Accurate and universal delineation of prokaryotic species

Daniel R Mende[1], Shinichi Sunagawa[1], Georg Zeller[1] & Peer Bork[1,2]

The exponentially increasing number of sequenced genomes necessitates fast, accurate, universally applicable and automated approaches for the delineation of prokaryotic species. We developed specI (species identification tool; http://www.bork.embl.de/software/specI/), a method to group organisms into species clusters based on 40 universal, single-copy phylogenetic marker genes. Applied to 3,496 prokaryotic genomes, specI identified 1,753 species clusters. Of 314 discrepancies with a widely used taxonomic classification, >62% were resolved by literature support.

The definition of prokaryotic species is one of the most debated topics among microbiologists, without any satisfying solution[1]. Phenotypic information, pathogenicity and environmental observations had originally guided the delineation of prokaryotic species. These approaches have been complemented by molecular techniques, such as the use of DNA-DNA hybridization (DDH) and sequencing of the 16S ribosomal RNA (rRNA) gene as a phylogenetic marker[2]. Currently, information from different genomic and phenotypic methods is combined to assign an organism to a species[3]. An *ad hoc* committee on the systematics of prokaryotes proposed to define species as "a category that circumscribes a (preferably) genomically coherent group of individual isolates (strains) sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions"[1,2]. In practice, a prokaryotic species is often defined as a group of organisms with a certain phenotypic consistency, a DDH value of over 70%, and a 16S rRNA gene nucleotide sequence identity of 97% or more using reference 'type strains'[4], which are representative strains for prokaryotic species chosen by experts and are used as a taxonomic reference.

Whereas DDH is still considered the 'gold standard'[4] for species assignments, its application is impractical because of experimental complexity, low throughput and often irreproducible results. Use of the 16S rRNA gene as a phylogenetic marker has consequently gained popularity[5], despite the fact that the routinely applied 97% sequence identity cutoff was not intended for this use and may be inaccurate[6,7]. A more stringent threshold range of 98.7–99% sequence identity has been proposed as a remedy[7]. However, additional complexity arises from genomes harboring multiple copies of the 16S rRNA gene, which can exhibit intragenomic diversity of up to ~5% (for example, *Desulfitobacterium hafniense*, **Supplementary Table 1**) or when 16S rRNA genes from different species (for example, *Aeromonas salmonicida* and *Aeromonas hydrophila*) are ~99% identical (**Supplementary Table 2**).

More recently, the calculation of the pair-wise average nucleotide identity (ANI) of whole genomes[4,8] has been proposed to computationally obtain results that mimic those from DDH (with intraspecies similarity of approximately >94% ANI)[8]. Even though ANI is easier to use and faster than DDH, it is still computationally expensive and thus has not been applied at large scale to refine the current taxonomy of sequenced prokaryotic genomes.

As an alternative, protein-coding genes have been suggested as markers for species delineation and clade-specific, but not universal, marker genes have already been used for this purpose[2,3]. Additionally, for the related task of placing genomes into an existing taxonomic classification, such as the one provided by the US National Center for Biotechnology Information (NCBI Taxonomy database), several methods that use universal marker genes have also been developed[9,10]. These include an approach based on 40 universal, single-copy phylogenetic marker genes (pMGs)[11,12] (**Supplementary Table 3**), which had been first identified to resolve the evolutionary history of organisms by building a phylogenetic tree that spans all three domains of life[11]. However, a comprehensive analysis of global prokaryotic genomic diversity at the species level has not been provided by any of these approaches because of both efficiency concerns and a lack of focus on the species level.

Here we present the specI method for fast and accurate delineation of prokaryotic species based on these 40 protein-coding pMGs. We used a clustering approach based on pair-wise nucleotide sequence dissimilarity between pMGs, bypassing the computational bottlenecks of constructing multiple sequence alignments and accurate phylogenetic trees. The computational efficiency of this approach enabled the analysis of a comprehensive set of 3,496 prokaryotic genomes with accuracy equal to or higher than that of existing approaches.

In a set of 3,496 high-quality prokaryotic genomes (see Online Methods for quality-control criteria), including 836 sequenced type strains (**Supplementary Table 4**), we identified 40 universal single-copy pMGs[11,12] by eggNOG (evolutionary genealogy of genes: nonsupervised orthologous groups) annotation[13]. To benchmark the methodology, we selected 943 genomes from species with sequenced type strains, for each of which all 40 pMGs and at least one full-length 16S rRNA gene could be detected. Type strains are used as a taxonomic reference and hence should be used to assess the performance of species-delineation tools[14]

[1]European Molecular Biology Laboratory, Heidelberg, Germany. [2]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. Correspondence should be addressed to P.B. (bork@embl.de).

