Determining virus-host interactions and glycerol metabolism profiles in geographically diverse solar salterns with metagenomics

Abraham G Moller [1, 2] , **Chun Liang** Corresp. [2]

[1] Department of Microbiology and Immunology, Emory University, Atlanta, Georgia, United States
[2] Department of Biology, Miami University of Ohio, Oxford, Ohio, United States

Corresponding Author: Chun Liang
Email address: liangc@miamioh.edu

Solar salterns are excellent model ecosystems for studying virus-microbial interactions because of their low microbial diversity, environmental stability, and high viral density. By using the power of CRISPR spacers to link viruses to their prokaryotic hosts, we explored virus-host interactions in geographically diverse salterns and related them to carbon cycling. Using taxonomic profiling, we identified hosts such as archaeal *Haloquadratum*, *Halorubrum*, and *Haloarcula* and bacterial *Salinibacter*, and we found that community composition related to not only salinity but also local environmental dynamics. Characterizing glycerol metabolism genes in these metagenomes suggested *Halorubrum* and *Haloquadratum* possess most dihydroxyacetone kinase genes while *Salinibacter* possesses most glycerol-3-phosphate dehydrogenase genes. We identified CRISPR spacers in the metagenomes with two different methods and found more spacers in the IC21 and C34 salterns compared with the SS19, SS33, and SS37 salterns, suggesting fewer types of CRISPR spacers in the *Haloquadratum*-majority salterns. After CRISPR detection, spacers were aligned against haloviral genomes to map virus to host. While most alignments linked viruses to *Haloquadratum walsbyi*, there were groups of interactions with the low abundance taxa *Haloarcula* and *Haloferax*. Further examination of the dinucleotide and trinucleotide usage differences between paired viruses and their hosts confirmed viruses and hosts had similar nucleotide usage signatures. Detection of *cas* genes in the salterns supported the possibility of CRISPR activity. Taken together, our studies suggest similar virus-host interactions exist in different solar salterns and that the glycerol metabolism gene dihydroxyacetone kinase is associated with *Haloquadratum* and *Halorubrum*.

Classification: Ecology, Genomics, Bioinformatics, Computational Biology, Microbiology

Title: **Determining virus-host interactions and glycerol metabolism profiles in geographically diverse solar salterns with metagenomics**

Authors: Abraham Moller[1] and Chun Liang[1,*]

Author Affiliation: [1]Department of Biology, Miami University, Oxford, Ohio 45056

[*]Corresponding Author: Chun Liang, liangc@miamioh.edu

Abstract

Solar salterns are excellent model ecosystems for studying virus-microbial interactions because of their low microbial diversity, environmental stability, and high viral density. By using the power of CRISPR spacers to link viruses to their prokaryotic hosts, we explored virus-host interactions in geographically diverse salterns. Using taxonomic profiling, we identified hosts such as archaeal *Haloquadratum*, *Halorubrum*, and *Haloarcula* and bacterial *Salinibacter*, and we found that community composition related to not only salinity but also local environmental dynamics. Characterizing glycerol metabolism genes in these metagenomes suggested *Halorubrum* and *Haloquadratum* possess most dihydroxyacetone kinase genes while *Salinibacter* possesses most glycerol-3-phosphate dehydrogenase genes. We identified CRISPR spacers in the metagenomes with two different methods and found more spacers in the IC21 and C34 salterns compared with the SS19, SS33, and SS37 salterns, suggesting fewer types of CRISPR spacers in the *Haloquadratum*-majority salterns. After CRISPR detection, spacers were aligned against haloviral genomes to map virus to host. While most alignments linked viruses to *Haloquadratum walsbyi*, there were groups of interactions with the low abundance taxa *Haloarcula* and *Haloferax*. Further examination of the dinucleotide and trinucleotide usage differences between paired viruses and their hosts confirmed viruses and hosts had similar nucleotide usage signatures. Detection of *cas* genes in the salterns supported the possibility of CRISPR activity. Taken together, our studies suggest similar virus-host interactions exist in different solar salterns and that the glycerol metabolism gene dihydroxyacetone kinase is associated with *Haloquadratum* and *Halorubrum*.

Introduction

Viruses, the most abundant entities on the planet, are critical for maintaining global biogeochemical cycles (Suttle, 2007; Rohwer, Prangishvili & Lindell, 2009). Many viruses infect and lyse microbial hosts, releasing stored nutrients back into the ecosystem through a mechanism called the viral shunt (Suttle, 2007; Weitz & Wilhelm, 2012). This relationship suggests viral-microbial interactions are important for environmental nutrient cycling. Clustered regularly interspaced palindromic repeat (CRISPR) arrays, which are found in many archaeal and bacterial genomes (Grissa, Vergnaud & Pourcel, 2007), can help us better understand the virus-host interactions that mediate nutrient cycling. Acting as an adaptive immune system against viral infection, CRISPRs incorporate short spacers cleaved from viral DNA into host genomes, providing a record of past infections and thus associations between viruses and prokaryotic hosts (Sorek, Kunin & Hugenholtz, 2008).

Virus-host interactions are particularly important in solar saltern ecosystems. These salt-collecting ponds are hypersaline ecosystems that vary from 10 up to 37% (saturation) salt concentration by weight (Ventosa et al., 2014). The salterns at Santa Pola near Alicante, Spain contain a low diversity community of halophile microbial hosts, making it a good model ecosystem (Ventosa et al., 2014). The square archaeon *Haloquadratum walsbyi* and bacterioidete *Salinibacter ruber* are the two major components of saltern microbial communities at the highest salinities (Ventosa et al., 2015). In addition, hypersaline environments are estimated to have the highest densities of viruses (10^9/mL) among studied aquatic environments (Santos et al., 2012). Reduced viral decomposition in hypersaline conditions and a lack of protozoan predation may

explain this extraordinarily high density (Santos et al., 2012). Though heterotrophic protozoans have been identified in Korean salterns (Park & Simpson, 2015), recent studies suggest these species primarily graze upon algae, not bacteria (Heidelberg et al., 2013). Though protozoa probably do not engulf bacteria, viruses may indeed prey upon halophilic bacteria and archaea.

While saltern microbial eukaryotes may not be predators of archaea and bacteria, they do serve as the primary producers in these ecosystems. The halotolerant, unicellular green algae *Dunaliella salina* synthesizes vast quantities of glycerol as an osmoprotectant in response to hyperosmotic shock (Chitlaru & Pick, 1991). Glycerol that leaks out of these algae may be the primary carbon source for the bacteria and archaea that inhabit these ecosystems (Wegmann, Ben-Amotz & Avron, 1980; Oren, 1993; Bardavid, Khristo & Oren, 2006). The rare saltern archaeon *Haloferax volcanii* converts glycerol to glycerol-3-phosphate and then dihydroxyacetone phosphate (DHAP) using glycerol kinase and glycerol-3-phosphate dehydrogenase, respectively (Sherwood, Cano & Maupin-Furlow, 2009; Rawls, Martin & Maupin-Furlow, 2011). Bacteria *Salinibacter ruber*, on the other hand, oxidizes glycerol to dihydroxyacetone (Sher et al., 2004; Bardavid & Oren, 2007). Both *Haloferax* and *Haloquadratum walsbyi* can convert dihydroxyacetone to glycolysis intermediate DHAP using dihydroxyacetone kinase (Bolhuis et al., 2006; Bardavid & Oren, 2007; Ouellette, Makkay & Papke, 2013). Considering this metabolic context, haloviruses that lyse halophile microbes may promote carbon cycling either by turning over glycerol metabolism intermediates or reducing rates of glycerol uptake within the ecosystem.

In recent years, whole metagenome sequencing has made it possible to study the total microbial DNA from solar salterns across the planet. Microbial DNA from the Santa Pola salterns has recently been sampled and sequenced (Fernandez et al., 2013), in addition to microbial DNA from salterns in Isla Cristina, Spain (Fernández et al., 2014a); Cahuil, Chile (Plominsky et al., 2014); and Chula Vista, California (Rodriguez-Brito et al., 2010). Metagenomics has also made it possible to study the vast haloviral populations that inhabit these ecosystems. A number of viral genomes from the CR30 (**cr**ystallizer, **30**% salt) saltern in Santa Pola have recently been isolated and sequenced using fosmid cloning (Santos et al., 2012). Many of these fosmid clones failed to hybridize with microbial DNA, suggesting most of isolated CR30 viruses are lytic (Santos et al., 2010). Haloviral genomes have also been assembled from viral metagenomes sequenced from Lake Tyrell in Australia (Emerson et al., 2013) and the Chula Vista salterns in California (Rodriguez-Brito et al., 2010).

While metagenomic sequencing has provided extraordinary glimpses into the viral and microbial diversity of hypersaline ecosystems, it has also enabled detection of microbial CRISPR arrays. Recent studies in a variety of ecosystems have exploited CRISPRs to link environmental viruses to their microbial hosts (Tyson & Banfield, 2008; Anderson, Brazelton & Baross, 2011; Berg Miller et al., 2012; Emerson et al., 2013; Zhang et al., 2013; Sanguino et al., 2015; Tschitschko et al., 2015). Among these, one study examined virus-archaea interactions in an ecosystem similar to salterns – the hypersaline Lake Tyrrell, Australia (Emerson et al., 2013). Although this study showed that CRISPRs in the lake were remarkably dynamic over time, it also indicated a number of viruses could be mapped to saltern microbes with CRISPR spacers (Emerson et al., 2013). In the

saltern context, the abundant community member *Haloquadratum walsbyi* encodes two clusters of *cas* genes (Bolhuis et al., 2006; Dyall-Smith et al., 2011) and contains CRISPR spacers that map to many Santa Pola haloviral genomes (Garcia-Heredia et al., 2012). These studies suggest CRISPRs have great potential to indicate virus-archaea interactions within saltern ecosystems.

In this study, we analyze saltern metagenomes to link haloviruses to their hosts and improve our understanding of carbon cycling amongst the prokaryotes that live in these environments. Furthermore, we incorporate knowledge gained from studying not one but many hypersaline ecosystems across the planet. We not only use metagenomic CRISPR sequences to map virus-host interactions across different saltern microbial populations, but also use relevant taxonomic and functional profiling tools to determine relative taxonomic abundances and relationships between taxa and glycerol metabolism functions. In this context, we find common sets of virus-host interactions and glycerol metabolism gene associations with particular taxa across these geographically distinct salterns.

Materials and Methods

Obtaining metagenomic data sets and haloviral genomes. The Santa Pola, Isla Cristina, and Cahuil microbial metagenomes were downloaded from the NCBI Sequence Read Archive (SRA). Technical details about these microbial metagenomes are provided in Table 1. Further details about sampling are described in associated publications and genome announcements (Ghai et al., 2011; Fernandez et al., 2013; Fernández et al., 2014a,b; Plominsky et al., 2014). Chula Vista microbial metagenomes were obtained from the iMicrobe Collaborative (Hurwitz, 2014) and have been previously examined in a larger

138 study of several aquatic biomes (Rodriguez-Brito et al., 2010). Chula Vista metagenomes
139 were grouped by corresponding salt concentration (low, medium, or high) and
140 concatenated to form single low (CV6-8), medium (CV12-14), and high (CV27-30) salt
141 concentration metagenomes, as shown in Table 2. The ways combined microbial
142 metagenomes were constructed and utilized in later analyses are outlined in Table 3.
143 Santa Pola and Isla Cristina metagenomes (SS13, SS19, IC21, SS33, and SS37) were
144 analyzed individually except for assembly and *de novo* CRISPR detection steps. The
145 Cahuil (C34) metagenome was analyzed individually throughout. All analyses performed
146 on the microbial metagenomes are outlined in Figure S1.

147 Haloviral genomes as well the Lake Tyrell viral contigs were downloaded from
148 NCBI GenBank (Table S1). Additionally, Chula Vista saltern viral metagenomes were
149 obtained through iMicrobe (Hurwitz, 2014) and then combined into one viral metagenome
150 (Table 3). This combined viral metagenome (Table 3) was then assembled with Newbler
151 into Chula Vista viral contigs that were also included in the haloviral library (Table S1).

152 **Profiling taxonomic and functional compositions of selected metagenomes.** The
153 selected individual metagenomes (Tables 1 and 2) were profiled for relative taxonomic
154 abundances with MetaPhyler (Liu et al., 2011a). MetaPhyler aligns reads against a library
155 of clade-specific marker genes. Reads that align against a marker gene above a particular
156 threshold alignment score are then binned to the corresponding taxon (e.g.,
157 *Haloquadratum*, for reads aligning to *Haloquadratum* marker genes). Taxonomic profiles
158 were further compared in MATLAB using principal component analysis (PCA). Clusters
159 identified with PCA were evaluated for statistical significance using the ADONIS R
160 function. The percent of reads mapping to markers for each metagenome were also

161 calculated and compared to detect differences in the proportions of reads used for
162 taxonomic profiling depending on metagenome.

163 The metagenomes were further profiled for functional abundances with ShotMAP
164 (Nayfach et al., 2015) and metAnnotate (Petrenko et al., 2015). ShotMAP determination
165 of gene family abundance was conducted with a Hidden Markov Model (HMM)-based
166 search algorithm against an indexed library of Pfam family HMMs. Abundances were
167 calculated from the numbers of metagenomic reads mapping to particular Pfam families
168 normalized by the length of each Pfam target gene in the gene family. We selected
169 abundances of nine gene families known to be related to glycerol metabolism for further
170 analysis – dihydroxyacetone kinase (Dak1 – PF02733; Dak1_2 – PF13684; Dak2 –
171 PF02734), NAD^+-dependent glycerol dehydrogenase (Fe-ADH – PF00465; Fe-ADH_2 –
172 PF13685), NAD^+-dependent glycerol-3-phosphate dehydrogenase (both N- and C-
173 terminal domains: NAD_Gly3p_dh_N – PF01210; NAD_Gly3p_dh_C – PF07479), and
174 glycerol kinase (FGGY_N – PF00370; FGGY_C – PF02782), listed both with Pfam family
175 names and ID numbers. Relative abundances were then analyzed for correlations with
176 respect to salinity using linear regression fits in Microsoft Excel.

177 metAnnotate (Petrenko et al., 2015) was used to determine the abundance of
178 particular Pfam gene families for particular taxa. Like ShotMAP, metAnnotate uses an
179 HMM-based search algorithm to find reads that match certain Pfam gene family HMMs.
180 Unlike ShotMAP, however, it also determines the taxonomic affiliations of these gene
181 family reads, and thus determines which taxa have which gene families. We used
182 metAnnotate to taxonomically profile nine aforementioned gene families known to be
183 related to glycerol metabolism. We then calculated the proportion of reads mapping to

each taxon among all those mapping to a gene family to determine the taxonomic profiles of these gene families. These gene family profiles were then visualized as heatmaps and hierarchically clustered with the R pheatmap package. Clusters were evaluated for statistical significance with the ADONIS R function.

***De novo* and reference-guided CRISPR detection.** CRISPR arrays were detected using two different methods (reference-guided and *de novo*). CRISPR arrays are highly difficult to assemble due to their repetitive nature. CRISPR arrays were thus identified purely from unassembled metagenomic reads. *De novo* detection was performed with Crass (Skennerton, Imelfort & Tyson, 2013) with default parameters, which identifies repeat-spacer-repeat patterns in individual reads, uses these as “seeds” to recruit other CRISPR-containing reads, and then assembles these reads into CRISPR arrays. We further clustered Crass spacers using CD-HIT (Li & Godzik, 2006, p.) with a similarity threshold of 0.9 and aligned them against known direct repeats from CRISPRdb (Grissa, Vergnaud & Pourcel, 2007) to determine the associated host organism.

Because Crass does not directly indicate the putative host encoding the detected CRISPRs, we developed a pipeline (MetaCRAST) to constrain spacer detection by expected host species (Moller & Liang, 2016). MetaCRAST searches each read for direct repeat (DR) sequences matching query DRs specified by the user. These query DRs are selected from known CRISPR DRs. The query DR sequences might be known to be present in CRISPR arrays found in genomes expected from the taxonomic profile; for example, known genomic *Haloquadratum walsbyi* DRs might be used to search a metagenome for *Haloquadratum walsbyi* CRISPRs. Direct repeats within a certain Levenshtein edit distance (i.e., number of insertions, deletions, or substitutions necessary

to convert one sequence to another) are identified in reads using the Wu-Manber multi-pattern search algorithm (Wu, Manber & Myers, 1995).

In the first step of the pipeline, reads that contain a single DR sequence matching the user-specified DR are detected with the Wu-Manber algorithm and collected for further analysis. In the second step of the pipeline, individual reads gathered in the first step are searched for two or more copies of the query direct repeat sequence within appropriate distances from each other. In the case of CRISPRs, direct repeats have been found to be between 23 and 47 bp long, while spacers, and thus the distances between DRs, have been found to be between 26 and 50 bp long (Skennerton, Imelfort & Tyson, 2013). In the third step, the sequence fragments between the direct repeats are extracted as the spacers. These spacers are then collected as a non-redundant set in a hash, and then further clustered into non-redundant sets using CD-HIT (Li & Godzik, 2006).

MetaCRAST was implemented in Perl and takes a set of query direct repeats and a metagenome (each in FASTA format) as file inputs. It is publically available for download at GitHub (https://github.com/molleraj/MetaCRAST), where all command line parameters are also documented. It has also been validated against simulated metagenomes and tested against two real metagenomes corresponding to acid mine drainage (AMD) and enhanced biological phosphorus removal (EBPR) ecosystems (Moller & Liang, 2016).

We performed reference-guided CRISPR detection with MetaCRAST using the following parameters. We used a group of 29 DR sequences previously detected in Haloferacales and Halobacteriales genomes (Table S4) in the genomic CRISPR database CRISPRdb (Grissa, Vergnaud & Pourcel, 2007) as query DRs. We thus

searched for reads containing sequences matching any of these 29 direct repeat sequences. The maximum allowed distance between detected direct repeats (i.e., maximum spacer length) was set to 60 bp. Acceptable edit distance - the allowed differences (insertions, deletions, or substitutions) between query and detected read DR sequences - were adjusted from 0 to 3, as shown in Figure 4, to see how variation in edit distance affected reference-guided CRISPR detection. All spacers detected for each condition (for each metagenome and acceptable error number) were then clustered using CD-HIT with a clustering similarity threshold of 0.9, which effectively clusters spacers together into a non-redundant set based on a pairwise similarity of 90% or greater.

