INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Palmer *et al.*, *Int. J. Syst. Evol. Microbiol.* 2020;70:2937–2948
DOI 10.1099/ijsem.0.004124

MICROBIOLOGY SOCIETY

All ANIs are not created equal: implications for prokaryotic species boundaries and integration of ANIs into polyphasic taxonomy

Marike Palmer[1,2,*], Emma T. Steenkamp[1], Jochen Blom[3], Brian P. Hedlund[2,4] and Stephanus N. Venter[1,*]

Abstract

In prokaryotic taxonomy, a set of criteria is commonly used to delineate species. These criteria are generally based on cohesion at the phylogenetic, phenotypic and genomic levels. One such criterion shown to have promise in the genomic era is average nucleotide identity (ANI), which provides an average measure of similarity across homologous regions shared by a pair of genomes. However, despite the popularity and relative ease of using this metric, ANI has undergone numerous refinements, with variations in genome fragmentation, homologue detection parameters and search algorithms. To test the robustness of a 95–96% species cut-off range across all the commonly used ANI approaches, seven different methods were used to calculate ANI values for intra- and interspecies datasets representing three classes in the *Proteobacteria*. As a reference point, these methods were all compared to the widely used BLAST-based ANI (i.e. ANIb as implemented in JSpecies), and regression analyses were performed to investigate the correlation of these methods to ANIb with more than 130000 individual data points. From these analyses, it was clear that ANI methods did not provide consistent results regarding the conspecificity of isolates. Most of the methods investigated did not correlate perfectly with ANIb, particularly between 90 and 100% identity, which includes the proposed species boundary. There was also a difference in the correlation of methods for the different taxon sets. Our study thus suggests that the specific approach employed needs to be considered when ANI is used to delineate prokaryotic species. We furthermore suggest that one would first need to determine an appropriate cut-off value for a specific taxon set, based on the intraspecific diversity of that group, before conclusions on conspecificity of isolates can be made, and that the resulting species hypotheses be confirmed with analyses based on evolutionary history as part of the polyphasic approach to taxonomy.

INTRODUCTION

In polyphasic taxonomy of prokaryotes, genomic cohesion is typically informed by similarity measures and metrics such as DNA–DNA hybridization (DDH) and average nucleotide identity (ANI) [1–6]. DDH was developed as a measure of genomic relatedness, where species boundary cut-offs were calibrated using data for known species (70% similarity and less than 5 °C melting temperature differences [1, 7–9]). However, due to the complexities associated with DDH [3, 9–11], and the increasing availability of whole genome sequence information for prokaryotic taxa [2–5, 8–12], a range of sequence-based measures or metrics that correspond to DDH have been proposed. They are generally grouped into overall genome relatedness indices (OGRI [5, 6, 13]) and in addition to ANI include genome-to-genome distances (or *in silico* DDH [4, 5, 14]), maximal unique matches index [4, 5, 12, 15] and tetranucleotide signatures [3, 5].

ANI serves as a useful indicator of the overall relatedness between species [1, 2, 12, 16], because it reflects the mean percentage similarity of shared genomic information between a pair of genomes [4, 6]. This indicator compensates for differences in genome content between different groups, as pair-wise comparisons allow the full complement of shared genomic information to be brought into consideration [2, 12]. Furthermore, by averaging the similarity across all the shared genomic information, it also alleviates issues associated with fast- or slow-evolving genomic regions [12]. As these comparisons are performed in a pair-wise manner, it typically results

Author affiliations: [1]Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa; [2]School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA; [3]Bioinformatics and Systems Biology, Justus-Liebig-University Giessen, Giessen, Germany; [4]Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, NV, USA.
***Correspondence:** Marike Palmer, marike.palmer@unlv.edu; Stephanus N. Venter, fanus.venter@fabi.up.ac.za
Keywords: average nucleotide identity; OGRI; genomics; taxonomy; systematics.
Abbreviations: AML, approximate maximum likelihood; ANI, average nucleotide identity; DDH, DNA–DNA hybridization; GSI, gene support index; NJ, neighbour-joining; OGRI, overall genome relatedness index; UBGC, Up-to-date Bacterial Core Gene.
Supplementary materials are available with the online version of this article.

in two values for each genome pair, where each genome is used as the reference and the query, respectively [1, 2, 12]. These values also differ slightly from one another, as the fragments that are detected in the first calculation are not necessarily the same as the genomic regions detected in the reciprocal comparison, due to the differences in fragmentation across the two genomes.

With the initial description of ANI as a genome similarity metric, the gene set (query) of each genome in a genome pair would be compared with the full sequence of the other genome (the reference). Genes were considered homologous when a gene segment matched with at least 60% identity (averaged across the entire gene) over 70% of the length of the gene in the reference genome [1, 12], using the Basic Local Alignment Search tool (BLAST) [17]. The identity values of these homologous genes were then averaged to obtain an overall value between the genomes [12].

Since its initial description, a number of adjustments have been made to either attempt to remove errors [2, 12] or to speed up the calculation of ANI values [13, 16, 18, 19]. This was done primarily by optimizing the search strategies for finding homologous fragments and the genome fragmentation strategy employed (Table 1). For example, methods such as OrthoANI [13] and Genome Matrix [20] involve the use of optimized search parameters, whereas ANIm [3], OrthoANIu [6], FastANI [19] and gANI [18] all employ algorithms aimed at speeding up the search process. In terms of genome fragmentation, a variety of approaches are available. Both genome pairs are used as the query and reference in reciprocal analyses, with the entire query genome being fragmented [2, 5, 12], which allows for the avoidance of inconsistencies associated with gene prediction [2, 12] and for the inclusion of homologous non-coding regions in the analyses. Examples of such methods include ANIb [3], Genome Matrix [20] and FastANI [19]. Alternatively, both genomes can be segmented simultaneously, either through artificial sectioning or through the use of all protein-coding nucleotide sequences. In the first instance, a single comparative step is required because there are no reciprocal analyses performed, as is done by OrthoANI [13] and OrthoANIu [6], while the latter involves the use of predicted protein-coding genes, as in gANI [18]. Among all of these methods, ANIb, as implemented in JSpecies [3], is the most frequently used, followed by ANIm [3]. Although use of the more recently developed approaches is increasing, the BLAST-based ANI approaches are still most widely employed (Table 1).

In light of the array of vastly different approaches available for calculating a single relatedness metric between pairs of genomes, it is unclear whether inferences drawn using the different ANI methods are comparable and provide the same conclusions regarding the conspecificity of isolates. Therefore, the aim of this study was to investigate the ANI methods currently in the public domain, and to compare them by employing the widely used ANIb (based on BLASTN searches of genomic fragments consisting of 1020 nt) implemented in JSpecies, as a reference point. For this purpose, we used genome sequence information for the diverse and well-sampled genera *Bradyrhizobium*, *Pantoea* and *Paraburkholderia*, which represent members of three well-known classes of *Proteobacteria*. We also evaluated the overall congruence between relationships inferred from the ANIb values with those recovered from phylogenomic analyses based on a marker set of 92 genes for each genus. Because of the wide variety of tools compared and taxonomic range considered in this study, its findings are invaluable for inferring broad-scale conclusions regarding the use of ANI in taxonomy and highlights the importance of using a polyphasic approach for taxon delineation.

