Fast genome-based delimitation of Enterobacterales species

Julie E. Hernández-Salmerón[1], Tanya Irani[1], and Gabriel Moreno-Hagelsieb[1,*]

[1]Department of Biology, Wilfrid Laurier University, 75 University Ave W, Waterloo, Ontario N2L 3C5, Canada

Average Nucleotide Identity (ANI) is becoming a standard measure for bacterial species delimitation. However, its calculation can take orders of magnitude longer than fast similarity estimates based on sampling of short nucleotides, compiled into so-called sketches. Though these estimates correlate well with ANI, they might miss some of what ANI would produce. We therefore compared the results of two fast programs, mash and dashing, against ANI, in delimiting species among publicly available Esterobacterales genomes. Receiver Operating Characteristic (ROC) curve analysis found that all three programs were highly accurate in species delimitation, with Area Under the Curve (AUC) values above 0.99, indicating almost perfect species discrimination. After sub-sampling from over-represented species, the AUC lowered to 0.94 with all three methods. In focused tests with ten genera represented by more than two species, all measures showed, again, almost identical results, with *Shigella* showing the lowest AUC values (0.68), followed by *Citrobacter* (0.79) and *Enterobacter* (0.91). The remaining genera, *Dickeya*, *Escherichia*, *Klebsiella*, *Pectobacterium*, *Proteus*, *Providencia* and *Yersinia*, produced AUC values above 0.97. Genome distance ranges varied among the species of these genera. The *E. coli* + *Shigella* group remains challenging to separate. The group had AUC values lower than 0.6 with all measures tested. Overall, our results suggest that fast estimates of genome similarity are as good as ANI for species delimitation. Therefore, these fast estimates might suffice for determining the role of genomic similarity in bacterial taxonomy.

Average Nucleotide Identity | MinHash | Enterobacterales | Bacteria | Species
Correspondence: ***gmorenohagelsieb@wlu.ca***

DRAFT

Supplementary Note 1: Introduction

Average Nucleotide Identity (ANI), seems to be becoming a standard in genome-based species delimitation (5, 10, 13, 16). The method involves the comparison of genome segments, around 1000 base-pairs long, against another genome, and, as the name implies, the calculation of the average identity of matching segments. While much faster than experimentally-based approaches, such as DNA-DNA hybridization, the method can take a while when comparing thousands of genomes, even when using an optimized program for the calculation, such as fastANI (7).

To produce faster comparisons, methods based on sampling k-mers, normally between 20 and 40 bp long, have been developed, with the most commonly used being the MinHash approach, programmed into the mash software (12). This program gathers k-mer samples, codified as hashes, into so-called sketches, which can be efficiently compared. The overall genome similarity is then estimated from the proportion of identical k-mers found between the compared sketches. MinHash can estimate genome similarities in a fraction of the time required for ANI (6, 12), and have been reported to highly correlate with ANI measurements (12). Another program, dashing, offers a few different approaches to codifying sketches, and can produce results very similar to those by mash, among other useful genome similarity estimates (3).

While the articles on mash and dashing show that their results correlate very well with those produced by ANI (3, 12), more where-the-rubber-meets-the-road analyses seem to be required to properly evaluate how well they would substitute overall genome similarity comparisons. We recently published a comparison of these programs against ANI in delimiting species in the *Klebsiella* genus (6). Since the classification of *Klebsiella* has involved ANI along other features (14, 18, 19), it is not surprising that ANI-based clusters had an excellent match with species annotations in this genus. The good news was that the sketch-based methods produced results of the exact same quality (6).

Here we present an expansion of those analyses towards all of the complete genomes available at NCBI's RefSeq genome database (10), classified into the Enterobacterales taxonomic order, to try and ensure that the sketch-based estimates can work as well as ANI in delimiting species beyond those of *Klebsiella*. We therefore tested fastANI, mash and dashing for delimiting species at the overall order level, as well as within the few genera represented by more than two species.

Supplementary Note 2: Methods

A. Genome sequences. To compile a genome sequences dataset, we downloaded all available genomes with a "Complete" status at NCBI's RefSeq genome database (10), by the end of August of 2022. From these, we selected all the genomes classified into the order Enterobacterales, as long as they were classified to the species level, for a total of 6,644 genome sequences (Supplementary Table S1). Type strains were identified from the annotations found in the corresponding NCBI's "GBFF" files.

