THE EVOLUTION OF NITROGEN FIXATION IN CYANOBACTERIA

SUPPLEMENTARY MATERIAL, REVISED VERSION, MARCH 2012

Table S1. See separate Supplementary File (Excel format).

Table S2 (in separate Supplementary File as tab-delimited text). The complete set of orthologous groups. The first column (headed 'group') gives the orthologous group number (unique within our study), or for singleton protein sequences, not assigned to any orthologous group, gives the sequence accession. Subsequent columns are headed by three-letter abbreviations for names of species or strains. For a key to these, see Table S1. Cells contain the count of sequences present in the orthologous group and species; and, if this is greater than zero, a colon followed by a comma-separated list of the corresponding protein sequence accessions.

Table S3. See separate Supplementary File (Excel format).

Genus	Taxa in GenBank	Taxa in our study	Taxa used (%)
Acaryochloris	12	1	8.3
Anabaena	390	2	0.5
Arcobacter	81	1	1.2
Arthrospira	176	1	0.6
Campylobacter	167	5	3.0
Crocosphaera	13	1	7.7
Cyanothece	37	7	18.9
Escherichia	1016	3	0.3
Gloeobacter	7	1	14.3
Helicobacter	315	1	0.3
Lyngbya	145	1	0.7
Microcystis	1041	1	0.1
Nautilia	12	1	8.3
Nitratiruptor	8	1	12.5
Nodularia	73	1	1.4
Nostoc	993	2	0.2
Prochlorococcus	203	12	5.9
Sulfurimonas	26	1	3.8
Sulfurospirillum	41	1	2.4
Sulfurovum	9	1	11.1
Synechococcus	829	16	1.9
Synechocystis	74	1	1.4
Thermosynechococcus	8	1	12.5
Trichodesmium	49	1	2.0
Wolinella	16	1	6.3
Total:	5741	65	1.1

Table S4. Taxon sampling within the genera in our study, compared to taxon sampling within these genera in GenBank. Counts in GenBank were obtained from the NCBI Taxonomy Database (http://www.ncbi.nlm.nih.gov/Taxonomy) in December 2011.

Supplementary Data 1 (separate Nexus-format file). The rooted cyanobacterial species phylogeny with the proteobacterial outgroup excluded.

Supplementary Data 2 (separate Nexus-format file). The rooted species phylogeny, including both the cyanobacterial ingroup and the proteobacterial outgroup.

