Functional analysis of the archaea, bacteria, and viruses from a halite endolithic microbial community

Alexander Crits-Christoph[1], Diego R. Gelsinger[1], Bing Ma[2], Jacek Wierzchos[3], Jacques Ravel[2], Alfonso Davila[4], M. Cristina Casero[3], and Jocelyne DiRuggiero[1§]

[1]Biology Department, The Johns Hopkins University, Baltimore, MD, USA; [2]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA; [3]Museo Nacional de Ciencias Naturales, MNCN - CSIC, Madrid, Spain; [4]Carl Sagan Center, SETI Institute, Mountain View, CA, USA

Running Title: halite metagenome

Keywords: halite, endoliths, cyanobacteria, archaea, viruses, extreme environment, Atacama Desert, metagenomics

[§]Corresponding author:
Jocelyne DiRuggiero
Johns Hopkins University
Biology Department
3400 N. Charles Street, Mudd Hall
Baltimore MD 21218, USA
Phone: 410-516-8498
Fax: 410-516-5213
Email: jdiruggiero@jhu.edu

The authors declare that there are no competing financial interests in relation to the work described here.

Abstract

Halite endoliths in the Atacama Desert represent one of the most extreme microbial ecosystems on Earth. Here we sequenced and characterized a shotgun metagenome from halite nodules collected in Salar Grande, Chile. The community is dominated by archaea and functional analysis attributed most of the autotrophic CO_2 fixation to a unique cyanobacterium. The assembled 1.1 Mbp genome of a novel nanohaloarchaeon, *Candidatus* Nanopetramus SG9, revealed a photoheterotrophic life style and a low median isoelectric point (pI) for all predicted proteins, suggesting a "salt-in" strategy for osmotic balance. Predicted proteins of the algae identified in the community also had pI distributions similar to "salt-in" strategists. The Nanopetramus genome contained a unique CRISPR/Cas system with a spacer that matched a partial viral genome from the metagenome. A combination of reference-independent methods identified over 30 complete or near complete viral or proviral genomes with diverse genome structure, genome size, gene content, and hosts. Putative hosts included *Halobacteriaceae*, *Nanohaloarchaea*, and *Cyanobacteria*. Despite the dependence of the halite community on deliquescence for liquid water availability, this study exposed an ecosystem spanning three phylogenetic domains, containing a large diversity of viruses, and a predominant "salt-in" strategy to balance the high osmotic pressure of the environment.

Introduction

In the most arid deserts on Earth, microorganisms find refuge inside rock substrates as a survival strategy (Pointing and Belnap, 2012, Wierzchos *et al.,* 2012b). The rock environment provides physical stability, protection from incident UV and excessive shifts in temperature, and enhances moisture availability (Chan *et al.,* 2012, Walker and Pace, 2007). The colonized substrates are translucent, allowing primary production to occur via photosynthesis (Walker and Pace, 2007, Wierzchos *et al.,* 2012b). These endolithic communities are typically composed of cyanobacteria associated with diverse heterotrophic bacteria and/or archaea, and sometimes eukaryotes (Chan *et al.,* 2013, Robinson *et al.,* 2015, Wierzchos *et al.,* 2012b). The diversity of rock habitats colonized by microorganisms has shown that life has found innovative ways to adapt to the extreme conditions of hyper-arid deserts (Friedmann 1982, DiRuggiero *et al.*, 2013, Pointing *et al.*, 2009, Wei *et al.*, 2015, Wierzchos *et al.*, 2012b). This is in stark contrast to soil, where microorganisms under extreme water stress and restricted access to nutrient must undergo long periods of stasis (Crits-Christoph *et al.,* 2013).

The Atacama Desert in Northern Chile is one of the oldest and driest deserts on Earth (Clarke 2006). In the hyper-arid zone of the desert, with decades between rainfall events and extremely low air relative humidity (RH) (mean yr^{-1} values <35%), deliquescence of ancient halite crusts of evaporitic origin was shown to provide sufficient moisture to sustain microbial communities (Davila *et al.,* 2008, de los Rios *et al.,* 2010, Robinson *et al.,* 2015, Wierzchos *et al.,* 2012a). Within the halite nodules, capillary condensation of water vapor at air RH as low as 50-55%, due to the presence of pores smaller than 100 nm surrounding large NaCl crystals inside the nodules, was reported as a potential source

of water for microorganisms (Davila *et al.*, 2008, Wierzchos *et al.*, 2012a). Under these conditions, the halite nodule interior in the Yungay area of the hyper-arid core remained wet for 5,362 hours yr^{-1} (Wierzchos *et al.*, 2012a). In contrast, in Salar Grande, located in the southwest area of the Tarapacá Region, coastal fogs are frequent (Cereceda *et al.*, 2008a, Cereceda *et al.*, 2008b) leading to constant moisture inside the nodules (Robinson *et al.*, 2015).

High-throughput culture-independent methods based on 16S rRNA gene sequencing have shown that the Atacama halite communities were dominated by archaea from the *Halobacteriaceae* family. The communities also contained a unique cyanobacterium related to *Halothece* species and diverse heterotrophic bacteria (de los Rios *et al.*, 2010, Robinson *et al.*, 2015). Halite communities exposed to costal fogs were more diverse and harbored a novel type of algae that was not found in the Yungay nodules, suggesting that the environmental conditions in this habitat might be too extreme for eukaryotic photosynthetic life (Robinson *et al.*, 2015).

Assimilation of atmospheric radiocarbon into the halite microbial community biomass showed that carbon cycling inside the halite nodules was ongoing, with carbon turnover times of less then a decade in Salar Grande (Ziolkowski *et al.*, 2013). Measurements of chlorophyll fluorescence using Pulse Amplitude Modulated (PAM) fluorometry recently demonstrated *in situ* active metabolism in halite endolithic communities (Davila et al., 2015). Photosynthetic activity was tightly linked to moisture availability and solar insolation and was sustained for days after a wetting event (Davila *et al.*, 2015). Radiolabelled experiments showed that the halite communities fixed CO_2 via

93 photosynthesis and further evidence of metabolic activity was supported by oxygen
94 production and respiration (Davila *et al.,* 2015).
95 To further characterize halite endoliths, we sequenced the pooled metagenome of a
96 microbial community associated with halite nodules. We found novel microorganisms,
97 community members from the three domains of life, and a large diversity of viruses. The
98 functional annotation of the metagenome revealed communities highly specialized to the
99 extreme salinity of the environment.
100

101 ## Materials and methods

102 *Sampling, DNA extraction, and sequencing*

103 The colonization zone from 5 halite nodules collected in Salar Grande (Fig. 1) was
104 harvested using aseptic conditions in a laminar flow hood and pooled together for DNA
105 extraction, as previously described (Robinson *et al.,* 2015). Sequencing was performed
106 on the Illumina HiSeq2500 at the University of Maryland School of Medicine Institute
107 for Genome Sciences (Baltimore, MD).

108 *Whole Metagenome Analysis*

109 Sequencing produced 95,230,365 paired-end reads. After quality control (see details in
110 supplementary material), reads were assigned taxonomy content using PhyloSift (Darling
111 *et al.*, 2014) and the functional content was characterized using the analysis server MG-
112 RAST (Meyer *et al.,* 2008). The metagenome was assembled using the IDBA-UD
113 assembler (Peng *et al.,* 2012). We use a *k* range of 20-100 with a pre-correction step
114 before assembly, producing a meta-assembly with a mean contig size of 1,060 bp, a max

contig size of 377,822 bp, and a contig n50 of 1,558 bp. Assembled contigs were grouped into potential draft genomes using tetranucleotide frequencies, abundance levels, and single-copy gene analysis with MaxBin (Wu *et al.,* 2014). Genomic bins were assigned taxonomic ranks with PhyloSift and Kraken (Darling *et al.,* 2014, Wood and Salzberg 2014).

Algae genome

Genomic bin 18 contained a high number of eukaryotic marker genes (~⅓ of all marker genes in the bin), all of which were identified to belong to a member of the eukaryotic green algae using BLASTP against the non-redundant (nr) database at NCBI. Contigs assembly and annotation are further described in supplementary material.

Nanohaloarchaea genome

Four of the largest assembled contigs were binned together with MaxBin (Wu *et al.,* 2014) and identified as the partial genome of a member of the Nanohaloarchaea. Reassembly methods (see details in supplementary material) resulted in a single assembled genomic contig. The completed assembly was uploaded to and annotated using the RAST server (Aziz *et al.,* 2008). A CRISPR/Cas system was annotated using RAST and CRISPRFinder (Grissa *et al.,* 2007). BLASTN was used to match CRISPR spacers to the assembled metagenome content.

Phylogenetic positioning of the novel *Candidatus* Nanopetramus SG9 genome was performed by extracting 12 ribosomal marker genes shared by archaeal genomes with PhyloSift (Darling et al 2014), aligning with MUSCLE (Edgar, 2004), and building a Maximum-Likelihood tree with FastTree (Price *et al.,* 2010) using concatenated conserved blocks from each alignment (Gblocks) (Castresana 2000). The G+C content of

the CRISPR/Cas system was calculated and compared to the average G+C content for 10 kbp windows across the entire genome. The phylogenetic position of the CRISPR/Cas system was determined by aligning Cas1 proteins from key archaeal species using MUSCLE and by building a Maximum-Likelihood tree using FastTree.

Viral genomes

VirSorter (Roux *et al.,* 2015) was used to extract viral genomic content from the assembled metagenome and viral contigs greater than 12 kbp were annotated and examined. All contigs greater than 5 kbp were checked for evidence of circularity using a custom Python script, CircMG (https://github.com/alexcritschristoph/CircMG). The RAST annotation server failed to annotate the majority of proteins encoded on the viral contigs. ORFs were predicted for each putative genome with Prodigal (Hyatt *et al.,* 2010). To compare relationships within the halite viral community, all predicted viral proteins were compared against all others using BLASTP and an e-value cutoff of 0.001, producing a protein-protein similarity network. This network was used to build a virus-virus weighted undirected network, where the edge weight between two viruses was determined by the sum of the amino acid percent identity of all protein-protein matches between two viruses, divided by the total number of predicted genes in both viral genomes.

Community finding was run on the virus-virus network using the Walktrap community finding algorithm (Pons and Latapy. 2005). All network analyses were done in R using the igraph package (Csardi and Nepusz, 2006). The largest clusters in the protein-protein network (representing conserved proteins) were annotated using BLASTP to the nr protein database (Gish and States 1993) and HHPred (Söding *et al.,* 2005).

Sequence data and availability

All sequences were deposited at the National Center for Biotechnology Information Sequence Read Archive under Bioproject PRJNA296403 and accession number SRP064713. The MG-RAST report for the data is available under ID 4600831.3 (http://metagenomics.anl.gov/linkin.cgi?metagenome=4600831.3).

