A computational pipeline for species- and strain-level classification of metagenomic sequences

Arthur Henrique Barrios Solano [1], **João Carlos Setubal** [12]

[1]Inter-Departmental Graduate Program in Bioinformatics
Universidade de São Paulo (USP) - São Paulo, SP - Brazil

[2]Department of Biochemistry, Institute of Chemistry
Universidade de São Paulo (USP) - São Paulo, SP - Brazil

arthur.barrios@usp.br, setubal@iq.usp.br

Abstract. *We present a pipeline for exploring genomic diversity in metagenomic datasets at the species and strain levels. To achieve accurate classifications independent of taxonomy labels, we introduce the concept of Genome Reference Set (GRS), modeled using the Maximal Independent Set problem for undirected graphs. For a given user-defined target genus, we build its GRS from GenBank genomes and use it for metagenomic contig classification using BLASTn. Additional phylogenetic processing allows the identification of putative novel species. We show that our pipeline can achieve better results than general-purpose tools, and apply the pipeline to the MetaSUB dataset, identifying two putative novel strains and one putative new species of* Acinetobacter.

1. Introduction

Most bacterial and archaeal species have not been cultivated in laboratories [Steen et al. 2019]. This means that the vast majority of microbes remain unknown (the so-called Microbial Dark Matter). The use of metagenomics techniques has contributed to shed considerable light into this Microbial Dark Matter. Over the last 20 years, a large number of metagenomic datasets have been generated, and today there are several thousands of publicly available shotgun metagenomic datasets sampled from different environments and hosts [Nayfach et al. 2021]. However, extracting accurate information about the presence and relative abundance of microbial taxa in metagenomic datasets containing millions of sequences is still a challenging task. One evidence is the competition Critical Assessment of Metagenome Interpretation (CAMI) [Meyer et al. 2022], which offers an opportunity for researchers to compare software for taxonomic classification using controlled metagenomic datasets.

Here we propose a novel pipeline for taxonomic classification of metagenomic contigs. Our aim is to provide a tool that provides more accurate results at the species and strain levels than general purpose classification tools such as Kraken2 [Wood et al. 2019] or MMseqs2 [Steinegger and Söding 2017]. We assume the user of our tool has a predefined list of target genera or target species, and he or she wants to determine the presence and relative abundance of species from those target genera or strains from target species in metagenomic datasets of interest. In its present state, the tool is capable of analyzing contigs only, that is, pre-assembled metagenomic datasets of raw reads.

The paper is structured as follows. In Section 2 we present details of the pipeline. In Section 3.1 we compare the performance of using BLAST [Altschul et al. 1997] with

the criteria we employ in the pipeline with the programs Kraken2 [Wood et al. 2019] and MMseqs2 [Steinegger and Söding 2017]. In Section 3.2 we present results running the pipeline on the MetaSUB dataset.

2. Methods

We assume the input to the pipeline is a list of target genera and a single file in FASTA format containing contig sequences originated from the assembly of metagenomic reads, from one or more samples. As a pre-processing step, we remove from the file any sequences shorter than 500 bp.

2.1. Genome Reference Set

For each target genus, we build a Genome Reference Set (GRS) obeying the following rule: the GRS should contain at least one representative of every species of the target genus with a publicly available genome sequence, subject to the following restriction: we do not include draft genomes with more than 100 contigs, reasoning that more fragmented genomes have lesser quality and therefore may negatively impact our classification accuracy.

Our method for GRS construction also includes the requirement that it should be as nonredundant as possible in terms of genome similarity (which means that we do not rely on taxonomy labels to reduce eventual redundancy, an important feature of our methodology). We model this redundancy minimization problem using the Maximum Independent Set (MIS) problem, as follows. Given an undirected graph $G(V, E)$, the MIS is a subset of nodes $V' \subseteq V$ such that: $\forall u, v \in V', (u, v) \notin E$; and V' has maximum size. MIS is an NP-hard problem [Garey and Johnson 1979]. However, if we relax the problem so that we require V' to be of maximal size instead of maximum size, a solution can be efficiently found using a greedy algorithm. One such greedy algorithm was proposed by Luby [Luby 1985], originally developed for distributed processing. Luby's algorithm finds the Maximal Independent Set (MLIS) in $O(\log n)$ iterations (where $n = |V|$), and can be written as shown in Algorithm 1.

