Title: A rapid and simple method for assessing and representing genome sequence relatedness.

Authors and affiliations : Briand M[1*], Bouzid M[1†], Hunault G[2], Legeay M[3], Fischer-Le Saux M[1], Barret M[1]

[1] IRHS-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49071, Beaucouzé, France

[2] Université d'Angers, Laboratoire d'Hémodynamique, Interaction Fibrose et Invasivité tumorale hépatique, UPRES 3859, IFR 132, F-49045 Angers, France

[3] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

Authors email addresses:

Martial Briand: martial.briand@inrae.fr

Mariam Bouzid : mariam.bouzid@icloud.com

Gilles Hunault : gilles.hunault@univ-angers.fr

Marc Legeay : legeay.marc@free.fr

Marion Fischer-Le Saux : marion.le-saux@inrae.fr

Matthieu Barret: matthieu.barret@inrae.fr

*Corresponding author

Phone: +33 (0)2 41 22 57 29

Fax: +33 (0)2 41 22 57 55

Abstract

Coherent genomic groups are frequently used as a proxy for bacterial species delineation through computation of overall genome relatedness indices (OGRI). Average nucleotide identity (ANI) is a widely employed method for estimating relatedness between genomic sequences. However, pairwise comparisons of genome sequences based on ANI is relatively computationally intensive and therefore precludes analyses of large datasets composed of thousands of genome sequences.

In this work we proposed a workflow to compute and visualize relationships between genomic sequences. A dataset containing more than 3,500 *Pseudomonas* genome sequences was successfully classified with an alternative OGRI based on *k*-mer counts in few hours with the same precision as ANI. A new visualization method based on zoomable circle packing was employed for assessing relationships among the 350 groups generated. Amendment of databases with these *Pseudomonas* groups greatly improved the classification of metagenomic read sets with *k*-mer-based classifier.

The developed workflow was integrated in the user-friendly KI-S tool that is available at the following address: https://iris.angers.inra.fr/galaxypub-cfbp.

Keywords : ANI, *k*-mers, circle packing, *Pseudomonas*, metagenome

Background

Species is a unit of biological diversity. Prokaryotic species delineation historically relies on a polyphasic approach based on a range of genotypic, phenotypic and chemo-taxonomic (e.g. fatty acid profiles) data of cultured specimens. According to the List of Prokaryotic Names with Standing in Nomenclature (LPSN), approximately 15,500 bacterial species names have been currently validated within this theoretical framework [1]. According to different estimates the number of bacterial species inhabiting planet Earth is predicted to range between 10^7 to 10^{12} species [2,3], the genomics revolution has the potential to accelerate the pace of species description.

Prokaryotic species are primarily described as cohesive genomic groups and approaches based on similarity of whole genome sequence, also known as overall genome relatedness indices (OGRI), have been proposed for delineating species. Genome Blast Distance Phylogeny (GBDP [4]) and Average nucleotide identity (ANI) are currently the most frequently used OGRI for assessing the relatedness between genomic sequences. Distinct ANI algorithms such as ANI based on BLAST (ANIb [5]), ANI based on MUMmer (ANIm [6]) or ANI based on orthologous genes (OrthoANIb [7]; OrthoANIu [8]; gANI,AF [9]), which differ in their precision but more importantly in their calculation times [8], have been developed. Indeed, improvement of calculation time for whole genomic comparison of large datasets is an essential parameter. As of November 2018, the total number of prokaryotic genome sequences publicly available in the NCBI database is 170,728. Considering an average time of 1 second for calculating ANI values for one pair of genome sequences, it would take approximately 1,000 years to obtain ANI values for all pairwise comparisons.

The number of words of length k (*k*-mers) shared between read sets [10] or genomic sequences [11] is an alignment-free alternative for assessing the similarities between entities. Methods based on *k*-mer counts, such as SIMKA [10], Kmer-db [12] or Mash [13] can quickly compute pairwise comparison of multiple metagenome read sets with high accuracy. In addition, specific *k*-mer profiles are now routinely employed by multiple read classifiers for estimating the taxonomic structure of metagenome read sets [14–16]. While these *k*-mer

based classifiers differ in terms of sensitivity and specificity [17], they rely on accurate genome databases for affiliating a read to a taxonomic rank.

The objective of the current work was to propose a workflow to quickly compute OGRI and visualize the outputs in an efficient manner. An alternative method based on *k*-mer counts was first evaluated to study species delimitation on extensive genome datasets. We employed *k*-mer counting to assess the similarity among genome sequences belonging to the *Pseudomonas* genus. Indeed, this genus contains an important diversity of species (n = 207), whose taxonomic affiliation is under constant evolution [18–24], and numerous genome sequences are available in public databases. We also proposed an original visualization tool based on D3 Zoomable Circle Packing (https://gist.github.com/mbostock/7607535) for assessing the relatedness of thousands of genome sequences. Finally, the benefit of taxonomic curation of reference database on the taxonomic affiliation of metagenomics read sets was assessed. The developed workflow was integrated in the user-friendly KI-S tool which is available in the galaxy toolbox of CIRM-CFBP (https://iris.angers.inra.fr/galaxypub-cfbp).

