Supplementary Methods

Seawater was collected and filtered from the top meter at the San Pedro Ocean time-series (SPOT) (33°33′ N, 118°24′ W) between March 12 and August 3, 2011, at a daily-to-weekly resolution as previously described (Needham and Fuhrman 2016). Methods for the physical and chemical measurements for the current study have been previously described and published (Needham & Fuhrman 2016), and all data are available via Figshare (https://figshare.com/s/ba6029898908bf7b80d5 or https://doi.org/10.6084/m9.figshare.4776643.v1). DNA was extracted from the 1-80 μm size fraction on a AE glass fiber (~ 1 μm pore size), and 0.22 μm Durapore filter and 0.02 μm Anotop filter as previously described (Needham & Fuhrman 2016; Steward & Culley 2010).

16S Sequencing

We PCR-amplified the V4 and V5 hypervariable regions of small sub-unit rRNA gene of bacteria, archaea, eukaryotes and of phytoplankton chloroplasts with the 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 926R (5'-CCGYCAATTYMTTTRAGTTT-3') primers as previously described (Needham and Fuhrman 2016).

g23 Sequencing Assay

To amplify and sequence the g23 major capsid protein gene on the Illumina MiSeq platform, PCR was performed first with g23 primers: T4-SuperF1 (5'-GAYHTIKSIGGIGTICARCCIATG-3') and T4-SuperR1 (5'-GCIYKIARRTCYTGIGCIARYTC-3') (Chow and Fuhrman 2012) as follows: 2 ng of 0.02 μm size fraction DNA template was amplified in triplicate in 25 μL reactions of 1x NEB Buffer, 2 units of Thermopol Taq polymerase (New England Biolabs, Catalog Number: M0267, Ipswich, MA, USA), 0.48 μM of each primer, 320 ng/μL BSA (Sigma-Aldrich, Catalog Number: A7030, St Louis, MO, USA), 0.2 μM dNTPs (Promega, Madison, WI, USA), and 2.8 mM $MgCl_2$ (Thermo Fisher, Catalog Number: R0971, Waltham, MA USA). Amplification proceeded as follows: initial denaturation of 180 s at 95 C; 35 cycles of 95 C for 30 s, 59 C for 45 s, and 72 C for 60 s; and a final extension for 300s at 72 C.

The g23 products were then prepared for Illumina sequencing by adding 1 μL of each PCR reaction to the previous reaction mix except g23-only primers were replaced with an Illumina compatible g23 primer construct (Supplementary Figure 1) and cycled for an additional 5 rounds of amplification. PCR products were confirmed by gel electrophoresis, purified with Ampure beads pooled in equimolar concentrations and sequenced via Illumina MiSeq 2x300bp at the University of California Davis Genome Center.

Mock community generation

16S "even" (10 clones, 10% each) and "staggered" (range 0.01% to 35%) mock communities were generated by mixing 16S environmental clone sequences at known concentrations, as previously described (Parada et al. 2016). Similarly, we generated an "even" (10 clones, 10% each) T4-like-myovirus mock communities from SPOT environmental g23 sequence clones (Chow and Fuhrman 2012). g23 clones were regrown and plasmids extracted with Zymo Miniprep as per manufacturer's instructions. Plasmids were diluted to 0.01 ng/µL and amplified for 25 cycles in the following reaction mixture of 25µL: M13F and M13R primer 0.5 µL (100 ng/µL), 1x HiFi Buffer (Invitrogen, Catalog Number: 11304011,Carlsbad, CA, USA), 0.16 mM dNTPs (Promega), 2 mM $MgS0_4$ (Invitrogen), 0.4 units HiFi Platinum Taq (Invitrogen). Amplified products were purified with Ampure beads, eluted in TE and pooled at 10% each. After mock community generation, clones were re-sequenced via Sanger sequencing to confirm consensus sequences (Supplementary Data Set 1). Contamination in one clone was detected. Upon sequencing of the mock community via MiSeq, we observed 2 unexpected g23 sequence clusters that were divergent (< 80% similar) to clones that were intentionally added, but similar to other SPOT g23 clones, thus were probably unintentionally added with the contaminated clone. Since these sequences were from lab-based contamination during mock community generation, we removed them and the contaminated sequence from our mock community analysis.