The construction process of the GRS for a given genus is illustrated in Figure 1. All available genomes in GenBank are downloaded; genomes with more than 100 contigs are removed. Genomes are then compared against each other using fastANI [Jain et al. 2018]. After that, a genome graph is built according to the following rules: (a) each genome is a node; and (b) if two genomes have ANI score greater or equal to 98%, they form an edge. Our implementation of Luby's algorithm as shown in Algorithm 1 then reduces the genome set to an MLIS of genomes. Finally, the genome of every species not represented in the MLIS (having been excluded from the MLIS by Luby´s Algorithm) is added back to the set, to ensure that the GRS does have every species represented. This final set is the GRS.

Data: Undirected graph: $G = (V, E)$
Result: Maximal Independent Set (MLIS): $I \subseteq V$
$I \leftarrow \emptyset$;
$G' = (V', E') \leftarrow G = (V, E)$;
while $V' \neq \emptyset$ **do**
  $I' \leftarrow \emptyset$ /* Temporary Independent Set */
  for $v \in V'$ **do**
    $r_v \leftarrow$ random number $\in [0, 1]$
  end
  for $v \in V'$ **do**
    $N(v) \leftarrow \{u \in V' \setminus \{v\} | (u, v) \in E'\}$;
    if $r_v > r_u, \forall u \in N(v)$ **then**
      add v to I'
    end
  end
  $I \leftarrow I \cup I'$;
  $Y \leftarrow I' \cup N(I')$;
  $G' = (V', E') \leftarrow G[V' \setminus Y]$ the subgraph induced by $V' \setminus Y$
end

Algorithm 1: Pseudo-code for Luby's Algorithm

Note that the GRS may have pairs of genomes (a, b) whose ANI value is greater than or equal to 98%; the existence of such pairs means that a and b belong to different species (or at least have been labeled as such) and yet their ANI value suggests that they should belong to the same species. Conversely, the GRS may contain genomes a and b such that a and b belong to the same species and yet their ANI value is less than 98%. We consider these features to be important properties of the GRS, making it less dependent on taxonomic labels.

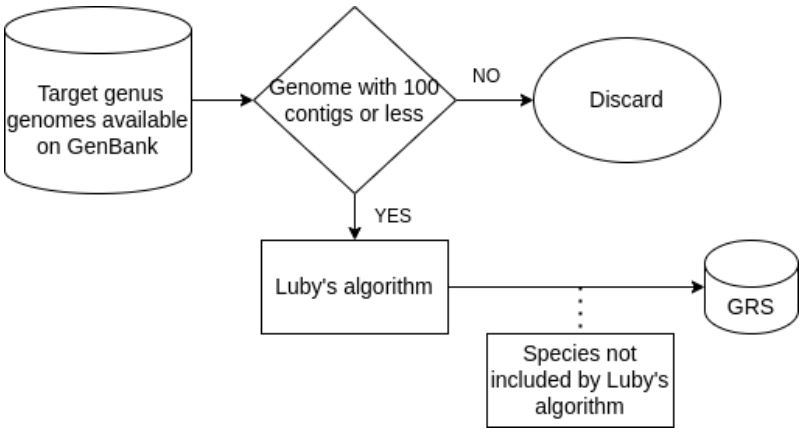

Figure 1. GRS construction workflow.

2.2. Identification Pipeline

Once the GRS is built, contig classification is done as illustrated in Figure 2 (upper path). The metagenomic dataset is searched with BLASTn against the GRS. Only the first hit is considered. If the first hit has query coverage at least 80% and e-value $\leq 10^{-10}$ then

it is classified according to the percent identity (PI) value. Sequences with PI $\geq 98\%$ are added to the "known species" set of contigs, with the same species label of the first BLAST hit. Sequences with PI $< 98\%$ are added to the "putative new species" set, without any taxonomic label.

