Concerted gene recruitment in early plant evolution

Jinling Huang   J. Peter Gogarten

1. Materials and methods

Data sources: Protein sequences for the red alga Cyanidioschyzon merolae were obtained from the Cyanidioschyzon Genome Project [41, 73]. EST sequences of several protists were obtained from TBestDB [74]. All other sequences were from the NCBI protein sequence database.

Identification of ancient HGT: Anciently acquired genes in this study include those horizontally acquired prior to the split of red algae and green plants. A list of ancient HGT candidates was first identified based on phylogenomic screening of the Cyanidioschyzon genome using PhyloGenie and the NCBI non-redundant protein sequence database. The vast majority of the genes on this list are predominantly identified in bacteria and archaea, and therefore are likely of prokaryotic origin. To reduce the complications arising from potential cases of IGT, we adopted an approach combining sequence comparison, phylogenetic analyses, and statistical tests. Each gene on the list was first used to search the NCBI protein sequence database. Because of the cyanobacterial origin of plastids and the α-proteobacterial origin of mitochondria, genes with cyanobacterial and plastid-containing eukaryotic homologs as top hits were considered as likely plastid-derived; those with proteobacterial and other eukaryotic homologs were considered as likely mitochondrion-derived. These potentially organelle-derived genes were removed from the candidate list and the remaining genes were subject to detailed phylogenetic analyses (see below). Gene tree topologies generated through detailed phylogenetic analyses were subject to careful inspections; any genes that formed a monophyly with cyanobacterial homologs or with proteobacterial and other eukaryotic sequences were also eliminated from further consideration. Additionally, alternative topologies representing various evolutionary scenarios for each gene were statistically evaluated based on AU tests [43] (see below); genes for which a straightforward IGT scenario (versus IGT followed by secondary transfers) could not be rejected (p-value > 0.05) were also removed from the HGT candidate list.

Detailed phylogenetic analyses: Sequences were sampled from representative groups (including major phyla of bacteria and major groups of eukaryotes) within each domain of life (bacteria, archaea, and eukaryotes). Because of the potential for sequence contaminations, eukaryotic EST sequences whose authenticity is suspicious (e.g., high nucleotide sequence percent identity with bacterial homologs and/or absence of homologs in genomes of closely related taxa) were not included in the analyses. Multiple protein sequence alignments were performed using MUSCLE [77] and clustalx [78], followed by cross-comparisons and manual refinement. Only unambiguously aligned sequence portions were used. Phylogenetic analyses were performed with a maximum likelihood method using PHYML [79], a Bayesian inference method using MrBayes [80], and a distance method using the program neighbor of PHYLIP version 3.65 [81] with maximum likelihood distances calculated using TREE-PUZZLE [82]. All maximum likelihood calculations were based on a substitution matrix determined using ProtTest [83] and a mixed model of 4 gamma-distributed rate classes plus invariant sites.
Maximum likelihood distances for bootstrap analyses were calculated using TREE-PUZZLE and PUZZLEBOOT v1.03 (by Michael E. Holder and Andrew J. Roger, available on the web [84]). Branch lengths and topologies of the trees depicted in all figures were calculated with PHYML. For the convenience of presentation, gene trees were rooted using archaeal (or archaeal + eukaryotic) sequences, or paralogous gene copies if ancient gene families were involved, as outgroups; otherwise, trees were rooted in a way that no top hits of the sequence similarity search were used as an outgroup. Nevertheless, all gene trees should be strictly interpreted as unrooted.

AU tests on alternative tree topologies: Following detailed phylogenetic analyses, alternative tree topologies for each remaining HGT candidates were assessed for their statistical confidence using Treefinder [85]. In most cases, multiple constraint trees were generated using Treefinder for each HGT candidate by enforcing a) monophyly of all eukaryotic sequences, b) monophyly of cyanobacterial, plant and other plastid-containing eukaryotic sequences, and c) monophyly of cyanobacterial, plant, and closely related bacterial sequences. These alternative topologies assumed that the subject gene in plants are not HGT-derived; they served as null hypotheses that all eukaryotic sequences have the same eukaryotic or mitochondrial origin or that plants acquired the subject gene from plastids, sometimes followed by secondary HGT to other bacterial groups. AU tests, which have been recommended for general tree tests, were performed on alternative tree topologies (non-HGT hypotheses) and the tree generated from detailed phylogenetic analyses (HGT hypothesis). In this study, topologies with a p-value < 0.05 were rejected.

Prediction of protein localization: Targeting signal of identified protein sequences was predicted using ChloroP [86] and TargetP [87]. Additional information about protein localization in green plants was obtained from The Arabidopsis Information Resource (TAIR).

2. Protein sequence alignment used for phylogenetic analyses and resulting phylogenetic trees. Each sequence name includes a GI number from GenBank (or ID number from other databases) followed by species name. Numbers above the branches of the gene tree show bootstrap support values for maximum-likelihood analyses and distance analyses, and posterior probability from Bayesian analyses, respectively. Asterisks indicate support values below 50%. N denotes genes whose homologs are rarely found in cyanobacteria and that likely possessed novel functions; E denotes genes for which plastid-derived homologs already exist in plants; D denotes genes for which a possible replacement of an endogenous homolog cannot be excluded.

1. GCN5-related N-acetyltransferase (E)

CLUSTAL X (1.83.1) multiple sequence alignment

| Sequence Name                  | Alignment         |
|-------------------------------|-------------------|
| 32444545_Rhodopirellula_baltic| KADTLIEAMGWRFRGKTTVRIKGLSSLEDREALQH1LLLVIFMETVGL |
| 21541979_Mycobacterium_leprae  | KAEVLAEALPWLQRIKVGVRVYSGNAMTDMLRRAFAADMAFLRNCGI |
| 488344901_Thermobifida_fusca  | KAKTLIEALPWLSAHHGKTVVYKGNAMIDEDLAAFAQDFFFVYVGL |
| 7106696_Streptomyces_coelic   | KAQTLIEALPWLQRIKVGVRVYSGNAMTDMLRRAFAADMAFLRNCGI |
| 30409279_Cyanidioschyzon_merol| RVQVLTEALPYQIKWNEIMVIRYGGAVVQKD---ADITKDIILFLXCGF |
Note: All top hits of GenBank Searches (using Cyanidioschyzon sequence and Arabidopsis accession number NP_974701 as queries) are from gamma and beta-proteobacteria. Multiple copies of this gene exist in plants and in bacteria. One of these copies in red algae and green plants forms a group with homologs from gamma and beta-proteobacteria with strong support, whereas the other copy groups with cyanobacterial sequences with modest support (the Cyanidioschyzon sequence in that group is encoded in the plastidic genome). In Arabidopsis, the protein product of the gamma and beta-proteobacteria-related gene copy is annotated as located in cytoplasm (GenBank accession number NP_974701 and TAIR locus AT4G37670). The EST sequence of glaucophyte Cyanophora was obtained from tbESTdb; this sequences groups with homologs of red algae and green plants in preliminary analyses, but was removed from the detailed phylogenetic analyses because of its very short length (only 32 aa).
Figure 1. Molecular phylogeny of GCN5-related N-acetyltransferase. P-value = 0.235 from AU test on the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of all red algal and green plant sequences, and (B) monophyly of cyanobacterial and all red algal, green plant sequences. These alternative topologies investigate if both copies of this gene in red algae and green plants have the same plastidic origin. P-values < 0.001 from AU tests on both alternative topologies.

2. Glycyl-tRNA synthetase (D)

Note: GlyRS in living organisms exists in forms of homodimer ($\alpha_2$) and heterotetramer ($\alpha_2\beta_2$); the former is distributed in eukaryotes, archaea, and many bacteria whereas the latter is only found in bacteria, red algae and green plants. Few bacteria contain both glyRS types. The $\alpha$- and $\beta$-subunits of tetrameric glyRS are usually encoded in separate genes. In a few groups, the two subunits are encoded in a single fused gene; these include actinomycetes, chlamydiae, red algae, and green plants. Not only are sequences from actinomycetes, red algae and green plants similar in gene structure, they also have the highest percent identity and share several conserved amino acid residues. Phylogenetic analyses of each of the subunits strongly suggest a common origin of actinomycete, red algal and plant sequences (Figures 2A-2B). Based on
the gene structure and molecular phylogeny, it is likely that primary photosynthetic eukaryotes acquired this gene from either actinomycetes or chlamydiae (Chlamydiae are the only bacterial group aside from actinomycetes and plants that possess a fused gene in our database searches). The second scenario requires an independent HGT event from photosynthetic eukaryotes or chlamydiae to actinomycetes. The *Arabidopsis* sequence (GenBank accession number NP_190394, TAIR locus AT3G48110) is experimentally determined to be targeted to both chloroplasts and mitochondria.

2A. Alpha subunit

**CLUSTAL X (1.83.1) multiple sequence alignment**

|       | Accession  | Species                  | Clusters | Sequence                                      |
|-------|------------|--------------------------|----------|-----------------------------------------------|
|       |            |                          |          |                                               |
|       |            |                          |          |                                               |
|       |            |                          |          |                                               |
|       |            |                          |          |                                               |

| Accession  | Species                  | Clusters | Sequence                                      |
|------------|--------------------------|----------|-----------------------------------------------|
|            |                          |          |                                               |
|            |                          |          |                                               |
|            |                          |          |                                               |
|            |                          |          |                                               |
**Figure 2A.** Molecular phylogeny of alpha subunit of glycyl-tRNA synthetase. P-value = 0.838 from AU test on the presented tree. AU tests were also performed for alternative topologies, including (A) monophyly of cyanobacterial and photosynthetic eukaryotic sequences, (B) monophyly of cyanobacterial, *Frankia*, *Trocheryma*, and photosynthetic eukaryotic sequences, and (C) monophyly of chlamydial, *Frankia*, *Trocheryma*, and photosynthetic eukaryotic sequences. These tests investigate different evolutionary scenarios, including a) plastidic origin of photosynthetic eukaryotic sequences, b) photosynthetic eukaryotes acquired this gene from plastids and subsequently spread to *Frankia* and *Trocheryma* by a secondary HGT event, c) photosynthetic eukaryotes acquired this gene from chlamydiae and subsequently spread to *Frankia* and *Trocheryma*. P-value < 0.001 from AU test for topology A whereas P-values = 0.247 and 0.287 for topologies B and C, respectively. Therefore, the scenario of a plastidic or chlamydial origin of this gene in red algae and green plants, though less parsimonious, cannot be confidently rejected.

