

## Evolutionary origins and diversification of proteobacterial mutualists

Joel L. Sachs, Ryan G. Skophammer, Nidhanjali Bansal and Jason E. Stajich

*Proc. R. Soc. B* 2014 **281**, 20132146, published 27 November 2013

---

### Supplementary data

["Data Supplement"](#)

<http://rsob.royalsocietypublishing.org/content/suppl/2013/11/26/rsob.2013.2146.DC1.html>

### References

[This article cites 45 articles, 17 of which can be accessed free](#)

<http://rsob.royalsocietypublishing.org/content/281/1775/20132146.full.html#ref-list-1>

### Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](#)



CrossMark  
click for updates

## Research

**Cite this article:** Sachs JL, Skophammer RG, Bansal N, Stajich JE. 2014 Evolutionary origins and diversification of proteobacterial mutualists. *Proc. R. Soc. B* **281**: 20132146. <http://dx.doi.org/10.1098/rspb.2013.2146>

Received: 16 August 2013

Accepted: 29 October 2013

### Subject Areas:

evolution, microbiology, taxonomy and systematics

### Keywords:

diversification, evolutionary transition, mutualism, parasitism, symbiont, virulence

### Author for correspondence:

Joel L. Sachs  
e-mail: [joels@ucr.edu](mailto:joels@ucr.edu)

<sup>†</sup>These authors are co-first authors.

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rspb.2013.2146> or via <http://rsob.royalsocietypublishing.org>.

# Evolutionary origins and diversification of proteobacterial mutualists

Joel L. Sachs<sup>1,3,†</sup>, Ryan G. Skophammer<sup>1,†</sup>, Nidhanjali Bansal<sup>1</sup>  
and Jason E. Stajich<sup>2,3</sup>

<sup>1</sup>Department of Biology, <sup>2</sup>Department of Plant Pathology and Microbiology, and <sup>3</sup>Institute for Integrative Genome Biology, University of California, Riverside, CA 92521, USA

Mutualistic bacteria infect most eukaryotic species in nearly every biome. Nonetheless, two dilemmas remain unresolved about bacterial–eukaryote mutualisms: how do mutualist phenotypes originate in bacterial lineages and to what degree do mutualist traits drive or hinder bacterial diversification? Here, we reconstructed the phylogeny of the hyperdiverse phylum Proteobacteria to investigate the origins and evolutionary diversification of mutualistic bacterial phenotypes. Our ancestral state reconstructions (ASRs) inferred a range of 34–39 independent origins of mutualist phenotypes in Proteobacteria, revealing the surprising frequency with which host-beneficial traits have evolved in this phylum. We found proteobacterial mutualists to be more often derived from parasitic than from free-living ancestors, consistent with the untested paradigm that bacterial mutualists most often evolve from pathogens. Strikingly, we inferred that mutualists exhibit a negative net diversification rate (speciation minus extinction), which suggests that mutualism evolves primarily via transitions from other states rather than diversification within mutualist taxa. Moreover, our ASRs infer that proteobacterial mutualist lineages exhibit a paucity of reversals to parasitism or to free-living status. This evolutionary conservatism of mutualism is contrary to long-standing theory, which predicts that selection should often favour mutants in microbial mutualist populations that exploit or abandon more slowly evolving eukaryotic hosts.

## 1. Introduction

An astonishing diversity of beneficial bacteria infect eukaryotes [1–3], but little is known about how these mutualists evolve [2]. Perhaps the biggest dilemma is to explain how mutualist phenotypes originate in bacterial lineages, and in particular whether mutualists evolve primarily from parasitic or free-living ancestors [1,4]. A classic paradigm from studies of pathogen virulence posits that beneficial bacteria evolve recurrently from parasitic ancestors [4–7]. Yet, these models predict that virulence is attenuated by vertical transmission among hosts [4,5], whereas biologists now believe that most beneficial bacteria are transmitted infectiously [1]. In contrast to virulence theory, comparative genomic analyses have suggested that differential patterns of gene loss in mutualists and parasites should hinder transitions between these states, leading researchers to propose that bacterial mutualism and parasitism most often represent independent origins of host association [8,9].

There is also intense debate about the evolutionary stability of mutualism [1,5,10–15], and in particular the degree to which mutualist phenotypes drive or hinder lineage diversification [2,16–19]. Early population-based models predicted that mutualist taxa are more vulnerable to extinction than other lifestyles [17,18] but few empirical data have supported these ideas [19]. In parallel, models of interspecific cooperation predict that bacteria—which often evolve rapidly—generate mutants that exploit or abandon their more slowly evolving eukaryotic hosts [14,15], thus predicting that mutualists commonly transition into other lifestyles [1,10]. Converse to these ideas, new theoretical frameworks argue that eukaryotic hosts must evolve mechanisms to control the interaction before stable mutualism can emerge [11–13], and thus that the mutualist lifestyle engenders evolutionary stability [2,13,16]. Despite the incredible prevalence and importance of mutualistic

bacteria in all aspects of eukaryote biology, the pathways to and the evolutionary diversification of such mutualisms are poorly understood [1,10].

Here, we investigated the evolution of mutualism in one of the most ecologically diverse and species-rich phyla of bacteria, the Proteobacteria. Proteobacteria encompass mutualists and pathogens of eukaryotes and diverse free-living species [20]. Phylogenetic relationships of 405 taxa (383 Proteobacteria and 22 outgroup taxa) were inferred using 47 protein markers mined from available whole-genome datasets [21,22]. We only sampled taxa with whole-genome sequences because this allows for robust phylogenetic reconstruction, and because these taxa are more often characterized with detailed phenotypic data. We used a data-gathering heuristic that incorporates ambiguous information to categorize host-association status for each of the bacterial taxa [23]. Proteobacterial taxa were categorized as free-living (no known association with eukaryotes) or host-associated (inhabitation of eukaryotes) and host-associated taxa were categorized as mutualist (neutral to beneficial to hosts), parasite (harmful) or dual lifestyle (evidence of mutualism and parasitism or ambiguous evidence). Host-associated taxa with no evidence for fitness effects on hosts (i.e. 'commensals') and dual-lifestyle taxa were independently analysed as mutualists and parasites to examine the effects of ambiguous categorization. We confirmed that each of our traits exhibit significant phylogenetic signal [24] and inferred ancestral host-association phenotypes using Markov chain Monte Carlo (MCMC; [25]), maximum likelihood (ML; [25]) and maximum parsimony (MP; [26]) on a posterior sample of Bayesian trees.

We used a multi-state speciation and extinction model (MuSSE) to infer trait-dependent extinction and speciation rates, as well as transition rates among states that take trait-dependent diversification rate into account [27]. MuSSE evaluates the potentially confounding effects of taxon sampling by inferring diversification and transition parameters under multiple schemes of extant taxon sampling. In order to test hypotheses about the drivers of mutualist origins, we compiled and analysed additional characteristics of mutualist taxa whenever possible, including information about habitat, host type, mode of transmission among hosts and mutualist services. Another recent study examined the evolution of mutualist traits across the Bacterial domain and found many origins of mutualism from both free-living and parasitic ancestors. But it did not include a quantitative analysis of transitions or their rates, nor did it provide robust validation of character states [1]. Our data here provide to our knowledge, the first quantitative analysis of the origins of proteobacterial mutualists. We demonstrated that proteobacterial mutualists are most often derived from parasitic ancestors. We uncovered negative diversification rates within mutualist taxa, which suggests that mutualism evolves primarily via transitions from other states. Contrary to the paradigm of mutualism instability, we found that mutualist taxa only rarely revert to parasitism or free-living status. Given that ancient evolutionary hypotheses are difficult to test empirically and that biases can be easily introduced into such studies, we provide extensive validation of our computational analyses.