Completed assemblies, annotation, and phylogenetic trees are available at http://figshare.com/s/02565916783811e58e4b06ec4bbcf141

Results

Halite metagenome taxonomic and functional analyses

We characterized at the molecular level the endolithic microbial community from halite nodules from Salar Grande (Fig. 1). Due to the difficulty in harvesting enough DNA from the colonization zone of a single halite nodule, the metagenome was obtained with DNA extracted from 5 different nodules. The metagenome sequence of the halite community was composed of 9.6 Gb of high quality, paired-end, metagenomic shotgun sequences. Taxonomic assignments of the metagenomic reads performed with PhyloSift (Darling *et al.* 2014) revealed a community dominated by Archaea (71%) and also composed of Bacteria (27%) and Eukarya (1%) (Fig. 2). *Halobacteria* represented the majority of the Archaea (90%) with a small representation of *Nanohaloarchaea* (~2%). Most bacteria belonged to the *Salinibacter* genera (63%) and cyanobacteria constituted 15% of the bacteria. Reconstruction of 16S rRNA gene sequences from the metagenomic dataset with EMIRGE (Miller *et al.*, 2011) provided full-length genes for all the major taxonomic groups and was consistent with our previous work using 16S rRNA gene

sequencing (Robinson *et al.*, 2015) (Fig. S1).

The functional composition of the halite metagenome was analyzed with MG-RAST (Meyer *et al.*, 2008) using total sequence reads (Fig. S2). Of the genes involved in carbon metabolism, only 8% were allocated to autotrophic CO_2 fixation (Fig. S3a) and the Calvin-Benson cycle (CB) was the only pathway for autotrophic CO_2 fixation (Fig. 3b). The majority of the assigned RubisCO type I genes (>91%) were attributed to members of the *Cyanobacteria* (Fig. S4). We identified a small number of RubisCO type III genes and those were all from members of the *Halobacteriaceae*. Other key enzymes of the CB pathway, phosphoribulokinase and sedoheptulose-1,7-bisphosphatase, were also present and more than 99% of those sequences belonged to *Cyanobacteria*. Pathways for CO_2 concentration (carboxysomes) and phosphoglycolate detoxification (photorespiration) were also present in the halite metagenome and were from *Cyanobacteria* (Fig. S3b).

With respect to photosynthesis, most of the genes for Photosystem I and II (PSI and PSII) major proteins, and for light harvesting complexes, belonged to *Cyanobacteria* with only a small fraction assigned to green algae (Fig. S5 and S6). For all photosystems, 24 to 26% of all sequences were not given a taxonomic rank and 3% of PSII sequences were assigned to unclassified viruses (Fig. S5b). Phototrophy was also supported via light-driven proton pumps that belong to a number of heterotrophs including *Roseiflexus* species (proteorhodopsin), *Salinibacter* (xanthorhodopsin), and *Halobacteriaceae* (bacteriorhodopsin) (Fig. S7). Surprisingly, no nitrogenase (*nif*) genes were detected (Fig. S8).

Osmotic adaptation of the algae from the halite community

We found 89 complete or partially complete genes identified as belonging to a eukaryotic alga with an average coverage of 7.2x and a maximum contig length of 3.9 kbp. The majority of these genes mapped closely to homologs in either *Ostreococcus tauri* or *Micromonas* sp. RCC299. Known genes included heat shock protein 70, DNA mismatch repair protein MSH4, and multiple RNA splicing factors. Eukaryotic translation initiation factors, RNA polymerase subunits, and multiple enzymes and ribosomal proteins were also identified (Table S2). Using concatenated sequences of chloroplast and mitochondrial genes, we found that the alga clustered with other members of the *Chlorophyta* (Fig. S9), grouping consistently with the *Micromonas and Ostreococcus* species and thereby confirming our previous phylogenetic position (Robinson *et al.*, 2015).

To reveal potential adaptations to high salt, we compared the isoelectric point (pI) of the halite alga predicted proteins with that of the translated proteomes for *Micromonas* sp. RCC299, *O. tauri*, and the reported proteins for *Dunaliella salina*, all belonging to halophilic algae (Fig. 3) (Paul *et al.*, 2008). Proteins from the halite alga had a statistically significantly lower mean pI (6.19) than the known reference proteins from *Micromonas* sp. RCC299 (6.97), *O. tauri* (7.60), and *D. salina* (7.7) (Mann-Whitney t-test; $p<0.001$). Isoelectric point distributions were compared using nonparametric statistical tests. Using a paired one-sided Wilcoxon t-test, we found that the predicted halite alga proteins had significantly lower pI when paired using BLASTP with *O. tauri* homologues ($n=85$; $p<0.001$; difference of means: 0.55) and with *Micromonas* sp. RCC299 homologues ($n=87$;$p=0.025$; difference of means: 0.24). A paired-sample bayesian model comparison, implemented with BEST (Bååth, 2014, Kruschke, 2013),

229 reported that the probability that the proteins of the halite alga had a lower mean pI
230 (difference of means less than 0) than that of *Micromonas* sp. RCC299 was 97.9%. This
231 analysis predicted that the halite alga might have one of the lowest protein pI
232 distributions of any reported eukaryote.

233 The halite metagenome also contained a number of genes from algal organelles. A 29.8
234 kbp contig with 39% G+C carried several chloroplast genes with high similarity to the *O.*
235 *tauri* chloroplast genome, which is 71.7 kbp and 39.9% G+C. Genes encoding for PSI
236 and PSII protein subunits, cytochrome protein subunits, ribosomal proteins, and for
237 rRNAs and tRNAs were also found on the 29.8 kbp contig. In the same genomic bin, we
238 also found 8 non-overlapping contigs that contained mitochondrial genes from algae. The
239 combined non-overlapping contigs were 45.4 kbp in length with 37.1% G+C, similar to
240 the 44.2 kbp mitochondrial genome of *O. tauri* (38.2% G+C). These genomic fragments
241 were found to be enriched in tRNAs and organelle genes. Predicted proteins for the
242 organelles had mean pI values above 8, which may indicate different environmental
243 conditions in the organelle than in the intracellular space (Table S2).

244 *The complete genome sequence for a novel Nanohaloarchaea*

245 Our genome assembly produced a nanohaloarchaeon genome of 1.1 Mbp long, encoding
246 for 1,292 genes, and with a G+C content of 46.4% (referred to as SG9) (Fig. 4). Although
247 read abundances show the assembled contigs represented only ~1-3% of the population,
248 these contigs likely assembled well because of the small genome size (1.1 Mbp), low
249 levels of micro-diversity, and a genome coverage around 20. A phylogenetic analysis,
250 using a set of 12 conserved concatenated genes, showed that *Candidatus* Haloredivivus
251 (Ghai *et al.*, 2011) was the closest known reference (Fig. 5). The 16S rRNA gene

252 sequence of SG9 was 91% identical to that of *Candidatus* Haloredivivus, and 90 and
253 88 % identical to *Candidatus* Nanosalina and *Candidatus* Nanosalinarum, respectively,
254 and in agreement with our concatenated protein phylogeny. We have named this new
255 microorganism, SG9, as *Candidatus* Nanopetramus SG9 (petramus: rock).
256 The SG9 genome was highly reduced and mostly composed of protein encoding genes,
257 similar to previously reported genomes for *Nanohaloarchaea* (Narasingarao *et al.,* 2012)
258 (Fig. 4). RAST annotation of the genome revealed that 79% of predicted proteins could
259 not be assigned to known function. We predict that SG9 has a photoheterotrophic life
260 style as indicated by the presence of genes for rhodopsin biosynthesis, archaeal genes for
261 carbohydrate metabolism and a phosphoenolpyruvate-dependent sugar
262 phosphotransferase system (PTS). The glucose-6-phosphate dehydrogenase gene,
263 essential to the Pentose-Phosphate pathway and reported in both *Candidatus* Nanosalina
264 and *Candidatus* Nanosalinarum, was absent (Narasingarao *et al.,* 2012). The presence in
265 the SG9 genome of three potassium uptake systems, Trk, Ktr and HKT, which in bacteria
266 are key components of osmotic regulation and resistance to high salinity (Becker *et al.,*
267 2014), along with systems for K homeostasis, indicated a potential "salt-in" strategy for
268 survival under high salt. This was supported by a low median pI for all *Candidatus*
269 Nanopetramus SG9 predicted proteins (pI 4.7) and a pI distribution similar to that of
270 *Haloarcula hispanica,* a salt-in strategist (Fig. S10). Potential motility was indicated by
271 the presence of genes for archaeal flagellar proteins. Genes for bacterial-like and
272 archaeal-like nucleotide excision repair pathways were also encoded in the SG9 genome,
273 together with a photolyase gene, and several homologs for the *radA* recombinase gene.

We also found genes for isoprenoid biosynthesis, a S-layer protein, and for DNA polymerases PolI and PolII, all genes typically found in archaea.

A Type I CRISPR/Cas system composed of eight CRISPR-associated proteins and a spacer/repeat region with 22 spacers was found in the SG9 genome (Fig. 6a). We found no evidence of a CRISPR system, or individual Cas proteins, in all publicly available nanohaloarchaea genomes using CRISPR-finder (Grissa *et al.,* 2007) and by searching for the Cas1 protein with a Cas1-HMM alignment using Hmmer3 (Wheeler and Eddy 2013). The 11 kbp CRISPR region of SG9 had a significantly lower G+C content of 41.5% than the whole genome (Fig. 6c). The difference in G+C%, along with the absence of any CRISPR loci in the genomes of its nearest neighbors, would indicate that SG9 acquired its CRISPR system via horizontal gene transfer (HGT) (Ochman *et al.,* 2005). To further elucidate the origins of the CRISPR/Cas system in SG9, the *cas*1 gene product was aligned with Cas1 proteins from diverse archaeal genomes (Fig. 6b). The resulting phylogeny showed that the Cas1 protein was rooted within the *Euryarchaeota*, making it likely that the locus was acquired from *Methanobacteria* or *Halobacteria* rather than the phylogenetically closer *Nanoarchaeota*. While *Halobacteria* genomes are characterized by high G+C%, the genomes of *Methanobacteria*, *Nanoarchaeota*, and that of the putative nanohaloarchaeal viruses (see below) all have low G+C% similar to the CRISPR locus of SG9.

A BLASTN analysis of all 22 spacers from the SG9 CRISPR array against the assembled halite metagenome returned a single hit for spacer 22 to contig Ha1987, which was found to be a partial viral genome. Spacer 22 was located adjacent to the Cas genes cluster indicating that it was the most recently added spacer to the CRISPR array.

