University of Puerto Rico, Mayagüez Campus
Biology Department

Data Visualization of Metagenomic Sequences from Las Salinas of Cabo Rojo, P.R.

Semester Project using RStudio

Roseliz Ríos Alemar
BIOL6694-096
Dr. Pablo E. Gutiérrez-Fonseca
December 16, 2025.

Introduction

Extreme ecosystems, such as hypersaline environments, are distributed across the world and serve as a home to a very selective group of microorganisms called extremophiles. These microorganisms typically include species of archaea and bacteria, but in halophilic environments virus-like-particles are usually present in extremely high concentrations making them the most abundant [5]. Nonetheless, very few haloviruses have been isolated and classified in comparison to halophilic species of archaea and bacteria.

Hypersaline environments, such as marine solar salterns and crystallizer ponds, are known to have high concentrations of salt (NaCl) above 3 M, which is very suitable for halophilic species that thrive at around 10-15% of NaCl [2]. *Salinibacter ruber* is a rod-shaped and red-orange pigmented bacteria, classified as part of the *Bacteroidetes*, commonly distributed worldwide across hypersaline environments [3]. It offers an excellent model for microbial diversity studies [6]. The *Salinibacter* genus is often found to be the second most abundant prokaryote species in crystallizer ponds, following haloarchaea species [1, 6], which also makes it a good model to understand virus-host relationships in these environments. *S. ruber* was not only the first bacterium confirmed to survive in higher salt concentrations, but its intraspecific diversity could be a response to intense viral predation in its own environment [4].

Viruses infecting *Salinibacter spp.* have been isolated in the past, but a study like this has never been done in Las Salinas of Cabo Rojo, Puerto Rico. Additionally, there are zero scientific reports published related to viral activity within this ecosystem on the island, even though literature shows a clear relationship between viruses and halophilic bacteria. In addition, although Las Salinas metagenome, carried out by Couto-Rodríguez and Montalvo-Rodríguez in 2019, demonstrates the presence of the genus *Salinibacter* [2], none of its species have been officially isolated in Puerto Rico. For this reason, even though the main goal of my thesis project is to isolate phages that infect *S. ruber*, obtaining clearer information on the relative abundance of *Salinibacter* is crucial before beginning.

Materials & Methods

The dataset used included the metagenome compiled by Couto-Rodríguez and Montalvo-Rodríguez (2019), whose SRA numbers were SRR8816317, SRR8816318, and SRR8816319, along with other two recent samples uploaded to the National Center for Biotechnology

Information (NCBI) database. These two samples corresponded to SRR27366713 and SRR27366714, which were published in 2024. However, for those sequences, a specific scientific article, providing further information on the process of obtaining the sequences, could not be identified. Therefore, a quality analysis was necessary before working with the selected dataset. This analysis was performed before uploading the final data files to RStudio.

To begin the visualization process in RStudio, it was required to work with the raw sequences in FASTQ format outside of the program first. For this, the Bash language was used in the command lines. Once the data was loaded into the university's Ubuntu server, assembled, and then contig binning was performed using Minimap2 and Rosella. After obtaining the bins, a CheckM2 analysis was ran on the refined bins to obtain information related to genome completeness and contamination percentage, allowing the filtering of lower-quality bins and the retention of the most reliable ones. Subsequently, a CAT/BAT operation was done to obtain the taxonomic classification of each bin. This made it possible to identify some species from their domain down to their genus or species level. Furthermore, this step allowed the selection of only those sequences corresponding to *Salinibacter ssp.* to evaluate their presence within Las Salinas. Finally, it was crucial to obtain a file, in TSV or CSV format, with this information and upload it to RStudio.

