DOI 10.1099/ijs.0.059774-0

Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes

Mincheol Kim,[1] Hyun-Seok Oh,[2] Sang-Cheol Park[2] and Jongsik Chun[1,2]

[1]School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

[2]Interdisciplinary Program in Bioinformatics and Bioinformatics Institute, Seoul National University, Seoul 151-742, Republic of Korea

Correspondence
Jongsik Chun
jchun@snu.ac.kr

Among available genome relatedness indices, average nucleotide identity (ANI) is one of the most robust measurements of genomic relatedness between strains, and has great potential in the taxonomy of bacteria and archaea as a substitute for the labour-intensive DNA–DNA hybridization (DDH) technique. An ANI threshold range (95–96 %) for species demarcation had previously been suggested based on comparative investigation between DDH and ANI values, albeit with rather limited datasets. Furthermore, its generality was not tested on all lineages of prokaryotes. Here, we investigated the overall distribution of ANI values generated by pairwise comparison of 6787 genomes of prokaryotes belonging to 22 phyla to see whether the suggested range can be applied to all species. There was an apparent distinction in the overall ANI distribution between intra- and interspecies relationships at around 95–96 % ANI. We went on to determine which level of 16S rRNA gene sequence similarity corresponds to the currently accepted ANI threshold for species demarcation using over one million comparisons. A twofold cross-validation statistical test revealed that 98.65 % 16S rRNA gene sequence similarity can be used as the threshold for differentiating two species, which is consistent with previous suggestions (98.2–99.0 %) derived from comparative studies between DDH and 16S rRNA gene sequence similarity. Our findings should be useful in accelerating the use of genomic sequence data in the taxonomy of bacteria and archaea.

INTRODUCTION

DNA–DNA hybridization (DDH) is a category of experimental methods which indirectly measure the overall similarity between two genome sequences (McCarthy & Bolton, 1963; Schildkraut *et al.*, 1961). Over the last 50 years, DDH has been the ‘gold standard’ for bacterial species demarcation as it provides a clear and objective numerical threshold for a species boundary, for which 70 % DDH was suggested and is widely used (Tindall *et al.*, 2010; Wayne *et al.*, 1987). However, due to the labour-intensive and error-prone nature of DDH experiments, there has been a continuous demand for an alternative genotype-based standard (Gevers *et al.*, 2005; Stackebrandt *et al.*, 2002).

The genome sequence of a microbial strain is the ultimate information for microbial taxonomy. As whole-genome sequencing has become more widely accessible due to the introduction of cost-effective high-throughput DNA sequencing technology, it is evident that genome sequence similarities have the potential to be a routine taxonomic parameter as a replacement for DDH. Much effort has been put into correlating DDH values with digital DDH-like similarity indices based on computational comparisons of two genome sequences. These include average nucleotide identity (ANI) (Konstantinidis *et al.*, 2006), genome BLAST distance phylogeny (GBDP) (Henz *et al.*, 2005) and the maximal unique matches index (MUMi) (Deloger *et al.*, 2009). Of these, ANI has been most widely used as a possible next-generation gold standard for species delineation (Chan *et al.*, 2012; Goris *et al.*, 2007; Grim *et al.*, 2013; Haley *et al.*, 2010; Konstantinidis & Tiedje, 2005; Richter & Rosselló-Móra, 2009; Yi *et al.*, 2012). ANI represents a mean of identity/similarity values between homologous genomic regions shared by two genomes. It is now generally accepted that ANI values of 95–96 % equate to a DDH value of 70 %, and can be used as a boundary for species delineation (Goris *et al.*, 2007; Richter & Rosselló-Móra, 2009).

An early comparative study between DDH and 16S rRNA gene sequence similarity revealed that 97 % 16S rRNA gene sequence similarity corresponded to 70 % DDH

Abbreviations: ANI, average nucleotide identity; DDH, DNA–DNA hybridization.

A supplementary figure and two supplementary tables are available with the online version of this paper.

