The ISME Journal (2017) 11, 1614–1629

www.nature.com/ismej

ORIGINAL ARTICLE

Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters

David M Needham, Rohan Sachdeva and Jed A Fuhrman
Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

Numerous ecological processes, such as bacteriophage infection and phytoplankton–bacterial interactions, often occur via strain-specific mechanisms. Therefore, studying the causes of microbial dynamics should benefit from highly resolving taxonomic characterizations. We sampled daily to weekly over 5 months following a phytoplankton bloom off Southern California and examined the extent of microdiversity, that is, significant variation within 99% sequence similarity clusters, operational taxonomic units (OTUs), of bacteria, archaea, phytoplankton chloroplasts (all via 16S or intergenic spacer (ITS) sequences) and T4-like-myoviruses (via g23 major capsid protein gene sequence). The extent of microdiversity varied between genes (ITS most, g23 least) and only temporally common taxa were highly microdiverse. Overall, 60% of taxa exhibited microdiversity; 59% of these had subtypes that changed significantly as a proportion of the parent taxon, indicating ecologically distinct taxa. Pairwise correlations between prokaryotes and myoviruses or phytoplankton (for example, highly microdiverse *Chrysochromulina sp.*) improved when using single-base variants. Correlations between myoviruses and SAR11 increased in number (172 vs 9, Spearman >0.65) and became stronger (0.61 vs 0.58, *t*-test: $P<0.001$) when using SAR11 ITS single-base variants vs OTUs. Whole-community correlation between SAR11 and myoviruses was much improved when using ITS single-base variants vs OTUs, with Mantel rho $=0.49$ vs 0.27; these results are consistent with strain-specific interactions. Mantel correlations suggested >1 μm (attached/large) prokaryotes are a major myovirus source. Consideration of microdiversity improved observation of apparent host and virus networks, and provided insights into the ecological and evolutionary factors influencing the success of lineages, with important implications to ecosystem resilience and microbial function.
The ISME Journal (2017) **11,** 1614–1629; doi:10.1038/ismej.2017.29; published online 11 April 2017

Introduction

Microbial communities are diverse and dynamic, and knowledge of the factors that control the success of individual populations is important to understanding microbial influence on ecosystems and biodiversity. Although some microbial community variation is broadly predictable using coarse phylogenetic characterization of microorganisms (Fuhrman *et al.*, 2006; Teeling *et al.*, 2016), other controlling factors, such as viral infection (Rodriguez-Brito *et al.*, 2010; Avrani *et al.*, 2011; Marston *et al.*, 2012; Thingstad *et al.*, 2014) and phytoplankton–prokaryote associations (Amin *et al.*, 2015, 2012), may require more finely resolved community characterizations. Significantly, we do not know the relevant ecological and taxonomic units that are useful for observing the dynamics of 'populations' in the environment, for example, do 16S ribosomal RNA (rRNA) sequences have the required resolution (and at what sequence similarity level) or are more highly resolving genes such as the intergenic spacer (ITS) better (as commonly used to distinguish different marine cyanobacteria (Rocap *et al.*, 2003)). To what extent should we expect the necessary resolution to vary between different controlling factors? These issues represent a subset of the thorny but fundamental problems in microbiology of knowing how resolving one needs to be in order to understand microbial processes and interactions. Although classical taxonomic terms like genus, species, subspecies and strain are common in the literature when referring to cultured organisms, there are no well-accepted rules for translating these names via sequence data from uncultured organisms in nature. In any case, we usually do not even know which level of resolution is most ecologically relevant in the first place; too resolving may differentiate taxa based on functionally irrelevant variation and decrease predictive power to observe

Correspondence: JA Fuhrman, Department of Biological Sciences, University of Southern California, 3616 Trousdale Pkwy, AHF 107, Los Angeles, CA 90089, USA.
E-mail: fuhrman@usc.edu
Received 30 September 2016; revised 13 January 2017; accepted 2 February 2017; published online 11 April 2017

environmental and biological community covariation (Lu *et al.*, 2016), whereas too coarse obscures important variation relating to particular ecological phenomena such as specific interactions. Dealing with these issues is a challenging problem, and we suggest that one of the best practical ways to do so with the natural spectrum of actual microbial variation is through empirical observations of natural dynamics, observing the relationships among organisms and environmental variation through time series measurements. Microbial communities in the surface ocean have turnover times of a few days or less (Fuhrman and Caron, 2016), therefore observing their relationship may benefit from daily-to-weekly resolution. Here, we examine the bacteria, archaea, phytoplankton and viruses over a 5-month interval following the decline of a phytoplankton bloom through to oligotrophic summer conditions, at a temporal resolution of days to weeks and at phylogenetic resolution down to specific single nucleotides of marker genes.