B. Genome comparisons. To calculate Average Nucleotide Identity (ANI), we selected the fastANI program (7), produced by the group that, as far as we know, most promoted the use of ANI in genome-wide classification studies (7–9, 21). Briefly, the original ANI calculations (4), started by breaking a query genome into 1020 base-pairs long segments, which were then compared against a target genome

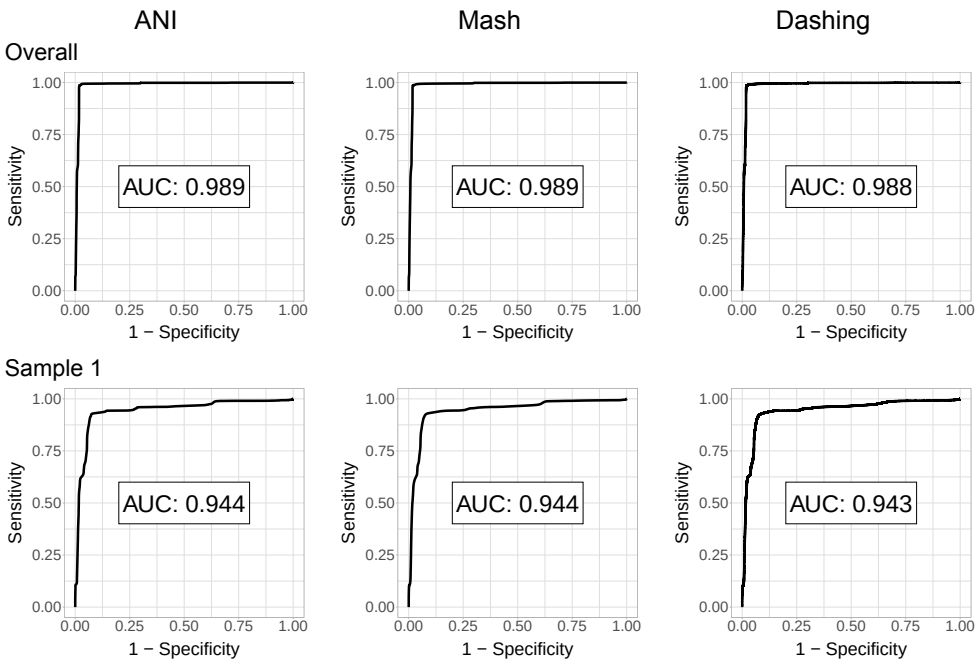

Fig. 1. Species separation across Enterobacterales. All three methods produced exactly the same overall quality of species separation with Area Under the Curve (AUC) values close to 0.99. After sampling genomes from over-represented species, to avoid bias, the AUC lowered to a bit above 0.94 in all cases (Sample 1 of 50 shown as an example).

using NCBI's blastn (2). As the name suggests, ANI was then calculated by taking the average identity of matching segments found by blastn. Instead of blastn, fastANI uses a MinHash approach to quickly estimate the similarity between the genome-derived DNA segments and the target genome (7). To run fastANI comparisons, we chose a fragment length of 1020 bp, instead of the default of 3000 bp (--fragLen 1020). Every other option was left at its default value. The ANI produced by fastANI is reported as a percent identity. Therefore, to transform ANI into distances, we subtracted the obtained ANI from 100 and divided the result by 100.

To calculate distances using the MinHash approach, we used two programs: mash (12) and dashing (3). For mash, version 2.3, we produced sketches of 5000 k-mers (-s 5000 option), rather than the default of 1000, because our preliminary tests showed that the default failed to produce distances for many genomes. We did not use more than 5000 sketches because our preliminary tests found the results to be almost identical to those produced using 10,000 k-mers (data not shown). Every other option was left to its default value. The distances produce by mash are reported as fractions, rather than percents.

We used dashing version 1.0.2. We changed the k-mer length to 21 bp, instead of its default of 31, because 21 is the k-mer size used by mash. We used a sketch size of 2^14 (-S 14), because this sketch size produced the best estimates of genomic Jaccard similarity in the original dashing publication (3). We only tested mash distances (-M option) because our prior report showed that results from other measures were identical (6). Every other option was left unchanged from its default value.

C. Accuracy and plotting. Plots and other calculations were produced using R (15). To test the accuracy of all programs, we performed Receiver Operating Characteristic (ROC) curve analyses as implemented in the R package cutpointr (20). We also used the cutpointr package to find optimal cutoffs. Trees were displayed using the R packages ggtree (23) and ggtreeExtra (22).

