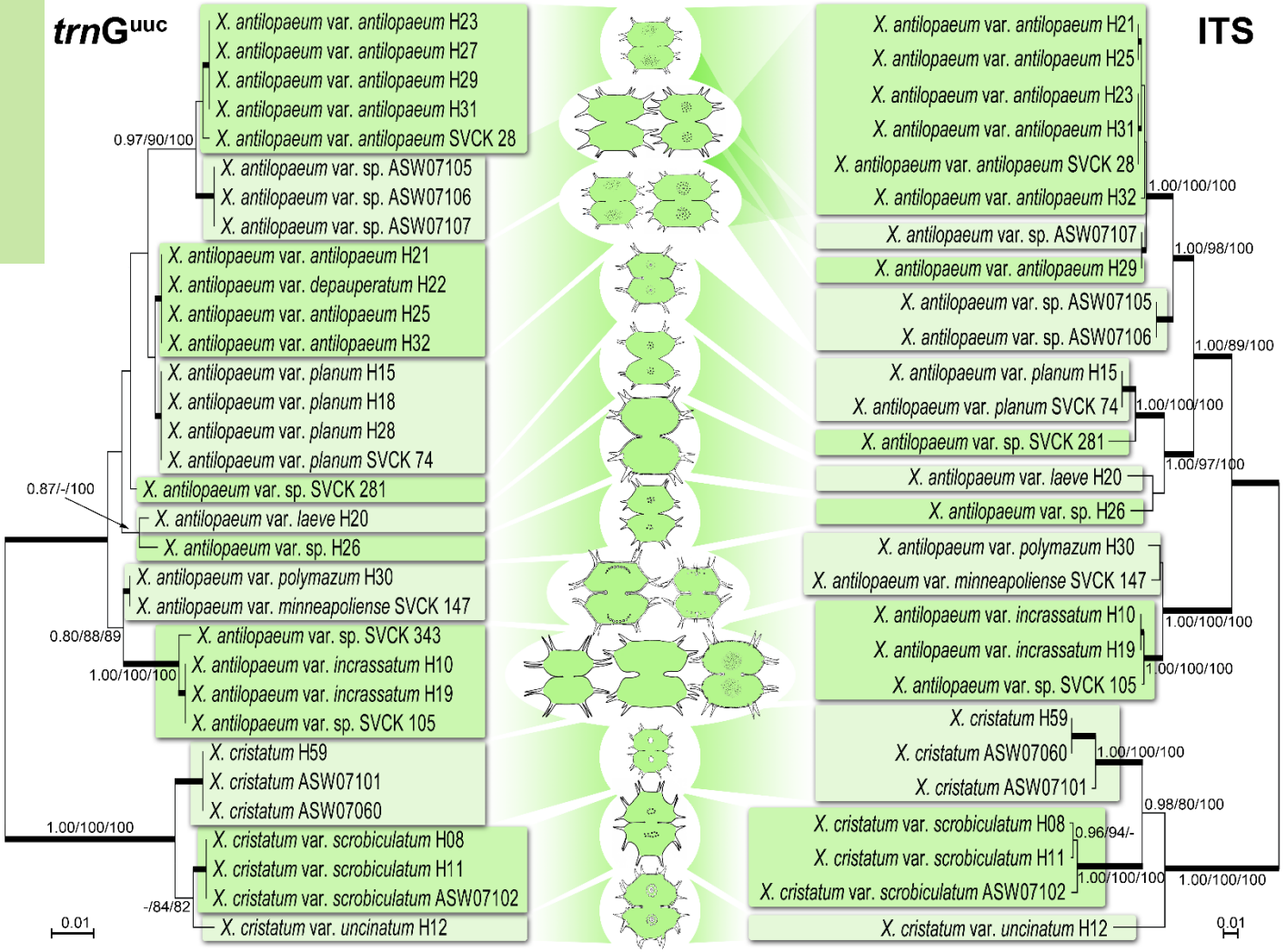
# *Typy fylogenetických analýz*

Distanční metody:
- Neighbor-Joining
- Minimum Evolultion
- UPGMA, ...

Maximum Likelihood

Bayesian Inference

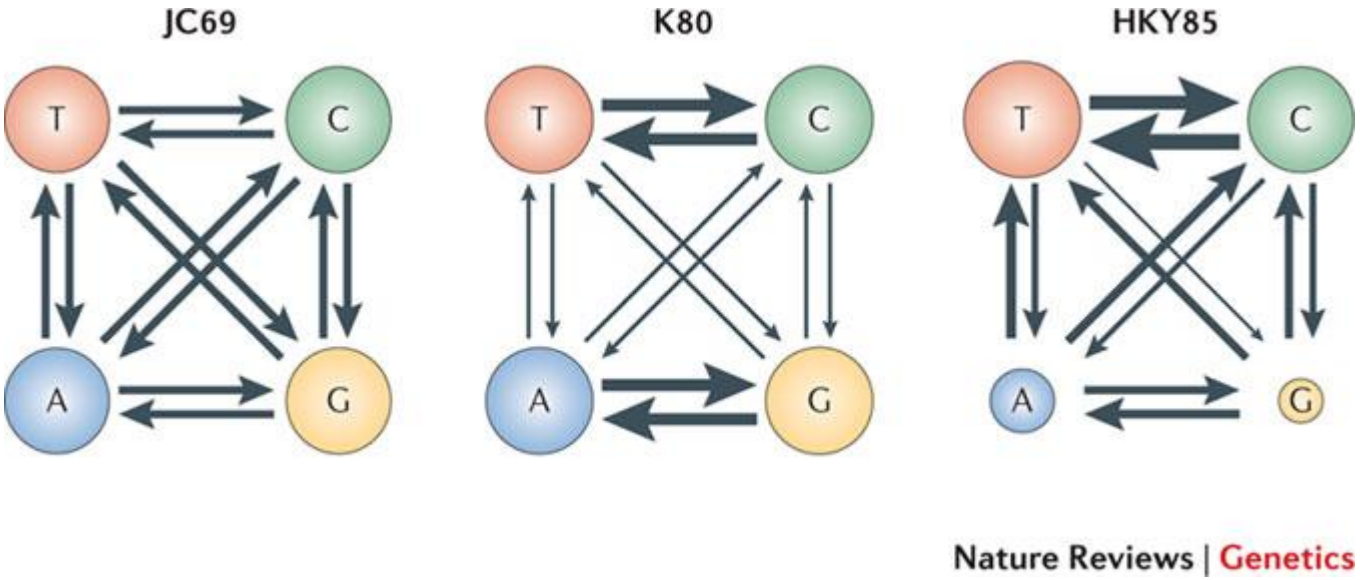
Maximum Parsimony



# *Genetické distance, substituční modely*

- pro výpočet fylogenetických analýz je nutné stanovit genetické (evoluční) distance mezi sekvencemi
- v případě molekulárních hodin je distance přímo úměrná času

- *p-distance*: prostý rozdíl sekvencí – výrazné podhodnocení reálných distancí (saturace)
- *substituční modely* – odhady jednotlivých substitučních rychlostí pomocí Markovových modelů + frekvence výskytu nukleotidů = Q matice



# *Substituční modely*

- Jukes-Cantor (JC69): během evoluce mají všechny nukleotidy stejnou pravděpodobnost substitucí i stejnou frekvenci výskytu bází (nst=1)

| | A | T | C | G |
|---|---|---|---|---|
| A | - | $\alpha$ | $\alpha$ | $\alpha$ |
| T | $\alpha$ | - | $\alpha$ | $\alpha$ |
| C | $\alpha$ | $\alpha$ | - | $\alpha$ |
| G | $\alpha$ | $\alpha$ | $\alpha$ | - |

- Felsenstein (F81): nukleotidy mají stejnou pravděpodobnost substitucí ale jinou frekvenci výskytu bází (nst=1)

| | A | T | C | G |
|---|---|---|---|---|
| A | - | $\alpha \Pi_T$ | $\alpha \Pi_C$ | $\alpha \Pi_G$ |
| T | $\alpha \Pi_A$ | - | $\alpha \Pi_C$ | $\alpha \Pi_G$ |
| C | $\alpha \Pi_A$ | $\alpha \Pi_T$ | - | $\alpha \Pi_G$ |
| G | $\alpha \Pi_A$ | $\alpha \Pi_T$ | $\alpha \Pi_C$ | - |

- Kimura (K80): jiné substituční rychlosti pro transice a transverze, shodné frekvence bází (nst=2)

| | A | T | C | G |
|---|---|---|---|---|
| A | - | $\beta$ | $\beta$ | $\alpha$ |
| T | $\beta$ | - | $\alpha$ | $\beta$ |
| C | $\beta$ | $\alpha$ | - | $\beta$ |
| G | $\alpha$ | $\beta$ | $\beta$ | - |

- Hasegawa-Kishino-Yano (HKY): jiné substituční rychlosti pro transice a transverze, různé frekvence bází (nst=2)

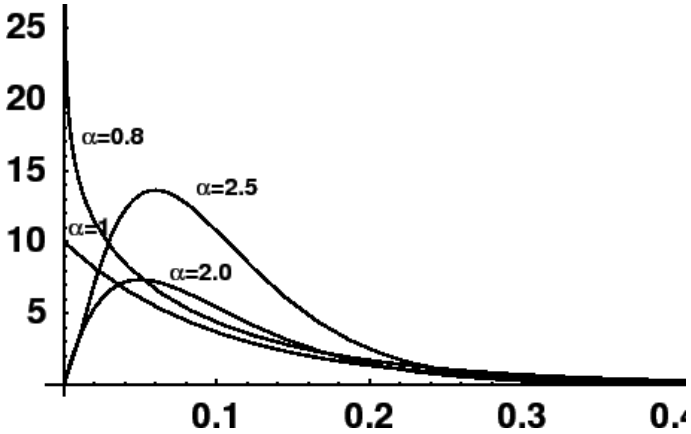
| | A | T | C | G |
|---|---|---|---|---|
| A | - | $\beta \Pi_T$ | $\beta \Pi_C$ | $\alpha \Pi_G$ |
| T | $\beta \Pi_A$ | - | $\beta \Pi_C$ | $\beta \Pi_G$ |
| C | $\beta \Pi_A$ | $\beta \Pi_T$ | - | $\beta \Pi_G$ |
| G | $\beta \Pi_A$ | $\beta \Pi_T$ | $\beta \Pi_C$ | - |

- General time reverdible (GTR): pravděpodobnosti substitucí a frekvence bází jsou specifikovány pro každou možnost (nst=6)

| | A | T | C | G |
|---|---|---|---|---|
| A | - | $\alpha \Pi_T$ | $\beta \Pi_C$ | $\gamma \Pi_G$ |
| T | $\alpha \Pi_A$ | - | $\delta \Pi_C$ | $\varepsilon \Pi_G$ |
| C | $\beta \Pi_A$ | $\delta \Pi_T$ | - | $\zeta \Pi_G$ |
| G | $\alpha \Pi_A$ | $\varepsilon \Pi_T$ | $\zeta \Pi_C$ | - |

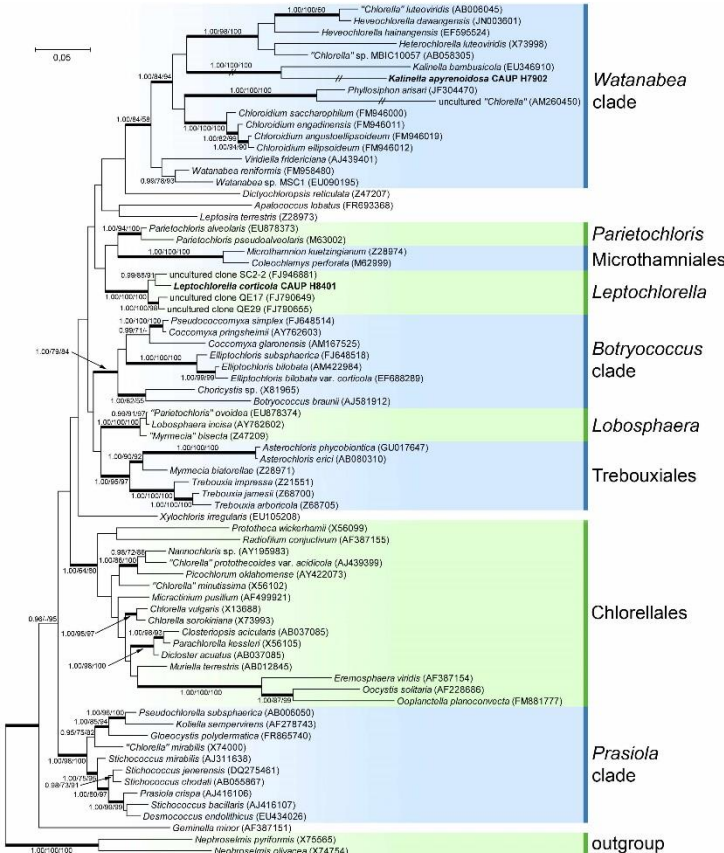
# *Substituční modely*

- Gamma distribuce (Γ): modeluje variabilitu v míře nukleotidových substitucí na různých pozicích alignmentu. Většinou se model zjednodušuje do 4 α kategorií



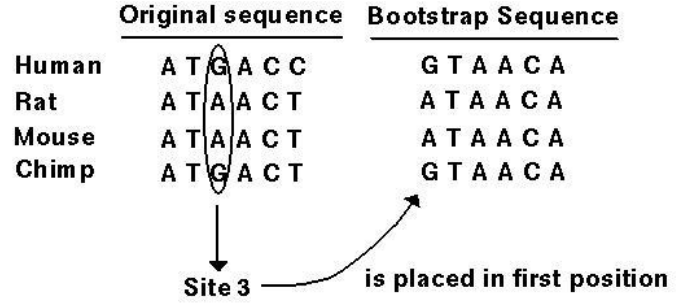

- Proporce nevariabilních míst (I): existence velkého množství nevariabilních pozic negativně ovlivňuje odhad genetických distancí. Aplikace I modelu je např. důležitá při současné přítomnosti krátkých a dlouhých větví

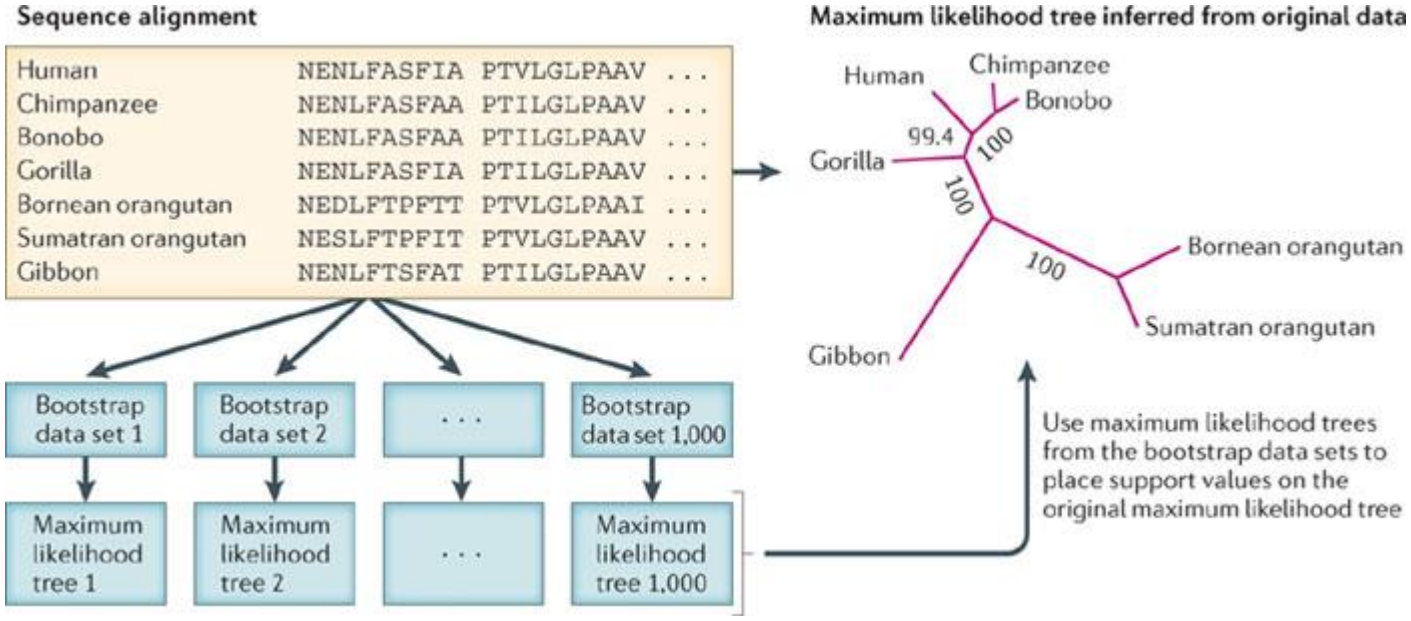
- Kovarion (cov): modeluje variabilitu v míře nukleotidových substitucí v závislosti na fylogenetické pozici dané sekvence



# Test topologie - bootstrapping

- výpočet stromů na základě nově generovaných alignmentů
- konstrukce majority-rule konsenzuálního stromu
- hodnoty bootstrapů je nutné zobrazit na topologii stromu zkonstruovaného na základě originálního alignmentu



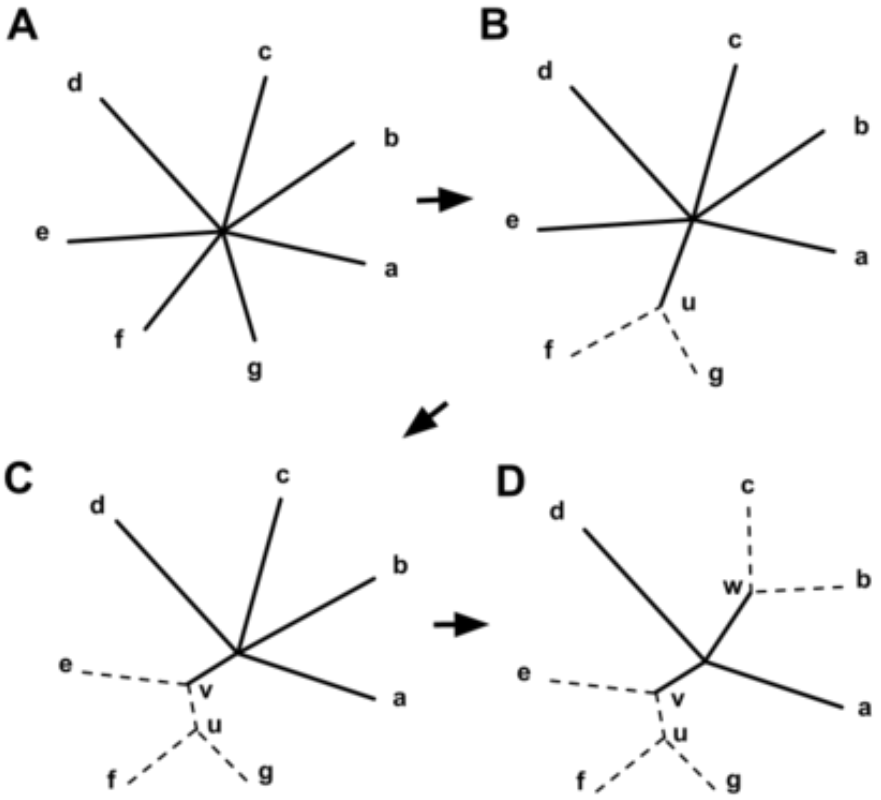
# *Fylogenetické analýzy na základě distančních matic*