With recent advances in methodological and sequencing technologies it is now possible to resolve single-nucleotide variation within marker genes of populations and communities (Eren *et al.*, 2013, 2014; Tikhonov *et al.*, 2015; Callahan *et al.*, 2016). Often these finely resolved taxa display distinct ecological dynamics, suggesting they are functionally different in ecologically relevant ways (Eren *et al.*, 2013, 2014; Tikhonov *et al.*, 2015). These single-base pair differences are obscured by standard 16S rRNA sequence clustering techniques (for example, Eren *et al.*, 2014; Reveillaud *et al.*, 2014; Newton *et al.*, 2015; Turlapati *et al.*, 2015). Here we examine the extent of significant sequence variations within sequence clusters to help understand the extent to which clusters appear to be made up of a single genetic entity, that is, 'clone-line', which is a consequence of the evolutionary and ecological forces that dictate the success of taxa. In addition, the amount of 'intra-species' variation is thought to positively influence productivity and resilience of lineages (Tesson *et al.*, 2014; Sjöqvist and Kremp, 2016), so observing how the extent of microdiversity varies between taxa will help understand the ecological resilience of particular lineages.

Diversity and dynamics of bacterial, archaeal and eukaryotic communities have often been characterized via the 'universal' SSU *rRNA* gene (16S for prokaryotes, 18S for eukaryotes). The strength and weakness of the SSU *rRNA* gene both derive from its conservation: although it allows exploration of nearly all cellular life, the resulting sequence diversity often does not allow distinguishing known species and strains. On the other hand, the 16S–23S ITS region is highly variable in sequence and length; however, ITS sequence length variability and diversity hinders the ability to develop a universal, or nearly universal assay, for high throughput sequencing. However, the ITS can and has been utilized to distinguish between taxa within a group such as the cyanobacteria (Rocap *et al.*, 2002) and the marine SAR11 clade (García-Martínez and Rodríguez-Valera, 2000; Rocap *et al.*, 2003; Brown and Fuhrman, 2005). Thus, to examine one group at a resolution higher than 16S, even with single-nucleotide resolution, we developed a novel ITS sequence analysis to interrogate the sequence diversity and dynamics of the most abundant microbial lineage in the ocean, the SAR11 clade (Morris *et al.*, 2002).

Unlike cellular communities, viruses contain no universal gene. However, the major capsid protein gene, (*g23*), of T4-like-myovirus superfamily has often been used to study viral dynamics as they are abundant in the ocean (Breitbart *et al.*, 2002; Filée *et al.*, 2005; Yooseph *et al.*, 2007) and infect a variety of marine bacteria including cyanobacteria (Sullivan *et al.*, 2003; Marston and Amrich, 2009; Ignacio-Espinoza and Sullivan, 2012), SAR11 (Zhao *et al.*, 2013) and gammaproteobacteria (Nolan *et al.*, 2006; Petrov *et al.*, 2010). A recent report suggests that marine Thaumarchaea may also be infected by these myoviruses (Labonté *et al.*, 2015). As assessed by molecular fingerprinting studies of g23, the T4-like-myovirus community composition can be relatively stable over days to weeks, for example, during summer (Needham *et al.*, 2013), but with seasonally distinctive communities over months to years (Chow and Fuhrman, 2012; Pagarete *et al.*, 2013; Goldsmith *et al.*, 2015).