ITS of SAR11 Clade

To design primers specific for SAR11 ITS, potential ITS sequences were selected from Sanger metagenomic reads from the Global Ocean Survey (GOS) (Venter et al. 2004; Rusch et al. 2007) and the Gulf of Maine (GoMA) (Tully et al. 2011). To find long reads with the ITS region which is adjacent to the 3' end of the 16S rRNA gene, the "universal" SSU rRNA gene primer 1492R (TACGGCTACCTTGTTACGACTT) was searched against the GOS and GoMA metagenomes using BLASTn (Altschul et al. 1990) (word-size = 7) and retained reads with less than 2 mismatches. These sequences were classified by BLASTn search against the SILVA database and sequences classified as members of the SAR11 clade were used to identify a marine SAR11 group specific primer. The forward primer "SAR11_ITS_F1" (5'-CCGTCCKCRYTTCTBTT-3') is located about 45 bases within the ITS sequence of SAR11 (near the 16S end) and reverse primer "SAR11_23S_R1" (5'-WBWGTGCCDAGGCATYC-3') is located about 45 bases inside the 23S ribosomal subunit, resulting in an *in silico* length range of 367-447bp, including primers. Primers with linkers suitable for direct Illumina-library generation were synthesized with the SAR11 ITS specific primers similarly to the 16S and g23 primer constructs (Supplementary Figure 1). Triplicate 25 µL PCR reactions proceeded as follows: 1x HiFi Buffer, 0.2 µM dNTPS (Promega), 0.4 µM of each SAR11 ITS primer with Illumina linkers, 2 mM $MgSO_4$, and 1 unit of HiFi Platinum Taq. Amplification had an initial denaturation at 95 C for 120 s; 35 rounds of 95 C for 30 s, 59 C for 45 s and 68 C for 90 s; and a final elongation at 68 C for 300 s. PCR products were confirmed by gel electrophoresis, purified with Ampure beads and sequenced via Illumina MiSeq 2x300 bp at the University of California Davis Genome Center.

Sequence Analysis

All sequence analysis commands and steps are available via Figshare (https://figshare.com/projects/Ecological_dynamics_and_co-occurrence_among_marine_phytoplankton_bacteria_and_myoviruses_shows_microdiversity_matters/16260 and main commands can be found at https://doi.org/10.6084/m9.figshare.3971709.v1). Briefly, g23 and SAR11 ITS sequences were trimmed via Trimmomatic (Bolger et al. 2014) to remove bases that caused the mean quality score to drop less than 30 within a sliding window of 5 bp and then sequences less than 250 bp were removed. Forward and reverse sequences were merged allowing 0 mismatches and a minimum merged length of 250 bp in USEARCH7 with *fastq_mergepairs* (Edgar 2013). g23 sequences were translated in the 3 forward frames via *transeq* in EMBOSS 6.6.0.0 (Rice et al. 2000) and sequences that contained a stop codon (*) or unidentified residues (X) in each of the 3 forward frames were removed (~ 7% of total). For SAR11 ITS and g23 sequences, chimeras were detected *de novo* via *identify_chimeric _seqs.py* within QIIME with USEARCH61 (settings: usearch61_minh 0.05, usearch 61_mindiffs 1, usearch 61_xn 2) and removed. 99% DNA sequence similarity OTU clusters were generated with *pick_otus.py* in QIIME with UCLUST.

16S sequence analysis was similar to that above (and was previously fully described (Needham and Fuhrman 2016)), with the following exceptions: merged sequences with an average quality score less than 25 across the full length were removed and chimera checking was supplemented with reference-based detection (via usearch61) against the SILVA gold database.

Decomposition of 99% OTUs

We used minimum entropy decomposition (MED) (Eren et al. 2014) to decompose (i.e., split into single-base variants) 99% OTUs that exceeding a threshold of 0.4% (relative abundance) on average or 2.5% on any day of the time series. All of the sequences from each of these 99% OTUs (10^3-10^5 sequences each) were aligned with MAFFT v7.123b (--retree 1 --maxiterate 0 --nofft --parttree) (Katoh et al. 2002) and then Shannon Entropy (a metric that assigns a value to a string of characters based on the amount of variation observed (Eren et al. 2014)), was used to discriminate significant sequence variations for each OTU independently using the *decompose* command of MED. To determine appropriate settings for the *decompose* command for our sample-to-sequence pipeline, we used the "staggered" 16S mock community as a guide, assuming each cloned mock community member contains a single sequence and therefore other sequences are erroneous. While the mean Shannon Entropy value for the 25 input clone sequences of the staggered mock community was 0.019 +/- 0.019 SD, we selected a conservative value of 0.25 in order to exceed the maximum (0.24) observed in any OTU (a low abundance SAR202 clone)(Supplementary Figure 2). As such, the alignments of environmental sequences that had entropy at sites greater than 0.25 were conservatively decomposed (into what we consider legitimate single-base variants) based on those positions, and decomposition continued until all positions had entropy less than

0.25. The following thresholds were also set to reduce the likelihood of erroneous sequences being considered Amplicon Sequence Variants (ASVs): the minimum number of the most abundant sequence within each ASV must exceed 50 and if ASVs did not exceed 1% of the parent 99% OTU's composition, on average, they were removed from analysis. Validation of our application of the MED method was performed on the g23 mock communities independently of environmental samples, but with the same settings. Entropy values for each position were obtained via the *entropy-analysis* command within MED.

Sequence Identification

16S sequence classification 99% OTU clusters were taxonomically classified via UCLUST assignment against the SILVA and Greengenes databases, as well as BLASTn search against the NCBI database, as previously described (Needham and Fuhrman 2016). Greengenes taxonomy was used to separate chloroplast 16S from prokaryotic sequences, and independent OTU tables were generated for chloroplasts and prokaryotic taxa. g23 sequences were classified using BLASTx (minimum e-value, 0.01) against viral proteins from genomes downloaded from NCBI in March 2015. Sequences obtained from the SAR11 ITS sequencing assay were classified by BLASTn search against both NCBI RefSeq genomic sequences (downloaded May 2015), and a custom ITS database that included environmental SAR11 16S-ITS-23S clone sequences (Brown and Fuhrman 2005; Brown et al. 2005), metagenomic reads (Brown et al. 2012), and SAR11 S4 sequences from SILVA119.

