Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data

Timothy G. Barraclough^{1,*}, Martin Hughes^{1,2},
Natalie Ashford-Hodges¹
and Tomochika Fujisawa^{1,2}

¹NERC Centre for Population Biology and Division of Biology,
Imperial College London, Silwood Park Campus, Ascot,
Berkshire SL5 7PY, UK

²Department of Entomology, Natural History Museum,
London SW7 5BD, UK

*Author for correspondence (t.barraclough@imperial.ac.uk).

By far the greatest challenge for diversity studies is to characterize the diversity of prokaryotes, which probably encompasses billions of species, most of which are unculturable. Recent advances in theory and analysis have focused on multi-locus approaches and on combined analysis of molecular and ecological data. However, broad environmental surveys of bacterial diversity still rely on single-locus data, notably 16S ribosomal DNA, and little other detailed information. Evolutionary methods of delimiting species from single-locus data alone need to consider population genetic and macroevolutionary theories for the expected levels of interspecific and intraspecific variation. We discuss the use of a recent evolutionary method, based on the theory of coalescence within independently evolving populations, compared with a traditional approach that uses a fixed threshold divergence to delimit species.

Keywords: bacteria; 16S rRNA; species; DNA barcodes; diversity

1. INTRODUCTION

Understanding the evolution of biological diversity remains a central goal for biologists. By far the greatest challenge lies in characterizing the diversity of prokaryotes. Molecular surveys have revealed a bewildering diversity of prokaryotes, perhaps many billions of species of which 99 per cent are unculturable (Acinas *et al.* 2004; Venter *et al.* 2004; Schloss & Handelsman 2006). Yet, prokaryotes harbour a wide diversity of biochemical processes and play a fundamental role in ecosystem function. Many problems of extreme importance to human society hinge on understanding prokaryotic diversity and how it will respond to change.

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsbl.2009.0091> or via <http://rsbl.royalsocietypublishing.org>.

One contribution of 11 to a Special Feature on 'Whole organism perspectives on understanding molecular evolution'.

Enormous progress has been made in prokaryotic systematics over recent years. After ongoing debate over the definition of bacterial species, there is now growing consensus on the range of processes that can cause divergence and the origin of independently evolving units of bacterial diversity. In general, geographical isolation and ecological divergence can cause independent evolution (Cohan 2001; Barraclough *et al.* 2003), leading to discrete units of genetic and phenotypic variation qualitatively similar to species in sexual eukaryotes (reviewed in Hanage *et al.* 2006; Achtman & Wagner 2008). Specific models have been developed to investigate how bacterial diversity patterns are shaped by niche specialization and periodic selection (likely to be especially important when effective population sizes are large and recombination rates are low; Koeppe *et al.* 2008) and by reproductive isolation (defined as limits to free recombination among separate groups, Fraser *et al.* 2007).

In addition, the increasing availability of molecular data has provided new opportunities for testing these ideas. Multi-locus data are now used widely to delimit bacterial species. Whole genome sequencing is becoming practical on the scale required for species delimitation, i.e. multiple samples from each of several putative species. However, these approaches remain unfeasible for environmental surveys of unculturable bacteria. Environmental shotgun sequencing projects have used shotgun assembly to infer whole genomes (Venter *et al.* 2004), but robust assembly of individual genomes down to the level of different closely related clones within populations is likely to be difficult from such data. Therefore, broad-scale studies of unculturable prokaryotic diversity still rely on single-locus data, notably the 16S ribosomal RNA (rRNA) gene. The ribosomal database project (RDP) online database (Cole *et al.* 2007) contained over 700 000 sequences in December 2008. Relying on a single locus has limitations: variation in 16S rRNA might not reflect variation at other loci, especially if there is extensive horizontal gene transfer, and 16S rRNA cannot be used to reveal functional diversity among isolates directly. However, 16S rRNA remains the common currency for studying broad-scale diversity patterns.