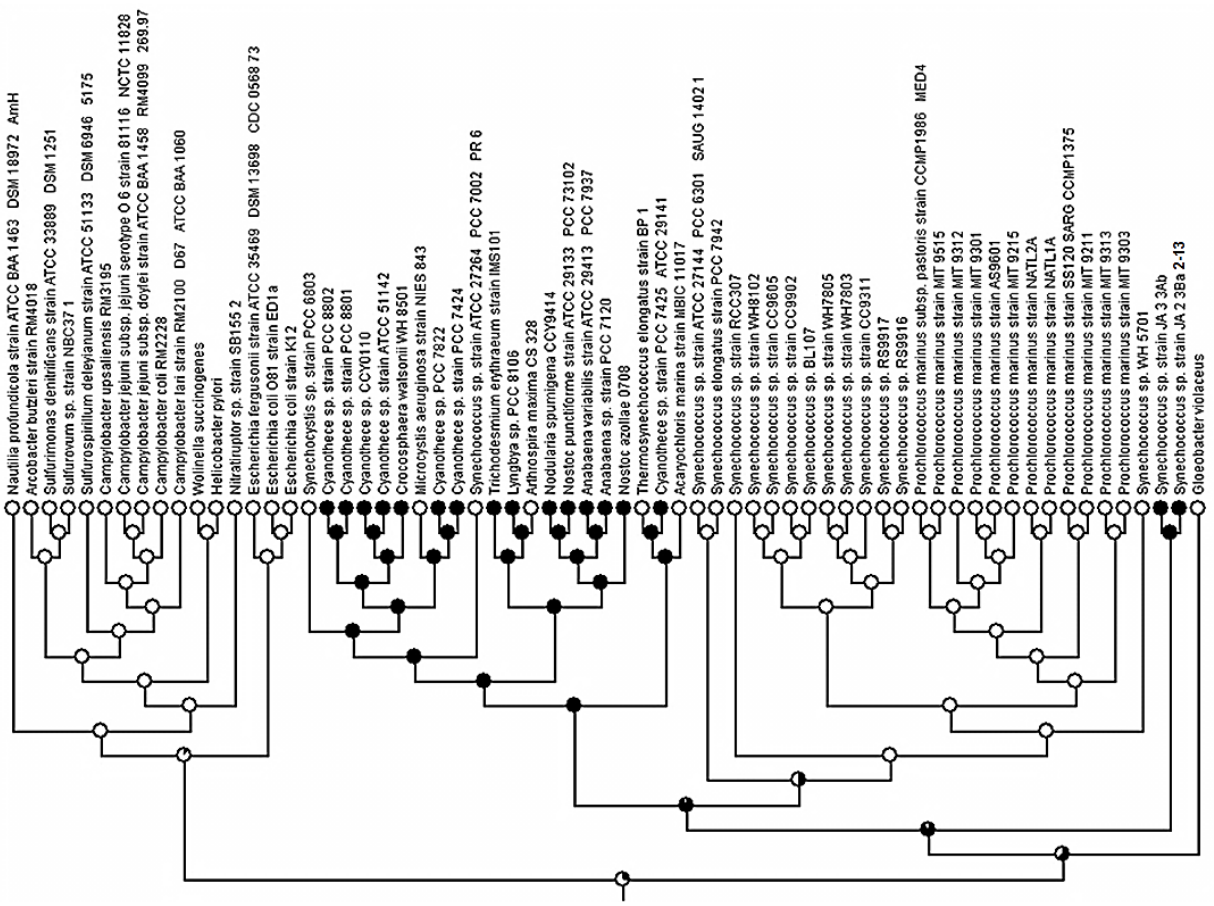

Figure S1. Rooted maximum likelihood phylogeny of species with the proteobacterial outgroup included. Pie charts represent the posterior probability of presence (black) and absence (white) of nitrogen fixation according to the Asymm.2 model of trait evolution.

