ARTICLES
https://doi.org/10.1038/s41559-018-0625-0
nature ecology & evolution

Bacterial diversification through geological time

Stilianos Louca[1,2*], Patrick M. Shih[3,4,5], Matthew W. Pennell[1,2], Woodward W. Fischer[6], Laura Wegener Parfrey[1,2,7] and Michael Doebeli[1,2,8]

Numerous studies have estimated plant and animal diversification dynamics; however, no comparable rigorous estimates exist for bacteria—the most ancient and widespread form of life on Earth. Here, we analyse phylogenies comprising up to 448,112 bacterial lineages to reconstruct global bacterial diversification dynamics. To handle such large phylogenies, we developed methods based on the statistical properties of infinitely large trees. We further analysed sequencing data from 60 environmental studies to determine the fraction of extant bacterial diversity missing from the phylogenies—a crucial parameter for estimating speciation and extinction rates. We estimate that there are about 1.4–1.9 million extant bacterial lineages when lineages are defined by 99% similarity in the 16S ribosomal RNA gene, and that bacterial diversity has been continuously increasing over the past 1 billion years (Gyr). Recent bacterial extinction rates are estimated at 0.03–0.05 per lineage per million years (lineage^{-1} Myr^{-1}), and are only slightly below estimated recent bacterial speciation rates. Most bacterial lineages ever to have inhabited this planet are estimated to be extinct. Our findings disprove the notion that bacteria are unlikely to go extinct, and provide a valuable perspective on the evolutionary history of a domain of life with a sparse and cryptic fossil record.

For over 3.5 Gyr, the geochemical composition of our planet has been shaped by the evolution and diversification of bacteria[1]. Most prominently, the Great Oxygenation Event was caused by cyanobacteria roughly 2.35 Gyr ago and dramatically altered Earth's surface environments and the subsequent evolution of life[2]. Despite the prominent role of bacteria in ancient and modern biospheres, little is known about the dynamics by which their diversity evolved over Earth's history. For many eukaryotes, the fossil record provides estimates of past diversity[3–5], revealing that extant global eukaryotic diversity only represents a small fraction of the total diversity that existed in the past. Analogous estimates for bacterial diversity are lacking, largely because their fossil record is extremely poor, and thus the clades that are known are those with extant representatives. Fortunately, past diversification dynamics also leave a footprint in molecular phylogenies of extant organisms[6]. Many approaches have been developed to infer past diversification dynamics from these patterns[7–9]. Despite these methodological advances, global bacterial diversification dynamics remain largely unresolved and much less studied than eukaryotic diversification. Previous studies only examined diversification within a single bacterial genus[10–12] or a single archaeal phylum[13], or phylogenies covering only a small and biased portion of diversity (~12,000 cultured bacterial and archaeal species)[14]. Many of these studies do not report absolute speciation or extinction rates[10,12,13]. Importantly, no previous study properly accounted for the incomplete sampling of bacterial diversity represented in the phylogenies. Knowledge of the 'sampling fraction', in addition to any phylogenetic information, is critical for estimating speciation and extinction rates from phylogenies, even to an order of magnitude[15]. As the extant global bacterial diversity was so far largely unknown, previous studies either assumed that the number of catalogued species was exhaustive[14] (an inaccurate assumption[16]), used local (rather than global) diversity estimates, such as for a small quantity of soil[14], or estimated the unknown sampling fraction directly from the phylogeny without additional information (an impossible task[15]). Consequently, there exists no rigorous estimate of global bacterial speciation rates, extinction rates or total diversity over time, and this uncertainty has clouded our interpretation of bacterial evolution over Earth's history. It is commonly hypothesized that bacterial extinction may not even occur at significant rates[11,17–20], partly due to their large population sizes and wide dispersal ranges[18,19], while others hypothesized that animal extinctions could cause substantial host-associated bacterial extinctions[21].

To address these questions, we examined bacterial phylogenies comprising up to hundreds of thousands of clades, using mathematical tools that we developed specifically for large phylogenies. To properly account for the fraction of undiscovered diversity in our methods, thus resolving a long-standing problem in bacterial phylogenetics, we independently estimated global bacterial diversity using massive DNA sequencing data from 60 studies in diverse environments across the world. To evaluate the robustness of our results, we used numerical simulations and examined several phylogenies constructed using alternative methods. Importantly, some of our phylogenies were constructed from environmental sequences retrieved using culture-independent methods, providing a less biased (and thus more suitable[22]) representation of bacterial diversity compared with previous studies[10,14]. We used our methods, as well as the independently estimated global bacterial diversity, to reconstruct global bacterial speciation, extinction and diversification (speciation minus extinction) rates over the past 1 Gyr.

We used two time-calibrated bacterial phylogenies ('timetrees') based on the 16S ribosomal RNA (rRNA) gene—a popular marker gene in microbial ecology and evolution (448,112 and 162,371 tips, respectively; see Supplementary Table 1 for an overview and the Methods for details). We also analysed cyanobacteria alone due to their great importance to Earth's evolution, using four 16S rRNA-based timetrees constructed with various methods (586, 6,308,

[1]Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada. [2]Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada. [3]Joint BioEnergy Institute, Emeryville, CA, USA. [4]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Department of Plant Biology, University of California, Davis, Davis, CA, USA. [6]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA. [7]Department of Botany, University of British Columbia, Vancouver, Brisitsh Columbia, Canada. [8]Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada.
*e-mail: louca.research@gmail.com

6,302 and 1,579 tips, respectively). In all cases, tips in the trees represent operational taxonomic units (OTUs); that is, clusters in the 16S rRNA gene delineated at 99% similarity—a common microbial 'species' measure[23,24]. We stress that bacterial OTUs only provide an approximate 'species' analogue to sexually reproducing organisms, and hence 'speciation' rates reported here should a priori only be interpreted as branching frequencies in 16S rRNA sequence space.

Estimating diversification dynamics from large timetrees

Our methods were derived from standard stochastic models for cladogenesis, in which extant lineages can split or go extinct randomly and independent of each other as time proceeds[25]. These models predict the total number of extant lineages (total diversity) at each time point, as well as the number of lineages represented in the final timetree comprising only extant and sampled taxa (lineages through time (LTT))[7]. Our methods can account for the effect of incomplete taxon sampling, as well as for speciation and extinction rates that vary over time. In contrast to most existing methods, our methods consider timetrees in the continuum limit of infinitely many lineages, which yields novel ways to extract information from timetrees (Supplementary Information section 1.3). Notably, given some LTT curve, one can estimate a quantity that is related to the diversification rate at each time point, and which we refer to as the pulled diversification rate (PDR):

$$r_{\mathrm{p}} = \lambda - \mu - \frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} \quad (1)$$

where t is time, λ is the instantaneous speciation rate and μ is the instantaneous extinction rate. The PDR partly resembles the diversification rate ($r = \lambda - \mu$), but is modified (pulled) by the term $\lambda^{-1}\mathrm{d}\lambda/\mathrm{d}t$, which represents the relative rate of change of λ over time and is small when λ varies slowly. In contrast to the diversification rate, the PDR can be estimated 'non-parametrically' from the curvature and slope of the LTT curve at any point in time. This approach does not require fitting a specific parameterized model[25], nor a priori assumptions on how λ and/or μ vary over time, nor assumptions about whether the PDR or diversification rate was positive or negative. More precisely, in the continuum limit, the PDR can be calculated using the LTT for any time t using the formula:

$$r_{\mathrm{p}}(t) = -\tilde{\nu}(t) - \frac{1}{\tilde{\nu}(t)}\frac{\mathrm{d}\tilde{\nu}}{\mathrm{d}t} \quad (2)$$

where $-\tilde{\nu}(t) = (1/\tilde{\mathrm{N}}(t))\mathrm{d}\tilde{\mathrm{N}}/\mathrm{d}t$ is the relative slope of the LTT and $\tilde{\mathrm{N}}(t)$ is the value of the LTT at time t. For finite trees, equation (2) is only an estimate.

Similar to the PDR, one can also estimate 'pulled' versions of other important variables, including the pulled extinction rate (PER),

$$\mu_{\mathrm{p}} = \mu + (\lambda_{\mathrm{o}} - \lambda) + \frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} \quad (3)$$

and the pulled total diversity (PTD),

$$N_{\mathrm{p}} = N\frac{\lambda_{\mathrm{o}}}{\lambda} \quad (4)$$

(estimation formulas provided in Supplementary Information section 1.3). Here, λ_{o} refers to the most recent speciation rate (that is, as observed near the tips of the tree) and N is the total diversity at any given point in time. The PER and PTD are equal to the extinction rate μ and the total diversity N, respectively, when λ is constant ($\lambda = \lambda_{\mathrm{o}}$). If λ varies slowly ($\lambda^{-1}\mathrm{d}\lambda/\mathrm{d}t \ll \mu$), the recent μ_{p} still resembles the recent extinction rate, although the difference increases for older ages. Rapid variations in λ and/or μ will usually lead to substantial variations in μ_{p} and r_{p}.

In contrast to conventional maximum-likelihood or Bayesian methods[26,27] for estimating λ, μ, r and N, the pulled variables μ_{p}, r_{p} and N_{p} can be estimated from the LTT for each past time point without any assumptions about how λ and μ varied over time, and without fitting a specific parameterized model. Model fitting is the current de facto standard in phylogenetics-based reconstruction of diversification[25,28], and is, in fact, included in the present study. However, it requires that a parameterized form be specified beforehand for λ and μ; for example, accounting for rate shifts at discrete time points, leading to well-known trade-offs between model realism and temporal resolution on the one hand versus model simplicity and confidence in parameter estimation on the other hand. The caveat is that μ_{p}, r_{p} and N_{p} are composite variables, and in general, solely knowing μ_{p}, r_{p} and N_{p} does not unambiguously determine the constituents μ, λ, r and N. This limitation can be traced back to the fact that extinction partly erases a clade's history[29] (further discussion in Supplementary Information section 5).