Supplementary Note 3: Results and Discussion

Though the focus of this report is on the quality of results, it is noteworthy that calculating ANI between all genomes, with the fastANI program, required the separation of all genomes into several datasets to divide the labour across three computers. The calculation took around two months. Calculating distances using mash or dashing took no more than a couple of hours, including the time to produce "presketch" files, which can be reused to produce results in a more timely manner. These fast programs were run in an iMac computer with 32G of RAM and an Intel processor. The programs ran using four of the available 8 cpu threads.

A. All methods discriminate Enterobacterales species with high accuracy. To test species/genus separation using all of the three tools, fastANI, mash and dashing, we produced tables with all *vs.* all pairwise distances, as calculated with all three methods (Supplementary Table S2). Pairs belonging to either the same species, or the same genus / different species, were used to test for species discrimination. A test with all same species and same genus pairs showed AUC values above 0.99 for all three methods (Fig. 1).
The species representation of the Enterobacterales genome

DRAFT

Area under the curve (AUC) values for species/genus separation across ten genera.

Genus	No. species	ANI	mash	dashing
Citrobacter	14	0.795	0.792	0.794
Dickeya	9	0.999	0.999	0.998
Enterobacter	14	0.916	0.916	0.896
*Escherichia**	4	1.000	1.000	1.000
*Klebsiella**	11	0.988	0.988	0.988
Pectobacterium	14	0.972	0.973	0.971
Proteus	6	0.995	0.995	0.995
Providencia	7	0.971	0.972	0.954
Shigella	4	0.689	0.670	0.675
Yersinia	15	0.984	0.983	0.981

Table 1. *The values for *Escherichia* and *Klebsiella* are the median AUC calculated from 20 sub-samples.

dataset is heavily biased, with representatives of *Escherichia coli* showing a number of representatives orders of magnitude larger than those of most other species. Therefore, the AUC values might be biased towards the distinction of *E. coli* (of all 3,836,048 same species pairs, 2,368,576, or 62%, are *E. coli* pairs). The abundance of *E. coli*, *Klebsiella pneumoniae* (774,390 pairs, 20%) and *Salmonella enterica* (635,628 pairs, 17%) covered approximately 99% of all same species pairs, though the dataset contained 248 species, 137 of them represented by more than one genome.

To avoid these biases, we obtained 50 subsamples each containing 200 genomes randomly sampled from these top three species. All other species kept their total genomes. Again, all three methods produced results of the same quality, though the AUC values averaged slightly above 0.94 (Fig. 1). Lower than the AUC values obtained in naïve test above, but still indicating very good accuracy in species separation.

B. Within genera tests vary in accuracy with all methods. The tests above combine all same species and same genus pairs, regardless of potential for confusion. Since the most difficult distinction should be between genomes from organisms of the same species from those in the same genus, we ran further tests choosing genera from the Enterobacterales, represented by more than three species, making sure that at least three species within the selected genera were represented by ten genomes or more. This process resulted in ten genera: *Citrobacter*, *Dickeya*, *Enterobacter*, *Escherichia*, *Klebsiella*, *Pectobacterium*, *Proteus*, *Providencia*, *Shigella*, and *Yersinia*.

Of the genera above, only two had severe biases in abundance of genomes representing a single species: *Escherichia* and *Klebsiella*. These were subject to subsampling to the number of genomes in the second most abundantly represented species.

The resulting AUC values for species discrimination between these genera varied (Table 1). The lowest values obtained were for *Shigella*, with AUC values of approx. 0.68. That value was followed by 0.79 for *Citrobacter*, followed in turn by 0.91 for *Enterobacter*. Most other AUC values were above 0.97 (Table 1). Again, the results were pretty much identical

