Integrated genome and gene diversity analyses of bacterial genera in metagenomic data

Mariana Louise Gabas[1], **Arthur Henrique Barrios Solano**[1], **João Carlos Setubal**[1]

[1]Institute of Chemistry, University of São Paulo, São Paulo, Brasil

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

Abstract. *We present an analysis of subspecies diversity of the bacterial genera Xanthomonas, Acinetobacter, and Stenotrophomonas present in the MetaSUB metagenomic dataset. We used the simulated rarefaction curve technique along with the non-parametrics estimators Chao1, Chao2, Jackknife, ICE, and ACE, to estimate the total number of subspecies (subgroups) within the MetaSUB dataset. Subgroup is an operational concept that can be used to distinguish between genomes from the same species. For Xanthomonas, our results suggest that MetaSUB may have more subgroups than currently known; for Acinetobacter, the number of estimated subgroups is lower than the known number; and for Stenotrophomonas, the estimates and the known number are the same. A comparison of the pangenomes of Xanthomonas obtained from our genome database and that from MetaSUB showed that the genic diversity in MetaSUB is much lower than the database diversity, suggesting that urban environments constrain the genic diversity compared to the biosphere as a whole.*

1. Introduction

The concept of *Microbial Dark Matter* [Rinke et al., 2013] is based on the observation that the vast majority of bacterial and archaeal species in the biosphere have not yet been cultivated in the laboratory [Steen et al., 2019]. The investigation of the biodiversity of these bacteria became possible thanks to metagenomics. The metagenomic surveys that have been conducted over the last 20 years have shed some light on microbial dark matter, and at the same time generated a large amount of data. Therefore, there is an opportunity to explore such datasets to obtain new knowledge. Among the possible explorations, we consider that the survey of the biodiversity of a given bacterial genus, as well as the exploration of its gene diversity, can bring significant contributions to scientists interested in that genus.

Due to space constraints, we present results from a single metagenomic dataset and three target genera. The metagenome dataset that we explored is MetaSUB [Danko et al., 2021], and is composed of 4,728 samples collected from train and subway stations in 60 cities around the world, with the aim of characterizing and exploring the diversity of the urban microbiome. The number of contigs with a size of at least 500 bp in this dataset is almost 80 million. We understand the MetaSUB dataset as a model for global datasets; the same methodology described here could be applied to other metagenomic datasets of large geographic scale (such as TARA Oceans [Sunagawa et al., 2020]).

The chosen genera (which we call the *target genera*) are *Xanthomonas, Acinetobacter,* and *Stenotrophomonas*. These choices were motivated mainly by the fact that these genera are already known to be abundant in the MetaSUB project samples [Danko et al., 2021]. The *Xanthomonas* biodiversity explored in the metagenomic data was complemented by the reconstruction of the pangenome, which is defined as the set of all gene families present in a given group of genomes belonging to a specific taxon [Tettelin et al., 2008].

2. Methodology

2.1. Defining the genome collection

In order to classify metagenomic sequences, we used a Genome Reference Set (GRS) [Solano & Setubal (2024)]. For each target genus, the number of genomes contained in the GRS and the corresponding number of species are shown in Table 1. The GRS is an economical way to represent the genome diversity of a given target genus; at the same time, it gives us a certain independence from taxonomic labels (which may be wrong or vary over time) and allows us to distinguish between genomes of the same species at the subspecies level. We use the term *subgroups* to refer to genomes from the same species that share greater similarity amongst each other than with other genomes of the same species. Each genome in the GRS belongs to a subgroup, i.e. could be grouped into a cluster of genomes that on average share 98% ANI to each other; every subgroup of the genus is represented by at least one genome in the GRS. This structure allows for classifying contigs with greater resolution while preserving the full extent of the known genome diversity. Contig taxonomic classification was done by the tool described by Solano and Setubal (2024).

