A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land

Fabia U. Battistuzzi[1] *and S. Blair Hedges*

Department of Biology, Pennsylvania State University

Evolutionary trees of prokaryotes usually define the known classes and phyla but less often agree on the relationships among those groups. This has been attributed to the effects of horizontal gene transfer, biases in sequence change, and large evolutionary distances. Furthermore, higher level clades of prokaryote phyla rarely are supported by information from ecology and cell biology. Nonetheless, common patterns are beginning to emerge as larger numbers of species are analyzed with sophisticated methods. Here, we show how combined evidence from phylogenetic, cytological, and environmental data support the existence of an evolutionary group that appears to have had a common ancestor on land early in Earth's history and includes two-thirds of known prokaryote species. Members of this terrestrial clade (Terrabacteria), which includes Cyanobacteria, the gram-positive phyla (Actinobacteria and Firmicutes), and two phyla with cell walls that differ structurally from typical gram-positive and gram-negative phyla (Chloroflexi and *Deinococcus–Thermus*), possess important adaptations such as resistance to environmental hazards (e.g., desiccation, ultraviolet radiation, and high salinity) and oxygenic photosynthesis. Moreover, the unique properties of the cell wall in gram-positive taxa, which likely evolved in response to terrestrial conditions, have contributed toward pathogenicity in many species. These results now leave open the possibility that terrestrial adaptations may have played a larger role in prokaryote evolution than currently understood.

Introduction

The evolutionary history of prokaryotes has been intensely studied using DNA and protein sequences, gene content, and sequence signatures (e.g., Gupta 1998; Wolf et al. 2001; Brochier et al. 2002; Battistuzzi et al. 2004; Ciccarelli et al. 2006; Lienau et al. 2006). Although the monophyly of most classes and phyla is well resolved, no consensus has been reached on relationships among those groups, especially among phyla. Horizontal gene transfer (HGT) has been considered at least part of the reason for this phylogenetic uncertainty (Doolittle and Bapteste 2007), although a working model holds that the tree can be resolved with a set of core genes (proteins) having reduced levels of HGT (Choi and Kim 2007). Core proteins are those shared by a set of species for which a major influence of HGT can be excluded. Based on different HGT detection methods and species sets, this core protein approach has identified overlapping sets of 20–40 proteins from complete genomes that are shared by eubacteria (also called "Bacteria"), archaebacteria (also called "Archaea"), and eukaryotes (e.g., Battistuzzi et al. 2004; Charlebois and Doolittle 2004; Ciccarelli et al. 2006). However, phylogenetic studies using core proteins often have differed in major ways from analyses of ribosomal RNA (rRNA) genes, leading to an overall uncertainty in prokaryote phylogeny. Here, we conducted sequence analyses of both types of genes to search for common patterns and reconcile the differences.

For our primary analysis, we constructed a core protein tree with 25 protein-coding genes from 218 species. For comparison with the protein tree, we also built an rRNA tree, from 189 species, that combined sequences of the small subunit (SSU), the gene traditionally used for analyses, and the rarely used large subunit (LSU). We subjected these data sets to a suite of sequence analyses and identified a sequence bias in the rRNA data that, when corrected, brings the rRNA and protein trees into closer agreement than in past studies. The trees reveal a large clade of phyla comprising two-thirds of the 9,740 recognized species of prokaryotes, including all gram-positive species and most species that form spores. Together with environmental data from culture-independent studies and molecular clock analyses, we show that this clade likely evolved on land early in the Precambrian, with some lineages later reinvading marine habitats. These results have implications for understanding the relations between the key adaptations of the terrestrial clade and the environment in which they evolved.

