# Genome-resolved viral ecology in a marine oxygen minimum zone (OMZ)

Dean Vik[1]*, Maria Consuelo Gazitúa[1,2]*, Christine L. Sun[1]*, Montserrat Aldunate[3,4], Margaret R. Mulholland[5], Osvaldo Ulloa[3,4], Matthew B. Sullivan[1,6,7,#].

[1]Department of Microbiology, The Ohio State University.

[2]Current affiliation: Viromica Consulting, Santiago, Chile.

[3]Department of Oceanography, Universidad de Concepción.

[4]Millennium Institute of Oceanography, Universidad de Concepción.

[5]Department of Ocean, Earth and Atmospheric Sciences, Old Dominion University.

[6]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University.

[7]Center of Microbiome Science, The Ohio State University

*Authors contributed equally.

[#]*Corresponding Author*: Matthew Sullivan; R914 Riffe Building, 496 W 12th Ave., Columbus, OH 43210; sullivan.948@osu.edu

**Running Title: Diversity and ecology of novel viruses from an OMZ**

Submitted to Environmental Microbiology as a Research Article

## Originality-Significance Statement:

Marine oxygen minimum zones (OMZs) are unique and important ocean ecosystems where microbes drive climate-altering nutrient transformations. This study provides a baseline, deeply sequenced viral metagenomic dataset and reference viral

genomes to assess ecological change and drivers across the oxygenated surface to de-oxygenated deep waters of the Eastern Tropical South Pacific (ETSP) OMZ. Community ecological assessment of the ETSP viromes reveals a relatively low diversity viral community with a high degree of endemic populations in the OMZ waters.

**Summary:**

Oxygen minimum zones (OMZs) are critical to marine nitrogen cycling and global climate change. While OMZ microbial communities are relatively well-studied, little is known about their viruses. Here we assess the viral community ecology of 22 deeply sequenced viral metagenomes along a gradient of surface oxygenated to anoxic waters (< 0.02 μmol/L $O_2$) in the Eastern Tropical South Pacific (ETSP) OMZ. We identified 46,127 viral populations (>5 kb), which augments the known viruses at this site by 10-fold. ETSP viral communities clustered into 6 groups that correspond to oceanographic features, with 3 clusters representing samples from suboxic to anoxic waters. Oxygen concentration was the predominant environmental feature driving viral community structure. Alpha and beta diversity of viral communities in the anoxic zone were lower than in surface waters, which parallels the low microbial diversity seen in other studies. Viruses were largely endemic as few (6% of viruses from this study) were found in at least another marine metagenome, and of those, most (77%) were restricted to other OMZs. Together these findings provide an ecological baseline for viral community structure, drivers and population variability in OMZs that will help future studies assess the role of viruses in these climate-critical environments.

## Introduction

Oxygen deficient regions of the ocean play a vital role in regulating the ocean nitrogen budget and greenhouse gas emission (Wright *et al.*, 2012). These low oxygen regions termed Oxygen Minimum Zones (OMZ) result from a combination of thermal stratification of the water column, low circulation, temperature- and wind-driven upwelling currents, and heterotrophic consumption of surface water primary production in deep water (Wyrtki, 1965; Kessler, 2006; Karstensen *et al.*, 2008; Paulmier & Ruiz-Pino, 2009; Czeschel *et al.*, 2011). Oxygen concentrations in OMZs can reach below the detection limit of a few nanomoles per liter in certain regions such as the Eastern Tropical South Pacific (ETSP) (Revsbech *et al*., 2009; Thamdrup *et al.*, 2012; Ulloa *et al.*, 2012), creating anoxic marine zones (AMZs). Problematically, OMZs have expanded over the last 50 years as a result of increased ocean stratification from rising surface water temperatures (Stramma *et al.*, 2008, Wright *et al.*, 2012, Ulloa *et al.*, 2012). As OMZs expand, the metabolisms found therein – anaerobic ammonium oxidation (anammox) and denitrification – result in large-scale nitrogen loss through the increased production of $N_2$ and the potent greenhouse gas $N_2O$ (Lam and Kuypers, 2011). Thus, the factors moderating microbe-mediated nutrient cycling in OMZs require rigorous examination to better understand the impact of OMZs on climatic trends.

The biological factors responsible for the development and maintenance of OMZs are almost exclusively a result of microbial activity (Zakem *et al.*, 2019). As oxygen is removed from the environment by heterotrophs, other electron acceptors (such as nitrate and sulfate) that have progressively lower electron affinities and energy potentials are used. This results in a gradient of electron acceptors and redox chemistry across the

depth gradient, referred to as the redoxycline. Due to the reduction in free energy, the biogeochemistry of OMZs is almost entirely dictated by the microbial metabolisms stratified along the redoxycline, rather than macrofaunal respiration (Hawley *et al.*, 2014). This reduction in energy may be expected to be associated with a reduction in community size and diversity possibly due to a depletion in niche space (Rabosky, 2009; Beman & Carolan, 2013). However, major trends in microbial diversity across OMZs remain unclear, as diversity has been shown to either increase or decrease with the reduction of oxygen concentration (Stevens & Ulloa, 2008; Bryant *et al.*, 2012). Nevertheless, it is plausible that diversity trends follow patterns in productivity, *i.e.*, higher relative diversity in the surface chlorophyll maximum (SCM) and the deep chlorophyll maximum (DCM) (Chase & Leibold, 2002; Walsh *et al.*, 2016), or the availability of niche space, which appears to peak at the interface between oxygenated and anoxic water, a transition environment where a wide array of metabolisms may persist (Bertagnolli & Stewart, 2018).

Importantly, viruses have also been shown to play major roles in both bottom-up and top-down mechanisms controlling microbial communities. In the surface oceans, viruses mediate microbial population dynamics and metabolism (Suttle, 2007; Hurwitz *et al.*, 2013) through viral lysis, which in addition to protist grazing, results in microbial mortality rates that are proportional to growth rates at ~1-2 $day^{-1}$ (Ducklow & Hill, 1985; Cole *et al.*, 1988; Suttle, 1994; Fuhrman & Noble, 1995). The result of this lysis is the redirection of fixed carbon away from macrofaunal production and into both microbial respiration and carbon export (Azam *et al.*, 1983; Fuhrman, 1992; Guidi *et al.*, 2016). In addition, viruses have been shown to encode host-derived auxiliary metabolic genes

(AMGs), with a few notable examples of viruses likely associated with sulfur and nitrogen cycling (Roux *et al.*, 2016; Ahlgren *et al.*, 2019). Though data to date suggest that OMZ viruses are likely important, a foundational ecological perspective on viruses in these habitats is lacking.

To date, only two genomic studies have explored community dynamics of viruses in OMZ systems (Cassman *et al.*, 2012; Roux *et al.*, 2014). Both studies were relatively small in terms of samples collected, sequencing available, and viruses recovered, but both led to significant advances in our understanding of OMZ viruses. Specifically, Roux *et al.* (2014) focused on viruses that were recovered from 127 single cell amplified genomes from uncultivated SUP05 bacteria in a model OMZ ecosystem while Cassman *et al.* (2012) examined the diversity and size of viral communities in the ETSP OMZ region through metagenomics. Fortunately, sequencing costs and analytics has improved considerably since these initial studies, which warrants re-investigation of these viral communities. Here, we deeply sequence viral metagenomes from 22 seawater samples to provide 420-fold more sequencing data from these environments to establish genome-resolved datasets for exploring viral community and population ecology along oxygenated to anoxic marine waters of the ETSP OMZ.

## Results and Discussion

Samples were collected from six stations spanning a transect from coastal to pelagic waters in the ETSP OMZ region off the coast of Lima, Peru (**Figure 1A**, see Experimental Procedures). The depths at which samples were collected were determined by distinct oceanographic features (**Figure 1B, Table S1, Figure S1**). At most stations a

sample was collected from the surface chlorophyll maximum, the upper oxycline, the upper OMZ (with or without a secondary deep chlorophyll maximum), and the core of the OMZ (**Figure 1**). Viral concentrates produced for each sample were sequenced with Illumina HiSeq 2000 to produce an average of 7.2 Gb bases per sample (**Table S2**). To provide context with other large studies, the Global Ocean Virome dataset had an average of 8.9 and 27.2 Gb per sample, on their first and second versions, respectively (Roux *et al*., 2016; Gregory *et al.*, 2019). Reads were assembled into scaffolds, from which viruses were identified and then clustered into viral populations (see Experimental Procedures) that represent approximately species level viral taxonomy (Brum *et al.*, 2015; Gregory *et al.*, 2016; Gregory *et al.*, 2019). In total, we recovered 46,127 viral populations of at least 5 kb in length and used these for all analyses in this study.