Mapping virus-host interactions with CRISPR spacers. CRISPR spacers extracted from the metagenomes with Crass and reference-guided methods and clustered with CD-HIT were then aligned against the library of haloviral genomes using NCBI BLAST (Altschul et al., 1990) with an e-value cutoff of 1e-03. All spacers detected with Crass for the combined Santa Pola dataset (see Table 3) were aligned against this library to create its corresponding interaction network. For several individual metagenomes (e.g., SS13, SS19), no Crass spacers mapped to the haloviral genomes, making the more sensitive reference-guided search (MetaCRAST) necessary. Thus, for individual metagenomes, spacers detected with MetaCRAST using a maximum Levenshtein edit distance of 3 were aligned for network creation. BLAST results from each method were then processed to form lists of virus-host pairs. Networks of virus-host interactions were drawn with Cytoscape (Shannon et al., 2003).

Comparative analysis of virus and host sequence characteristics. Viruses and hosts matched through CRISPR mapping were further compared in nucleotide k-mer

253 frequencies to determine whether virus and host frequencies were similar. Dinucleotide
254 and trinucleotide usages were determined using the function *nmercount* from the
255 MATLAB Bioinformatics Toolbox. Dinucleotide and trinucleotide frequencies were
256 compared for a non-redundant list of viruses associated to each of the six detected hosts
257 listed below. All viruses were associated to possible hosts using our in-house reference-
258 based CRISPR detection method. *Haloquadratum walsbyi* C23, *Haloferax volcanii* DS2,
259 *Halogeometricum borinquense* DSM 11551, *Halorhabdus utahensis* DSM 12940,
260 *Haloarcula marismortui* ATCC 43049 (chromosome 1), and *Haloarcula hispanica* N601
261 (chromosome 1) host genomes were analyzed along with the matched haloviral
262 genomes. To measure the sequence characteristic differences among virus-host pairs,
263 each virus dinucleotide and trinucleotide frequency was subtracted from each
264 corresponding host genome frequency and squared. All these squared distances were
265 summed for possible dinucleotides and for possible trinucleotides to generate an overall
266 Euclidean measure of distance between virus and host.

267 **Assembling metagenomes and detecting halobacterial *cas* genes in metagenomic**
268 **contigs.** Combined Santa Pola, combined Chula Vista, and Cahuil metagenomes (Table
269 3 outlines combined metagenomes) were assembled into contigs to detect *cas* operons.
270 Combined Santa Pola, combined Chula Vista, and Cahuil metagenomes were assembled
271 both with Newbler and Velvet, creating a total of six sets of contigs (Zerbino & Birney,
272 2008). Both Newbler and Velvet were used to compare the effects of Overlap-Layout-
273 Consensus (Newbler) and de Bruijn graph (Velvet) assembly methods. Newbler assembly
274 was performed with default parameters, while Velvet assembly was performed with a k-
275 mer size of 31.

Newbler and Velvet contigs assembled from each combined metagenome were first aligned against *cas* genes annotated in *Haloferax volcanii* DS2, *Halorubrum lacusprofundi* ATCC 49239, and *Haloquadratum walsbyi* C23 genomes using NCBI blastn (Altschul et al., 1990) with an e-value cutoff of 1e-06 and NCBI tblastx with an e-value cutoff of 1e-30. Color-coded *cas* detection results for genes in the *Haloquadratum*, *Halorubrum*, and *Haloferax cas* operons were drawn with the synthetic biology tool Pigeon (Bhatia & Densmore, 2013). A library of *cas* genes from the three major halobacterial orders (the Haloferacales, Halobacteriales, and Natrialbales) were downloaded from NCBI Genbank (see Tables S13-S15) and aligned against each set of assembled contigs (assembled with each method and from each combined metagenome) to determine the taxonomic affiliation of matching *cas* genes. BLAST alignments were performed with the same parameters described above. Taxonomy information was extracted from matching *cas* gene Genbank records with the NCBI Efetch utility to determine the taxonomic affiliations of the assembled *cas* contigs.

Results

Taxonomic profiling identifies predominant members of saltern microbial communities. We profiled the taxonomic abundances of metagenomes sequenced from **C**ahuil (C34), **C**hula **V**ista (CV6-8, CV12-14 and CV27-30), **I**sla **C**ristina (IC21), and **S**anta Pola **S**altern (SS13, SS19, SS33, SS37) sites (where the ID numbers indicate salinity percentages) to better understand the population structures of saltern microbial communities and to compare host abundances amongst sites. We found the low salinity Chula Vista (CV6-8) and SS13 salterns had different microbial communities from the other salterns, containing large proportions of bacterial Actinobacteria, Proteobacteria

299 and Rhodobacteraceae (7.64, 14.28, and 20.69% in SS13 and 6.6, 10.05, and 11.13% in
300 CV6-8, respectively – see Figure 1A). In addition, we found *Haloquadratum* was the major
301 taxon identified in the SS19, SS33, and SS37 salterns (88.32, 65.55, and 78.06%
302 abundance, respectively), while *Halorubrum* and Halobacteriaceae were major
303 components of the Cahuil and IC21 salterns (31.79 and 64.55% in Cahuil; 32.5 and
304 51.63% in IC21, respectively - see Figure 1A). Bacterial *Salinibacter* was identified in the
305 SS19, SS33, and SS37 salterns (2.18, 7.16, and 6.82% relative abundances,
306 respectively). While *Haloquadratum* was the major identified taxon in the higher salinity
307 Santa Pola salterns, Halobacteriaceae composed the most abundant identified taxon
308 identified in the medium and high salinity Chula Vista salterns (51.46 and 77.02%,
309 respectively). We realize that the Halobacteriaceae component of the saltern
310 communities may also include species in the new order Haloferacales (Gupta, Naushad
311 & Baker, 2015), but this taxon was defined after the MetaPhyler marker library was
312 constructed (Liu et al., 2011a). These taxonomic profiles and clustering results further
313 guided our analyses of host CRISPR arrays, as we used taxonomy to guide our searches
314 for CRISPR spacers.

315 We further compared the taxonomic profiles through principal component analysis,
316 which found that the first two principal components accounted for 88% of variation
317 amongst the examined taxonomic profiles. PCA grouped the SS19, SS33, and SS37
318 profiles; low salinity Chula Vista (CV6-8) and SS13 profiles; and medium salinity Chula
319 Vista (CV12-14), high salinity Chula Vista (CV27-30), IC21, and C34 profiles (Figure 1B).
320 These groups were found to be statistically significant in their between-group differences
321 (ADONIS; F=19.817; p=0.002).

We also determined the percentage of reads mapping to phylogenetic markers to detect read length-dependent differences in taxonomic profiling among metagenomes (Figure S2). The percentage mapping was similar among Santa Pola, Isla Cristina, and Cahuil (C34) metagenomes (an average of 0.51% of reads with a standard deviation of 0.0483% among these metagenomes), but lower among the Chula Vista metagenomes (an average of 0.21% of reads with a standard deviation of 0.041% among these metagenomes) due to lower coverage (see Tables 1 and 2).

Functional profiling reveals abundances of glycerol metabolism genes and their taxonomic affiliations. Glycerol is the main carbon source for saltern microbes (Oren, 1993). We used two functional profiling tools to study the glycerol metabolism genes critical for carbon cycling in these ecosystems. We used the tool ShotMAP to determine abundances of glycerol metabolism gene families (Nayfach et al., 2015) and metAnnotate to study the taxonomic distribution of reads mapping to these gene families (Petrenko et al., 2015). As with MetaPhyler, Cahuil, Isla Cristina, and Santa Pola metagenomic reads mapped to the ShotMAP reference library in higher proportions than the Chula Vista reads (Figure S3). Using these tools, we examined the abundances and taxonomic affiliations of the glycerol kinase (FGGY_N and FGGY_C), NAD^+-dependent glycerol-3-phosphate dehydrogenase (NAD_Gly3P_dh_N and NAD_Gly3P_dh_C), glycerol dehydrogenase (iron-containing alcohol dehydrogenase – Fe-ADH and Fe-ADH_2), and dihydroxyacetone kinase (Dak1, Dak1_2, and Dak2) Pfam gene families.

We first determined gene family abundances with ShotMAP and then examined relationships between these abundances and salinity levels. Among the gene families we examined, we noticed both negative and positive correlations between relative gene

family abundance and salinity (Figure 2). While several gene families showed weak correlations ($R^2 < 0.10$) with salinity (Dak1, $R^2 = 0.0487$; Fe-ADH_2, $R^2 = 0.0208$; FGGY_C, $R^2 = 0.0437$), two showed positive correlations (Dak2, $R^2 = 0.3264$; FGGY_N, $R^2 = 0.1473$) and several showed negative correlations with salinity (NAD_Gly3P_dh_N, $R^2 = 0.5856$; NAD_Gly3P_dh_C, $R^2 = 0.619$; Fe-ADH, $R^2 = 0.2984$). We also noted family abundances calculated for the Chula Vista metagenomes to be larger than those found for any other metagenome. More short reads mapping to these gene families may account for these higher family abundances than those found for the Santa Pola, Isla Cristina, and Cahuil metagenomes.

We next determined the taxonomic affiliations of reads mapping to these gene families using metAnnotate. We generated heat maps of these affiliations hierarchically clustered both with regard to metagenome (Figure 3A) and gene family (Figure 3B). As with taxonomic profiles, the taxonomic affiliations of glycerol metabolism genes in the *Haloquadratum*-rich SS19, SS33, and SS37 metagenomes group together, as do those of the Chula Vista metagenomes, and the SS13, IC21, and Cahuil (C34) metagenomes (Figure 3A). Across gene family, taxonomic affiliations of the dihydroxyacetone kinase families were closest to each other, as well as those of the alcohol dehydrogenase (Fe-ADH) and glycerol kinase (FGGY) gene families (Figure 3B). The two NAD^+-dependent glycerol-3-phosphate dehydrogenase families grouped together, but many reads mapping to these families were found to be unclassified (Figure 3B). These described gene family and metagenome clusters were found to be statistically significant (ADONIS; F=7.5468 and 2.8768; p=0.014 and 0.013, for gene family and metagenome clusters, respectively).

We also analyzed these taxonomic affiliations for each gene family across each individual metagenome (Figure 3C). *Haloquadratum* associated with dihydroxyacetone kinase gene families in the SS19, SS33, and SS37 metagenomes while *Halorubrum* associated with these families in the IC21 and Cahuil metagenomes (Figure 3C). The two NAD^+-dependent glycerol-3-phosphate dehydrogenase families, on the other hand, associated mainly to *Salinibacter* in the Cahuil (C34), high salinity Chula Vista, SS19, SS33, and SS37 metagenomes and to *Spiribacter* in the SS13 and IC21 metagenomes. Unlike the previous families, alcohol dehydrogenase and glycerol kinase had less clear associations across the metagenomes, with association to *Haloquadratum* and *Halorubrum* but also many unclassified reads. In addition, glycerol kinase families associated to *Spiribacter* (SS13 and IC21) and the alcohol dehydrogenase family to *Halobellus* (C34, CV27-30, SS19, SS33, and SS37) across several metagenomes.

Metagenomic CRISPRs detected with two independent methods map haloviruses to saltern hosts. We detected CRISPR arrays in metagenomes using both *de novo* and reference-guided methods. We used Crass for *de novo* detection (Skennerton, Imelfort & Tyson, 2013) and a custom pipeline (MetaCRAST) for reference-guided detection (see Materials and Methods). We used a combination of Halobacteriales and Haloferacales CRISPR direct repeats from CRISPRdb as our query for MetaCRAST analyses across all metagenomes.

Overall, more spacers were detected in IC21 and Cahuil (C34) metagenomes compared with the high salinity Chula Vista, SS13, SS19, SS33, and SS37 metagenomes (Figure 4) both with MetaCRAST (3 errors allowed) and Crass. More spacers were detected with Crass compared to reference-guided methods in the combined Santa Pola,

391 SS13, and SS33 metagenomes (377 vs. 277, 29 vs. 9, and 146 vs. 80 spacers,
392 respectively). For the reference-guided detection analyses, the number of spacers
393 identified increased as the allowable error (insertions, deletions, or substitutions different
394 from the query DR sequence) in alignment was increased from 0 to 3 (Figure 4). No
395 spacers were detected with either method in the low or medium salinity Chula Vista
396 metagenomes. When we normalized by total number of metagenomic reads, we found
397 that the Cahuil (C34) metagenome had the largest number of spacers detected per million
398 reads (Figure S4).

399 We then compared the detected direct repeats and spacers among metagenomes
400 to discover common CRISPR array types or spacers (Figures 5 and 6). Remarkably,
401 common spacers were identified amongst the SS19 and SS37 metagenomes (13 with
402 Crass and 18 with MetaCRAST among these two, respectively) using both detection
403 methods (Figures 5A and 5B). Amongst the metagenomes with a higher number of
404 detected spacers (IC21 and Cahuil/C34), common spacers were detected with both
405 methods (5 in IC21 and 24 in Cahuil – see Figures 5C and 5D). Amongst the direct
406 repeats detected with Crass in the combined Chula Vista, combined Santa Pola, and
407 Cahuil metagenomes, there was one direct repeat (DR) sequence common to both Cahuil
408 and combined Santa Pola (Figure 6). We also searched the combined Santa Pola direct
409 repeats against a full database of direct repeats from CRISPRdb to determine possible
410 taxonomic affiliations. We found that 14 of the 67 detected combined Santa Pola DRs
411 matched DRs in the CRISPRdb database (Table S3). Of these 14 matching DRs, 10
412 matched haloarchaeal DRs, while 4 matched bacterial DRs.

We then aligned the spacers detected with both methods against a library of haloviral genomes (Table S1) to determine virus-host interactions in these salterns. Unlike CRISPR spacers detected with Crass in several individual Santa Pola metagenomes, spacers detected in a combination of all Santa Pola and Isla Cristina metagenomes did align to haloviral genomes (Figure 7). In this Crass analysis of the combined Santa Pola dataset, we detected virus interactions most likely with *Haloquadratum* (G467), *Haloarcula* (G42 and G82), and Pseudonocardiales (G57) hosts, based on our previous comparison of Crass DRs (listed by group number, such as G467) against CRISPRdb DRs (Figure 7). For many of the interactions (e.g., G467, G42), we could determine a possible host by such comparison. For several notable groups of interactions, however, we could not determine possible hosts (e.g., G85, G1290).

While Crass only provided a picture of virus-host interactions in the combined Santa Pola metagenome, MetaCRAST detected spacers that mapped to haloviral genomes in each of the individual Santa Pola, Isla Cristina, and Cahuil metagenomes (Figure 8). Only one interaction was detected in the SS13 dataset (Figure 8A), but there were strikingly similar groups of virus-host interactions detected in the SS19, SS33, and SS37 metagenomes (Figures 8B, 8C, and 8D). While many groups of interactions were similar to those expected from the *Haloquadratum walsbyi* strain C23 CRISPR array (Figure S5), no C23 spacers matched any detected metagenomic spacers, while at least three *Haloquadratum walsbyi* strain DSM 16790 spacers did match spacers detected in SS19, IC21, SS33, and SS37 metagenomes. Aside from *Haloquadratum*, in the SS33 metagenome, we also detected *Haloarcula*-virus interactions with eHP-4, eHP-8, and eHP-13 (Figure 8C). Quite surprisingly, we noticed remarkably different virus-

Halobacteriaceae interactions in the IC21 and Cahuil (C34) metagenomes (Figures 8E and 8F), perhaps due to differences in taxonomic composition at the genus level.

Nucleotide usage comparison suggests biological plausibility of virus-host associations. To determine whether our virus-host mappings are biologically plausible, we compared the dinucleotide and trinucleotide frequencies of CRISPR-determined virus-host pairings. Over evolutionary time, viruses' nucleotide usage steadily approaches that of the host, reflecting adaptation to host codon usage, tRNA availability, intracellular nucleotide pools, and host restriction modification systems (Edwards et al., 2015). We compared these nucleotide usage frequencies both among the predicted hosts and analyzed metagenomes (Figure 9). Among predicted hosts, *Haloquadratum walsbyi*, *Haloferax volcanii*, and *Haloarcula marismortui* showed the smallest overall differences in dimer and trimer usage, while *Haloarcula hispanica* and *Halorhabdus utahensis* showed the highest overall differences (Figures 9A and 9B). Only one virus associated to *Halogeometricum borinquense*, and it had a greater difference in trimer frequency compared with dimer frequency relative to the other predicted hosts. Among the metagenomes, the smallest differences were found in the SS33 metagenome; the next smallest differences were found in the Cahuil (C34), SS19, and SS37 metagenomes, while the largest difference was found in the single virus-host pair detected in the SS13 metagenome (Figures 9C and 9D).

Assembled metagenomic contigs contain halobacterial *cas* genes. To further confirm our CRISPR detection and mapping results, we searched through metagenomic contigs for the *cas* genes required for CRISPR activity. Because all microbial metagenomes studied were sequenced with 454 technology, and Newbler is designed to

assemble 454 reads, we expected longer contigs to be assembled with Newbler and found this in our assembly results (Table S2). We both searched for contigs that matched several different haloarchaeal *cas* operons (Figure 10A) and profiled the taxonomic affiliations of haloarchaeal *cas* genes that best matched contigs (Figure 10B). We were able to assemble almost all genes in *cas* operons in three different metagenomes (Cahuil, combined Chula Vista, and combined Santa Pola) from several different Type I-B and Type I-D *cas* operons (Figure 10A).

We also searched a library of *cas* genes from three different haloarchaeal orders (Halobacteriales, Haloferacales, and Natrialbales) and generated profiles of taxonomic affiliations for the genes that best matched contigs (Figure 10B). *Haloarcula* composed a large proportion of Halobacteriales contig affiliations, especially amongst the Cahuil (C34) contigs (71.4% among Newbler contigs; 41.4% among Velvet contigs). On the other hand, *Haloquadratum* and *Halorubrum* composed large proportions of the Haloferacales annotations, especially for the combined Santa Pola dataset (14.5% and 25.8% for Newbler contigs; 23.5% and 26.6% for Velvet contigs). We also calculated the total number of hits in the search against the *cas* gene libraries (Figure S6) and profiled the taxonomic affiliations of all *cas* gene alignments to the contigs (Figure S7). There were smaller *Haloarcula*, *Haloferax*, *Halorubrum*, and *Haloquadratum* proportions in the total hit profiles (Figure S7) compared with the best hit profiles (Figure 10B).