(Stackebrandt & Goebel, 1994). This threshold of 16S rRNA gene sequence similarity has been widely used in bacterial classification as an alternative to the tedious DDH, and it is now generally accepted that DDH is only required when 16S rRNA gene sequence similarity between two strains is over 97 % (Tindall *et al.*, 2010), even though higher thresholds of 98.7–99.0 % have also been used (Stackebrandt & Ebers, 2006). However, in many cases, it is hard to differentiate two species using 16S rRNA gene sequences alone, as some species share a high level of 16S rRNA gene sequence similarity (>99 %), even though they are clearly separated by DDH (Ash *et al.*, 1991; Rosselló-Mora & Amann, 2001). On the basis of a comparison of 380 pairs of strains, Stackebrandt & Ebers (2006) suggested 98.7–99.0 % 16S rRNA gene sequence similarity as a threshold, which is higher than the original value of 97 % proposed by Stackebrandt & Goebel (1994). More recently, a similar range of threshold (98.2–99.0 %) was recommended using information derived from 571 pairs of strains (Meier-Kolthoff *et al.*, 2013). All of these studies focused on the comparison of DDH and 16S rRNA gene sequence similarity. However, correlation of ANI and 16S rRNA gene sequence similarity has not been thoroughly investigated, except for a small-scale study based on 70 genomes (Konstantinidis & Tiedje, 2005).

In this study, we used a large set of genome sequences ($n=6787$) to determine the optimal threshold of 16S rRNA gene sequence similarity for species delineation that corresponds to 95–96 % ANI. In addition, the overall distribution of ANI values was investigated to review the potential for the use of ANI in the classification of prokaryotes.

METHODS

Data collection and genome selection. A set of 11 995 genome sequences of prokaryotes was retrieved from the GenBank database (as of 30 July 2013). Taxonomic identification of each genome sequence was achieved by calculating 16S rRNA gene sequence similarity and ANI between the query genome sequences and those of type strains. Low-quality genomes, which have >2000 contigs, and draft genomes stemming from single-cell genomics using multiple displacement amplifications were omitted due to their low genome coverage and poor quality. Genomes that do not contain any extractable full-length 16S rRNA gene sequence were also removed from the dataset. The resulting set of 6787 genomes was used for further analysis.

Calculation of pairwise ANI and 16S rRNA gene sequence similarities. Among all qualified genomes of prokaryotes, ANI values were calculated between all genomes of strains belonging to the same family. Computation between strains belonging to different families was not considered as it produced low values (<60 % ANI) which are insignificant. All ANI calculations were performed using BLASTN and in-house software, following the algorithm described by Goris *et al.* (2007). Calculations were only performed in cases where there were at least three genomes and two species within a family. Statistics on the distribution of ANI values were analysed using the R statistical software (http://www.r-project.org). To get normalized ANI values, a mean of pairwise ANI values was calculated from those obtained for all pairs between strains belonging to the same species. The rRNAselector (Lee *et al.*, 2011) program was used to obtain full-length 16S rRNA gene sequences from each genome. The degree of completeness of an rRNA gene sequence was checked as described previously (Kim *et al.*, 2012). Pairwise similarities between 16S rRNA gene sequences were calculated based on robust global sequence alignment algorithms as used in the EzTaxon server (Chun *et al.*, 2007). The possible association between ANI and 16S rRNA gene sequence similarity was tested by logarithmic regression.