The relative abundances of the viral populations in each sample were calculated and normalized across all samples (see Experimental Procedures) and used as input for biogeographical and diversity analyses (**Table S3**). We hypothesized that viruses would cluster into distinct communities from the anoxic and oxic waters due to the environmental differences of these marine habitats (Bertagnolli & Stewart, 2018). This would be similar to previous studies that have shown OMZs have unique microbial communities, compared to more oxygenated surface waters (Madrid *et al.*, 2001; Fuchs *et al.*, 2005; Stevens & Ulloa, 2008; Wright *et al.*, 2012). Overall, 6 main clusters were detected via hierarchical clustering (**Figure 2**), which is relatively consistent with those predicted from other statistics including a gap statistic (**Figure S2**) and an affinity propagation analysis (**Figure S3**). As expected, the anoxic waters of the OMZ were significantly distinct from the rest of the samples (**Figure 2**). Within the OMZ samples, there were 3 sub-clusters:

139 samples from coastal OMZs (OMC_C cluster), samples from the upper OMZ with a DCM
140 (uOMZd cluster), and samples from the remaining upper and core OMZ from pelagic
141 waters (OMZ_P cluster) (see Experimental Procedures for cluster name information).
142 Five of the samples from OMZ_P cluster also have relatively high concentrations of nitrite
143 (**Table S1** and **Figure S1F**), and so represent an anoxic marine zone (AMZ) (Thamdrup
144 *et al.*, 2012; Ulloa *et al.,* 2012). However, for the rest of the OMZ_P cluster samples, nitrite
145 measurements are missing (see Experimental Procedures). Other clustering differences
146 showed that pelagic surface waters differed from the oxycline and coastal surface waters.
147 The main clusters support the conclusion that viral communities that exist in OMZs are
148 relatively distinct from those communities found in the ocean surface.

149 We next evaluated our viral communities with various diversity measures, including
150 evenness, alpha diversity, and beta diversity (**Figure 3A, 3B, 3C**, respectively). The
151 diversity metrics used were specifically selected to minimize the impact of varying
152 sequencing depth on diversity estimations. In terms of the entire system, species
153 accumulation analysis revealed only a 2% increase in species recovery in the last of the
154 22 randomly permuted sub-samples, indicating sequencing approached saturation
155 (**Figure S4)**. Statistically robust species accumulation analysis was not possible at the
156 individual community scale due to a lack of samples within a given habitat. Evenness did
157 not significantly differ between samples from oxygenated regions and the OMZ (**Figure**
158 **3A**, Kruskal-Wallis p-value = 0.742). The evenness was nearly 1 in all cases (range 0.965-
159 0.978, mean 0.974), which indicates that no community or sample cluster had a high
160 relative proportion of dominant viral populations.

Alpha diversity, a measure of diversity within a community, was calculated (**Figure 3B**) using the inverse Simpson's concentration (see Experimental Procedures), which facilitates a relatively unbiased comparison of alpha diversity across communities despite uneven sequencing depth (Haegeman *et al.*, 2013). Alpha diversity did not differ between communities (Kruskal-Wallis p=NS) except for the OMZ_P cluster, which exhibited a significantly lower alpha diversity (Kruskal-Wallis p = 0.01). Beta diversity, a measure of the amount of the total diversity accounted for by a given community, was estimated (**Figure 3C**) via multivariate dispersion, which leverages distance-based ordination techniques to derive the average distance of all samples in a community from the community centroid (Anderson *et al.*, 2006). Beta diversity was significantly lower in the OMZ regions than in the surface waters regarding both population composition (modified Gower's distance $\log_{10}$ p = 0.005) and abundance (modified Gower's distance $\log_2$, p = 0.006) (Anderson *et al.*, 2006). The lower alpha and beta diversities for OMZ samples, consistent with a previous study (Cassman *et al.*, 2012), indicate that niche space is reduced as energetics of the system decreases along the redoxycline. A similar trend has also been suggested for microbial community distributions in previous studies (Bryant *et al.*, 2012; Beman & Carolan, 2013; Bertagnolli & Stewart, 2018).

Because the hierarchical clustering and diversity measures indicated that OMZ viral communities were distinct and relatively low in diversity, we next sought to identify the environmental features driving these patterns. To this end, we created ordination plots for the samples, as well as for the environmental features data (**Figure 4A, 4B,** *respectively*; stress plots **Figure S5**). Non-metric multidimensional scaling analysis (NMDS) with Bray-Curtis dissimilarity revealed that the 6 viral clusters (**Figure 2**) were

184 distinct (**Figure 4A,** stress plots **Figure S5**). These findings were also supported
185 statistically by ANOSIM (community R stat 0.855, p = 0.001 after 999 permutations) and
186 by MRPP (distances within groups 0.528, between groups 0.872, overall 0.754, chance
187 corrected within group agreement 0.351, p = 0.001). Together, these findings suggest
188 ecologically distinct viral communities exist in samples within our dataset.

189 In order to determine whether this separation between clusters could be explained
190 by the measured environmental features, we compared ordinations based on viral
191 populations (**Figure 4A**) and environmental features (**Figure 4B**). Similarities among the
192 structures of these ordination plots and their underlying dissimilarity/distance matrices
193 would indicate an environmental influence on the distribution of the viral communities.
194 The structure of the viral community and environmental features ordination plots
195 (Procrustes sum of squares 0.194, correlation 0.898, p = 0.001) (**Figure S6**) and trends
196 in the dissimilarity/distance matrices (Mantels R = 0.675 p=0.001) were similar with the
197 main structural difference being that the OMZ_P cluster was collapsed in the ordination
198 created from the environmental features rather than separated into different OMZ sub
199 groups. These results indicate that environmental features impact viral community
200 distributions. While paired microbial community data were not available for comparison
201 here, we posit, as done previously for surface ocean viral communities (Brum *et al.,* 2013,
202 Brum *et al.*, 2015), that this reflects the environment associated biogeographical
203 distribution of the resident microbial populations rather than a direct environmental impact
204 on the viral populations.

205 Temperature and oxygen were the most descriptive gradients implicated in driving
206 the biogeographical distributions of the viral communities (GAM and Pearson correlation;

207 temperature p<0.0001, oxygen p<0.001). The co-variation of these two factors is
208 expected in our system, as both decrease with depth, from surface to core OMZ waters
209 (**Table S1**, **Figure S1A**, and **Figure S1B**). We addressed this co-variance by comparing
210 the structure and agreement between NMDS ordinations of the environmental features
211 and the viral community distributions again, but with the iterative removal of these
212 parameters (removal of temperature in **Figure 4C,** removal of oxygen in **Figure 4D**) to
213 determine which of these features was most important in retaining the similarity between
214 these ordination analysis. The removal of temperature had an almost negligible impact
215 on the relationship between the environmental features and the community distributions
216 (Procrustes sum of squares 0.195, correlation 0.897, p = 0.001) indicating a relatively
217 lower overall impact on the viral community structure. However, removal of oxygen
218 reduced the relationship considerably (Procrustes sum of squares 0.385, correlation
219 0.784, p = 0.001), suggesting that, not surprisingly, oxygen was the most important driver
220 of viral community composition (particularly the distinction between the communities
221 found in the surface oxygenated water and the OMZ). Again, presumably this is due to
222 the effect of oxygen on microbial populations rather than oxygen directly impacting the
223 viruses.

224 Previous studies have indicated that OMZs have unique microbial and viral
225 communities compared to the rest of the ocean (Madrid *et al.*, 2001; Fuchs *et al.*, 2005;
226 Steven and Ulloa, 2008; Cassman *et al.*, 2012; Wright *et al.*, 2012). In order to determine
227 to what extent the ETSP viral communities overlap with other oceanic viral communities,
228 we evaluated whether our ETSP OMZ virus populations were among the ~488K viral
229 populations available in the Global Ocean Virome version 2.0 dataset (Gregory *et al.*,

2019), and if so, assessed their biogeography. Using MMseq2, we identified viral populations from our study that shared 95% identity (over 50% of the ETSP query protein coding sequence) with the GOV2.0 populations (see Experimental Procedures). In total 2,763 of our 46,127 ETSP viral populations were also observed in the GOV2.0 dataset (**Figure 5A**), with about half (1,466) from OMZ samples (**Table S4**). Among these shared ETSP OMZ viruses, most (77%) were only found in other OMZ samples ($O_2$ concentration below 10 µmol/L) (Helly & Levin, 2004) (**Table S4**). This shows that virtually all of our ETSP OMZ viral populations are endemic to OMZs, which is consistent with prior work (Cassman *et al.*, 2012) where viral genotypes were evaluated (rather than viral populations) and where the geographical context was drastically reduced (only 4,552 viral genotypes were available for comparison as opposed to the 46,127 assessed here).