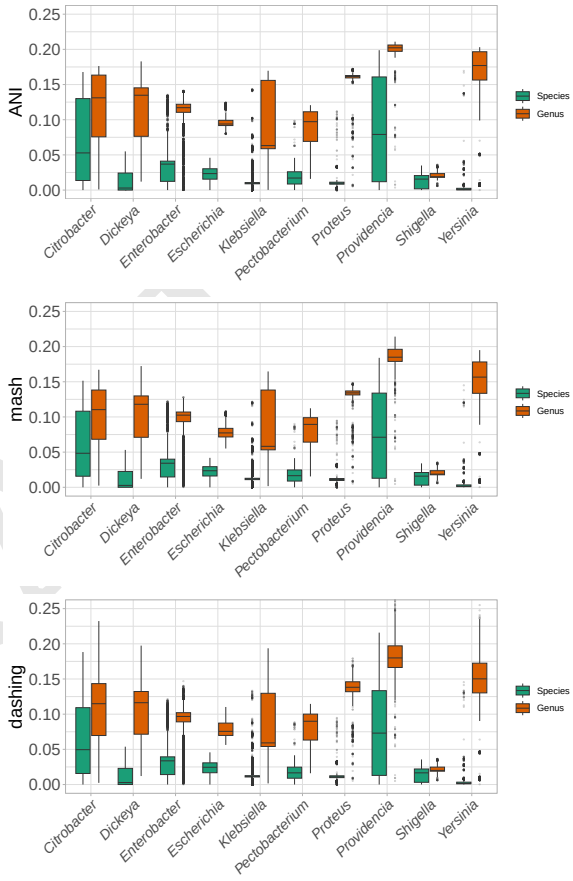

Fig. 2. Species distances. The range of species distances varied from one genus to the next, suggesting that there is no universal species delimitation. Though results are very similar for all measures, dashing seemed to produce more outliers.

with all three measures.

Klebsiella is the genus with the best species representation of all ten, with seven species represented by more than ten genomes, six of them by more than 30. The AUC for species discrimination in this genus remained close to 0.99 in all subsamples with all programs tested (Table 1). As explained above, though ANI has been part of the taxonomic assignments in the genus, which explains the high accuracy with this measure, it is still important to note that mash and dashing produced results of the same quality. We have presented a more focused work about species delimitation in this genus before (6).

Distance ranges varied among the species in the ten genera (Fig. 2). *Yersinia* showed the lowest values, while *Providencia* showed values overlapping the within-Genus distributions of most other genera. This variation agrees with a previous study focused on the ANI distributions of several bacterial species (11), but somewhat contradicting a previously published notion of a universal genetic species delimitation (7, 17). This result seems to also conflict with the high accuracy for species delimitation across Enterobacterales that we presented above. However, the apparent conflict is resolved by the fact that the species in seven of these genera have distances below, roughly, 0.05, combined with within-genus distances above the same mark.

C. The *E. coli* + *Shigella* group remains challenging to separate. The separation of species in the *Escherichia* genus was perfect for all samples tested (AUC: 1), while that for *Shigella* was the worst (AUC: 0.68, see above). However, the clustering of *Shigella* within the *E. coli* species is a classic taxonomy *vs.* tradition issue. Accordingly, by cutting hierarchical clusters at an ANI distance of 0.052, or a mash distance of 0.048 (the cutoffs suggested for the whole dataset when optimising for the F1 score using the cutpointr package), we obtained a cluster containing all *E. coli* genomes combined with all the genomes representing the four *Shigella* species in the dataset (Fig. 3). No other species belonging to different genera had within species distances. We therefore decided to examine this group more closely.

To avoid biases due to over-representation of *E. coli* genomes, we produced subsamples from this *E. coli* + *Shigella* group containing 80 randomly sampled *E. coli* genomes each. The species *vs.* genus AUC values varied from 0.52 to 0.61 with all three methods (ANI, mash and dashing). Examination of the hierarchical cluster produced from either distance measure showed some clean clusters of *Shigella* species, with several more *Shigella* strains interspersed with clear *E. coli* clusters (Fig. 3). A few genomes classified as *E. coli* were inserted into *Shigella* clusters as well. The latter two cases could be due to mislabeled genomes.

A prior analysis, founded on mash distances and supported by further analyses, suggested that the *E. coli* + *Shigella* group, could be divided into 14 phylogroups (1). It is therefore possible that the anomalously clustered genomes need to be relabeled according to the main names within each apparent phylogroup.

To test for this possibility, we looked for a mash distance cutoff that could separate the observed clusters that coincide with *Shigella* species. A mash distance cutoff of 0.021 seemed to work for this purpose. In other words, this cutoff produced four *Shigella* clusters, corresponding each to a *Shigella* species. Three of these were clean single-species clusters: *S. flexneri*, *S. sonnei* and *S. dysenteriae*, the latter including the *S. dysenteriae* type srain ATCC 13313. The *S. boydii* cluster contained the type strain of the species, FDA ARGOS 1139, plus five *E. coli* genomes and two *S. dysenteriae* ones. However, this cutoff produced 21 more clusters for a total of 25 (Fig. 3, Supplementary Table S3). Of these, 17 were clean *E. coli* clusters. The four apparently contaminated *E. coli* clusters were mixed with *S. flexneri*, with one of them also containing a *S. sonnei* genome.

