Université de Montréal

Assessing the effects of compositional heterogeneity on phylogenomic analyses

Denis Baurain[1,2,3], Robert G. Beiko[2], and Mark A. Ragan[2]

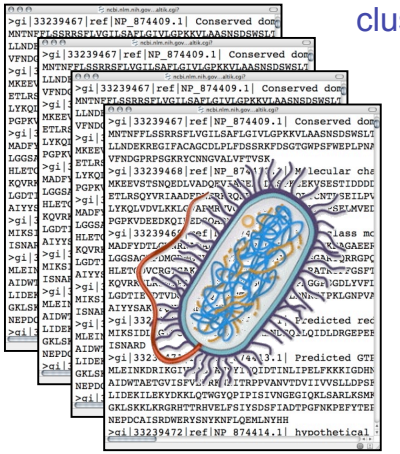
[1] Université de Liège / FNRS, Belgium
[2] University of Queensland / ARC Centre in Bioinformatics, Australia
[3] Université de Montréal

November 4th, 2005

THE ROBERT-CEDERGREN BIOINFORMATICS COLLOQUIUM

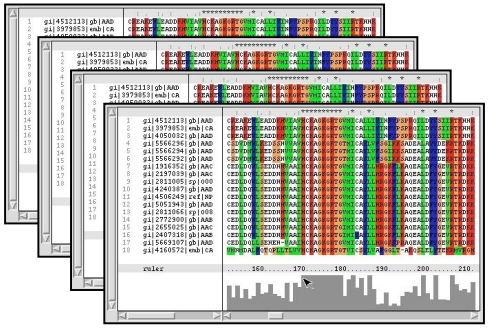
Background

Beiko *et al.* (2005) PNAS **102**:14332-14337

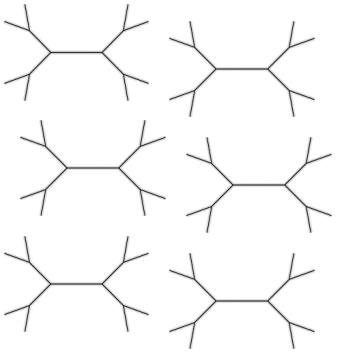


144 bacterial genomes
422,971 proteins

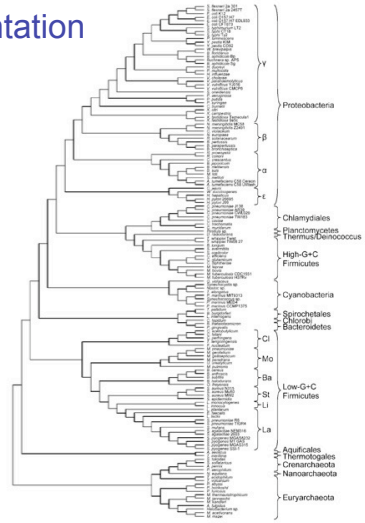
clustering & alignment → Bayesian inference → matrix representation with parsimony



22,432 alignments
220,240 proteins (52.1%)

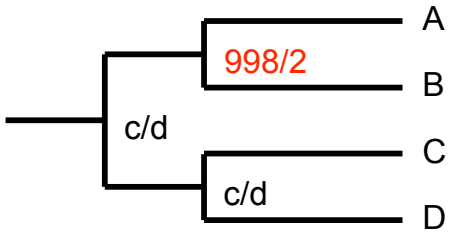


22,432 unrooted trees
95,194 bipartitions (PP ≥ 0.95)

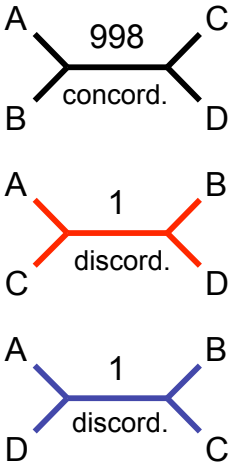


1 supertree

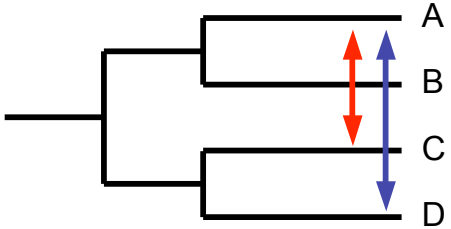
detection of LGT events



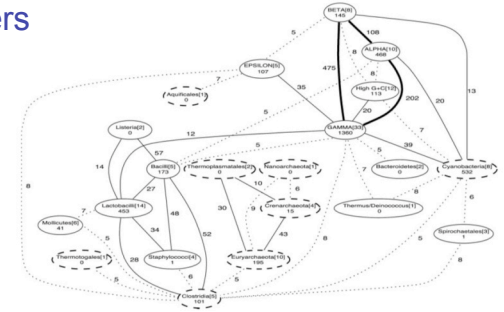
supertree topology



identification of exchange partners



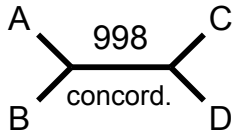
2 LGT events



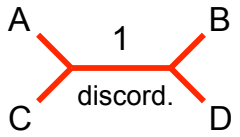
inheritance pattern largely vertical with ‘highways’ of gene sharing

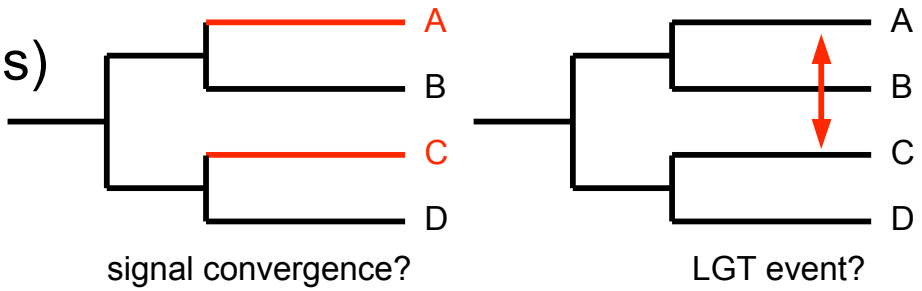
LGT or methodological issues?