To further explore how the identified ETSP viral populations compared to known viruses in the RefSeq database, we used gene sharing networks where viral clusters (VCs) are approximately genus level taxonomy (Bolduc *et al.*, 2017, Jang *et al.*, 2019). With the sequences from viral RefSeq (v85) and the 10kb and larger viral populations in this study, these analyses clustered 10,632 viral populations into 1,465 VCs (**Figure 5B**), 4,020 viral populations into outliers (where populations were assigned to a VC but shared fewer similar proteins than the bulk of the cluster), and 482 viral populations into singletons (populations that did not cluster with any other sequences). Only 27 VCs included known reference viral sequences, which suggests that 98% (1438/1465) of the VCs derived from the ETSP OMZ dataset likely represent novel viral genera. If true, this is a 5-fold expansion of viral genus sequence space recovered from our analysis, as compared to RefSeq. Within the ETSP, 28% of the VCs identified in the OMZ sample

were not present in any of the surface or oxycline samples further suggesting that the OMZ sample is distinct from the oxic habitats

Finally, we sought to use read mapping against our expanded dataset of ETSP viral populations to provide a very gross metric of population stability in these systems as assessed against the previous shallow viral metagenomic sequencing (Cassman *et al.* 2012). Less than 3% of the reads from *Cassman et al.* recruited to the ETSP viral populations, which corresponded to either as little as 1 or as much as 698 ETSP viral populations (using conservative vs permissive coverage cut-offs, see Experimental Procedures) being present in the prior dataset. This may represent high turnover in viral populations, but the inference does suffer from ascertainment bias due to the minimal sequencing available in the prior study.

## Conclusions

OMZs have been expanding over the last 50 years as a result of rapidly escalating anthropogenic carbon dioxide emissions increasing atmospheric temperatures, which in turn has increased ocean temperatures and stratification – features that select for OMZ formation and expansion (Schmidtko *et al.* 2017). With these ocean changes, and the fact that the oceans are a major carbon dioxide sink where microbes control that carbon's fate, it becomes critical to understand how microorganisms will respond to and impact such changes (Cavicchioli *et al*., 2019). Viruses that infect these microbes also become important to understand. In this study, we present the largest survey of viruses from an OMZ – 46,127 unique viral populations across 6 stations at the ETSP OMZ (**Figure 1**). In ETSP, OMZ viral communities were distinct and relatively low in diversity compared to

276 oxygenated, surface waters (**Figure 2**; **Figure 3),** with oxygen as the most important
277 driver of viral community composition (**Figure 4D**). These viruses are more similar to
278 viruses from other OMZs and are novel (**Figure 5A**, **Figure 5B**). This is congruent with
279 previous studies that have shown OMZs have unique and low diversity microbial
280 communities, compared to the rest of the ocean (Madrid *et al.*, 2001; Fuchs *et al.*, 2005;
281 Wright *et al.*, 2012), which may result from the reduced redox potential of the prevalent
282 electron acceptors in OMZs.

283 Though a large study, limitations are as follows. First, we cannot link the viruses
284 to their microbial hosts because there is a lack of metagenomic samples from which we
285 could construct metagenomically assembled genomes (MAGs), and such co-sampled
286 MAGs improve virus-host linkages typically 5-fold or more (Emerson *et al*, 2018). Though
287 AMGs are important in the surface oceans (reviewed in Rosenwasser *et al.,* 2016, Hurwitz
288 & U'Ren, 2016), they were not studied here as they are the focus of a parallel study from
289 the same dataset that revealed viral genomes that contain AMGs associated with the
290 denitrification, nitrification, and other nitrogen cycle processes, suggesting that these
291 OMZ viruses influence the nitrogen cycle (Gazitua *et al, submitted*). Future work in OMZs
292 should be enabled by our current findings and the vast sequence database of reference
293 virus genomes that will empower a new generation of researchers to evaluate viral roles
294 in modulating microbial population dynamics and biogeochemical cycling climate-critical
295 OMZs as they expand due to climate change.

296

297 Acknowledgements.

We thank Sullivan Lab members and Heather Maughan for comments on the manuscript, and the crew of the *R/V Atlantis* for the sampling opportunity and support at sea. This work was funded in part by awards from NSF Biological Oceanography to MRM (#1356056), from the Agouron Institute to OU and MBS, a Gordon and Betty Moore Foundation Investigator Award (#3790) and NSF Biological Oceanography Awards (#0940390, #1536989) to MBS.

## Experimental Procedures

### *Sample collection*

On December 31, 2014 – January 22, 2015, six stations spanning a transect from coastal to pelagic waters in the ETSP OMZ region (off the coast of Peru) were sampled during the cruise AT-2626 aboard the *R/V Atlantis*. Volumes of 20 liters were collected using a pump profiling system (PPS), equipped with a Seabird SBE 25 Conductivity Temperature Depth (CTD), a WET Labs ECO-AFL/FL fluorometer, a Seabird SBE 43 dissolved oxygen sensor and a STOX sensor for nanomolar scale measurements of oxygen concentrations (detection limit of of 1-10 nmol $L^{-1}$ $O_2$). Oxygen detection limits using this sensor was about 0.02 µmol/L. High nitrite concentrations are found in waters with <50 nM of oxygen (Thamdrup *et al.,* 2012). However, nitrite values in our sampling were only available for 3 samples (**Table S1**). In the cases where nitrite values were not obtained for a sample, nitrite values from adjacent depths (±10m) were used if available.

Concentrations of dissolved oxygen, nitrite, and other metadata can be found in **Table S1**. Sampling depths were selected according to variation in oxygen and chlorophyll concentrations, such as the surface chlorophyll maximum, the suboxic upper

oxycline, the anoxic upper OMZ (with or without a deep chlorophyll maximum), and the core of the OMZ (**Figure 1**, **Table S1**). Samples for nitrite were filtered using a 0.2 µm cartridge filter. Filtrate was collected into sterile Falcon™ tubes and stored upright at -20°C until analysis. Nitrite concentrations were measured using an Astoria-Pacific autoanalyzer and standard colorimetric methods (Parsons *et al.*, 1984), with a limit of detection (LOD) of 0.02 µM $NO_2^-$, (3σ, n = 7) (Selden *et al., submitted*).

Viral particles of the 22 samples were concentrated from the filtrate by iron chloride flocculation (John *et al.*, 2011; Duhaime & Sullivan, 2012). Viral concentrates were then collected on a 1.0µm, 142mm, polycarbonate (PC) membrane (GE Water and Process Technologies, Trevose, PA, USA; Cat. #K10CP14220) and stored at 4 °C. The viral-iron precipitates were resuspended overnight in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mgcl, 0.2 M ascorbic acid, pH 6.0), rotating in the dark at 4°C. DNaseI at 100U $ml^{-1}$ concentration was added to the final viral concentrate to remove any free DNA (Hurwitz *et al.*, 2013). Viral DNA was then extracted using a Wizard DNA purification kit (Promega) with 1 ml resin to 0.5 ml sample. Samples yielding more than 1µg DNA (7 out of 22) were further purified using CsCl buoyant density gradients (Hurwitz *et al.*, 2013). Viral contigs detected in the CsCl purified samples were retained only if they clustered into populations with viruses from the non-purified samples and became the representative contig of the population (the longest contig) (see *Assembly and processing*). Ecological analyses were then performed using the only 22 DNase-purified samples and representative contigs from all samples. DNA samples were submitted to JGI for library preparation and Nextera sequencing on an Illumina Hiseq 2000.

*Assembly and processing*

Data processing and metagenomic analyses were performed using high-memory computer nodes from the Ohio State Supercomputer Center (Ohio State Supercomputer Center). Trimmomatic version 0.33 was used to remove Nextera adapters, to split reads into paired and unpaired groups, and to trim reads with low quality regions below a Phred score threshold of 15, using a sliding window of 4 bases (Bolger *et al.*, 2014). Reads from each sample, with or without the CsCl purification step, were then assembled with Spades version 3.11.1, using the --meta option with paired end reads and the --sc and --careful options with unpaired reads, both with kmers of 21, 33, and 55 bases (Nurk *et al.*, 2017). The resulting scaffolds were then clustered into population scale groups at 95% ANI over 80% of the shorter sequence using an in-house wrapper script for nucmer, run with default settings (Kurtz *et al.*, 2004; Brum *et al.*, 2015).

*Viral identification*

Population contigs larger than 5 kb were processed with the viral identification tools VirSorter and Virfinder (Roux *et al.,* 2015; Ren *et al.*, 2017), and CAT (Cambuy *et al.,* 2016), based on the steps described in Gregory *et al.* (2019). Populations with VirSorter categories 1 or 2, or with a VirFinder score ≥ 0.9 and a p-value < 0.05 were considered to be viral, as well those with VirSorter categories 3 to 6 and a VirFinder score >= 0.7 and a p-value < 0.05. Contigs with VirSorter categories 4 and 5 and a VirFinder score < 0.7 were manually curated to check if they were misannotated as prophages. If so, they were re-assigned to categories 1 or 2, respectively, and considered viral. Contigs with a VirFinder score between 0.7 and 0.9 (p-value < 0.05), without a VirSorter category, were

run through CAT and those with < 60% of the genome classified as bacterial, archaeal, or eukaryotic (based on an average gene size of 1000) were considered viral. **Table S5** shows the VirSorter, VirFinder, and CAT assignments for each viral population.