297 *Halite community viral diversity*

298 Analysis of contigs reported by VirSorter (Roux *et al.,* 2015) and of circular viral contigs
299 (CircMG) resulted in the identification of over 30 complete or near complete viral or
300 proviral genomes in the halite metagenome (Tables 1 and S1). These viruses are novel
301 with a majority of viral protein products that were either hypothetical or had no homologs
302 in the nr database. Complete genome sizes ranged from 12 kbp to 70 kbp (Table 1).
303 While members of the *Halobacteriales* were the predicted hosts for a majority of the
304 viruses, the *Nanohaloarchaea* were predicted as putative hosts for viral genomes Ha139
305 (G+C%, host genes) and Ha1987 (G+C%, spacer match), and the cyanobacterium
306 *Halothece* for viral genomes Ha238 (host genes, G+C%) and Ha322 (host genes, G+C%),
307 based on the evidence indicated.
308 A majority of the annotated viral genomes shared few to no genes with other genomes,
309 making a phylogenetic or tree-based analysis impractical. To elucidate relationships
310 among viruses and the structure of the viral diversity in the halite community, we used a
311 network-based approach in which genomes were linked in a weighted network by the
312 proportion and percent identity of predicted proteins they shared with every other genome
313 (Fig. 7). The Walktrap community finding algorithm was used to then classify viral
314 genomes into one of six communities (I-VI), and representative members of each cluster
315 were further annotated and curated (Table 1)
316 The protein similarity network used to build the network of viral relationships was
317 examined for large clusters, which represented highly conserved proteins in the viral
318 population. The eight largest clusters were extracted and annotated using both HHPred
319 and BLASTP (Gish and States 1993, Söding *et al.,* 2005) (Figure 8). This analysis

confirmed that a majority of the viruses, particularly those in community I, had a head-tail structure. Distinct BLASTP hits to putative host transcriptional regulators were found throughout all the viral genomes and they often contained an HNH endonuclease domain. DNA polymerases were common in the larger genomes and DNA helicases were abundant in community III genomes.

Among haloviruses, the largest linear genomes, Ha32 and Ha68 (community II), were structurally similar to each other and shared several genes with HSTV-2, an icosahedrally symmetric Myovirus with host *Halorubrum* (Table 1). Both contained genes for tail assembly proteins, baseplate and portal proteins, and a prohead protease, all characteristics of large head-tail viruses. The circular Ha38 (community I) genome shared few proteins with Ha32/Ha68, and most of the predicted protein products had no homologs to haloviruses in the nr protein database. Ha38 shared several core phage genes with Myovirus HGTV-1, including a major capsid protein and a prohead protease (Table 1). The genome of Ha38 also encoded DNA repair proteins (Rad25 and Hef nuclease) and a transcription initiation factor TFIIB similar to archaeal hosts. Other large, circular community I viruses (Ha86, Ha92) were not clearly related to any known halovirus and shared a small number of proteins with HGTV-1, HSTV-1, and HHTV-1, while the majority of encoded proteins had no known homologs (Table 1). Ha92 encoded for several haloarchaeal and bacterial transcriptional regulators. The Ha86 genome encoded for an ArsR family transcriptional regulator of the putative archaeal host and a *cas*6 gene from a haloarchaea. Integrase genes were not found in Ha32 and Ha68, but were found in Ha216, Ha68, and Ha1987.

342 Genome Ha687 and the community III-associated genome Ha929 shared several genes
343 for the structural proteins VP2 and VP4 with published *Halorubrum* pleomorphic viruses,
344 including HRPV-1, a ssDNA halovirus (Table 1) (Pietila *et al.,* 2010). Both Ha966 and
345 Ha934 (community IV) shared a majority of their proteins with each other and few with
346 other identified viruses from the literature or this analysis. The two genomes were both
347 circular and approximately 12 kbp in length. They had similar structure and gene
348 composition but surprisingly their G+C content differed by 6.6% (Table 1). Annotated
349 proteins in their genomes included the plasmid partition protein ParB, Zn-finger domain
350 proteins, and DNA and RNA polymerase subunits.
351 Genomes Ha322, Ha238, Ha1987, and Ha139 uniquely had G+C content below 48%,
352 while haloviruses typically have GC% above 50% (Table 1) (Klein *et al.,* 2002, Oren,
353 2006, Pagaling *et al.,* 2007, Pietila *et al.,* 2013b, Pietila *et al.,* 2013c, Sencilo *et al.,* 2013,
354 Tang *et al.,* 2002, Tang *et al.,* 2004). Spacer 22 from the CRISPR-Cas array above had an
355 exact BLAST match to the linear and partially complete genome Ha1987, linking this
356 virus to a Nanohaloarchaeal host. In addition, several of Ha1987 viral genes had close
357 hypothetical homologs to the published *Candidatus* Haloredivivus sp. G17 genome (Ghai
358 *et al.,* 2011). Ha139 was also putatively assigned to a Nanohaloarchaeal host because of
359 its low G+C%, a shared hypothetical gene product with *Candidatus* Nanosalina, and
360 because large number of gene products shared with environmental Halophage eHP-23
361 and eHP-35 and two viral genomes previously assigned a Nanohaloarchaeal host (Garcia-
362 Heredia *et al.,* 2012).
363 Both linear and incomplete genomes Ha322 and Ha238 had their closest BLASTP hits to
364 multiple *Synechococcus* phage and cyanophage proteins, and had homologs for multiple

365 genes in several cyanobacteria. These findings suggest that they might be novel
366 cyanophages targeting the *Halothece* members of the community. Despite multiple close
367 BLASTP hits to cyanophage proteins in public datasets, Ha322 and Ha238 contigs shared
368 no homologous proteins with each other and did not cluster in the network analysis.
369

370 ## Discussion

371 Halite nodules in fossil continental evaporites from the Atacama Desert are at the
372 extreme of salt concentrations for hypersaline environments. Microorganisms inhabiting
373 this environment must balance the osmotic pressure of their cytoplasm with that of the
374 outside milieu. One osmotic strategy is the "salt in" strategy where ions, mainly K^+ and
375 Cl^-, are accumulated in the cell and the entire intracellular enzymatic machinery is
376 adapted to high salt (Oren, 2008). To remain soluble, the proteins of "salt-in" strategists
377 have an increased number of acidic amino acid residues on their surface, resulting in a
378 proteome with a low pI. This strategy is used by halophilic archaea and one extremely
379 halophilic bacterium, *Salinibacter*, a member of the *Bacteroidetes* (Oren, 2008). Other
380 halophilic and halotolerant microorganisms balance the high osmotic pressure of their
381 environment by synthesizing compatible solutes, a strategy called "salt-out" (Galinski,
382 1995, Oren, 2008).

383 *Adaptation of the halite community to high salt*

384 The phylogenetic composition of the halite community reflected the extreme salinity of
385 its environment with Archaea greatly outnumbering Bacteria (Ghai *et al.,* 2011, Podell *et*
386 *al.,* 2013, Robinson *et al.,* 2015). In addition, most of the bacteria in the community
387 belonged to the genus *Salinibacter*, a "salt-in" strategist (Oren 2008). We found only one

388 cyanobacteria, *Halothece,* as previously described using high-throughput 16S rRNA gene
389 sequencing (Robinson *et al.,* 2015). Many species of cyanobacteria are adapted to high
390 salt and they often form dense benthic mats in saline and hypersaline environments,
391 where they are the main primary producers (Oren, 2015). However, above 25% NaCl
392 only cyanobacteria of the *Aphanothecee-Halothece-Euhalothece* cluster have been found
393 so far (de los Rios *et al.,* 2010, Garcia-Pichel *et al.,* 1998, Robinson *et al.,* 2015,
394 Wierzchos *et al.,* 2006). A property of this cluster is the production of abundant
395 extracellular polysaccharides (EPS) (de los Rios *et al.,* 2010, Oren, 2015). We previously
396 reported that, in the halite nodules, *Halothece* formed cell aggregates surrounded by a
397 thick sheath embedded in EPS (de los Rios *et al.,* 2010, Robinson *et al.,* 2015). It is likely
398 that these structural components play a significant role in the desiccation tolerance of
399 *Halothece* and its ability for photosynthetic O_2 evolution (Tamaru *et al.,* 2005).
400 We found evidence of autotrophic CO_2 fixation via the CB pathway by cyanobacteria in
401 the halite metagenome. Although we also found RubisCO type III gene in several archaea,
402 it is not clear whether this enzyme participate in autotrophic CO_2 fixation or in a novel
403 AMP recycling pathway (Falb *et al.,* 2008, Sato *et al.,* 2007). Our findings indicate that
404 the unique cyanobacteria is likely responsible for most of the CO_2 fixed in the halite
405 community. In addition, we recently reported *in situ* carbon fixation through oxygenic
406 photosynthesis in halite nodules supporting the idea that cyanobacteria are the major
407 primary producers in this ecosystem (Davila *et al.,* 2015). A number of organisms
408 encoded light-driven proton pumps in their genomes, carrying out photoheterotrophy and
409 significantly increasing the energy budget from light.

We assembled the partial genes of the alga previously detected in the halite community, together with large regions of the genomes of its mitochondria and chloroplast. The pI of the alga predicted proteins was one of the lowest pI reported for any eukaryote (Kiraga *et al.,* 2007). In addition, the bimodal distribution of the proteins pI was similar to that of "salt-in" strategists, suggesting that lower eukaryotes might potentially use intracellular salt as a mean to balance osmotic pressure in hypersaline environments.c

A novel nanohaloarchaeal genome with a CRISPR array

We report here the first nanohaloarchaeon genome assembled into a single scaffolded contig from metagenome data of hypersaline environments (Ghai *et al.,* 2011, Martinez-Garcia *et al.,* 2014, Narasingarao *et al.,* 2012, Podell *et al.,* 2013). The *Candidatus* Nanopetramu SG9 genome of 1.1 Mb is very similar in size to that of previously reported Nanohaloarchaea and its genomic G+C content is intermediate in the reported range (Ghai *et al.,* 2011, Narasingarao *et al.,* 2012, Podell *et al.,* 2013). Evidence from its genome support the idea that *Candidatus* Nanopetramus SG9 has a photoheterotrophic life-style and that it uses the "salt-in" strategy to counterbalance the high salt of its environment. A low proteome pI has also been reported for *Candidatus* Nanoredivirus, suggesting that the "salt-in" strategy might be a ubiquitous feature of Nanohaloarchaea (Ghai *et al.,* 2011). A unique attribute of *Candidatus* Nanopetramus SG9 was the presence of a CRISPR array on its genome, demonstrating that adaptive immunity against viruses is also a feature of Nanohaloarchaeal genomes. This is the first report of annotated CRISPR-associated features in a nanohaloarchaeal genome and documented acquisitions of CRISPR/Cas systems via HGT in the Archaea (Godde and Bickerton,

2006, Portillo and Gonzalez, 2009, Brodt *et al.,* 2011) support our hypothesis that SG9 acquired its CRISPR system via HGT.

The viral component of the halite community

With up to 10^{10} virus-like particles per ml, hypersaline systems harbor some of the highest viral concentrations of any aquatic environments (Baxter *et al.,* 2011, Boujelben *et al.,* 2012). In these extreme environments, with very few eukaryotes, haloviruses are likely to play an important role in shaping community structure through predation. We have assembled over 30 complete to near complete viral genomes from a metagenome obtained from the cellular fraction of the halite samples, restricting access to viruses that were either contained inside cells at the time of sampling or associated with particle surfaces. Despite this limitation, we found great viral diversity in the halite community in terms of genome structure, genome size, G+C%, and gene composition. These viruses were novel with a majority of the viral protein products having no characterized homologs. The viral genomes were not integrated within their host genomes in the metagenome assembly, suggesting that most of the viruses were lytic rather than lysogenic, in contrast to viruses found in high temperature environments (Anderson *et al.,* 2015).