Here we report on our application of the high-resolution sequence analyses to the 16S rRNA of bacteria, archaea, 16S chloroplast sequences of eukaryotic phytoplankton, ITS of SAR11 and g23 sequencing of T4-like-myoviruses during and following a marine spring phytoplankton bloom. Such blooms, representing somewhat predictable and large ecological disturbance, happen annually over much of the world ocean and are particularly important events to study to more fully understand global nutrient and energy cycles (Ducklow *et al.*, 2001; Teeling *et al.*, 2012, 2016). We use mock communities generated from environmental sequence clones to examine the precision and accuracy of the 16S and g23 assays and to validate the application of the high-resolution sequencing analysis methods. From environmental sequences, we assess the extent that 99% sequence clusters, operational taxonomic units (OTUs) can be 'decomposed' (that is, broken into finer resolved taxa), and the extent to which these finer resolved taxa are ecologically distinct. We then examine the ecological dynamics of these finely resolved taxa and the extent that fine resolution improves observation of apparent ecological associations.

Materials and methods

Sampling and processing

Seawater was collected and filtered from the top meter at the San Pedro Ocean Time-series (33°33' N,

118°24' W) between March 12 and September 28, 2011, at a daily-to-weekly resolution as previously described (Needham and Fuhrman, 2016). Methods and data for the physical, chemical and seawater size fractionation for community composition, and 16S PCR amplification have been previously described and are published (Needham and Fuhrman, 2016). We studied two size fractions, calling the organisms caught on the Gelman type A/E glass filter (approximate pore size 1 μm) 'large and/or attached,' and those passing that filter and caught on the 0.22 μm pore size Durapore filter 'small and free-living.'

g23 and SAR11 sequencing assays

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′-GAYHTIKSIGGIG-TICARCCIATG-3′) and T4-SuperR1 (5′-GCIYKIARRT-CYTGIGCIARYTC-3′) (Chow and Fuhrman, 2012), followed by additional rounds of PCR with Illumina sequencing adapters (Supplementary Methods).

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). This was done by searching for the 'universal' SSU *rRNA* gene primer 1492R (5′-TACGGCTACCTTGTTACGAC TT-3′) against the GOS and GoMA metagenomes using BLASTn (Altschul *et al*., 1990) (word size = 7) and retaining reads with <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 367–447 bp (see Supplementary Methods for PCR cycling conditions).

Mock community generation

16S 'even' (10 clones, 10% each) and 'staggered' (range 0.01–35%) 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' T4-like-myovirus mock communities from San Pedro Ocean Time-series environmental g23 sequence clones (Chow and Fuhrman, 2012; Supplementary Methods).

Sequence analysis

All sequence analysis commands and steps are available via Figshare (https://doi.org/10.6084/m9.figshare.4546813) and see Supplementary Methods. Briefly, g23 and SAR11 ITS sequences were trimmed via Trimmomatic (Bolger *et al*., 2014) merged length in USEARCH7 with *fastq_mergepairs* (Edgar, 2013). g23 sequences that had stop codons or unidentified residues in all translation frames were removed (Rice *et al*., 2000). For SAR11 ITS and g23 sequences, chimeras were detected *de novo* via *identify_chimeric_seqs.py* within QIIME (Caporaso *et al*., 2010) with USEARCH61 (settings: usearch61_minh 0.05, usearch 61_mindiffs 1, usearch 61_xn 2) and removed. Ninety nine percent 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)); see Supplementary Methods for details.

Decomposition of 99% OTUs

We used minimum entropy decomposition (MED) (Eren *et al*., 2014) to decompose (that is, 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. MED utilizes sequence alignment positions of high variation, higher than background sequence variation, to partition sequences into homogenous types that can be differentiated from close relatives by as little as a single-specific base. Although MED was designed to partition the full data set at once, we performed the analysis on individual OTUs because we were interested, specifically, in the amount of variation obscured by standard cluster OTUs; our approach is similar to the related Oligotyping approach, which is the predecessor of the MED program (Eren *et al*., 2013). 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 we determined Shannon entropy (a metric that assigns a value to a string of characters based on the amount of variation observed (Eren *et al*., 2014)). We used the 16S mock communities, made from cloned environmental sequences, to validate the OTU sequence decomposition approach that we used. Sequence variations from the pure clones, from a combination of PCR, cloning and sequencing error, were generally minor and randomly distributed throughout the alignments (Supplementary Figures 2 and 3). We used this variation within what should have been identical sequences in the 16S mock community to assign a threshold for determining significant sequence variations at specific positions in environmental sequences. We did this by observing the Shannon entropy associated with each position in the alignments of the 25 clones retrieved after sequencing. Shannon entropy is zero for a position that always has the same base, 1.0 for a position that has 50% each of two bases and a maximum of 2.0 with 25% each of all four bases. The threshold value we used to distinguish 'real' variants from background, 0.25 (~1 base different every 24 bases within