Methods

Genomic dataset

All genome sequences (*n*=3,623 as of April 2017) from the *Pseudomonas* genus were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/).

Calculation of Overall Genome Relatedness Indices

The percentage of shared *k*-mers between genome sequences was calculated with Simka version 1.4 [10] with the following parameters (abundance-min 1 and *k*-mer lengths of 11, 13, 15, 17 and 19). Distances between genomic sequences were also calculated using *k*-mers based methods including FastANI [25], Kmer-db [12] (with the following parameters -k 15, distance min) and Mash [13]. The percentage of shared *k*-mer was compared to ANIb values calculated with PYANI version 0.2.3 [26]. Due to the computing time required for ANIb calculation, only a subset of *Pseudomonas* genomic sequences (*n*=934) were selected for this comparison. This subset was composed of genome sequences containing less than 150 scaffolds.

Development of KI-S workflow

An integrative workflow named KI-S (**K**inship relationships **I**dentification with **S**hared *k*-mers) was developed. At the first step, the percentage of shared *k*-mers (*k* can be selected by the user) between genome sequences is calculated with Simka [10] or estimated with Kmer-db [12]. A custom R script is then employed to cluster the genome sequences according to their connected components at different selected thresholds (e.g. 50% of shared 15-mers). The clustering result is visualized with Zoomable Circle Packing representation with the D3.js JavaScript library (https://gist.github.com/mbostock/7607535). The source code of the KI-S workflow is available at the following address: https://sourcesup.renater.fr/wiki/ki-s. A wrapper

for accessing KI-S in a user-friendly Galaxy tool is also available at the following address: https://iris.angers.inra.fr/galaxypub-cfbp.

Taxonomic inference of metagenomic read sets

Classification of nine metagenomic read sets derived from seed, germinating seeds and seedlings of common bean (*Phaseolus vulgaris* var. Flavert) were estimated with Clark version 1.2.4 [16] at the species level. Clark is a method based on a supervised sequence classification using discriminative *k*-mers [16]. These metagenomic datasets were selected because of the high relative abundance of reads affiliated to *Pseudomonas* [27]. The following Clark parameters –k 31 –t <minFreqTarget> 0 and -o <minFreqtObject> 0 were used for the taxonomic profiling. Indeed reducing *k* increase the number of read assignments but also increase the probability of misclassification [16]. Three distinct Clark databases were employed: (*i*) the original Clark database from NCBI/RefSeq (*ii*) the original Clark database supplemented with the 3,623 *Pseudomonas* genome sequences and their original NCBI taxonomic affiliation (*iii*) the original Clark database supplemented with the 3,623 *Pseudomonas* genome sequences whose taxonomic affiliation was corrected according to the reclassification based on the number of shared *k*-mers. For this third database, genome sequences were clustered at >50% of 15-mers, which corresponded to the species level.

Results

Selection of optimal *k*-mer size, *k*-mers indexing softwares and percentage of shared *k*-mers

Using the percentage of shared *k*-mers as an OGRI for species delineation first required the determination of the optimal *k*-mer size. This was performed by comparing the percentage of shared *k*-mers with SIMKA [10] to a widely employed OGRI, ANIb [5], among 934 *Pseudomonas* genome sequences. Since the species delineation threshold was initially proposed following the observation of a gap in the distribution of pairwise comparison values [28], the distribution profiles obtained with *k*-mer lengths 11, 13, 15, 17 and 19 were compared to ANIb values. Short *k*-mers (k =11) were evenly shared by most strains and therefore non-discriminatory (**Fig. 1**). As the length of the *k*-mer increased ($k \geq 13$), a multimodal distribution based on several peaks was observed (**Fig. 1**). This multimodal distribution reflects a genetic discontinuity previously observed in several studies [6, 9, 25]. For example, for k=15, the first peak observed (at 15% of shared *15*-mers) is related to the genome sequences that do not belong to the same species. Then, the peaks at 50 and 80% shared *15*-mers corresponded to genome sequences associated to the same species and subspecies, respectively. The last peak at 100% of shared *k*-mers was related to identical genome sequences. Since increasing *k*-mer lengths beyond 15 did not improve the resolution of the multimodal distribution but leads to a more rapid drop in the percentage of shared *k*-mers between strains, we selected k=15 for subsequent analyses.

Several tools using *k*-mers for estimating genomes relatedness (e.g. FastANI, Simka, Kmer-db, Mash) were available at the time of analysis. Mash outperformed all others tools in term of computation time (**Table 1**). Indeed computation time of *15*-mers for 934 genome sequences was 7 min with Mash on a DELL Power Edge R510 server, while it took approximately 3 months for obtaining all ANIb pairwise comparisons. The outputs of these different softwares were next compared to ANIb (**Fig. 2**). FastANI was the best estimator of OGRI as indicated by the strong linear relationship of average pairwise similarities with ANIb values (**Fig. 2**). However, one small caveat is that sequence similarity values were ignored

by FastANI when below 0.76 value, which artificially improved the linear relationship. According to these linear relationships, Simka and Kmer-db performed reasonably well for ANIb values above 0.9, while MASH was restricted to ANIb values above 0.95. Hence, Kmer-db and Simka were selected in KI-S workflow since these tools are the best compromise between quick computation time and robust genome relatedness indexes (**Table 1** and **Fig. 2**).

