bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

From the Andes to the desert: First overview of the bacterial community in the Rimac river, the main source of water for Lima, Peru

Pedro E. Romero[a], Erika Calla-Quispe[b], Camila Castillo-Vilcahuaman[a], Mateo Yokoo[c] Hammerly Lino Fuentes-Rivera[b], Jorge L. Ramirez[d], Alfredo J. Ibáñez[b], Paolo Wong[c]

a. Facultad de Ciencias y Filosofía. Universidad Peruana Cayetano Heredia. Lima, Peru.
b. Instituto de Ciencias Ómicas y Biotecnología Aplicada (ICOBA). Pontificia Universidad Católica del Peru. Lima, Peru.
c. Facultad de Medicina Humana. Universidad de Piura. Lima, Peru.
d. Facultad de Ciencias Biológicas. Universidad Nacional Mayor de San Marcos. Lima, Peru.

Corresponding authors
Pedro E. Romero, pedro.romero@upch.pe
Paolo Wong, paolo.wong@udep.pe

Abstract
Background
The Rimac river is the main source of water for Lima, Peru's capital megacity. The river is constantly affected by different types of contamination including mine tailings in the Andes and urban sewage in the metropolitan area. We aim to produce the first characterization of bacterial communities in the Rimac river using a 16S rRNA amplicon sequencing approach which would be useful to identify bacterial diversity and potential understudied pathogens.
Results
We report a higher diversity in bacterial communities from the Upper and, especially, Middle Rimac compared to the Lower Rimac (Metropolitan zone). Samples were generally grouped according to their geographical location. Bacterial classes Alphaproteobacteria, Bacteroidia, Campylobacteria, Fusobacteriia, and Gammaproteobacteria were the most frequent along the river in different proportions. *Arcobacter cryaerophilus* (Campylobacteria) was the most frequent species in the Lower Rimac while an unidentified Leptotrichiaceae (Fusobacteriia) and *Flavobacterium succinicans* (Bacteroidia) were the most predominant in the Upper zone. Predicted metabolic functions in the microbiota include bacterial motility, quorum sensing and xenobiotics metabolism. Additional metabolomic analyses showed the presence natural flavonoids and antibiotics in the Upper Rimac and herbicides in the Lower Rimac.
Conclusions
The dominance in the Metropolitan area of *Arcobacter cryaerophilus*, an emergent pathogen associated with fecal contamination and antibiotic multiresistance, but that is not usually reported in traditional microbiological quality assessments, highlights the necessity to apply next-generation sequencing tools to improve pathogen surveillance. We believe that our study will the integration of omics sciences in Peru and its application on current environmental and public health issues.

Keywords
Bacterial communities, Rimac, Peru, river pollution, 16S rRNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ethical approval

The study did not involve experiments in humans nor animal subjects.

1. Introduction

Worldwide, water quality problems are associated with poverty conditions and lack of efficient sanitation, especially in developing countries [1]. These problems are recurrent in highly populated cities where most of the waste is directly washed to nearby rivers [2]. Lima, the capital city of Peru, is the second largest desert city in the world [3]. Its Metropolitan area is inhabited by more than 10 million people creating enormous challenges for environmental and public health [4]. For instance, it has been shown that water quality in Lima is a significant risk factor for pathogenic infections in children [5].

The Rimac river is the main source of drinking water for the Lima Metropolitan area. Recently, more than 700 pollution sources were identified by the National Authority of Water [6]. The river is constantly polluted by mine tailings in its Upper zone close to the central Peruvian Andes, by agricultural wastewater in its Middle zone, and by industrial wastewater and urban sewage in its Lower zone within the Metropolitan area nearby the Pacific Ocean [7]. The lack of an efficient wastewater treatment in the Metropolitan region promotes the presence of potentially pathogenic bacteria such as *Escherichia coli*, *Salmonella typhi*, or *Vibrio cholerae*, associated with fever and diarrhea symptoms [2,8].

Traditionally, assessments of water quality and bacterial contamination in Peru are focused on evaluating the presence of common coliforms (e.g. *Citrobacter*, *Escherichia*, *Enterobacter*, *Klebsiella*) [9]. However, there are a plethora of likely pathogens that are not currently studied using classic methods because of taxonomic assignation problems or the existence of non-culturable phenotypes [10]. In recent years, advances in the study of bacterial communities and diversity have occurred because of the improvement of next-generation sequencing (NGS) methodologies. One of these techniques, 16S rRNA amplicon sequencing, uses a fragment of the 16S ribosomal gene to obtain a diversity profile of the bacterial community in a specific environment. Therefore, it has been used to study different types of samples from the internal microbiota of several species, including humans to environmental community surveys [11].

16S rRNA amplicon sequencing has been also used to study bacterial communities from several rivers. For instance, a study in the Danube river, which crosses many countries from central Europe, found a higher microbial community richness in the Upper basin, while in the Lower basin, there was a predominance of only a few free-living and particle associated bacteria [12]. Studies in rivers have also looked for the occurrence of potential pathogens. A report from the river Tama (Tokyo, Japan) showed that the predominant bacteria genus was *Flavobacterium* (Bacteroidia), a freshwater fish pathogen [13]. Recently, a study in the Pinheiros river (Sao Paulo, Brazil), one of the most polluted Brazilian rivers, found the predominance of *Arcobacter cryaerophylus* (Campylobacteria). This species is considered an emerging pathogen and an indicator of fecal contamination [14,15].

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In addition, information from metabolomics provides additional insight into chemical contaminants in water that may influence bacterial diversity and would be useful to indicate the health status of an aquatic ecosystem and understand interactions between microbial communities and their environment [16,17]. For instance, a study in the Brisbane river (Australia) found that human interference was associated with a population size increment of Actinomycetes and Pseudomonadales, This increment was linked to higher levels of sugar alcohols, short-chain fatty acids and aromatic amino acids which contribute towards biofilm production [16].

In our work, we attempt to provide for the first time an overview of the microbiota from the Rimac river, comparing diversity patterns from Andean to Metropolitan areas and focusing on potential pathogenic bacteria.

2. Material and Methods

2.1 Area of study and collection

In January 2020, we sampled river water in 13 points along the extension of the Rimac river, within provinces of Huarochiri, Lima and Callao (Figure 1). Five sampling sites were from the Upper Rimac (1-5: Chicla, San Mateo City, Tamboraque, Huanchor and Chacahuaro); three, from the Middle Rimac (6-8: Santa Eulalia, Chaclacayo and Huachipa Bridge; and five, from the Lower Rimac near main bridges in the Metropolitan area (9-13: Libertadores, Nuevo, Universitaria, Faucett and Gambetta). Sterile plastic bottles were used to sample approximately 1.5 L of superficial water. When we sampled from a bridge, we attached the plastic bottle to another flask linked to a 30 m rope. In this case, we sampled superficial water (ca. 50 cm below the surface) from the middle of the bridge. Water samples were transported to the lab in coolers and were processed in the same day. We filtered river water using 0.2 µM Sterivex filters.

2.2 DNA isolation, amplification, and preparation of genomic libraries

Total DNA was isolated from the Sterivex filter using the PowerWater DNA kit (Qiagen) following manufacturer's instructions. Then, we quantified DNA quality using a Qubit 3.0 (Thermo Fisher Scientific, USA) fluorimeter. For the preparation of genomic libraries, we followed the Illumina protocol [18] which consists in the amplification of the 16S rRNA V3-V4 region (ca. 460 nt). A second amplification is performed to attach oligo adapters (indexes) to each amplicon sample. Indexes will be informative to differentiate samples after the sequencing step. Each step was followed by amplicon cleaning using AMPure XP beads (Beckman Coulter, USA). In the final step, all samples were pooled and sequenced in an MiSeq (Illumina, USA) instrument. Raw sequences from the genomic libraries are deposited in the NCBI Bioproject database (accession number: PRJNA646070).

