DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–5
doi: 10.1093/database/baw084
Database tool

Database tool

ANItools web: a web tool for fast genome comparison within multiple bacterial strains

Na Han[1,2], Yujun Qiang[1,2] and Wen Zhang[1,2,*]

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

*__Corresponding author:__ Tel: (86)-010-61739446; Email: zhangwen@icdc.cn

Received 25 November 2015; Revised 5 April 2016; Accepted 28 April 2016

Abstract

Background: Early classification of prokaryotes was based solely on phenotypic similarities, but modern prokaryote characterization has been strongly influenced by advances in genetic methods. With the fast development of the sequencing technology, the ever increasing number of genomic sequences per species offers the possibility for developing distance determinations based on whole-genome information. The average nucleotide identity (ANI), calculated from pair-wise comparisons of all sequences shared between two given strains, has been proposed as the new metrics for bacterial species definition and classification.
Results: In this study, we developed the web version of ANItools (http://ani.mypathogen.cn/), which helps users directly get ANI values from online sources. A database covering ANI values of any two strains in a genus was also included (2773 strains, 1487 species and 668 genera). Importantly, ANItools web can automatically run genome comparison between the input genomic sequence and data sequences (Genus and Species levels), and generate a graphical report for ANI calculation results.
Conclusion: ANItools web is useful for defining the relationship between bacterial strains, further contributing to the classification and identification of bacterial species using genome data.

Database URL: http://ani.mypathogen.cn/

Background

Rapid and accurate classification of bacterial isolates is the most important task in medical microbiology, especially during infectious disease outbreaks with national or global spreading threat (1). However, the current classification methods all have shortcomings at the resolution level (2), not only the methods based on phenotypic similarities and chemical characteristics, but also modern genetic methods based on fragment nucleotide sequences (16S and multilocus sequence typing [MLST]) (3–5). The molecular structure of 16S rRNA is too conserved to distinguish between closely related species (>97% similarity) (6–8). Additionally, early classification of prokaryotes was based solely on phenotypic similarities and chemical characteristics, which are to some extent affected by environmental factors, such as temperature and pH, which can cause possible biases (4). Classification using 16S rRNA and MLST methods could be also biased by one or more sequencing errors.

Most recently, the average nucleotide identity (ANI), calculated from pair-wise comparisons of all sequences shared between any two strains, has been proposed as the new metrics for bacterial species definition and classification. In 2005, Pro. Konstantinidis firstly assessed 70 related species and found ANI of the shared genes between two strains to be a robust means for comparing genetic relatedness among strains; ANI values of ~94% were shown to correspond to the traditional 70% DNA–DNA reassociation standards of the current species definition (9–11). In 2012, using 38 strains in the genus *Acinetobacter* as a test case, Chan further proved that ANI results are congruent with the core genome phylogeny and traditional approaches, and also compatible with the existing taxonomy (12). In our previous work (2), we calculated and listed the precise ANI values of any two genome comparisons in 1226 bacterial strains, indicating that species classification based on ANI is in excellent agreement with the NCBI's bacterial taxonomy. This work proved ANI to be useful for bacterial taxonomy, representing a powerful candidate method for the definition for existing as well as novel bacterial species (2).