The traditional approach for estimating bacterial diversity has been to use a threshold of sequence divergence: individuals separated by more than 3 per cent pairwise distance in 16S rRNA were treated as separate species (Schloss & Handelsman 2006). Recent work has recommended 1 per cent as more appropriate for the species level (Acinas *et al.* 2004). However, there are two problems with assuming a universal threshold. First, substitution rates vary among lineages, meaning that *X* per cent in one group may not reflect the same period of divergence as *X* per cent in another group. Second, levels of variation within and between species are expected to vary widely among lineages, for example owing to differences in effective population size or speciation rates (Achtman & Wagner 2008). An alternative is to test directly for genetic clusters indicative of independent evolution. Acinas *et al.* (2004) used a graphical approach to demonstrate the presence of genetic clusters in bacterial 16S rDNA from the ocean and varied the level of threshold to infer an appropriate

Table 1. Estimated numbers of operational taxonomic units (OTUs; 3 and 1% thresholds) and ESUs (GMYC method) across five datasets. (Ranges in brackets indicate estimates within two log-likelihood units of the maximum-likelihood solution. All models were preferred over a null model of a single ESU (log-likelihood ratio tests, all $p < 0.001$). The datasets included all sequences for the following phyla: 1, Aquificae, Chlamydiae, Thermodesulfobacteria, BRC1; 2, Fusobacteria, Fibrobacteres, OP10, SR1; 3, Nitrospira, Deferribacteres, Dehalococcoides, OD1, Chrysiogenetes; 4, Deinococcus-Thermus, TM7, Lentisphaerae, OP11, Dictyoglomi; 5, Gemmatimonadetes, Thermotogae, Chlorobi, WS3, Thermomicrobia.)

dataset	no. sequences	estimated no. of OTUs		estimated no. of ESUs
		3% threshold	1% threshold	UPGMA + GMYC
1	1127	263	516	243 (239–253)
2	1064	330	541	288 (241–309)
3	1054	339	565	350 (342–360)
4	978	421	561	488 (464–507)
5	1028	434	637	541 (493–560)
total	5251	1787	2820	1910

cut-off for the delimitation of those clusters. However, statistical methods are needed that allow species limits to be optimized based on evolutionary predictions and that provide a framework for hypothesis testing and estimation of species richness.

Recent studies have developed evolutionary models for bacterial species delimitation, but they have either focused on multi-locus data (Fraser *et al.* 2007), required additional ecological data (Hunt *et al.* 2008), or are impractical for analysing thousands of sequences (Koeppel *et al.* 2008, electronic supplementary material). These limitations make them unsuited for the extreme case considered here: only single-locus data available for large taxonomic samples. Here, we apply a new method, previously applied only to a few eukaryote clades (Pons *et al.* 2006; Fontaneto *et al.* 2007), to the delimitation of 16S rDNA sequences from the RDP database. The method considers a general evolutionary model of the effects of independent evolution, whatever its causes, on the shape of a gene tree for a single locus. It requires no information except a gene tree, making it particularly suitable for 16S rDNA data from broad surveys. We use the model to test whether significant genetic clusters are present in 16S rDNA from the RDP database and, if so, whether estimates of numbers of evolutionarily significant units (ESUs) match those using threshold values of 3 or 1 per cent.

2. MATERIAL AND METHODS

(a) Sequence data and tree reconstruction

We downloaded the 16S rRNA sequences from 23 bacterial phyla chosen for illustrative purposes. Sequences were downloaded as the RDP alignment in five separate alignment files to speed up later analyses (table 1). Each matrix contained 1214 aligned bases. The default archaeal outgroup in RDP was used. Identical sequences were pruned to retain just one representative of each. Later analyses require ultrametric trees. For speed, we used the unweighted pair group method with arithmetic mean (UPGMA) on Kimura two-parameter distances to reconstruct the tree topology and branch lengths. More rigorous methods of phylogeny reconstruction and correcting for molecular rate variation might improve our estimates (electronic supplementary material), but UPGMA is sufficient for current purposes.