2.3 Bioinformatics and data analyses

Sequencing reads per each sample were analyzed using DADA2 [19] in R v3.6.3 [20]. We filtered reads using the parameter maxEE=2 for quality threshold as suggested before [21,22] . DADA2 was used to infer Amplicon Sequence Variants (ASV) which are groups of identical sequences. We did the taxonomic assignation, comparing the sequences against the ribosomal database SILVA v.138 [23] using the DECIPHER package [24]. For the identification to the species (or multispecies) level, we used DADA2 function "assignTaxonomy". Then, we estimated Chao1, Shannon and Simpson alpha diversity indexes using the package phyloseq [25]. Significant

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

differences among regions were evaluated using a paired Wilcoxon test. To estimate beta diversity, we normalized the ASV matrix using the variance stabilizing transformation (VST) method as suggested before for microbiome analyses [26]. This method is available in the package DESeq2 [27]. Then, we used a multidimensional scaling (MDS) approach to compare beta diversity among localities.

In addition, we used the taxonomic and occurrence information to produce a stack barplot of the most predominant classes. Then, we explored the most predominant genera using the package ampvis2 [28]. Next, we selected ASV with more than 1 000 counts and a taxonomic assignation until species level. We used this information to look for potential human and other animal pathogens in the Risk Group database of the American Biological Safety Association, specifically if species are pathogens for humans or other animals [29]. Finally, we used Piphillin [30,31] to predict functional content based on the frequency of the 16S rRNA sequences comparing them to annotated genomes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Predicted KEGG orthologues (KO) occurrence was retrieved from KEGG (October 2018) using a 97% cutoff threshold to create a gene feature table. We used the 20 most frequent unique KO per locality, then we grouped localities in three regions (Upper, Middle and Lower Rimac) and created a Venn diagram to look for shared KOs for each region using Venny [32].

2.4 Metabolomic analyses

Ten milliliters of each water sample and Milli Q water (Blank) were filtered through a Nalgene™ 0.22 µm syringe filter (Thermo Scientific) to remove suspended particulates. Each water sample was immediately extracted with dichloromethane (ACS grade, J.T. Baker) (3 times x 10 mL). After that, dichloromethane extracts were concentrated *in vacuo* and placed in a vial with 150 µL of dichloromethane and 2 µL of 26.7 ppb of toluene (GC grade, Sigma-Aldrich) (internal standard). The standard solution consisted in 2 ppm of toluene in dichloromethane, 2 µL of this solution was diluted with 150 µL dichloromethane. All water samples were analyzed except Santa Eulalia (6), Chaclacayo (7), Huachipa (8) and Gambetta (13) because these water samples were used completely in the step 2.1. Water samples, blanks and internal standards were analyzed by gas chromatography (GC) coupled to an APPI-Q-Exactive HF mass spectrometer (Thermo Fisher Scientific, USA). GC was equipped with a DB-5 column (30 m x 0.25 mm i.d., 0.25 µm thickness film). The oven temperature was programmed as follows: aliquots of 2 µL sample were injected at 45 °C for 2 min, increase at 10 °C/min until 270 °C, and hold for 7.5 min. Injections were made in splitless-mode with helium as the carrier gas (1.5 mL/min), injector temperature at 270 °C, and detector temperature at 270 °C. High-accuracy MS data were acquired in positive data-dependent acquisition (DDA) with scan range m/z 50–750. Raw data from this GC-MS experiment was first converted into ABF format. Then, peak spotting was performed by exploring retention time and accurate mass. MS-DIAL v4.16 provided peak alignments of all samples and normalizes data based on TIC (total ion current). Final filtering was performed with a principal component analysis (PCA, p-anova < 0.05) using MATLAB vR2019b. Additionally, the software Compound Discovery was used for tentative identification using MS/MS data and comparing our results against several databases such as ChemBioFinder, Chemspider, Kegg, LipidBank, LipidMaps, Metlin and NIST. Besides, a PCA was performed to observe differences in mass profiles among samples group.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3. Results

3.1 Bacterial community characterization

16S rRNA amplicon sequencing produced an average of 265 625 [235 989-287 437] reads per sample. After discarding low quality sequences and merging forward and reverse reads, we obtained an average of 99 770 [76 129-122 999] final reads per sample (Supplementary Table 1). From this subset, we identified a total of 15 059 ASV. The first 151 ASV were frequent in more than 1 000 reads and the first eight, in more than 10 000 reads (Supplementary Table 2). Chao1, Shannon and Simpson diversity indexes were consistently higher in the Upper and Middle Rimac compared to the Lower Rimac (Figure 2). Alpha diversity values for each sample can be found in Supplementary Table 3, The information from the significance of pairwise comparisons can be found in Supplementary Table 4. In the MDS analysis, samples were grouped by their geographical location (Upper, Middle and Lower Rimac) except for Chacahuaro, a locality from the Upper Rimac that appeared closer to the Middle Rimac samples (Figure 3).

Phyla Bacteroidota, Campilobacterota, Fusobacteriota, Proteobacteria represented 49.34% (7 443 from 15 059) of the bacterial diversity in the ASV (Supplementary Table 5). Taxonomic identification of ASV was similar for phylum, class, and order ranks (approx. 60%), and a bit lower for family rank (48.93) (Supplementary Table 5). Classes Alphaproteobacteria, Bacteroidia, Campylobacteria, Fusobacteriia, and Gammaproteobacteria were the most common ones along all zones (Figure 4). Fusobacteriia declines in sampling sites 5 to 8 (Chacahuaro in the Upper Rimac zone and Middle Rimac samples). Campylobacteria especially rises in the Lower Rimac zone (Metropolitan area: sampling sites 9-13) but is also common in the Upper and Middle Rimac zones. The most frequent genera (Figure 5) in the Upper Rimac (sampling sites 1-4) was an unidentified Leptotrichiaceae (ASV2) while *Flavobacterium* (Bacteroidia) was predominant in Chacahuaro and in the Middle Rimac. In the Lower Rimac, we found a dominance of *Arcobacter* (Campylobacteria). Human pathogenic bacteria identified until species level rank were *Arcobacter cryaerophylus* and *Prevotella copri* (Supplementary Table 6), and until the genus level rank were *Aeromonas*, *Escherichia*, *Pseudomonas* and *Shigella*. Other animal pathogens identified were *Flavobacterium succinicans* y *Faecalibacterium prausnitzii*.

Piphillin functional prediction of the top KEGG orthologies (KO) for each locality revealed 32 unique KO in the Upper Rimac, 26 in the Middle Rimac, and 27 in the Lower Rimac (Supplementary Table 7). The Venn diagram shows nine KO shared by the three zones, 15 KO were only shared between the Upper and Middle Rimac and four, only between the Upper and Lower Rimac (Supplementary Figure 1). The most abundant predicted pathways shared by the three zones (Supplementary Table 8) were related to bacterial chemotaxis (methyl-accepting chemotaxis protein/MCP, K03406), metabolism of xenobiotics (glutathione S-transferase, K00799), and aminoacyl-tRNA biosynthesis (K14218, K14228). Between the Upper and Middle Rimac the most frequent pathways predicted were ATP-binding cassette transporters and quorum sensing (Leucine-Isoleucine-Valine *liv* gene cluster, K01996 - K01999).