We next investigated what percentage of shared *15*-mers corresponds to an ANIb value of 0.95, a threshold commonly employed for delineating bacterial species [5]. Fifty percent of *15*-mers were closed to an ANIb value of 0.95 (**Fig. 2**). More precisely, the median percentage of shared *15*-mers was 49% [34%-66%] for ANIb value ranging from 0.94 to 0.96. The 934 genomic sequences were clustered in 329 and 315 groups at an ANIb value of 0.95 and 50% of *15*-mers, respectively. The composition of these groups was identical between the two approaches for 302 groups that contained 808 genomic sequences. The 27 additional groups obtained with ANIb were nestled within the 13 additional groups derived from 50% of shared 15-mers.

Classification of *Pseudomonas* genome sequences

The percentage of shared *15*-mers was then used to investigate the relatedness between 3,623 *Pseudomonas* publicly available genome sequences. At a threshold of 50% of *15*-mers, we identified 350 groups. The group containing the most abundant number of genome sequences was related to *P. aeruginosa* (*n* = 2,341), followed by the phylogroups PG1 (*n* = 111), PG3 (*n* = 92) and PG2 (*n* = 74) of *P. syringae* species complex ([19]; **Table S1**). At the clustering threshold employed, 185 groups were composed of a single genome sequence, therefore highlighting the high *Pseudomonas* strain diversity. Graphical representation of -clustering by dendrogram for a large dataset is generally not optimal. Here we employed Zoomable circle packing as an alternative to dendrogram for representing similarity between genome sequences (**Fig. 3** and **FigS1.html**). The different clustering thresholds that can be

superimposed on the same graphical representation allow the investigation of inter- and intra- groups relationships (**Fig. 3** and **FigS1.html**).

Improvement of taxonomic affiliation of metagenomic read sets.

The taxonomic composition of metagenome read sets is frequently estimated with *k*-mer based classifiers. While these *k*-mer based classifiers differ in terms of sensitivity and specificity, they all rely on accurate genome databases for affiliating a read to taxonomic rank. Here, we investigated the impact of database content and curation on taxonomic affiliation. Using Clark [16] as a taxonomic profiler with the original Clark database, we classified metagenome read sets derived from bean seeds, germinating seeds and seedlings [27]. Adding the 3,623 *Pseudomonas* genome sequences with their original taxonomic affiliation from NCBI to the original Clark database did not increase the percentage of classified reads (**Fig. 4**). However, adding the same genome sequences reclassified in groups according to their percentage of shared *k*-mers (*k*=15; threshold= 50%) increased 1.4-fold on average the number of classified reads (**Fig. 4**).

Discussion

Classification of bacterial strains on the basis on their genome sequence similarities has emerged over the last decade as an alternative to the cumbersome DNA-DNA hybridizations [4, 29]. Although ANIb is one widely employed method for investigating genomic relatedness, its intensive computational time prohibits its use for comparing large genome datasets [8]. In contrast, investigating the percentage of shared *k*-mer is scalable for comparing thousands of genome sequences.

In a method based on *k*-mer counts, choosing the length of *k* is a compromise between accuracy and speed. The distribution of shared *k*-mer values between genome sequences is impacted by *k* length. For *k* = 15, four peaks were observed at 15%, 50%, 80% and 100% of shared *k*-mers. The second peak is close to an ANIb value of 0.95 and falls in the so called grey or fuzzy zone [29] where taxonomists might decide to split or merge species. Hence, according to our working dataset, it seems that 50% of *15*-mers is a good proxy for estimating *Pseudomonas* group. Despite the diverse range of habitats colonized by different *Pseudomonas* populations [22], it is likely that the percentage of shared *k*-mers has to be adapted when investigating other bacterial genera. Indeed, since population dynamics, lifestyle and location impact molecular evolution, it is somewhat illusory to define a fixed threshold for species delineation [30]. While *15*-mers is a good starting point for investigating infra-specific to infra-generic relationships between genome sequences, the computational speed of KI-S offers the possibility to perform large scale genomic comparisons at different *k* sizes to select the most appropriate threshold.

Genomic relatedness using whole genome sequences has become the standard method for bacterial strain identification and bacterial taxonomy [4,29,31]. This is primarily motivated by fast and inexpensive sequencing of bacterial genomes together with the limited availability of cultured specimen for performing classical polyphasic approach. Whether full genome sequences should represent the basis of taxonomic classification is an ongoing debate between systematicians [32]. While this consideration is well beyond the objectives of this work, obtaining a classification of bacterial genome sequences into coherent groups is of

general interest. Indeed, the number of misidentified genome sequences is exponentially growing in public databases. A number of initiatives such as Digital Protologue Database (DPD [33]), Microbial Genomes Atlas (MiGA [34]), Life Identification Numbers database (LINbase [35]) or the Genome Taxonomy Database (GTDB [31]) proposed services to classify and rename bacterial strains based on ANIb values or single copy marker proteins. Using the percentage of shared *k*-mers between unknown bacterial genome sequences and reference genome sequences associated to these databases could provide a rapid complementary approach for bacterial classification. Moreover, KI-S includes a friendly visualization interface that could help systematicians to curate whole genome databases. Indeed, zoomable circle packing could be employed for highlighting (*i*) misidentified strains, (*ii*) bacterial taxa that possess representative type strains or (*iii*) bacterial taxa that contain few genome sequences.

