The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation

Julie Marin,[1,2,*] Fabia U. Battistuzzi,[3] Anais C. Brown,[3] and S. Blair Hedges[1,*]

[1]Center for Biodiversity, Temple University, SERC Suite 502, 1925 N 12th Street, Philadelphia, PA

[2]Institut de Systématique, Evolution, Biodiversité UMR 7205, Département Systématique et Evolution, Muséum National d'Histoire Naturelle, Sorbonne-Universités, Paris Cedex 05 , France

[3]Department of Biological Sciences, Oakland University, Rochester, MI

*Corresponding authors: E-mails: juliemarin46@gmail.com; sbh@temple.edu

Associate editor: Naoko Takezaki

Abstract

The increasing size of timetrees in recent years has led to a focus on diversification analyses to better understand patterns of macroevolution. Thus far, nearly all studies have been conducted with eukaryotes primarily because phylogenies have been more difficult to reconstruct and calibrate to geologic time in prokaryotes. Here, we have estimated a timetree of 11,784 'species' of prokaryotes and explored their pattern of diversification. We used data from the small subunit ribosomal RNA along with an evolutionary framework from previous multi-gene studies to produce three alternative timetrees. For each timetree we surprisingly found a constant net diversification rate derived from an exponential increase of lineages and showing no evidence of saturation (rate decline), the same pattern found previously in eukaryotes. The implication is that prokaryote diversification as a whole is the result of the random splitting of lineages and is neither limited by existing diversity (filled niches) nor responsive in any major way to environmental changes.

Key words: **diversification, timetree, prokaryote, evolution**

Introduction

Large and nearly complete molecular phylogenies of eukaryotes have allowed evolutionary biologists to better understand patterns of macroevolution in recent years (Hedges et al. 2015). The expansion model states that diversity increases without limit and depends only on time and diversification rate, which is the balance between speciation and extinction rates (Cornell 2013). On the other hand, density-dependent species production, as from competition or resource limitation, will lead to saturated diversity characterized by a null diversification rate (Cornell 2013). Evidence supporting saturated (Rabosky et al. 2012; Rabosky 2013) or expanding diversity (Morlon et al. 2010; Venditti et al. 2010; Jetz et al. 2012) has been found in recent years for selected groups of eukaryotes. A more inclusive study with a wider sampling coverage (50,455 species) found support for a constant rate of diversification over time in eukaryotes (Hedges et al. 2015). That study also found evidence for saturation and accelerating or decelerating diversification in several eukaryote clades, suggesting that the globally constant rate overall may be the product of averaging many small and random pulses of diversification (Ricklefs 2014). Unfortunately, a species-level timetree of prokaryotes has not been available to conduct similar analyses in those organisms. Instead, existing timetrees are relatively small and primarily involve higher taxonomic groups (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009a, b, c; Jun et al. 2010; Loren et al. 2014; Gubry-Rangin et al. 2015; Hedges et al. 2015).

Prokaryote diversification patterns are expected to differ from those of eukaryotes. For example, horizontal gene transfer (HGT), a widespread mechanism in prokaryotes, can favor the movement of a gene variant, whether adaptive or neutral, between species thereby lowering rates of extinction and confounding boundaries between species (Young 1989; Cohan 2001). Debate continues over the definition of a prokaryote "species", with some suggesting that the term will eventually be abandoned (Doolittle and Zhaxybayeva 2009) while others see value in the concept, albeit redefined (Staley 2013), and new evidence (Bendall et al. 2016; Cohan 2016) adds fuel to the debate. Because it continues to be used in the literature and databases, we also use the term here, but do not imply any special meaning, and therefore our use of "species" is equivalent to "operational taxonomic unit".

The few studies that have focused on the macroevolutionary diversification of prokaryotes have done so with small groups (15–153 species) and have obtained mixed results, with constant (Martin et al. 2004; Loren et al. 2014; Gubry-Rangin et al. 2015) or decreasing (Morlon et al. 2012) diversification rates over time. Some experimental observations support the second result, an explosive radiation followed by decay in diversification rates (MacLean 2005; Kassen 2009). However, they were conducted on small taxonomic groups and cannot be extrapolated to all prokaryotes. Indeed, small groups of species are more likely to show patterns of saturated diversity more often than larger groups (Hedges et al. 2015). To explore the global diversification rate of

Article

prokaryotes over time, a large and comprehensive timetree is needed.

In order to construct the most complete prokaryote timetree (PTT), we used a comprehensive small subunit (SSU) data set of 11,269 species (Munoz et al. 2011) supplemented by SSU sequences of 684 species of cyanobacteria from the National Center for Biotechnology Information (NCBI) database and in the Ribosomal Database project (RDP; Cole et al. 2013). The SSU gene sequences may be the most accurate way to establish genealogical relationships (Yarza et al. 2008). However, the SSU is subject to base compositional biases that can affect phylogeny, unless corrected (Battistuzzi and Hedges 2009a). Because of this, we constrained the deepest nodes, between families and above, according to phylogenies obtained with multi-protein datasets (Battistuzzi and Hedges 2009a,b,c). For comparison, we also used two other higher level topologies based on many protein orthologs to constrain the relationships (Lang et al. 2013; Rinke et al. 2013).

We studied the diversification patterns of prokaryotes with three approaches. First, we explored variation in net diversification rate over time, using two methods for the prokaryotes as a whole, as well as subclades. We also timed a multi-gene Bacilli phylogenetic tree to compare the diversification patterns obtained with one gene (SSU) versus many genes. Second, we evaluated branch length distribution, which is another way of testing whether the data are clock-like (exponential distribution) or non-clock like (other distributions). We also compared the branch-lengths of prokaryotes and eukaryotes in order to further explore their difference. Finally, because the SSU gene is the only marker available for all described species, we used simulations to investigate how a limited number of variable sites could influence our results.