The first step in RStudio required the installation of two main R packages, "tidyverse," and "pheatmap". All analyses were conducted in R using a reproducible tidyverse workflow. Data was imported, cleaned, and transformed using functions from the dplyr and tidyr packages. All plots were produced with ggplot2, ensuring proper labeling and formatting. The analysis was organized in an R Project, and the full workflow, including code, output, and written interpretation, was executed and documented in a R Script file. This ensures transparency, reproducibility, and a clear record of each processing step. In this case, "tidyverse" served to read, filter, and organize data using functions such as read_csv(), filter(), mutate(), and ggplot(), among others. The "pheatmap" package allowed the creation of a heatmap to evaluate the quality features of the *Salinibacter* related bins found in the dataset. This was done using mainly the pheatmap() function. Below, **Table 1** provides a summary of results for each variable, as well as notes and frequency if it applied. It is important to note that the font size for this table was reduced to allow a better view of all the variables included.

Table 1. Description of the dataset variables, including type, description, example values, and summary information.

Variable Name	Type	Description/Units	Example Values	Notes/Frequency/Summary
Bin ID	Factor	Unique identifier for each metagenome assembled genome (MAG) bin	rosella_refined_0_440, rosella_refined_0_279	One per recovered genome
Completeness	Numeric	Estimated genome completeness (%)	34.78, 5.38, 7.66	To assess MAG quality and classify *Salinibacter spp*. bins
Contamination	Numeric	Estimated genome contamination (%)	2.3, 0.11, 0.04	To assess MAG quality and understand contamination levels
Coding Density	Numeric	Proportion of coding sequence per genome	0.867, 0.824, 0.866	To compare genome compactness in *Salinibacter* bins heatmap
Contig N50	Numeric	N50 contig length (bp)	10293, 12523, 5240	Contiguity metric in *Salinibacter* bins heatmap
Lineage	Character	BAT Taxonomic Classification	root;d_Bacteria;p_Bacteroidota;g_Salinibacter	Information about Domain, Phylum, and Genus
Domain	Factor	Highest-level taxonomic classification	Archaea, Bacteria	Only two Domains present
Phylum	Factor	Phylum-level taxonomic classification	Bacteroidota, Nanohaloarchaeota, Bacillota	Phylum-level organized plot
Genus	Factor	Phylum-level taxonomic classification	*Salinibacter, Escherichia, Halovenus*	Genus-level organized plot
Group	Factor	MAG bin category for data visualization	Other bins, *Salinibacter* (<30%), *Salinibacter* (>=30%)	Used for MAGs quality plot

Results & Discussion

The original pre-proposal for this project included the use of "vegan" for statistical measures of diversity and relative abundance of specific microorganisms, like those associated with the *Salinibacter* genus. However, this was impossible to incorporate due to the MAGs lacking crucial pieces of information like taxa abundance, presence or absence per sample, relative abundance per sample, etc. Additionally, since the data was produced a few weeks

advance, it was not possible to go back into the server and fix the issue, because access had been restricted to run bigger datasets. In order to account for that issue, only data visualization of MAGs was done in RStudio in the form of plots.

A crucial first step was to visually represent the results from CheckM2 quality assessment. In **Fig. 1**, the plot shows information about all MAG bins in terms of Completeness (%) vs Contamination (%), but only *Salinibacter* related bins were given emphasis by being classified by color based on their completeness percentage. In this case, *Salinibacter* genus bins with low completeness were represented in blue, while those with higher completeness were represented in blue. The rest of the pink-transparent dots on the plot represent the rest of the MAGs and were chosen to be represented this way for a better view of *Salinibacter* bins positioning vs the rest of the MAG sequences.

For the next part of the analysis, understanding Taxonomic Classification seemed important, especially after conducting a BAT analysis as part of the CAT/BAT pack program ran. For this reason, two plots were generated with the purpose of having a visual idea of Phylum and Genus present in the refined bins, while also dividing them based on their respective Domain. In both **Fig. 2** and **Fig. 3** is clear that only two Domains were found, Bacteria and Archaea, for which both are prokaryotic Domains. For both plots, the red represented Archaea, while the light pink represented Bacteria related genomes. In **Fig. 2**, the total amount of Phylum classifications present in all MAG bins is 8, from which only two were associated with Archaea. Similarly, in **Fig. 3** the total amount of Genus classifications present was 25, from which only 8 of those were Archaea. These findings were specifically accurate with literature as it is well known that halophilic environments are home primarily to species of haloarchaea and halobacteria [6]. Additionally, *Salinibacter* genus had at least 10 bins in the overall dataset, following *Halomonas* as the second most abundant bacteria. This also correlates to previous literature findings placing *Salinibacter* as one of the most abundant bacteria across hypersaline environments [1, 2].