### 2B. Beta subunit

**CLUSTAL X (1.83.1) multiple sequence alignment**

| Accession | Organism                        | Sequence                                                                 |
|-----------|---------------------------------|--------------------------------------------------------------------------|
| 68172611  | *Frankia sp.*                   | LFEIGTEELPAEVTREVTRVEAVRAGLVERLAATRILTGTIPRIVALIVATDEVA                 |
| 68230323  | *Frankia*                       | -------------- --------------------------------------------------------------- |
| 28572421  | *Trocheryma whippl*             | LFEIGTEEMPSS-TQEITTTVNLKRELGRSLQKDATPRIVIYKMH                         |
| 46399968  | *Parachlamydia sp.*             | VIEIGSELPLAPSVIGQGMQLLERALLKEGISGTFPRIVAYYLYLS                        |
| C_130039  | *Chlamydomonas reinh*           | VLEIGVEELPDDVYASVQRLERYPVAKKALGLSKEGTPRIVALVYEPAAL                    |
| 15863647  | *Chlamydomphila pneumo*         | LIEEGSEELPFFVPIQGLQESLARKVQDNLGSPRALTLLVYNVA                         |
| 15659330  | *Chlamydia tracho*              | VIEIGTEELPPADLDAEVRQLEHQLQDNRLRATPRHAILVUOGVA                         |
| 55296761  | *Oryza sativa*                  | LIEIGEELPLFRLPHLSKSWTTLKQRMETYFTPTVLRLVLSQFAP                       |
| 30629278  | *Arabidopsis thalia*            | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| CM269     | *Cyanidioschyzon merolae*       | LIEIGEELPPADLDAEVRQLEHQLQDNRLRATPRHAILVUOGVA                         |
| 34482366  | *Wolinella succin*              | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 15611973  | *Helicobacter pylori*           | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 15607085  | *Aquifex aeolic*                | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 66855634  | *Anaeromyxobacter deha*         | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 46580306  | *Desulfuvibrio vulgar*          | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 4990715   | *Thermotoga marit*              | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 68207212  | *Desulfotobacterium ha*         | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 50914778  | *Streptococcus pyogen*          | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 17133246  | *Nostoc sp.*                    | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 45508785  | *Anabaena variabilis*           | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 72382003  | *Prochlorococcus marinus*       | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 46907686  | *Listeria monocytogenes*        | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 23099403  | *Oceanococcus thieyanus*        | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 68179078  | *Desulforudomona acetox*        | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 39982449  | *Geobacter sulfurreducens*      | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 51891713  | *Symbio bacterium thermophilus*  | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 67931182  | *Solibacter usitatimuricus*     | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 1573946   | *Haemophilus influenzae*        | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 26986005  | *Pseudomonas putida*            | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 13476182  | *Mesorhizobium loti*            | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 39944122  | *Rhodopseudomonas palustris*    | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |
| 23013833  | *Magnetospirillum gryphiswaldense* | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                    |
| 68213210  | *Methyllococcus flagellatus*    | LIEIGEELPPQDNQVTNLLECQRLQVLVYDANS                                        |

**Table continued...**
**Figure 2B.** Molecular phylogeny of beta subunit of glycyl-tRNA synthetase. P-value = 0.731 from AU test for the presented tree. AU tests were also performed on alternative topologies, including (A) monophyly of cyanobacterial and photosynthetic eukaryotic sequences, and (B) monophyly of chlamydial, *Frankia*, *Trophieryma*, and photosynthetic eukaryotic sequences. These tests investigate different scenarios, including a) plastid origin of photosynthetic eukaryotic sequences, and b) photosynthetic eukaryotes acquired this gene from chlamydiae and subsequently spread to *Frankia* and *Trophieryma*. P-values = 0.085 and 0.446 from AU tests for these alternative topologies respectively. Therefore, the scenario of a plastidic or chlamydial origin of this gene in red algae and green plants cannot be confidently rejected based on these statistical tests. However, given that the two glycyl-tRNA subunits likely acquired from a single event, we have chosen to present this gene here as a likely case of HGT.

---

3. Dihydrodipicolinate synthase (dapA) (D)

CLUSTAL X (1.83.1) multiple sequence alignment
| Accession    | Organism                                    | Sequence                          |
|--------------|---------------------------------------------|-----------------------------------|
| A0A0240141    | Pseudoalteromonas_sp.                       | SLLFQANIPPVKLMHQLQRGLTSVLRLPLTHL  |
| A0A0240142    | Alteromonadales_bact                        | SLFANIPPVKLMHQLQRGLTSVLRLPLTHL    |
| A0A0240143    | Pseudoalteromonas_sp.                       | SLLFANIPPVKLMHQLQRGLTSVLRLPLTHL  |
| A0A0240144    | Aspergillus_oryzae                          | WLFCEPNPIAINTALMMTWGAVKPVFRLPYVPL |
| A0A0240145    | Ostreococcus_tauri                          | WLFCEPNPIAINTALMMTWGAVKPVFRLPYVPL |
| A0A0240146    | Oryza_sativa                                | WLFCEPNPIAINTALMMTWGAVKPVFRLPYVPL |
| A0A0240147    | Arabidopsis_thalialist                      | WLFCEPNPIAINTALMMTWGAVKPVFRLPYVPL |
| A0A0240148    | Pseudomonas_aeruginosa                      | SLFQANIPPVKLMHQLQRGLTSVLRLPLTHL  |
| A0A0240149    | Mycobacterium_avium                        | AMARLGGVTMSKAGLRLQGIDVGDPRLPQVPA  |
| A0A0240150    | Streptomyces_coei                           | GMFRTQGVMTTKGALALQGLPAGPLRAPMVGL |
| A0A0240151    | Bacillus_halodul                           | GLFTAPNPTCVKAALQGLPAGPLRAPMVGL   |
| A0A0240152    | Cyanidioschyzon_merolae                     | ALFVMANPIPAINTALMLGVPNVLQLRPLPQ  |
| A0A0240153    | Crocosphaera_watson                         | VLFCTSNPIPAINTALMLGVPNVLQLRPLPQ  |
| A0A0240154    | Prochlorococcus_marinus                     | SLFATTNPIPVKAAITLMLGVPNVLQLRPLPQ |
| A0A0240155    | Gloeobacter_violace                         | GLFWEANIPPVKAAITLMLGVPNVLQLRPLPQ |
| A0A0240156    | Cytophaga_hutchi                           | LMYEEQGNPVGKSLMLMGCSSEVRLPQLVA   |
| A0A0240157    | Bacteroides_fragil                          | LLFVDGNPAGVKSMLNAMGDIENKRPLAPV   |
| A0A0240158    | Thermotoga_maritil                         | ALFVETNPIPAINTALMLGVPNVLQLRPLPQ  |
| A0A0240159    | Haloarcula_marism                          | AMFVETNPIPAINTALMLGVPNVLQLRPLPQ  |
| A0A0240160    | Methanosarcina_acetic                      | ALFLETNPIPAINTALMLGVPNVLQLRPLPQ  |
| A0A0240161    | Chlorobium_limicola                        | LNFIESNVPVYKALMLGMIEEVRPLPVKL    |
| A0A0240162    | Acidobacteria_bacter                        | ANFLESNPGVXAMMKRFEYNNRPLPVKL     |
| A0A0240163    | Rhodopirellla_baltica                      | KMLLATNPIPAKAMQNGQTVDTEMLRPLPVKL|
| A0A0240164    | Brucella_suis                              | ALFLEPNPVSQTKALGRLKHIENLRPLPVKL  |
| A0A0240165    | Caulobacter_sp.                            | ALFLEPNPVSQTKALGRLKHIENLRPLPVKL  |
| A0A0240166    | Leptospira_interr                          | TFLATNPIPAKAMQNGQTVDTEMLRPLPVKL  |
| A0A0240167    | Clostridium_diffic                         | ALFIEQNPVTKALMLGMIEEVRPLPVKL     |
| A0A0240168    | Guillardia_theta                           | AMFCTENPIPAKAMQNGQTVDTEMLRPLPVKL |
| A0A0240169    | Euglena_gracilis                           | AMFCTENPIPAKAMQNGQTVDTEMLRPLPVKL |
| A0A0240170    | Aquifex_aeolic                             | VLFIEQNPVTKALMLGMIEEVRPLPVKL     |
| A0A0240171    | Geobacter_uranii                           | AMFIEQNPVTKALMLGMIEEVRPLPVKL     |
Figure 3. Molecular phylogeny of dihydrodipicolinate synthase (dapA). The red algal Cyanidioschyzon sequence forms a group with cyanobacterial and Dehalococcoides homologs with strong support whereas green algal and glaucophyte sequences form another strongly supported group with gamma-proteobacterial homologs. This could be explained by an HGT from gamma-proteobacteria and subsequent gene losses among primary photosynthetic eukaryotes. See text for explanations. Sequences of Cyanophora, Guillardia, Hartmanella, and Euglena were obtained from TBestDB. P-value = 0.199 from the AU test for the presented tree.