Determination of an optimal 16S rRNA gene sequence similarity threshold for species demarcation. To determine the optimal threshold in 16S rRNA gene sequence similarity corresponding to a given ANI threshold for species demarcation, precision-recall and F score were calculated for all similarities of 16S rRNA gene sequences at an interval of 0.01 %. The F measure was originally introduced for measuring classification performance in information retrieval processes (van Rijsbergen, 1979) and has been used frequently in assessing the performance of binary or multilabel classifiers (Lan *et al.*, 2012; Read *et al.*, 2011). Recently, this performance assessment strategy was applied successfully to determine the optimal sequence similarity threshold for species demarcation of prokaryotes (Mende *et al.*, 2013) and to find coherence between a given taxonomy and its phylogenetic placement (McDonald *et al.*, 2012). We assumed that two strains showing pairwise ANI values below a given threshold (95 or 96 %) belong to different species and otherwise to the same species (Goris *et al.*, 2007; Richter & Rosselló-Móra, 2009). All pairwise 16S rRNA gene sequence similarity values were assigned into four categories given thresholds of X ANI and Y 16S rRNA gene sequence similarity: true positives (TP) consisted of cases with ANI $\geqslant X$ and 16S rRNA gene sequence similarity $\geqslant Y$; false negatives (FN) with ANI $\geqslant X$ and 16S rRNA gene sequence similarity $<Y$; false positives (FP) with ANI $<X$ and 16S rRNA gene sequence similarity $\geqslant Y$ and true negatives (TN) with ANI $<X$ and 16S rRNA gene sequence similarity $<Y$. After each round of evaluations with ANI (95 or 96 %) and 16S rRNA gene sequence similarity (95–100 % at 0.01 % intervals) thresholds, the full dataset was randomly split into two subsets for twofold cross-validation and the precision and recall values were estimated for each subset using the four categories defined above. The optimal threshold was obtained by maximizing the sensitivity (recall) while minimizing the false discovery rate (1−specificity). Afterwards, precision and recall values were used to calculate the F score, which is a harmonic mean of precision and recall and represents an accuracy of the test. The highest F score among the series of 16S rRNA gene sequence similarity thresholds was selected as the final optimal cut-off for each subset, and the performance of each cut-off was evaluated by applying it to the respective holdout subset.

RESULTS AND DISCUSSION

Overall ANI distribution

The final dataset consisted of 1 044 179 pairwise values of 16S rRNA gene sequence similarity and ANI, covering 22 phyla, 41 classes, 93 orders, 202 families, 655 genera and 1738 species. The ANI values calculated between strains belonging to the same family showed uneven distribution, having an obvious low-frequency area in the range of 81.0–96.0 % ANI and high frequency at >96 % ANI (Fig. 1a). Most of the intraspecies ANI values were found to be over 96 %, which is within the range previously recommended for species demarcation (Richter & Rosselló-Móra, 2009). A similar pattern was also observed

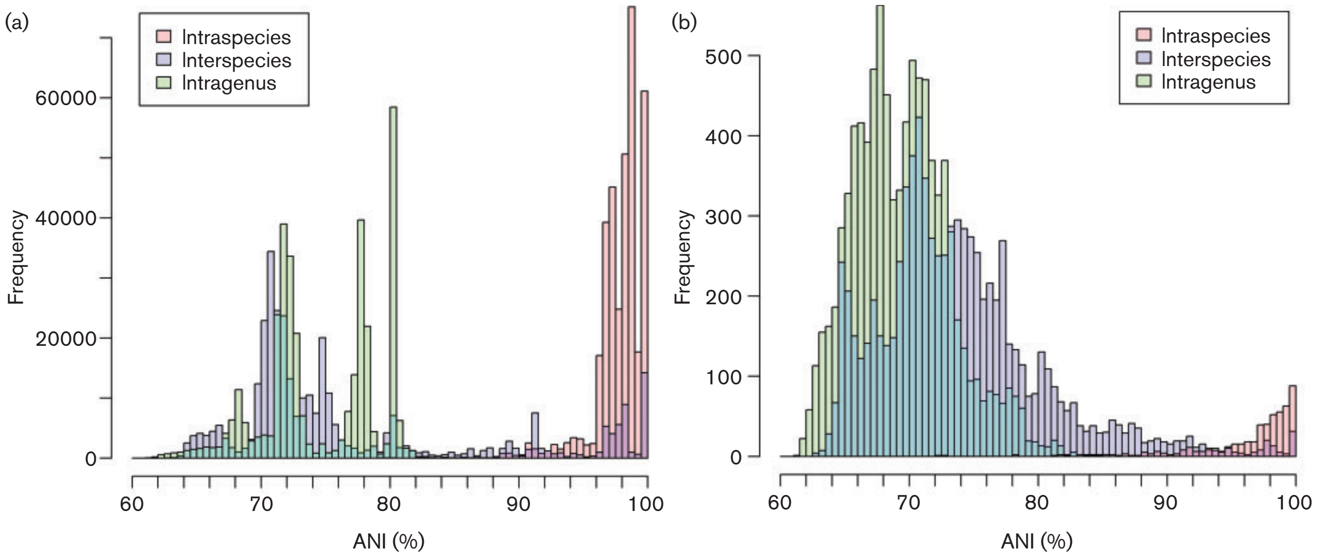


Fig. 1. ANI distribution based on all genomes (a) and a normalized set of genomes (b). Each colour represents pairwise ANI calculations between strains belonging to different taxonomic ranks: between two strains belonging to the same species (orange), belonging to different species (green) and belonging to different genera (purple). Mean ANI values per species are represented in (b).

