Evolution of b-type cytochromes in prokaryotes

Vassiliki Lila Koumandou[a,1], Sophia Kossida[a,2]

[a]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Soranou Efessiou 4, Athens 11527, Greece

[1]Present address: Department of Biotechnology, Agricultural University of Athens, Iera Odos 75, Athens 118 55, Greece

[2]Present address: Institute of Human Genetics, IMGT, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpelier cedex 5, France. E-mail : Sofia.Kossida@igh.cnrs.fr

Correspondence to: Vassiliki Lila Koumandou, Department of Biotechnology, Agricultural University of Athens, Iera Odos 75, Athens 118 55, Greece, Tel: + 30 210 529 4645, E-mail: koumandou@aua.gr

Keywords:

Molecular evolution, bioenergetics, bacteria, archaea, cytochrome b

Abstract:

Prokaryotes use a wide variety of bioenergetic pathways but the order of emergence of these pathways and their evolutionary relationships are still unresolved issues. In this study we focus on the evolutionary relationships of different families of b-type cytochromes, which form part of a variety of bioenergetic enzymes (the cytochrome b_6f complex, ubiquinol and menaquinol reductases, formate dehydrogenase, Ni/Fe-hydrogenase, and succinate dehydrogenase). We use data from 272 species of fully sequenced bacteria and archaea, which represent the full diversity of prokaryotic lineages and multiple bioenergetic modes, to examine the distribution of these cytochromes across lineages, and ask the question of whether sequences from different species cluster by cytochrome-b family, by bioenergetic mode or by taxonomic group. Different cytochrome-b types are found in many lineages of the bacteria and archaea, and form distinct groups in phylogenetic analysis, which indicates an ancient origin for this complex, and diversification of different cytochrome-b types before the diversification of lineages. We find that species do not cluster based on bioenergetic mode. We also re-examine data from previous studies using this expanded sample of organisms spanning the full diversity of prokaryotic lineages. Concerning the b_6f complex of photosynthetic organisms, our expanded phylogenies do not show significant bootstrap support for a "green clade" of cytochrome b_6; also, the split form of b_6 is not monophyletic, indicating that the split form arose independently multiple times. We also present data on the similarities between prokaryotic and eukaryotic cytochrome b561 sequences, in light of the recently reported structure of eukaryotic cytochrome b561.

1. Introduction:

Prokaryotes use a wide variety of bioenergetic pathways depending on the energy sources available in their chosen habitat (Todar) However, the order of emergence of these pathways and their evolutionary relationships are still unresolved issues. Much attention has been focused on the emergence of oxygenic photosynthesis, as this likely had far-reaching implications for life on Earth (Olson and Blankenship, 2004, Xiong, 2006, Blankenship, 2010). 16S rRNA phylogeny reveals a patchy picture of bioenergetic pathways in different taxonomic groups, suggestive of rampant horizontal gene transfer (HGT) (Castresana, 2001, Reysenbach and Shock, 2002, Boucher et al., 2003, Koumandou and Kossida, 2014). Molecular evolution studies of the electron transport chains of different pathways could help to clarify this picture. Each pathway depends on an electron transport chain (ETC) which comprises protein complexes acting as electron donors and electron acceptors, with a cytochrome bc-type complex and mobile electron carriers between them; the ETC generates a proton gradient across the bioenergetic membrane which is then harnessed by the ATP synthase complex to generate ATP. Recently we reported that the ATP synthase from a variety of organisms, representing nine bioenergetic pathways, shows an ancient origin and no propensity to horizontal gene transfer (Koumandou and Kossida, 2014). Indeed, the evolution of the ATP synthase subunits closely follows taxonomic groups, indicating that there is not a special enzyme associated with each ETC. In this study we focus on the cytochrome bc-type complexes, which are shared by most ETCs, and ask whether their evolution can shed light on the evolution of different ETCs. Previous studies have looked at the evolution of cytochrome bc but have used a relatively limited set of sequences which, in turn, represent only a portion of the true prokaryotic bioenergetic diversity. In recent years,

the number of complete genomes available has increased greatly, allowing us to expand the analysis to more lineages.

Schutz *et al* reviewed the evolution of cytochrome bc complexes 15 years ago, showing that the core of the complex was likely the cytochrome-b and the Rieske Iron Sulfur Protein (ISP), while the cytochrome c subunit is highly divergent and does not have a monophyletic origin (Schutz et al., 2000). Molecular evolution analysis of cytochrome-b and the ISP in 30 species (including chloroplasts, mitochondria, proteobacteria, cyanobacteria, firmicutes, chlorobi, *Deinocccus*, and some archaea) showed global agreement with the 16S rRNA phylogeny, with the exception of one HGT event into aquificae (Schutz et al., 2000). A subsequent analysis by the same group including more species, showed similar results, but also indicated that evolutionary reconstruction of the Rieske protein had to be guided by structure-based alignments to avoid pitfalls suggestive of multiple HGT events (Lebrun et al., 2006). More recently, with expanded genome sampling, a grouping of the b-type cytochromes from cyanobacteria, chlorobi and heliobacteria was suggested, all of which share RCI-type photosystems (Nitschke et al., 2010).

Integral membrane di-heme b-type cytochromes form part of many key enzymes of bioenergetic pathways, and were argued to be part of a pre-LUCA construction kit from which many energy-conserving enzymes evolved (Baymann et al., 2003, Schoepp-Cothenet et al., 2013). However, other researchers have refuted this conclusion, again based on expanded phylogenies, suggesting an early origin for a b_6f-type complex, and later fusions and HGT to form the present picture of distribution of these enzymes (Dibrova et al., 2013). Recent evidence from genomic data showing the presence of multiple Rieske/cytb proteins in many species further complicates the picture (Dibrova et al., 2013, ten Brink et al., 2013).

Previous analyses (Schutz et al., 2000, Lebrun et al., 2006, Nitschke et al., 2010, Dibrova et al., 2013) have attempted to derive a complete phylogeny of cytochrome-b from a diverse set of prokaryotes, but they mostly focus on the comparison of cytochrome b_6 of the b_6f complex of photosynthesis and the cytochrome bc complex of respiration. Also, in those analyses only some cytochrome-b domain containing proteins were included while some other were excluded; this might skew the results, and thus the conclusions. In this study we start with a full survey of all genes with predicted cytochrome-b domains. We examine all the orthology groups included in the KEGG database which include proteins with the characteristic cytochrome-b domains, as well as genes which contain these domains but are not part of an orthology group. This expanded analysis allows us to re-examine hypotheses about the early evolution of split cytochrome b_6, relationships between more distantly related b-type cytochromes, as well as cytochrome b561 in prokaryotes.

2. Methods:

2.1 Organism selection

272 species of bacteria and archaea, whose genomes have been completely sequenced and are available at NCBI, were chosen so as to cover the full diversity of bacterial and archaeal lineages (Wu et al., 2009) (http://tolweb.org/tree/) and the full bioenergetic diversity per lineage, as previously described (Koumandou and Kossida, 2014). For lineages with many sequenced genomes, the tree of (Wu et al., 2009) was used to pick species so as to cover as much phylogenetic diversity as possible with a

limited number of species. The set of species selected represent 131 clusters with a genome similarity score (GSS) threshold of 0.5 (Moreno-Hagelsieb et al., 2013); of those, 24 belong to single-species "clusters", and 63 are the sole representatives from their cluster. Information on the metabolic mode of all species was retrieved from NCBI and the IMG database (Markowitz et al., 2012). Each species name was assigned an 8-character abbreviation for better data handling during the phylogenetic analysis, by keeping the first two letters of the genus name and the first three letters of the species name, as well as a 2-3 letter ending, denoting the bioenergetic mode. Details of all the organisms analyzed and of the species names abbreviations are given in Table S1.

2.2 Sequence retrieval and phylogenetic analysis

Sequence accession numbers were retrieved using the ortholog tables from the KEGG database: these are regularly updated and are based on RefSeq annotations, sequence similarity and best-hit searches, as well as tools for operon-like consistency and completeness of pathway modules and complexes (http://www.kegg.jp/kegg/ko.html). This was supplemented by synteny considerations when only parts of the Rieske/cytochrome-bc-type operon were annotated. For sequences not belonging to an orthology group, putative orthologs in other species were identified by BLAST (best reciprocal blastp hits). For each species, the list of genome annotations were also individually checked in KEGG to collect any sequence annotated as having a cytochrome-b domain, which was not part of an orthology group. The accession numbers of all sequences analyzed are given in Table S1. Sequences were downloaded from KEGG in fasta format using a custom perl script. Alignments were created using MUSCLE (Edgar, 2004). Only unambiguous homologous regions were retained for phylogenetic analysis by manually inspecting,

masking and trimming the sequences in McClade (alignments available upon request). ProtTest (Abascal et al., 2005) was used to estimate the appropriate model of sequence evolution (Blosum62+I+G for datasets R3, R5 and R7; Blosum62+G for all the other datasets).