AU tests were also performed on alternative topologies, including (A) monophyly of all eukaryotic sequences except Cyanidioschyzon that appears to be plastid-derived, (B) monophyly of Cyanidioschyzon and cyanobacterial sequences, and (C) monophyly of all plastid-containing eukaryotic and cyanobacterial sequences. P-values < 0.001 from AU test for topologies A and C, suggesting that green plant and glaucophyte sequences are unlikely of the same origin (mitochondrial or plastidic origin) with other eukaryotic sequences. P-value = 0.801 from AU test for topology B, confirming the likely plastidic origin of the Cyanidioschyzon sequence.
4. ThiC family protein (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Species | Sequence |
|-----------|---------|----------|
| 15922177 | Sulfolobus tokoda | MAIIDDARKG II I TDEM KKKKL SIEK LEKISEK KPVK RIVEK GKM I RNEK YS K |
| 18311872 | Pyrobaculum aeroph | VVAIAG GLKSTI NVL GTS TEEV DLEK SVKVEVNA KGK DTVMD SG |
| 15789982 | Halobacterium sp. | NTTIIR AEGR KEKDMK IERAEGSKE KLEKRAKE KQA VYHRNWS K |
| 34541691 | Porphyromonas gingiv | NTTIIR AEGR KEKDMK IERAEGSKE KLEKRAKE KQA VYHRNWS K |
| 48766428 | Rhodospirillum rubrum | PTQLY ARAQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 71863643 | Pelobacter propio | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 67938921 | Chlorobium phaeob | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 54072970 | Nocardia farcin | RQCLAR QIVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 45658515 | Leptosira intern | MQHFMKEAKG I EYAR N FE Y RD E VAI R GAI RAINPinHEE |
| 67977524 | Thermococcus geoth | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 73593919 | Halobacterium sp. | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 19236156 | Arabidopsis thalia | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| CP000022869 | Cyanophora paradox | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 32454237 | Rhodoopisellula baltic | VTMQIYARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 67930575 | Solibacter usitad | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 15605979 | Aquifex aeolic | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 33241220 | Prochlorococcus marin | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 71635503 | Synnechocystis sp. | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 51758962 | Thermococcus kodakla | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 18977903 | Pyrococcus furios | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 71831809 | Pelobacter propio | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 19712998 | Fusobacterium nuclea | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |
| 67875148 | Clostridium thermod | VTQLYAARQG IVTDMV VAINENRTPFEVQVGEAVGAI RAINPinHEE |

5. ThiC family protein (E)

| Accession | Species | Sequence |
|-----------|---------|----------|
| 16331858 | Synechocystis sp. | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 149920854 | Plesiocystis pacific | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 48855688 | Cytophaga Hutchi | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 26250765 | Escherichia coli | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 26991600 | Pseudomonas putida | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 5402178 | Nocardia farcin | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 67938921 | Chlorobium phaeob | VTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 71836834 | Pelobacter propio | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 48764428 | Rhodospirillum rubrum | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 67938921 | Chlorobium phaeob | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |

6. ThiC family protein (F)

| Accession | Species | Sequence |
|-----------|---------|----------|
| 15922177 | Sulfolobus tokoda | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 18311872 | Pyrobaculum aeroph | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 15789982 | Halobacterium sp. | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 34541691 | Porphyromonas gingiv | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 48766428 | Rhodospirillum rubrum | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 71863643 | Pelobacter propio | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 67938921 | Chlorobium phaeob | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 54072970 | Nocardia farcin | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 45658515 | Leptosira intern | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 19712998 | Fusobacterium nuclea | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |

7. ThiC family protein (G)

| Accession | Species | Sequence |
|-----------|---------|----------|
| 15922177 | Sulfolobus tokoda | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 18311872 | Pyrobaculum aeroph | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 15789982 | Halobacterium sp. | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 34541691 | Porphyromonas gingiv | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 48766428 | Rhodospirillum rubrum | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 71863643 | Pelobacter propio | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 67938921 | Chlorobium phaeob | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 54072970 | Nocardia farcin | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 45658515 | Leptosira intern | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| 19712998 | Fusobacterium nuclea | YTSQGHRARE GRIDDE MRKIA E XEKKERK RYRERSTV AAVKAK RAINPinHEE |
| Gene Name               | Accession     |
|------------------------|---------------|
| Plesiocystis pacifica  | ASMCGHFCSMKLTQDV |
| Ralstonia eutropica    | ASMCGKFCSMKITQEV |
| Deinococcus geotherma  | ASMC           |
| Leptospira inter media| ASMCGHFCSMNLTQEL |
| Nocardia farcinii      | ASMCGKFCSMRISADV |
| Chlorobium phaeo bob   | GMTMCGDFCSMKKSQEV |
| Pelobacter propionic   | DSMCGDFCAMRITRNI |
| Rhodospirillum rubrum  | HSMCGKFCSMKIS |
| Porphyromonas gingival | HSMCGHFCSMRANKNF |
| Halobacterium sp.      | YSMCGDFCSMRIDQDA |
| Pyrobaculum aerophilus | KTMCGGYCPMNMVIQQ |
| Clostridium thermo taxon | DIAKGIKGAREWDYQMSEARRNLDWNRMFELAIDREKAERYRKSSMPEDE |
| Fusobacterium nucleatum | DLAKGHPGAQVRDDALSKARFEFRWDQFALSDFLTDPRQTHERDETMPEA |
| Sulfolobus tokodai     | HITSAIGGAIAAAAGADFLCYVTPAEHLGLPDVQDVREDVQRFAVIAKIAAHAV |
| Deinococcus geotherma  | DLAKGHPGAQVRDDALSKARFEFRWDQFALSDFLTDPRQTHERDETMPEA |
| Rhodopirellula baltica | HIAGAIGGALAGWAGAAMLCYVTPKEHLGLPNEDVRMTREYHDATLPADG |
Note: All top hits in GenBank searches (using Cyanidioschyzon sequence and Arabidopsis GI 22136156 as queries) are from proteobacteria, firmicutes and spirochaetes. Primary photosynthetic eukaryotic sequences share many conserved residues with non-cyanobacterial sequences, and likely are not of cyanobacterial origin. The same is also supported by the phylogenetic analyses. The Arabidopsis sequence (GenBank accession number NP_180524 and TAIR locus AT2G29630) is annotated as a chloroplast precursor. The glaucophyte Cyanophora sequence was obtained from TBESTDB.
**Figure 4.** Molecular phylogeny of ThiC. P-value = 0.539 from AU test for the presented tree. AU tests were also performed on alternative topologies, including (A) monophyly of red algal, green plant and cyanobacterial sequences, and (B) monophyly of red alga, green plant and archaeal sequences. These tests investigate if red algal and green plant sequences has a plastidic or an archaeal (or eukaryotic) origin. P-values < 0.001 from AU tests for both alternative topologies.

5. Diaminopimelate decarboxylase (lysA) (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Organism                          | Sequence                                              |
|-----------|-----------------------------------|-------------------------------------------------------|
| 66825991  | Dicyostelium discoi               | TPFHIYNGDIKKEGELLMNYFAYKATPNPSILKLLKEIGMKGVDCSSLAE   |
| 24213747  | Leptospira_interr                 | TPVFYVRERIEKCSVEALVRYAMKANPNRTVEIMKRKGIGIDASESEYE    |
| 116056027 | Ostreococcus tauri                 | TPTYVYDIAPTEARAAAVKVRAMKAPNNAIKFKEKLHVDAATAGE       |
| 87306324  | Blastopirellula_marin             | TPFYYDAKVERIEDLRIRIAYAQRKACSNILDLMLMKEGIVDAYSAME    |
| 21224744  | Streptomyces_coelic               | TPLYYLSQVQRGQQLAYACKANTNAIKLMEKGGIDAVIGE             |
| 11835090  | Tetrahymena_thermo                 | TPLYVYDEIIKRQQTITVLYACKANTNNAIKLMEKGGIDAVIGE         |
| 145525334 | Paramecium_tetrau                 | TPLYYVSVQIKERCQQLAYACKANTNAIKLMEKGGIDAVIGE           |
| 48477088  | Picrophilus_torrid                | TPLITYVSQIKERCQQLAYACKANTNAIKLMEKGGIDAVIGE           |
| 126007985 | Ferrobacillus_marina              | TPFVYNMARVREIIRIRYAVSHYKSNPFIYVSQIKEETGIADAVINE      |
| 60682355  | Bacteroides_fragil                | TPFYYDTKVRLDTCLVKHNYVAKNANPKVILTIREGLGADYSGE         |
| 123416580 | Trichomonas_vagina                | TPLYYVSETQTVSNFTRYVISFAYKANTNLAFTLSSKLGCADYSAGE      |
| 45358763  | Methanococcus_mariposa            |-------------------------------------------------------|
Note: All top hits in GenBank searches (using *Cyanidioschyzon* sequence and *Arabidopsis* GI 15231844 as queries) are from various non-cyanobacterial groups. The donor of the acquired gene in primary photosynthetic eukaryotes (upper part of the tree) is difficult to pinpoint because of the lack of sufficient internal support on the gene tree, but it is unlikely from cyanobacteria based on AU tests. The *Arabidopsis* sequence (GenBank GI 15231844 and TAIR locus AT3G14390) is annotated as a chloroplast precursor.

**Figure 5.** Molecular phylogeny of lysA. See text for detailed explanation. P-value = 0.235 from AU tests for the presented tree. AU tests were also performed on alternative topologies, including (A) monophyly of all eukaryotic sequences, (B) monophyly of cyanobacterial and primary photosynthetic eukaryotic sequences (except for the *Ostreococcus* sequence that appears to be distant). These tests investigate if red algal, green plant and glaucophyte sequences have the same origin (mitochondrial or eukaryotic) with other eukaryotic sequences and if red algae, green plants and glaucophytes acquired this gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

---

6. MGDG synthase (N)

CLUSTAL X (1.83.1) multiple sequence alignment
Figure 6. Molecular phylogeny of MGDG synthase. See text for detailed explanation. P-value = 0.235 from AU test on the presented tree. AU tests were also performed an alternative topology enforcing a monophyly of photosynthetic eukaryotic sequences with the rare cyanobacterial homolog (i.e., *Gloeobacter* in this case). P-value < 0.001 from AU test for the alternative topology.