*Viral relative abundances*

In order to determine the relative abundance of each viral population larger than 5kb, the final viral populations were concatenated and then used as a database to recruit the quality trimmed reads using a custom wrapper script for bowtie2, which automatically determines groupings of paired and unpaired reads (Langmead & Salzberg, 2012). The resulting coverage files were then converted into a relative abundance table with the per population coverages using a custom wrapper script for BamM (https://github.com/ecogenomics/BamM). Coverages were calculated using the tpmean algorithm and adjusted coverages were calculated based on the coverage of each viral population per Gb of metagenome sequenced. Relative coverages were only reported if more than 75% of the population had at least 5x coverage, with at least 90% identity over 90% of the read. For reference, **Figure S7** shows the sequencing depth and number of reads that mapped to viruses for each sample.

In order to determine the fraction of the ETSP viral population that was also identified in the Cassman *et al.* (2012) study, the high quality, trimmed reads from Cassman *et al.* 2012 were downloaded from MGrast and recruited to the ETSP populations as described above for the abundance estimates. Due to a low number of ETSP viral populations being recovered with the stringent coverage threshold above, we

389 eliminated the viral population coverage threshold to allow for more permissive read
390 recruitment with the reads from Cassman *et al.* (2012).
391
392 *Cluster identification*
393 Clusters were inferred using a combination of affinity propagation using the R
394 function APCluster with options negDistMat(r=2) (Frey & Dueck, 2007; Bodenhofer *et al.*,
395 2011) and a gap statistic using the R function clusGap with options kmeans, 10, and B =
396 100, (Tibshirani *et al.*, 2001), resulting in an estimation of between 5 and 8 statistically
397 supported groups. The relative abundances of the viral populations were then used in
398 multiple permutations of a hierarchical clustering analysis with minkowski distances (p=2)
399 to identify an approximation of the viral communities (Suzuki & Shimodaira, 2006).
400 Viromes clustered with an approximately unbiased bootstrap value of 100% were
401 considered viral communities. Viral population distributions among the viral communities
402 were visualized with a heatmap plotted using the R package heatmap3. Bray Curtis
403 distances were then plotted in a similar fashion to further validate the observed clustering
404 patterns (Bray & Curtis, 1957).
405 The names of the 6 clusters in **Figure 2** were generated to be as descriptive as
406 possible, using similar abbreviations from **Figure 1**. The clusters 'OXY_SCM_1' and
407 'OXY_SCM_2' denote clusters consisting of samples from oxycline and surface
408 chlorophyll maximum. The 'SCM' cluster has samples from the surface chlorophyll
409 maximum only. The 'OMZ_C' cluster has OMZ samples from the coastal St16, the
410 'uOMZD' cluster has samples from the upper OMZ with deep chlorophyll maximum, and
411 the 'OMZ_P' has samples from the pelagic OMZ core.

412

413 *Comparison with environmental features*

414 Distances within and between viral communities, as defined by the hierarchical
415 clustering, were evaluated using nonmetric multidimensional scaling with Bray Curtis
416 distances and 999 permutations or until convergence using the R package vegan and the
417 function metaMDS (Field *et al.*, 1982; Oksanen *et al.*, 2018). The statistical significance
418 of the viral community groups was then validated by comparing the within community and
419 between community distances with MRPP and ANOSIM (Mielke *et al.*, 1976; Clarke,
420 1993). Standardized Z-score and raw environmental feature measurements were then
421 correlated with the viral community ordination using maximum linear and GAM non-linear
422 algorithms using the R package envfit and odrisurf respectively (Clarke & Ainsworth,
423 1993). The environmental features that were used for these correlations are found in
424 **Table S1**. Note that nitrite values were not used because they were only available from
425 3 samples directly (other samples had nitrite values taken from adjacent depths).

426 Known co-correlations between significant environmental features were then
427 addressed by comparing distances between samples according to the relative
428 abundances of the viral populations and the measured environmental features
429 (Sunagawa *et al.*, 2015). The standardized and raw environmental features were
430 represented in ordination space using NMDS and Manhattan distances with 999
431 permutations until convergence (Field *et al.*, 1982). Relationships between the
432 standardized or raw environmental features and viral community ordinations were then
433 evaluated, with and without the removal of a specific environmental feature of interest,
434 using a Procrustes analysis and Mantel test (Mantel, 1967; Jackson, 1995). Analyses

conducted with the standardized and raw environmental features were congruent, but more easily interpreted with the raw environmental features, and thus, results from the raw numbers are reported in the main text.

*Alpha diversity, beta diversity, and evenness*

Diversity estimates were based on the relative abundance tables generated via read recruitment. Evenness, as a measure of the relative similarity among population abundances within a community, was calculated manually using equation H/ln(S) where H is the Shannon Wiener diversity index per community, calculated with the vegan diversity application in R using the option index = "shannon", and S is the observed species richness of the community, calculated using the vegan application specnumber in R (Shannon, 1948; Pielou, 1966). Simpson concentration indices were calculated per viral community using the R package vegan and the application diversity with the option index = "invsimpson" (Simpson, 1949). Alpha diversities were represented as the inverse simpson concentration in order to facilitate the representation of statistically significant differences between communities (Jost, 2006) and to mitigate the uncertainties in diversity estimates due to variations in sampling effort (Haegeman *et al.*, 2013).

We then compared beta-diversity among the 6 communities as a measure of the amount of the total diversity within a system accounted for by a given habitat, using a multivariate dispersion analysis. This approach facilitates attributing the observed diversity to only population composition or to population composition and abundance based on modified Gower distances (Anderson *et al.*, 2006). Raw normalized relative abundance tables were first log transformed using the R application “decostand” and

458 options method = "log" and logbase = 2 or 10. Distances were then calculated using the
459 vegan application “vegdist” and the options method = "altGower" in R. The multivariate
460 dispersion analysis was then performed on these distance matrices using the vegan
461 application “betadisper”, “anova”, and “permutest” in R with defined groups from the
462 hierarchical clustering analysis above.
463
464 *Endemism within ETSP*
465 Viral populations were clustered into approximately genus level taxonomic groups
466 using the network analytic vConTACT2 (Jang *et al.*, 2019) in order to determine the
467 relative proportion of each viral genus found within a community or shared across ETSP
468 OMZ communities. Viral ORFs were first predicted and translated from the viral
469 populations larger than 10 kb using Prodigal version 2.6.3 with the -p meta option (Hyatt
470 *et al.*, 2010). These predicted ORFs were then used to cluster the 10 kb populations
471 amongst themselves and with viral Refseq version 85 using vConTACT2 with default
472 parameters. Specific viral genera in each community were evident from the resulting
473 network, so the relative abundance of each genus was determined by summing the
474 relative abundances of the viral populations included in each genus. Genus abundance
475 data were then tabulated and visualized using the R packages ggplots (Wickham, 2016).
476 Sequence comparisons were then used to determine the amount of ETSP viral
477 populations larger than 5 kb that were identified in other regions of the ocean. MMseq2
478 using the easy-search command and with a 95% identity over 50% of the query protein
479 coding sequence was used to compare the ETSP viruses with the 488k viral populations
480 identified in the GOV2.0 database (Hauser *et al.*, 2016; Gregory *et al.*, 2019) (**Table S4**).

The abundance and distribution of each GOV2.0 population identified was then evaluated to determine the habitats in which these populations were found. A stacked bar chart was then created to show the proportional abundance of each ETSP population with significant similarity to a population in GOV2.0, and the habitat wherein each GOV2.0 population was identified.

*Data availability*

All high-quality reads and assembled contigs are available on iVirus (CyVerse, https://doi.org/10.25739/mmj5-kt58). Requests for further information should be directed to MBS at sullivan.948@osu.edu.

References

Ahlgren, N. A., Fuchsman, C. A., Rocap, G., & Fuhrman, J. A. (2019) Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *The ISME journal* **13:** 618-631.

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006) Multivariate dispersion as a measure of beta diversity. *Ecology Letters* **9:** 683-693.

Azam, F., Fenchel, T., Field, J. G., Gray, J. S., Meyer-Reil, L. A., & Thingstad, F. (1983) The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series* **10:** 257-263.