Phylogenetic analysis was performed by three separate methods. To obtain the Bayesian tree topology and posterior probability values, the program MrBayes version 3.1.2 was used (Ronquist and Huelsenbeck, 2003). Analyses were run for 1–5 million generations (as needed per dataset), removing all trees before a plateau established by graphical estimation. All calculations were checked for convergence and had a splits frequency of <0.1. Maximum-likelihood (ML) analysis was performed using PhyML (Guindon and Gascuel, 2003) and RAxML (Stamatakis, 2006) with 100 bootstrap replicates. Nodes with better than 0.95 posterior probability and 80% bootstrap support were considered robust, and nodes with better than 0.80 posterior probability and 50% bootstrap support are shown. Tree files were processed in Figtree v1.4 and Adobe Illustrator to highlight monophyletic groups, sequences belonging to the same orthology group, and/or color-code species names based on bioenergetic mode.

3. Results and Discussion:

3.1 Multiple b-type cytochromes exist in various lineages

Previous analyses (Lebrun et al., 2006, Dibrova et al., 2013, Kao and Hunte, 2014) have mostly focused on the comparison of cytochrome b from the cytochrome bc complex of the respiratory ETC, and cytochrome b_6 of the b_6f complex of the photosynthetic ETC. These are grouped into eight orthology groups in the KEGG

database: K02635 petB (fused to K02637 petD if in the split form), K00412 CYTB, K00410 fbch, K03887 MQCRB (fused to K03888 MQCRC), K03891 qcrB, and K15879 narC. However, three more orthology groups in the KEGG database include proteins with the same pfam domains (K12262: cytb 561 of *E. coli*; K00127: fdoI of formate dehydrogenase; K03620: hyaC of Ni/Fe-hydrogenase; Table 1). The presence of these pfam domains indicates homology, so we included all sequences from the orthology groups previously analyzed, but also added the additional three orthology groups for completeness. In addition, there are many sequences which contain these domains but are not part of an orthology group; these were also included in the analysis. In total, we collected all annotated cytochrome-b sequences within these orthology groups, and all sequences with a cytochrome-b domain for a set of 272 species. As previously reported, this set of species was chosen to represent all the major prokaryotic lineages as evenly as possible given the sequencing bias (e.g. for proteobacteria), as well as the full variety of bioenergetics pathways used by prokaryotes, (Koumandou and Kossida, 2014).

Some lineages lack b-type cytochromes altogether (e.g. fusobacteria, mollicutes, dictyoglomi), some have a limited set (e.g. chloroflexi, chlamydia, planctomycetes), while others have a great variety of them (notably acidobacteria, clostridia, alpha- and gamma-proteobacteria; Table 2). However, this may partly be a reflection of the much lower number of genomes available for certain lineages (Table 2). Overall, 65 of the species analyzed here lack cytochrome-b domain containing proteins (highlighted in grey in Table S1). For the set of archaea we examined, most lack b-type cytochromes altogether, but at least some representatives from multiple lineages do have a cytochrome-b type protein (Table 2, Table S1). A pre-LUCA presence of bioenergetic enzymes was argued based on the presence of these enzymes

in both bacteria and archaea (Baymann et al., 2003). With more genomes available, it is now clear that many of the orthology groups analyzed here are present in both bacteria and archaea.

Sequences belonging to each orthology group were marked with a suffix from "A" to "L", so as to distinguish them in the phylogenetic analysis, and different subsets of the data were analyzed to address different questions regarding the evolutionary relationships between and within these orthology groups (Table 1). We first attempted to reconstruct a phylogeny for all sequences and all species (dataset cytb-all.R1 comprising 627 sequences total). Cytochrome b5 and succinate dehydrogenase cytochrome b556 (K00241) were included as outgroups. The resulting phylogeny had very bad resolution overall (results not shown). Removal of the K00241 sequences (dataset cytb.R2 comprising 472 sequences total) improved the resolution, so that most orthology groups appeared tightly clustered, albeit not with significant bootstrap support (Figure S1). Dataset cytABCD.R7 (section 3.2) was analyzed to better resolve questions raised by previous analyses (Lebrun et al., 2006, Dibrova et al., 2013) on the evolution of cytochrome b_6. Dataset cytBHL.R9 was used to address the relationships of more distant cytochrome-b types (section 3.3); it includes the b-type cytochromes of the bc complex (K00412: CYTB and K00410: fbch), the formate dehydrogenase (K00127) and the Ni/Fe dehydrogenase (K03620). Cytochrome b556 of the succinate dehydrogenase (K00241) sequences were analyzed in combination with these three, (cytBHL-K00241.R10) and as a separate group (dataset K00241.R4 including 155 sequences). Datasets cytb561.R3 and cytb561.R5 were used to address the question of the relationship between prokaryotic and eukaryotic cytochrome b561 (section 3.4). Finally, phylogenetic analysis of datasets

cytb561.R6 and cytA-FI.R8 aimed at the assignment of hypothetical cytochrome b sequences to known orthology groups (section 3.5).

3.2 Cytochrome-b of the *b_6f* and the bc complex

The b_6f complex of photosynthetic ETCs, and the bc complex of respiratory ETCs have long been known to be homologous (Schutz et al., 2000). Interestingly, the b_6f complex features a split cytochrome b_6 in many species. Previous analyses of cytochrome-b in prokaryotes have suggested (a) a grouping of cyanobacteria, chlorobi and heliobacteria, all of which share RCI-type photosystems (Nitschke et al., 2010), (b) an early origin of the split form of cytochrome b_6 and the *b_6f*-type complex, and later fusions and HGT to form the present picture of distribution of these enzymes (Dibrova et al., 2013), and (c) that cytochrome-b proteins form part of a pre-LUCA construction kit from which many energy-conserving enzymes evolved (Baymann et al., 2003, Schoepp-Cothenet et al., 2013). In order to better resolve the position of cytochrome-b in these two complexes, we analyzed sequences from our set of species, corresponding to cytochrome b_6 of the *b_6f* complex (A: petB-K02635 fused to petD-K02637 when split), and to the cytochrome-b subunit of the cytochrome bc complex (B: fbch-K00410 and CYTB-K00412; C: MQCRB-K03887 fused to MQCRC-K03888; D: qcrb-K03891).

The results indicate a separation of the four orthology groups, but only the "D" (K03891) and the "C" groups (K03887/8) are monophyletic with good bootstrap support (Figure 1); this is likely influenced by the fact that these groups are small. Halobacteria sequences group with "D" sequences of the actinobacteria, with (or without in the PhyML analysis) the inclusion of two divergent "B" sequences from delta-proteobacteria and firmicutes (Figure 1). The overall tree structure is similar to Figure 1 of (Nitschke et al., 2010), based on which a grouping of the cyanobacteria,

heliobacteria and chlorobi was suggested. However, these three taxa only cluster in our PhyML analysis (not supported by high bootstrap values), whereas in the MrBayes (Figure 1) and RaxML trees, the chlorobi are in between the "A" (cytochrome b_6) and "B" (cytochrome b subunit of the ubiquinol cytochrome-c reductase) group of sequences. Notably, sequences from the cyanobacteria and the chlorobi, each form a very tight cluster, indicating little room for variation in the cytochrome b_6 in these two lineages, except for a second copy in *Gliobacter* which is very divergent. Species-specific duplications are seen in five other species (marked with a red "-" or ">" next to the species name in Figure 1). About 75% of the "A" group of sequences correspond to a split cytochrome b_6 (marked with a red asterisk at the corresponding nodes), where petD encodes the C-terminal subunit 4; in these cases petB and petD sequences were concatenated before the multiple alignment and phylogenetic analysis. There is, however, not significant bootstrap support for grouping all the split cytochromes together, and thus inferring an early origin of the split cytochrome b_6 from which the fused cytochrome-b of the bc complex evolved, as suggested previously (Dibrova et al., 2013). Our results are in agreement with studies which have suggested that the split between the N- and C-terminal regions of cytochrome b_6 occurred several times independently in different clades (Baymann et al., 2012, Kao and Hunte, 2014).

The "A" cytochrome b_6 group also includes a small cluster with sequences from a mix of lineages (*Pirellula staleyi* - planctomycetes, *Candidatus Solibacter usitatus* - acidobacteria, *Herpetosiphon aurantiacus* - chloroflexi, and *Desulfovibrio* sp. ND132 - delta-proteobacteria), which are more divergent and likely the result of HGT. Other delta-proteobacteria confidently cluster with the "B" group of ubiquinol cytochrome c reductase sequences. Notably, "B" sequences from the aquificae cluster

with the epsilon-proteobacteria, suggestive of horizontal gene transfer, as reported previously (Schutz et al., 2000). Apart from the cluster of "B" sequences from the proteobacteria and the aquificae, there is a cluster of divergent "B" sequences from a mix of lineages (*Opitutus terrae* -verrucomicrobia, *Acidobacterium capsulatum* and *Candidatus Solibacter usitatus* - acidobacteria, *Gemmatimonas aurantiaca* - gemmatimonadetes, *Candidatus Nitrospira defluvii* - nitrospirae) which group with the divergent *Gliobacter* cytochrome b_6 "A" sequence.