METHODS AND MATERIALS

Datasets

All species of *Bradyrhizobium*, *Pantoea* and *Paraburkholderia* with effectively published names and with whole genome sequence data available for the type strains (on 7 August 2019) were included in the analysis. These taxa were identified and selected based on the Genome Taxonomy Database [GTDB v. 04-RS89 [21]; https://gtdb.ecogenomic.org/; accessed 7 August 2019], in order to ensure that only well-circumscribed species with high-quality genome information were used to limit confounding factors such as contamination. To allow intraspecific comparisons, we used all species for which the genomes of at least two additional isolates were available in the public domain. All genome sequences were obtained from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/). Table S1 (available in the online version of the paper) provides strain numbers, NCBI accession numbers and other relevant information.

Phylogenetomic analyses

Phylogenomic analysis based on approximate maximum likelihood (AML [22]) was conducted for each of the respective genera. These phylogenies were rooted based on the most recent and complete phylogenomic hypotheses for each taxon [23–25]. The AML analysis was based on 92 conserved genes as implemented in the Up-to-date Bacterial Core Gene set (UBCG [26]; https://www.ezbiocloud.net/tools/ubcg). Shimodaira–Hasegawa (SH [27, 28]) branch support was inferred from 1000 replicates. Phylogenetic trees were visualized using MEGA-X [29] and edited with Inkscape 0.92.

For confirming conspecificity of isolates used for the calculation of intraspecies ANI values, individual gene trees obtained from the UBCG pipeline was used. In order to identify isolates as belonging to the same species based on evolutionary history, genealogical concordance principles [30] were applied when interpreting gene support index (GSI) values. Based on these principles it is expected that most of the individual gene trees would recover monophyletic groups for species hypotheses.

ANI analyses

ANI values were calculated for the three respective datasets using seven different ANI methods (Table 1). The first programs utilized were the widely used ANIb and ANIm

Table 1. Information regarding the different ANI approaches compared in this study

Type of ANI	Algorithm used	Search strategy	Genome fragmentation	Estimated calculation time (per pair)*	Number of search results*	References	Times cited†
ANIb	BLAST	Each genome in a pair serves as query and reference sequence in separate analyses. The query genome is fragmented while the reference genome is kept intact. Homologous regions are considered to show ≥30% identity over ≥70% alignment length. The identity values normalized over the length of the various homologous regions are then averaged. All fragments ≤1020 bp are used in the analysis.	1020 bp	~20 min	>700	Goris *et al.*, 2007 [2]; Richter and Rosselló-Móra, 2009 [3]‡; Arahal 2014 [12]	2512
OrthoANI (OAT)	BLAST	Both genomes are simultaneously fragmented, and a single BLAST analysis is performed for each genome pair. Homologous regions are identified as the reciprocal best BLAST hits with ≥35% aligned over the total length of the fragment. The ANI is then calculated as the averaged identity values of all homologous fragments between a pair of genomes. Fragments <1020 bp are discarded from the analysis.	1020 bp	~2 min	>200	Lee *et al.*, 2016 [13]‡	675
Genome matrix	BLAST	Each genome in a pair serves as query and reference sequence in respective analyses. The query genome is fragmented while the reference genome is kept intact. Homologous regions are considered to show ≥70% identity over ≥700 bp of the alignment length. The mean identity of the values normalized over the length of the various homologous regions represents ANI. Shorter fragments are discarded from the analysis, as the minimum alignment length used for calculation is ≥700 bp.	1000 bp	Dependent on job queue; online platform	>150	Rodriguez-R and Konstantinidis, 2016 [20]‡; https://github.coc/lmrodriguezr/eenveomic; http://enve-omics.ce.gatech.edu/g-matrix/index	297
ANIm	MUMmer	Sequence alignments are produced using suffix trees from multiple reference or query sequences. For further processing, the data of all alignments against all alignments are used (all-vs-all). For alignment lengths, the distance of the entire region between insertions or deletions producing the best-scoring alignments are used. The ANIm value is determined by subtracting the number of non-identical sites from the alignment length, and the percentage nucleotide identity for each fragment is then determined. ANIm represents the mean of all these fragment identity values. This approach requires fully sequenced genomes for reliable estimates.	No fragmentation required. Fragment length determined based on detection of indels.	~1 min	>700	Richter and Rosselló-Móra, 2009 [3]‡; Arahal 2014 [12]	2512

Continued

Table 1. Continued

Type of ANI	Algorithm used	Search strategy	Genome fragmentation	Estimated calculation time (per pair)*	Number of search results*	References	Times cited†
OrthoANIu (OAU)	USEARCH	This approach is calculated in precisely the same manner as OrthoANI, with the exception of USEARCH being used as search algorithm. USEARCH also employs searching of short sequence stretches; however, weaker matches are not considered as the database sorts the different sequences based on the number of words in common between sequences, allowing exclusion of weaker matches. Shorter fragments are also excluded as in OrthoANI.	1020 bp	~1 min	18	Yoon *et al.*, 2017 [6]‡	590
fastANI	MashMap	A query genome is fragmented and mapped to an intact reference genome through an alignment-free mapping approach based on MinHash searching. Fragments mapping in close proximity to each other are placed in a bin. The percentage identity for reciprocally mapped fragments in each bin is then averaged and an overall mean is determined. Fragments <3000 bp are not included as the minimum read length is 3000 bp.	3000 bp	~5 s	38	Jain *et al.*, 2018 [19]‡	121
gANI	NSimScan	The protein-coding sequences are extracted from each genome in a genome pair. Each set of ORFs are then used as query and reference respectively. NSimScan is used to identify potential homologues. Homology is determined based on bidirectional best hits with ≥70% identity over ≥70% alignment coverage of the shorter sequence. The percentage identities are then multiplied by the alignment lengths of the best hits and summed, divided by the sum of all the alignment lengths.	Protein-coding ORFs (~950 bp)	~2 min	399	Varghese *et al.*, 2015 [18]‡	337

*Based on Google Scholar search results containing first author name, tool name and publication year
†Number of citations as determined on Google Scholar (30 January 2020).
‡Article used for indicating number of citations.

