J. Gen. Appl. Microbiol., **60**, 75−78 (2014)
doi 10.2323/jgam.60.75

jGAM

Full Paper

Whole-genome sequence comparison as a method for improving bacterial species definition

(Received November 12, 2013; Accepted January 15, 2014)

Wen Zhang,[1,2] Pengcheng Du,[1,2] Han Zheng,[1] Weiwen Yu,[1,2] Li Wan,[1] and Chen Chen[1,2,*]

[1] *National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control, Beijing 102206, China*
[2] *Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease, Hangzhou 310003, China*

We compared pairs of 1,226 bacterial strains with whole genome sequences and calculated their average nucleotide identity (ANI) between genomes to determine whether whole genome comparison can be directly used for bacterial species definition. We found that genome comparisons of two bacterial strains from the same species (SGC) have a significantly higher ANI than those of two strains from different species (DGC), and that the ANI between the query and the reference genomes can be used to determine whether two genomes come from the same species. Bacterial species definition based on ANI with a cut-off value of 0.92 matched well (81.5%) with the current bacterial species definition. The ANI value was shown to be consistent with the standard for traditional bacterial species definition, and it could be used in bacterial taxonomy for species definition. A new bioinformatics program (ANItools) was also provided in this study for users to obtain the ANI value of any two bacterial genome pairs (http://genome.bioinfo-icdc.org/). This program can match a query strain to all bacterial genomes, and identify the highest ANI value of the strain at the species, genus and family levels respectively, providing valuable insights for species definition.

Key words: ANI; bacteria; definition; genome; species

Introduction

Prokaryotes are the original inhabitants of this planet. Both Archaea and Bacteria evolved somewhere more than 3 billion years ago. Prokaryotes diversified widely throughout their long, long time of existence. The characterization of a strain is a key element in prokaryote systematic study (Rappe and Giovannoni, 2003). Early classification of prokaryotes was based solely on phenotypic similarities (Tindall et al., 2010), but modern prokaryote characterization has been strongly influenced by advances in genetic methods. Various classification methodologies based on nucleotide sequences, such as 16S ribosomal rRNA sequence comparison (Fox et al., 1977), have been developed in recent years (Rappe and Giovannoni, 2003; Moore et al., 2010).

The 16S rRNA sequence-based division into higher taxa is currently the most widely used classification system for prokaryotes. However, the sequence of 16S rRNA is too conserved to distinguish between closely related species (＞97% similarity) (Rosselló-Mora and Amann, 2001; Richter and Rossello-Mora, 2009; Tindall et al., 2010). The other most commonly used method is multiple-locus sequence typing (MLST) (Maiden et al., 1998). In this method, several housekeeping genes common to all strains are selected and sequenced. Phylogenetic relationship of the strains is derived from the sequence alignments of the selected housekeeping genes. MLST has the same resolution problem. The effectiveness of using the phylogenetic relationship based on a small number of loci to represent the phylogenies of the whole genome is yet to be determined. In addition, the comparability of results derived from different sets of genes and criteria for choosing good phylogenetic markers remains unknown (Konstantinidis et al., 2006). There is no gold standard as to which and how many genes should be used in MLST for different species. Finally, sequencing errors, even a single nucleotide sequencing error, can have a significant effect on the outcome of MLST.

With the fast development of sequencing technology, the availability of an ever greater number of genomic sequences per species offers the possibility for developing distance determinations based on whole-genome information

*Corresponding author: Dr. Chen Chen, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, China CDC, State Key Laboratory for Infectious Disease Prevention and Control, P.O. Box 5, Changping Beijing, China 102206. Tel/Fax: (+86) 10 61739459 E-mail: chenchen@icdc.cn

None of the authors of this manuscript has any financial or personal relationship with other people or organizations that could inappropriately influence their work.