A cutoff mash distance of 0.023 produced 14 clusters. Only two of these were *Shigella* clusters, in apparent agreement with the two *Shigella* phylogroups suggested by (1). One of the *Shigella* clusters was mainly composed of genomes named after all four species: *S. flexneri*, *S. sonnei*, *S. boydii* and *S. dysenteriae*, in order of abundance. This *Shigella* cluster also contained the five *E. coli* genomes found within the *S. boydii* cluster produced with the mash cutoff of 0.021 above. Accordingly, the cluster also contained the *S. boydii* type strain (FDA ARGOS 1139). The second *Shigella* cluster was composed exclusively of ten *S. dysenteriae* genomes, one of them the type strain ATCC 13313. Three of the remaining twelve *E. coli* clusters contained genomes named after *Shigella* species (Fig. 3, Supplementary Table S3).

The results presented in this section suggest that some genomes in the group are mislabelled, with many genomes still showing some phylogenomic coherence. Deciding whether the *Shigella* clusters should correspond to two or more phylogroups requires further analyses that are not within the scope of this research.

Supplementary Note 4: Conclusion

Overall, our results suggest that using fast, sketch-based estimates of genome similarity, can be as accurate for bacterial taxonomy as the whole-genome ANI. Therefore, these estimates should suffice for further determination of the role that genome similarity should play in bacterial taxonomy.

Acknowledgements

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant: RGPIN:2018-06180.

Funding

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant: RGPIN:2018-06180.

References

1. Abram, K. *et al.* (2021). Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. *Communications Biology*, **4**(1), 117.