(using nucmer in MUMmer v. 3.23 [31]) as implemented in JSpecies v. 1.2.1 to calculate ANIs for all pair-wise comparisons. Second, OrthoANI (OAT using BLASTN in BLAST 2.7.1+) and OrthoANIu (OAU using USEARCH v. 11.0.667 [32]) were used to calculate pair-wise ANI values for the respective genome sets. As the command-line version of OAT was developed for running single pair-wise analyses, a custom python script was developed (Supplementary data) to calculate the values of all genomes against a single reference genome at a time, in a sequential manner. Furthermore, Genome Matrix (available at http://enve-omics.ce.gatech.edu/g-matrix/index), which also utilizes BLASTN, and FastANI v. 1.1 [19], which uses Mashmap as search algorithm [19, 33], were also employed for the calculation of ANI for pair-wise comparisons.

We also used ANIcalculator v. 1, which calculates gANI from the nucleotide sequences of protein-coding regions in the genomes using NSimScan [18]. For this purpose, the protein-coding genes of all genomes were extracted from NCBI. As the ANIcalculator was developed for single pair-wise analyses, a custom python script (Supplementary data) was developed to run the protein-coding regions of all genomes against a single reference at a time, in a sequential manner.

We also attempted to evaluate how much of a particular genome is taken into account when ANI is calculated. For this purpose, the genus *Paraburkholderia* was targeted and the alignable proportion of genomes used in each pair-wise comparison was noted for the programs that provided an indication thereof. These were ANIb, Genome Matrix, FastANI and gANI. All ANI values obtained with the various approaches were compared to the respective ANIb values for each of the respective sets of genomes. These data points were plotted against the ANIb values and linear regression analyses were performed in Microsoft Excel 365 ProPlus.

For comparisons with the phylogenomic trees, distances between taxa were calculated from ANIb values for the three respective datasets. The distances were used to construct neighbour-joining (NJ) phylogenies using the Neighbor algorithm in Phylip v. 3.69 [34] using a randomized input of sequences. NJ trees were also visualized using MEGA-X and Inkscape.

RESULTS

ANI at the interspecific level

The search and fragmentation strategies of the various ANI approaches differed extensively (Table 1). The strategies employed also caused vast differences in the calculation time of each pair-wise comparison. Calculation time for ANI in these assemblages of bacterial genomes, encompassing medium (*ca*. 5 Mb) to large (*ca*. 9 Mb) genomes, ranged from roughly 5 s (FastANI) to approximately 20 min (ANIb) per genome pair. All values for pair-wise comparisons obtained from the different ANI approaches, as well as the alignable fractions or percentages of *Paraburkholderia* are available in File S1.

For all the genome datasets analysed, more than 130000 values were determined using the seven different ANI approaches examined. For comparative purposes, ANIb, with the suggested species cut-off of 95–96% [3, 12], was used as the basis against which the other approaches were compared (Figs 1 and S1). This was done merely to simplify comparisons as none of the respective methods were considered superior to any other. Overall, it appeared that the various approaches correlated well with ANIb, with R^2 values above 0.9 for all methods used (Figs 1 and S1). However, OrthoANI (based on BLASTN), OrthoANIu (based on USEARCH) and gANI correlated best with the BLAST-based ANIb (R^2=0.9984, 0.9905 and 0.9917, respectively), where the slopes of the linear regression lines approached 1.

A positive relationship was observed between the alignable portion of the *Paraburkholderia* genome and the relatedness of these taxa. In other words, the more closely related two species were (based on ANIb), the higher the shared genomic fraction (Fig. 2). This trend was consistent across all ANI methods evaluated and for all taxa in *Paraburkholderia* analysed. However, between more distantly related species much less of their genomes were alignable. As little as *ca*. 10% of the genome was shared with some methods between taxa with ANIb values at *ca*. 76%. Also, the proportions of the genomes analysed with the different approaches were not consistent across genome pairs, with the biggest fluctuations ranging from 20% to more than 50% between approaches (Fig. 2 and File S1). For example, in comparisons between *Par. acidophila* as reference and *Par. phenoliruptrix* JPY366_1 as query, only *ca*. 16% was alignable using Genome Matrix, while *ca*. 68% was alignable in the FastANI analysis. Overall, the data produced by Genome Matrix consistently showed the smallest portion of the genome being used, while gANI and FastANI consistently showed the largest.

The interspecific relationships based on ANI were congruent with the robust AML phylogenies inferred from the 92 conserved marker genes for the three genera (Figs 3, S2 and S3). In the AML trees, all branches depicting relationships above the species level were well supported based on SH-support and GSI values (Fig. S3). This agreement was particularly evident in both the *Pantoea* and *Paraburkholderia* datasets, where all relationships among species were congruent between the AML and NJ phylogenies. The majority of the relationships among *Bradyrhizobium* species were generally also consistent, with the exception of the placement of the group containing *B. arachidis* and *B. stylosanthis*, and the placement of *B. embrapense* and *B. mercantei*.

ANI at the intraspecific level

The taxa used for intraspecific comparisons were confirmed to belong to single species based on the concordance between the individual gene trees. In the phylogenomic AML trees (Fig. S2), the various groups representing species were supported by high GSI scores, which indicated that these groupings were concordant among most of the single gene trees. For example, among the intraspecies comparisons, the