(Deloger et al., 2009). Average nucleotide identity (ANI) among the conserved genes of a pair of genomes was first described by Konstantinidis and Tiedje (2005). By calculating the ANI of 70 pairs of related species, Konstantinidis and Tiedje found that ANI is a robust means to compare genetic relatedness among strains, and ANI values of ~94% corresponded to the traditional 70% DNA-DNA reassociation standard of the current species definition (Konstantinidis and Tiedje, 2005; Konstantinidis et al., 2006; Goris et al., 2007). Several programs are available for calculating the ANI; for example, JSpecies can be found at the website: http://www.imedea.uib.es/jspecies/ (Richter and Rossello-Mora, 2009). In this study, we calculated ANI values of any two genomes from 1,226 bacterial strains, 18 times more than those in the previous study. Our findings showed that the species classification based on ANI is in good agreement with the bacterial taxonomy from the National Center for Biotechnology Information (NCBI). Our study further showed that most bacterial strains within the same species have an ANI value higher than 0.92, thus demonstrating that ANI is a robust means for defining both cultured and uncultured bacterial species.

Materials and Methods

All available genomic sequences of 1,226 bacterial strains from 466 genera were downloaded from the database of the National Center for Biotechnology Information (NCBI: ftp://ftp.ncbi.nih.gov/genomes). For whole-genome sequence comparison between a pair of genomes, Blast similarity searches were performed using local BLAST software (Altschul et al., 1990). Based on the Blast results, ANI was calculated for each pair of genomes using the following protocol. First, coding sequences (CDSs) from the query genome were searched against the reference genome. Alignments with a BLAST match of at least 60% of the overall sequence identity and an alignable region more than 70% of their length were kept, whereas the other CDSs were considered genome-specific and filtered out (Konstantinidis and Tiedje, 2005). Second, genome comparisons with a total alignable region less than 50% of the query genome length were also filtered out. Third, for genes with multiple alignments, the alignment with the highest identical sites was kept for ANI calculation. ANItools, a new program written in PERL, was used to calculate the ANI of each pair of compared genomes. Genome comparisons with both the query and the reference genomes from the same species are referred to as SGC, whereas genome comparisons with the query and the reference genomes from different species of a genus are referred as DGC in this study. The total ANI values of 2730 SGC and 1729 DGC were calculated using ANItools. The genomes used in this study and the ANIs of SGC and DGC are summarized in Tables S1 and S2.

Results

ANI of 1,226 bacteria strains

Using ANItools, we compared the genome sequences and calculated the ANI values of any two genomes among 1,226 bacterial strains in this study. ANI value can be used to represent the evolutionary distance between two strains. Our results showed that two strains with a closer evolutionary relationship also have a higher ANI value. The average ANI for two strains in the same species is 0.936, which is larger than that for two strains in the same genus (0.836), and also larger than that for two strains in the same family (0.789). ANI values of strain pairs in the same species (SGC) are usually higher than those of strain pairs from different species in a genus (DGC), which is shown in Table S1, S2 and Fig. 1. Take *Escherichia coli* as an example. Among 1,226 bacterial strains, there are 30 *Escherichia coli* strains. The average ANI of 30 *E. coli* strains was 0.943, which is significantly higher than average ANI of *E. coli* strains compared with other bacterial strains in *Escherichia genus*, such as *Escherichia fergusonii*. The average ANI between *E. coli* and *Escherichia fergusonii* strains was only 0.872.

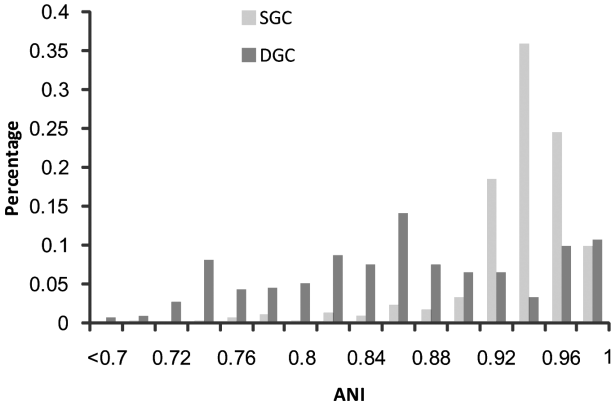

Fig. 1. Distribution of ANI values for 2730 SGC (Blue column) and 1729 DGC (Red column).