(b) Delimiting evolutionarily significant units

We used the generalized mixed Yule coalescent (GMYC) method of Pons *et al.* (2006). The null model is that the sample derives from a single interacting population: the pattern of variation

should conform to the genealogy of a single population, for example a neutral coalescent. The alternative is that the sample represents several independently evolving populations, i.e. selection and drift operate independently in different populations (Cohan 2001). In this case, coalescence occurs separately in different populations, leading over time to the appearance of discrete genetic clusters, separated from each other by longer internal branches (Barraclough *et al.* 2003; Acinas *et al.* 2004). Branching within clusters reflects population genetic processes, whereas branching between clusters reflects the timing of population divergence (Barraclough *et al.* 2003). Unlike other recent studies (e.g. Koeppel *et al.* 2008), we do not consider a specific model of bacterial speciation but rather a generic model of independent evolution: any process causing drift and selection to operate separately in different groups can generate distinct clusters detectable by our method (electronic supplementary material).

The method is implemented in an R package called GMYC (available at <http://r-forge.r-project.org/projects/splits/>) that compares the null and alternative hypotheses and infers the location and number of genetic clusters. We used the simplest version: a single threshold is optimized to delimit nodes defining the most recent common ancestors of species. Nodes younger than the threshold are assumed to be branching events within species. Note that this differs from the traditional threshold approach: it is applied on an ultrametric tree, therefore correcting for variation in substitution rates (albeit crudely here), and it optimizes the threshold for the given dataset rather than assuming a global cut-off value. We refer to the delimited entities as ESUs. For comparison, we used patristic distances on a neighbour-joining tree to delimit clusters using a 3 and 1 per cent furthest-neighbour rule (Schloss & Handelsman 2006). We refer to these entities as operational taxonomic units (OTUs).

3. RESULTS

Lineage-through-time (LTT) plots for the UPGMA trees (electronic supplementary material) show that a recent shift to a higher apparent branching rate has occurred near the tips (figure 1). This matches the predicted signature of independently evolving ESUs and leads to a sharp peak in the likelihood surface for the GMYC model (figure 1). Plotting the observed and fitted values for branching intervals between nodes confirms that the model provides a reasonable fit to the data (figure 2; electronic supplementary material). The GMYC method estimates 1910 ESUs across the five datasets (table 1). The 3 and 1 per cent threshold methods delimit 1787 and 2820 OTUs, respectively. In three datasets, the GMYC estimate falls between the 1 and 3 per cent threshold solutions, but closer to the 3 per cent solution. In datasets 1 and 2, the GMYC estimate is lower than the 3 per cent threshold solution.

Figure 1. Log LTT plots and likelihood surfaces for datasets (a) 1 and (b) 2. LTT plots show the number of lineages present at successive time points along the root-to-tip axis. The likelihood surface shows the likelihood of the GMYC model as the location of the threshold varies along the root-to-tip axis. The dashed line on the LTT plots shows the maximum-likelihood location for the threshold.

Figure 2. Observed branching intervals (grey circles) and fitted values under the maximum likelihood GMYC model (solid curve) for datasets (a) 1 and (b) 2. Branching interval is the time elapsed between successive branching events and position is the ranked order of nodes along the axis from the root to the tips. Dashed curves show the 95% confidence intervals for data simulated assuming the fitted GMYC model. Details are provided in the electronic supplementary material.

4. DISCUSSION

Our results demonstrate a new evolutionary method for delimiting units of bacterial diversity from single-locus DNA data. The approach is similar to previous

methods that varied the percentage of cut-off in pairwise similarity for grouping individuals into species (e.g. [Acinas *et al.* 2004](#)). Unsurprisingly then, we found a signature of genetic clustering within the data falling mostly between the 1 and 3 per cent thresholds previously advocated. However, the main advantage is that, by relying on a statistical model of branching rates, the method allows optimization, assignment of confidence limits and hypothesis testing. Also, by separating out phylogenetic reconstruction from the clustering algorithm, the approach allows the use of more rigorous phylogenetic methods to deal with complex evolution of the chosen molecule, should these be desired. Models that incorporate secondary structure of 16S rRNA or a relaxed-clock method of dating might yield improved phylogenetic reconstruction. A disadvantage in the present version is the use of a single threshold to delimit ESUs: versions that relax this assumption are in development.

One limitation affecting evolutionary models of bacterial diversity is the nature of the sample. The equations behind the GMYC method assume random sampling of individuals within any true species clusters that exist. In reality, the RDP database is far from random and often contains bacteria from a single physical sample, which may tend to include close relatives. For disease-causing bacteria, members of the same epidemic may be commonly sampled. A more intriguing possibility is that multiple levels of structure reflect a true feature of bacterial diversity, rather than an artefact of sampling ([Achtman & Wagner 2008](#); [Barraclough *et al.* 2008](#)). More complex evolutionary

methods allowing hierarchical levels of evolutionary independence might deal with these problems (electronic supplementary material).