RECEIVED 10 MAY; ACCEPTED 24 JUNE; PUBLISHED ONLINE 28 JULY 2013; DOI:10.1038/NMETH.2575

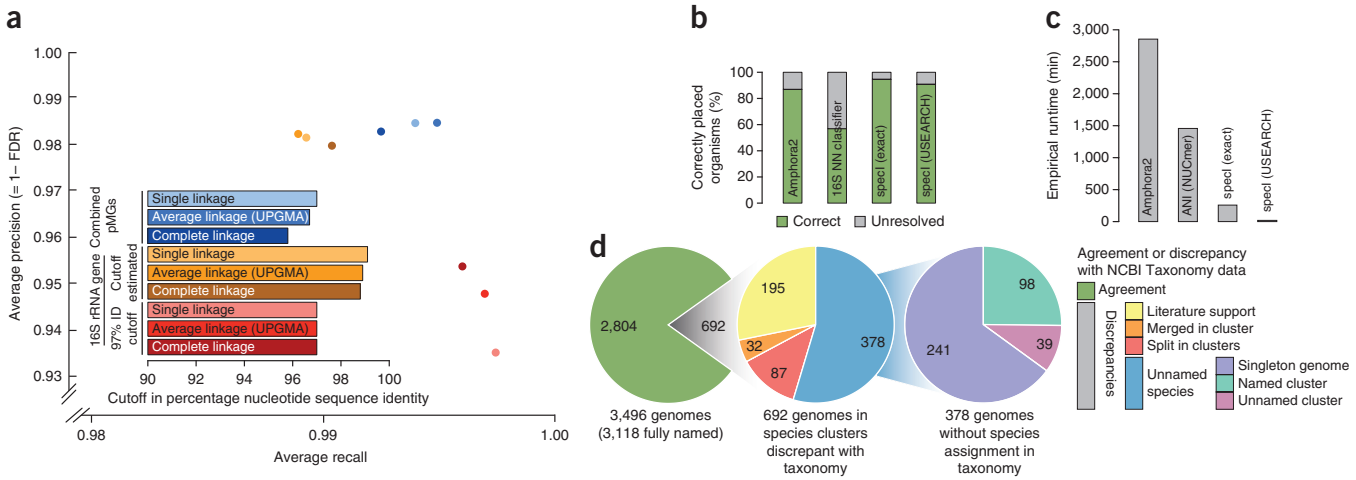

Figure 1 | Comparative performance assessment of specI. (**a**) Performance of the 40 pMGs (combined pMGs) and 16S rRNA gene species-level clustering cutoffs (cutoff estimated) as well as the classical 16S rRNA gene 97% nucleotide sequence identity cutoff (97% ID cutoff) in terms of average precision and recall using indicated clustering algorithms. The cutoffs are reported in the inset. (**b**) Accuracy of species placements compared to type strains using specI (two implementations, using different alignment algorithms; Online Methods), Amphora2 and a 16S rRNA nearest-neighbor classifier (16S NN classifier; which included an optimized 99% ID cutoff) based on 130 holdout genomes, whose taxonomy is not disputed. (**c**) Empirical runtimes of specI, Amphora2 and ANI calculations for analysis of 100 randomly chosen genomes. For ANI calculations, only the MUMmer alignment step of JSpecies was benchmarked (Online Methods). (**d**) Large-scale application of the species-level clustering method to 3,496 high-quality genomes. The genomes were either in agreement with the NCBI Taxonomy data or showed discrepancies. In these cases, the species-level clustering either was supported by literature, or implied splits or merges of named species. Genomes that were not taxonomically identified were subdivided into different types according to their species clusters.

(see Online Methods, **Supplementary Fig. 1** and **Supplementary Table 4** for details on type strain classification). For each pMG, we calculated distances between all pairs of genomes as inverse nucleotide sequence identities. We used nucleotide sequences as they provide better resolution than amino acid sequences when comparing closely related organisms[15]. Next, we combined the distances using a length-weighted average (Online Methods), which confers robustness to potential errors in individual pMGs (for example, due to rare duplication, loss and horizontal transfer events or misidentification of one or more pMGs in a genome).

Based on the combined distance metric, we partitioned genomes into species clusters, exploring single, average and complete linkage-clustering algorithms. We applied a twofold cross-validation approach to determine the sequence identity cutoffs, which provided optimal species delineations in comparison to a taxonomic classification based on type strains[14]. Finally, we assessed the accuracy of our clustering approach as cross-validation error using the average false discovery rate (FDR; equals 1 – precision) among the clusters and the average recall among the named species in the type-strain taxonomy (Online Methods). A nucleotide identity cutoff of 96.5% (**Fig. 1a** and **Supplementary Fig. 2**) for average linkage clustering yielded optimal results (average recall of 98.9% and average FDR of 1.5%, **Fig. 1a**; see **Supplementary Table 5** for a list of discrepant genomes). Cutoffs for the other clustering algorithms only deviated slightly (**Fig. 1a**).