Genome comparisons with both the query and the reference genomes from the same species are referred to as SGCs, whereas other genome comparisons with the query and the reference genomes from different species of a genus are called DGCs in this study.

Can we use ANI to identify Bacteria?

Konstantinidis et al. first calculated average nucleotide identity among conserved genes between pairs of genomes selected from 70 related species and discovered that ANI is a robust means to compare genetic relatedness among strains, and that ANI values of ~94% corresponded to the traditional 70% DNA-DNA reassociation standard of the current species definition (Konstantinidis and Tiedje, 2005). In this study, we developed a new bioinformatics method to calculate the ANI of any two genomes in 1,226 bacterial strains and found that only 70.2% of bacterial strains within the same species have ANI values higher than 0.94 (Fig. 1). The accuracy rate is only 72.6% when the ANI cut-off value of 0.94 is used for species identification. A further test of different ANI cut-off values found that the peak accuracy rate (81.5%) was reached at the cut-off value of 0.92, as shown in Fig. 2. Strains in several species, such as *Bacillus amyloliquefaciens* or *Klebsiella pneumonia* (Table S1), can be identified with the cut-off value of 0.92, but not with 0.94. In our study, species definition based on the 0.92 ANI cut-off is the most consistent with the current species definition using traditional standard in NCBI taxonomy.

However, using the ANI cut-off value of 0.92 alone is not sufficient for some species definition in 14 genera of this study, such as *Brucella* sp., *Rickettsia* sp., *Shigella* sp. and *Yersinia* sp. Some DGC strains have ANI values higher than

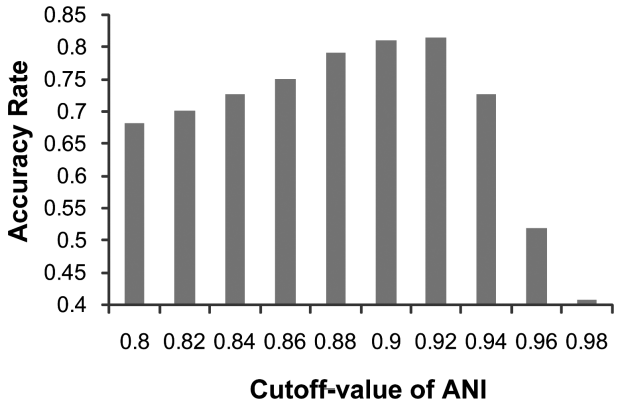

Fig. 2. Accuracy rate of bacterial species definition with different cut-off values for ANI.

0.92, which cannot be correctly distinguished at the species level using an ANI cut-off value of 0.92. All these strains are listed in Table S2 (false positive cases). Some SGC strains have ANI values lower than 0.92, which cannot be correctly assigned to a species based on the ANI cut-off value of 0.92. These species are listed in Table 1 and Table S1 (false negative cases). These strains cannot be directly identified using the ANI cut-off value of 0.92, because the ANI values within a species vary significantly. Table 1 shows that strains of *Leptospira biflexa* have the highest average ANI of SGC (0.998), whereas the lowest average ANI of SGC obtained is that of *Polynucleobacter necessaries* (0.779).

Among 117 species with more than one strain in this study, only 86 species (73.5%) have an average ANI higher than 0.92. The species with significantly lower SGC ANI (<0.92) or higher DGC ANI (>0.92) could not be directly distinguished from other strains by using the ANI cut-off value of 0.92. Definition of strains in these species can be achieved by determining their best-matched species in the bacterial database using ANI. For these strains that cannot be directly defined using an ANI cut-off value of 0.92, we developed a new bioinformatics program (ANItools, http://genome.bioinfo-icdc.org/) for users to calculate the ANI of two strains. This program can match a strain to all bacterial genomes, and identify the highest ANI value of this strain at the species, genus and family level. So for strains that cannot be distinguished using the ANI cut-off value of 0.92, this program can provide valuable information for species definition.