Beman, J. M., & Carolan, M. T. (2013) Deoxygenation alters bacterial diversity and community composition in the ocean's largest oxygen minimum zone. *Nature Communications* **4:**

504 Bertagnolli, A. D., & Stewart, F. J. (2018) Microbial niches in marine oxygen minimum
505 zones. *Nature Reviews Microbiology* **16:** 723-729.
506 Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011) APCluster: an R package for
507 affinity propagation clustering. *Bioinformatics* **27:** 2463-2464.
508 Bolduc, B., Jang, H.B., Doulcier, G., You, Z., Roux, S., Sullivan, M.B. (2017) vConTACT:
509 an iVirus tool to classify double-stranded DNA viruses that infect Archaea and
510 Bacteria. *PeerJ* **5**: e3243.
511 Bolger, A. M., Lohse, M., & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina
512 sequence data. *Bioinformatics* **30:** 2114-2120.
513 Bray, J. R., & Curtis, J. T. (1957) An Ordination of the Upland Forest Communities of
514 Southern Wisconsin. *Ecological Monographs* **27:** 325-349.
515 Brum, J., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S., Alberti, A. et al.
516 (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348:**
517 1261498.
518 Brum, J. Schenck, R. & Sullivan, M.B. (2013). Global morphological analysis of marine
519 viruses shows minimal regional variation and dominance of non-tailed viruses.
520 *ISME J* **7**:1738–1751.
521 Bryant, J. A., Stewart, F. J., Eppley, J. M., & DeLong, E. F. (2012) Microbial community
522 phylogenetic and trait diversity declines with depth in a marine oxygen minimum
523 zone. *Ecology* **93:** 1659-1673.
524 Cambuy, D.D., Coutinho, F.H. & Dutilh, B.E (2016) Contig annotation tool CAT robustly
525 classifies assembled metagenomic contigs and long sequences. bioRxiv.
526 https://doi.org/10.1101/072868.

527 Cassman, N., Prieto-Davo, A., Walsh, K., Silva, G. G., Angly, F., Akhter, S. et al. (2012)
528 Oxygen minimum zones harbour novel viral communities with low diversity.
529 *Environmental Microbiology* **14:** 3043-3065.
530 Cavicchioli, R., Ripple, W.J., Timmis, K.N, Azam, F., Bakken, L.R., Baylis, M. et al. (2019)
531 Scientists' warning to humanity: microorganisms and climate change. *Nat Rev*
532 *Microbiol* **17**: 569–586.
533 Chase, J. M., & Leibold, M. A. (2002) Spatial scale dictates the productivity–biodiversity
534 relationship. *Nature* **416:** 427-430.
535 Clarke, K. R. (1993) Non-parametric multivariate analyses of changes in community
536 structure. *Austral Ecology* **18:** 117-143.
537 Clarke, K. R., & Ainsworth, M. (1993) A method of linking multivariate community structure
538 to environmental variables. *Marine Ecology Progress Series* **92:** 205-219.
539 Cole, J. J., Findlay, S., & Pace, M. L. (1988) Bacterial production in fresh and saltwater
540 ecosystems: a cross-system overview. *Marine Ecology Progress Series* **43:** 1-10.
541 Czeschel, R., Stramma, L., Schwarzkopf, F. U., Giese, B. S., Funk, A., & Karstensen, J.
542 (2011) Middepth circulation of the eastern tropical South Pacific and its link to the
543 oxygen minimum zone. *Journal of Geophysical Research* **116:**
544 Ducklow, H. W., & Hill, S. M. (1985) The Growth of Heterotrophic Bacteria in the Surface
545 Waters of Warm Core Rings. *Limnology and Oceanography* **30:** 239-259.
546 Duhaime, M. B., & Sullivan, M. B. (2012) Ocean viruses: Rigorously evaluating the
547 metagenomic sample-to-sequence pipeline. *Virology* **434:** 181-186.

548 Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B. et al.
549 (2018). Host-linked soil viral ecology along a permafrost thaw gradient. *Nature*
550 *Microbiology* **3**: 870-880.
551 Field, J. G., Clarke, K. R., & Warwick, R. M. (1982) A Practical Strategy for Analysing
552 Multispecies Distribution Patterns. *Marine Ecology Progress Series* **8:** 37-52.
553 Frey, B. J., & Dueck, D. (2007) Clustering by Passing Messages Between Data Points.
554 *Science* **315:** 972-976.
555 Fuchs, B., Woebken, D., Zubkov, M., Burkill, P., & Amann, R. (2005) Molecular
556 identification of picoplankton populations in contrasting waters of the Arabian Sea.
557 *Aquatic Microbial Ecology* **39:** 145-157.
558 Fuhrman. (1992) *Bacterioplankton roles in cycling of organic matter: the microbial food*
559 *web.*
560 Fuhrman, J. A., & Noble, R. T. (1995) Viruses and Protists Cause Similar Bacterial
561 Mortality in Coastal Seawater. *Limnology and Oceanography* **40:** 1236-1242.
562 Gregory, A. C., Solonenko, S. A., Ignacio-Espinoza, J. C., LaButti, K., Copeland, A.,
563 Sudek, S. et al. (2016) Genomic differentiation among wild cyanophages despite
564 widespread horizontal gene transfer. *BMC genomics* **17:** 930.
565 Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A.
566 et al. (2019) Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*
567 **177:** 110-1123.e14.
568 Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S. et al. (2016)
569 Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532:**
570 465-470.

571 Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., & Weitz, J. S. (2013)
572 Robust estimation of microbial diversity in theory and in practice. *The ISME journal*
573 **7:** 1092-1101.
574 Hauser, M., Steinegger, M., & Söding, J. (2016) MMseqs software suite for fast and deep
575 clustering and searching of large protein sequence sets. *Bioinformatics (Oxford,*
576 *England)* **32:** 1323-1330.
577 Hawley, A. K., M. Brewer, H., Norbeck, A. D., Paša-Tolić, L., & Hallam, S. J. (2014)
578 Metaproteomics reveals differential modes of metabolic coupling among
579 ubiquitous oxygen minimum zone microbes. *Proceedings of the National Academy*
580 *of Sciences of the United States of America* **111:** 11395-11400.
581 Helly, J. J., & Levin, L. A. (2004) Global distribution of naturally occurring marine hypoxia
582 on continental margins. *Deep-Sea Research Part I* **51:** 1159-1168.
583 Hurwitz, B. L., Hallam, S. J., & Sullivan, M. B. (2013) Metabolic reprogramming by viruses
584 in the sunlit and dark ocean. *Genome Biology* **14:** R123.
585 Hurwitz, B. L. & U'ren, J. M. (2016) Viral metabolic reprogramming in marine ecosystems.
586 *Current Opinion in Microbiology* **31**: 161-168.
587 Hyatt, D., Chen, G., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010)
588 Prodigal: prokaryotic gene recognition and translation initiation site identification.
589 *BMC bioinformatics* **11:** 119.
590 Jackson, D. A. (1995) PROTEST: A PROcrustean Randomization TEST of community
591 environment concordance. *Écoscience* **2:** 297-303.

592 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D. et al.
593 (2018). vegan: Community Ecology Package. R package version 2.5-2.
594 https://CRAN.R-project.org/package=vegan
595 Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M. et al. (2019)
596 Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by
597 gene-sharing networks. *Nature biotechnology* **37:** 632-639.
598 John, S. G., Mendez, C. B., Deng, L., Poulos, B., Kauffman, A. K. M., Kern, S. et al. (2011)
599 A simple and efficient method for concentration of ocean viruses by chemical
600 flocculation. *Environmental microbiology reports* **3:** 195-202.
601 Jost, L. (2006) Entropy and diversity. *Oikos* **113:** 363-375.
602 Karstensen, J., Stramma, L., & Visbeck, M. (2008) Oxygen minimum zones in the eastern
603 tropical Atlantic and Pacific oceans. *Progress in oceanography* **77:** 331-350.
604 Kessler, W. S. (2006) The circulation of the eastern tropical Pacific: A review. *Progress*
605 *in Oceanography* **69:** 181-217.
606 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., &
607 Salzberg, S. L. (2004) Versatile and open software for comparing large genomes.
608 *Genome biology* **5:** R12.
609 Lam, P. and Kuypers, M. M.M. (2011) Microbial Nitrogen Cycling Proceses in Oxygen
610 Minimum Zones. *Annual Review of Marine Science* **3**:317-345.
611 Langmead, B., & Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2.
612 *Nature Methods* **9:** 357-359.

Madrid, V. M., Taylor, G. T., Scranton, M. I., & Chistoserdov, A. Y. (2001) Phylogenetic Diversity of Bacterial and Archaeal Communities in the Anoxic Zone of the Cariaco Basin. *Applied and Environmental Microbiology* **67:** 1663-1674.

Mantel, N. (1967) The Detection Approach.

Mielke, P. W., Berry, K. J., & Johnson, E. S. (1976) Multi-response permutation procedures for a priori classifications. *Communications in Statistics - Theory and Methods* **5:** 1409-1424.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome research* **27:** 824-834.

Ohio Supercomputer Center. Ohio Supercomputer Center. Columbus OH: Ohio Supercomputer Center 1987.

Parsons, T.R., Maita, Y., and Lalli, C.M. (1984) A manual of biological and chemical methods for seawater analysis. Oxford: Pergamon Press.

Paulmier, A., & Ruiz-Pino, D. (2009) Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography* **80:** 113-128.

Pielou, E. C. (1966) The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* **13:** 131-144.

Rabosky, D. L. (2009) Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters* **12:** 735-743.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5:** 69.