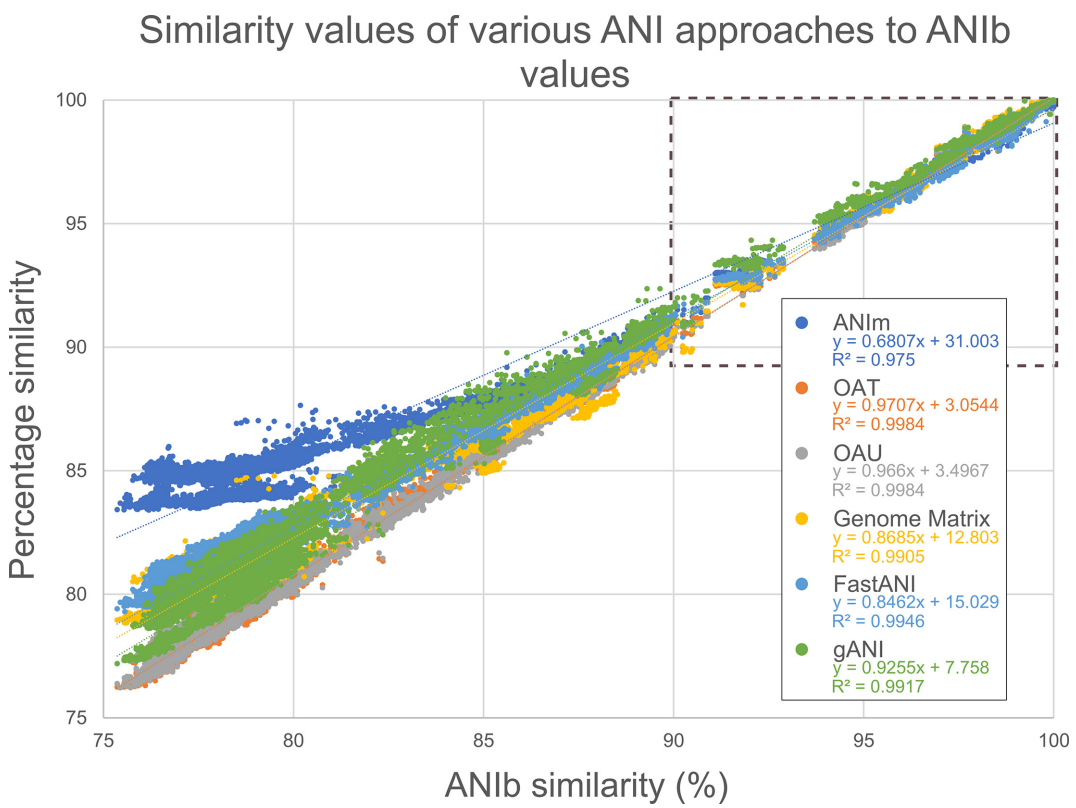

Fig. 1. Comparison of the different methods used for calculating ANI values compared to the widely used BLAST-based approach, ANIb. This graph was generated using the genome sequences for the type strains of all effectively published species of *Bradyrhizobium*, *Pantoea* and *Paraburkholderia*. The *x*-axis indicates the ANIb similarity percentage, while the *y*-axis indicates the percentage similarity for each of the six alternative approaches (for details of the various ANI methods, see Table 1). The equation for the linear regression lines and the R^2-values are indicated in the key under each of the approaches compared. The region indicated by the dotted line represents values near to or at the suggested species boundary and is interrogated further in Fig. 4.

clade representing *Bradyrhizobium elkanii* had the lowest GSI, indicating that it was supported by 78 of the 92 individual gene trees. Conversely, a number of species groups were recovered in all 92 of the single gene phylogenies; these included *B. arachidis*, *B. ottawaense*, *Pan. allii*, *Par. phenoliruptrix*, *Par. caribensis* and *Par. caballeronis*. Furthermore, the species assignment of the isolates used for the intraspecies ANI comparisons also corresponded with those previously suggested and implemented in GTDB [21].

Comparison of ANI values revealed that the ANIb values obtained were always lower than those obtained with the other methods (Table 2). By combining all ANI values for the three different genera, a clear discontinuity in ANIb values (i.e. between ANIb of 92 and 94%) was observed (Fig. 1). To investigate the possibility that this discontinuity is caused by the 'species boundary' thought to exist at the species–population interface ([35]; indicated with dotted block in Fig. 1), we interrogated the individual comparisons of the ANI methods and the different taxon sets (Fig. 4). These data showed that a discontinuity occurs at different levels in the different taxon sets (Fig. 4). In *Bradyrhizobium* this discontinuity occurred at ANIb values from 92.23 to 93.71%, whereas the range for *Pantoea* was between 90.49 and 95.36%, and that for *Paraburkholderia* was between 95.68 and 96.92%. These ANIb ranges were drastically different from those obtained using the other methods, with the biggest differences recovered using gANI (i.e. 93.21–95.07% for *Bradyrhizobium*, 90.92–96.09% for *Pantoea* and 96.47–98% for *Paraburkholderia*). Also, in *Bradyrhizobium*, some interspecies comparisons produced values above this discontinuity (Fig. 4), particularly in *B. japonicum*, whereas in *Paraburkholderia* some intraspecies comparisons were below this discontinuity (Fig. 4), specifically for members of *Par. caribensis* and *Par. caledonica*.

For the intraspecies comparisons, different levels of genomic cohesion, as reflected by ANIb and the other approaches, were observed across the various species investigated (Table 2). For example, for *B. japonicum* isolates ANIb values of 93.96–100% corresponded to FastANI values of 94.80–100%, but for its well-known relative, *B. elkanii*, values of respectively 94.68–100% and 94.91–100% were obtained with the two methods. Additionally, many of the ANI values obtained for intraspecies comparisons, irrespective of the method used, fell below the suggested species cut-off of 95–96% [3, 12].

Differences in the alignable proportion of the genomes for intraspecific comparisons were markedly less prominent