We extended the assessment to include analysis of the 16S rRNA gene. The most frequently used species cutoff for operational taxonomic units of 97% sequence identity[6] yielded clusters that were less accurate than those derived from the method based on pMGs, with more than threefold larger FDR (**Fig. 1a**). When we optimized the 16S rRNA sequence identity cutoff in the same way as for the 40 pMGs, identities between 98.7% and 99.1% (depending on clustering algorithm) distinguished prokaryotic species best, in agreement with a previously recommended range of 98.7–99% sequence identity[7] (**Fig. 1a**), although the average FDR and average recall were worse than for the combined 40 pMGs.

We performed global species clustering with the aim of revising the current species classification in an accurate and consistent manner. In addition, we offer the specI tool for the automated placement of newly sequenced genomes into our global clustering (Online Methods). The accuracy of specI for assigning new genomes into named species clusters was compared to existing taxonomic placement methods Amphora2[10], JSpecies[4] (implementing the ANI method) and 16S rRNA nearest-neighbor classification, using the type strain classification (**Fig. 1b**,**c**; genomes used are listed in **Supplementary Table 6**). specI was more accurate than Amphora2 and the 16S rRNA nearest-neighbor classifier (**Fig. 1b**,**c** and Online Methods). It also was considerably faster than Amphora2 and JSpecies, including the speed-optimized MUMmer-based implementation[4] (**Fig. 1c**). Because a large-scale comparison to assess the overall accuracy of species assignments by JSpecies (ANI) was too time-consuming (Online Methods), we selected several clades; in all of these, the species groupings were very similar between specI and ANI (**Supplementary Tables 7–13**).

Applying the optimized cutoff (96.5% sequence identity for combined pMGs) to the set of 3,496 prokaryotic high-quality genomes resulted in 1,753 species clusters (**Supplementary Table 14**), which we examined for inconsistencies with the current NCBI Taxonomy information (a widely used classification that is less well curated than the type strain classification). Although not all of these 3,496 genomes contained all 40 pMGs (mostly due to incomplete genome sequences; **Supplementary Table 15**), we expect this to have only negligible influence on the results, as specI is robust to missing pMGs (**Supplementary Fig. 3**).

After removing 378 genomes not assigned to a species in the NCBI Taxonomy database, 2,804 of the remaining 3,118 genomes

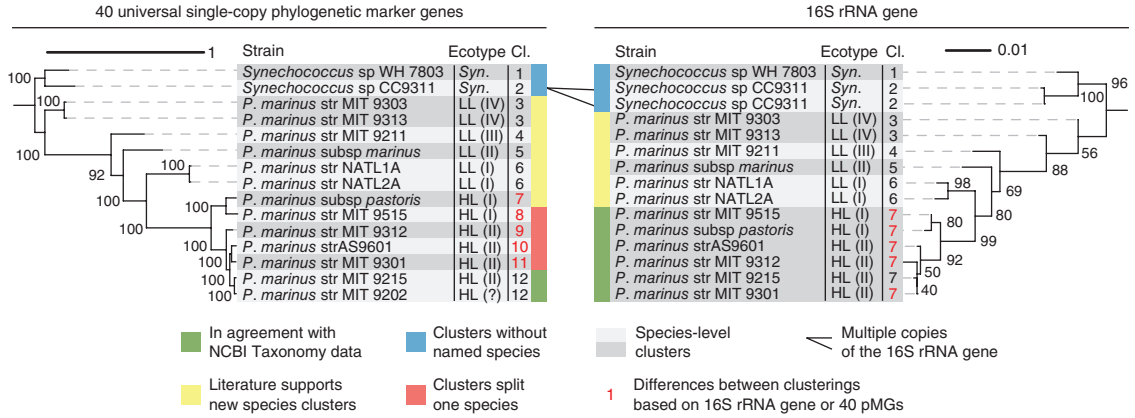

Figure 2 | Phylogenetic trees and species-level clustering of *Prochlorococcus* displaying discrepancies with the NCBI Taxonomy data. Trees were independently built from concatenated alignments of the 40 universal single-copy phylogenetic marker genes and from the 16S rRNA gene (see **Supplementary Figs. 5** and **6** for additional examples). Species-level clusterings and phylogenetic trees of the combined 40 pMGs and the 16S rRNA gene suggest that *Prochlorococcus marinus* ecotypes form individual species. Cl., specI cluster.