Statistical Analyses

All commands used for statistical analyses are available via Figshare (https://figshare.com/projects/Ecological_dynamics_and_co-occurrence_among_marine_phytoplankton_bacteria_and_myoviruses_shows_microdiversity_matters/16260 and main commands can be found at https://doi.org/10.6084/m9.figshare.3971709.v1). We identified monotonic increases and decreases of OTUs and ASVs using the Mann-Kendall test within the “Kendall” package in R. To calculate the estimated abundances of the various ASVs we multiplied the fraction of the ASV within an OTU by the overall parent OTU relative abundances for each day. For example, in a given sample, if a ASV is 33% of the parent and the parent was 3% of the whole community, then we used 1%. Pairwise correlations between estimated abundance of ASVs and OTUs were then determined using extended Local Similarity Analysis (Xia et al. 2011, 2013) on the types that were present >0.05% (relative abundance) on average and on at least 25% of sample dates. Network visualizations of correlation matrices were generated in Cytoscape_v3.0.1(Shannon et al. 2003). Mantel tests were performed in R via the Vegan package on only fully overlapping set of data, i.e., if a sample date did not have a value for all types of data, that sample date was removed. The total number of dates for Mantel tests was 32.

Sequence alignments for heatmap phylogeny was generated in via default MUSCLE v3.8.31 (Edgar 2004) settings with 100 iterations, and consensus phylogeny was generated via PhyML (Guindon and Gascuel 2003) with 100 bootstraps. To calculate dN/dS for each g23 OTU, ASVs for each OTU were translated in frame 1 using *transeq*, aligned with MAFFT v7.123b, and converted to codon alignments using tranalign in EMBOSS 6.6.0.0. dN/dS ratios were then generated using KaKs_Calculator 2.0 (-c 11 -m MS $p < 0.05$) using codon alignments(Wang et al. 2010).

Nucleotide sequence accession numbers

Sequences from this study are available via EMBL project numbers PRJEB14228 (major capsid protein g23), PRJEB12108 (SAR11 ITS), and PRJEB10834 (SSU rRNA).

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**: 403–10.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Brown, M. V., and J. A. Fuhrman. 2005. Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. Aquat Microb Ecol **41**: 15–23.

Brown, M. V., F. M. Lauro, M. Z. DeMaere, L. Muir, D. Wilkins, T. Thomas, M. J. Riddle, J. A. Fuhrman, C. Andrews-Pfannkoch, J. M. Hoffman, J. B. McQuaid, A. Allen, S. R.

Rintoul, and R. Cavicchioli. 2012. Global biogeography of SAR11 marine bacteria. Mol Syst Biol **8**: 595.

Brown, M. V., M. S. Schwalbach, I. Hewson, and J. A. Fuhrman. 2005. Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. Environ Microbiol **7**: 1466–79.

Chow, C.-E. T., and J. A. Fuhrman. 2012. Seasonality and monthly dynamics of marine myovirus communities. Environ Microbiol **14**: 2171–83.

Countway, P. D., P. D. Vigil, A. Schnetzer, S. D. Moorthi, and D. A. Caron. 2010. Seasonal analysis of protistan community structure and diversity at the USC Microbial Observatory (San Pedro Channel, North Pacific Ocean). Limnol Oceanogr **55**: 2381–2396.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**: 1792–1797.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods **10**: 996–8.

Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin. 2014. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J **9**: 968–979.

Guindon, S., and O. Gascuel. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol **52**: 696–704.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**: 3059–3066.

Needham, D. M., and J. A. Fuhrman. 2016. Pronounced daily succession of phytoplankton , archaea and bacteria following a spring bloom. , doi:10.1038/NMICROBIOL.2016.5

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite (2000). Trends Genet **16**: 276–277.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. J. Williamson, S. Yooseph, D. Wu, J. a Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H.

Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol **5**: e77.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res **13**: 2498–504.

Steward, G. F., and A. I. Culley. 2010. Extraction and purification of nucleic acids from viruses, p. 154–165. *In* S. Wilhelm, M. Weinbauer, and C. Suttle [eds.], Manual of Aquatic Viral Ecology. American Society of Limnology and Oceanography.

Tully, B. J., W. C. Nelson, and J. F. Heidelberg. 2011. Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. Environ Microbiol , doi:10.1111/j.1462-2920.2011.02628.x

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. J. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**: 66–74.

Wang, D., Y. Zhang, Z. Zhang, J. Zhu, and J. Yu. 2010. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics, Proteomics Bioinforma **8**: 77–80.

Xia, L. C., D. Ai, J. A. Cram, J. A. Fuhrman, and F. Sun. 2013. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. Bioinformatics **29**: 230–7.

Xia, L. C., J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun. 2011. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. BMC Syst Biol **5 Suppl 2**: S15.