Journal of Microbiology (2021) Vol. 59, No. 5, pp. 476–480
DOI 10.1007/s12275-021-1154-0
eISSN 1976-3794
pISSN 1225-8873

Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity[§]

Dongwook Kim[†], Sein Park[†], and Jongsik Chun[*]

Interdisciplinary Program in Bioinformatics, Institute of Molecular Biology & Genetics, School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

(Received Mar 22, 2021 / Revised Apr 13, 2021 / Accepted Apr 15, 2021)

The average amino acid identity (AAI) is an index of pairwise genomic relatedness, and multiple studies have proposed its application in prokaryotic taxonomy and related disciplines. AAI demonstrates better resolution in elucidating taxonomic structure beyond the species rank when compared with average nucleotide identity (ANI), which is a standard criterion in species delineation. However, an efficient and easy-to-use computational tool for AAI calculation in large-scale taxonomic studies is not yet available. Here, we introduce a bioinformatic pipeline, named EzAAI, which allows for rapid and accurate AAI calculation in prokaryote sequences. The EzAAI tool is based on the MMSeqs2 program and computes AAI values almost identical to those generated by the standard BLAST algorithm with significant improvements in the speed of these evaluations. Our pipeline also provides a function for hierarchical clustering to create dendrograms, which is an essential part of any taxonomic study. EzAAI is available for download as a standalone JAVA program at http://leb.snu.ac.kr/ezaai.

***Keywords*:** average amino acid identity, comparative genomics, phylogeny, software suite

Introduction

Genomics has provided a reliable and objective approach for the classification and identification of bacteria and archaea during prokaryotic taxonomy type evaluations (Chun *et al.*, 2018). Advances in next-generation sequencing (NGS) technology have facilitated the exponential growth of the overall amount of prokaryotic genome sequence data in public databases often overcoming the time-consuming, expensive, and labor-intensive process of conventional Sanger sequencing. This massive quantity of data has made it possible to use genome sequences and direct comparisons in microbial taxonomy, even in general laboratories (Chun and Rainey, 2014).

Given this it is unsurprising that we have developed several indices for evaluating the overall genome relatedness of sequences, often referred to as the overall genome relatedness indices (OGRIs). These indices allow researchers to evaluate the similarity between two genome sequences without the need to employ DNA-DNA hybridization (DDH), which has been regarded as the gold standard for prokaryote species delineation and which indirectly measures genome similarity (Wayne *et al.*, 1987). Although there are several OGRIs 16S rRNA gene sequence similarity and average nucleotide identity (ANI) are the ones most commonly applied to identify novel species (Chun *et al.*, 2018). This is likely the result of their high correlation with DDH (Konstantinidis and Tiedje, 2005a; Goris *et al.*, 2007) and the availability of several improved algorithms for their evaluation (Richter and Rosselló-Móra, 2009; Lee *et al.*, 2016; Richter *et al.*, 2016; Yoon *et al.*, 2017b).

Despite the robustness of species level delineation using ANI, it has been reported that ANI shows poor resolution at higher taxonomic levels (Qin *et al.*, 2014). On the other hand, the average amino acid identity (AAI) has been shown to be a useful measurement for genus delineation (Konstantinidis and Tiedje, 2005b). Several studies have developed boundaries that delimit genera, such as *Chryseobacterium* (Nicholson *et al.*, 2020), *Prochlorococcus* (Walter *et al.*, 2017), and *Lactobacillus* (Zheng *et al.*, 2020) based on the discontinuous distribution of their AAI values. To expand this strategy to other genera it is necessary to develop an easy-to-use, high-performance software tool designed to calculate AAI values for large-scale studies.

For the AAI calculation, BLAST program (Altschul *et al.*, 1997) or other packages based on BLAST algorithm, such as the enveomics collection (Rodriguez-R and Konstantinidis, 2016), were used to identify homologous sequences and compute the identity. However, it is not appropriate to perform the BLAST algorithm to calculate AAI values for a large dataset because of the computational speed limitation. Therefore, to expand this AAI-based delineation strategy to the diverse suprageneric taxa, it is necessary to develop an easy-to-use, high-performance software tool for large-scale studies.