a single alignment position), is above the maximum background variation value observed in the 16S mock communities (0.24) and 13× higher than the average of 'background' Shannon entropy (0.019 ± 0.019 s.d.) sequence variation, and about 3× the default used in the MED program (Eren *et al.*, 2014). Many of the 16S mock community sequence clusters, when observed in the environmental samples, had numerous wild relatives with positions of high entropy and were thus decomposed into multiple amplicon sequence variants (ASVs) (Supplementary Figure 2). The alignments of environmental sequences that had entropy at sites >0.25 were decomposed based on those positions, and decomposition continued until all positions had entropy <0.25. The following additional thresholds were also set to reduce the likelihood of erroneous sequences being considered 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.

For the analyses where we investigated prevalence, abundance and taxonomic patterns related to microdiversity, we randomly subsampled the sequences from each OTU 10 times to the minimum number of sequences observed for any OTU in the 16S chloroplast, 16S prokaryote or g23 data sets (that is, relative abundance threshold of 0.4% on average or above 2.5% on any day of the time series, as previously described). The minimum sequences for the different data sets were 724 for prokaryote 16S, 979 for chloroplast 16S and 1085 for g23. The minimum number of sequences for a SAR11 ITS OTU was 4288, but to be comparable to the others we subsampled the SAR11 ITS data to 1000 reads. As an extra test to examine the prevalence and microdiversity, we also subsampled to 2500 sequences for each OTU (a threshold not met by 5, 13 and 8 OTUs from the prokaryote, phytoplankton and g23 sequences, respectively, so these were removed for this analysis). Then, each subset was decomposed with MED as defined above. The average of the 10 replicate sequence sets for each OTU is reported, with s.e.m.

Sequence identification

16S sequence classification. Ninety nine percent sequence similarity 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). 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 December 2016), 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://doi.org/10.6084/m9.figshare.4546813). We identified monotonic increases and decreases of OTUs and ASVs using the Mann–Kendall test within the 'Kendall' package in R (R Core Team, 2015). 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 an ASV is 50% of the parent and the parent was 5% of the whole community, then we used 2.5%. Pairwise correlations between estimated abundance of ASVs and OTUs were then determined via Spearman correlations ($P<0.001$, $Q<0.001$) as implemented in the local similarity analysis program (Xia *et al.*, 2013, 2011) on the types that were present >0.05% (relative abundance) on average and on at least 25% of sample dates. We used local similarity analysis 'mixed' method for determining the correlation significance, which determines *P*-values based on permutations only for correlations with explicitly determined *P*-values <0.05 (Xia *et al.*, 2013, 2011). False discovery rate, *Q*-values (<0.001), were used to account for multiple testing (Storey & Tibshirani 2003).

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, that is, 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.

The sequence alignment for heatmap phylogeny was generated 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 the nonsynonymous to synonymous mutation ratio (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 (Rice *et al.*, 2000). 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).

Results

Assessment of sequence analysis methods

To assess the precision and accuracy of the 16S and g23 PCR-to-analysis pipeline, we analyzed mock communities of environmental sequence clones pooled in known proportions. The staggered 16S community (25 clones, abundance range 0.1–33%) had an r^2 of expected vs observed of 0.97 (Supplementary Figure 2), similarly high to the 0.95 previously reported (Parada *et al*., 2016). We used the 16S mock communities (which should have identical sequences within each OTU) to distinguish what we consider significant specific sequence variations vs background sequence and PCR error within each OTU. We set the threshold for calling 'real' native sequence variants higher than the observed variation in any of the mock community clones, and 13 × the average background variation level (see Materials and methods section). For the g23 mock community (9 OTUs, 11% each, which ranged in length from 276 to 387 bp (mean length 349 ± 30 (s.d.)), the observed abundances ranged from 4 to 20% (Supplementary Figures 4a–c) and resulted in a single ASV for each clone (with one exception, possibly due to an early PCR error). There was a weak negative correlation between g23 sequence length and observed relative abundance in the even g23 mock community (Supplementary Figure 4d). Thus, while we had no mock community for SAR11 ITS sequences, we expect that length, in addition to primer bias, may play a small role in determining the relative abundance of SAR11 ITS (and g23) sequences observed (Supplementary Figure 4d).