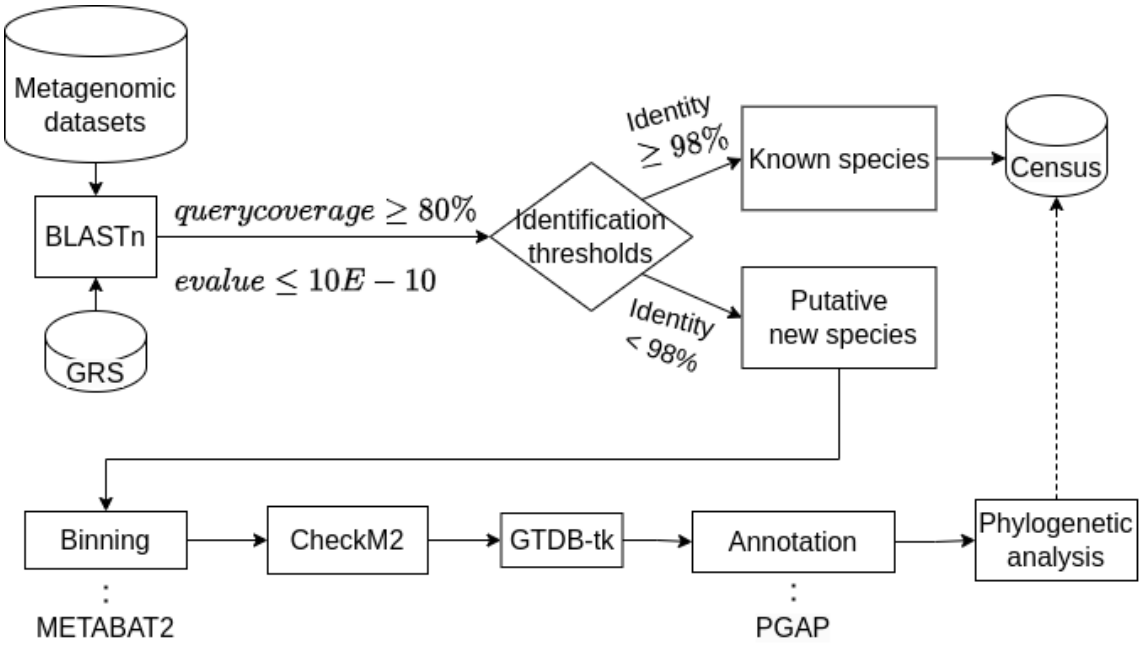

Figure 2. Identification pipeline workflow.

The known species set constitutes the catalog of known species in the input metagenomic dataset, possibly incremented with additional elements depending on the results of the next step, represented by the lower path in Figure 2. In that path, the contigs are clustered with the MetaBAT2 binning program [Kang et al. 2019]. Then, all resulting bins are evaluated in terms of completeness and contamination with CheckM2 [Chklovski et al. 2023].

Because the input metagenomic dataset may contain DNA from several separate samples, we have found it necessary to separate contigs in a bin by originating sample. We call the new bins obtained in this way *sub-bins*, and the process is illustrated in Figure 3. Note that the number of originating samples in bins varies according to bin. Sub-bins are classified with GTDB-tk [Chaumeil et al. 2022]. Sub-bins classified as a known species are assigned to the known species set (rightmost vertical arrow in Figure 2). Otherwise, the sequences belonging to sub-bins with at least 50% completeness are annotated with PGAP [Tatusova et al. 2016] and phylogenetically analyzed.

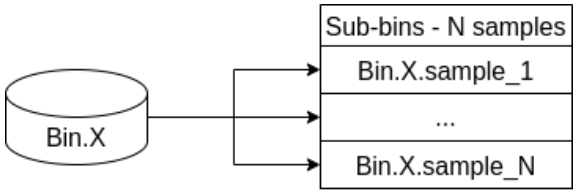

Figure 3. Sub-binning process illustration considering a pool of N samples.

For each sub-bin obtained that passes the completeness threshold, a phylogenetic tree is inferred using the core-genome of the sub-bin and representative genomes of the target genus, selected from the GRS. Orthologous genes are determined with GetHomologues [Contreras-Moreira and Vinuesa 2013], using the OrthoMCL [Li et al. 2003] algorithm and coverage and identity thresholds of 80%. Orthologous genes are aligned with MAFFT [Katoh et al. 2002] with a maximum of 1,000 iterations. Alignment regions with gaps are manually removed and the final alignments are concatenated. The phylogenetic tree is then constructed with IQtree2 [Minh et al. 2020], using the substitution model defined by ModelFinder [Kalyaanamoorthy et al. 2017] and 1,000 bootstrap replicates.