Materials and Methods

Data Assembly and Sequence Analyses

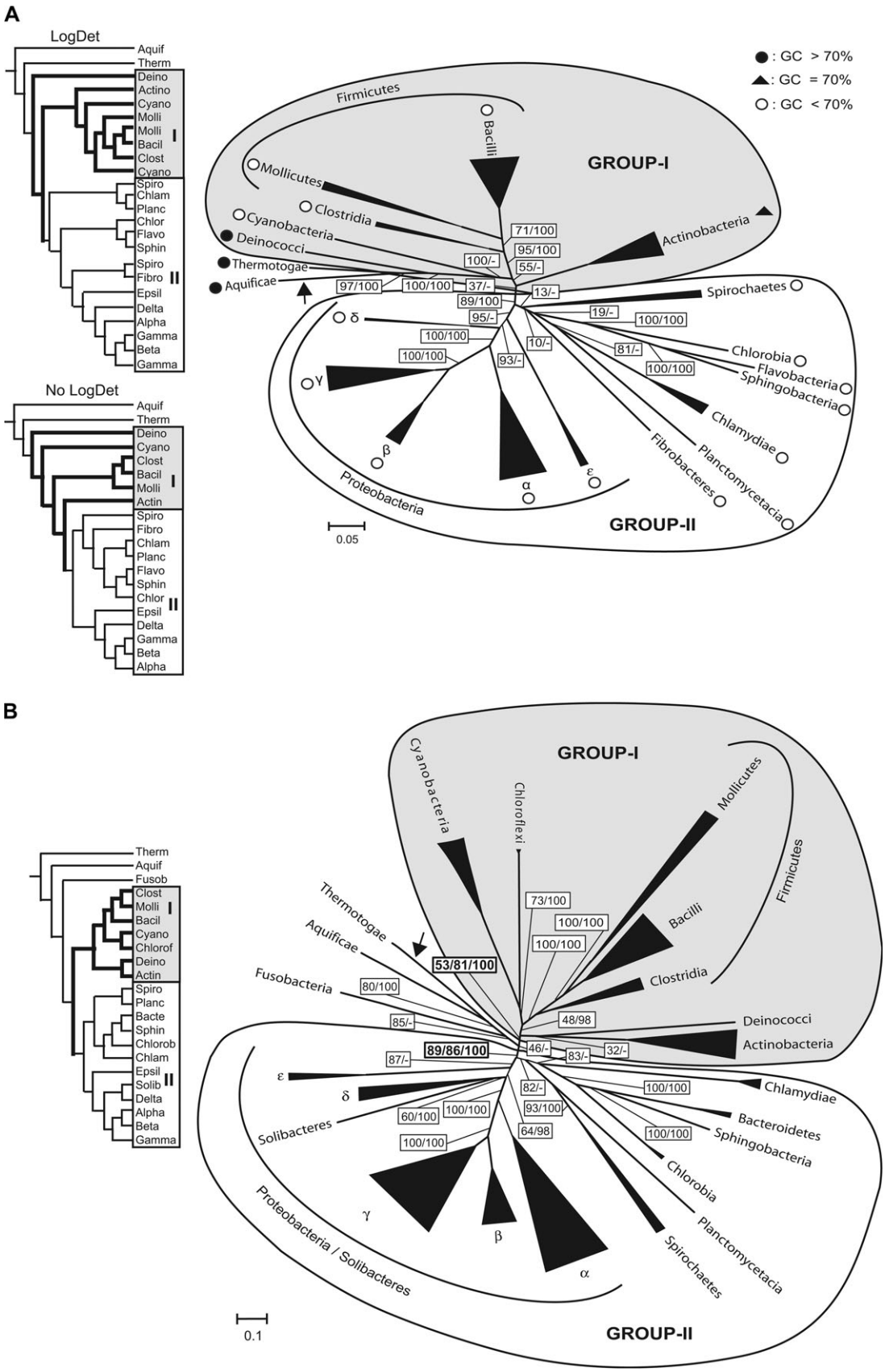
For our primary analysis, we constructed a protein tree with 25 protein-coding genes. These correspond to a subset of previously identified orthologous core proteins (Battistuzzi et al. 2004) that were used as queries for a similarity search (Altschul et al. 1997) against 311 fully sequenced genomes of Eubacteria and Archaebacteria (supplementary table S1, Supplementary Material online). Given the large number of species analyzed, a few species-specific gene losses are expected even in widely distributed genes. To maximize the number of protein-coding genes, 28 species showing such losses were omitted resulting in a data set of 283 species. In doing so, we created a complete matrix of genes and species and avoided any potential bias of missing data. We chose classes as our working taxonomic level because species of the same class are obtained in our and other phylogenies in highly supported monophyletic clusters (Ciccarelli et al. 2006; Pisani et al. 2007). The omitted species are members of monophyletic classes already represented. The retrieved sequences were aligned for each protein by ClustalX (Thompson et al. 1994). Distance and maximum likelihood (ML) single-protein phylogenies were built in the program MEGA4 (Tamura et al. 2007) (Neighbor-Joining, model JTT + gamma = 0.5, 1, and 1.5, complete deletion of gaps) and the program RAxML (Stamatakis 2006) (ML, model JTT + estimated gamma),

[1] Present address: Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University.

Key words: phylogeny, Bacteria, Archaea, Terrabacteria, Hydrobacteria, molecular clock.

E-mail: sbh1@psu.edu.

Mol. Biol. Evol. 26(2):335–343. 2009
doi:10.1093/molbev/msn247
Advance Access publication November 6, 2008


A
LogDet
Aquif
Therm
Deino
Actino
Cyano
Molli
Molli
Bacil
Clost
Cyano
Spiro
Chlam
Planc
Chlor
Flavo
Sphin
Spiro
Fibro
Epsil
Delta
Alpha
Gamma
Beta
Gamma
I
II
No LogDet
Aquif
Therm
Deino
Cyano
Clost
Bacil
Molli
Actin
Spiro
Fibro
Chlam
Planc
Flavo
Sphin
Chlor
Epsil
Delta
Gamma
Beta
Alpha
I
II
● : GC > 70%
▲ : GC = 70%
○ : GC < 70%
Firmicutes
Bacilli
Mollicutes
Clostridia
Cyanobacteria
Deinococci
Thermotogae
Aquificae
Actinobacteria
GROUP-I
71/100
95/100
100/-
55/-
97/100
100/100
37/-
13/-
89/100
95/-
19/-
100/100
δ
100/100
10/-
81/-
100/100
93/-
γ
β
α
ε
Spirochaetes
Chlorobia
Flavobacteria
Sphingobacteria
Chlamydiae
Planctomycetacia
Fibrobacteres
Proteobacteria
GROUP-II
0.05
B
Therm
Aquif
Fusob
Clost
Molli
Bacil
Cyano
Chlorof
Deino
Actin
Spiro
Planc
Bacte
Sphin
Chlorob
Chlam
Epsil
Solib
Delta
Alpha
Beta
Gamma
I
II
GROUP-I
Cyanobacteria
Chloroflexi
Mollicutes
Firmicutes
Bacilli
Clostridia
Thermotogae
Aquificae
Fusobacteria
73/100
100/100
100/100
53/81/100
80/100
85/-
48/98
Deinococci
Actinobacteria
89/86/100
87/-
46/-
83/-
32/-
ε
δ
Solibacteres
60/100
100/100
100/100
82/-
93/100
64/98
γ
β
α
Chlamydiae
Bacteroidetes
Sphingobacteria
Chlorobia
Planctomycetacia
Spirochaetes
Proteobacteria / Solibacteres
GROUP-II
0.1

respectively, to check for orthology and possible HGT events. Genes with nested domains (Eubacteria and Archaebacteria) and/or highly supported ($\geq$95% bootstrap) nesting of one class within another were considered as candidates for nonvertical inheritance and deleted from the data set.