Association between a taxonomic group and its distribution across a range of habitats is useful for inferring the role of this taxa on its host or environment. For instance, community profiling approaches based on molecular marker such as hypervariable regions of 16S rRNA gene have been helpful for highlighting correlations between host fitness and microbiome composition. Higher taxonomic resolution of microbiome composition could be achieved with metagenomics through *k*-mer based classification of reads. Using the *Pseudomonas* genus as a use-case, we showed that increasing the breadth of genomic database without investigating the relatedness of genome sequences did not improved the proportion of classified reads. Worse, an unresolved classification may limit the number of species-specific *k*-mers identified by CLARK and therefore the number of classified reads. Interestingly, an inverse relationship between the number of genome sequences in NCBI RefSeq database and the number of classified reads at the species level was also recently highlighted with other *k*-mer-based read classifiers [36]. On the contrary, prior classification of the genomic database improve the number of classified reads at the species level. Hence, investigating the relationships between bacterial genome sequences not only benefits bacterial taxonomy but also indirectly microbial ecology.

Competing interests

The authors declare that they have neither competing interests nor conflict of interest.

Funding

This research was supported by grant awarded by the Region des Pays de la Loire (metaSEED, 2013 10080).

Acknowledgements

The authors wish to thank Claire Lemaitre and Guillaume Rizk for their assistances with the SIMKA software and Jason Shiller for manuscript assessment and for editing the English.

References

1. Parte AC. LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. Int J Syst Evol Microbiol. 2018;68:1825–9.

2. Amann R, Rosselló-Móra R. After All, Only Millions? MBio. 2016;7:e00999-16.

3. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. PNAS. 2016;113:5970–5.

4. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14:60.

5. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:81–91.

6. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA. 2009;106:19126–31.

7. 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;66:1100–3.

8. 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;110:1281–6.

9. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43:6761–71.

10. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, et al. Multiple Comparative Metagenomics using Multiset k-mer Counting. PeerJ Computer Science. 2016 ; 2:e94

11. Déraspe M, Raymond F, Boisvert S, Culley A, Roy PH, Laviolette F, et al. Phenetic Comparison of Prokaryotic Genomes Using k-mers. Mol Biol Evol. 2017;34:2716–29.

12. Deorowicz S, Gudy A, Długosz M, Kokot M, Danek A. Kmer-db: instant evolutionary distance estimation. Bioinformatics. 2019; 35: 133–136.

13. Ondov BD, Treangen TJ, Melsted P, . Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016; 17: 132.

14. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15:R46.

15. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016; 26: 1721-1729.

16. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236.

17. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software. Nat Methods. 2017;14:1063–71.

18. Peix A, Ramírez-Bahena M-H, Velázquez E. Historical evolution and current status of the taxonomy of genus *Pseudomonas*. Infect Genet and Evol. 2009;9:1132–47.

19. Berge O, Monteil CL, Bartoli C, Chandeysson C, Guilbaud C, Sands DC, et al. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. PLoS ONE. 2014;9:e105547.

20. Gomila M, Busquets A, Mulet M, García-Valdés E, Lalucat J. Clarification of Taxonomic Status within the *Pseudomonas syringae* Species Group Based on a Phylogenomic Analysis. Front Microbiol. 2017;8:2422.

21. Gomila M, Peña A, Mulet M, Lalucat J, García-Valdés E. Phylogenomics and systematics in *Pseudomonas*. Front Microbiol. 2015;6:214.

22. Peix A, Ramírez-Bahena M-H, Velázquez E. The current status on the taxonomy of *Pseudomonas* revisited: An update. Infect Genet Evol. 2018;57:106–16.

23. Garrido-Sanz D, Meier-Kolthoff JP, Göker M, Martín M, Rivilla R, Redondo-Nieto M. Genomic and Genetic Diversity within the *Pseudomonas fluorescens* Complex. PLoS ONE. 2016;11:e0150183.

24. Hesse C, Schulz F, Bull CT, Shaffer BT, Yan Q, Shapiro N, et al. Genome-based evolutionary history of *Pseudomonas* spp. Environ Microbiol. 2018; doi: 10.1111/1462-2920.14130

25. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. 2018 ; 9 : 5114.

26. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Analytical Methods. 2016 ; 8, 12-24.

27. Torres-Cortés G, Bonneau S, Bouchez O, Genthon C, Briand M, Jacques M-A, et al. Functional Microbial Features Driving Community Assembly During Seed Germination and Emergence. Front Plant Sci. 2018;9:902.