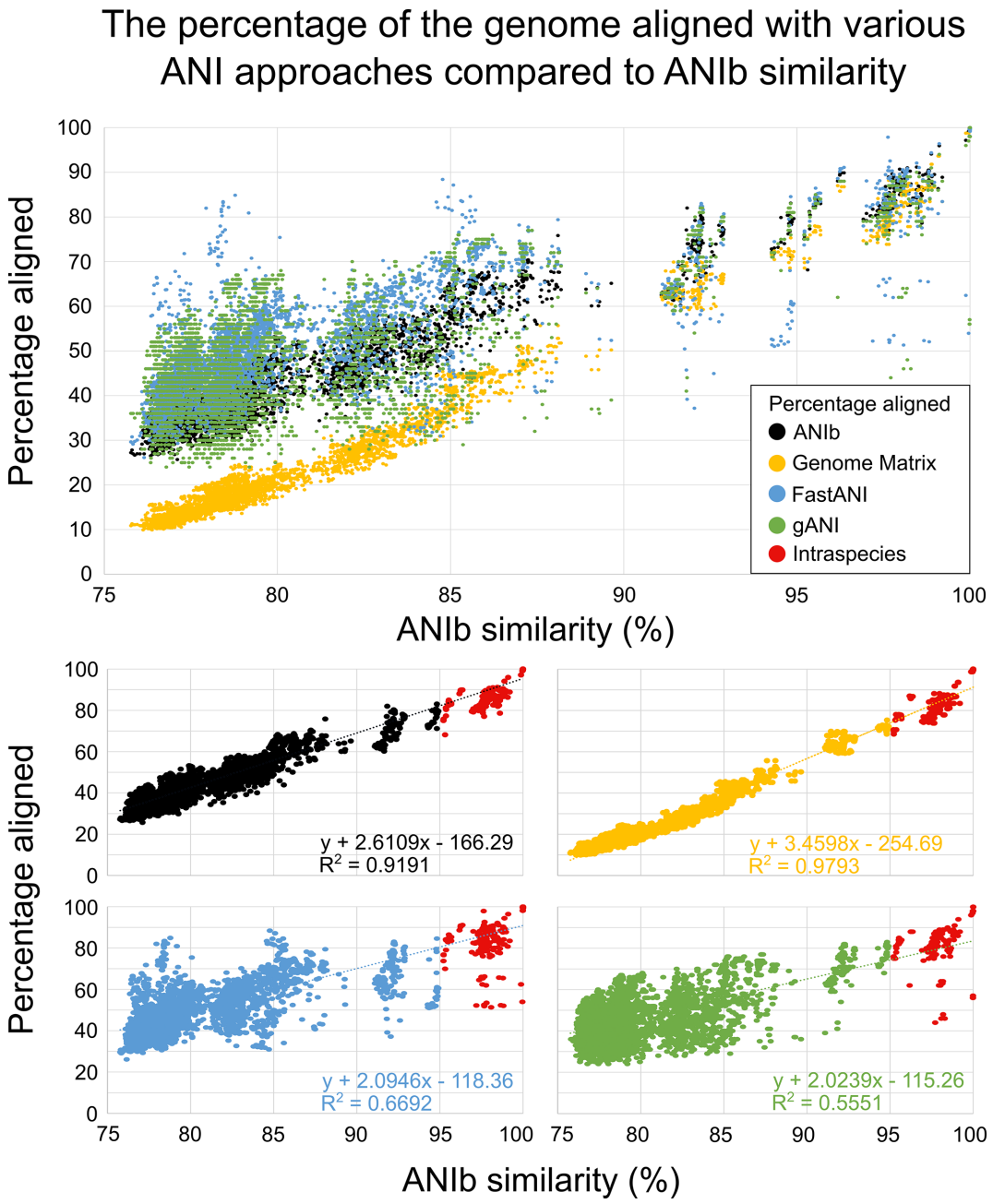

Fig. 2. The percentage of the genomes that were alignable in each pair-wise comparison for *Paraburkholderia*. Only ANIb, Genome Matrix, FastANI and gANI provided the percentage aligned or the alignment fraction of the genomes compared. Intraspecific comparisons are indicated in red, and the equation and R^2-value for the linear regression line is indicated for each metric.

than at the interspecies level (Fig. 2). In fact, it appears as if the more closely related the taxa, the lower the differences observed between methods (File S1). For conspecific taxa with an ANIb value of *ca.* 95%, the alignable portions of the genomes typically ranged between 50 and 80%, whereas an ANIb value around 98% showed an alignable portion of *ca.* 80–90% with all approaches. In general, ANIb values above 99.8% resulted in almost 100% of the genomes being used with all methods.

Overall, the intrageneric relationships inferred from the ANI data were considerably less congruent with the AML relationships inferred from the marker genes (Fig. 3, S3). In fact, the main topological differences observed between the phylogenomic AML trees and the ANI-based NJ trees were the placement of strains belonging to the same species. Congruent with the low GSI values from the individual gene trees, this discordance was particularly evident for strains of *B. ottawaense*, *B. yuanmingense*, *Pan. agglomerans*, *Pan.*

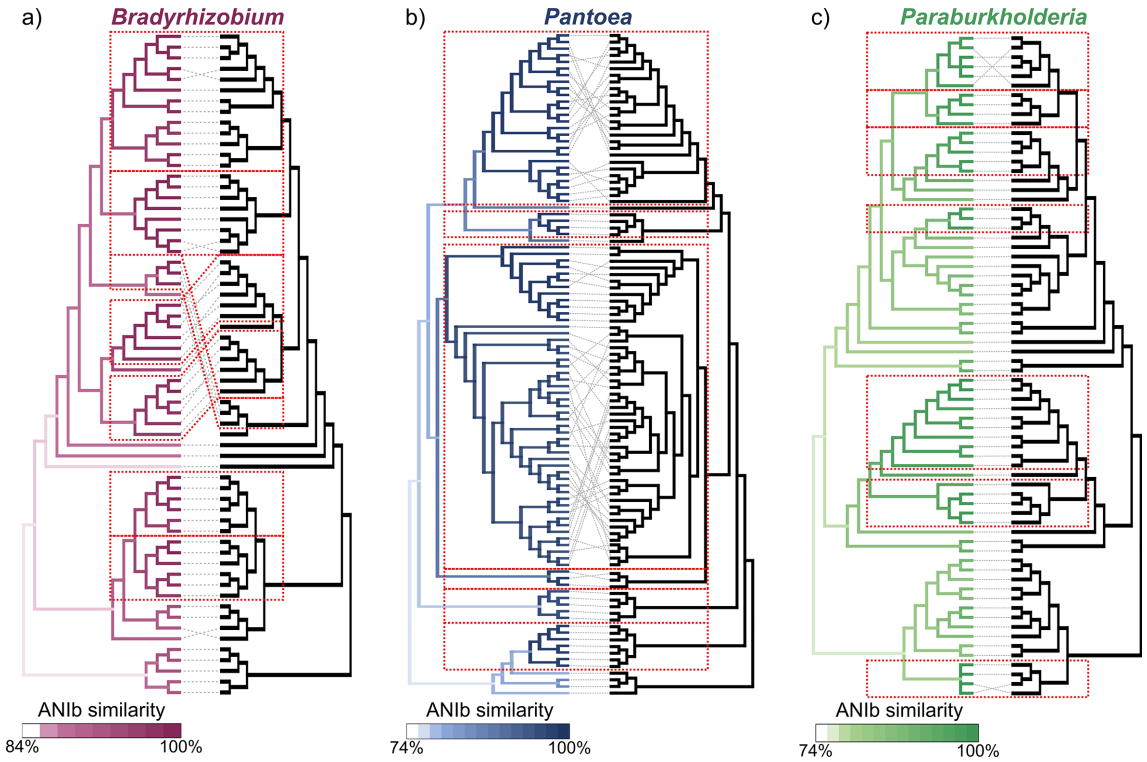

Fig. 3. Comparison of cladograms inferred from phylogenomic AML and ANI-based NJ (indicated in black) phylogenies for (a) *Bradyrhizobium*, (b) *Pantoea* and (c) *Paraburkholderia*. AML analyses were performed with the concatenated supermatrix of 92 shared protein sequences using UBCG [26], and individual phylogenies are presented in Figs S2 and S3. NJ trees were inferred using distances calculated from ANIb values. Similarity values obtained with ANIb are indicated as a colour range on branches of the corresponding AML cladograms, with all pair-wise comparisons between members (inter- and intraspecies) having an ANIb value of at least 76% within a genus. Multi-strain species are indicated with red dotted lines, while specific taxa are connected with grey dotted lines.

ananatis, *Par. caballeronis* and *Par. graminis*. This is not surprising, as it has early on been well documented that distance-based phylogenetics often do not recover relationships obtained with maximum-likelihood analyses of molecular data [36, 37].

DISCUSSION

Our findings showed that the array of ANI approaches available does not provide directly comparable values of genetic relatedness or genomic cohesion. We also showed that these approaches fail to provide consistent results for taxonomic purposes. All the available ANI approaches were compared to the widely used ANIb, and none were perfectly correlated with ANIb, particularly at the level of the species boundary. If one were to use a single pre-determined cut-off value for taxonomic purposes across all the different ANI methods (i.e. to delineate biologically meaningful groups with any ANI tool using a single value), perfect correlation among the different ANI methods would be required, specifically in the genetic similarity range where species and generic boundaries are located. This means that conclusions regarding relatedness are not directly transferrable if different approaches and different taxon sets are employed.