Based on the study of 1,226 bacterial strains, we found that an ANI cut-off value of 0.92 and/or the best match search can be used for the rapid identification of bacterial strains. This whole genome-comparison method can be combined with traditional biochemical and molecular methods to improve the reliability of the techniques used for bacterial species definition, especially for an uncultured bacterial species.

Discussion

We compared the pairs of 1,226 complete genome sequences to determine whether ANI-based species definition is consistent with the current species definition based on the traditional methods. We found that most SGCs have significantly higher ANI values than those of DGCs. Previous results also indicated that ANI values corresponded to the traditional DNA-DNA reassociation and 16S rRNA standard of species definition (Fox et al., 1977; Konstantinidis and Tiedje, 2005). Thus the ANI between the query and the reference genomes can be used as a parameter to determine whether two genomes come from the same species. The ANI values of any two complete genomes can be obtained using ANItools. A draft genome or a partial genome can also be analyzed using ANItools if the query genome has more than 1,500 genes (50% of the average number of genes in bacterial genomes). In a simulation test, calculation of an ANI value for a 6.5 Mb query genome and a typical bacterial genome takes about 30 s using one CPU. It takes about 10 CPU h to calculate ANI values comparing the query genome and 1,000 bacterial genomes in the database. Thus, ANItools can be used to compare thousands of prokaryotic genome sequences that are anticipated to become available in the near future.

The results of using the ANI cut-off value of 0.92 for species definition were shown to match the predefined species taxonomy better than those using the cut-off value of 0.94. However, some traditional species definitions based on phenotypic classification can be different from species definitions based on the modern technology of genetic sequence analysis. For example, *Escherichia* vs. *Shigella* bacteria, which are traditionally classified into different genera for historical reasons related to clinical diagnosis, share $>94\%$ ANI among their genomic sequences. Some *Escherichia* genomes are more divergent among themselves than compared with *Shigella* genomes. Groups of strains sharing $>94\%$ ANI are rarely classified into different

Table 1. Top 10 and Bottom 10 of Average ANI of SGCs among 124 bacterial species.

Top 10			Bottom 10		
Species	Strain No.	Average ANI	Species	Strain No.	Average ANI
Leptospira biflexa	4	0.998	*Polynucleobacter necessarius*	2	0.779
Treponema pallidum	2	0.995	*Enterobacter cloacae*	2	0.787
Brucella abortus	4	0.993	*Pseudomonas fluorescens*	3	0.790
Yersinia pestis	8	0.991	*Ralstonia eutropha*	4	0.797
Streptococcus suis	5	0.989	*Dickeya dadantii*	3	0.820
Synechococcus elongatus	2	0.989	*Rhodopseudomonas palustris*	6	0.824
Chlamydia trachomatis	6	0.988	*Mycobacterium bovis*	3	0.829
Rickettsia rickettsii	2	0.988	*Prochlorococcus marinus*	12	0.838
Brucella melitensis	6	0.987	*Wolbachia endosymbiont*	3	0.838
Brucella suis	4	0.986	*Mycobacterium tuberculosis*	5	0.839

ANI is the average nucleotide identity between two genomes, and SGC is a genome comparison between two strains within the same species.

species nowadays (for new species). The best ANI cut-off value that fits the traditional or current designation remains yet to be determined.

ANI, as the new method for bacterial species definition, provides several benefits unavailable in other methods. First, the use of ANI for bacterial species definition can directly identify the species difference at the nucleotide level. The early classification of prokaryotes was based solely on phenotypic similarities and chemical characteristics (Tindall et al., 2010). However, the chemical characteristics and phenotype of bacteria species are to some extent affected by environmental factors, such as temperature and pH, which can cause possible biases. For example, two isolates of one species living in different environments may have different phenotypes or chemical characteristics, causing misplacement in bacterial species definition. If bacteria species definition is based on genome difference, this misplacement due to environmental factors can be effectively avoided. Compared with other technologies also based on nucleotide sequences, such as 16S rRNA and MLST, ANI analysis based on whole genome comparison between two stains has higher resolution and can avoid the bias caused by sequence selection. Even two closely related bacterial species can be distinguished based on their DNA divergence at the genomic level, and one or a few sequencing errors can be easily adjusted with the help of the depth coverage of sequence reads.