Results

Phylogenies and Timetrees

The recently released SSU dataset, along with topological constraints (Battistuzzi and Hedges 2009b,c), were combined to produce a species-level PTT of 11,784 species (Topology A; fig. 1). Topological constraints from other studies produced timetrees of 11,771 species (Topology B; Lang et al. 2013) and 11,774 species (Topology C; Rinke et al. 2013). To evaluate our time estimates, we compared our results, node estimates of the PTT, with a timetree of 98 representative prokaryote species built with a different phylogenetic and timing method (Sheridan et al. 2003). They calibrated a Neighbor Joining distance tree, built from SSU rRNA sequences, using a minimum time of 2,650 Ma for the emergence of cyanobacteria. The tree used to calibrate the PTT (Battistuzzi and Hedges 2009b,c) was also built with different genes and calibration points (see the "Materials and Methods" section). We tested the relationship of 32 common node estimates between the PTT and the timetree from Sheridan et al. (2003) (supplementary fig. S2 and table S6, Supplementary Material online). Over the 32 common nodes, 11 were not used as calibration points in the timing process of the PTT because they were not reported in Battistuzzi and Hedges (2009b,c) and none of the 32 nodes was used to calibrate the timetree from Sheridan et al. (2003). When using the 32 common nodes we obtained a significant correlation (regression by origin: P-value $< 2.2 \times 10^{-16}$, $r^2 = 0.95$, slope $= 0.95$; unconstrained regression: P-value $= 1.71 \times 10^{-06}$, $r^2 = 0.54$, slope $= 0.53$). A strong correlation was also obtained with only the 11 nodes when we constrained the regression through the origin (P-value $= 1.297 \times 10^{-6}$, $r^2 = 0.91$, slope $= 1.11$). However, without constraint the correlation was not significant (P-value $= 0.23$, $r^2 = 0.15$, slope $= 0.23$). Constraining regressions through the origin reflects the age of tips (0 Ma) in both data sets. Without this constraint, the regressions showed a weaker or no correlation which might be explained by the different phylogenetic and timing methods used in both studies.

Diversification Analyses

A significant positive gamma statistic (Pybus and Harvey 2000) obtained for the main PTT did not indicate a decline in diversification rate through time (gamma statistic $= 94.5$, P-value $< 2.2 \times 10^{-16}$). This result was confirmed by our analyses on diversification rates through time using the programs BAMM (Rabosky et al. 2014) and TreePar (Stadler 2013). The first analysis (BAMM: Bayesian Analysis of Macroevolutionary Mixtures) showed a constant net diversification rate over the major part of the PTT (fig.. 2*a*) as well as for the multi-gene Bacilli timetree (fig.. 2*c*). Concerning the PTT (fig.. 2*a*) we also detected a sharp increase in net diversification rate around 30 Ma but this is explained by sampling bias (see section below).

The rate shift analyses (TreePar) gave us the same constant net diversification pattern for the major part of the PTT (between 100 and 3,720 Ma; fig.. 2*b*) and the multi-gene Bacilli timetree (between 50 and 1,884 Ma; fig.. 2*d*) but other sections of the tree showed rates shifts. For the PTT, a model with five rate shifts was not rejected in favor of a model with six rate shifts (P-value $= 0.133$) (supplementary table S1, Supplementary Material online). The five shifts were detected at 20, 40, 100, 3,720, and 4,180 Ma (supplementary table S1, Supplementary Material online). The parameters obtained for the five-shifts model for the section between 100 and 3,720 Ma were $\lambda = 0.0483$ and $\mu = 0.0469$ with λ being the speciation rate and μ the extinction rate. For the multi-gene Bacilli timetree, a model with one rate shift was not rejected in favor of a model with two rate shifts (P-value $= 0.439$). The shift was detected at 50 Ma (supplementary table S1, Supplementary Material online). The parameters obtained for the one-shift model for the section between 50 and 1,884 Ma were $\lambda = 0.0719$ and $\mu = 0.0704$ (fig.. 2*d*). Similar results were obtained with the alternative topologies (B and C), when using 500,000 as the number of prokaryote species and when removing the archaea that might suffer for a higher bias regarding the estimation of the number of species (Castelle et al. 2015) (supplementary fig. S3, Supplementary Material online). The maximum *a posteriori* probability shift configuration of the PTT was determined with the program BAMM, showing 215 shifts between lineages. We also evaluated rate shifts within subclades of the PTT. We did not take into account plot intervals with <30 nodes involved (supple

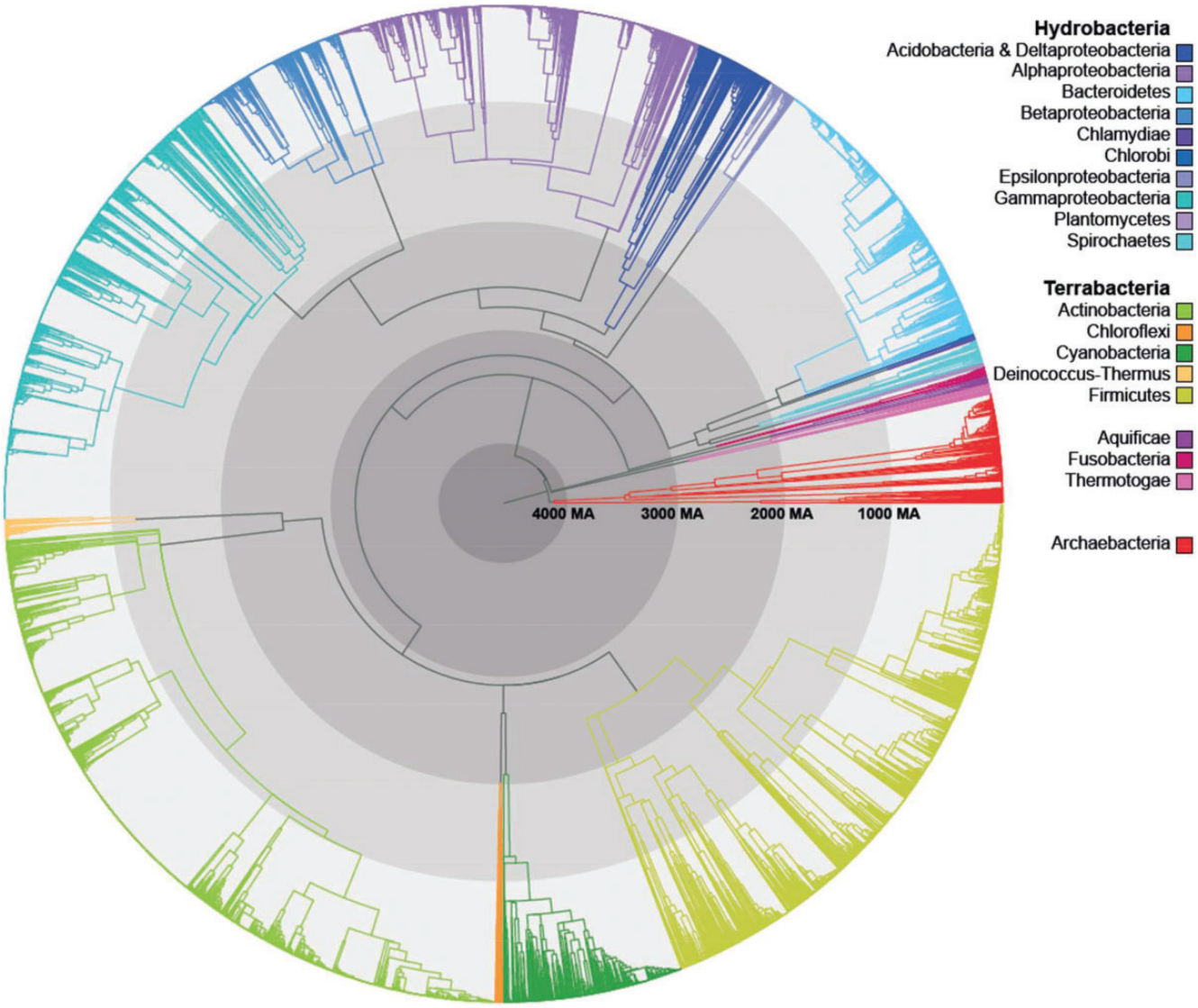

Fig. 1. PTT (topology A; 11,784 species) based on the SSU gene. Divergence times were estimated with the program RelTime and 87 calibration points (listed in supplementary table S3, Supplementary Material online). Ma: millions of years ago.