- tests for systematic biases
  - clustering strategy (e.g. cluster size)
  - alignment quality (e.g. sequence length variation)
  - phylogenetic inference (e.g. alignment size)



- conflicting signals in the data
  - history
  - composition (e.g. GC bias)
  - rate (e.g. LBA)
  - other signals



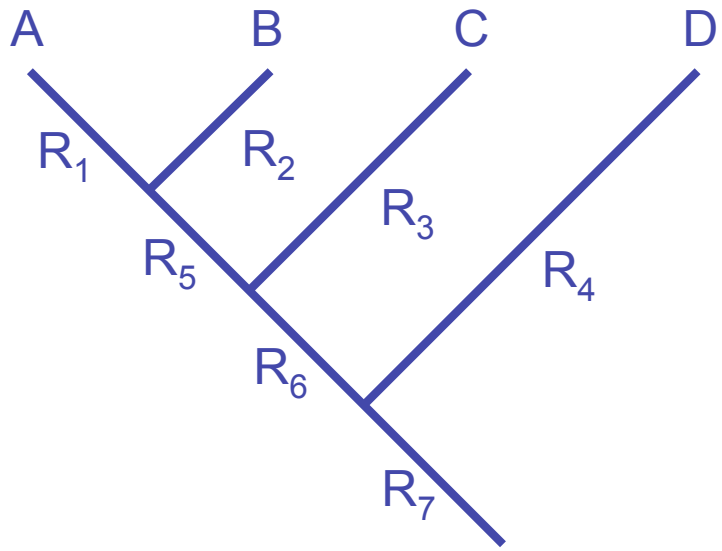


  - « *Any signal having experienced convergence in nonsister lineages will affect recovery of the historical signal.* »

Beiko *et al.* (2005) PNAS **102**:14332-14337; Ho and Jermiin 2004 Syst Biol **53**:623-637

Phylogenetic assumptions

- sites evolve independently and identically using a Markov process given by *R* (e.g. GTR model)
- for practical reasons, we assume that the sites have evolved under reversible, homogeneous and stationary conditions
- **note** – new models allow to safely violate some of these assumptions



$$R_k = \begin{bmatrix} -\sum_{j \neq A} \pi_j \alpha_{Aj} & \pi_C \alpha_{AC} & \pi_G \alpha_{AG} & \pi_T \alpha_{AT} \\ \pi_A \alpha_{AC} & -\sum_{j \neq C} \pi_j \alpha_{Cj} & \pi_G \alpha_{CG} & \pi_T \alpha_{CT} \\ \pi_A \alpha_{AG} & \pi_C \alpha_{CG} & -\sum_{j \neq G} \pi_j \alpha_{Gj} & \pi_T \alpha_{GT} \\ \pi_A \alpha_{AT} & \pi_C \alpha_{CT} & \pi_G \alpha_{GT} & -\sum_{j \neq T} \pi_j \alpha_{Tj} \end{bmatrix}$$

$\pi_A, \ldots, \pi_T$ – nucleotide frequencies

α_{kij} – conditional rates of change

$\alpha_{kij} = \alpha_{kji}$ – reversibility

$R_1 = R_2 = \ldots = R_6 = R_7$ – homogeneity

$\pi_{kj} = f_{0j}$ – stationarity

adapted from Lars S. Jermiin, University of Sydney; Lio and Goldman (1998) Genome Res **8**:1233-1244

Statistical tests for stationarity

S_1 ATGGTACAATGCGGCATGTACTCGCGATATCGACGATACG
S_2 ATCGAACGATGTGGCGTACACTCACGTTACCGACACGACG

	A	C	G	T	Σ
A	7	0	3	1	11
C	1	8	1	2	12
G	2	0	7	1	10
T	1	1	0	5	7
Σ	11	9	11	9	**40**

two-way contingency table

- matched-pair tests of homogeneity
  - Bowker's (1948) test for symmetry

$$\chi^2_{bowker} = \sum_{i<j} \frac{(x_{ij} - x_{ji})^2}{x_{ij} + x_{ji}}$$

  - Stuart's (1955) test for marginal homogeneity

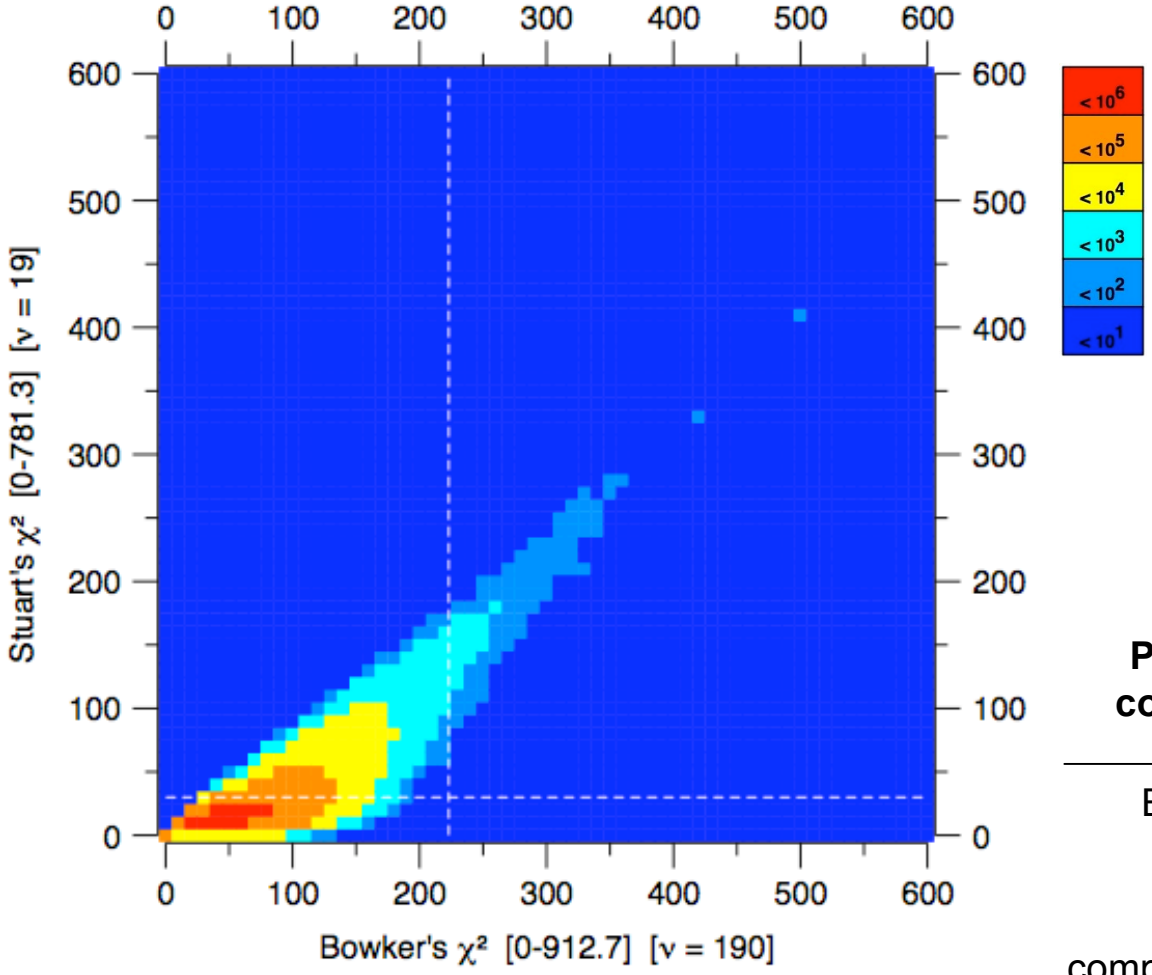
$$\chi^2_{stuart} = d' S^{-1} d \quad with\ d = x_{i\cdot} - x_{\cdot i}$$
$$and\ s_{ii} = x_{i\cdot} + x_{\cdot i} - 2x_{ii}, s_{ij} = -(x_{ij} + x_{ji}), i \neq j$$