- stromy se počítají na základě distancí, spočtených pro každou dvojici sekvencí
- aplikují se substituční modely
- vhodné pro rychlou analýzu velkého množství sekvencí (v řádu několika stovek až tisíc)

- Minimum Evolution (ME): hledá se strom o nejmenším součtu délek všech větví

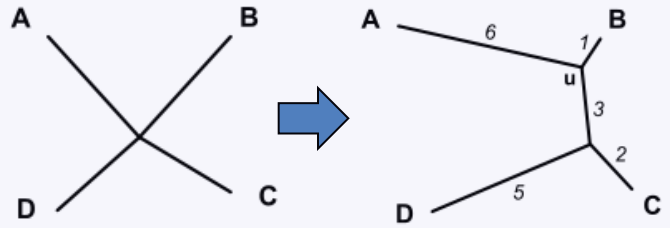
- Neighbor Joining (NJ): heuristický algoritmus na rychlé nalezení ME stromu (začíná se u hvězdicovitého stromu)

- BioNJ: lepší přesnost topologie u vzdáleně příbuzných sekvencí



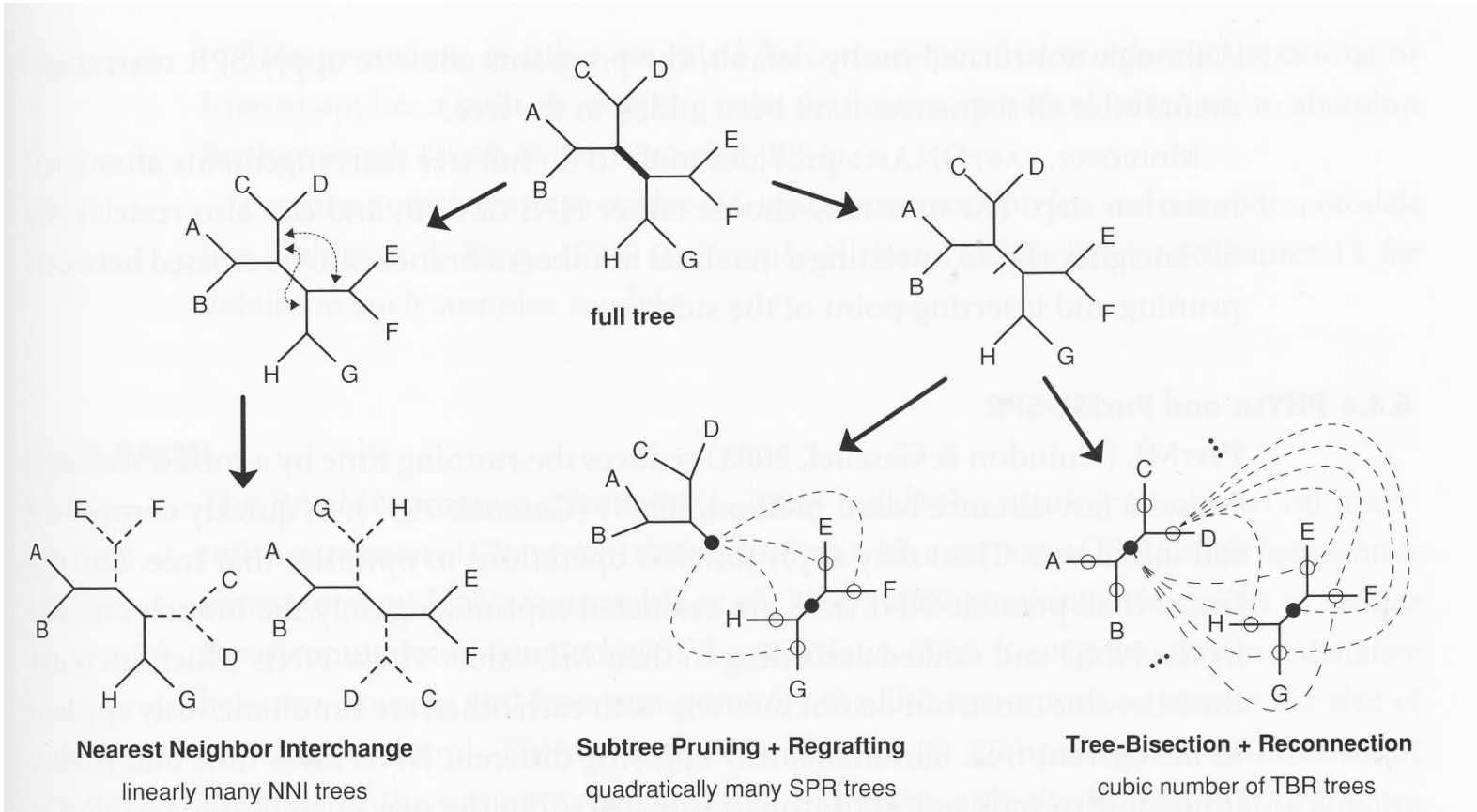

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 7 | 11 | 14 |
| B | 7 | 0 | 6 | 9 |
| C | 11 | 6 | 0 | 7 |
| D | 14 | 9 | 7 | 0 |

| | u | C | D |
|---|---|---|---|
| u | 0 | 5 | 8 |
| C | 5 | 0 | 7 |
| D | 8 | 7 | 0 |



# *Maximum Likelihood (ML)*

- hledání nejpravděpodobnějšího stromu odrážejícího evoluci sekvencí
- posuzování pravděpodobností nekonečně velkého množství stromů (různé topologie, délky větví, parametrů substitučních modelů, ...)
- heuristické metody pro hledání struktury stromu:
  - Nearest Neighbor Interchange (NNI)
  - Subtree Pruning + Regrafting (SPR)
  - Tree-Bisection + Reconnection (TBR)

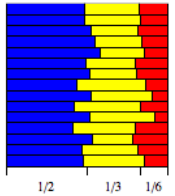


# *Bayesova analýza (BI)*

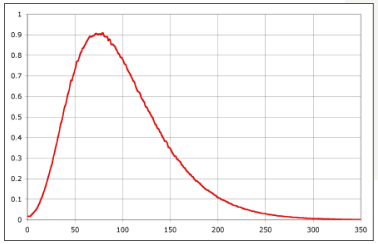
- tradiční otázka: pokud je v košíku stejně modrých a červených kuliček, jaká je pravděpodobnost, že si vytáhnu 3 modré a 3 červené kuličky?

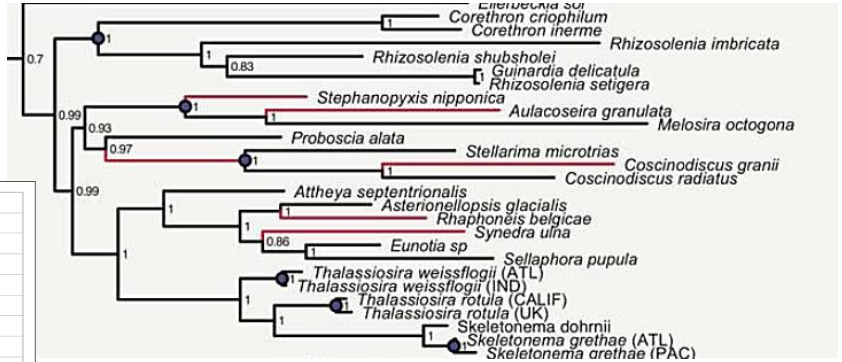
- Bayesovská otázka: pokud si vytáhnu 3 modré a 3 červené kuličky, jaká je pravděpodobnost, že je v košíku stejně modrých a červených kuliček?
  ➡ *podmíněná pravděpodobnost* (posterior)

- existuje nekonečné množství předpokladů (víc modrých, víc červených,...), které ale mohou mít různou pravděpodobnost ➡ tzv. priors
  - uniform priors – stejná pravděpodobnost, žádné předpoklady (*topologie*)
  - exponencial priors – např. *délky větví* (likelihood je negativní exponenciální funkcí)
  - dirichlet priors – pravděpodobnosti oscilují okolo dané hodnoty (*frekvence bází, substituční modely, I, ...)*
  - lognormal priors – např. *kalibrace stromu fosilními daty*



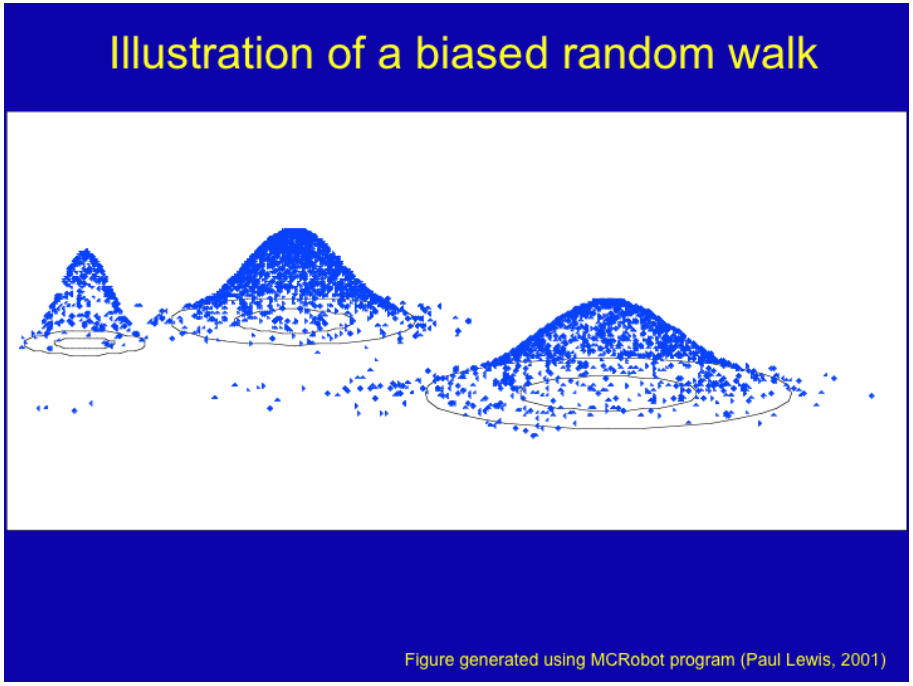

- priors se během analýzy mění na základě analyzovaných dat (alignment sekvencí), pomocí stochastických modelů
  ➡ získáme posteriorní pravděpodobnosti

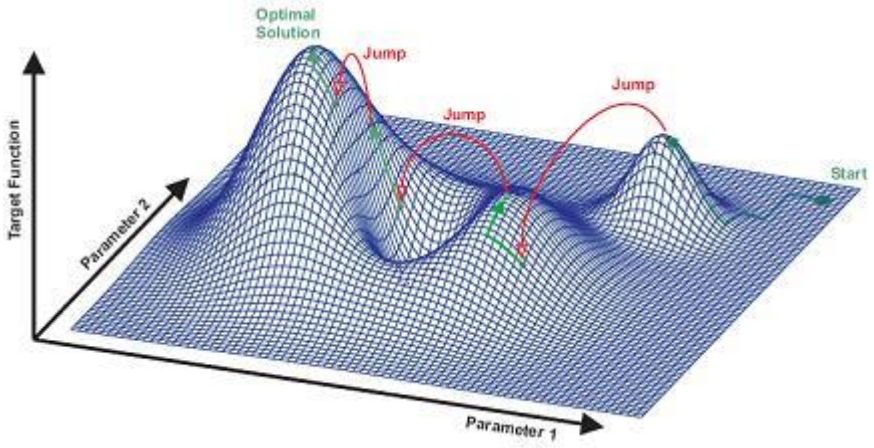


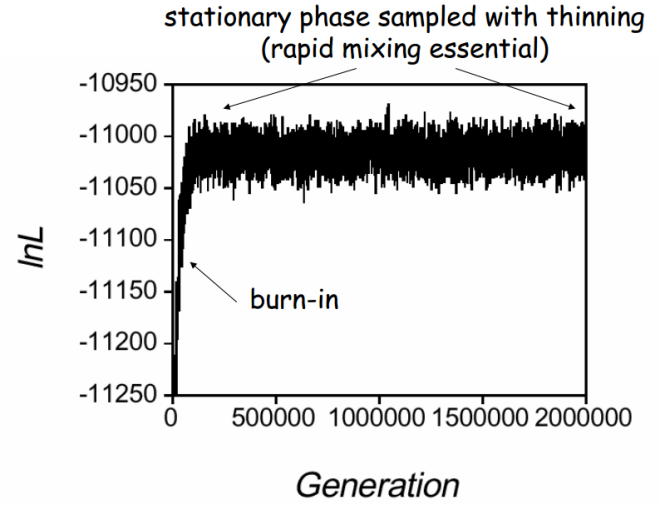


# *Bayesova analýza (BI)*

- Markov chain Monte Carlo (MCMC) sampling
  - výpočet posteriorních pravděpodobností pomocí náhodných změn prior parametrů a ty buď zamítnout či přijmout na základě jejich pravděpodobností
  - Metropolic coupling MCMC = $(MC)^3$ – 1 studený a 3 horké řetězce



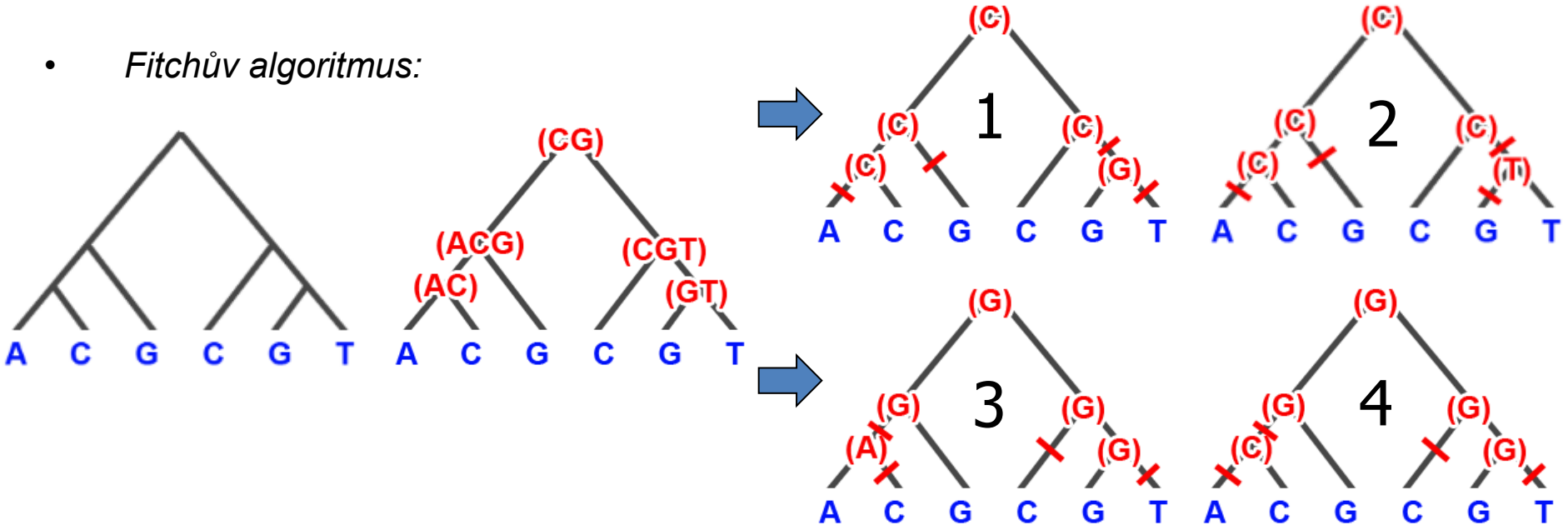




- burn-in: odstranění iniciální fáze MCMC
- výsledná topologie: konsenzuální strom posteriorních topologií

# *Maximální parsimonie (MP)*

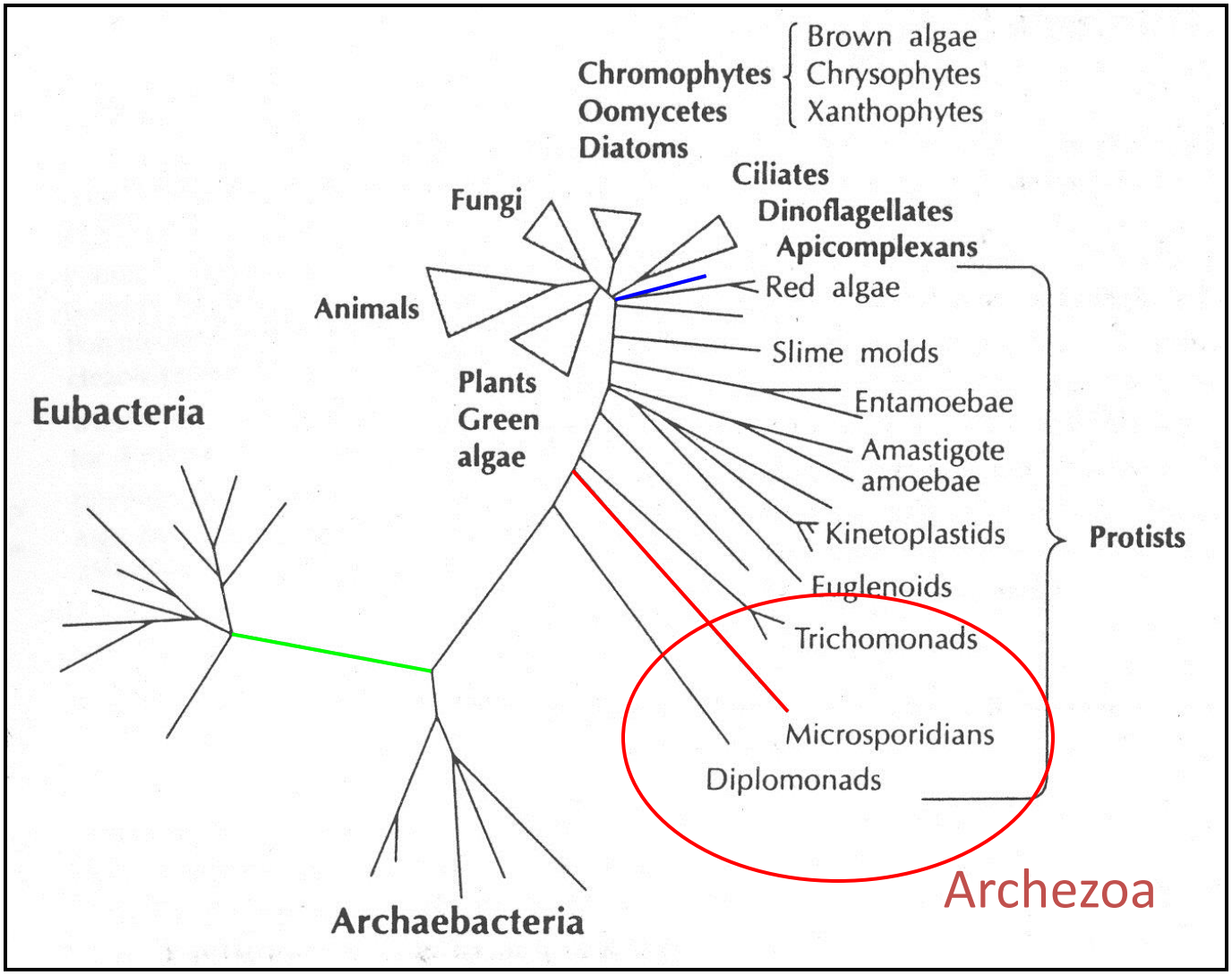
- hledání stromu s co nejmenším počtem evolučních kroků
- heuristické prohledávání stromů stejné jako u ML: NNI, SPR, TBR