in the distribution of DDH values, which showed a 'hollow' shape around the value of 70 % (Meier-Kolthoff *et al.*, 2013). In contrast to the pattern at the intraspecies level, there was huge variation in ANI values obtained by interspecies and intergenus comparisons. ANI values between different species within a given genus were distributed widely over the range 62–100 %.

Since the number of publicly available genome sequences is unequal among taxonomic groups, we normalized data by obtaining a single averaged ANI value per species to alleviate the bias resulting from uneven taxon sampling of genome sequences. The normalized ANI distribution (Fig. 1b) showed a similar pattern to the one based on all genomes (Fig. 1a), except for substantially lower frequencies observed in the range 96–100 % ANI. This is because much effort has been made in genome sequencing projects related to medically important species. For example, 512 genome sequences for *Escherichia coli* alone were used in this study.

In some cases, genomes showing ANI values higher than 96 % were from different species with validly published names. *Escherichia coli–Shigella* species, *Burkholderia mallei–Burkholderia pseudomallei* and *Bordetella bronchiseptica–Bordetella parapertussis–Bordetella pertussis* are cases that are already known (Richter & Rosselló-Móra, 2009). Through more comprehensive comparisons in this study, we found additional cases, including *Bacillus anthracis–Bacillus thuringiensis–Bacillus cereus*, *Yersinia pseudotuberculosis–Yersinia pestis*, *Brucella* species, *Lactobacillus casei–Lactobacillus paracasei*, *Mycobacterium tuberculosis–Mycobacterium bovis*, *Leptospira kirschneri–Leptospira interrogans* and *Treponema paraluiscuniculi–Treponema pallidum*. These sets of species require further taxonomic investigation to see whether they merit independent taxonomic status.

Determination of the optimal 16S rRNA gene sequence similarity threshold for species demarcation

The overall relationship between ANI and 16S rRNA gene sequence similarity was found to be non-linear (Fig. S1A, available in IJSEM Online), which is in good agreement with a previous study based on a small dataset (n=70) (Konstantinidis & Tiedje, 2005). However, linear correlation could be seen when logarithmic transformation was applied, which is expressed by the equation ln(1.001−16S rRNA gene sequence similarity)=−0.0097ln(ANI)−0.602, with r^2=0.805 and P<0.001 (Fig. S1B). Similarly, a non-linear association was observed between the original DDH and 16S rRNA gene sequence similarity in previous studies, but high correlation coefficients were achieved after log (Devereux *et al.*, 1990) or log–log (Keswani & Whitman, 2001) transformation on both variables.

To find the optimal threshold of 16S rRNA gene sequence similarity corresponding to a level of 95 or 96 % ANI, which has been recommended as a boundary for species demarcation, precision-recall and F score were calculated for all 16S rRNA gene sequence similarities (at 0.01 % intervals) at each of the given ANI cut-offs. Twofold cross-validation revealed that 98.65 % 16S rRNA gene sequence similarity showed the highest F score at both ANI cut-offs (precision=0.922, recall=0.986 at 95 % ANI cut-off; precision=0.912, recall=0.985 at 96 % ANI cut-off) (Fig. 2). This 16S rRNA gene sequence similarity threshold

