Associations between depth and micro-diversity within marine viral communities revealed through metagenomics

Coutinho, FH^{1*}; Rosselli, R¹; Rodríguez-Valera F¹

Author affiliations

1 – Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Campus San Juan, San Juan, Alicante 03550, Spain

* Address correspondence to: fhernandes@umh.es

Abstract

Viruses are extremely abundant and diverse biological entities that contribute to the functioning of marine ecosystems. Despite their recognized importance no studies have addressed trends of micro-diversity in marine viral communities across depth gradients. To fill this gap we obtained metagenomes from both the cellular and viral fractions of Mediterranean seawater samples spanning the epipelagic to the bathypelagic zone at 15, 45, 60 and 2000 meters deep. The majority of viral genomic sequences obtained were derived from bacteriophages of the order *Caudovirales*, and putative host assignments suggested that they infect some of the most abundant bacteria in marine ecosystems such as *Pelagibacter*, *Puniceispirillum* and *Prochlorococcus*. We evaluated micro-diversity patterns by measuring the accumulation of synonymous and non-synonymous mutations in viral genes. Our results demonstrated that the degree of micro-diversity differs among genes encoding metabolic, structural, and replication proteins and that the degree of micro-diversity

increased with depth. These trends of micro-diversity were linked to the changes in environmental conditions observed throughout the depth gradient, such as energy availability, host densities and proportion of actively replicating viruses. These observations allowed us to generate hypotheses regarding the selective pressures acting upon marine viruses from the epipelagic to the bathypelagic zones.

Running title: Viral micro-diversity at different depths

Introduction

Viruses are increasingly recognized as important players in the functioning of marine ecosystems[1, 2]. In recent years many efforts were undertaken to describe associations between viral biodiversity and spatial[3], temporal[4], and ecological[5] gradients. The taxonomic composition and functioning of host communities respond to such changes in environmental parameters across such gradients[6, 7]. In response, viruses adapt to those changes to guarantee their survival. The depth gradient of stratified water masses displays marked changes in environmental conditions mainly driven by light availability and temperature[8]. Thus it is an ideal habitat to study associations between environmental parameters, viruses, and their hosts.

In stratified waters, temperature decreases with depth while the concentration of inorganic nutrients increases. The micro-habitat at the thermocline provides photosynthetic microorganisms with ideal conditions of temperature, nutrient availability and light irradiation. The intense proliferation of photosynthetic microbes there leads to a peak of chlorophyll concentration and microbial cell density, known as the deep chlorophyll maximum (DCM). In the stratified water column, the DCM often exhibits the highest densities of prokaryotic cells and viral particles[9, 10]. Moving towards the aphotic zone, the concentrations of inorganic nitrogen and phosphorus increase, but the gradual decrease of light hampers productivity, thus leading to much lower cell

50 densities than observed in the surface or the DCM. Below the DCM both viral and bacterial
51 abundances decrease, and deeper waters of the bathypelagic zone often display the lowest densities
52 of both bacteria and viruses [9, 11].

53 Previous studies have used metagenomics to assess changes in the taxonomic and
54 functional composition of viral communities throughout depth gradients[4, 12]. Nevertheless,
55 studies addressing patterns of micro-diversity, i.e. accumulation of mutations within genomes,
56 through the stratified water column are lacking. Investigating patterns of micro-diversity can help to
57 elucidate the selective pressures acting upon viral genomes. For example, in co-evolution
58 experiments in which bacteriophages and hosts are cultured together over multiple generations,
59 viruses tend to preferentially accumulate mutations in genes that affect their host range and the
60 productivity of viral particles[13, 14]. These discoveries provided insightful information regarding
61 the processes by which viruses adapt to more efficiently infect their hosts in cultures. Yet no studies
62 have addressed this topic in free-living marine viral communities through culturing-independent
63 approaches. These are necessary because the selective pressures acting on viral genomes in cultures
64 and environmental communities might be drastically distinct.

65 Here we sought to investigate micro-diversity patterns in the environment to generate
66 hypotheses about the selective pressures acting on marine viral communities throughout the depth
67 gradient. We selected a site at the Mediterranean sea off the coast of Spain during a period of water
68 stratification (October 2015). Seawater samples were retrieved from multiple depths ranging from
69 the epipelagic at 15, 45 (DCM), 60 (DCM) to the bathypelagic at 2000 meters deep, and used for
70 preparing both cellular and viral metagenomes (viromes). Viromes were assembled to obtain
71 complete or partial viral genomes and cellular metagenomes were assembled and binned to obtain
72 metagenome assembled genomes (MAGs). Next, reads from both the viral and cellular fractions
73 were mapped against the assembled viral scaffolds to calculate the level of micro-diversity for each
74 viral protein. Our rationale was that the changes taking place in microbial communities at surface,

DCM and aphotic habitats would subject the associated viral communities to different constraints, which would be reflected in the micro-diversity patterns within viral genomes.