28. Patrick A Grimont. Use of DNA reassociation in bacterial classification. Canadian J Microbiol; 1988; 34:541-6.

29. Rosselló-Móra R, Amann R. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol. 2015;38:209–16.

30. Bromham L. Why do species vary in their rate of molecular evolution? Biol Lett. 2009;5:401–4.

31. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotech. 2018;36:996–1004.

32. Garrity GM. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? J Clinic Microbiol. 2016;54:1956–63.

33. Rossello-Mora R, Sutcliffe IC. Reflections on the introduction of the Digital Protologue Database — A partial success? System Appl Microbiol. 2019; 42:1-2.

34. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 2018;46:W282–8.

35. Vinatzer BA, Tian L, Heath LS. A proposal for a portal to make earth's microbial diversity easily accessible and searchable. Antonie van Leeuwenhoek. 2017;110:1271–9.

36. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of *k*-mer-based lowest common ancestor species identification. Genome Biol. 2018; 19, 165.

Figures and Supplemental files

Figure 1: Distribution of shared *k*-mers values. Relatedness between genome sequences were estimated with ANIb (green) or shared *k*-mers (blue). The *x* axis represents ANIb or percentage of shared *k*-mers while the *y* axis represents the number of values by class in the subset of 934 *Pseudomonas* genomic comparison.

Figure 2: Comparison of average pairwise similarity between genomes sequences. Overall genome relatedness indexes of 934 *Pseudomonas* genomes sequences were calculated with PYANI [26] and four different *k*-mers indexing softwares : FastANI [25], Simka [10], Kmer-db [12] and MASH [13].

Figure 3: Clustering of *Pseudomonas* genome sequences. Zoomable circle packing representation of *Pseudomonas* genome sequences (*n* = 3,623). Similarities between genome sequences were assessed by comparing the percentage of shared 15-mers. Each dot represents a genome sequence, which is colored according to its group of species [19,24]. These genome sequences have been grouped at three distinct thresholds for assessing infraspecific (0.75), species-specific (0.5) and interspecies relationships (0.25).

Figure 4: Percentage of classified reads. Classification of metagenome read sets derived from bean seeds, germinating seeds and seedlings with Clark [16]. Three distinct databases were employed for read classification: the original Clark database (red), Clark database supplemented with 3,623 *Pseudomonas* genome sequences (green) and the Clark database supplemented with 3,623 *Pseudomonas* genome sequences that were classified according to their percentage of shared *k*-mers (blue).

Table 1 : Comparison of softwares that estimate genome relatedness.

Software	Time	RAM	Reference
PYANI	3 months	ND	26
FastANI	11 days	15 Go	25
Simka	4 hours	18 Go	10
Kmer-db	40 minutes	25 Go	12
Mash	7 minutes	26 Mo	13

TableS1.csv : *Pseudomonas* groups. Description of the 350 groups obtained after clustering at 50% of shared 15-mers. For each group, the *Pseudomonas* group [24] and subgroup [19,24] are displayed.

FigureS1.html: Zoomable circle packing representation of *Pseudomonas* genome sequences. Similarities between genome sequences were assessed by comparing the percentage of shared 15-mers. Each dot represents a genome sequence, which is colored according to its group of species [19,24]. These genome sequences have been grouped at three distinct thresholds for assessing infraspecific (0.75), species-specific (0.5) and interspecies relationships (0.25).