Here, we introduce EzAAI, a suite of workflows for improved AAI calculation when compared to the standard BLAST based algorithm (Altschul *et al.*, 1997), which includes a hierarchical clustering analysis tool. MMseqs2 (Steinegger and Söding, 2017) was used for the fast calculation while achieving a sensitive protein sequence search. We went on to evaluate the accuracy and throughput of the EzAAI work-

[†]These authors contributed equally to this work.
*For correspondence. E-mail: jchun@snu.ac.kr; Tel.: +82-2-880-8153; Fax: +82-2-874-1206
[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

flow by comparing its performance when investigating a set of quality-controlled bacterial genome pairs. Our results suggest the potential utility of our novel suite as a toolkit for large-scale studies of prokaryotic taxonomy using AAI.

Materials and Methods

Pipeline implementation

We developed a novel pipeline named EzAAI, which significantly enhances the throughput of the AAI calculation process while maintaining the accuracy associated with the BLAST procedure (Altschul *et al.*, 1997). The EzAAI pipeline can be separated into three modules: *extraction*, *calculation*, and *clustering* (Fig. 1). In the extraction module, EzAAI predicts and extracts coding sequences from a prokaryotic genome using the Prodigal program (Hyatt *et al.*, 2010). These gene sequences are then converted into the database format used in the MMSeqs2 program (Steinegger and Söding, 2017). The *calculation* module speeds up the evaluation process by using MMSeqs2 to create a reciprocal hit profile. MMSeqs2 compares the predicted amino acid sequences extracted from a pair of prokaryotic genomes and searches for a sequence pair with given amino acid identity and length coverage from both directions. Default values for these parameters are given as 40% identity and 50% coverage, which have been utilized by the previous study (Nicholson *et al.*, 2020); However, since AAI values are dependent on these parameters, we have provided the options (-id, -cov) for a flexible calibration. EzAAI then stores these valid pairs in Seqalign (Text ASN.1) format and automatically obtains the AAI value from a pair by calculating the mean identities in each pair of amino acid sequences, which we refer to as AAIm (AAI calculated using MMSeqs2). Based on this process, EzAAI produces a matrix of AAIm values from a given set of prokaryotic taxa in a tab-separated text format and optional Matrix Market format (-mtx). Finally, EzAAI uses its integral hierarchical clustering module, which implements the unweighted pair group method with arithmetic mean (UPGMA) method, to produce a matrix with pairwise AAI values. In this way our pipeline simplifies the AAI process and supports the large-scale calculation of AAI matrices and UPGMA dendrograms, which are then provided as a Newick format file.

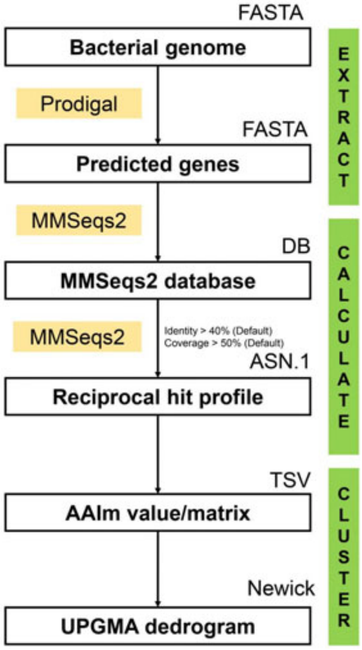

Fig. 1. EzAAI Workflow. Each of the file formats is written in the top right of the box, indicating the file processing steps. The external programs are displayed in yellow. The names of the corresponding module of each step were indicated in the green boxes on the right.

Benchmarks

To benchmark EzAAI, we compared the accuracy and throughput of its core computational algorithm, MMSeqs2, against the standard AAI computation algorithm, BLASTp+ (AAIb, AAI calculated with BLASTp+) (Altschul *et al.*, 1997). Pairs of quality-controlled whole-genome sequences were selected from the EzBioCloud database (http://www.ezbiocloud.net) (Yoon *et al.*, 2017a), including 1,347,262 pairs from 1,642 type strain bacterial genomes under Actinobacteria phylum and 63,190 pairs from 356 archaeal type strain genomes. For both algorithms, AAI values and wall-clock time consumption were then evaluated for the entirety of the pairs. We measured the computational benchmark under same thresholds of 40% identity and 50% coverage, and identical conditions using a single core from computers with an Intel Core i7-4790 3.60 GHz processor. The average nucleotide identity (ANI) values on a set of 8,385 bacterial pairs from 130 genomes under family *Microbacteriaceae* were also calculated using OrthoANIu (Yoon *et al.*, 2017b) for additional comparison.