The remaining genes (25) were concatenated in a final alignment of 18,586 amino acid sites. From this alignment, site homology was further refined (Castresana 2000) using monophyly of classes as an approximation of the strength of the phylogenetic signal in progressively reduced data sets (i.e., a stronger signal results in more monophyletic classes). Based on this analysis, nonconserved sites were omitted, resulting in a final concatenated alignment of 6,884 amino acids and 218 nonredundant (i.e., one strain per species) species, which were used in nonpartitioned and partitioned analyses. For comparison, we built a phylogeny with all available nonredundant species (189 total; 19 eubacterial classes, 10 archaebacterial classes) from the European Ribosomal RNA Database (see supplementary table S1 in Supplementary Material online). The initial rRNA alignment based on secondary structure (Wuyts et al. 2004) was modified to include only conserved sites using the same approach applied to proteins to select a threshold between number of sites and phylogenetic signal (Castresana 2000). The final alignment included a total of 3,786 conserved nucleotides (60% of the original alignment) from the concatenation of SSU and LSU rRNA genes. We made little modifications to the species composition of the rRNA alignments to preserve the original secondary structure alignment; only two species (*Methanopyrus kandleri* and *Nanoarchaeum equitans*) that were absent from the database were added because they represented additional classes.

Phylogenetic analyses of aligned sequences were conducted with ML and Bayesian methods (Ronquist and Huelsenbeck 2003; Stamatakis 2006) on partitioned data sets in order to allow the optimization of parameters for each gene. Phylogenetic confidence was estimated with 100 bootstrap replicates in the ML phylogeny and by posterior probability (PP) in the Bayesian approach. Additional analyses were carried out on the protein and rRNA data set with a method (Brinkmann and Philippe 1999) designed to identify slow-evolving sites. For the primary phylogenetic analyses, the root was set between eubacteria and archaebacteria, which is the current consensus based on duplicate gene evidence (Zhaxybayeva et al. 2005). In the rRNA analyses, we also used a modified version (Tamura and Kumar 2002) of the LogDet analysis (Lockhart et al. 1994) for modeling base compositional differences, as implemented in the program MEGA4 (Tamura and Kumar 2002); this was carried out on the complete data set with 100 bootstrap replicates.

Times of divergence were estimated using the protein and rRNA data sets separately, ML phylogenies, and three methods: nonparametric rate smoothing (Sanderson 1997), penalized likelihood (Sanderson 1997), and Bayesian analysis (partitioned and nonpartitioned data sets) (Thorne and Kishino 2002). Separate analyses were carried out with eubacteria and archaebacteria using reciprocal rooting. Branch lengths were estimated with a JTT + gamma model for the protein data set and Felsenstein 84 (F84) model (Kishino and Hasegawa 1989; Felsenstein and Churchill 1996) with estimation of gamma distribution and transition/transversion ratio for the rRNA data set; this was accomplished with the programs Estbranches (Thorne and Kishino 2002) and PAML (Yang 1997). We used six calibration points from the geologic and biomarker records, including the earliest habitable time at 4.2 billion years ago (Ga) based on ocean-boiling impact probabilities (such impacts also may have occurred as late as 3.8 Ga during the late heavy bombardment) (Sleep et al. 1989; Zahnle et al. 2007), earliest continents at 4.0 Ga (Rosing et al. 2006), earliest methanogens at 3.46 Ga (Bapteste et al. 2005; Ueno et al. 2006), earliest oxygen at 2.3 Ga (Holland 2002), divergence of Chlorobia and Bacteroidetes at 1.64 Ga (Brocks et al. 2005), and of Gammaproteobacteria and Betaproteobacteria at 1.64 Ga (Brocks et al. 2005). Additional details on parameter specifications for each analysis are in the Supplementary Material online.

Species Counts

A list of validly published bacterial names was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ), the German Collection of Microorganisms and Cell Culture (www2.dsmz.de). From this list, all subspecies and synonymous names were removed to obtain a total count of prokaryote species. Cyanobacteria were not included in the DSMZ list because they have been historically associated with algae in taxonomic treatments. We retrieved information regarding this phylum from Algaebase (www.algaebase.org). Furthermore, we integrated the genera listed in DSMZ with those present in National Center for Biotechnology Information

←

FIG. 1.—Unrooted ML phylogenies of the rRNA tree (*A*) and protein tree (*B*) for Eubacteria. Each panel has an inset showing the relationship of the trees rooted with Archaebacteria. Insets in panel *A* show phylogenies before (no LogDet) and after (LogDet) the correction for compositional biases. Triangles on branches are proportional to the number of sequences analyzed within each lineage (total = 189 and 218, respectively). ML confidence values (left of slash) and Bayesian PPs are shown at each node; nodes supporting the two major groups in (*B*) are bold, with middle support value from ML analysis of slow-evolving sites. Filled circles next to clade name in (*A*) indicate >70% GC content of the conserved sites for each lineage; filled triangle indicates 70%; open circles indicate <70%. Dashes represent groups not present in the Bayesian phylogeny. The Greek letters indicate the five classes of the phylum Proteobacteria. Lineages in insets are abbreviated. Actino, Actinobacteria; Alpha, Alphaproteobacteria; Aquif, Aquificae; Bacil, Bacilli; Bacte, Bacteroidetes; Beta, Betaproteobacteria; Chlam, Chlamydiae; Chlor, Chlorobia; Chlorof, Chloroflexi; Clost, Clostridia; Cyano, Cyanobacteria; Deino, *Deinococcus–Thermus*; Delta, Deltaproteobacteria; Epsilon, Epsilonproteobacteria; Fibro, Fibrobacteres; Flavo, Flavobacteria; Fusob, Fusobacteria; Gamma, Gammaproteobacteria; Molli, Mollicutes; Planc, Planctomycetacia; Solib, Solibacteres; Spiro, Spirochaetes; Sphin, Sphingobacteria; and Therm, Thermotogae. Some classes appear multiple times in the tree because their representative species are nonmonophyletic. The arrow points to the root.

(http://www.ncbi.nlm.nih.gov/) (e.g., Dehalococcoides). A breakdown of the number of species in each major category is given in supplementary table S3 (Supplementary Material online).