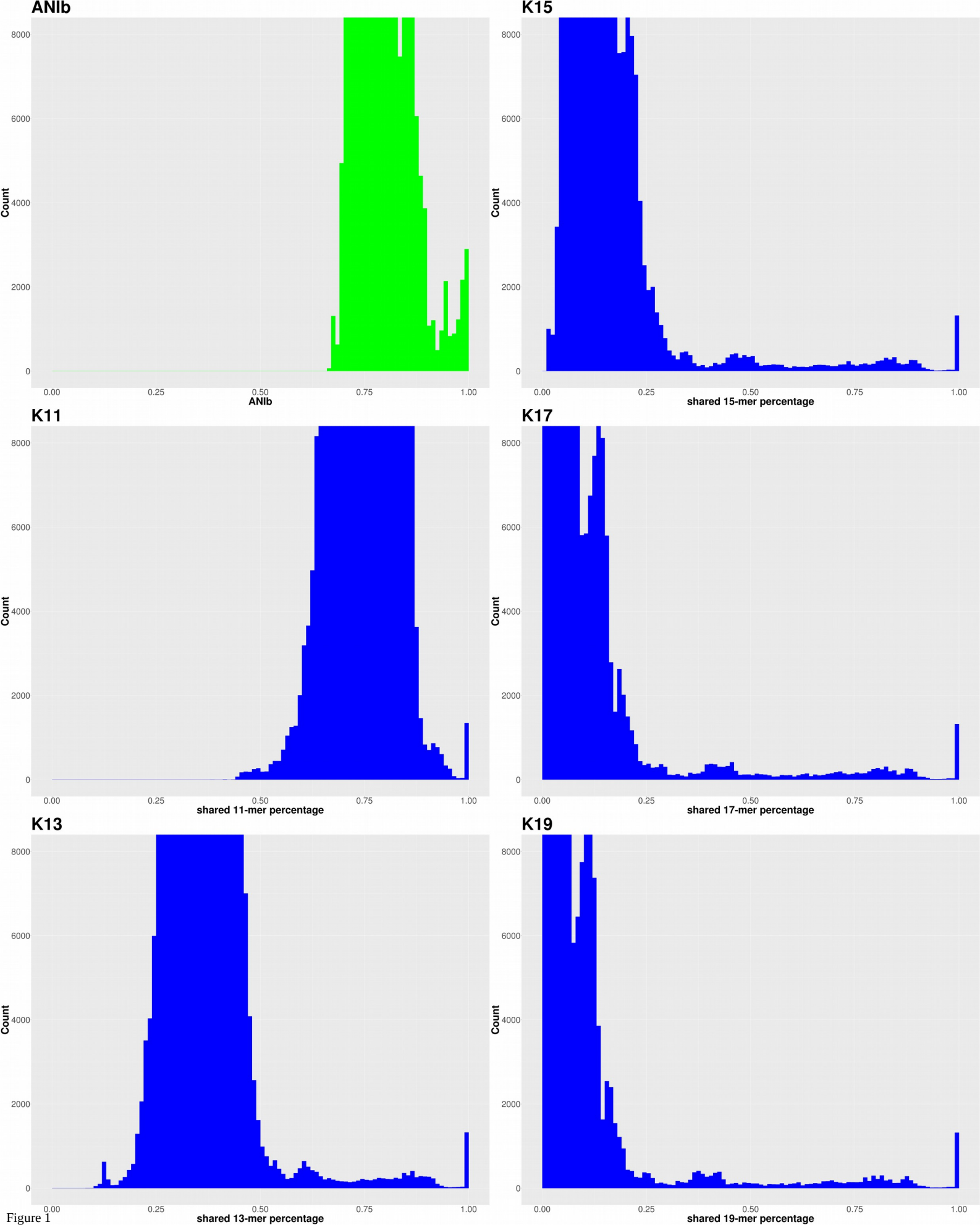
ANIb
K15
K11
K17
K13
K19
Count
8000
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
ANIb
shared 15-mer percentage
shared 11-mer percentage
shared 17-mer percentage
shared 13-mer percentage
shared 19-mer percentage

Figure 1

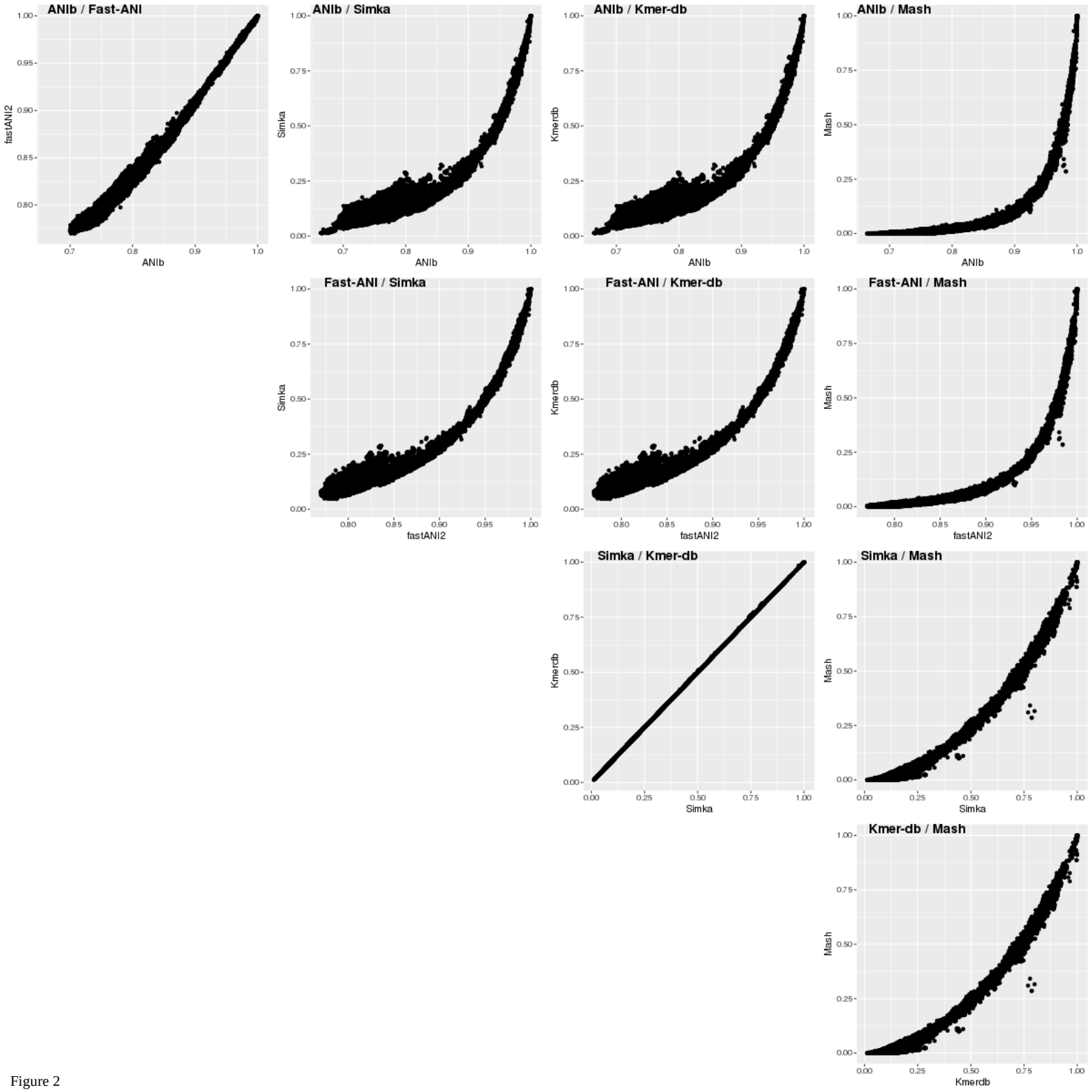
ANIb / Fast-ANI
fastANI2
ANIb
ANIb / Simka
Simka
ANIb
ANIb / Kmer-db
Kmerdb
ANIb
ANIb / Mash
Mash
ANIb
Fast-ANI / Simka
Simka
fastANI2
Fast-ANI / Kmer-db
Kmerdb
fastANI2
Fast-ANI / Mash
Mash
fastANI2
Simka / Kmer-db
Kmerdb
Simka
Simka / Mash
Mash
Simka
Kmer-db / Mash
Mash
Kmerdb