Results

Assembled viral genomes and predicted hosts

Assembly of viral metagenomes yielded 10,263 genomic sequences of length equal or greater than 5 kbp, within which 133,352 protein encoding genes were identified (Table 1). A total of 7,164 (69.8%) scaffolds were classified as *bona fide* viral sequences based on the annotation of their protein encoding genes (see methods). Among these, 21 scaffolds with length equal or above 10 Kbp (average length = 44 Kbp) and with overlapping ends were identified, which likely represent complete viral genomes. Computational host predictions were obtained for the *bona fide* viral sequences by scanning viral and prokaryote genomes for three signals of virus-host association: homology matches (i.e. long genomic segments sharing high nucleotide identity), shared tRNA genes, and matches between CRISPR spacers and viral sequences. These approaches have previously been benchmarked and shown to provide accurate host predictions, specially at higher taxonomic ranks such as phylum and class[15, 16]. In addition, we manually curated host-predictions by investigating the gene content of viral genomic sequences. Host predictions were obtained for 171 of the *bona fide* viral sequences (Table S1 and Figure 1A). Among those, the majority were predicted to infect *Proteobacteria* (99 sequences), particularly *Alphaproteobacteria* of the genera *Pelagibacter* (52 sequences) and *Puniceispirillum* (38), followed by *Cyanobacteria* (58) of the genera *Prochlorococcus* and *Synechococcus*.

Taxonomic classification of the assembled scaffolds identified most of them as tailed bacteriophages from the order *Caudovirales* (Figure 1B), specifically as members of the families *Myoviridae*, *Podoviridae* and *Siphoviridae*. Some of the scaffolds from the epipelagic samples were

classified as *Phycodnaviridae*, viruses that infect Eukaryotic algae. Scaffolds annotated as *Microviridae* bacteriophages were exclusively retrieved from the bathypelagic sample.

Viral community composition

Grouping viral abundances according to predicted host revealed differences among samples of the depth gradient (Figure 2A). Scaffolds predicted to infect *Proteobacteria* were among the most abundant in all depths with abundances ranging from 0.5% to 2.4% of mapped reads. Scaffolds predicted to infect *Cyanobacteria* and *Euryarchaeota* displayed their highest abundances at the 15m and 45m samples while those predicted to infect *Bacteroidetes* were abundant only at the 45m sample. The 2000m displayed a unique profile with abundant scaffolds predicted to infect *Firmicutes* and *Actinobacteria*.

Previous investigation of the metagenomes from the cellular fraction revealed shifts in taxonomic composition of prokaryotic communities throughout the depth gradient[8]. These were dominated, at all depths, by *Proteobacteria*, mostly from the classes *Alphaproteobacteria* and *Gammaproteobacteria*. The taxonomic composition of viral communities also displayed shifts according to depth (Figure 2B). The families of tailed bacteriophages *Myoviridae*, *Podoviridae* and *Siphoviridae* within the order *Caudovirales* were dominant in all samples, and together accounted for 15% to 45% of the annotated reads. Bacteriophages from the family *Microviridae* were abundant in the bathypelagic sample only, while eukaryotic viruses from the family *Phycodnaviridae* were detectable only at the epipelagic samples, although at lower abundances.

Mediterranean viruses actively replicating in the cellular fraction

Read mapping revealed that many of the viral scaffolds assembled from viromes could also be detected in the cellular metagenomes (Figure 3A). We assumed that the viral sequences that were abundant in the cellular metagenomes are derived from actively replicating viruses undertaking lytic

infections, which lead to high copy numbers of their genomes inside host cells. Alternatively, viral sequences in the cellular fraction could be the result of lysogenic infections. Yet those are not expected to produce the high copy numbers of viral genomes inside host cells that could lead to the observed abundance patterns.