- 'traditional' homogeneity tests
  - compositional χ^2

$$\chi^2_{compos.} = \sum_{i=1}^{K} \frac{(x_{i\cdot} - e_i)^2}{e_i} \qquad with\ e_i = \frac{x_{i\cdot} + x_{\cdot i}}{2}$$

Tavaré (1986) Lect Math Life Sci **17**:57-86; Jermiin *et al.* (2004) Syst Biol **53**:638-643

Correlations among all three tests



failure statistics

[AA, n = 2815041, p = 0.05]

Bowker's χ^2	0.23%
Stuart's χ^2	27.53%
compositional χ^2	2.47%

Pearson's correlation	Bowker's χ^2	Stuart's χ^2	compositional χ^2
Bowker's χ^2	–	0.752	0.712
Stuart's χ^2	0.752	–	0.946
compositional χ^2	0.712	0.946	–

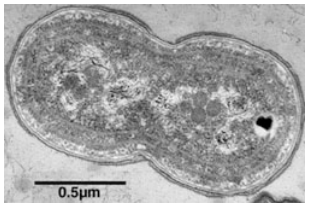
Ranking of the worst players

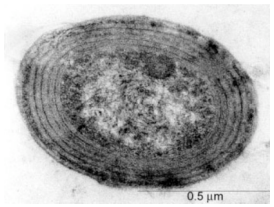
rank	fail. (#)	pairs (#)	fail. (%)	organism a	organism b
1	317	1274	24.88	*Synechococcus* WH8102	*Prochlorococcus* MED4
2	269	1245	21.61	*Prochlorococcus* MIT9313	*Prochlorococcus* MED4
3	204	835	24.43	*Thermosynechococcus* BP-1	*Prochlorococcus* MED4
4	203	749	27.10	*Gloeobacter violaceus*	*Prochlorococcus* MED4
5	168	868	19.35	*Synechocystis* PCC6803	*Prochlorococcus* MED4
6	160	390	41.03	*Wigglesworthia brevipalpis*	*Pseudomonas* PA01
…	…	…	…	…	…
12	152	1301	11.68	*Synechococcus* WH8102	*Prochlorococcus* SS120
…	…	…	…	…	…
29	131	774	16.93	*Gloeobacter violaceus*	*Prochlorococcus* SS120
…	…	…	…	…	…

[Bowker's χ^2, AA, n = 2815041, 10296 organism pairs]

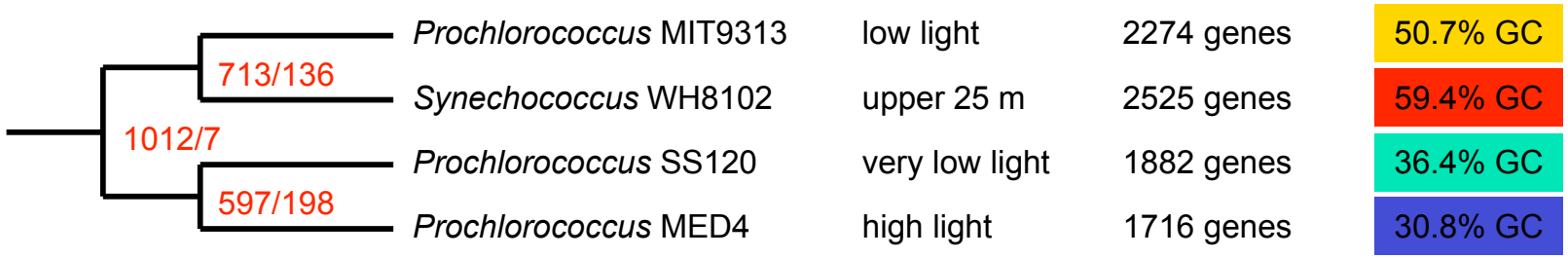
Who are the picocyanobacteria?

- both *Prochlorococcus* and *Synechococcus* are part of the picophytoplankton (tiny organisms: cell Ø < 1 µm)
  - they account for 1/3 of Earth's primary biomass production
  - all known members of this group are 96% similar at the rRNA level (but have quite different gene contents)
  - *Synechococcus* found 25 ya / *Prochlorococcus* found 15 ya