Discussion

Our studies have not only confirmed prior understandings of saltern taxonomic and functional diversity but also provided new insights into virus-host interactions in many salterns worldwide. We have built upon prior metagenomics studies to show that

482 *Haloquadratum* is the majority taxon in many salterns (Oh et al., 2009), particularly those
483 that are environmentally stable with high salinity. The Santa Pola and Isla Cristina
484 metagenomes have previously been taxonomically profiled through mapping of reads to
485 16S rRNAs obtained from the Ribosomal Database Project (Fernandez et al., 2013;
486 Fernández et al., 2014a; Ventosa et al., 2014). These 16S rRNA profiling studies revealed
487 that the lowest salinity saltern (SS13) contained a community of bacterial Actinobacteria
488 and Proteobacteria and archaeal Euryarchaeota, while members of the Euryarchaeota
489 increased in abundance with increasing salinity (Fernández et al., 2014b,c; Ventosa et
490 al., 2015). Compared with the Santa Pola salterns, Cahuil and Isla Cristina salterns may
491 be more dynamic, with more rapidly changing daily and yearly temperatures and higher
492 rainfall (Fernández et al., 2014b). This may lead to lower concentrations of potassium,
493 magnesium, and sulfate in those salterns, thus leading to lower abundances of
494 *Haloquadratum*, and higher abundances of other haloarchaea (Podell et al., 2014;
495 Fernández et al., 2014b).

496 Our functional studies found rather similar levels of glycerol metabolism genes
497 across high salinity salterns. Prior studies of the Santa Pola and Isla Cristina salterns
498 found that glycerol metabolism genes, such as dihydroxyacetone kinase, glycerol kinase,
499 and NAD^+-dependent glycerol-3-phosphate dehydrogenase, increased in abundance
500 with salt concentrations (Fernández et al., 2014b,c). Taxonomic affiliations, however,
501 differed amongst salterns and different gene families. Dihydroxyacetone kinase most
502 strongly associated with *Haloquadratum*, except in the IC21 and Cahuil (C34) salterns,
503 where it associated strongly with *Halorubrum* instead. NAD^+-dependent glycerol-3-
504 phosphate dehydrogenase, on the other hand, associated with *Salinibacter*. Based on

prior understanding of saltern glycerol cycling (Oren, 1993, 1994; Sher et al., 2004), these results together suggest archaea such as *Haloquadratum* and *Halorubrum* consume the glycerol metabolism intermediate dihydroxyacetone in these ecosystems, while bacteria such as *Salinibacter* may first phosphorylate and then oxidize glycerol to metabolize it (Sher et al., 2004).

Our CRISPR detection results showed a smaller number of CRISPR spacers found in salterns where *Haloquadratum* was identified as the majority taxon compared with salterns where Halobacteriaceae was found to be the majority taxon. Furthermore, comparison of detected metagenomic spacers to those found in the genomes of two different *H. walsbyi* strains (C23 and DSM 16790) suggests variation in CRISPR array length in the salterns. Strain C23 contains a full set of *cas* genes and several CRISPR arrays, while strain DSM 16790 lost all its *cas* genes and all but five spacers (Dyall-Smith et al., 2011). While we could not identify any C23 spacers amongst our metagenomes, we could find at least three DSM 16790 spacers in several metagenomes. A successful mobile element integration event in spite of CRISPR defense is thought to have promoted loss of spacers in *Haloquadratum* genomes (Dyall-Smith et al., 2011). Taken together, this evidence suggests genomic CRISPR arrays in *Haloquadratum* species may vary in length.

Our CRISPR mapping results not only showed consistent viral interactions with *Haloquadratum* across salterns but also numerous interactions with less abundant taxa such as *Haloferax* and *Haloarcula*. As expected from our comparison of taxonomic and functional profiles, there were consistent sets of interactions amongst SS19, SS33, and SS37 salterns, where *Haloquadratum* species form the majority of the microbial

528 population. These interactions suggest similar viral infection patterns across these
529 environments, consistent with “global pond” theories of microbial and viral exchange
530 (Dyall-Smith et al., 2011; Atanasova et al., 2012). Unclear biological requirements for
531 CRISPR spacer incorporation make it difficult to make any conclusion about the
532 abundances of the viruses. Though some models suggest viruses must have reached a
533 critical number within their hosts to induce CRISPR spacer incorporation into host
534 genomes (Garrett, Vestergaard & Shah, 2011), characterization of the *Haloferax volcanii*
535 CRISPR systems suggests plasmid copy number has little effect on incorporation (Maier
536 et al., 2013).

537 There were also interactions detectable with Crass that could not map to any
538 known hosts. Some of these hosts may be poorly characterized members of the
539 Nanohaloarchaea, as recent studies have suggested for eHP-12 and eHP-16 (Martínez-
540 García et al., 2014). Surprisingly, we did not detect any spacers or virus-host interactions
541 with *Halorubrum* with either CRISPR detection method. Either limitations of our detection
542 tools or a lack of *Halorubrum* CRISPRs could account for this observation.

543 Taken together, our results suggest common sets of virus-host interactions across
544 different salterns and variation in CRISPR antiviral defenses amongst saltern microbes.
545 While common sets of viruses infect haloarchaea across salterns, these hosts may have
546 strikingly different levels of CRISPR defenses. *Haloquadratum*, it seems, may exist in
547 either virus-resistant or virus-sensitive strains depending on the presence or absence of
548 its CRISPR/Cas system. CRISPR/Cas systems may be part of the previously examined
549 *Haloquadratum* accessory gene pool (Legault et al., 2006). While it is surprising that virus-
550 sensitive strains would be present in such a virus-rich environment, even these forms

551 may have some innate halovirus resistance through extensive genomic variation of phage
552 recognition proteins (Cuadros-Orellana et al., 2007). These patterns may not apply to
553 other haloarchaea and saltern inhabitants, however. While *Haloquadratum* encodes
554 CRISPRs chromosomally, *Haloarcula* and *Haloferax* encode CRISPR elements on both
555 their chromosomes and plasmids, suggesting the possibility of horizontally transferred
556 CRISPR systems (Baliga et al., 2004; Grissa, Vergnaud & Pourcel, 2007; Hartman et al.,
557 2010; Liu et al., 2011b; Han et al., 2012). Viral predation of *Salinibacter* remains far more
558 elusive, as this bacterium lacks a CRISPR system (Grissa, Vergnaud & Pourcel, 2007).
559 Nonetheless, the variation of CRISPR systems and resulting variation in
560 *Haloquadratum* phage resistance could carefully control its predation. Together both
561 haloviral predation and CRISPR defense could tightly regulate the population of
562 *Haloquadratum*, which is both the major component of many salterns and consumer of
563 the glycerol metabolism intermediate dihydroxyacetone. Levels of CRISPR defense may
564 thus account for the turnover of this organism. Future studies should further analyze the
565 number and genomic arrangement of spacers in different *Haloquadratum* strains and
566 examine how abundances of these populations change over time. Such work would build
567 upon recent studies of the temporal dynamics of virus-host interactions in Lake Tyrrell,
568 an Australian hypersaline ecosystem (Emerson et al., 2013). Our studies and such future
569 projects will help us better understand how viruses regulate halophile populations across
570 the planet.

571 **Acknowledgments**

572 Thanks to Michael Crowder and Gary Lorigan (Miami University) for feedback on the
573 project and manuscript.

References

Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Anderson RE., Brazelton WJ., Baross JA. 2011. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage: CRISPR spacers reveal hosts of marine vent viral assemblage. *FEMS Microbiology Ecology* 77:120–133. DOI: 10.1111/j.1574-6941.2011.01090.x.

Atanasova NS., Roine E., Oren A., Bamford DH., Oksanen HM. 2012. Global network of specific virus-host interactions in hypersaline environments: Haloviruses are promiscuous. *Environmental Microbiology* 14:426–440. DOI: 10.1111/j.1462-2920.2011.02603.x.

Baliga NS., Bonneau R., Facciotti MT., Pan M., Glusman G., Deutsch EW., Shannon P., Chiu Y., Weng RS., Gan RR., Hung P., Date SV., Marcotte E., Hood L., Ng WV. 2004. Genome sequence of Haloarcula marismortui: A halophilic archaeon from the Dead Sea. *Genome Research* 14:2221–2234. DOI: 10.1101/gr.2700304.

Bardavid RE., Khristo P., Oren A. 2006. Interrelationships between Dunaliella and halophilic prokaryotes in saltern crystallizer ponds. *Extremophiles* 12:5–14. DOI: 10.1007/s00792-006-0053-y.

Bardavid RE., Oren A. 2007. Dihydroxyacetone metabolism in Salinibacter ruber and in Haloquadratum walsbyi. *Extremophiles* 12:125–131. DOI: 10.1007/s00792-007-0114-x.

596 Berg Miller ME., Yeoman CJ., Chia N., Tringe SG., Angly FE., Edwards RA., Flint HJ.,
597 Lamed R., Bayer EA., White BA. 2012. Phage-bacteria relationships and
598 CRISPR elements revealed by a metagenomic survey of the rumen microbiome:
599 Rumen viral metagenome. *Environmental Microbiology* 14:207–227. DOI:
600 10.1111/j.1462-2920.2011.02593.x.
601 Bhatia S., Densmore D. 2013. Pigeon: A Design Visualizer for Synthetic Biology. *ACS*
602 *Synthetic Biology* 2:348–350. DOI: 10.1021/sb400024s.
603 Bolhuis H., Palm P., Wende A., Falb M., Rampp M., Rodriguez-Valera F., Pfeiffer F.,
604 Oesterhelt D. 2006. The genome of the square archaeon Haloquadratum
605 walsbyi : life at the limits of water activity. *BMC Genomics* 7:169. DOI:
606 10.1186/1471-2164-7-169.
607 Chitlaru E., Pick U. 1991. Regulation of Glycerol Synthesis in Response to Osmotic
608 Changes in Dunaliella. *Plant Physiology* 96:50–60. DOI: 10.1104/pp.96.1.50.
609 Cuadros-Orellana S., Martin-Cuadrado A-B., Legault B., D'Auria G., Zhaxybayeva O.,
610 Papke RT., Rodriguez-Valera F. 2007. Genomic plasticity in prokaryotes: the
611 case of the square haloarchaeon. *The ISME Journal* 1:235–245. DOI:
612 10.1038/ismej.2007.35.
613 Dyall-Smith ML., Pfeiffer F., Klee K., Palm P., Gross K., Schuster SC., Rampp M.,
614 Oesterhelt D. 2011. Haloquadratum walsbyi : Limited Diversity in a Global Pond.
615 *PLoS ONE* 6:e20968. DOI: 10.1371/journal.pone.0020968.
616 Edwards RA., McNair K., Faust K., Raes J., Dutilh BE. 2015. Computational
617 approaches to predict bacteriophage–host relationships. *FEMS Microbiology*
618 *Reviews*:fuv048. DOI: 10.1093/femsre/fuv048.

619 Emerson JB., Andrade K., Thomas BC., Norman A., Allen EE., Heidelberg KB., Banfield
620 JF. 2013. Virus-Host and CRISPR Dynamics in Archaea-Dominated Hypersaline
621 Lake Tyrrell, Victoria, Australia. *Archaea* 2013:1–12. DOI: 10.1155/2013/370871.
622 Fernandez AB., Ghai R., Martin-Cuadrado AB., Sanchez-Porro C., Rodriguez-Valera F.,
623 Ventosa A. 2013. Metagenome Sequencing of Prokaryotic Microbiota from Two
624 Hypersaline Ponds of a Marine Saltern in Santa Pola, Spain. *Genome*
625 *Announcements* 1:e00933-13. DOI: 10.1128/genomeA.00933-13.
626 Fernández AB., León MJ., Vera B., Sánchez-Porro C., Ventosa A. 2014a. Metagenomic
627 Sequence of Prokaryotic Microbiota from an Intermediate-Salinity Pond of a
628 Saltern in Isla Cristina, Spain. *Genome Announcements* 2:e00045-14. DOI:
629 10.1128/genomeA.00045-14.
630 Fernández AB., Vera-Gargallo B., Sánchez-Porro C., Ghai R., Papke RT., Rodriguez-
631 Valera F., Ventosa A. 2014b. Comparison of prokaryotic community structure
632 from Mediterranean and Atlantic saltern concentrator ponds by a metagenomic
633 approach. *Frontiers in Microbiology* 5. DOI: 10.3389/fmicb.2014.00196.
634 Fernández AB., Ghai R., Martin-Cuadrado A-B., Sánchez-Porro C., Rodriguez-Valera
635 F., Ventosa A. 2014c. Prokaryotic taxonomic and metabolic diversity of an
636 intermediate salinity hypersaline habitat assessed by metagenomics. *FEMS*
637 *Microbiology Ecology* 88:623–635. DOI: 10.1111/1574-6941.12329.
638 Garcia-Heredia I., Martin-Cuadrado A-B., Mojica FJM., Santos F., Mira A., Antón J.,
639 Rodriguez-Valera F. 2012. Reconstructing Viral Genomes from the Environment
640 Using Fosmid Clones: The Case of Haloviruses. *PLoS ONE* 7:e33802. DOI:
641 10.1371/journal.pone.0033802.

642 Garrett RA., Vestergaard G., Shah SA. 2011. Archaeal CRISPR-based immune
643 systems: exchangeable functional modules. *Trends in Microbiology* 19:549–556.
644 DOI: 10.1016/j.tim.2011.08.002.
645 Ghai R., Pašić L., Fernández AB., Martin-Cuadrado A-B., Mizuno CM., McMahon KD.,
646 Papke RT., Stepanauskas R., Rodriguez-Brito B., Rohwer F., Sánchez-Porro C.,
647 Ventosa A., Rodríguez-Valera F. 2011. New Abundant Microbial Groups in
648 Aquatic Hypersaline Environments. *Scientific Reports* 1. DOI:
649 10.1038/srep00135.
650 Grissa I., Vergnaud G., Pourcel C. 2007. The CRISPRdb database and tools to display
651 CRISPRs and to generate dictionaries of spacers and repeats. *BMC*
652 *Bioinformatics* 8:172. DOI: 10.1186/1471-2105-8-172.
653 Gupta RS., Naushad S., Baker S. 2015. Phylogenomic analyses and molecular
654 signatures for the class Halobacteria and its two major clades: a proposal for
655 division of the class Halobacteria into an emended order Halobacteriales and two
656 new orders, Haloferacales ord. nov. and Natrialbales ord. nov., containing the
657 novel families Haloferacaceae fam. nov. and Natrialbaceae fam. nov.
658 *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY*
659 *MICROBIOLOGY* 65:1050–1069. DOI: 10.1099/ijs.0.070136-0.
660 Han J., Zhang F., Hou J., Liu X., Li M., Liu H., Cai L., Zhang B., Chen Y., Zhou J., Hu
661 S., Xiang H. 2012. Complete Genome Sequence of the Metabolically Versatile
662 Halophilic Archaeon Haloferax mediterranei, a Poly(3-Hydroxybutyrate-co-3-
663 Hydroxyvalerate) Producer. *Journal of Bacteriology* 194:4463–4464. DOI:
664 10.1128/JB.00880-12.

665 Hartman AL., Norais C., Badger JH., Delmas S., Haldenby S., Madupu R., Robinson J.,
666 Khouri H., Ren Q., Lowe TM., Maupin-Furlow J., Pohlschroder M., Daniels C.,
667 Pfeiffer F., Allers T., Eisen JA. 2010. The Complete Genome Sequence of
668 Haloferax volcanii DS2, a Model Archaeon. *PLoS ONE* 5:e9605. DOI:
669 10.1371/journal.pone.0009605.
670 Heidelberg KB., Nelson WC., Holm JB., Eisenkolb N., Andrade K., Emerson JB. 2013.
671 Characterization of eukaryotic microbial diversity in hypersaline Lake Tyrrell,
672 Australia. *Frontiers in Microbiology* 4. DOI: 10.3389/fmicb.2013.00115.
673 Hurwitz B. 2014. iMicrobe: Advancing Clinical and Environmental Microbial Research
674 using the iPlant Cyberinfrastructure. In: Plant and Animal Genome,.
675 Legault BA., Lopez-Lopez A., Alba-Casado JC., Doolittle WF., Bolhuis H., Rodriguez-
676 Valera F., Papke RT. 2006. Environmental genomics of “Haloquadratum walsbyi”
677 in a saltern crystallizer indicates a large pool of accessory genes in an otherwise
678 coherent species. *BMC Genomics* 7:1–13. DOI: 10.1186/1471-2164-7-171.
679 Li W., Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
680 protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. DOI:
681 10.1093/bioinformatics/btl158.
682 Liu B., Gibbons T., Ghodsi M., Treangen T., Pop M. 2011a. Accurate and fast
683 estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC*
684 *Genomics* 12:S4. DOI: 10.1186/1471-2164-12-S2-S4.
685 Liu H., Wu Z., Li M., Zhang F., Zheng H., Han J., Liu J., Zhou J., Wang S., Xiang H.
686 2011b. Complete Genome Sequence of Haloarcula hispanica, a Model

687 Haloarchaeon for Studying Genetics, Metabolism, and Virus-Host Interaction.
688 *Journal of Bacteriology* 193:6086–6087. DOI: 10.1128/JB.05953-11.
689 Maier L-K., Lange SJ., Stoll B., Haas KA., Fischer SM., Fischer E., Duchardt-Ferner E.,
690 Wöhnert J., Backofen R., Marchfelder A. 2013. Essential requirements for the
691 detection and degradation of invaders by the Haloferax volcanii CRISPR/Cas
692 system I-B. *RNA Biology* 10:865–874. DOI: 10.4161/rna.24282.
693 Martínez-García M., Santos F., Moreno-Paz M., Parro V., Antón J. 2014. Unveiling
694 viral–host interactions within the “microbial dark matter.” *Nature Communications*
695 5:4542. DOI: 10.1038/ncomms5542.
696 Moller AG., Liang C. 2016. MetaCRAST: Reference-guided extraction of CRISPR
697 spacers from unassembled metagenomes. *PeerJ Preprints* 4:e2278v2. DOI:
698 10.7287/peerj.preprints.2278v2.
699 Nayfach S., Bradley PH., Wyman SK., Laurent TJ., Williams A., Eisen JA., Pollard KS.,
700 Sharpton TJ. 2015. Automated and Accurate Estimation of Gene Family
701 Abundance from Shotgun Metagenomes. *PLoS Comput Biol* 11:e1004573. DOI:
702 10.1371/journal.pcbi.1004573.
703 Oh D., Porter K., Russ B., Burns D., Dyall-Smith M. 2009. Diversity of Haloquadratum
704 and other haloarchaea in three, geographically distant, Australian saltern
705 crystallizer ponds. *Extremophiles* 14:161–169. DOI: 10.1007/s00792-009-0295-6.
706 Oren A. 1993. Availability, uptake and turnover of glycerol in hypersaline environments.
707 *FEMS Microbiology Ecology* 12:15–23. DOI: 10.1111/j.1574-
708 6941.1993.tb00012.x.