mentary fig. S4, Supplementary Material online) because of lack of data (Jetz et al. 2012). For 7 out of the 10 trees tested we detected rate shifts near the present time.

Branch Length Distribution

For a given tree an expected branch length under the Yule process can be calculated from its speciation rate (Steel and Mooers 2010). For the PTT the expected branch length was 10.5 My using λ estimated with the program TreePar, between 100 and 3,720 Ma. Similarly, the expected branch length of eukaryotes was 6.8 My (Hedges et al. 2015), with λ estimated between 151 and 2,100 Ma using TreePar. For all trees tested, including the species level prokaryote timetrees (topologies A, B and C), four prokaryote subtrees, and the eukaryote timetree (Hedges et al., 2015), the branch length distributions followed an exponential or variable rate distribution, with the latter corresponding to variation of the exponential model with lineages having different constant rates (supplementary table S2, Supplementary Material online).

Influence of the Number of Variable Sites on Diversification Rates

The diversification rate through time analysis of prokaryotes revealed a sharp peak around 100 Ma (fig.. 2*a* and *b*). In order to understand the possible causes of this peak, we simulated three sets of 1,000 sequences with rate matrix parameters and base frequencies estimated from the prokaryote alignment and three different shape parameters α (0.01, 0.1, and 0.6) of the gamma distribution of substitution rate across sites. The net diversification rate plot (BAMM; fig.. 3*a*) showed a similar constant diversification rate for the three trees until a sharp increase around 3 Ma for the tree built with the lowest shape parameter α (0.01) and to a lesser degree for the tree built with the intermediate shape parameter α (0.1), that is, with a low number of variable sites compared with the third one (with $\alpha = 0.6$). Empirical data showed the same pattern (fig.. 3*b*) with a constant diversification rate until the very recent time where the PTT revealed a sharp increase toward

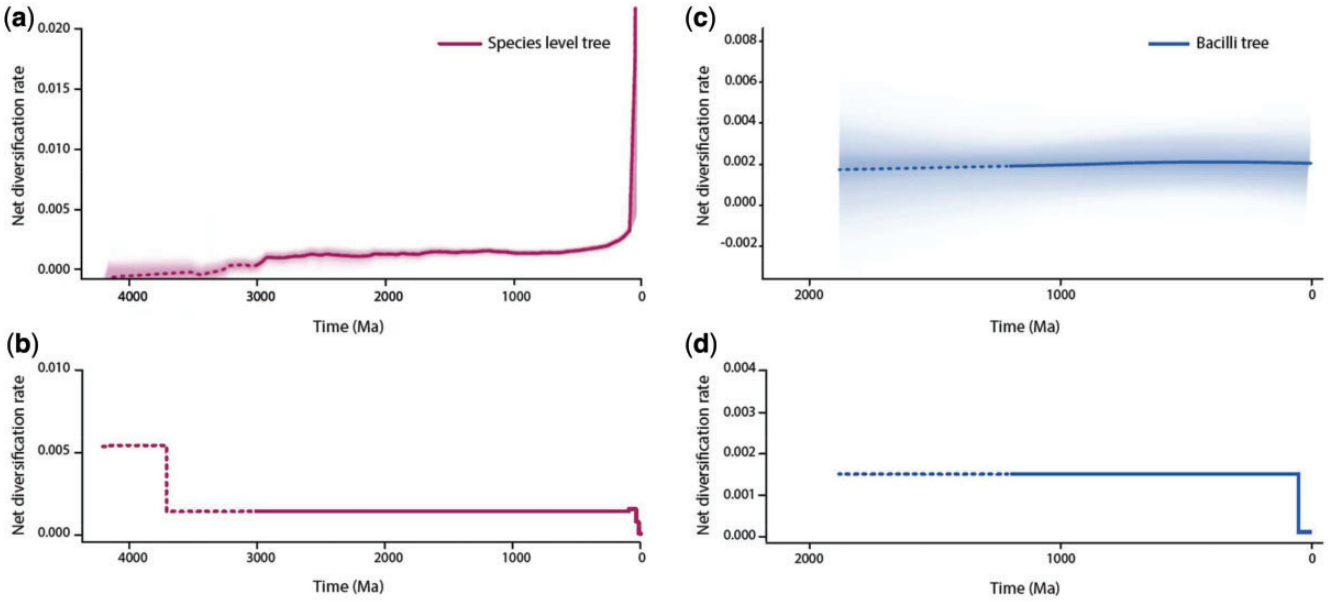

Fig.2. Net diversification rate plots of the PTT (topology A) obtained with the programs BAMM *(a)* and TreePar *(b)*. Net diversification rate plots of the multi-geneBacilli timetree obtained with the programs BAMM *(c)*, and TreePar *(d)*. Shading denotes confidence on evolutionary rate at ± 95%. Dotted lines represent tree section with <10 nodes involved.