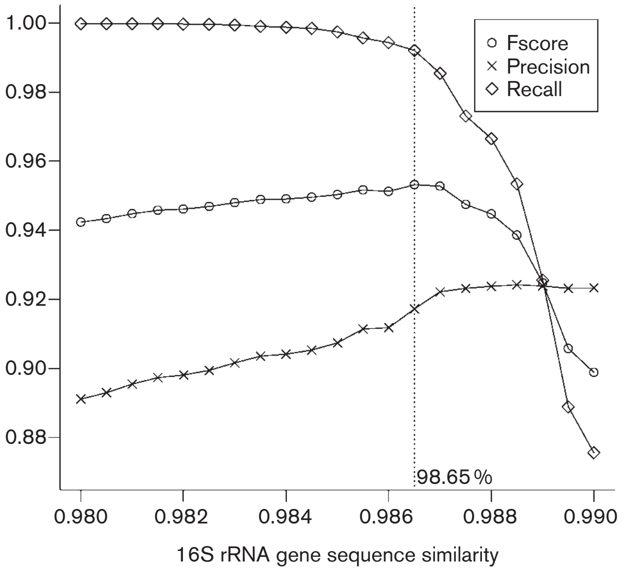


Fig. 2. Precision-recall curve with F measure. When the species boundary was set at 95 % ANI, the highest F score was found at 98.65 % 16S rRNA gene sequence similarity.

level is within the threshold range (98.2–99.0 %) previously suggested on the basis of comparative studies between DDH and 16S rRNA gene sequence similarity (Meier-Kolthoff *et al.*, 2013; Stackebrandt & Ebers, 2006).

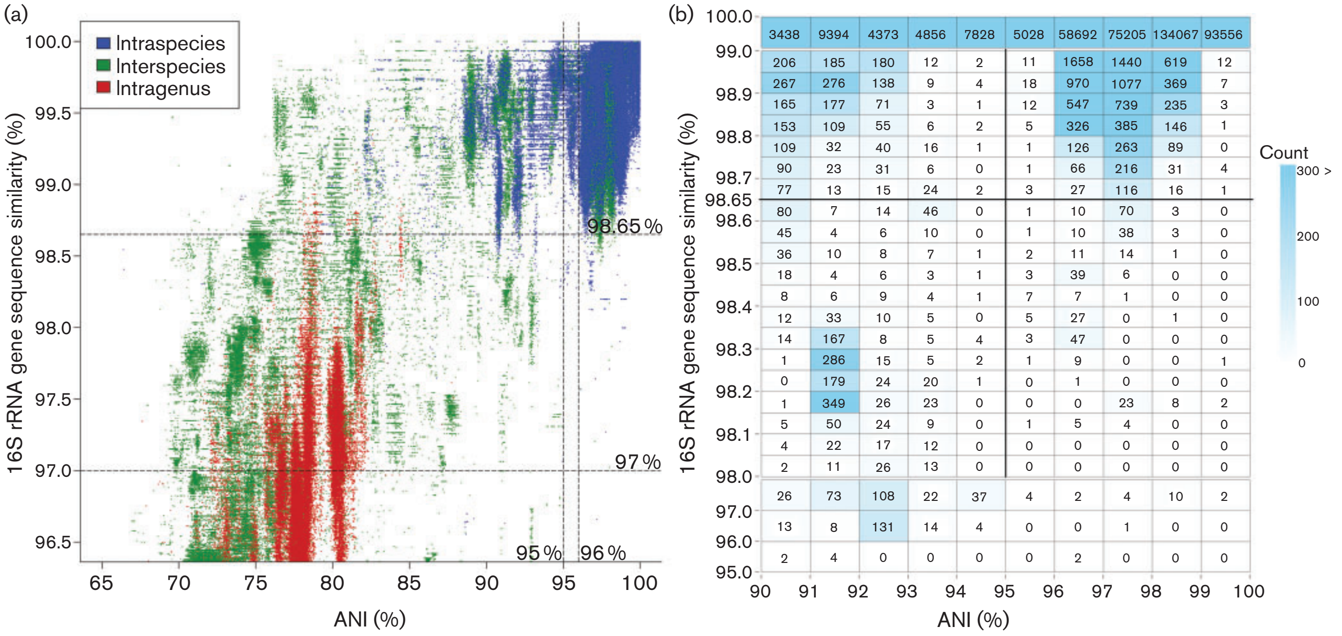
We applied this threshold value to our dataset in order to check its validity for the current taxonomy. The overall distribution of genome pairs with ANI and 16S rRNA gene sequence similarity is summarized in Fig. 3. Given the 16S rRNA gene sequence similarity threshold of <98.65 %, we divided all ANI values into three categories (<95 %, 95–96 % and ⩾96 %) (Fig. 3a). At <95 % ANI (667 705 comparisons), there were only two exceptional cases in which strains of the same species fall into this range, *Fusobacterium nucleatum* subsp. *fusiforme* ATCC 51190^T and *F. nucleatum* subsp. *animalis* ATCC 51191^T (with 16S rRNA gene sequence similarity/ANI of 98.51/91.80 %) and *F. nucleatum* subsp. *animalis* ATCC 51191^T and *F. nucleatum* subsp. *nucleatum* ATCC 25586^T (with 16S rRNA gene sequence similarity/ANI of 98.56/90.64 %).