Abundances of viral sequences in the cellular fraction differed between samples. The average ratios of cellular/viral abundances were highest for the 45 and 60m samples, followed by 15m and lastly the 2000m sample (Figure 3B). Likewise, the abundance of raw reads annotated as viral in the cellular fraction metagenomes followed the same trend. Thus, there were more viruses actively replicating at the DCM samples than at any other depth, followed by the 15m sample and lastly the 2000m sample, which displayed the lowest proportion of actively replicating viruses. In addition, the DCM samples displayed the lowest values for the Shannon diversity index (5.55 and 5.61), while these values were higher for the 15m (7.21) and 2000m (7.26) samples. The high proportion of actively replicating viruses, and the low Shannon diversity observed at the DCM suggest that the intense viral replication taking place at these depths lead to a highly clonal community, with many nearly-identical viral genomes co-existing at high densities.

140

141 *Levels of micro-diversity shift throughout the depth gradient and across functional categories*

We evaluated micro-diversity patterns by measuring the pN/pS ratios of protein encoding genes identified in the *bona fide* viral scaffolds. The pN/pS ratio is a measure analogous to dN/dS that does not require specific haplotypes to be identified, and therefore can be applied to metagenomic datasets to provide a population level measure of micro-diversity[17–19]. Briefly, reads from the metagenomes were mapped to the assembled scaffolds to detected mutations, specifically single nucleotide polymorphisms. Next, pN and pS were calculated by respectively dividing the observed counts of non-synonymous and synonymous mutations by the expected frequencies of these mutations under a neutral model.

150 The majority of proteins displayed pN/pS values below 1, regardless of sample, meaning
151 that the frequencies of non-synonymous mutations was below that which was expected by chance.
152 Thus purifying selection was a major driving force regulating frequencies on mutations among viral
153 genes. Nevertheless, 117 proteins displayed pN/pS above 1 in the cellular fraction metagenomes,
154 and 1,092 in the viral fraction metagenomes. Most of these proteins were retrieved from the 15m
155 sample (755), followed by 2000m (239), 45m (148) and 60m (67) samples. Although the majority
156 of these genes had no assigned functions, some were identified as: recombinase/nuclease proteins
157 (21), oxygenases (17), lysins (16), methylases (13), and tail fibers (11).

158 We observed a negative association between depth and the median pN/pS ratio of each
159 sample (Figure 4A). The highest median of pN/pS values was observed for the 2000m sample,
160 followed by 60m, 45m and lastly the 15m sample. These trends of pN/pS and depth were observed
161 for viral sequences detected in the metagenomes from both the viral and cellular fractions. Because
162 the coverage of proteins in the viral fraction metagenomes was much higher and spanned many
163 more of the viral proteins, we focused subsequent analysis of pN/pS using viral fraction
164 metagenomes only.

165 Due to the many unknown proteins present in marine viral genomes, our capacity to
166 annotate these genes and predict their function is limited[20]. Nevertheless, we observed marked
167 differences of median pN/pS ratios among proteins according to functional categories (Figures 4B
168 and 4C). Genes involved in genome replication (e.g. DNA polymerase, DNA primase and genes of
169 the nucleotide metabolism) displayed the lowest median pN/pS values compared to other
170 categories. Structural viral proteins (e.g. capsid, neck and tail) showed intermediate median pN/pS
171 values. Finally, proteins associated with altering host metabolism (e.g. ferrochelatases, thioredoxins
172 and oxygenases) displayed the highest median pN/pS values. A positive association between pN/pS
173 and depth was also observed when grouping proteins according to broad functional categories
174 (Figure 4B). A notable exception was the median pN/pS ratio of structural proteins, which was

175 highest for the DCM samples.

176   These differences of pN/pS among functional categories are associated with their roles
177 during the viral infection cycle. Genes involved in genome replication must operate at high fidelity
178 and efficiency, thus deleterious non-synonymous mutations in these proteins are readily removed
179 from the population by purifying selection. Meanwhile, structural proteins are fundamental for
180 adequate particle assembly, encapsulation of the viral genome, and host recognition. Deleterious
181 mutations in structural genes can also compromise viral infections, but not as much as errors during
182 genome replication. Finally, metabolic genes are responsible for re-directing host metabolism
183 towards pathways that favour viral particle production[21, 22]. Thus, lower efficiency of metabolic
184 genes due to deleterious mutations is likely to reduce viral productivity but not to compromise it as
185 much as deleterious mutations in the genome replication or structural modules.