In **Fig. 4,** a heatmap is shown to understand the *Salinibacter* related bacteria present. This was created using RStudio by filtering only 10 bins associated with the Genus, allowing a better understanding of each bin in terms of Completeness, Contamination, Contig_Density, and Contig_N50. The heatmap ranges in colors to represent higher (warm-tones) and lower(cool-tones) amounts for each category. In this case, only one bin exhibited higher completeness and

contamination, while the rest of the bins are mostly incomplete and have lower (almost zero) percentage of contamination. Additionally, values like Contig_N50 (assembly contiguity) provided valuable information about the quality of assembly. On the other hand, Coding_Density showed the fraction of the MAG bin that was coding for genes. Overall, *Salinibacter* genus seemed to have mostly higher quality bins, although their completeness was very little. This hints the possibility of these MAGs being useful for future bioinformatic analyses and approaches.

For future studies, more detailed data could be obtained, since it would improve the information already found through this visualization process. In addition, doing previous Anti-SMASH or Bakta analyses could produce better plots to understand metabolic pathways, proteins present, G-C% content, among other values. Moreover, to address the project's current limitation with the diversity analysis, data containing information related to taxa abundance could be useful to run "vegan" on RStudio and commence statistical analyses. Although, all the main goals of this semester project were met successfully.

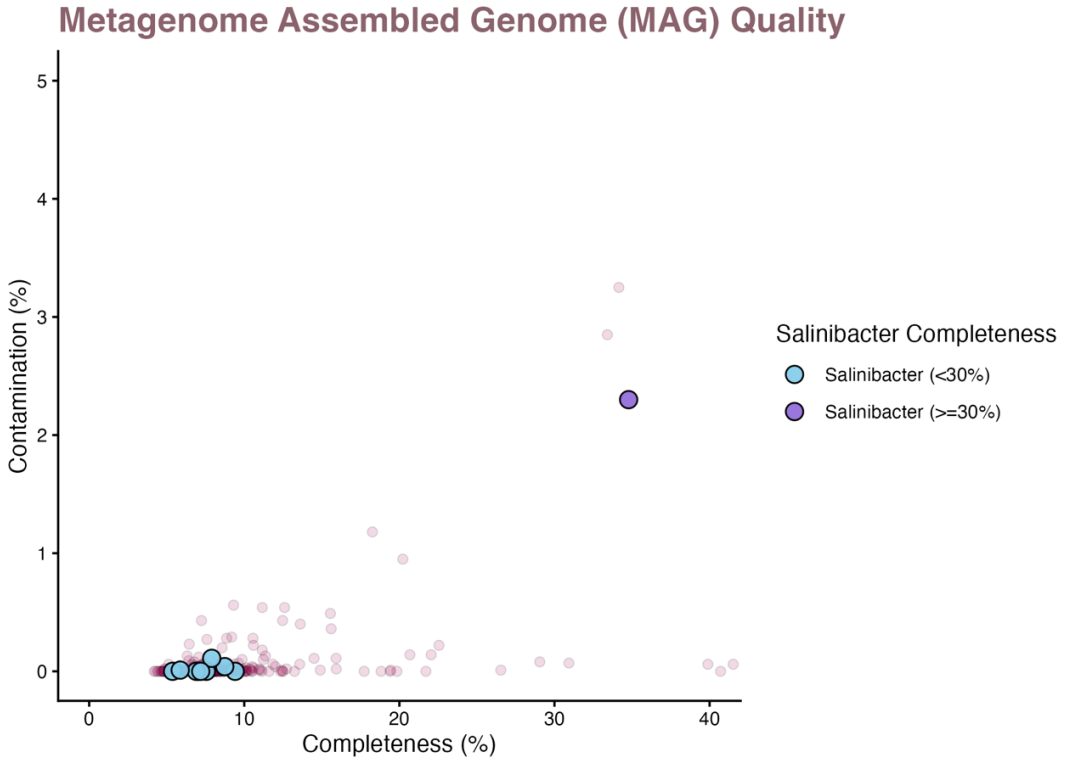

Figure 1. Assessment of MAG bins quality in terms of Completeness (%) vs Contamination (%). The plot shows specific information for *Salinibacter* related genomes based on completeness, as well as data distribution for all other refined bins.