709 Oren A. 1994. The ecology of the extremely halophilic archaea. *FEMS Microbiology*
710 *Reviews* 13:415–439. DOI: 10.1111/j.1574-6976.1994.tb00060.x.
711 Ouellette M., Makkay AM., Papke RT. 2013. Dihydroxyacetone metabolism in Haloferax
712 volcanii. *Frontiers in Microbiology* 4. DOI: 10.3389/fmicb.2013.00376.
713 Park JS., Simpson AGB. 2015. Diversity of Heterotrophic Protists from Extremely
714 Hypersaline Habitats. *Protist* 166:422–437. DOI: 10.1016/j.protis.2015.06.001.
715 Petrenko P., Lobb B., Kurtz DA., Neufeld JD., Doxey AC. 2015. MetAnnotate: function-
716 specific taxonomic profiling and comparison of metagenomes. *BMC Biology*
717 13:92. DOI: 10.1186/s12915-015-0195-4.
718 Plominsky AM., Delherbe N., Ugalde JA., Allen EE., Blanchet M., Ikeda P., Santibanez
719 F., Hanselmann K., Ulloa O., De la Iglesia R., von Dassow P., Astorga M.,
720 Galvez MJ., Gonzalez ML., Henriquez-Castillo C., Vaulot D., Lopes do Santos A.,
721 van den Engh G., Gimpel C., Bertoglio F., Delgado Y., Docmac F., Elizondo-
722 Patrone C., Narvaez S., Sorroche F., Rojas-Herrera M., Trefault N. 2014.
723 Metagenome Sequencing of the Microbial Community of a Solar Saltern
724 Crystallizer Pond at Cahuil Lagoon, Chile. *Genome Announcements* 2:e01172-
725 14-e01172-14. DOI: 10.1128/genomeA.01172-14.
726 Podell S., Emerson JB., Jones CM., Ugalde JA., Welch S., Heidelberg KB., Banfield
727 JF., Allen EE. 2014. Seasonal fluctuations in ionic concentrations drive microbial
728 succession in a hypersaline lake community. *The ISME Journal* 8:979–990. DOI:
729 10.1038/ismej.2013.221.
730 Rawls KS., Martin JH., Maupin-Furlow JA. 2011. Activity and Transcriptional Regulation
731 of Bacterial Protein-Like Glycerol-3-Phosphate Dehydrogenase of the

732 Haloarchaea in Haloferax volcanii. *Journal of Bacteriology* 193:4469–4476. DOI:
733 10.1128/JB.00276-11.
734 Rodriguez-Brito B., Li L., Wegley L., Furlan M., Angly F., Breitbart M., Buchanan J.,
735 Desnues C., Dinsdale E., Edwards R., Felts B., Haynes M., Liu H., Lipson D.,
736 Mahaffy J., Martin-Cuadrado AB., Mira A., Nulton J., Pašić L., Rayhawk S.,
737 Rodriguez-Mueller J., Rodriguez-Valera F., Salamon P., Srinagesh S., Thingstad
738 TF., Tran T., Thurber RV., Willner D., Youle M., Rohwer F. 2010. Viral and
739 microbial community dynamics in four aquatic environments. *The ISME Journal*
740 4:739–751. DOI: 10.1038/ismej.2010.1.
741 Rohwer F., Prangishvili D., Lindell D. 2009. Roles of viruses in the environment.
742 *Environmental Microbiology* 11:2771–2774. DOI: 10.1111/j.1462-
743 2920.2009.02101.x.
744 Sanguino L., Franqueville L., Vogel TM., Larose C. 2015. Linking environmental
745 prokaryotic viruses and their host through CRISPRs. *FEMS Microbiology Ecology*
746 91:fiv046. DOI: 10.1093/femsec/fiv046.
747 Santos F., Yarza P., Parro V., Briones C., Antón J. 2010. The metavirome of a
748 hypersaline environment: The metavirome of a hypersaline environment.
749 *Environmental Microbiology* 12:2965–2976. DOI: 10.1111/j.1462-
750 2920.2010.02273.x.
751 Santos F., Yarza P., Parro V., Meseguer I., Rosselló-Móra R., Antón J. 2012. Culture-
752 Independent Approaches for Studying Viruses from Hypersaline Environments.
753 *Applied and Environmental Microbiology* 78:1635–1643. DOI:
754 10.1128/AEM.07175-11.

755 Shannon P., Markiel A., Ozier O., Baliga NS., Wang JT., Ramage D., Amin N.,
756 Schwikowski B., Ideker T. 2003. Cytoscape: A Software Environment for
757 Integrated Models of Biomolecular Interaction Networks. *Genome Research*
758 13:2498–2504. DOI: 10.1101/gr.1239303.
759 Sher J., Elevi R., Mana L., Oren A. 2004. Glycerol metabolism in the extremely
760 halophilic bacterium Salinibacter ruber. *FEMS Microbiology Letters* 232:211–215.
761 DOI: 10.1016/S0378-1097(04)00077-1.
762 Sherwood KE., Cano DJ., Maupin-Furlow JA. 2009. Glycerol-Mediated Repression of
763 Glucose Metabolism and Glycerol Kinase as the Sole Route of Glycerol
764 Catabolism in the Haloarchaeon Haloferax volcanii. *Journal of Bacteriology*
765 191:4307–4315. DOI: 10.1128/JB.00131-09.
766 Skennerton CT., Imelfort M., Tyson GW. 2013. Crass: identification and reconstruction
767 of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*
768 41:e105–e105. DOI: 10.1093/nar/gkt183.
769 Sorek R., Kunin V., Hugenholtz P. 2008. CRISPR — a widespread system that provides
770 acquired resistance against phages in bacteria and archaea. *Nature Reviews*
771 *Microbiology* 6:181–186. DOI: 10.1038/nrmicro1793.
772 Suttle CA. 2007. Marine viruses — major players in the global ecosystem. *Nature*
773 *Reviews Microbiology* 5:801–812. DOI: 10.1038/nrmicro1750.
774 Tschitschko B., Williams TJ., Allen MA., Páez-Espino D., Kyrpides N., Zhong L., Raftery
775 MJ., Cavicchioli R. 2015. Antarctic archaea–virus interactions: metaproteome-led
776 analysis of invasion, evasion and adaptation. *The ISME Journal* 9:2094–2107.
777 DOI: 10.1038/ismej.2015.110.

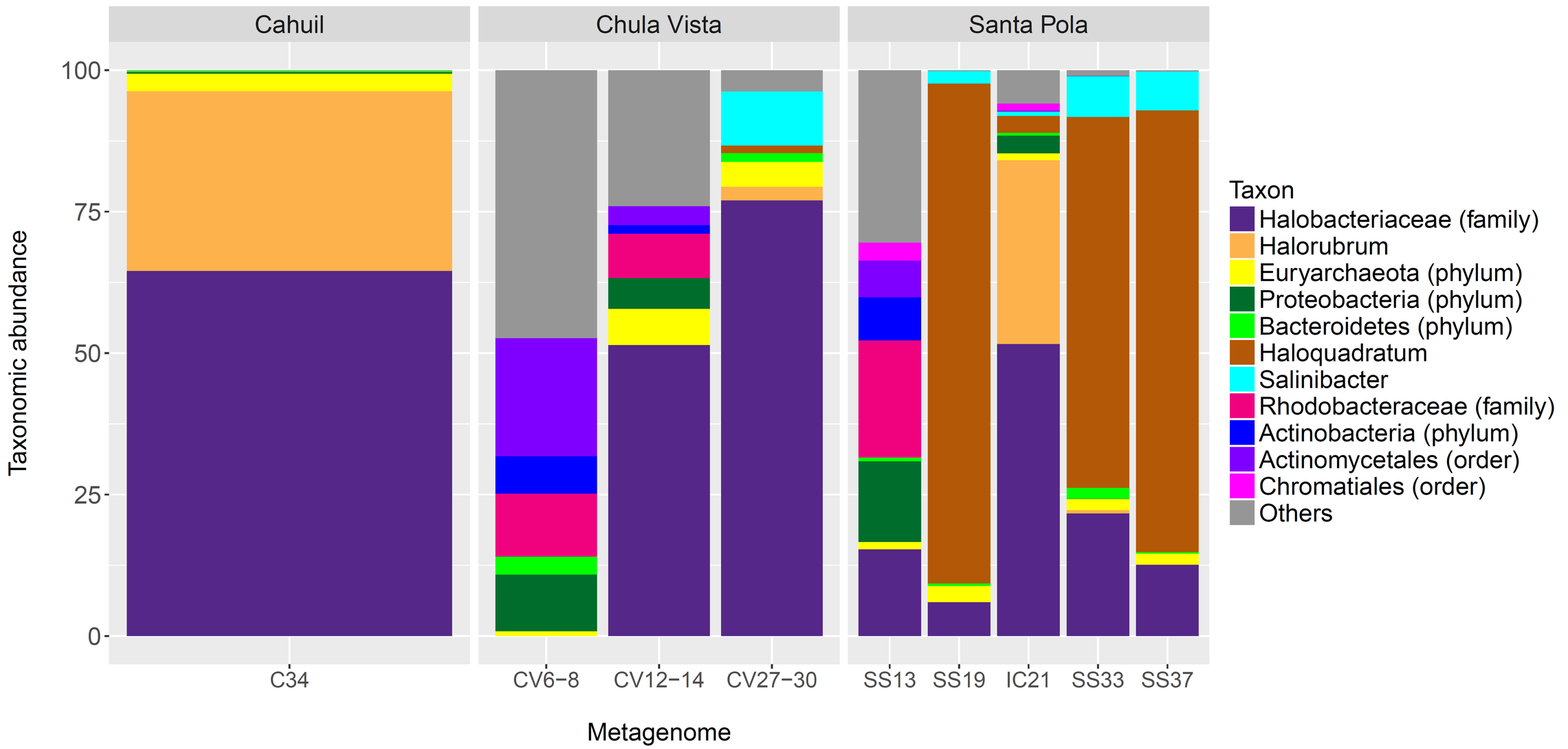
778 Tyson GW., Banfield JF. 2008. Rapidly evolving CRISPRs implicated in acquired
779 resistance of microorganisms to viruses. *Environmental Microbiology* 10:200–
780 207. DOI: 10.1111/j.1462-2920.2007.01444.x.
781 Ventosa A., Fernández AB., León MJ., Sánchez-Porro C., Rodriguez-Valera F. 2014.
782 The Santa Pola saltern as a model for studying the microbiota of hypersaline
783 environments. *Extremophiles* 18:811–824. DOI: 10.1007/s00792-014-0681-6.
784 Ventosa A., de la Haba RR., Sánchez-Porro C., Papke RT. 2015. Microbial diversity of
785 hypersaline environments: a metagenomic approach. *Current Opinion in*
786 *Microbiology* 25:80–87. DOI: 10.1016/j.mib.2015.05.002.
787 Wegmann K., Ben-Amotz A., Avron M. 1980. Effect of Temperature on Glycerol
788 Retention in the Halotolerant Algae Dunaliella and Asteromonas. *Plant*
789 *Physiology* 66:1196–1197. DOI: 10.1104/pp.66.6.1196.
790 Weitz JS., Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities
791 and biogeochemical cycles. *F1000 Biology Reports* 4. DOI: 10.3410/B4-17.
792 Wu S., Manber U., Myers E. 1995. A Subquadratic Algorithm for Approximate Regular
793 Expression Matching. *Journal of Algorithms* 19:346–360. DOI:
794 10.1006/jagm.1995.1041.
795 Zerbino DR., Birney E. 2008. Velvet: Algorithms for de novo short read assembly using
796 de Bruijn graphs. *Genome Research* 18:821–829. DOI: 10.1101/gr.074492.107.
797 Zhang Q., Rho M., Tang H., Doak TG., Ye Y. 2013. CRISPR-Cas systems target a
798 diverse collection of invasive mobile genetic elements in human microbiomes.
799 *Genome Biology* 14:R40. DOI: 10.1186/gb-2013-14-4-r40.
800

Figure 1(on next page)

Taxonomic profiles (A) and principal component analysis (B) of selected saltern metagenomes.

Metagenomes were profiled using MetaPhyler and resulting profiles clustered using MATLAB. Profiles show the relative abundances (in percent) of 11 different taxa identified with MetaPhyler (A). Profiles are grouped by geographic locations – Cahuil, Chula Vista, and Santa Pola (A). The first two principal components, which together account for 88% of variation in the profiles, are shown in the PCA results (B). Three different groups of similar profiles were identified with PCA (blue – CV6-8, SS13; orange – CV12-14, IC21, C34, CV27-30; red – SS33, SS37, SS19). These groups were found to be statistically significant in their between-group differences (ADONIS; F=19.817; p=0.002).

A


Cahuil
Chula Vista
Santa Pola
Taxonomic abundance
100
75
50
25
0
C34
CV6−8
CV12−14
CV27−30
SS13
SS19
IC21
SS33
SS37
Metagenome
Taxon
Halobacteriaceae (family)
Halorubrum
Euryarchaeota (phylum)
Proteobacteria (phylum)
Bacteroidetes (phylum)
Haloquadratum
Salinibacter
Rhodobacteraceae (family)
Actinobacteria (phylum)
Actinomycetales (order)
Chromatiales (order)
Others

B

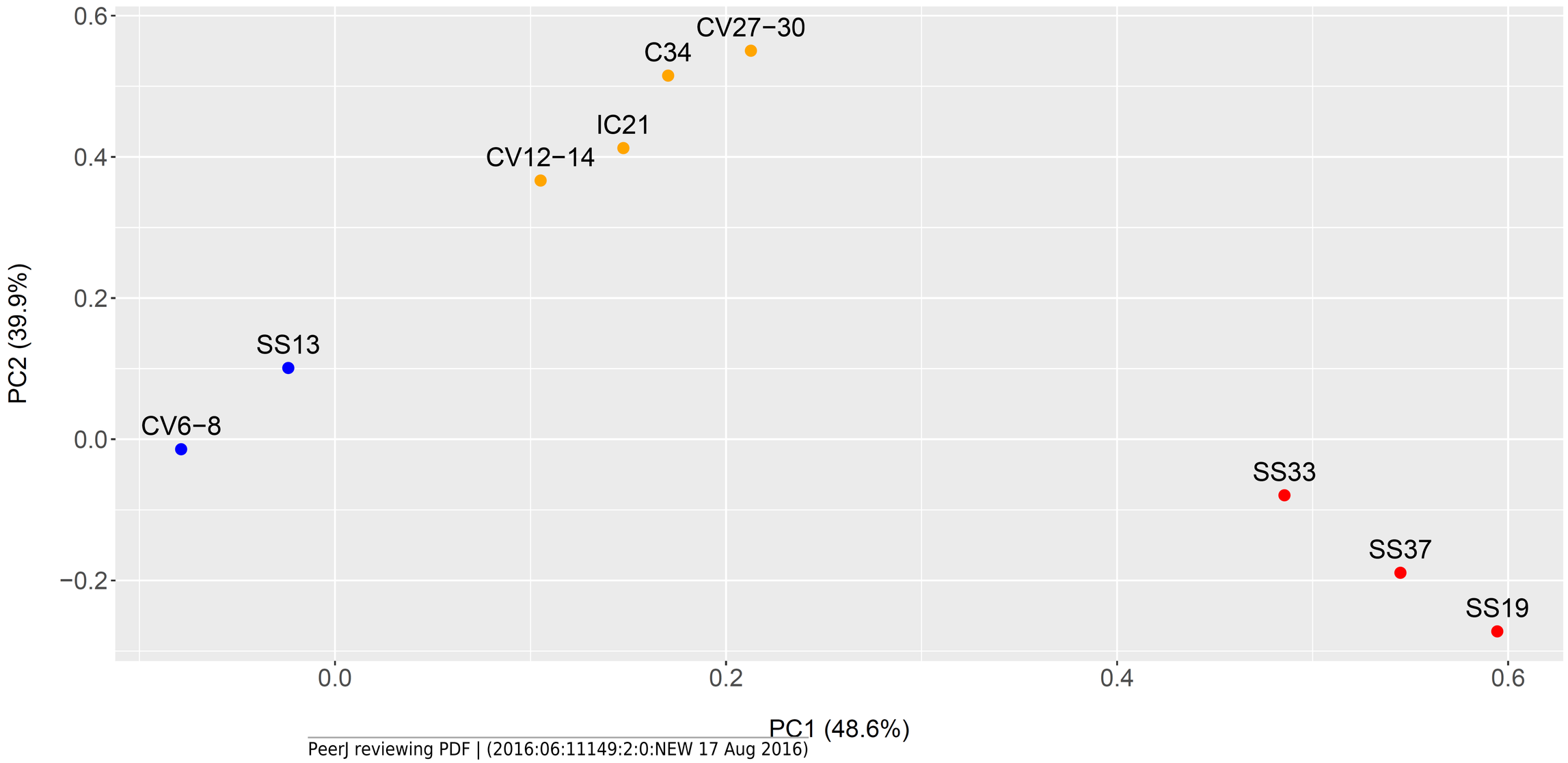

0.6
0.4
0.2
0.0
−0.2
PC2 (39.9%)
CV27−30
C34
IC21
CV12−14
SS13
CV6−8
SS33
SS37
SS19
0.0
0.2
0.4
0.6
PC1 (48.6%)

Figure 2(on next page)

Analysis of correlations between percent salinity and relative gene family abundances determined with ShotMAP.

We determined relative abundances of nine different gene families using ShotMAP (glycerol kinase, FGGY_N and FGGY_C; NAD+-dependent glycerol-3-phosphate dehydrogenase, NAD_Gly3P_dh_N and NAD_Gly3P_dh_C; glycerol dehydrogenase/iron-containing alcohol dehydrogenase, Fe-ADH and Fe-ADH_2; and dihydroxyacetone kinase, Dak1, Dak1_2, and Dak2). Relative gene family abundance (listed on the y-axis) was calculated as the sum of hits to target genes within a gene family each normalized by target gene length. Linear regression lines between relative gene family abundance and percent salinity and R^2 values were determined with Microsoft Excel.