636 Resvbech, N. P., Larsen, L. H., Gundersen, J., Dalsgaard, T., Ulloa, O., & Thamdrup, B.
637 (2009) Determination of ultra-low oxygen concentrations in oxygen minimum
638 zones by the STOX sensor. *Limnology and Oceanography Methods* **7**(5):371-381.
639 Rosenwasser, S., Carmit, Z., Graff van Creveld, S., & Vardi, A. (2016) Virocell
640 Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean.
641 *Trends in Microbiology* **24**: 821-832.
642 Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A. et al. (2016)
643 Ecogenomics and potential biogeochemical impacts of globally abundant ocean
644 viruses. *Nature* **537:** 689-693.
645 Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015) VirSorter: mining viral signal
646 from microbial genomic data. *PeerJ* **3:** e985.
647 Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas,
648 R. et al. (2014) Ecology and evolution of viruses infecting uncultivated SUP05
649 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3:** e03125.
650 Schmidtko, S., Stramma, L. & Visbeck, M. (2017) Decline in global oceanic oxygen
651 content during the past five decades. *Nature* **542:** 335–339.
652 Selden, C. R., Mulholland, M. R., Widner, B., Bernhardt, P. W., Chang, B., and A.
653 Jayakumar. $N_2$ fixation in the Eastern Tropical South Pacific oxygen deficient zone:
654 Implications for the range of marine diazotrophs. Frontiers in Microbiology (in
655 review).
656 Shannon, C. E. (1948) A Mathematical Theory of Communication. **27:** 3790423.
657 Simpson, E. (1949) Measurement of diversity. *Nature* **163:** 688.

Stevens, H., & Ulloa, O. (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environmental Microbiology* **10:** 1244-1259.

Stramma, L., Johnson, G. C., Sprintall, J., & Mohrholz, V. (2008) Expanding Oxygen-Minimum Zones in the Tropical Oceans. *Science* **320:** 655-658.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G. *et al.* (2015) Structure and function of the global ocean microbiome. *Science* **348:** 1261359.

Suttle, C. A. (1994) The Significance of Viruses to Mortality in Aquatic Microbial Communities. *Microbial Ecology* **28:** 237-243.

Suttle, C. A. (2007) Marine viruses - major players in the global ecosystem. *Nature Reviews Microbiology* **5:** 801-812.

Suzuki, R., & Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22:** 1540-1542.

Thamdrup, B., Dalsgaard, T., & Revsbech, N.P. (2012) Widespread functional anoxia in the oxygen minimum zone of the Eastern South Pacific. *Deep Sea Research Part 1: Oceanographic Research Papers* **65**:36-45.

Tibshirani, R., Walther, G., & Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society* **63:** 411-423.

Ulloa, O., Canfield, D. E., DeLong, E.F., Letelier, R. M., & Steward, F. J. (2012) Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences of the USA* **109**(40): 15996-16003.

679 Walsh, E. A., Kirkpatrick, J. B., Rutherford, S. D., Smith, D. C., Sogin, M., & D'Hondt, S.
680 (2016) Bacterial diversity and community composition from seasurface to
681 subseafloor. *The ISME journal* **10:** 979-989.
682 Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York,
683 2016.
684 Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012) Microbial ecology of expanding oxygen
685 minimum zones. *Nature Reviews Microbiology* **10:** 381-394.
686 Wyrtki, K. (1965) Surface Currents of the Eastern Tropical Pacific Ocean. *Inter-American*
687 *Tropic Tuna Commission* **IX**:
688 Zakem, E.J., Mahadevan, A., Lauderdale, J.M., & Follows, M. J. (2019) Stable aerobic
689 and anaerobic coexistence in anoxic marine zones. *The ISME Journal* **14**: 288-
690 301.
691

692 Table and Figure Legends

693 **Figure 1. Map of the study area and vertical characterization of the sampling**
694 **stations. A.** Location of stations 7, 8, 14, 16, 17, and 18, off the coast of Peru in the ETSP
695 OMZ. The map was created with Ocean Data View (http://odv.awi.de). **B.** Oxygen (solid
696 blue line) and fluorescence/chlorophyll (solid green line) depth profiles from each station.
697 Samples were collected at depths indicated by lines with the colored circles to depict the
698 sample type: surface chlorophyll maximum in yellow, oxycline in orange, upper OMZ with
699 deep chlorophyll maximum (DCM) in green and without DCM in light blue, and OMZ core
700 in dark blue.
701

**Figure 2. Hierarchical clustering of samples based on normalized relative abundances of viral populations revealed 6 clusters.** Each row represents a different sample, labelled by station and sample type (scm = surface chlorophyll maximum, oxy = oxycline, uomzD = upper OMZ with deep chlorophyll maximum, uomz = upper OMZ without DCM, and omz = omz core). Each column represents a different viral population (≥ 5kb), where the normalized relative abundance values ($log_{10}$ transformed) are shown in grayscale. RPKM means Reads Per Kilobase, per Million mapped reads. "Approximately unbiased" bootstrapping values are represented as a proportion of 100 permutations.

**Figure 3**. **Diversity measures for the 6 main viral community clusters. A.** Evenness across the 6 clusters. **B.** Alpha Diversity. **C.** Beta Diversity, where red corresponds to the beta diversity differences resulting from abundance while blue represents the beta diversity differences related to composition. The top and bottom of each box show the standard deviation while the line inside the box shows the mean. Points within each box represent the number of samples per community.

**Figure 4. Environmental influence on the distribution of viral communities. A.** Viral community ordination using NMDS and Bray-Curtis dissimilarities. Each of the 6 distinct communities are incorporated into the outlined clusters and color coded. **B.** Environmental feature ordination using NMDS and Euclidean distance. The colors are the same as in A. **C.** A comparison between the viral community ordination and environmental feature ordination where temperature has been removed from the latter dataset. Each

arrow represents a sample's spatial distance between ordinations, again with the same color coding. The statistical similarity between ordinations is represented as the Procrustes sum of squares where a lower value is more significant. D. The same comparison between the viral community ordination and environmental feature ordination, but with oxygen removed from the latter dataset.

**Figure 5. Sequence similarity to GOV2.0. A.** ETSP sequence matches to GOV2.0 sequences, separated and color coded by GOV2.0 sample type. Each bar along the x-axis represents an ETSP community and each color along the y-axis represents the percent of relative abundance per GOV2.0 hit. Relative proportions within each community were calculated by summing relative abundances of each population within a GOV2.0 hit location, and then dividing that sum by the total sum of relative abundances with GOV2.0 hits (see Experimental Procedures). **B.** Distribution of viral clusters (viral genera) across each of the 6 communities. Each bar along the x-axis represents one viral cluster, colored by the percent of VC found in given community.

**Table S1. Metadata for the 22 samples collected and sequenced.** Location and measurements of the selected environmental features sampled per site and depth. Samples are organized and labeled according to their respective cluster from the hierarchical clustering of the viral population abundances.

**Table S2. Sequencing information for each sample.** For each of the samples used in this study, this table lists the number of raw read, number of reads following quality

control, the number of viral populations identified in each sample, and the number of reads that map to viral populations.

**Table S3. Relative abundances of the viral populations in each sample.** The abundances were normalized across samples via number of viral sequencing reads and by the length of each sequence.

**Table S4. Comparison of viruses from ETSP and GOV2.** Statistics for the ETSP viral populations with significant protein coding sequence similarity with GOV2.0, according to MMseq2 “easy-search”. Only sequences sharing 95% identity across 50% of the sequence are reported.

**Table S5. Categorization of viral populations.** For each of the 46,127 viral populations, this table contains the VirSorter, VirFinder, and CAT assignments.

**Figure S1. Vertical distribution of oxygen, temperature and salinity of the 6 sampled stations.** Oxygen, temperature and salinity depth profiles of the first 300 meters of each station. Colored circles indicate the depths where each of the 22 samples were collected: surface chlorophyll maximum in yellow (“scm”), oxycline in orange (“oxy”), upper OMZ with deep chlorophyll maximum (DCM) in green (“uomzD”) and without DCM in light blue (“uomz”), and OMZ core in dark blue (“omz”).

**Figure S2. Gap statistic for the number of significant sample clusters.** The cluster size, in number of samples, is where the observed within cluster distance is the smallest and yields the highest "gap" between expected within group distances. This was calculated using a null model, and observed within group distances. Here, clusters were derived by kmeans with 100 bootstraps and a maximize cluster size of 10.

