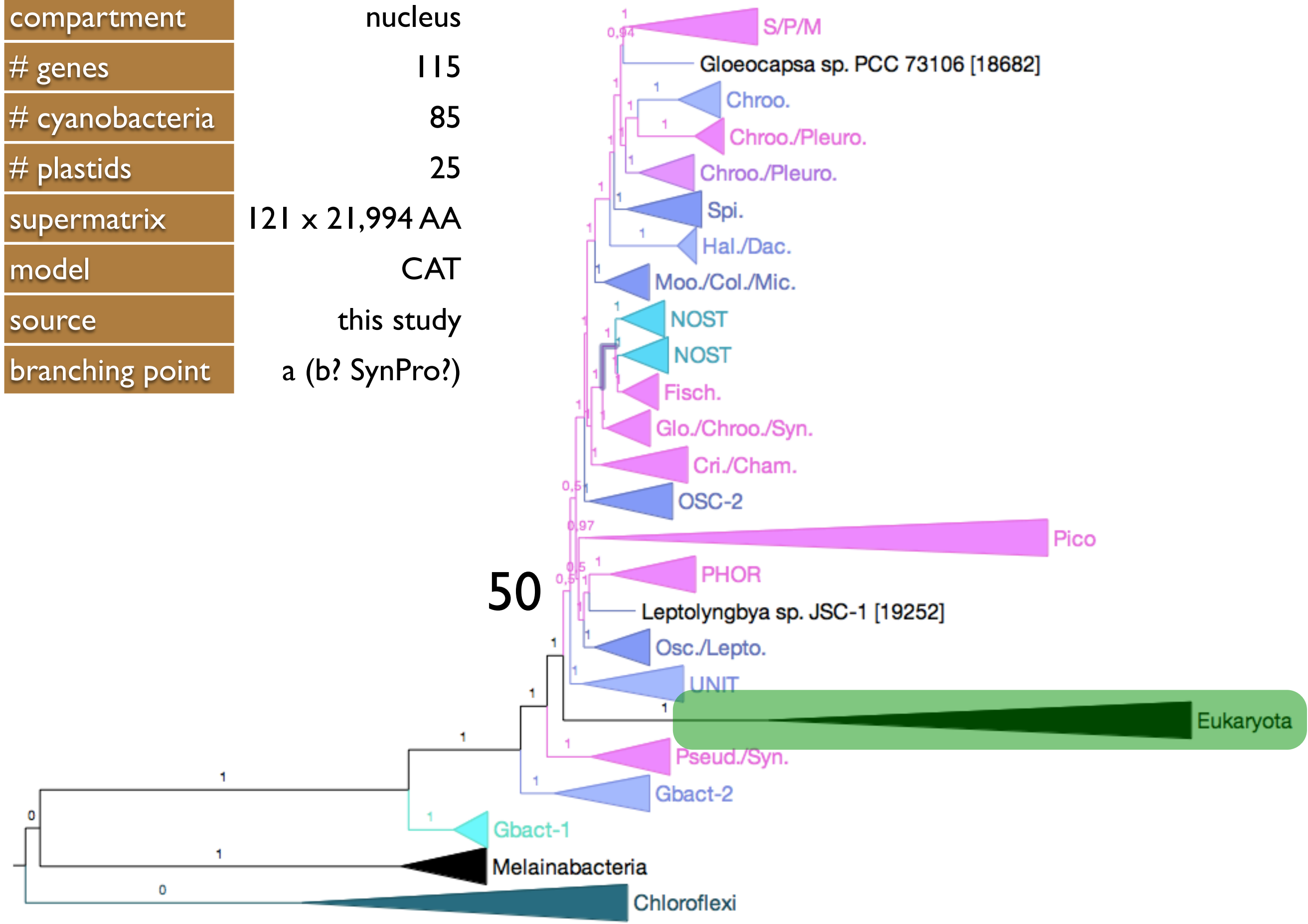
compartment
genes
cyanobacteria
plastids
supermatrix
model
source
branching point

nucleus
 84
 85
 25
 121 x 16,166 AA
 CAT
 this study
 SynPro (b?)