## 2. Results

### (a) Phylogenetic reconstruction

Our phylogenetic reconstruction (figure 1; electronic supplementary material, S1 and S2) represents an extremely robust

proteobacterial tree, with more than 96% of the nodes on the consensus reconstruction having greater than or equal to 0.95 posterior support (see the electronic supplementary material, S2). The consensus Bayesian tree recovered monophyletic clades for each of the five proteobacterial classes, with the exception of *Acidithiobacillus ferrooxidans*-ATCC 23270 being placed outside the Gammaproteobacteria, as was previously found [28,29]. One clade on the tree is probably a result of phylogenetic artefact. The obligate intracellular symbionts in the Gammaproteobacteria—that have small A-T-rich genomes—are present on long branches that can cause them to incorrectly comprise a single lineage [28].

### (b) Host-association phenotypes and trait evolution

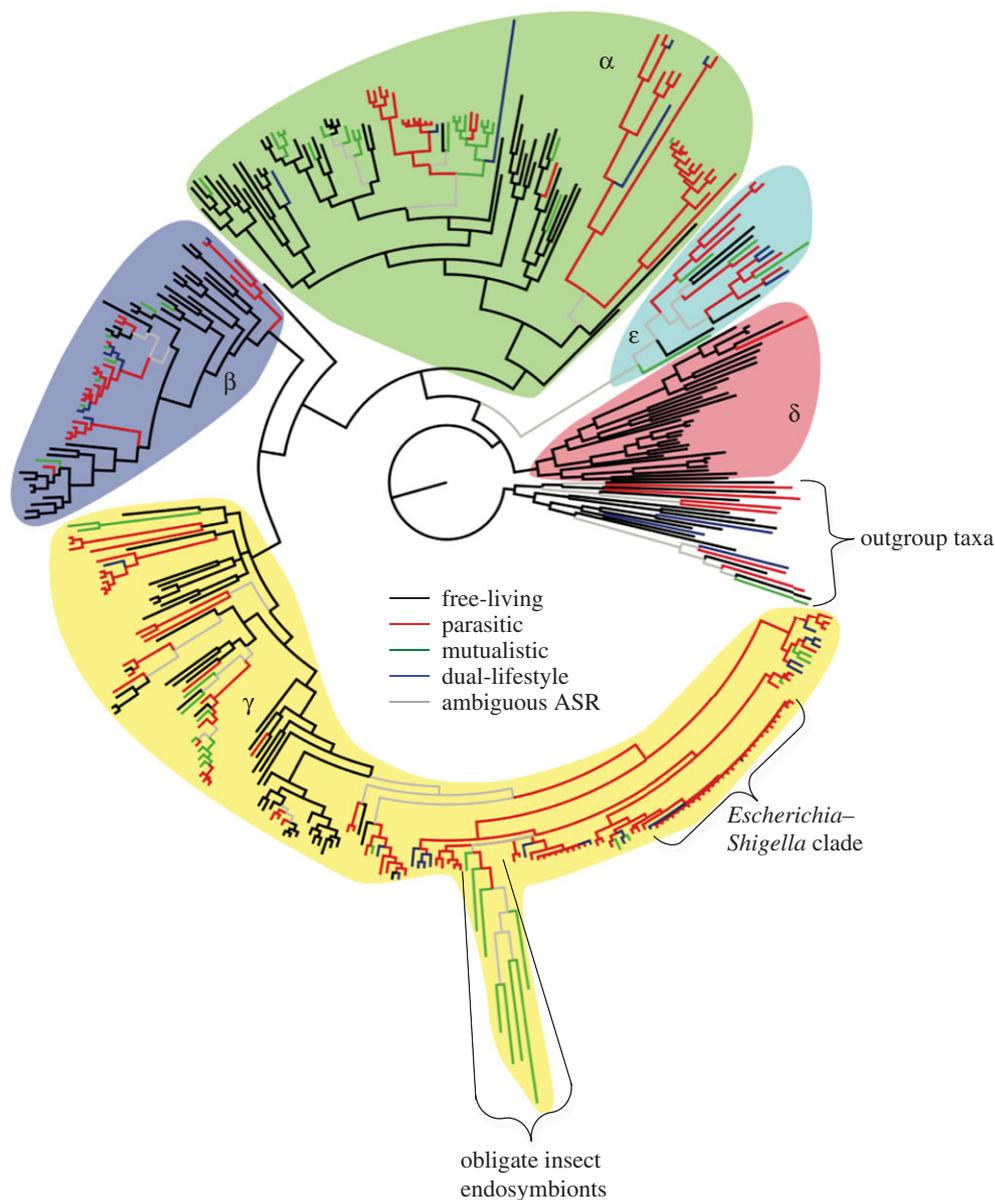
Among the 405 taxa, our information-gathering heuristic categorized 162 taxa as free-living, 62 as mutualists, 33 as dual lifestyle and 148 as parasites (see the electronic supplementary material, S3). In total, 43 taxa exhibited some ambiguity in their host-association status, because they were labelled as commensal or dual lifestyle. To examine the effect of ambiguity in the trait assignments, these 43 taxa were lumped into mutualists or parasites in independent analyses. We confirmed that host-association phenotypes exhibit significant phylogenetic signal—a prerequisite for ancestral state reconstruction (ASR)—by quantifying Pagel's lambda ( $\lambda$ ; [24]) for each character state classification. We found that the binary traits of host association, mutualism and parasitism all exhibit significant phylogenetic signal (ML estimates,  $\lambda > 0$ ,  $p < 0.05$ ; electronic supplementary material, S4).

We compared the fit of eight different evolutionary models of trait evolution using an ML approach [27,30]. Trait-dependent speciation and extinction rates, and transition rates among traits were either fixed to be equal or were allowed to have separate rates for each category. To minimize the number of parameter estimates, we focused only on transition types that we sought to test hypotheses about, including transitions between free-living status and mutualism ( $F \rightarrow M$ ,  $M \rightarrow F$ ), and between parasitism and mutualism ( $P \rightarrow M$ ,  $M \rightarrow P$ ). Using the Akaike information criterion (AIC), and pairwise  $\chi^2$  tests, we found no significant difference in fit between a model in which all parameters were allowed to vary and a model in which  $P \rightarrow M$  and  $M \rightarrow P$  transitions were constrained to be equal (see the electronic supplementary material, S5). Given that we wanted to explicitly test the hypothesis of a rate asymmetry between mutualism and parasitism, we chose to use the more complex model [31].

### (c) Evolutionary origins of proteobacterial mutualism

The deepest nodes of the tree were inferred to be free-living (see the electronic supplementary material, S2), consistent with the Proteobacteria anciently predating their eukaryotic hosts [1]. The most recent common ancestor (MRCA) of all Proteobacteria was decisively inferred (Bayes factor (BF)  $> 5$ ; [32]) to be free-living (BF = 6.550), as were the MRCAs of four of five proteobacterial classes (Alphaproteobacteria, BF = 7.505; Beta-, BF = 5.228; Gamma-, BF = 7.736; Delta-, BF = 11.240; Epsilon- is more ambiguous, BF = 1.131). Based on a consensus ASR, we inferred 38 origins of mutualism from free-living and parasitic ancestors (table 1).

