Supplementary Notes

A. Previous methods for delimiting bacterial species from single locus data

1) The program DOTUR by Schloss & Handelsmann (2005) estimates species richness using % cut-offs and allows comparison of different % cut-offs. Plotting species richness estimated under varying thresholds can reveal a transition point after which the estimated richness increases at a higher rate. The same approach was used by Acinas et al. (2004) and as a summary of the observed sequence diversity pattern by Koeppel et al. (2007, see point 2). The transition point detected by these methods is essentially the same feature being detected by GMYC. However, GMYC is based on general theory for the cause of the transition point and allows its optimization without visual inspection of graphs (contrary to the approach of Acinas et al. 2004).

2) Koeppel et al. (2007) implement a simulation approach, called Ecosim, to model bacterial diversity under the framework of periodic selective sweeps developed by Cohan and colleagues. The model considers the effects of periodical selection, drift and ecotype formation on patterns of sequence diversity. The model is similar to the GMYC method in that it looks for clustering caused by independent evolution. Potential benefits of the approach are that parameters relate to a specific model of bacterial speciation: estimates might reveal something about the processes involved. Also, the simulation does not require all speciation events to be older than all within-species coalescence events, which is a simplifying assumption made by the current method of optimizing species limits using GMYC [Note that the GMYC model itself does not assume this, just the current method of optimizing species limits by sliding a threshold]. However, the key disadvantage is that the method is simulation-based and therefore much more cumbersome than GMYC. It takes at least a week for 200 sequences and an unspecified longer time for the 2000 sequences considered in each dataset here (http://fcohan.web.wesleyan.edu/ecosim/). Therefore, although it might prove a useful tool for case studies, it is impractical for the broad surveys being considered here.

(Another difference is that Ecosim assumes that mutations follow a Poisson process whereas GMYC assumes that phylogenetic methods have already been applied to control for stochastic variation due to the underlying substitution process. The latter provides more flexibility for complex substitution models and substitution rate calibrations to be applied).
3) Hunt et al. (2008) developed a method for identifying ecotypes based on the identification of clades associated with ecological measures: the size fraction of particles on which the bacteria grow (with free-living bacteria being sampled in the smallest size fraction) and seasonality. In principle, incorporation of ecological data could greatly enhance detection of species entities, compared to the use of arbitrary DNA markers. However, it assumes that relevant ecological attributes can be quantified: if not, independently evolving species might be detectable by the cluster-based methods that go undetected based on ecological association. Basic ecological data could be compiled for the 16S rRNA sequences studied here by referring back to source publications, but this would be extremely time-consuming.

B. Supplementary methods

*Phylogenetics.* We used UPGMA as the fastest method able to generate an ultrametric tree suitable for GMYC. UPGMA is known to introduce biases in topology, but this might not be a problem for present purposes as long as the structure of genetic divergence among sequences is preserved: our method relies primarily on the pattern of branch lengths rather than accurate topology. The opposite end of the methodological spectrum would be to implement a model of secondary structure of 16S rRNA and a relaxed molecular clock to correct for substitution rate variation.

*Further details on GMYC.* Koeppel et al. (2007) in their supplementary information stated that GMYC assumes that drift is the only processes causing coalescence. This is incorrect. The equations underlying the GMYC model were derived initially assuming that the genealogy within species conforms to a neutral coalescent with effective population size \( N_e \). However, the generalized method is also applicable to cases in which the marker or linked regions are under diversifying selection. Selective sweeps affecting the study marker will depress \( N_e \) and distort the genealogy of the population: a relative excess of branching occurs immediately after the selective sweep, generating a star-like genealogy. The ‘Generalized’ part of the GMYC introduced scaling parameters for branching rates in both the diversification and coalescent parts of the model. Therefore, if selective sweeps have occurred across species, the scaling parameter for the coalescent process will be optimized as less than 1, indicating a relative excess of deeper branching events within species. Hence, through the combined inference of lower \( N_e \) and scaling parameter less than 1, GMYC can model clusters originating by either drift or by divergent selection. It is true that GMYC deals with these
different processes anonymously rather than including specific parameters for their effects, but this does not mean that it assumes that those processes are not operating.

C. Supplementary results

Figure 2 plots the observed and fitted values for branching intervals, i.e. the times elapsed between successive nodes along the root to tip axis. Fitted values are the expected branching intervals under the fitted GMYC model, which equals the reciprocal of the exponential rate parameter $b^*$ in equation 6 of Pons et al. (2006). The 95% confidence intervals for the fitted model were simulated using 10000 replicates, each drawing branching intervals from an exponential distribution with rate parameter $b^*$. The main departure of the observed values from expected values is an excess of near zero branching intervals, especially near the tips. This likely reflects the discrete nature of genetic distances (the minimum non-zero distance is 1 transition out of 1214 base pairs) compared to the continuous time assumed by the model: the shortest intervals are either 1/1214 or zero, not values in between. Model fitting does not seem unduly influenced by these near zero distances: the fitted line encompasses the main block of data rather than being pulled to shorter internodes.

D. Supplementary Discussion

Significant clustering in 16S rRNA, as observed by Acinas et al. (2004) and in the present paper, is consistent with the existence of independently evolving entities at a level detectable by 16S rRNA. However, many studies have shown that 16S rRNA cannot always distinguish ecotypes of bacteria distinguishable by other means: the molecule has evolved too slowly to accumulate variation. How can we reconcile these observations? One possibility is that the majority of bacteria sampled by broad surveys do comprise ecotypes that are detectable using 16S rRNA: the counter-examples could represent a minority of cases. An intriguing alternative, which we prefer, is that multiple hierarchical levels of independent evolution are present: related ecotypes comprise an evolutionary guild of species that have diverged somewhat in ecological niche, but still share a common evolutionary fate at a higher-level of organization sufficient to generate clustering at that level. Barraclough et al. (2008) discuss how independent evolution might act at multiple levels to generate higher-level units of diversity. The important point here is that significant clustering in 16S rRNA needs to be explained, even if those clusters are not believed to represent all distinct ecotypes or ‘species’.
Deciding which level to assign the term ‘species’ is an important task for systematics, but characterising diversity patterns and understanding the evolutionary forces behind their origins can proceed even if assignment of the term ‘species’ to a particular level is difficult.

**Additional References**

**Supplementary Figures**

Figure S1
Log lineage-through-time (LTT) plots and likelihood surfaces for datasets 3 and 4. LTT plots show the number of lineages along successive time points from the root to the tips of the tree (units of genetic divergence shown on the bottom row of graphs). The likelihood surface shows the likelihood of the GMYC model when the threshold is placed at points along the root-to-tip axis. The red line on the LTT plots shows the maximum likelihood location for the threshold. UPGMA trees with clusters indicated are provided as Supplementary Figures.

Figure S2
As figure S1 but for dataset 5.

Figure S3. UPGMA tree for datasets 1 to 5 with clusters optimized by GMYC indicated. Black branches = between-cluster branching; red branches = within-cluster branching.