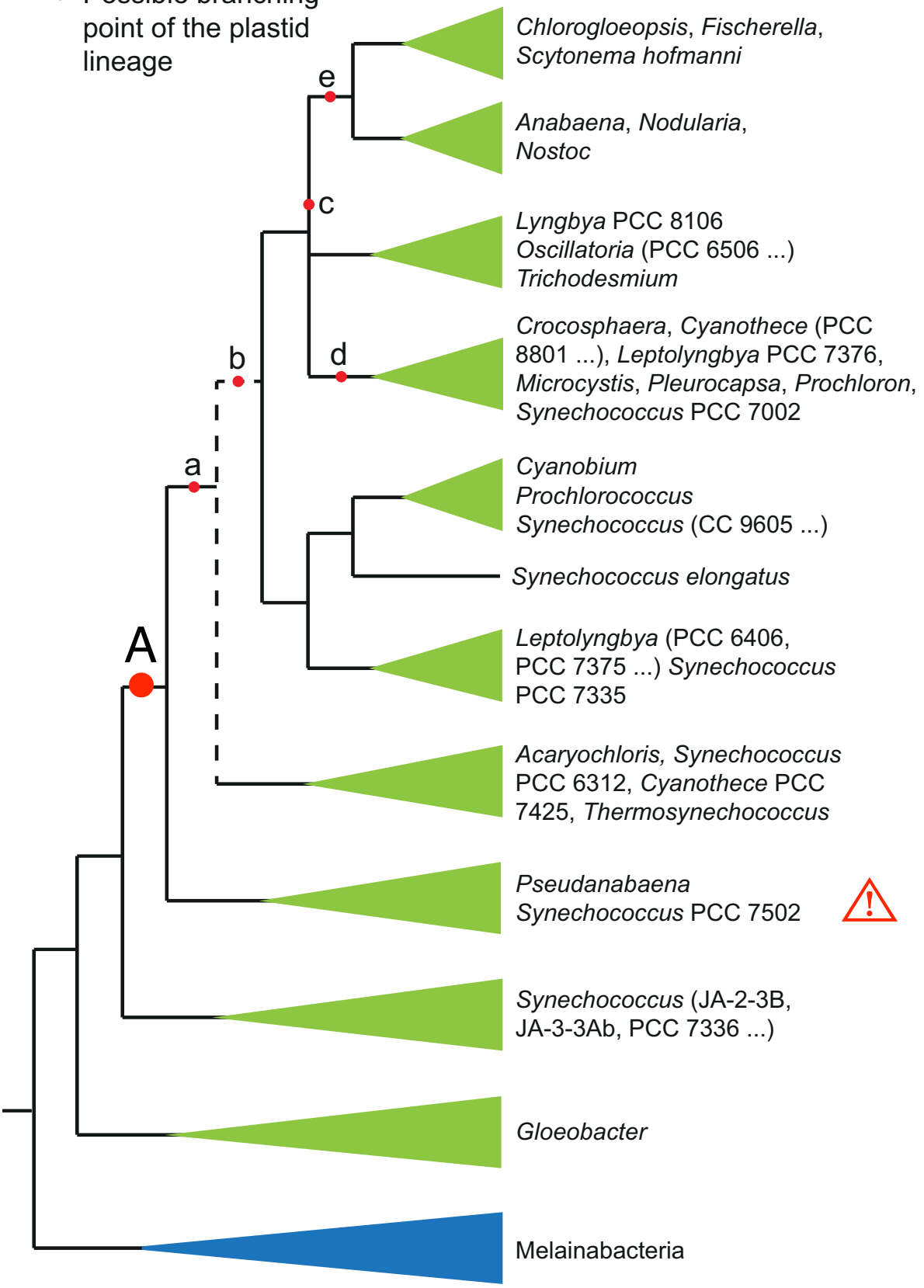


compartment
genes
cyanobacteria
plastids
supermatrix
model
source
branching point

nucleus
 115
 85
 25
 121 x 21,994 AA
 CAT
 this study
 a (b? SynPro?)