3.2 Metabolomic analyses

Results from the GC-MS experiment can be found in Supplementary Table 9. PCA performed from these results (Supplementary Figure 2), first showed that the Upper and Lower Rimac were different from blank and internal standard samples. Some

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

localities from the Upper zone (Tamboraque, Huanchor and Chacahuaro) were well differentiated from the other areas. However, other Upper Rimac samples (Chicla and San Mateo) were clustered within the Lower Rimac because of similar chemical composition. Additionally, GC-MS (APPI) in positive mode allowed to detect eleven compounds (p-anova < 0.05, total score > 95%) using Compound Discovery and MS-DIAL softwares (Supplementary Table 10). The loading plot of PCA coefficients from metabolomics analysis (Supplementary Figure 3) showed that the Upper Rimac localities, Tamboraque, Huanchor and Chacahuaro are influenced by the following compounds: nanaomycin derivates, natural bacterial antibiotics, and hispidulin, a natural plant flavonoid (Supplementary Table 10). Whereas the other Lower Rimac samples were influenced by propazine-2-hydroxy, an herbicide. Additionally, a natural flavonoid, isopeonol, and other pharmaceutical compounds such as desloratadine, meprylcaine and N-(4-ethoxyphenyl)acetamide were identified in both zones (Supplementary Table 10).

4. Discussion

Our study provides the first overview of the bacterial community diversity in the Rimac river (Lima, Peru) using a 16S rRNA amplicon sequencing approach. We described bacterial community shifts along an altitude gradient from the Upper Rimac zone (1 500 – 4 500 msl) to the Middle (300 – 1 500 msl) and Lower Rimac zones (0 – 300 msl). The Lower Rimac crosses the Lima Metropolitan area, a mega-city situated on the hyper-arid Peruvian desert [4,33], being an essential water source for this city.

Previous studies in rivers have shown different patterns of diversity along altitudinal gradients. For instance, a diversity decline downriver was found in the Danube in free-living and particle-associated communities [12], while an increase was found in the Mississippi especially in organisms related to suspended particles [34]. In our case, different estimators of alpha diversity (Chao1, Shannon and Simpson) provided evidence of a decline of diversity in the Lower Rimac (Metropolitan area). The greatest mean diversity values were found in the Middle Rimac and may be influenced by confluences with minor rivers such as the Santa Eulalia river (sampling site 6). It appears necessary to sample more localities between sites 5 to 8 to provide a better assessment of local diversity in the Middle Rimac. Besides, we sampled in January 2020, rainy season in the Andes, so it would be desirable also collecting during the dry season to compare temporal diversity patterns.

Bacterial communities in high-altitude aquatic environments are still not thoroughly studied. We found only few characterizations using 16S rRNA sequencing, especially in lakes. For instance, community profiling of Tibetan and Pyrenean lakes showed dominance of classes Actinobacteria, Alphaproteobacteria, Betaproteobacteria, and phyla Bacteroidota [35–37]. Our samples did have a major presence of the Bacteroidia class, part of the latter phylum, and class Alphaproteobacteria. However, we found no Betaproteobacteria. Additional surveys in high-altitude rivers are necessary to have a broader view of bacterial communities in these kinds of habitats. On the other hand, samples from the Lower Rimac looked more similar among each other than samples in the Upper or Middle Rimac (MDS, Figure 3). Community composition in the Lower Rimac is probably influenced by similar pollution constrains such as human feces and domestic sewage that are produced in the Metropolitan area [38].

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Furthermore, we found a constant occurrence of the phyla Bacteroidota, Campilobacterota, Fusobacteriota and Proteobacteria along the Rimac river basin. This result is similar to a previous report in the polluted Pinheiros river in Sao Paulo, Brazil [14]. The authors associated the high presence of these phyla with freshwater environments and domestic sewage sludges. In addition, they also found *Arcobacter cryaerophilus*, as the predominant species in the river. This species has been reported as very frequent in other rivers such as the Yangtze (China) [38], many rivers in Nepal [39], and the Llobregat (Spain) [40]. As mentioned before, *Arcobacter* is an emergent enteropathogen and indicator of fecal contamination. It can also occur in water treatment and sewage systems [41] and survive adverse conditions imposed by food processing and storage [42]. *Arcobacter* is also abundant in effluent from wastewater treatment plants [43]. In Lima, the main water treatment plant is situated in the Lower Rimac nearby the sampling site 9 (Figure 1). Thus, our results corroborate previous reports of the endurance of this genus and the necessity of a better water quality assessment based in amplicon sequencing [43] not only for environmental samples but also for water samples taken after treatment.

Moreover, *Arcobacter* has shown resistance to several antibiotics [42] and it has been proposed to be involved in the exchange of resistance genes between gram-negative and gram-positive phyla [44]. Notwithstanding, reports in Peru on the pathogenicity of this genus are scarce, e.g. we found only one published study that mentioned *Arcobacter* in stool from children with diarrhea [45]. In particular, *Arcobacter cryaerophilus*, the most frequent ASV in the Metropolitan area, has been associated with severe diarrhea and found to carry several virulence genes [46].

In the Upper Rimac, we found a predominance of an unidentified Leptotrichiaceae, a rarely isolated bacterial family associated to internal tracts of human and animals [47]. This result may show that even in a region far from the Metropolitan area, human pollution is constantly affecting the composition of bacterial river communities. The genus *Flavobacterium* was common in the sampling site 5 (Chacahuaro), and it has been related to several issues in fish such as “cold water” and *columnaris* disease [48]. In particular, *Flavobacterium succinicans*, the most frequent *Flavobacterium*, has been associated to bacterial gill disease in trout [49].

The high occurrence of bacterial chemotaxis pathway, a necessary function to move towards nutrients or away toxins, would be connected to bacterial growth and survival aquatic environments [50]. A study in the Pinheiros river found similar high occurrence of bacterial chemotaxis and flagellar assembly functions and hypothesized that this could be influenced by the elevated concentration of nutrients, e.g. ammonia and phosphates [14]. In our study, the most predicted KO in our samples was the MCP protein (Supplementary Table 8). This protein is important for bacterial motility because it triggers the activation of flagella, and has been also encountered in other highly polluted rivers such as the Yamuna, a major tributary of the Ganges river in India [51]. In addition, we also found a high occurrence of predicted ABC transporters which are ubiquitous proteins involved in several processes including nutrient uptake and chemotaxis [52]. According to the KEGG annotation these proteins also take part in quorum sensing functions. Quorum sensing has been profusely studied in free-living bacteria in laboratory conditions. However, it is now known that, in nature, bacteria can use quorum sensing

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mechanisms in fluid environments such as rivers, streams, intertidal and marine areas by forming biofilms [53]. Another abundant function was related to the Glutathione metabolism. Glutathione S-transferase is involved in biodegradation of xenobiotics, defense against chemical and oxidative stress, and antibiotic resistance [54]. Rivers have been shown as reservoirs of genes related to antibiotic resistance influenced by anthropogenic causes [51,55]. Further research is necessary to obtain environmental metagenomes from the Rimac and to look for possible genes linked to resistance to antibiotics in the Lower and Middle Rimac or resistance to heavy metals which may be frequent in the Upper Rimac.

Metabolomics analysis by GC-MS showed differences in chemical composition between some localities from the Upper Rimac with respect to the Lower Rimac. This variance may be due to less population density and distance from the Metropolitan area. However, we also found that some localities from the Upper Rimac are as contaminated as the Lower Rimac. This apparently contradictory results highlight that more work should be done to understand the pollution patterns caused by human influence. For instance, nanaomycin derivates are actinomycete metabolites, biosynthesized by *Streptomyces* [56]. These compounds had a higher frequency in the Upper Rimac. However, we did not identify any *Streptomyces* in our data (Supplementary Table 5). This could be because *Streptomyces* are generally found in the river sediment [57,58]. Further sampling efforts should also consider different portions of the water column and sediment.