Diversity and microdiversity within microbial communities

The sequences from the novel SAR11 ITS sequencing assay, in general, had high similarity to previously published genome and metagenomic SAR11 sequences (Supplementary Figure 5) and the g23 sequences, while often very divergent from published genome sequences, were predominantly identifiable as capsid protein sequences (Supplementary Figure 6). We clustered the 16S, SAR11 ITS and g23 data sets at a relatively highly resolved threshold of 99% similarity as a starting point for investigating finer-scale sequence resolution. Because high-resolution sequence analyses depend on high coverage to separate actual sequence differences from sequencing errors, we focused on the decomposition of OTUs on the taxa that were 'abundant' (>0.4% on average or >2.5% on at least 1 day). The percentage of OTUs that decomposed into at least two ASVs varied between the different genes: SAR11 ITS most (100%), 16S rRNA second (59%) and g23 least (49%) (overall mean: 60%; Table 1). Likewise, the average number of ASVs per OTU varied between the different genes with the

Table 1 Summary of OTU decomposition of 16S bacteria, archaea, chloroplasts of phytoplankton, g23 of T4-like-myoviruses and SAR11 ITS OTUs

OTU taxon and gene used	*Decomposed OTUs/total*	*ASVs per OTU*	*Number of ASVs whose proportions vary non-randomly over time (P<0.001, P<0.005)*	*Number OTUs with >1 ASV observed as 'most abundant' (within an OTU) on some day during study*
Bacteria and Archaea 16S	*52/78*	*2.9*	*20, 21*	*21*
Actinobacteria	0/2	1	NA, NA	NA
Bacteroidetes	20/30	2.2	9, 12	8
Cyanobacteria	2/3	3.7	1, 2	2
Euryarchaeota	4/4	3	3, 3	4
Planctomycetes	1/1	3	0, 0	0
Proteobacteria	22/28	4.2	11, 14	7
Unassigned	1/4	1.3	1, 1	0
Verrucomicrobia	2/6	1.3	1, 1	0
Phytoplankton chloroplast 16S	32/64	2.5	16, 25	26
Prasino-clade 7	0/1	1	NA	NA
Chlorophyta	2/8	1.4	1, 2	0
Cryptophyta	2/3	2.0	1, 1	2
Haptophyceae	16/26	3.5	8, 12	13
Stramenopiles	12/26	2.1	6, 10	11
T4-like-viruses g23	35/72	1.9	11,15	7
SAR11 via ITS	26/26	12.4	24, 25	19
SAR11 via 16S	4/4	7	3, 4	0
Total	145/240	3.5	71, 86	80

Abbreviations: ASV, amplicon sequence variant; ITS, intergenic spacer sequence; NA, not applicable.

ITS, 16S and g23 being decomposed into an average of 12.4, 2.7 and 1.9 ASVs, respectively (Table 1). For the g23 sequences, 18 out of 21 of OTUs with >2 ASVs resulted in synonymous mutations, that is, indicative of purifying selection (Supplementary Table 1).

To try to explain the differing amounts of microdiversity within the OTUs, we examined whether temporal ubiquity, or prevalence, was related to the amount of microdiversity. To control for the fact that more abundant taxa have a greater chance to display rare variants (and thus exceed the minimum threshold for calling new ASVs), we subsampled the prokaryotic, chloroplast and g23 data sets to the minimum number of sequences present within an OTU for each data set (724, 979, 1085, respectively). For all groups studied, we observed that OTUs with the highest microdiversity occurred in taxa that were highly prevalent (present every day) although some prevalent taxa had little microdiversity (Figure 1). In this subsampled data set, the majority of taxa below 0.7 prevalence had only one ASV, with a maximum of 2. Above 0.7 prevalence, a larger fraction have 2 and some have three or more (up to 6) ASVs. When the OTUs are subsampled to higher numbers of sequences (2500), though a few rare OTUs have been excluded, the overall trend holds (Supplementary Figure 7). The data suggest that it is necessary to be prevalent (temporally common) in order to have high microdiversity, but other factors could limit the microdiversity even for the most prevalent taxa. Showing that microdiversity is not related to abundance alone, when plotted vs OTU abundance, microdiversity was highest for intermediate, not the highest abundance taxa (Figure 1; Supplementary Figure 7). Taxonomy was also a poor predictor of microdiversity as the amount of microdiversity varied within lineages (Supplementary Figures 7 and 8; Table 1; Supplementary Table 3).