Figure 2

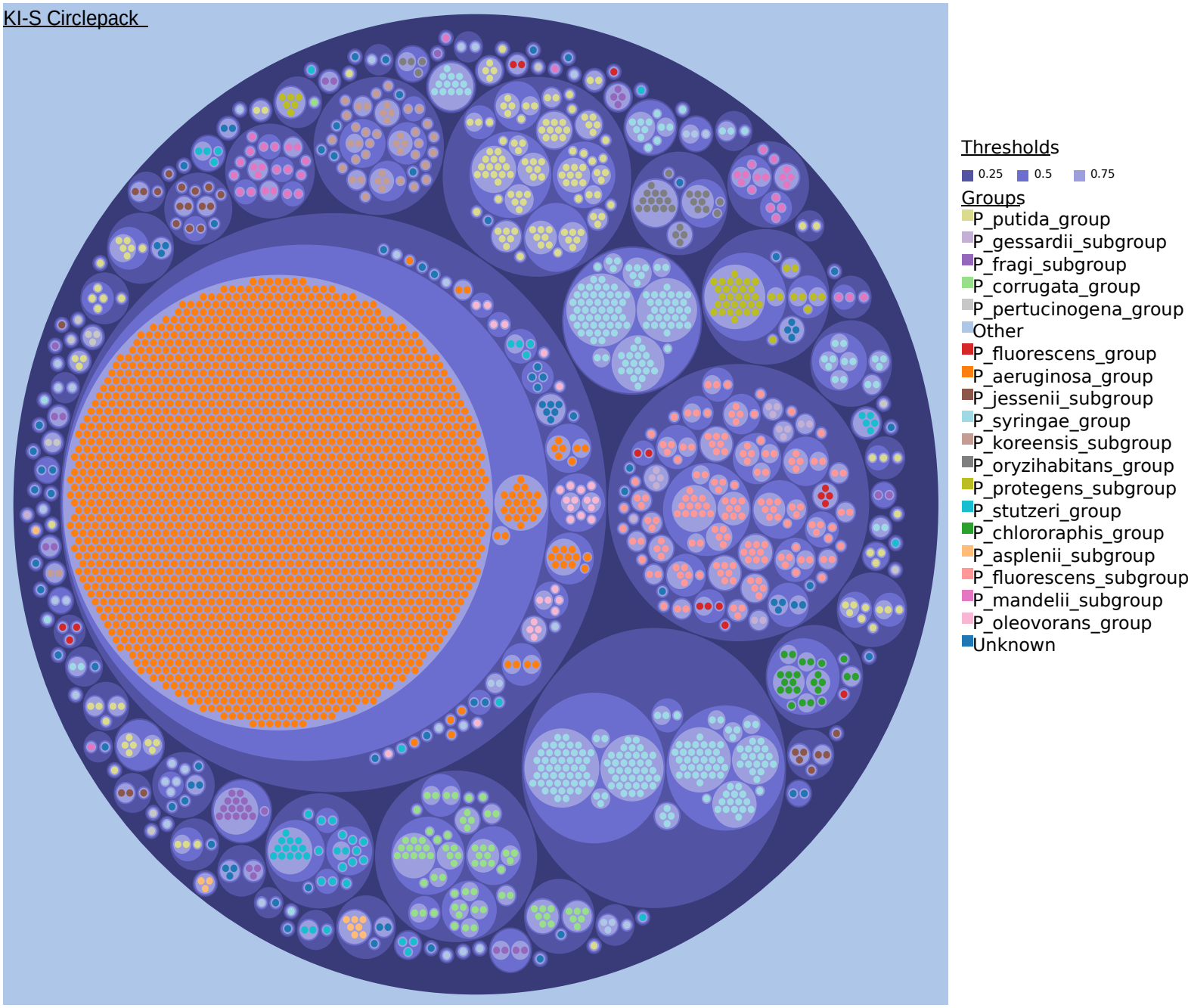
KI-S Circlepack
Thresholds
0.25
0.5
0.75
Groups
P_putida_group
P_gessardii_subgroup
P_fragi_subgroup
P_corrugata_group
P_pertucinogena_group
Other
P_fluorescens_group
P_aeruginosa_group
P_jessenii_subgroup
P_syringae_group
P_koreensis_subgroup
P_oryzihabitans_group
P_protegens_subgroup
P_stutzeri_group
P_chlororaphis_group
P_asplenii_subgroup
P_fluorescens_subgroup
P_mandelii_subgroup
P_oleovorans_group
Unknown