Strikingly, proteobacterial mutualists were inferred to originate from parasitic ancestors almost twice as frequently as from free-living lineages (figure 2). This difference is significant



**Figure 1.** Inferred evolutionary history of host-association traits in Proteobacteria. Mutualist traits exhibit diverse and frequent origins in Proteobacteria from both parasitic and free-living ancestors. Branch colours represent host-associated traits on the tips of the tree and confidently inferred states on ancestral nodes. Inferred ancestral states were considered confident if both MCMC and ML analyses inferred the same state with BF scores greater than or equal to 2 (309 of 332 internal nodes). Two taxa are highlighted: the obligate insect endosymbionts (including the genera *Buchnera*, *Blochmannia*, *Hamiltonella*, *Riesia*, *Sodalis* and *Wigglesworthia*) and the *Escherichia-Shigella* clade. Both of these taxa were pruned from the tree in some analyses to test whether these densely sampled clades were biasing the results. A version with taxon labels is included as electronic supplementary material, S1.

when analysing variation in transition frequencies among posterior phylogenetic reconstructions, irrespective of statistical ASR framework (table 1). We also found the same patterns for differences in origination rates when using a MuSSE [27] that accounts for each character's influence on net diversification rate. MuSSE estimated that  $P \rightarrow M$  transitions occur more than 10 times as often as  $F \rightarrow M$  transitions when diversification rates are held constant (figure 3a; electronic supplementary material, S6). To corroborate these results, we also considered three-state models (F, M, P) in which dual lifestyle (MP) and commensal taxa were alternatively recategorized as mutualists or parasites. Irrespective of how the ambiguous taxa were categorized, we consistently inferred that  $P \rightarrow M$  transitions occurred more frequently (see the electronic supplementary material, S7) and at higher rates than  $F \rightarrow M$  transitions (see the electronic supplementary material, S6). Finally, given that uneven or incomplete taxon sampling can bias results about evolutionary transitions [27], we performed MuSSE under 16

alternate taxon-sampling schemes, varying both estimated taxon sampling (100%, 10%, 1% and 0.1%) and the inclusion of two densely sampled taxa; the *Escherichia-Shigella* clade and the obligate insect endosymbionts (figure 1). For instance, one concern is that  $P \rightarrow M$  transitions are common in these well-studied Gammaproteobacterial taxa, creating a bias. Yet, when estimated taxon sampling was adjusted or when each of these clades was individually pruned off the tree,  $P \rightarrow M$  transitions consistently occurred at higher rates than  $F \rightarrow M$  (see the electronic supplementary material, S6).

We inferred  $P \rightarrow M$  transitions to be most frequent in Gammaproteobacteria, including animal-associated bacteria with diverse mutualist services (18 transitions; electronic supplementary material, S1, S8 and S9). By contrast,  $F \rightarrow M$  transitions were common in Alpha- and Betaproteobacteria (six and four transitions, respectively; electronic supplementary material, S8) dominated by nitrogen-fixing plant symbionts (10 out of 12 descendent taxa fix nitrogen;

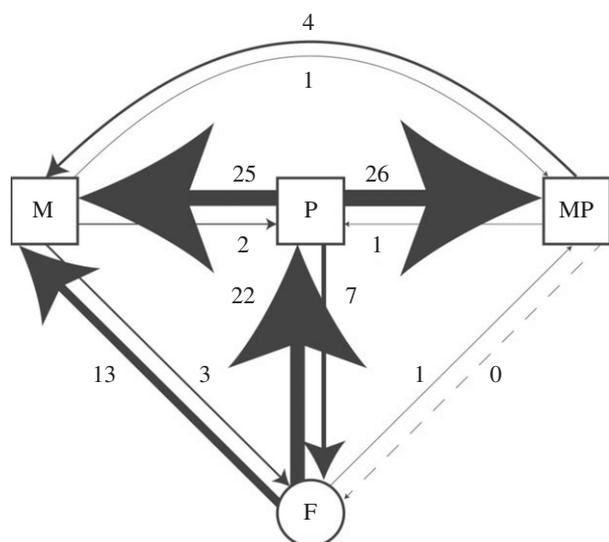
**Table 1.** Frequencies of host-association transitions estimated with different ASR frameworks. (Mean transition frequencies  $\pm$  standard deviation and minimum, maximum values are listed for the 723 posterior Bayesian trees (MCMC, ML and consensus) and for the 10 most parsimonious reconstructions of the Bayesian consensus tree (MP). For transition types, the ancestral and derived host-association phenotypes of each transition type are listed, respectively; host-associated phenotypes: F, free-living; M, mutualist; MP, dual lifestyle; P, parasite. The MCMC/ML consensus reconstruction optimizes the BF at each node using information from both the MCMC and ML ASRs.)

transition type	ASR frameworks			
	MCMC	ML	MCMC/ML consensus	MP
F $\rightarrow$ M <sup>a,b</sup>	14.00 $\pm$ 0.00 (14,14)	12.70 $\pm$ 0.46 (12,13)	13.00 $\pm$ 0.00 (13,13)	17 $\pm$ 0.00 (17,17)
F $\rightarrow$ P	23.00 $\pm$ 0.00 (23,23)	22.00 $\pm$ 0.00 (22,22)	22.00 $\pm$ 0.00 (22,22)	24 $\pm$ 0.00 (24,24)
F $\rightarrow$ MP	1.00 $\pm$ 0.00 (1,1)	1.00 $\pm$ 0.00 (1,1)	1.00 $\pm$ 0.00 (1,1)	1.00 $\pm$ 0.00 (1,1)
M $\rightarrow$ F	4.06 $\pm$ 0.24 (4,5)	4.00 $\pm$ 0.00 (4,4)	3.06 $\pm$ 0.24 (3,4)	3.00 $\pm$ 0.00 (3,3)
M $\rightarrow$ P	5.00 $\pm$ 0.00 (5,5)	2.00 $\pm$ 0.00 (2,2)	2.00 $\pm$ 0.00 (2,2)	6.3 $\pm$ 0.48 (6,7)
M $\rightarrow$ MP	1.00 $\pm$ 0.00 (1,1)	1.00 $\pm$ 0.00 (1,1)	1.00 $\pm$ 0.00 (1,1)	3.7 $\pm$ 0.48 (3,4)
P $\rightarrow$ F	6.94 $\pm$ 0.24 (6,7)	7.00 $\pm$ 0.00 (7,7)	6.94 $\pm$ 0.24 (6,7)	2.00 $\pm$ 0.00 (2,2)
P $\rightarrow$ M <sup>a,c</sup>	23.33 $\pm$ 1.00 (19,26)	25.52 $\pm$ 0.79 (25,28)	25.41 $\pm$ 0.89 (21,26)	17.7 $\pm$ 0.48 (17,18)
P $\rightarrow$ MP	26.09 $\pm$ 0.28 (26,27)	28.00 $\pm$ 0.00 (28,28)	26.01 $\pm$ 0.11 (26,27)	25.3 $\pm$ 0.48 (25,26)
MP $\rightarrow$ M	4.12 $\pm$ 0.42 (4,7)	1.00 $\pm$ 0.00 (1,1)	4.02 $\pm$ 0.22 (4,6)	0.00 $\pm$ 0.00 (0,0)
MP $\rightarrow$ P	1.06 $\pm$ 0.23 (1,2)	0 $\pm$ 0.00 (0,0)	1.00 $\pm$ 0.00 (1,1)	0.00 $\pm$ 0.00 (0,0)
MP $\rightarrow$ F	0 $\pm$ 0.00 (0,0)	0 $\pm$ 0.00 (0,0)	0 $\pm$ 0.00 (0,0)	0.00 $\pm$ 0.00 (0,0)

<sup>a</sup>P  $\rightarrow$  M is significantly greater than E  $\rightarrow$  M in both MCMC and ML ASRs when analysing variation among posterior trees using a two-tailed *t*-test (MCMC,  $t = 251.6$ ,  $p < 0.0001$ ; ML,  $t = 312.3$ ,  $p < 0.0001$ ) and is also significant among MPRs ( $t = 11.1$ ,  $p < 0.0001$ ).