Variation observed in ANI values is entirely due to the differences in parameters and search algorithms implemented by the respective ANI methods. Similarity searches in BLAST use a hashing approach, which entails a heuristic search of short sequence stretches (or words) between two sequences acting like anchors followed by attempts to extend the alignments of the sequences from these areas of similarity [12, 17, 31, 38]. NSimScan and USEARCH also use these sequence searches, but they employ filters to ensure the exclusion of weak matches from further analyses [18, 39, 40]. In contrast to these sequence search-based approaches, MUMmer employs suffix trees to identify potential anchor points for an alignment [12, 31], but is less sensitive toward detecting lower similarity matches and it becomes unreliable with draft or incomplete genomes [3, 31]. FastANI uses fast approximate read mapping with Mapmash, based on MinHash alignment identity estimates [19, 33, 41]. Accordingly, the values obtained for the pair-wise calculations with the various methods differ, as stricter search criteria produce higher values due to the exclusion of more variable homologous regions. In turn, this causes variation in the proportion of the genomes being analysed and ultimately each ANI method generates ANI values that are only comparable to other ANI values calculated with that specific method.

Table 2. Lower limit for the ANI values (%) obtained within species analysed using the seven ANI methods compared

Multi-strain species*	Number of Strains	ANIb	ANIm	OAT	OAU	Genome matrix	FastANI	gANI
Bradyrhizobium								
B. arachidis	3	98.23	98.42	98.51	98.54	98.70	98.42	98.74
B. diazoefficiens	8	98.05	98.32	98.42	98.37	98.51	97.98	98.61
*B. elkanii**	6	94.68	95.30	95.22	95.03	95.44	94.91	95.93
*B. japonicum**	13	93.96	95.20	94.75	94.74	95.23	94.80	95.28
B. ottawaense	6	97.99	98.44	98.28	98.42	98.72	98.19	98.55
*B. pachyrhizi**	6	94.60	95.36	95.18	94.95	95.42	95.24	95.81
B. yuanmingense	6	96.12	96.61	96.46	96.41	96.69	96.43	97.03
Pantoea								
Pan. agglomerans	26	96.82	97.15	97.02	97.02	97.07	96.73	97.29
Pan. allii	3	98.29	98.65	98.54	98.54	98.59	98.37	98.71
Pan. ananatis	49	95.68	96.04	96.07	96.07	96.02	95.91	96.49
Pan. brenneri	4	98.90	99.16	99.13	99.13	99.18	99.06	99.29
Pan. eucrina	7	98.19	98.66	98.50	98.50	98.50	98.14	98.60
Pan. septica	5	95.36	95.86	95.87	95.87	95.80	95.73	96.09
Paraburkholderia								
*Par. caribensis**	5	94.77	95.59	95.92	96.00	95.95	95.96	96.56
Par. hospita	10	96.92	97.31	97.78	97.88	98.03	97.62	98.00
Par. caledonica	5	95.36	95.91	95.89	95.89	95.66	95.89	96.45
Par. fungorum	3	97.71	98.31	98.27	98.37	98.35	98.30	98.45
Par. graminis	6	98.03	98.16	98.14	98.16	98.18	98.19	96.45
Par. phenoliruptrix	4	97.34	98.12	97.94	98.01	98.15	98.05	98.25
P. caballeronis	4	99.98	99.72	99.98	99.95	99.99	99.98	100

*Species with ANI values lower than the expected intraspecific values (i.e. <95%) using at least one method are indicated with an asterisk and corresponding ANI values are indicated with orange blocks.

Based on our results, the proportion of the genome considered to be homologous may also not be a reliable approach for inferring species relatedness and species boundaries. It was suggested previously that relatedness should be inferred from both genome content and similarity [18], but the homologous proportion between genomes is often highly variable [1]. Consider for example the comparison between *Par. oxyphila* and the type strain of *Par. caledonica*, where the genome of the latter was used as the subject and the *Par. oxyphila* genome as query. In this instance there was a large difference between the proportions of the *Par. caledonica* genome analysed with the different approaches, ranging between 11 and 52%. This again relates to the parameters and algorithms employed by these different approaches, especially because some approaches are stricter in their homology detection. There was also a vast difference between the reciprocal analyses. More than 50% of the protein-coding sequences of *Par. caledonica* were analysed with gANI whereas only 36% of the protein-coding sequences of *Par. oxyphila* were analysed in the reciprocal gANI calculation. This suggests that gANI is particularly sensitive to genome size differences, because the number of protein-coding sequences for *Par. caledonica* is 6571 as opposed to 9156 in *Par. oxyphila*. Furthermore, where closely related species occur in the same or overlapping niches, they may share a very high proportion of their gene content [42–44], resulting in a much larger proportion that is homologous among their genomes than what is typically expected between species. Thus, although valuable information is captured by analysing and quantifying the homologous fraction between the genomes of species, this measure alone does not provide a reliable indication of relatedness between genomes.

The data obtained from this study indicated that calculation of ANI values with the more recently developed algorithms allow for massive decreases in computation time. As ANI