Viruses infecting haloarchaea come in a variety of virion morphotypes, including spindle-shaped, pleomorphic, icosahedral, and head-and-tail (Atanasova *et al.,* 2012, Pietila *et al.,* 2013a, Roine and Oksanen, 2011). To date, 43 haloarchaeal tailed viruses have been reported (Atanasova *et al.,* 2012, Kukkaro and Bamford, 2009, Sabet, 2012) and 17 completely sequenced genomes, comprised approximately 1.2 Mb of sequence information (Klein *et al.,* 2002, Pagaling *et al.,* 2007, Pietila *et al.,* 2013b, Pietila *et al.,*

2013c, Sencilo *et al.,* 2013, Tang *et al.,* 2002, Tang *et al.,* 2004). Our analysis confirmed that the majority of the viruses we identified in the halite metagenome also had a head-tail structure, as previously found in other hypersaline environments (Garcia-Heredia *et al.,* 2012). The viral genetic diversity uncovered here hints that a significant portion of the diversity of both head-tail and other viruses in halophilic environments still remains largely unexplored.

Our network-based approach allowed us to analyze viral relationships in the community and identify core protein encoding genes with similarity to previously described haloviruses and cyanophages (Klein *et al.,* 2002, Oren, 2006, Pagaling *et al.,* 2007, Pietila *et al.,* 2013b, Pietila *et al.,* 2013c, Sencilo *et al.,* 2013, Tang *et al.,* 2002, Tang *et al.,* 2004). Haloviruses genomes have rather high G+C content (above 50% on average), also characteristic of haloarchaea, and the halite viruses followed that trend. The exceptions were putative cyanophages and Nanohaloarchaea viruses that exhibited a lower G+C content, consistent with previous findings (Emerson *et al.,* 2012, Martinez-Garcia *et al.,* 2014, Podell *et al.,* 2013).

A number of partial halovirus genomes have been obtained from metagenomes from crystallizer ponds and the hypersaline Lake Tyrrell in Australia with potential hosts including *Haloquadratum walsbyi*, *Nanohaloarchaea*, and the bacterium *Salinibacter* (Emerson *et al.,* 2012, Garcia-Heredia *et al.,* 2012). Here we described new groups of viruses that prey on members of the *Halobacteriales,* the cyanobacterium *Halothece,* and on a newly described Nanohaloarchaeon, *Candidatus* Nanopetramus SG9. CRISPR sequences found in the newly described SG9 genome where also present on a partial viral

477 genome, providing a direct connection between virus and host. However, these
478 predictions remain to be tested (Martinez-Garcia *et al.*, 2014).
479 Despite the extreme stress of the halite environment, this work demonstrates that this
480 endolithic community is exquisitely adapted to the challenges of its environment.
481 Metagenomic analysis revealed a relatively complex community with members from the
482 three domains of life. It also revealed trophic levels, from photosynthetic primary
483 producers to heterotrophs, viral predations, and specific physiological adaptations to the
484 high osmotic pressure of the halite milieu. Further understanding of the major taxonomic
485 groups and essential metabolic pathways underlying the functioning of this unique
486 community will require a combination of field-based measurements of metabolic activity
487 coupled with meta-transcriptomics, to capture gene expression levels for specific function
488 and taxonomic groups.
489

490 Acknowledgements

491 This work was funded by grant EXOB08-0033 from NASA and grant NSF-0918907
492 from the National Science foundation to JDR, by grant CGL2013-42509P from MINECO
493 (Spain) to JW and by grant NNX12AD61G from NASA to AD. We thank Octavio
494 Artieda for support in the field and Carmen Ascaso for valuable discussions.
495
496

497 Conflicts of Interests

The authors declare that there are no competing financial interests in relation to the work described here.

References

Anderson RE, Sogin ML, Baross JA (2015). Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. *FEMS Microbiol Ecol* **91:** 1-11.

Atanasova NS, Roine E, Oren A, Bamford DH, Oksanen HM (2012). Global network of specific virus-host interactions in hypersaline environments. *Environ Microbiol* **14:** 426-440.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75.

Bååth R (2014). Bayesian First Aid: A Package that Implements Bayesian Alternatives to the Classical *. test Functions in R. *UseR 2014*; Los Angeles, USA.

Baxter BK, Mangalea MR, Willcox S, Sabet S, Nagoulat MN, Griffith JD (2011). Haloviruses of Great Salt Lake: A Model for Understanding Viral Diversity In: Ventosa A, Oren A, Ma Y (eds). *Halophiles and Hypersaline Environments*. Springer-Verlag: Berlin Heidelberg. pp 173-190.

Becker EA, Seitzer PM, Tritt A, Larsen D, Krusor M, Yao AI *et al.* (2014). Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet* **10:** e1004784.

519 Boujelben I, Yarza P, Almansa C, Villamor J, Maalej S, Anton J *et al.* (2012).
520 Virioplankton community structure in Tunisian solar salterns. *Appl Environ*
521 *Microbiol* **78:** 7429-7437.
522 Brodt A, Lurie-Weinberger MN, Gophna U (2011). CRISPR loci reveal networks of gene
523 exchange in archaea. *Biol Direct* **6:** 65.
524 Castresana J (2000). Selection of conserved blocks from multiple alignments for their use
525 in phylogenetic analysis. *Mol Biol Evol* **17:** 540-552.
526 Cereceda P, Larrain H, Osses P, Farías M, Egaña I (2008a). The spatial and temporal
527 variability of fog and its relation to fog oases in the Atacama Desert, Chile.
528 *Atmospheric Res* **87:** 312-323.
529 Cereceda P, Larrain H, Osses P, Farías M, Egaña I (2008b). The climate of the coast and
530 fog zone in the Tarapacá Region, Atacama Desert, Chile. *Atmospheric Res* **87:** 301-
531 311.
532 Chan Y, Lacap DC, Lau MC, Ha KY, Warren-Rhodes KA, Cockell CS *et al.* (2012).
533 Hypolithic microbial communities: between a rock and a hard place. *Environ*
534 *Microbiol* **14:** 2272-2282.
535 Chan Y, Van Nostrand JD, Zhou J, Pointing SB, Farrell RL (2013). Functional ecology
536 of an Antarctic Dry Valley. *Proc Natl Acad Sci U S A* **110:** 8990-8995.
537 Clarke JDA (2006). Antiquity of aridity in the Chilean Atacama Desert. *Geomorphology*
538 **73:** 101-114.
539 Crabtree J, Agrawal S, Mahurkar A, Myers GS, Rasko DA, White O (2014). Circleator:
540 flexible circular visualization of genome-associated data with BioPerl and SVG.
541 *Bioinformatics* **30:** 3125-3127.

542 Crits-Christoph A, Robinson CK, Barnum T, Fricke WF, Davila AF, Jedynak B *et al.*
543 (2013). Colonization patterns of soil microbial communities in the Atacama Desert.
544 *Microbiome* **1:** 28.
545 Csardi G, Nepusz T (2006). The igraph software package for complex network research.
546 *Inter J Complex Systems* **1695** http://igraph.org.
547 Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM, Eisen JA (2014). PhyloSift:
548 phylogenetic analysis of genomes and metagenomes. *Peer J* **2:** e243.
549 Davila AF, Gomez-Silva B, de los Rios A, Ascaso C, Olivares H, McKay CP *et al.*
550 (2008). Facilitation of endolithic microbial survival in the hyperarid core of the
551 Atacama Desert by mineral deliquescence. *J Geophys Res* **113:** G01028,
552 doi:01010.01029/02007JG000561.
553 Davila AF, Hawes I, Garcia J, Gelsinger DR, DiRuggiero J, Ascaso C *et al.* (2015). In
554 situ metabolism in halite endolithic microbial communities of the hyperarid Atacama
555 Desert. *Frontiers Microbiol* http://dx.doi.org/10.3389/fmicb.2015.01035.
556 de los Rios A, Valea S, Ascaso C, Davila AF, Kastovsky J, McKay CP *et al.* (2010).
557 Comparative analysis of the microbial communities inhabiting halite evaporites of
558 the Atacama Desert. *Int Microbiol* **2:** 79-89.
559 DiRuggiero J, Wierzchos J, Robinson CK, Souterre T, Ravel R, Artieda O *et al.* (2013).
560 Microbial Colonization of Chasmoendolithic Habitats in the Hyper-arid Zone of the
561 Atacama Desert. *Biogeosciences* **10:** 2439-2450.
562 Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high
563 throughput. *Nucleic Acids Res* **32:** 1792-1797

564 Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF (2012).
565 Dynamic viral populations in hypersaline systems as revealed by metagenomic
566 assembly. *Appl Environ Microbiol* **78:** 6309-6320.
567 Falb M, Muller K, Konigsmaier L, Oberwinkler T, Horn P, von Gronau S *et al.* (2008).
568 Metabolism of halophilic archaea. *Extremophiles* **12:** 177-196.
569 Friedmann EI (1982). Endolithic Microorganisms in the Antarctic Cold Desert. *Science*
570 **215:** 1045-1053.
571 Galinski EA (1995). Osmoadaptation in bacteria. *Adv Microb Physiol* **37:** 272-328.
572 Garcia-Heredia I, Martin-Cuadrado AB, Mojica FJ, Santos F, Mira A, Anton J *et al.*
573 (2012). Reconstructing viral genomes from the environment using fosmid clones: the
574 case of haloviruses. *PLoS One* **7:** e33802.
575 Garcia-Pichel F, Nubel U, Muyzer G (1998). The phylogeny of unicellular, extremely
576 halotolerant cyanobacteria. *Arch Microbiol* **169:** 469-482.
577 Ghai R, Pasic L, Fernandez AB, Martin-Cuadrado AB, Mizuno CM, McMahon KD *et al.*
578 (2011). New abundant microbial groups in aquatic hypersaline environments. *Sci*
579 *Rep* **1:** 135.
580 Gish W, States DJ (1993). Identification of protein coding regions by database similarity
581 search. *Nature Genet* **3:** 266-272.
582 Godde JS, Bickerton A (2006). The repetitive DNA elements called CRISPRs and their
583 associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62:**
584 718-729.

585 Grissa I, Vergnaud G, Pourcel C (2007). The CRISPRdb database and tools to display
586 CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8:**
587 172.
588 Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010). Prodigal:
589 prokaryotic gene recognition and translation initiation site identification. *BMC*
590 *Bioinformatics* **11:** 119.
591 Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N *et al.* (2007).
592 The relationships between the isoelectric point and: length of proteins, taxonomy and
593 ecology of organisms. *BMC Genomics* **8:** 163.
594 Klein R, Baranyi U, Rossler N, Greineder B, Scholz H, Witte A (2002). Natrialba
595 magadii virus phiCh1: first complete nucleotide sequence and functional
596 organization of a virus infecting a haloalkaliphilic archaeon. *Mol Microbiol* **45:** 851-
597 863.
598 Kruschke JK (2013). Bayesian estimation supersedes the t test. *J Exp Psych Gen* **142:**
599 573-603.
600 Kukkaro P, Bamford DH (2009). Virus-host interactions in environments with a wide
601 range of ionic strengths. *Environ Microbiol Rep* **1:** 71-77.
602 Martinez-Garcia M, Santos F, Moreno-Paz M, Parro V, Anton J (2014). Unveiling viral-
603 host interactions within the 'microbial dark matter'. *Nat Commun* **5:** 4542.
604 Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The
605 metagenomics RAST server - a public resource for the automatic phylogenetic and
606 functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386.