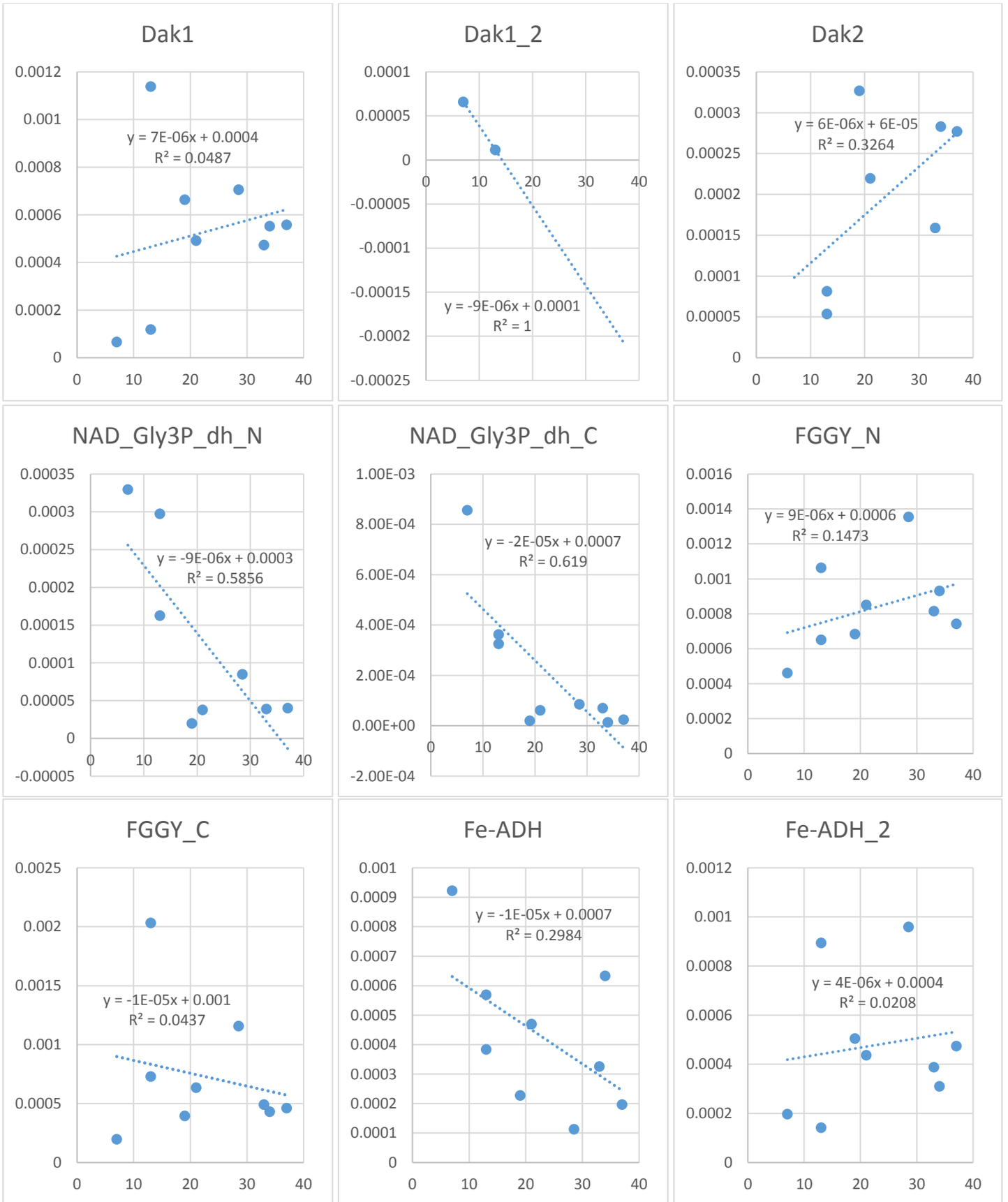

Dak1
y = 7E-06x + 0.0004
R² = 0.0487
Dak1_2
y = -9E-06x + 0.0001
R² = 1
Dak2
y = 6E-06x + 6E-05
R² = 0.3264
NAD_Gly3P_dh_N
y = -9E-06x + 0.0003
R² = 0.5856
NAD_Gly3P_dh_C
y = -2E-05x + 0.0007
R² = 0.619
FGGY_N
y = 9E-06x + 0.0006
R² = 0.1473
FGGY_C
y = -1E-05x + 0.001
R² = 0.0437
Fe-ADH
y = -1E-05x + 0.0007
R² = 0.2984
Fe-ADH_2
y = 4E-06x + 0.0004
R² = 0.0208
Relative Gene Family Abundance
Salinity

Figure 3(on next page)

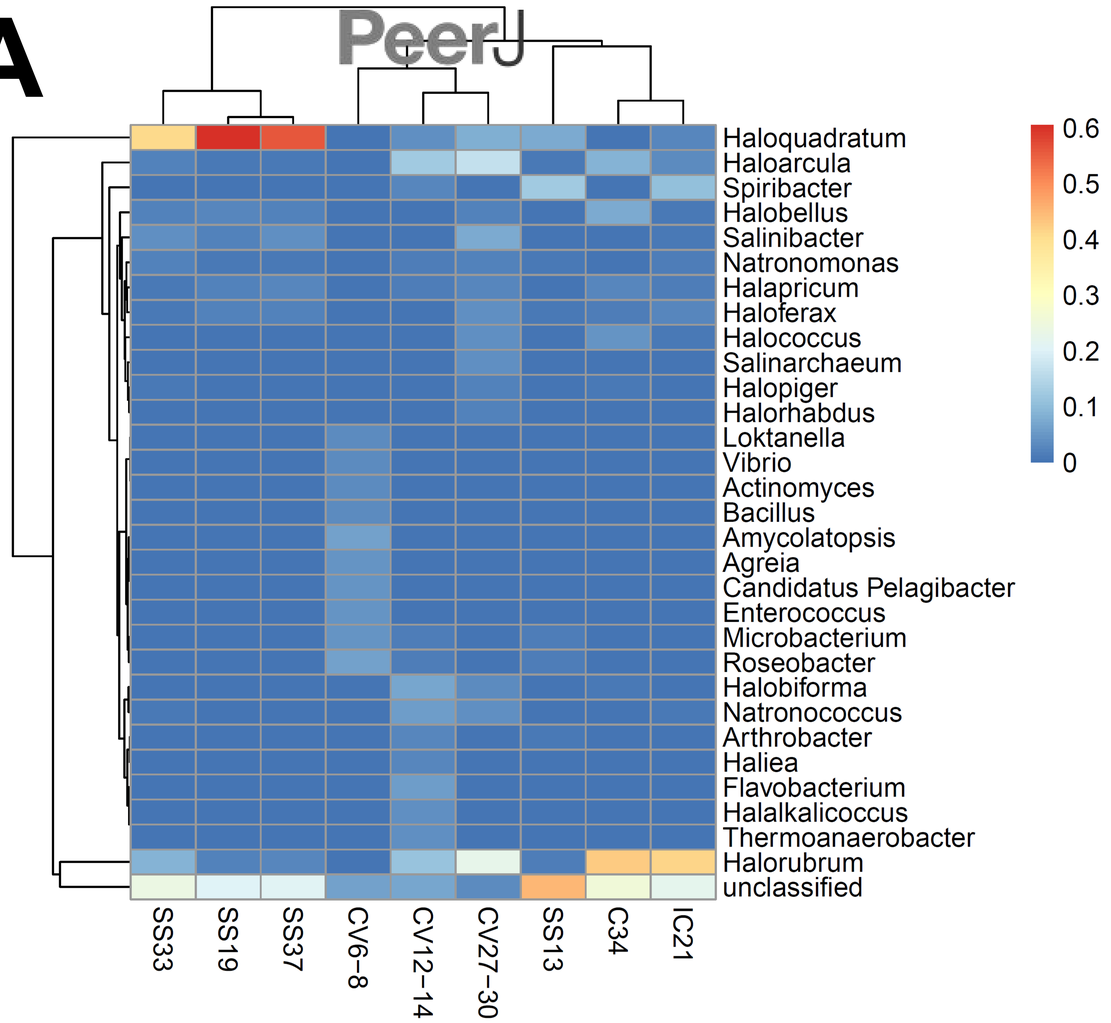
Heat map visualizations of glycerol metabolism taxonomic profiles determined with metAnnotate.

We determined taxonomic profiles of nine different gene families using metAnnotate (glycerol kinase, FGGY_N and FGGY_C; NAD+-dependent glycerol-3-phosphate dehydrogenase, NAD_Gly3P_dh_N and NAD_Gly3P_dh_C; glycerol dehydrogenase/iron-containing alcohol dehydrogenase, Fe-ADH and Fe-ADH_2; and dihydroxyacetone kinase, Dak1, Dak1_2, and Dak2). Relative gene family abundance here is represented as the fraction of all hits for a gene family mapping to a particular taxon. Abundances were hierarchically clustered by genus and site (A), by genus and gene family (B), and by genus, site, and gene family (C). The primary gene family and metagenome clusters determined (see Results) were found to be statistically significant (ADONIS; F=7.5468 and 2.8768; p=0.014 and 0.013, for gene family and metagenome clusters, respectively). The first two heat maps (A and B) were generated with the heatmap R package, while the last heat map (C) was generated with ggplot2.

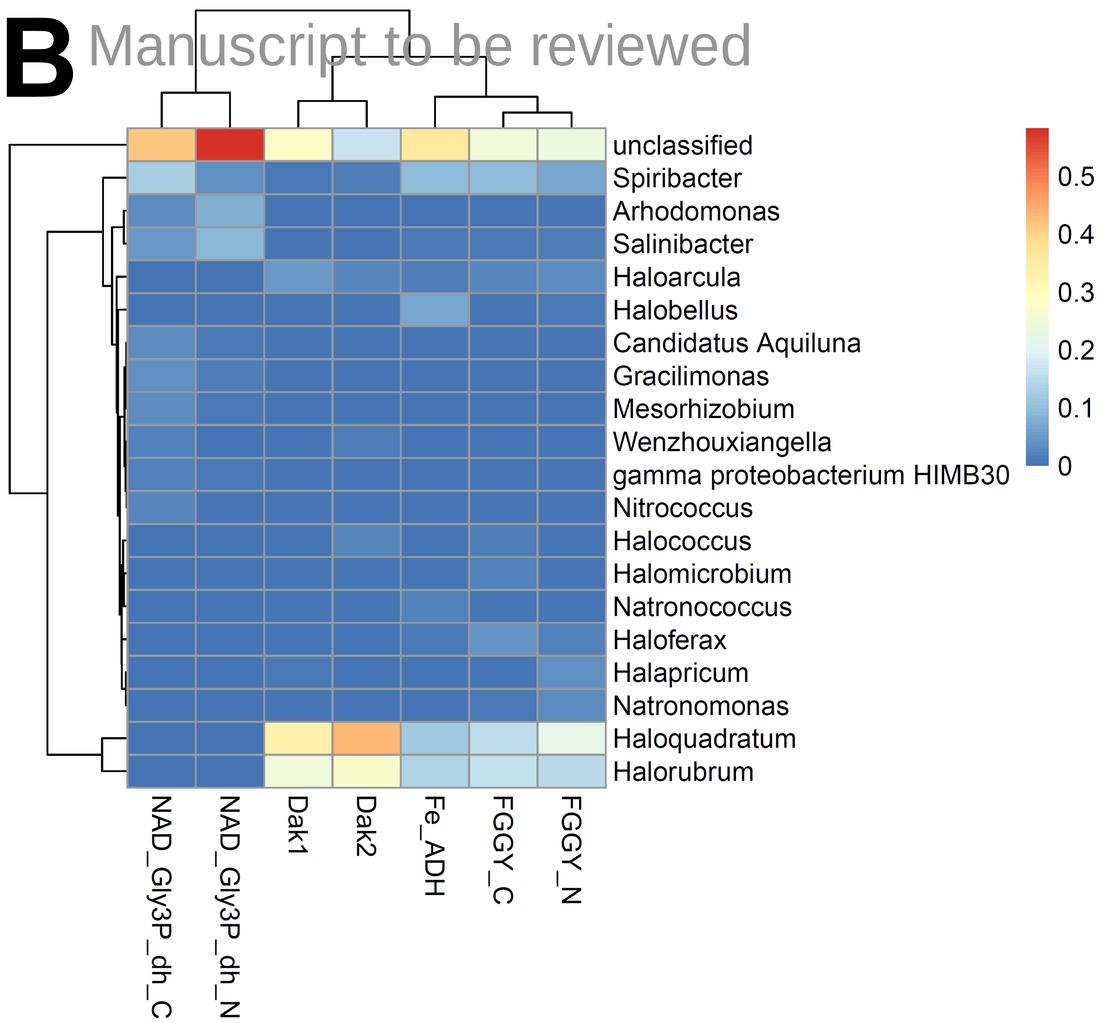
PeerJ

Manuscript to be reviewed

A



B



C

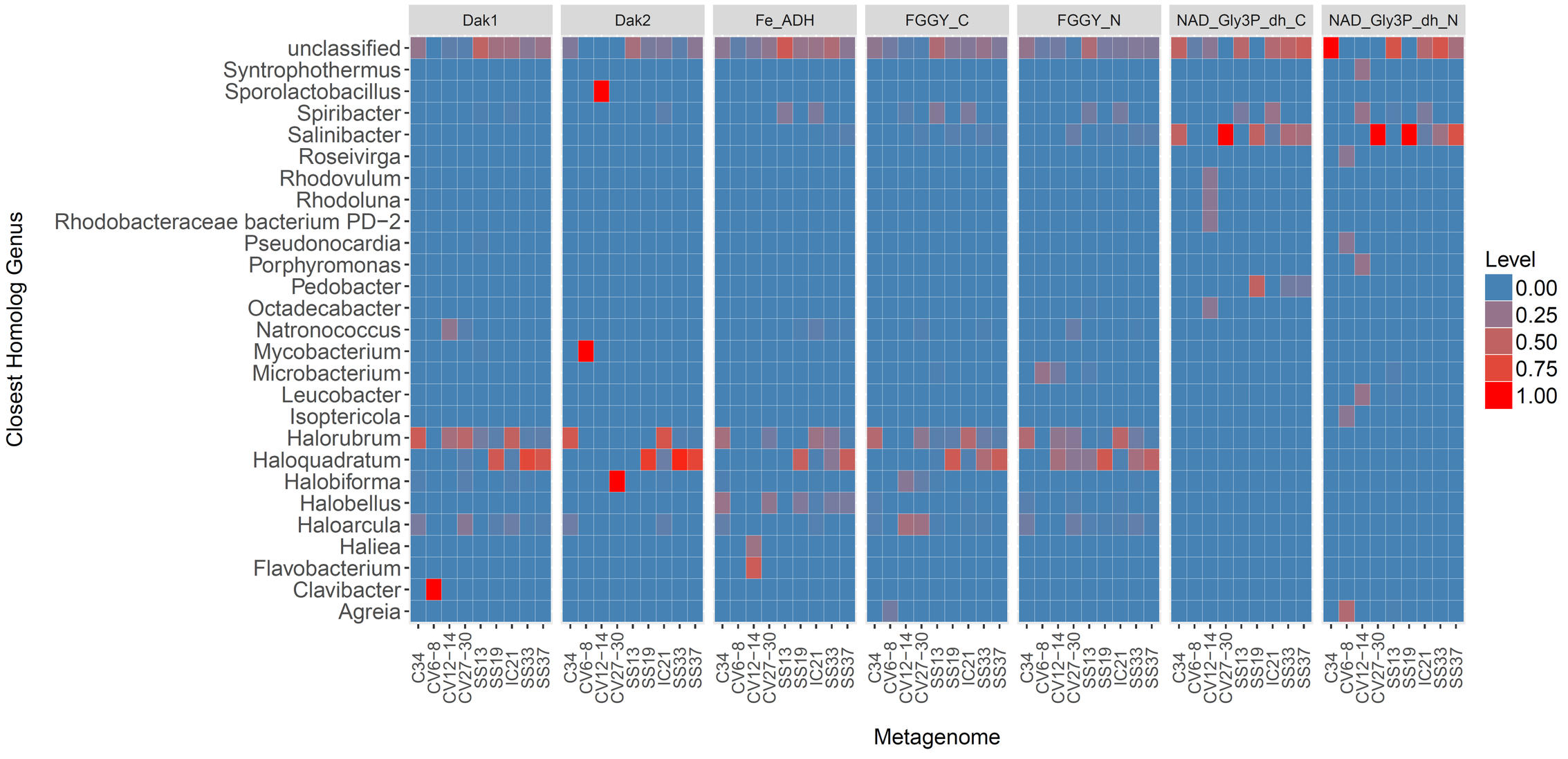


Figure 4(on next page)

Number of CRISPR spacers detected in each metagenome after clustering with CD-HIT.

Spacers were detected either with the *de novo* detection method (Crass) or the reference-guided method (MetaCRAST) with corresponding maximum edit distances (from 0 to 3) described in Materials and Methods that used a query of 29 halobacterial CRISPR direct repeat sequences. All spacer counts are reported after clustering initially detected spacers with CD-HIT (with a clustering similarity threshold of 0.9).