186

187 *The DCM is a micro-diversity hot-spot for viral receptor binding proteins*

188   The DCM samples displayed the highest median pN/pS values for structural proteins (Figure
189 4B). Specifically, structural proteins that encoded baseplate, capsid, tail, and tail fiber genes
190 displayed pN/pS values higher than their counterparts in the remaining samples (Figure 4C).
191 Interestingly all of these proteins either are or interact directly with receptor binding proteins that
192 mediate host recognition, a fundamental step for successful viral infection[23, 24]. The enhanced
193 pN/pS observed for these genes at the DCM provides evidence that this habitat is a micro-diversity
194 hot-spot for viral receptor binding proteins.

195   Adaptation to sub-optimal hosts is a major driver of genomic diversification for viruses,
196 which is associated with the quick accumulation of non-synonymous mutations in tail fiber
197 proteins[14]. A single nucleotide polymorphism in tail fiber gene can be sufficient to alter viral host
198 range[13, 25]. Consistent with those findings, we observed multiple cases of tail proteins in which
199 non-synonymous mutations were concentrated in small segments of these gene (Figure 5). These

200 sites that accumulate non-synonymous mutations at higher frequencies than the other codons are
 201 likely those that confer a selective advantage to the virus at their specific habitat according to the
 202 availability of hosts. These trends are consistent with a scenario where, on the one hand, positive
 203 selection acts on tail fiber proteins to expand host range, while on the other hand, purifying
 204 selection removes mutations from other sites where they cause loss of function or restrict the host
 205 range instead of expanding it[14, 26].

206

207 **Discussion**

208 *Different selective pressures determine levels of micro-diversity throughout the depth gradient*

209 Major changes take place among prokaryotic and viral communities throughout the depth
 210 gradient, affecting their taxonomic composition and virus-host interactions [4, 8, 9, 27, 28]. These
 211 differences in cell densities and frequency of replication events impact micro-diversity because the
 212 rate at which viral genomes accumulate mutations is density dependent, meaning that they adapt
 213 faster in conditions with higher host density, in which more infection events take place[30]. Our
 214 results demonstrated that the DCM viral communities had the highest proportions of actively
 215 replicating viruses but were the least diverse. We propose that this scenario leads to intense intra-
 216 species competition between viruses for suitable hosts, creating a selective pressure that favours
 217 viruses with mutations in receptor binding proteins which provide them with a different host-range,
 218 allowing them to exploit a distinct niche (Figure 6). The high micro-diversity observed among
 219 receptor binding proteins and the clonal populations observed within DCM samples suggests that
 220 many strains of viruses with distinct host ranges co-exist at this habitat. It follows that host strains
 221 with different patterns of viral-susceptibility are also co-existing in these sites. This is in agreement
 222 with the constant diversity theory[29], which postulates that the trade-offs between ecological
 223 fitness and viral susceptibility are responsible for avoiding that a single bacterial clone dominates
 224 the community through clonal sweeps, thus preserving the taxonomic and functional diversity of

225 these communities[13, 26].

226   Meanwhile, a different pattern was observed for the bathypelagic sample. In this habitat
 227 both viral and cell densities are much lower than in the DCM or the surface[9, 11]. Due to the lower
 228 availability of hosts at 2000 meters, chance encounters between viruses and hosts are expected to
 229 occur less often. Thus, less infection events take place at 2000m compared to shallower depths with
 230 higher host densities, as evidenced by the differences in abundances of viruses actively replicating
 231 in the cellular fraction. Interestingly, the bathypelagic sample displayed the highest Shannon
 232 diversity but lowest proportion of actively replicating viruses. This finding suggests that in the
 233 energy-limited bathypelagic zone, intra-species competition for hosts is expected to be less relevant
 234 than it is at the DCM, where a highly clonal population with high density was observed. Instead, the
 235 major constraint faced by viruses at this depth could be the efficient production of viral progeny,
 236 since in this scenario a lower reproductive fitness is more likely to lead to local extinction than in
 237 the highly productive conditions of the euphotic zone. Consistent with that, we observed the highest
 238 pN/pS values of proteins encoding metabolic functions (e.g. oxygenases and thioredoxin) and
 239 transcriptional regulators in the 2000m sample. We postulate that the higher micro-diversity
 240 observed among these genes in the bathypelagic sample is evidence of positive selection acting on
 241 proteins that increase the capacity of viruses to generate progeny by using a diverse array of
 242 auxiliary metabolic genes and transcriptional regulators to fine-tune host metabolism to enhance the
 243 production of viral particles under conditions of low energy availability and productivity (Figure
 244 6).