607 Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ *et al.*
608 (2012). De novo metagenomic assembly reveals abundant novel major lineage of
609 Archaea in hypersaline microbial communities. *ISME J* **6:** 81-93.
610 Ochman H, Lerat E, Daubin V (2005). Examining bacterial species under the specter of
611 gene transfer and exchange. *Proc Natl Acad Sci U S A* **102:** 6595-6599.
612 Oren A (2006). The order halobacteriales. In: Dworkin M, Falkow S, Rosenberg E,
613 Schleife RK-H, Stackebrandt E (eds). *The Prokaryotes*. Springer;: Singapore. pp
614 113-164.
615 Oren A (2008). Microbial life at high salt concentrations: phylogenetic and metabolic
616 diversity. *Saline Systems* **4:** doi:10.1186/1746-1448-1184-1182.
617 Oren A (2015). Cyanobacteria in hypersaline environments: biodiversity and
618 physiological properties. *Biodivers Conserv* **24:** 781–798.
619 Pagaling E, Haigh RD, Grant WD, Cowan DA, Jones BE, Ma Y *et al.* (2007). Sequence
620 analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia,
621 China. *BMC Genomics* **8:** 410.
622 Pandit AS, Joshi MN, Bhargava P, Shaikh I, Ayachit GN, Raj SR *et al.* (2015). A
623 snapshot of microbial communities from the Kutch: one of the largest salt deserts in
624 the World. *Extremophiles* **19:** 973-987.
625 Paul S, Bag SK, Das S, Harvill ET, Dutta C (2008). Molecular signature of hypersaline
626 adaptation: insights from genome and proteome composition of halophilic
627 prokaryotes. *Genome Biol* **9:** R70.

628 Peng Y, Leung HC, Yiu SM, Chin FY (2012). IDBA-UD: a de novo assembler for
629 single-cell and metagenomic sequencing data with highly uneven depth.
630 *Bioinformatics* **28:** 1420-1428.
631 Pietila MK, Laurinavicius S, Sund J, Roine E, Bamford DH (2010). The single-stranded
632 DNA genome of novel archaeal virus halorubrum pleomorphic virus 1 is enclosed in
633 the envelope decorated with glycoprotein spikes. *J Virol* **84:** 788-798.
634 Pietila MK, Atanasova NS, Oksanen HM, Bamford DH (2013a). Modified coat protein
635 forms the flexible spindle-shaped virion of haloarchaeal virus His1. *Environ*
636 *Microbiol* **15:** 1674-1686.
637 Pietila MK, Laurinmaki P, Russell DA, Ko CC, Jacobs-Sera D, Butcher SJ *et al.* (2013b).
638 Insights into head-tailed viruses infecting extremely halophilic archaea. *J Virol* **87:**
639 3248-3260.
640 Pietila MK, Laurinmaki P, Russell DA, Ko CC, Jacobs-Sera D, Hendrix RW *et al.*
641 (2013c). Structure of the archaeal head-tailed virus HSTV-1 completes the HK97
642 fold story. *Proc Natl Acad Sci U S A* **110:** 10604-10609.
643 Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg KB, Allen EE (2013).
644 Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLoS*
645 *One* **8:** e61692.
646 Pointing SB, Chan Y, Lacap DC, Lau MC, Jurgens JA, Farrell RL (2009). Highly
647 specialized microbial diversity in hyper-arid polar desert. *Proc Natl Acad Sci U S A*
648 **106:** 19964-19969.
649 Pointing SB, Belnap J (2012). Microbial colonization and controls in dryland systems.
650 *Nat Rev Microbiol* **10:** 551-562.

651 Pons P, Latapy M (2005). Computing communities in large networks using random walks
652 *arXiv:physics/0512106 [physicssoc-ph]*.
653 Portillo MC, Gonzalez JM (2009). CRISPR elements in the Thermococcales: evidence
654 for associated horizontal gene transfer in Pyrococcus furiosus. *J. Appl Genet* **50:**
655 421-430.
656 Price MN, Dehal PS, Arkin AP (2010). FastTree 2--approximately maximum-likelihood
657 trees for large alignments. *PLoS One* **5:** e9490.
658 Robinson CK, Wierzchos J, Black C, Crits-Christoph A, Ma B, Ravel J *et al.* (2015).
659 Microbial diversity and the presence of algae in halite endolithic communities are
660 correlated to atmospheric moisture in the hyper-arid zone of the Atacama Desert.
661 *Environ Microbiol* **17:** 299-315.
662 Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF,
663 Rohwer F *et al.* (2009). Explaining microbial population genomics through phage
664 predation. *Nat Rev Microbiol* **7:** 828-836.
665 Roine E, Oksanen HM (2011). Viruses from the hypersaline environments: current
666 research an future trends. In: Ventosa A, Oren A, Ma Y (eds). *Halophiles and*
667 *Hypersaline Environments*. Springer: Heidelberg. pp 153–172
668 Roux S, Enault F, Hurwitz BL, Sullivan MB (2015). VirSorter: mining viral signal from
669 microbial genomic data. *Peer J* **3:** e985.
670 Sabet S (2012). Halophilic viruses. In: Vreeland R (ed). *Advances in Understanding the*
671 *Biology of Halophilic Microorganisms*. Springer: New York, NY. pp 81–116

672 Santos F, Yarza P, Parro V, Meseguer I, Rossello-Mora R, Anton J (2012). Culture-
673 independent approaches for studying viruses from hypersaline environments. *Appl*
674 *Environ Microbiol* **78:** 1635-1643.
675 Sato T, Atomi H, Imanaka T (2007). Archaeal type III RuBisCOs function in a pathway
676 for AMP metabolism. *Science* **315:** 1003-1006.
677 Sencilo A, Jacobs-Sera D, Russell DA, Ko CC, Bowman CA, Atanasova NS *et al.* (2013).
678 Snapshot of haloarchaeal tailed virus genomes. *RNA Biol* **10:** 803-816.
679 Sencilo A, Roine E (2014). A Glimpse of the genomic diversity of haloarchaeal tailed
680 viruses. *Front Microbiol* **5:** 84.
681 Söding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein
682 homology detection and structure prediction. *Nucl Acids Res* **33:** W244-W248;
683 doi:210.1093/nar/gki1040.
684 Tamaru Y, Takani Y, Yoshida T, Sakamoto T (2005). Crucial role of extracellular
685 polysaccharides in desiccation and freezing tolerance in the terrestrial
686 cyanobacterium Nostoc commune. *Appl Environ Microbiol* **71:** 7327-7333.
687 Tang SL, Nuttall S, Ngui K, Fisher C, Lopez P, Dyall-Smith M (2002). HF2: a double-
688 stranded DNA tailed haloarchaeal virus with a mosaic genome. *Mol Microbiol* **44:**
689 283-296.
690 Tang SL, Nuttall S, Dyall-Smith M (2004). Haloviruses HF1 and HF2: evidence for a
691 recent and large recombination event. *J Bacteriol* **186:** 2810-2817.
692 Walker JJ, Pace NR (2007). Endolithic microbial ecosystems. *Annu Rev Microbiol* **61:**
693 331-347.

694 Wei STS, Fernandez-Martinez M, Chan Y, Van Nostrand JD, de los Rios-Murillo A,
695 Chiu JMY *et al.* (2015). Diverse metabolic and stress-tolerance pathways in
696 chasmoendolithic and soil communities of Miers Valley, McMurdo Dry Valleys,
697 Antarctica. *Polar Biol* **38:** 433-443.
698 Wheeler TJ, Eddy SR (2013). nhmmer: DNA Homology Search With Profile HMMs.
699 *Bioinformatics* **28:** 2487-2489.
700 Wierzchos J, Ascaso C, McKay CP (2006). Endolithic cyanobacteria in halite rocks from
701 the hyperarid core of the Atacama Desert. *Astrobiology* **6:** 415-422.
702 Wierzchos J, Davila AD, Sánchez-Almazo IM, Hajnos M, Swieboda R, Ascaso C
703 (2012a). Novel water source for endolithic life in the hyperarid core of the Atacama
704 Desert. *Biogeosci Discuss* **9:** 3071-3098.
705 Wierzchos J, de los Ríos A, Ascaso C (2012b). Microorganisms in desert rocks: the edge
706 of life on Earth. *Inter Microbiol* **15:** 173-183.
707 Wood DE, Salzberg SL (2014). Kraken: ultrafast metagenomic sequence classification
708 using exact alignments. *Genome Biol* **15:** R46.
709 Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW (2014). MaxBin: an automated
710 binning method to recover individual genomes from metagenomes using an
711 expectation-maximization algorithm. *Microbiome* **2:** 26.
712 Ziolkowski LA, Wierzchos J, Davila AF, Slater GF (2013). Radiocarbon evidence of
713 active endolithic microbial communities in the hyper-arid core of the Atacama
714 Desert,. *Astrobiology* **13:** 607-616.
715

716 **Figure legends**

717 Figure 1: (a) Shaded relief digital map of the northern Atacama Desert, Chile, with the
718 Salar Grande sampling location (triangle); (b) Salar Grande halite nodule field; (c)
719 Section of a halite nodule with a back arrow indicating the green diffuse colonization
720 zone.

721

722 Figure 2: Taxonomic assignments of the halite metagenome sequence reads using
723 Phylosift and displayed with Krona.

724

725 Figure 3: Comparison of isoelectric point profiles of the predicted proteomes for the
726 halite alga (blue) and 3 closely related halophilic algae, *Micromonas sp. RCC299 (red),*
727 *Ostreococcus tauri (purple),* and *Dunaliella salina (orange).* All reported protein
728 sequences in the nr database were used for *Micromonas sp. RCC299* and *O. tauri.*
729 Sequences from the UniProt Protein database were used for *D. salina*.

730

731 Figure 4: Circular representation of the SG9 genome using the Circleator tool (Crabtree
732 *et al.*, 2014). The G+C% of a 10 kbp window is displayed on the outermost circle (G+C
733 scale: 40 to 50%). Following circles represent predicted genes on the forward strand and
734 reverse strands, respectively; genes related to potassium homeostasis and uptake are in
735 red, genes related to a heterotrophic lifestyle are in green, genes related to DNA repair
736 are in purple, and Cas genes are in orange. The position of the ribosomal rRNA genes is
737 indicated in grey.