In the analysis by Dibrova et al (Dibrova et al., 2013), narC sequences were also included (K15879 here assigned as "F"), as well as certain cytochrome-b domain containing proteins, which are not part of an orthology group. This set of hypothetical cytochrome b sequences, and their orthologs in the other species examined in the present study, are assigned as group "E". There are, however, a number of other cytochrome-b domain containing proteins that are not part of an orthology group in many of these species, and we assigned this as group "I" (see Table 1). When we add these cytochrome-b domain containing proteins to our analysis, the clustering of the core groups of the "A" and "B" sequences is confirmed (Figure S2), but some of the more divergent "A" and "B" sequences end up in different clusters than in Figure 1. Namely, cyanobacterial, heliobacterial and chlorobi cytochrome b_6 "A" sequences cluster, but not with significant bootstrap support. The mixed group of divergent cytochrome b_6 "A" sequences from Figure 1 ends up clustering with hypothetical "E" sequences mostly of the delta-proteobacteria. Ubiquinol cytochrome c reductase "D" sequences confidently cluster with narC "F" sequences. The halobacterial "A" sequences cluster with this "D" and "F" group, but not with high bootstrap support (in contrast to the statistically supported grouping of the halobacteria with the actinobacteria in Figure 1). "B" sequences of the alpha- beta- gamma- delta- epsilon-

proteobacteria and aquificae group confidently, while most of the divergent "B" sequences end up in a separate cluster which also includes the hypothetical "E" sequence of the planctomycete *Pirellula staleyi*. The small menaquinol cytochrome c reductase "C" cluster ends up closest to this set of divergent "B" sequences. The hypothetical "E" an "I" sequences of archaea (thermoprotei and thermoplasmata) also confidently cluster together. The rest of the "I" sequences form a separate cluster, not with significant bootstrap support, and with quite long branches relative to the rest of the sequences (Figure S2). Further analysis showed that these are more similar to cytochrome b561-like sequences as discussed below (section 3.5, Figure S5).

3.3 Cytochrome-b subunit of ubiquinol-cytochrome c reductase, formate dehydrogenase and Ni/Fe-hydrogenase

Apart from the cytochrome bc and b_6f complexes, b-type cytochromes form part of a number of other enzymes (Baymann et al., 2003, Schoepp-Cothenet et al., 2013). Given that the overall phylogeny of all cytochrome-b groups failed to give good resolution, we analyzed separately three groups of sequences, corresponding to the cytochrome-b subunit of (i) the bc complex (ubiquinol-cytochrome c reductase, "B" set of sequences), (ii) formate dehydrogenase ("H" sequences), and (iii) Ni/Fe hydrogenase ("L" sequences). B-type cytochromes of the bc complex and formate dehydrogenase belong to the same scop superfamily [SCOP: 81342], while those of formate dehydrogenase and Ni/Fe hydrogenase are also structurally related (Berks et al., 1995, Jormakka et al., 2002).

The analysis shows that the three orthology groups are clearly separated. The bc complex sequences cluster together with high bootstrap support in the MrBayes analysis (blue "B" group in Figure 2, contrast to Figure 1). All but one of the Ni/Fe-hydrogenase sequences also cluster together with good bootstrap support (grey "L"

group in Figure 2), while the formate dehydrogenase sequences (pink "H" sequences) form two groups in between "B" and "L": one compact group with high bootstrap support ends up closer to the "B" group, and a more loose group closer to the "L" cluster (Figure S3). Interestingly, archaeal formate dehydrogenase sequences are placed well within the "H" cluster, while the divergent "H" group includes sequences from the alpha- and beta-proteobacteria, verrucomicrobia and gemmatimonadetes. This clustering may be influenced by the fact that certain species use tungsten instead of molybdenum in the active site of the pyranopterin cofactor for formate dehydrogenase (Maia et al., 2015). However, overall within each orthology group, species largely cluster based on taxonomy. This indicates that these three types of cytochromes evolved before the diversification of lineages, and are thus an ancient features of prokaryotes, in agreement with previous analyses (Baymann et al., 2003).

The formate dehydrogenase and Ni/Fe-hydrogenase clusters are more closely clustered in the tree, than the cytochrome bc complex (Figure S3A). If we assume an early origin for the Ni/Fe-hydrogenase, as has been suggested previously (Baymann et al., 2003), the cytochrome-b of the bc complex evolved later than the formate dehydrogenase cytochrome (Figure 2, Figure S3A), and forms a less-divergent group overall, perhaps constrained by its position and/or function in the bc complex. This early origin was argued based on the presence of both archaeal and bacterial sequences for typeI hydrogenases (Baymann et al., 2003), but with more genomes available, it is now clear that all three orthology groups analyzed here are present in both bacteria and archaea (Table 2). In an attempt to "root" the tree, sequences of the cytochrome b556 subunit of the succinate dehydrogenase (K00241) were added to the analysis as an outgroup; these share the same scop fold [SCOP: 81344] as b-type cytochromes of the bc complex and formate dehydrogenase, but belong to a different

superfamily [SCOP: 81343]. The phylogenetic analysis shows that succinate dehydrogenase (K00241) sequences end up between the "B" cluster and the "H" and "L" cluster, i.e. it enforces the close relationship of formate dehydrogenase and Ni/Fe-hydrogenase, and hints at an early split between these and the cytochrome bc complex (Figure S3B). In addition, the more divergent sequences of formate dehydrogenase cluster with succinate dehydrogenase, although the bootstrap values for this grouping are not significant. The bootstrap values in the deep nodes are generally low, thus do not allow a distinction of the exact relationships of these different groups of cytochromes. Our data does however suggest an early presence of succinate dehydrogenase, as inferred for the other enzymes, in prokaryotes.

Notably, the succinate dehydrogenase (K00241) sequences also form a number of distinct subgroups in Figure S3B. Six subgroups are also seen when succinate dehydrogenase sequences are analyzed separately (Figure 3; dataset K00241.R4 including 155 sequences). Two of them (groups V and VI) have a mix of sequences from various bacterial lineages, while two (groups I and IV) include sequences from archaea and bacteria; groups II and III only include alpha- beta- and gamma-proteobacteria sequences (Figure 3). These results indicate horizontal gene transfer of succinate dehydrogenase cytochrome b556 (K00241) among bacterial and archaeal lineages, e.g. between thermoprotei and cyanobacteria in group I. Notably, lineages which normally cluster tightly in phylogenetic analysis have representatives in more than one group, e.g. cyanobacteria and chlorobi end up in groups I and VI, chloroflexi in groups IV and VI, and delta-proteobacteria in groups I, IV and VI. A previous analysis of the soluble subunits of succinate dehydrogenase distinguished five different types of the enzyme (Lemos et al., 2002). Although the trees are difficult to compare, as different species were included in each, in general terms,

groups I and VI of our analysis correspond to type E, groups II and III correspond to type C, group IV corresponds to type A, and group V corresponds to type B.

3.4 Prokaryotic cytochrome b561

The eukaryotic ascorbate-dependent oxidoreductase cytochrome b561 (cyt-b561), plays an important role in ascorbate recycling and iron absorption. Recently, the structure of the *Arabidopsis* cytochrome b561 was reported, and based on the fact that the protein family is highly conserved in eukaryotes, a general mechanism for the function of the protein was suggested (Lu et al., 2014). A previous study had reported that cytochrome b561 is absent from prokaryotes and fungi (Verelst and Asard, 2003), although fungal representatives were reported in a later study (Tsubaki et al., 2005). Indeed, a family of annotated prokaryotic cyt-b561 also exists in proteobacteria but their precise role is still unclear (Murakami et al., 1986). We therefore examined the evolutionary relationship between eukaryotic and prokaryotic cyt-b561. Annotated cyt-b561 sequences were collected for selected eukaryotic and prokaryotic species from the corresponding KEGG orthology groups (K08360 and K12262, respectively). The multiple alignment shows overall good conservation between the prokaryotic and eukaryotic sequences, along with some residues shared between both groups (Figure S4). Eukaryotic cytochrome b561 sequences have six helices, with helices 2-5 contributing the four histidines that co-ordinate the hemes (Lu et al., 2014). Prokaryotic cytochrome b561 sequences (K12262) are predicted by PSI-Pred to have only four helices, and as indicated by the multiple alignment these correspond to helices 1, 2, 3, and 5 of eukaryotes. Consequently, only 3 of the four histidines are conserved (Figure S4). However, another 3 histidines are conserved among widely divergent prokaryotic species, and one or more of these could help co-ordinate the heme irons. Lu *et al* identified a number of other residues important for forming H-

bonds with the heme groups, as well as for substrate binding and catalysis. These are not generally conserved between eukaryotes and prokaryotes, but a number of other residues - mostly leucines and glycines - are conserved between the two kingdoms. Importantly, Lys^{81} and His^{106} (*Arabidopsis* numbering) were identified in the Lu *et al* structure as crucial for activity, with additional evidence from previous studies supporting the importance of Lys^{81} (Lu et al., 2014). Lys^{81} is partly conserved across some prokaryotes, but His^{106} is not, and seems to be replaced by glycine in almost all prokaryotes (Figure S4).