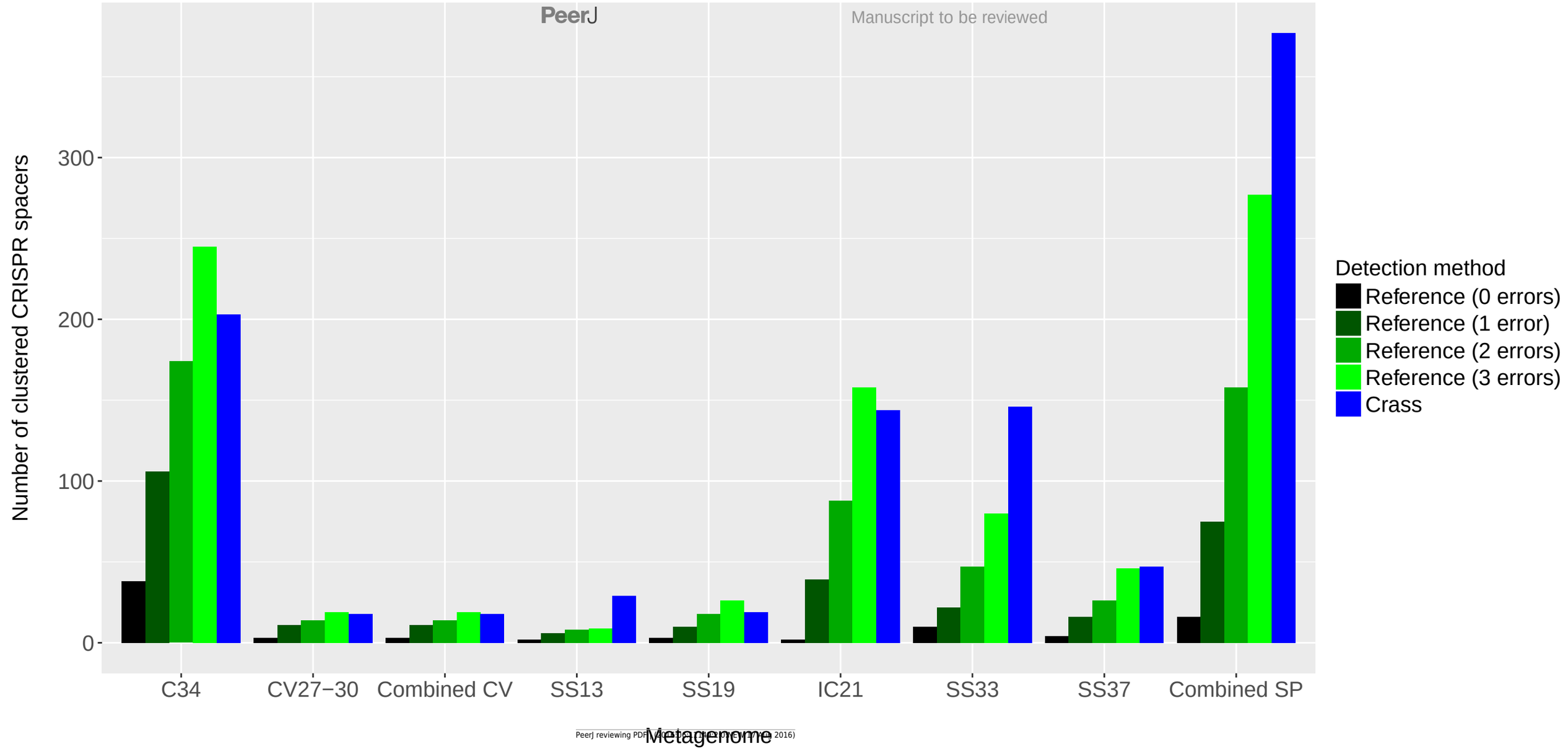
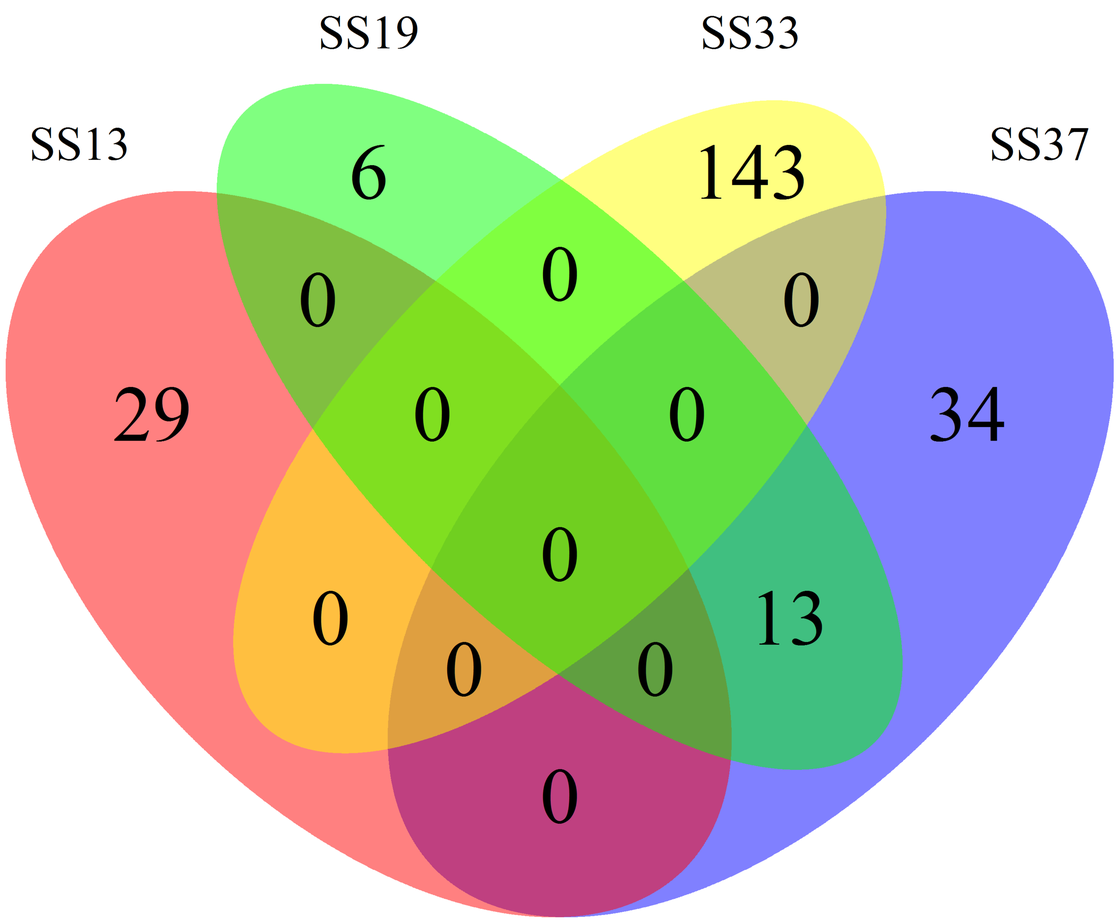

Number of clustered CRISPR spacers
0
100
200
300
C34
CV27−30
Combined CV
SS13
SS19
IC21
SS33
SS37
Combined SP
Metagenome
Detection method
Reference (0 errors)
Reference (1 error)
Reference (2 errors)
Reference (3 errors)
Crass

Figure 5(on next page)

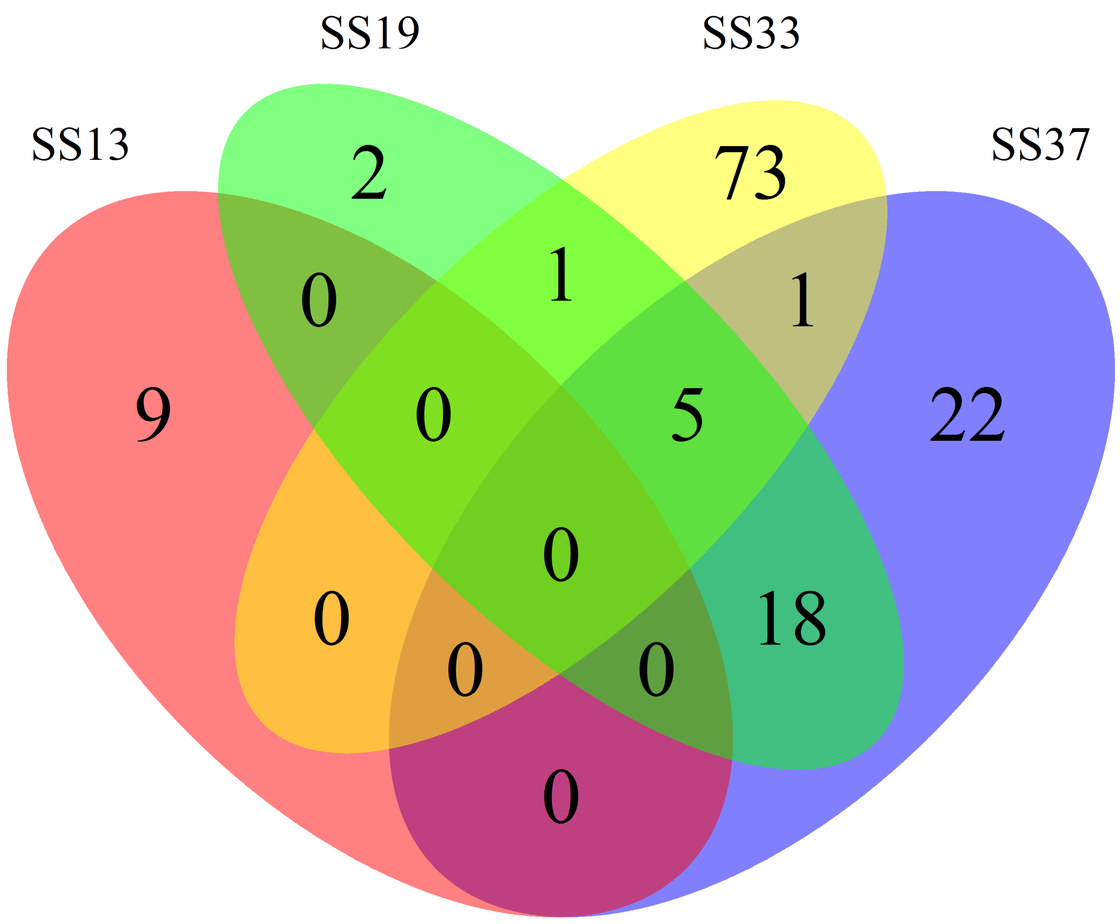
Comparison of CRISPR spacers detected with *de novo* (Crass) and reference-guided (MetaCRAST) methods.

All reference-guided spacers compared here with Crass spacers were detected with a maximum allowed edit distance of 3 (insertions, deletions, or mismatches with query direct repeat sequence). CRISPR spacers detected in SS13, SS19, SS33, and SS37 metagenomes with *de novo* (A) and reference-guided (B) methods were compared among sites to determine identical spacers among sets. Across IC21 (C) and Cahuil/C34 (D) data sets, spacers detected were compared across the two detection methods.

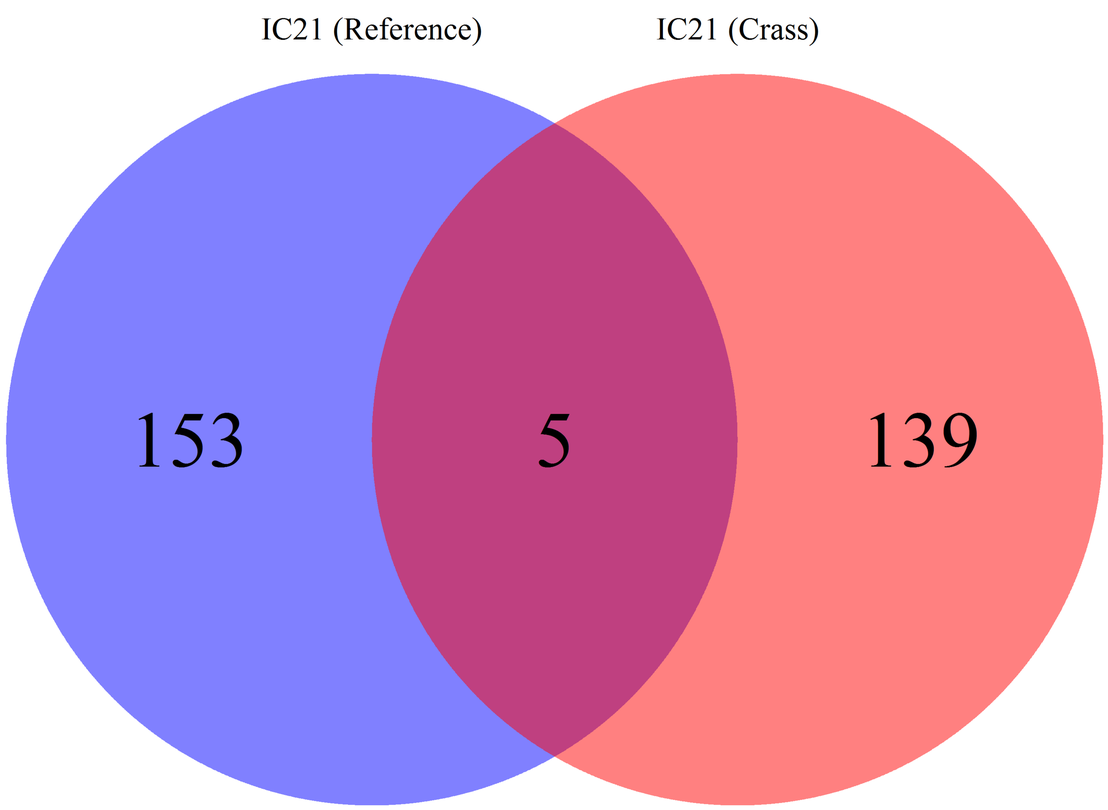
A


SS19
SS33
SS13
SS37
6
143
0
0
0
29
0
0
34
0
0
0
13
0
0

B


SS19
SS33
SS13
SS37
2
73
1
0
1
9
0
5
22
0
0
0
18
0
0

C


IC21 (Reference)
IC21 (Crass)
153
5
139

D

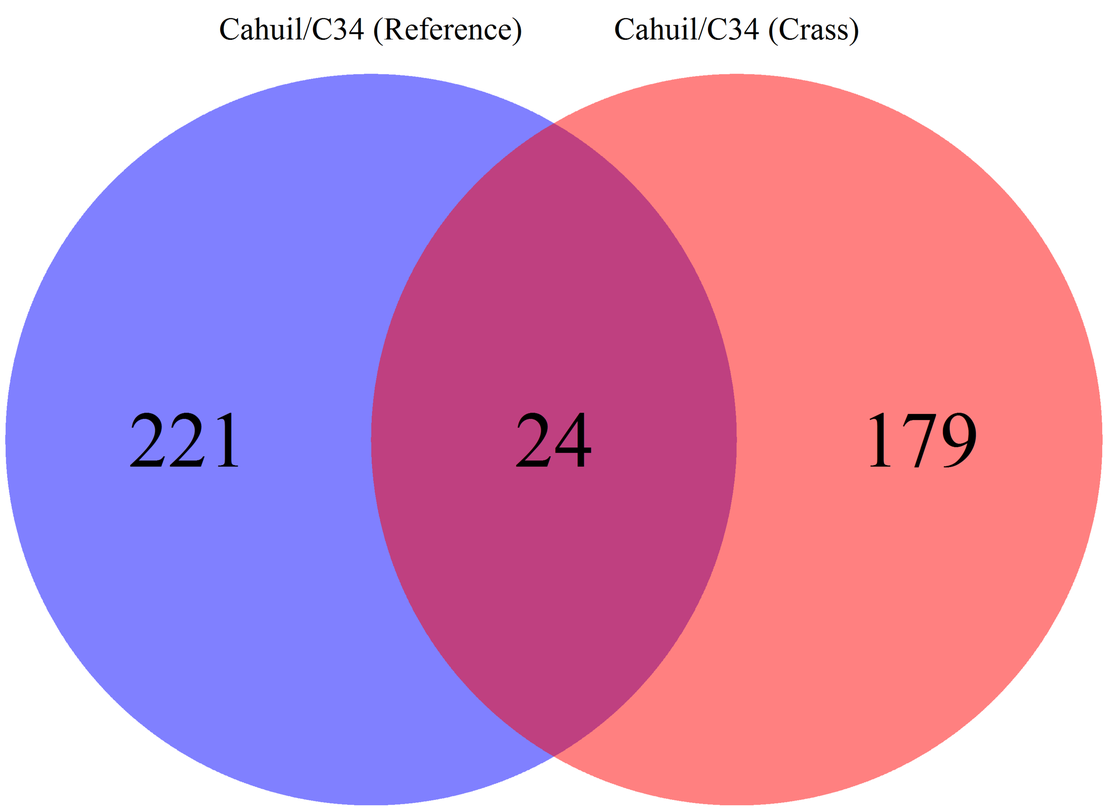

Cahuil/C34 (Reference)
Cahuil/C34 (Crass)
221
24
179

Figure 6(on next page)

Comparison of CRISPR direct repeats detected with the *de novo* (Crass) method.

Direct repeats detected with Crass were compared amongst Cahuil (C34), combined Chula Vista, and combined Santa Pola and Isla Cristina metagenomes listed in Table 3.

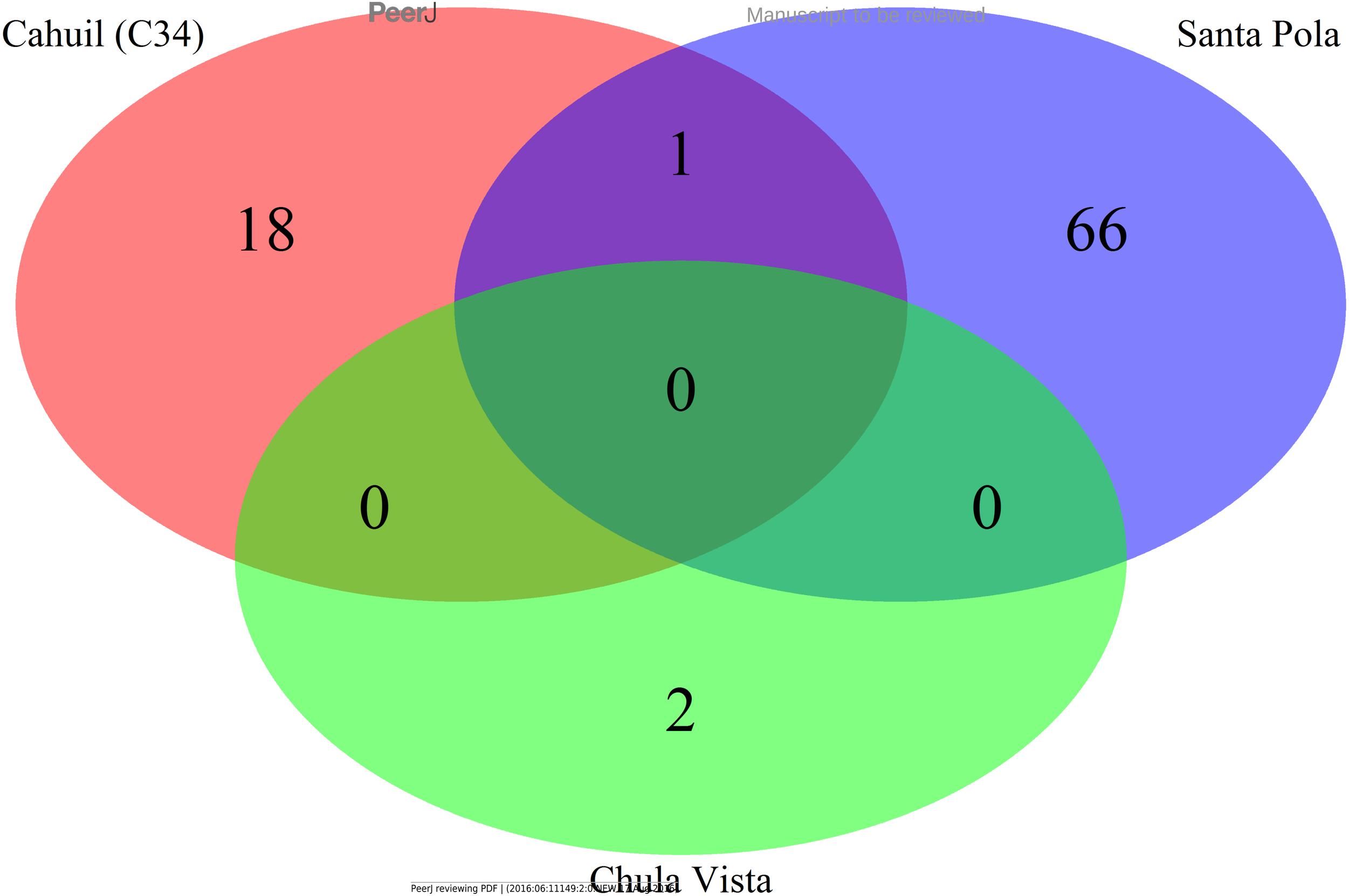

Cahuil (C34)
Santa Pola
Chula Vista
18
1
66
0
0
0
2

Figure 7 (on next page)

Map of virus-host interactions generated by aligning Crass spacers detected in the combined Santa Pola and Isla Cristina metagenome (Table 3) against a library of haloviral genomes.