738

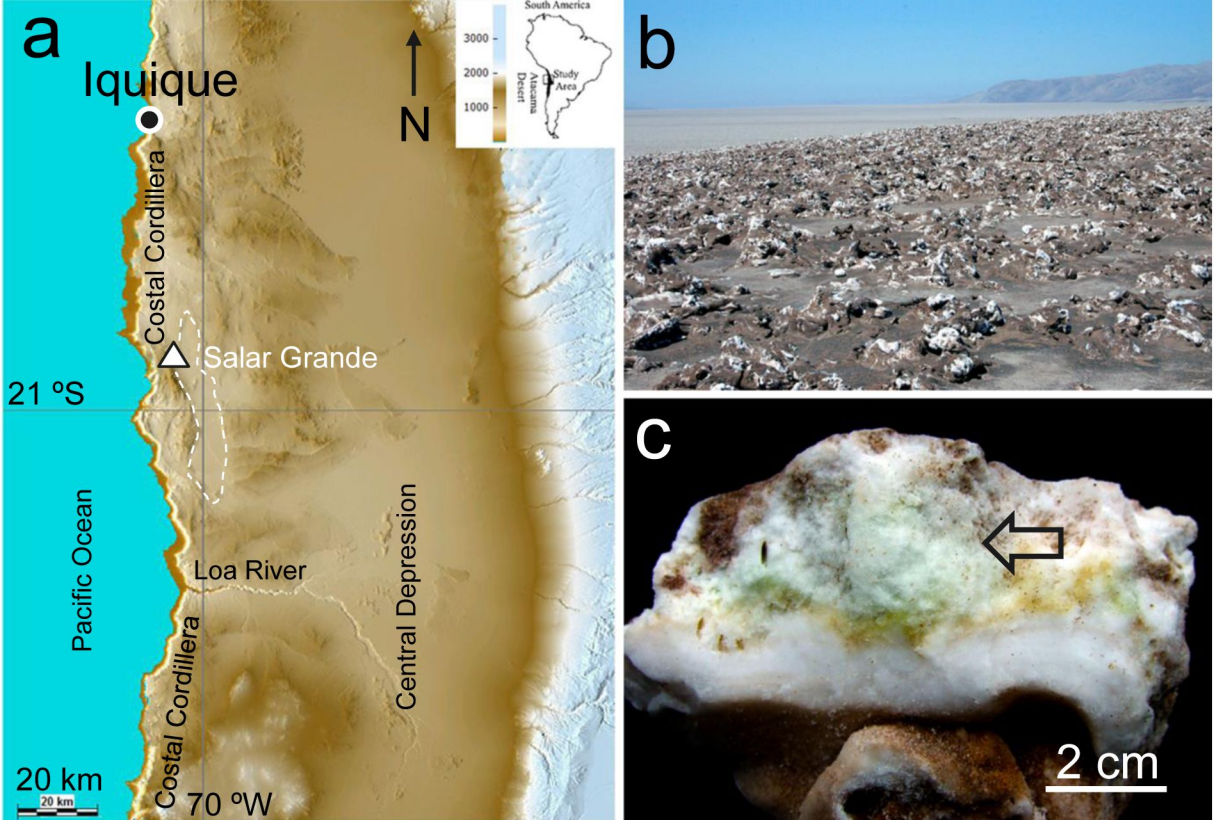
739 Figure 5: Phylogenetic position of the novel *Candidatus* Nanopetramus SG9 genome
740 within the Archaea. The tree was built with alignments of concatenated genes for *rpsB*,
741 *rplA*, *IF-2*, *rpsI, S5, S7, rplF, rplE, rpsK, S8, L18P/L5E*, and *rplM*. Euryarcheota are in
742 Red, the TACK phyla in purple, Nanoarchaeota in blue, and the Nanohaloarcheota in
743 green. Bacterial species were used as an outgroup. The scale bar represents 0.2 %
744 sequence divergence. Bootstrap values (1000 replicates) are shown at nodes.
745
746 Figure 6: (a) A diagram of the CRISPR/Cas system found in the *Candidatus*
747 Nanopetramus SG9 genome; (b) phylogenetic analysis of Cas1 gene products. Bootstrap
748 values (1000 replicates) are shown at nodes; and (c) G+C% of a 10 kbp windows, with a
749 1 kbp step size, across the SG9 genome with the CRISPR 10 kbp region marked by a red
750 line.
751
752 Figure 7: Phage-phage similarity network visualizing relationships between viral
753 genomes. Edges are weighted based on the proportion and % ID of shared genes between
754 two genomes. Viruses are colored according to network communities predicted by the
755 Walktrap community finding algorithm: (I), blue; (II) red; (III) yellow; (IV) green; (V)
756 pink; (VI) orange.
757
758 Figure 8: Protein-protein similarity networks of the largest clusters of viral proteins,
759 colored by the cluster assigned to the derivative virus genome of each protein.
760

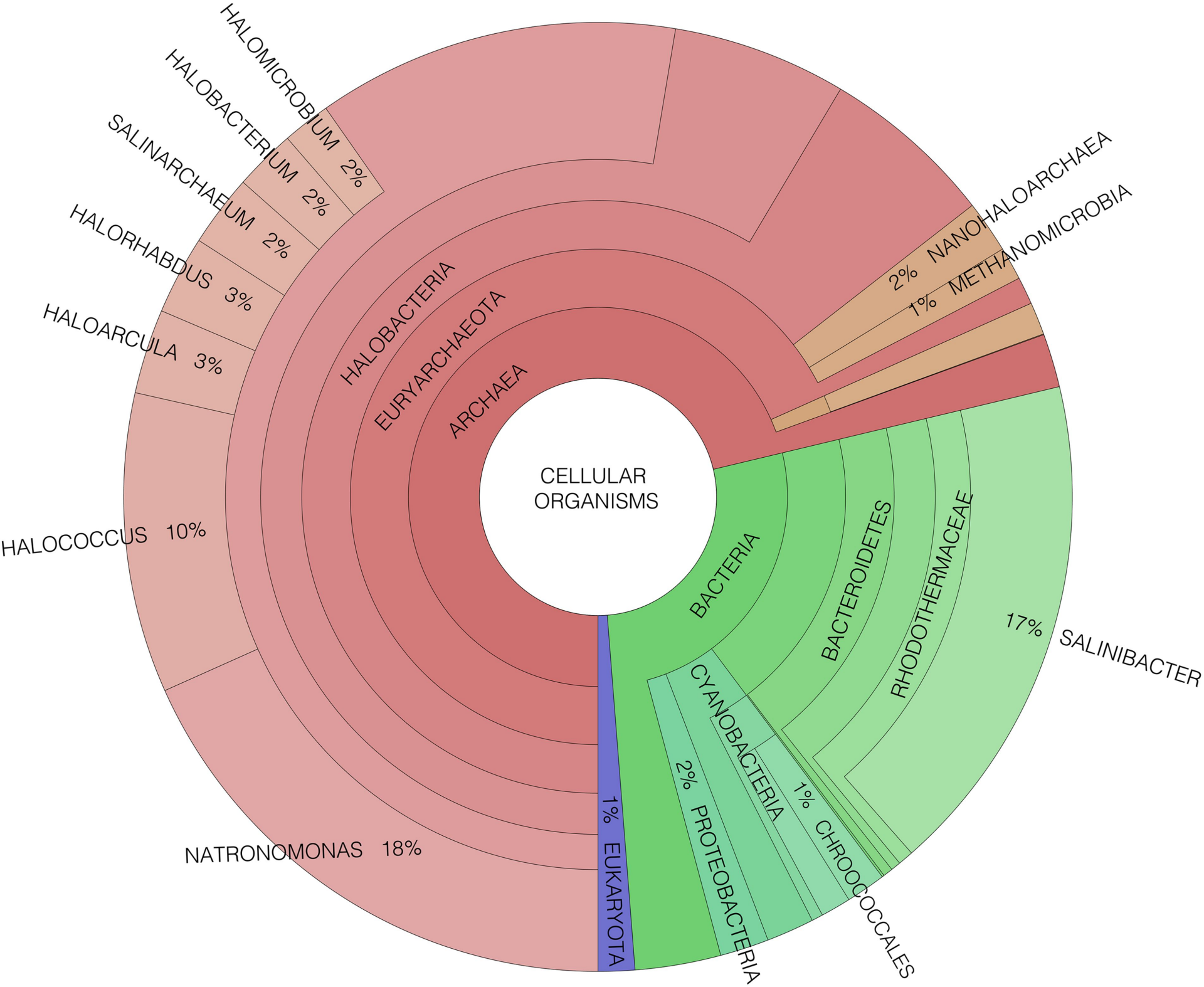
761 Table 1: Viral genome composition and structure diversity.
762

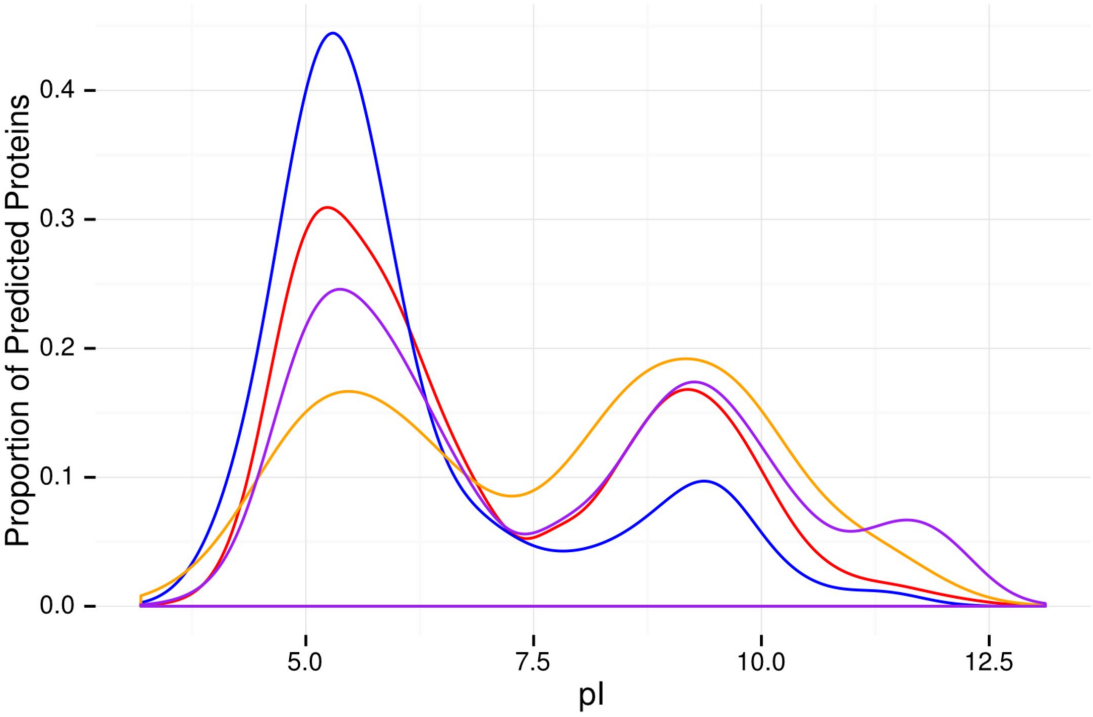
Contig	Size	GC%	Genome Structure	Putative Host	Features	Virus group
32* (II)	70.0	51.5	Linear	Halobacteria	DNA pol II, B, RNA ligase, tail assembly, baseplate J, major capside, prohead protease, portal protein.	Myovirus (similar to HSTV-2)
38* (I)	64.0	54.9	Circular	Halobacteria	DNA primase/helicase, HNH, tape measure, major capsid, prohead protease, phage head morphogenesis, terminase, DNA pol II.	Myovirus (similar to HGTV-1)
68* (II)	52.7	50.3	Linear	Halobacteria	Portal protein, tail assembly baseplate J, RNA ligase, prohead protein, integrase, DNA pol II small subunit.	Myovirus (similar to HSTV-2)
92* (I)	44.5	63.9	Circular	Halobacteria	HNH, DNA/RNA Helicase, PadR TR, terminase, major capsid, tape measure, Zn finger.	Myovirus-like
86 (I)	46.2	63.8	Linear / Incomplete	Halobacteria	Cas6, SNase-like protein, HNH, major capsid, tail protein, baseplate assembly J.	Myovirus-like (shares 2 genes with HHTV-1)
139* (I)	34.1	43.3	Circular	Nanohaloarchaea	capsid protein, phage terminase, excinuclease,	Similar to Environmental Halophage eHP-23 and eHP-35
216*	26.4	65	Circular	Halobacteria	Integrase, RNA pol sigma 24 subunit, resolvase.	No head/tail, similar to *Halosimplex carlsbadense* provirus
238 (I)	24.0	47.2	Linear / Incomplete	Halothece	DNA primase/helicase, DNA pol I, tail fiber, tape measure	Cyanophage; shares genes with *Synechococcus* phage S-CBS4 and cyanophage PSS2