<sup>b</sup>F  $\rightarrow$  M  $>$  M  $\rightarrow$  F (MCMC,  $t = 1093.9$ ,  $p < 0.0001$ ; ML,  $t = 510.1$ ,  $p < 0.0001$ ; MPRs,  $t = \infty$ ,  $p = 0.0$ ).

<sup>c</sup>P  $\rightarrow$  M  $>$  M  $\rightarrow$  P (MCMC,  $t = 494.3$ ,  $p < 0.0001$ ; ML,  $t = 804.0$ ,  $p < 0.0001$ ; MPRs,  $t = \infty$ ,  $p = 0.0$ ).



**Figure 2.** Path diagram of transitions in host-association phenotypes reveals frequent origins of mutualist phenotypes in Proteobacteria, but a paucity of reversals. Transitions among four proteobacterial host-association phenotypes are inferred on the pool of 723 posterior Bayesian trees. Transition frequencies are reported from the consensus ASR (figure 1; see Material and methods). F, free-living; M, mutualist; MP, dual lifestyle (mutualist and parasite, or ambiguous); P, parasite. We reconstructed no transitions from dual-lifestyle to free-living status. Arrow sizes are scaled to the frequency of transitions between host-association phenotypes.

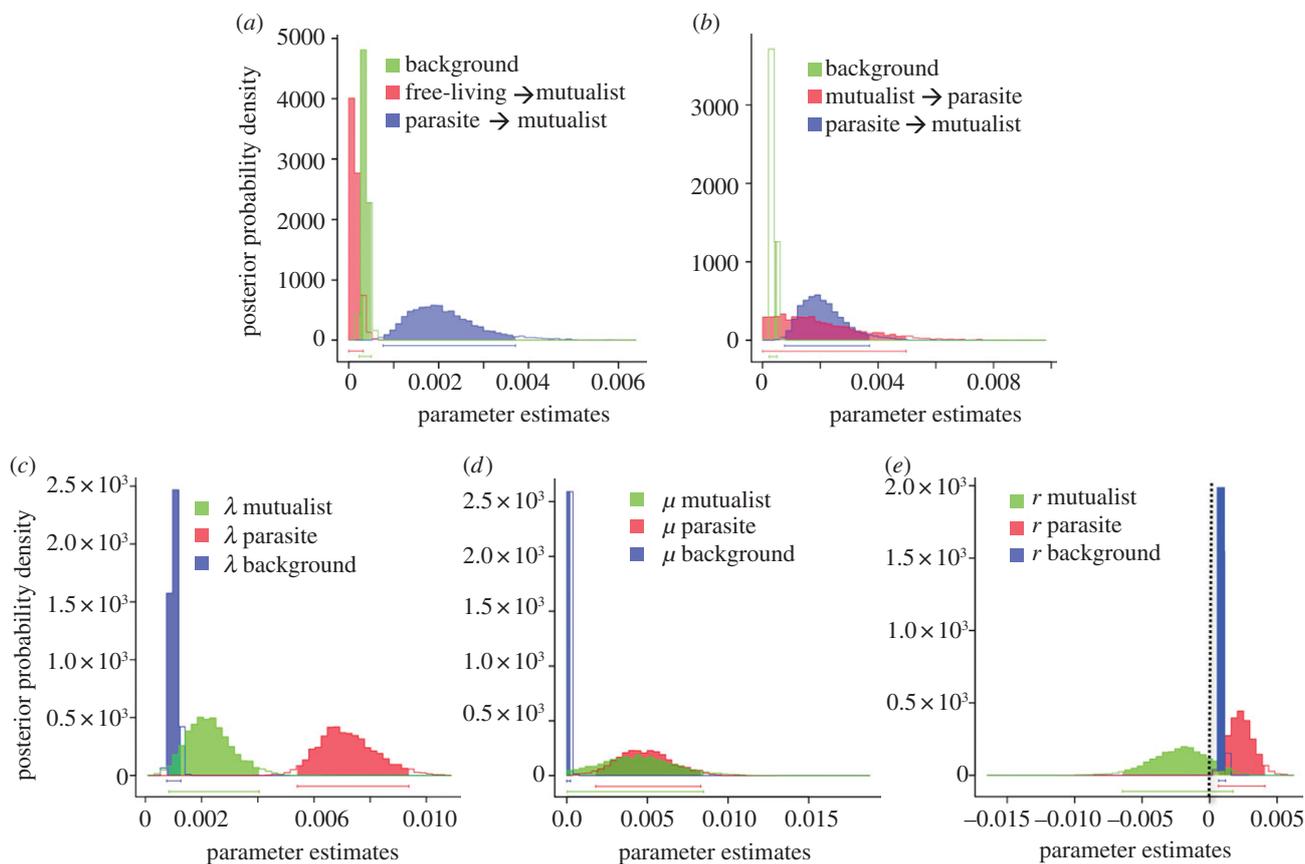
electronic supplementary material, S1 and S8). Theoreticians have predicted that bacterial mutualists only originate directly from free-living taxa if their ancestors carry traits that can provide immediate benefits to hosts that outweigh

the initial costs of infection [1,4]. Bacterial nitrogen fixation fits this prerequisite in most ecological settings. Moreover, nitrogen fixation traits in proteobacteria are often encoded on genomic islands or plasmids, consistent with the hypothesis that horizontal gene transfer (HGT) of host-beneficial traits represents a rapid route to the origins of novel mutualisms and potentially equally rapid loss [1].

Both P  $\rightarrow$  M and F  $\rightarrow$  M transitions occurred most frequently in lineages in which the descendent taxa were only transmitted horizontally among hosts. Among the 38 origins of mutualism on the consensus reconstruction (see the electronic supplementary material, S1), 29 of the transitions exhibited no evidence of vertical transmission in any of the inclusive taxa (see the electronic supplementary material, S8). These data reject the hypothesis that vertical transmission is a key prerequisite for transitions from parasitism to mutualism [4] and reveal a gap in theory to explain the evolutionary origins of mutualist bacterial traits.