Finally, our study could be complemented with metabarcoding analyses of other pathogenic protozoa [7] or invertebrates which are also used as indicators of water quality [59]. We believe that this initial work will promote more similar studies in our country. Further work that integrates -omics and environmental sciences with public health studies will be beneficial for future public policies especially focused on rural and sub-urban areas in Lima which depend on the Rimac river water.

5. Acknowledgments

PER thanks to Yulissa Estrada and Grecia Valdivia (Universidad de Ingeniería y Tecnología. Lima, Peru) for their support during sampling in the Upper Rimac. In addition, PER thanks to Raúl Condori and Enrique Santisteban for their kind support during sampling in the Metropolitan area. AJI thanks Thermo Fisher Scientific for its support in the metabolomic analyses. PER and PW thank to Arturo Gonzáles and Stefany Infante (Universidad de Piura) for their help in the lab work. All authors thank to Guillermo Trujillo (GenLab) for his support in the preparation of genomic libraries.

6. Financial support

This work was funded by the Facultad de Medicina Humana (Universidad de Piura) [grant number PI2002, “Metabarcoding del Río Rímac, el principal afluente de la ciudad de Lima”]; the company GenLab del Perú SAC [Programa de Incentivos 2019, “Aplicaciones de la secuenciación masiva (NGS) en metagenómica y secuenciación de genes”]; and the Max Planck Society [Max Planck Partner Groups: Chemical-Ecology + Pontificia Universidad Católica del Perú]. Additionally, PER was funded by the Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (Fondecyt - Perú) [grant number 34-2019, “Proyecto de Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica”]. JLR received a grant from Concytec - Banco Mundial,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

through the Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (Fondecyt) [grant number 022-2019].

References

[1] Chalchisa D, Megersa M, Beyene A. Assessment of the quality of drinking water in storage tanks and its implication on the safety of urban water supply in developing countries. Environ Syst Res 2018;6:12. https://doi.org/10.1186/s40068-017-0089-2.

[2] Abraham WR. Megacities as sources for pathogenic bacteria in rivers and their fate downstream. Int J Microbiol 2011;2011:1–13. https://doi.org/10.1155/2011/798292.

[3] Edelman DJ. Managing the Urban Environment of Lima, Peru. Adv Appl Sociol 2018;08:233–84. https://doi.org/10.4236/aasoci.2018.83014.

[4] Silva J, Rojas J, Norabuena M, Molina C, Toro RA, Leiva-Guzmán MA. Particulate matter levels in a South American megacity: the metropolitan area of Lima-Callao, Peru. Environ Monit Assess 2017;189:635. https://doi.org/10.1007/s10661-017-6327-2.

[5] Wesley I V. Helicobacter and Arcobacter: Potential human foodborne pathogens? Trends Food Sci Technol 1997;8:293–9. https://doi.org/10.1016/S0924-2244(97)01050-9.

[6] Observatorio del Agua Chillón Rímac Lurín (OA CHIRILU). Diagnóstico inicial para el Plan de gestión de recursos hídricos en el ámbito de las cuencas Chillón, Rímac, Lurín y Chilca. Lima, Peru: 2019.

[7] Bautista M, Bonatti TR, Fiuza V, Terashima A, Canales-Ramos M, Jose J, et al. Occurrence and molecular characterization of Giardia duodenalis cysts and Cryptosporidium oocysts in raw water samples from the Rimac River, Peru. Env Sci Pollut Res Int 2018;25:11454–67. https://doi.org/10.1007/s11356-018-1423-6.

[8] Grothen DC, Zach SJ, Davis PH. Detection of Intestinal Pathogens in River, Shore, and Drinking Water in Lima, Peru. J Genomics 2017;5:4–11. https://doi.org/10.7150/jgen.18378.

[9] Buccheri MA, Salvo E, Coci M, Quero GM, Zoccarato L, Privitera V, et al. Investigating microbial indicators of anthropogenic marine pollution by 16S and 18S High-Throughput Sequencing (HTS) library analysis. FEMS Microbiol Lett 2019;366. https://doi.org/10.1093/femsle/fnz179.

[10] Cannon M V, Craine J, Hester J, Shalkhauser A, Chan ER, Logue K, et al. Dynamic microbial populations along the Cuyahoga River. PLoS One 2017;12:e0186290. https://doi.org/10.1371/journal.pone.0186290.

[11] Bukin YS, Galachyants YP, Morozov I V, Bukin S V, Zakharenko AS, Zemskaya TI. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data 2019;6:190007. https://doi.org/10.1038/sdata.2019.7.

[12] Savio D, Sinclair L, Ijaz UZ, Parajka J, Reischer GH, Stadler P, et al. Bacterial diversity along a 2600 km river continuum. Env Microbiol 2015;17:4994–5007. https://doi.org/10.1111/1462-2920.12886.

[13] Reza MS, Mizusawa N, Kumano A, Oikawa C, Ouchi D, Kobiyama A, et al. Metagenomic analysis using 16S ribosomal RNA genes of a bacterial community in an urban stream, the Tama River, Tokyo. Fish Sci 2018;84:563–77. https://doi.org/10.1007/s12562-018-1193-6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[14] Godoy RG, Marcondes MA, Pessoa R, Nascimento A, Victor JR, Duarte A, et al. Bacterial community composition and potential pathogens along the Pinheiros River in the southeast of Brazil. Sci Rep 2020;10:9331. https://doi.org/10.1038/s41598-020-66386-y.

[15] McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN, Sogin ML. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. Env Microbiol 2010;12:378–92. https://doi.org/10.1111/j.1462-2920.2009.02075.x.

[16] Beale DJ, Karpe A V, Ahmed W, Cook S, Morrison PD, Staley C, et al. A Community Multi-Omics Approach towards the Assessment of Surface Water Quality in an Urban River System. Int J Env Res Public Heal 2017;14. https://doi.org/10.3390/ijerph14030303.

[17] Yang L, Li Y, Su F, Li H. A study of the microbial metabolomics analysis of subsurface wastewater infiltration system. RSC Adv 2019;9:39674–83. https://doi.org/10.1039/C9RA05290A.

[18] 16S Metagenomic Sequencing Library Preparation 2020. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf (accessed July 19, 2020).

[19] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.

[20] R: A language and environment for statistical computing. 2020. https://www.r-project.org/ (accessed July 19, 2020).

[21] Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 2015;31:3476–82. https://doi.org/10.1093/bioinformatics/btv401.

[22] A full example workflow for amplicon data n.d. https://astrobiomike.github.io/amplicon/dada2_workflow_ex.

[23] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;41:D590-6. https://doi.org/10.1093/nar/gks1219.

[24] Wright ES. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. R J 2016;8:352–9.

[25] McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8:e61217. https://doi.org/10.1371/journal.pone.0061217.

[26] McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.

[27] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.

[28] ampvis2: an R package to analyse and visualise 16S rRNA amplicon data 2020. https://madsalbertsen.github.io/ampvis2/ (accessed July 19, 2020).

[29] Risk Group Database 2020. https://my.absa.org/tiki-index.php?page=Riskgroups (accessed July 19, 2020).

[30] Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, et al. Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Human Microbiomes. PLoS One 2016;11:e0166104. https://doi.org/10.1371/journal.pone.0166104.