- *Fitchův algoritmus:*

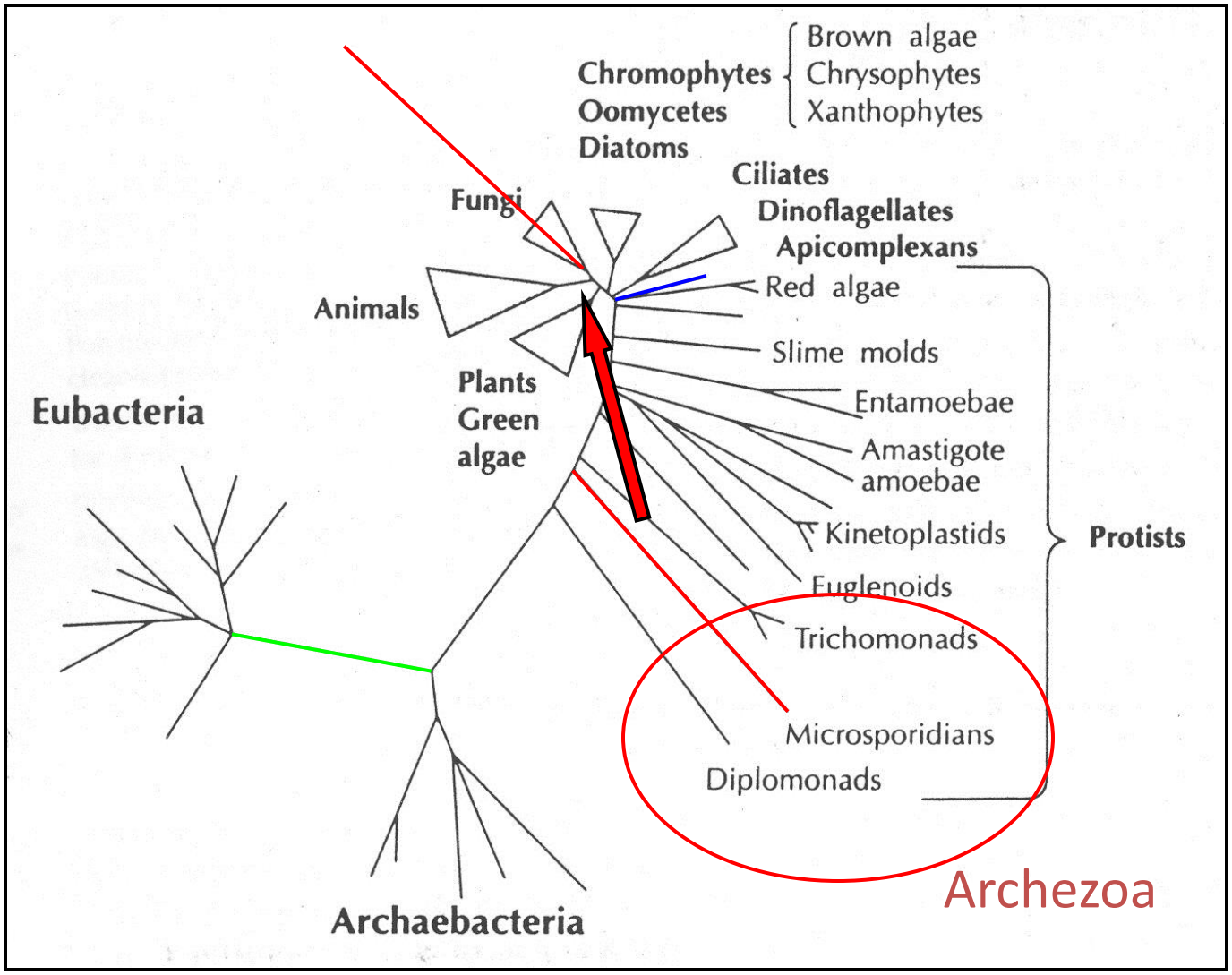



- vážená parsimonie (wMP):
  - MP výpočet parsimonního skóre: 1 pro substituci, 0 pro žádnou změnu
  - wMP výpočet: každý typ substituce je vážen např. pomocí frekvence jeho výskytu (*rescaled consistency index* - méně častým mutacím je dáno vyšší skóre)

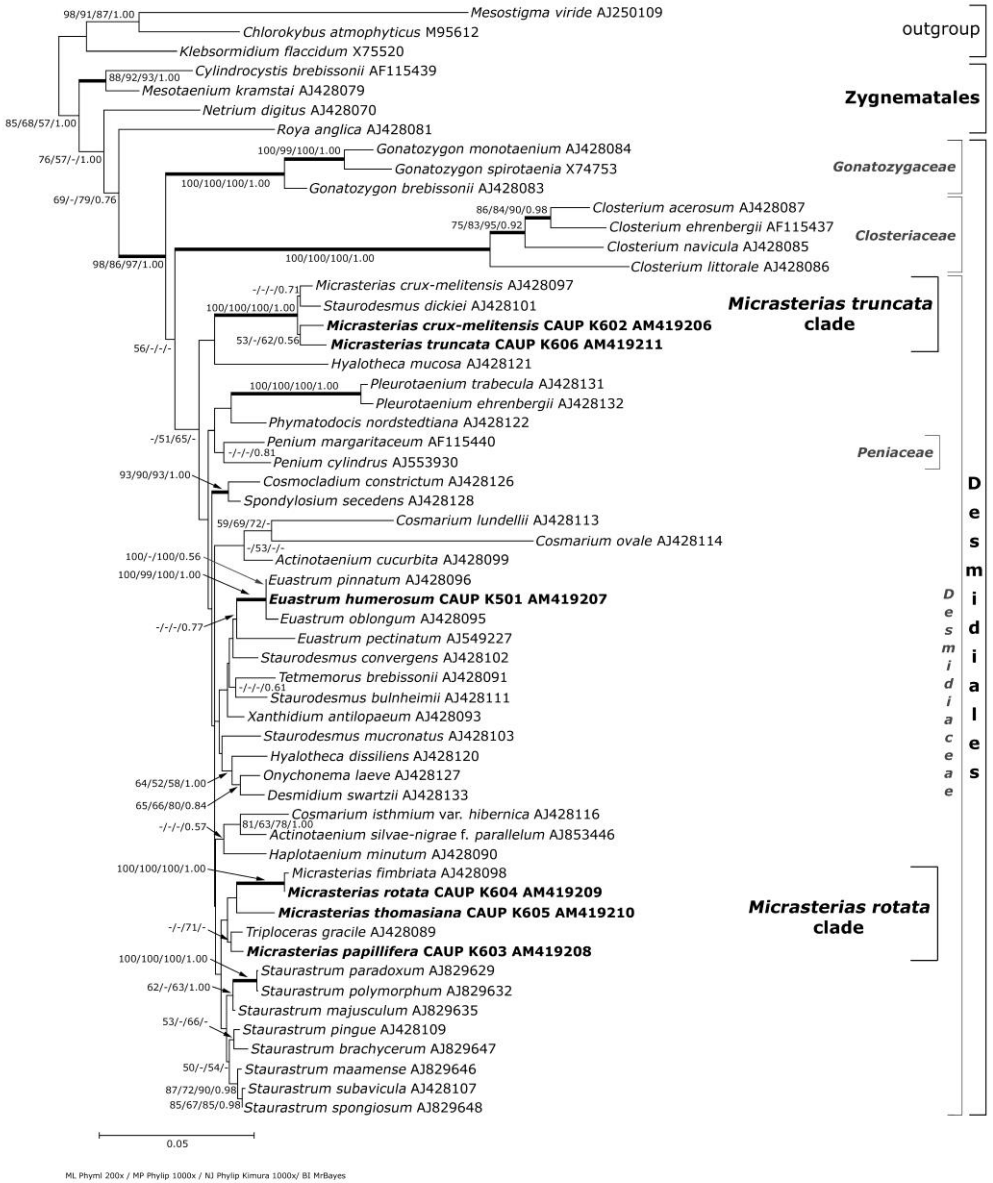
# Long branch attraction



# Long branch attraction

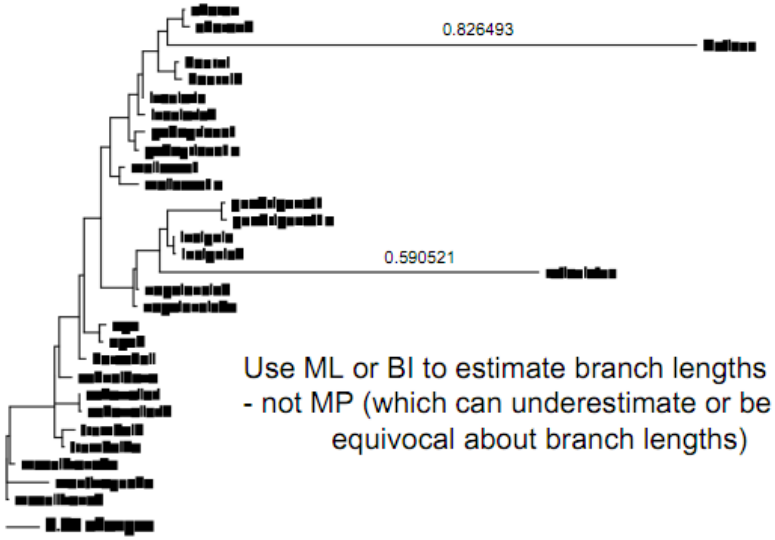


# Long branch attraction



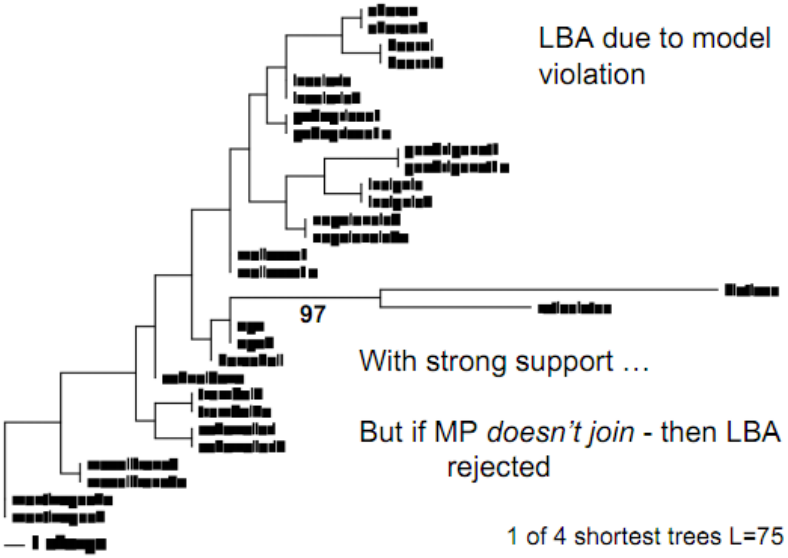
# Long branch attraction

**1. Are there long branches?**




Use ML or BI to estimate branch lengths
- not MP (which can underestimate or be equivocal about branch lengths)

**2. Are they joined in Parsimony analyses?**




LBA due to model violation

With strong support …

But if MP *doesn't join* - then LBA is rejected

1 of 4 shortest trees L=75

# kontaminace, špatná identifikace, ...

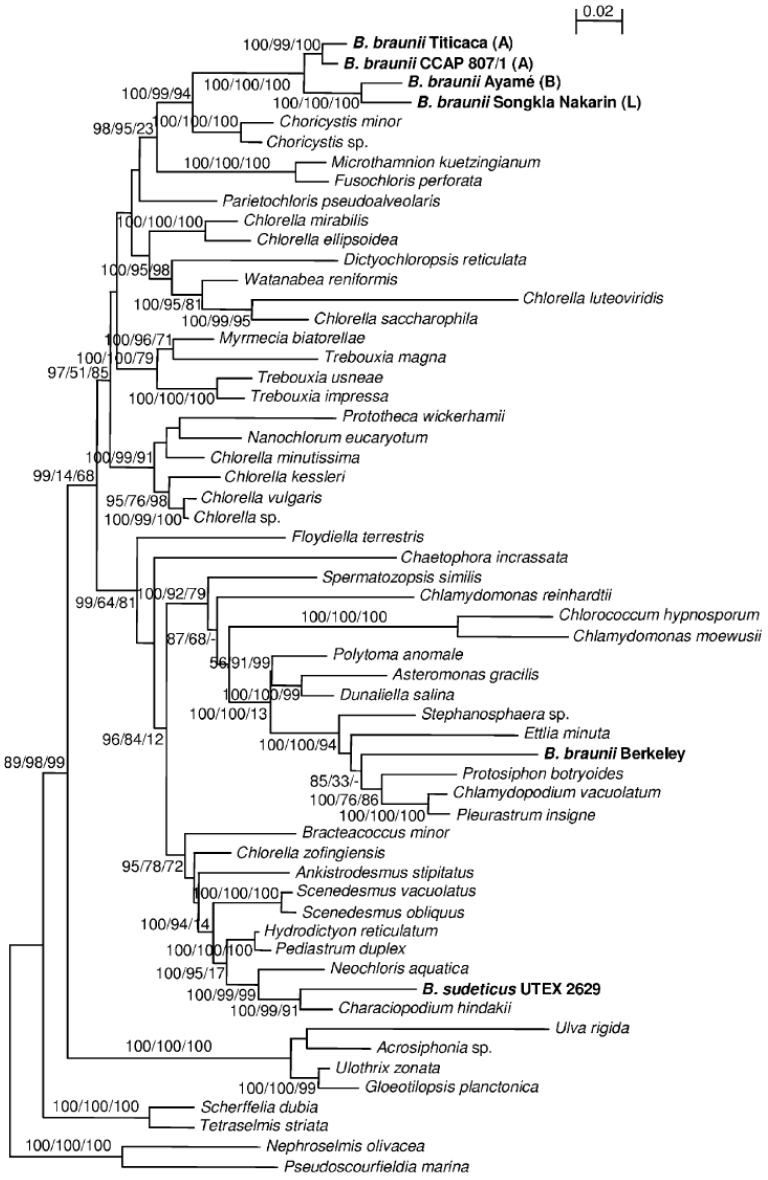


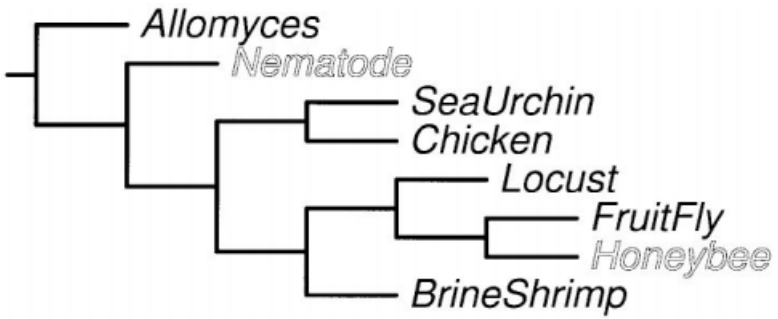

FIG. 2. Phylogenetic tree of 18S rRNA sequences inferred with the maximum likelihood method. Numbers shown at the branches are the credibility percentages using Bayesian inference (left), and bootstrap percentages using weighted parsimony (middle), and minimum evolution (right) methods. Percentages are shown only for branches that have credibility values above 50%. The horizontal lengths are proportional to the estimated number of substitutions per site. The scale bar is for 0.02 substitutions per site. Taxa shown in bold letters are the isolates whose 18S rRNA sequences were determined in this study.

# base composition bias

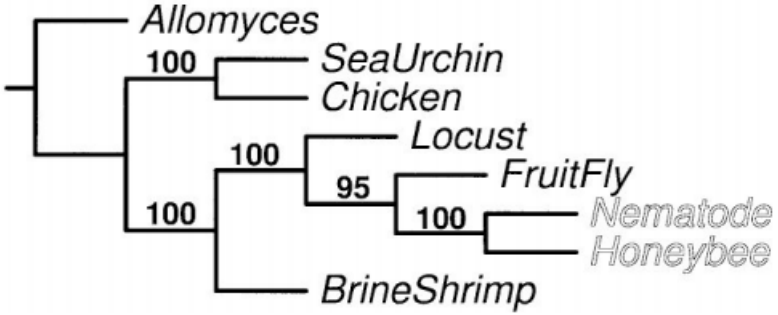
sekvence jsou shlukovány  podle velkého GC obsahu

Foster, P. G. & Hickey D. A. (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48: 284-290.