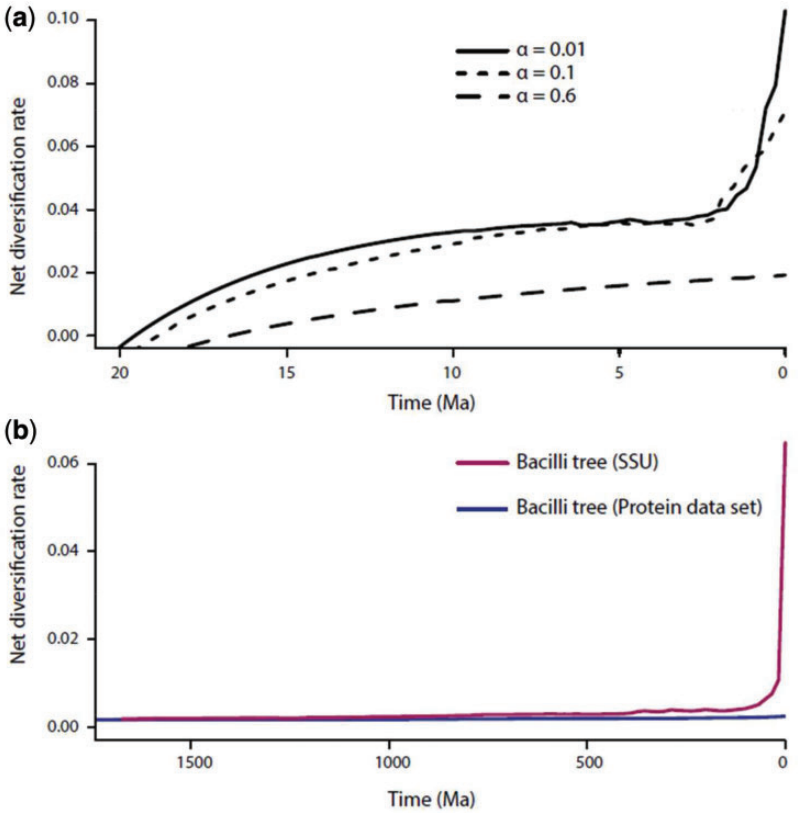

Fig.3. Influence of the number of variable sites on the net diversification rate of timetrees. Simulated timetrees from sequences with shape parameters α set at 0.6, 0.1 and 0.01 *(a)*. Bacilli timetrees reconstructed with the SSU or 30 orthologous genes *(b)*.

0 Ma in contrast to the protein timetree which remained flat (fig.. 3*b*). According to these results, increasing the number of variable sites (higher shape parameter α) tends to reduce the sharp recent peak without affecting the shape of the diversification curve (here constant) until this peak. In other words, it is an artifact (bias).

Sparse Node Artifact Simulations

Because we detected early shifts in the diversification rate of the PTT, at 3,720 and 4,180 Ma involving only four or fewer nodes, we tested this potential sampling bias called here the "sparse nodes" artifact by simulations. We recorded the age and the number of nodes involved in the earliest rate shifts detected by TreePar for 10 simulated trees under a constant birth–death model. Significant shifts were detected for 6 trees out of 10, at 52, 209, 792, 338, 151, and 252 Ma involving 19, 7, 2, 5, 14, and 2 nodes, respectively.

Discussion

Here, we were able to time the most complete phylogenetic tree of prokaryotes (11,784 species) providing us the opportunity to explore their pattern of diversification. Despite a limited number of nucleotide sites, the PTT is consistent with a subset, the Bacilli timetree built with 30 orthologous genes, in terms of pattern of diversification. We found evidence for a constant diversification rate over the major part of prokaryote evolution in the PTT as well as for two alternative topologies, in line with previous results obtained on taxonomically restricted datasets (Martin et al. 2004; Loren et al. 2014; Gubry-Rangin et al. 2015). The same constant pattern of diversification rate was obtained with fewer taxa and calibration points, and more genes (fig.. 2*d*). These results indicate that the constant-rate pattern detected for the PTT was not an artifact from limited nucleotide sites available, or from a high number of calibrations used here to constrain the phylogenetic tree. We also detected significant shifts in diversification rate early in the timetree, and near zero time, with the program TreePar, and a sharp increase toward zero time with the program BAMM. These changes in diversification rate result from three biases: the "sparse nodes artifact", the "taxonomic artifact", and the "sparse sites artifact".

The earlier shifts detected at 3,720 and 4,180 Ma for the PTT (fig.. 2*b*), the sparse nodes artifact, involved only four or fewer timetree nodes, which are too few for statistical significance. Our simulations showed that when only few (<20) nodes are involved, rate shifts can be detected even if the trees were simulated under a birth-death process. The decrease in rate toward zero for both the PTT and the Bacilli timetree (fig.. 2*b* and *d*) is another sampling bias, in this case caused by the omission of lineages below the species level, called "taxonomic artifact" or bias (Hedges and Kumar 2009). When genera, families, or other taxa are selected, the lower level clades are omitted creating an artificial end in the diversification process resulting in a drop of the rate. For example, adding many strains for each species would have shifted this final drop closer to zero. We also detected a sharp increase of the net diversification rate just prior to zero time for the PTT (fig.. 2*a*), as was found for eukaryotes (Hedges et al. 2015). Our simulations showed that it can result from a small number of variable sites and not affect the earlier shape of the diversification rate slope (fig.. 3*a*). This hypothesis was further confirmed by the net diversification rate plots of Bacilli (fig.. 3*b*), where the SSU timetree, but not the protein timetree (with its higher number of variable sites), displayed the same sharp increase as in the PTT. Therefore, the constant rate of diversification in the PTT, prior to zero time, is unbiased whereas the sharp increase near zero time, the "sparse sites artifact" results from use of a small number of nucleotide sites. When the number of variable sites is limited, the sister species have a high probability of sharing the same nucleotide or protein sequence. As a consequence of identical sequences between close relatives, the estimated time will be close to zero and will result in a sharp increase of the diversification rate near zero time. Having more variable sites will increase the probability of differentiating the sequences between close relatives and thus spread out the time nodes, attenuating or eliminating the sharp increase. Our simulations corroborate this statement, however more investigation in this area is required in order to better understand the influence of the variation level of genetic markers on diversification rates as there are also other factors that could contribute to the development of this artifact (Hedges et al. 2015).

In addition, our results showed that the branch lengths of the prokaryote timetrees (topologies A–C) and three subtrees followed an exponential distribution (supplementary table S2, Supplementary Material online), further supporting the random nature of lineage-splitting expected to underlie a constant rate of diversification. Variable rates within Firmicutes were detected (variant of the exponential density distribution), meaning that lineages within Firmicutes are evolving under different between lineages but constant through time diversification rates. This might explain why we detected shifts with the program TreePar probably due to the emergence of lineages with a higher or lower diversification rate than other lineages within the same timeframe. We also detected shifts in diversification for prokaryote subclades (supplementary fig. S4, Supplementary Material online) essentially toward age zero reflecting the sparse sites artifact and the taxonomic artifact (lineages not sampled below species level) discussed above. Over the 10 sub-clades analysed, three showed some variation in diversification rate throughout the subclade history (Clade A; Chloroflexi and Cyanobacteria; Firmicutes). Finally, when considering the global PTT we detected 215 shifts between lineages, reflecting the different diversification rates observed for each prokaryote subtree (supplementary fig. S4, Supplementary Material online). Overall, our results for each subclade and for the whole tree are consistent with results obtained for eukaryotes, where different diversification dynamics, such as hyper-, hypo-diversification, and saturation were detected in particular subclades whereas the overall average was a constant diversification rate over time (Hedges et al. 2015). Ricklefs (2014) proposed a similar model, whereby diversification rate variability is scale dependent.