The other useful feature of ANI analysis is for the definition of uncultured bacterial species. To date, only about half of the phyla of bacteria have species that can be grown in the laboratory (Rappe and Giovannoni, 2003) and more than 99% of the bacterial species on Earth are uncultured. Without the clone of a strain, any traditional taxonomy methodology based on phenotype/chemical characteristic is not applicable. Without strain clones, we cannot get sufficient DNA for species definition using 16S rRNA or MLST. With advances in single-cell genome sequencing technology (Kalisky and Quake, 2011), the whole genome sequence of a single cell of an uncultured bacteria strain can be obtained without any culture or clone. The taxonomy of this unknown uncultured bacterial strain could be characterized based on the ANI of its genome sequence compared with those of other bacterial strains. Then the species definition of this novel bacterial strain could be inferred based on its ANI values with other genomes. Thus, ANI as a powerful species definition tool can be used to identify the abundant resources of microbiology on Earth. For bacteria that can grow in the laboratory, this ANI technology is time-efficient, bypassing the culture and cloning steps. Therefore, ANI as a new method for species definition can be used for rapid identification of the target pathogen during possible future epidemics.

To date, ANI, as the new bacterial definition method, may only be a supplemental to other methods. However, the explosion of genome data and rapid development of sequencing technology provide an opportunity for the ANI technology to be widely accepted and used in prokaryotes classification.

Availability

The ANItools software is free and available at http://genome.bioinfo-icdc.org.

Authors' Contribution

WZ performed the genome analysis and wrote the manuscript. CC designed the project and edited the manuscript. All authors edited the manuscript. WY developed the website. The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 812111251, no. 81201322 and no. 81301402), and the Priority Project on Infectious Disease Control and Prevention (2013ZX10003006-002 and 2013ZX10004-221).

Supplementary Materials

Table S1. List of ANI of genome comparisons of two strains in the same species (SGC).

Strains without underlines can be directly defined by ANI method using the cut-off value of 0.92, while others can not. For these strains marked by underlines, they could be defined on species, genus and family level by finding a companion strain with the highest score of ANI in our bacterial genome database.

Table S2. List of ANI of genome comparisons of strains from different species in the same genus (DGC).

Strains without underlines can be directly defined by ANI method using cut-off value 0.92, while others can not. For these strains marked by underlines, they could be defined on species, genus and family level by finding a companion strain with the highest score of ANI in our bacterial genome database.

Supplementary tables are available in our J-STAGE site (http://www.jstage.jst.go.jp/browse/jgam).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403−410.

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

Fox, G. E., Pechman, K. R., and Woese, C. R. (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int. J. Syst. Bacteriol.,* **27**, 44−57.

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

Kalisky, T. and Quake, S. R. (2011) Single-cell genomics. *Nat. Meth.,* **8**, 311−314.

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl. Environ. Microbiol.,* **72**, 7286−7293.

Konstantinidis, K. T. and Tiedje, J. M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA,* **102**, 2567−2572.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E. et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA,* **95**, 3140−3145.

Moore, E. R. B., Mihaylova, S. A., Vandamme, P., Krichevsky, M. I., and Dijkshoorn, L. (2010) Microbial systematics and taxonomy: relevance for a microbial commons. *Res. Microbiol.,* **161**, 430−438.

Rappe, M. S. and Giovannoni, S. J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.,* **57**, 369−394.

Richter, M. and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA,* **106**, 19126−19131.

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

Tindall, B. J., Rossello-Mora, R., Busse, H. J., Ludwig, W., and Kampfer, P. (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.,* **60**, 249−266.