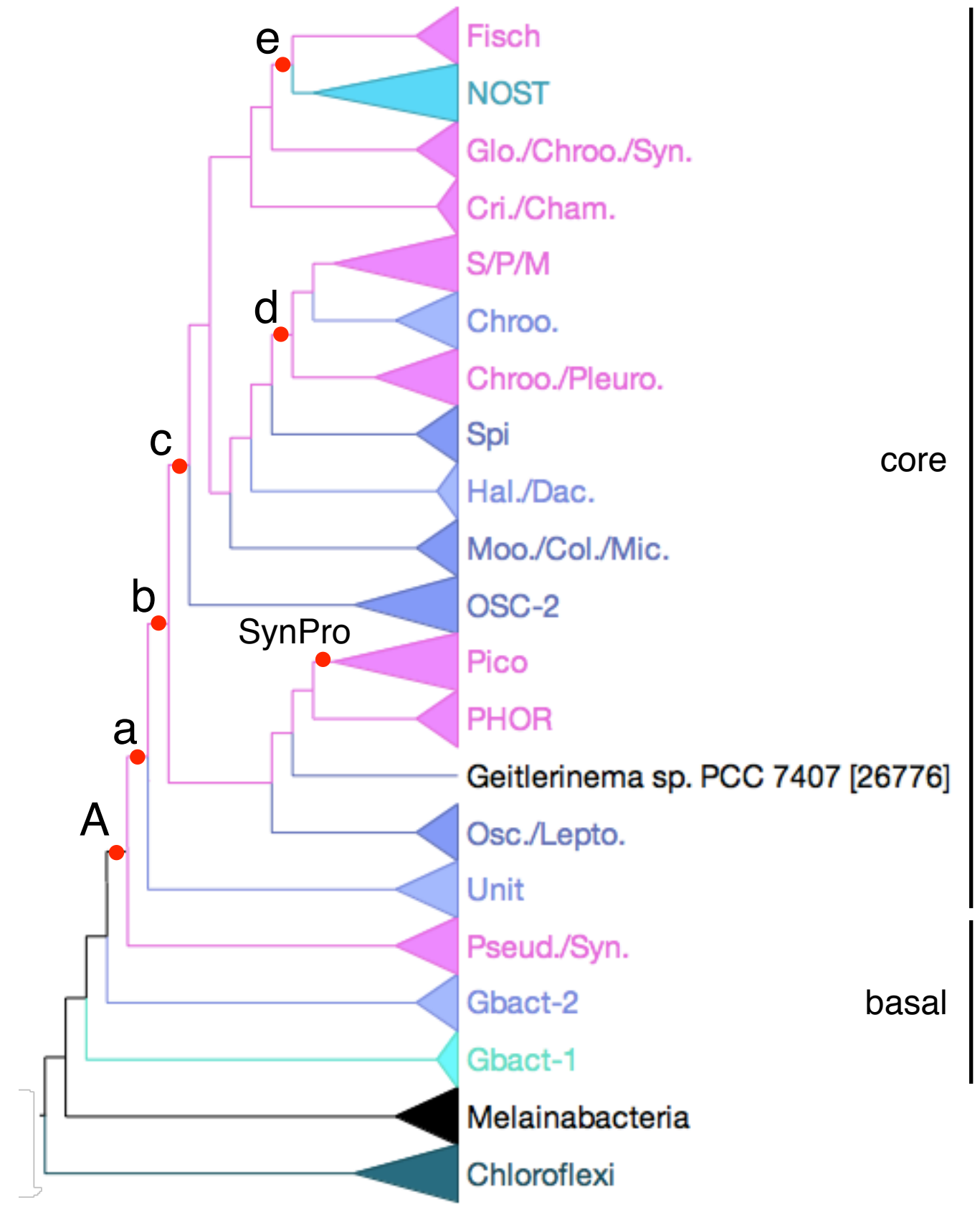


● Possible branching point of the plastid lineage



Core cyanobacteria

Basal cyanobacteria



Our work suggests a *not-so-early* branching point (a-b).



Even if the number of possibilities has lowered, additional analyses are needed to pinpoint the origin of eukaryotic plastids among Cyanobacteria.