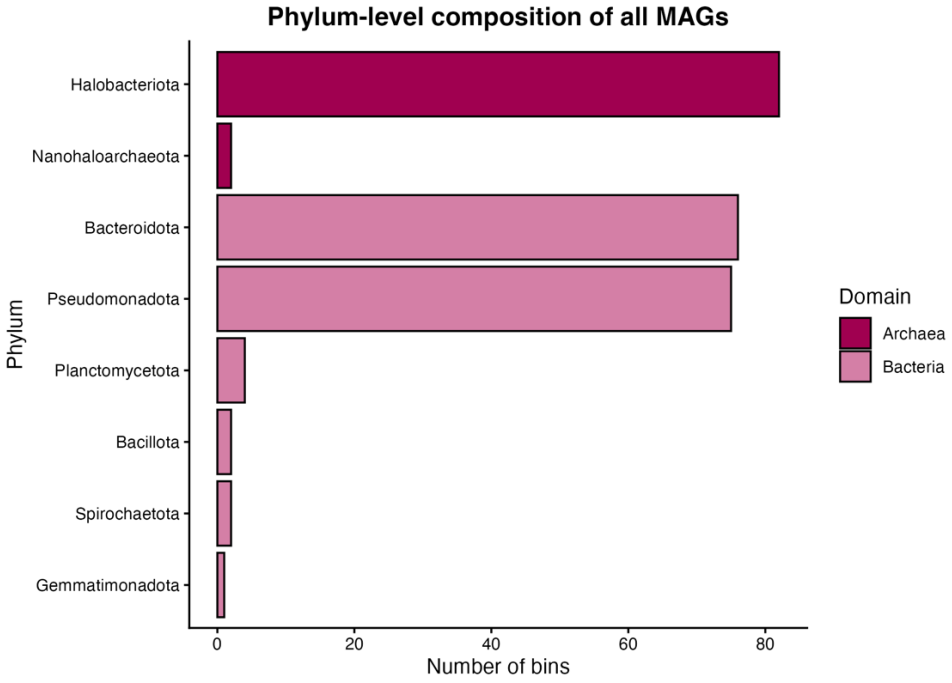

Figure 2. Plot illustrates 8 Phylum-level taxonomic classifications inside the dataset and classifies them based on the Domain Bacteria or Archaea.