Table 1: Genomes and species for each target genus

Target genus	Total species	Number of genomes in the GRS
Xanthomonas	46	226
Acinetobacter	93	1010
Stenotrophomonas	34	183

2.2. Subgroup estimation

We used the simulated sampling method to obtain estimates for the total number of subgroups for a given target genus and the MetaSUB contigs. This method is based on the methodology described by Gotelli, N. & Colwell, Robert (2011), and it uses non-parametric estimators to obtain total diversity values. Since the data used in this study belong to a single database (MetaSUB), there is the limitation that the inference of subgroup diversity can only be made for the environments sampled by MetaSUB.

In order to estimate the total diversity of the target genera sampled in the MetaSUB set, we used the rarefaction curve technique, as well as total richness estimators (estimated

asymptotic value of the rarefaction curve). There are different ways to obtain the asymptotic value, described and compared in the literature [Gotelli & Colwell, 2011]. According to them, the estimation of the asymptotic value can be done with parametric estimators (asymptotic curve fitting and extrapolation) or nonparametric estimators. However, comparative analyses between the two techniques indicate that parametric methods perform worse than non-parametric ones. This is due to the following reason: different asymptotic curves can be fitted to the same species accumulation curve (or subgroups), arriving, however, at variable conclusions about the asymptotic value.

Therefore, this work aims to construct rarefaction curves and obtain from them the total richness values with non-parametric estimators. We used two types of estimators, considering the differences between the types of data used in the analysis, dividing them into *incidence data* (with information on the presence of a given subgroup in the sample) and *abundance data* (with information on the relative abundances of the subgroups found in the sample). The estimators used were: Chao1, ACE and Jackknife (for abundance data) and Chao2, ICE and Jackknife (for incidence data) [Gotelli & Colwell, 2011 and Elena Schmitz & Rahmann, 2025].

2.3. Pangenome reconstruction for *Xanthomonas*

The tool chosen to reconstruct the pangenome of *Xanthomonas* was Panaroo version 1.5.2 [Tonkin-Hill et al., 2020]. The reconstruction of the pangenome was performed from the list of genomes present in the GRS and the set of MetaSUB contigs classified as *Xanthomonas* by the tool described by Solano and Setubal (2024). The genomes and the contigs were annotated with Prokka [Seemann, 2014].

3. Results and Discussion

3.1. Subgroup estimation

The observed and estimated subgroup values in the MetaSUB dataset for the target genera are shown in Table 2.

Table 2: Subgroup estimates for target genera in MetaSUB.

			Abundance Estimators				**Incidence Estimators**			
Target genus	**GRS subgroups**	**observed subgroups**	**Chao1**	**Jackknife**		**ACE**	**Chao 2**	**Jackknife**		**ICE**
				1	2			1	2	
Xanthomonas	210	203	217	216	223	210	217	216	222	209
Acinetobacter	807	793	796	803	795	797	796	803	795	795
Stenotrophomonas	170	170	170	170	170	170	170	170	170	170

Of 46 *Xanthomonas* species present in its GRS, we found 42 in the MetaSUB dataset. This is surprising, given that this genus is known to be mostly plant-associated [An et al., 2019]. On the other hand, Table 2 shows that 96.7% (203/210) of the *Xanthomonas* subgroups were observed. The estimated numbers of subgroups suggest that, if additional MetaSUB samples were obtained, between 210 and 223 *Xanthomonas* subgroups would be found, indicating a minimum detection rate with current samples of 91.0% (203/223). These results suggest that new *Xanthomonas* subgroups could be found in urban environments, but it appears unlikely that new *Xanthomonas* species could be detected, even with additional sampling.

Of 93 *Acinetobacter* species present in its GRS, we found 92 in the MetaSUB dataset (in 1,552,960 contigs). In terms of subgroups, we found 98.3% (793/807). The number of estimated subgroups present in the dataset reached a maximum of 803, below the total in the GRS, suggesting that the urban environments sampled by MetaSUB have a lower subgroup diversity than environments in general, for the *Acinetobacter* genus.