At a range of 95–96 % ANI, 0.53 % (27 of 5104) showed <98.65 % 16S rRNA gene sequence similarity (see Table S1 for details). Among the cases with ⩾96 % ANI, only 0.10 % (n=356) of the total cases (n=371 370) showed <98.65 % 16S rRNA gene sequence similarity; details of these exceptional cases are given in Table S2. Since these are exceptional cases in which the 98.65 % 16S rRNA gene sequence similarity cut-off may not guarantee different genomic species status based on ANI, we went on to examine each case in more detail.

Over half of the exceptional cases resulted from relatively higher levels of intragenomic (interoperon) heterogeneity of 16S rRNA genes. Of 383 exceptional cases with >98.65 % 16S rRNA gene sequence similarity and <95 % ANI, 150 cases were found between strains of the *E. coli*–*Shigella* group. *E. coli* is known to have high intraspecific variation (1.10 %) between multiple 16S rRNA genes in the genome (Pei *et al.*, 2010). Other species (13 cases) also showed a high degree of intragenomic variation among 16S rRNA gene copies, including *Lactobacillus rhamnosus* (0–7.67 %), *Caldanaerobacter subterraneus* (0.03–6.23 %), *Desulfitobacterium hafniense* (0.06–3.73 %), *Bacteroides ovatus* (0.07–3.30 %), *Yersinia enterocolitica* (0–2.67 %) and *Desulfitobacterium dehalogenans* (0–2.14 %). It appears that higher levels of intraspecific divergence of 16S rRNA gene sequences may result in deviation from the optimal 16S rRNA gene sequence similarity threshold, at least in the strains mentioned above. There are some other exceptional cases that show lower levels of intragenomic variation of 16S rRNA gene sequences (<0.1 %). For example, strains of *Aggregatibacter actinomycetemcomitans* (44 cases) showed relatively low 16S rRNA gene sequence similarity (98.00–98.20 %), despite sharing a higher level of ANI (97.43–99.09 %). At present, it is not clear why a relatively higher level of 16S rRNA gene sequence diversity is maintained between these strains. Ecological and physiological traits, together with the evolutionary history of this species, may provide clues to this discrepancy with further investigation.

About 36 % (138 cases) of the exceptional cases were found between the genome sequences of *Neisseria meningitidis* alpha704 (GenBank accession no. CAJS00000000) and other *N. meningitidis* strains with 97.00–98.64 % 16S rRNA gene sequence similarity and 96.68–98.12 % ANI. The quality of the genome sequence of *N. meningitidis* alpha704 is questionable, as it is at the draft stage of sequencing, containing 42 contigs and only one full-length 16S rRNA gene sequence. Normally, *N. meningitidis* strains possess four *rrn* operons. The reason for this exceptional case needs to be explored.