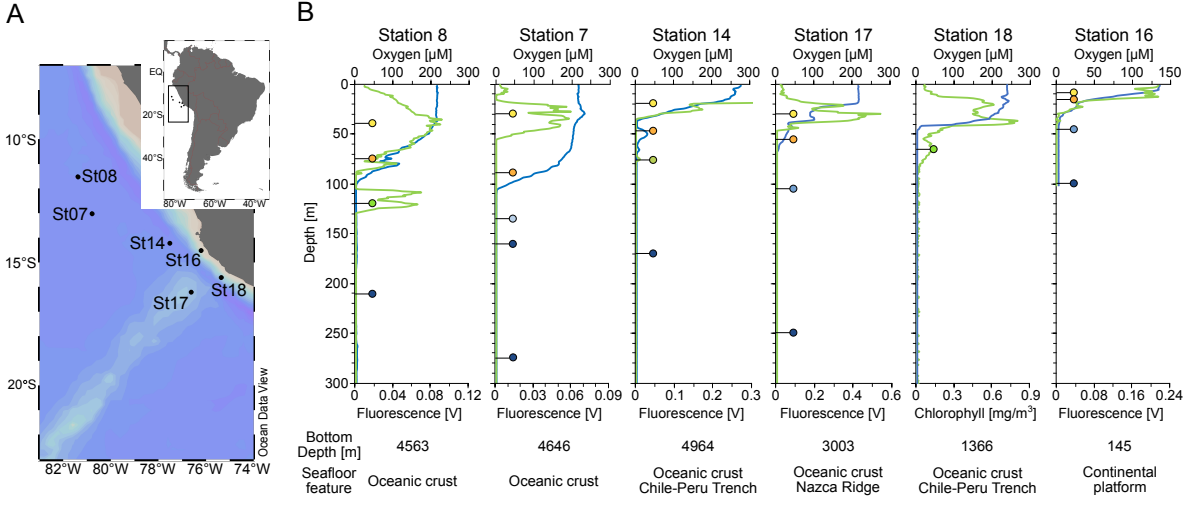
**Figure S3. Affinity propagation analysis.** Clustering of samples according to negative squared Euclidean distances, using default input and exemplar preferences. The lighter yellow color corresponds with a higher similarity score, and each cluster is represented in the dendrograms with color coded bars.

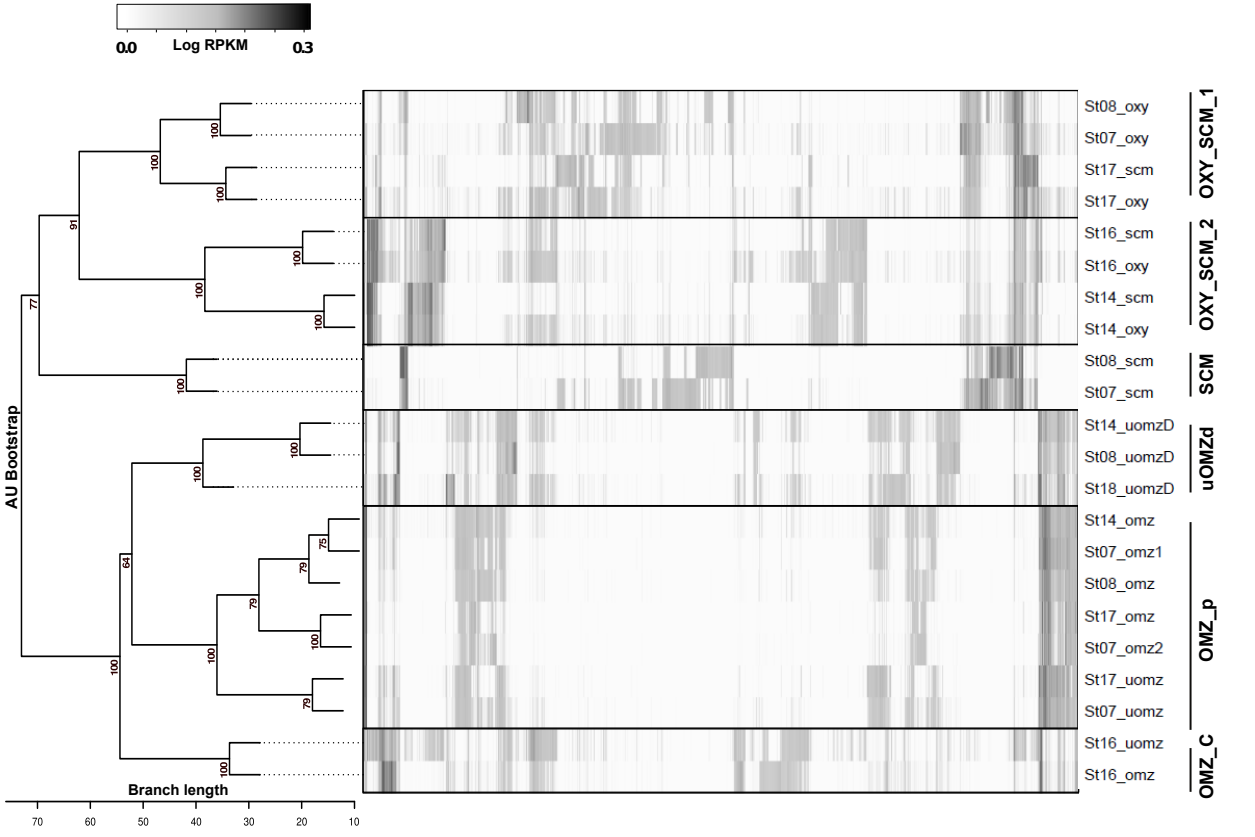
**Figure S4. Species accumulation curve.** The number of species (viral populations) identified by 100 random sub-samplings of each of the pooled 22 samples across sampling depth and stations. Species richness estimations were computed using the jacknife2 estimator.

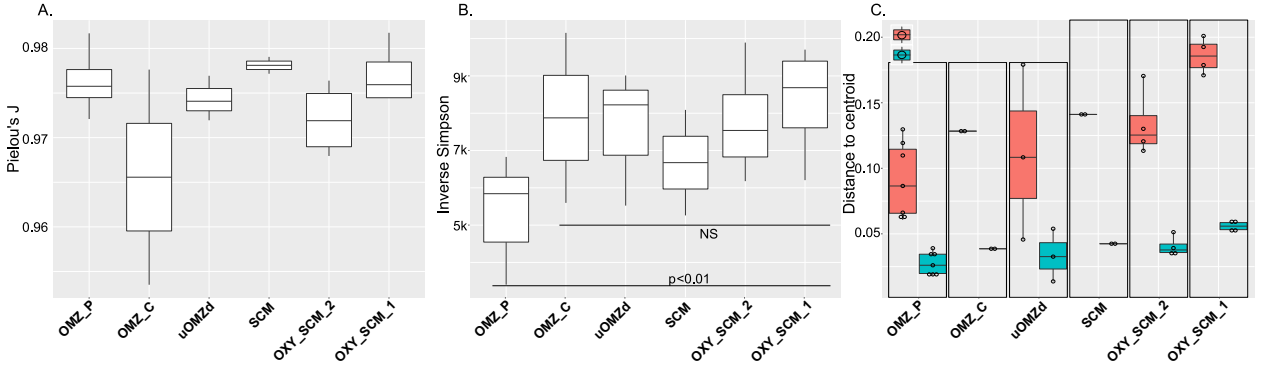
**Figure S5. NMDS stress plots of viral communities and environmental features.** A comparison of the fit of the distances displayed in the NMDS ordination plot with the true Bray Curtis dissimilarities between each station.

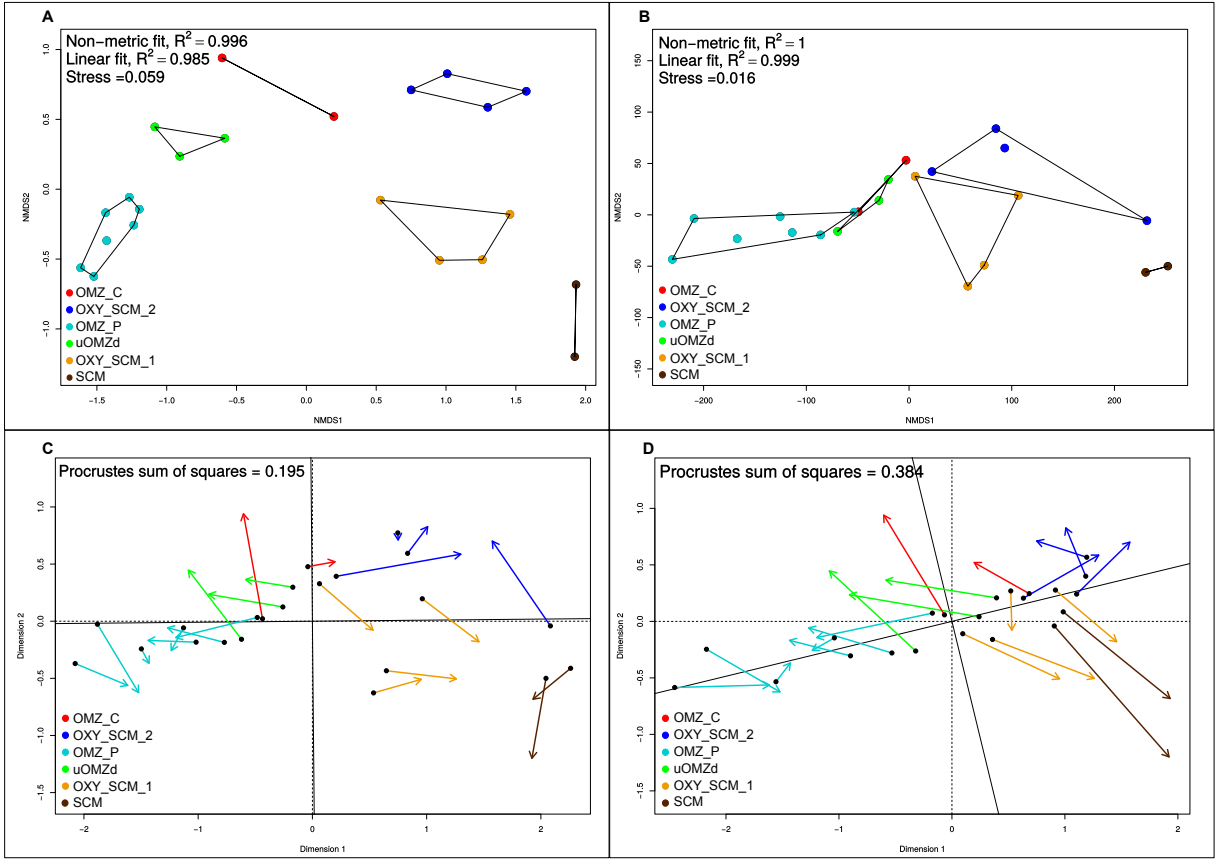
**Figure S6. Ordination plot of environmental features. Procrustes comparison of the environmental features and viral populations.** An alignment of the NMDS ordination plot created for the viral populations (Bray-Curtis dissimilarity) and environmental features