Measuring taxonomic utility

We devised an index, named Proteome coverage, which represents the degree of overlap between the proteomes of two strains. Proteome coverage from a given pair of genomes is defined as the proportion of reciprocally matched pairs of proteins, which are used to obtain AAI by calculating the mean of their identity values (i.e., pairs exceeding 40% identity and 50% query coverage), on the entire proteome originating from this pair of genome assemblies. This can be more accurately described by saying that the proteome coverage of a given pair consisting of the proteome X_p and Y_p extracted from the genomes X and Y, respectively, can be calculated using the following formula:

$$Coverage(X_p, Y_p) = \frac{Size((X_p \rightarrow Y_p) \cap (Y_p \rightarrow X_p)) \times 2}{Size(X_p) + Size(Y_p)}$$

Where $X_p \rightarrow Y_p$ describes the set of matched proteins from the query proteome X_p onto the reference proteome Y_p, and *vice versa*.

The EzAAI pipeline contains a clustering module that produces a UPGMA dendrogram based on a matrix consisting of all-against-all AAIm values from each set of taxa. To confirm the taxonomic utility of this module, we created dendrograms that show the taxonomic relationships for small sets of taxa belonging to multiple genera in both the Bacteria and Archaea. The dataset containing the type strains of the 29 species under six genera of bacterial family *Microbacteriaceae*

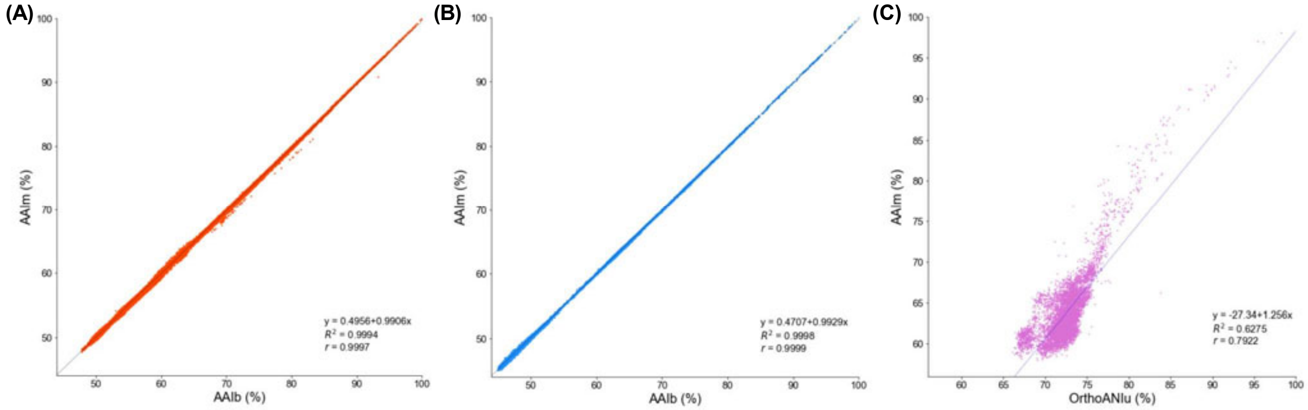

Fig. 2. Correlation between AAIm values and other measures of genomic relatedness. (A) AAI using MMSeqs2 (AAIm) and AAI using BLAST (AAIb) values were obtained using 1,347,262 pairs of bacterial type strain genomes. (B) Comparison of AAIm and AAIb values for 63,190 pairs of archaeal type strain genomes. (C) AAIm and OrthoANIu values for 8,385 bacterial pairs from the *Microbacteriaceae*. Each dataset was visualized as a scatter plot and evaluated using linear regression and Pearson's correlation (r).

and 30 species under archaeal domain, respectively, were randomly selected for the analysis.

Programming language and statistics

The EzAAI pipeline and the benchmarking algorithms were written using JAVA programming language (https://www.java.com/) and executed on a Linux operating system. Statistical analyses and visualization were performed using Python scripts (https://www.python.org/), based on the SciPy (Virtanen *et al.*, 2020) package and matplotlib (Hunter, 2007) package.