In addition to exceptions at the intraspecies level mentioned above, exceptional cases were also found at the interspecies level. They include *Halomicrobium katesii* DSM 19301^T vs *Halomicrobium mukohataei* DSM 12286^T (96.71 % ANI and 95.59 % 16S rRNA gene sequence similarity), *Thermoanaerobacter* species (95.03–98.16 % ANI and 97.02–98.54 % 16S rRNA gene sequence similarity) and *Caldicellulosiruptor kronotskyensis* 2002^T vs *Caldicellulosiruptor bescii* DSM 6725^T (95.59 % ANI and 97.99 % 16S rRNA gene sequence similarity). Species belonging to *Thermoanaerobacter* and *Caldicellulosiruptor* are known to display higher intercistronic variation among multiple 16S rRNA gene copies (Acinas *et al.*, 2004; Pei *et al.*, 2010). *Halomicrobium*, a halophilic archaeal genus, was also reported to show a higher level of intraspecies divergence (9.0–9.8 %) of 16S rRNA gene sequences (Cui *et al.*, 2009). The current classification of these species as separate species requires further attention, given the exceptionally low 16S rRNA gene sequence similarity.



3438	9394	4373	4856	7828	5028	58692	75205	134067	93556
206	185	180	12	2	11	1658	1440	619	12
267	276	138	9	4	18	970	1077	369	7
165	177	71	3	1	12	547	739	235	3
153	109	55	6	2	5	326	385	146	1
109	32	40	16	1	1	126	263	89	0
90	23	31	6	0	1	66	216	31	4
77	13	15	24	2	3	27	116	16	1
80	7	14	46	0	1	10	70	3	0
45	4	6	10	0	1	10	38	3	0
36	10	8	7	1	2	11	14	1	0
18	4	6	3	1	3	39	6	0	0
8	6	9	4	1	7	7	1	0	0
12	33	10	5	0	5	27	0	1	0
14	167	8	5	4	3	47	0	0	0
1	286	15	5	2	1	9	0	0	1
0	179	24	20	1	0	1	0	0	0
1	349	26	23	0	0	0	23	8	2
5	50	24	9	0	1	5	4	0	0
4	22	17	12	0	0	0	0	0	0
2	11	26	13	0	0	0	0	0	0
26	73	108	22	37	4	2	4	10	2
13	8	131	14	4	0	0	1	0	0
2	4	0	0	0	0	2	0	0	0

Fig. 3. Association plot (a) and table (b) between ANI values and 16S rRNA gene sequence similarities. The number of strain pairs is displayed in each category square when divided by several intervals in 16S rRNA gene sequence similarity and 1.0 in ANI.

Overall discussion and conclusions

It has been estimated that there are millions of species of prokaryotes on Earth, of which only a fraction has been formally recognized. Currently, the number of species of prokaryotes with validly published names is only about 11 000. The role of the 16S rRNA gene sequence similarity threshold in recognizing novel species has been crucial since its introduction in the taxonomy of prokaryotes. The originally proposed threshold value of 97 % (Stackebrandt & Goebel, 1994) has been challenged by higher thresholds (Meier-Kolthoff *et al.*, 2013; Stackebrandt & Ebers, 2006). However, these studies are based on a limited number, i.e. hundreds, of DDH and 16S rRNA gene sequence similarity comparisons. Moreover, the sampling of DDH values used is biased, as most recently published DDH values were obtained from publications describing novel species, in which only DDH values less than 70 % were considered. In this study, we carried out over a million comparisons between genome and 16S rRNA gene sequence similarities, which do not suffer from this bias.

Using ANI as a substitute for DDH, our proposed threshold of 98.65 % would greatly speed up the process of recognizing novel species. However, this threshold should be applied carefully under the following conditions. In this study, we used 16S rRNA gene sequences extracted from whole-genome data, which are based on multiple depths of sequencing coverage; in general, 7 × for Sanger and >20 × for next-generation sequencing methods are applied. Therefore, sequences used in 16S rRNA gene sequence similarity calculation must be as accurate as possible. As a minimum, (i) both strands should be sequenced (equivalent to >2 × depths) and (ii) 16S rRNA gene sequences should be complete, as used in this study. Here, we used the full-length sequence between the two most popular PCR primers, 27f (5′-AGAGTTTGATCMTGGCTCAG-3′) and 1492r (5′-TACCTTGTTACGACTT-3′). In addition, it needs to be pointed out that the new threshold of 16S rRNA gene sequence similarity should not be applied directly to microbial community studies based on next-generation sequencing, as this new technology has native errors. In conclusion, the new threshold of 16S rRNA gene sequence similarity for defining species of prokaryotes proposed here should accelerate the rate of discovery of novel species in a more objective manner, especially when combined with genome-based relatedness methods such as ANI.