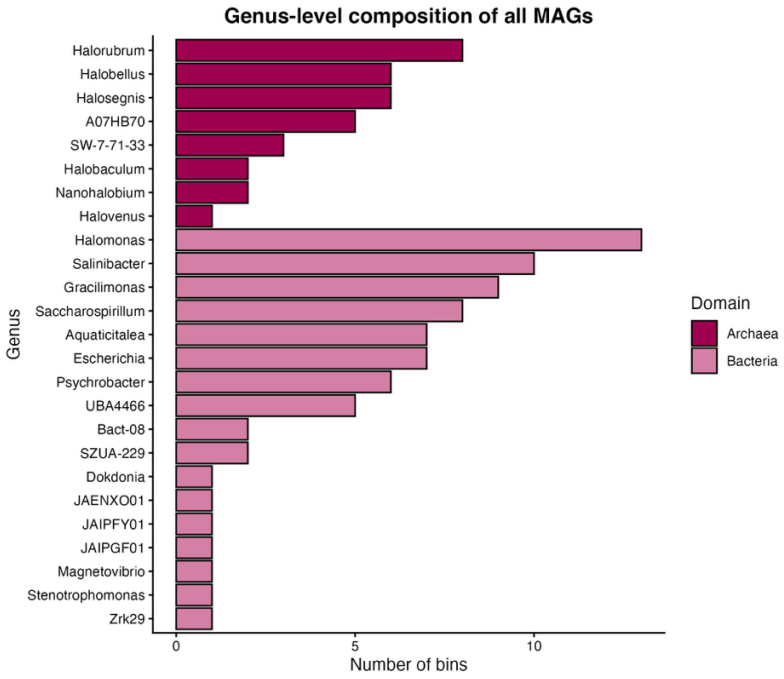

Figure 3. Plot illustrates 25 Genus-level taxonomic classifications inside the dataset and classifies them based on the Domain Bacteria or Archaea.

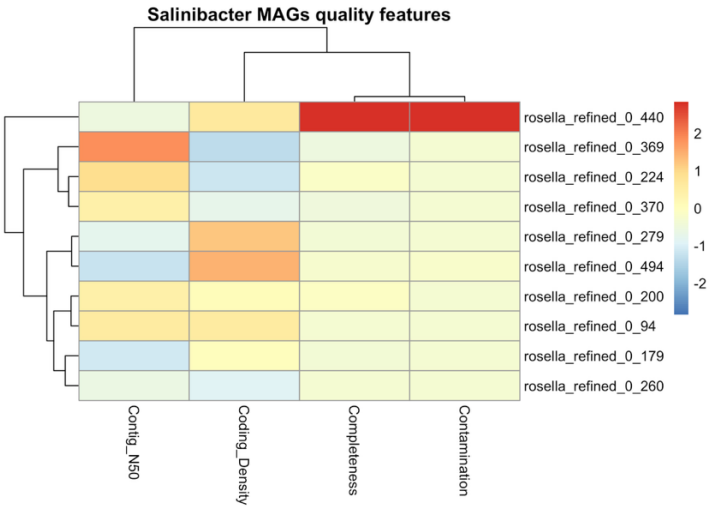

Figure 4. Heatmap for 10 specific *Salinibacter* bins

References

[1] Atanasova, N. S., Oren, A., Bamford, D. H., & Roine, E. (2013). Diverse antimicrobial interactions of halophilic archaea and bacteria extend over geographical distances and cross the domain barrier. *MicrobiologyOpen, 2*(6), 809–817. https://doi.org/10.1002/mbo3.115

[2] Couto-Rodríguez, R. L., & Montalvo-Rodríguez, R. (2019). Temporal Analysis of the Microbial Community from the Crystallizer Ponds in Cabo Rojo, Puerto Rico, Using Metagenomics. *Genes, 10*(6), 422. https://doi.org/10.3390/genes10060422

[3] Oren A. (2013). *Salinibacter:* an extremely halophilic bacterium with archaeal properties. *FEMS microbiology letters*, *342*(1), 1–9. https://doi.org/10.1111/1574-6968.12094

[4] Sanchez-Martinez, R., Arani, A., Krupovic, M., Weitz, J. S., Santos, F., & Antón, J. (2025). Episomal virus maintenance enables bacterial population recovery from infection and promotes virus-bacterial coexistence. *The ISME journal*, *19*(1), wraf066. https://doi.org/10.1093/ismejo/wraf066

[5] Santos, F., Yarza, P., Parro, V., Meseguer, I., Rosselló-Móra, R., & Antón, J. (2012). Culture-independent approaches for studying viruses from hypersaline environments. *Applied and environmental microbiology*, *78*(6), 1635–1643. https://doi.org/10.1128/AEM.07175-11

[6] Villamor, J., Ramos-Barbero, M. D., González-Torres, P., Gabaldón, T., Rosselló-Móra, R., Meseguer, I., Martínez-García, M., Santos, F., & Antón, J. (2018). Characterization of ecologically diverse viruses infecting co-occurring strains of cosmopolitan hyperhalophilic *Bacteroidetes*. *The ISME journal*, *12*(2), 424–437. https://doi.org/10.1038/ismej.2017.175