Results and Discussion

First, we compared the correlation between AAIm and AAIb values (Fig. 2A and B). The results show that both algorithms produced similar AAI values with a significant degree of correlation ($R^2 > 0.999$, $r > 0.999$) across the entire range of AAI values in both the bacterial and archaeal datasets. These results suggest that the AAIm values from EzAAI have both statistical and taxonomical confidence and can be used to substitute the values calculated using BLASTp+. When we compared the EzAAI values with the ANI (OrthoANIu) values, AAIm tended to report lower identity for genome pairs where the ANI value was below 80% (Fig. 2C). Nevertheless, AAIm and OrthoANIu showed a positive correlation (r = 0.7922), which supports the use of AAIm as an OGRI.

We then estimated the elapsed wall-clock time for the AAI calculation processes using MMSeqs2 and BLASTp+ under identical circumstances (Fig. 3). Statistics were derived by considering the wall-clock time consumed for the AAI calculation between a single pair as a data point. Processes unrelated to AAI calculation (e.g., CDS extraction, data formatting, and visualization) were excluded to allow for an independent comparison. As a result, the mean run-time of the AAI calculations completed using MMSeqs2 (63.1 sec for bacteria; 26.6 sec for archaea) was 3.7–6.9 × faster than that of the standard BLASTp+ (434 sec for bacteria; 98.6 sec for archaea) calculations. The throughput can be improved up to 16× against BLASTp+ when using 30 CPU cores for each calculation (Supplementary data Fig. S1).

We then used the same set of pairs initially applied to benchmark our pipeline to determine the proteome coverage in each dataset using the formula described above. As a result, AAI values and proteome coverages showed a significant correlation (r = 0.87 for bacteria; r = 0.97 for archaea), while the median coverage was 23.4% and 11.3% for bacteria and archaea, respectively (Fig. 4). Although taxonomically distant pairs showed relatively low proteome coverages, principal coordinate analysis using AAI successfully discriminated the taxa belonging to the various phyla, which is not the case of ANI (Supplementary data Fig. S2).

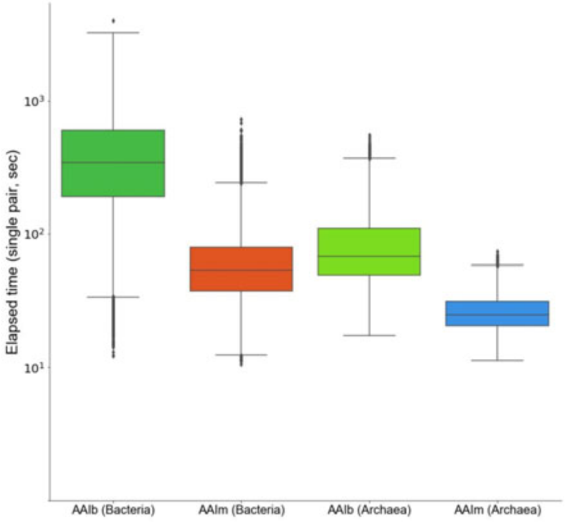

Fig. 3. Performance of BLASTp+ (AAIb) and MMSeqs2 (AAIm) for both bacterial and archaeal AAI calculation tasks. Wall-clock time elapsed to calculate AAI between each pair of bacterial and archaeal genomes were displayed as a box-and-whisker plot. Boxes indicate 1st quartile, median, and 3rd quartile of each dataset, while whiskers denote 1.5 IQR and data points indicate the outliers exceeding 1.5 IQR. Note that the amount of time elapsed is expressed on a log-scale.