(89.9%) formed clusters in agreement with the NCBI Taxonomy database (**Fig. 1d**). Manual investigation of the 314 assignments that disagreed indicated that for 195 of these (62.1%), the results of our method were supported by classification proposals in recent literature (**Fig. 1d** and **Supplementary Table 16**). We found no publications on the taxonomy for most of the 119 unsupported disagreements. When including all literature-supported cases, we estimate at least 96.2% of our assignments are likely to be correct. Of the remaining discrepancies, data for 87 of 119 genomes (73.1%) suggest a split of the original species according to our clustering, whereas merges of different species into the same cluster occurred less frequently (**Fig. 1d**). The rate of correct assignments was even higher for the type strains, and nearly all discrepancies were supported by ANI values (**Supplementary Table 7**). The discrepancies with the NCBI Taxonomy data were not concentrated in certain phyla (**Supplementary Fig. 4**), indicating that a single, globally adjusted criterion for species definition appears to work well throughout the archaeal and bacterial domains of life.

We observed systematic discrepancies between our consistent and universally applied species delineation criteria and the NCBI Taxonomy database for clinically relevant genera because these tend to have a more fine-grained species taxonomy (for example, *Mycobacterium tuberculosis*[16]), resulting in merges in our clusters. In contrast, more coarse-grained naming is often applied in understudied clades, which can be resolved at a finer resolution with our approach (see below; **Supplementary Table 16**). Finally, some species are inherently difficult to delineate based on molecular markers and consequently pose a source of error for automated approaches. These organisms can be identified as extreme cases exhibiting low coherence between species assignments based on individual pMGs, as the coherence between pMGs is generally very high (87% of genomes have a coherence score >90%; **Supplementary Fig. 4** and Online Methods).

To better understand the reasons for discrepancies between the specI results and the current taxonomic classification, we investigated three clades in detail (**Supplementary Table 17**). For these, we built phylogenetic trees using concatenated alignments of the 40 pMGs and the 16S rRNA gene and compared them with trees constructed from concatenated alignments of all one-to-one orthologs (derived using eggNOG routines[13]). These one-to-one orthologs represent a considerable fraction of the analyzed genomes (12.3–65.9%), and thus we expected them to give a very robust signal. This tree-based approach can resolve the phylogenetic history of the clades and reveal whether named species or species clusters are monophyletic (as is desired for consistent taxonomy) or not. The phylogenetic trees had identical topologies (**Fig. 2**, and **Supplementary Figs. 5** and **6**). The ANI[4,8] values we calculated for these clades also coincided with the species clusters and trees (**Supplementary Tables 8–13**). These results demonstrate the power of the 40 pMGs for species delineation. In contrast, the trees generated from 16S rRNA gene alignments had differences in topologies and typically much lower bootstrap support at the species level (**Fig. 2**, and **Supplementary Figs. 5** and **6**). However, splits between different genera defining monophyletic clades also had high bootstrap support, confirming the usefulness of the 16S rRNA gene analysis for inferring genus membership (as provided by tools such as Greengenes[17]).

We investigated two genera known to be co-differentiating long-time endosymbionts of insects: *Buchnera* and *Wigglesworthia* (**Supplementary Figs. 5a** and **6b**)[18,19]. In an earlier study based on ANI values, the authors proposed the division of *Buchnera* into host-specific taxa[4], and our results support a co-speciation scenario by both clustering and phylogenetic trees for these genera, at a resolution that could not be obtained with 16S rRNA–based analysis.

specI can be used to find and resolve errors in the NCBI Taxonomy database. For example, we propose a reclassification of *Serratia odorifera* 4Rx13 to *Serratia plymuthica,* which is also supported by phylogenetic trees as well as ANI calculations. Our species clusters of the genus *Serratia* also suggest the inclusion of two unnamed organisms in the *S. plymuthica* species (**Supplementary Figs. 5b** and **6c**). Polyphyly is a frequent type of inconsistency in the NCBI Taxonomy database, as highlighted by *S. odorifera*, which was readily detected by specI (**Supplementary Fig. 5b** and **6c**). Other examples of polyphyletic species include known cases from the genus *Escherichia–Shigella*[20] (**Supplementary Fig. 7**), for which one specI cluster represented the whole *Escherichia–Shigella* clade except *Escherichia albertii* (**Supplementary Note**).

Finally, we examined a major discrepancy between our results and the NCBI Taxonomy database for the genus *Prochlorococcus*.