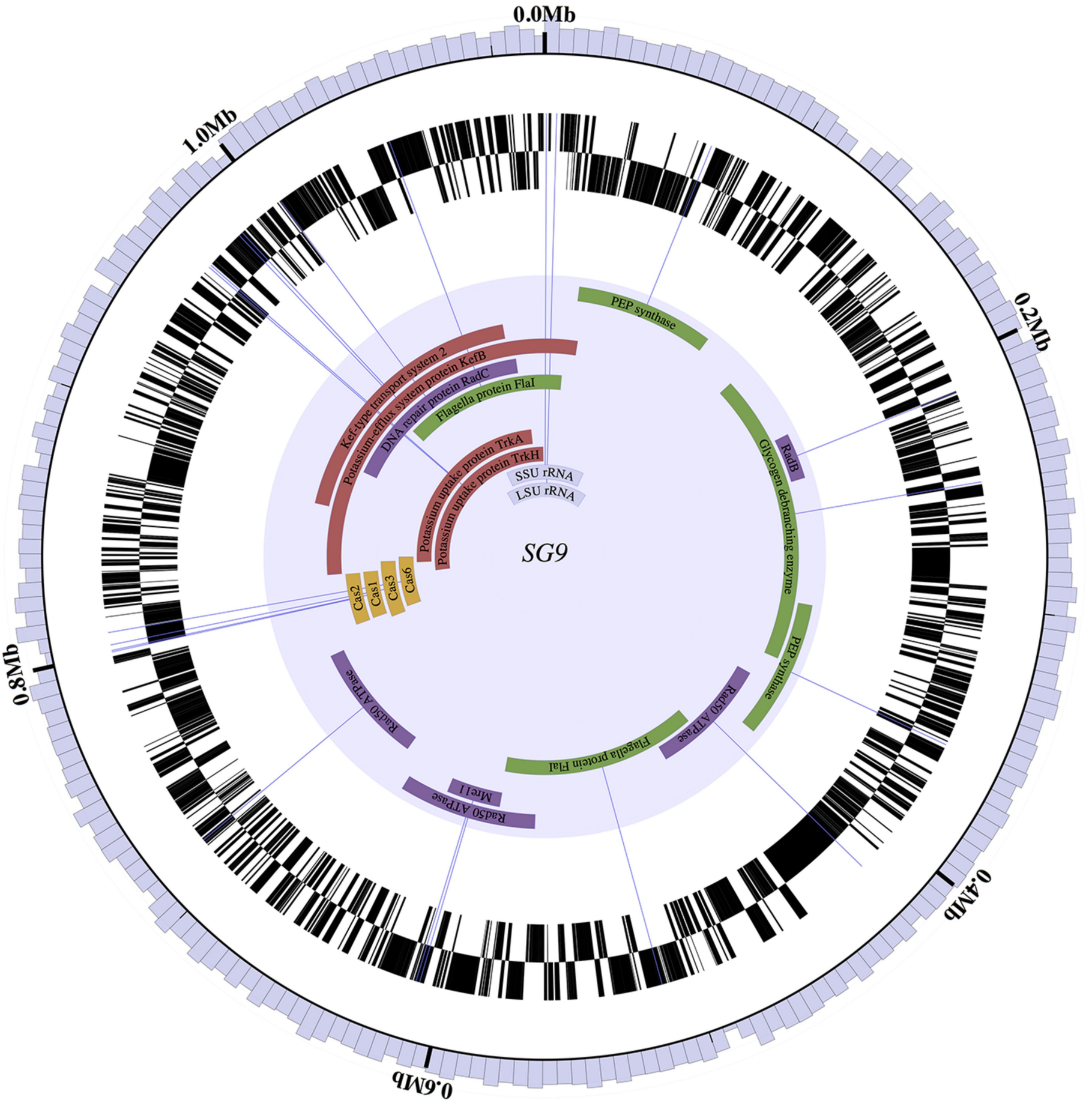
257* (V)	23.3 63.2	Circular	Halobacteria	Phage repressor, replication protein, ATPase.	No head/tail; shares genes with *Halorubrum* phage GNf2
322 (VI)	21.0 45.2	Linear / Incomplete	Halothece	Phage tail protein, phage baseplate protein.	Cyanophage; Shares genes with *Synechococcus* phage S-ShM2 and *Prochlorococcus* phage P-RSM4
687* (III)	14.4 61.4	Circular	Halobacteria	VP2, VP4, replication protein.	Shares genes with *Halobacteria* pleomorphic viruses
929*	12.3 62.5	Circular	Halobacteria	RNA polymerase sigma-24 subunit, integrase, PhiH1 repressor, VP1.	Shares genes with *Haloarcula hispanica* icosahedral virus and *Halorubrum* pleomorphic virus 3
934 (IV)*	12.3 57.6	Circular	Halobacteria	DNA pol subunits, ParB, Zn-finger domain.	Shares gene with HSTV-1 (Podovirus)
966* (IV)	12.0 63.9	Circular	Halobacteria	ParB, Zn-finger domain.	Shares gene with HSTV-1 (Podovirus)
1987 (III)	8.4 46.1	Linear / Incomplete	Nanohaloarchaea	CRISPR spacer match, Integrase, VP4 precursor, CopG TF.	

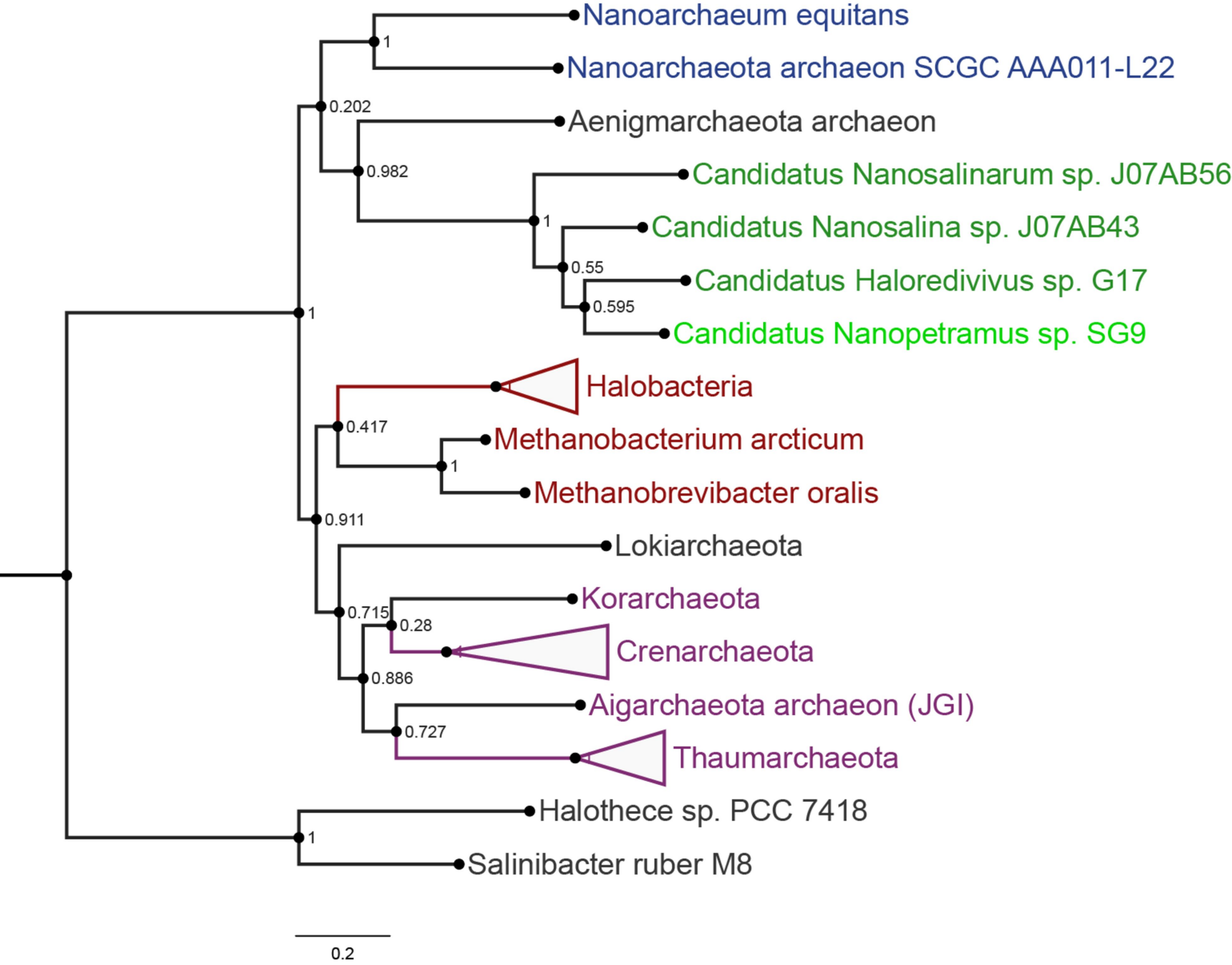
763 *Annotation curated and submitted; community type in parenthesis