245
 246 *Micro-diversity patterns differ between pure cultures and environmental samples*

247   In laboratory experiments of phage-bacteria co-evolution, mutations usually accumulate in
 248 genes involved in host specificity such as tail proteins[14, 31]. In contrast, we observed a broader
 249 distribution of mutations that spanned all functional categories within viral genomes. We attribute

250 this to the differences between the selective pressures imposed over viruses in co-evolution
 251 experiments versus in the environment. In the former, the only selective pressure is to effectively
 252 infect one single host derived from a clonal population. In the latter, viruses have a multitude of
 253 hosts available, each with their specific viral receptors and resistance mechanisms (e.g. CRISPR
 254 and restriction modification systems). In cultures, once resistance mutations appear their prevalence
 255 quickly rises within bacterial populations[14]. In the environment, the frequency of resistance
 256 mutations is simultaneously regulated by a trade-off of viral resistance and the fitness cost brought
 257 by these resistance mutations[26]. These differences in selective pressures faced by viruses in
 258 environmental communities is likely to lead to the accumulation of mutations throughout the
 259 entirety of viral genomes, and not just at the sites associated with host recognition and infection.

260

261 *Concluding remarks*

262 Light, depth and temperature are main factors structuring the taxonomic and functional
 263 composition of marine viral communities[5, 32]. These variables are major determinants of the
 264 energy available across the ecosystem, and they shift drastically throughout the depth gradient from
 265 which our samples were retrieved[8]. Our data shows that these parameters not only shape the
 266 taxonomic composition of viral communities but also influence how the genomes of these viruses
 267 accumulate mutations and evolve. To our knowledge this is the first study assessing patterns of
 268 micro-diversity within marine viromes. The obtained results allowed us to postulate hypotheses
 269 about the selective pressures acting upon marine viruses from the community to the amino acid
 270 level. Furthermore, we demonstrated that the frequencies of non-synonymous mutations differed
 271 among functional categories and depth. Finally, free-living viruses displayed patterns of mutation
 272 accumulation different from those observed in laboratory conditions, which has important
 273 implications for how the latter should be interpreted. Here we set a stepping stone for investigating
 274 patterns of micro-diversity among environmental viral communities. Further research will be

275 necessary to determine if the patterns presented here are also present in other marine habitats as
276 well as different ecosystems (such as host-associated, freshwater and soils), and to determine the
277 driving forces behind them.

278

279 **Materials and Methods**

280 *Sampling and sample processing*

281 Four samples from different depths, 15, 45, 60 and 2000 meters were collected on October
282 15th 2015 from aboard the research vessel “Garcia del Cid” [8]. The sampling site was located at
283 approximately 60 nautical miles off the coast of Alicante, Spain, at 37.35361° N - 0.286194° W. Sea
284 water samples were filtered for Eukaryote and Prokaryote fractions through 20 µm, 5 µm and 0.22
285 µm pore size polycarbonate filters (Millipore). Two technical replicates (50 L for each depth) were
286 ultra-filtered on board through a Millipore Prep/Scale-TFF-6 filter, yielding 250 ml of viral
287 concentrate stocks. Each stock was purified through Sterivex 0.22 filters (Millipore), stored at 4°C
288 and subsequently reduced to 1,5 mL using Ultra-15 Centrifugal Filter Units (Amicon).

289 To minimize the carry-over of free-residual nucleic acids, stocks were treated with 2,5 U of
290 DNase-I at 37°C for 1h, followed by inactivation with EDTA (0,5 mM). Total viral DNA was
291 extracted with PowerViral Environmental RNA/DNA Isolation Kit (MoBio). Quality and quantity
292 of extracted DNA were determined using the ND-1000 Spectrophotometer (NanoDrop, Wilmington,
293 USA) and Qubit Fluorometer (Thermofisher). The absence of prokaryotic DNA was tested through
294 PCR using 16S universal primers on aliquots from each sample. Multiple Displacement
295 Amplification (MDA) was performed using Illustra GenomiPhi V2 DNA Amplification Kit (GE
296 Healthcare, Life Sciences).

297

298 *Sequencing, Assembly and Binning*

299 Metaviromes were sequenced using Illumina Hiseq-4000 (150 bp, paired-end reads) by
300 Macrogen (Republic of Korea). Reads from metaviromes were pre-processed using

301 Trimmomatic[33] in order to remove low-quality bases (Phred-quality score of 20 in 4-base sliding
302 windows) and reads shorter than 30 bases. Each metagenome was individually assembled through
303 SPAdes[34] using default parameters for the metagenomic mode. Sequences shorter than 5 Kbp
304 were discarded. Both raw reads and assembled scaffolds were deposited at ENA under project
305 ERP113162. Taxonomic and functional annotation of proteins were performed by querying PEGs
306 against the NCBI-nr database using Diamond[35], and against the pVOGs[36] database using
307 hmmer[37].