This genus had been identified relatively recently[21] and classified as a single species (*Prochlorococcus marinus*) based on the 97% sequence identity cutoff for the 16S rRNA marker, with a subdivison into 'ecotypes' proposed later[22]. specI splits this conglomeration of 13 genomes into several species clusters that correspond to ecological preference with respect to light availability (ecotypes HL and LL), regardless of whether clusters are based on pMGs or the 16S rRNA gene (**Fig. 2** and **Supplementary Fig. 6a**). The more fine-grained species distinction suggested by clusters based on pMGs was independently supported by whole-genome comparisons using ANI (**Supplementary Tables 8** and **9**).

The robustness and accuracy of specI allowed us to confidently assign a large portion of the 378 unidentified organisms to a named species or to propose new species. We assigned 98 genomes to named species (**Supplementary Table 18**). In addition, 39 genomes, which were grouped into 15 unclassified species clusters with two or more members (**Supplementary Table 19**), and 241 singleton genomes require new species names (**Fig. 1d**). To evaluate the accuracy of these assignments, we classified an additional 100 recently sequenced genomes (not included for global species delineation of 3,496 genomes), 92 of which were assigned to the same species as in the NCBI Taxonomy database (**Supplementary Fig. 8**), but most of the remaining eight genomes were clearly distinct from their named species clusters in terms of divergence of pMGs (**Supplementary Table 20**). Classification of new genomes with specI is completely automatic. The only manual intervention needed is the selection of a name for newly discovered species. Over 10% of all sequenced genomes (378 of 3,496) are unclassified at the species level today, and we expect this number to increase, highlighting the need for automated species assignment.

In conclusion, we developed a method based on 40 pMGs[11,12] to address the need for accurate, universal, computationally efficient and automated approaches for species assignment of existing and newly sequenced genomes. specI is applicable to any newly sequenced genome (via a dedicated web server), as it uses universal, single-copy phylogenetic marker genes that are automatically identified. The approach facilitates large-scale applications because it only uses ~1% of a prokaryotic genome and is thus faster than existing methods that use the whole genome, while maintaining high accuracy and robustness. We believe that specI has the potential to help make prokaryotic taxonomy, which has been called "the most subjective branch in any biological discipline,"[23] a more objective field. specI provides reliable guidelines for species classification, could in principle be extended to other taxonomic levels (such as genera) and could serve as a basis for other large-scale applications that rely on a consistent taxonomy guided by molecular markers in comparative genomics and metagenomics.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the members of the Bork group for helpful discussions and Y. Yuan and members of the European Molecular Biology Laboratory information technology core facility for managing the high-performance computing resources. We acknowledge funding provided by the CancerBiome project (European Research Council project reference 268985), the 'METACARDIS' project (FP7-HEALTH-2012-INNOVATION-I-305312) and the International Human Microbiome Standards project (HEALTH-F4-2010-261376).

AUTHOR CONTRIBUTIONS

P.B., D.R.M., S.S. and G.Z. designed the study. D.R.M. developed and implemented the program, D.R.M. and G.Z. performed the experiments, D.R.M., S.S. and G.Z. analyzed the data, and D.R.M., S.S., G.Z. and P.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Rosselló-Mora, R. & Amann, R. *FEMS Microbiol. Rev.* **25**, 39–67 (2001).
2. Stackebrandt, E. *et al. Int. J. Syst. Evol. Microbiol.* **52**, 1043–1047 (2002).
3. Kämpfer, P. & Glaeser, S. *Environ. Microbiol.* **14**, 291–317 (2012).
4. Richter, M. & Rosselló-Móra, R. *Proc. Natl. Acad. Sci. USA* **106**, 19126–19131 (2009).
5. Chun, J. *et al. Int. J. Syst. Evol. Microbiol.* **57**, 2259–2261 (2007).
6. Stackebrandt, E. & Goebel, B.M. *Int. J. Syst. Bacteriol.* **44**, 846–849 (1994).
7. Stackebrandt, E. & Ebers, J. *Microbiol. Today* **33**, 152 (2006).
8. Konstantinidis, K. & Tiedje, J. *Proc. Natl. Acad. Sci. USA* **102**, 2567–2572 (2005).
9. von Mering, C. *et al. Science* **315**, 1126–1130 (2007).
10. Wu, M. & Scott, A. *Bioinformatics* **28**, 1033–1034 (2012).
11. Ciccarelli, F. *et al. Science* **311**, 1283–1287 (2006).
12. Creevey, C.J. *et al. PLoS ONE* **6**, e22099 (2011).
13. Powell, S. *et al. Nucleic Acids Res.* **40**, 9 (2012).
14. Murray, R.G.E. *Int. J. Syst. Bacteriol.* **46**, 831 (1996).
15. Konstantinidis, K. & Tiedje, J. *J. Bacteriol.* **187**, 6258–6264 (2005).
16. Kremer, K. *et al. J. Clin. Microbiol.* **37**, 2607–2618 (1999).
17. McDonald, D. *et al. ISME J.* **6**, 610–618 (2012).
18. Jousselin, E., Desdevises, Y. & Coeur d'acier, A. *Proc. Royal Soc. B Biol. Sci.* **276**, 187–196 (2009).
19. Chen, X., Li, S. & Aksoy, S. *J. Mol. Evol.* **48**, 49–58 (1999).
20. Brenner, D.J. *Bergey's Manual of Systematic Bacteriology* **1**, 408–420 (The Williams & Wilkins Co., 1984).
21. Chisholm, S.W. *et al. Nature.* **334**, 340–343 (1988).
22. Moore, L., Rocap, G. & Chisholm, S. *Nature* **393**, 464–467 (1998).
23. Cowan, S. *J. Gen. Microbiol.* **67**, 1–8 (1971).