Phylogenetic analysis confidently separates the eukaryotic from the prokaryotic group of sequences (Figure 4; dataset cytb561.R3 including 22 eukaryotic cytochrome b561 sequences (K08360) plus 61 prokaryotic homologs (K12262) comprising 83 sequences total). A species-specific duplication is seen in *Emiliana huxleyi*. Multiple duplications are seen in *Populus trichocarpa* and *Oryza sativa*; two of these can be traced to their common ancestor, with further species-specific duplications in *P. trichocarpa*. Multiple duplications are also evident in various prokaryotes, none of which appear to be species-specific, and thus might be attributed to HGT, although the bootstrap values for most groupings are weak, so that even alpha- beta- and gamma-proteobacterial lineages are not clearly separated.

One more orthology group annotated as prokaryotic cytochrome b561 exists (narC, K15879) with representatives in deinococci and in the archaeon *Haloarcula marismortui*. Sequences from this small group form a tight cluster within the K12262 sequence cluster (Figure 5); however, previous analyses have grouped these sequences with cytochrome b sequences of ubiquinol cytochrome c reductases (Dibrova et al., 2013), where they are also confidently placed in our analysis of the respective dataset (Figure S2).

3.5 Assigning hypothetical cytochrome-b domain containing proteins to an orthology group

For the species analyzed, 57 other prokaryotic sequences are annotated as cyt-b561 but are not part of an orthology group (group "J" in Table 1). Seven of these (corresponding to species: *Nitrosococcus oceani*, *Nitrosococcus halophilus*, *Gallionella capsiferriformans*, *Sideroxydans lithotrophicus*, *Rhodobacter sphaeroides*, *Dinoroseobacter shibae*, *Roseobacter denitrificans*) can be assigned to the K11262 orthology group, as they cluster with K11262 sequences in the phylogenetic analysis (Figure 5). The rest of the "J" sequences form a group with three subclusters, distinct from K12262 and K15879, and further removed from the eukaryotic sequences (Figure 5). Many species have multiple hypothetical sequences annotated as cytb-561 which end up in the different subclusters, e.g. *Rhodoferax ferrireducens* and *Allochromatium vinosum* have sequences in all three subclusters of the "J" group. The archaeal sequence from *Candidatus Methanoregula boonei* groups with the eukaryotes.

We also included in the analysis "E" and "I" sequences with the cytochrome-b domain, which were not assigned to a particular orthology group. When these were analyzed alongside cytochrome b sequences from ubiquinol cytochrome c reductases and the b6f complex, there was a divergent group of "I" sequences with very long branches (Figure S2). In the phylogenetic analysis with cytochrome b-561-like sequences, all of these divergent "I" sequences (except the two sequences corresponding to *Nostoc azollae* and *Desulfarculus baarsii*) cluster either with cytochrome b-561-like "J" sequences, or with eukaryotic cytochrome b561 sequences (Figure S5). In some cases, the bootstrap values at the nodes of clusters containing both "I" and "J" sequences are significant, and branch lengths are shorter than in the

tree of Figure S2, allowing a more confident placement of these sequences closer to cyt561. In particular, there is significant bootstrap support for the assignment of the "I" sequences of the following species to the "J" group: *Desulfovibrio magneticus*, *Gallionella capsiferriformans*, *Geobacter bemidjiensis*, *Geobacter metallireducens*, *Geobacter sulfurreducens*, *Magnetospirillum magneticum*, *Pelodictyon luteolum* and *Sulfuricurvum kujiense*. In Figure S2 there is a well-supported group of "E" and "I" sequences from the Thermoprotei and the Thermoplasmata, which is also seen in Figure S5. Finally, a group of "E" sequences from delta-proteobacteria cluster with "A" cytochrome b_6 sequences in Figure S2, and this cluster persists and ends up closer to "F" (K15879) sequences in Figure S5. Therefore, the assignment of "E" sequences in Figure S2 is robust. Overall, the phylogenetic analysis allows the assignment of some of the hypothetical cytochrome-b sequences to an annotated orthology group.

4 Conclusions:

B-type cytochromes are part of a variety of bioenergetic enzymes. Different cytochrome-b types are found in many lineages of the bacteria and archaea, and form distinct groups in phylogenetic analysis, largely consistent with orthology group assignments in the KEGG database. This indicates a divergence of cytochrome-b types before the diversification of prokaryotic lineages. B-type cytochromes are thus an ancient feature of the prokaryotes. In the phylogenetic analysis, some of the cytochrome-b orthology groups split into a number of tightly grouped subclusters; notably, subclusters of the succinate dehydrogenase cytochrome b556 subunit sequences include representatives from various lineages of the bacteria and archaea, suggesting rampant horizontal gene transfer. Interestingly, cytochrome -b sequences

do not group according to bioenergetic mode. In our expanded phylogenetic analysis of cytochrome-b of the b_6f complex and the bc complex, there is not significant bootstrap support for the "green" clade of cytochrome b_6. Also, the split form of cytochrome b_6 is not monophyletic, suggesting multiple fission events throughout evolution. Analysis of annotated and hypothetical prokaryotic cytochrome b561 sequences indicates many similarities to their eukaryotic homologs. Finally, a number of hypothetical cytochrome b sequences can be assigned to an orthology group based on phylogeny.

Acknowledgments:

We thank Ioanna Karamichali for providing the custom perl script used for collecting the sequences in fasta format from the KEGG database. We are indebted to the CamGrid computational resource on which the phylogenetic analyses were performed. This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme (FP7-PEOPLE-2011-IEF proposal No 298890).

References:

Abascal, F., Zardoya, R., and Posada, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. 21:2104-5.

Baymann, F., Lebrun, E., Brugna, M., Schoepp-Cothenet, B., Giudici-Orticoni, M.T., and Nitschke, W. 2003. The redox protein construction kit: pre-last universal common ancestor evolution of energy-conserving enzymes. *Philos Trans R Soc Lond B Biol Sci*. 358:267-74.

Baymann, F., Schoepp-Cothenet, B., Lebrun, E., van Lis, R., and Nitschke, W. 2012. Phylogeny of Rieske/cytb complexes with a special focus on the Haloarchaeal enzymes. *Genome Biol Evol*. 4:720-9.

Berks, B.C., Page, M.D., Richardson, D.J., Reilly, A., Cavill, A., Outen, F., and Ferguson, S.J. 1995. Sequence analysis of subunits of the membrane-bound nitrate reductase from a denitrifying bacterium: the integral membrane subunit provides a prototype for the dihaem electron-carrying arm of a redox loop. *Mol Microbiol*. 15:319-31.

Blankenship, R.E. 2010. Early evolution of photosynthesis. *Plant Physiol*. 154:434-8.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E., Nesbo, C.L., Case, R.J., and Doolittle, W.F. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet*. 37:283-328.

Castresana, J. 2001. Comparative genomics and bioenergetics. *Biochim Biophys Acta*. 1506:147-62.

Dibrova, D.V., Cherepanov, D.A., Galperin, M.Y., Skulachev, V.P., and Mulkidjanian, A.Y. 2013. Evolution of cytochrome bc complexes: from membrane-anchored dehydrogenases of ancient bacteria to triggers of apoptosis in vertebrates. *Biochim Biophys Acta*. 1827:1407-27.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792-7.

Guindon, S., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696-704.

Jormakka, M., Tornroth, S., Byrne, B., and Iwata, S. 2002. Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science*. 295:1863-8.

Jun, S.R., Sims, G.E., Wu, G.A., and Kim, S.H. 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc Natl Acad Sci U S A*. 107:133-8.

Kao, W.C., and Hunte, C. 2014. The molecular evolution of the Qo motif. *Genome Biol Evol*. 6:1894-910.

Koumandou, V.L., and Kossida, S. 2014. Evolution of the F0F1 ATP synthase complex in light of the patchy distribution of different bioenergetic pathways across prokaryotes. *PLoS Comput Biol*. 10:e1003821.

Lebrun, E., Santini, J.M., Brugna, M., Ducluzeau, A.L., Ouchane, S., Schoepp-Cothenet, B., Baymann, F., and Nitschke, W. 2006. The Rieske protein: a case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. *Mol Biol Evol*. 23:1180-91.

Lemos, R.S., Fernandes, A.S., Pereira, M.M., Gomes, C.M., and Teixeira, M. 2002. Quinol:fumarate oxidoreductases and succinate:quinone oxidoreductases: phylogenetic relationships, metal centres and membrane attachment. *Biochim Biophys Acta*. 1553:158-70.

Lu, P., Ma, D., Yan, C., Gong, X., Du, M., and Shi, Y. 2014. Structure and mechanism of a eukaryotic transmembrane ascorbate-dependent oxidoreductase. *Proc Natl Acad Sci U S A*. 111:1813-8.

Maia, L.B., Moura, J.J., and Moura, I. 2015. Molybdenum and tungsten-dependent formate dehydrogenases. *J Biol Inorg Chem*. 20:287-309.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N.N., Kyrpides, N.C. 2012. IMG: the Integrated

Microbial Genomes database and comparative analysis system. *Nucleic Acids Res*. 40:D115-22.

Moreno-Hagelsieb, G., Wang, Z., Walsh, S., and ElSherbiny, A. 2013. Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. *Bioinformatics*. 29:947-9.