Correct tree:




ML tree - wrong due to base composition bias



# použití špatných markerů

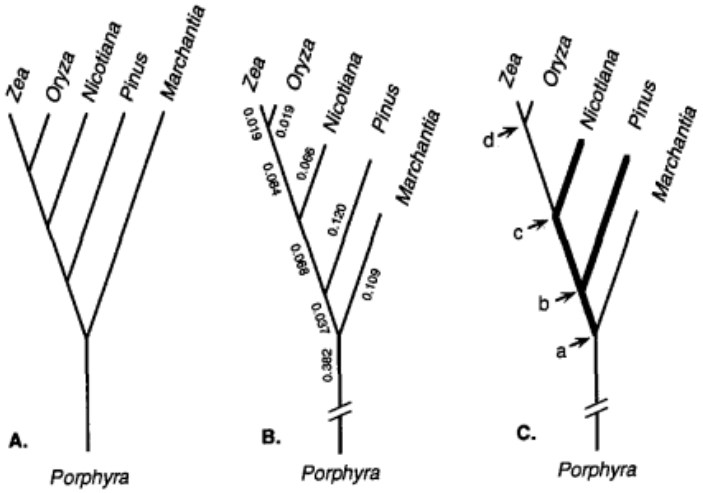


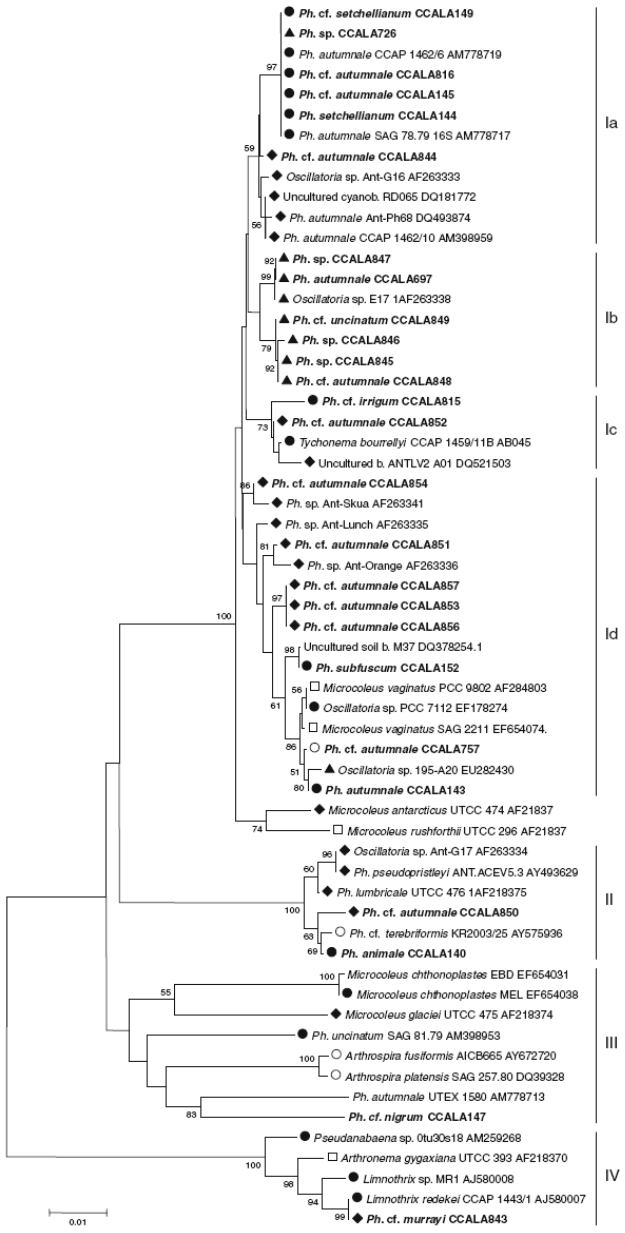

Fig. 1. Rooted 5-species trees for the taxa considered in this paper. *A* The true biological tree. *B* The NJ-tree of Dayhoff distances on the basis of the complete concatenated 14295 amino acid data set. Branch lengths are indicated. The *Porphyra* branch is not drawn to full length for convenience. *C* The strategy for estimating *Nicotiana*–*Gramineae* and angiosperm–*Pinus* divergence. Nodes *a*, *b*, *c*, and *d* denote branches referred to in the text. The portion of the tree shown in bold lines indicates branches used to estimate divergence times

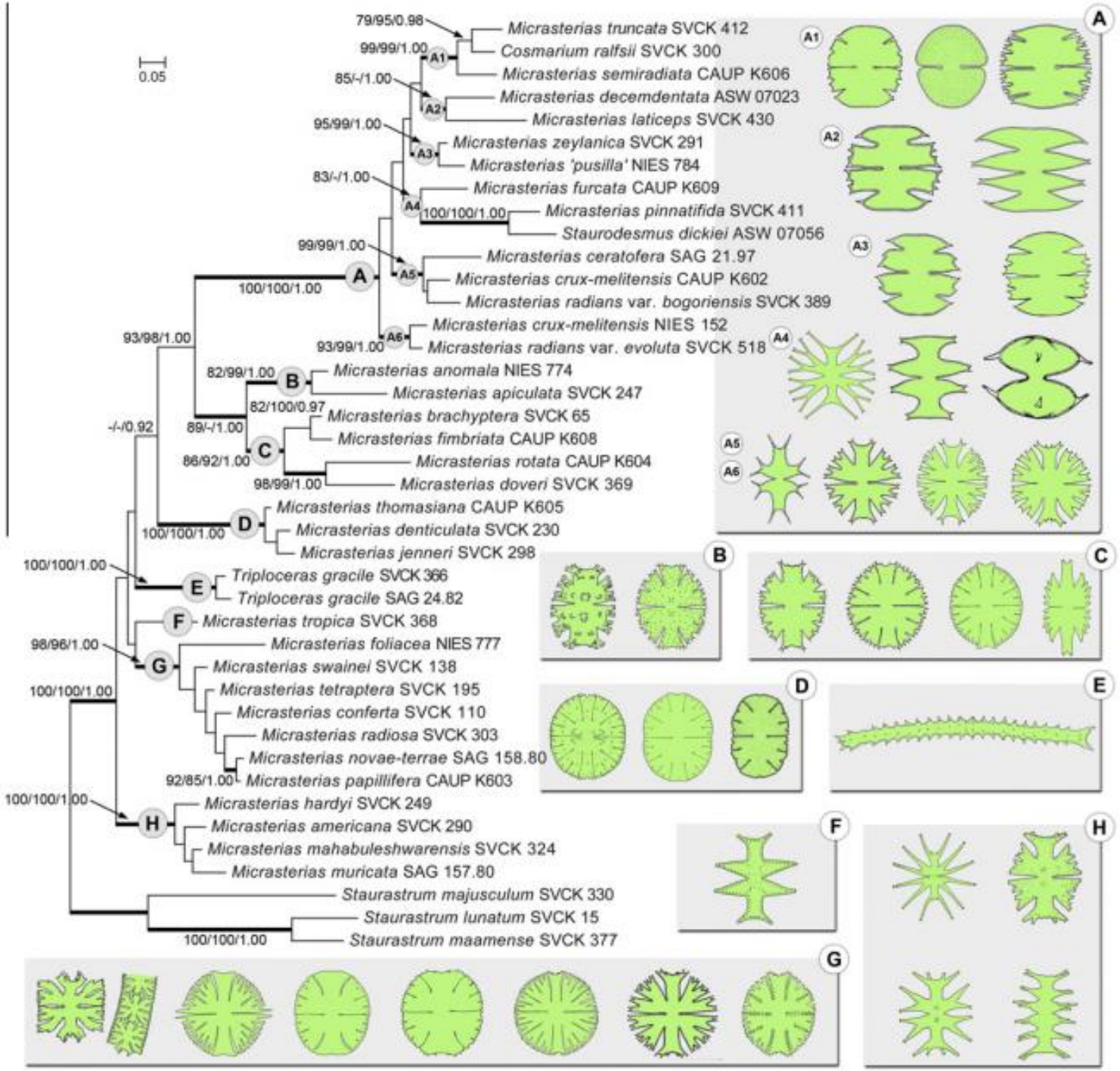
| Gene name | Land plant tree length[1] | True tree[2] |
|---|---|---|
| psbF[3] | 0.0359 | no |
| atpH[3] | 0.0432 | no |
| petB | 0.0919 | no |
| psbD | 0.0935 | yes |
| petD | 0.1121 | yes |
| psbC | 0.1209 | no |
| psbL | 0.1485 | yes |
| psbA | 0.1490 | yes |
| psbE | 0.1506 | yes |
| psaB | 0.1511 | yes |
| psbI | 0.1538 | no |
| psaA | 0.1548 | yes |
| psaC | 0.1747 | no |
| psbB | 0.1847 | yes |
| rbcL[3] | 0.1889 | no |
| orf29 | 0.2092 | no |
| atpB | 0.2360 | no |
| rps12 | 0.2442 | yes |
| psbN | 0.2447 | yes |
| psbT | 0.2447 | yes |
| psbJ | 0.2584 | no |

# špatná bootstrapová podpora

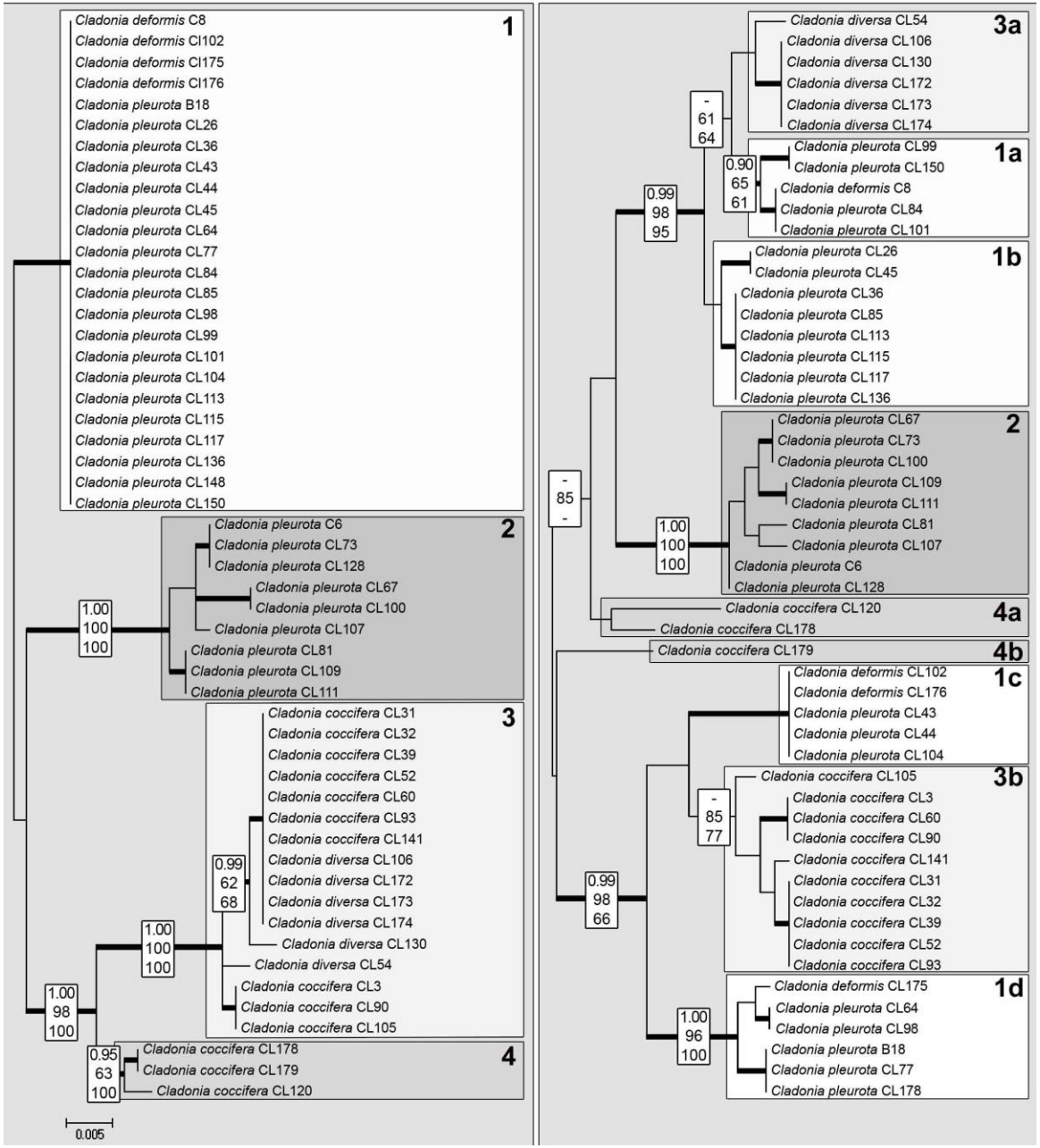


**Fig. 3** Phylogenetic relationships of *Phormidium* and *Phormidium*-like cyanobacteria estimated by neighbour-joining of 16S rRNA gene. Original sequences from this study are in *bold*. The symbols denote the following: *filled circle* European, *filled diamond* Antarctic, *filled triangle* Arctic, *open square* American and *open circle* other temperate zones origin. The origin of species without *symbol* is unknown. The evolutionary distances were computed using the maximum composite likelihood method, and values indicate nodes with bootstrap values higher than 50%



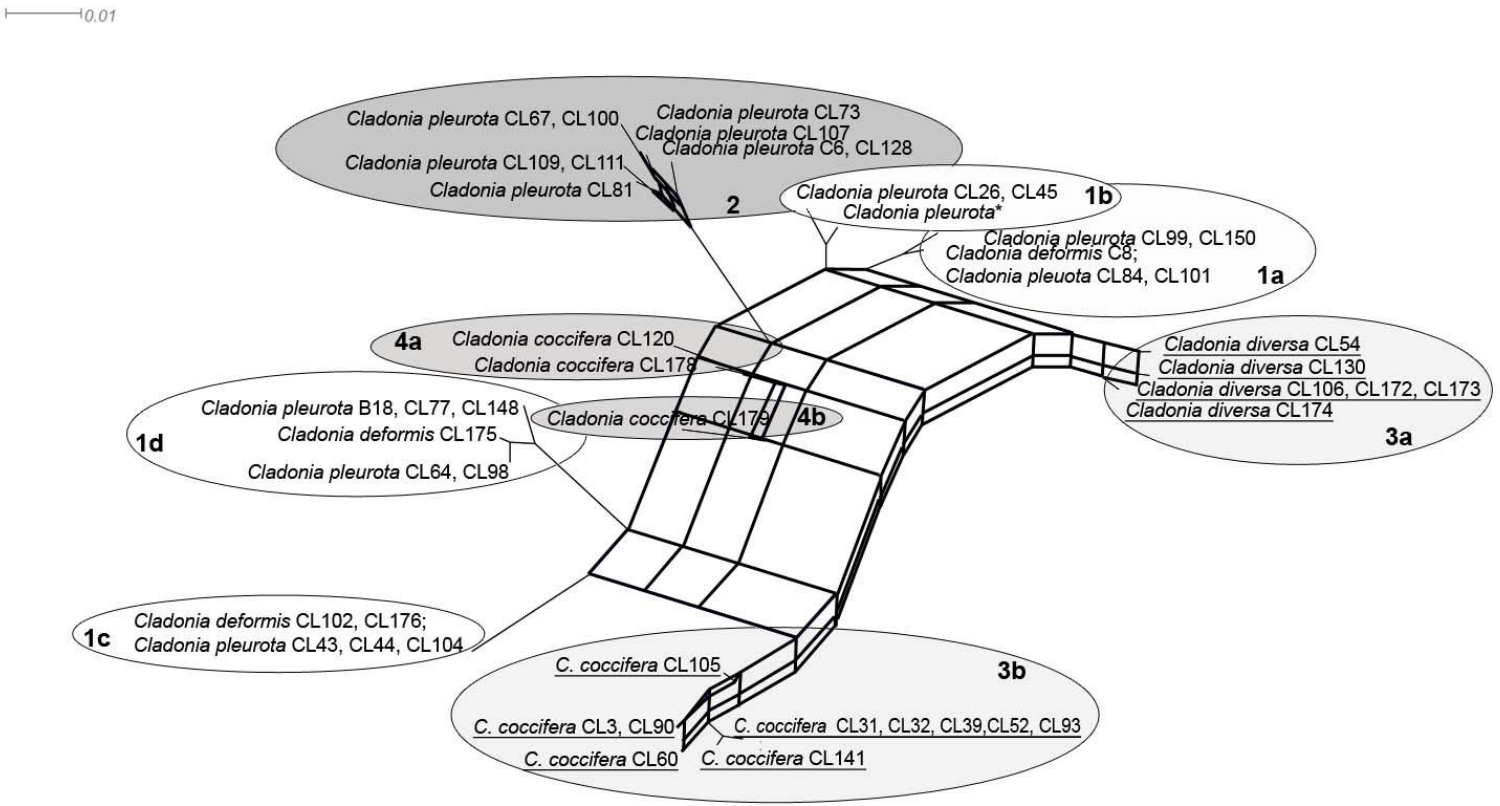

0.05
79/95/0.98
Micrasterias truncata SVCK 412
99/99/1.00
A1
Cosmarium ralfsii SVCK 300
Micrasterias semiradiata CAUP K606
85/-/1.00
A2
Micrasterias decemdentata ASW 07023
Micrasterias laticeps SVCK 430
95/99/1.00
A3
Micrasterias zeylanica SVCK 291
Micrasterias 'pusilla' NIES 784
83/-/1.00
A4
Micrasterias furcata CAUP K609
100/100/1.00
Micrasterias pinnatifida SVCK 411
Staurodesmus dickiei ASW 07056
99/99/1.00
A5
Micrasterias ceratofera SAG 21.97
A
100/100/1.00
Micrasterias crux-melitensis CAUP K602
Micrasterias radians var. bogoriensis SVCK 389
A6
Micrasterias crux-melitensis NIES 152
93/98/1.00
93/99/1.00
Micrasterias radians var. evoluta SVCK 518
82/99/1.00
B
Micrasterias anomala NIES 774
Micrasterias apiculata SVCK 247
82/100/0.97
-/-/0.92
89/-/1.00
Micrasterias brachyptera SVCK 65
C
86/92/1.00
Micrasterias fimbriata CAUP K608
Micrasterias rotata CAUP K604
98/99/1.00
Micrasterias doveri SVCK 369
Micrasterias thomasiana CAUP K605
D
100/100/1.00
Micrasterias denticulata SVCK 230
Micrasterias jenneri SVCK 298
100/100/1.00
E
Triploceras gracile SVCK 366
Triploceras gracile SAG 24.82
F
Micrasterias tropica SVCK 368
98/96/1.00
Micrasterias foliacea NIES 777
G
Micrasterias swainei SVCK 138
100/100/1.00
Micrasterias tetraptera SVCK 195
Micrasterias conferta SVCK 110
Micrasterias radiosa SVCK 303
Micrasterias novae-terrae SAG 158.80
92/85/1.00
Micrasterias papillifera CAUP K603
100/100/1.00
H
Micrasterias hardyi SVCK 249
Micrasterias americana SVCK 290
Micrasterias mahabuleshwarensis SVCK 324
Micrasterias muricata SAG 157.80
Staurastrum majusculum SVCK 330
Staurastrum lunatum SVCK 15
100/100/1.00
Staurastrum maamense SVCK 377
A
A1
A2
A3
A4
A5
A6
B
C
D
E
F
G
H