308 Scaffolds from the cellular fraction of the 2000m sample were binned with MetaBat[38] to
309 obtain Metagenome Assembled Genomes (MAGs) of Bacteria and Archaea. Quality of MAGs was
310 assessed through CheckM[39]. MAGs were manually curated to improve completeness and reduce
311 potential contamination. Protein encoding genes were identified using the metagenomic mode of
312 Prodigal[40].

313

314 *Computational host prediction*

315 Putative hosts were assigned to viral scaffolds through homology matches, CRISPR spacers
316 and shared tRNAs as previously described[41]. These were performed using two datasets: The
317 NCBI RefSeq genomes of Bacteria and Archaea (June 2017 release), and the MAGs obtained from
318 the binning of scaffolds from the cellular fraction metagenomes obtained from the same samples
319 from which the viromes were derived[8]. Putative hosts were manually assigned for sequences that
320 displayed high similarity to RefSeq bacteriophage genomes as measured by proportion of shared
321 genes and synteny between genomes. Ambiguous host predictions, i.e., derived from viral
322 sequences predicted to infect more than a single taxa were removed from further analyses.

323

324 *Abundance profiles and Micro-diversity analysis*

325 Sequencing reads from the cellular and viral metagenomes were mapped to assembled viral

scaffolds using Bowtie2 in sensitive-local mode[42]. The number of reads mapped was used to estimate the relative abundances of the viral sequences in both fractions. To estimate mutational frequencies on viral genomes, raw reads were mapped to assembled scaffolds using the sensitive-mode of Bowtie2. Next, the generated bam files were analysed through Diversitools (<http://josephhughes.github.io/DiversiTools/>) to obtain counts of synonymous and non-synonymous mutations in each protein. Codon mutations were only considered valid if they were detected at least 4 times, in at least 1% of the mapped reads, and if the codon coverage was equal or above 5x. Only the mutations that passed the aforementioned filters were considered to estimate the percentage pN/pS ratios, which were calculated as described in [19].

Acknowledgments: FHC, RR and FRV were supported by the grants “VIREVO” CGL2016-76273-P [AEI/FEDER, EU], (co-funded with FEDER funds); Acciones de dinamización “REDES DE EXCELENCIA” CONSOLIDER- CGL2015-71523-REDC from the Spanish Ministerio de Economía, Industria y Competitividad. FHC was supported by APOSTD/2018/186 Post-doctoral fellowship from Generalitat Valenciana.

Competing interests: The authors declare they have no competing interests.

Figure Legends:

Figure 1: Taxonomic affiliation and predicted hosts of at the *bona fide* viral scaffolds. A) Bubble plot depicting computational host predictions obtained for viral scaffolds. B) Bubble plot depicting taxonomic assignments of the scaffolds based on percentage of matched proteins and average amino acid identity to protein sequences from viral families in the NCBI-nr database.

Figure 2: Viral community composition profile across the depth gradient. Bar plots depicting abundances in viromes based on raw read annotation against the database of assembled viral scaffolds. A) Scaffold abundances were grouped according to the phylum level putative hosts of

353 viral scaffolds. B) Scaffold abundances were grouped according to the family level taxonomic
354 affiliation of viral scaffolds. Only taxa that displayed relative abundances equal or above 0.1% are
355 shown.

356

357 **Figure 3:** Viral scaffold abundances in viral and cellular metagenomes from the depth gradient. A)
358 Scatter-plots depicting the relative abundances of viral sequences in the viral (Y axis) and cellular
359 (X axis) metagenomes. B) Boxplots depicting the ratio between abundances in the cellular and viral
360 fractions for each sample. Boxes depict the median, the first and third quartiles. Whiskers extend to
361 1.5 of the interquartile ranges. Outliers are represented as dots above or below whiskers.