Comparing with other methods, ANI analysis based on whole-genome comparison between two strains has higher resolution and can avoid the bias caused by sequence selection and errors. 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 depth coverage of sequence reads (2). Besides ANItools, the other program is available for ANI value calculation (JSpecies) (6,13). However, the use of our previous version ANItools still requires the installation as well as that of several appended programs (such as BLAST and Hmmer) on personal computer, in addition to parameter adjustments. Additionally, no ANI database is available currently to the public for thousands of bacterial genomes. Although JSpeciesWS (13) also support a web version for calculating ANI values between several bacterial strains, the strain number limitation (a maximum of 15 genomes) hinder the possibility to get the ANI matrix on genus or species level, and there is also no phylogenetic tree result to graphically show the relationship among strains in the same genus or species. Besides ANItools, there is another ANI value calculation program available named JSpecies (6,13). However, the both tools are not as perfect as we expect. ANItools still needs to be installed locally in personal computer or sever, which certain Add-In programs like BLAST and Hmmer, are also essential for in the meanwhile, always accompanied by the related parameters set up or adjustment works. That means not so friendly to the users who has no background about IT or Bioinformatics. When it comes to JSpeciesWS (13), the first tool can be used online to calculate ANI, doesn't require any kind of parameter set up or adjustment works. But due to the limited capacities (maximum 15) of strain number in comparison, users have no chance to get the ANI matrix on genus or species level, when they are required to analyze too many strains in the meantime. Moreover, there is no phylogenetic tree result either to graphically show the relationship among strains in the same genus or species.

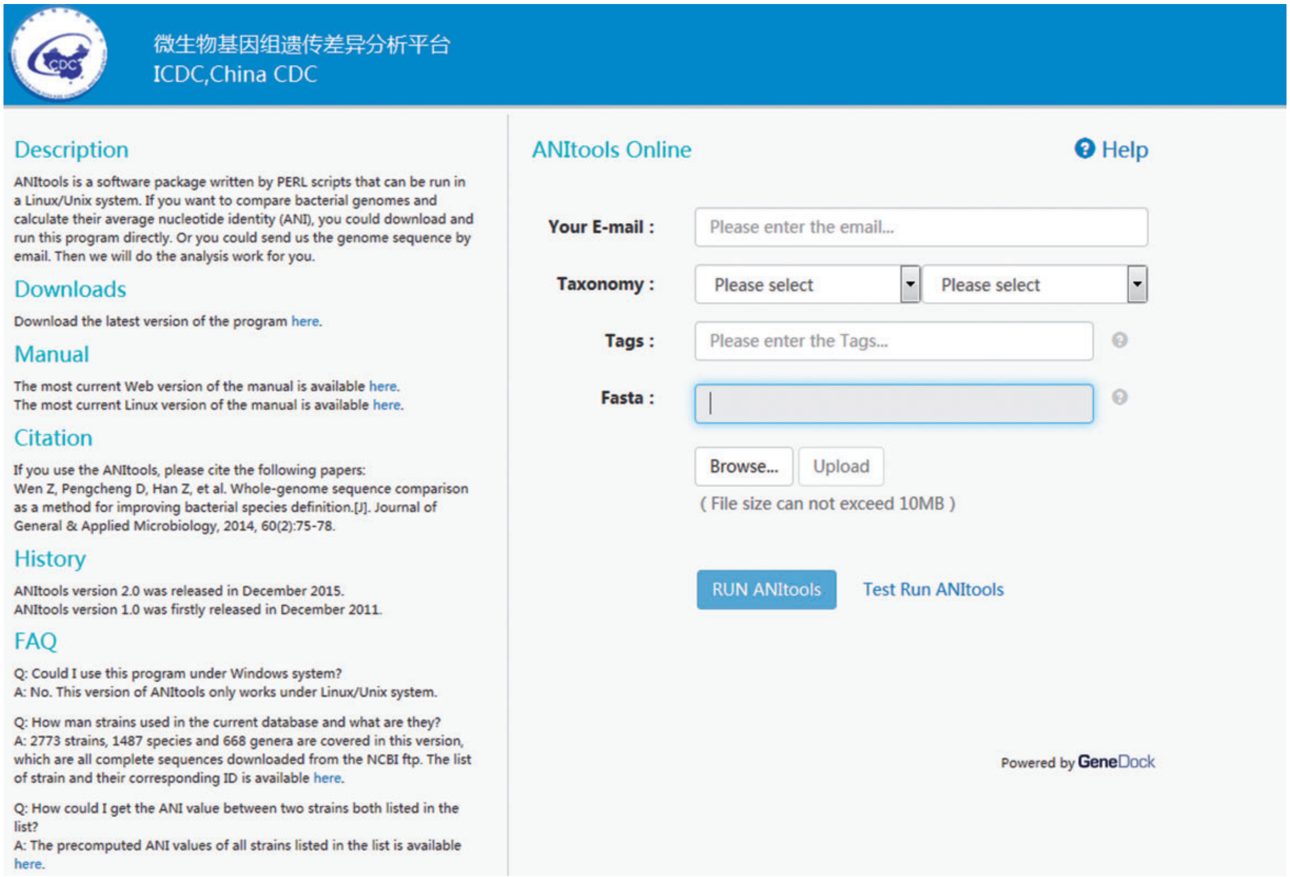
Therefore, we finally programmed web version of ANItools 2.0 (http://ani.mypathogen.cn/) to get rid of all disadvantages of current tools in line with the conclusion above. ANItools web version helps users directly obtain ANI values online and increases the number of genomes examined comparing to previous Linux version. A database covering ANI values of any two strains in a genus was included in this database (2773 strains, 1487 species and 668 genera). ANItools web is useful to define the relationship of bacterial strains, and helpful for the classification and identification of bacterial species using genome data. Compared with currently available software, ANItools web reduces users' involvement to a minimum level: only genomic sequence uploading and genus data selection are required. It can automatically run genome comparison between the input genomic and data sequences, and generate a graphical report for ANI calculation results.