Figure 3

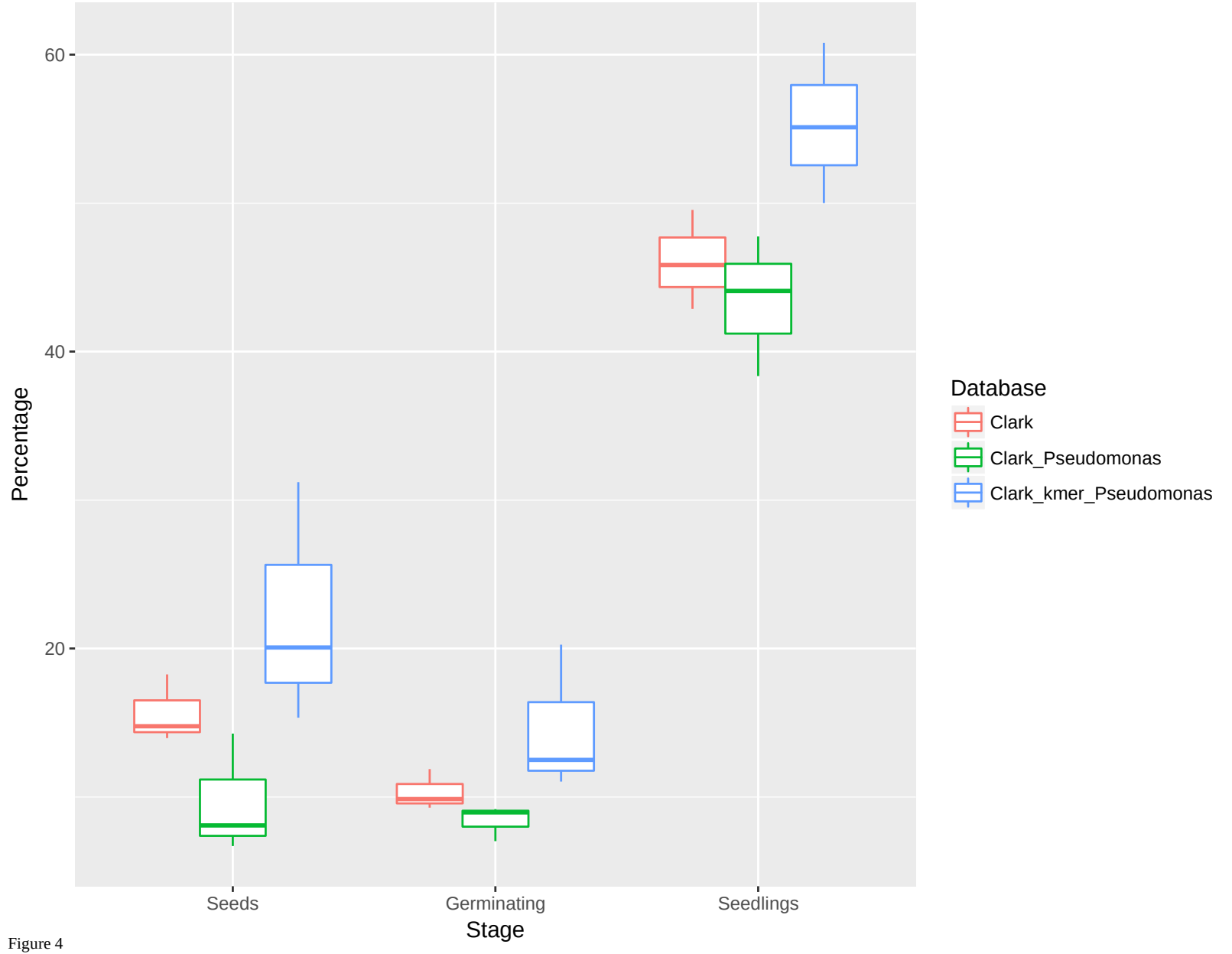
60
40
20
Percentage
Seeds
Germinating
Seedlings
Stage
Database
Clark
Clark_Pseudomonas
Clark_kmer_Pseudomonas

Figure 4