Environmental Evidence

Information on the natural habitat of families or single genera was retrieved from the literature. Lineages were categorized as terrestrial if their known habitat is strictly nonmarine (e.g., soil or rock on continents), freshwater (e.g., lakes, rivers, and springs), or if their host is a nonmarine species. Marine lineages have their primary habitat in saltwater environments (e.g., sea surface, water column, sea floor, and deep-sea vent) or are associated with marine hosts. ML family-level phylogenies for each of the classes Actinobacteria, Cyanobacteria, and Deinococci were estimated from an SSU alignment (secondary structure) using one representative per family. One member of each of the other classes in the terrestrial clade (Group I) was used as outgroup. The class-level phylogeny of Firmicutes (fig. 1*B*) and an existing phylogeny of Chloroflexi (Costello and Schmidt 2006) were used. The habitat assignments of the lineages and of the common ancestor were estimated with maximum parsimony (MP) and ML (Maddison WP and Maddison DR 1989, 2008). Evidence supporting Groups I and II was drawn from phylogenetic analyses (this study) and the literature for gram staining and spore production (Holt 1984; Garrity 2001). For quantitative estimates of Group I versus Group II sequences from different environments (table 1), only culture-independent studies were considered, to avoid biases introduced by culturing methods, although other biases may be present. Information for four diverse habitat classifications was retrieved from the literature: 1) deep sea (Tringe et al. 2005; Sogin et al. 2006; Huber et al. 2007; Lauro and Bartlett 2008), 2) sea surface (DeLong 2005; Rusch et al. 2007), 3) humid soils (Tringe et al. 2005; Roesch et al. 2007; Aislabie et al. 2008), and 4) arid (warm and dry) soils (Chanal et al. 2006; Connon et al. 2007). Additional details are available in the Supplementary Material online.

Results and Discussion

Phylogenetic Evidence

The ML phylogeny obtained with the concatenated data set of SSU and LSU rRNA genes from 189 species (fig. 1*A*) is similar to earlier SSU-only phylogenies in identifying a single large group of classes and phyla, supported here by 89% ML bootstrap probability (BP) and 100% Bayesian PP. The group contains Bacteroidetes, Chlamydiae, Chlorobia, Fibrobacteres, Planctomycetacia, Proteobacteria, and Spirochaetes. The tree was rooted with Archaebacteria and the remaining classes stem in a ladderlike fashion from the rooted tree (fig. 1*A*, insets). The hyperthermophilic classes Aquificae and Thermotogae are the most basal branches, followed by *Deinococcus–Thermus* and Cyanobacteria. An ML phylogeny built from an alignment with only slow-evolving sites and a Bayesian analysis of all sites both formed the identical large group of classes and phyla and showed the same topology at the base of the tree. Furthermore, they differed only at nodes that were poorly supported in both trees (see Supplementary Material online).

The protein tree (fig. 1*B*) is similar to the rRNA tree in supporting the same cluster of classes and phyla, at 89% BP and 100% PP. It differs from the rRNA tree in placing all other eubacteria, except for the hyperthermophiles and Fusobacteria, in an even larger group (Group I), supported by 53% BP and 100% PP, rather than in a stepwise branching order near the root. Members of Group I include the phyla Actinobacteria, Chloroflexi, Cyanobacteria, *Deinococcus–Thermus*, and Firmicutes. An ML phylogeny built from an alignment with only slow-evolving sites was identical and showed increased support for Group I (81% BP) (fig. 1*B*). Trees showing similar major groupings of phyla have been found in the past (Gupta and Johari 1998; Brochier et al. 2002; Wolf et al. 2002; House et al. 2003; Battistuzzi et al. 2004; Lienau et al. 2006) indicating stability with increased taxon sampling and application of diverse methods. Nonetheless, most relationships of the phyla within Group I and the other smaller group (Group II) remain uncertain.

Although the rooted versions of the two trees (rRNA and protein tree) are different in the order of their earliest branches (fig. 1, insets), the overall similarity of the unrooted trees suggested that a base compositional bias present in the rRNA sequences might explain the difference, especially given the high GC ratio of SSU and LSU in taxa near the root of the rRNA tree (Deinococci, Aquificae, and Thermotogae; fig. 1*A*). When methods designed to compensate for such biases have been used on rRNA gene data in the past (Brochier et al. 2002), they did not fully reproduce Group I but nonetheless supported major components of Group I. For example, the high GC taxon of Group I, *Deinococcus–Thermus*, that typically clusters with other high GC taxa (hyperthermophiles) near the root instead clustered with the Group I taxon Cyanobacteria (Brochier et al. 2002).

When we used a nucleotide substitution model (Tamura and Kumar 2002) to compensate for compositional biases in the combined SSU–LSU rRNA data set, all components of Group I were obtained (69% BP) except *Deinococcus–Thermus*. Group II was also obtained, albeit with a lower support (41% BP) (fig. 1*A* and Supplementary Material online). Nonetheless, the deep position of the high GC *Deinococcus–Thermus* lineage probably reflects the susceptibility of rRNA data sets to compositional biases even when ameliorating methods are applied. As is typical of most sequence analyses of these deeply divergent groups (Brochier et al. 2002), none of these trees are strongly supported, except with Bayesian PPs. Although further resolution and support of the GC-bias hypothesis may not be possible, this evidence suggests that it has affected several key nodes in the prokaryote rRNA phylogeny, placing greater emphasis on the protein phylogeny (fig. 1*B*). Despite the small number of nodes affected in the rRNA phylogeny, it appears to have delayed general recognition of a major evolutionary clade, Group I.