7. Alpha amylase (D)

**CLUSTAL X (1.83.1) multiple sequence alignment**

```
CM1405_Cyanidioschyzon_merolae      YGCPTVNYAIYSTEQRHMSLIIYRMKGVKHAVAQAGYRMYRVRGKG-
33635247_Prochlorococcus_marin     LGSTIGVNSVATATKAVELVLLFHRSGSVGVEVMETGCGYFRGFPH-
149177467_Plantomycyes_maris      LGETWYMNFIYSKHEASVSLFFKNTAEWKCRIPPAYYQIEGPSD
32390865_Rhodopirellula_baltic     LGASWGFNFSYRHTAVHLHLYNKSVPWCRVREWAYAIYRVDGAPD
33236242_Chlamydophila_pneumo     LASKRVRɣLAMYQATSEVEILALMRTGAIWHIIEQGSYAFRNGF-S
42525620_Treponema_dentic        LAGKLVNFSVSNRKPEVHLHFTKGTGWHVVFNSWAFYLTADGFD
19714344_Fusobacterium_nuclea     LGANLGFSYALYANLVNSLSQFFHKLGDFWIFLEEGTVLWY-
150017328_Clostridium_bijer       LGASQGINFAVFSQATSCODELFGTVHFLKGYRFDGKFD
48894542_Trichoscladium_eurthr     FGTAGYNFNSIFSSYTTSTCVLFLFGRIGNYCHVFENIEYRGMDEN
16330244_Synechocystis_sp.        LGATIGVNSVATATKAVELVLLFHRSGSVGVEVMETGCGYFRGFPH-
CM4050_Cyanidioschyzon_merolae     LGVSLGPMNFIYSTQATSEVEILALMRTGAIWHIIEQGSYAFRNGF-
22328517_Arabidopsis_thalia       LQDSQINFAVNTSQSTCDELFAGTWHFLKGYRFDGKFD
15225595_Arabidopsis_thalia       LGASWGFNFSYRHTAVHLHLYNKSVPWCRVREWAYAIYRVDGAPD
46446740_Parachlamydia_sp.        FMRPNKYNFLYKNIKESLCLNTGVNTHAIYEPYLYAFR-
1166210_Solibacter_usitatus       LGATGNFNSYKQAAGVFLLRPFARDKLWBCAVRQAGQYKVGSHEN-
48785636_Rhodospirillum_rubrum     LQAMAGVENAVFSEARQIDCLCLARTDGIWGLFLPPGLVGLRAGFY-
24370369_Shewanella_oneida        LGATGNFNSYKHAYKESLCLNTGVNTHAIYEPYLYAFR-
1707700_Sulfofolobus_solfat       LGSNWGNFSYKSAENKELVLLKNTGDKVQPYGLQAYRIGFVY-
48784580_Burkholderia_fungor       LGASWGFNFSYRHTAVHLHLYNKSVPWCRVREWAYAIYRVDGAPD
17548459_Ralstonia_solana         LGANLTVGFVNSYRRTAVHLHLYNKSVPWCRVREWAYAIYRVDGAPD
118048461_Chloroflexus_aggrega     LGATGNYVAFYSAHKVRELHFSAVIMRTAVIDCLFCDTVEWGLFLPPGLVGLRAGFY-
```

25028566_Corynebacterium_efficacis
108803168_Rubrobacter_xylanophilus
64576111_Streptomyces_coelicoflavus
25028566_Corynebacterium_efficacis
32398065_Rhodopirellula_baltica
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
143186879_Pseudomonas_putida
116622006_Solibacter_usitatus
46446740_Parachlamydia_sp.
33635247_Picrobacter_maris
33236242_Chloroflexus_agregatus
46446740_Parachlamydia_sp.
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
16329290_Synechocystis_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622206_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
35430870_Cyanidioschyzon_merolae
16329290_Synechocystis_sp.
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
48894542_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
21223814_Streptomyces_coelicoflavus
116622006_Solibacter_usitatus
48893487_Trichodesmium_erythraeum
149177467_Planctomyces_maris
16330244_Synechocystis_sp.
Note: Top hits in GenBank searches (using Cyanidioschyzon and Arabidopsis sequences) are from chlamydiae and cyanobacteria. The Arabidopsis sequences are annotated as chloroplast precursors.
Figure 7. Molecular phylogeny of alpha amylase. P-value = 0.983 from AU test for the presented gene tree. AU tests were also performed on alternative topologies, including (A) monophyly of red algal and green plant sequences with different copies of cyanobacterial homologs, and (B) monophyly of red algal, green plants, cyanobacterial, and chlamydial sequences. These tests investigate if red algal and green plant sequences are plastid-derived and if red algae and green plants acquired the gene from plastids and subsequently spread to chlamydiae. P-values < 0.001 from AU tests for scenarios A whereas the P-value = 0.017 for scenario B.

8. 3-dehydroquinate synthase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Organism | Sequence |
|-----------|----------|----------|
| 45508724 | Anabaena variabilis | IELRFFITITEFDTIKTKQRSSRFVADFKVKEPVLVVGGLITDVV |
| KM000293952 | Karlodinium | IKFTRKFLVSGNEADKDIORDVERILVAKLEIGLARNEPLLVVGGGIADIA |
**Note:** All top hits in GenBank searches (using *Cyanidioschyzon* and *Arabidopsis* sequences as queries) are from gamma-proteobacteria. Phylogenetic analyses show that red algal and green plant sequences group with beta and gamma-proteobacterial homologs with strong support. The *Arabidopsis* sequence (GenBank accession number AAM98284 and TAIR locus AT5G66120) is annotated as a chloroplast precursor. Sequences of *Euglena* and *Karlodinium* were obtained from TBestDB.

**Figure 8.** Molecular phylogeny of 3-dehydroquinate synthase. P-value = 0.997 from AU test on the presented tree. AU tests were also performed on alternative topologies, including (A) monophyly of all eukaryotic sequences (except for *Karlodinium*, which appears to be of cyanobacterial origin), (B) monophyly of cyanobacterial, green plant...
and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

9-10. 2-methylthioadenine synthetase (E, D)

CLUSTAL X (1.83.1) multiple sequence alignment

and red algal sequences. These tests investigate if red algal and green plant sequences have the same origin (mitochondrial or eukaryotic) and if they acquired the gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.
| Accession | Organism Name                  | Abbreviation |
|-----------|--------------------------------|--------------|
| 15836008  | Chlamydia pneumoniae           | GLIGE        |
| 24213088  | Leptospira interrata           | DKEGT        |
| 66576257  | Chlorobium tepidum             | QLHGR        |
| 19705459  | Porphyromonas gingivalis       | --SGE        |
| 12663485  | Flavobacterium bacterium       | SVRLQ        |
| 11063813  | Cytophaga hutchii              | LVEIE        |
| 32477342  | Rhodopirellula baltica         | RLELA        |
| 16079957  | Bacillus subtilis              | YNEQG        |
| 118444753 | Clostridium novyi               | YAVGK        |
| 15605955  | Aquiex aeolic                  | --VGK        |
| 32475869  | Rhodopirellula baltica         | TLIGE        |
| 76261232  | Chloroflexus auranti           | SLQGV        |
| 94905469  | Deinococcus geothermalis       | MLYGR        |
| 116625754 | Solibacter usitats             | SLVGE        |
| 66807387  | Dictyostelium discoidei        | TLKGR        |
| 118402077 | Tetrahymena thermophilus       | SLFCT        |
| 116061814 | Ostreococcus tauri             | TLVAE        |
| CM1685    | Cyanidioschyzon merolae        | RLIKE        |
| 24213805  | Leptospira interrata           | TLKGR        |
| Jakaba bahamensis                       | TLRRAE       |
| 28872782  | Homo sapien                    | TLRGH        |
| 21674804  | Chlorobium tepidum             | TLIGE        |
| 30690642  | Arabidopsis thaliana           | SLFGE        |
| 34540755  | Porphyromonas gingivalis       | TLIGE        |
| 126663275 | Flavobacterium bacterium       | TLIGE        |
| 110638221 | Cytophaga hutchii              | TLIGE        |
| 119885210 | Thermotoga petrop              | PLYGK        |
| 108759982 | Myxococcus xanthus             | QLAGK        |
| 34581161  | Rickettsia sibirica            | SLIGE        |
| 94417083  | Pseudomonas aerugine           | SLIGE        |
| CM4402    | Cyanidioschyzon merolae        | SLIGE        |
| 33862596  | Prochlorococcus marinus        | SLIGE        |
| 16329745  | Synechocystis sp.              | SLIGE        |
| 11225071  | Frankia alni                   | HLTAD        |
| 19703810  | Fusobacterium nucleatum        | TLIGE        |
| 15605821  | Aquiex aeolic                  | SLIGE        |
| 118443818 | Clostridium novyi              | SLIGE        |
| 51892897  | Symbiobacterium thermophilus   | TLEGRE       |
| 16078764  | Bacillus subtilis              | SLIGE        |
| 71075619  | Giardia lama                   | ALIGS        |
| 13542169  | Thermoplasma volcanii          | SLIGE        |
| 71029160  | Theileria parva                | HLECR        |
| 71661088  | Trypanosoma cruzi              | SVVGR        |
| 93277076  | Homo sapien                    | FMKGQ        |
| 118373032 | Tetrahymena thermophilus       | HIEGE        |
| 116056662 | Ostreococcus tauri             | SCKAR        |
| 18409989  | Arabidopsis thaliana           | SVFGE        |
| 121522286 | Methanococcococcus maripal      | GLSGK        |
| 18978284  | Pyrococcus furios              | YLLGE        |
| 15920732  | Sulfolobus tokodai             | DLRGE        |
| 119871859 | Pyrolobus island                | YLYGE        |
Figure 9. Molecular phylogeny of 2-methylthioadenine synthetase. P-value = 0.983 from AU test for the presented tree. See text for a more detailed discussion. AU tests were also performed on alternative topologies, including (A) miaB1 and miaB2 forming a monophyly, (B) miaB1 and miaB2 forming a monophyly that in turn groups with archaeal sequences, and (C) miaB2 forming a monophyly with proteobacterial sequences from the top part of the tree. These tests investigate if (a) miaB1 and miaB2 have the same origin, (b) miaB1 and miaB2 have an eukaryotic origin and are related to archaeal homologs, and (c) miaB2 has a mitochondrial origin. P-values < 0.001 from AU tests for these alternative topologies.