#### (d) Evolution and diversification of proteobacterial mutualists

In our analysis of transition frequencies, we found that proteobacterial mutualist clades exhibit an extreme paucity of reversals to other states. In particular, origins of mutualism from parasitism occur 25 times on the ASR consensus tree, but only two reversals to parasitism are inferred (figure 2). Similarly, we inferred 13 origins of mutualism from free-living ancestors but only three reversals to free-living status. In both instances, the difference in origins versus reversals of mutualism was significant when analysing



**Figure 3.** MuSSE plots of transition rate and diversification parameters. Plots of the posterior probability density of the parameter estimates are shown for (a) the origins of mutualism ( $F \rightarrow M$  versus  $P \rightarrow M$ ), (b) transitions between parasitism and mutualism ( $P \rightarrow M$  versus  $M \rightarrow P$ ), as well as (c) speciation rate, (d) extinction rate and (e) diversification rate (speciation minus extinction). Transition rate parameters are estimated while keeping the diversification rate for each phenotype constant. Transition rate parameters are shown for 100% taxon sampling. For the diversification rate estimates, the dotted line represents zero net diversification (speciation equals extinction). The 95% credibility intervals for each parameter are shaded and indicated by bars along the  $x$ -axis.

variation in transition frequencies among posterior phylogenetic reconstructions (table 1). Moreover,  $P \rightarrow M$  and  $F \rightarrow M$  transitions were more frequent than reversals in all sampled posterior topologies regardless of ASR method. The two  $M \rightarrow P$  reversals are well supported in the consensus ASR (see the electronic supplementary material, S1) and appear to be driven by HGT events; the plant pathogen taxa *Pseudomonas syringae* and *Agrobacterium* spp. have probably evolved from mutualists via HGT of Type-III secretion systems and other key virulence loci [33,34]. By contrast,  $M \rightarrow F$  transitions occur on nodes with ambiguous ASRs (figure 1; electronic supplementary material, S1 and S2), so we cannot reject the null hypothesis that no such reversals occurred. To deal with uneven taxon sampling, two key densely sampled taxa were experimentally pruned from the tree in some analyses (*Escherichia* spp., insect endosymbionts; see Material and methods). Even when these clades are individually pruned off the tree,  $P \rightarrow M$  and  $F \rightarrow M$  transitions were still significantly more frequent than reversals, irrespective of ASR method (table 1). In three-state models ( $F, M, P$ ) in which dual lifestyle (MP) and commensal taxa were alternatively recategorized as M or P, both  $P \rightarrow M$  and  $F \rightarrow M$  transitions occurred more frequently than reversals irrespective of how ambiguous taxa were categorized (see the electronic supplementary material, S7).

Using MuSSE, we infer that mutualist taxa have a negative net diversification rate, i.e. an extinction rate greater than speciation rate (figure 3; electronic supplementary material,

S6). This pattern was consistent among almost all taxon-sampling schemes (e.g. whether or not the *Buchnera* and *Escherichia* clades were included and irrespective of estimated extant taxon sampling; 100%, 10%, 1% and 0.1%). Moreover, mutualist diversification rate was also inferred to be marginally lower than the diversification rates of either free-living or parasite taxa ( $M < F$ ,  $p = 0.0638$ ;  $M < P$ ,  $p = 0.0516$ , respectively). Yet, the latter comparisons were sensitive to taxon sampling and were significant in only approximately 60% of the sampling schemes (see the electronic supplementary material, S6), so these conclusions are preliminary. Finally, classic theory predicted that mutualists are particularly vulnerable to extinction [17,18], but MuSSE inferred that both mutualists and parasites exhibit similarly elevated extinction rates compared to free-living taxa (figure 3; electronic supplementary material, S6). We recognize the uncertainty in estimating extinction rates from phylogenies [30], so we treat this conclusion with some caution.

We also used MuSSE to examine transition rates between mutualists and parasites that control for the trait-specific differences in diversification rate. When we account for variation in trait-dependent diversification rate, we failed to find any significant asymmetry in  $M \rightarrow P$  versus  $P \rightarrow M$  (see the electronic supplementary material, S6) consistent with our initial model testing using AIC (see the electronic supplementary material, S5). Hence, even though we consistently found asymmetric transition frequencies between mutualism and

parasitism, the asymmetry can be accounted for by the low diversification rate of mutualists compared with parasitic proteobacterial lifestyles.

### 3. Discussion

#### (a) Validation of computational analyses

Hypotheses about ancient evolutionary events can be difficult to test empirically, hence we must use caution when drawing conclusions using phylogenetic inference. Our goal is to test hypotheses about the origins and diversification of mutualism and our results are both consistent and robust in this regard. As we describe below, our results are robust to the topological reconstructions of the tree, statistical ASR frameworks, phenotype classification protocols, character-coding schemes and the taxon sampling. Nonetheless, inferring the states of individual ancestral nodes on the tree must be treated with some caution, and we appreciate that some ancestral nodes on the tree do not always agree with dominant views of bacterial evolution. Consistent with this caution, two key densely sampled taxa were experimentally pruned from the tree in some analyses (*Escherichia* spp., insect endosymbionts), and tests showed that removal of these taxa had negligible effects on the overall results.

We vetted all key aspects of our inferential approach. For the tree reconstruction, we used a Bayesian framework in which hypothesis testing does not rely on any particular topology. But nonetheless, our consensus tree is extremely well supported and topologically matches phylogenies recently recovered by other investigators [20,28,29]. In terms of the ASR, we used three different statistical frameworks to reconstruct ancestral characters (MCMC, ML and MP). These independent approaches resulted in identical state reconstructions at more than 90% of nodes (see the electronic supplementary material, S2) and similar patterns of transition frequencies (table 1). For our categorization of host-association characters, our protocol placed ambiguous taxa into separate categories (commensal and dual lifestyle), which included species that have varying effects on hosts, context-specific effects or for which the effects on hosts are poorly understood. Yet, in some cases these taxa might be more accurately described as mutualists that are only opportunistically pathogenic (e.g. *Klebsiella pneumoniae*; [35]) or as parasites that are rarely avirulent (e.g. *Anaplasma centrale*; [36]). To deal with ambiguity, both dual-lifestyle taxa as well as all commensals were recategorized and all analyses were repeated. Importantly, the transition analysis remained relatively unchanged in these analyses (see the electronic supplementary material, S6 and S7). In terms of our character-coding scheme, the ML and MCMC analyses using binary coding (to estimate transition frequencies) produced qualitatively similar results relative to the multi-state coding in MuSSE ([27]; used to analyse transition rates; electronic supplementary material, S6). To deal with the potential biasing effects of incomplete taxon sampling, we ran all transition rate analyses under 16 different sampling schemes in which the estimated extant taxon sampling was varied (100%, 10%, 1% and 0.1%) and two densely sampled clades were pruned. Finally, although our data suggest that HGT can drive some important transitions among host-association lifestyles, these transfers do not undermine our inferences of ancestral states, because analysis of Pagel's lambda [24] inferred significant phylogenetic constraint in each trait.

#### (b) Diverse and numerous origins of proteobacterial mutualisms

Our reconstruction of the origins of proteobacterial mutualisms exposes a surprising ease with which these bacteria can evolve beneficial associations with eukaryote hosts. A range of 34–39 mutualist origins were inferred using a consensus ASR method, and mutualist associations were found to have evolved multiple times in all major clades of the Proteobacteria, except the Deltaproteobacteria. Proteobacterial mutualists were inferred to originate most frequently from parasites, in support of the classic paradigm of virulence theory [4–6]. A recent reconstruction of the domain Bacteria found mutualists to be more commonly derived from free-living taxa and uncovered only 9–10 origins of mutualism within the Proteobacteria [1]. Yet, taxon sampling was much less dense in that dataset and only parsimony was used to reconstruct ancestral characters. Nonetheless, both our dataset and the bacterial study suggest that evolutionary routes to mutualism can vary widely across different bacterial taxa.

Our transition frequency and rate data support the hypothesis that origins of mutualism from free-living ancestors are more difficult than transitions from parasitism to mutualism [4]. Ewald [4] made this prediction with the reasoning that origins of mutualism from free-living ancestors should require bacteria to simultaneously evolve to associate with and provide significant benefits to hosts. Mutualists inferred to descend from free-living ancestors frequently exhibited host-association traits gained via HGT (e.g. 12 out of 20 taxa exhibit nitrogen fixation; electronic supplementary material, S8), which is less common in mutualists descended from parasitic ancestors (8 out of 43 taxa). Gain of host-association genes through HGT is probably a common mechanism for mutualist origins [1], despite the paradigm that these transitions often exhibit patterns of net gene loss [8,9].