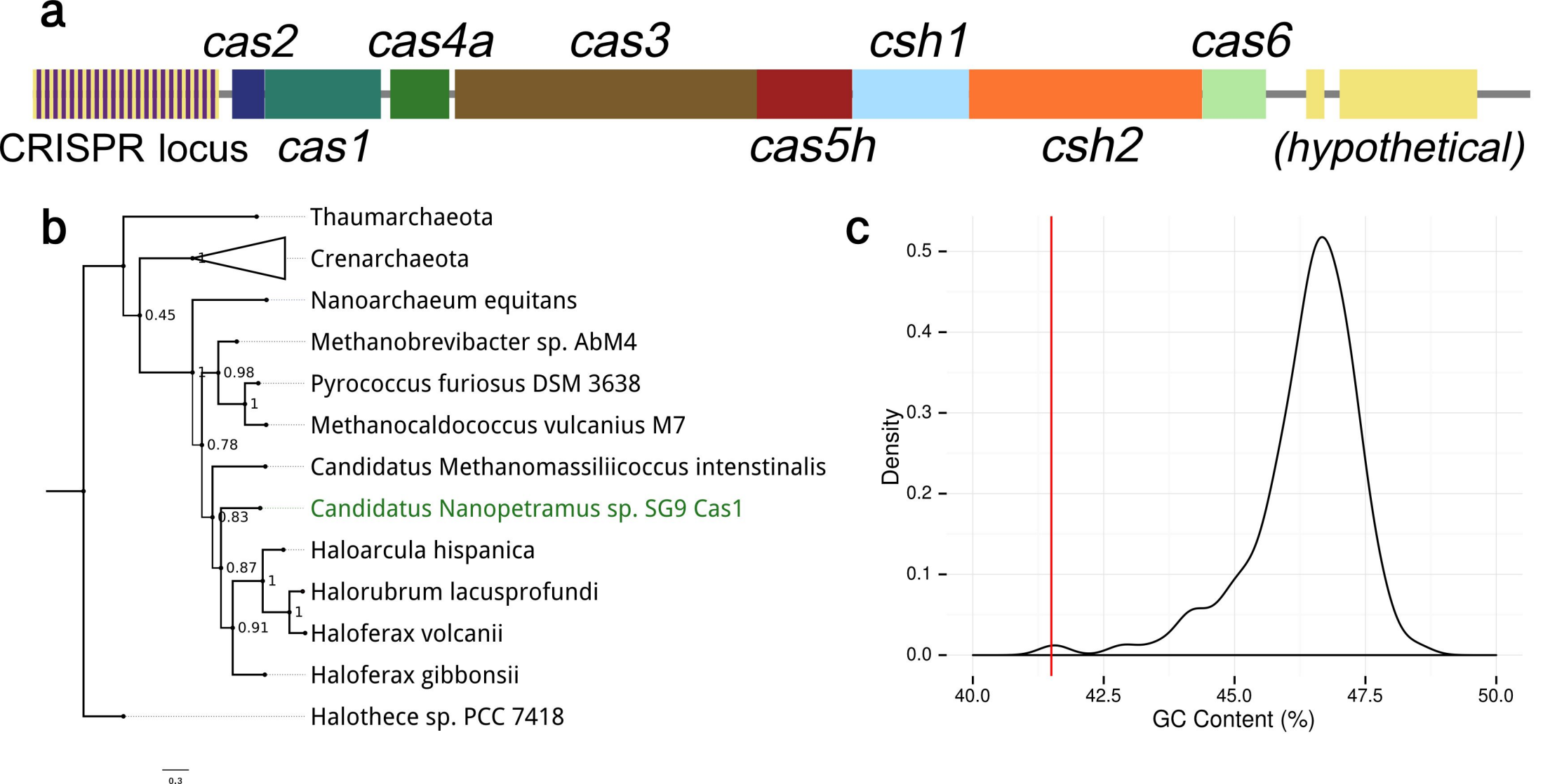

a
Iquique
N
3000
2000
1000
South America
Study Area
Atacama Desert
Costal Cordillera
Salar Grande
21 ºS
Pacific Ocean
Loa River
Central Depression
Costal Cordillera
20 km
20 km
70 ºW
b
c
2 cm

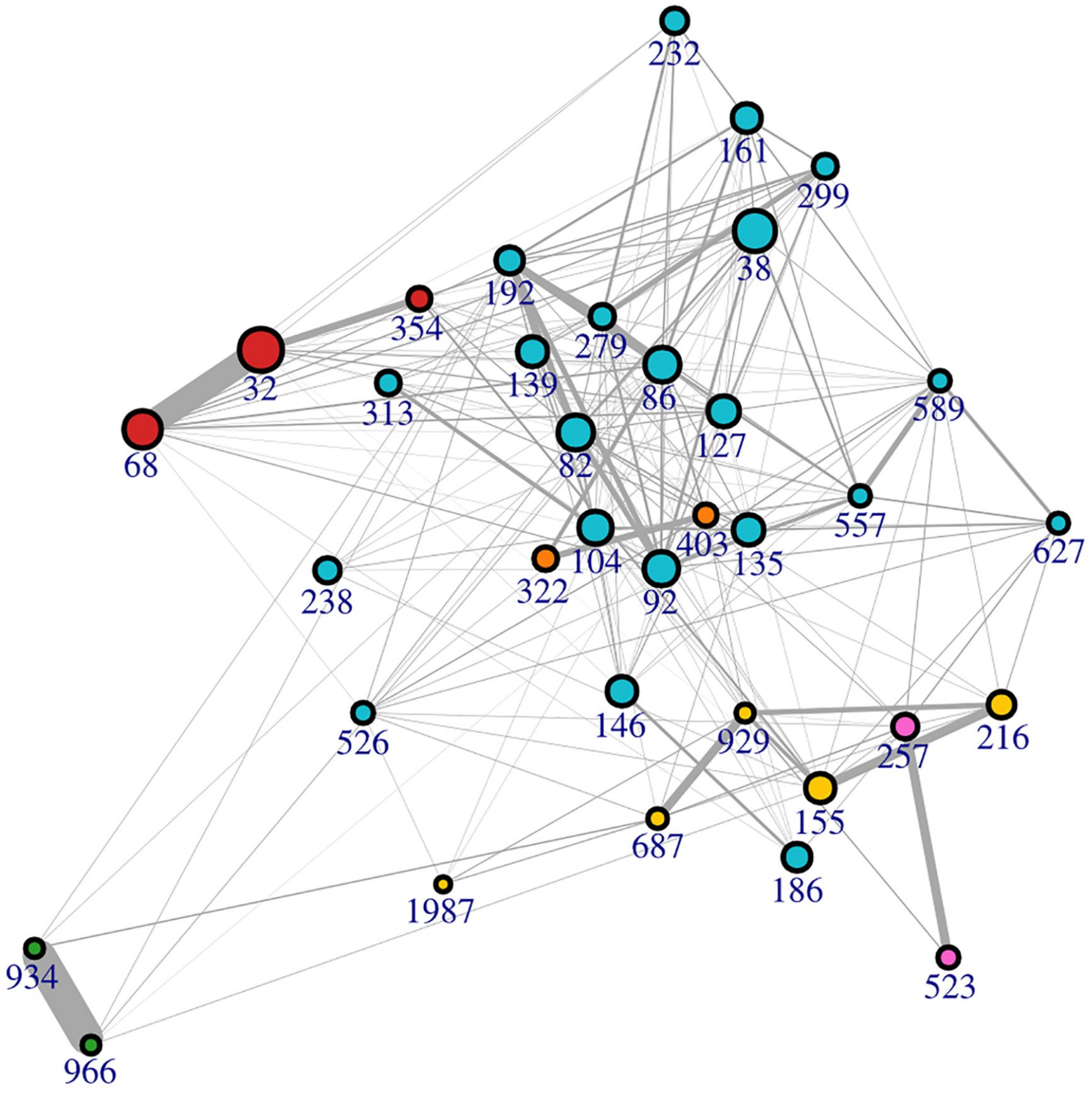

CELLULAR ORGANISMS
ARCHAEA
EURYARCHAEOTA
HALOBACTERIA
NATRONOMONAS 18%
HALOCOCCUS 10%
HALOARCULA 3%
HALORHABDUS 3%
SALINARCHAEUM 2%
HALOBACTERIUM 2%
HALOMICROBIUM 2%
2% NANOHALOARCHAEA
1% METHANOMICROBIA
BACTERIA
BACTEROIDETES
RHODOTHERMACEAE
17% SALINIBACTER
CYANOBACTERIA
1% CHROOCOCCALES
2% PROTEOBACTERIA
1% EUKARYOTA


Proportion of Predicted Proteins
0.4
0.3
0.2
0.1
0.0
5.0
7.5
10.0
12.5
pI


0.0Mb
0.2Mb
0.4Mb
0.6Mb
0.8Mb
1.0Mb
PEP synthase
Kef-type transport system 2
Potassium-efflux system protein KefB
DNA repair protein RadC
Flagella protein FlaI
Potassium uptake protein TrkA
Potassium uptake protein TrkH
SSU rRNA
LSU rRNA
SG9
Cas2
Cas1
Cas3
Cas6
Glycogen debranching enzyme
RadB
PEP synthase
Rad50 ATPase
Flagella protein FlaI
Rad50 ATPase
Rad50 ATPase
MreJ1


Nanoarchaeum equitans
Nanoarchaeota archaeon SCGC AAA011-L22
Aenigmarchaeota archaeon
Candidatus Nanosalinarum sp. J07AB56
Candidatus Nanosalina sp. J07AB43
Candidatus Haloredivivus sp. G17
Candidatus Nanopetramus sp. SG9
Halobacteria
Methanobacterium arcticum
Methanobrevibacter oralis
Lokiarchaeota
Korarchaeota
Crenarchaeota
Aigarchaeota archaeon (JGI)
Thaumarchaeota
Halothece sp. PCC 7418
Salinibacter ruber M8
1
0.202
0.982
1
0.55
0.595
1
0.417
1
0.911
0.715
0.28
0.886
0.727
1
0.2


a
cas2
cas4a
cas3
csh1
cas6
CRISPR locus
cas1
cas5h
csh2
(hypothetical)
b
Thaumarchaeota
Crenarchaeota
Nanoarchaeum equitans
Methanobrevibacter sp. AbM4
Pyrococcus furiosus DSM 3638
Methanocaldococcus vulcanius M7
Candidatus Methanomassiliicoccus intenstinalis
Candidatus Nanopetramus sp. SG9 Cas1
Haloarcula hispanica
Halorubrum lacusprofundi
Haloferax volcanii
Haloferax gibbonsii
Halothece sp. PCC 7418
0.45
0.98
0.78
0.83
0.87
0.91
0.3
c
Density
0.0
0.1
0.2
0.3
0.4
0.5
40.0
42.5
45.0
47.5
50.0
GC Content (%)

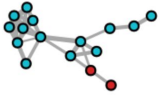

232
161
299
38
192
354
279
32
139
86
313
589
68
127
82
557
403
627
104
135
322
92
238
146
526
929
216
257
155
687
186
1987
934
523
966

DNA polymerase B

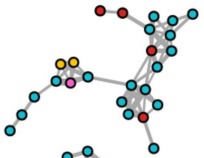
Phage head assembly

DNA primase/helicase

Terminase large subunit



HNH endonuclease associated



Tail tape measure

Terminase

Prohead protease