DRAFT

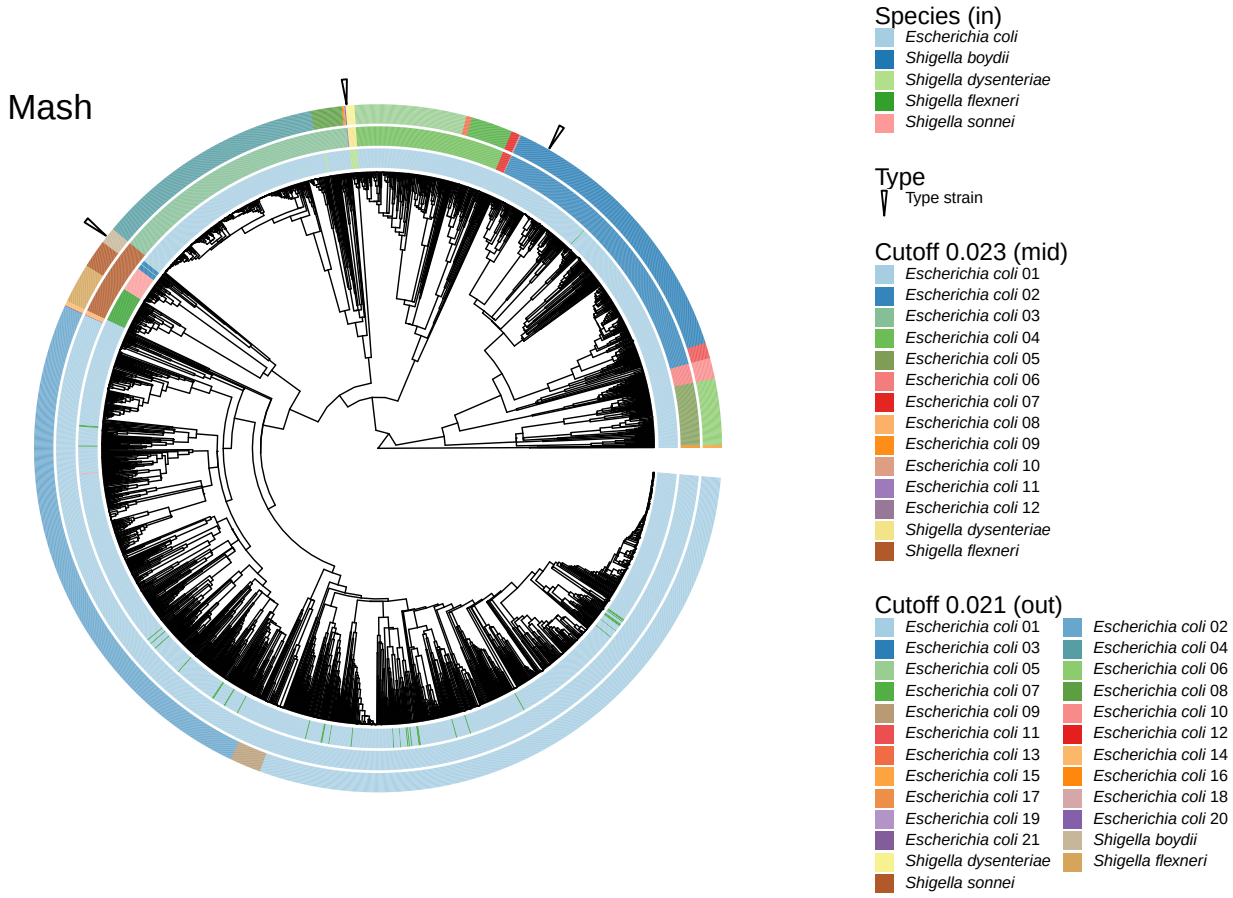

Fig. 3. The *Escherichia coli* + *Shigella* group. The hierarchy shows that several *Shigella* genomes clustered together forming coherent groups, though many are interspersed into groups dominated by genomes labeled as *E. coli* (inner circle, or circle closest to the dendogram). A mash distance cutoff of 0.023 produced 14 groups, two of them mainly composed of *Shigella* genomes (middle circle). A lower cutoff, 0.021, separated *Shigella* into all four represented species, but it also divided the whole group into a total of 25 clusters (external circle).

2. Altschul, S. F. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.
3. Baker, D. N. and Langmead, B. (2019). Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biology*, **20**(1), 265.
4. Goris, J. *et al.* (2007-01). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, **57**(1), 81–91.
5. Hedlund, B. P. *et al.* (2022). SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nature Microbiology*, pages 1–7.
6. Hernández-Salmerón, J. E. and Moreno-Hagelsieb, G. (2022). FastANI, Mash and Dashing equally differentiate between Klebsiella species. *PeerJ*, **10**, e13784.
7. Jain, C. *et al.* (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, **9**(1), 5114.
8. Konstantinidis, K. T. and Tiedje, J. M. (2005-09). Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology*, **187**(18), 6258–6264.
9. Konstantinidis, K. T. *et al.* (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**(1475), 1929–1940.
10. Li, W. *et al.* (2020). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, **49**(D1), gkaa1105–.
11. Murray, C. S. *et al.* (2021). Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nature Communications*, **12**(1), 4059.
12. Ondov, B. D. *et al.* (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, **17**(1), 132.
13. Parks, D. H. *et al.* (2021). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, **50**(D1), D785–D794.
14. Potter, R. F. *et al.* (2018). Population Structure, Antibiotic Resistance, and Uropathogenicity of Klebsiella variicola. *mBio*, **9**(6), e02481–18.
15. R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
16. Rinke, C. *et al.* (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology*, **6**(7), 946–959.
17. Rodriguez-R, L. M. *et al.* (2021). Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species". *Nature Communications*, **12**(1), 4060.
18. Rodríguez-Medina, N. *et al.* (2019). Klebsiella variicola: an emerging pathogen in humans. *Emerging Microbes & Infections*, **8**(1), 973–988.
19. Saxenborn, P. *et al.* (2021). Genotypic Characterization of Clinical Klebsiella spp. Isolates Collected From Patients With Suspected Community-Onset Sepsis, Sweden. *Frontiers in Microbiology*, **12**, 640408.
20. Thiele, C. and Hirschfeld, G. (2021). cutpointr: Improved estimation and validation of optimal cutpoints in R. *Journal of Statistical Software*, **98**(11), 1–27.
21. Varghese, N. J. *et al.* (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, **43**(14), 6761–6771.
22. Xu, S. *et al.* (2021). ggtreeextra: Compact visualization of richly annotated phylogenetic data. *Molecular Biology and Evolution*, **38**, 4039–4042.
23. Yu, G. *et al.* (2018). Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Molecular Biology and Evolution*, **35**(12), 3041–3043.