# konflikt genů




Cladonia

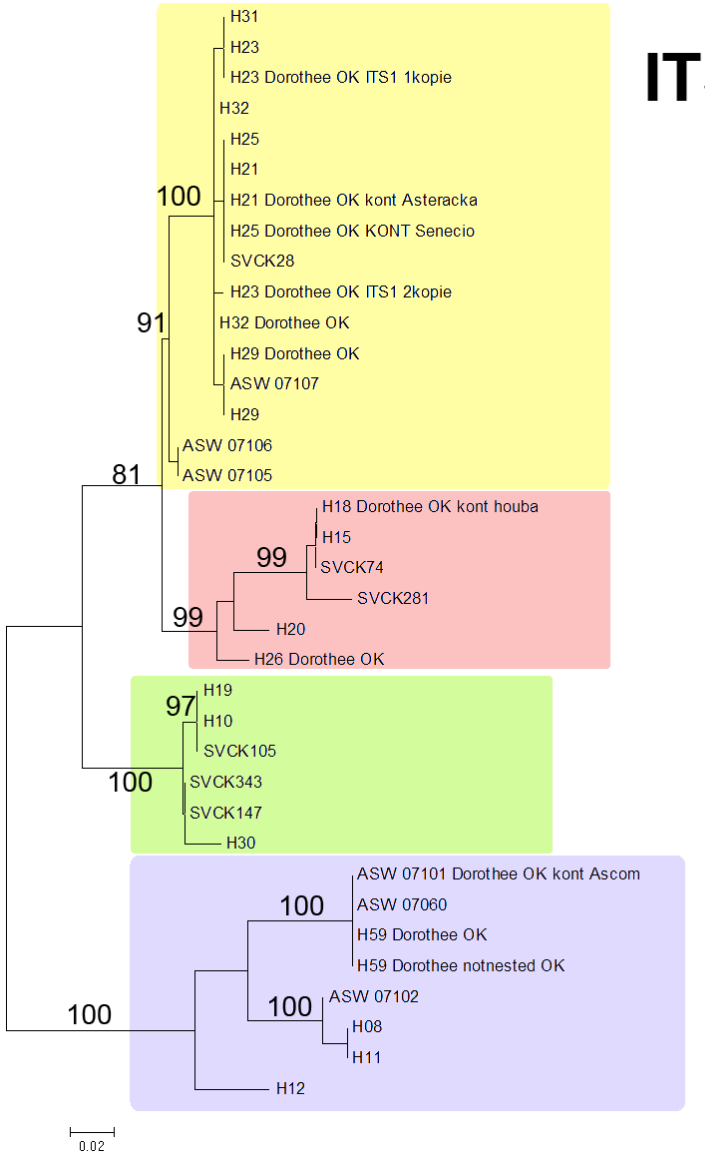
# konflikt genů

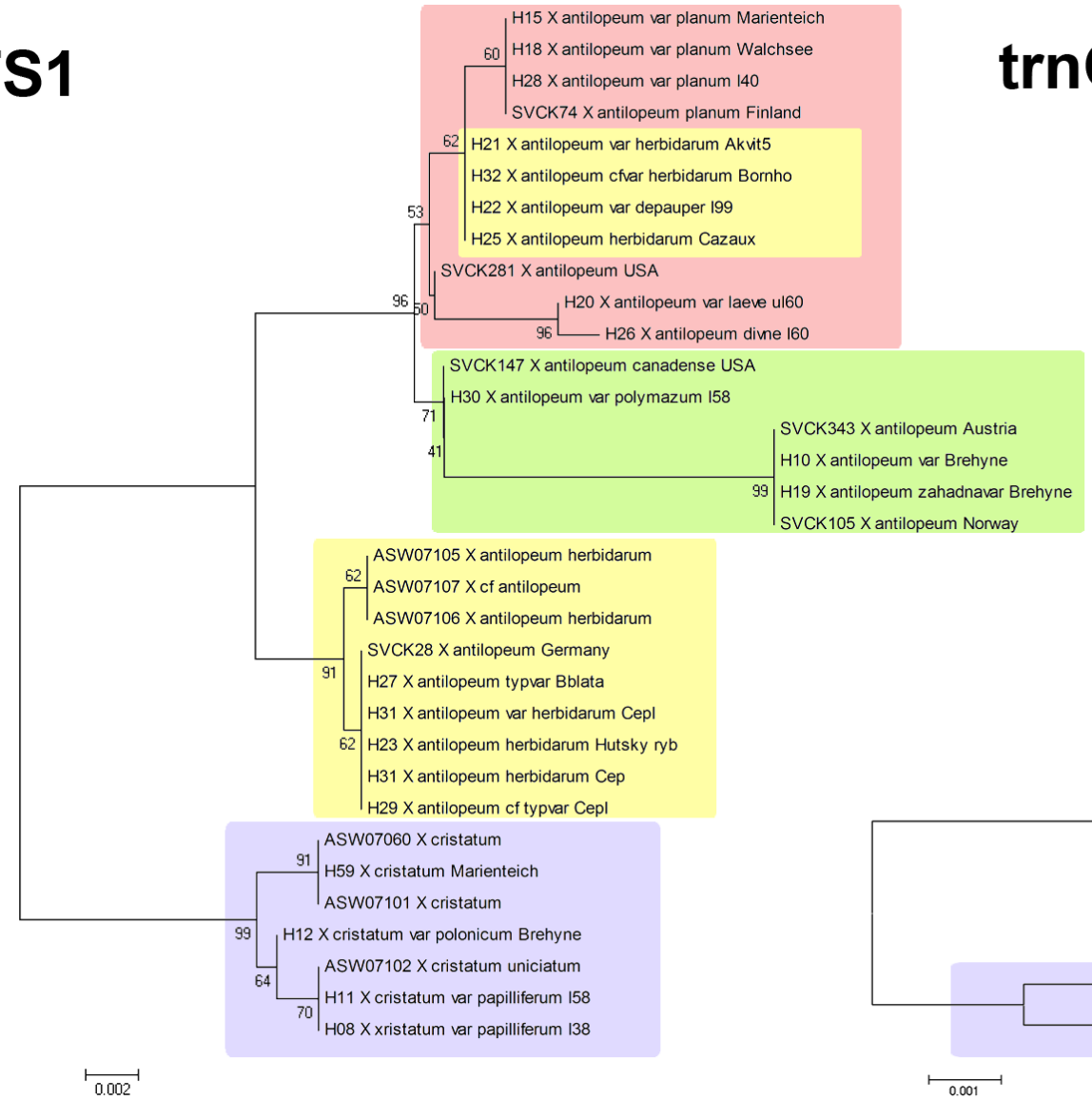



* CL36, CL85, CL113, CL115, CL117, CL136

Cladonia

# konflikt genů

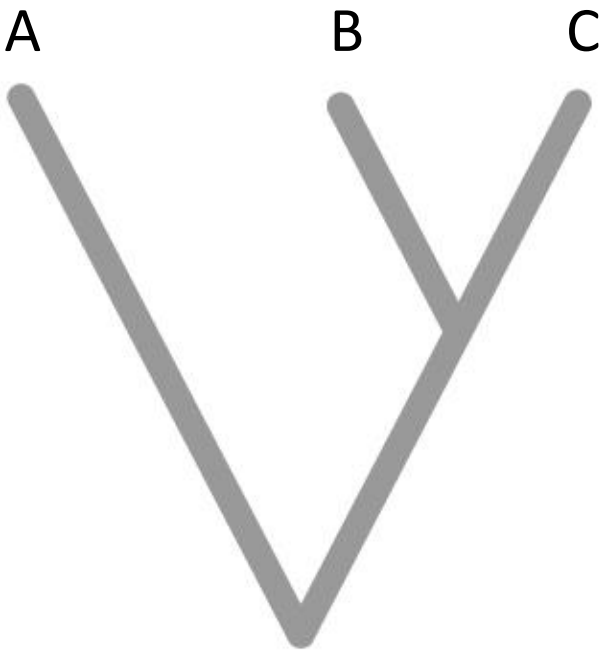



Xanthidium

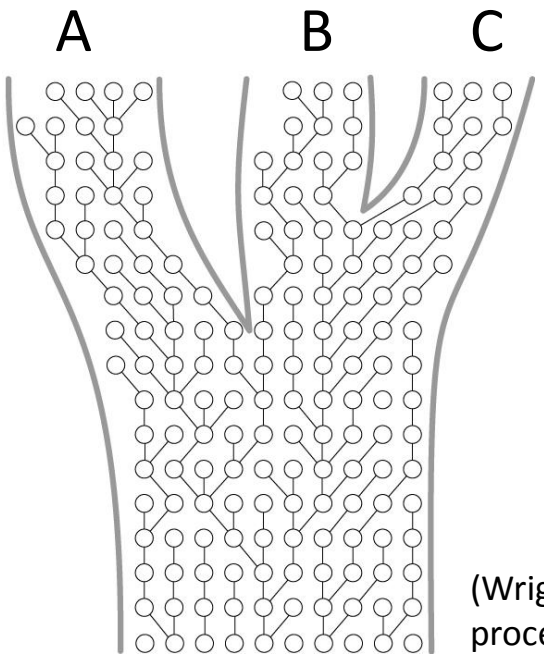
# Phylogeny at the level of populations and species

**Species are lineages**




Species phylogeny

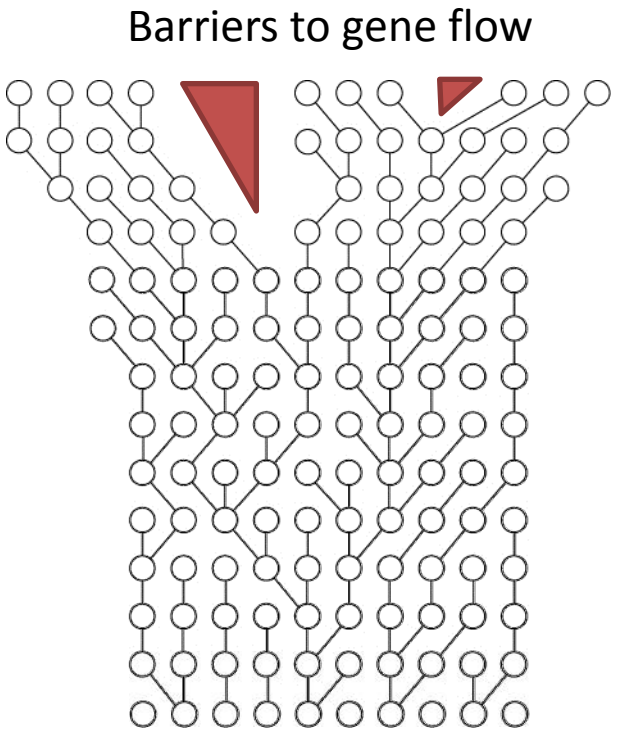
>>



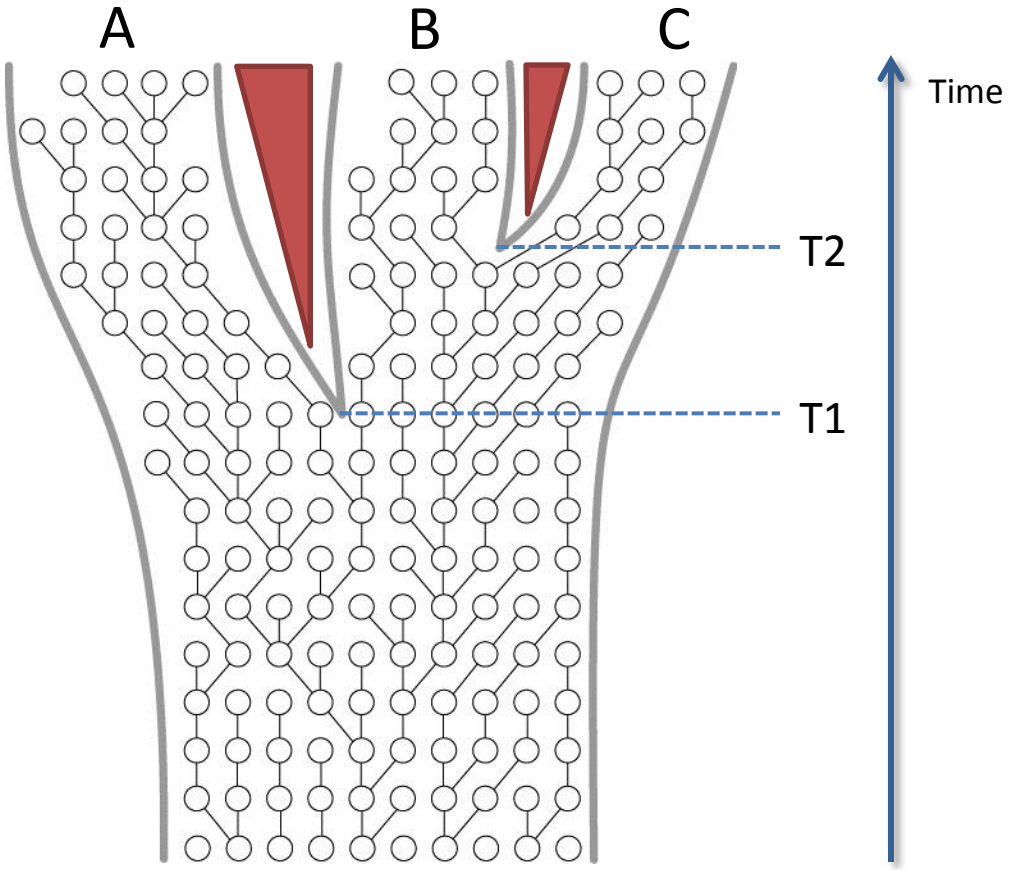

(Wright-Fisher process)

Population genetics:
**coalescence process**

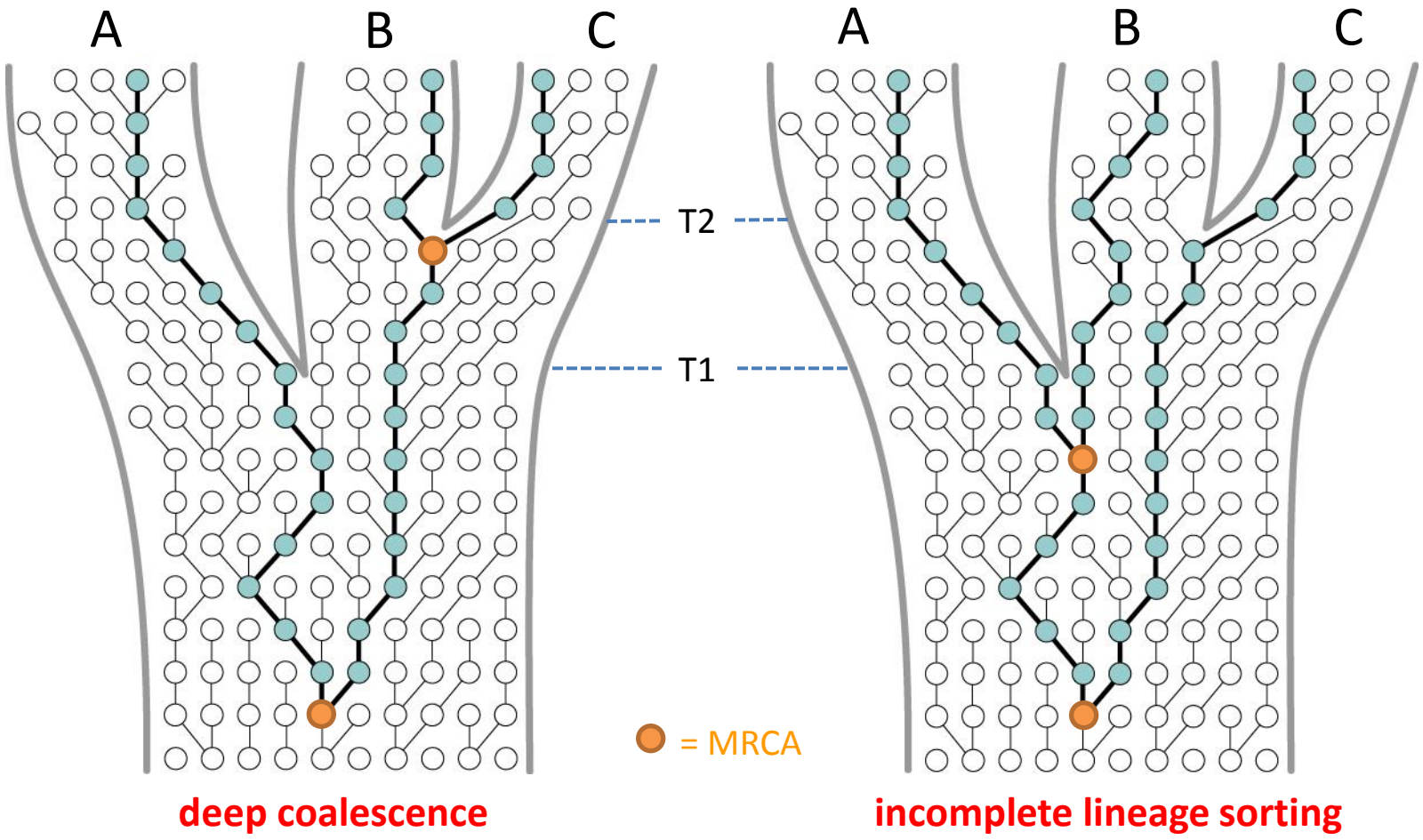
# Phylogeny at the level of populations and species



# Phylogeny at the level of populations and species



# Phylogeny at the level of populations and species



# AFLP

## Advantage

- high variability – many loci
- many independent loci (*multilocus method*)
- covering „whole“ genome
- statistical apparat for data analysis

## Drawbacks

- anonymous marker
- asymetry in probability of loss and gain of fragments – yes/no?
- dominant – impossible to distinguish homozygotes and heterozygotes
- evaluation subjectivity
- unknown rate of mutation accumulation (impossible to use molecular clock)
- problematic (impossible) addition of further samples

# microsatellites

## Advantage

- usually high variation – many alleles
- codominant – distinguish among homozygotes and heterozygotes, allelic frequencies
- models of allele evolution – „known“ relationships among alleles
- more objective evaluation
- statistical apparat for data analysis
- possible to add further samples

## Drawbacks

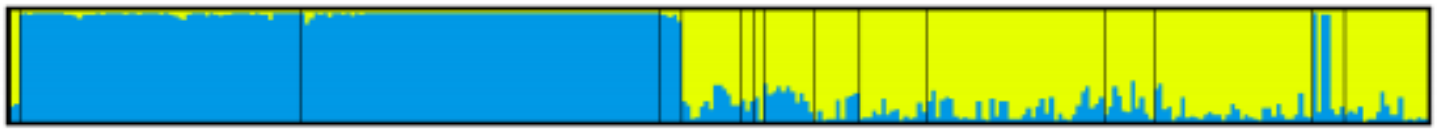
- species-specific markers
- simultaneous analysis of limited number of loci
- more limited representation of the „whole“ genome

# STRUCTURE results evaluation

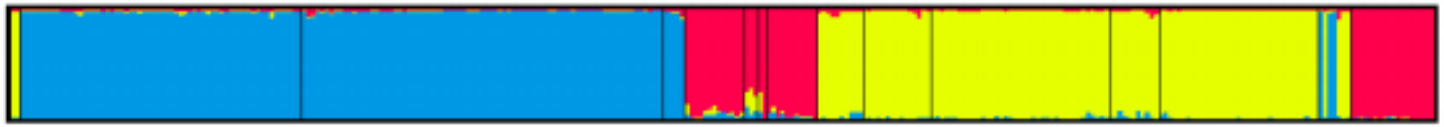
Distruct (Rosenberg 2004)