The library of haloviral genomes screened is listed in Table S1. Nodes represent either viruses or spacers, while edges represent BLAST alignments linking spacers to viruses. Viruses are marked in orange and spacers in blue. Visualization was performed with Cytoscape.

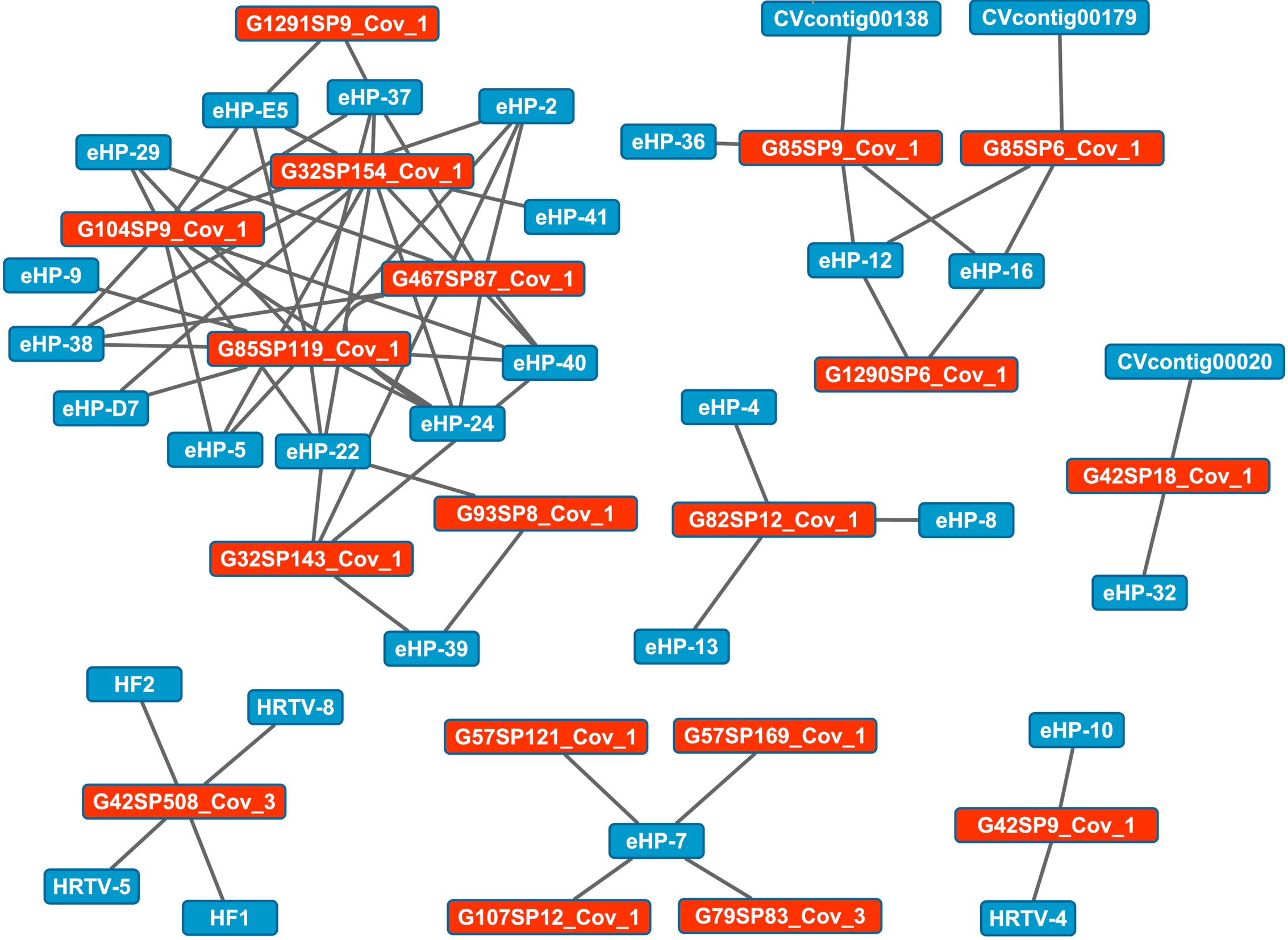

G1291SP9_Cov_1
eHP-E5
eHP-37
eHP-2
eHP-29
G32SP154_Cov_1
eHP-41
G104SP9_Cov_1
eHP-9
G467SP87_Cov_1
eHP-38
G85SP119_Cov_1
eHP-40
eHP-D7
eHP-24
eHP-5
eHP-22
G93SP8_Cov_1
G32SP143_Cov_1
eHP-39
CVcontig00138
CVcontig00179
eHP-36
G85SP9_Cov_1
G85SP6_Cov_1
eHP-12
eHP-16
G1290SP6_Cov_1
eHP-4
G82SP12_Cov_1
eHP-8
eHP-13
CVcontig00020
G42SP18_Cov_1
eHP-32
HF2
HRTV-8
G42SP508_Cov_3
HRTV-5
HF1
G57SP121_Cov_1
G57SP169_Cov_1
eHP-7
G107SP12_Cov_1
G79SP83_Cov_3
eHP-10
G42SP9_Cov_1
HRTV-4

Figure 8(on next page)

Map of virus-host interactions generated by aligning spacers detected with reference-guided methods in SS13 (A), SS19 (B), SS33 (C), SS37 (D), IC21 (E), and Cahuil/C34 (F).

The library of haloviral genomes screened is listed in Table S1. Nodes represent either viruses or spacers, while edges represent BLAST alignments linking spacers to viruses. Viruses are marked in orange and spacers in blue. Visualization was performed with Cytoscape.

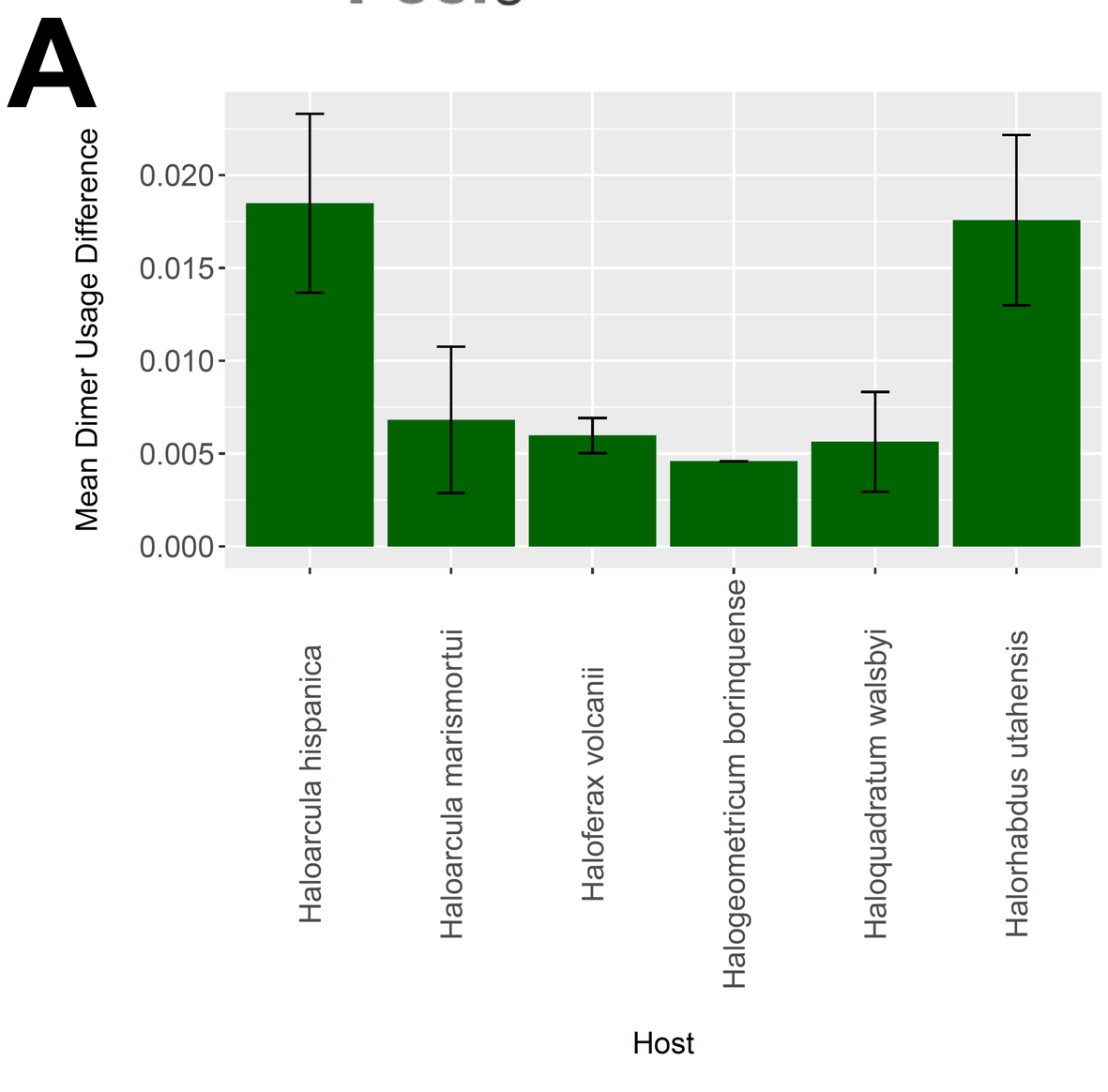
PeerJ

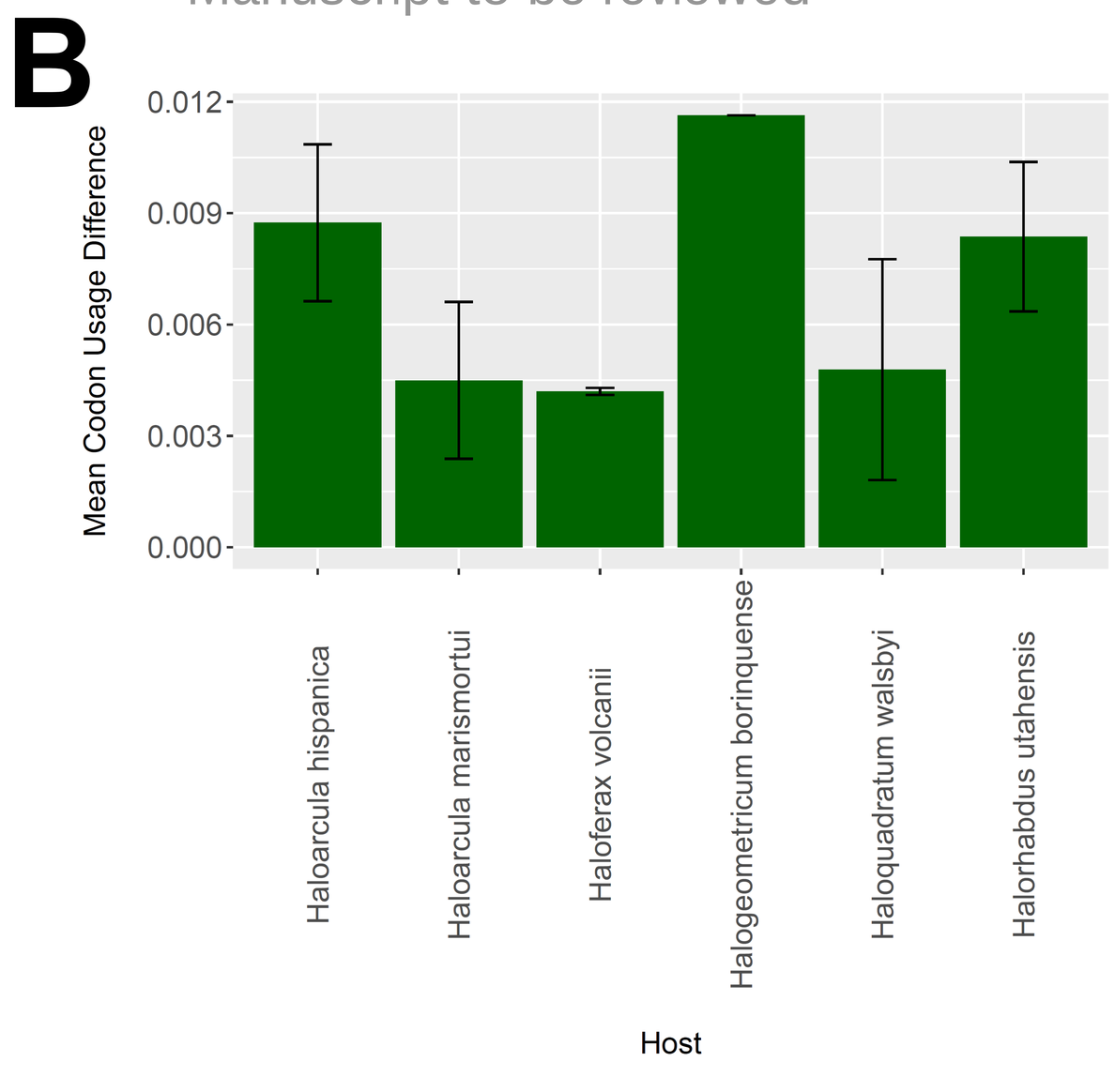
Manuscript to be reviewed

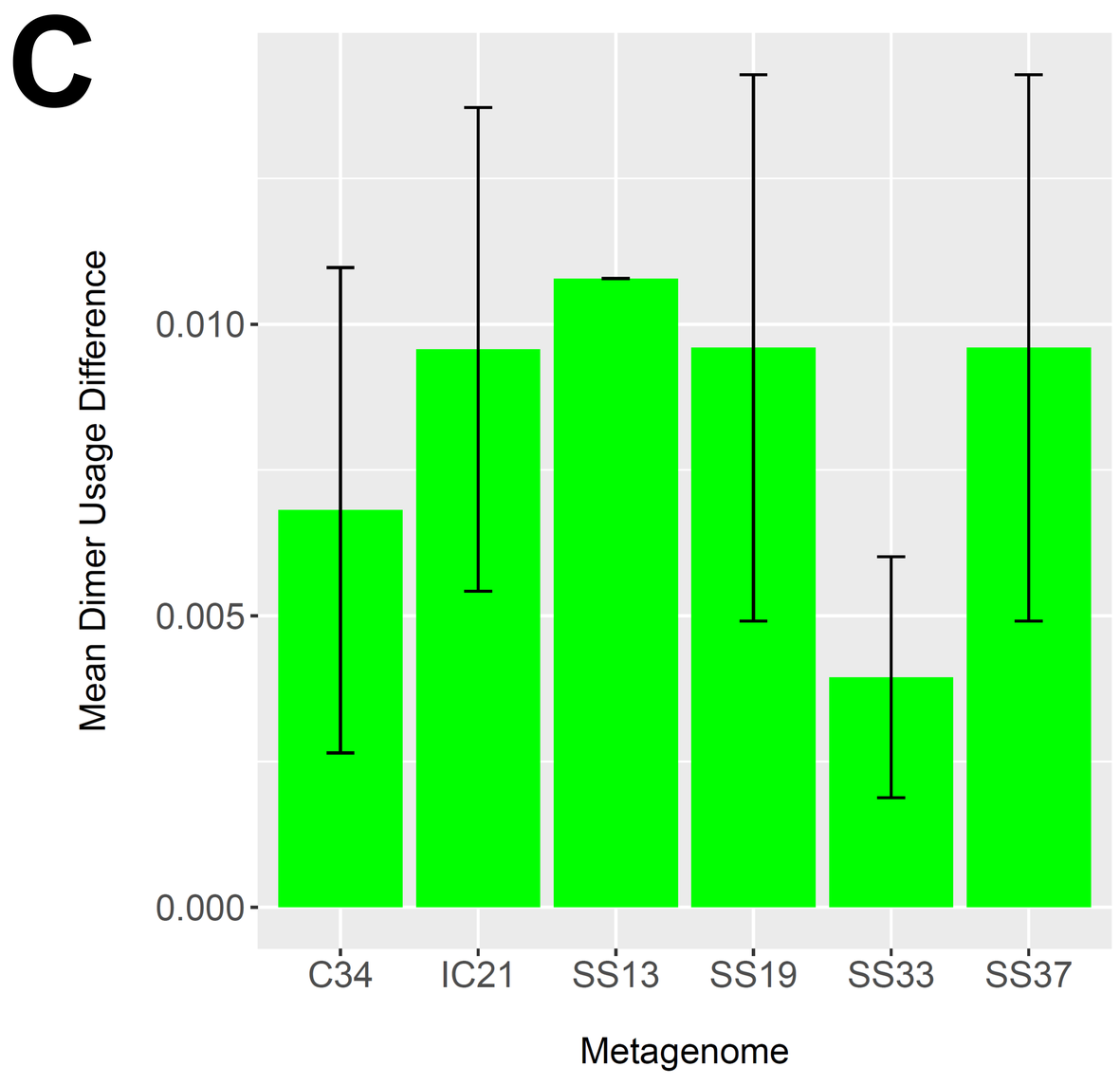
Figure 9(on next page)

Measures of difference in dimer and codon usages between host and virus genomes matched with CRISPR spacers.

The differences between each possible dimer and codon usage was subtracted, squared, and summed across all possible dimers or codons to generate a measure of dimer or codon difference. The average dimer usage and codon usage was calculated for all combinations of viruses with a particular host (A and B, respectively) and all virus-host combinations detected for a particular metagenome (C and D, respectively). All averages are reported with a 95% confidence interval (within two standard error) in black.