The deepest (most basal) nodes in the protein and rRNA trees are occupied by the hyperthermophiles, Groups

Table 1
Multiple Evidence Supporting Two Major Groups of Eubacteria (Groups I and II)

	Phylogeny				Environmental Surveys			
Phylum or Lineage	Protein	rRNA	Gram Stain	Spores	Deep Sea	Sea Surface	Humid Soils	Arid Soils
Actinobacteria	I	I	P	Yes	5%	1%	13%	64%
Chloroflexi	I	—	P/N	No	4%	1%	5%	1%
Cyanobacteria	I	I	N	Yes	<1%	6%	4%	—
Deinococcus–Thermus	I	III	P	No	—	<1%	<1%	1%
Firmicutes	I	I	P	Yes	2%	6%	6%	1%
Group I, total (min–max)					12% (0–23%)	14% (7–20%)	28% (7–41%)	67% (32–99%)
Acidobacteria	II	—	N	No	<1%	—	13%	1%
Bacteroidetes	II	II	N	No	8%	9%	19%	2%
Chlamydiae	II	II	N	No	—	—	—	—
Chlorobi	II	II	N	No	—	—	—	—
Fibrobacteres	—	II	N	No	—	—	—	—
Planctomycetes	II	II	N	No	1%	13%	<1%	1%
Proteobacteria	II	II	N	No	79%	64%	40%	22%
Spirochaetes	II	II	N	No	—	<1%	—	—
Group II, total (min-max)					88% (77–100%)	86% (80–93%)	72% (59–93%)	33% (1–68%)
Fusobacteria	I/III	—	N	No	—	—	—	—
Aquificae	IV	V	N	No	—	—	—	—
Thermotogae	V	IV	N	No	—	—	—	—

NOTE.—P, gram-positive stain; N, gram-negative stain; *Deinococcus–Thermus* stains P but has a cell wall structurally similar to that of gram-negative taxa. Percentages refer to average taxonomic composition of sequences across multiple geographic sites; see Supplementary Material online for references. Spores in Proteobacteria are confined to one order in the Deltaproteobacteria. Dashes indicate that no data were available.

IV and V (Aquificae and Thermotogae), a position that has been criticized based mostly on compositional biases dictated by their lifestyle (Brochier and Philippe 2002). However, contrary to previous phylogenies (Brochier and Philippe 2002; Ciccarelli et al. 2006; Pisani et al. 2007), the use of multiple methods to compensate for this and other biases (e.g., analysis of only slow-evolving sites) did not change the phylogenetic position of these two lineages in either the protein or rRNA trees, increasing the confidence in an early origin of the hyperthermophiles. The phylum Fusobacteria (Group I/III) appears in the protein tree of eubacteria basal to Groups I and II and above the hyperthermophiles. Although this lineage has generally been considered a close relative of Firmicutes (Mira et al. 2004), alternative positions have been found, often associated with hyperthermophiles, in large phylogenetic studies (Gupta 2003; Ciccarelli et al. 2006; Pisani et al. 2007). Furthermore, in a Bayesian analysis of the protein data set, Fusobacteria is placed within Group I with 100% PP. Based on this phylogenetic evidence and on the extensive HGT history of this lineage (Mira et al. 2004), the position of Fusobacteria remains uncertain.

Organismal Evidence

The cytological and physiological characteristics of eubacteria (table 1) lend support to the recognition of these two major groups. Group I phyla Actinobacteria and Firmicutes (including the classes Bacilli, Clostridia, and Mollicutes) are gram positive and as such have a thick peptidoglycan layer; they also include mostly terrestrial taxa (see below). Group II (ancestrally marine, see below) includes most of the gram-negative taxa, many of which are also terrestrial. These include members of Proteobacteria, Acidobacteria, and the Cytophaga–Flavobacteria–Bacteroidetes group (Connon et al. 2007). However, experiments have shown that gram-negative species that are terrestrial decrease in abundance after soil drying, whereas gram positives (Actinobacteria and Firmicutes) increase (Rokitko et al. 2001), suggesting an ancestral function (desiccation resistance) of the peptidoglycan layer. Furthermore, the gram-positive taxa and Cyanobacteria produce resting stages (e.g., spores), albeit not evolutionarily related, which confer resistance to multiple stresses typical of terrestrial habitats such as desiccation, ultraviolet radiation, and high salt concentration (Potts 1994; Nicholson et al. 2000). Only one other type of spore is known in prokaryotes and it is constrained to one order (i.e., derived) within the Group II Class Deltaproteobacteria (Myxococcales) (Nicholson et al. 2000).

There is confusion in the literature over the number of described species of prokaryotes. Often, the number reported is approximately 6,000 (Oren 2004) but our preliminary survey showed this number to be an underestimate by as much as 30–40%. We found that there are 9,740 recognized species of prokaryotes, of which Group I comprises 63% and Group II comprises 33%. The most species-rich lineages are Actinobacteria and Cyanobacteria (Group I) and Gammaproteobacteria (Group II), with more than 1,000 known species in each taxon (Supplementary Material online). Many pathogens of humans and other terrestrial eukaryotes are gram positive and therefore are members of Group I (Holt 1984; Fischetti et al. 2006). The structural characteristics of gram-positive prokaryotes, such as the lack of an outer membrane and presence of a thick peptidoglycan layer, have led to novel adaptations for pathogenicity including unique surface proteins, toxins, and enzymes (Fischetti et al. 2006). Thus, aspects of their pathogenicity are probably related to a terrestrial ancestry, either directly or indirectly. Similarly, radiation tolerance of *Deinococcus* is likely related to selection for desiccation tolerance (Mattimore and Battista 1996).