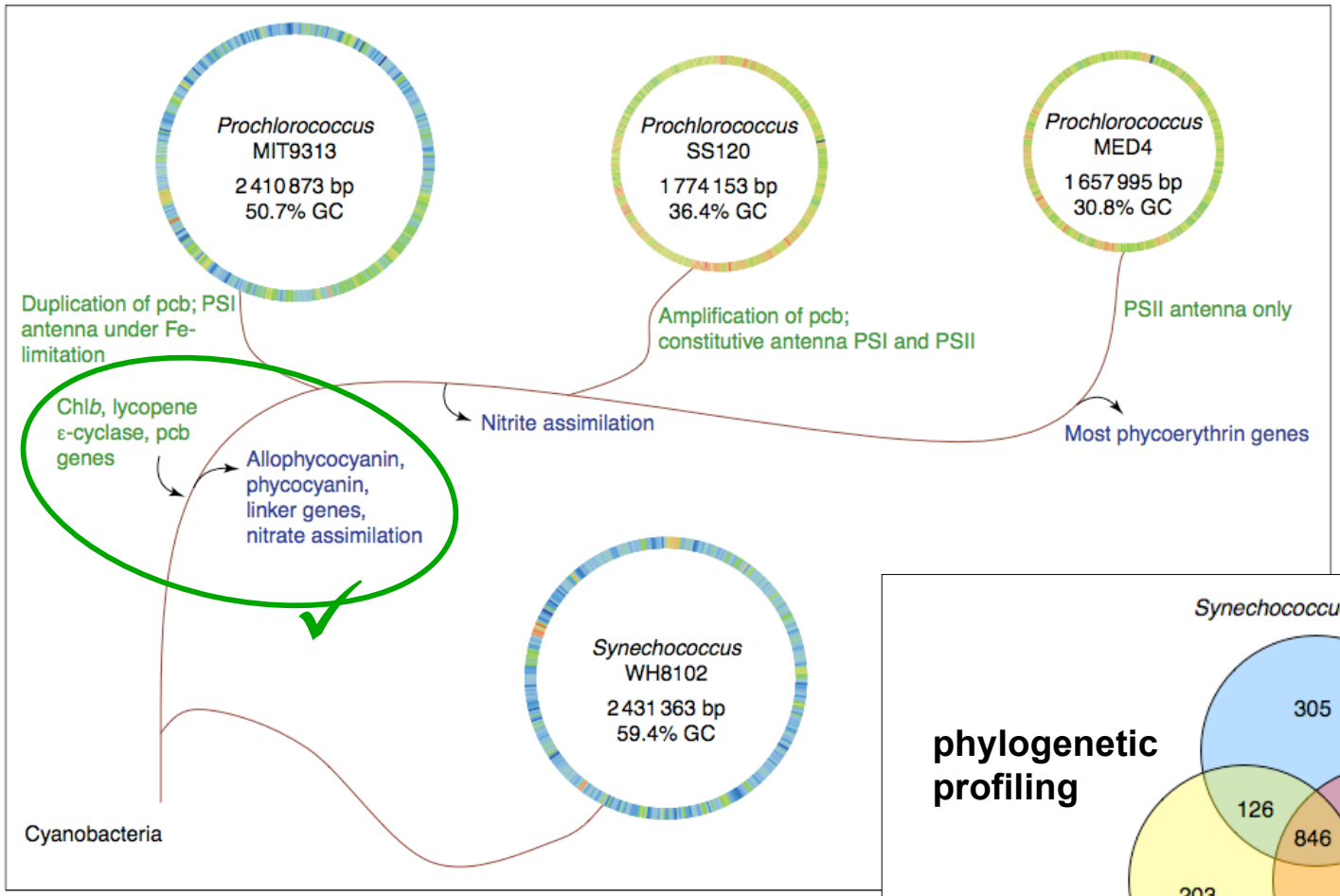


- we had 4 genomes in our 144-species dataset



Bryant (2003) PNAS **100**:9647-9649; Hess (2004) Curr Opin Biotech **15**:191-198

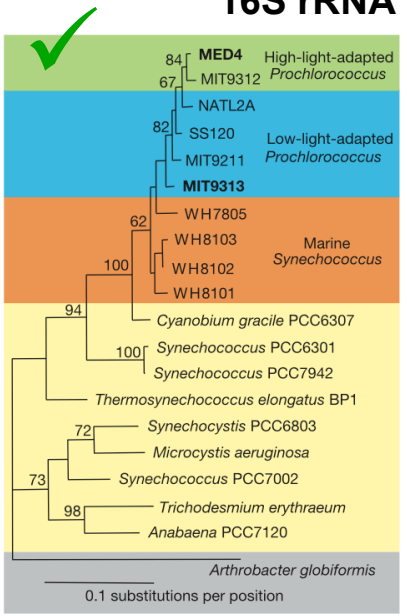
Is *Prochlorococcus* monophyletic?



photosynthetic apparatus

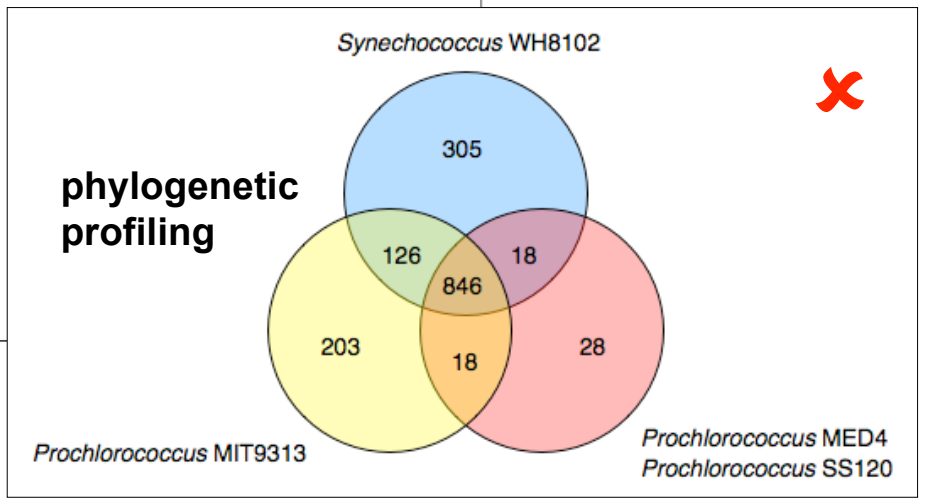
Hess (2004) Curr Opin Biotech **15**:191-198