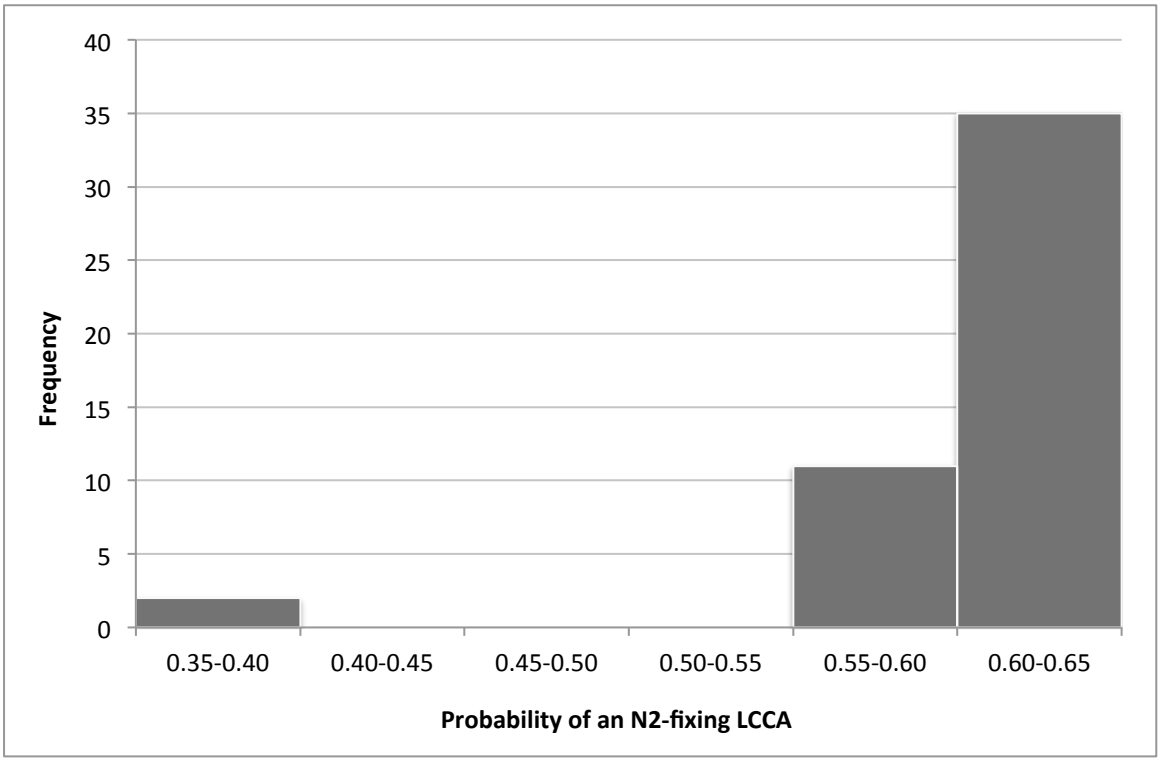

Figure S2. Distribution of the probabilities of a nitrogen-fixing cyanobacterial last common ancestor across 48 bootstrap replicate phylogeny reconstructions (n = 48, SD = 0.050, range 0.363 to 0.636).

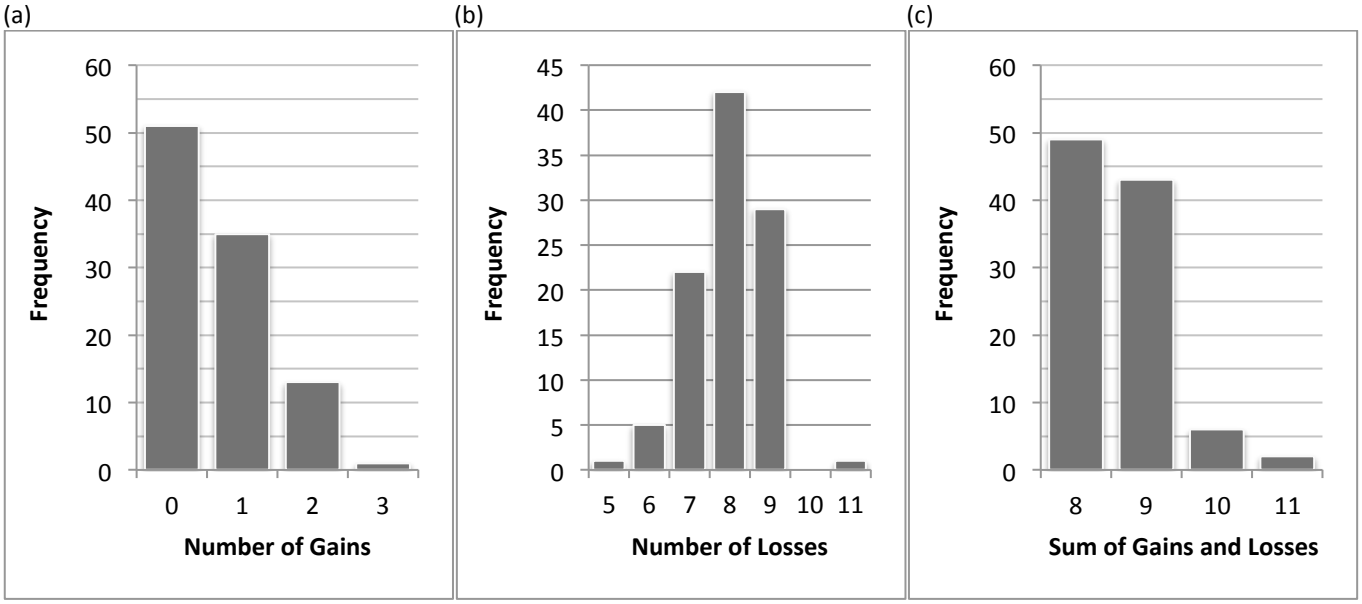

Figure S3. The frequency of counts of (a) gains, (b) losses, and (c) their sum, for the nitrogen-fixing trait on the ML cyanobacterial phylogeny for 100 stochastic mappings (excluding change and reversal on the same branch).