Implementation

ANItools web was built around two public programs, Glimmer 3 (14) and ANItools (2). The website interface is written in Java. ANItools web can analyze nucleotide sequence in 'strict' FASTA format (a first line with a sequence identifier preceded by '>', followed by a second line with the sequence).

The analysis process consists of the following steps:

1. Gene prediction using Glimmer 3 (14) for the query nucleotide sequence. The parameters for the software used for CDS prediction Glimmer are -o50 -g110 -t30.
2. Acquisition of an ANI value matrix from the ANI database based on the species or genus name selected by the user.
3. Comparison of all predicted gene sequences of the query sequence with the target genome sequences using BLASTN. Target genomes are nucleotide sequences of



微生物基因组遗传差异分析平台
ICDC,China CDC

Description

ANItools is a software package written by PERL scripts that can be run in a Linux/Unix system. If you want to compare bacterial genomes and calculate their average nucleotide identity (ANI), you could download and run this program directly. Or you could send us the genome sequence by email. Then we will do the analysis work for you.

Downloads

Download the latest version of the program here.

Manual

The most current Web version of the manual is available here.
The most current Linux version of the manual is available here.

Citation

If you use the ANItools, please cite the following papers:
Wen Z, Pengcheng D, Han Z, et al. Whole-genome sequence comparison as a method for improving bacterial species definition.[J]. Journal of General & Applied Microbiology, 2014, 60(2):75-78.

History

ANItools version 2.0 was released in December 2015.
ANItools version 1.0 was firstly released in December 2011.

FAQ

Q: Could I use this program under Windows system?
A: No. This version of ANItools only works under Linux/Unix system.

Q: How man strains used in the current database and what are they?
A: 2773 strains, 1487 species and 668 genera are covered in this version, which are all complete sequences downloaded from the NCBI ftp. The list of strain and their corresponding ID is available here.

Q: How could I get the ANI value between two strains both listed in the list?
A: The precomputed ANI values of all strains listed in the list is available here.