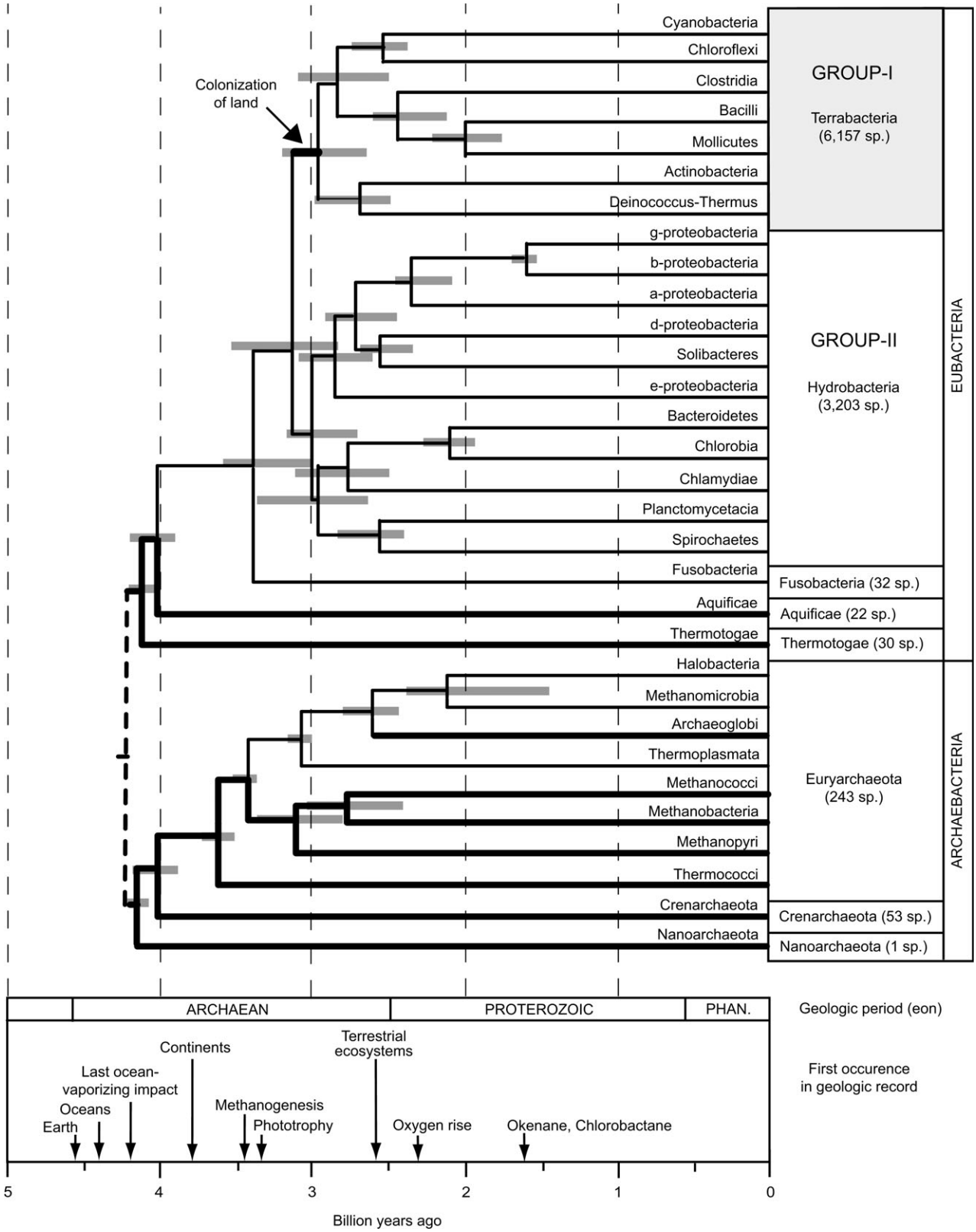


FIG. 2.—Timescale of prokaryote evolutionary history. The timetree shows divergences for Eubacteria and Archaebacteria (ML, protein data set) with particular attention to major groups: Hydrobacteria and Terrabacteria (Eubacteria) and Euryarchaeota and Crenarchaeota (Archaebacteria). First occurrences of major events in the geologic record are represented by arrows on the timescale. The timescale is in billion years ago. Each horizontal line represents a class; exceptions are the phyla Bacteroidetes (which includes two classes), Cyanobacteria, and Nanoarchaeota. Thicker lines are lineages that include hyperthermophilic species. Gray bars show the range of time estimates for each node, from each of the four estimation methods. For source of species counts and methods, see Supplementary Material online.

Environmental Evidence

The environment occupied by species in these two groups is consistent with the evolution of desiccation-resistant traits in Group I. Culture-independent sampling of prokaryotes, including metagenomic studies, shows that marine samples have the lowest fraction of Group I taxa and that continental (terrestrial) samples have the highest fraction (table 1). At the extremes of the marine and terrestrial environments, some deep-sea sampling (Tringe et al. 2005) reveals a virtual absence (0–1%) of Group I sequences whereas hyperarid desert samples are comprised almost exclusively (99%) of Group I sequences (Connon et al. 2007). Near-surface marine samples (Rusch et al. 2007) have on average a higher fraction (14%) of Group I sequences than those from the deep sea, and samples of arid soils (Chanal et al. 2006) usually have a higher fraction than those of humid soils (Tringe et al. 2005). Viral communities also parallel this pattern, with viruses of Group I species dominating terrestrial samples and those of Group II dominating marine samples (Fierer et al. 2007). Despite these general trends, the composition of soil communities is phylogenetically and structurally complex, with different phyla dominating based on the location, type, and structure of the soil (Mummey et al. 2006).

Ancestor analysis provides additional support by showing that the earliest branching lineages of each phylum in Group I are terrestrial (supplementary figs. S6 and S7, Supplementary Material online). In agreement with previous studies, these include Gloeobacteria (Cyanobacteria) and Rubrobacteriales (Actinobacteria) which are found exclusively in terrestrial environments (Stackebrandt et al. 1997; Ludwig and Klenk 2001; Seo and Yokota 2003; Gao et al. 2006; Tomitani et al. 2006; Kunisawa 2007) and most of Clostridia (Firmicutes) which inhabit soil or are parasites of terrestrial hosts. There are only three known families in *Deinococcus–Thermus*; two of them (Deinococcaceae and Trueperaceae) are terrestrial and the third contains both marine and terrestrial species. Finally, terrestriality is widespread in the Phylum Chloroflexi with evidence of the earliest branches living in terrestrial habitats (Costello and Schmidt 2006). Parsimony and ML ancestral state reconstructions show support (MP: 100%, ML: 73%) for a terrestrial habitat preference in the ancestor of Group I. Although the natural habitat and distribution of most species of prokaryotes is not well known, the combined evidence from phylogenetic, organismal, and environmental analyses supports a terrestrial origin of Group I (table 1).