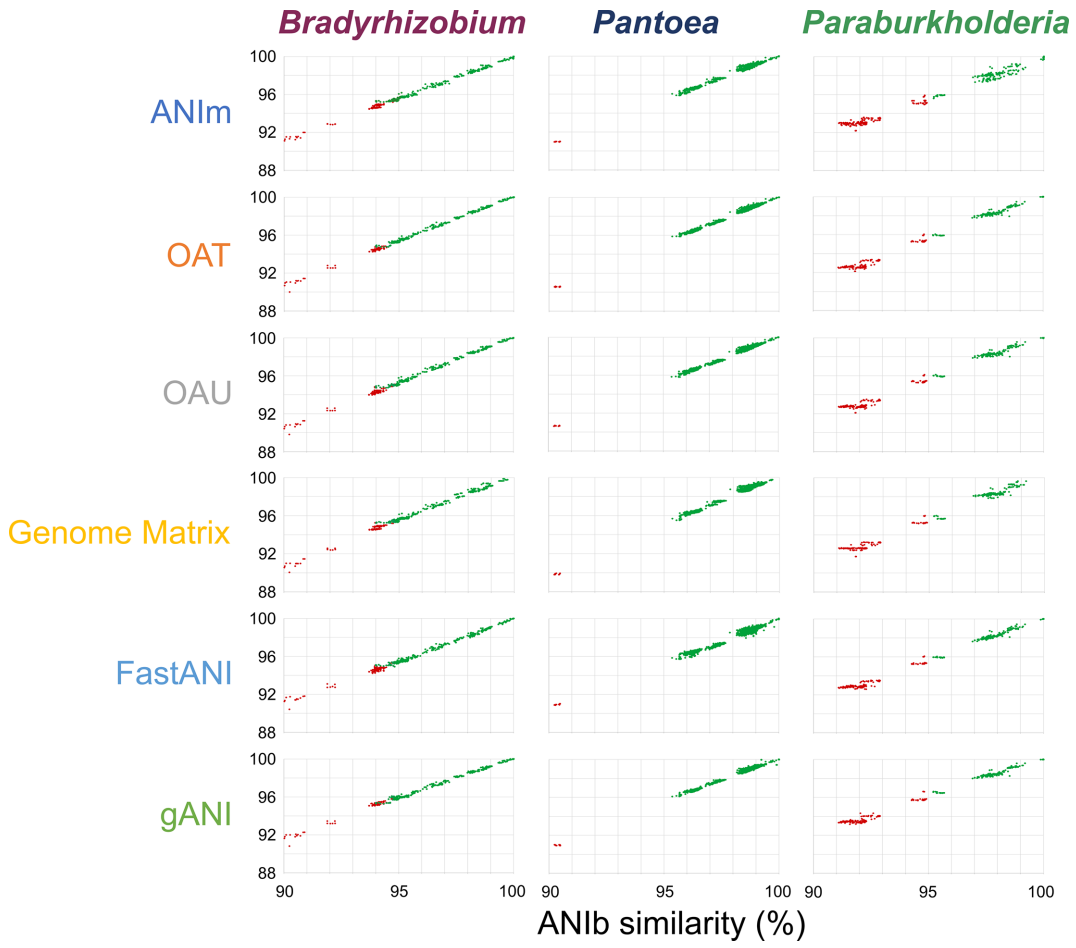

Fig. 4. ANI values near the species boundary in the ANIb similarity range of 90–100% for the three taxon sets. In these analyses, the widely used BLAST-based approach, ANIb, was compared to each of the six alternative methods for determining ANI (for details of the various ANI methods, see Table 1). In each case, the *x*-axis indicates the ANIb similarity percentage, while the *y*-axis indicates the percentage similarity for each of the six alternative approaches. Intraspecies comparisons are indicated with green dots, whereas red dots denote interspecies comparisons.

comparisons are also performed on large sets of data, for instance in metagenomic analyses, the development of faster approaches for investigating genomic relatedness is crucial. However, in most instances there appeared to be a trade-off between correlation to ANIb and the speed of calculations [18, 19]. OrthoANIu, employing USEARCH, showed a marked decrease in calculation time compared to ANIb (i.e. ANIb is approximately five times slower than OrthoANIu). Although ANIm showed similar calculation time decreases, many of these large-scale analyses deal with draft or incomplete genomes, which is not ideal for ANIm calculations [3]. Among the seven ANI methods tested, FastANI was by far the most time-efficient; ANIb is more than 200 times slower than FastANI per comparison.

The data presented here strongly support integration of ANI into a polyphasic approach for delineating bacterial species. This is because a single species cut-off value cannot be directly applied for all the different ANI methods and/or taxon sets. This notion is supported by previous studies where it was suggested that the species cut-off for different approaches (e.g. gANI [18]) as well as for different taxon sets [12, 16] should be adjusted. For the taxa investigated and for ANIb values between 90 and 100%, ANIm correlated most closely to ANIb, despite the apparent weak correlation of ANIm to ANIb overall. With all other methods tested, conclusions regarding the conspecificity of isolates would not necessarily be congruent to those drawn using a polyphasic delineation approach. It would therefore be useful to extend ANI analyses to sets of strains to obtain an indication of the level of diversity one might expect within a particular species, without which one might be unable to determine whether multiple species are present in the taxon set [12]. This is especially important for taxa where unexpectedly low values are observed between apparently conspecific members.

A good example where the use of ANI-based species cut-off values could have caused taxonomic confusion, had it not been for the incorporation of this metric in the polyphasic approach, was observed in the *Bradyrhizobium* dataset. The

ANI values obtained among members of *B. elkanii* and *B. pachyrhizi* were within the same range as values obtained among members of *B. japonicum* and only slightly lower than those obtained within these same species groups. However, their recognition as distinct species is supported by a range of other types of data (e.g. phenotypic traits, geographical distribution, plant host range), which includes phylogenetic information [e.g. multi-locus sequence analysis (MLSA) and genealogical concordance] [45]. The inverse may also be true (see *Par. caledonica* or *Par. caribensis*) where intraspecies comparisons below the typical ANI discontinuity for the genus may suggest the existence of potentially separate species groups, although in this case limited additional data exists to support their separation. Such discrepancies regarding conclusions on conspecificity of isolates between ANI and other approaches are the consequence of the intrinsically variable nature of prokaryotic species [46], which render their accurate diagnosis using a universal cut-off metric unlikely. Hence, examples where the typical ANIb-based species cut-off range fail to accurately delineate species are fairly common among prokaryotes [47].

Because ANI is dependent on the evolutionary rate, fate and tendencies of the taxa being studied [4, 16, 46], it has been argued that by enforcing a specific cut-off value across all taxa, one may obtain the same level of intraspecies genetic diversity across all prokaryotes [18]. However, the taxa delineated in this way are unlikely to reflect natural diversity patterns [35, 46]. If species are naturally occurring entities, kept together by a variety of evolutionary forces, they should be circumscribed as such and not according to a general cut-off value of relatedness [46]. In fact, ANI is a product of evolution, and given that all species are distinct and subject to their own unique sets of evolutionary fates and tendencies, the application of predetermined species cut-off measures is fundamentally flawed.

Although ANI has proven a valuable measure of genomic relatedness, we argue that the predetermined cut-off range should be interpreted as a guideline. Holistically, polyphasic data need to be in agreement regarding cohesion at the phylogenetic, genomic and phenotypic levels [5, 8, 9, 12]. If the conspecificity of isolates is questioned, evolutionary history should be used for identifying the putative species boundary and species hypotheses, which can subsequently be subjected to the polyphasic approach for ultimately generating robustly supported species groups that approximate those occurring in nature [30, 46]. Also, if any of the faster and high-throughput approaches for the calculation of ANI are used, the inferred species boundary should be calibrated for the particular method and the specific taxon set of interest, before conspecificity is inferred from cut-offs.

In conclusion, the development of new ANI algorithms has drastically decreased the computation time required for large-scale genome comparisons. However, these different approaches do not produce similar results to the original approaches that were used to establish the suggested ANI thresholds. This is in part because these methods are based on different computational procedures. Their application therefore does not produce directly comparable ANI data, particularly for analyses at, or near, the inferred species boundary. More importantly, however, deviation from the suggested ANI thresholds is a direct consequence of species evolution. All species have evolved via independent evolutionary trajectories subjected to different evolutionary forces, resulting in cohesion among individuals occurring at different levels of similarity. Therefore, application of a universal metric or species cut-off ANI is irrational. Appropriate interpretation of ANIs is thus dependent on the methodology used to determine them, and the evolutionary history of the taxa being investigated.