Despite the consistency in the pattern of diversification that we obtained from different datasets, it would be useful to enhance the PTT in the future with an increased number of genes and calibrations, for increased precision of recent divergence times. The increased site-sampling should reduce or eliminate the sampling artifact that causes a spike in diversification rate near zero time.

Prokaryote diversification patterns have been linked to the number of available niches, increasing with the appearance of animals (Sepkoski et al. 2002; Loren et al. 2014). However, our results of a constant rate instead support the random survival of isolated lineages, as suggested for eukaryotes (Hedges et al 2015). Examples of factors that can promote the emergence and survival of prokaryote lineages are spatial isolation (Petursdottir et al. 2000; Papke et al. 2003), mutation, and HGT from other ecotypes, providing functional innovation (Ochman et al. 2000; Cohan 2001; Treangen and Rocha 2011). When acquiring an entirely new metabolic function, by mutation or HGT, a nascent lineage will escape the "periodic selection", that is, competition with its former population (Cohan 2001).

Despite the globally constant lineage diversification observed for both prokaryotes and eukaryotes, they have different rates of evolution. When only considering the constant diversification rate section, that is, the major part of prokaryote and eukaryote evolution (between 100 and 3,720 Ma, and between 151 and 2,100 Ma, respectively), the diversification rate for eukaryotes is 2.1 times faster than that of prokaryotes. A force of cohesion specific to asexual (or rarely sexual) bacteria, periodic selection (Cohan 2001), might be responsible for this slower diversification compared with eukaryotes. Because of periodic selection, a bacterial lineage expands its diversity until an adaptive mutant outcompetes all other strains leading to the purge of nearly all its diversity at all loci through natural selection (Cohan 2001). This process might explain why the mean branch length of prokaryotes, 10.5 My, is longer than in eukaryotes, 6.8 My. Recent evidence for cohesion in prokaryotes has been found (Bendall et al. 2016) although more work needs to be conducted to determine the generality of this mechanism (Cohan 2016).

In conclusion, we produced a timetree of most described prokaryote 'species' that revealed a constant diversification rate (fig.. 2*a* and *b*) remarkably similar in that respect to eukaryotes and probably resulting from the same mechanism, the random nature of lineage survival over millions of years. The rate of diversification that is 2.1 times slower than in eukaryotes is probably the result of the periodic selection, which provides genetic cohesion of lineages, slowing their genetic divergence. The overall similarity in these important aspects of the evolution of prokaryotes and eukaryotes lends support to the idea that species of prokaryotes may be real evolutionary units (Cohan 2001, 2016) much like eukaryotic species, although the generality of bacterial cohesion needs to be investigated.

Materials and Methods

Phylogenies and Timetrees

The aligned SSU dataset (11,269 sequences) (November 2014) was downloaded from the SILVA project website (http://www.arb-silva.de/projects/living-tree/) (Munoz et al. 2011) corresponding to all type strains of all species with validly published names up to July 2014. A total of 670 supplementary species have been added since then, representing 5% of our dataset. Because of the unsuitability of SSU in phylogenetic reconstruction of deep nodes we used the topology (supplementary fig. S1, Supplementary Material online) obtained from a protein data set to constrain the phylogeny (Battistuzzi and Hedges 2009b,c). A total of 169 species (beyond the 11,269) were removed from our data set: 113 species did not belong to any of the 11 groups (listed in the supplementary table S5, Supplementary Material online), 40 of them showed unusually long branches in preliminary phylogenies, 6 were listed twice, and the 10 cyanobacteria were replaced by 684 Cyanobacteria sequences (see below). Only ten cyanobacteria sequences were available in the SILVA database; however, 2,852 species were listed in the CyanoDB (database of cyanobacteria genera) (Komárek and Hauer 2013). The difference is caused by the difficulty in validly publishing new names of Cyanobacteria under the rules of the International Code of Nomenclature of Prokaryotes (ICNP) (Oren 2011). We searched for those sequences in the NCBI and RDP databases (Cole et al. 2013) and downloaded 684 16S rRNA sequences (accession numbers are listed supplementary table S4, Supplementary Material online). The final data set contained 11,784 species (11,110 prokaryote species from SILVA and 684 cyanobacteria from additional databases) and is available on the Center for Biodiversity site (www.biodiversitycenter.org).

As some families were not represented in the study used to constrain the phylogeny (Battistuzzi and Hedges 2009b,c) we divided the bacteria data set to avoid those families to be grouped by mistake in another phylum or division. One to nine outgroups were added for each sub-dataset (listed in the supplementary table S5, Supplementary Material online). Eleven sub-trees were built: Actinobacteria and Deinococcus-Thermus (2,851 species), Alphaproteobacteria (1,389 species), Betaproteobacteria (586 species), Clade A containing Chlamydiae, Chlorobi, Bacteroidetes, Plantomycetes, and Spirochaetes (1,259 species), Chloroflexi and Cyanobacteria (715 species), Deltaproteobacteria and Acidobacteria (304 species), Epsilonproteobacteria (107 species), Firmicutes (2,264 species), Gammaproteobacteria (1,787 species), Archaea (415 species), and backbone (Aquificae, Fusobacteria, Thermotogae and one representative of each subgroup—117 species).

A 40% partial deletion cut-off was applied (as recommended elsewhere; Munoz et al. 2011) resulting in an alignment of 1,493 nucleotides and 11 phylogenies using a maximum likelihood (ML) method were built. The relationships between families were constrained as depicted in supplementary fig. S1, Supplementary Material online (Battistuzzi and Hedges 2009b,c). The families belonging to 1 of the 11 groups (described above) but not included in the protein phylogenies were included in the alignment files but not in the constraints files, allowing each of these species to branch anywhere in the tree. The phylogenies were constructed with RAxML 8.1.11 (Stamatakis 2014) assuming GTR (general time reversible) model with 1,000 bootstrap replicates. We used the CAT model (GTRCAT) to take into account rate heterogeneity among sites. Trees were visualized with FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). Over the 11 phylogenies 40% of the nodes (4,764 over 11,839) were supported by bootstrap probabilities $>70\%$.

Each phylogenetic tree was timed with RelTime (Tamura et al. 2012) implemented in MEGA 6.0.6 (Tamura et al. 2013).