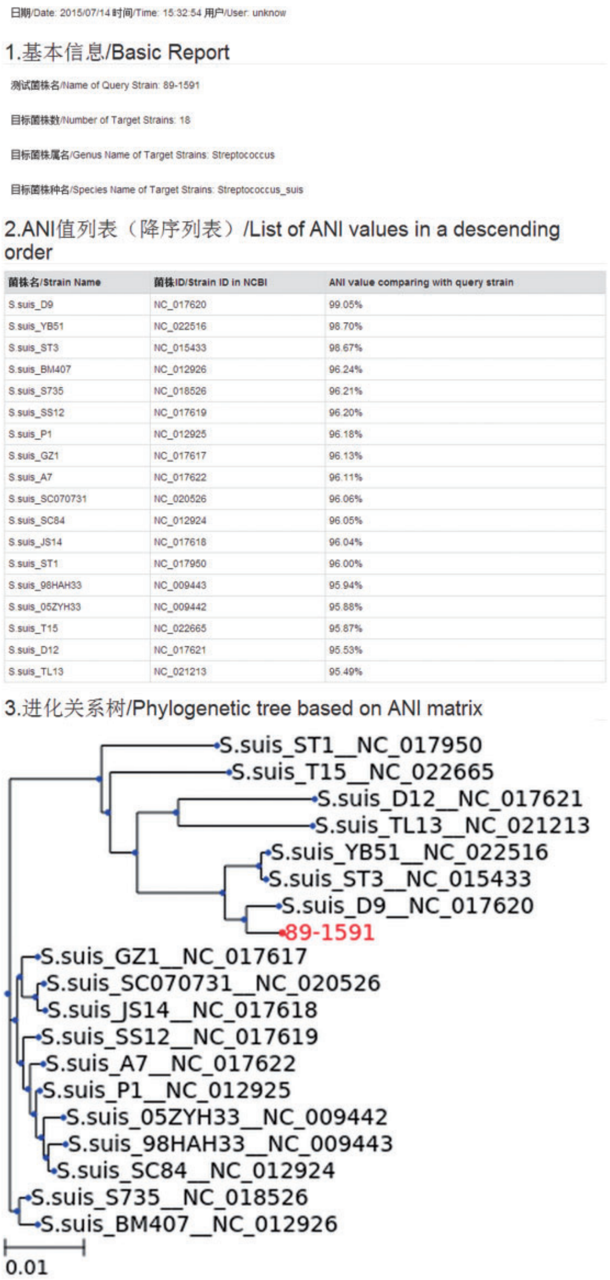
Figure 1. Interface of the input page.

bacterial species in a genus (if user-defined genus) or species (if defined species) from the genome database. The current genome database covers 2773 strains, 1487 species and 668 genera. The genome sequences of 2773 bacterial strains from 668 genus were downloaded from the database of National Center for Biotechnology Information (NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). All these sequences are complete genome and will be updated once every 3 months.

4. Based on BLAST results, ANI was calculated between the query sequence and each target genome. First, CDSs from the query genome were searched against the reference genome. With a BLAST match of at least 60% overall sequence identity and an alignable region >70% of their length, these alignments could be kept, and the remaining CDSs considered to be genomic specific and filtered out (9). Second, genome comparisons with total alignable region <50% of the query genome length should also be filtered out. Third, for genes with multi-alignments, only alignments with highest identical sites should be kept.
5. Acquisition of a new ANI matrix combining new ANI results and covering the query sequences and target genomes in the data. Using Trex 3 (15,16), the matrix was converted to a phylogenetic tree which represents the evolutionary relationship of the query strain with the target genomes.

Results and Discussion

We have developed a web-based computational method to quickly compare bacterial strains. The use of traditional biochemistry methods and 16S sometimes only allows species distinction at the genus level. With the help of ANItools, users can obtain a list of ANI values between the query strain and every strain in the same genus, and identify the best match. Based on a large scale survey in our previous study (2), ANI values of strain pairs in the same species are usually higher than those of strain pairs from different species in a genus. Thus, the list of ANI values in the report page is useful to users in classifying previous undefined bacterial species at the genome level, combining with the next-generation sequencing (NGS) technology. Using *Streptococcus* as a model, users select ‘*Streptococcus*’ and ‘*Streptococcus suis*’ in the Taxonomy list, then upload a genome sequence in the input page (Figure 1) and click ‘Run ANItools’. Several minutes later (5–20 min), a report page is displayed (Figure 2). As shown in Figure 1, a strain sampled