[31] Narayan NR, Weinmaier T, Laserna-Mendieta EJ, Claesson MJ, Shanahan F, Dabbagh K, et al. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. BMC Genomics 2020;21:56. https://doi.org/10.1186/s12864-019-6427-1.

[32] Venny. An interactive tool for comparing lists with Venn's diagrams. n.d. https://bioinfogp.cnb.csic.es/tools/venny/ (accessed July 22, 2020).

[33] Arana C, Carlo TA, Salinas L. Biological soil crust in Peru: first record and description. Zo Áridas 2016;16:112. https://doi.org/10.21704/za.v16i1.632.

[34] Payne JT, Millar JJ, Jackson CR, Ochs CA. Patterns of variation in diversity of the Mississippi river microbiome over 1,300 kilometers. PLoS One 2017;12:e0174890. https://doi.org/10.1371/journal.pone.0174890.

[35] Casamayor EO. Towards a Microbial Conservation Perspective in High Mountain Lakes. In: J. C, J. N, M A, editors. High Mt. Conserv. a Chang. World. Adv. Glob. Chang. Res., 2017, p. 157–80. https://doi.org/10.1007/978-3-319-55982-7_7.

[36] Sommaruga R, Casamayor EO. Bacterial 'cosmopolitanism' and importance of local environmental factors for community composition in remote high-altitude lakes. Freshw Biol 2009;54:994–1005. https://doi.org/10.1111/j.1365-2427.2008.02146.x.

[37] Xing P, Hahn MW, Wu QL. Low Taxon Richness of Bacterioplankton in High-Altitude Lakes of the Eastern Tibetan Plateau, with a Predominance of Bacteroidetes and Synechococcus spp. Appl Environ Microbiol 2009;75:7017–25. https://doi.org/10.1128/AEM.01544-09.

[38] Cui Q, Huang Y, Wang H, Fang T. Diversity and abundance of bacterial pathogens in urban rivers impacted by domestic sewage. Environ Pollut 2019;249:24–35. https://doi.org/10.1016/j.envpol.2019.02.094.

[39] Shrestha RG, Tandukar S, Bhandari D, Sherchan SP, Tanaka Y, Sherchand JB, et al. Prevalence of Arcobacter and Other Pathogenic Bacteria in River Water in Nepal. Water 2019;11:1416. https://doi.org/10.3390/w11071416.

[40] Collado L, Kasimir G, Perez U, Bosch A, Pinto R, Saucedo G, et al. Occurrence and diversity of Arcobacter spp. along the Llobregat River catchment, at sewage effluents and in a drinking water treatment plant. Water Res 2010;44:3696–702. https://doi.org/10.1016/j.watres.2010.04.002.

[41] Fisher JC, Levican A, Figueras MJ, McLellan SL. Population dynamics and ecology of Arcobacter in sewage. Front Microbiol 2014;5:525. https://doi.org/10.3389/fmicb.2014.00525.

[42] Ferreira S, Queiroz JA, Oleastro M, Domingues FC. Insights in the pathogenesis and resistance of Arcobacter□: A review. Crit Rev Microbiol 2015:1–20. https://doi.org/10.3109/1040841X.2014.954523.

[43] Kristensen JM, Nierychlo M, Albertsen M, Nielsen PH. Bacteria from the Genus Arcobacter Are Abundant in Effluent from Wastewater Treatment Plants. Appl Environ Microbiol 2020;86. https://doi.org/10.1128/AEM.03044-19.

[44] Jacquiod S, Brejnrod A, Morberg SM, Abu Al-Soud W, Sorensen SJ, Riber L. Deciphering conjugative plasmid permissiveness in wastewater microbiomes. Mol Ecol 2017;26:3556–71. https://doi.org/10.1111/mec.14138.

[45] Zerpa Larrauri R, Alarcón Villaverde JO, Lezama Vigo PE, Patiño Gabriel L, Reyes Dioses A, Valencia Ramírez AM, et al. Identificación de Arcobacter en heces de niños y adultos con/sin diarrea y en reservorios animales. An La Fac

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Med 2014;75. https://doi.org/10.15381/anales.v75i2.8389.

[46] Barboza K, Cubillo Z, Castro E, Redondo-Solano M, Fernández-Jaramillo H, Echandi MLA. First isolation report of Arcobacter cryaerophilus from a human diarrhea sample in Costa Rica. Rev Inst Med Trop Sao Paulo 2017;59. https://doi.org/10.1590/s1678-9946201759072.

[47] Eisenberg T, Fawzy A, Nicklas W, Semmler T, Ewers C. Phylogenetic and comparative genomics of the family Leptotrichiaceae and introduction of a novel fingerprinting MLVA for Streptobacillus moniliformis. BMC Genomics 2016;17:864. https://doi.org/10.1186/s12864-016-3206-0.

[48] Loch TP, Faisal M. Emerging flavobacterial infections in fish: A review. J Adv Res 2015;6:283–300. https://doi.org/10.1016/j.jare.2014.10.009.

[49] Good C, Davidson J, Wiens GD, Welch TJ, Summerfelt S. Flavobacterium branchiophilum and F. succinicans associated with bacterial gill disease in rainbow trout Oncorhynchus mykiss (Walbaum) in water recirculation aquaculture systems. J Fish Dis 2015;38:409–13. https://doi.org/10.1111/jfd.12249.

[50] Pandey G, Jain RK. Bacterial Chemotaxis toward Environmental Pollutants: Role in Bioremediation. Appl Environ Microbiol 2002;68:5789–95. https://doi.org/10.1128/AEM.68.12.5789-5795.2002.

[51] Mittal P, Prasoodanan PK V, Dhakan DB, Kumar S, Sharma VK. Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes. Environ Microbiome 2019;14:5. https://doi.org/10.1186/s40793-019-0345-3.

[52] Detmers FJ., Lanfermeijer FC, Poolman B. Peptides and ATP binding cassette peptide transporters. Res Microbiol 2001;152:245–58. https://doi.org/10.1016/S0923-2508(01)01196-2.

[53] Emge P, Moeller J, Jang H, Rusconi R, Yawata Y, Stocker R, et al. Resilience of bacterial quorum sensing against fluid flow. Sci Rep 2016;6:33115. https://doi.org/10.1038/srep33115.

[54] Allocati N, Federici L, Masulli M, Di Ilio C. Glutathione transferases in bacteria. FEBS J 2009;276:58–75. https://doi.org/10.1111/j.1742-4658.2008.06743.x.

[55] Amos GCA, Zhang L, Hawkey PM, Gaze WH, Wellington EM. Functional metagenomic analysis reveals rivers are a reservoir for diverse antibiotic resistance genes. Vet Microbiol 2014;171:441–7. https://doi.org/10.1016/j.vetmic.2014.02.017.

[56] Takahashi Y, Nakashima T. Actinomycetes, an Inexhaustible Source of Naturally Occurring Antibiotics. Antibiotics 2018;7:45. https://doi.org/10.3390/antibiotics7020045.

[57] Sun B, Yuan L, Xia Z, Wan C, Zhang L. Streptomyces albicerus sp. nov., a novel actinomycete isolated from the sediments of the Tailan River in Xinjiang, China. Arch Microbiol 2020;202:1639–46. https://doi.org/10.1007/s00203-020-01871-6.

[58] Leiva P S, Yáñez S M, Zaror C L, Rodríguez S H, García-Quintana H. Actividad antimicrobiana de actinomycetes aislados desde ambientes acuáticos del sur de Chile. Rev Med Chil 2004;132. https://doi.org/10.4067/S0034-98872004000200003.