RelTime estimates the relative divergence times for each node of the tree. It has been shown that when applied to large dataset with variable evolutionary rates between lineages (under autocorrelated and uncorrelated models) this method outperformed other methods (e.g., MCMCTree) (Tamura et al. 2012). We used confidence intervals of the 90 calibration points listed in supplementary table S3, Supplementary Material online (Battistuzzi and Hedges 2009b,c) as minimum and maximum times to convert the relative times into absolute times. Those calibration times correspond to time estimates for 218 Bacteria and Archaea species built from 25 protein coding genes (15 ribosomal proteins (RPL1, 2, 3, 5, 6, 11, 13, 16; RPS2, 3, 4, 5, 7, 9, 11), four genes (RNA polymerase alpha, beta, and gamma subunits, Transcription antitermination factor NusG) from the functional category of Transcription, three proteins (Elongation factor G, Elongation factor Tu, Translation initiation factor IF2) of the Translation, Ribosomal Structure and Biogenesis functional category, one protein (DNA polymerase III, beta subunit) of the DNA Replication, Recombination and repair category, one protein (Preprotein translocase SecY) of the Cell Motility and Secretion category, and one protein (O-sialoglycoprotein endopeptidase) of the Posttranslational Modification, Protein Turnover, Chaperones category). The divergence times of the bacteria timetree (Battistuzzi and Hedges 2009b) were obtained using three calibration points: 1) a minimum of 1,640 Ma for the origin of Chromatiaceae (Brocks et al. 2005); 2) a minimum of 1,640 Ma for the divergence of Chlorobi and Bacteroidetes (Brocks et al. 2005); and 3) a maximum of 4,000 Ma for the earliest land-dwelling taxa corresponding to the presence of continents (Rosing et al. 2006). The divergence times of the archaea timetree (Battistuzzi and Hedges 2009c) were obtained using two calibration points: 1) a minimum of 3,460 Ma for the origin of methanogenesis (Bapteste et al. 2005; Ueno et al. 2006) and 2) a maximum of 4,200 Ma for the first divergence within archaea (Sleep et al. 1989). The divergence between archaea and bacteria was set at a maximum of 4,200 Ma (Battistuzzi and Hedges 2009b). To assemble the final timetree of prokaryotes, the representatives of each group in the backbone tree were replaced by the corresponding timetree, resulting in a timetree of 11,784 prokaryote species. Because more topological constraints (inter-family and above) and more calibration points were available for topology A, we refer to it as our main topology.

The same methodology as described above was applied using the Lang et al. (2013) topology to constrain the nodes at the family level or above (topology B). Three groups that were not part of the Battistuzzi and Hedges (2009b) tree were added to the backbone tree resulting in a dataset of 11,860 species: Deferribacteres (13 species), Synergistetes (21 species), and Verrucomicrobia (42 species). To time the 11 phylogenetic trees (Actinobacteria and Deinococcus-Thermus (2,851 species), Alphaproteobacteria (1,389 species), Betaproteobacteria (586 species), Clade A containing Chlamydiae, Chlorobi, Bacteroidetes, Plantomycetes and Spirochaetes (1,259 species), Chloroflexi and Cyanobacteria (715 species), Deltaproteobacteria and Acidobacteria (304 species), Epsilonproteobacteria (107 species), Firmicutes (2,264 species), Gammaproteobacteria (1,787 species), Archaea (415 species), and backbone (Aquificae, Deferribacteres, Fusobacteria, Synergistetes, Thermotogae and Verrucomicrobia, and one representative of each subgroup—193 species) we used confidence intervals of the 41 calibration points listed in supplementary table S3, Supplementary Material online (Battistuzzi and Hedges 2009b,c) as minimum and maximum times to convert the relative times into absolute times.

The same methodology as described earlier was also applied using the Rinke et al. (2013) topology to constrain the nodes at the family level or above (topology C). One group, Fusobacteria (37 species) that did not appear in the Rinke et al. (2013) tree was removed and 12 that were not part of the Battistuzzi and Hedges (2009b) tree were added (Elusimicrobia [5 species], Fibrobacteres [3 species], Gemmatimonadetes [1 species], Lentisphaerae [4 species], and Verrucomicrobia [42 species]) to the Clade A and (Armatimonadetes [3 species], Chrysiogenetes [4 species], Deferribacteres [13 species], Dictyoglomi [2 species], Nitrospira [8 species], Synergistetes [21 species], and Thermodesulfobacteria [8 species]) to the backbone tree resulting in a dataset of 11,861 species. The Epsilonproteobacteria were placed into Clade A (see above) and the Acidobacteria within the backbone group. To time the 10 phylogenetic trees (Actinobacteria and Deinococcus-Thermus (2,851 species), Alphaproteobacteria (1,389 species), Betaproteobacteria (586 species), Clade A containing Chlamydiae, Chlorobi, Bacteroidetes, Elusimicrobia, Epsilonproteobacteria, Fibrobacteres, Gemmatimonadetes, Lentisphaerae, Plantomycetes, Spirochaetes and Verrucomicrobia (1,421 species), Chloroflexi and Cyanobacteria (715 species), Deltaproteobacteria and Acidobacteria (304 species), Firmicutes (2,264 species), Gammaproteobacteria (1,787 species), Archaea (415 species) and backbone (Aquificae, Armatimonadetes, Chrysiogenetes, Deferribacteres, Dictyoglomi, Nitrospira, Synergistetes, Thermodesulfobacteria, and Thermotogae and one representative of each subgroup—138 species) we used confidence intervals of the 42 calibration points listed in supplementary table S3, Supplementary Material online (Battistuzzi and Hedges 2009b,c) as minimum and maximum times to convert the relative times into absolute times. The additional 132 sequences used to built the topologies B and C are also available on the Center for Biodiversity site (www.biodiversitycenter.org).

Species Level—Bacilli

A phylogenetic tree of Bacilli was also reconstructed using 30 orthologous genes (5,262 amino acids) shared by 129 species. The tree was built with the model LG + CAT + I using the rapid hill-climbing tree-search algorithm with RAxML 8.1.11 (Stamatakis 2014) on an ungapped alignment. The divergence times were estimated with RelTime (Tamura et al. 2012) (see above) using the confidence intervals of four

calibration points (node 30, 50, 54, and 61, supplementary table S3, Supplementary Material online).