As we demonstrate here, pulled variables are a powerful tool for obtaining insight into past diversification dynamics and for testing model assumptions. Using timetrees simulated under realistic scenarios, we found that pulled variables can reveal past changes in diversification rates, such as those due to mass extinction events, oscillating speciation rates and short temporary spikes in the speciation rate, as well as diversity-dependent speciation and extinction rates (Fig. 1, Supplementary Figs. 1–6 and Supplementary Information section 2). In particular, our simulations revealed that changes in the speciation and/or extinction rate usually lead to similarly strong changes in μ_{p} and that, reciprocally, a constant μ_{p} over time is a strong indication that both λ and μ were constant or varied only slowly over time (details in Supplementary Information section 4). Our simulations also revealed that the magnitude of the PDR is usually comparable to the magnitude of the diversification rate, and in fact, in all of our simulations, the two closely resembled each other. Furthermore, we found that N_{p} provides a quick way to roughly estimate past total diversities to order-of-magnitude accuracy (Fig. 1a–e), provided λ does not change drastically over time (that is, by orders of magnitude).

Estimating extant global bacterial diversity

Estimating speciation and extinction rates and past total diversities from a timetree requires knowledge of the fraction of extant diversity represented in the tree[15]. Substantial uncertainty currently exists regarding the extent of extant bacterial diversity, with estimates ranging from a few million OTUs[30] to trillions of OTUs[31]. To better constrain extant bacterial diversity, we examined 165,422 bacterial OTUs recovered de novo from 16S rRNA sequences amplified from various environments, such as animal guts, the ocean, lakes and soils (60 distinct studies comprising 6,303 samples). De novo OTUs covered ~200 base pairs (bp) in the V4 region of the 16S rRNA gene—a region commonly targeted in microbial taxonomic surveys[32]. We calculated the overlap of these de novo OTUs with SILVA—one of the largest 16S rRNA sequence databases[33]—to estimate the fraction of extant bacterial and cyanobacterial OTUs covered by SILVA, as well as the total number of extant OTUs. Our approach is analogous to traditional mark–recapture approaches for estimating population sizes[34], whereby the number of individuals found in a second survey (analogous to the number of OTUs in SILVA) is divided by the fraction of individuals marked in a first survey (analogous to our de novo dataset) that were recaptured by the second survey. We found that SILVA covers ~33% of de novo OTUs at 99% similarity. Based on the presence of 448,112 16S rRNA clusters in SILVA (that is, obtained by clustering SILVA's full-length 16S rRNA sequences at 99% similarity), we estimate that there exist globally ~1.4 million bacterial full-length OTUs (overview in Supplementary Table 2).

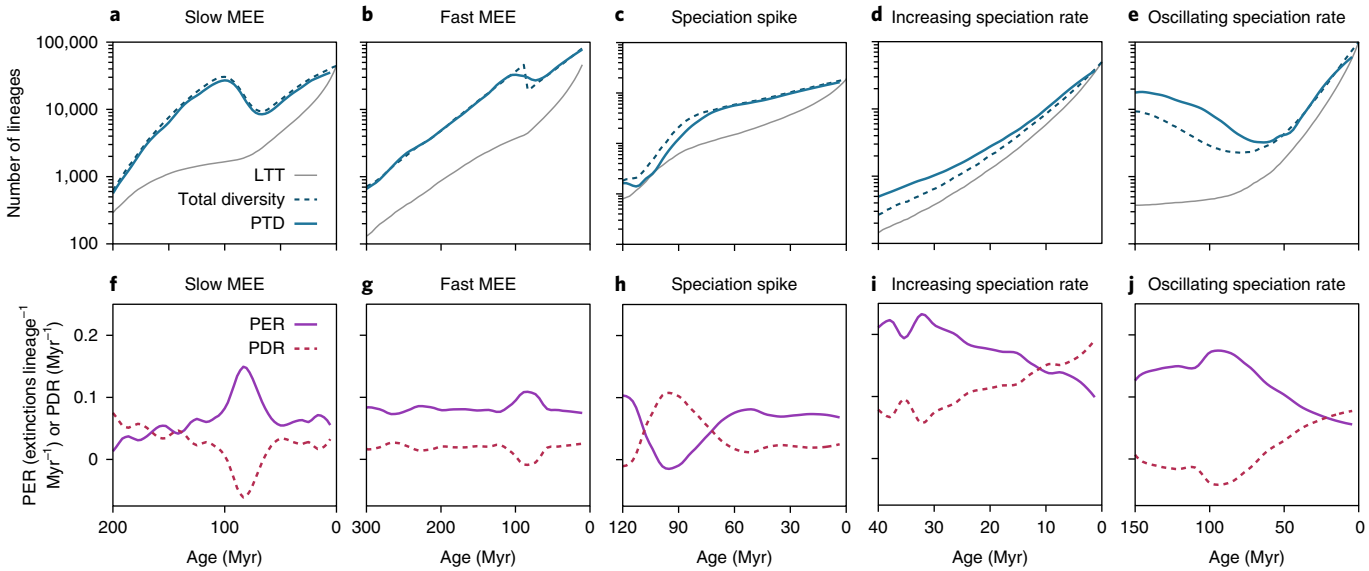

Fig. 1 | Non-parametric methods capture complex diversification scenarios. a–e, LTT (grey continuous curves), total diversities (dashed curves) and non-parametrically estimated PTDs for trees simulated under various realistic scenarios, including a slow mass extinction event ~80 Ma (**a**), a fast mass extinction event ~90 Ma (**b**), a speciation spike ~90 Ma (**c**), a gradually increasing diversity-dependent speciation rate (**d**) and an oscillating speciation rate (**e**). **f–j**, PERs (solid curves) and PDRs (dashed curves), estimated non-parametrically from the LTT in **a–e** under the same scenarios. In all scenarios, PER and diversification rates reveal changes in extinction and/or speciation rates, and PTDs approximately resemble the true total diversities (known in this case, since trees were generated via simulations). In **g**, the estimated PER and PDR are damped due to our noise filter, which blurs short (~1 Ma) fluctuations, although the mass extinction event's footprint is still clearly visible in the LTT and PTD (**b**). For more details and additional simulation examples, see Supplementary Information section 2 and Supplementary Figs. 1–6.

This estimate is robust (in order of magnitude) to variation in the methodology, sequencing depth and datasets used (all estimates are within 1.4–1.9 million; Supplementary Information section 3), and is comparable in order of magnitude to recent estimates by Schloss et al.[30]. Furthermore, based on the number of partial-length clusters in SILVA (that is, obtained by clustering the V4 region only), we estimate that there exist globally ~451,000 bacterial partial-length (V4) OTUs.

Estimating bacterial speciation and extinction rates

To estimate recent bacterial speciation and extinction rates, we fitted parametric models to the LTT of the timetrees over a relatively short recent time interval (200 Myr). To gain further insight into past diversification dynamics and to scrutinize model assumptions, we also fitted models over a more extended time interval (1,000 Myr) and used non-parametric methods to estimate PDRs, PERs and PTDs. We found that simple models with constant speciation and extinction rates fitted bacterial and cyanobacterial LTTs well over the past 1 Gyr (mean relative deviation (MRD) below 5% in all cases; Fig. 2a,b and Supplementary Fig. 21a–d). This indicates that overall speciation and extinction rates were roughly constant over time. This conclusion is supported by our observation that fitted speciation and extinction rates change only moderately (by less than 35% in all cases; Supplementary Fig. 18) when models are fitted over the shorter time interval (200 Ma). Our conclusion is also consistent with our estimates that PERs were almost constant over time and almost identical to the extinction rates fitted in the models (less than 10% difference at any time point; Fig. 2d–f and Supplementary Fig. 18). As mentioned previously, a constant PER in and of itself is indicative of constant or only slowly varying speciation and extinction rates, because a rapidly and strongly varying speciation and/or extinction rate usually results in a varying PER (see simulations in Fig. 1 and discussion in Supplementary Information section 2). Based on the small variation observed in the PERs, speciation and extinction rates must have had relative rates of change below ~0.005 Myr^{-1} (see explanation in Supplementary Information section 4). In comparison, estimated PERs for birds and vascular plants vary substantially over time (Fig. 2i and Supplementary Fig. 19c, using previously published timetrees[35,36]).

Our findings suggest that, during the past 1 Gyr, global bacterial speciation and extinction rates were not substantially affected during the mass extinction events seen in eukaryotic fossil records[3–5]. This conclusion does not support previous speculations that extinctions of plant- and animal-associated bacteria—resulting from extinction of their hosts—may contribute substantially to bacterial extinction rates[21]. The frequent existence of multiple ecotypes within single OTUs[37,38] may have facilitated bacterial lineage persistence during environmental perturbations and eukaryotic mass extinctions. Even if bacteria experienced extinctions at local scales because of environmental perturbations[39], these extinctions may have been largely buffered at global scales due to wide dispersal ranges[40]. Our findings also suggest that overall bacterial speciation and extinction rates were not dramatically altered by eukaryotic radiation events, such as the radiation of animals ~600 Myr ago (Ma) or the emergence of land plants ~465 Ma. It is possible that diversification within individual bacterial clades may have been influenced by eukaryotic radiations and extinctions[11,12], and that these cases are overshadowed when considering all bacteria together. We also cannot rule out slow effects on speciation and extinction rates (at time scales of billions of years), nor brief fluctuations (shorter than ~1 Mya) with little effect on total diversity, both of which could be missed by our methods.