#### (c) The evolutionary diversification of proteobacterial mutualists

Classic mutualism models predict evolutionary instability of mutualist phenotypes, either because mutualists experience increased extinction risk [10,17–19] or because cheater mutants can invade mutualist populations and drive transitions from mutualism to parasitism [14,15]. Yet, our data are inconsistent with these models. Instead, our analyses suggest that mutualist traits engender evolutionary stasis, in the sense that mutualist taxa exhibit little evidence of adaptive diversification and show a paucity of transitions to other lifestyles. A previous analysis of transitions to and from mutualism in Bacteria found a qualitatively similar pattern, with more origins than reversals of mutualist traits [1]. The topological pattern on our tree suggests how the low net diversification rate has shaped mutualist taxa, which in contrast to the other phenotypes, are not inferred in any of the deeper ancestral nodes on the tree (see the electronic supplementary material, S1 and S2). Concomitant with this pattern and the negative diversification rate of mutualist taxa, our dataset suggests that mutualist phenotypes can only be maintained in the Proteobacteria by recurrent transitions from other host-association states.

Theoretical [11,12] and empirical research [1,13,16] is shaping a new paradigm for the evolution of bacterial mutualists [2]. Recent models have posited that eukaryotic–bacterial

mutualisms only originate when hosts exhibit mechanisms to prevent exploitation [11–15]. Mechanisms of host-control can be diverse and include ‘capture’ of bacterial mutualists via strict vertical transmission to offspring [1,4] or for horizontally acquired mutualists, through efficient manipulation of bacterial infection and proliferation within host tissues [11–15]. Consistent with this paradigm of stability and control by hosts, new molecular data often find evidence of overlapping fitness interests between bacterial mutualists and their hosts [2,13,16], meaning that there are few if any opportunities for these bacterial mutualists to exploit the host interaction. Our data support the paradigm of host control and suggest that once bacteria evolve to be mutualists, they are unlikely to undergo transitions into any other lifestyle.

## 4. Material and methods

### (a) Taxon selection and trait categorization

Proteobacterial taxa were chosen based upon sequence availability on the National Center for Biotechnology Information server (NCBI) as well as genome similarity. Multiple strains per species were sampled only if they differed in host-association phenotypes or content of type-III secretion system loci, which encode key host-associated functions [1,33]. Otherwise, single strains were chosen per species that exhibited maximal content of the 47 proteins used for phylogenetic reconstruction. If two or more strains were identical in phenotype, T3SS content and marker protein content, one was arbitrarily chosen for analysis. Outgroup taxa with completely sequenced genomes were sampled from 12 related eubacterial phyla (Actinobacteria, Bacteroidetes, Chlamydiae, Chlorobi, Cyanobacteria, Deinococcus-Thermus, Firmicutes, Planctomycetes, Spirochaetes, Tenericutes, Thermotogae and Verrucomicrobia; [20]). Within these phyla, taxa were chosen to maximize diversity in host-association phenotypes and genome content for the marker proteins. *Candidatus Hodgkinia cicadicola* DSEM and *Candidatus Carsonella ruddii* PV were pruned from our dataset to optimize phylogenetic support, because they nested incorrectly [28] within the outgroup clade in initial phylogenetic reconstructions (probably because of long branch effects [37,38]).

We used an information-gathering heuristic to collect data on host-association traits for the Proteobacteria and outgroup taxa. The heuristic, called a positive test strategy, searches for and automatically accepts affirmative information [23], in this case from the following set of trusted sources: primary literature that is indexed in the ‘Web of Knowledge’ ([www.webofknowledge.com](http://www.webofknowledge.com)), the DOE-JGI sequencing website (<http://genome.jgi-psf.org/programs/bacteria-archaea/index.jsf>), the NCBI-Entrez Genome Projects (<http://www.ncbi.nlm.nih.gov/genome>) and the High-quality Automated and from the Manual Annotation of Microbial Proteomes website (<http://hamap.expasy.org/>). Within each source database, the following search terms were used with each taxon or strain name to search for affirmative information about the taxon’s status as commensal (avirulent, commensal, epibiont, no effect), free-living (aquatic, free-living, environment, environmental isolate, soil), mutualistic (beneficial, complementary, fitness enhancing, growth promoting, mutualist, nutrient exchange, symbiont, symbiotic) or parasitic (causal, causative agent, deleterious, harmful, parasite, parasitic, pathogen, toxic, toxin, virulence, virulent). All references with any of these terms were read in full to manually curate all trait assignments.

The positive test heuristic is useful for host-association traits because it can incorporate mixed sources of information as well as ambiguous information. Taxa with evidence of free-living status but no evidence of commensalism, mutualism or parasitism were assigned as free-living. Taxa with any evidence of commensalism, mutualism or parasitism were assigned as host-

associated. Taxa with evidence only for mutualism or parasitism were assigned to these categories, and taxa with evidence of both mutualism and parasitism or both commensalism and parasitism were categorized as dual lifestyle (see the electronic supplementary material, S3). Taxa with evidence only of commensalism were initially categorized as mutualists, because we reasoned that it is easier to uncover harmful than beneficial effects upon hosts. However, to examine the effects of ambiguous trait categorizations, two alternate versions of a three-state categorization were also analysed (free-living, mutualist, parasite). In one version, commensal and dual-lifestyle taxa were categorized as mutualists and in the other they were categorized as parasites.

### (b) Phylogenetic reconstruction

Protein sequences for phylogenetic reconstruction were selected based on conservation and lack of horizontal transfer among taxa [21,22]. The following genes were selected: *dnaG*, *frz*, *gcp*, *infC*, *leuS*, *nusA*, *pgk*, *pheS*, *pyrG*, *rplA*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*, *rplL*, *rplM*, *rplN*, *rplO*, *rplP*, *rplR*, *rplS*, *rplT*, *rplV*, *rpmA*, *rpoA*, *rpoB*, *rpsB*, *rpsC*, *rpsD*, *rpsE*, *rpsG*, *rpsH*, *rpsI*, *rpsJ*, *rpsK*, *rpsL*, *rpsM*, *rpsO*, *rpsQ*, *rpsS*, *secY*, *serS*, *smpB*, *tsf* and *ychF*. Orthologues were identified through complementary methods. First, we searched for proteins annotated in the Kyoto Encyclopaedia of Genes and Genomes Orthology database (KEGG; <http://www.genome.jp/kegg/ko.html>). For organisms in our study that did not have entries in the entire KEGG database (e.g. multiple strains of the same species), we identified orthologues through searches of the NCBI Protein database. These searches were supplemented with reciprocal BLASTs, using proteins sequences from closely related organisms as the initial queries. This was intended to distinguish between multiple annotations of the same gene in an organism and also to confirm the orthologous relationship of the proteins.

Orthologous sequences were downloaded from the Batch-Entrez website and aligned using default settings on MUSCLE [39]. Alignments were concatenated using the BiOPERL script `concat_aln` [40] and were trimmed with the program TRIMAL [41] using the ‘strict’ setting, resulting in a concatenation of 7828 amino acids. Missing proteins were represented by gaps. MCMC phylogenies were reconstructed with MRBAYES v. 3.1.2 [42] using a fixed rate model of evolution [43] selected by an MCMC sampler that explored multiple models. Three MCMC runs of  $10^6$  generations each converged on a stationary distribution (average standard deviation of split frequencies less than 0.01). One tree out of 100 for each of the final 24 100 generations (postburn-in) was sampled in each run and a majority-rules consensus tree was generated from this pool of 723 trees.