362

363 **Figure 4:** pN/pS values of viral genes differ among functional categories.. A) Barplots depict the
364 median pN/pS values of the functional categories of each sample for the cellular and viral fractions
365 B) Median pN/pS values of proteins grouped by sampling site and broad functional category for the
366 viral fraction only. C) Median pN/pS values of proteins grouped by sampling site and specific
367 functional category for the viral fraction only. Only proteins derived from the set of *bona fide* viral
368 sequences were included in these analyses. When calculating medians only proteins that displayed
369 pN and pS values above 0 were included. Also, only proteins with a total number of polymorphic
370 sites equal or above 1 and percentage of polymorphic sites equal or above 1% were included, so to
371 avoid estimating pN/pS values based only on a small fraction of protein length. Median values
372 obtained from less than three proteins were omitted.

373

374 **Figure 5:** Micro-diversity patterns within a group of homologous tail proteins. X axis depicts the
375 amino acid position along proteins. Y axis depicts the frequency of the reference amino acid among
376 the viral population from each sample. Valleys in the plot represent areas that concentrate non-
377 synonymous mutations, possibly driven by positive selection favouring mutations that modify or

378 expand host-range.

379

380 **Figure 6:** Conceptual model summarizing the observed patterns of micro-diversity in marine viral
 381 genomes across the depth gradient. Different capsid colours represent different viral species.
 382 Different colours for receptor binding proteins, auxiliary metabolic genes and replication proteins
 383 represent different isoforms of the same protein created by non-synonymous mutations. Surface
 384 samples have intermediate densities of viral particles and intermediate species diversity, this sample
 385 displayed the lowest degree of micro-diversity for all functional categories. DCM samples have the
 386 highest density of viral particles but the lowest species diversity. These samples displayed the
 387 highest degree of micro-diversity among receptor binding proteins. Deep samples have the lowest
 388 density of viral particles but highest species diversity. This sample displayed the highest degree of
 389 micro-diversity among metabolic and replication proteins.

390

391 **Table 1:** Characteristics of virome assemblies.

392	Depth	15m	45m	60m	2000m
	Scaffolds	6038	1801	1419	1005
393	N50 (Kbp)	10.7	9	9.1	9.4
	Max Scaffold Length (Kbp)	110.8	121.6	54.4	56.2
394	Assembly size (Mbp)	61.5	16.5	13	9.3
	PEGs	80599	21749	18478	12526
395	Mean Scaffold GC%	36.9	33.4	36.2	45.8

396

References

1. Breitbart M. Marine viruses: truth or dare. *Mar Sci* 2012; **4**: 425–448.
2. Suttle CA. Viruses in the sea. *Nature* 2005; **437**: 356–361.
3. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. *Science* 2015; **348**: 1261498.
4. Luo E, Aylward FO, Mende DR, Delong EF. Bacteriophage Distributions and Temporal Variability in the Ocean’s Interior. *MBio* 2017; **8**: 1–13.
5. Hurwitz BL, Brum JR, Sullivan MB. Depth-stratified functional and taxonomic niche specialization in the ‘core’ and ‘flexible’ Pacific Ocean Virome. *ISME J* 2015; **9**: 472–484.
6. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science (80-)* 2015; **348**: 1–10.
7. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Determinants of community structure in the global plankton interactome. *Science (80-)* 2015; **348**: 1262073.
8. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A, Rodriguez-Valera F. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. *Microbiome* 2018; **6**: 128.
9. Lara E, Vaqué D, Sà EL, Boras JA, Gomes A, Borrull E, et al. Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci Adv* 2017; **3**: e1602565.
10. Winter C, Moeseneder MM, Herndl GJ, Weinbauer MG. Relationship of geographic distance, depth, temperature, and viruses with prokaryotic communities in the eastern tropical Atlantic Ocean. *Microb Ecol* 2008; **56**: 383–389.
11. De Corte D, Sintes E, Yokokawa T, Reinthaler T, Herndl GJ. Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect. *ISME J* 2012; **6**: 1566–1577.
12. Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* 2013; **8**: e57355.
13. Martiny JBH, Riemann L, Marston MF, Middelboe M. Antagonistic Coevolution of Marine Planktonic Viruses and Their Hosts. *Ann Rev Mar Sci* 2014; **6**: 393–414.
14. Enav H, Kirzner S, Lindell D, Mandel-gutfreund Y, Beja O. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. *Nat Commun* 2018; **9**: 1–27.
15. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* 2016; **40**: 258–272.