We emphasize that our results do not imply that speciation and extinction rates are homogeneous across clades ('clock-like'). For example, Marin et al.[14] found variable diversification rates across lineages of the Firmicutes bacterial phylum. Using simulations, we found that timetrees, in which speciation and extinction rates are evolving heritable traits and are thus not clock-like, can be fitted well by models with homogenous rates; in these cases, fitted speciation and extinction rates approximately correspond to the

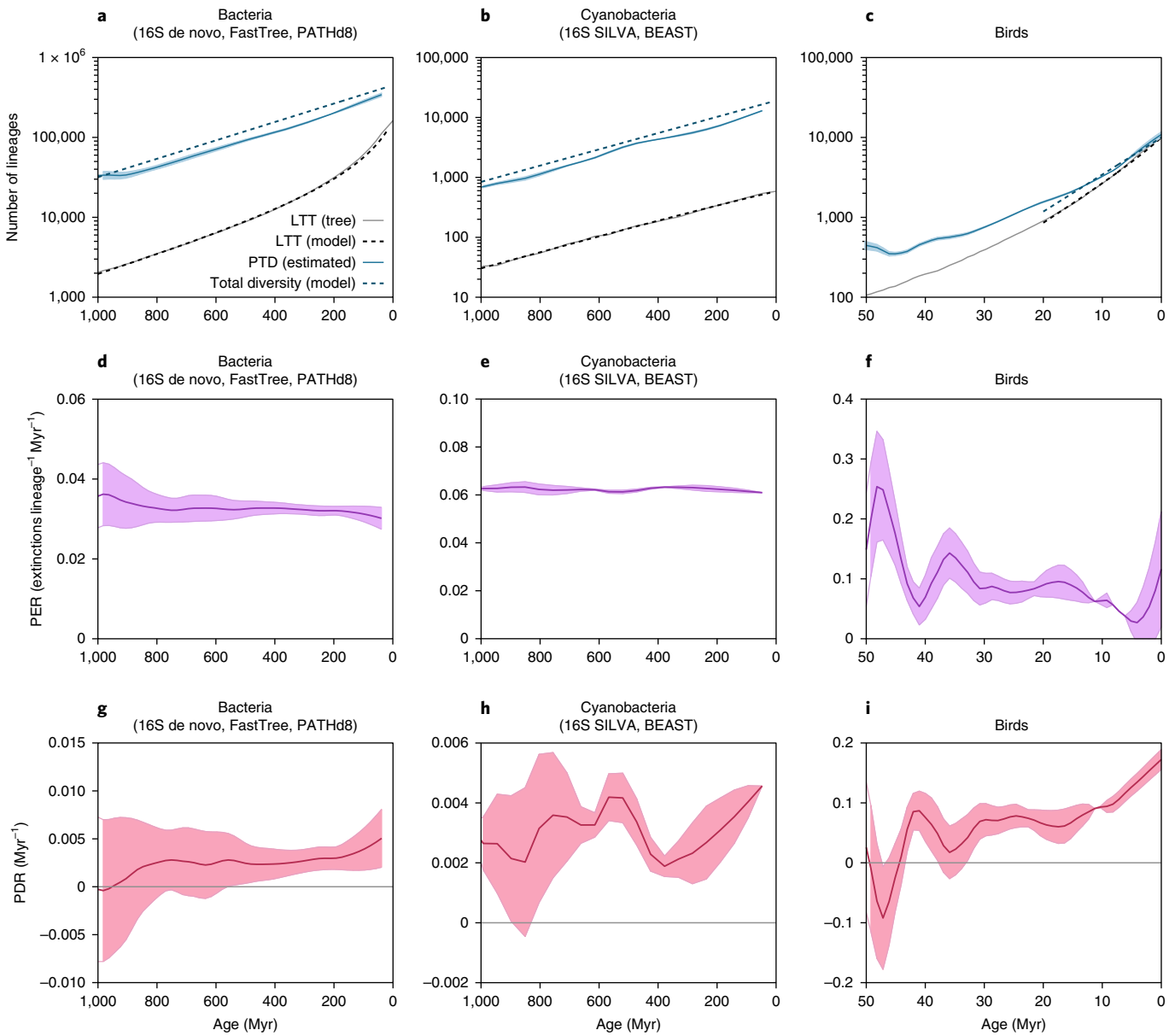

Fig. 2 | Bacterial, cyanobacterial and bird diversification dynamics through time. a–c, LTT (grey solid curves) for bacteria (**a**), cyanobacteria (**b**) and birds (**c**), compared with speciation–extinction models fitted over the past 1 Gyr (grey dashed curves). Blue solid curves show non-parametrically estimated PTDs, while blue dashed curves show total diversities predicted by the fitted models. Note that each tree only comprises a subset of extant taxa; thus, the LTT does not coincide with total diversity at age 0 (the right-most point on the blue curve). The total diversity at age 0 is only an estimate, based on the fraction of de novo OTUs covered by SILVA (details in main text; overview in Supplementary Table 2). Also note the different time scales shown for birds (50 Myr) compared with bacteria and cyanobacteria (1,000 Myr). **d–f**, PERs (equation (13)), estimated non-parametrically from the same trees as in **a–c**. The roughly constant PERs in **d** and **e** are indicative of constant or only slowly varying speciation and extinction rates, consistent with the fitted models. **g–i**, PDRs (equation (1)), estimated non-parametrically from the same trees, respectively, as in **a–c**. In all panels, shading indicates standard errors of noise-filtered estimates. Summaries of timetree construction methods are indicated in the labels at the top. The bird tree was obtained from Jetz et al.[36]. For analogous figures using alternative timetrees, see Supplementary Fig. 21. For analogous figures for vascular plants, see Supplementary Fig. 19.

average speciation and extinction rates over all lineages (details in Supplementary Information section 4). This means that if bacterial speciation and extinction rates deviated from clock-like behaviour, our clock-like models would not necessarily be able to detect this deviation. Hence, despite our finding of roughly constant overall bacterial speciation and extinction rates over time, we cannot rule out potential differences between clades at any given time point. We also point out that our results only pertain to overall global bacterial diversification and do not distinguish between environments (for example, terrestrial versus marine).

Our fitted models suggest that recent overall extinction rates are 0.03–0.05 extinctions lineage^{-1} Myr^{-1} for bacteria and 0.02–0.06 extinctions lineage^{-1} Myr^{-1} for cyanobacteria (Fig. 3b and Supplementary Fig. 18b). These estimates are consistent with estimated recent PERs (Fig. 2d–f and Supplementary Figs. 21d–f and 22a). Our estimates are robust (to an order of magnitude) against variations in the dating of our timetrees (for example, 0.015–0.05 extinctions lineage^{-1} Myr^{-1} for bacteria; Supplementary Fig. 23). For comparison, using the same methods, we also estimated global extinction rates for vascular plants (~0.35 extinctions lineage^{-1} Myr^{-1}) and birds

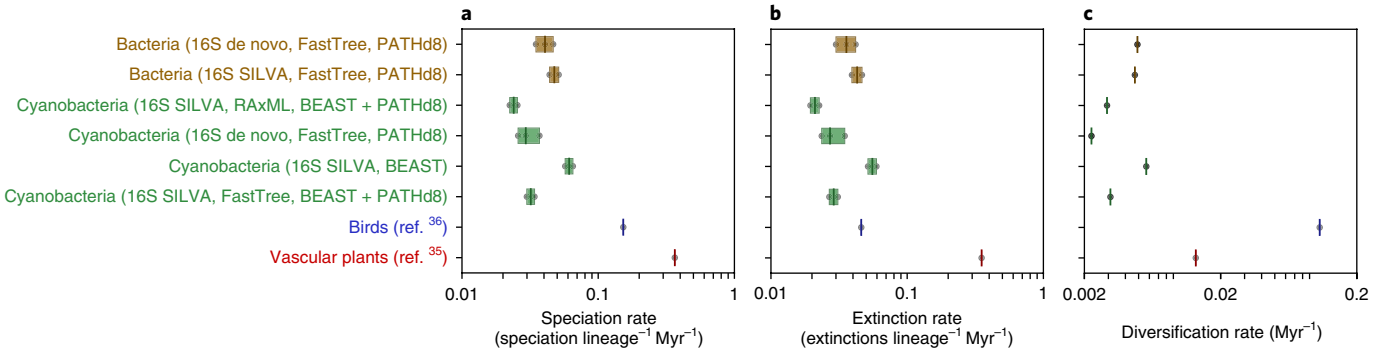

Fig. 3 | Estimated recent speciation, extinction and diversification rates. a–c, Recent speciation rates (**a**), extinction rates (**b**) and diversification rates (**c**), estimated for various taxa and using various timetrees (one box per timetree). Estimates were obtained by fitting cladogenic models over a recent time interval of 200 Myr (for estimates over longer time intervals, see Supplementary Fig. 18). For each bacterial or cyanobacterial timetree, various alternative estimates of incomplete sampling fractions were used (Supplementary Tables 2–4); boxes span the results from all alternatives. Tree labels and boxes are coloured by taxon. Summaries of timetree sources and construction methods are indicated in brackets (see Methods for details).

(~0.05 extinctions lineage^{-1} Myr^{-1}), reproducing previous estimates for plants and animals (order of magnitude: ~0.1 extinctions lineage^{-1} Myr^{-1})[41]. We once more point out that bacterial OTUs are only approximately analogous to plant and animal species[24,42]; hence, comparisons between the two should be treated with caution.

We further found that bacterial speciation rates are only slightly above extinction rates—an observation commonly made for larger organisms[29]. Specifically, fitted bacterial diversification rates (~0.004–0.005 Myr^{-1}) are much lower than fitted extinction and speciation rates (Fig. 3c). These values are consistent with similarly low estimated PDRs (~0.003–0.004 Myr^{-1}; Fig. 2g and Supplementary Figs. 21i and 22b). Because bacterial extinction rates are so close to speciation rates, most bacterial lineages that ever existed are now extinct. Based on our fitted models, for each extant bacterial and cyanobacterial OTU there have been ~10–14 and ~5–24 extinctions over the past 1 Gyr, respectively. The conclusion that only a small fraction of bacterial lineages survived to the present resembles analogous observations for plants and animals[29]. Our finding of substantial bacterial extinction contrasts with reports that diversification within the *Aeromonas* bacterial genus is best explained without extinction[11], although most analyses yielding zero extinction rates are arguably probably wrong[29]. Nevertheless, at this point, we cannot exclude the possibility that some younger clades (for example, genera) may exhibit much lower extinction rates than the bacterial average.