- graphical representation of sample assignment to individual clusters

K2

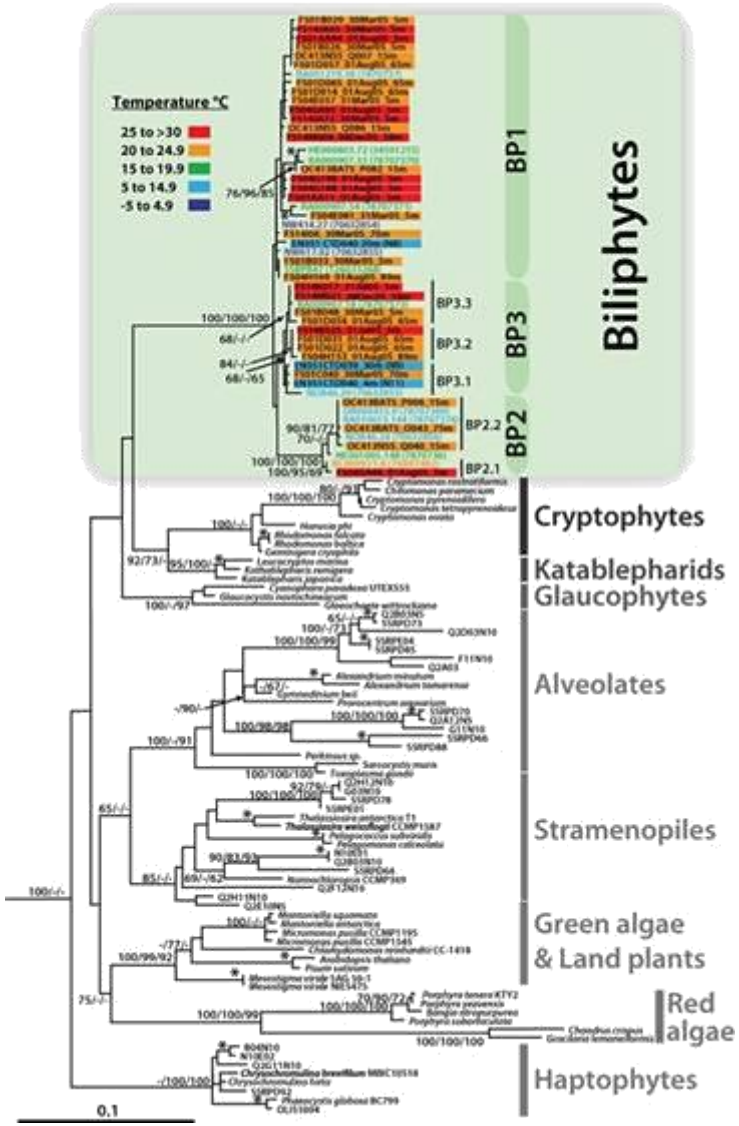


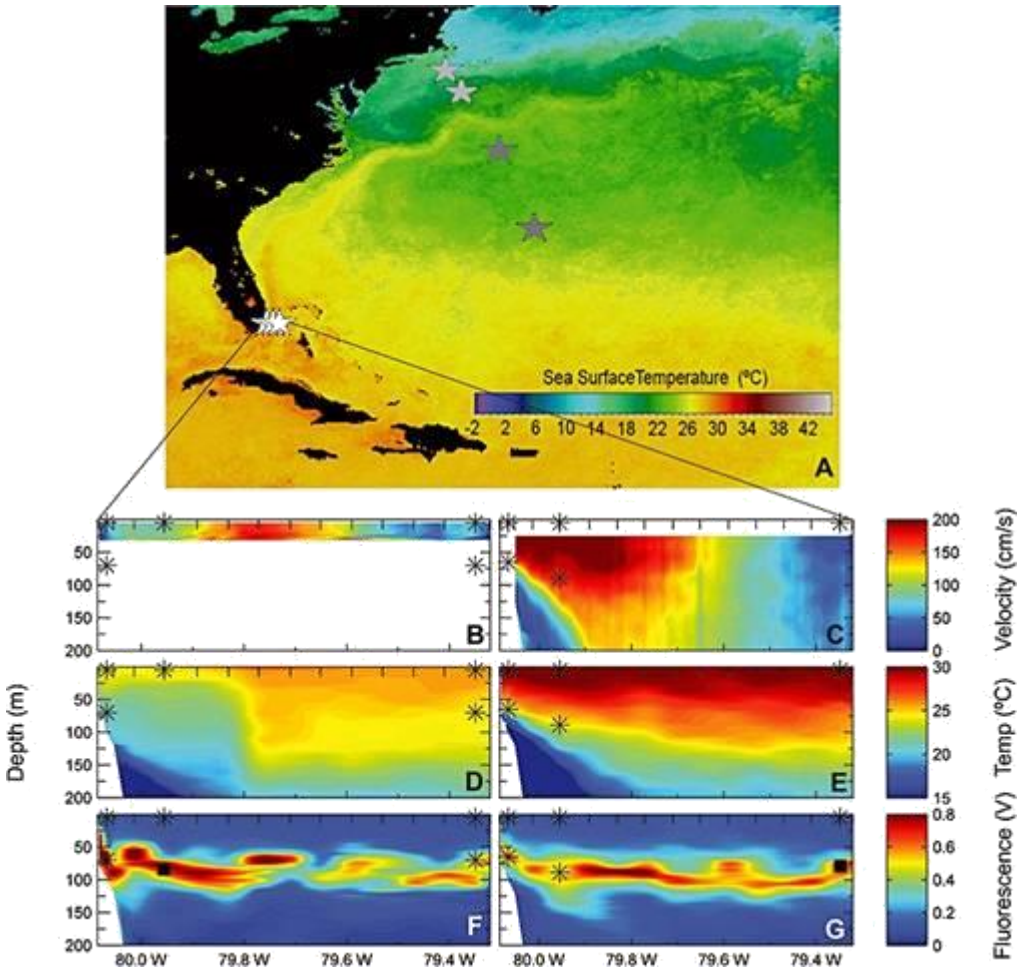
K3



# NGS

- Environmentální sekvenování – nyní více jak 300 000 SSU rDNA sekvencí v databázi GenBank



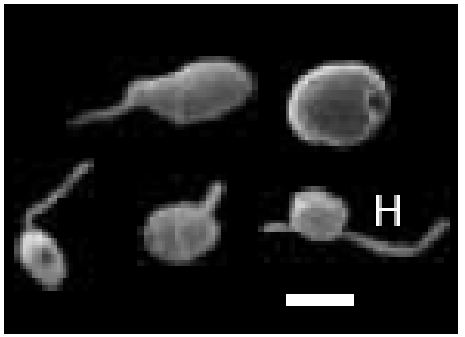

Cuvelier et al. - 2008

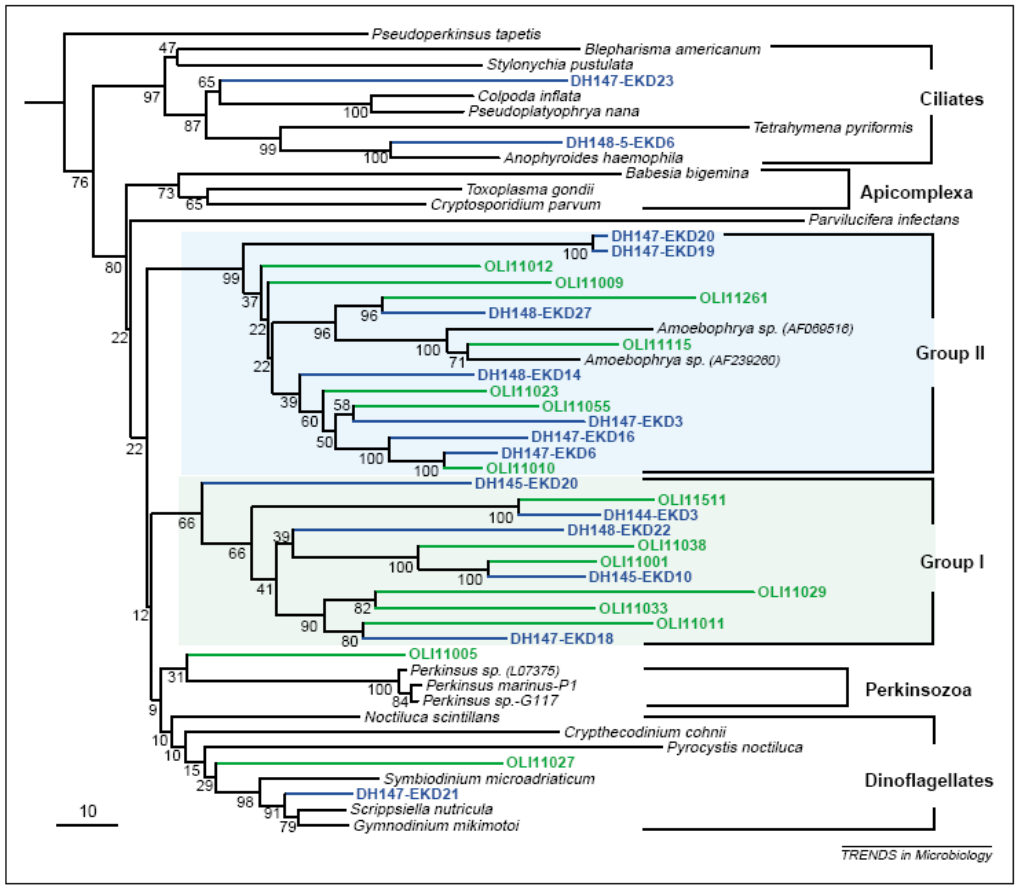
# NGS

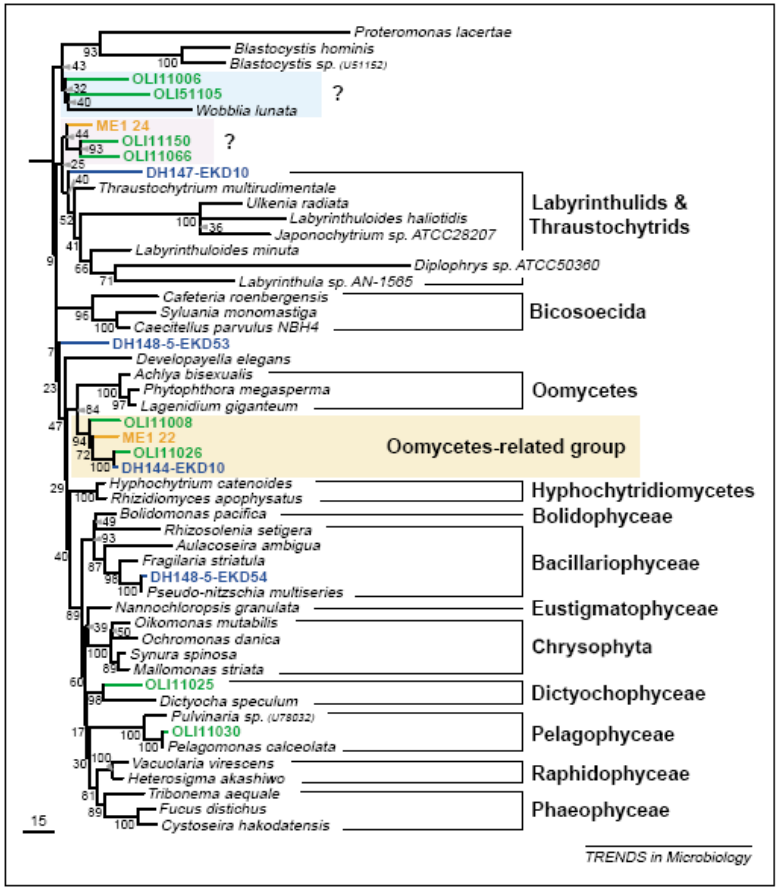
Review TRENDS in Microbiology Vol.10 No.1 January 2002 31

# The molecular ecology of microbial eukaryotes unveils a hidden world

David Moreira and Purificación López-García



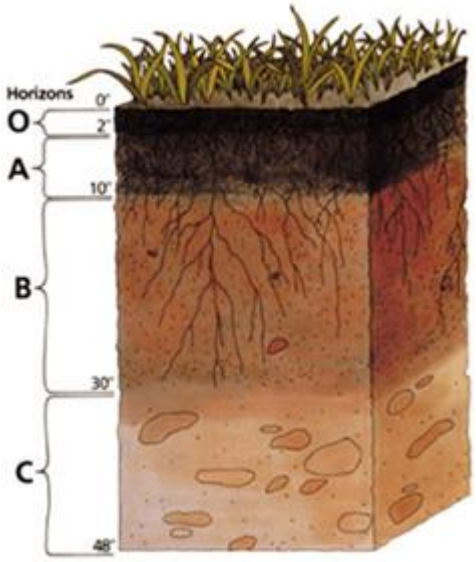



# NGS

Applied and Environmental Microbiology, Sept. 2005, p. 5544–5550 Vol. 71, No. 9
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.9.5544–5550.2005

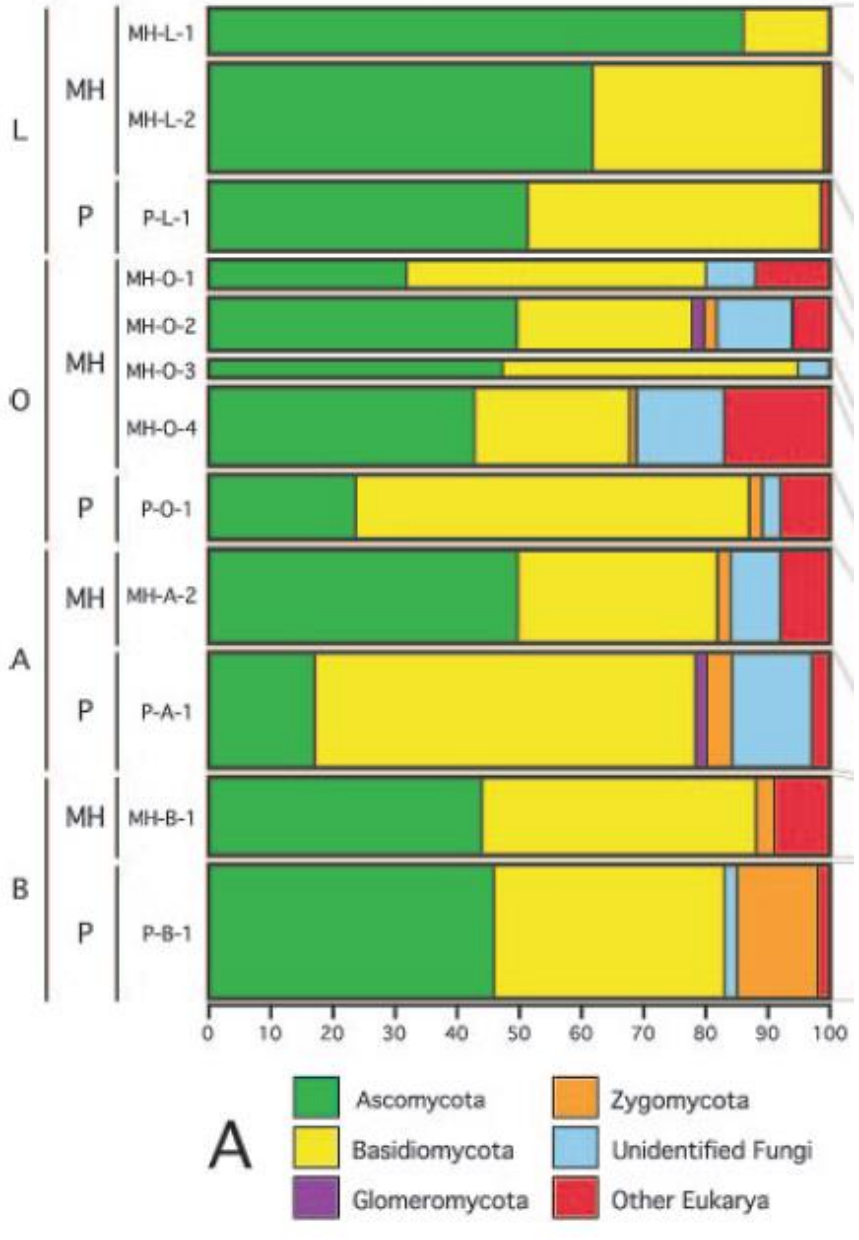

## Fungal Community Analysis by Large-Scale Sequencing of Environmental Samples†

Heath E. O'Brien,[1]* Jeri Lynn Parrent,[1] Jason A. Jackson,[1] Jean-Marc Moncalvo,[2] and Rytas Vilgalys[1]




L – litter
O – organic
A – A horizon
B – B horizon

P – pine MH – mixed hardwood



# NGS

## Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis

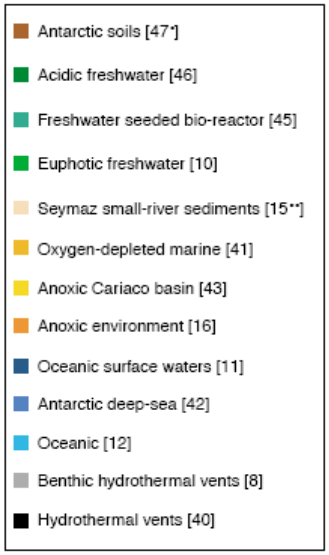
Thomas A Richards[1,2] and David Bass[2]

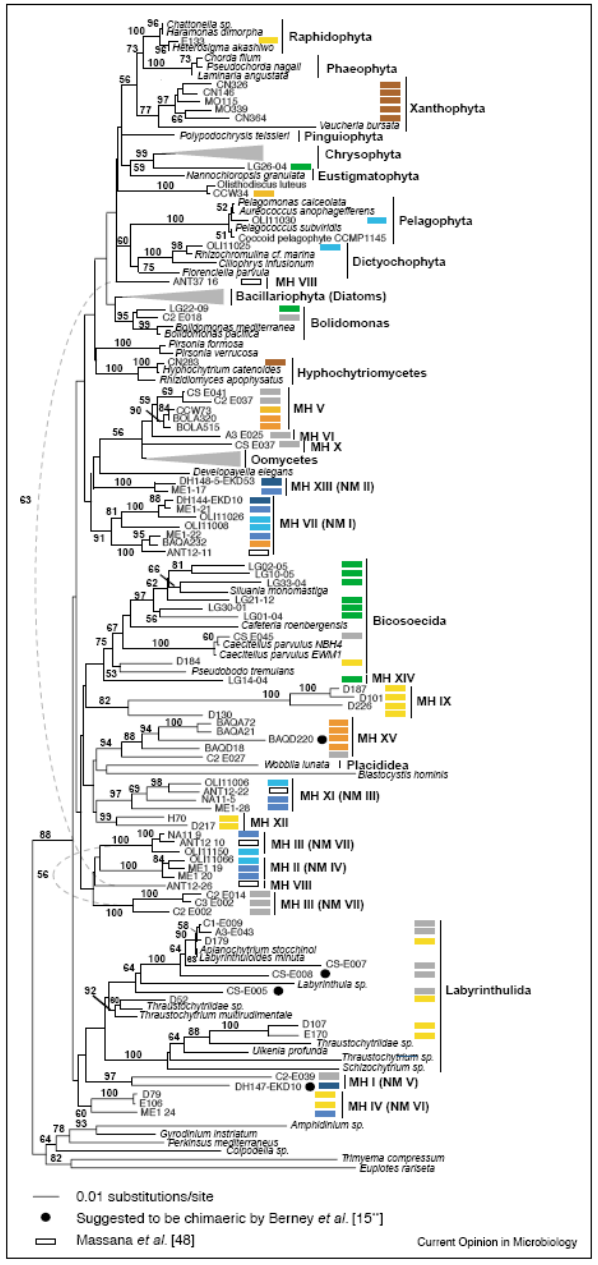
Current Opinion in Microbiology 2005, 8:240–252