ONLINE METHODS

Genome selection. All prokaryotic genome sequences listed at http://www.ncbi.nlm.nih.gov/genome/browse/ were downloaded in conjunction with their predicted genes and proteins from the NCBI Nucleotide database and European Molecular Biology Laboratory (EMBL) Bank website on 23 February 2012. To filter out low-quality genomes, we removed genomes that had more than 300 contigs and a length for which the collection of all contigs of that length or longer contain at least half of the total of the lengths of the contigs (N50) of less than 10 kilobases (removing 213 genomes). Next, we extracted the 40 pMGs[11,12,24] from all genomes and removed all incomplete genomes for which less than 30 pMGs were found (removing 68 genomes). This quality filtering retained 1,789 complete genomes and 1,707 high-quality incomplete genomes, and 281 genomes were removed. The resulting set of 3,496 genomes was used for the global analysis (**Fig. 1d**).

From this comprehensive collection of genomes, we extracted a subset of genomes for the estimation of optimal species delineation cutoffs. This set contains 943 genomes, which contained all 40 pMGs, at least one full-length copy of the 16S rRNA gene (**Supplementary Table 15**) and whose named species was represented by a sequenced type strain as a ground truth for taxonomic classification.

One hundred additional annotated genome sequences (including predicted genes and proteins) were downloaded on 24 May 2013 and used to evaluate the specI classification accuracy. The results are shown in **Supplementary Figure 8** and **Supplementary Table 20**.

Extraction of 40 universal single-copy phylogenetic marker genes. We annotated the 40 pMGs of the 3,496 prokaryotic genomes using SMASH routines (version 1.6)[25] and sequences of the 40 pMGs obtained from eggNOG version 3 (v3; ref. 13). For this, the genes of the 3,496 genomes were screened against eggNOG v3 using Blast[26], filtered for best hits and annotated using SMASH.

Extraction of 16S rRNA gene. Full-length 16S rRNA gene sequences were identified in the 3,496 genomes using rna_hmm.py[27] with customized hidden Markov models generated using hmmbuild[28] and alignments downloaded from the SILVA database (version SSU r104; alignment quality = 100; sequence length ≥ 1,200; sequence quality ≥ 75; restrict search to SILVA)[29]. Sequences with a length <1,250 base pairs[30] and a bootstrap support of <90% at the genus level were removed from the analysis.

Extraction of one-to-one orthologs. We generated orthologous groups of all genes from the genomes shown in **Figure 2** and **Supplementary Figure 5** using the eggNOG pipeline[31]. We then extracted all orthologous groups that contained exactly one gene member from each genome. We found 629 one-to-one orthologs in the *Prochlorococcus-Synechococcus* clade used for the analysis in **Supplementary Figure 6a**, 666 one-to-one orthologs for the *Buchnera-Wigglesworthia* clade used for **Supplementary Figure 6b** and 161 one-to-one orthologs for the *Serratia-Rahnella* clade used for **Supplementary Figure 6c**.

Species type strains. We collected information about the type strains of all listed species from the 'List of prokaryotic names with standing in nomenclature' hosted at http://www.bacterio.net/. These were then assigned to NCBI Taxonomy identifiers (downloaded on 24 July 2012) by matching the culture collection names of the type strains with the NCBI Taxonomy names including all synonyms. The results were manually inspected, and mismatching names were excluded. All type strains that could be assigned to an NCBI Taxonomy identifier were classified depending on whether a genome sequence for this NCBI Taxonomy identifier was available or not (**Supplementary Fig. 1** and **Supplementary Table 5**).

NCBI Taxonomy curation. We used the NCBI Taxonomy database (downloaded on 24 July 2012), marked all species that were not assigned a proper species name in the NCBI Taxonomy database (for example, "*Vibrio* sp. Ex25"), to exclude these genomes when calculating the accuracy of species classification.