Global bacterial diversity increases over time

According to all fitted models, bacterial and cyanobacterial diversification rates have been positive over the past 1 Gyr, suggesting an increase in the total diversity over time. Consistent with this, estimated PDRs are also mostly positive over the past 1 Gyr (Fig. 2g,h and Supplementary Fig. 21g–i), and PTDs (N_p) increase roughly exponentially over time (Fig. 2a–c and Supplementary Fig. 21a–c). In principle, a positive PDR and an exponentially increasing N_p could be a mere result of a decreasing speciation rate over time (that is, $\lambda > \lambda_o$ in equation (14)), rather than reflecting a truly increasing total diversity. This scenario seems unlikely because it would imply that λ (or the ratio λ/N, if N also varied) decreased substantially and approximately exponentially over the past 1 Gyr, and that μ followed in a similar way (since $\lambda - \mu \sim 0$). It is hard to imagine a simple scenario that would lead to these specific trajectories in λ and μ (see discussion in Supplementary Information section 4). Instead, a simpler explanation for an exponentially increasing N_p is that λ and μ were approximately constant and thus N increased fairly steadily over time. This interpretation is also supported by the fact that when plotted on a logarithmic axis over time, PTDs exhibit a similar slope to the total diversities predicted by the fitted constant-rate models. A continuous increase in bacterial diversity has been observed previously in a smaller dataset[14]. Similarly, Gubry-Rangin et al.[13] found a stably high diversification rate within the Thaumarchaeota archaeal phylum over the past 400–700 Myr. A continuous increase in bacterial diversity is also comparable to the continuous increase of diversity observed in many eukaryotic taxa during the past 200 Myr[4]. However, we emphasize that diversification rates reconstructed here a priori only reflect 16S branching dynamics and may not reflect ecological diversification[43].

Conclusions

Our analysis sheds light on bacterial diversification over geological time. We found evidence that global bacterial diversity has mostly increased over the past 1 Gyr, with roughly constant or only slowly changing overall speciation and extinction rates when averaged over all clades. This conclusion has implications for how life unfolded over Earth's history, since bacteria are the most ancient and the most ubiquitous form of life on Earth[44]. We estimated that global bacterial extinction rates are only slightly below their speciation rates, and that only a small fraction of bacterial lineages that ever existed survived to the present. This has important implications for how we interpret records of ancient life. Some authors have interpreted morphological similarities between microfossils and extant bacterial taxa as signs of 'extreme evolutionary stasis' and absence of speciation and extinction[17,20], while others even consider cyanobacteria to be living fossils that do not go extinct[19]. Our finding that lineage turnover is an important aspect of bacterial evolution suggests that it is possible that ancient microfossils belonged to extinct lineages, regardless of whether morphology was conserved or convergent[45], although these extinct lineages may be stem lineages of extant groups. In a similar fashion, it is possible that some ancient molecular biomarkers, such as fossil lipids[46], were produced by lineages that have gone extinct.

Our work extends various empirical palaeontological 'laws' of macrobial evolution[29] to bacteria—namely, that: extinction is an integral part of evolution; lineages are short-lived at geological time scales; the number of extinct species far exceeds the number of extant species; and speciation and extinction rates are typically similar in magnitude. Despite the high diversity of extant microorganisms, this diversity only represents a snapshot of the microbial diversity ever to have inhabited our planet.

Methods

Estimating the total number of extant OTUs. To estimate the total number of extant bacterial OTUs, we used two alternative approaches. In the first approach,

a large random set of partial-length OTUs (99% identity in the V4 region of the 16S rRNA gene), recovered de novo from environmental samples, was compared with the SILVA SSU database (release 128)[47], and the total number of extant OTUs was estimated based on the overlap between the de novo OTUs and SILVA. In the second approach, variable-length 16S rRNA sequences extracted from metagenome-assembled genomes (MAGs)[48] were compared with SILVA, and again the total number of extant OTUs was estimated based on the overlap between the MAG 16S rRNA sequences and SILVA. In both cases, non-redundant (NR99) full-length 16S rRNA sequences in the SILVA database were first clustered at 99% identity using uclust version 1.2.22 (ref. [49]) (options: –usersort –nucleo). Furthermore, to assess the potential phylogenetic bias of OTUs represented in SILVA, and how this bias affects the representation of older clades compared with random OTU sampling, we constructed and dated a phylogenetic tree of the de novo OTUs and counted the fraction of lineages in the tree over time that was represented in SILVA. Below, we describe these procedures in detail.

Generating de novo OTUs. We downloaded public raw Illumina reads of 16S rRNA gene amplicons (V4 region) from the European Nucleotide Archive (https://www.ebi.ac.uk/ena) for 6,303 samples from 60 studies across the globe, including animal guts, marine sediments and water columns, soils, bioreactors, lakes, and phytotelmata (henceforth referred to as the 'de novo dataset'; accession numbers provided in Supplementary File 1). We focused particular effort on the inclusion of soils (1,067 samples), which are thought to host a large fraction of Earth's bacterial diversity[32]. Any paired-end reads were merged using flash version 1.2.11 (ref. [50]) (options: -min-overlap=20 -max-mismatch-density 0.25 -phred-offset=33 -allow-outies). Merged and single-end reads were trimmed and quality-filtered using vsearch version 2.4.3 (ref. [51]), keeping only reads at least 200 bp long after trimming (options: -fastq_ascii 33 -fastq_minlen 200 -fastq_qmin 0 –fastq_maxee 1 -fastq_truncee 1 -fastq_maxee_rate 0.005 -fastq_stripleft 7). Samples with more than 200,000 quality-filtered reads were rarefied down to 200,000 reads to reduce the computation time, by randomly picking reads without replacement. Rarefied reads were chimera-filtered de novo, separately for each sample, using vsearch (options: --abskew 1.9 -mindiv 0.5 -minh 0.1), yielding 265,640,818 quality-filtered and chimera-filtered reads in total, with a mean length of 262 nucleotides. Pooled reads from all samples were error-filtered and clustered de novo at 99% similarity using cd-hit-otu version 0.0.1 (ref. [52]), yielding 345,229 OTUs. OTUs were subsequently filtered anew for chimeras using vsearch (same options as before), yielding 216,707 OTUs. Lastly, any OTUs found in fewer than 2 samples were omitted to further reduce spurious OTUs, leaving us with 185,620 OTUs for downstream analysis.

We note that the above quality filters were chosen to be quite stringent in order to minimize the recovery of spurious OTUs (for example, stemming from sequencing errors or chimeras); however, this conservatism potentially came at the cost of also removing real OTUs. Minimizing the recovery of spurious OTUs is important for both a correct estimation of global bacterial diversity and improving the quality of the phylogenetic tree generated from the OTUs (see below). We emphasize that falsely omitting real OTUs does not affect our estimation of global bacterial diversity, as long as the inclusion or omission of an OTU is independent of the OTU's presence in SILVA. For similar reasons, while our omission of OTUs found in fewer than two samples may bias our census towards more cosmopolitan OTUs, it should not affect our estimation of global bacterial diversity. Indeed, when we also included OTUs found only in a single sample, our estimate of global bacterial diversity changed by less than 10%.

We point out that our mark–recapture-type approach may have underestimated global extant bacterial diversity if some bacterial OTUs are generally much more difficult to detect than others (for example, if they are only present in very specialized environments). In particular, extreme environments (such as hot springs) are under-represented in our de novo dataset, although these environments host relatively low diversity compared with other environments, such as soils. Future, more exhaustive sequencing studies, including a greater variety of environments, will undoubtedly improve estimates of extant bacterial diversity. An underestimation of extant bacterial diversity in the present study would mean that bacterial extinction rates are even higher than estimated here[15]. However, it would not affect our conclusions regarding the constancy of overall bacterial speciation and extinction rates over time, nor our conclusions regarding the continuous nearly exponential increase in bacterial diversity, unless the missed diversity was strongly phylogenetically biased and clustered within the tree—a scenario we view as unlikely.

Fraction of de novo OTUs represented in SILVA. De novo OTUs were taxonomically identified using a consensus approach based on the first 10 hits in SILVA at a similarity threshold of at least 70%. Specifically, OTUs were globally aligned to the SILVA non-redundant (NR99) SSU reference database using vsearch version 2.4.3 (ref. [51]) at a minimum similarity of 70% (options: --id 0.7 --strand both --iddef 2 --maxaccepts 10 --uc_allhits), while keeping track of the taxonomies provided by SILVA for each hit. For any given OTU, if at least one hit had a similarity of 100%, all hits with a similarity of 100% were considered candidates for forming a consensus taxonomy. Otherwise, if at least one hit had a similarity of $s \geq 70\%$, all hits with a similarity of at least $(s-3\%)$ were used as candidates for a consensus taxonomy. For any candidate set of hits, the consensus taxonomy was defined as the taxon at the lowest taxonomic level possible, containing all of the candidate hit taxonomies. If an OTU did not match any SILVA entry at or above 70% similarity, or did not form a consensus taxonomy even at the domain level, it was considered unidentified and was subsequently omitted. A total of 171,816 OTUs could be identified at some taxonomic level. OTUs identified as eukaryotes, chloroplasts and mitochondria were omitted from all subsequent analyses. Taxonomically identified OTUs were matched to the SILVA non-redundant (NR99) set using vsearch (options: --iddef 2 --strand both) at a similarity threshold of 99%. For any given focal taxon (for example, bacteria, archaea or cyanobacteria), we estimated the fraction (ρ) of extant OTUs represented in SILVA as the fraction of taxonomically identified de novo OTUs that could be aligned to the clustered SILVA database (clustered at 99% identity) at a similarity of ≥99%.