日期/Date: 2015/07/14 时间/Time: 15:32:54 用户/User: unknow

1.基本信息/Basic Report

测试菌株名/Name of Query Strain: 89-1591

目标菌株数/Number of Target Strains: 18

目标菌株属名/Genus Name of Target Strains: Streptococcus

目标菌株种名/Species Name of Target Strains: Streptococcus_suis

2.ANI值列表（降序列表）/List of ANI values in a descending order

菌株名/Strain Name	菌株ID/Strain ID in NCBI	ANI value comparing with query strain
S.suis_D9	NC_017620	99.05%
S.suis_YB51	NC_022516	98.70%
S.suis_ST3	NC_015433	98.67%
S.suis_BM407	NC_012926	96.24%
S.suis_S735	NC_018526	96.21%
S.suis_SS12	NC_017619	96.20%
S.suis_P1	NC_012925	96.18%
S.suis_GZ1	NC_017617	96.13%
S.suis_A7	NC_017622	96.11%
S.suis_SC070731	NC_020526	96.06%
S.suis_SC84	NC_012924	96.05%
S.suis_JS14	NC_017618	96.04%
S.suis_ST1	NC_017950	96.00%
S.suis_98HAH33	NC_009443	95.94%
S.suis_05ZYH33	NC_009442	95.88%
S.suis_T15	NC_022665	95.87%
S.suis_D12	NC_017621	95.53%
S.suis_TL13	NC_021213	95.49%

3.进化关系树/Phylogenetic tree based on ANI matrix

Figure 2. Interface of the report page. All information are shown in two languages: Chinese and English.

from a diseased pig (89–1591) in Canada with serotype 2 showed highest similarity with *S. suis* D9 (NC_017620) with serotype 7. The previous genome typing method also supports this result (1): both strains were in the Minimal Core Genome Group 4 (MCGG4) group.

A phylogenetic tree is also included in the report page, graphically representing the evolutionary relationship among bacterial strains; it is helpful for determining the pathogenic strain source in an epidemical outbreak research. Based on our previous research, even in the same pathogenic species, the pathogenicity level is also variable for the strains carrying different pathogenic genes or with variable genotype (1,17–21). Still using *Streptococcus suis* as an example, *S. suis* strains could be divided into seven groups based on minimal core genome, and Minimal Core Genome Group 1 (MCGG1) strains had higher virulence compared with those in other groups (1). Similar genetic differences within bacterial strains are also shown in our ANItools, which has the fastest calculation rate (~10 min for result generation).

To protect the privacy of the users, the uploaded sequence and analysis results will not be kept in our database. The genome sequences in this ANI database will be updated once every 3 months for users to get more information in time.

In the current version of ANItools, the analysis is restricted to the genus or the species that users choose. And the reference genome found in elsewhere or users sequenced by themselves could not be analyzed neither. We will upgrade the ANItools as soon as possible to address these limitations in next version.

Conclusions

To facilitate effective and fast genome comparison among bacterial strains, we have developed ANItools web, which is accessible at a website (http://ani.mypathogen.cn/). Website stability was tested by online website tools (http://www.websitepulse.com). For users interested in using ANItools on their own computer, an installation package for ANItools is also available for download.

Currently, ANItools web is being used to compare bacterial strains at the genus and species levels. This will provide further clues to define bacterial strain at the genome level and graphically represent the complex relationship among strains, which is helpful for finding a cluster of strains with high similarity (candidate pathogen strains causing an outbreak) in an epidemic study.

Availability and Requirements
Project Name: ANItools web.
Project home page: http://ani.mypathogen.cn/.
Operating system(s): Platform independent.
Programming language: Java.
Other requirements: Java 1.3.1 or higher.
License: GNU GPL.
Any restrictions to use by non-academics: License needed.