To investigate discrepancies between our species-level clustering and the existing NCBI Taxonomy database in detail, we performed a 'targeted' curation of the latter based on a literature survey of all discrepant genomes. This extensive curation resulted in the reclassification of 182 genomes found in the NCBI Taxonomy database (**Supplementary Table 16**).

Calculation of distances. We calculated all symmetrical pair-wise distances for the 40 pMGs as well as the 16S rRNA gene using glsearch of the fasta package (version 36)[32] by globally aligning the shorter sequence to the longer one for every pair of sequences. We then extracted pair-wise percentage nucleotide sequence identities. To generate distance matrices as input for the clustering procedure, we transformed identities using the formula: distance = 1 – identity. Subsequently, we calculated the gene-length–weighted average percentage sequence identity and distance of the combination of all pMGs (emulating a distance derived from concatenation). The increased phylogenetic resolution of the combined set of pMGs has been shown before[11].

Clustering. Clustering was performed using the SciPy package hcluster (version 0.2.0) (Eads, D. https://code.google.com/p/scipy-cluster/) for the Python programming language. Distance matrices were generated as described above and then clustered using single, average and complete linkage. The hierachical clusterings were then transformed into discrete species-level clusters through the application of distance cutoffs. To select optimal cutoffs, we regenerated clusterings with various distance cutoffs ranging between 0% and 100% dissimilarity in 1,000 linear increments.

Evaluation of clustering results. Congruence between clusters and a reference taxonomy was computed using the average FDR of all clusters and the average recall of all named species. FDR of a cluster was defined as the proportion of genomes not belonging to the majority species within that cluster, and recall of a named species was defined as the largest proportion of genomes from this species that were grouped together in a single cluster. Both FDR and recall had values between 0 and 1, with low FDR values and high recall values indicating good agreement with the taxonomy. For each clustering generated with a certain distance cutoff, we computed the average FDR and recall as the mean of the FDR and recall values across all clusters weighted by the number of genomes per cluster or species, respectively. We used 10 as a

maximum weight even if more representative genomes existed, to decrease bias (toward very well-studied species such as *Escherichia coli*, for which >100 genomes have been sequenced).

Selection of cutoffs. To select optimal nucleotide identity cutoffs for the 40 pMGs and the 16S rRNA gene, we first selected a set of 943 genomes in which all 40 pMGs as well as a full-length 16S rRNA gene could be identified and whose named species was represented by a sequenced type strain. This set was randomly split into two subsets for twofold cross-validation. For both subsets, we calculated the cutoffs with the highest average *F* score (defined as $F1 = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$ with precision = 1 − FDR) for each of the pMGs individually as well as the combined set. Next we evaluated the performance of these cutoffs by applying them on the respective holdout subset and calculating the average FDR and recall on all clusters from both sets (**Fig. 1a**). To obtain a global cutoff, we calculated the average of the two cutoffs from the individual subsets. Also, over-fitting regarding the existing taxonomic classification is unlikely as our cross-validation procedure guarded against it. The resulting optimal cutoffs were then applied to the complete set of 3,496 genomes to obtain comprehensive species-level clusters, which were compared in detail with the NCBI taxonomic classification (**Fig. 1d**).

Assessment of discrepancies between clustering results and NCBI Taxonomy database. We extracted discrepancies between the NCBI Taxonomy database and the clustering of the combined 40 pMGs applied to the set of 3,496 genomes. We manually inspected these and performed a literature survey to classify all observed discrepancies into one of the following types: (i) literature supports the clustering rather than the NCBI Taxonomy database, (ii) clusters merged genomes from more than one species into a single cluster and (iii) splits corresponding to cases where several clusters contained representative genomes of the same species. This was only possible for genomes with a proper species name in the NCBI Taxonomy database. The remaining genomes without explicit species names (for example, genomes that are taxonomically not assigned to a named species such as "*Pusillimonas* sp. T7-7" or "uncultured Termite group 1 bacterium") were classified depending on whether they were grouped together with named genomes, which suggested the transfer of that species name (**Fig. 1d**).