Total number of extant OTUs (based on overlap with SILVA). Since de novo OTUs only cover a fraction of the 16S rRNA gene (~200 bp from the V4 hypervariable region), correctly estimating the number of extant partial-length (V4) OTUs requires knowledge of the number of V4 OTUs already contained in SILVA. Therefore, we extracted and clustered the part of 16S rRNA sequences in SILVA corresponding to the region covered by de novo OTUs. Specifically, we aligned de novo OTUs to SILVA using the QIIME script parallel_align_seqs_pynast.py[53], and using a reduced set of the SILVA alignments (clustered at 90% similarity) as a template. We then identified the first nucleotide position in the OTU alignments that had a gap fraction below 0.9, and extracted the part of the NR99 SILVA alignments starting at that nucleotide position and extending 200 bp in the $5' \rightarrow 3'$ direction (omitting gaps). Extracted partial SILVA sequences were then clustered at 99% identity using uclust version 1.2.22 (ref. [49]), yielding 161,070 bacterial and archaeal partial-length clusters.

The global number of extant V4 OTUs in the focal taxon was estimated as N_{V4}/ρ, where N_{V4} is the number of V4 clusters within the focal taxon in SILVA, and ρ is the previously estimated fraction of extant OTUs within the focal taxon represented in SILVA. To estimate the number of extant full-length OTUs, we multiplied the estimated number of extant V4 OTUs by the ratio N_{FL}/N_{V4}, where N_{FL} is the number of full-length clusters within the focal taxon in SILVA. Estimated fractions of V4 OTUs represented in SILVA at 99% similarity, as well as total numbers of extant V4 OTUs and full-length OTUs are listed in Supplementary Table 2. We note that one important assumption of the above estimation method is that the presence or absence of an OTU in the de novo dataset is independent of its presence or absence in SILVA. This assumption is probably approximately met, since OTUs in the de novo dataset were recovered without the use of any reference database and the de novo dataset covers a wide range of environments. We emphasize that this assumption does not imply that SILVA is phylogenetically unbiased (that is, we do not assume that OTUs co-occur in SILVA regardless of their phylogenetic relatedness). In fact, we detected substantial phylogenetic bias in SILVA when comparing the observed representation of deeply branching clades with the expectation under random unbiased sampling (Supplementary Fig. 7). This bias can explain why, despite the overall high fraction of extant OTUs already represented in SILVA, recent studies have discovered new deeply branching clades (for example, phyla) not represented in SILVA[16,48].

Fraction of MAG 16S rRNA sequences represented in SILVA. To further verify our estimates of global extant bacterial and cyanobacterial OTUs using a separate method not constrained by potential primer bias, we also used 16S rRNA sequences from MAGs. Specifically, we downloaded 16S rRNA sequences for 2,853 MAGs[48] from https://data.ace.uq.edu.au/public/misc_downloads/uba_genomes/ on 25 October 2017. Any sequences shorter than 500 bp were omitted, leaving us with 1,166 sequences for downstream analyses. Sequences were taxonomically identified and globally aligned to the clustered SILVA using the same approach as the de novo OTUs (see above). The global number of extant OTUs in each focal taxon was estimated as N_{FL}/ρ, where N_{FL} is the number of full-length clusters within the focal taxon in SILVA, and ρ is the fraction of MAG 16S rRNA sequences that could be aligned to the clustered SILVA at a similarity threshold of 99%. The results are shown in Supplementary Table 2. While this set of 16S rRNA sequences is much smaller than the de novo OTUs, and obtained from a much smaller set of samples, it can serve as a rough verification of the estimates obtained from de novo OTUs. For bacteria as well as cyanobacteria, estimates based on MAGs (Supplementary Table 3) are similar to estimates based on de novo OTUs (Supplementary Table 2). To assess the robustness of our estimated recent speciation, extinction and diversification rates (Fig. 3 and Supplementary Fig. 18), we performed our analyses based on the total number of extant OTUs estimated from de novo OTUs, as well as from MAGs (Supplementary Tables 2–4).

Total number of extant OTUs (based on overlap with the Earth Microbiome Project (EMP)). To obtain an additional independent estimate of the number of extant OTUs, we repeated our analysis by considering the overlap between our de novo OTUs and V4 OTUs recovered from a dataset published by the EMP[32]. Specifically, we downloaded raw reads generated by the EMP based on the run accession numbers provided on the EMP GitHub (https://github.com/biocore/emp/blob/master/code/download-sequences/download_ebi_fasta.sh). Project number ERP010098 was omitted as the sequencing instrument was unspecified.

EMP reads were processed similarly to the de novo dataset described above, with the following differences: the minimum allowed read length (after quality filtering) was reduced to 100 bp to accommodate the much shorter EMP reads, and the number of quality-filtered reads per sample was limited to 20,000. This yielded 400,528 chimera-filtered OTUs, representing 195,291,388 reads from 18,034 samples across 44 studies. OTUs were taxonomically identified as before, and any OTUs identified as eukaryotes, chloroplasts or mitochondria were omitted. OTUs found in fewer than 2 samples were also omitted, leaving us with 343,743 taxonomically identified OTUs. To calculate the fraction of EMP OTUs represented by our de novo OTUs, we matched the EMP OTUs against the de novo OTUs using vsearch (options: --iddef 2 --strand both) at a similarity threshold of 99%. The total number of extant V4 OTUs within a focal taxon was estimated as N_{dn}/ρ, where N_{dn} is the number of de novo OTUs within the focal taxon, and ρ is the fraction of EMP OTUs within the focal taxon that could be matched to a de novo OTU. The total number of extant full-length OTUs was estimated as before; that is, by multiplying the estimated number of extant V4 OTUs by the ratio N_{FL}/N_{V4}, where N_{FL} and N_{V4} are the number of full-length and V4 clusters, respectively, within the focal taxon in SILVA.

Building a tree from de novo OTUs. Representative sequences of taxonomically identified bacterial and archaeal de novo OTUs, excluding chloroplasts and mitochondria, and with a length of at least 200 nucleotides, were aligned using the QIIME script parallel_align_seqs_pynast.py[53], and using a reduced set of the SILVA alignments (clustered at 90% similarity) as a template. A total of 171,510 OTUs could be successfully aligned. Alignments were reduced by first removing nucleotide positions with >95% gaps, and then removing the top 5% most entropic nucleotide positions. Taxonomic identities at the domain, phylum, class and order level from the preceding step were used to create split constraints for FastTree[54] by constraining each taxon to be on a single side of a split. Taxa with fewer than two OTUs were omitted from the constraints. A total of 603 constraints were defined. Using the alignments and the taxonomically generated constraints, we constructed a phylogenetic tree with FastTree (options: -spr 4 -gamma -no2nd -constraintWeight 100000). The phylogenetic tree was re-rooted so that bacteria and archaea were split at the root. The resulting tree was then dated with PATHd8 using the following dating anchors:

- GOE (secondary constraint). Most recent common ancestor (MRCA) of Oxyphotobacteria and Melainabacteria constrained by the Great Oxygenation Event[55] and based on a molecular clock analysis by Shih et al.[56] (table 3 therein).
- Ri. MRCA of Rickettsiales constrained to before the earliest known appearance of mitochondria[57].
- CB. MRCA of Chlorobium and Bacteroidetes constrained to before the first known Chlorobium-specific biomarkers[58].
- Chr. MRCA of Chromatiaceae constrained to before the first known purple sulfur bacterial biomarkers[46,58].
- LUCA. MRCA of archaea and bacteria constrained to before the earliest known stromatolites and after the late heavy bombardment[59,60].

The dating anchors are summarized in Supplementary Table 5.

Fraction of extant lineages represented in SILVA through time. To assess the phylogenetic bias of SILVA, we calculated the fraction of lineages in the de novo tree represented in SILVA over time (that is, the fraction of discovered lineages (FDL); Supplementary Fig. 7a). Specifically, we extracted the de novo subtree comprising only the OTUs matched to SILVA at 99% similarity (see previous paragraph), counted the remaining LTT and divided those by the LTT in the full de novo tree. Note that this fraction loosely corresponds to the probability that a lineage at any age would be represented in SILVA, provided that it has not gone extinct. To examine how the FDL differed in the case of the null model of random independent sampling of OTUs, we replaced the OTUs represented in SILVA with a new, equally sized set of randomly chosen OTUs and recalculated the FDL. The variability of the outcome was assessed by repeating this procedure 100 times. To further examine, based both on SILVA and the null model, how the FDL through time depends on the fraction of OTUs discovered (sampling fraction), we subsampled SILVA down to various fractions (for example, 10 and 1%) and repeated the previous analysis (Supplementary Fig. 7b,c). We found that the FDL through time deviated substantially from the null model, indicating phylogenetically correlated representation of OTUs in SILVA, with some clades being over-represented or under-represented compared with random OTU sampling. For strong subsampling of SILVA ($\lesssim$1%; Supplementary Fig. 7c), this deviation diminished towards recent ages.

Comparison with the 97% similarity threshold. For comparison purposes, we repeated some of the above calculations for de novo OTUs clustered at 97% similarity. Specifically, we re-clustered de novo OTUs at a coarser similarity threshold of 97% using vsearch (options: --cluster_fast --usersort –id 0.97 --iddef 2 --strand plus) and used the selected centroids as new representative sequences. Taxonomic identification was done as described above for de novo OTUs. Coverage by SILVA, as well as the total number of extant 97% OTUs, were estimated similarly to above, the difference being that SILVA was clustered at 97% similarity and coverage by SILVA was calculated at 97% similarity. A timetree comprising 31,231 de novo 97% OTUs was constructed as above. We do not discuss these results in the main text, but provide them in Supplementary Table 6.