Murakami, H., Kita, K., and Anraku, Y. 1986. Purification and properties of a diheme cytochrome b561 of the Escherichia coli respiratory chain. *J Biol Chem*. 261:548-51.

Nitschke, W., van Lis, R., Schoepp-Cothenet, B., and Baymann, F. 2010. The "green" phylogenetic clade of Rieske/cytb complexes. *Photosynth Res*. 104:347-55.

Olson, J.M., and Blankenship, R.E. 2004. Thinking about the evolution of photosynthesis. *Photosynthesis Research*. 80:373-386.

Reysenbach, A.L., and Shock, E. 2002. Merging genomes with geochemistry in hydrothermal ecosystems. *Science*. 296:1077-82.

Ronquist, F., and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19:1572-4.

Schoepp-Cothenet, B., van Lis, R., Atteia, A., Baymann, F., Capowiez, L., Ducluzeau, A.L., Duval, S., ten Brink, F., Russell, M.J., and Nitschke, W. 2013. On the universal core of bioenergetics. *Biochim Biophys Acta*. 1827:79-93.

Schutz, M., Brugna, M., Lebrun, E., Baymann, F., Huber, R., Stetter, K.O., Hauska, G., Toci, R., Lemesle-Meunier, D., Tron, P., Schmidt, C., Nitschke, W. 2000. Early evolution of cytochrome bc complexes. *J Mol Biol*. 300:663-75.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688-90.

ten Brink, F., Schoepp-Cothenet, B., van Lis, R., Nitschke, W., and Baymann, F. 2013. Multiple Rieske/cytb complexes in a single organism. *Biochim Biophys Acta*. 1827:1392-406.

Todar, K. The Diversity of Procaryotic Metabolism. *Online Textbook of Bacteriology*. http://www.textbookofbacteriology.net/, Accessed 22 January 2014.

Tsubaki, M., Takeuchi, F., and Nakanishi, N. 2005. Cytochrome b561 protein family: expanding roles and versatile transmembrane electron transfer abilities as predicted by a new classification system and protein sequence motif analyses. *Biochim Biophys Acta*. 1753:174-90.

Verelst, W., and Asard, H. 2003. A phylogenetic study of cytochrome b561 proteins. *Genome Biol*. 4:R38.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., Hooper, S.D., Pati, A., Lykidis, A., Spring, S., Anderson, I.J., D'Haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J.F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E.M., Kyrpides, N.C., Klenk, H.P., Eisen, J.A. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*. 462:1056-60.

Xiong, J. 2006. Photosynthesis: what color was its origin? *Genome Biol*. 7:245.

Figure legends:

Table 1: Orthology groups of cytochrome-b sequences in the KEGG database.

N/A refers to sequences which are not part of an orthology group, but contain the characteristic cytochrome-b domains. Sequence handling for the phylogenetic analysis per group is shown in the "notes" column, as well as the number of sequences per group for the set of species analyzed; see text for more details. For each of the different datasets used in the phylogenetic analysis (R1-R10), shaded boxes indicate that the dataset contains the respective orthology group (white boxes indicate that it does not); the number of the Figure(s) in the paper where the corresponding results can be found, is also indicated.

Table 2: Distribution of cytochrome-b families across bacterial lineages.

Filled boxes with a plus sign indicate presence in at least one species from that lineage in the taxonomy view of the orthology groups in the KEGG database; white sectors indicate that the protein is not found in that lineage (based on KEGG orthology assignments). Some cytochrome-b families are restricted to only a few lineages (e.g. K03888, K03891), while others are much more widely distributed (e.g. K00412, K12262, K00241). Some lineages (e.g. fusobacteria, mollicutes, dictyoglomi) lack b-type cytochromes altogether, some (e.g. chloroflexi, chlamydia, planctomycetes) have a limited set, while others, notably acidobacteria, clostridia, alpha- and gamma-proteobacteria, have a great variety of them. However, this may be a reflection of the much lower number of genomes available for certain lineages. The total number of species per lineage included in the KEGG database is shown in parentheses next to the lineage name. Bacterial lineages are grouped and color-coded

based on the "whole-proteome phylogeny of prokaryotes by feature frequency profiles" (Figure 2 of (Jun et al., 2010)). Actinobacteria and planctomycetes which were not part of a specific group in that phylogeny are shown in black/grey, and lineages not analyzed by Jun *et al* are shown at the end, also in black/grey.

Figure 1: Phylogenetic reconstruction of cytochrome-b of the b_6f and the bc complex.

The tree shown is the best Bayesian topology, based on 122 sequences total, including cytochrome b_6 of the cytochrome b_6f complex fused to subunit IV (K02635/7), cytochrome-b of the ubiquinol cytochrome c reductase (K00410 combined with K00412), cytochrome-b of the menaquinol cytochrome c reductase of firmicutes (K03887/8), and cytochrome-b of the haloarchaeal ubiquinol cytochrome c reductase (K03891). K02635/7 sequences are marked with an "A" at the end, K00410/2 sequences with a "B", K03887/8 sequences with a "C", and K03891 with a "D" (see Table 1 for details), and the corresponding groupings in the tree are boxed and shaded grey. Species names are color-coded based on their bioenergetic mode, as indicated in the insert. Numerical values at the nodes of the tree (x/y/z) indicate statistical support by MrBayes, PhyML and RAxML (posterior probability, bootstrap and bootstrap, respectively). Values for highly supported nodes have been replaced by symbols, as indicated. Duplicates are indicated with a red line after the name, or with a ">" sign in the case of species-specific duplications. Red asterisks at the nodes indicate a split cytochrome b_6. For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1.

Figure 2: Phylogenetic reconstruction of the cytochrome-b subunit of ubiquinol-cytochrome c reductase (bc complex), formate dehydrogenase and Ni/Fe dehydrogenase.

The tree shown is the best Bayesian topology of 205 sequences total, including cytochrome-b of the bc complex (B: K00410/2), formate dehydrogenase (H: K00127), and Ni/Fe hydrogenase (L: K03620). All sequences are marked with a letter at the end to distinguish different orthology groups (see Table 1), and are also color-coded accordingly, as shown in the insert. Numerical values at the nodes of the tree (x/y/z) indicate statistical support by MrBayes, PhyML and RAxML (posterior probability, bootstrap and bootstrap, respectively). Values for highly supported nodes have been replaced by symbols, as indicated. Archaeal species are marked with an asterisk (red for arhaeoglobi and black for thermoprotei).

Figure 3: Phylogenetic reconstruction of prokaryotic succinate dehydrogenase cytochrome b556 subunit.

The tree shown is the best Bayesian topology, based on 155 SdhC sequences from K00241. For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1. Species names are color-coded based on their bioenergetic mode, as indicated in the insert. Numerical values at the nodes of the tree (x/y/z) indicate statistical support by MrBayes, PhyML and RAxML (posterior probability, bootstrap and bootstrap, respectively). Values for highly supported nodes have been replaced by symbols, as indicated. Duplicates are indicated with a red line after the name (or a ">" sign in the case of Ba_sel, as they are both within the same group). The tree distinguishes six groups of sequences: group I has a mix of sequences from archaea and bacteria, group II includes mostly alpha-proteobacterial sequences (and two gamma-), group III includes only sequences from

the beta- and gamma-proteobacteria, groups V and VI have a mix of sequences from various bacterial lineages, and group IV has three subgroups, two of which are tightly clustered archaeal sequences, and the other a mix of various bacterial lineages. These results indicate rampant horizontal gene transfer for this subunit among bacterial and archaeal lineages.

Figure 4: Phylogenetic reconstruction of eukaryotic and prokaryotic cytochrome b561.

The tree shown is the best Bayesian topology, based on 22 eukaryotic cytochrome b561 (K08360) and 61 prokaryotic cytochrome b561 (K12262) sequences (from the alpha- beta- and gamma-proteobacteria). For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1. Prokaryotic sequences are marked with a "G" at the end (see Table 1), and are color-coded based on their bioenergetic mode, as indicated in the insert. Numerical values at the nodes of the tree (x/y/z) indicate statistical support by MrBayes, PhyML and RAxML (posterior probability, bootstrap and bootstrap, respectively). Values for highly supported nodes have been replaced by symbols, as indicated. Duplicates are indicated with a red line after the name, or with a ">" sign in the case of species-specific duplications. The tree confidently separates the eukaryotic from the prokaryotic group of sequences. Multiple duplications are seen in *Populus trichocarpa* and *Oryza sativa*; some of these can be traced to their common ancestor, with further species species-specific duplications in *P. trichocarpa*. A species-specific duplication is seen in *Emiliana huxleyi*. Multiple duplications are also evident in various prokaryotes, none of which appear to be species-specific, and thus might be attributed to HGT, although the bootstrap values for most groupings are weak, and the alpha- beta- and gamma-proteobacterial lineages are not clearly separated.