11. Uroporphyrinogen III synthase (D)
**Note:** This sequence has identifiable homologs only in green plants and bacteria (using *Cyanidioschyzon* and *Arabidopsis* GI 20196944 as queries). The top hits in GenBank searches are from firmicute and *Deinococcus* homologs. The *Arabidopsis* sequence (GenBank accession number NP_565625 and TAIR locus AT2G26540) is annotated as a chloroplast precursor, but appears not likely to be plastid-derived. Cyanobacterial sequences have a 1-aa insertion not shared by any other sequences.

**Figure 11.** Molecular phylogeny of Uroporphyrinogen-III synthase. P-value = 0.959 from the AU test on the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of red algal, green plant and cyanobacterial sequences, and (B) red algal, green plant and *Deinococcus* sequences forming a monophyly that in turn groups with cyanobacterial homologs. These tests investigate if a) red algae and green plants acquired this genes from plastids, and b) red algae and green plants acquired the genes from plastids and subsequently spread to *Deinococcus* by secondary HGT. P-value < 0.001 from AU test for topology A and p-value = 0.04 for topology B.
Note: Identifiable homologs of this gene (using Cyanidioschyzon sequence and Arabidopsis GI 9758449 as queries) are restricted to red algae, green plants, proteobacteria and firmicutes with low sequence similarities (Evalue e-4 and lower). The annotation of the
Arabidopsis sequence (TAIR locus AT5G04740) indicates that the protein is located on the chloroplast thylakoid membrane.

Figure 12. Molecular phylogeny of ACT-domain containing protein. Since no other eukaryotic or cyanobacterial homologs were identified, no statistical test on alternative topologies was given.

13. Queuine tRNA-ribozymyltransferase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Organism                     | Sequence                                                                 |
|-----------|------------------------------|--------------------------------------------------------------------------|
| 42520578  | Wolbachia_endosy             | SGSARVGTIKTPNGSVETPAFIFCAPATKAA1KAADIERISTRGQ1LNSNTYH                   |
| 58613533  | Heterocapsa_trique           |--------------------------------------------------------------------------|
| 45656892  | Leptospira_interr            |----------------------------------------------------------------------------|
| 19068923  | Encephalitozoon_cunic       |----------------------------------------------------------------------------|
| 3881825   | Caenorhabditis_elegans      |----------------------------------------------------------------------------|
| 56473322  | Entamoeba_histol             |----------------------------------------------------------------------------|
| 46229743  | Cryptosporidium_parvulus    |----------------------------------------------------------------------------|
| 12597314  | Homo_sapiens                |----------------------------------------------------------------------------|
| 60463331  | Dicyostelium_discoi         |----------------------------------------------------------------------------|
| 62360604  | Trypanosoma_brucei          |----------------------------------------------------------------------------|
| 50900348  | Oryza_sativa                |----------------------------------------------------------------------------|
| JBO00061016 | Jakoba_bahamensis         |----------------------------------------------------------------------------|
| 28850382  | Dicyostelium_discoi         |----------------------------------------------------------------------------|
| 116059209 | Ostreococcus_tauri          |----------------------------------------------------------------------------|
| 2419163   | Cyanidioschyzon_merolae     |----------------------------------------------------------------------------|
| 46444428  | Parachlamydia_sp            |----------------------------------------------------------------------------|
| 76788915  | Chlamydia_trachomat            |----------------------------------------------------------------------------|
| 29840339  | Chlamydomena_caviae          |----------------------------------------------------------------------------|
| 11499080  | Archaeoglobus_fulgid        |----------------------------------------------------------------------------|
| 21674218  | Chlorobium_tepidum          |----------------------------------------------------------------------------|
| 48855228  | Cytophaga_hutchin           |----------------------------------------------------------------------------|
| 53713592  | Bacteroides_fragsil         |----------------------------------------------------------------------------|
| 76258015  | Chloroflexus_auranti        |----------------------------------------------------------------------------|
| 15644309  | Thermotoga_maritii          |----------------------------------------------------------------------------|
| 6460406   | Deinococcus_radiod          |----------------------------------------------------------------------------|
| 48891133  | Trichodesmium_erythrophy    |----------------------------------------------------------------------------|
| 56751072  | Synechococcus_elongata      |----------------------------------------------------------------------------|

Arabidopsis sequence (TAIR locus AT5G04740) indicates that the protein is located on the chloroplast thylakoid membrane.

Figure 12. Molecular phylogeny of ACT-domain containing protein. Since no other eukaryotic or cyanobacterial homologs were identified, no statistical test on alternative topologies was given.

13. Queuine tRNA-ribozymyltransferase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Organism                     | Sequence                                                                 |
|-----------|------------------------------|--------------------------------------------------------------------------|
| 42520578  | Wolbachia_endosy             | SGSARVGTIKTPNGSVETPAFIFCAPATKAA1KAADIERISTRGQ1LNSNTYH                   |
| 58613533  | Heterocapsa_trique           |--------------------------------------------------------------------------|
| 45656892  | Leptospira_interr            |----------------------------------------------------------------------------|
| 19068923  | Encephalitozoon_cunic       |----------------------------------------------------------------------------|
| 3881825   | Caenorhabditis_elegans      |----------------------------------------------------------------------------|
| 56473322  | Entamoeba_histol             |----------------------------------------------------------------------------|
| 46229743  | Cryptosporidium_parvulus    |----------------------------------------------------------------------------|
| 12597314  | Homo_sapiens                |----------------------------------------------------------------------------|
| 60463331  | Dicyostelium_discoi         |----------------------------------------------------------------------------|
| 62360604  | Trypanosoma_brucei          |----------------------------------------------------------------------------|
| 50900348  | Oryza_sativa                |----------------------------------------------------------------------------|
| JBO00061016 | Jakoba_bahamensis         |----------------------------------------------------------------------------|
| 28850382  | Dicyostelium_discoi         |----------------------------------------------------------------------------|
| 116059209 | Ostreococcus_tauri          |----------------------------------------------------------------------------|
| 2419163   | Cyanidioschyzon_merolae     |----------------------------------------------------------------------------|
| 46444428  | Parachlamydia_sp            |----------------------------------------------------------------------------|
| 76788915  | Chlamydia_trachomat            |----------------------------------------------------------------------------|
| 29840339  | Chlamydomena_caviae          |----------------------------------------------------------------------------|
| 11499080  | Archaeoglobus_fulgid        |----------------------------------------------------------------------------|
| 21674218  | Chlorobium_tepidum          |----------------------------------------------------------------------------|
| 48855228  | Cytophaga_hutchin           |----------------------------------------------------------------------------|
| 53713592  | Bacteroides_fragsil         |----------------------------------------------------------------------------|
| 76258015  | Chloroflexus_auranti        |----------------------------------------------------------------------------|
| 15644309  | Thermotoga_maritii          |----------------------------------------------------------------------------|
| 6460406   | Deinococcus_radiod          |----------------------------------------------------------------------------|
| 48891133  | Trichodesmium_erythrophy    |----------------------------------------------------------------------------|
| 56751072  | Synechococcus_elongata      |----------------------------------------------------------------------------|
| Genus               | Species                          | Sequence     |
|---------------------|----------------------------------|--------------|
| Helicobacter        | pylori                           | SLHNLHFYLELVKNARNAI |
| Bdellovibrio        | bacter                            | TIHNIHFYMKVMEKAREAIAQGRW |
| Rubrobacter         | xylano                           | SLHNVRFVTELCRSARREILAGTY |
| Nostoc              | sp                               | SIHNITELIRFTQKIREAILSDRF |
| Synechococcus       | elongata                         | SIHNITE       |
| Trichodesmium       | erythra                          | SLHNVTELISFTQRIRDAILKDRF |
| Deinococcus         | radiod                           | SLHNLRYLHRLVERMRVAINGQQF |
| Thermotoga          | maritii                          | TIHNINFMISLMKEVRRSIESGTF |
| Bacteroides         | fragil                           | SLHNLAFYLWLVGEARKHIIAGDF |
| Cytophaga           | hutchi                           | TIQNISFYLWLMREARKQIVAGTF |
| Chlorobium          | tepidu                           | TMQNLSFYLWLTRTAREHIAAGDF |
| Archaeoglobus       | fulgid                           | TYHNIYFVVKLMERIRESIADGSF |
| Chlamydia           | trachomat                        | GMILPLKINNQYRSLLNPSCASLCAGQITRAYLHLFVKEHPPNAI |
| Chlamydia           | trachomat                        | GTALILHRSLFKASKEPIDAQRHCMVUCHEYMTGRHLQACED---GELI |
| Rickettsia          | prowaz                           | GQAFTVNIRNSKYADDKEPLEHDCKCPACTNYTKAYL |
| Silicibacter        | sp                               | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Rhodopseudomonas    | palau                            | GLAFPINLRNAKHADDPRPLDEESDWPSARTYSRAYLHHLVRSSETLMLL |
| Leptospira          | interrup                         | GTCMRLVIKNAKFTHDFRPIDENCDCYTCKKNYSRAYLHLRMSGKCEILRLN |
| Bacillus            | clausi                           | GTCMRLVVRNAKYARDFSPLDEKCDCHVCRTYSRAYIRHLIKVQEVLRLT |
| Pseudomonas         | fluores                          | ATLFKIRLTHRNYRDFYVPDCTYCYTRCTNYRAYLHLFVNEIATLL |
| Geobacter           | sulfur                           | GTLYVYDIRHKSEFSDPLPDECDYTCRNFNSKAYLHLFVAEESVLN |
| Cyanidioschyzon     | merolae                          | GTLLRIQIKSRKWERVFEPVDTTCEGFVSQNHLYTAYHLHLVHNEMILGA |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Silicibacter        | sp                               | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
| Cytophaga           | hutchi                           | GMLFIINIRNKKWFVTDPDQLDYLNDYTYKAYHRVLHNVNMELQA |
| Chlorobium          | tepidu                           | GTLYVVDIRHSKWKEDFSPLDPECDYTCKNFTRSYIHHLFVAGKELTLL |
| Archaeoglobus       | fulgid                           | GTALKLNLKASYNKRSLEPVDERGCGTCKNTRTSYHLHLFVAGKELTLL |
| Rickettsia          | prowaz                           | GQAFVVNIKNARHQDDPRPLDENCSCPACSNYSRAYLHHVFRSNEMIMLL |
| Thermotoga          | maritii                          | GMFLFVNRMLRDDRQPLQECPCYTCHRSAYLHLFVKEHPPNAI |
| Dicystostelium      | discoi                           | TQHNLRAMSRLMENYRNLKQEVL |
| Bacteroides         | fragil                           | GAALRWNLKNAQFREDQPLDECDNCYCyCNQFSRAYLHLREISITLL |
Note: This is an intriguing case of ancient HGT. The Dictyostelium sequence forms a group with homologs of green and red algae as well as chlamydiae. This specific affiliation of Dictyostelium and plant sequences has been observed in multiple cases (Huang, unpublished data), likely resulting from plant-Dictyostelium transfer. A plausible explanation is that primary photosynthetic eukaryotes acquired this gene from chlamydiae and then further spread to Dictyostelium via secondary gene transfer. Heterocapsa sequence is a chloroplast precursor based on the original GenBank annotation. Heterocapsa and Wolbachia sequences also share indels and many conserved residues. The remaining eukaryotic sequences are much more similar to bacterial than to archaeal homologs, and they are likely of bacterial origin. One possible explanation is that they are derived from mitochondria. Nevertheless, most of these eukaryotic sequences lack a N-terminal extension. Sequences of Trimastix and Jakoba were obtained from TBestDB.