16S rRNA



Rocap *et al.* (2003) Nature **424**:1042-1047

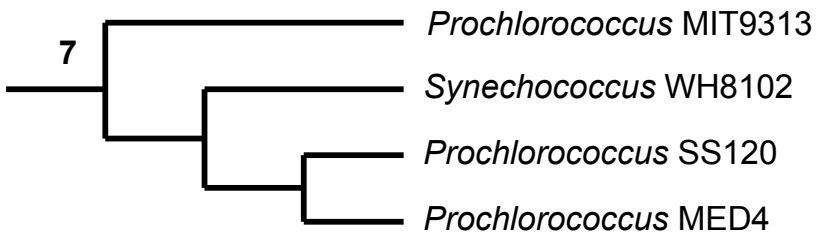
phylogenetic profiling

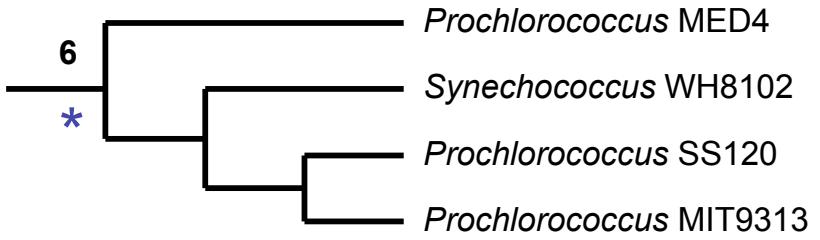


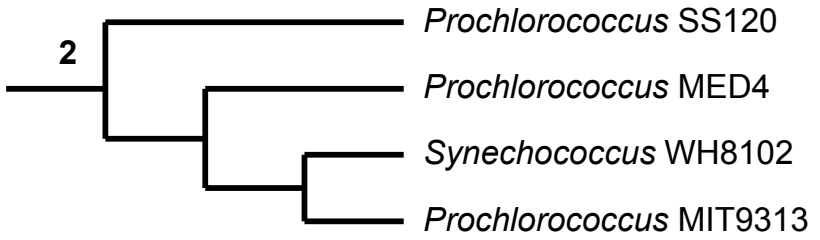
7 rooted trees…

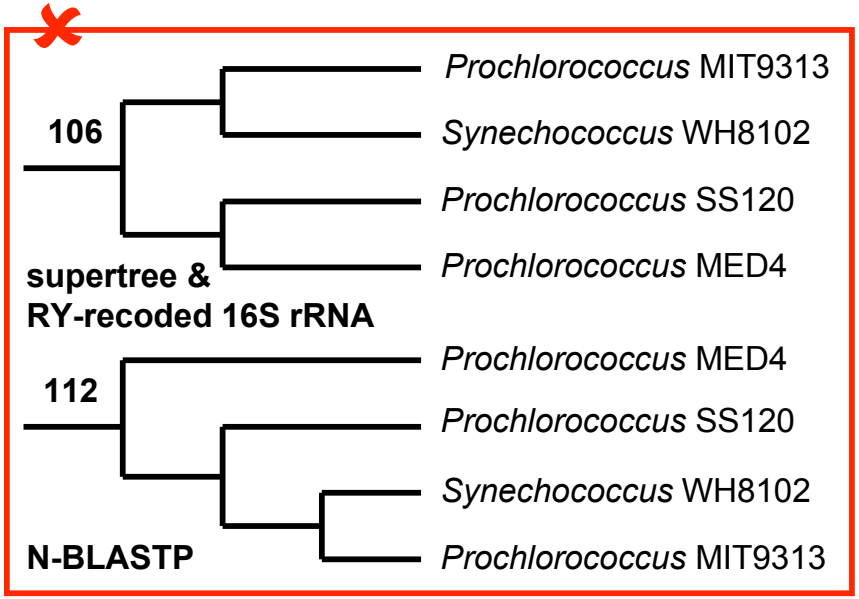
alignments of [size ≥ 6] for which the monophyly of the four picocyanobacteria is highly supported (n = 819)
only fully resolved topologies are considered (n = 310)

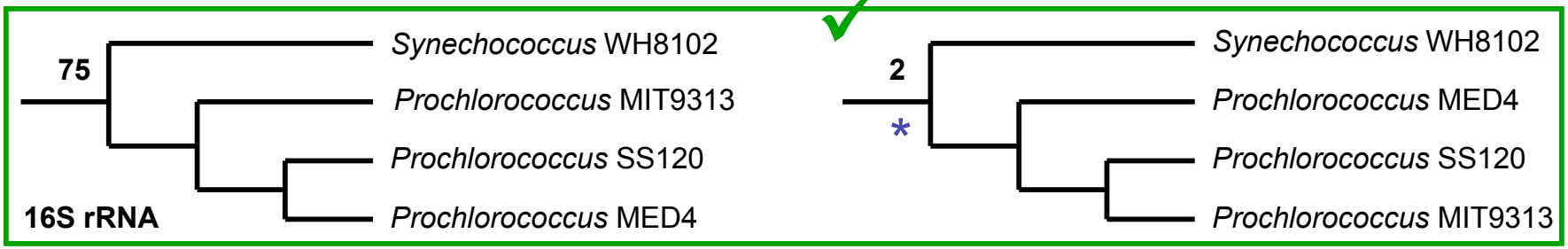
blue stars (*) denote *really* discordant topologies





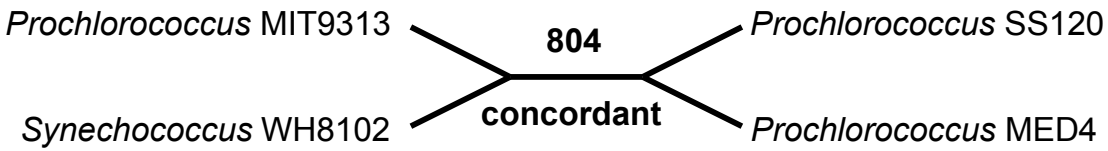


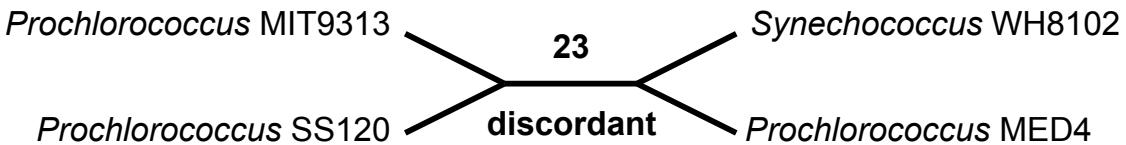


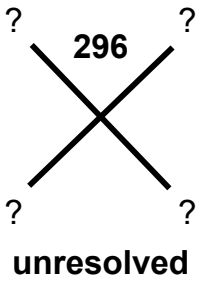


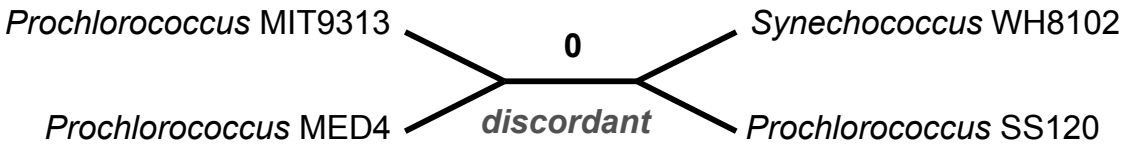
… fold to 2 unrooted topologies

same dataset + all alignments of [4 ≤ size ≤ 5] that include the four picocyanobacteria (n = 304)
all topologies are considered (n = 1123)