Authors' Contribution

N.H. has made contributions to acquisition of data, analysis and interpretation of data; Y.Q. have been involved in the drafting the manuscript and W.Z. contribute to the design, and write the manuscript.

Funding

National Natural Science Foundation of China (no. 81301402), 863 Project 2014AA021505, 2013ZX10004221 and 2013ZX10004-101-002.

Conflict of interest. None declared.

References

1. Chen,C., Zhang,W., Zheng,H. *et al*. (2013) Minimum core genome sequence typing of bacterial pathogens: a unified approach for clinical and public health microbiology. *J. Clin. Microbiol*., 51, 2582–2591.
2. Zhang,W., Pengcheng,D., Han,Z. *et al*. (2014) Whole-genome sequence comparison as a method for improving bacterial species definition. *J. Gen. Appl. Microbiol*., 60, 75–78.
3. Fox,G.E., Pechman,K.R. and Woese,C.R. (1977) Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. *Int. J. Syst. Bacteriol*., 27, 44–57.
4. Tindall,B.J., Rossello-Mora,R., Busse,H.J. *et al*. (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol*., 60, 249–266.
5. Moore,E.R.B., Mihaylova,S.A., Vandamme,P. *et al*. (2010) Microbial systematics and taxonomy: relevance for a microbial commons. *Res. Microbiol*., 161, 430–438.
6. 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.
7. Stackebrandt,E. and 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.
8. Rosselló-Mora,R. and Amann,R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev*., 25, 39–67.
9. 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.
10. Goris,J., Konstantinidis,K.T., Klappenbach,J.A. *et al*. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*., 57, 81–91.
11. Konstantinidis,K.T., Ramette,A. and Tiedje,J.M. (2006) The bacterial species definition in the genomic era. *Phil. Trans. R. Soc. B Biol. Sci*., 361, 1929–1940.
12. Chan,J.Z.-M., Halachev,M.R., Loman,N.J. *et al*. (2012) Defining bacterial species in the genomic era: insights from the genus Acinetobacter. *BMC Microbiol*., 12, 302.
13. Richter,M., Rossello-Mora,R., Oliver Glockner,F. *et al*. (2015) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*, 32, 1367–4811 (Electronic).
14. Delcher,A.L., Bratke,K.A., Powers,E.C. *et al*. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23, 1367–4811. (Electronic).
15. Makarenkov,V. and Lapointe,F.J. (2004) A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20, 2113–2121.
16. Boc,A., Diallo,A.B. and Makarenkov,V. (2012) T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res*., 40, 1362–4962 (Electronic).
17. Zhang,W., Rong,C., Chen,C. *et al*. (2012) Type-IVC secretion system: a novel subclass of type IV secretion system (T4SS) common existing in gram-positive genus *Streptococcus*. *PLoS One*, 7, e46390.
18. Li,M. (2011) GI-type T4SS-mediated horizontal transfer of the 89K pathogenicity island in epidemic *Streptococcus suis* serotype 2. *Mol. Microbiol*., 79, 1670–1683.
19. Du,P., Cao,B., Wang,J. *et al*. (2014) Sequence variation in tcdA and tcdB of *Clostridium difficile*: ST37 with truncated tcdA is a potential epidemic strain in China. *J. Clin. Microbiol*., 52, 3264–3270.
20. Jing,H., Chen,C., Zheng,X. *et al*. (2011) Identification of genes and genomic islands correlated with high pathogenicity in *Streptococcus suis* using whole genome tiling microarrays. *PLoS One*, 6, 589.
21. Jiang,Y., Liu,H., Wang,H. *et al*. (2013) Polymorphism of antigen MPT64 in *Mycobacterium tuberculosis* strains. *J. Clin. Microbiol*., 51, 1558–1562.