A
Mean Dimer Usage Difference
0.020
0.015
0.010
0.005
0.000
Haloarcula hispanica
Haloarcula marismortui
Haloferax volcanii
Halogeometricum borinquense
Haloquadratum walsbyi
Halorhabdus utahensis
Host


B
Mean Codon Usage Difference
0.012
0.009
0.006
0.003
0.000
Haloarcula hispanica
Haloarcula marismortui
Haloferax volcanii
Halogeometricum borinquense
Haloquadratum walsbyi
Halorhabdus utahensis
Host


C
Mean Dimer Usage Difference
0.010
0.005
0.000
C34
IC21
SS13
SS19
SS33
SS37
Metagenome

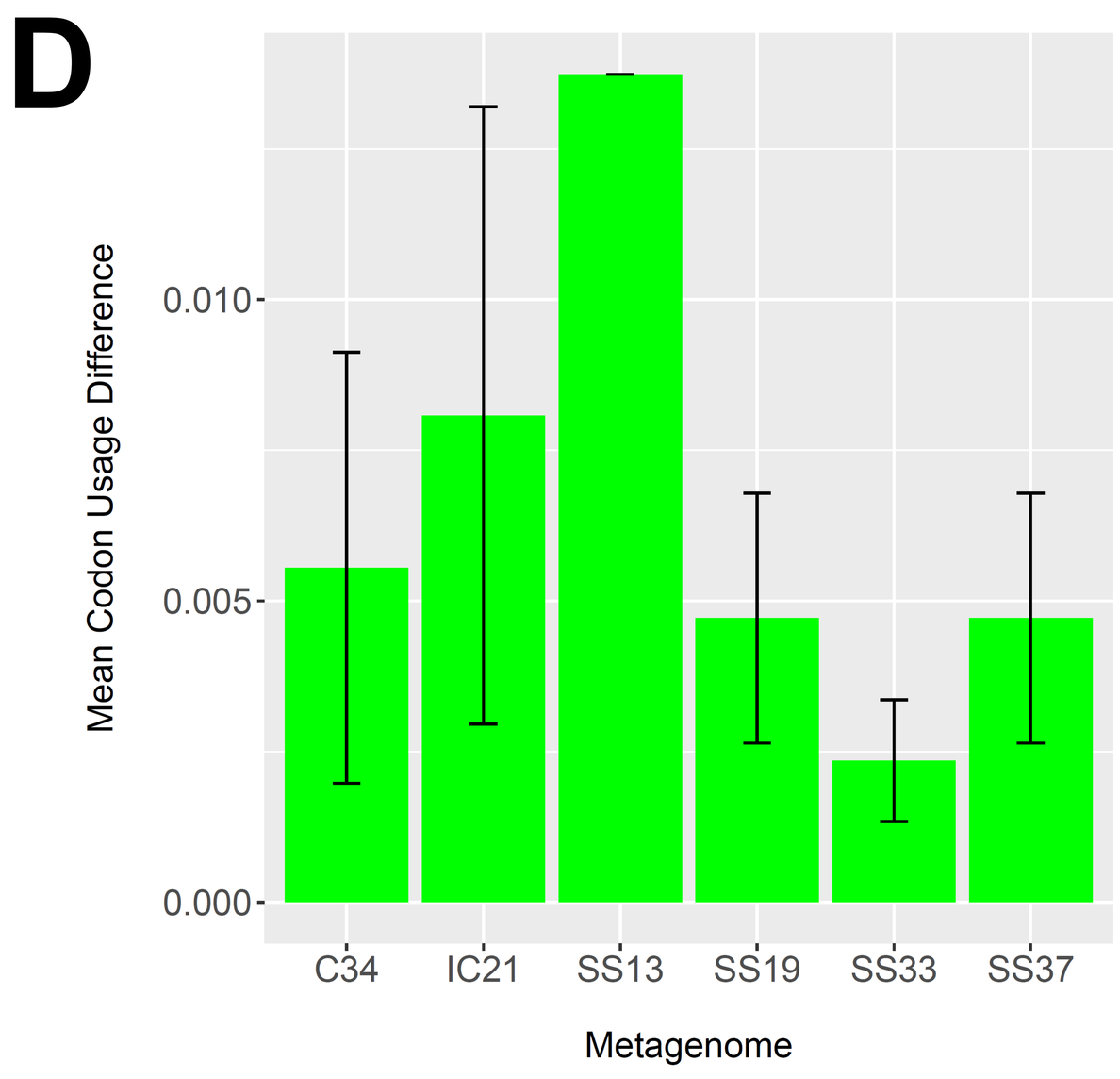
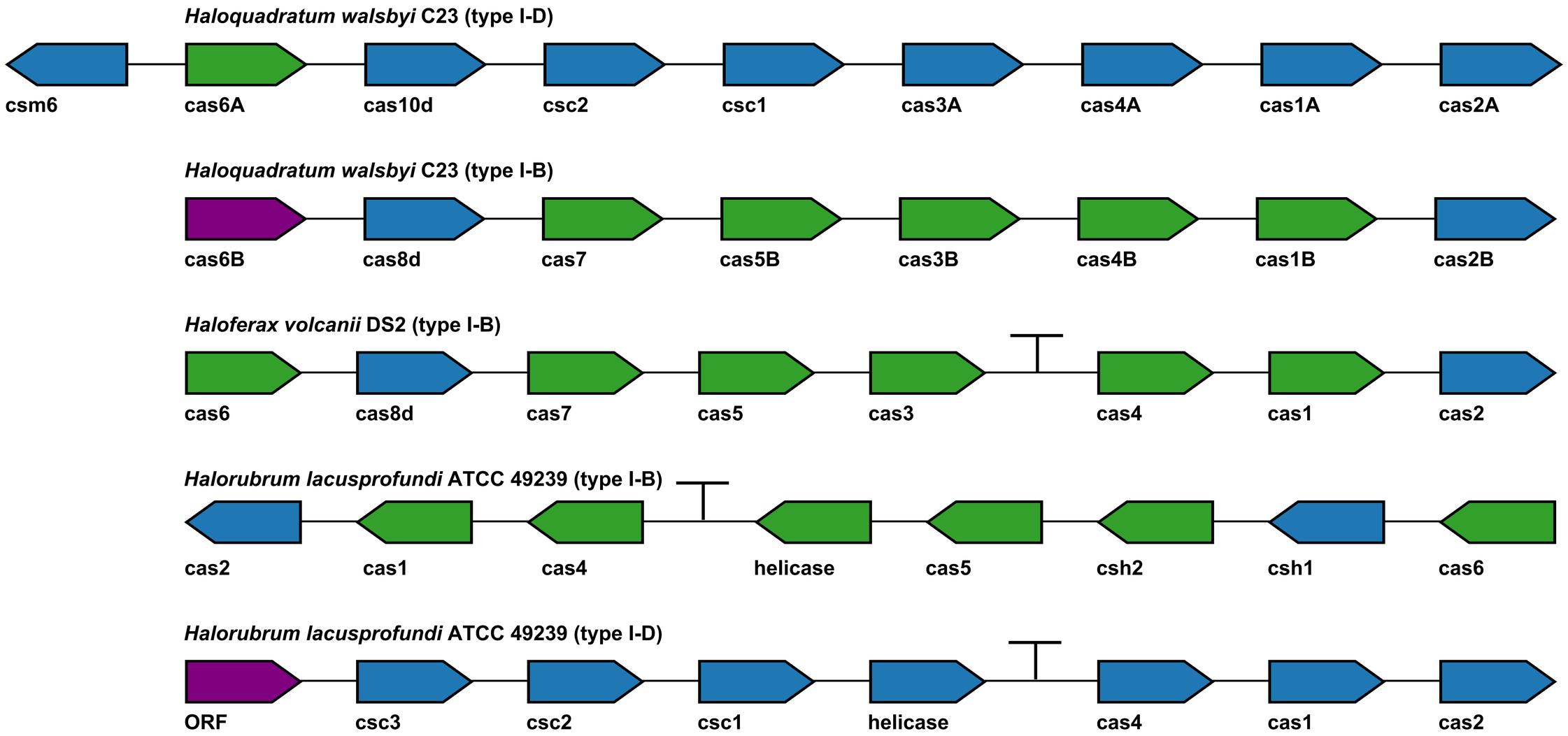

D
Mean Codon Usage Difference
0.010
0.005
0.000
C34
IC21
SS13
SS19
SS33
SS37
Metagenome

Figure 10(on next page)

Detection of *cas* operons in metagenomes and affiliation of detected genes to representative taxa.

Three metagenomes (Cahuil/C34, combined Chula Vista, and combined Santa Pola and Isla Cristina) were assembled and searched for contigs matching *cas* genes. Detection results are reported for five halobacterial *cas* operons (A). *cas* genes detected in no metagenomes are shown in red; genes detected in one metagenome are shown in blue; genes detected in two metagenomes are shown in green (A). Detected contigs assembled with Newbler and Velvet were further aligned against a library of cas genes from three halobacterial orders (Halobacteriales, Haloferacales, and Natrialbales). The taxonomic affiliations of the best hits for each contig were tabulated into profiles for each order and assembly method (B).

A


Haloquadratum walsbyi C23 (type I-D)
csm6
cas6A
cas10d
csc2
csc1
cas3A
cas4A
cas1A
cas2A
Haloquadratum walsbyi C23 (type I-B)
cas6B
cas8d
cas7
cas5B
cas3B
cas4B
cas1B
cas2B
Haloferax volcanii DS2 (type I-B)
cas6
cas8d
cas7
cas5
cas3
cas4
cas1
cas2
Halorubrum lacusprofundi ATCC 49239 (type I-B)
cas2
cas1
cas4
helicase
cas5
csh2
csh1
cas6
Halorubrum lacusprofundi ATCC 49239 (type I-D)
ORF
csc3
csc2
csc1
helicase
cas4
cas1
cas2

B

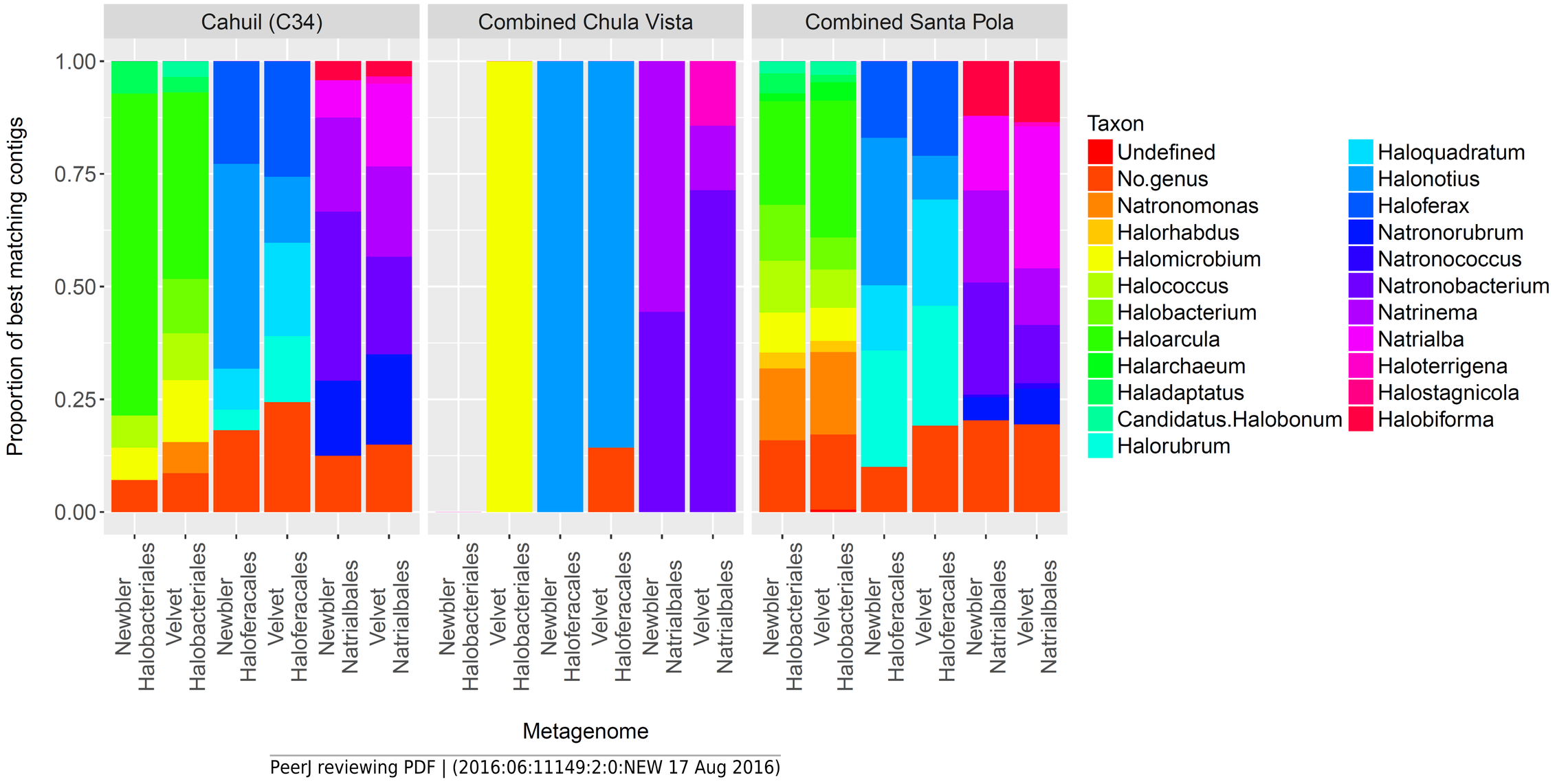

Cahuil (C34)
Combined Chula Vista
Combined Santa Pola
Proportion of best matching contigs
1.00
0.75
0.50
0.25
0.00
Newbler Halobacteriales
Velvet Halobacteriales
Newbler Haloferacales
Velvet Haloferacales
Newbler Natrialbales
Velvet Natrialbales
Metagenome
Taxon
Undefined
No.genus
Natronomonas
Halorhabdus
Halomicrobium
Halococcus
Halobacterium
Haloarcula
Halarchaeum
Haladaptatus
Candidatus.Halobonum
Halorubrum
Haloquadratum
Halonotius
Haloferax
Natronorubrum
Natronococcus
Natronobacterium
Natrinema
Natrialba
Haloterrigena
Halostagnicola
Halobiforma

Table 1(on next page)

Additional technical information about the Santa Pola, Isla Cristina, and Cahuil solar saltern metagenomes.

This information is provided in the NCBI Sequence Read Archive (SRA) and associated genome announcements (Fernandez et al., 2013; Fernández et al., 2014; Plominsky et al., 2014). All metagenomes were sequenced using 454 technology. Site names (as listed in the Figures) are listed in bold. NCBI SRA accession numbers (SRX for experiment and SRR for run) are included in the table. Combinations of these metagenomes used for further analyses are listed in Table 3.

Table 1: Additional technical information about the Santa Pola, Isla Cristina, and Cahuil solar saltern metagenomes provided in the NCBI Sequence Read Archive (SRA) and associated genome announcements (Fernandez et al., 2013; Fernández et al., 2014; Plominsky et al., 2014). All metagenomes were sequenced using 454 technology. Site names (as listed in the Figures) are listed in bold. NCBI SRA accession numbers (SRX for experiment and SRR for run) are included in the table. Combinations of these metagenomes used for further analyses are listed in Table 3.

Site	Salinity	Reference	SRX number	Sequencing technology	Number of reads (millions)	Average read length (bp)
SS13	13%	Fernandez et al., 2013	SRX328504	454	1.5	479
SS19	19%	Ghai et al., 2011	SRX024859	454	0.293	504
IC21	21%	Fernandez et al., 2014	SRX352042	454	1.2	508
SS33	33%	Fernandez et al., 2013	SRX347883	454	0.971	429
SS37	37%	Ghai et al., 2011	SRX090229	454	0.741	511
Cahuil (**C34**)	34%	Plominsky et al., 2014	SRX680116	454	0.222	523

Table 2(on next page)

Additional information about the Chula Vista saltern microbial metagenomes obtained from the iMicrobe Collaborative and combined into three metagenomes based on salt concentration.

Site names (as listed in the Figures) are in bold. The datasets can be found at this web address: http://data.imicrobe.us/project/view/58.

Table 2: Additional information about the Chula Vista saltern microbial metagenomes obtained from the iMicrobe Collaborative and combined into three metagenomes based on salt concentration. Site names (as listed in the Figures) are in bold. The datasets can be found at this web address: http://data.imicrobe.us/project/view/58.

Salt content (and label in figures in bold)	Sampling dates corresponding to included metagenomes	Sequencing technology	Number of reads (millions)	Average read length (bp)	Total length (Mbp)
Low (6-8%) **(CV6-8)**	07/2004 11/10/2005 11/28/2005	454	0.352	95	33.4
Medium (12-14%) **(CV12-14)**	11/10/2005 11/11/2005 11/16/2005 11/28/2005	454	0.223	99	22.1
High (27-30%) **(CV27-30)**	07/2004 11/28/2005	454	0.618	102	63.2

Table 3(on next page)

Combined metagenomes, the individual metagenomes that compose them, and how they were used in following analyses. <!--RZ2N-->

The Cahuil metagenome (C34) was never included in any combined metagenome used for further analyses. The combined metagenomes were mainly used to improve *de novo* CRISPR detection, virus-host mapping using these *de novo*-detected CRISPR spacers, and assembly into contigs for detection of *cas* operons. These analyses required higher depth metagenomes; for example, no *de novo*-detected CRISPR spacers in several individual Santa Pola metagenomes mapped to any viruses in the halovirus library. The combined Chula Vista viral metagenome was used to assemble viral contigs included in the haloviral library (Table S1).

Table 3: Combined metagenomes, the individual metagenomes that compose them, and how they were used in following analyses. The Cahuil metagenome (C34) was never included in any combined metagenome used for further analyses. The combined metagenomes were mainly used to improve *de novo* CRISPR detection, virus-host mapping using these *de novo*-detected CRISPR spacers, and assembly into contigs for detection of *cas* operons. These analyses required higher depth metagenomes; for example, no *de novo*-detected CRISPR spacers in several individual Santa Pola metagenomes mapped to any viruses in the halovirus library. The combined Chula Vista viral metagenome was used to assemble viral contigs included in the haloviral library (Table S1).

Combined metagenome	Individual metagenomes included	How they were analyzed
Santa Pola and Isla Cristina (Combined SP)	SS13, SS19, IC21, SS33, and SS37	• *De novo* CRISPR detection with Crass (**Figure 4**) • Reference-guided CRISPR detection (**Figure 4**) • Comparing DRs detected with the *de novo* CRISPR detection method Crass (**Figure 6**) • Virus-host mapping using *de novo*-detected CRISPR spacers (**Figure 7**) • Assembly to detect *cas* operons (**Figure 10**)
Chula Vista (Combined CV)	CV6-8, CV12-14, and CV27-30	• *De novo* CRISPR detection with Crass (**Figure 4**) • Reference-guided CRISPR detection (**Figure 4**) • Comparing DRs detected with the *de novo* CRISPR detection method Crass (**Figure 6**) • Assembly to detect *cas* operons (**Figure 10**)
Chula Vista combined viral metagenome	Chula Vista viral metagenomes	• Assembly of Chula Vista viral contigs included in haloviral library (**Figures 7 and 8; Table S1**)