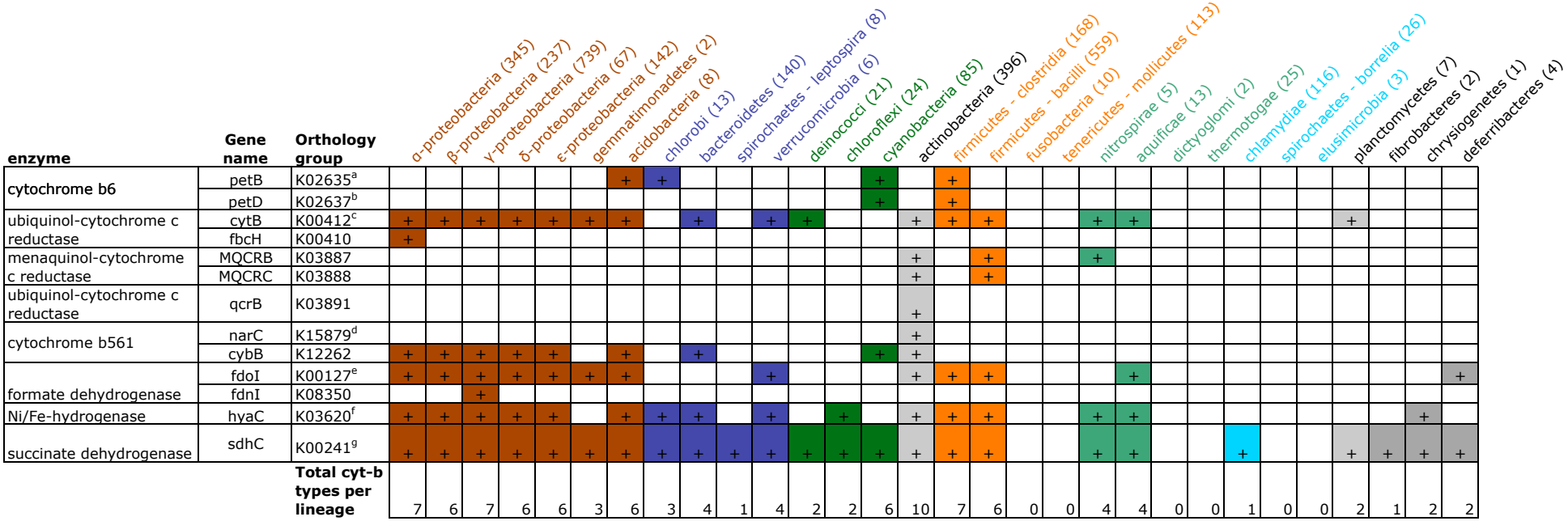
Figure 5: Phylogenetic reconstruction of eukaryotic and prokaryotic cytochrome b561, including sequences annotated as b561 but not belonging to an orthology group.

The tree shown is the best Bayesian topology, based on 22 eukaryotic cytochrome b561 (K08360), 61 prokaryotic cytochrome b561 (K12262) sequences (alpha- beta- and gamma-proteobacteria), 4 narC (K15879) sequences, and 57 prokaryotic sequences annotated as cyt-b561 but which are not part of an orthology group (see Table 1). K12262 sequences are marked with a "G" at the end, K15879 sequences with an "F" and the ones not belonging to an orthology group with a "J"; these three groups are also color coded as indicated in the insert (F: purple, G: brown, J: light blue). For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1. Numerical values at the nodes of the tree (x/y/z) indicate statistical support by MrBayes, PhyML and RAxML (posterior probability, bootstrap and bootstrap, respectively). Values for highly supported nodes have been replaced by symbols, as indicated. As in Figure 4, the tree confidently separates the eukaryotic from the prokaryotic group of sequences, except for the archaeal sequence from *Candidatus Methanoregula boonei*. The "F" group of sequences is also clearly separated (and indeed in the PhyML analysis is distinct from the "G" group, Figure S5), while "J" sequences form two distinct clusters, and are more distant from the eukaryotic group than group "G".

Enzyme	Orthology group	Gene name	Definition	pfam domains	notes	datasets for phylogenetic analysis: R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
cytochrome b6	K02635	petB	cytochrome b6	1, 2, 3	fused with K02637 (C-term) "A" group of sequences (44) associated Rieske (K02636 petD)										
	K02637	petD	cytochrome b6-f complex subunit 4	4	fused with K02635 (N-term) "A" group of sequences (33) associated Rieske (K02636 petD)										
ubiquinol-cytochrome c reductase	K00412	cytB	ubiquinol-cytochrome c reductase cytochrome b subunit	1, 4	combined with K00410 "B" group of sequences (67) associated Rieske (K00411 UQCRFS1)										
	K00410	fbcH	ubiquinol-cytochrome c reductase cytochrome b/c1 subunit	1, 2, 4	combined with K00412 "B" group of sequences (2) associated Rieske (K00411 UQCRFS1)										
menaquinol-cytochrome c reductase	K03887	MQCRB	menaquinol-cytochrome c reductase cytochrome b subunit	1, 2, 3	fused with K03888 (C-term) "C" group of sequences (2) associated Rieske (K03886 MQCRA)										
	K03888	MQCRC	menaquinol-cytochrome c reductase cytochrome b/c subunit	4	fused with K03887 (N-term) "C" group of sequences (2) associated Rieske (K03886 MQCRA)										
ubiquinol-cytochrome c	K03891	qcrB	ubiquinol-cytochrome c reductase cytochrome b subunit	1, 2, 4	"D" group of sequences (7) associated Rieske (K03890 qcrA)										
	N/A*		cytochrome b domain containing	1, 3	E group of sequences (24)										
cytochrome b561	K15879	narC	cytochrome b-561	1, 2, 3	"F" group of sequences (4) associated Rieske (K15878 narB)										
	K12262	cybB	cytochrome b-561 of E.coli	2, 3, 5	"G" group of sequences (61)			**		**	**				
formate dehydrogenase	K00127	fdoI	formate dehydrogenase subunit gamma	2, 3	associated Rieske (K00124 fdoH)										
	K08350***	fdnI	cytochrome b of formate dehydrogenase-like	2, 3	associated Rieske (K08349 fdnH) grouped as "H" group of sequences (83)										
	N/A		cytochrome b/b6 domain-containing protein	2, 3	"I" group of sequences (55)										
	N/A		annotated as cytochrome b561 but not part of K0	2, 3	"J" group of sequences (57)										
	N/A		cytochrome b5	6	"K" group of sequences (13)										
Ni/Fe-hydrogenase	K03620	hyaC	Ni/Fe-hydrogenase 1 B-type cytochrome subunit	2, 3	"L" group of sequences (53)										
succinate dehydrogenase	K00241	sdhC	succinate dehydrogenase cytochrome b556 subunit	7	associated Rieske (K00240 sdhB) 155 sequences										
						627	472	83	155	144	223	122	205	205	360
						total number of sequences per dataset									
					Figure	**N/A**	**S1**	**4/S4**	**3**	**5**	**S5**	**1**	**S2**	**2/S3**	**S3**

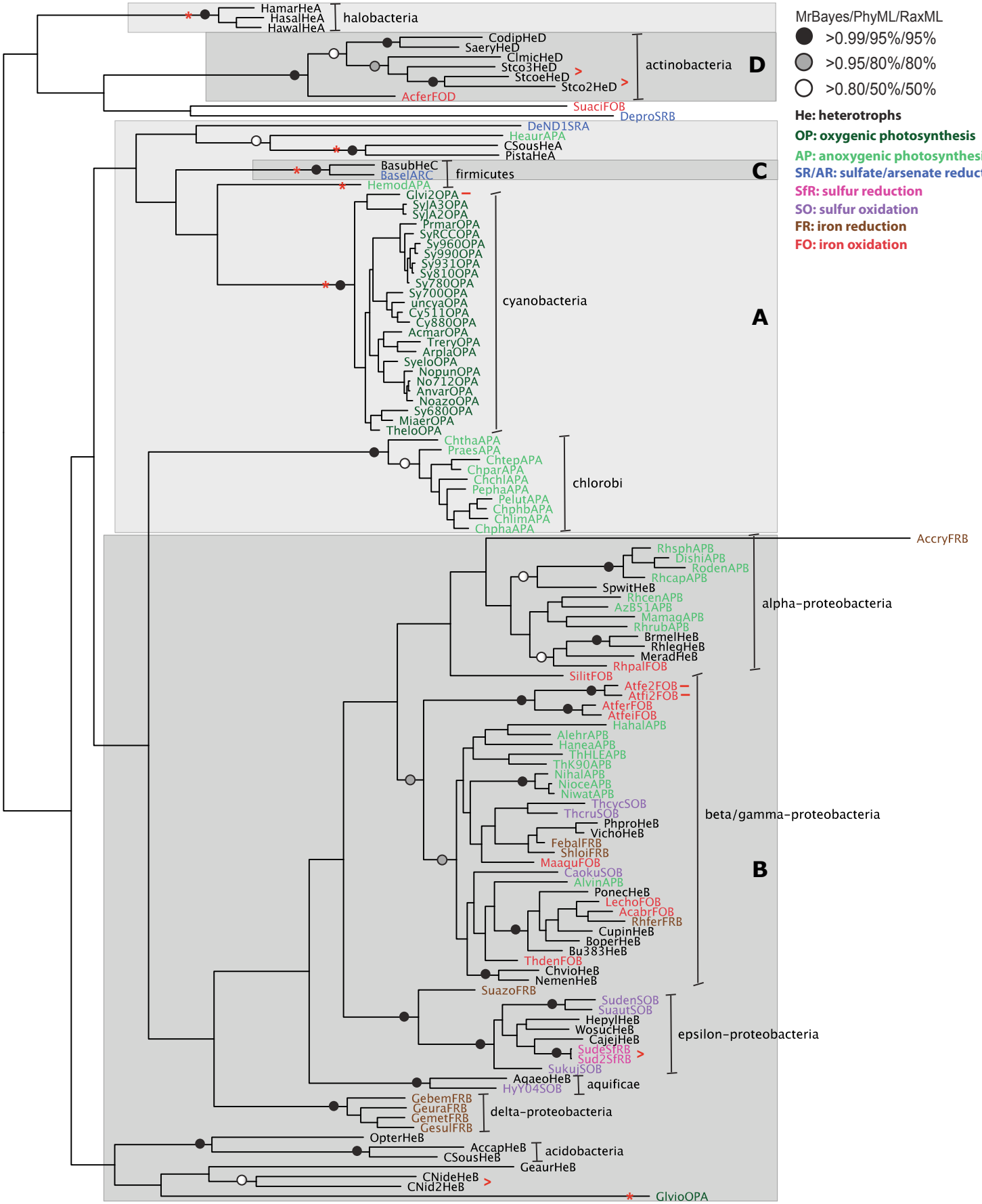
* sequence set from Figure 1 of Dibrova2013
** dataset also includes eukaryotic cytb561 sequences from orthology group K08360
*** only one sequence (from *E. coli*) in our set of species