ACKNOWLEDGEMENTS

We thank Jenny Tan for editing the manuscript. This work was supported by the National Research Foundation of the Republic of Korea (grants 2013-035122 and 2012M3A9D1054622).

REFERENCES

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* **186**, 2629–2635.

Ash, C., Farrow, J. A. E., Dorsch, M., Stackebrandt, E. & Collins, M. D. (1991). Comparative analysis of *Bacillus anthracis*, *Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int J Syst Bacteriol* **41**, 343–346.

Chan, J. Z. M., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. (2012). Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol* **12**, 302.

Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K. & Lim, Y. W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int J Syst Evol Microbiol* **57**, 2259–2261.

Cui, H. L., Zhou, P. J., Oren, A. & Liu, S. J. (2009). Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, *Haloarcula* and *Halomicrobium*. *Extremophiles* **13**, 31–37.

Deloger, M., El Karoui, M. & Petit, M. A. (2009). A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* **191**, 91–99.

Devereux, R., He, S. H., Doyle, C. L., Orkland, S., Stahl, D. A., LeGall, J. & Whitman, W. B. (1990). Diversity and origin of *Desulfovibrio* species: phylogenetic definition of a family. *J Bacteriol* **172**, 3609–3619.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P. & other authors (2005). Opinion: re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**, 733–739.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91.

Grim, C. J., Kotewicz, M. L., Power, K. A., Gopinath, G., Franco, A. A., Jarvis, K. G., Yan, Q. Q., Jackson, S. A., Sathyamoorthy, V. & other authors (2013). Pan-genome analysis of the emerging foodborne pathogen *Cronobacter* spp. suggests a species-level bidirectional divergence driven by niche adaptation. *BMC Genomics* **14**, 366.

Haley, B. J., Grim, C. J., Hasan, N. A., Choi, S. Y., Chun, J., Brettin, T. S., Bruce, D. C., Challacombe, J. F., Detter, J. C. & other authors (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiol* **10**, 154.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* **21**, 2329–2335.

Keswani, J. & Whitman, W. B. (2001). Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. *Int J Syst Evol Microbiol* **51**, 667–678.

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S. C., Jeon, Y. S., Lee, J. H. & other authors (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* **62**, 716–721.

Konstantinidis, K. T. & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**, 2567–2572.

Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**, 1929–1940.

Lan, Y. M., Wang, Q., Cole, J. R. & Rosen, G. L. (2012). Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS ONE* **7**, e32491.

Lee, J. H., Yi, H. & Chun, J. (2011). rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J Microbiol* **49**, 689–691.

McCarthy, B. J. & Bolton, E. T. (1963). An approach to the measurement of genetic relatedness among organisms. *Proc Natl Acad Sci U S A* **50**, 156–164.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**, 610–618.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P. & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nat Methods* **10**, 881–884.

Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., Jin, Z., Lee, P., Yang, L. & other authors (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol* **76**, 3886–3897.

Read, J., Pfahringer, B., Holmes, G. & Frank, E. (2011). Classifier chains for multi-label classification. *Mach Learn* **85**, 333–359.

Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126–19131.

Rosselló-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Schildkraut, C. L., Marmur, J. & Doty, P. (1961). The formation of hybrid DNA molecules and their use in studies of DNA homologies. *J Mol Biol* **3**, 595–617.

Stackebrandt, E. & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **33**, 152–155.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A. D., Kämpfer, P., Maiden, M. C., Nesme, X., Rosselló-Mora, R., Swings, J. & other authors (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**, 1043–1047.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W. & Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* **60**, 249–266.

van Rijsbergen, C. J. (1979). *Information Retrieval*, 2nd edn. London: Butterworths.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E. & other authors (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Yi, H., Cho, Y. J., Yoon, S. H., Park, S. C. & Chun, J. (2012). Comparative genomics of *Neisseria weaveri* clarifies the taxonomy of this species and identifies genetic determinants that may be associated with virulence. *FEMS Microbiol Lett* **328**, 100–105.