**Table 1**

**Comparison of sampling strategies used in the 13 studies included in this meta-analysis.**

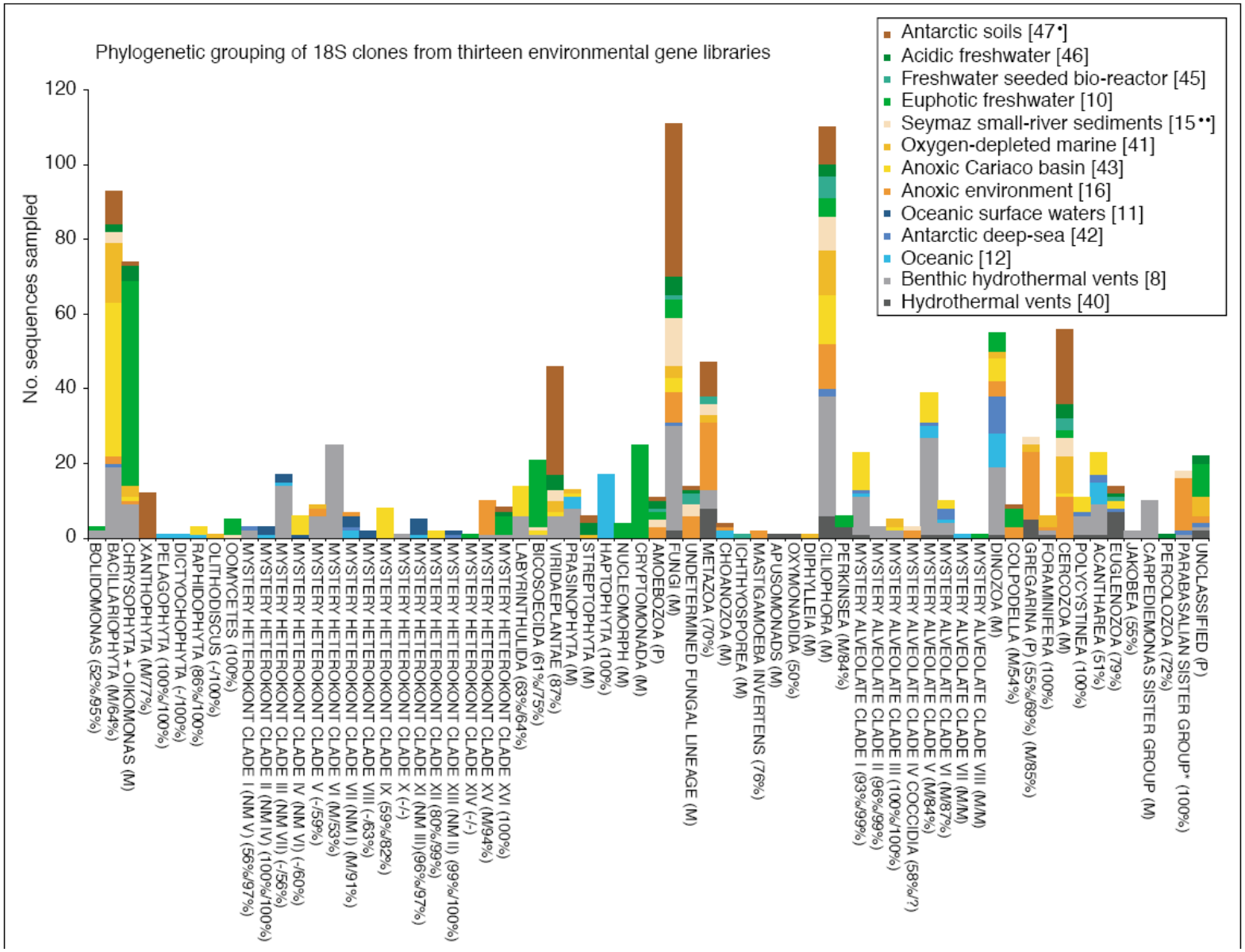
| Environment | Reference |
|---|---|
| Synthetic microbial bio-reactor fed by monocultures to create a detritus environment seeded with the microbes from the waters of lake Ketelmeer (Netherlands), a small turbid lake. | [45] |
| Marine Antarctic polar fronts –250, –500, –2000 and –3000*m (cold oligotrophic, highly oxygenated waters). | [42] |
| Equatorial Pacific Ocean at a depth of 75 m; oligotrophic. | [12] |
| Marine surface samples; Mediterranean, Antarctica and North Atlantic. | [11] |
| Everest mound region from the Guaymas hydrothermal vent. Sediment cores were sampled from sediment/seawater interface, and sediment. | [8] |
| Rio Tinto Spain, pH 2 freshwater river with high concentration of heavy metals (3–20 g iron/litre). | [46] |
| Anoxic sediments: two marine (1–3 cm deep) plus one freshwater (5 cm deep). Samples were taken at >1 cm below surface of black reducing sediment. | [16] |
| Suboxic waters and anoxic sediments from a well protected intertidal pool in the Great Sippewisset salt marsh, Cape Cod, MA, USA. The sediment cores were sampled at the sediment/water interface and at a depth of ~10 cm. | [41] |
| Cariaco basin in the Caribbean Sea on the northern continental margin of Venezuela: a large permanently anoxic basin. Vertical stratification of microbial communities and a clear transition from oxic to anoxic are seen. Three depths sampled at 270-340-900 m corresponding to oxic, oxic/anoxic interface and the anoxic component of the water column. | [43] |
| Rainbow hydrothermal sediment (depth 2264 m). Vent-fluid/seawater mixtures from Lucky Strike (depth 1695 m) and Rainbow Chimneys. Micro-colonisers exposed close to vent-fluid emissions for 15 days (Lucky Strike site). | [40] |
| Soil samples were collected from six sites on the Antarctic continent; sites corresponded to a latitudinal and environmental gradient between approximately 60 and 87°S. | [47•] |
| Sediment from a small freshwater river (Seymaz, Geneva, Switzerland). | [15••] |
| Oligotrophic freshwater lake in upstate New York, USA. Four sampling sites up to a depth of 20 m (euphotlc portion of the water column). | [10] |



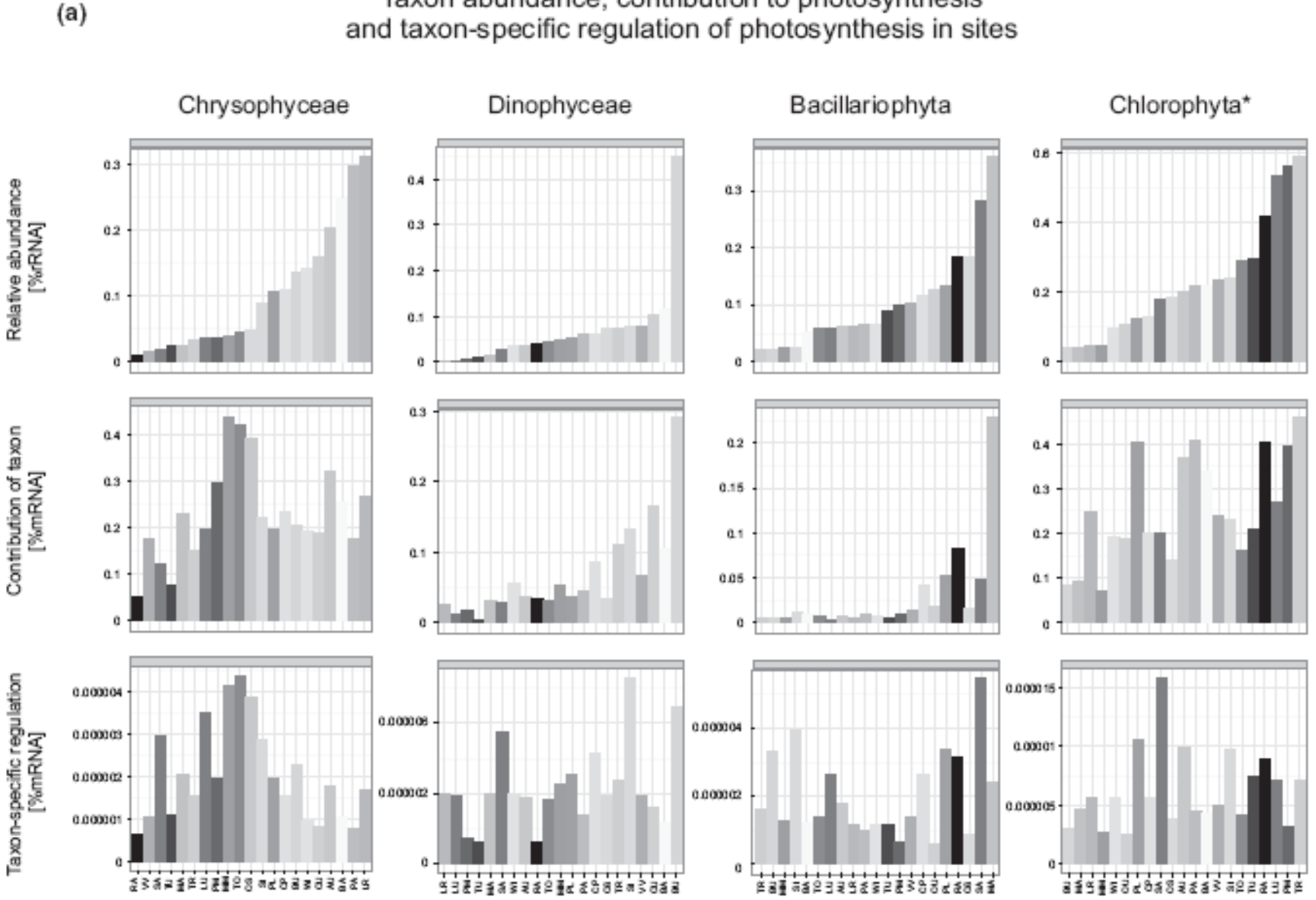



# NGS

- Ekologická struktura linií




8 freshwater
31 marine
12 anoxic
2 deep-sea


(a)
Taxon abundance, contribution to photosynthesis
and taxon-specific regulation of photosynthesis in sites
Chrysophyceae
Dinophyceae
Bacillariophyta
Chlorophyta*
Relative abundance
[%rRNA]
Contribution of taxon
[%mRNA]
Taxon-specific regulation
[%mRNA]

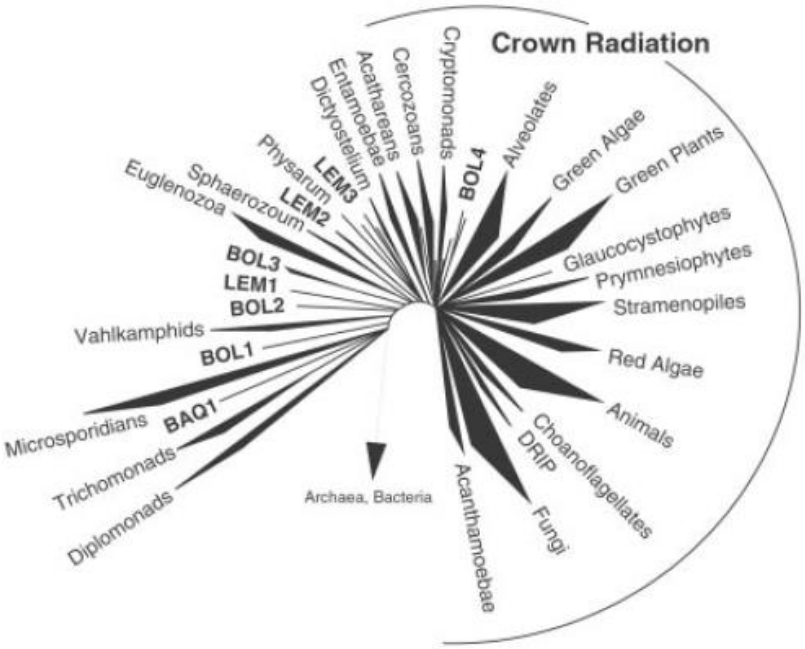
# Skryté pasti environmentálního sekvenování

## Novel kingdom-level eukaryotic diversity in anoxic environments

8324–8329 | PNAS | June 11, 2002 | vol. 99 | no. 12

Scott C. Dawson† and Norman R. Pace‡§

†Department of Molecular and Cell Biology, 345 LSA Building, University of California, Berkeley, CA 94720; and ‡Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309




Objev 8 nových říší

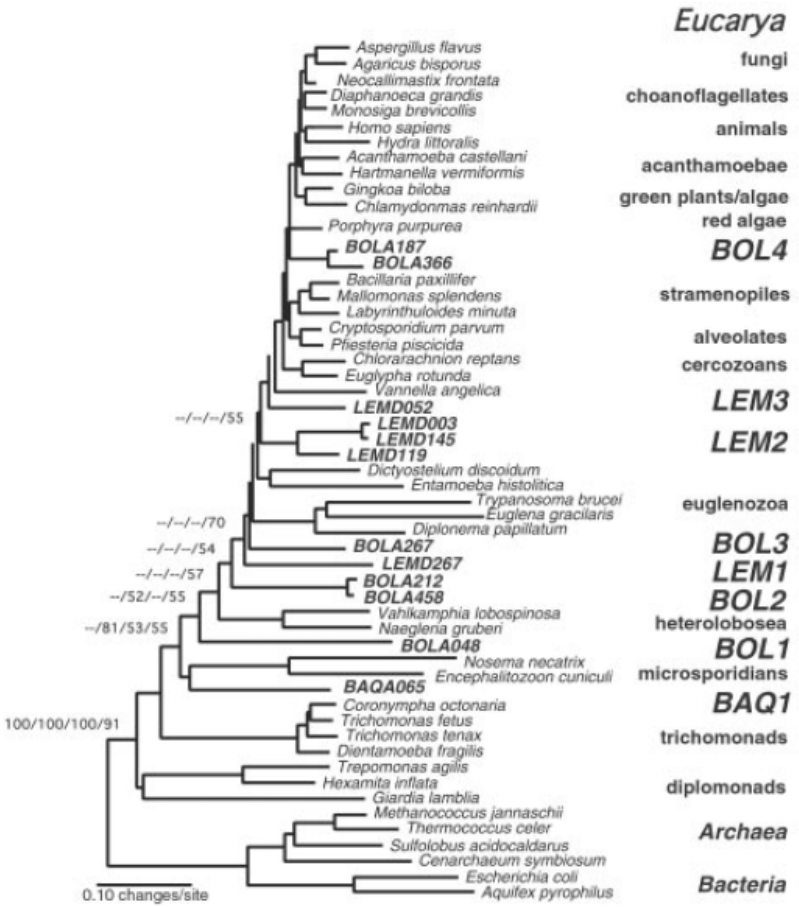



Fig. 4. Molecular phylogeny of novel kingdom-level lineages in Eucarya.

# Skryté pasti environmentálního sekvenování

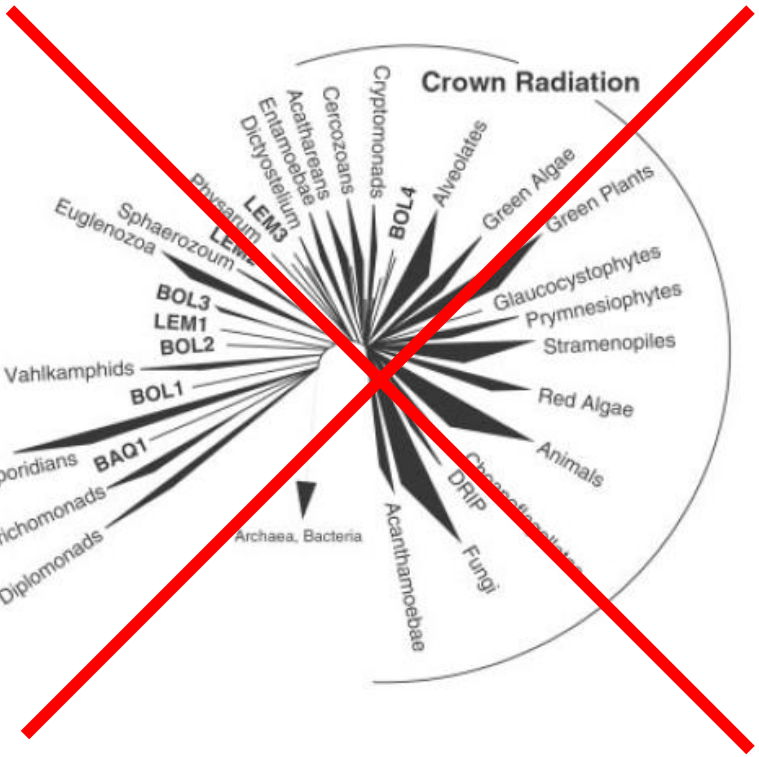
- Nedostatečný taxon sampling

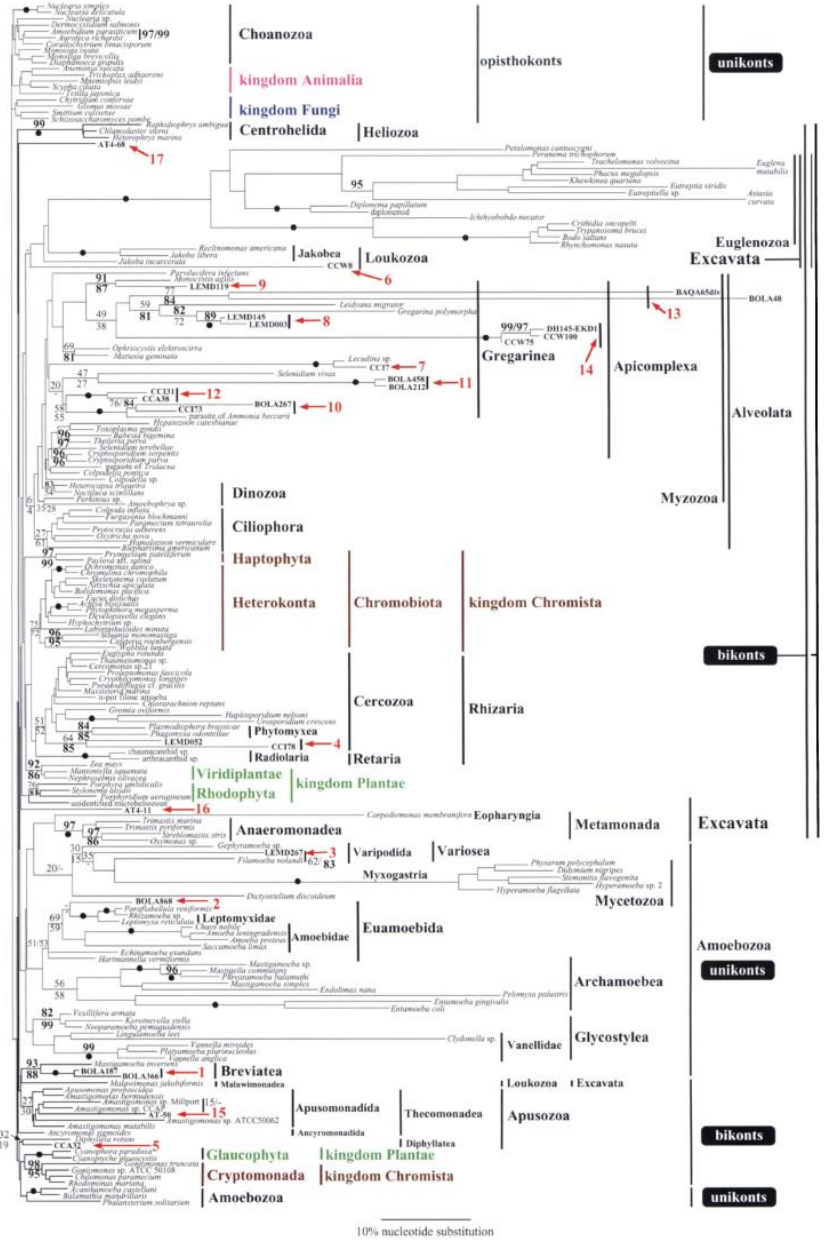
THE ROYAL SOCIETY

*Received* 13 November 2003
*Accepted* 3 February 2004
*Published online* 17 May 2004