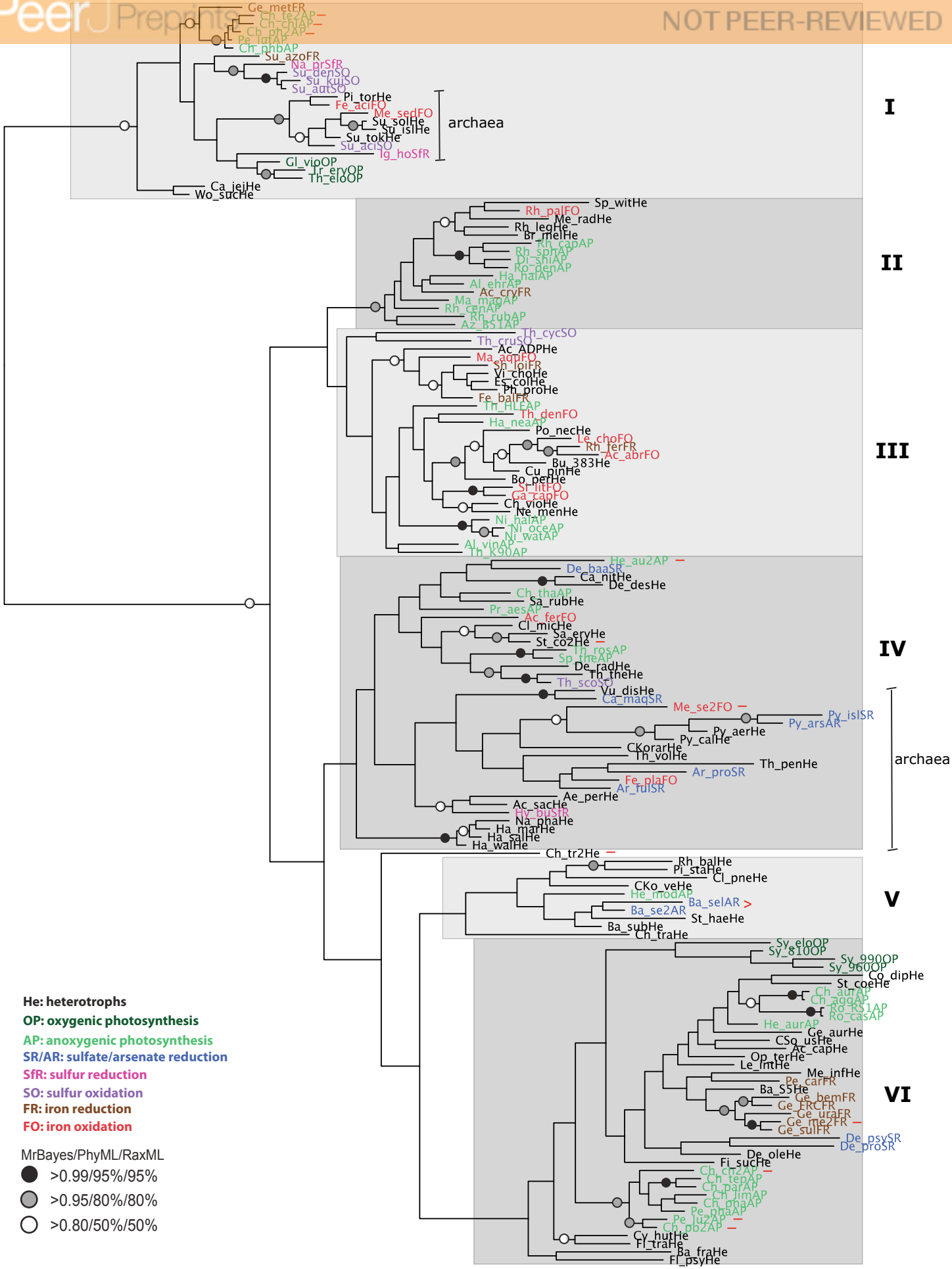
pfam domains
1 = Cytochrom_B_N_2 (PF13631)
2 = Cytochrom_B_N (PF00033)
3 = Ni_hydr_CYTB (PF01292)
4 = Cytochrom_B_C (PF00032)
5 = Cytochrom_B561 (PF03188)
6 = Cyt-b5 (PF00173)
7 = Sdh_cyt (PF01127)