3. Results

3.1. Classification performance comparisons

Comparisons between BLASTn classification and taxonomic classification tools Kraken2 and MMseqs2 were performed on marine and plant-associated datasets from CAMI2 challenges, available at (https://frl.publisso.de/data/frl:6425521/). For each tool, the classifications were based on local databases built from GRS sequences. For MMseqs2, two different modules were tested: easy-taxonomy (using the Lowest Common Ancestor approach) and easy-search (using the local alignment approach). The datasets were downloaded and filtered for contigs with a minimum length of 500 bp. Tests were run on two genera of interest: *Acinetobacter* (from the marine dataset) and *Stenotrophomonas* (from the plant-associated dataset). The performances were evaluated and analyzed at the species level, using specificity, precision, recall, accuracy, and F1 score as metrics, defined below:

$$specificity = \frac{TN}{TN + FP} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively. A classification was considered "positive" if it reached the species level. If the classification did not reach the species level or the contig remained unclassified, it was considered "negative". Classification at genus level or unspecified species (sp.) level were also considered "negative" classifications.

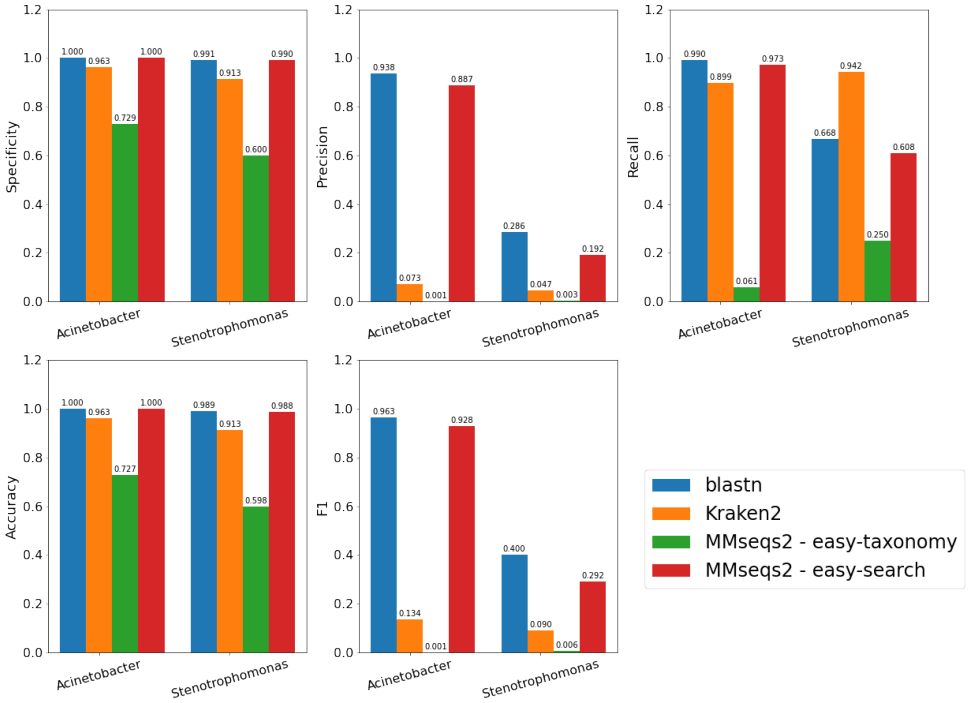

Figure 4. Specificity, precision, recall, accuracy and F1 comparisons between BLASTn, Kraken2 and MMseqs2 performances on CAMI2 marine and plant-associated datasets. Custom databases based on GRS of *Acinetobacter* and *Stenotrophomonas*.

Results are shown in Figure 4. The BLASTn method outperformed Kraken2 and MMseqs2 across all metrics, except recall for the *Stenotrophomonas* genus. Although MMseqs2 (with easy-search module) produced competitive results, BLASTn (running at a slightly slower speed) consumed less "on process" disk space (an advantage for large datasets searches) and performed better overall. Therefore, we found that BLASTn alignments are the most suitable for the pipeline here described.