**Only six kingdoms of life**

**Thomas Cavalier-Smith**



# Skryté pasti environmentálního sekvenování

- Nedostatečný taxon sampling

**BMC Biology**

BioMed Central

Research article

Open Access

**How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys**

Cédric Berney*, José Fahrni and Jan Pawlowski

Address: Department of Zoology and Animal Biology, University of Geneva, CH – 1211 Geneva 4, Switzerland

Email: Cédric Berney* - cedric.berney@zoo.unige.ch; José Fahrni - jose.fahrni@zoo.unige.ch; Jan Pawlowski - jan.pawlowski@zoo.unige.ch

* Corresponding author

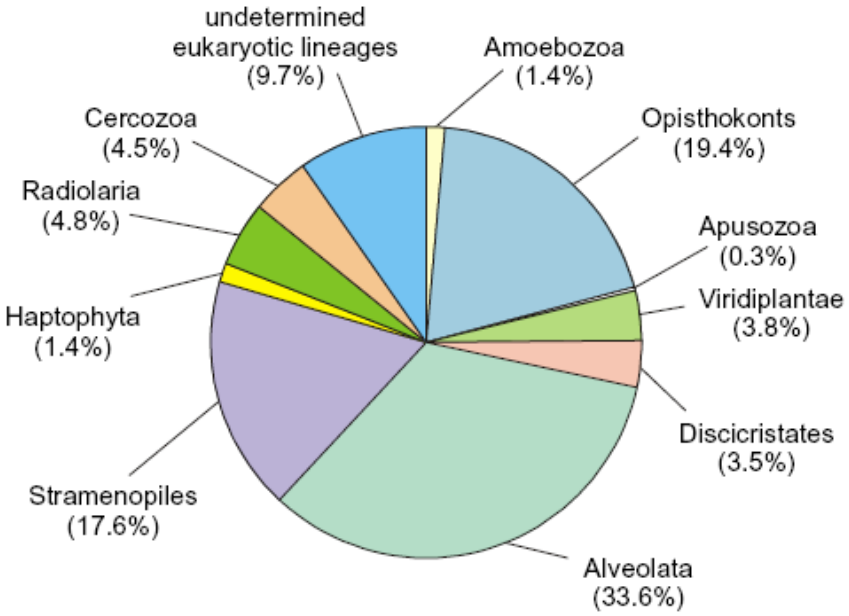
Published: 04 June 2004

*BMC Biology* 2004, 2:13

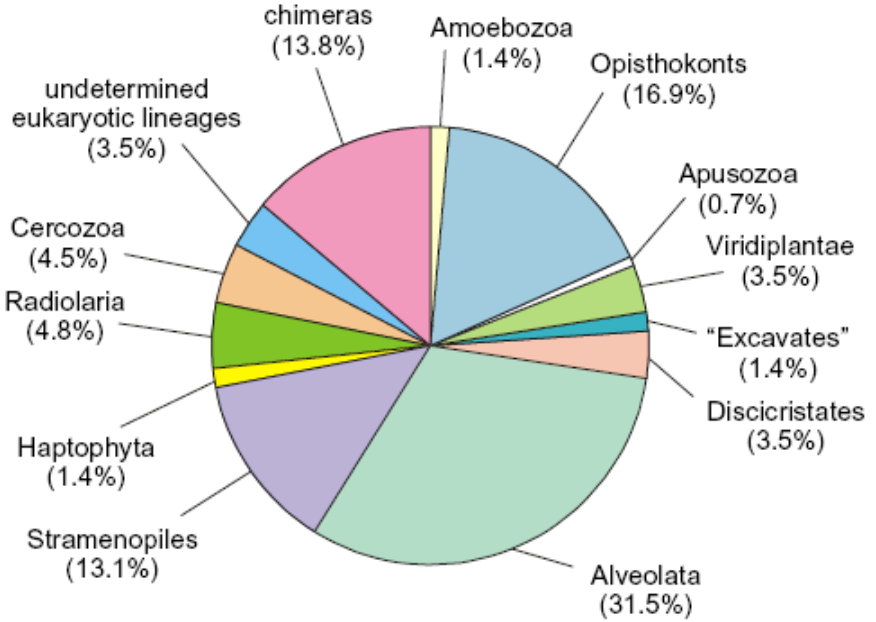
Received: 05 February 2004

Accepted: 04 June 2004

publikované určení sekvencí

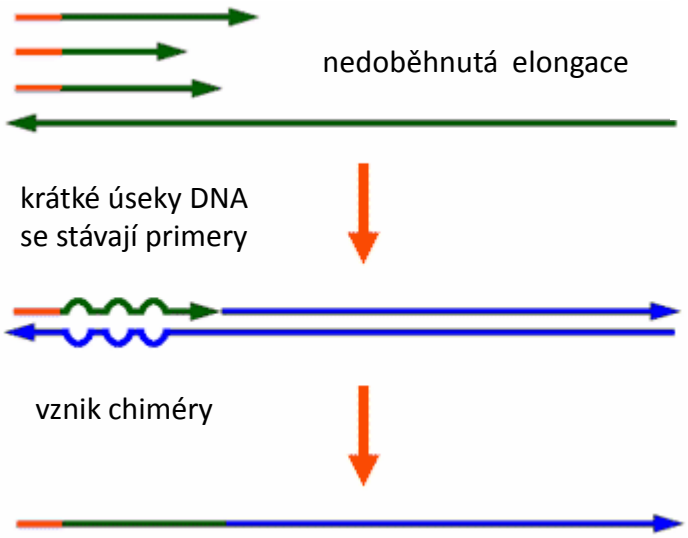


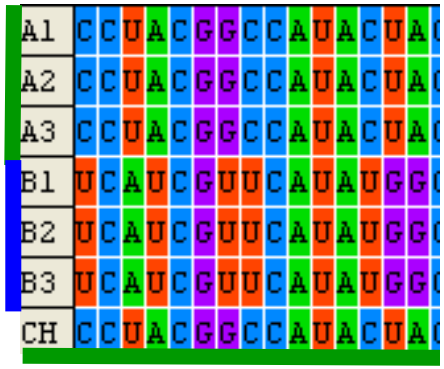

reanalyzované určení sekvencí



# Skryté pasti environmentálního sekvenování

- Chimerické sekvence




APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2003, p. 2657–2663
0099-2240/03/$08.00+0 DOI: 10.1128/AEM.69.5.2657–2663.2003

Vol. 69, No. 5

## Novel Eukaryotic Lineages Inferred from Small-Subunit rRNA Analyses of Oxygen-Depleted Marine Environments†

Thorsten Stoeck* and Slava Epstein

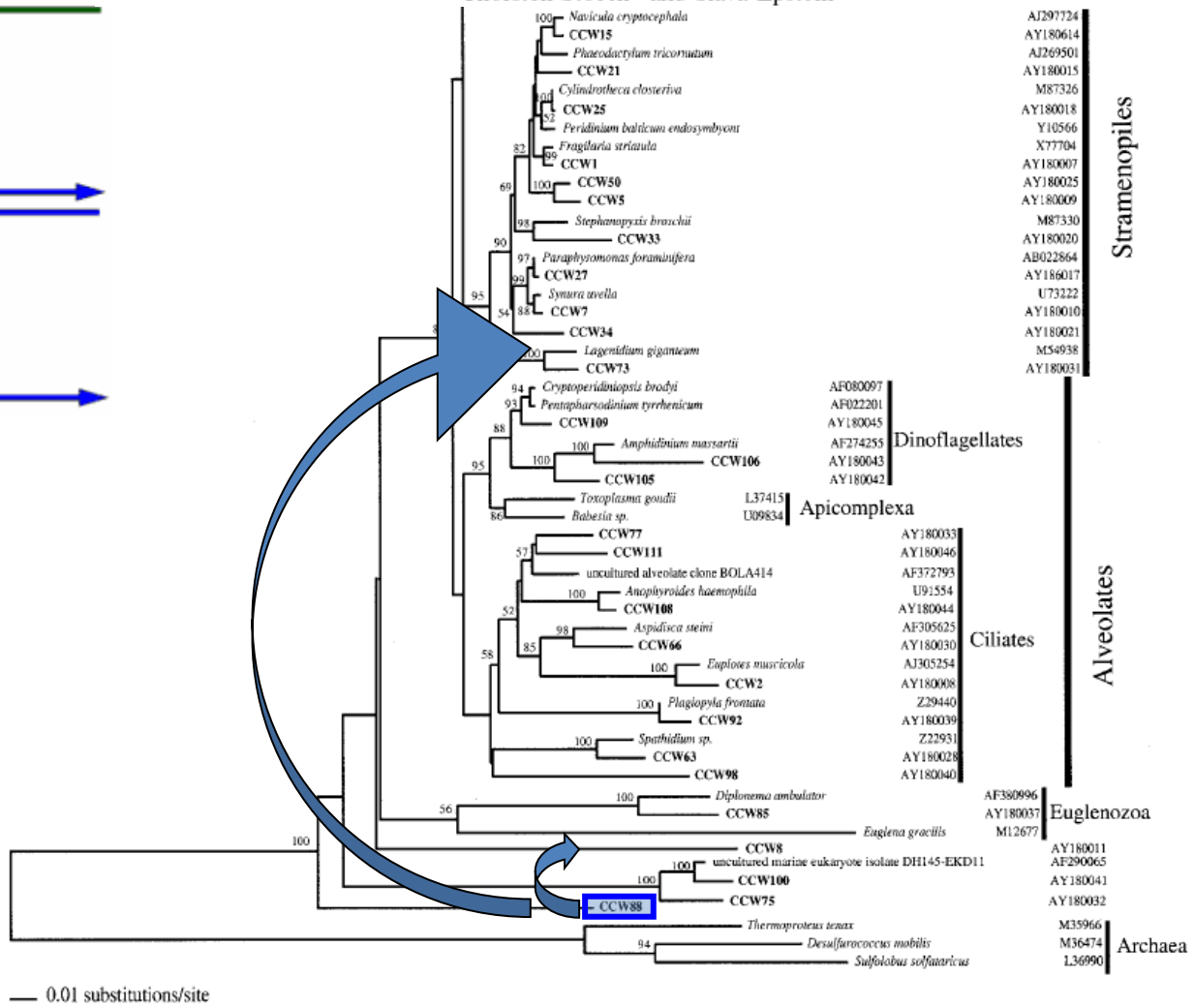
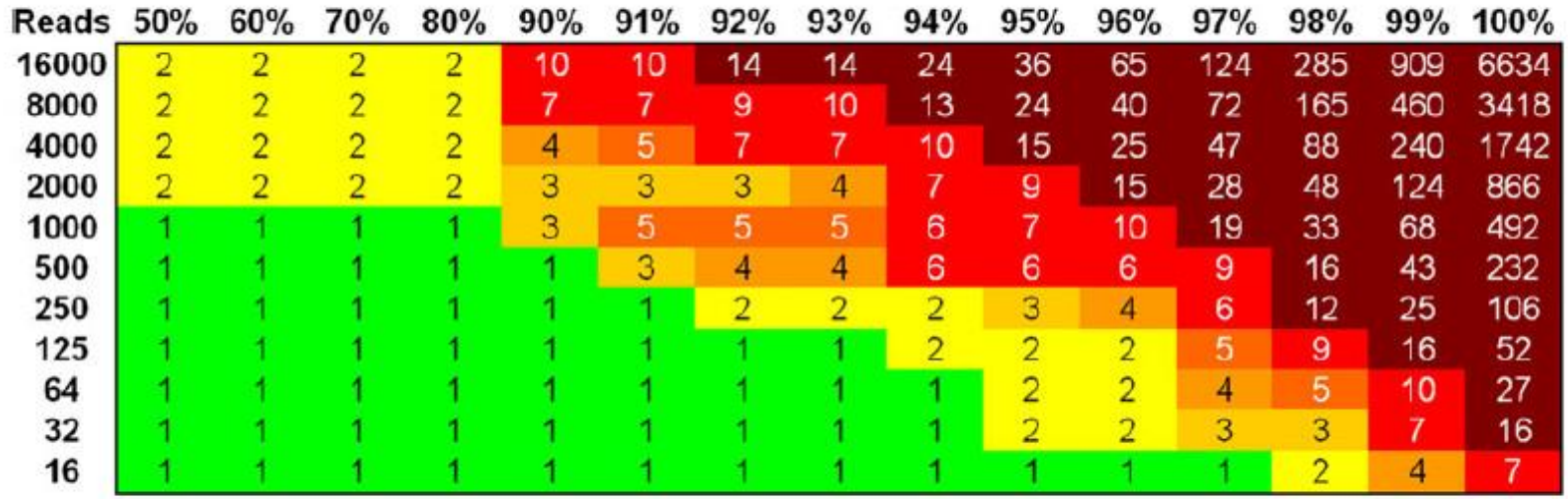


**Table 5** Summary of 31 heteropolymer insertion sites present in ten or more reads, sorted in descending order by number of reads in which insertion was present

| Pos. | Reg. | Base call | No. of reads | Proport. of total reads (%) | R1 149 | R1 150 | R1 154 | R2A 149 | R2A 150 | R2A 154 | R2B 149 | R2B 150 | R2B 154 | Alternat. base call | No. of reads | No. of reads with gaps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 47 | ITS1 | A | 931 | 5.58 | 159 | 471 | 273 | 1 | 10 | 1 | 2 | 9 | 5 | C | 2 | 15732 |
| 331 | ITS1 | A | 574 | 3.44 | 67 | 199 | 207 | 19 | 13 | 18 | 17 | 15 | 19 | C | 1 | 16090 |
| 277 | ITS1 | C | 451 | 2.70 | 70 | 223 | 110 | 5 | 6 | 11 | 5 | 11 | 10 | – | – | 16214 |
| 151 | ITS1 | G | 303 | 1.82 | 19 | 231 | 38 | 2 | 2 | 4 | 1 | 1 | 5 | – | – | 16362 |
| 194 | ITS1 | T | 203 | 1.22 | 13 | 162 | 24 | 0 | 1 | 2 | 0 | 0 | 1 | – | – | 16462 |
| 420 | ITS1 | A | 108 | 0.65 | 4 | 19 | 47 | 7 | 7 | 5 | 5 | 11 | 3 | – | – | 16557 |
| 134 | ITS1 | A | 106 | 0.64 | 9 | 65 | 24 | 0 | 2 | 1 | 1 | 3 | 1 | C | 1 | 16558 |
| 453 | ITS1 | C | 106 | 0.64 | 31 | 35 | 10 | 3 | 8 | 6 | 1 | 6 | 6 | – | – | 16559 |
| 515 | 5.8S | A | 63 | 0.38 | 0 | 8 | 1 | 12 | 10 | 6 | 10 | 7 | 9 | – | – | 16602 |
| 264 | ITS1 | C | 53 | 0.32 | 3 | 4 | 3 | 4 | 5 | 11 | 6 | 6 | 11 | A/G | 3 | 16609 |
| 283 | ITS1 | C | 49 | 0.29 | 11 | 4 | 28 | 0 | 0 | 1 | 1 | 1 | 3 | – | – | 16616 |
| 329 | ITS1 | G | 39 | 0.23 | 13 | 13 | 12 | 0 | 0 | 0 | 1 | 0 | 0 | – | – | 16626 |
| 418 | ITS1 | C | 35 | 0.21 | 1 | 5 | 21 | 2 | 1 | 0 | 3 | 2 | 0 | – | – | 16630 |
| 111 | ITS1 | G | 33 | 0.20 | 2 | 21 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | – | – | 16632 |
| 209 | ITS1 | A | 30 | 0.18 | 1 | 24 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | – | – | 16635 |
| 310 | ITS1 | T | 30 | 0.18 | 10 | 10 | 9 | 0 | 0 | 1 | 0 | 0 | 0 | A | 1 | 16634 |

Lucking, 2014, pyrosequencing



| Reads | 50% | 60% | 70% | 80% | 90% | 91% | 92% | 93% | 94% | 95% | 96% | 97% | 98% | 99% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16000 | 2 | 2 | 2 | 2 | 10 | 10 | 14 | 14 | 24 | 36 | 65 | 124 | 285 | 909 | 6634 |
| 8000 | 2 | 2 | 2 | 2 | 7 | 7 | 9 | 10 | 13 | 24 | 40 | 72 | 165 | 460 | 3418 |
| 4000 | 2 | 2 | 2 | 2 | 4 | 5 | 7 | 7 | 10 | 15 | 25 | 47 | 88 | 240 | 1742 |
| 2000 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 7 | 9 | 15 | 28 | 48 | 124 | 866 |
| 1000 | 1 | 1 | 1 | 1 | 3 | 5 | 5 | 5 | 6 | 7 | 10 | 19 | 33 | 68 | 492 |
| 500 | 1 | 1 | 1 | 1 | 1 | 3 | 4 | 4 | 6 | 6 | 6 | 9 | 16 | 43 | 232 |
| 250 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 4 | 6 | 12 | 25 | 106 |
| 125 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 5 | 9 | 16 | 52 |
| 64 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 4 | 5 | 10 | 27 |
| 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 7 | 16 |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 4 | 7 |

**Fig. 2** Read frequency (randomly selected out of total of 16,665 reads, with ten repeats per frequency level) and resulting estimate of biological diversity. *Colors* indicate level of correctness compare to real biological diversity (a single species) (Color figure online)