### (c) Inference of ancestral state evolution

Both binary and multi-state phenotype-coding regimes were used to infer ancestral states. The binary coding scheme required three dichotomous classifications (1, free-living/host-associated; 2, parasite/non-parasite and; 3, mutualist/non-mutualist), whereas the multi-state coding included either four (e.g. free-living, dual lifestyle, mutualist and parasite) or three discrete states (free-living, mutualist and parasite). We used the binary coding for the ASRs described below, whereas the multi-state coding was used to estimate trait-dependent diversification and transition rates in MuSSE [27]. The ultrametric phylogenetic tree for MuSSE was generated by randomly choosing a full resolved tree from the sampled posterior distribution of 723 MRBAYES trees and rate smoothing with *r8s* [44].

Host-association phenotypes were confirmed to exhibit significant phylogenetic signal by quantifying Pagel’s  $\lambda$  [24] for each trait. We used the `fitDiscrete` function in the Geiger [45,46] package in R to calculate ML values of Pagel’s  $\lambda$  on 10 randomly selected postburn-in MCMC trees. Likelihood ratio tests were used to compare

the ML values of  $\lambda$  to models with a  $\lambda$  value of zero (no phylogenetic signal) and models with a  $\lambda$  value of one (shared trait values are proportional to genetic distance).

We used an ML based, model testing approach to examine whether host-association phenotypes had an impact on diversification rates as well as the rates of key transitions. We tested eight models in which speciation, extinction and transition rates were either constrained to be equal or were allowed to be independent among host-association phenotypes.  $\chi^2$ -values (ChiSq) and their significance (Pr-ChiSq) were calculated by performing pairwise comparisons between the fit of each model to the most complex model. The most complex model examined had independent speciation, extinction transition rates for all host-association phenotypes (20 parameters; electronic supplementary material, S5).

ASR was performed using MCMC, ML and MP. BAYESTRAITS [25] was used for MCMC and ML ASRs, which fits continuous-time Markov models to character data with discrete states and provides marginal likelihoods of two models for comparison. For the MP ASR, we used the 'trace character history' function of MESQUITE v. 2.74 [26] and used the 'most parsimonious reconstructions' option to generate a pool of 10 random MPRs. ASR was performed in a stepwise fashion; nodes reconstructed as free-living were not examined further, whereas host-associated nodes were labelled as parasites, mutualists or both (dual lifestyle). The MCMC analysis ran for  $10^6$  iterations, consisting of random samples from the pool of 723 trees; the ML analysis computed the likelihood for each tree separately and the mean marginal likelihood for the MCMC and ML analyses was computed. The likelihood for each inferred ancestral character state was evaluated at all 4719 nodes present in the 723 posterior MCMC trees, including 363 nodes present on the consensus tree. To compare the magnitude of the evidence for one state versus another at each node, the fossil command in BAYESTRAITS [25] was used to fix nodes at a particular state, one node-trait-state combination at a time. Each node was evaluated for both states of each trait, resulting in 28 314 values in both the MCMC and ML analyses.

A single ASR consensus hypothesis was created for display and discussion purposes (figures 1 and 2; electronic supplementary material, S1). On this ASR, character states and nodes were considered confident if both MCMC and ML analyses inferred the same state 'decisively' [32] ( $BF \geq 2$ ). Decisive evidence was found for 309 of 332 internal nodes in the consensus tree. The remaining nodes were considered ambiguous and a protocol was used to generate an optimal ASR hypothesis. If the states reconstructed by MCMC and ML differed, the test with the higher magnitude of difference was accepted. For nodes in

which MCMC and ML did not find positive evidence for parasitism or mutualism, we chose the state with the smaller negative evidence. If significant evidence rejecting one state but not the other existed for both analyses, then that reconstruction was chosen. Cases in which significant parasitic nodes also provided significant evidence for mutualism in one test but not significant evidence in the other were marked as significant for both states.

To test hypotheses about transition frequencies, two-tailed *t*-tests were used to compare different transitions types among the 723 posterior MCMC trees. To provide complementary information, MuSSE [27] was used to examine trait-dependent diversification and transition rates. Transition rates estimated with MuSSE hold diversification rates of each trait as constant. MuSSE was used to specifically test hypotheses about relative rates for the different origins of mutualism ( $P \rightarrow M$  versus  $F \rightarrow M$ ) and about bias transition asymmetry between mutualism and parasitism ( $P \rightarrow M$  versus  $M \rightarrow P$ ). A random MCMC tree from the postburn-in dataset was made ultrametric using penalized likelihood in APE 2.7 [47,48] with arbitrary scaling. For the MuSSE analyses, ancestral states were constructed using the multi-state coding. We calculated MuSSE model parameters from the ultrametric ML tree using ML in the software DIVERSITREE [27]. To be conservative, hypotheses about relative transition rates were tested under four schemes of estimated extant taxon sampling (100%, 10%, 1% and 0.1%). We also attempted to use the multi-state coding to reconstruct transition frequencies using Multi-state in BAYESTRAITS [25]. However, consistent with other researchers [49,50] we found that the parameter-rich multi-state models produced inconclusive results in BAYESTRAITS, so these results were not reported.

**Acknowledgements.** We are grateful to A. E. Arnold, K. Gano, J. Gatesy, L. Harmon, A. Hollowell, L. Nunney and M. Springer for helpful comments on the manuscript. R. Bantay, C. Fenster, J. Eastman, R. Meredith and M. Pennell offered assistance or advice on the analyses.

**Data accessibility.** Data archived in Dryad include (i) BAYESTRAITS input files and scripts associated with the analyses, (ii) DIVERSITREE input data, script files and search outputs, (iii) r8s smoothed input treefile and analysis parameters, (iv) MRBAYES input, and output files, including the consensus tree, and posterior sampled trees, and (v) trimmed gene alignment and MRBAYES consensus tree (doi:10.5061/dryad.6v06v). Data archived in TREEBASE include the trimmed gene alignment and the MRBAYES consensus tree (<http://purl.org/phylo/treebase/phylo/study/TB2:S14910>).

**Funding statement.** J.L.S. and R.G.S. were supported by grant no. 0816663 from the NSF. J.E.S. was supported by a grant from the Alfred P. Sloan Foundation.