3.2. Pipeline results for the MetaSUB dataset

The MetaSUB dataset [Danko et al. 2021] is a highly diverse metagenomic dataset consisting of 4,728 samples collected from trains and subway stations across 60 cities worldwide, with the objective of characterizing and exploring the urban microbiome diversity. For the purposes of this work we used an assembled version of the dataset (i.e., a set of contigs). We removed contigs shorter than 500 bp; the resulting set has 79,742,596 contigs, and a total length of 96,865,804,315 bp.

Five target genera were selected among the genera previously identified in the MetaSUB dataset. The list of target genera is: *Xanthomonas*, *Stutzerimonas*, *Moraxella*, *Acinetobacter* and *Stenotrophomonas*.

For each target genus, its GRS was built, with relevant statistics shown in Table 1.

Table 1. GRS numbers for each target genus.

Genus	# RefSeq genomes	# genomes in GRS	# GRS / # RefSeq (%)
Xanthomonas	2,734	192	7
Stutzerimonas	411	148	36
Moraxella	394	81	21
Acinetobacter	12,912	945	7.3
Stenotrophomonas	1,520	190	12.5

The pipeline was run and resulted in the "known species" and "putative new species" sets shown in Table 2.

Table 2. Number of contigs for the sets "known species" and "putative new species".

Genus	# contigs	
	Known species	putative new species
Xanthomonas	76,020	1,081,123
Stutzerimonas	908,548	1,401,992
Moraxella	286,964	391,385
Acinetobacter	1,243,548	659,844
Stenotrophomonas	529,810	916,272

A survey of bacterial species for the target genera was conducted, yielding values of relative abundance. Table 3 shows the most abundant species for each target genus explored. Despite the high relative abundance of certain species, there is a significant diversity within this taxonomic classification that remains hidden. Our method can uncover this hidden diversity as exemplified by results for *Acinetobacter*. In the case of this genus, the most abundant species is *Acinetobacter lwoffii* and our pipeline yielded the result that it has a taxonomic diversity of five strains in the MetaSUB dataset: *Acinetobacter lwoffii*, *Acinetobacter lwoffii* NIPH 715, *Acinetobacter lwoffii* ATCC 9957, *Acinetobacter lwoffii* NIPH 478, and *Acinetobacter lwoffii* NCTC 5866. However, our results go even further, as shown in Table 4, revealing a hidden diversity of 26 non-redundant genomes, with any pair of genomes not having more than 98% pairwise ANI.

Table 3. Most abundant taxa for the target genera and their relative abundance. in MetaSUB dataset.

Genus	Most abundant Taxa	Relative abundance (%)
Xanthomonas	*Xanthomonas campestris*	77.8
Stutzerimonas	*Stutzerimonas stutzeri*	65.5
Moraxella	*Moraxella osloensis*	69.3
Acinetobacter	*Acinetobacter lwoffii*	13.0
Stenotrophomonas	*Stenotrophomonas maltophilia*	68.9

Table 4. Genomes diversity found for *Acinetobacter lwoffii* taxon in MetaSUB dataset.

Accession Number	Taxon	Relative Abundance (%)
GCF_000369125.1	*Acinetobacter lwoffii ATCC 9957*	0,31
GCF_900444925.1	*Acinetobacter lwoffii*	0,25
GCF_019787625.1	*Acinetobacter lwoffii*	0,24
GCF_019048525.1	*Acinetobacter lwoffii*	0,57
GCF_022967985.1	*Acinetobacter lwoffii*	0,37
GCF_024129555.1	*Acinetobacter lwoffii*	0,38
GCF_019787645.1	*Acinetobacter lwoffii*	1,28
GCF_031455115.1	*Acinetobacter lwoffii*	0,49
GCF_000369145.1	*Acinetobacter lwoffii NIPH 478*	0,58
GCF_009730095.1	*Acinetobacter lwoffii*	1,65
GCF_013349205.1	*Acinetobacter lwoffii*	0,63
GCF_022809915.1	*Acinetobacter lwoffii*	0,6
GCF_000369105.1	*Acinetobacter lwoffii NCTC 5866*	0,08
GCF_024129855.1	*Acinetobacter lwoffii*	0,58
GCF_019343495.1	*Acinetobacter lwoffii*	0,53
GCF_963518635.1	*Acinetobacter lwoffii*	0,59
GCF_002321025.1	*Acinetobacter lwoffii*	0,93
GCF_012393445.1	*Acinetobacter lwoffii*	0,23
GCF_024129635.1	*Acinetobacter lwoffii*	0,58
GCF_035788095.1	*Acinetobacter lwoffii*	0,28
GCF_000368165.1	*Acinetobacter lwoffii NIPH 715*	0,19
GCF_024129435.1	*Acinetobacter lwoffii*	0,38
GCF_024129685.1	*Acinetobacter lwoffii*	0,34
GCF_963516025.1	*Acinetobacter lwoffii*	0,33
GCF_015602705.1	*Acinetobacter lwoffii*	0,27
GCF_024129715.1	*Acinetobacter lwoffii*	0,27