[59] Kuntke F, de Jonge N, Hesselsøe M, Lund Nielsen J. Stream water quality assessment by metabarcoding of invertebrates. Ecol Indic 2020;111:105982. https://doi.org/10.1016/j.ecolind.2019.105982.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.16.252965; this version posted August 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1. Area of study and sampling localities. The Rimac basin (in gray) is situated between the central Peruvian Andes and the Pacific Ocean. The river runs across Callao, Huarochiri and Lima provinces in the Lima region. Upper Rimac (>1 500 msl): 1, Chicla; 2, San Mateo; 3, Tamboraque; 4, Huanchor; 5, Chacahuaro. Middle Rimac (300 - 1 500 msl): 6, Santa Eulalia; 7, Chaclacayo; 8, Huachipa. Lower Rimac (Metropolitan area, 0 – 300 msl): 9, Libertadores; 10, Nuevo; 11, Universitaria; 12, Faucett; 13, Gambetta.

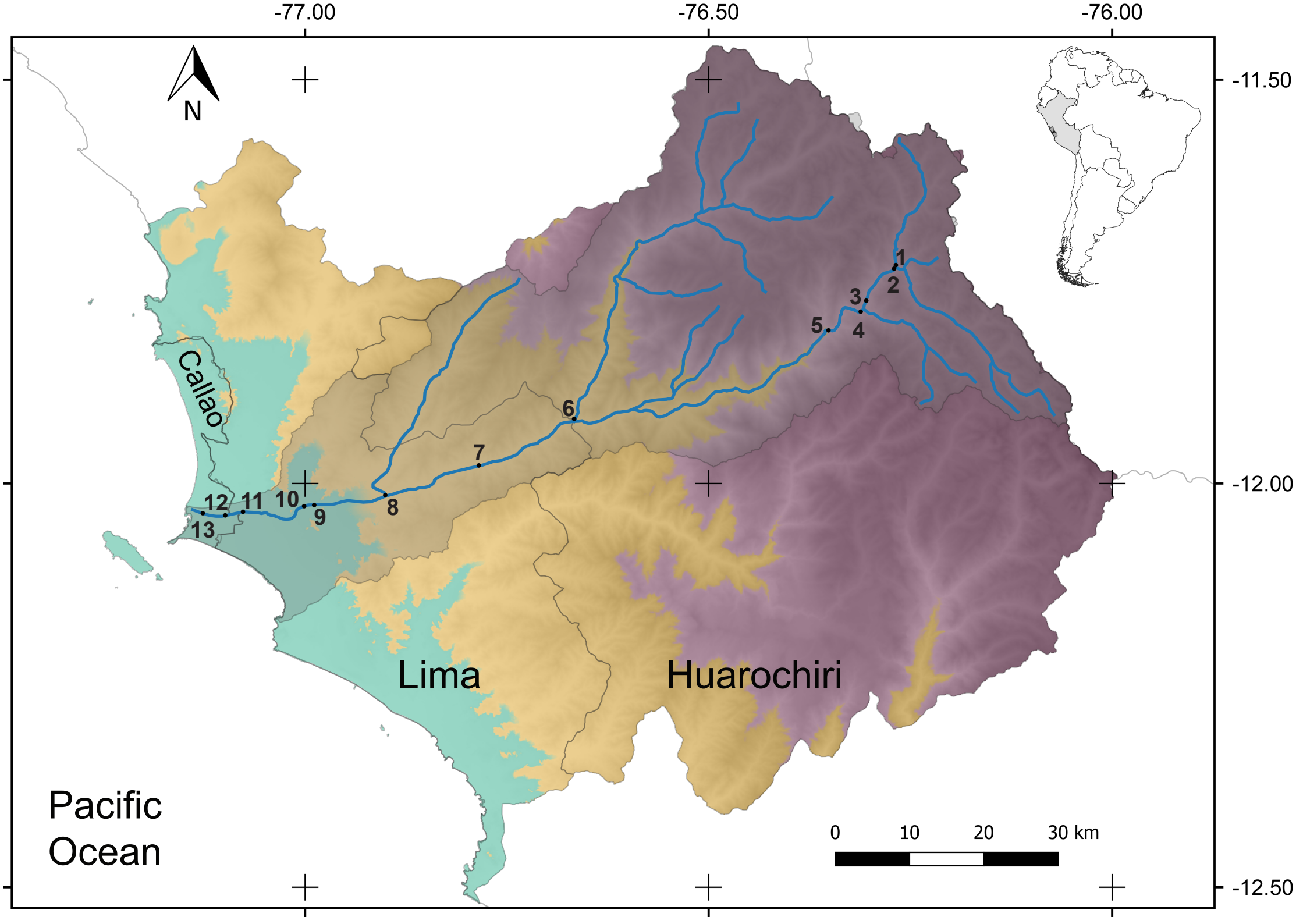
Figure 2. Alpha diversity indexes. Significant differences among the groups are denoted by ** ($p<0.05$), * ($p<0.01$), ns: non-significant.

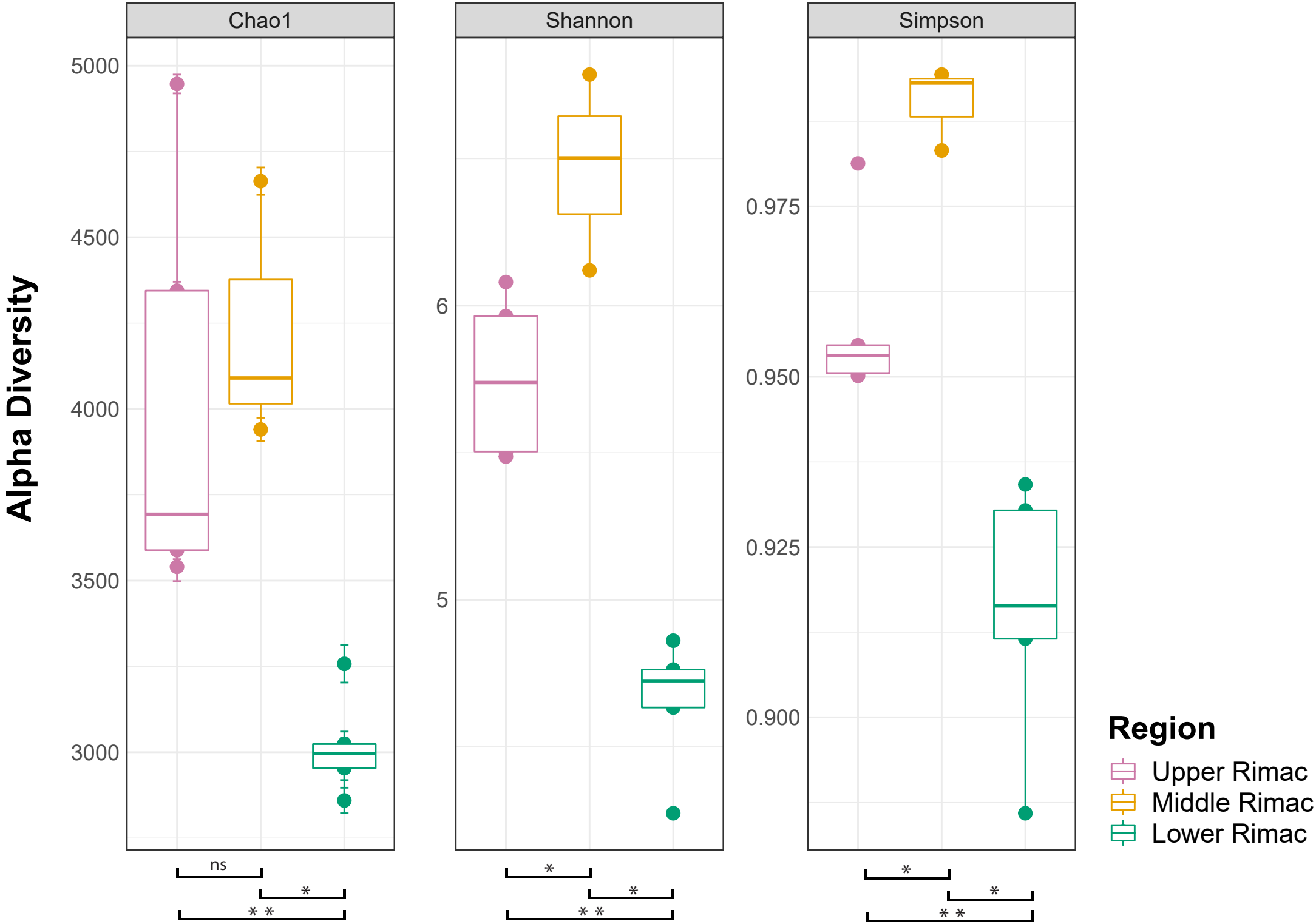
Figure 3. Multi-dimensional scaling (MDS). Principal coordinate analysis (PCoA) of ASV shows differences among groups of samples. Upper Rimac (light purple), Middle Rimac (orange), Lower Rimac (green).

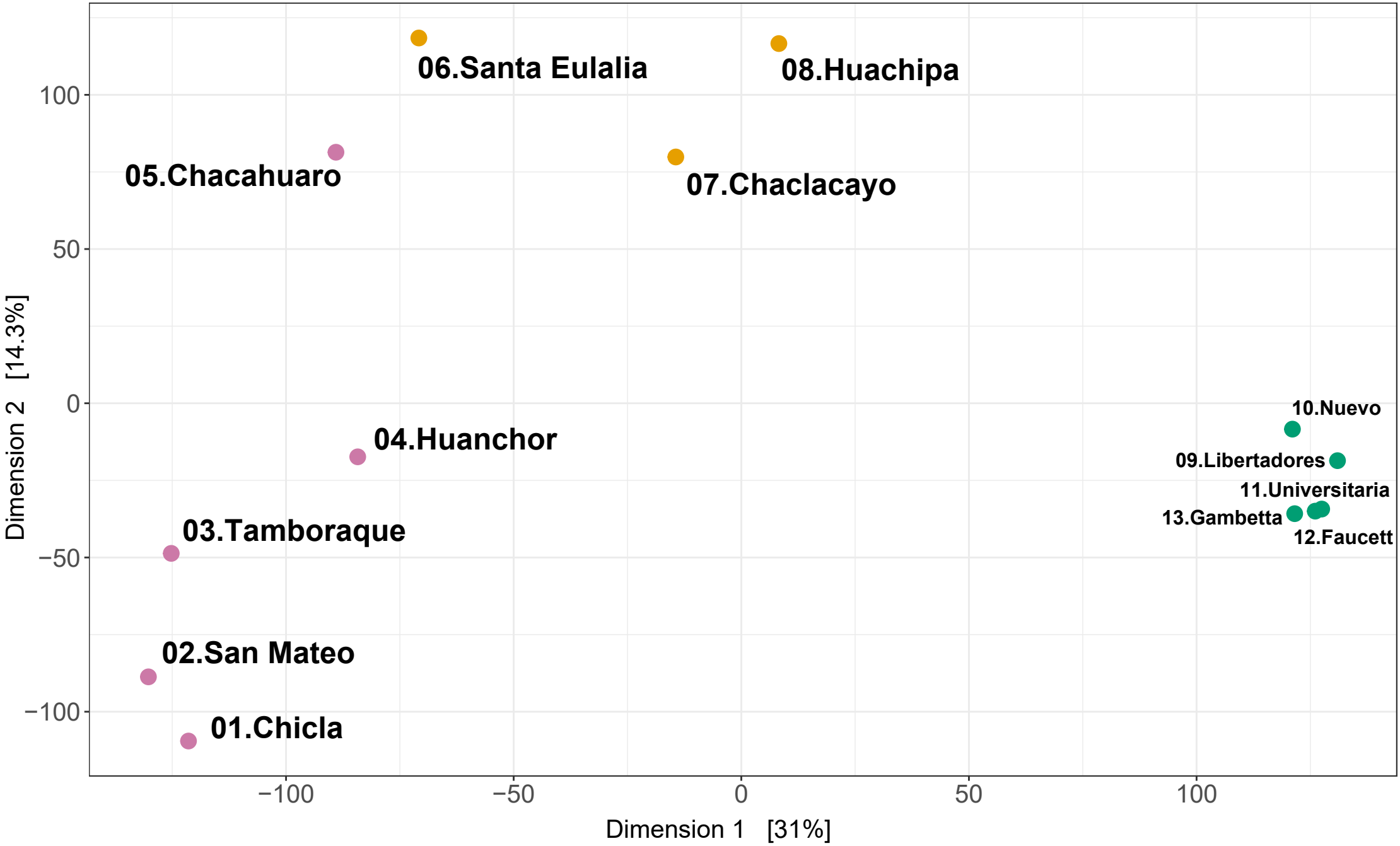
Figure 4. Stack barplot of the frequencies from the most common bacterial classes in each locality. Samples are numbered according to the information in Figure 1.

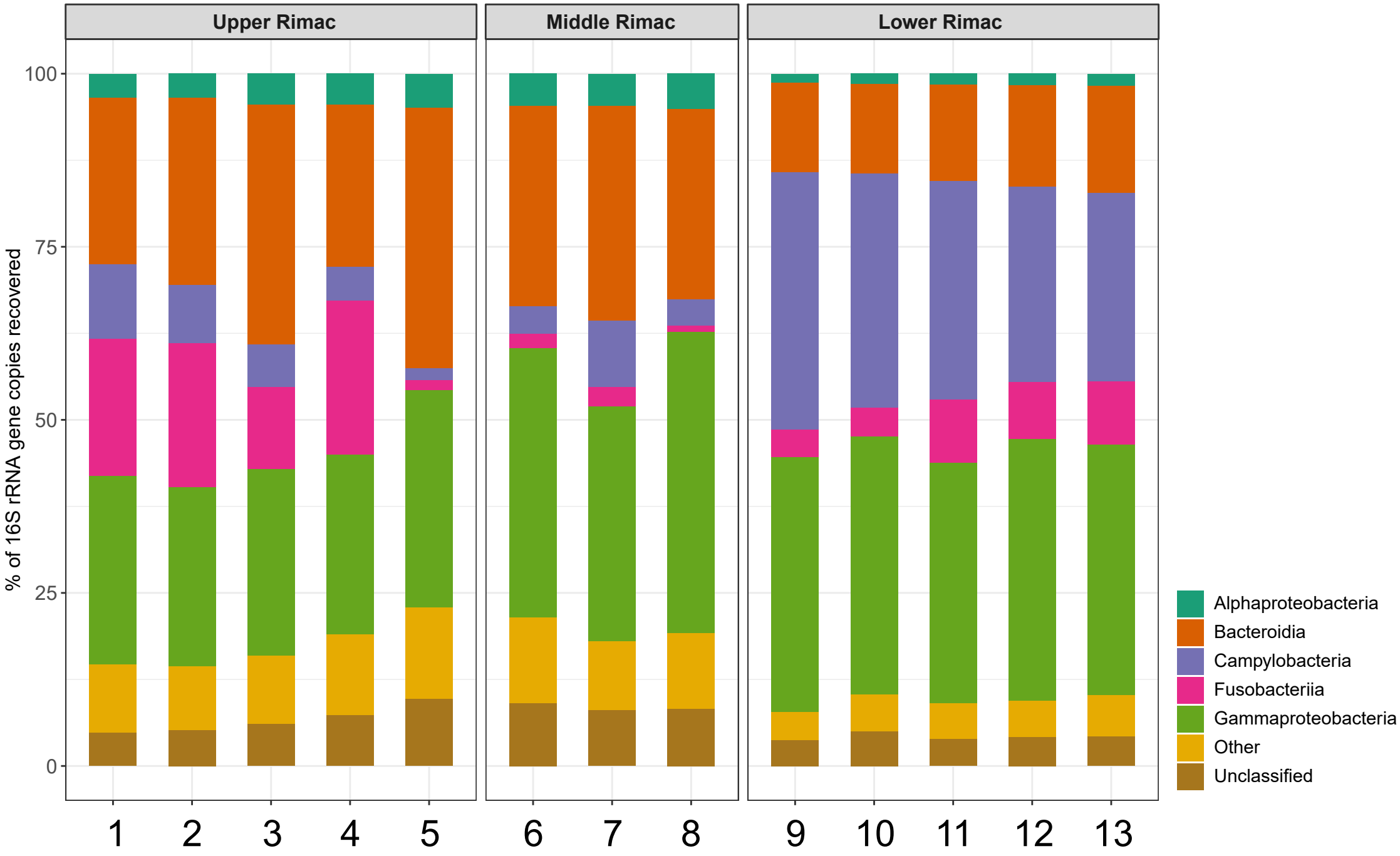
Figure 5. Heatmap of the abundance from the ten most frequent bacterial genera ordered from the most frequent to the less frequent. Classes codes: B, Bacterioidota; F, Fusobacteriia; G, Gammaproteobacteria. Three ASV could not be assigned to the genus level: Leptotrichiaceae (ASV2), Paludibacteraceae (ASV5), and Comamonadaceae (ASV8). Samples are numbered according to the information in Figure 1.