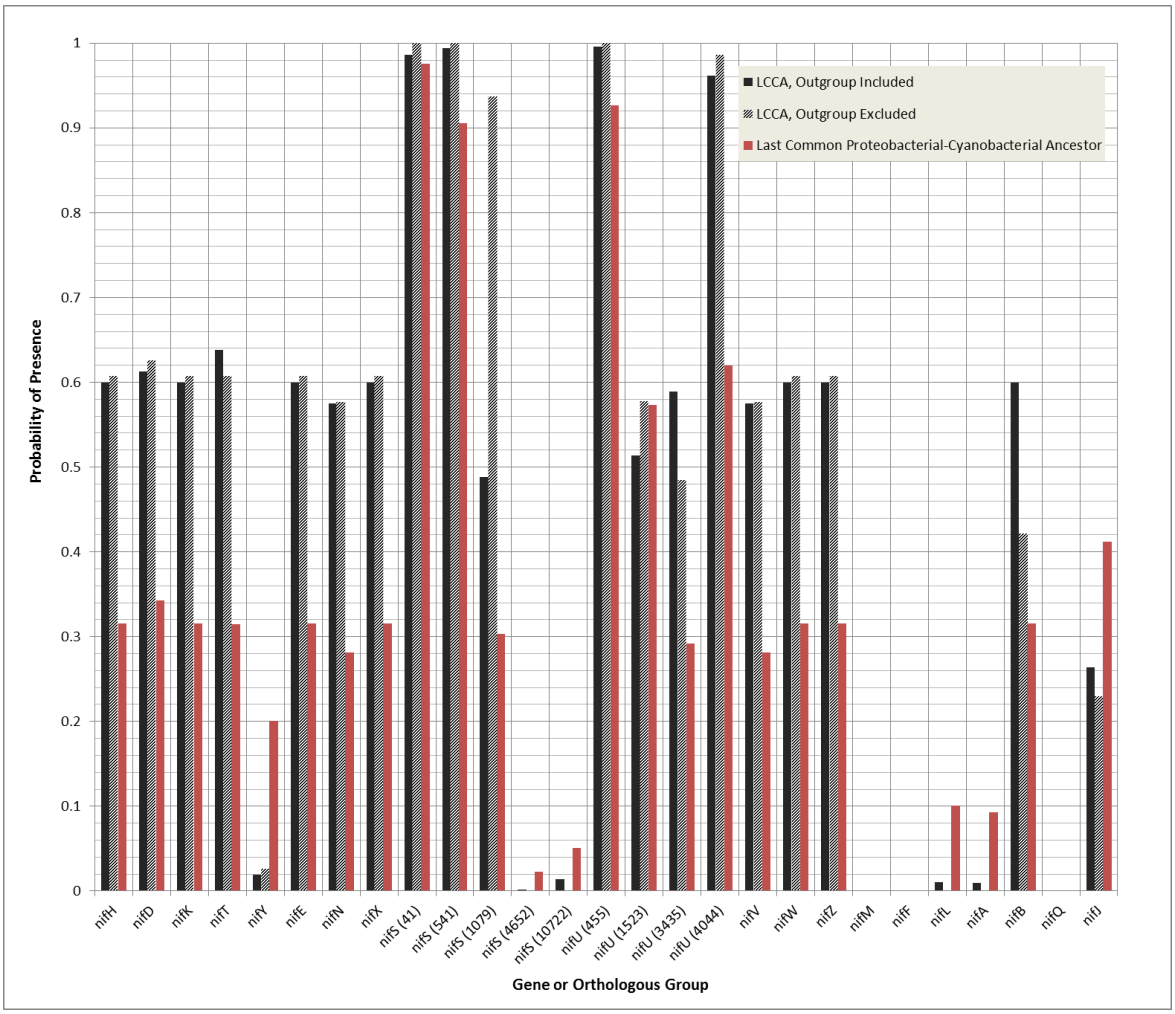

Figure S4. The results of the ancestral state reconstructions for each of the *nif* gene or orthologous group, showing the results of both the purely cyanobacterial phylogeny and the cyanobacterial phylogeny with the proteobacterial outgroup included. The reconstructions for the N_2 fixing trait are identical to those for *nifHKEXWZ,* with which the trait shares a pattern of presence and absence in the extant species and strains. See Supplementary Table S2 for the complete list of orthologous groups in cyanobacteria and Supplementary Table S3 for the accession numbers of species-specific *nif* genes.

S1 PHYLOGENY RECONSTRUCTION

Multiple alignment of each universal, single-copy orthologous group was performed using MAFFT in 'E-INS-I' mode with 1000 iterations (Katoh and Toh 2008). For a concatenation of these multiple alignments, a phylogenetic model, LG+Γ, was selected using the Bayesian Information criterion in MODELGENERATOR (Keane *et al*., 2006), with 4 rate categories. Phylogeny was estimated using 'best' rearrangements in PhyML (Guindon *et al*., 2010). 50 bootstrap replicates of the concatenation were generated using seqboot in the PHYLIP package (Felsenstein 2008); sequence order within each replicate was randomized, and the phylogeny of each was estimated using LG+Γ in PhyML with 'NNI' rearrangements. Bootstrap support for clades was obtained using consense in the PHYLIP package (version 3.68; J. Felsenstein, http://evolution.gs.washington.edu/phylip.html).

S2 ANCESTRAL STATE RECONSTRUCTIONS

For reconstructions excluding the outgroup, we predict that the complement of *nif* genes in LCCA (inferred as individual posterior probability of presence > 0.5) is as follows: *nifH*, *nifD*, *nifK*, *nifT*, *nifE*, *nifN*, *nifX*, *nifS(41)*,

nifS(541), *nifS(1079)*, *nifU(455)*, *nifU(1523)*, *nifU(4044)*, *nifV*, *nifW* and *nifZ* (Figure S4). We consider this set of reconstructions to be our most robust.