The processing of the "putative new species" set generated sub-bins, with completeness distributions shown in Figure 5. *Acinetobacter* was the target genus with best results, showing several sub-bins with completeness greater than 50% and contamination under 10%. Three sub-bins are shown in Table 5, along with completeness and contamination values assessed using CheckM2, and their taxonomy assignment performed with GTDB-tk. These three sub-bins were phylogenetically analyzed, as shown in Figure 6. The analysis includes 92 genomes from the *Acinetobacter* genus.

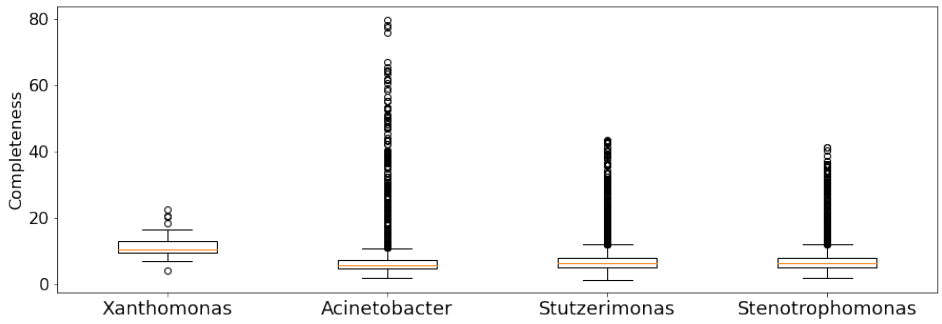

Figure 5. Completeness box plots of sub-bins generated for the genera *Xanthomonas*, *Acinetobacter*, *Stenotrophomonas* and *Stutzerimonas*.

Table 5. *Acinetobacter* sub-bins SB01-A, SB02-A and SB03-A. Completeness and contamination analysis executed with CheckM2 and taxonomic classification performed with GTDB-tk.

Sub-bin code	Completeness	Contamination	GTDB-tk classification
SB01-A	53,09	4,59	g__Acinetobacter;s__
SB03-A	58,37	6,03	g__Acinetobacter;s__
SB02-A	53,3	7,46	g__Acinetobacter;s__

The sub-bins were also processed with fastANI against the GRS of *Acinetobacter*. The results are that SB01 and SB03 have ANI score above 95% with genomes of *Acinetobacter variabilis* and *Acinetobacter lwoffii*, respectively. However, for SB02, no genome reached 95% ANI score, with *Acinetobacter lwoffii* being the closest one at 94.54%. These findings suggest that SB02 potentially represents a new species, while SB01 and SB03 probably are new strains of known species of *Acinetobacter*.

4. Conclusions

The results presented show that the pipeline here described can uncover hidden genomic diversity that would otherwise remain hidden if only general-purpose taxonomic classification tools are used. This hidden diversity can be classified in three categories: 1) the tool can show the presence in the samples of many separate strains for a given species; 2) it can show the presence of new strains; and 3) it can find putative new species for a given genus.

Although common tools for taxonomy classification, such as Kraken2 and MMseqs2, can in principle recover some of this same diversity, in order to do so they would have to use the appropriate GRS, which is itself a contribution of this work.