- 431 16. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the marine virosphere using
432 metagenomics. *PLoS Genet* 2013; **9**: e1003987.
- 433 17. Hannigan GD, Zheng Q, Meisel JS, Minot SS, Bushman FD, Grice EA. Evolutionary and
434 functional implications of hypervariable loci within the skin virome. *PeerJ* 2017; **5**: e2959.
- 435 18. Rubino F, Carberry C, M Waters S, Kenny D, McCabe MS, Creevey CJ. Divergent functional
436 isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome.
437 *ISME J* 2017; **11**: 932–944.
- 438 19. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic
439 variation landscape of the human gut microbiome. *Nature* 2013; **493**: 45–50.
- 440 20. Brum JR, Ignacio-Espinoza JC, Kim E-H, Trubl G, Jones RM, Roux S, et al. Illuminating
441 structural proteins in viral “dark matter” with metaproteomics. *Proc Natl Acad Sci* 2016; **113**:
442 2436–2441.
- 443 21. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, et al. Phage auxiliary
444 metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl*
445 *Acad Sci* 2011; **108**: E757–E764.
- 446 22. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine
447 microbial realm. *Nat Microbiol* 2018; 1.
- 448 23. Roos WH, Ivanovska IL, Evilevitch A, Wuite GJL. Viral capsids: Mechanical characteristics,
449 genome packaging and delivery mechanisms. *Cell Mol Life Sci* 2007; **64**: 1484–1497.
- 450 24. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R, et al. Targeting
451 mechanisms of tailed bacteriophages. *Nat Rev Microbiol* 2018.
- 452 25. De Sordi L, Khanna V, Debarbieux L. The Gut Microbiota Facilitates Drifts in the Genetic
453 Diversity and Infectivity of Bacterial Viruses. *Cell Host Microbe* 2017; **22**: 801–808.e3.
- 454 26. Marston MF, Pierciey FJ, Shepard A, Gearin G, Qi J, Yandava C, et al. Rapid diversification
455 of coevolving marine *Synechococcus* and a virus. *Proc Natl Acad Sci* 2012; **109**: 4544–4549.
- 456 27. Nunoura T, Takaki Y, Hirai M, Shimamura S, Makabe A, Koide O, et al. Hadal biosphere:
457 Insight into the microbial ecosystem in the deepest ocean on Earth. *Proc Natl Acad Sci U S A*
458 2015; **112**: E1230-1236.
- 459 28. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and
460 dark ocean. *Genome Biol* 2013; **14**: R123.
- 461 29. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF,
462 Rohwer F, et al. Explaining microbial population genomics through phage predation. *Nat Rev*
463 *Microbiol* 2009; **7**: 828–836.
- 464 30. Wright RCT, Brockhurst MA, Harrison E. Ecological conditions determine extinction risk in
465 co-evolving bacteria-phage populations. *BMC Evol Biol* 2016; **16**: 1–6.
- 466 31. Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, et al. Antagonistic

467 coevolution accelerates molecular evolution. *Nature* 2010; **464**: 275–278.

468 32. Coutinho FH, Silveira CB, Gregoracci GB, Edwards RA, Brussaard CPD, Dutilh BE, et al.
469 Marine viruses discovered through metagenomics shed light on viral strategies throughout
470 the oceans. *Nat Commun* 2017; **8**: 1–12.

471 33. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence
472 data. *Bioinformatics* 2014; **30**: 2114–2120.

473 34. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: A
474 New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J*
475 *Comput Biol* 2012; **19**: 455–477.

476 35. Buchfink B, Xie C, Huson DH. Fast and Sensitive Protein Alignment using DIAMOND. *Nat*
477 *Methods* 2015; **12**: 59–60.

478 36. Grazziotin AL, Koonin E V., Kristensen DM. Prokaryotic Virus Orthologous Groups
479 (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids*
480 *Res* 2017; **45**: D491–D498.

481 37. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web
482 server: 2015 update. *Nucleic Acids Res* 2015; **43**: W30--W38.

483 38. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
484 reconstructing single genomes from complex microbial communities. *PeerJ* 2015; **3**: e1165.

485 39. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the
486 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
487 *Genome Res* 2015; **25**: 1043–55.

488 40. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic
489 gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**:
490 119.

491 41. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et
492 al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the
493 oceans. *Nat Commun* 2017; **8**.

494 42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;
495 **9**: 357–9.
496

Host

Taxon

Microviridae
Phycodnaviridae
Siphoviridae
Podoviridae
Myoviridae

15m

45m

60m

2000m

Sample

log10(bp_Assigned)

log2(Sequences_Assigned)

Amino acid position

Surface

bioRxiv preprint doi: <https://doi.org/10.1101/551291>; this version posted February 15, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DCM

Deep