Including the proteobacterial outgroup when reconstructing states for LCCA introduces two arbitrary elements that influence the reconstructions. Firstly, proteobacteria were not the focus of our study and our proteobacterial taxon sampling is incomplete and patchy. Secondly, there is no reason to believe the position of the root really is at the mid-point of the branch (in the unrooted tree) between ingroup and outgroup; yet for ancestral state reconstruction, it must be placed somewhere. However, the reconstruction of LCCA is generally robust to inclusion of the outgroup. Compared to the ingroup-only reconstruction of gene content, the effects on reconstructions of *nif* gene presence (inferred as individual posterior probability > 0.5) are the removal of *nifS(1079)* and the addition of *nifU(3435)* and *nifB*. LCCA is still reconstructed as N_2 fixing (Figures S1, S4).

We expect our reconstructions for the proteobacterial-cyanobacterial ancestor to be even more sensitive to taxon sampling within the proteobacteria. Our reconstructions suggest the proteobacterial-cyanobacterial ancestor was not nitrogen fixing (Figure S1) but did possess *nifS(41)*, *nifS(541)*, *nifU(455)*, *nifU(1523)* and *nifU(4044)* (Figure S4).

S3 REFERENCES

Guindon,S. *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.,* **59**, 307-321.

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286-298.

Keane,T.M. *et al.* (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.*, **6**, 29.

S4 CHANGES SINCE PREVIOUS VERSION

In the supplementary text published on the *Bioinformatics* Web site in January 2012, Figure S4 and its caption were in error concerning *nifD*. In the current document, Figure S4 and its caption have been corrected.

For all other supplementary files, the versions published in January 2012 are the correct ones, and may be found on the *Bioinformatics* Web site at:

http://bioinformatics.oxfordjournals.org/cgi/content/abstract/bts008?ijkey=upbeUx8GGv5yM5T&keytype=ref

or:

http://dx.doi.org/10.1093/bioinformatics/bts008