Tree construction and dating. To verify the robustness of our results, we examined a multitude of bacterial and cyanobacterial timetrees, constructed using various alternative methods, described below. Trees were constructed using either full-length 16S rRNA alignments from the SILVA reference database (release 128; non-redundant set)[33] or partial-length alignments of de novo clustered OTUs from public amplicon sequences from various environmental samples. Unless otherwise mentioned, tips in all 16S-based trees corresponded to OTUs clustered at 99% similarity. We note that bacterial OTUs were historically delineated using a similarity threshold of 97%. However, modern genomics revealed that taxa defined on this basis are usually underspeciated, and that a greater similarity threshold ($\gtrsim$99%) is required for distinguishing ecologically differentiated organisms[23,24,42,61]. In this study, we thus delineated OTUs at 99% similarity, but we also performed comparisons at 97% similarity (provided in Supplementary Figs. 8 and 9, and Supplementary Table 6). We stress that bacterial OTUs (even at the similarity threshold of 99%) only provide an approximate 'species' analogue, and any given OTU may still comprise multiple closely related strains with different genomic contents and ecological strategies[38]. Even formally named bacterial 'species' can display strong genomic and phenotypic strain diversity[37]. Hence, 'speciation' rates reported here probably represent a conservative estimate of the rate at which bacteria differentiate ecologically. Whether and how bacterial species can ever be reasonably defined remains an open question[62]; hence, the 16S rRNA gene remains a popular marker for cataloguing bacterial diversity and describing evolutionary relationships[42] in a well-defined and reproducible manner.

Birds. The bird tree was constructed and dated by Jetz et al.[36], and was downloaded from the project's website (http://litoria.eeb.yale.edu/bird-tree/archives/Stage2) on 1 August 2017 ('Hackett_backbone_stage2_tree_0001.tre'). We assumed a sampling fraction (ρ) of 1, since almost all bird species have probably already been discovered[66].

Vascular plants. The vascular plants tree was constructed and dated by Zanne et al.[35], and was downloaded from the Dryad Digital Repository (http://datadryad.org/resource/doi:10.5061/dryad.63q27.2). We assumed a sampling fraction of 0.724, according to estimates by Mora et al.[67] on the fraction of plant species discovered.

Bacteria (16S SILVA, FastTree, PATHd8). Non-redundant, full-length 16S rRNA gene alignments of 448,112 bacteria, 3 chloroplasts and one archaeon *Methanococcales* (for dating purposes), representing OTUs at 99% similarity, were extracted from SILVA. Alignments were reduced by first removing nucleotide positions with >95% gaps, and then removing the top 5% most entropic nucleotide positions. SILVA taxonomies provided for OTUs were used to define topological constraints for FastTree at the domain, phylum, class or order level, by constraining each taxon to be a monophyletic group in the new tree. Taxa with fewer than two OTUs were not constrained. A total of 625 constraints were defined. Using the reduced alignments, with the SILVA guide tree as a starting tree and using the taxonomic constraints, a new tree was generated with FastTree (options: -spr 4 -gamma -no2nd). The generated tree was re-rooted such that bacteria and archaea split at the root. The re-rooted tree was dated with PATHd8 using the anchors GOE, Chl, Ri, CB, Chr and LUCA, as listed in Supplementary Table 5. All archaea, chloroplasts and mitochondria were subsequently removed from the dated tree. An overview of the tree is shown in Supplementary Fig. 10.

Bacteria (16S SILVA, 97%, FastTree, PATHd8). This tree was created similarly to the previous bacterial tree (16S SILVA, FastTree, PATHd8), but with OTUs clustered at 97% similarity. An overview of the tree is shown in Supplementary Fig. 11.

Bacteria (16Sde novo, FastTree, PATHd8). As described above, a dated phylogenetic tree comprising bacterial and archaeal partial-length OTUs (99% identity in the V4 region) was constructed from de novo clustered 16S rRNA gene amplicon sequences from a wide range of environments. From this 'de novo' tree, we extracted the subtree comprising those OTUs identified as bacterial. An overview of the tree is shown in Supplementary Fig. 12.

Bacteria (16S de novo, 97%, FastTree, PATHd8). This tree was created similarly to the previous bacterial tree (16S de novo, FastTree, PATHd8), but with OTUs clustered at 97% similarity. An overview of the tree is shown in Supplementary Fig. 13.

Cyanobacteria (16S SILVA, FastTree, BEAST + PATHd8). This tree was constructed with 16S rRNA sequences from SILVA, then dated using secondary constraints inferred from a previously dated multigene cyanobacterial tree[56], as follows. Non-redundant full-length 16S rRNA alignments of all non-chloroplast cyanobacteria and representative chloroplasts were extracted from SILVA. Alignments were pre-processed and used to construct a tree with FastTree, in the same way as described above for the bacteria (16S SILVA, FastTree, PATHd8). The generated tree was

re-rooted such that the root separates the Melainabacteria from the rest of the tree[56]. Next, a previously published dated multigene tree, including 60 cyanobacterial and 37 plastid taxa, was obtained from Shih et al.[56] (run T65). The multigene tree is based on full-length 16S rRNA gene sequences and 10 additional marker genes, and was dated using BEAST. To link tips in our 16S rRNA-based tree to tips in the multigene tree, the original SILVA 16S rRNA alignments were de-aligned (all gap characters removed) and mapped to the 16S rRNA sequences of the strains in the multigene tree, via global alignment using vsearch (options: --id 1.0 --iddef 2 --strand both --maxhits 1 --maxaccepts 1). A total of 60 tips could be mapped. From this mapping, and based on the divergence times of nodes in the multigene tree, secondary dating constraints were generated for the 16S rRNA tree using the congruency method by Eastman et al.[68]. This method was developed for generating very large timetrees based on a smaller previously dated 'reference' timetree (in our case, the multigene tree), by identifying nodes that are concordant in the target tree (in our case, the non-dated 16S rRNA-based tree) and the reference tree. Using the congruency method, which was performed with the R package castor[69], a total of 17 concordant node pairs were identified. The divergence times of concordant nodes were then used as fixed constraints for dating the 16S rRNA-based tree using PATHd8. All chloroplasts were subsequently removed from the dated tree. This yielded a dated tree for 6,308 non-chloroplast cyanobacteria. An overview of the tree is shown in Supplementary Fig. 14.

Cyanobacteria (16S SILVA, BEAST). Non-redundant 16S rRNA alignments of non-chloroplast cyanobacteria and representative chloroplasts were subsampled randomly down to 586 OTUs (including 30 chloroplasts), and a tree was constructed and dated using BEAST, as follows. A log-normal relaxed molecular clock model was implemented using BEAST with the GTR + G substitution model on the 16S rRNA dataset. Chloroplast taxa from land plants were constrained to a normal distribution of 477 ± 70 Ma based on Smith et al.[70]. A uniform prior on the base of crown group cyanobacteria ranged from 1,909–2,450 Ma[56]. The younger boundary of the age constraint is based on the conservatively younger age reported by Shih et al.[56], while the older boundary represents a geological constraint on the origins of oxygenic photosynthesis and cyanobacteria based on the Great Oxygenation Event. A uniform prior was used to enable flexibility by allowing the MCMC search to agnostically converge on a date that best fit the data. Two separate MCMC chains were generated for 50 million generations, sampling every 10,000 generations, with the first 20 million generations discarded as burn-in. An overview of the tree is shown in Supplementary Fig. 15.

Cyanobacteria (16S SILVA, RAxML, BEAST + PATHd8). Non-redundant full-length 16S rRNA gene alignments of 6,302 cyanobacteria and 28 representative chloroplasts, representing OTUs at 99% similarity, were extracted from SILVA. Alignments were reduced by first removing nucleotide positions with >95% gaps, and then removing the top 5% most entropic nucleotide positions. Using the reduced alignments, and using the SILVA guide tree as a starting tree, a new tree was generated with RAxML Stamatakis2014RAxML (options: -m GTRCAT -p 34612 -f d). The generated tree was re-rooted such that the root separates the Melainabacteria from the rest of the tree[56]. The re-rooted tree was dated using PATHd8, based on secondary constraints extracted from a previously published multigene timetree[56], as described above (cyanobacteria, '16S SILVA, FastTree, BEAST + PATHd8'). All chloroplasts were subsequently removed from the dated tree. An overview of the tree is shown in Supplementary Fig. 16.

Cyanobacteria (16S de novo, FastTree, PATHd8). This tree was extracted from the de novo tree (see Methods), similarly to the bacterial (16S de novo, FastTree, PATHd8) tree. An overview of the tree is shown in Supplementary Fig. 17.

Some bacterial trees were constructed and dated simultaneously using BEAST[63], and some trees were first constructed using FastTree version 2.1.10 (ref. [54]) or RAxML version 8.2.9 (ref. [64]) and subsequently dated using PATHd8 version 1.0 (ref. [65]), depending on computational feasibility. Some trees were dated using primary dating anchors (summarized in Supplementary Table 5), while others were dated using secondary dating anchors extracted from timetrees previously constructed with BEAST. Note that, because PATHd8 (ref. [65]) requires at least one anchor with a fixed age, for timetrees dated using PATHd8 and primary anchors, the GOE anchor (split between Oxyphotobacteria and Melainabacteria) was fixed to an age of 2.55 Ga[56]. Below, we describe the source or construction method for each timetree in detail. An overview of all considered timetrees is provided in Supplementary Table 1.

Sampling fractions of trees (ρ) were calculated by dividing the number of tips in each tree by the total number of extant full-length OTUs, extant partial-length (V4) OTUs or extant species (whichever was appropriate) in the corresponding taxon (as estimated in this study; overview in Supplementary Tables 2–4). Figure 2 and Supplementary Fig. 21 were generated using the estimates listed in Supplementary Table 2. The same approach for fitting speciation and extinction models, and estimating speciation and extinction rates (see Methods), was applied to all timetrees. The timetrees analysed in Fig. 2 are bacteria (16S de novo, FastTree, PATHd8), cyanobacteria (16S SILVA, BEAST) and vascular plants[35]. Analogous results are shown for additional timetrees in Supplementary Figs. 19 and 21.