For Group I, the appropriate name Terrabacteria is available, previously applied to a subset of phyla (Actinobacteria, Cyanobacteria, and *Deinococcus–Thermus*) in a study involving fewer sequences (Battistuzzi et al. 2004). The current analysis differs in defining a larger land clade (expanded to include Bacilli, Chloroflexi, Clostridia, and Mollicutes), reconciling rRNA and protein tree differences, and integrating cytological and environmental data. Fusobacteria may be an additional member of Terrabacteria because its position varied from below the major Group I/Group II split in the ML protein tree (weakly supported) to within Group I in the Bayesian tree (strongly supported). Members of Group II occupy diverse environments from marine to terrestrial (Madigan et al. 2003). However, the limited ecological information indicates that terrestrial adaptations of Group II are mostly restricted to low taxonomic levels (species and genera) rather than higher (derived) levels. This would suggest an aquatic ancestor for this group as a whole; and thus, we propose the name Hydrobacteria (from the Greek, *hydro*, water) in allusion to the moist environment inferred for the common ancestor of these species. Although specific environments appear to have influenced the early evolutionary history of each of the two major groups, many descendant species living today are adapted to other environments.

Early Evolution

The earliest evidence of life in the fossil record is from marine environments, 3.5 Ga (Schopf et al. 2007), whereas ancient soils from South Africa (2.6 Ga) record the earliest terrestrial ecosystems (Watanabe et al. 2000). Later in the Precambrian, there is abundant evidence of terrestrial life (Horodyski and Knauth 1994; Schwartzman 1999). To better constrain the timing of the colonization of land, we estimated divergence times among lineages using Bayesian and ML methods. The divergence of Terrabacteria and Hydrobacteria was estimated to have occurred in the mid-Archean, 3.18 Ga (2.83–3.54 Ga) (fig. 2), which is consistent with both the origin of continents that occurred earlier (4.0–3.8 Ga) (Hawkesworth and Kemp 2006; Rosing et al. 2006) and the first evidence of terrestrial ecosystems that occurred later (2.6 Ga). Alternatively, assuming that the Earth's surface was not habitable until as late as 3.8 Ga (instead of 4.2 Ga), the resulting estimates are ~4% to 5% younger. A recent study on the effects of UV fluxes for terrestrial life (Cockell and Raven 2007) suggests that colonization of land was possible even before the establishment of a protective ozone layer. This scenario agrees with our evolutionary hypothesis of a land clade (Terrabacteria) in which Cyanobacteria and, thus, oxygenic photosynthesis (Raymond and Blankenship 2008) evolved after the colonization of land (3.54–2.66 Ga). Although it is too soon to conclude that all of the major adaptations of Terrabacteria—including oxygenic photosynthesis and resistance to environmental hazards—necessarily evolved on land, these results now leave open the possibility that terrestrial adaptations may have played a larger role in prokaryote evolution than currently understood.

Supplementary Material

Supplementary methodological information, tables S1–S4, and figures S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank J. G. Ferry, J. F. Kasting, S. Kumar, J. Macalady, and H. Ohmoto for discussion. This work was supported by grants from the National Science Foundation and National Aeronautics and Space Administration (to S.B.H.).

Literature Cited

Aislabie JM, Jordan S, Barker GM. 2008. Relation between soil classification and bacterial diversity in soils of the Ross Sea region, Antarctica. Geoderma. 144:9–20.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bapteste E, Brochier C, Boucher Y. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. Archaea. 1:353–363.

Battistuzzi FU, Feijão A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol. 4:44.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol. 16:817–825.

Brochier C, Babteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18:1–5.

Brochier C, Philippe H. 2002. Phylogeny—a non-hyperthermophilic ancestor for bacteria. Nature. 417:244.

Brocks JJ, Love GD, Summons RE, Knoll AH, Logan GA, Bowden SA. 2005. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. Nature. 437:866–870.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chanal A, Chapon V, Benzerara K, Barakat M, Christen R, Achouak W, Barras F, Heulin T. 2006. The desert of tataouine: an extreme environment that hosts a wide diversity of microorganisms and radiotolerant bacteria. Environ Microbiol. 8:514–525.

Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. 14:2469–2477.

Choi IG, Kim SH. 2007. Global extent of horizontal gene transfer. Proc Natl Acad Sci USA. 104:4489–4494.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Cockell CS, Raven JA. 2007. Ozone and life on the Archaean Earth. Philos Trans A Math Phys Eng Sci. 365:1889–1901.

Connon SA, Lester ED, Shafaat HS, Obenhuber DC, Ponce A. 2007. Bacterial diversity in hyperarid Atacama Desert soils. J Geophys Res. 112:G04S17.

Costello EK, Schmidt SK. 2006. Microbial diversity in alpine tundra wet meadow soil: novel Chloroflexi from a cold, water-saturated environment. Environ Microbiol. 8:1471–1486.

DeLong EE. 2005. Microbial community genomics in the ocean. Nature Rev Microbiol. 3:459–469.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 104:2043–2049.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 13:93–104.

Fierer N, Breitbart M, Nulton J, et al. (13 co-authors). 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol. 73:7059–7066.

Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Rood JI. 2006. Gram-positive pathogens. Washington (DC): ASM Press.

Gao B, Paramanathan R, Gupta RS. 2006. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. Antonie Van Leeuwenhoek. 90:69–91.