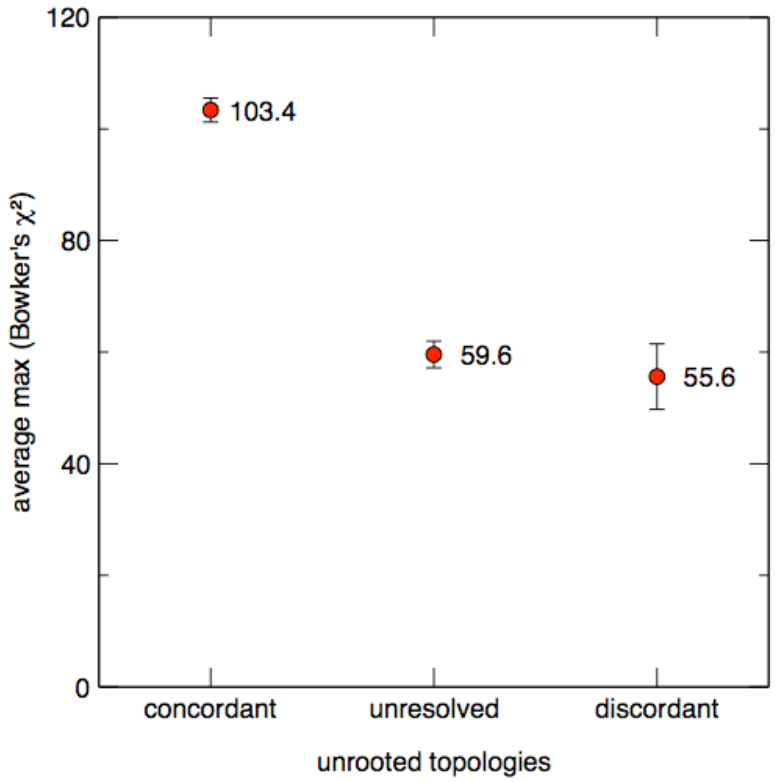
having binned all alignments leading to one given topology, we look for features of the compositional signal (compositional bias) that would be characteristic of that topology

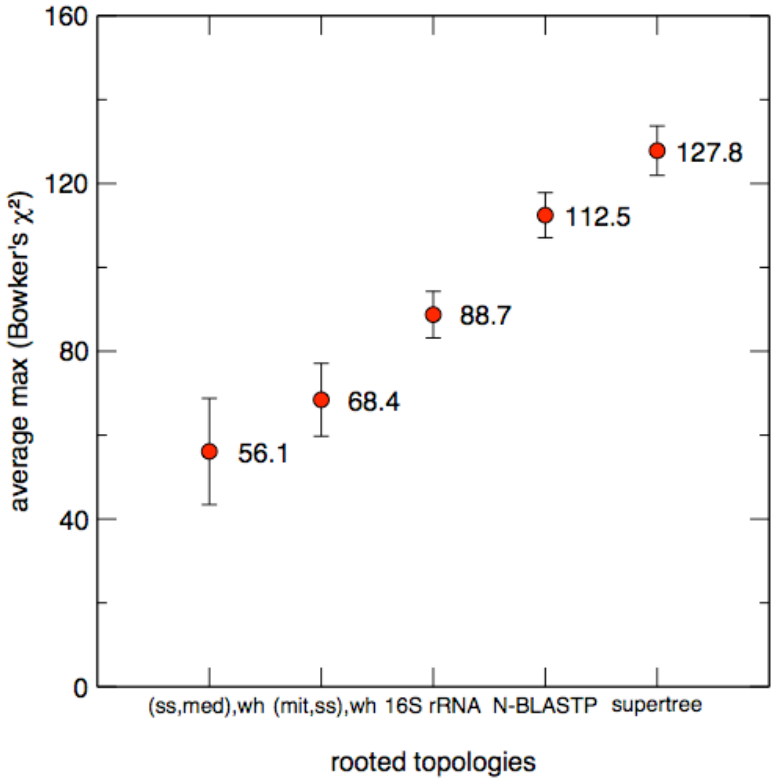
Signals that work together

$$\chi^2_{bowker} = \sum_{i<j} \frac{(x_{ij} - x_{ji})^2}{x_{ij} + x_{ji}}$$

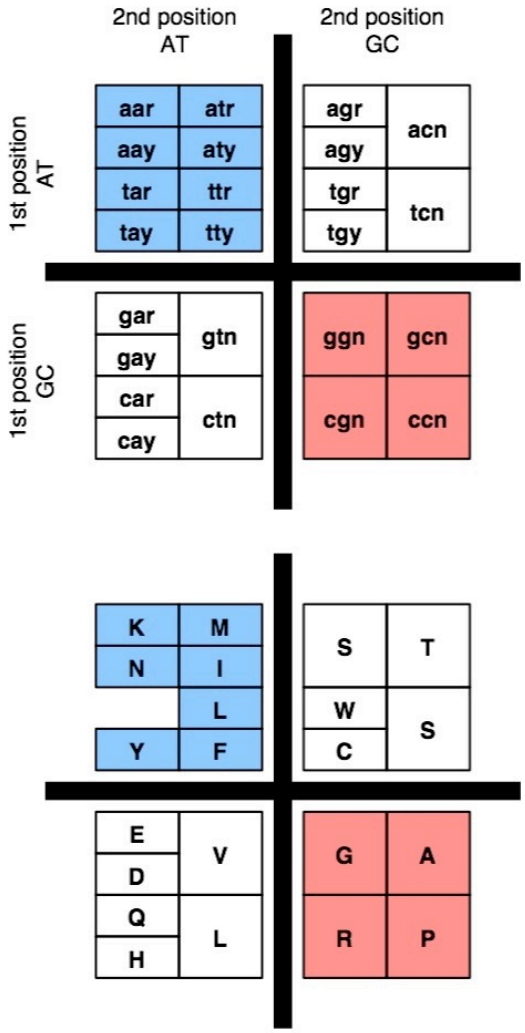
only bins of [size ≥ 6] are considered
bars denote standard error