Diversification Analyses

The diversification analyses were performed in R (http://www.r-project.org/) using several packages: APE (Paradis et al. 2004), BAMMtools (Rabosky et al. 2014), LASER (Rabosky and Schliep 2013) and TreePar (Stadler 2013). The BAMMtools package and the BAMM program (Rabosky et al. 2014) were used to estimate the diversification rate through time of three timetrees: the timetree of prokaryote species (11,784 species) and two Bacilli timetrees, one extracted from our timetree (1,361 species) and the multi-genes based timetree (129 species). The function "setBAMMpriors" was used to generate a prior block that matched the "scale" (e.g., depth of the tree) of our data. Both λ and μ rates were allowed to vary through time and across lineages, and MCMC chains were run for 500,000,000 iterations. A species specific sampling fraction was set for the Bacilli timetrees by genus corresponding to the fraction of species present in the tree over the number of species present in the SILVA database (Munoz et al. 2011), comprising all species with validly published names up to July 2014, for a given genus. A species-specific sampling was also set for the PTT at 0.24 for the Cyanobacteria (684 species in our tree over 2,852 species listed in CyanoDB; Komárek and Hauer 2013) and at 0.93 for the remaining species (11,110 species in our tree over 11,929 non-cyanobacterial species listed in SILVA). The number of shifts between lineages was obtained with the function "getBestShiftConfiguration", corresponding to the shift configuration with the highest posterior probability. In order to evaluate the effect of the underestimation of the prokaryote species richness and by consequence the sampling fraction on diversification rates we specified a global sampling fraction for the PTT at 0.023 corresponding to the number of species in our tree over a high estimate of prokaryote species richness (i.e., 500,000 species; Dykhuizen 1998). We discarded 50% of the MCMC chains to reach the convergence for the PTT, which was checked by calculating the effective sample size of the log-likelihood and of the number of shifts events present in each sample that should be over 200 as recommended by the authors of the program. Diversification rate plots were obtained with the function "plotRateThroughTime". The gamma statistic (Pybus and Harvey 2000) was employed to detect decrease in λ over the history of the PTT (LASER package). A negative gamma value reveals a concentration of branching times near the root, meaning a decelerated diversification.

In addition, the TreePar (Stadler 2013) package was used to estimate the number of significant changes in diversification rate under a birth-death model. We used a sampling fraction accordingly to the number of Cyanobacteria species (sampling fraction at 0.24) and the number of non-cyanobacterial species (sampling fraction at 0.93) for each tree. We analysed 15 timetrees using the greedy approach as described in Stadler (2011): the three prokaryote timetrees (topology A–C), the bacteria timetree, 10 subtrees of the PTT described below, and the multi-gene Bacilli timetree. We estimated rates in 20 My steps between 0 and 4,210 My for the PTT and the bacteria timetree (archaea were removed from the PTT), in 50 My steps between 0 and 4,210 My for the topologies B and C (for computing time reasons) and in 10 My steps between 0 and the maximum divergence time for the multi-gene Bacilli timetree and 10 subtrees (Actinobacteria, Alphaproteobacteria, Archae, Betaproteobacteria, Clade A, Chloroflexi-Cyanobacteria, Deltaproteobacteria, Epsilonproteobacteria, Firmicutes, and Gammaproteobacteria). As the TreePar analysis is computationally time demanding and because the topologies B and C are secondary analyses we decided to evaluate their shifts every 50 Myr instead of 20 Myr for the main topology. We used smaller (10 Myr steps) for subclades to be able to capture their shifts because they evolve over smaller periods of time and because they contain fewer species. The birth-death shift model (without mass extinction) was used and we obtained maximum-likelihood rate estimates for the different datasets for zero to six rate shifts.

Branch Length Distribution

Branch times were estimated using the splitting rate (λ) obtained with the program TreePar for the main section of the PTT (between 100 and 3,720 Ma) with the formula: $\frac{1}{2*\lambda}$ corresponding to the expected length of a random interior edge length under the Yule process (Steel and Mooers 2010) because we wanted true times unaffected by extinction to better reflect the lineage-splitting process.

We plotted the frequency of branch lengths of the PTT, of four subtrees of the PTT (Actinobacteria and Deinococcus-Thermus; Archae; Chlamydiae, Chlorobi, Bacteroidetes, Plantomycetes, Verrucomicrobia and Spirochaetes and Firmicutes), and of the smoothed eukaryote timetree (Hedges et al. 2015). Then we fitted five models described elsewhere (Venditti et al. 2010) to their distributions and selected the best model using AIC score. According to their models, branch length distributions will reflect the interaction of the potential factors of speciation. Thereby a normal density distribution will be the result of factors combining additively and a log-normal density distribution of factors combining multiplicatively to induce a speciation event. If branch lengths are distributed randomly, then an exponential density distribution will be observed. A variant of the exponential model exists allowing lineages to have different constant rates. Other distributions do not support a constant rate model.

Influence of the Number of Variable Sites on Diversification Rates

The Bacilli trees, built with the SSU data or with the protein data, showed a different diversification pattern toward zero time (see "Results" section). Those two data sets are characterized by a similar shape parameter α (0.48 for the SSU data and 0.54 for the protein data) but differed by the length of sequences and the number of taxa, 1,361 sequences for 1,493 nucleotides for the SSU data and 129 sequences for 5,262 amino acids for the protein data. Therefore, the number of variable sites is lower for the SSU data (less sites and more

sequences). Because simulating the same number of taxa and sites required too much computational time we decided instead to set three different parameters α in order to evaluate the influence of the variation level of genetic markers on diversification rates. We simulated three sets of sequences (PhyloSim package, Sipos et al. 2011) of 1,000 nucleotides and the corresponding trees of 100 tips. The sequences were simulated from the same tree of 101 tips generated with the function "rcoal" (100 tips and 1 outgroup). To define the model of DNA evolution for the simulations we used the parameters estimated by modeltest (function "modelTest", package phangorn in R; Schliep 2011) on the prokaryote alignment (model: GTR model; rate matrix: $a = 4.22$, $b = 10.42$, $c = 5.94$, $d = 4.42$, $e = 17.27$, $f = 4.38$; base frequencies: 0.24, 0.22, 0.23, 0.31). The shape parameters α were set at 0.6, 0.1, and 0.01. The phylogenetic constructions were performed with RAxML 8.1.11 (Stamatakis 2014) with 1,000 bootstrap replicates. Relative times estimated by the program RelTime were converted to absolute time using one calibration (100 Myr) for the ingroup node to scale the timetrees. Their diversification rates were estimated as described earlier (BAMMtool package).

Sparse Node Artifact Simulations

We simulated 10 trees of 100 tips under a birth-death model with the function "sim.bdtree" (GEIGER package; Harmon et al. 2008) using the speciation rate and extinction rate estimated with the program TreePar for the prokaryote tree between 100 and 3,720 Ma. We set the sampling fraction at the same value as that for the PTT. Next, we used the method described above to estimate the number of significant changes in diversification rate under a birth-death model with the program TreePar package in 1 My steps between 0 and the maximum divergence time of each tree.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgements

We thank Sudhir Kumar for comments, Michael Suleski for programming assistance, and Jaanki Dave for preliminary analyses. This work was supported by grants from the US National Science Foundation (grant number 1136590 and 1262481 to S.B.H) and Oakland University to F.U.B.