Garrity GM. 2001. Bergey's manual of systematic bacteriology, 2nd ed. New York: Springer.

Gupta RS. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev. 62:1435–1491.

Gupta RS. 2003. Evolutionary relationships among photosynthetic bacteria. Photosynth Res. 76:173–183.

Gupta RS, Johari V. 1998. Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. J Mol Evol. 46:716–720.

Hawkesworth CJ, Kemp AIS. 2006. Evolution of the continental crust. Nature. 443:811–817.

Holland HD. 2002. Volcanic gases, black smokers, and the great oxidation event. Geochim Cosmochim Acta. 21:3811–3826.

Holt JG. 1984. Bergey's manual of systematic bacteriology, 1st ed. Baltimore (MD): Williams & Wilkins.

Horodyski RJ, Knauth LP. 1994. Life on land in the Precambrian. Science. 263:494–498.

House CH, Runnegar B, Fitz-Gibbon ST. 2003. Geobiological analysis using whole genome-based tree building applied to the bacteria, archaea and eukarya. Geobiology. 1:15–26.

Huber JA, Mark Welch D, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. Science. 318:97–100.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 29:170–179.

Kunisawa T. 2007. Gene arrangements characteristic of the phylum Actinobacteria. Antonie Van Leeuwenhoek. 92:359–365.

Lauro FM, Bartlett DH. 2008. Prokaryotic lifestyles in deep sea habitats. Extremophiles. 12:15–25.

Lienau EK, DeSalle R, Rosenfeld JA, Planet PJ. 2006. Reciprocal illumination in the gene content tree of life. Syst Biol. 55:441–453.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.

Ludwig W, Klenk H-P. 2001. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholtz RW, editors. Bergey's manual of systematic bacteriology. Berlin (Germany): Springer-Verlag. p. 49–65.

Maddison WP, Maddison DR. 1989. Interactive analysis of phylogeny and character evolution using the computer program MacClade. Folia Primatol (Basel). 53:190–202.

Maddison WP, Maddison DR. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5. [Internet] Available from: http://mesquiteproject.org.

Madigan MT, Martinko JM, Parker J. 2003. Brock biology of microorganisms. Saddle River (NJ): Prentice-Hall Inc.

Mattimore V, Battista JR. 1996. Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. J Bacteriol. 178:633–637.

Mira A, Pushker R, Legault BA, Moreira D, Rodriguez-Valera F. 2004. Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics. BMC Evol Biol. 4:50.

Mummey D, Holben W, Six J, Stahl P. 2006. Spatial stratification of soil bacterial populations in aggregates of diverse soils. Microbial Ecol. 51:404–411.

Nicholson WL, Munakata N, Horneck G, Melosh HJ, Setlow P. 2000. Resistance of *Bacillus* endospores to extreme terrestrial

and extraterrestrial environments. Microbiol Mol Biol Rev. 64:548–572.
Oren A. 2004. Prokaryote diversity and taxonomy: current status and future challenges. Philos Trans Roy Soc B Biol Sci. 359:623–638.
Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24:1752–1760.
Potts M. 1994. Desiccation tolerance of prokaryotes. Microbiol Rev. 58:755–805.
Raymond J, Blankenship RE. 2008. The origin of the oxygen-evolving complex. Coord Chem Rev. 252:377–383.
Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. 1:283–290.
Rokitko PV, Romanovskaya VA, Malashenko YR, Chernaya NA, Gushcha NI, Mikheev AN. 2001. Soil drying as a model for the action of stress factors on natural bacterial populations. Microbiology. 72:756–761.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Rosing MT, Bird DK, Sleep NH, Glassley W, Albarede F. 2006. The rise of continents—an essay on the geologic consequences of photosynthesis. Palaeogeogr Palaeoclimatol Palaeoecol. 232:99–113.
Rusch DB, Halpern AL, Sutton G, et al. (40 co-authors). 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5:e77.
Sanderson M. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol. 14:1218–1231.
Schopf JW, Kudryavtsev AB, Czaja AD, Tripathi AB. 2007. Evidence of archean life: stromatolites and microfossils. Precambrian Res. 158:141–155.
Schwartzman D. 1999. Life, temperature, and the Earth. New York: Columbia University Press.
Seo PS, Yokota A. 2003. The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, gyrB, rpoC1 and rpoD1 gene sequences. J Gen Appl Microbiol. 49:191–203.
Sleep NH, Zahnle KJ, Kasting JF, Morowitz HJ. 1989. Annihilation of ecosystems by large asteroid impacts on the early Earth. Nature. 342:139–142.
Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA. 103:12115–12120.
Stackebrandt E, Rainey FA, WardRainey NL. 1997. Proposal for a new hierarchic classification system, Actinobacteria classis nov. Int J Syst Bacteriol. 47:479–491.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596–1599.
Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol Biol Evol. 19:1727–1736.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst Biol. 51:689–702.
Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. Proc Natl Acad Sci USA. 103:5442–5447.
Tringe SG, von Mering C, Kobayashi A, et al. (13 authors). 2005. Comparative metagenomics of microbial communities. Science. 308:554–557.
Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature. 440:516–519.
Watanabe Y, Martini JE, Ohmoto H. 2000. Geochemical evidence for terrestrial ecosystems 2.6 billion years ago. Nature. 408:574–578.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–479.
Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol. 1:8.
Wuyts J, Perriere G, de Peer YV. 2004. The European ribosomal RNA database. Nucleic Acids Res. 32:D101–D103.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS. 13:555–556.
Zahnle K, Arndt N, Cockell C, Halliday A, Nisbet E, Selsis F, Sleep NH. 2007. Emergence of a habitable planet. Space Sci Rev. 129:35–78.
Zhaxybayeva O, Lapierre P, Gogarten JP. 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma. 227:53–64.

Koichiro Tamura, Associate Editor

Accepted October 27, 2008