Coherence between different phylogenetic marker genes. We calculated a score reflecting the coherence between clusterings based on individual pMGs. This was done by pair-wise comparisons of cluster memberships of the 3,496 genomes. If two genomes were found in the same cluster in at least half of all 40 independent clusterings, we counted cases in which these genomes did not cluster together as false negatives and consistent clusterings as true positives. If a pair of genomes co-clustered in less than half of the single-pMG clusterings, all instances in which we observed co-clustering nonetheless were counted as false positives, otherwise as true negatives. For each genome we summed up these numbers and calculated the *F* score as the final cluster-coherence score. The results are visualized using iTOL[33] as shown in **Supplementary Figure 4**.

specI web server. The specI tool is provided as an easy to use web application at http://www.bork.embl.de/software/specI/. Users can upload the predicted genes sequences of a genome and their respective protein sequences. The 40 pMGs are then automatically extracted from these sequences using the fetchMG tool (http://vm-lux.embl.de/~mende/fetchMG/). Next the length-weighted average distance to all 1,753 species clusters is determined using usearch[34]. For the 'accurate' algorithm, the distances to all clusters closer than 85% nucleotide sequence identity are realigned using glsearch. In the last step, the most similar species cluster is determined and the input genome is assigned to the species cluster if its length-weighted average nucleotide sequence identity is higher than 96.5%.

Comparison of taxonomic placement methods. We compared specI to Amphora2 (ref. 10) and a 16S rRNA gene–based nearest-neighbor classifier (calling a species match, if the maximum pair-wise nucleotide identity between full-length 16S rRNA genes was at least 98%). For a fair method comparison, we generated a specI reference clustering including only those genomes that are also present in the original Amphora2 reference tree. We then compiled a test set consisting of genomes that were not included in the reference tree or clustering, but for which another genome belonging to the same type strain species was. Additionally, we required that at least one full-length 16S sequence could be identified in these genomes, resulting in a set of 212 genomes. From these, we excluded genomes whose taxonomy is debated, which yielded a final test set of 130 genomes (**Supplementary Table 6**). These genomes were classified using the three approaches mentioned above (**Fig. 1b**).

Runtime comparison of taxonomic placement methods. We compared the runtime of specI to that of Amphora2 and JSpecies for the placement of 100 randomly chosen genomes (**Fig. 1c**). As the JSpecies GUI (graphical user interface) seemed to be very slow when using that many genomes, we only benchmarked the speed of the MUMmer tool used to for the fast calculation of ANIm values[35].

Accuracy estimations for missing phylogenetic marker genes. To study the impact of missing pMGs, we randomly removed 10, 20 and 30 pMGs from the analysis and used specI to taxonomically place the test set of 130 genomes that was generated for the above described method comparison. We repeated the analysis ten times with differing sets of pMGs.

Phylogenetic trees. Phylogenetic trees were built for concatenated DNA alignments of the 40 pMGs, the 16S rRNA gene and a concatenation of all one-to-one orthologs found in the clades used to generate the tree (see above for ortholog identification). Alignments were made with AQUA (version 1.1)[36] using standard parameters and masked using Gblocks (version 0.91b) with the 'relaxed' parameter settings ("minimum number of sequences for a flank position": 9; "maximum number of contiguous nonconserved positions": 10; "minimum length of a block": 5; "allowed gap positions": "with half")[37]. Maximum likelihood phylogenetic trees were built with RAxML 7.2.8 (ref. 38) using the 'GTRGAMMA' model and default parameters with 100 bootstraps per data set. Phylogenetic trees were visualized using iTOL (version 2.2)[33].

ANI calculation. We used JSpecies (V1.2.1)[4] to calculate ANIb and ANIm values for a number of genomes (**Supplementary Tables 7–13**).

doi:10.1038/nmeth.2575

24. Sorek, R. *et al. Science* **318**, 1449–1452 (2007).
25. Arumugam, M., Harrington, E., Foerstner, K., Raes, J. & Bork, P. *Bioinformatics* **26**, 2977–2978 (2010).
26. Altschul, S. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
27. Huang, Y., Gilna, P. & Li, W. *Bioinformatics* **25**, 1338–1340 (2009).
28. Finn, R., Clements, J. & Eddy, S. *Nucleic Acids Res.* **39**, 37 (2011).
29. Pruesse, E. *et al. Nucleic Acids Res.* **35**, 7188–7196 (2007).
30. Caporaso, J. *et al. Bioinformatics* **26**, 266–267 (2010).
31. Jensen, L. *et al. Nucleic Acids Res.* **36**, D250–D254 (2008).
32. Pearson, W. & Lipman, D. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).
33. Letunic, I. & Bork, P. *Nucleic Acids Res.* **39**, W475–W478 (2011).
34. Edgar, R. *Bioinformatics* **26**, 2460–2461 (2010).
35. Delcher, A., Salzberg, S. & Phillippy, A. *Curr. Protoc. Bioinformatics* **10**, 10.3 (2003).
36. Muller, J., Creevey, C., Thompson, J., Arendt, D. & Bork, P. *Bioinformatics* **26**, 263–265 (2010).
37. Talavera, G. & Castresana, J. *Syst. Biol.* **56**, 564–577 (2007).
38. Stamatakis, A. *Bioinformatics* **22**, 2688–2690 (2006).