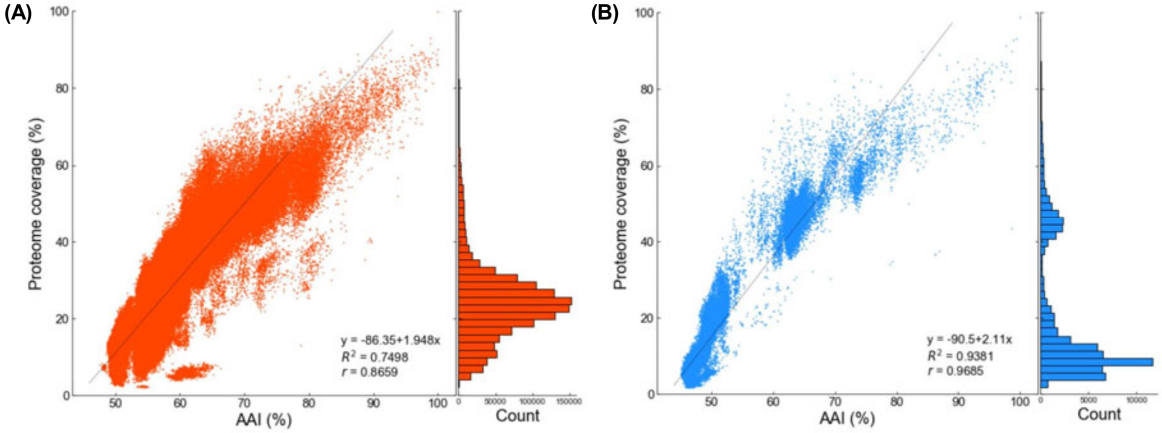

Fig. 4. Proteome coverage of AAIm values from the previously analyzed pairs. (A) Scatter plot of coverage and histogram from the 1,347,262 bacterial pairs described in Fig. 2A. (B) Data from the 63,190 archaeal pairs described in Fig. 2B. Proteome coverage of each pair was plotted along with its AAI value. Histograms binned with the coverage values were drawn on the right to visualize the density of the plot. Linear regression line and statistics along with Pearson's correlation (*r*) has been displayed on the bottom-right.

The utility of the EzAAI pipeline was demonstrated using a dataset containing 29 bacterial taxa belonging to six genera from the *Microbacteriaceae*, an actinobacterial taxon that is well established by chemotaxonomy and molecular systematics (Fig. 5A). Similarly, archaeal genomes were successfully analyzed using EzAAI, and the current classification was recovered in the hierarchical clustering (Fig. 5B). We additionally generated the UPGMA dendrograms driven from a more significant number of genome pairs, which also corresponded to the current taxonomy (Supplementary data Fig. S3).