Estimated recent speciation, extinction and diversification rates are summarized in Fig. 3 and Supplementary Fig. 18. As seen in Fig. 3 and Supplementary Fig. 18, all bacterial and cyanobacterial timetrees yielded similar estimates for speciation, extinction and diversification rates. This reproducibility underlines the robustness of our estimates, despite potential inaccuracies in tree construction due to short sequence alignments (in the case of the de novo dataset) and due to heuristic algorithms (for example, FastTree) used for some of the trees for computational feasibility. This result may not be surprising given that our estimates are entirely based on the LTT curve, which is a high-level summary statistic and which, for larger trees, may be rather invariant to uncertainties in tree topology.

We mention that very few geological anchors are currently available for dating bacterial phylogenies and, in principle, 16S rRNA nucleotide substitution rates could vary strongly between clades[71]. This variation could therefore introduce errors when translating phylogenetic distances to temporal distances. However, such errors are generally expected to further increase the deviations of the resulting trees from the simple cladogenic models fitted here. Hence, our conclusion that constant speciation and extinction rates are adequate models for global-scale bacterial diversification dynamics over geological time, as discussed in the main text, is actually conservative.

Comparing tree topologies. To quantify the variation in tree topologies obtained using the different tree construction methods, and to compare that variation with previously published trees, we proceeded as follows. In all cases, trees were either pruned to only include bacteria (excluding chloroplasts and mitochondria) or cyanobacteria (excluding chloroplasts), as appropriate. A 16S rRNA gene-based and manually curated guide tree (release 128, non-redundant set) was downloaded from the SILVA database[33], and a tree of amplicon sequences previously published by the EMP (release 1, 'deblurred; 150-bp sequences)[32] was obtained from ftp://ftp.microbio.me/emp/release1/otu_info/deblur/emp150.5000_1000_rxbl_placement_pruned75.tog.tre. Each pair of trees (among our trees, the SILVA guide tree and the EMP tree) was compared after pruning trees to the set of tips shared by both trees, and using the Robinson–Foulds metric[72]. The Robinson–Foulds metric is widely used for comparing tree topologies, based on the number of tip clusters (sets of tips descending from internal nodes) that are unique to each tree. Robinson–Foulds distances were calculated using the R package castor version 1.3.3 (ref. [69]) (option: normalized=TRUE). To match tips across trees, all tips were renamed to SILVA sequence accessions whenever possible. Tips in the EMP tree, as well as tips in our de novo trees, were mapped to SILVA via global alignment at a similarity threshold of 99.5% using vsearch (options: --iddef 2 --strand both --maxhits 1 --maxaccepts 1)[51]; any unmapped tips were omitted from the comparisons. In cases where multiple tips matched the same SILVA entry, the nearest match was kept. The number of tips considered in each tree comparison, and calculated pairwise tree distances, are listed in Supplementary Table 7. To visualize relative distances between trees, we used multidimensional scaling ordination plots[73], in which each tree is represented by a single point, and where the distance between any two points approximately corresponds to the original Robinson–Foulds distance (Supplementary Fig. 24). As can be seen in Supplementary Fig. 24, our trees fall within the typical range of variation of trees of this magnitude.

Sensitivity analysis with respect to dating. To assess the sensitivity of our rate estimates to uncertainties in tree dating, we analysed variants of our timetrees created by randomly varying dating constraints. Specifically, we created random variants of our bacterial and cyanobacterial PATHd8-dated timetrees for which we had originally used the primary dating anchors listed in Supplementary Table 5 (bacterial '16S SILVA, FastTree, PATHd8', bacterial '16S de novo, FastTree, PATHd8' and cyanobacterial '16S de novo, FastTree, PATHd8'). For each random variant, we chose ages randomly and independently within the original age intervals of the dating anchor according to a triangular distribution, whose mode was set to the anchor node's calibrated age in the original timetree. These randomly drawn ages (one per anchor node) were then used as fixed dating constraints for dating the molecular phylogeny anew with PATHd8, thus obtaining a random variant. For each original timetree, we created ten random variants (see Supplementary Fig. 23 for example LTT calculated from these variants). For each timetree, we used its random variants to estimate speciation, extinction and diversification rates using the same methods as for the original tree. This yielded, for each timetree, a set of ten slightly different rate estimates (Supplementary Fig. 25), whose spread can be seen as a measure for estimation uncertainty due to errors in tree dating.

Fitting models and estimating speciation, extinction and diversification rates. Parameterized speciation and extinction models were fitted to the LTT of each considered timetree using the general approach described in Supplementary Information section 1.2. Models were fitted for three purposes: (1) to estimate recent speciation rates $\lambda(\tau=0)$; (2) to assess whether a constant speciation and extinction rate provide an adequate description of the observed LTT within some sufficiently small age interval; and (3) to predict past total diversity, $N(\tau)$, using the fitted model within the considered age interval. Models were integrated backwards in time using the differential equation listed in Supplementary Information section 1.2 and with the initial condition $N(0) = \tilde{N}(0)/\rho$, where ρ is the

previously estimated sampling fraction (fraction of extant OTUs or species included in the tree) and $\tilde{N}(0)$ is the number of tips in the tree. Model parameters were fitted by minimizing the MRD of the model's predicted LTT ($\tilde{N}_m$) from the real LTT:

$$\mathrm{MRD} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{N}(\tau_i)}\,|\tilde{N}_m(\tau_i) - \tilde{N}(\tau)| \quad (5)$$

where $\tau_1, \ldots, \tau_n$ are discrete ages at which the model's predicted LTT is compared with the real LTT. This fitting objective, which is based on relative rather than absolute errors, was chosen so as to increase the importance of earlier time points in the tree, where the LTT can be orders of magnitude lower than at the tips. The τ_i were chosen on a regular grid, comprising 100–200 points (depending on the size of the tree) and spanning the minimum and maximum ages considered for the tree. We avoided the most recent part of the LTT, where incomplete and phylogenetically biased taxon sampling can lead to deviations from the assumption of uncorrelated speciations, extinctions and discoveries, and where the choice of the OTU similarity threshold strongly affects branching frequencies (and thus the slope of the LTT). The last few time points (up to 20 Myr; overview in Supplementary Table 1) were therefore ignored during model fitting, in accordance with common practice[74]. For bacteria, the omitted age interval corresponds to ~1–2% divergence in the 16S rRNA gene[75,76], and hence to one or two expected branching events in the timetrees. Age intervals considered for fitting are listed in Supplementary Table 1. We fitted each model by minimizing the MRD using the optimization function stats::nlminb in R. We repeated the optimization 1,000 times with random start values to avoid non-global local optima. This complete approach has been implemented in the R package castor[69]. All models assumed constant speciation and extinction rates, λ and μ, which were fitted as described above. All fitted bacterial and cyanobacterial models achieved very good agreement with the observed LTT (MRD below 5% in all cases; overview in Supplementary Table 1).

Furthermore, to validate model assumptions and gain additional insight into past diversification dynamics, we estimated PERs (μ_p), PDRs (r_p) and PTDs (N_p) using the non-parametric methods described in Supplementary Information section 1.3. Before any non-parametric estimation of pulled variables (which involves derivatives of the LTT), the LTT was noise-filtered (smoothened) using a quadratic Savitzky–Golay filter. A noise filter is essential before estimating derivatives from the LTT, because finite-difference derivative estimators tend to amplify high-frequency noise in time series. Estimated PTDs, PERs and PDRs were also smoothened using local polynomial regression fitting (LOESS) to reduce noise, using the R function msir::loess.sd (span 0.2, degree 2)[77]. Standard errors of the smoothened estimates were calculated from the confidence intervals provided by loess.sd. As discussed in the main text, we found that μ_p was almost constant over time for all prokaryotic trees examined (Fig. 2d–f), supporting the assumption of a roughly constant (or only slowly varying) λ and μ made in the models (see discussion in Supplementary Information section 4).

Here, we have restricted our analyses to the most recent 1 Gyr because, for older time points, the smaller number of lineages in the tree, and thus the greater stochasticity and deviation from the continuum limit, lead to increased uncertainties in the estimated diversification dynamics (Supplementary Fig. 23). While future larger phylogenies will allow more accurate reconstruction of diversification dynamics for even more ancient times, we generally advise against using our non-parametric methods near the origin of a tree or clade of interest.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The R script for performing the diversification analyses on the timetrees, as well as the simulations discussed in Supplementary Information sections 2 and 4, is included as Supplementary File 3. The non-parametric methods introduced in the manuscript are implemented in the R package castor—a package for efficient phylogenetic analyses on very large trees[69] available on The Comprehensive R Archive Network (CRAN).

Data availability. Amplicon sequencing data used to recover de novo OTUs are publicly available under the accession numbers listed in Supplementary File 1. Accession numbers for sequencing data used from the EMP[32] are listed in Supplementary File 2. R code used in this study is provided as Supplementary File 3. Timetrees and undated phylogenetic trees constructed in this study are provided as Supplementary File 4. Taxonomic classifications of de novo OTUs are provided as Supplementary File 5.