The GMYC method is designed for cases in which only single-locus sequence data are available for delimiting species. Clearly, methods incorporating additional data, where available, will improve the detection of evolutionary species and insights into the processes generating diversity. With the growing number of alternative methods for delimiting bacterial species, an important task for future work will be the comparison of those methods in case-study groups in which detailed information is available. The correspondence of GMYC clusters to ecologically distinct entities could then be explored. The 16S rRNA molecule is rather conservative for surveying bacterial diversity: ecologically distinct strains that represent putative species can be indistinguishable by their 16S rRNA (Koeppel *et al.* 2008). However, our results show that 16S rRNA is significantly clustered and that gaps between clusters are greater than expected under an even branching model: sampling issues aside, the only explanation is that these clusters represent evolutionarily coherent entities that evolve independently of one another (electronic supplementary material).

We expect that future studies will sequence multiple genes or whole genomes of individuals at sample sizes equivalent to ours, for example by extracting DNA from individual cells isolated from the environment rather than from pooled samples. It will be possible then to apply more sophisticated methods. However, in the meantime, single-locus data remain the common currency for studies of environmental bacterial diversity. Evolutionary approaches can provide a deeper understanding of the nature and extent of bacterial diversity than possible using simple universal thresholds.

We thank Lindell Bromham and two anonymous referees for helpful suggestions on previous versions of the manuscript.

Achtman, M. & Wagner, M. 2008 Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microb.* **6**, 431–440. (doi:10.1038/nrmicro1872)

- Acinas, S. G., Klepac-Ceraj, V., Hunt, D. E., Pharino, C., Ceraj, I., Distel, D. L. & Polz, M. F. 2004 Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551–554. (doi:10.1038/nature02649)
- Barraclough, T. G., Birky Jr, C. W. & Burt, A. 2003 Diversification in sexual and asexual organisms. *Evolution* **57**, 2166–2172. (doi:10.1554/02-339)
- Barraclough, T. G., Fontaneto, D., Herniou, E. A. & Ricci, C. 2008 The evolutionary nature of diversification in sexuals and asexuals. In *Speciation and ecology* (eds R. Butlin, J. Bridle & D. Schluter), 29–45. Cambridge, UK: Cambridge University Press
- Cohan, F. M. 2001 Bacterial species and speciation. *Syst. Biol.* **50**, 513–524. (doi:10.1080/10635150118398)
- Cole, J. R. *et al.* 2007 The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. *Nucleic Acids Res.* **35**, D169–D172. (doi:10.1093/nar/gkl889)
- Fontaneto, D., Herniou, E. A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C. & Barraclough, T. G. 2007 Independently evolving species in asexual bdelloid rotifers. *PLoS Biol.* **5**, e87. (doi:10.1371/journal.pbi.0050087)
- Fraser, C., Hanage, W. P. & Spratt, B. G. 2007 Recombination and the nature of bacterial speciation. *Science* **315**, 476–480. (doi:10.1126/science.1127573)
- Hanage, W. P., Fraser, C. & Spratt, B. G. 2006 Sequences, sequence clusters and bacterial species. *Phil. Trans. R. Soc. B* **361**, 1917–1927. (doi:10.1098/rstb.2006.1917)
- Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J. & Polz, M. F. 2008 Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081–1085. (doi:10.1126/science.1157890)
- Koeppel, A. *et al.* 2008 Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl Acad. Sci. USA* **105**, 2504–2509. (doi:10.1073/pnas.0712205105)
- Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D. & Vogler, A. P. 2006 Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* **55**, 595–609. (doi:10.1080/10635150600852011)
- Schloss, P. D. & Handelsman, J. 2006 Toward a census of bacteria in soil. *PLoS Comput. Biol.* **2**, 786–793. (doi:10.1371/journal.pcbi.0020092)
- Venter, J. C. *et al.* 2004 Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74. (doi:10.1126/science.1093857)