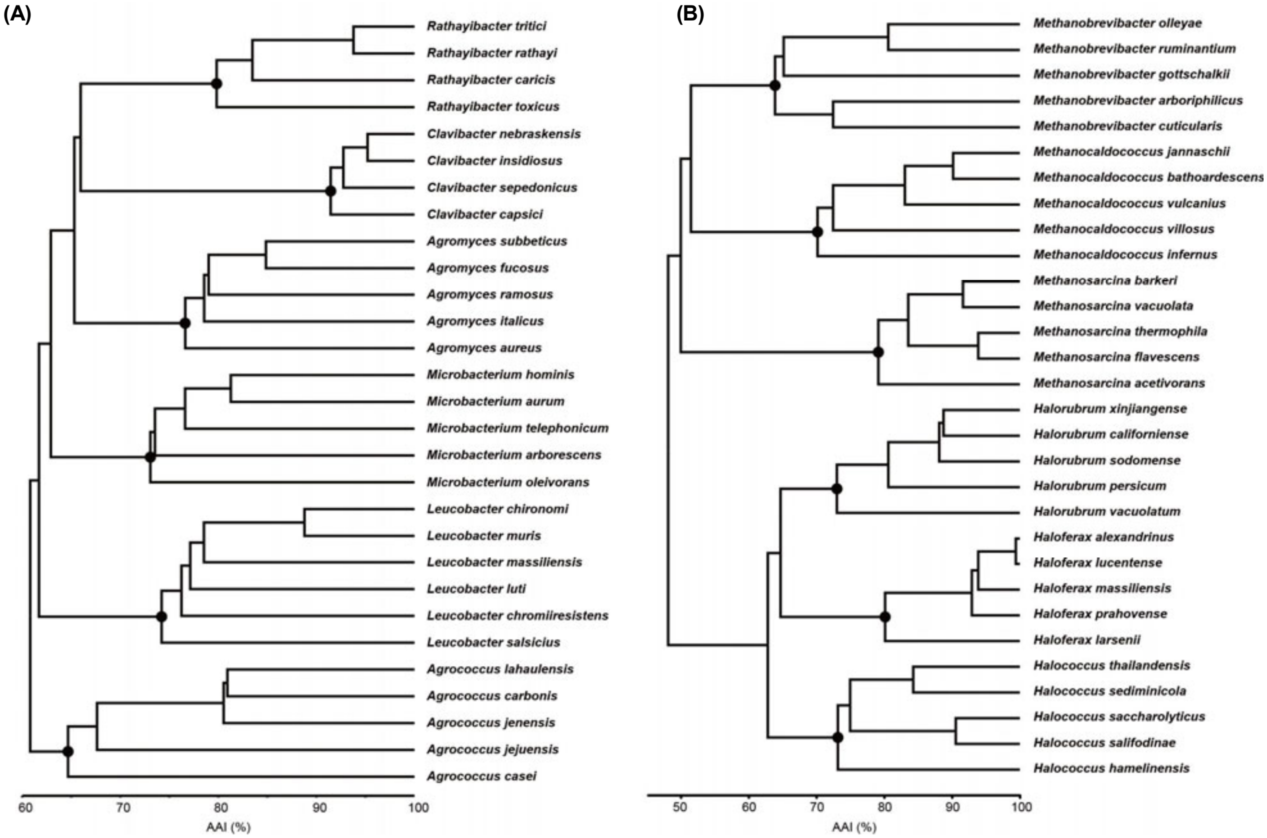

Fig. 5. UPGMA dendrograms from the collection of prokaryotic species, constructed using AAIm values and the EzAAI clustering module. (A) Dendrogram constructed using 29 bacterial species in 6 genera in the family *Microbacteriaceae*. (B) Dendrogram constructed using 30 archaeal species in 6 genera. The scale of the AAI values is shown under each dendrogram. Nodes representing a monophyletic genus are marked with a circular dot.

In conclusion, we present a novel software pipeline (EzAAI) that allows for faster and organized calculation of the AAI values in large-scale taxonomic studies. The easy-to-use interface and its hierarchical clustering functionality should facilitate the use of AAI algorithms in the taxonomy of both Bacteria and Archaea, as well as other microbiological disciplines. A standalone JAVA pipeline for EzAAI is available for download at http://leb.snu.ac.kr/ezaai.

Acknowledgments

This research was supported by the Korean Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) of Korea (grant number 918013-04-4-SB010), and the Collaborative Genome Program for Fostering New Post-Genome Industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science ICT and Future Planning (grant number NRF-2014M3C9A3063541).

Conflict of Interest

The authors have no conflicts of interest to declare.

Open Access

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P., Yi, H., Xu, X.W., De Meyer, S., *et al.* 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **68**, 461–466.

Chun, J. and Rainey, F.A. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int. J. Syst. Evol. Microbiol.* **64**, 316–324.

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

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

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

Konstantinidis, K.T. and Tiedje, J.M. 2005b. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264.

Lee, I., Kim, Y.O., Park, S.C., and Chun, J. 2016. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* **66**, 1100–1103.

Nicholson, A.C., Gulvik, C.A., Whitney, A.M., Humrighouse, B.W., Bell, M.E., Holmes, B., Steigerwalt, A.G., Villarma, A., Sheth, M., Batra, D., *et al.* 2020. Division of the genus *Chryseobacterium*: Observation of discontinuities in amino acid identity values, a possible consequence of major extinction events, guides transfer of nine species to the genus *Epilithonimonas*, eleven species to the genus *Kaistella*, and three species to the genus *Halpernia* gen. nov., with description of *Kaistella daneshvariae* sp. nov. and *Epilithonimonas vandammei* sp. nov. derived from clinical specimens. *Int. J. Syst. Evol. Microbiol.* **70**, 4432–4450.

Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., Oren, A., and Zhang, Y.Z. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.* **196**, 2210–2215.

Richter, M. and Rosselló-Móra, R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA* **106**, 19126–19131.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. 2016. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **32**, 929–931.

Rodriguez-R, L.M. and Konstantinidis, K.T. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *Peer J. Preprints* **4**, e1900v1.

Steinegger, M. and Söding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272.

Walter, J.M., Coutinho, F.H., Dutilh, B.E., Swings, J., Thompson, F.L., and Thompson, C.C. 2017. Ecogenomics and taxonomy of Cyanobacteria phylum. *Front. Microbiol.* **8**, 2132.

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., Stackebrandt, E., *et al.* 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Evol. Microbiol.* **37**, 463–464.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., and Chun, J. 2017a. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617.

Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., and Chun, J. 2017b. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek* **110**, 1281–1286.

Zheng, J., Wittouck, S., Salvetti, E., Franz, C.M.A.P., Harris, H.M.B., Mattarelli, P., O'Toole, P.W., Pot, B., Vandamme, P., Walter, J., *et al.* 2020. A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int. J. Syst. Evol. Microbiol.* **70**, 2782–2858.