Received: 1 February 2018; Accepted: 28 June 2018;
Published online: 30 July 2018

References

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
2. Fischer, W. W., Hemp, J. & Johnson, J. E. Evolution of oxygenic photosynthesis. *Annu. Rev. Earth Planet. Sci.* **44**, 647–683 (2016).
3. Raup, D. M. & Sepkoski, J. J. Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
4. Signor, P. W. Biodiversity in geological time. *Am. Zool.* **34**, 23–32 (1994).
5. McElwain, J. C. & Punyasena, S. W. Mass extinction events and the plant fossil record. *Trends Ecol. Evol.* **22**, 548–557 (2007).
6. Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B* **344**, 305–311 (1994).
7. Nee, S., Holmes, E. C., May, R. M. & Harvey, P. H. Extinction rates can be estimated from molecular phylogenies. *Phil. Trans. R. Soc. Lond. B* **344**, 77–82 (1994).
8. Sanderson, M. J. & Donoghue, M. J. Reconstructing shifts in diversification rates on phylogenetic trees. *Trends Ecol. Evol.* **11**, 15–20 (1996).
9. Morlon, H. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
10. Morlon, H., Kemps, B. D., Plotkin, J. B. & Brisson, D. Explosive radiation of a bacterial species group. *Evolution* **66**, 2577–2586 (2012).
11. Lorén, J. G., Farfán, M. & Fusté, M. C. Molecular phylogenetics and temporal diversification in the genus *Aeromonas* based on the sequences of five housekeeping genes. *PLoS. ONE* **9**, 1–15 (2014).
12. Lebreton, F. et al. Tracing the Enterococci from paleozoic origins to the hospital. *Cell* **169**, 849–861 (2017).
13. Gubry-Rangin, C. et al. Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. *Proc. Natl Acad. Sci. USA* **112**, 9370–9375 (2015).
14. Marin, J., Battistuzzi, F. U., Brown, A. C. & Hedges, S. B. The timetree of prokaryotes: new insights into their evolution and speciation. *Mol. Biol. Evol.* **34**, 437–446 (2017).
15. Stadler, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261**, 58–66 (2009).
16. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
17. Schopf, J. W. Disparate rates, differing fates: tempo and mode of evolution changed from the Precambrian to the Phanerozoic. *Proc. Natl Acad. Sci. USA* **91**, 6735–6742 (1994).
18. Dykhuizen, D. E. Santa rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**, 25–33 (1998).
19. Butterfield, N. Macroevolution and macroecology through deep time. *Palaeontology* **50**, 41–55 (2007).
20. Schopf, J. W. et al. Sulfur-cycling fossil bacteria from the 1.8-Ga duck creek formation provide promising evidence of evolution's null hypothesis. *Proc. Natl Acad. Sci. USA* **112**, 2087–2092 (2015).
21. Weinbauer, M. G. & Rassoulzadegan, F. Extinction of microbes: evidence and potential consequences. *Endang. Species Res.* **3**, 205–215 (2007).
22. Höhna, S., Stadler, T., Ronquist, F. & Britton, T. Inferring speciation and extinction rates under different sampling schemes. *Mol. Biol. Evol.* **28**, 2577–2589 (2011).
23. Stackebrandt, E. & Ebers, J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today* **33**, 152–155 (2006).
24. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
25. Sanmartn, I. & Meseguer, A. S. Extinction in phylogenetics and biogeography: from timetrees to patterns of biotic assemblage. *Front. Genet.* **7**, 35 (2016).
26. Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl Acad. Sci. USA* **108**, 6187–6192 (2011).
27. Silvestro, D., Schnitzler, J. & Zizka, G. A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evol. Biol.* **11**, 311 (2011).
28. Stadler, T. How can we improve accuracy of macroevolutionary rate estimates? *Syst. Biol.* **62**, 321–329 (2013).
29. Marshall, C. R. Five palaeobiological laws needed to understand the evolution of the living biota. *Nat. Ecol. Evol.* **1**, 165 (2017).
30. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the archaeal and bacterial census: an update. *mBio* **7**, e00201-16 (2016).
31. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).
32. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
33. Glöckner, F. O. et al. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J. Biotechnol.* **261**, 169–176 (2017).
34. Krebs, C. J. *Ecological Methodology* (Benjamin Cummings, San Francisco, CA, 1999).
35. Zanne, A. E. et al. Three keys to the radiation of angiosperms into freezing environments. *Nature* **505**, 89–99 (2014).
36. Jetz, W. et al. Global distribution and conservation of evolutionary distinctness in birds. *Curr. Biol.* **24**, 919–930 (2014).
37. Welch, R. A. et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).

38. Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* **22**, 235–247 (2014).
39. Xie, S., Pancost, R. D., Yin, H., Wang, H. & Evershed, R. P. Two episodes of microbial change coupled with Permo/Triassic faunal mass extinction. *Nature* **434**, 494–497 (2005).
40. Gibbons, S. M. et al. Evidence for a persistent microbial seed bank throughout the global ocean. *Proc. Natl Acad. Sci. USA* **110**, 4651–4655 (2013).
41. Pimm, S. L. et al. The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752 (2014).
42. Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346–351 (2014).
43. Straub, T. J. & Zhaxybayeva, O. A null model for microbial diversification. *Proc. Natl Acad. Sci. USA* **114**, E5414–E5423 (2017).
44. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
45. Butterfield, N. J. Proterozoic photosynthesis—a critical review. *Palaeontology* **58**, 953–972 (2015).
46. Brocks, J. J. & Banfield, J. Unravelling ancient microbial history with community proteogenomics and lipid geochemistry. *Nat. Rev. Microbiol.* **7**, 601–609 (2009).
47. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
48. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
49. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
50. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
51. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
52. Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.* **13**, 656–668 (2012).
53. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
54. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
55. Rasmussen, B., Fletcher, I. R., Brocks, J. J. & Kilburn, M. R. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**, 1101–1104 (2008).
56. Shih, P. M., Hemp, J., Ward, L. M., Matzke, N. J. & Fischer, W. W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19–29 (2017).
57. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108**, 13624–13629 (2011).
58. Brocks, J. J. et al. Biomarker evidence for green and purple sulphur bacteria in a stratified palaeoproterozoic sea. *Nature* **437**, 866–870 (2005).
59. Walter, M., Buick, R. & Dunlop, J. Stromatolites 3,400–3,500 Myr old from the North Pole area, Western Australia. *Nature* **284**, 443–445 (1980).
60. Ryder, G., Koeberl, C. & Mojzsis, S. J. in *Origin of the Earth and Moon* (eds Canup, R. & Kevin Righter, K.) 475–492 (Univ. Arizona Press, Tucson, AZ, 2000).
61. Dykhuizen, D. Species numbers in bacteria. *Proc. Calif. Acad. Sci.* **56**, 62–71 (2005).
62. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).
63. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
64. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
65. Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S. & Bremer, K. Estimating divergence times in large phylogenetic trees. *Syst. Biol.* **56**, 741–752 (2007).
66. May, R. M. How many species inhabit the earth? *Sci. Am.* **267**, 42–49 (1992).
67. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. & Worm, B. How many species are there on Earth and in the ocean? *PLoS Biol.* **9**, e1001127 (2011).
68. Eastman, J. M., Harmon, L. J., Tank, D. C. & Paradis, E. Congruification: support for time scaling large phylogenetic trees. *Methods Ecol. Evol.* **4**, 688–691 (2013).
69. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2017).
70. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl Acad. Sci. USA* **107**, 5897–5902 (2010).
71. Kuo, C.-H. & Ochman, H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol. Direct* **4**, 35 (2009).
72. Day, W. H. E. Optimal algorithms for comparing trees with labeled leaves. *J. Classif.* **2**, 7–28 (1985).
73. Borg, I., Groenen, P. J. F. & Mair, P. *Applied Multidimensional Scaling* (Springer, Berlin, 2013).
74. Ricklefs, R. E. Estimating diversification rates from phylogenetic information. *Trends Ecol. Evol.* **22**, 601–610 (2007).
75. Ochman, H. & Wilson, A. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**, 74–86 (1987).
76. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc. R. Soc. Lond. B* **253**, 167–171 (1993).
77. Scrucca, L. Model-based SIR for dimension reduction. *Comput. Stat. Data Anal.* **55**, 3010–3026 (2011).

Acknowledgements

We thank D. H. Parks for providing the 16S rRNA sequences from MAGs[48]. S.L. was supported by an NSERC grant and a postdoctoral fellowship from the Biodiversity Research Centre, University of British Columbia. M.W.P., M.D. and L.W.P. were supported by NSERC Discovery Grants. P.M.S. was supported by The Branco Weiss Fellowship – Society in Science. W.W.F. acknowledges support from the Simons Collaboration on the Origins of Life and NASA Exobiology award number NNX16AJ57G.

Author contributions

S.L., L.W.P. and M.D. conceived the project. S.L. developed the mathematical methods, performed the diversification analyses and wrote the first draft of the manuscript. P.M.S. performed the molecular clock analyses of the BEAST trees, provided the cyanobacterial multigene tree and contributed to the development of the project ideas. All authors helped to interpret the results, advised on methodological improvements and contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0625-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Stilianos Louca

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection. All data analyzed were already available on public repositories.
Data analysis	Any custom software used is either included as Supplemental material, or freely available at the CRAN R package repository (https://cran.r-project.org). Any 3rd party software used is publicly available and cited in the Methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Amplicon sequencing data used to recover de novo OTUs are publicly available under the accession numbers listed in Supplementary File 1. Accession numbers for sequencing data used from the Earth Microbiome Project are listed in Supplementary File 2. R code used in this study is provided as Supplementary File 3. Timetrees

as well as undated phylogenetic trees constructed in this study are provided as Supplementary File 4. Taxonomic classifications of de novo OTUs are provided as Supplementary File 5.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We reconstruct diversification dynamics of bacteria using publicly available 16S rRNA gene sequence data and novel phylogenetic methods.
Research sample	All analyses were based on existing, publicly available DNA sequence data. Accession numbers are provided as Supplementary Material.
Sampling strategy	Amplicon sequencing data, used to generate our de-novo trees, were chosen so as to represent as wide of an environmental range as possible, under the constraint that they must cover at least 200 bp of the 16S V4 region. For the SILVA trees, all OTUs available within the appropriate taxon were used. For the BEAST-computed trees, trees were sub-sampled randomly to enable computational feasibility. Any sub-sampling has been accounted for in our calculations.
Data collection	No novel data was collected.
Timing and spatial scale	No novel data was collected.
Data exclusions	No relevant data was explicitly excluded.
Reproducibility	All raw sequencing data used are available at public repositories under accession numbers provided as Supplemental material. SILVA 16S sequences are publicly available at the SILVA project website (https://www.arb-silva.de). All trees created in this study, and computer code needed to analyze them, are provided as supplemental material. All other software used in our analyses are publicly available and cited in our Methods section.
Randomization	This study does not include experimental groups.
Blinding	This study does not include experimental groups.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging