



OPEN

# Assessment of the plasmidome of an extremophilic microbial community from the Diamante Lake, Argentina

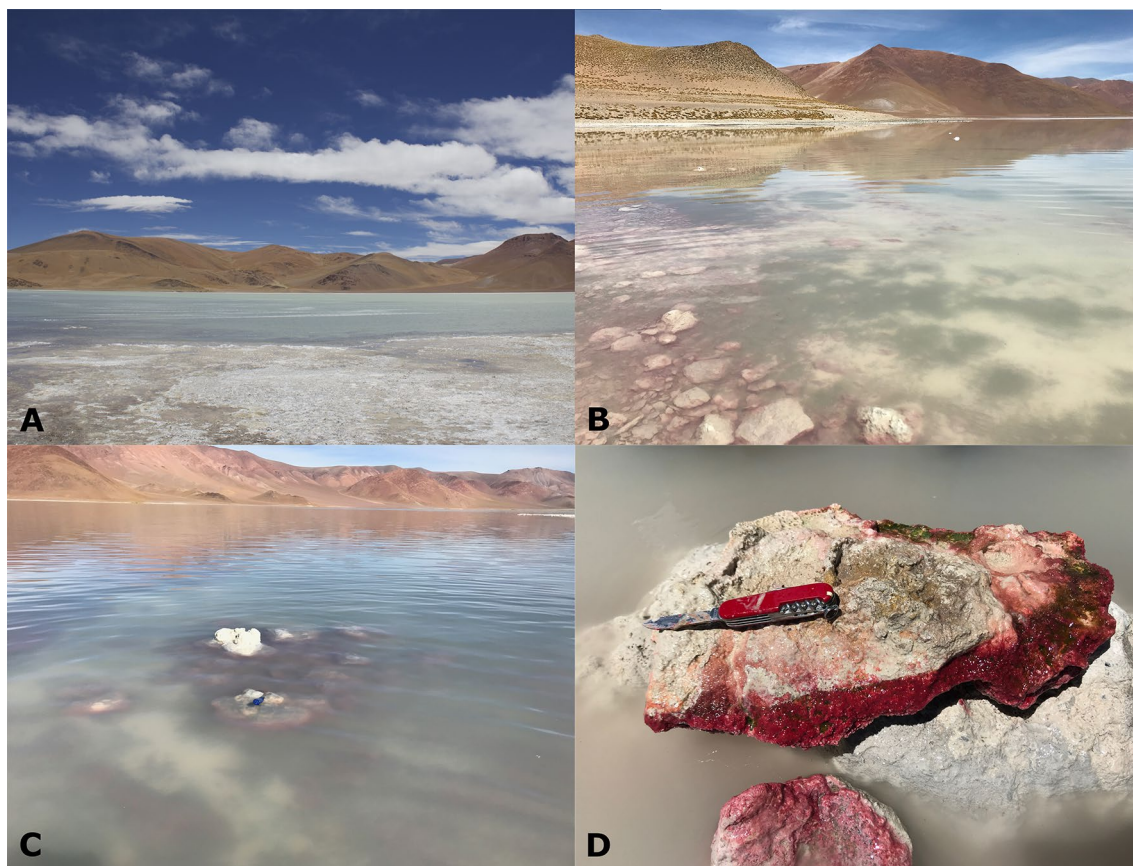
María Florencia Perez<sup>1</sup>, Luis Alberto Saona<sup>1</sup>, María Eugenia Farías<sup>1</sup>, Anja Poehlein<sup>2</sup>, Friedhelm Meinhardt<sup>3</sup>, Rolf Daniel<sup>2</sup>✉ & Julián Rafael Dib<sup>1,4</sup>✉

Diamante Lake located at 4589 m.a.s.l. in the Andean Puna constitutes an extreme environment. It is exposed to multiple extreme conditions such as an unusually high concentration of arsenic (over 300 mg L<sup>-1</sup>) and low oxygen pressure. Microorganisms thriving in the lake display specific genotypes that facilitate survival, which include at least a multitude of plasmid-encoded resistance traits. Hence, the genetic information provided by the plasmids essentially contributes to understand adaptation to different stressors. Though plasmids from cultivable organisms have already been analyzed to the sequence level, the impact of the entire plasmid-borne genetic information on such microbial ecosystem is not known. This study aims at assessing the plasmidome from Diamante Lake, which facilitates the identification of potential hosts and prediction of gene functions as well as the ecological impact of mobile genetic elements. The deep-sequencing analysis revealed a large fraction of previously unknown DNA sequences of which the majority encoded putative proteins of unknown function. Remarkably, functions related to the oxidative stress response, DNA repair, as well as arsenic- and antibiotic resistances were annotated. Additionally, all necessary capacities related to plasmid replication, mobilization and maintenance were detected. Sequences characteristic for megaplasmids and other already known plasmid-associated genes were identified as well. The study highlights the potential of the deep-sequencing approach specifically targeting plasmid populations as it allows to evaluate the ecological impact of plasmids from (cultivable and non-cultivable) microorganisms, thereby contributing to the understanding of the distribution of resistance factors within an extremophilic microbial community.

Plasmids are self-replicating, extrachromosomal, mobile, genetic elements of ecological importance, as they may confer functions or beneficial traits enabling their hosts to thrive in a given environment and—equally important—they can act as horizontal gene transfer vehicles<sup>1</sup> thereby contributing to the spread of genetic information within a microbial community. Thus, plasmids act as important evolutionary driving force by accelerating genome innovation and allowing acquisition of evolutionary novelty<sup>2</sup>.

There are myriads of studies addressing bacterial and archaeal plasmids, which eventually revealed the typical functions ensuring plasmid replication, mobilization and maintenance, as well as other accessory characteristics<sup>3–12</sup>. Only quite recently, however, it was recognized that to characterize them as a whole is necessary to understand their ecological impact in microbial ecosystems<sup>13–18</sup>. In this regard, the total plasmid DNA present in an ecological niche was defined as the “plasmidome”<sup>19</sup>. Indeed, plasmidome studies based on independent microbial culture methods substantiated the significance of extrachromosomal genetic elements with respect to different environments, such as the bovine rumen and wastewater treatment plants<sup>13–15</sup>. We have recently reported the study of the plasmidome of the Puquío de Campo Naranja, an extreme environment of

<sup>1</sup>Planta Piloto de Procesos Industriales Microbiológicos, Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Tucumán, Argentina. <sup>2</sup>Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany. <sup>3</sup>Institut für Molekulare Mikrobiologie und Biotechnologie, Westfälische Wilhelms Universität Münster, Münster, Germany. <sup>4</sup>Instituto de Microbiología, Facultad de Bioquímica, Química y Farmacia, Universidad Nacional de Tucumán, San Miguel de Tucumán, Tucumán, Argentina. ✉email: rdaniel@gwdg.de; jdib@conicet.gov.ar



**Figure 1.** (A) Panoramic photography of the Diamante Lake, located in the Catamarca province, Argentina. (B–C) Submerged microbialites. (D) Red biofilms attached to gaylussite crystals at the bottom of the submerged microbialite.

the Puna Argentina<sup>18</sup>. However, despite the ecological relevance of plasmidomes such studies were only rarely performed, which is possibly due to the complexity of sample processing and the need of tailor-made bioinformatics tools for analysis.

Andean Puna constitutes a large reservoir of Andean Microbial Ecosystems (AMEs) including biofilms, microbial mats, microbialites and endoevaporites, all of which exposed to multiple extreme conditions such as high UV irradiation, and—due to the high altitude—low oxygen pressure, large thermal fluctuations, high dryness, hypersalinity, alkalinity, high concentrations of heavy metals and metalloids such as arsenic<sup>20,21</sup>. The highest concentration of As (up to 347 mg L<sup>-1</sup> in the summertime) was reported for the Diamante Lake<sup>22</sup>, which is located inside the Galán Volcano boiler (40 km diameter) at an altitude of 4589 m above sea level in the Catamarca province, Argentina (Fig. 1). Physico-chemical parameters include high pH-values (9–11), elevated salinity (270 g L<sup>-1</sup>, 217 mS cm<sup>-1</sup>), strong UV irradiation (84 W m<sup>-2</sup> of UV-AB at noon) and vast day–night temperature ranges (–20 °C to +20 °C)<sup>23</sup>. Microbial communities facing such an environment need to develop capacities ensuring survival. As stated above, extrachromosomal genetic elements could procure the genes enabling microbes to withstand harsh environmental conditions<sup>24</sup>. Rascovan et al.<sup>25</sup> reported the discovery of a red biofilm consisting of 94% archaeal representatives, with members belonging to the class of the haloarchaea dominating. In the conventional metagenomic studies already carried out on such microbial communities<sup>25,26</sup>, a high abundance of genes related to anaerobic arsenate respiration (*arr*) as well as arsenite oxidation (*aio*) was detected, strongly supporting the assumption that arsenic is used in bioenergetic processes. In addition, genes related to detoxifying mechanisms for removal of intracellular arsenic were also found, e.g. the *acr3* gene and the *arsABCRD* operon. However, in those studies, the impact of plasmids is possibly underestimated because it is difficult to discriminate them from the chromosomal DNA, due to their low number of copies and corresponding low proportion. In this study, we assessed the plasmidome of the red biofilm from Diamante Lake (Fig. 1D), and compare it to the previously reported plasmidome of microbialites from Puquio de Campo Naranja<sup>18</sup>. In addition, we compared it to the plasmidome from a wastewater-treatment-plant containing effluents of the chemical/pharmaceutical industry (WWTP Visp, Switzerland) as it concerns an environment similarly loaded with high concentrations of metals<sup>15</sup>. Potential hosts as well as encoded functions were identified.

## Materials and methods

**Sampling and biomass purification.** Samples were aseptically taken from red biofilms attached to geylussite crystals at the bottom of submerged microbialites in the Diamante Lake, Catamarca, Argentina (26°00′51.04″S, 67°01′46.42″W) in September 2019 (Fig. 1). Microbialites were found at a distance of 2 m from the lake shore. Samples from three randomly chosen sites were taken and pooled to ensure representativeness. Stored in sterile plastic flasks at 4 °C such pooled samples were further processed within a week. Permission for sample collection was granted by the Secretaría de Medio Ambiente, Catamarca, Argentina (No. 22935/2016).

Microorganisms were separated from the sample by using the protocol described by Perez et al.<sup>18</sup>. Pellets obtained from biomass purification were kept at –20 °C until plasmid DNA extraction.

**DNA extractions.** The plasmid DNA was isolated by using the Large-Construct kit as recommended by the manufacturer (Qiagen, Hilden, Germany).

In parallel, metagenomic DNA was extracted from red biofilm samples by using the FastDNA Spin kit for soil as recommended by the manufacturer (MP Biomedicals, CA, USA). The extracted metagenomic DNA served as template for 16S rRNA gene amplicon sequencing.

**Chromosomal DNA removal.** The plasmid DNA was subjected to overnight digestion at 37 °C with exonuclease V Rec BCD (New England Biolabs, Massachusetts, USA) to remove chromosomal DNA. PCR reactions using universal primers covering the V4 region of 16S rRNA gene were performed to check for chromosomal DNA contamination. The primers used for bacteria were F5′-CCTACGGGNGGCWGCAG-3′ (Bac\_341F) and R5′-GGATTAGATACCCBDGTAGTC-3′ (Bac\_785R)<sup>27</sup>; and for archaea F5′-CCCTAYGGGGYGCASCAG-3′ (Arc\_340F) and R5′-ATTAGAKACCCSNGTAGTCC-3′ (Arc\_806R)<sup>28</sup>. The plasmid DNA was purified using the SureClean Plus kit (Bioline, London, UK).

**Sequencing, quality control and assembly.** Illumina shotgun paired-end sequencing libraries were generated from isolated plasmid DNA using the Nextera XT sample preparation kit as recommended by the manufacturer (Illumina, CA, USA). The MiSeq system together with the MiSeq reagent kit version 3 (600-cycle) was used for the plasmidome sequencing as recommended by the manufacturer (Illumina). The quality control of raw sequence reads was carried out with FastQC v0.11.9 and the reads were quality-filtered using Trimmomatic v0.38.0<sup>29</sup>. Finally, the reads were de novo assembled by using SPAdes software v3.9.0 with the -meta parameter to call the metaSPAdes module<sup>30</sup>. Recycler algorithm was used to assemble cyclic sequences, which are likely plasmids, phages and other circular elements from the assembly graphs provided by SPAdes<sup>31</sup>. The bioinformatic analysis pipeline described by Kothari et al.<sup>17</sup> was also used to identify the complete closed circular contigs. The circular elements obtained in both cases were compared to DoriC 10, a database of replication origins in prokaryotic genomes including chromosomes and plasmids<sup>32</sup>.

**Bioinformatic analysis.** The reads generated by sequencing of the plasmid DNA were aligned with the metagenome contigs using the Bowtie2 tool<sup>33</sup>. Metagenome contigs were assembled using SPAdes v3.9.0 from sequencing data of three independent red biofilm samples taken on another occasion and published by Saona et al.<sup>26</sup>.

Annotation and labeling of all the relevant genomic characteristics on plasmidome contigs were done with Prokka v1.14.5<sup>34</sup>. The assembled plasmidome dataset was submitted to the MG-RAST server<sup>35</sup> for functional and taxonomic analysis. Comparisons with the SEED subsystem database were performed by using a maximum *E*-value of 10<sup>-5</sup>. The deduced functional profile of the red biofilm plasmidome was compared with the one derived from the metagenome mentioned above<sup>26</sup> by employing the software STAMP (Statistical Analysis of Metagenomic Profiles)<sup>36</sup>. Comparisons with other plasmidomes were also performed<sup>15,18</sup>.

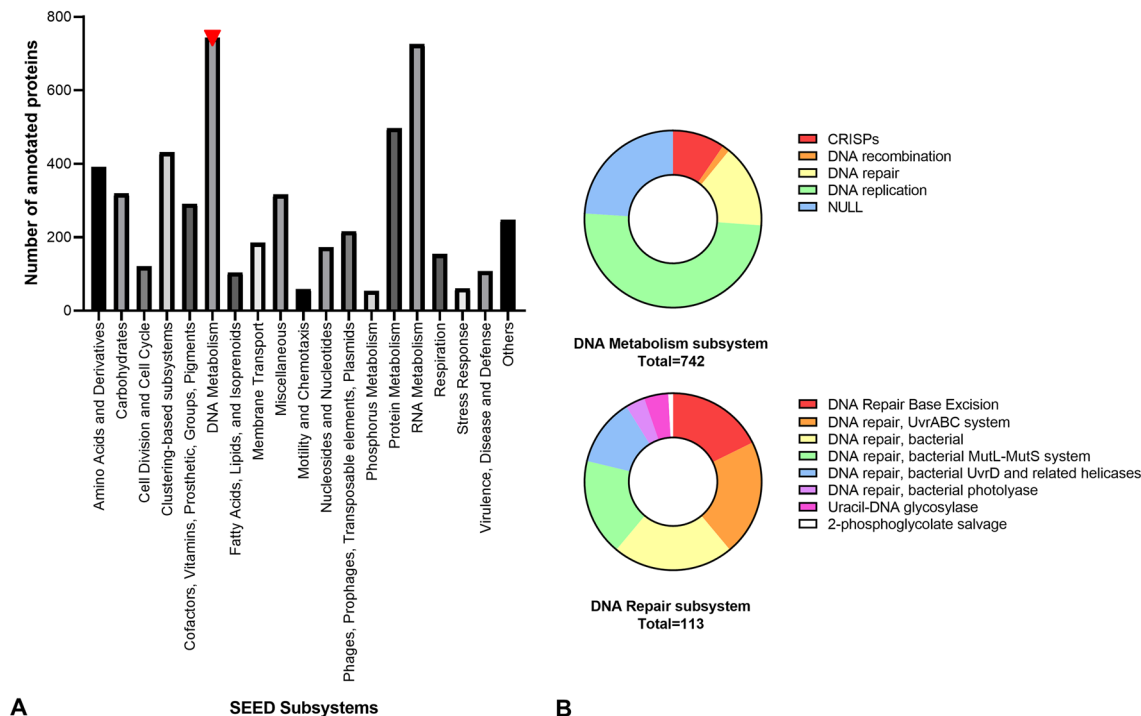
Both the known plasmid sequences from NCBI database and the domains related to plasmid replication and mobilization were assessed as described previously by Perez et al.<sup>18</sup>. In addition, the plasmidome contigs were compared to TADB 2.0 database by blastn in order to identify toxin–antitoxin (TA) systems<sup>37</sup>. Furthermore, the Prokka annotation file of the plasmidome was subjected to Conditional Reciprocal Best BLAST (*crb-blast*) against plasmid genes sequences from the ACLAME database with an *E*-value ≤ 10<sup>-338</sup>. Hits with an identity ≥ 70% and an alignment coverage ≥ 90% were selected. Similarly, putative genes encoding metal resistance and virulence factors were also searched by using the BacMet<sup>39</sup> and VFDB databases<sup>40</sup>, respectively.

Due to the high arsenic concentration found in the lake<sup>22</sup>, arsenic resistance-related genes were separately annotated. For this purpose, the amino acid sequences were downloaded from Uniprot and were subjected to Position-Specific Iterated BLAST (PSI-BLAST)<sup>41</sup>. CD-HIT v4.8.1 was used for creating non-redundant datasets<sup>42</sup> and Clustal Omega v1.2.4 for sequence alignments<sup>43</sup>. Profiles Hidden Markov model were build and searched for in the plasmidome translated gene sequences identified with Prokka by using HMMER 3.3 (cut-off *E*-value < 10<sup>-3</sup>)<sup>44</sup>.

The Resistance Gene Identifier (RGI) software was employed for prediction of antibiotic-resistance genes using the Comprehensive Antibiotic Resistance database (CARD)<sup>45</sup> as a reference.

The ISEScan software pipeline was used to search for mobile elements such as insertion sequences<sup>46</sup>, and the HMM profiles downloaded from TnpPred web<sup>47</sup> for prediction of prokaryotic transposases by HMMER 3.3.

**Amplicon sequencing and taxonomic analysis.** 16S rRNA gene amplicon sequencing was performed using the above-described primers partially covering the 16S rRNA gene sequence. The MiSeq system together with MiSeq reagent kit version 3 (600-cycle) was used for sequencing of the amplicons as recommended by



**Figure 2.** Predicted functional profile of the Diamante Lake plasmidome. **(A)** Number of annotated proteins of each of the major SEED subsystems by using a maximum  $E$ -value of  $10^{-5}$ . SEED subsystems with less than 50 hits were grouped into “Others”. **(B)** On top, “DNA Metabolism” subsystem level 2 classification of the SEED database. On the bottom, “DNA Repair” subsystem level 3 classification of the SEED database ( $E$ -value  $\leq 10^{-5}$ ).

the manufacturer (Illumina). Data quality control and analysis were performed using the QIIME software<sup>48</sup>. First, paired-end reads were joined with PEAR v0.9.6<sup>49</sup>. Quality-filtering was performed using the split\_libraries\_fastq.py script. Forward and reverse primers were removed by using cutadapt v1.16<sup>50</sup>. USEARCH v11<sup>51</sup> was used for zero-radius operational taxonomic unit (zOTU) determination. Taxonomy was assigned against Silva 132 database<sup>52</sup>.

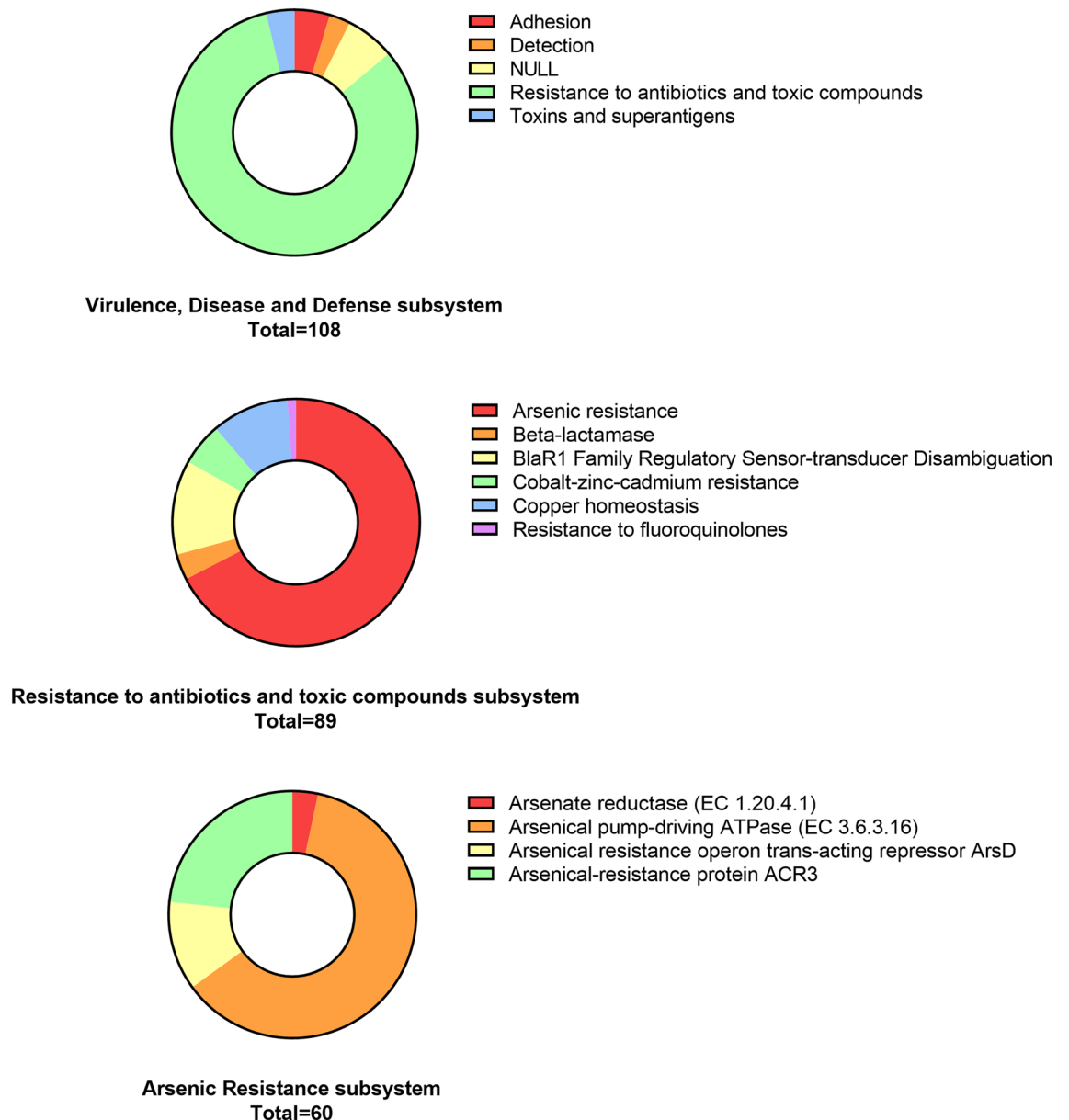
## Results and discussion

**Sequencing and assembly output.** Illumina sequencing from the Diamante Lake plasmidome generated 1,071,941 paired-end reads, of which 941,587 passed quality-filtering. The SPAdes assembler produced 13,492 contigs (> 500 bp) corresponding to roughly 16.9 Mb (largest contig 20.415 bp) (Supplementary Table S1). It is smaller than the previously reported one for another similar extremophile community of Puquico de Campo Naranja (135,813 contigs, 127.9 Mb)<sup>18</sup>.

Thirty-nine closed replicons were predicted by the Recycler software (> 1000 bp), the largest consisting of 3313 bp, but none displayed a known plasmid origin of replication when compared to the DoriC 10.0 database. The bioinformatic pipeline used to detect circularity produced 20 circular contigs. The largest comprised 8295 bp and the smallest 2025 bp. Only one of them showed a known plasmid origin of replication (87% similarity, alignment length of 70 nt), corresponding to pSN found in *Haloterrigena thermotolerans* strain H13 (DoriC ID: pORI00000477). Exclusively hypothetical proteins could be annotated from the open reading frames present in the circular elements.

**Functional analysis.** MG-RAST analysis revealed that the Diamante Lake plasmidome contains a large fraction of unknown DNA as 18,314 sequences (41.5%) code for predicted proteins with known functions but 25,800 sequences (58.5%) encode putative proteins of unknown function. Thus, such deep-sequencing approaches are suited to detect novel proteins with so far undiscovered functions. In our previous plasmidome analysis from an AME, 39% of the predicted proteins could not be functionally annotated<sup>18</sup>. The difference between the two environments displaying rather similar environmental characteristics is possibly due to the large proportion of archaea present in the Diamante Lake, as archaeal genomes contain typically a higher fraction of “dark matter” when compared to bacterial genomes. The isolation and cultivation of most of the archaea, and accordingly, the experimental characterization of archaeal gene products, is challenging<sup>53</sup>. Likewise, Sentchilo et al.<sup>15</sup> also reported 52 and 66% of coding sequences without assigned function in two wastewater treatment plant plasmidomes.

From the functional SEED assignment, only 5196 predicted proteins were annotated (28.4%), most of them covering basic metabolic functions such as DNA, RNA and protein metabolism (Fig. 2A). It is noteworthy that among the proteins involved in DNA metabolism those related to DNA repair were rather diverse (Fig. 2B). It



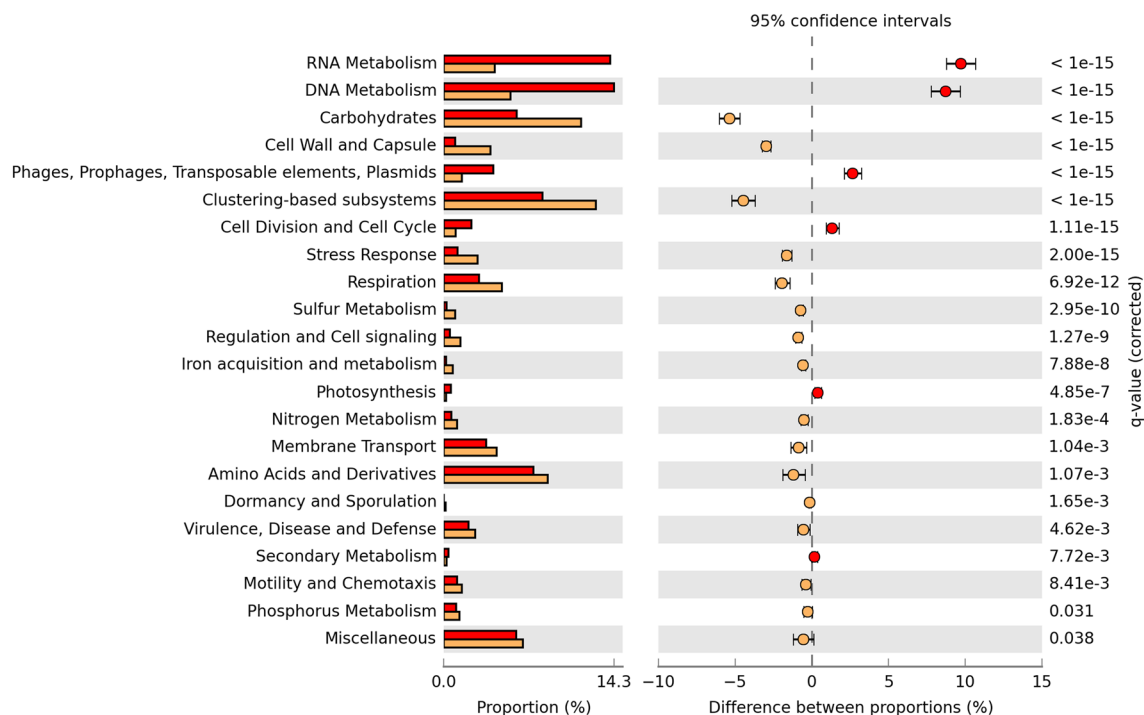
**Figure 3.** Predicted functional profile of the Diamante Lake plasmidome. On top, “Virulence, Disease and Defense” subsystem level 2 classification of the SEED database. In the middle, “Resistance to antibiotics and toxic compounds” subsystem level 3 classification of the SEED database. On the bottom, “Arsenic resistance” subsystem level 4 classification of the SEED database ( $E$ -value  $\leq 10^{-5}$ ).

is well known that DNA repair plays a key role as an adaptive mechanism to withstand the high UV irradiation in the Andean Puna<sup>54–58</sup>.

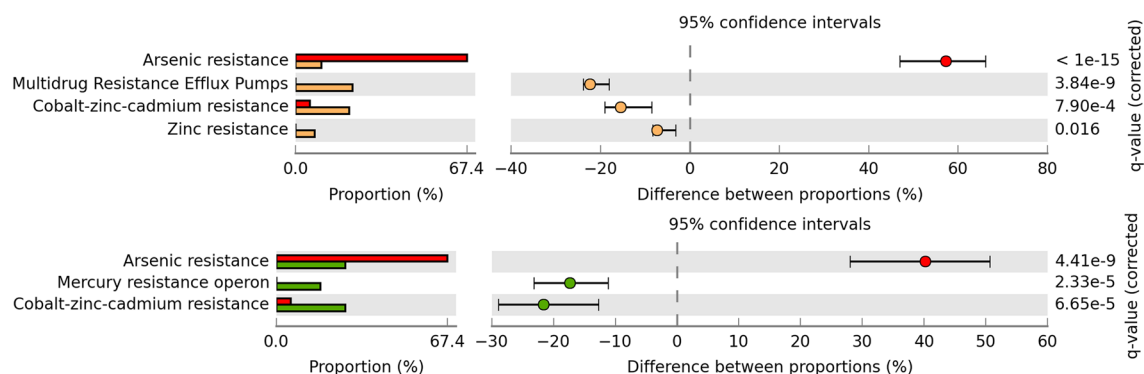
Although on a smaller scale, the subsystems “Stress Response” and “Virulence, Disease and Defense” were represented (Fig. 2A). With respect to stress response, predicted proteins involved in the response to oxidative stress were most frequently found (57.4%) (Supplementary Fig. S1). As for the DNA repair mechanisms, oxidative stress response systems contribute to protect microorganisms from UV-mediated damage. 82.4% of the assignments to the above mentioned second group corresponded to the subsystem “Resistance to antibiotics and toxic compounds”, with arsenic resistance systems prevailing (67.4%). The arsenic resistance included predicted proteins for an arsenate reductase (ArsC), an arsenical pump-driving ATPase (ArsA), an arsenical resistance operon trans-acting repressor (ArsD) and an arsenical-resistance protein ACR3 (Fig. 3).

The unique characteristics of this environment, such as its location at a high altitude, the exposure to extreme conditions and the peculiarities of its microbial composition, which includes a major proportion of archaea, may explain the relatively few functional annotations during grouping into SEED categories.

**Functional comparison between plasmidomes.** The predicted functional profile of the Diamante Lake plasmidome was compared to the one derived from the Puquio de Campo Naranja plasmidome<sup>18</sup>. “RNA



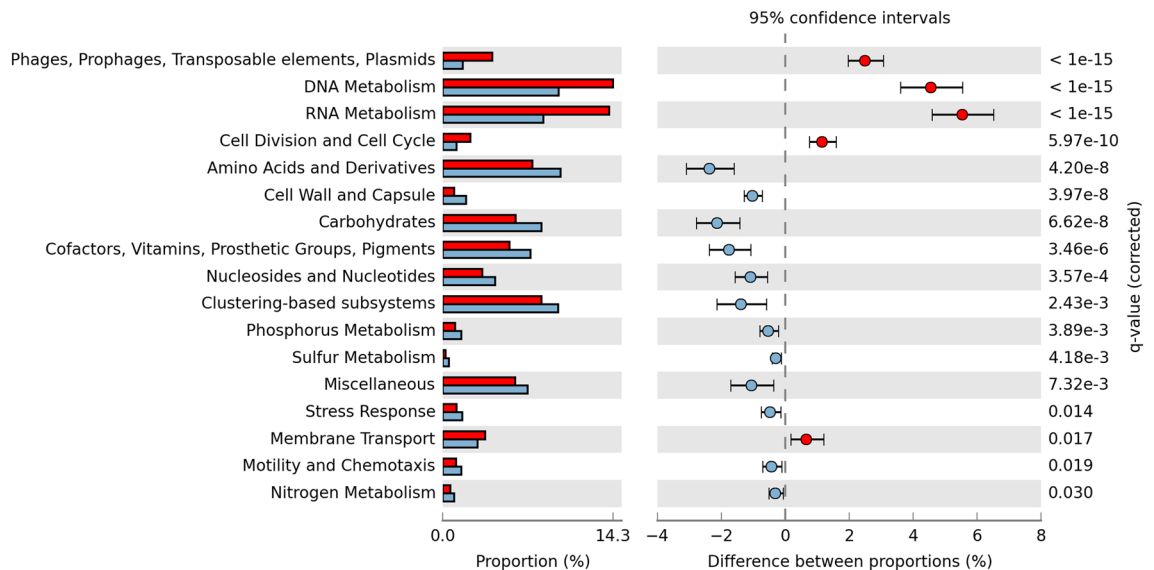
**Figure 4.** Predicted functional profiles derived from the Diamante Lake plasmidome (red bars) and the Puquio de Campo Naranja plasmidome (orange bars). Percent SEED categorizable protein-encoding genes and pairwise proportional differences calculated using STAMP. Fisher's exact test was used and corrected P-values were calculated using Storey's FDR. Only the statistically significant SEED subsystems are shown ( $q < 0.05$ ).



**Figure 5.** Predicted functional profiles derived from the Diamante Lake plasmidome (red bars), the Puquio de Campo Naranja plasmidome (orange bars) and the one derived from the wastewater treatment plant plasmidome in Visp, Switzerland (green bars). Percent “Resistance to antibiotics and toxic compounds” categorizable protein coding genes belonging to the “Virulence, Disease and Defense” SEED subsystem, and pairwise proportional differences calculated using STAMP. Fisher's exact test was used and corrected P-values were calculated using Benjamini–Hochberg's FDR. Only the statistically significant SEED subsystems are shown ( $q < 0.05$ ).

Metabolism”, “DNA Metabolism”, “Phages, Prophages, Transposable elements, Plasmids” and “Cell Division and Cell Cycle” subsystems were more frequently represented in Diamante Lake than in the other, while “Carbohydrates”, “Cell Wall and Capsule”, “Clustering-based subsystems”, “Stress Response”, “Respiration” were more abundant in Puquio de Campo Naranja. No significant differences in the abundances of other subsystems were observed, suggesting a certain degree of similarity of the predicted functional profiles for both of the AMEs (Fig. 4).

Diamante Lake is known to be among the aqueous environments displaying high arsenic concentrations<sup>22,25,59</sup>, reinforcing the repeatedly mentioned presence of genes related to its resistance. Proportional differences between “Resistance to antibiotics and toxic compounds” categorizable protein coding genes belonging to the “Virulence, Disease and Defense” SEED subsystem showed that arsenic resistance is more abundant in the Diamante Lake plasmidome than in that of the Puquio de Campo Naranja (Fig. 5). In addition, the comparison of the former



**Figure 6.** Predicted functional profiles derived from the plasmidome (red bars) and the metagenome (blue bars) of Diamante Lake red biofilm. Percent SEED categorizable protein-encoding genes and pairwise proportional differences calculated using STAMP. G-test (w/Yates') was used and corrected P-values were calculated using Benjamini–Hochberg's FDR. Only the statistically significant SEED subsystems are shown ( $q < 0.05$ ).

with that of a wastewater treatment plant containing effluents from chemical/pharmaceutical industries (WWTP Visp, Switzerland)<sup>15</sup> revealed that arsenic resistance traits were again more abundant in the Diamante Lake plasmidome (Fig. 5).

**Plasmid-purification advantage.** Plasmids usually represent only a small fraction of the total DNA in a given environment, due to their low rate of occurrence and number of copies. Hence, though they are casually recorded by conventional metagenomic sequencing methods, experimental plasmid-purification prior to sequencing allows for an analysis specifically targeting plasmid populations in a culture-independent manner, at best without losing information<sup>60</sup>. Obtained results are in line with such notion, as only 52% of plasmidome reads aligned with metagenome contigs described above in Saona et al.<sup>26</sup>. The same applies to the Puquío de Campo Naranja plasmidome, in which alignment reached only 30%<sup>18</sup>. Thus, our study strengthens that plasmid-purification prior to sequencing more satisfactorily meets the requirements to comprehensively assess the ecological importance of plasmid-borne sequences.

The pairwise comparison aiming at distinguishing the plasmid gene pool from the metagenomic one (Fig. 6) accordingly revealed that “Phages, Prophages, Transposable elements, Plasmids” and “Membrane Transport” subsystems are more frequently represented in the plasmidome.

**Plasmid backbone functions: replication, mobilization and maintenance.** In order to identify plasmid-like traits within the plasmidome, we focused on the search for Pfam domains related to plasmid replication and MOB-type relaxase families, which are related to plasmid mobilization.

RHH\_1 and DUF1424 were detected as the main Pfam domains of plasmid replication in the Diamante Lake plasmidome, followed by RepL (Table 1). Likewise, RHH\_1 and RepL protein families were also the most abundant in the plasmidome from Puquío de Campo Naranja<sup>18</sup>. It was not the case of DUF1424, which is a family of several archaeal proteins that seems to be present exclusively in *Halobacterium* and *Haloferax* species. Although the function of the latter family is unknown, its members are probably rep proteins due to the presence of conserved functional motifs<sup>61,62</sup>.

Rep\_1 and Rep\_3 are the major families of replication initiation proteins. They have been reported among the most abundant in plasmidomes from wastewater treatment plants and a rat cecum<sup>15,16,63,64</sup>. In this study, domains belonging to the Rep\_1 and Rep\_3 families were also detected, but with a lower hit rate. Domains of replication initiation proteins from other known families were not detected (Table 1). Possibly, there are replication systems for which the molecular details and the mechanisms are currently unknown, particularly as most of the contributing microorganisms were not cultured and because of the taxonomic composition dominated by specific taxa, i.e. halobacteria<sup>25</sup>.

The most abundant relaxase families in the plasmidome were MOB<sub>C</sub> and MOB<sub>P</sub>, for which 41 and 18 protein domain matches, respectively, were counted. MOB<sub>T</sub>, MOB<sub>V</sub> and MOB<sub>M</sub> were also present (3, 2 and 1 protein domain matches, respectively) (Table 2). Mobilization elements have been reported in most of the previous plasmidome analyses<sup>14–16,63</sup>, however, the classification in relaxase MOB families proposed by Garcillán-Barcia et al.<sup>65,66</sup> was not performed. Meanwhile, 29 protein domain matches were counted for MOB<sub>T</sub> family in the

Pfam name	Pfam code	Description	Hits
Rep_1	PF01446.17	Replication protein	45
Rep_2	PF01719.17	Plasmid replication protein	1
Rep_3	PF01051.21	Initiator Replication protein	5
RepL	PF05732.11	Firmicute plasmid replication protein (RepL)	85
TrfA	PF07042.11	TrfA protein	3
RepA_C	PF04796.12	Plasmid encoded RepA protein	2
Rep_trans	PF02486.19	Replication initiation factor	0
RHH_1	PF01402.21	Ribbon-helix-helix protein, copG family	344
Rop	PF01815.16	Rop protein	0
RP-C	PF03428.13	Replication protein C N-terminal domain	32
RPA	PF10134.9	Replication initiator protein A	2
RepA_N	PF06970.11	Replication initiator protein A (RepA) N-terminus	5
RepC	PF06504.11	Replication protein C (RepC)	4
Replicase	PF03090.17	Replicase family	5
IncFII_repA	PF02387.15	IncFII RepA protein family	0
PriCT_1	PF08708.11	Primase C terminal 1 (PriCT-1)	4
DUF1424	PF07232.11	Putative rep protein (DUF1424)	179
Phage_CRI	PF05144.14	Phage replication protein CRI	0
PriCT_2	PF08707.11	Primase C terminal 2 (PriCT-2)	1
Prim-Pol	PF09250.11	Bifunctional DNA primase/polymerase, N-terminal	12
SSB	PF00436.25	Single-strand binding protein family	0
RepA1_leader	PF08048.12	Tap RepA1 leader peptide	0
DUF1738	PF08401.11	Domain of unknown function (DUF1738)	0
UvrD-helicase	PF00580.21	UvrD/REP helicase N-terminal domain	12
RepB-RCR_reg	PF10723.9	Replication regulatory protein RepB	30
Replic_Relax	PF13814.6	Replication-relaxation	73
KorB_C	PF06613.11	KorB C-terminal beta-barrel domain	0
KorB	PF08535.10	KorB domain	6
Activator-TraM	PF11657.8	Transcriptional activator TraM	1
pRN1_helical	PF13010.6	Primase helical domain	0
KORA	PF16509.5	TrfB plasmid transcriptional repressor	17
Rol_Rep_N	PF18106.1	Rolling Circle replication initiation protein N-terminal domain	0

**Table 1.** Plasmid replication-related Pfam in the Diamante Lake plasmidome.

Relaxase MOB family	Profile HMM	Hits
MOB <sub>B</sub>	T4SS_MOBB	0
MOB <sub>C</sub>	T4SS_MOBC	41
MOB <sub>F</sub>	Profile_MOBF	0
MOB <sub>H</sub>	T4SS_MOBH	0
MOB <sub>P</sub>	T4SS_MOBP1	18
	T4SS_MOBP2	0
	T4SS_MOBP3	0
MOB <sub>Q</sub>	T4SS_MOBQ	0
MOB <sub>T</sub>	Profile_MOBT	3
MOB <sub>V</sub>	T4SS_MOBV	2
MOB <sub>M</sub>	Profile_MOBM	1

**Table 2.** Relaxase MOB families in the Diamante Lake plasmidome.

Puquico de Campo Naranja plasmidome<sup>18</sup>, and Kothari et al.<sup>17</sup> reported the MOB<sub>Q</sub> and MOB<sub>P</sub> families as the most abundant in circular plasmids from groundwater plasmidomes.

In addition to the above plasmid replication and mobilization entries, sequences harboring genes involved in plasmid maintenance such as loci corresponding to toxin-antitoxin (TA) systems were identified (identity and coverage at least 85%). All of the TA systems belong to type II TA-loci (Supplementary Table S2). Only a



single complete system could be annotated, i.e. a toxin with the respective antitoxin (T2787-AT2787), both located in the same contig (NODE\_2880). It corresponds to the VapBC family, where the toxin is a PIN-domain ribonuclease (145 aa) and the antitoxin is a transcription factor (98 aa)<sup>67</sup>. Interestingly, this system is known from *Haloquadratum walsbyi* DSM 16790, a halophilic archaeon that was isolated from a solar saltern in Brac del Port (Alicante, Spain), and it was found to dominate most of the thalassic NaCl-saturated environments<sup>68</sup>.

In the previous plasmidome studies, TA systems were not taken into consideration. Only Kothari et al.<sup>17</sup> reported the YoeB-YefM and RelE/StbE-RelB/StbD type II TA systems in some circular plasmids from ground-water plasmidomes. YoeB and RelE are ribosome-dependent RNase toxins that bind directly to the A site of the ribosome, where they cleave ribosome-associated mRNA<sup>69</sup>.

**Plasmid accessory functions: antibiotic resistance and arsenic resistance.** Sequence analysis of the plasmidome from the Puquio de Campo Naranja revealed that antibiotic resistance traits are widespread in this extreme pristine environment, as 123 putative antibiotic resistance genes (ARGs) were annotated<sup>18</sup>. In the present study, only 8 ARGs could be classified, conveying resistance to 10 drug classes, among them macrolides, carbapenems, cephalosporins, penams (Supplementary Table S3). Such noticeable difference with respect to the number of ARGs found in similar extreme environments is probably due to microbial regional distinctions. The metabolic processes and the cell walls of bacteria and archaea display significant differences, offering an explanation for the fact that a number of antibiotics are effective against the former but do not threaten the latter<sup>70</sup>. Moreover, studying antibiotic-resistance mechanisms is—due to the clinical relevance—of much more necessity in the bacterial domain, as pathogenic archaea have not yet been identified<sup>71</sup>. So far, only a relationship between the periodontal-disease-severity and the relative abundance of the archaeon *Methanobrevibacter oralis* was reported<sup>72</sup>.

The bias necessarily introduced by the existing databases and developed from the information currently available, as well as the dominance of the archaea in the microbial community studied interfere with the analysis of the resistome encoded by the Diamante Lake plasmidome. Thus, it cannot be excluded that the lack of relevant knowledge is the reason for the low number of identifiable ARGs and virulence factors.

Regarding the resistances to metals, a respective search in the BacMet database produced no hits, possibly also due the above reason as the database consists solely of bacterial entries. However, our manual annotation disclosed the presence of arsenic resistance genes. Arsenic hits various microorganisms, however, several bacteria and archaea possess detoxification systems enabling growth even under high As-concentrations<sup>73</sup>. The most common resistance system is encoded by the *ars* operon for which different genetic organizations were described among prokaryotes<sup>74</sup>. In addition to the extrusion systems, composed of the gene-products encoded by *arsA*, *arsB*, *arsC*, *arsD*, *arsR*, and *acr3*, another mechanism involving a putative arsenite(III)-methyltransferase (*ArsM*) was reported in *Halobacterium* sp. NRC-1<sup>75</sup>. We identified 28 proteins possibly related to arsenic resistance; ten of them had been automatically annotated as "hypothetical proteins" by Prokka (Supplementary Table S4). Hence, automatic annotation bears the risk of less accurate results or the disclosure of a fewer number of genes than actually exist. It is to be emphasized that genes enabling microorganisms to conduct anaerobic arsenate respiration (*arr*) and arsenite oxidation (*aio*) were not detectable in the plasmidome, as these are usually encoded by the chromosome.

As already mentioned, the genetic organization of the *ars* operons can vary among diverse microorganisms. In the Diamante Lake plasmidome, *arsC* (arsenate reductase) and *acr3* (arsenite efflux transporter) genes were present twice in close proximity, but the most relevant was the *arsADR* gene cluster of contig 116 (Supplementary Table S4). The genetic arrangement agrees with that described for pHLAC01 and pNRC100 of *Halorubrum lacusprofundi* ATCC 49239 and *Halobacterium* sp. NRC-1, respectively (Supplementary Fig. S2). In all cases, the *arsDA* and *arsR* genes are transcribed in opposite directions and the absence of the arsenite transporter *ArsB*-encoding gene is noticeable. This otherwise unusual operon structure is apparently characteristic for the haloarchaea<sup>74,75</sup>.

**Plasmid databases: NCBI and ACLAME.** Sequences belonging to 24 megaplasmids described in 13 strains of halophilic archaea isolated from different saline environments were found (Table 3). Most of the matches were with the plasmid of *Halobacterium* sp. DL1 (315 kb), which was isolated from a freshwater pond (NZ\_CP007061.1). Fourteen matches were detected between the plasmidome sequences and the sequences of the plasmid pHLAC01 (431 kb) of *Halorubrum lacusprofundi* ATCC 49239 that was isolated from the Deep Lake, a hypersaline Antarctic site. It is noteworthy that one of these matches (identity 92%, length 4123 bp) corresponds to contig 116 of the plasmidome, which harbors the *arsDA* genes (Supplementary Fig. S3). Thirteen sequences match with plasmid pHTIA (330 kb) of *Halorhabdus tiamatea* SARL4B, which was isolated from the Shaban deep-sea hypersaline anoxic lake in the Red Sea<sup>76</sup>. Thus, such plasmid sequences are evidently preserved in different high salinity environments.

When the red biofilm plasmidome genes annotated by Prokka were compared with plasmid genes of the Aclame database, 125 matches corresponded to genes of 11 megaplasmids of five different halophilic archaeal and one actinobacterial strain (*Rhodococcus* sp. RHA1) (Supplementary Table S5). Most of them were related to DNA metabolism, transposition and recombination. Gene related to arsenic resistance, DNA repair and plasmid partitioning were also identified. Thus, for many of the hypothetical proteins identified by Prokka a function was attributed, but not to all of them. Anyway, the comparison provided further justification of the practice as the presence of known plasmid-associated genes in our plasmidome dataset was proven.

**Mobile genetic elements: transposases and insertion sequences.** A total of 532 insertion sequences (IS) were identified in the Diamante Lake plasmidome, with IS200/IS605 and IS5 like elements being the most frequent ones followed by members of the IS4, IS6, IS630 and ISH3 families (Supplementary Table S6).

Plasmid	Host	Size (kb)	Isolation source	Matches	NCBI accession number	References
pNRC100	<i>Halobacterium</i> sp. NRC-1	191	NE*	6	NC_001869.1	Ng et al. (2000) <sup>82</sup>
pNRC200		365		8	NC_002608.1	
pNG500	<i>Haloarcula marismortui</i> ATCC 43,049	132	Dead Sea	1	NC_006393.1	Baliga et al. (2004) <sup>83</sup>
pNG600		155		7	NC_006394.1	
PHS1	<i>Halobacterium salinarum</i> R1	147	NE	3	NC_010366.1	Pfeiffer et al. (2008) <sup>84</sup>
PHS4		41		1	NC_010367.1	
PHS3		284		4	NC_010368.1	
PHS2		195		2	NC_010369.1	
pHLAC01	<i>Halorubrum lacusprofundi</i> ATCC 49,239	431	Deep Lake, Antarctica	14	NC_012030.1	NA**
pHTUR01	<i>Haloterrigena turkmenica</i> DSM 5511	698	Saline soil in Turkmenistan	1	NC_013744.1	Saunders et al. (2010) <sup>85</sup>
pHV3	<i>Haloferax volcanii</i> DS2	438	Bottom sediment of the Dead Sea	3	NC_013964.1	Hartman et al. (2010) <sup>86</sup>
pHV4		635		4	NC_013966.1	
pHV1		85		1	NC_013968.1	
p1	<i>Halalkalicoccus jeotgali</i> B3	406	Salt-fermented seafood from South Korea	2	NC_014298.1	Roh et al. (2010) <sup>87</sup>
p2		363		1	NC_014299.1	
pHBOR02	<i>Halogeometricum borinquense</i> DSM 11,551	339	Solar salterns of Cabo Rojo, Puerto Rico	1	NC_014731.1	Malfatti et al. (2009) <sup>88</sup>
pHBOR01		362		1	NC_014735.1	
pHBOR03		210		1	NC_014736.1	
pHALXA01	<i>Halopiger xanaduensis</i> SH-6	436	Saline sediment of Lake Shangmata, China	2	NC_015658.1	Anderson et al. (2012) <sup>89</sup>
pHALXA02		181		1	NC_015667.1	
pHTIA	<i>Halorhabdus tiamatea</i> SARL4B	330	Shaban deep-sea hypersaline anoxic lake in the Red Sea	13	NC_021913.1	Werner et al. (2014) <sup>76</sup>
pHH126	<i>Haloarcula hispanica</i> N601	125	Solar saltern in Alicante, Spain	1	NC_023011.1	Ding et al. (2014) <sup>90</sup>
pXH-48	<i>Halostagnicola larsenii</i> XH-48	131	Saline sediment of Lake Xilinhot, China	1	NZ_CP007057.1	Castillo et al. (2006) <sup>91</sup>
NA	<i>Halobacterium</i> sp. DL1	315	Fresh water, Ponds	23	NZ_CP007061.1	NA

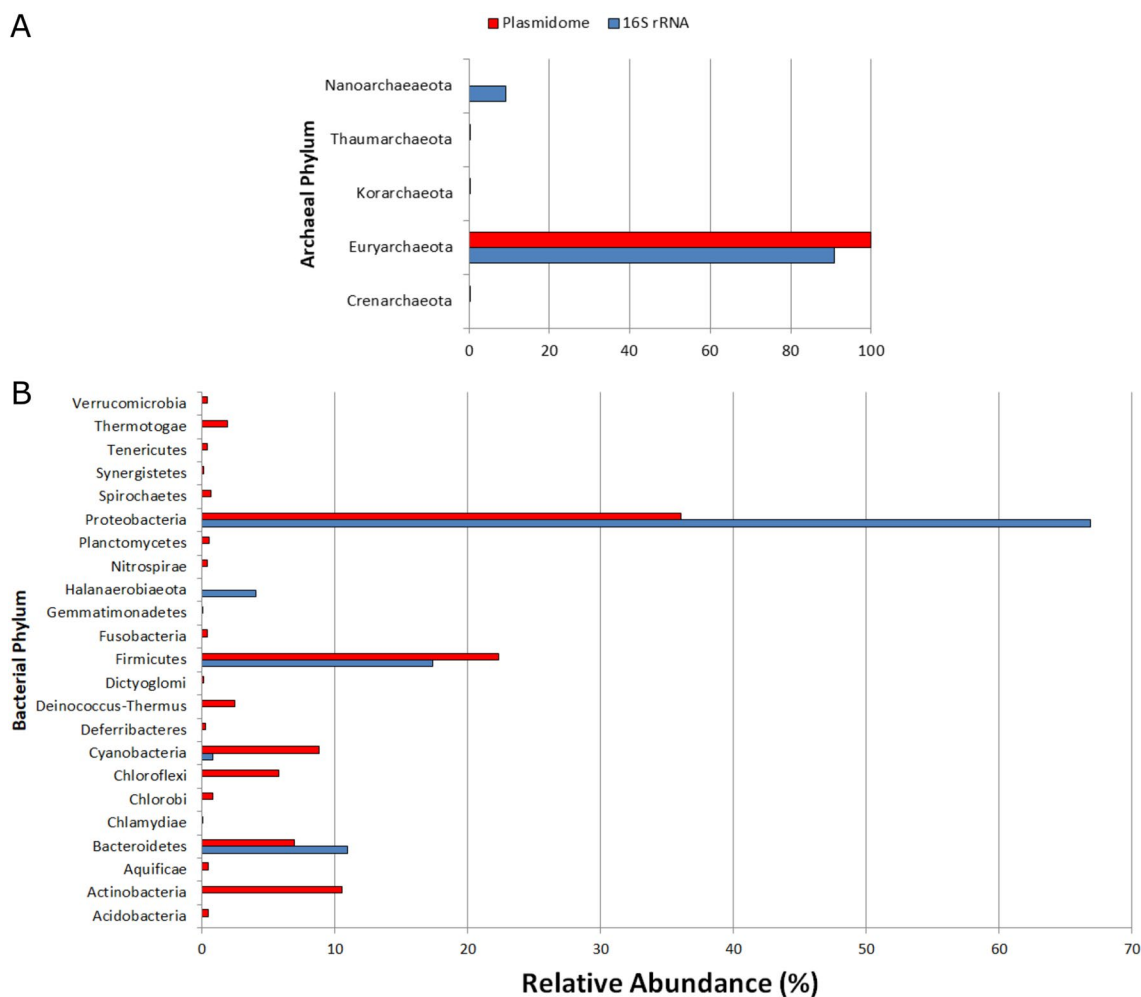
**Table 3.** Plasmid matches of the Diamante Lake plasmidome with entries derived from the NCBI database. \*NE not specified, \*\*NA not available.

The first are the five main IS families spreading in Halobacteria, which is the dominating class of the studied community<sup>77</sup>. Most of the archaeal IS fall into families detected in Bacteria, while others are restricted to Archaea such as members of the ISH3 family<sup>78</sup>. Two new potential IS not attributable to any of the known families were classified as well (GenBank accession OK172335 and OK172336). In the Puquio de Campo Naranja plasmidome, a much lower number of IS (28) was reported. Again, most of them were assigned to the IS5, IS630 and IS4 families<sup>18</sup>. The absence of Tn3 family transposases in both of the plasmidomes is conspicuous, as it represents one of the most abundant families in bacterial genomes, and Tn3 elements preferentially transpose into plasmids<sup>79</sup>.

The presence of so many IS elements is in line with the notion that the plasmidome substantially contributes to genome evolution as well as adaptation processes by facilitating the acquisition of novel genes and beneficial traits<sup>80,81</sup>.

**Taxonomic analysis.** Although plasmids can be transferred between different microorganisms, the taxonomic assignment of the plasmidome contigs allows an estimation of potential hosts. Eighty-eight percent were assigned to Archaea, while 11% were assigned to Bacteria and the remaining 1% to Eukarya. Among the Archaea, the phylum Euryarchaeota (99.85%) is dominating.

When the phylum distribution of the plasmidome was compared with 16S rRNA sequencing data from the corresponding metagenomic DNA sample, the phylum Euryarchaeota again stood out with the highest relative zOTU abundance (Fig. 7A). In both, the class of the halobacteria dominated. With respect to the Bacteria, the Proteobacteria (36%), the Firmicutes (22.3%) and the Actinobacteria (10.5%) comprised most of the contigs assigned in the plasmidome. Also, in the metagenome, the Proteobacteria (66.9%) and the Firmicutes (17.3%) account for the phyla with the highest relative zOTU abundance with the Bacteroidetes, however, ranking third (10.9%) (Fig. 7B). Notably, plasmid contigs of Actinobacteria, Chloroflexi and Deinococcus-Thermus were obtained, whereas the 16S rRNA analysis did not disclose members of these phyla. A possible explanation might be horizontal transfer of plasmid-borne genes between bacterial phyla, or the existence of plasmids with a wide host range. On the other hand, the 16S rRNA analysis indicates the presence of members of the phylum Halan-aerobiaeota but no plasmidome contig could be assigned to the latter, which is possibly due to a bias in plasmid databases when a phylum is not well represented or to the absence of plasmids within the taxon.



**Figure 7.** Red biofilm taxonomic analysis from Diamante Lake. Red bars show the relative abundance of each archaeal (A) or bacterial (B) phylum in the plasmidome by MG-RAST analysis using similarity to the RefSeq database ( $E$ -value  $\leq 10^{-5}$ ). Blue bars show the relative abundance of each archaeal (A) or bacterial (B) phylum by metagenomic DNA analysis using 16S rRNA gene amplicon sequencing. Phyla with abundance less than 1% in both datasets were not included.

## Conclusions

It is currently possible to study plasmid elements in the course of a conventional metagenomic analysis, but an approach to specifically target plasmid populations allows to overcome the inherent constraints of the bioinformatic tools applied for the analysis of plasmids from total community DNA. Under this perspective, and from the comparison with the metagenome of the same community, this study showed that part of the plasmid information will not be detected when the experimental plasmid-purification is not carried out prior to sequencing. Furthermore, a large fraction of genes with an unknown function was present in the plasmidome dataset, as at least 58.5% of the predicted proteins were hypothetical. In addition, the percentages of SEED assignments were even lower. The relatively few functional annotations may accord with the peculiarities of the extreme environment, which harbors a microbial community that is dominated by archaea. On the other hand, functions related to the response to oxidative stress and DNA repair were annotated, which agrees with the requirement of adaptive mechanisms enabling the hosts to withstand the exposure to the high UV irradiation in the Andean Puna.

Comparison of the Diamante Lake plasmidome to that of Puquio de Campo Naranja, revealed a certain degree of similarity between the predicted functional profiles of both AMEs. However, striking differences with respect to antibiotic and arsenic resistance were detected. Sequences pointing to arsenic resistance are more abundant in the Diamante Lake plasmidome, a fact that also accounts for the plasmidome derived from a wastewater treatment plant that contains large quantities of effluents of the chemical/pharmaceutical industry. Our results reflect the high amount of arsenic present in the environment under investigation. Traits expected to be found in a plasmid pool were detected, such as Pfam domains related to plasmid replication, MOB-type relaxase families related to plasmid mobilization and genes belonging type II toxin-antitoxin systems related to plasmid maintenance. Moreover, there are sequences known from megaplasmids of halophilic archaea isolated from different saline environments, which provides further evidence for known plasmid-associated genes in the obtained dataset.

The results presented here along with the detection of numerous IS elements favors the opinion that the plasmidome facilitates the mobility and the transfer of genes within such extreme microbial communities.

### Data availability

The sequence data of the Diamante Lake plasmidome and 16S rRNA gene amplicon have been deposited at NCBI (National Center for Biotechnology Information) under the accession numbers SRR13795604, SRR13795605 and SRR13795606.

Received: 21 May 2021; Accepted: 15 October 2021

Published online: 02 November 2021

### References

- Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Baptiste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci.* **107**, 127–132 (2010).
- Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**, 1598–1602 (2003).
- Greve, B. O., Jensen, S., Brügger, K., Zillig, W. & Garrett, R. A. Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea* **1**, 231–239 (2004).
- Van Kranenburg, R. *et al.* Functional analysis of three plasmids from *Lactobacillus plantarum*. *Appl. Environ. Microbiol.* **71**, 1223–1230 (2005).
- Sekine, M. *et al.* Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. *Environ. Microbiol.* **8**, 334–346 (2006).
- Kuenne, C. *et al.* Comparative analysis of plasmids in the genus *Listeria*. *PLoS One* **5**, e12511 (2010).
- Dib, J. R. *et al.* Complete genome sequence of pAP13, a large linear plasmid of a *Brevibacterium* strain isolated from a saline lake at 4,200 meters above sea level in Argentina. *Genome Announc.* **1**, e00878–e913 (2013).
- Dib, J. R. *et al.* First complete sequence of a giant linear plasmid from a *Micrococcus* strain isolated from an extremely high-altitude lake. *Genome Announc.* **1**, e00885–13 (2013).
- Dib, J. R. *et al.* Complete genome sequence of the linear plasmid pJD12 hosted by *Micrococcus* sp. D12, isolated from a high-altitude volcanic lake in Argentina. *Genome Announc.* **3**, e00627–15 (2015).
- Dib, J. R. *et al.* Complete genome sequences of pLMA1 and pLMA7, two large linear plasmids of *Micrococcus* strains isolated from a high-altitude lake in Argentina. *Genome Announc.* **6**, e00010–18 (2018).
- Salto, I. P. *et al.* Comparative genomic analysis of *Acinetobacter* spp. plasmids originating from clinical settings and environmental habitats. *Sci. Rep.* **8**, 1–12 (2018).
- Dib, J. R., Liebl, W., Wagenknecht, M., Fariás, M. E. & Meinhardt, F. Extrachromosomal genetic elements in *Micrococcus*. *Appl. Microbiol. Biotechnol.* **97**, 63–75 (2013).
- Zhang, T., Zhang, X. X. & Ye, L. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLoS One* **6**, 1–7 (2011).
- Kav, A. B. *et al.* Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci.* **109**, 5452–5457 (2012).
- Sentchilo, V. *et al.* Community-wide plasmid gene mobilization and selection. *ISME J.* **7**, 1173–1186 (2013).
- Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One* **9**, 1–9 (2014).
- Kothari, A. *et al.* Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. *MBio* **10**, e02899–e2918 (2019).
- Perez, M. F. *et al.* First report on the plasmidome from a high-altitude lake of the Andean Puna. *Front. Microbiol.* **11**, 1343 (2020).
- Dib, J. R., Wagenknecht, M., Fariás, M. E. & Meinhardt, F. Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements?. *Front. Microbiol.* **6**, 463 (2015).
- Fariás, M. *Microbial Ecosystems in Central Andes Extreme Environments: Biofilms, Microbial Mats, Microbialites and Endoevaporites* (Springer Nature, 2020).
- Vignale, F. A. *et al.* Lithifying and non-lithifying microbial ecosystems in the wetlands and salt flats of the Central Andes. *Microb. Ecol.* <https://doi.org/10.1007/s00248-021-01725-8> (2021).
- Saona Acuña, L. A., Soria, M. N., Villafaña, P. G., Stepanenko, T. & Fariás, M. E. Arsenic and its biological role: From early Earth to current Andean microbial ecosystems. In *Microbial Ecosystems in Central Andes Extreme Environments* (ed. Fariás, M.) 275–284 (Springer, 2020).
- Ordoñez, O. F., Rasuk, M. C., Soria, M. N., Contreras, M. & Fariás, M. E. Haloarchaea from the Andean Puna: Biological role in the energy metabolism of arsenic. *Microb. Ecol.* **76**, 695–705 (2018).
- Wagenknecht, M., Pérez, M. F. & Dib, J. R. Linear megaplasmids spreading in the Andean resistome. In *Microbial Ecosystems in Central Andes Extreme Environments* (ed. Fariás, M. E.) 261–273 (Springer, 2020).
- Rascovan, N., Maldonado, J., Vazquez, M. P. & Eugenia Fariás, M. Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea. *ISME J.* **10**, 299–309 (2016).
- Saona, L. A. *et al.* Analysis of co-regulated abundance of genes associated with arsenic and phosphate metabolism in Andean Microbial Ecosystems. *bioRxiv.* <https://doi.org/10.1101/870428> (2019).
- Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
- Gantner, S., Andersson, A. F., Alonso-Sáez, L. & Bertilsson, S. Novel primers for 16S rRNA-based archaeal community analyses in environmental samples. *J. Microbiol. Methods* **84**, 12–18 (2011).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
- Rozov, R. *et al.* Recycler: An algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**, 475–482 (2016).
- Luo, H. & Gao, F. DoriC 10.0: An updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. *Nucleic Acids Res.* **47**, D74–D77 (2019).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
- Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
- Meyer, F. *et al.* The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **9**, 386 (2008).
- Parks, D. H. & Beiko, R. G. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26**, 715–721 (2010).

37. Xie, Y. *et al.* TADB 2.0: An updated database of bacterial type II toxin–antitoxin loci. *Nucleic Acids Res.* **46**, D749–D753 (2018).
38. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* **38**, D57–D61 (2010).
39. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. & Larsson, D. G. J. BacMet: Antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* **42**, D737–D743 (2014).
40. Chen, L. *et al.* VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
41. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
42. Li, W., Jaroszewski, L. & Godzik, A. Sequence clustering strategies improve remote homology recognitions while reducing search times. *Protein Eng.* **15**, 643–649 (2002).
43. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
44. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).
45. Jia, B. *et al.* CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, gkw1004 (2017).
46. Xie, Z. & Tang, H. ISEScan: Automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340–3347 (2017).
47. Riadi, G., Medina-Moenne, C. & Holmes, D. S. TnpPred: A web service for the robust prediction of prokaryotic transposases. *Comp. Funct. Genomics* **2012**, 678761 (2012).
48. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Publ. Gr.* **7**, 335–336 (2010).
49. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd merger. *Bioinformatics* **30**, 614–620 (2014).
50. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
51. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
52. Yilmaz, P. *et al.* The SILVA and ‘all-species Living Tree Project (LTP)’ taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
53. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Towards functional characterization of archaeal genomic dark matter. *Biochem. Soc. Trans.* **47**, 389–398 (2019).
54. Albarracín, V. H. *et al.* Extremophilic *Acinetobacter* strains from high-altitude lakes in Argentinean Puna: Remarkable UV-B resistance and efficient DNA damage repair. *Orig. Life Evol. Biosph.* **42**, 201–221 (2012).
55. Fernández Zenoff, V., Siñeriz, F. & Fariás, M. E. Diverse responses to UV-B radiation and repair mechanisms of bacteria isolated from high-altitude aquatic environments. *Appl. Environ. Microbiol.* **72**, 7857–7863 (2006).
56. Flores, M. R., Ordoñez, O. F., Maldonado, M. J. & Fariás, M. E. Isolation of UV-B resistant bacteria from two high altitude Andean lakes (4,400 m) with saline and non saline conditions. *J. Gen. Appl. Microbiol.* **55**, 447–458 (2009).
57. Ordoñez, O. F., Flores, M. R., Dib, J. R., Paz, A. & Fariás, M. E. Extremophile culture collection from Andean lakes: Extreme pristine environments that host a wide diversity of microorganisms with tolerance to UV radiation. *Microb. Ecol.* **58**, 461–473 (2009).
58. Kurth, D. *et al.* Genomic and proteomic evidences unravel the UV-resistome of the poly-extremophile *Acinetobacter* sp. Ver3. *Front. Microbiol.* **6**, 328 (2015).
59. Sancho-Tomás, M. *et al.* Geochemical evidence for arsenic cycling in living microbialites of a high altitude Andean lake (Laguna Diamante, Argentina). *Chem. Geol.* **549**, 119681 (2020).
60. Li, L. L., Norman, A., Hansen, L. H. & Sørensen, S. J. Metamobilomics—Expanding our knowledge on the pool of plasmid encoded traits in natural environments using high-throughput sequencing. *Clin. Microbiol. Infect.* **18**, 5–7 (2012).
61. Bateman, A., Coggill, P. & Finn, R. D. DUFs: Families in search of function. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1148–1152 (2010).
62. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
63. Schlüter, A., Krause, L., Szczepanowski, R., Goesmann, A. & Pühler, A. Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. *J. Biotechnol.* **136**, 65–76 (2008).
64. Shi, Y., Zhang, H., Tian, Z., Yang, M. & Zhang, Y. Characteristics of ARG-carrying plasmidome in the cultivable microbial community from wastewater treatment system under high oxytetracycline concentration. *Appl. Microbiol. Biotechnol.* **102**, 1847–1858 (2018).
65. Garcillán-Barcia, M. P., Francia, M. V. & De La Cruz, F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687 (2009).
66. Garcillán-Barcia, M. P., Redondo-Salvo, S., Vielva, L. & de la Cruz, F. MOBscan: Automated annotation of MOB relaxases. *Methods Mol. Biol.* **2075**, 295–308 (2020).
67. Arcus, V. L., McKenzie, J. L., Robson, J. & Cook, G. M. The PIN-domain ribonucleases and the prokaryotic VapBC toxin–antitoxin array. *Protein Eng. Des. Sel.* **24**, 33–40 (2011).
68. Bolhuis, H., Te Poele, E. M. & Rodríguez-Valera, F. Isolation and cultivation of Walsby’s square archaeon. *Environ. Microbiol.* **6**, 1287–1291 (2004).
69. Page, R. & Peti, W. Toxin–antitoxin systems in bacterial growth arrest and persistence. *Nat. Chem. Biol.* **12**, 208–214 (2016).
70. Khelaifia, S. & Drancourt, M. Susceptibility of archaea to antimicrobial agents: Applications to clinical microbiology. *Clin. Microbiol. Infect.* **18**, 841–848 (2012).
71. Cavicchioli, R., Curmi, P. M. G., Saunders, N. & Thomas, T. Pathogenic archaea: Do they exist?. *BioEssays* **25**, 1119–1128 (2003).
72. Lepp, P. W. *et al.* Methanogenic Archaea and human periodontal disease. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6176–6181 (2004).
73. Andres, J. & Bertin, P. N. The microbial genomics of arsenic. *FEMS Microbiol. Rev.* **40**, 299–322 (2016).
74. Ben Fekih, I. *et al.* Distribution of arsenic resistance genes in prokaryotes. *Front. Microbiol.* **9**, 2473 (2018).
75. Wang, G., Kennedy, S. P., Fasiludeen, S., Rensing, C. & DasSarma, S. Arsenic resistance in *Halobacterium* sp. strain NRC-1 examined by using an improved gene knockout system. *J. Bacteriol.* **186**, 3187–3194 (2004).
76. Werner, J. *et al.* *Halorhabdus tiamatea*: Proteogenomics and glycosidase activity measurements identify the first cultivated euryarchaeon from a deep-sea anoxic brine lake as potential polysaccharide degrader. *Environ. Microbiol.* **16**, 2525–2537 (2014).
77. Filée, J., Siguier, P. & Chandler, M. Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.* **71**, 121–157 (2007).
78. Craig, N. *Mobile DNA III* (2020).
79. Szuplewska, M., Czarnecki, J. & Bartosik, D. Autonomous and non-autonomous Tn 3-family transposons and their role in the evolution of mobile genetic elements. *Mob. Genet. Elem.* **4**, 1–4 (2015).
80. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: The agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
81. Hülter, N. *et al.* An evolutionary perspective on plasmid lifestyle modes. *Curr. Opin. Microbiol.* **38**, 74–80 (2017).
82. Ng, W. V. *et al.* Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12176–12181 (2000).
83. Baliga, N. S. *et al.* Genome sequence of *Haloarcula marismortui*: A halophilic archaeon from the Dead Sea. *Genome Res.* **14**, 2221–2234 (2004).

84. Pfeiffer, F. *et al.* Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1. *Genomics* **91**, 335–346 (2008).
85. Saunders, E. *et al.* Complete genome sequence of *Haloterrigena turkmenica* type strain (4k T). *Stand. Genomic Sci.* **2**, 107–116 (2010).
86. Hartman, A. L. *et al.* The complete genome sequence of *Haloferax volcanii* DS2, a model Archaeon. *PLoS One* **5**, e9605 (2010).
87. Roh, S. W. *et al.* Complete genome sequence of *Halalkalicoccus jeotgali* B3T, an extremely halophilic archaeon. *J. Bacteriol.* **192**, 4528–4529 (2010).
88. Malfatti, S. *et al.* Complete genome sequence of *Halogeometricum borinquense* type strain (PR3T). *Stand. Genomic Sci.* **1**, 150–158 (2009).
89. Anderson, I. *et al.* Complete genome sequence of *Halopiger xanaduensis* type strain (SH-6 T). *Stand. Genomic Sci.* **6**, 31–42 (2012).
90. Ding, J. Y., Chiang, P. W., Hong, M. J., Dyll-Smith, M. & Tang, S. L. Complete genome sequence of the extremely halophilic archaeon *Haloarcula hispanica* strain N601. *Genome Announc.* **2**, 178–192 (2014).
91. Castillo, A. M. *et al.* *Halostagnicola larsenii* gen. nov., sp. Nov., an extremely halophilic archaeon from a saline lake in Inner Mongolia, China. *Int. J. Syst. Evol. Microbiol.* **56**, 1519–1524 (2006).

## Acknowledgements

This research was partially funded by PICT 2015-3825 and 2018-2545 PRESTAMO BID. We are grateful to Mr. Luis Ahumada for their assistance in the field trips, the native communities in Antofalla, El Peñón, Antofagasta de la Sierra, and Secretaría de Medio Ambiente from Catamarca Province for their support. We also want to thank Dr. Mariano Pistorio for valuable discussions and Melanie Heinemann for technical support. J.R.D. thanks the support by the Alexander von Humboldt Foundation.

## Author contributions

J.R.D. and M.F.P. planned the work and drafted the manuscript. M.F.P., J.R.D. and A.P. performed the experiments and analyzed the data. M.E.F. and L.A.S. contributed to the lake samples for the study. R.D., F.M. and J.R.D. contributed to the conception and edited the manuscript. All authors interpreted the results and reviewed and revised the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-00753-1>.

**Correspondence** and requests for materials should be addressed to R.D. or J.R.D.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021