Of 34 *Stenotrophomonas* species present in its GRS, we found 32 in the MetaSUB dataset (in 681,401 contigs). The subgroup simulation results in this case yielded a single number, which is the same as the number of observed subgroups (170). This result suggests that the subgroup diversity of *Stenotrophomonas* was exhausted by the sampling done.

3.2. Pangenome reconstruction for the *Xanthomonas* genus

In addition to exploring the *genomic* diversity of target genera in the MetaSUB dataset, we also wanted to explore the *genic* diversity. We did this only for *Xanthomonas*, and we chose the technique of pangenome reconstruction. We computed the GRS pangenome and the MetaSUB pangenome (using the 75,192 contigs classified as *Xanthomonas*). We then compared the two results, and obtained the Venn diagram shown in Figure 1.

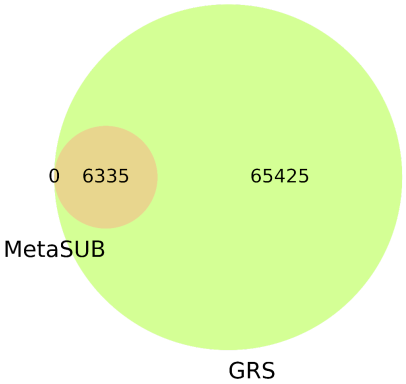

Figure 1: Venn diagram of *Xanthomonas* gene families found in the GRS and in the MetaSUB contigs.

This result shows that no MetaSUB-specific gene families were found; moreover, the the genic diversity found in MetaSUB corresponds to only 9,7% of the genic diversity found in the GRS, despite the fact that we could find in MetaSUB 42 out of 46 species and 96,7% of its subgroups. This result can be explained in part due to the fragmented nature of metagenomic data; but another reason might be related to constraints of the urban environments sampled by MetaSUB: *Xanthomonas* species found there may have a more limited gene repertoire compared to *Xanthomonas* in general.

4. Conclusions

The main results of this work are: 1) The MetaSUB dataset has a high number of species and subgroups for all the target genera explored; 2) Urban environments may contain novel *Xanthomonas* subgroups; 3) The genic diversity of *Xanthomonas* found in urban environments is much lower than for environments in general, in contrast to conclusion (1).

Acknowledgments

This work was made possible in part by a grant from CNPq (award #2024-2807) and by FAPESP fellowships (awards #2024/21751-9 and #2024/01729-9).

References

An, S.-Q., Potnis, N., Dow, M., Vorhölter, F.-J., He, Y.-Q., Becker, A., Teper, D., Li, Y., Wang, N., Bleris, L., & Tang, J.-L. (2019). Mechanistic insights into host adaptation, virulence and epidemiology of the phytopathogen Xanthomonas. *FEMS Microbiology Reviews*, *44*(1), 1–32.

Danko, D., Bezdan, D., Afshin, E. E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., Donnellan, D., Hecht, J., Jackson, K., Kuchin, K., Karasikov, M., Lyons, A., Mak, L., Meleshko, D., Mustafa, H., Mutai, B., Neches, R. Y., Ng, A., & Nikolayeva, O. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, *0*(0).

Elena Schmitz, J., & Rahmann, S. (2025). A comprehensive review and evaluation of species richness estimation. *Briefings in Bioinformatics*, *26*(2).

Gotelli, N.J. and Colwell, R.K. (2011) Chapter 4. Estimating Species Richness. In: Magurran, A.E. and McGill, B.J., Eds., Biological Diversity: Frontiers in Measurement and Assessment, Oxford University Press, New York, 39-54.

Magurran A. E. (2007). Species abundance distributions over time. *Ecology letters*, *10*(5), 347–354.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., & Rubin, E. M. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437.

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Frontiers in Genetics*, *8*.

Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, *30*(14), 2068–2069.

Solano, A., & Setubal, J. (2024). A computational pipeline for species- and strain-level classification of metagenomic sequences. In *Proceedings of the 17th Brazilian Symposium on Bioinformatics*, pp. 155-166. Porto Alegre: SBC.

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

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., & de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, *18*.

Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current opinion in microbiology*, *11*(5), 472–477.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1).