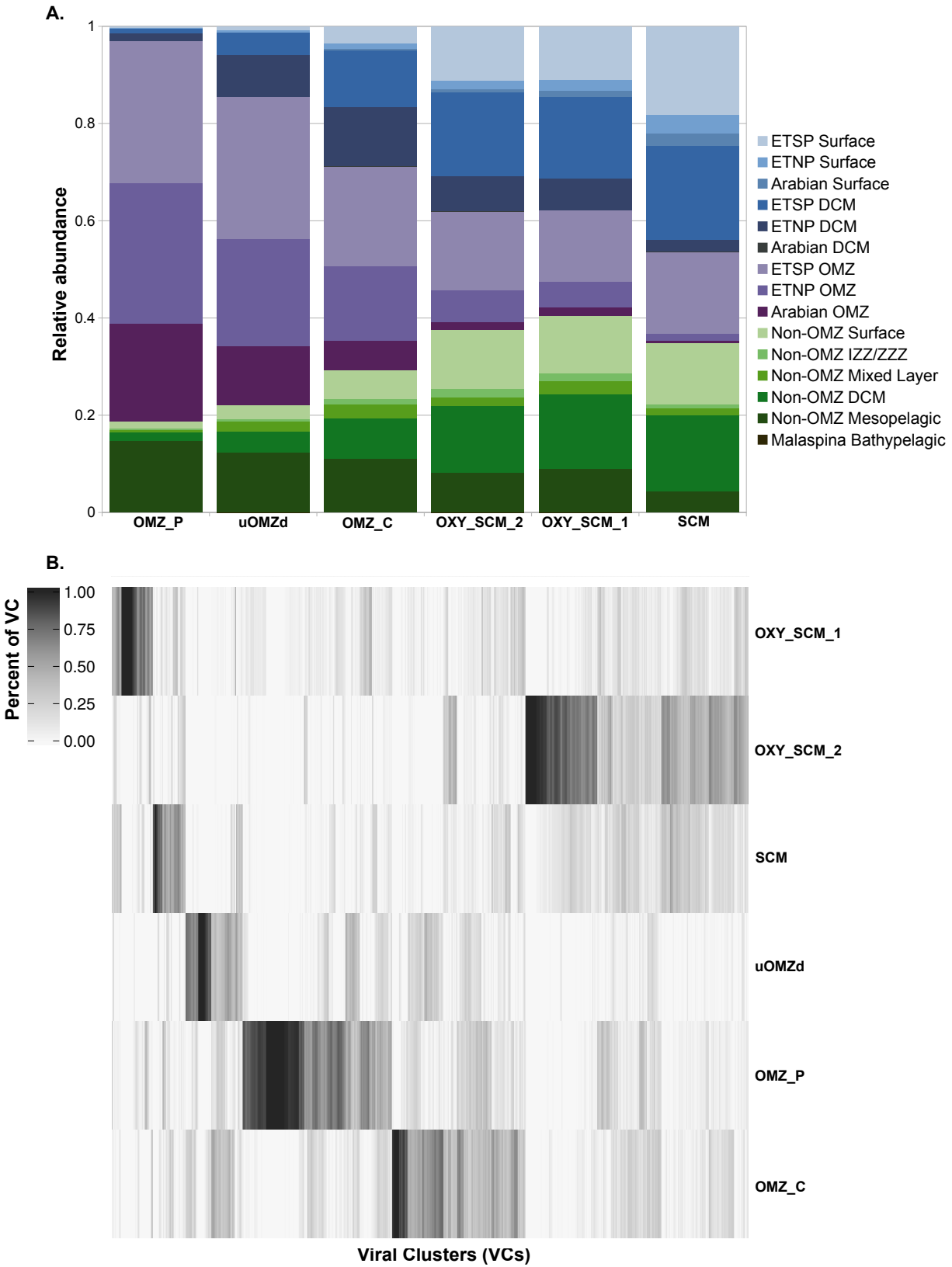
793 (Euclidean distance). No environmental features were removed from this comparison.
794 Procrustes sum of squares is 0.194.
795
796 **Figure S7. Sequencing depth.** Plot comparing the sequencing depth (gray bars), in
797 terms of the number of post-quality control paired end reads, to the number of reads from
798 that recruited to the identified viral populations, pooled from all samples (black bars).
799


A
B
St08
St07
St14
St16
St17
St18
10°S
15°S
20°S
82°W
80°W
78°W
76°W
74°W
Ocean Data View
EQ
20°S
40°S
80°W
60°W
40°W
Station 8
Station 7
Station 14
Station 17
Station 18
Station 16
Oxygen [µM]
Depth [m]
Fluorescence [V]
Chlorophyll [mg/m3]
Bottom Depth [m]
4563
4646
4964
3003
1366
145
Seafloor feature
Oceanic crust
Oceanic crust
Oceanic crust Chile-Peru Trench
Oceanic crust Nazca Ridge
Oceanic crust Chile-Peru Trench
Continental platform


0.0
Log RPKM
0.3
AU Bootstrap
Branch length
70
60
50
40
30
20
10
St08_oxy
St07_oxy
St17_scm
St17_oxy
St16_scm
St16_oxy
St14_scm
St14_oxy
St08_scm
St07_scm
St14_uomzD
St08_uomzD
St18_uomzD
St14_omz
St07_omz1
St08_omz
St17_omz
St07_omz2
St17_uomz
St07_uomz
St16_uomz
St16_omz
OXY_SCM_1
OXY_SCM_2
SCM
uOMZd
OMZ_p
OMZ_C


A.
Pielou's J
0.98
0.97
0.96
OMZ_P
OMZ_C
uOMZd
SCM
OXY_SCM_2
OXY_SCM_1
B.
Inverse Simpson
9k
7k
5k
NS
p<0.01
OMZ_P
OMZ_C
uOMZd
SCM
OXY_SCM_2
OXY_SCM_1
C.
Distance to centroid
0.20
0.15
0.10
0.05
OMZ_P
OMZ_C
uOMZd
SCM
OXY_SCM_2
OXY_SCM_1


A
Non−metric fit, $R^2 = 0.996$
Linear fit, $R^2 = 0.985$
Stress =0.059
NMDS1
NMDS2
B
Non−metric fit, $R^2 = 1$
Linear fit, $R^2 = 0.999$
Stress =0.016
NMDS1
NMDS2
C
Procrustes sum of squares = 0.195
Dimension 1
Dimension 2
D
Procrustes sum of squares = 0.384
Dimension 1
Dimension 2
OMZ_C
OXY_SCM_2
OMZ_P
uOMZd
OXY_SCM_1
SCM


A.
Relative abundance
1
0.8
0.6
0.4
0.2
0
OMZ_P
uOMZd
OMZ_C
OXY_SCM_2
OXY_SCM_1
SCM
ETSP Surface
ETNP Surface
Arabian Surface
ETSP DCM
ETNP DCM
Arabian DCM
ETSP OMZ
ETNP OMZ
Arabian OMZ
Non-OMZ Surface
Non-OMZ IZZ/ZZZ
Non-OMZ Mixed Layer
Non-OMZ DCM
Non-OMZ Mesopelagic
Malaspina Bathypelagic
B.
Percent of VC
1.00
0.75
0.50
0.25
0.00
OXY_SCM_1
OXY_SCM_2
SCM
uOMZd
OMZ_P
OMZ_C
Viral Clusters (VCs)