## References

- Sachs JL, Skophammer RG, Regus JU. 2011 Evolutionary transitions in bacterial symbiosis. *Proc. Natl Acad. Sci. USA* **108**, 10 800–10 807. (doi:10.1073/pnas.1100304108)
- Sachs JL, Essenberg CJ, Turcotte MM. 2011 New paradigms for the evolution of beneficial infections. *Trends Ecol. Evol.* **26**, 202–209. (doi:10.1016/j.tree.2011.01.010)
- Huttenhower C *et al.* 2012 Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214. (doi:10.1038/nature11234)
- Ewald PW. 1987 Transmission modes and evolution of the parasitism–mutualism continuum. *Ann. NY Acad. Sci.* **503**, 295–306. (doi:10.1111/j.1749-6632.1987.tb40616.x)
- Yamamura N. 1993 Vertical transmission and evolution of mutualism from parasitism. *Theor. Pop. Biol.* **44**, 95–109. (doi:10.1006/tpbi.1993.1020)
- Dubos R. 1965 *Man adapting*. New Haven, CT: Yale University Press.
- Corsaro D, Venditti D, Padula M, Valassina M. 1999 Intracellular life. *Crit. Rev. Microbiol.* **25**, 39–79. (doi:10.1080/10408419991299167)
- Moran NA, Wernegreen JJ. 2000 Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol. Evol.* **15**, 321–326. (doi:10.1016/S0169-5347(00)01902-9)
- Gil R, Latorre A, Moya A. 2004 Bacterial endosymbionts of insects: insights from comparative genomics. *Environ. Microbiol.* **6**, 1109–1122. (doi:10.1111/j.1462-2920.2004.00691.x)
- Sachs JL, Simms EL. 2006 Pathways to mutualism breakdown. *Trends Ecol. Evol.* **21**, 585–592. (doi:10.1016/j.tree.2006.06.018)
- Weyl EG, Frederickson ME, Yu DW, Pierce NE. 2010 Economic contract theory tests models of mutualism. *Proc. Natl Acad. Sci. USA* **107**, 15 712–15 716. (doi:10.1073/pnas.1005294107)
- Archetti M *et al.* 2011 Economic game theory for mutualism and cooperation. *Ecol. Lett.* **14**, 1300–1312. (doi:10.1111/j.1461-0248.2011.01697.x)
- Friesen ML. 2012 Widespread fitness alignment in the legume–rhizobium symbiosis. *New Phytol.*

- 194, 1096–1111. (doi:10.1111/j.1469-8137.2012.04099.x)
14. Bull JJ, Rice WR. 1991 Distinguishing mechanisms for the evolution of cooperation. *J. Theor. Biol.* **149**, 63–74. (doi:10.1016/S0022-5193(05)80072-4)
  15. Sachs JL, Mueller UG, Wilcox TP, Bull JJ. 2004 The evolution of cooperation. *Q. Rev. Biol.* **79**, 135–160. (doi:10.1086/383541)
  16. Kimbrel JA *et al.* 2013 Extreme evolutionary stasis of type III effector genes in mutualistic *Sinorhizobium fredii* and *Bradyrhizobium japonicum*. *PLoS Pathog.* **9**, e1003204. (doi:10.1371/journal.ppat.1003204)
  17. May RM. 1976 Models for two interacting populations. In *Theoretical ecology: principles and applications* (ed. RM May), pp. 47–71. Philadelphia, PA: W. B. Saunders.
  18. Vandermeer JH, Boucher DH. 1978 Varieties of mutualistic interaction in population models. *J. Theor. Biol.* **74**, 549–558. (doi:10.1016/0022-5193(78)90241-2)
  19. Briand F, Yodzis P. 1982 The phylogenetic distribution of obligate mutualism: evidence of limiting similarity and global instability. *Oikos* **39**, 273–275. (doi:10.2307/3544496)
  20. Gupta RS. 2000 The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol. Rev.* **24**, 367–402. (doi:10.1111/j.1574-6976.2000.tb00547.x)
  21. Wu M, Eisen JA. 2008 A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, 10. (doi:10.1186/gb-2008-9-10-r151)
  22. Ciccarelli FD *et al.* 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287. (doi:10.1126/science.1123061)
  23. Klayman J, Ha YW. 1987 Confirmation, disconfirmation, and information in hypothesis-testing. *Psych. Rev.* **94**, 211–228. (doi:10.1037/0033-295X.94.2.211)
  24. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884. (doi:10.1038/44766)
  25. Pagel M, Meade A, Barker D. 2004 Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53**, 673–684. (doi:10.1080/10635150490522232)
  26. Maddison WP, Maddison DR. 2010 MESQUITE: a modular system for evolutionary analysis, v. 2.74. See: <http://mesquiteproject.org>.
  27. FitzJohn RG. 2012 DIVERSITREE: comparative phylogenetic analyses of diversification in R. *Meth. Ecol. Evol.* **3**, 1084–1092. (doi:10.1111/j.2041-210X.2012.00234.x)
  28. Williams KP *et al.* 2010 Phylogeny of Gammaproteobacteria. *J. Bacteriol.* **192**, 2305–2314. (doi:10.1128/JB.01480-09)
  29. Lang JM, Darling AE, Eisen JA. 2013 Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS ONE* **8**, e62510. (doi:10.1371/journal.pone.0062510)
  30. Rabosky DL. 2010 Extinction rates should not be estimated from molecular phylogenies. *Evolution* **64**, 1816–1824. (doi:10.1111/j.1558-5646.2009.00926.x)
  31. Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference: a practical information theoretic approach*. New York, NY: Springer.
  32. Kass RE, Raftery AE. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.1080/01621459.1995.10476572)
  33. Ma WB, Dong FFT, Stavrinos J, Guttman DS. 2006 Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genet.* **2**, 2131–2142. (doi:10.1371/journal.pgen.0020209)
  34. Slater SC *et al.* 2009 Genome sequences of three agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria. *J. Bacteriol.* **191**, 2501–2511. (doi:10.1128/JB.01779-08)
  35. Lau HY, Huffnagle GB, Moore TA. 2008 Host and microbiota factors that control *Klebsiella pneumoniae* mucosal colonization in mice. *Microbes Infection* **10**, 1283–1290. (doi:10.1016/j.micinf.2008.07.040)
  36. Kuttler KL. 1984 Anaplasma infections in wild and domestic ruminants: a review. *J. Wildl. Dis.* **20**, 12–20. (doi:10.7589/0090-3558-20.1.12)
  37. Felsenstein J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410. (doi:10.2307/2412923)
  38. Kueck P, Mayer C, Waegle JW, Misof B. 2012 Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS ONE* **7**, e36593. (doi:10.1371/journal.pone.0036593)
  39. Edgar RC. 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 1–19. (doi:10.1186/1471-2105-5-113)
  40. Stajich JE *et al.* 2002 The BioPERL toolkit: perl modules for the life sciences. *Genome Res.* **12**, 1611–1618. (doi:10.1101/gr.361602)
  41. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973. (doi:10.1093/bioinformatics/btp348)
  42. Ronquist F, Huelsenbeck JP. 2003 MrBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)
  43. Dayhoff M, Schwartz R, Orcutt B. 1978 A model of evolutionary change in proteins. I. In *Atlas of protein sequence and structure* (ed. M Dayhoff), pp. 345–352. Washington, DC: National Biomedical Research Foundation.
  44. Sanderson MJ. 2003 r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302. (doi:10.1093/bioinformatics/19.2.301)
  45. Pagel M. 1994 Detecting correlated evolution on phylogenies: a general-method for the comparative-analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37–45. (doi:10.1098/rspb.1994.0006)
  46. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008 GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131. (doi:10.1093/bioinformatics/btm538)
  47. Sanderson MJ. 2002 Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109. (doi:10.1093/oxfordjournals.molbev.a003974)
  48. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
  49. Hibbett DS. 2004 Trends in morphological evolution in homobasidiomycetes inferred using maximum likelihood: a comparison of binary and multistate approaches. *Syst. Biol.* **53**, 889–903. (doi:10.1080/10635150490522610)
  50. Nagy LG *et al.* 2010 The evolution of autodigestion in the mushroom family Psathyrellaceae (Agaricales) inferred from maximum likelihood and Bayesian methods. *Mol. Phylog. Evol.* **57**, 1037–1048. (doi:10.1016/j.ympev.2010.08.022)