Figure 13. Molecular phylogeny of queuine tRNA-ribosyltransferase. P-value = 0.369 from the AU test for the presented gene tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of red algal, green...
14. SAM-dependent methyltransferase (N)

**CLUSTAL X (1.83.1) multiple sequence alignment**

| Accession | Organism                  | Sequence                                                                 |
|-----------|---------------------------|--------------------------------------------------------------------------|
| 2350973   | Plasmodium falciparum     | INGFNWLSKDKIEYENENIGVGYRNSISISSLFEFFNIR1KRAYEKLE                           |
| 16080887  | Bacillus subtilis         | KGYPILQKEDL6SLSKQARLGOEYESGKNNQGNNLSSFQCLFEEISREKTT                        |
| 29375834  | Enterococcus faecalis     | KQGGFLQICSERGLGCaBFCFLGAGYNEHSL1NYRVL5LITHR1QASL1RVRK                          |
| 54294134  | Legionella pneumo         | LCGHPWFPFKA2SKEG5GFLCGAGYNEHSL1NYRVL5LITHR1QASL1RVRK                          |
| 42571865  | Arabidopsis thalia        | KDSGMPWVQ5GZEPNG5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| CM2295    | Cyanidioschyzon merolae   | RQVRPLIFGRAITPFPVGLGFYNPD9MYRVLRLTFLPAWLRLEH                        |
| 42522494  | Bdellovibrio bacteri       | KQGHWPYFWL5CNEYEGN5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 59980234  | Geobacter sulfurreducens  | QJLGDWYPIFA5GFLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 34763641  | Rhodopirellula baltica    | LNFPN1KDE5ISDMLKQ5IARAAEQVLRK                                         |
| 6457710   | Deinococcus radiodurans   | RDGHPWYESSVRDFLAI5GFYDPSPRLRFVWLRDAAALARRAT                               |
| 21110809  | Xanthomonas axonopodii    | RS6HWPQFKL6V6GEGN5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 53797508  | Chloroflexus aurantii     | QCGHPWYNLR5CGLTVGL5DQ5ESPQG5IARAAEQVLRK                                         |
| 22960169  | Rhodobacter sphaericus    | RGHMPW5FVGGN5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 56679104  | Silicibacter pomerorum     | RGHAGPWF5VGRG5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 15650998  | Aquifex aeolicus          | KTHF5MPW5ORE5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 34564576  | Wolinella succinogenes     | RTLPAW5YL5GEGN5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 64546677  | Parachlamydia sp.         | RNYBHWP5GQ5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 53712322  | Bacteroides fragilis      | IQ6HPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 48584326  | Cytophaga cynctica        | NRHMPWS5FV6GNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 20094503  | Methanopyrus kandler       | RSGALSVP5AV5GNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 14591657  | Pyrococcus horikoshii     | K6GHVPFK5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 15607307  | Aquifex aeolicus          | K6GHVPFK5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 26246990  | Escherichia coli          | LRRHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 29142267  | Salmonella enterica       | LRRHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 47573089  | Rubrivivax gelatin        | LRRHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 48770361  | Rasoplasteria metallica   | LRRHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 48786349  | Burkholderia fungorum     | LRRHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 21674038  | Chlorobium tepidum        | V5GLHMPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 15644445  | Thermotoga maritima       | SG6LMPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 39997720  | Geobacter sulfurreducens  | RSG6HPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 4619218   | Thermus thermus           | LSRHHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 48857855  | Clostridium thermus       | K6GHWP5FGR5GNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 37523934  | Gloeobacter violaceus     | KRTHHPW5FPSA5GGLGNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |
| 51894272  | Symbiobacterium thermosphaer | QAGHPWIFQ5GNSW5GCMGLNSVSVMCFVRLMLKQ5IARAAEQVLRK                          |

**Tests for both alternative topologies.**

- Values < 0.001 from AU tests for both alternative topologies.

Plant, *Dictyostelium*, and cyanobacterial sequences. These tests investigate if all euukaryotic sequences have the same origin (mitochondrial or eukaryotic) and if red algal and green plant sequences are likely derived from plastids. P-values < 0.001 from AU tests for both alternative topologies.
Note: This gene has identifiable homologs only in prokaryotes and plastid-containing eukaryotes (red algae, green plants, and apicomplexan Plasmodium) in our GenBank and TBestDB searches (using Cyanidioschyzon and Arabidopsis sequences as queries). Homologs of the gene are rarely found in cyanobacteria. All top hits are from gamma and beta-proteobacteria.
Figure 14. Molecular phylogeny of SAM-dependent methyltransferase. P-value = 0.921 from AU test for the presented gene tree. AU tests were also performed on alternative topologies including (A) monophyly of red algal, green plant and Plasmodium sequences, (B) monophyly of red algal, green plant, Plasmodium and cyanobacterial sequences, and (C) monophyly of red algal, green plant, Plasmodium, and archaeal sequences. Topology A investigates if red algal, green plant, and Plasmodium (which also has a relict plastid) sequences have the same origin. Topology B investigates if the three plastid-containing groups acquired the genes from plastids. Topology C investigates if these plastid-containing eukaryotic sequences have an archaeal or eukaryotic origin. P-value = 0.082 from AU test for topology A whereas p-values < 0.001 for topologies B and C.

15. Semialdehyde dehydrogenase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Species                     | Sequence                                                                 |
|-----------|-----------------------------|--------------------------------------------------------------------------|
| NU000054019 | Capsaspora owczar           | RVGLIGARGFTGGMVLVRLIDCHPMNSSTNGKFPITTEFPQLKPKENIDGW                    |
| 60466242   | Dictyostelium discoi        | RVGLIGARGFTGGMVLVRLIDCHPMNSSTNGKFPITTEFPQLKPKENIDGW                    |
| 32408531   | Neurospora crassa           | RVALIGARGTYQELIDSHPNRELGKLEGYNQKVSPEDVIDCW                               |
| 5155       | Saccharomyces cerevis       | RVALIGARGTYQELIDSHPNRELGKLEGYNQKVSPEDVIDCW                               |
| 2421478    | Leptosira interr            | EISILGAGGLTGKELLMFSRQKRESKLAGKTISETVEFSVFKKKVL                        |
| 21673935   | Chlorobium tepidum          | TVSVIGASGYSAGELVLRMKMFQAMTQARFTDNLTPQTPYSCTDV                        |
| 27360910   | Vibrio vulnif               | KTTIIGAGSGAELLMWSRTVEPAGTKMCSQKEVACELL                                 |
| 5711832    | Bacteroides fragil          | KAGIIGGAGTAGEELLMWSRTVEPAGTKMCSQKEVACELL                                 |
| 34558370   | Wolinella succin            | PAVVAGASGAGGELLWSRTVEPAGTKMCSQKEVACELL                                 |
| 47529651   | Bacillus anthra             | KVAVAGASGAGGELLWSRTVEPAGTKMCSQKEVACELL                                 |
| 38233757   | Corynebacterium diphtheriae | KVAIVGASGAGGELLWSRTVEPAGTKMCSQKEVACELL                                 |
| 29610427   | Streptomyces avermii        | KVAVAGASGAGGELLWSRTVEPAGTKMCSQKEVACELL                                 |
Note: This gene is found in green plants, red algae, fungi, Capsaspora and prokaryotes. All top hits in GenBank searches are from alpha-proteobacteria. Phylogenetic analyses also support an alpha-proteobacterial origin of the gene in red algae and green plants. Fungal and Capsaspora sequences are apparently different (with an about 500 aa N-terminal extension and many conserved residues shared between them) from red algal, green plant and prokaryotic sequences. The *Neurospora* sequence (GenBank accession number P54898) is experimentally determined to be a mitochondrial precursor. Red algal and green plant sequences share many conserved residues with alpha-proteobacterial homologs. The *Arabidopsis* sequences (GenBank accession numbers and TAIR loci NP_565461, AT2G19940 and NP_649993, AT2G19940 respectively) are annotated to be located in cytoplasm. Sequence of *Capsaspora* was obtained from TBestDB.

Figure 15. Molecular phylogeny of semialdehyde dehydrogenase. P-value = 0.981 from AU test for the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of cyanobacterial, red algal, green plant, and alpha-proteobacterial sequences. Topology A investigates if all eukaryotic sequences have the same origin (mitochondrial or eukaryotic) whereas topology B investigates if red algae and green plants acquired this gene from plastids and further spread to alpha-proteobacteria. P-value = 0.045 from AU test for topology A and p-value < 0.001 for topology B.

16. Dihydrodipicolinate reductase (dapB) (D)
Figure 16. Molecular phylogeny of dihydrodipicolinate reductase (dapB). P-value = 0.995 from AU test for the presented tree. AU test was also performed on an alternative topology that enforces a monophyly of
Desulfovoccus, cyanobacterial (i.e., Prochlorococcus since this gene is very rarely identified in cyanobacteria in our similarity searches), and photosynthetic eukaryotic sequences. P-value < 0.001 from AU test for the alternative topology. See text for more detailed explanation.

17. Leucyl-tRNA synthetase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Sequence Name                      | Alignment |
|------------------------------------|-----------|
| _Dictyostelium discoi_             | FYSLSQFPYPSGA-LHMGHVRVYTISDCAIRLRKMQGVDYIDHPWYDAGF |
| _Homo sapien_                      | FYYLVMFPYPSGK-LHMGHVRYTISDIARQFMRQMGVQINHPWYDAGF |
| _Mycobacterium_leprae_             | LFVQDMFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |
| _Ostreococcus tauri_               | LFVQDMFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |
| _Rhodobacter sphaeroides_          | YYVLSQFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |
| _Rhodospirillum rubrum_             | YYVLSQFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |
| _Thermotoga_maritima_              | YYVLSQFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |
| _Nexisseria meningitidis_          | YYVLSQFPYPSGAGDGHVQHGLYIAUSDYAYKQRKQNFVYDSFAG |

< 0.001 from AU test

---

Since this gene is
Rhodobacter sphaeroi

Deinococcus radiodurans

Helicobacter pylori

Ostreococcus tauri

Bacillus halodurans

Victivallis vadensis

Salinibacter ruber

Blastopirellula marin

Cyanidioschyzon merolae

Treponema denticola

Pavlova lutheri

Chlamydia pneumonitis

Clostridium phytofermentans

Mycobacterium leprae

Halobacterium sp.

Tetrahit yessoensis

Homo sapiens

Dictyostelium discoideum

Geobacter metallireducens

Parachlamydia sp.

Rhodospirillum rubrum

Chlamydia trachomatis

Bacteroides fragilis

hymenaea thermoformis

Escherichia coli

Neisseria meningitidis

Nitrosomonas europa

Escherichia coli

Anaerobacter anginosus

* ***...* ; * ; * ; **; ; **; ** **:* *:  :           **   :  ** :   .:*.*
60463472_Dicyostelium_discoi
61963964_Homo_sapien
89306112_Tetrahymena_thermophila
15791045_Halobacterium_sp.
11595633_Neurospora_crassa
53715865_Bacteroides_fragilis
78678931_Chamydia_trachomatis
32325997_Chamyphila_pneumoniae
46446681_Parachlamydia_sp.
503072_Cyanobacteria
87311335_Blastopirellula_marina
42527840_Treponema_denticola
21674468_Chlorobium_tepidum
15826893_Mycobacterium_leprae
10175904_Bacillus_halodurans
116058126_Ostreococcus_tauri
73351135_Blastocystis
83816700_Salinitabacter_ruber
48846878_Geobacter_metallireducens

Note: All top hits in GenBank searches (using the Cyanidioschyzon sequence and Arabidopsis NP_192344 as queries) are from firmicutes, CFB bacteria and lenti/phaerae. These sequences also differ from others in sharing several unique residues and a 1-bp insertion. Sequences of Neurospora and Homo are mitochondrial precursors according to GenBank annotations. In Arabidopsis (GenBank accession number NP_192344 and TAIR locus AT4G04350), the protein product is targeted to both chloroplasts and mitochondria. The plant sequences are distantly related to other eukaryotic homologs based on both sequence similarity and phylogenetic analyses. Sequence of Pavlova was obtained from TBestDB.
Figure 17. Molecular phylogeny of leucyl-tRNA synthetase. P-value = 0.235 from AU test for the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of cyanobacterial and photosynthetic eukaryotic sequences. Topology A investigates if all eukaryotic sequences have the same origin (e.g. mitochondrial) whereas topology B investigates if photosynthetic eukaryotes acquired this gene from plastids. P-values < 0.001 from AU tests for both alternative topologies.

18. Ribosomal protein L11 methyltransferase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession | Organism                  | Sequence                                                                 |
|-----------|---------------------------|--------------------------------------------------------------------------|
| 24214091  | Leptospira interr          | EYEAXKYEFYKFISSYRYV1PTWFPYLLVNPGLAFGTGHETTRLVGLRNG                     |
| 91206171  | Rickettsia bellii          | DMVAYQNLVPQTRFFICTTSLILIEASRAFGTGHETTSGCIALE                            |
| 67940209  | Chlorobium phaeob          | NNNRAWEANLPVEIRIN1VQSSLIEINPKGSFGTGYHATTTLMLQIE                         |
| 34540441  | Porphyromonas gingiv       | NNNQWKKFPIRKCLVRAPFLEIISPQMAFGTGHETTSLMSYLL                            |
| 108866182 | Aedes aegypti              | NNNRAWEKNIPINVKLIRAESFEEIIOPKMSFGTGHATTLMIQQMM                         |
| RA00324465 | Reclinomonas ameri         | NNNRAWEWSFEPV1FVVAAVHRHFQIEITPRMSFGTGHATTWMMIRSIEME                   |
| 787778668 | Thiomicrospira denitr       | DNNVYQDDSTPLRHKFY1HPTWINAIDPALAPGTGHATTTASALRAIA                       |
| 33240884  | Prochlorococcus marinus    | DNNSSWKKFKNKVDPKILSPVWIKLDPGSFGTGHATPTLRCLLEDLE                       |
| 77464128  | Rhodobacter sphaer         | DNNVKVRRELSPVFEARFFVFYGSIALQIEATVAFGTVGRHTTGLCLRALD                   |
Note: The identifiable homologs of this gene (using *Cyanidioschyzon* and *Arabidopsis* sequences as queries) are found only in bacteria and eukaryotes, with all top hits being from beta and gamma-proteobacteria. This disjunct distribution suggests that the eukaryotic sequences are likely of bacterial origin. Phylogenetic analyses show a common ancestry of sequences from red algae, green plants and gamma, beta-proteobacteria. The gene was likely transferred independently to two groups of eukaryotes, one to the ancestor of red algae and green plants, another to the bacteriotrophic *Reclinomonas*. *Reclinomonas* and *Aedes* share several conserved residues and their grouping together likely resulted from eukaryote-to-eukaryote gene transfer. The *Arabidopsis* gene product is localized in cytoplasm according to GenBank annotation. Sequence of *Reclinomonas* was obtained from TBESTDB.

**Figure 18.** Molecular phylogeny of ribosomal protein L11 methyltransferase. P-value = 0.959 from AU test for the presented gene tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of red algal, green plant, and cyanobacterial sequences. P-value < 0.001 from AU tests for topology A and p-value = 0.041 for topology B.

19. GTP binding protein typA (D)

**CLUSTAL X (1.83.1) multiple sequence alignment**
Note: This gene is distributed in bacteria and various eukaryotes, including several protists. Sequences from TBestDB are too short for phylogenetic analyses. There are multiple versions of this gene in photosynthetic eukaryotes, one of which forms a monophyletic group with homologs GNS bacterial sequences with strong support (middle part of the tree). The other eukaryotic sequences group with alpha-proteobacterial homologs, likely resulting from intracellular gene transfer from mitochondria to the nucleus. Protein products of GNS bacteria-related genes (GIs: 30684514 and 30684509) are targeted to chloroplasts and intracellular components in Arabidopsis.
Figure 19. Molecular phylogeny of GTP binding protein, typA. P-value = 0.811 from AU test for the presented gene tree. AU tests were performed on alternative topologies including (A) monophyly of all eukaryotic sequences, (B) monophyly of red algal, green plant and cyanobacterial sequences, and (C) monophyly of red algal, green plant, cyanobacterial, and chloroflexi sequences. Topologies A and B investigate if red algae and green plants acquired the gene from mitochondria or plastids respectively, without invoking scenarios of secondary HGT. Topology C investigates if red algae and green plants acquired their genes from plastids and subsequently spread to chloroflexi through secondary HGT events. P-values < 0.001 from AU tests for topologies A and B whereas p-value = 0.195 for topology C. Therefore, the scenario of a plastidic origin in red algae and green plants and secondary HGT, although less parsimonious, cannot be confidently rejected.

20. Histidinol-phosphate transaminase (D)

CLUSTAL X (1.83.1) multiple sequence alignment
Identifiable homologs of this gene are only found in bacteria, primary photosynthetic euukaryotes and fungi. Protein sequences of primary photosynthetic euukaryotes and chloroflexi have the highest percent identity and they also share many conserved residues. As expected, these sequences also form a well-supported group in phylogenetic analyses. Fungal sequences have the highest percent identity with those of CFB bacteria and form another group with them. It is likely that primary photosynthetic euukaryotes and fungi acquired this gene from chloroflexi and CFB bacteria, respectively. The Arabidopsis sequence (GenBank accession number NP_568226 and TAIR locus AT5G10330) is annotated to be a chloroplast precursor. Sequence of Cyanophora paradoxa was obtained from TBESTDB.
Figure 20. Molecular phylogeny of histidinol-phosphate transaminase. P-value = 0.999 from AU test for the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of red algal, green plant, glaucophyte, and different versions of cyanobacterial sequences. P-values < 0.001 from AU tests for both alternative topologies.

21. tRNA methyltransferase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

21357477_Drosophila_melanogaster  RVVIGMSGGVDSSVAAHLLLHLYHSLVFMYKDMRQFVYKDFVY
39997644_Geobacter_sulfurreducens  RVVIGMSGGVDSSAVALLLYFNVVYNQETFQVQSVKDFVY
IEELKQMGNTPSDIDPFPCNMRKFGAPGYPDLYTVGTHYAKLTKQQAKKIDQKSYQ
LKDQYKMDPLPDVWNCVFKFDLVDFVYATHVGHNLIRRAIDKQFTQF
KVAKASAPTPNPCLDLCNRSVCKGFALGLDATHAVERTLRRLGVIDETQKSY
INAYEVEQTPNPNCALCNLMKFGALACXATHAVYRIQVLEALDIKQSY
KROLKLTPNPCACHRCFVKFGYLFDAFASGHYAIKIKVGLDLKQSY
IEDEYQGRTPNQPVCMNCLFLFRRMLCDIATHAVYRLIVAGDDDKQSY
EFLNQKQGPPQPCNLCCVFNQVTRPVIFGLKLDLATKASGHYAIKIKVGLDLKQSY
IEDEYQGRTPNQPVCMNCLFLFRRMLCDIATHAVYRLIVAGDDDKQSY
EFLNQKQGPPQPCNLCCVFNQVTRPVIFGLKLDLATKASGHYAIKIKVGLDLKQSY
IEDEYQGRTPNQPVCMNCLFLFRRMLCDIATHAVYRLIVAGDDDKQSY
EGLYRQGTTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
VFLHGTLTPNPCVCRQCKVTFELADPLATHAVYRRLGLDLKQSY
Note: This gene appears to be restricted to bacteria and eukaryotes. All top hits of GenBank searches are from Borrelia, Lentisphaeria, and CFb bacteria. Protein product of the Arabidopsis sequence (GenBank accession number NP_175542 and TAIR locus AT1G51310) is localized in both chloroplasts and cytoplasm. The major eukaryotic sequence group contains some mitochondrial precursors (e.g. Homo sequence) and is likely of mitochondrial origin, although it is not particularly related to alpha-proteobacterial homologs. Please also note that the green alga Ostreococcus contains two versions of this gene, one of which groups with the eukaryotic mitochondrial clade while the other version with sequences of red algae, Theileria, CFb bacteria and spirochaetes. Sequences of Reclinomonas and Hartmanella were obtained from TbBestDB; these two sequences formed a group with Homo, Dictyostelium, and other eukaryotic sequences in preliminary phylogenetic analyses, but were excluded from detailed phylogenetic analyses because of their short length. Theileria is an apicomplexan parasite containing a plastid derived from an algal endosymbiont.
Figure 21. Molecular phylogeny of tRNA methyltransferase. P-value = 0.977 from AU test for the presented tree. AU tests were also performed on alternative topologies including (A) monophyly of all eukaryotic sequences, and (B) monophyly of cyanobacterial, red algal, and green plant sequences. These alternative topologies investigate if red algae and green plants acquired the genes from mitochondria or plastids respectively. P-values < 0.001 from AU tests for both alternative topologies.

22. Isoleucyl-tRNA synthetase (D)

CLUSTAL X (1.83.1) multiple sequence alignment

| Accession     | Species                        | Sequence  |
|---------------|--------------------------------|-----------|
| 118193839     | Cenarchaeum symbio             | VRSHLEGADL-DRNIVMFIEGPFTWNGPHAGHLGRVRKIDLWYRXNTLR |
| 6325217       | Saccharomyces mt               | LVYKEQLRDFFEEFSFILHDGPPYANGELHLGHANLKKLTDINRYQLSQQ |
| 66800699      | Dictyostelium mt              |           |
| 118751539     | Marinomonas sp.                |           |
| 46129096      | Haemophilus influenza          |           |
| 75237181      | Escherichia coli              |           |
| 47572014      | Rubrivivax gelati             |           |
| 1174519       | Pseudomonas fluorescens       |           |
| 83594296      | Rhodospirillum rubrum          |           |
| 94968888      | Acidobacteria bacter           |           |
| 15644113      | Thermotoga maritima           |           |
| 15605834      | Aquifex aeolic                |           |
| 116747573     | Syntrophobacter fumar           |           |
| 116507018     | Ostreococcus tauri             |           |
| CM546_Cyanidioschyzon_merolae |                | IQRFWENRIYEGLYFLLHDGPPYANGSLHMGHLNKLKDINRYFILC |
45023744 Nocardia farcin

19074910 Entecophilozoon_cunic

1599084 Borrelia garini

42528163 Treponema dentica

51599084 Borrelia garini

46461128 Protochloramylae_aeomaeba

89989154 Chlamydia_felis

11874931 Marinomonas_sp.

63193874 Escherichia_coli

57157197 Trichoma_viridula

84477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium

36680069 Dictyostelium_montanum

63193874 Escherichia_coli

57157197 Trichoma_viridula

48477483 Picrophilus_torrid

80429573 Thermoanaerobacter_vanilli

60681847 Bacteroides_fragil

81185002 Salinibacter_halofermen

21673155 Chlorobium_tepidum

11122641 Frankia_alni

28210025 Clostridium
| Genus/Moniker | Sequence          |
|--------------|------------------|
| Salinibacter_ruber | AARAVEDFVEELSNWHLRRSRPRFWYQTIHECLAATAKLMSPIAPFFGEW |
| Chlorobium_tepidu | ACRLIGDFVDDLSNWYIRRSRKRFWYQTLSTVLETLAKLMAPFVPFIAEK |
| Frankia_alni | AGRRIARFIDDLSNWYVRRSRRRFWYTLYTCLALTATRAFFPFPFLTW |
| Blastopirellula_marin | ACGALTIFVDAALSNWVRSSRDRFWYTLTVLYECLITCCKLIAAPFTPFLAEAG |
| Deinococcus_geothede | GGRALERFVHDLSNWYVRRNRSRFWYATLHEALLTVSQQLAPFPTFLAEAG |
| Clostridium_symbio | LYSILLEGW |
| Saccharomyces_tetani | VWNPMRGKW |
| Dictyostelium_tetani | VFSVFAHGW |
| Marinomonas_sp. | IWSVFLETW |
| Syntrophobacter_fuma | VSFVFAHGW |
| Bacillus_fundu | AREIAVLDEVSNNYVRSSRNRFYWTLEHVLTSSKLIAAPFTPFLAEAG |
| Clostridium_tetani | AALAIEDFVDELSNWYVRRNRSRFWYVTLYKVLTTVSLAIPFVTFLAEAG |
| Eubacterium_fundu | VWSVQLTDM |
| Homo_mt | VFSVFRTGW |
| Oroccobacter_fundu | VSFVFRTGW |
| Pyrobaculum_aeroph | LWSVHLAQY |
| Sulfolobus_acidoc | IYSISMEKI |
| Archaeoglobus_fundu | FYSIFMEEY |
| Pyrococcus_abyssi | IYSVHMLDW |
| Pseudoalteromonas_fundu | IWSVFQRYW |
| Prochlorococcus_marin | IWSVFQRYW |
| Thalassospira_fundu | IWSVFQRYW |
| Synechocystis_praeternalis | LYSIFQNGW |
| Thalassospira_fundu | IWSVFQRYW |
| Arthrobacter_fundu | VWSVHLTDW |
| Leishmania_major | MYSIHFWAV |
| Vibrionaceae_fundu | LYSIFQNGW |
| Aeromonas_fundu | VWSVHLTDW |
| Heliobacter_fundu | LYSIFQNGW |
| Clostridium_fundu | VWSVHLTDW |
| Eubacterium_fundu | VWSVHLTDW |
| Thalassospira_fundu | IWSVFQRYW |
| Synechocystis_praeternalis | LYSIFQNGW |
| Thalassospira_fundu | IWSVFQRYW |
| Arthrobacter_fundu | VWSVHLTDW |
| Leishmania_major | MYSIHFWAV |
| Vibrionaceae_fundu | LYSIFQNGW |
| Aeromonas_fundu | VWSVHLTDW |
| Heliobacter_fundu | LYSIFQNGW |
| Clostridium_fundu | VWSVHLTDW |
Note: There are two eukaryotic sequence clades for this gene, each of which clusters within bacterial homologs with strong support. The minor eukaryotic sequence clade (lower part of the tree) contains chloroplast precursors from plants and mitochondrial precursors from opisthokonts. Sequences in the major eukaryotic sequence clade are cytosolic. It is likely that the major eukaryotic sequence clade resulted from an ancient HGT event prior to the split of most eukaryotic super groups. An alternative explanation is that the common ancestor of cellular organisms contained two copies of this gene, which were differentially retained among lineages.

Figure 22. Molecular phylogeny of isolecucyl-tRNA synthetase. P-value = 0.235 from AU test for the presented tree. AU test was also performed on an alternative topology enforcing a sequence monophyly of archaea and the major eukaryotic group. Such an alternative topology is based on the common belief that archaea and eukaryotes are more closely related than each is to bacteria. P-value < 0.001 from AU test for the alternative topology.
23. IspD (D)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

24. Polynucleotide phosphorylase (D)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

25. ATP/ADP translocase (N)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

26. Glycerol-3-phosphate acyltransferase (N)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

27. Sodium:hydrogen antiporter (N)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

28. GcpE (D)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

29. Beta-ketoacyl-ACP synthase (fabF) (D)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

30. Aspartate aminotransferase (D)
See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.
31. Tyrosyl-tRNA synthetase (D)

See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

32. Cu-ATPase (D)

See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

33. IspE (D)

See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

34. Enoyl-ACP reductase (fabI) (D)

See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

35. Florfenicol resistance protein (E)

See Huang and Gogarten 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends in Genetics 22:361-366.

36. 23S rRNA (Uracil-5')-methyltransferase (D)

See Huang and Gogarten 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology 8:R99.

37. Topoisomerase 6 subunit B (TOP6B) (N)

See Huang and Gogarten 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends in Genetics 22:361-366.