For future work, we plan to automate certain steps of the pipeline, so that it can be automatically run from beginning to end; and then run the pipeline on additional datasets, to obtain more presence and relative abundance results relative to the target genera mentioned here.

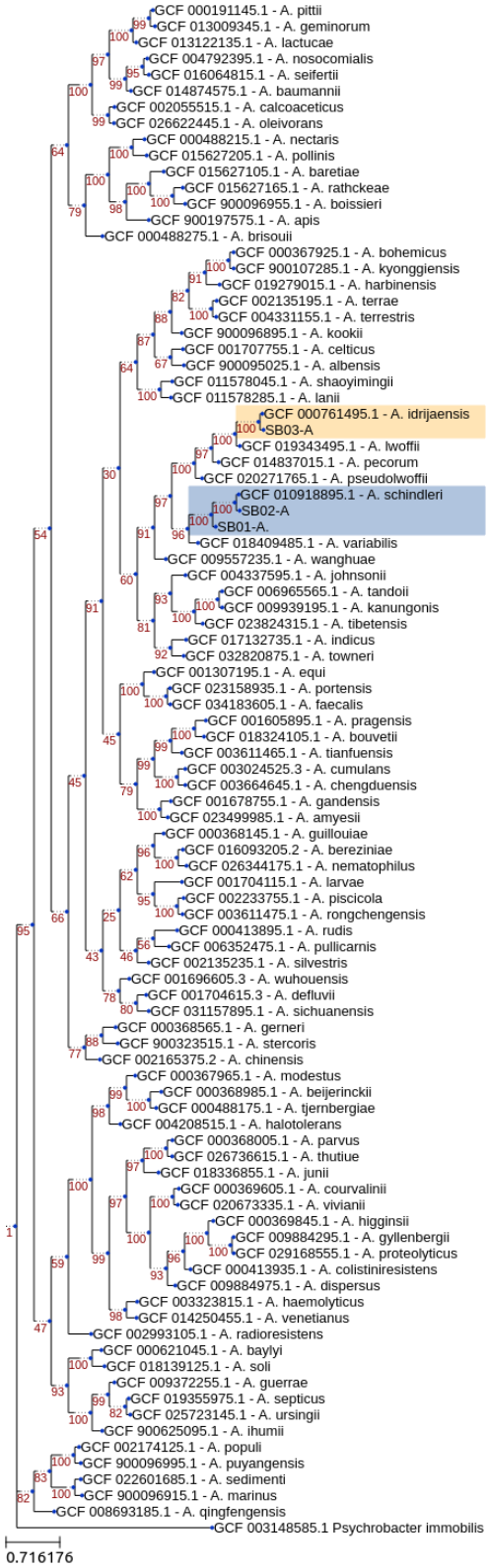

Figure 6. Phylogenetic tree for *Acinetobacter* sub-bins SB01A, SB02A and SB03A. *Psychrobacter immobilis* genome used as outgroup.

5. Acknowledgements

This work was made possible in part by a grant from CNPq (award #440230/2022-5) and by a FAPESP PhD fellowship to A.H.B.S. (award #2024/01729-9).

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2022). Gtdb-tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23):5315–5316.

Chklovski, A., Parks, D. H., Woodcroft, B. J., and Tyson, G. W. (2023). Checkm2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8):1203–1212.

Contreras-Moreira, B. and Vinuesa, P. (2013). Get_homologues, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology*, 79(24):7696–7701.

Danko, D., Bezdan, D., Afshin, E. E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., Donnellan, D., Hecht, J., Jackson, K., et al. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, 184(13):3376–3393.

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability*, volume 174. freeman San Francisco.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature communications*, 9(1):5114.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., and Jermiin, L. S. (2017). Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189.

Luby, M. (1985). A simple parallel algorithm for the maximal independent set problem. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 1–10.

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., et al. (2022). Critical assessment of metagenome interpretation: the second round of challenges. *Nature methods*, 19(4):429–440.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021). A genomic catalog of earth's microbiomes. *Nature biotechnology*, 39(4):499–509.

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., and Thrash, J. C. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, 13(12):3126–3130.

Steinegger, M. and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. (2016). Ncbi prokaryotic genome annotation pipeline. *Nucleic acids research*, 44(14):6614–6624.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13.