References

Bapteste E, Brochier C, Boucher Y. 2005. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* 1(5): 353–363.

Battistuzzi FU, Feijao 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.

Battistuzzi FU, Hedges SB. 2009a. A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol*. 26:335–343.

Battistuzzi FU, Hedges SB. 2009b. Archaebacteria. In: Hedges SB, Kumar S, editors. The Timetree of Life. New York: Oxford University Press. p. 101–105.

Battistuzzi FU, Hedges SB. 2009c. Eubacteria. In: Hedges SB, Kumar S, editors. The Timetree of Life. New York: Oxford University Pres. p. 106–115.

Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations revealed by time-series metagenomics. *ISME J*. 10:1589–1601.

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(7060): 866–870.

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM, et al. 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol*. 25(6): 690–701.

Cohan FM. 2001. Bacterial species and speciation. *Syst Biol*. 50:513–524.

Cohan FM. 2016. Bacterial speciation: genetic sweeps in bacterial species. *Curr Biol*. 26:R102–R124.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2013. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 42:D633–D642.

Cornell HV. 2013. Is regional species diversity bounded or unbounded?. *Biol Rev*. 88:140–165.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Res*. 19(5): 744–756.

Dykhuizen DE. 1998. Santa Rosalia revisited: why are there so many species of bacteria?. *Antonie van Leeuwenhoek* 73:25–33.

Gubry-Rangin C, Kratsch C, Williams TA, McHardy AC, Embley TM, Prosser JI, Macqueen DJ. 2015. Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. *Proc Natl Acad Sci U S A*. 112(30):9370–9375.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. *Bioinformatics* 24(1):129–131.

Hedges SB, Kumar S. 2009. Discovering the timetree of life. In: Hedges SB, Kumar S, editors. The Timetree of Life. New York: Oxford University Press. p. 3–18.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*. 32(4):835–845.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. *Nature* 491:444–448.

Jun SR, Sims GE, Wu GHA, Kim SH. 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*. 107:133–138.

Kassen R. 2009. Toward a general theory of adaptive radiation. *Ann NY Acad Sci*. 1168.1:3–22.

Komárek J, Hauer T. 2013. CyanoDB.cz - On-line database of cyanobacterial genera. - World-wide electronic publication, Univ. of South Bohemia & Inst. of Botany AS CR. Available from: http://www.cyanodb.cz

Lang JM, Darling AE, Eisen JA. 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One* 8:e62510.

Loren JG, Farfan M, Fuste MC. 2014. Molecular phylogenetics and temporal diversification in the genus *Aeromonas* based on the sequences of five housekeeping genes. *PLoS One* 9:e88805.

MacLean RC. 2005. Adaptive radiation in microbial microcosms. *J Evol. Biol*. 18:1376–1386.

Martin AP, Costello EK, Meyer AF, Nemergut DR, Schmidt SK. 2004. The rate and pattern of cladogenesis in microbes. *Evolution* 58:946–955.

Morlon H, Kemps BD, Plotkin JB, Brisson D. 2012. Explosive radiation of a bacterial species group. *Evolution* 66:2577–2586.

Morlon H, Potts MD, Plotkin JB. 2010. Inferring the dynamics of diversification: a coalescent approach. *PLoS Biol*. 8:e1000493.

Munoz R, Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R. 2011. Release LTPs104 of the all-species living tree. *Syst Appl Microbiol*. 34:169–170.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

Oren A. 2011. Naming Cyanophyta/Cyanobacteria - a bacteriologist's view. *Fottea* 11:9–16.

Papke RT, Ramsing NB, Bateson MM, Ward DM. 2003. Geographical isolation in hot spring cyanobacteria. *Environ Microbiol*. 5:650–659.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Petursdottir SK, Hreggvidsson GO, Da Costa MS, Kristjansson JK. 2000. Genetic diversity analysis of Rhodothermus reflects geographical origin of the isolates. *Extremophiles* 4:267–274.

Pybus OG, Harvey PH. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc R Soc Lond B Biol Sci*. 267:2267–2272.

Rabosky DL. 2013. Diversity-dependence, ecological speciation, and the role of competition in macroevolution. *Annu Rev Ecol Evol Syst*. 44:481–502.

Rabosky DL, Grundler M, Anderson C, Title P, Shi JJ, Brown JW, Huang H, Larson JG. 2014. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Meth Ecol Evol*. 5:701–707.

Rabosky DL, Schliep K. 2013. LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Vienna (Austria): Comprehensive R Archive Network [Cited 2015 Nov 21]. Available from: http://cran.r-project.org/package=laser.

Rabosky DL, Slater GJ, Alfaro ME. 2012. Clade age and species richness are decoupled across the Eukaryotic tree of life. *PLoS Biol*. 10:e1001381. CrossRef][10.1371/journal.pbio.1001381]

Ricklefs RE. 2014. Reconciling diversification: Random pulse models of speciation and extinction. *Am Nat*. 184:268–276.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.

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(2):99–113.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4):592–593.

Sepkoski JJ, Jablonski D, Foote M. 2002. A compendium of fossil marine animal genera. *Bull Am Paleo* 363:1–560.

Sheridan PP, Freeman KH, Brenchley JE. 2003. Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol J*. 20:1–14.

Sipos B, Massingham T, Jordan GE, Goldman N. 2011. PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 12:104.

Sleep NH, Zahnle KJ, Kasting JF, Morowitz HJ. 1989. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* 342(6246):139–142.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proc Natl Acad Sci U S A*. 108:6187–6192.

Stadler T. 2013. TreePar: estimating birth and death rates based on phylogenies. Vienna (Austria): Comprehensive R Archive Network [Cited 2015 Nov 21]. Available from: http://CRAN.R-project.org/package=TreePar.

Staley JT. 2013. Transitioning towards a universal species concept for the classification of all organisms. In: Igor Ya. Pavlinov, editor. The Species Problem - Ongoing Issues. InTech Open, DOI: 10.5772/53218.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Steel M, Mooers A. 2010. The expected length of pendant and interior edges of a Yule tree. *Appl Math Lett*. 23:1315–1319.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci U S A*. 109:19333–19338.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 30:2725–2729.

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*. 7(1):e1001284.

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(7083):516–519.

Venditti C, Meade A, Pagel M. 2010. Phylogenies reveal new interpretation of speciation and the Red Queen. *Nature* 463:349–352.

Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glockner FO, Rossello-Mora R. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol*. 31:241–250.

Young JPW. 1989. The population genetics of bacteria. In Hopwood DA, Chater KF, editors. Genetics of Bacterial Diversity. Academic Press, London. p. 417–438.