Funding information
M.P. received funding from the NRF-DST Centre of Excellence in Tree Health Biotechnology (CTHB) at the Forestry and Agricultural Biotechnology Institute (FABI), the University of Pretoria and M.P. and B.H. received funding from the US National Science Foundation, Division of Environmental Biology (DEB1557042). Informatics infrastructure was funded by the NRF National Bioinformatics Functional Genomics Grant (No: 93668). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions
M.P. was involved in conceptualization, methodology, software, validation, formal analyses, investigation, interpretation of findings, writing, reviewing, editing and visualization of the work. E.S. was involved in conceptualization, interpretation of findings, writing, editing, visualization of the work. J.B. was involved in the provision of software and resources, as well as data curation. B.H. was involved in provision of resources, review and editing of the work. S.V. was involved in conceptualization, provision of resources, interpretation of findings, writing, review and editing of the work.

Conflicts of interest
The authors declare that they have no conflicts of interests.

References

1. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 2005;102:2567–2572.
2. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P *et al*. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.
3. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 2009;106:19126–19131.
4. Sentausa E, Fournier P-E. Advantages and limitations of genomics in prokaryotic taxonomy. *Clin Microbiol Infect* 2013;19:790–795.
5. Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the bacteria and archaea. *Int J Syst Evol Microbiol* 2014;64:316–324.
6. Yoon S-H, Ha S-M, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 2017;110:1281–1286.
7. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR *et al*. Report of the AD hoc Committee on reconciliation of approaches to bacterial Systematics. *Int J Syst Evol Microbiol* 1987;37:463–464.
8. Thompson CC, Chimetto L, Edwards RA, Swings J, Stackebrandt E *et al*. Microbial genomic taxonomy. *BMC Genomics* 2013;14:913.
9. Rosselló-Móra R, Amann R. Past and future species definitions for bacteria and archaea. *Syst Appl Microbiol* 2015;38:209–216.

10. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 2007;10:504–509.
11. Whitman WB. The need for change: embracing the genome. In. In: Goodfellow M, Sutcliffe I, Chun J (editors). *Methods in Microbiology*. Academic Press; 2014. pp. 1–12.
12. Arahal DR. Whole-genome analyses: average nucleotide identity. *Methods in microbiology*. Elsevier; 2014. pp. 103–.122.
13. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 2016;66:1100–1103.
14. Auch AF, Klenk H-P, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* 2010;2:142–148.
15. Deloger M, El Karoui M, Petit M-A. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* 2009;191:91–99.
16. Federhen S, Rosselló-Móra R, Klenk H-P, Tindall BJ, Konstantinidis KT *et al*. *Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015)*. BioMed Central; 2016.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
18. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015;43:6761–6771.
19. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
20. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes: PeerJ Preprints. Report No.: 2167-9843; 2016.
21. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996–1004.
22. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
23. Avontuur JR, Palmer M, Beukes CW, Chan WY, Coetzee MPA *et al*. Genome-informed *Bradyrhizobium* taxonomy: where to from here? *Syst Appl Microbiol* 2019;42:427–439.
24. Estrada-de Los Santos P, Palmer M, Chávez-Ramírez B, Beukes C, Steenkamp ET *et al*. Whole genome analyses Suggests that *Burkholderia* sensu lato contains two additional novel genera (*Mycetohabitans* gen. nov., and *Trinickia* gen. nov.): implications for the evolution of Diazotrophy and nodulation in the *Burkholderiaceae*. *Genes* 2018;9:389.
25. Palmer M, Steenkamp ET, Coetzee MPA, Chan W-Y, van Zyl E *et al*. Phylogenomic resolution of the bacterial genus *Pantoea* and its relationship with *Erwinia* and *Tatumella*. *Antonie van Leeuwenhoek* 2017;110:1287–1309.
26. Na S-I, Kim YO, Yoon S-H, Ha S-M, Baek I, S-I N, S-m H *et al*. UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J Microbiol* 2018;56:280–285.
27. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 1999;16:1114–1116.
28. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;59:307–321.
29. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.
30. Venter SN, Palmer M, Beukes CW, Chan W-Y, Shin G *et al*. Practically delineating bacterial species with genealogical concordance. *Antonie van Leeuwenhoek* 2017;110:1311–1325.
31. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:R12.
32. Edgar RC. Search and clustering orders of magnitude faster than blast. *Bioinformatics* 2010;26:2460–2461.
33. Jain C, Dilthey A, Koren S, Aluru S, Phillippy AM. *A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. international Conference on Research in Computational Molecular Biology*. Springer; 2017.
34. Felsenstein J. *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author Department of Genome Sciences*. Seattle: University of Washington; 2005.
35. Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. *Microbe* 2014;9:111–118.
36. Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. *J Mol Evol* 1991;32:443–445.
37. Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol Biol Evol* 1995;12:843–849.
38. McGinnis S, Madden TL. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 2004;32:W20–W25.
39. Kaznadzey A, Alexandrova N, Novichkov V, Kaznadzey D. PSimScan: algorithm and utility for fast protein similarity search. *PLoS One* 2013;8:e58505.
40. Novichkov V, Kaznadzey A, Alexandrova N, Kaznadzey D. NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. *Bioinformatics* 2016;32:2380–2381.
41. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
42. Purushe J, Fouts DE, Morrison M, White BA, Mackie RI *et al*. Comparative genome analysis of *Prevotella ruminicola* and *Prevotella bryantii*: insights into their environmental niche. *Microb Ecol* 2010;60:721–729.
43. Grim CJ, Kotewicz ML, Power KA, Gopinath G, Franco AA *et al*. Pan-genome analysis of the emerging foodborne pathogen *Cronobacter* spp. suggests a species-level bidirectional divergence driven by niche adaptation. *BMC Genomics* 2013;14:366.
44. Palmer M, de Maayer P, Poulsen M, Steenkamp ET, van Zyl E *et al*. Draft genome sequences of *Pantoea agglomerans* and *Pantoea vagans* isolates associated with termites. *Stand Genomic Sci* 2016;11:23.
45. Ramírez-Bahena MH, Peix A, Rivas R, Camacho M, Rodríguez-Navarro DN *et al*. *Bradyrhizobium pachyrhizi* sp. nov. and *Bradyrhizobium jicamae* sp. nov., isolated from effective nodules of *Pachyrhizus erosus*. *Int J Syst Evol Microbiol* 2009;59:1929–1934.
46. Palmer M, Venter SN, Coetzee MPA, Steenkamp ET. Prokaryotic species are *sui generis* evolutionary units. *Syst Appl Microbiol* 2019;42:145–158.
47. Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K *et al*. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* 2018;68:2386–2392.