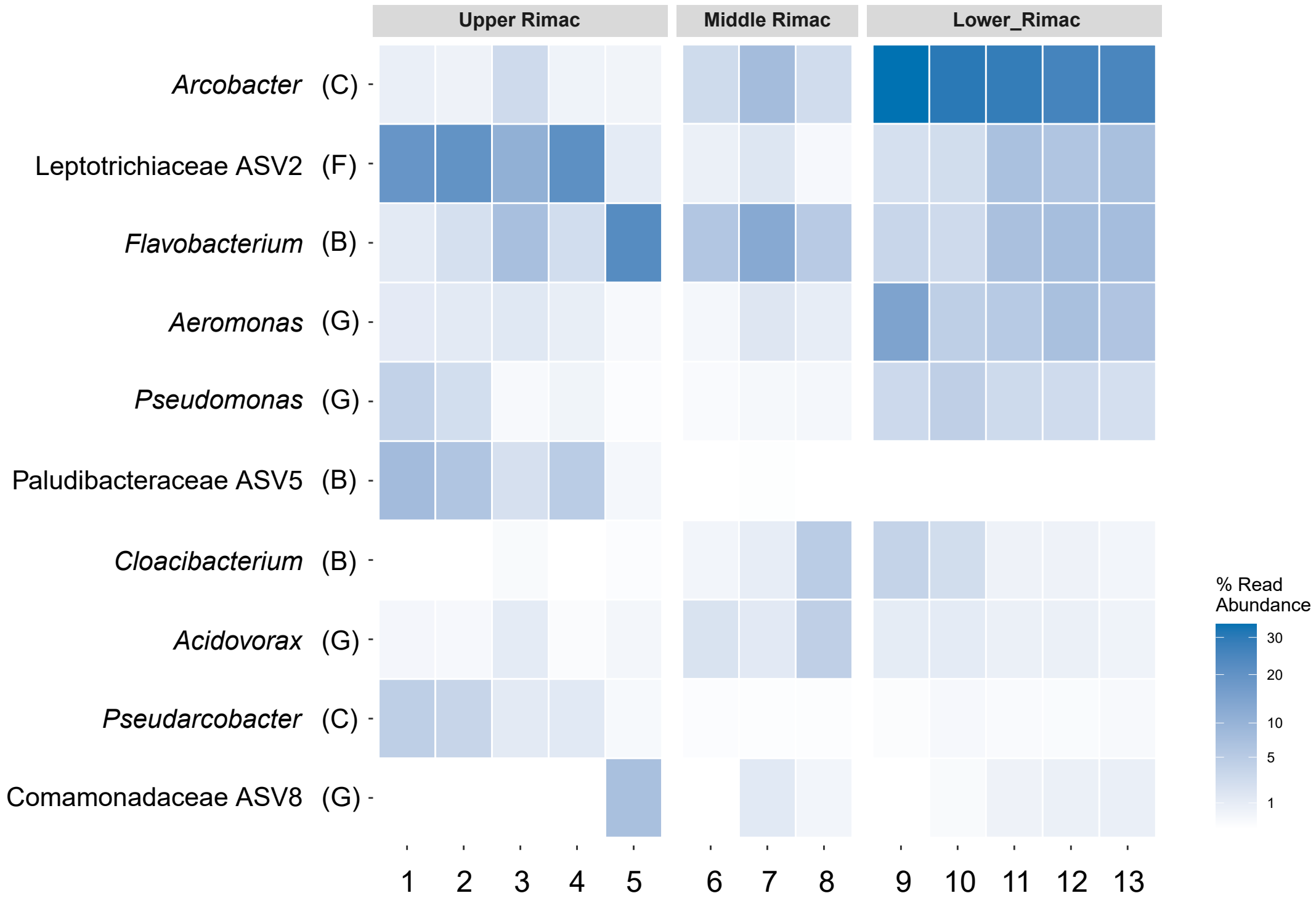
Supplementary material can be requested to pedro.romero@upch.pe


-77.00
-76.50
-76.00
-11.50
-12.00
-12.50
N
Callao
Lima
Huarochiri
Pacific
Ocean
1
2
3
4
5
6
7
8
9
10
11
12
13
0
10
20
30 km


Chao1
Shannon
Simpson
Alpha Diversity
5000
4500
4000
3500
3000
6
5
0.975
0.950
0.925
0.900
ns
*
* *
Region
Upper Rimac
Middle Rimac
Lower Rimac


06.Santa Eulalia
08.Huachipa
05.Chacahuaro
07.Chaclacayo
04.Huanchor
03.Tamboraque
02.San Mateo
01.Chicla
10.Nuevo
09.Libertadores
11.Universitaria
13.Gambetta
12.Faucett
100
50
0
−50
−100
Dimension 2 [14.3%]
−100
−50
0
50
100
Dimension 1 [31%]


Upper Rimac
Middle Rimac
Lower Rimac
% of 16S rRNA gene copies recovered
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Alphaproteobacteria
Bacteroidia
Campylobacteria
Fusobacteriia
Gammaproteobacteria
Other
Unclassified


Upper Rimac
Middle Rimac
Lower_Rimac
Arcobacter (C)
Leptotrichiaceae ASV2 (F)
Flavobacterium (B)
Aeromonas (G)
Pseudomonas (G)
Paludibacteraceae ASV5 (B)
Cloacibacterium (B)
Acidovorax (G)
Pseudarcobacter (C)
Comamonadaceae ASV8 (G)
1 2 3 4 5 6 7 8 9 10 11 12 13
% Read Abundance
30
20
10
5
1