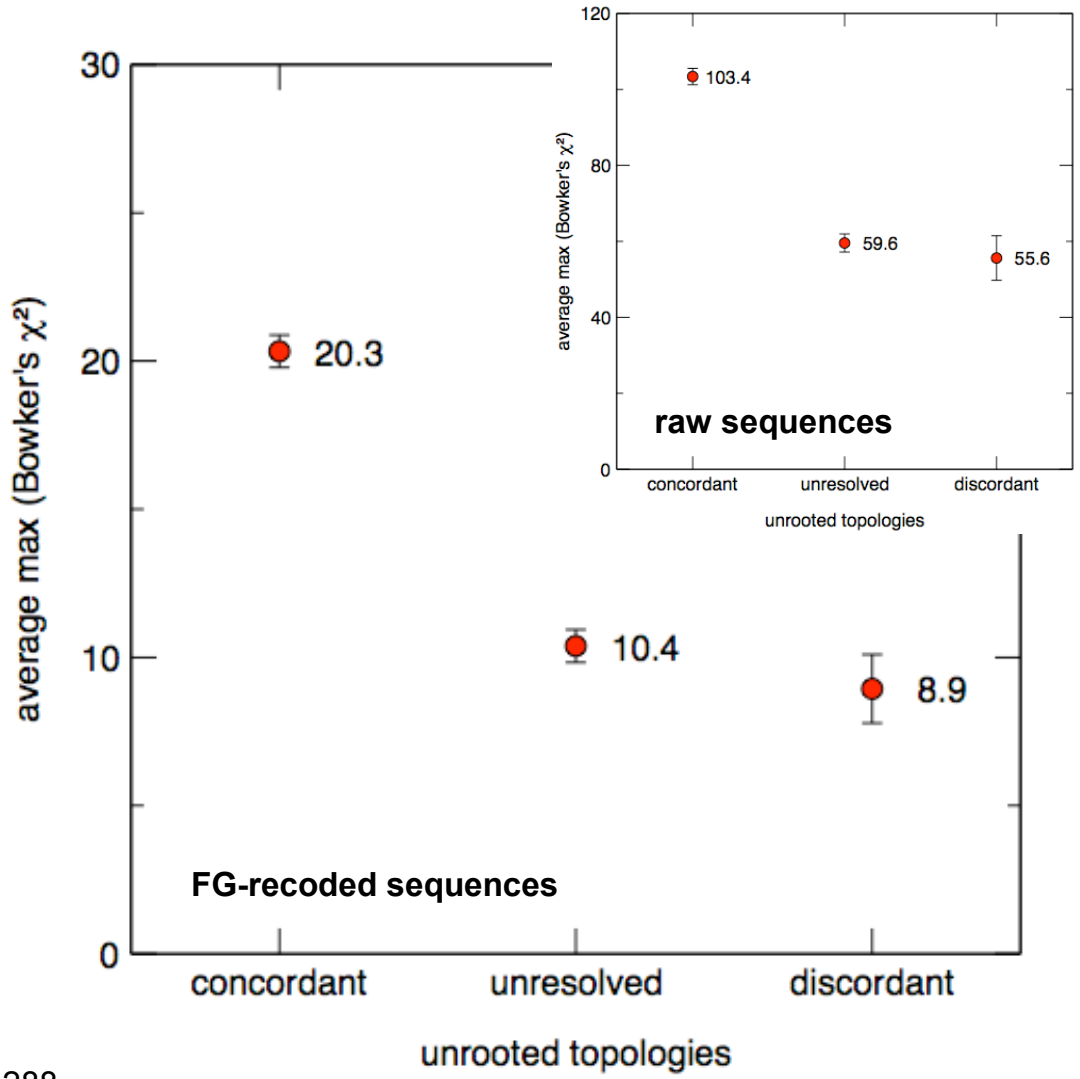
topologies are sorted on y values





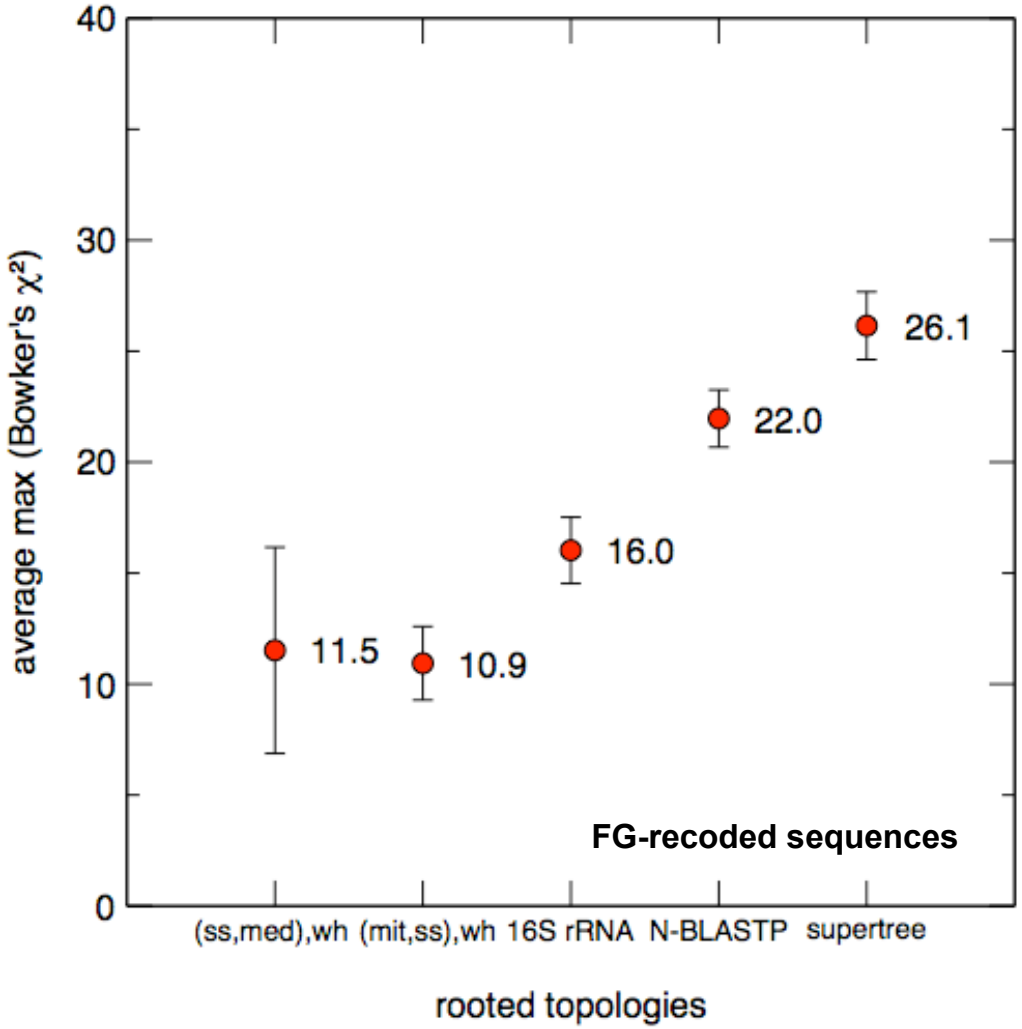
GC bias at the protein level?