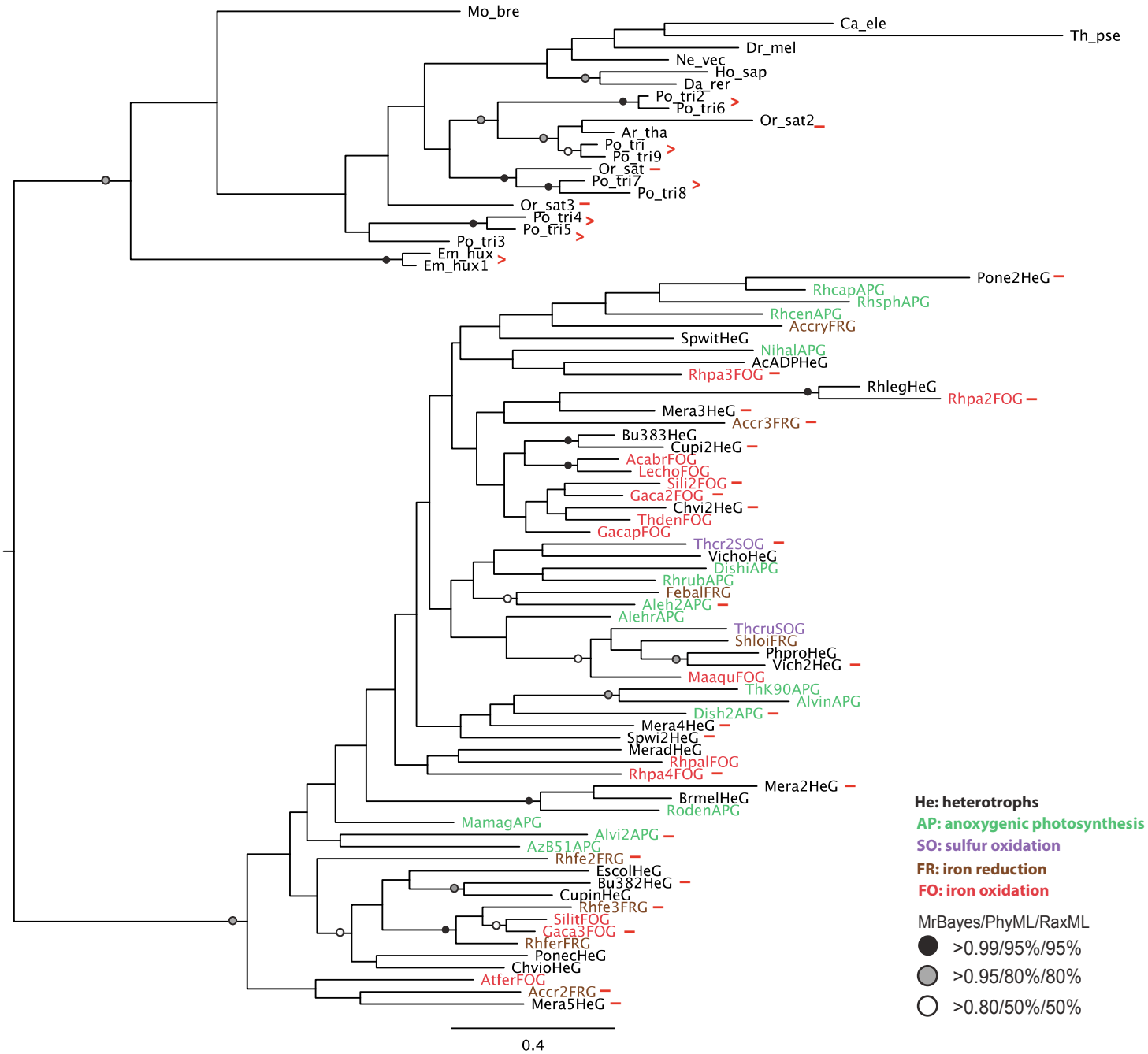


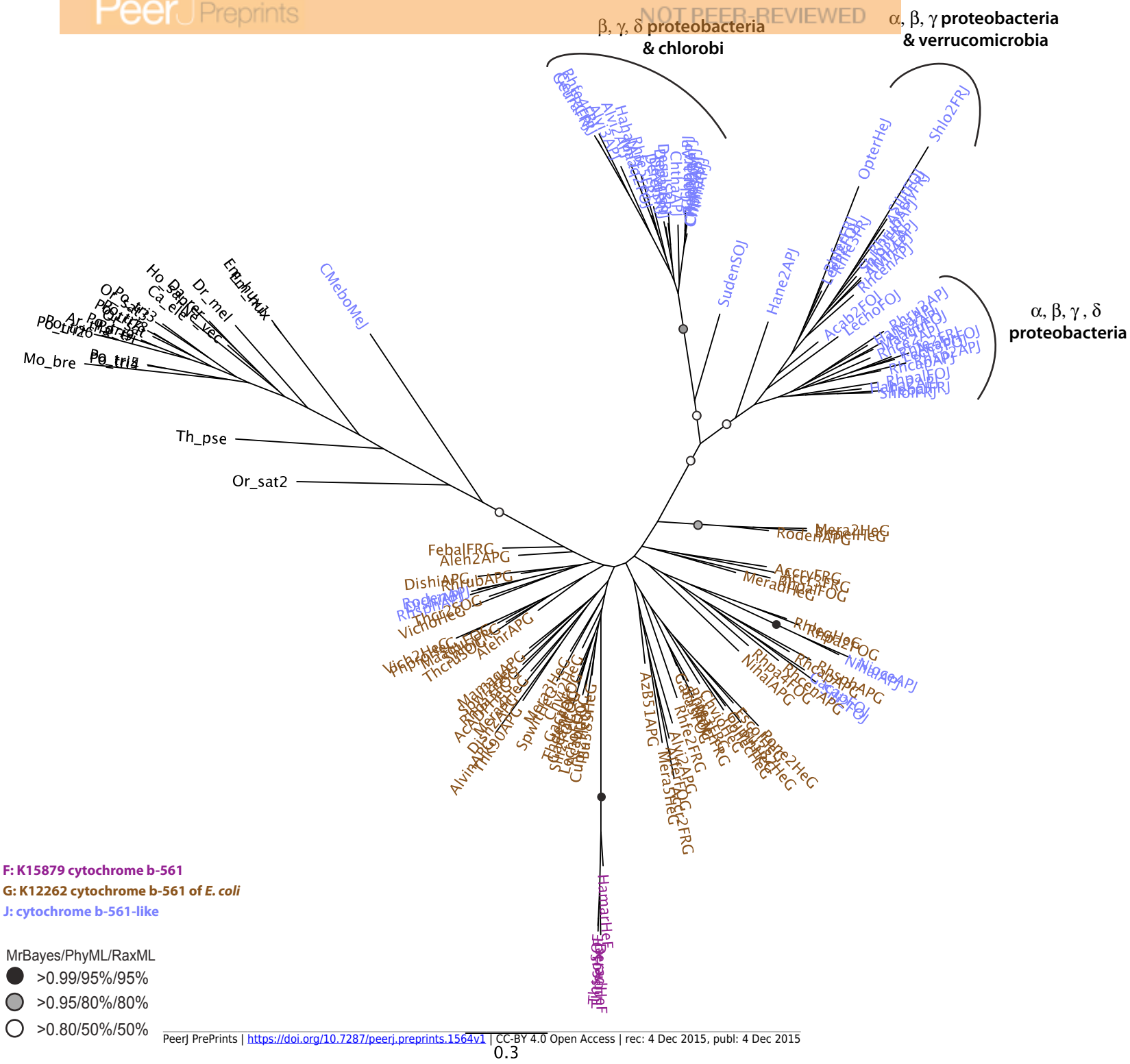
enzyme	Gene name	Orthology group	α-proteobacteria (345)	β-proteobacteria (237)	γ-proteobacteria (739)	δ-proteobacteria (67)	ε-proteobacteria (142)	gemmatimonadetes (2)	acidobacteria (8)	chlorobi (13)	bacteroidetes (140)	spirochaetes - leptospira (8)	verrucomicrobia (6)	deinococci (21)	chloroflexi (24)	cyanobacteria (85)	actinobacteria (396)	firmicutes - clostridia (168)	firmicutes - bacilli (559)	fusobacteria (10)	tenericutes - mollicutes (113)	nitrospirae (5)	aquificae (13)	dictyoglomi (2)	thermotogae (25)	chlamydiae (116)	spirochaetes - borrelia (26)	elusimicrobia (3)	planctomycetes (7)	fibrobacteres (2)	chrysiogenetes (1)	deferribacteres (4)
cytochrome b6	petB	K02635[a]							+	+						+		+														
	petD	K02637[b]														+		+														
ubiquinol-cytochrome c reductase	cytB	K00412[c]	+	+	+	+	+	+	+		+		+	+			+	+	+				+	+					+			
	fbcH	K00410	+																													
menaquinol-cytochrome c reductase	MQCRB	K03887															+		+				+									
	MQCRC	K03888															+		+													
ubiquinol-cytochrome c reductase	qcrB	K03891															+															
cytochrome b561	narC	K15879[d]															+															
	cybB	K12262	+	+	+	+	+		+		+					+	+															
formate dehydrogenase	fdoI	K00127[e]	+	+	+	+	+	+	+				+				+	+	+					+								+
	fdnI	K08350			+																											
Ni/Fe-hydrogenase	hyaC	K03620[f]	+	+	+	+	+		+	+	+		+		+		+	+	+				+	+							+	
succinate dehydrogenase	sdhC	K00241[g]	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+		+			+	+	+	+
		Total cyt-b types per lineage	7	6	7	6	6	3	6	3	4	1	4	2	2	6	10	7	6	0	0	4	4	0	0	1	0	0	2	1	2	2

a: also found in plants, diatoms, Methylomirabilis, Caldiarchaeum
b: also found in plants, diatoms
c: also found in animals, plants, fungi, amoebozoa, alveolata, Fimbriimonas, Thermodesulfobacteria, Thaumarchaeota
d: also found in Euryarchaeota (Haloarcula, Halorubrum, Haloferax, Halorhabdus, Halomicrobium, Halogeometricum)
e: also in Thermodesulfobacteria, Euryarchaeota (Ferroglobus, Natronobacterium), Crenarchaeota (Hyperthermus, Pyrobaculum, Thermoporteum, Thermofilum)
f: also in Thermodesulfobacteria
g: also in Thermobaculum, Halyomorpha symbiont, Euryarchaeota, Crenarchaeota, Korarchaeum






Mo_bre
Ca_ele
Th_pse
Dr_mel
Ne_vec
Ho_sap
Da_rer
Po_tri2
Po_tri6
Or_sat2
Ar_tha
Po_tri
Po_tri9
Or_sat
Po_tri7
Po_tri8
Or_sat3
Po_tri4
Po_tri5
Po_tri3
Em_hux
Em_hux1
Pone2HeG
RhcapAPG
RhsphAPG
RhcenAPG
AccryFRG
SpwitHeG
NihalAPG
AcADPHeG
Rhpa3FOG
RhlegHeG
Rhpa2FOG
Mera3HeG
Accr3FRG
Bu383HeG
Cupi2HeG
AcabrFOG
LechoFOG
Sili2FOG
Gaca2FOG
Chvi2HeG
ThdenFOG
GacapFOG
Thcr2SOG
VichoHeG
DishiAPG
RhrubAPG
FebalFRG
Aleh2APG
AlehrAPG
ThcruSOG
ShloiFRG
PhproHeG
Vich2HeG
MaaquFOG
ThK90APG
AlvinAPG
Dish2APG
Mera4HeG
Spwi2HeG
MeradHeG
RhpalFOG
Rhpa4FOG
Mera2HeG
BrmelHeG
RodenAPG
MamagAPG
Alvi2APG
AzB51APG
Rhfe2FRG
EscolHeG
Bu382HeG
CupinHeG
Rhfe3FRG
SilitFOG
Gaca3FOG
RhferFRG
PonecHeG
ChvioHeG
AtferFOG
Accr2FRG
Mera5HeG
0.4
He: heterotrophs
AP: anoxygenic photosynthesis
SO: sulfur oxidation
FR: iron reduction
FO: iron oxidation
MrBayes/PhyML/RaxML
>0.99/95%/95%
>0.95/80%/80%
>0.80/50%/50%



PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1564v1 | CC-BY 4.0 Open Access | rec: 4 Dec 2015, publ: 4 Dec 2015