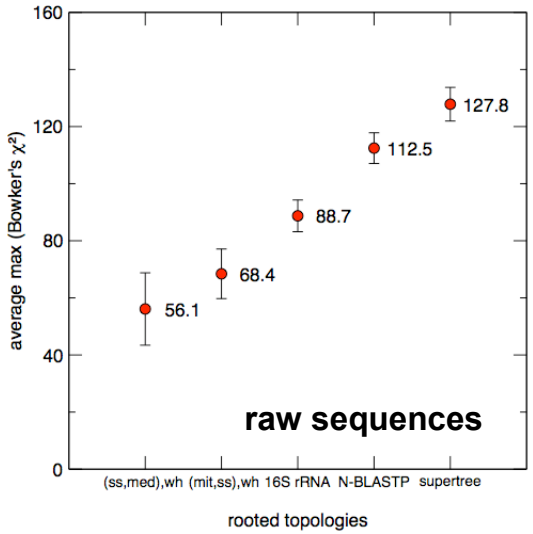




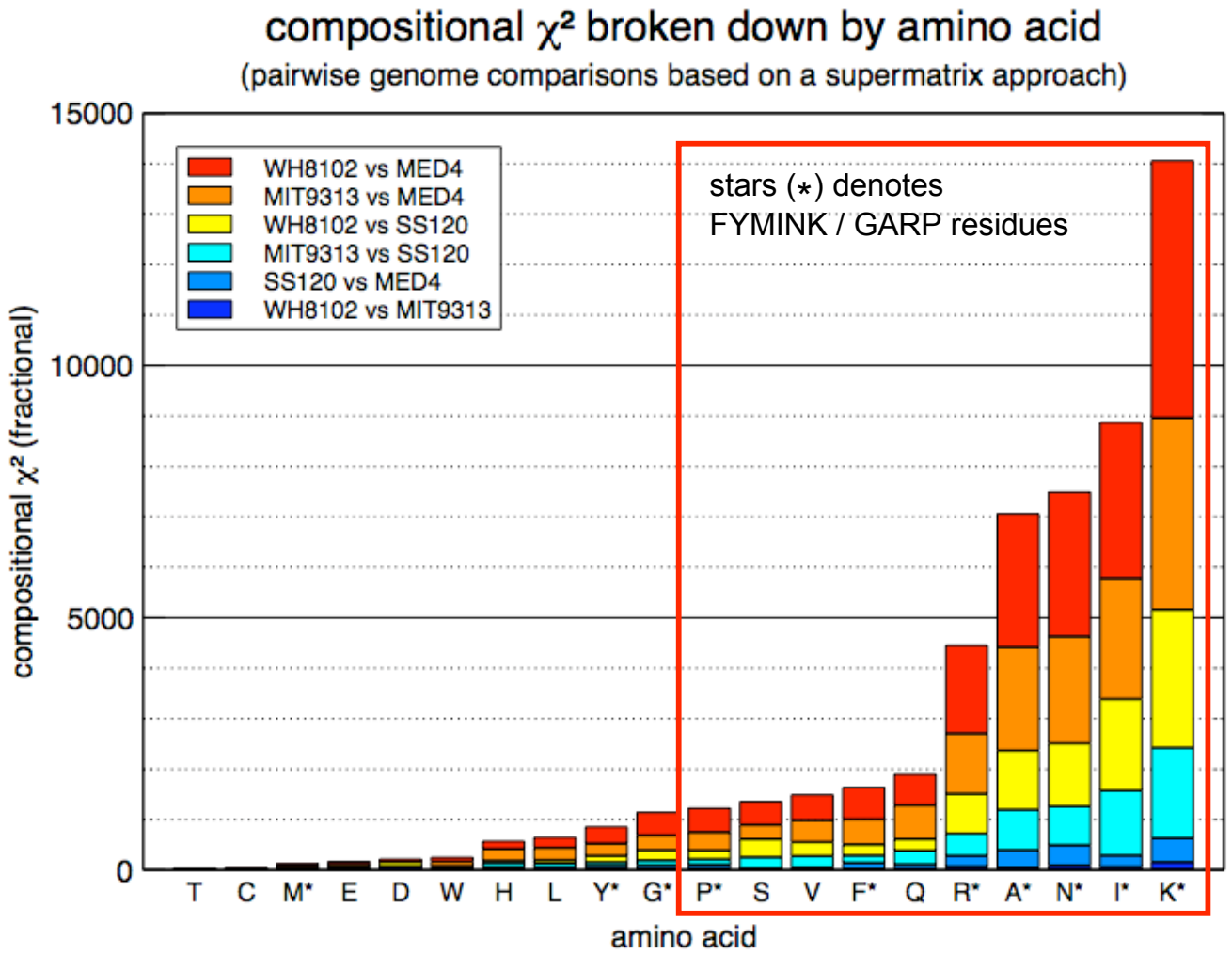
adapted from Foster *et al.* (1997) J Mol Evol **44**:282-288

GC bias at the protein level? (2)





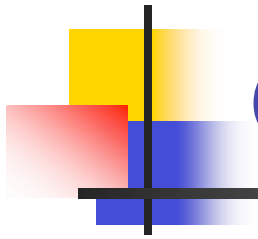
The case for a closer look



GC-content rankings

WH8102	59.4% GC
MIT9313	50.7% GC
SS120	36.4% GC
MED4	30.8% GC

'supergene' made of 1,485 genes found in at least two (out of four) organisms (469,682 positions; 15% missing)



Conclusions

- the compositional heterogeneity definitely has an impact on phylogenomic analyses
  - here, the compositional signal likely exaggerates the historical signal (i.e. non-monophyly of *Prochlorococcus*)
  - in other circumstances, it could be the opposite
- while the compositional heterogeneity is obvious at both the DNA and protein levels, the propagation of the GC bias is not limited to FYMINK / GARP codons
- in their 'holy war' against systematic biases, phylogenomic analyses would certainly benefit from newer evolutionary models that can deal with the violation of the stationarity assumption
  - e.g. Galtier and Gouy (1998) Mol Biol Evol **15**:871-879

Acknowledgments

- data, help and advice
  - Robert Charlebois (NeuroGadgets Inc.)
  - Nick Hamilton (UQ)
  - Tim Harlow (UQ / ACB)
  - Lars S. Jermiin (USyd)
- funding
  - FNRS / ULg (Belgium)
  - ACB (Australia)

The University Of Queensland AUSTRALIA | IMB Institute for Molecular Bioscience

ULg UNIVERSITÉ de Liège

FNRS Fonds National de la Recherche Scientifique

ACB Australian Research Council Centre in Bioinformatics