To assess potential ecological differences between ASVs, we looked for evidence that the ASVs changed in relative proportions over time. To increase the observational power for this analyses we used all the sequences, not just the random subsets. Overall, 59% of the taxa that contained >1 ASV had at least one ASV which significantly changed monotonically as a proportion of the parent OTU over the 5 months of study (Mann–Kendall test, $P<0.005$) (Table 1). Furthermore, most decomposed taxa had more than one ASV, which became (at some date) most abundant within an OTU: phytoplankton, prokaryotes and T4-like-myoviruses had 26, 21 and 7, respectively (Table 1). 19 SAR11 ITS OTUs had more than one ASV that became most abundant, but no SAR11 16S OTUs had more than 1 most abundant ASV. Furthermore, many OTUs had three or more ASVs that became over 50% of an OTU over the 5 months of study, including eight phytoplankton OTUs (Figure 2), three prokaryotic OTUs in both size fractions (Figure 3) and eight prokaryotic OTUs in either size fraction (Supplementary Figure 9). Ten SAR11 ITS OTUs had multiple ASVs that became most abundant within their respective OTU. In contrast, the g23 data set had no OTUs that had three or more ASVs become most abundant, but seven with more than one (Supplementary Figure 10). Notably, there was a remarkably large number (>30) of co-occurring and dynamic ASVs in the haptophyte genus *Chrysochromulina* (Supplementary Figure 11). We also noted that when an organism peaked in abundance, it generally tended to be strongly dominated by a single ASV

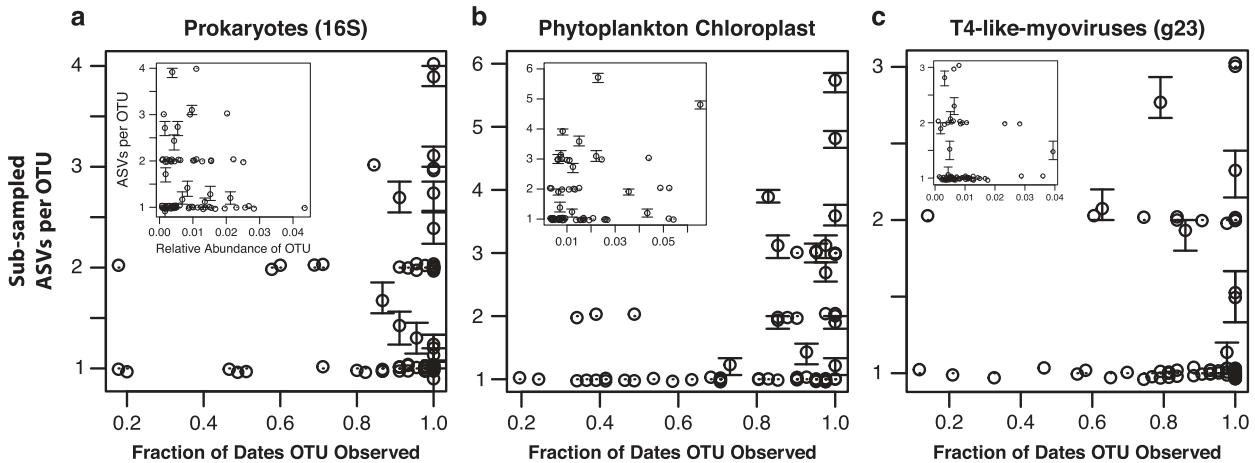

Figure 1 Number of ASVs associated with OTUs of various temporal ubiquity for (**a**) prokaryotes (**b**) phytoplankton via chloroplasts and (**c**) T4-like-myoviruses. Each data set was subsampled 10 times to the minimum number of sequences observed for a given OTU within the data set, and the average number of ASVs per OTU is shown along with the s.e.m. ASV data for prokaryotes include reads from both 1–80 and 0.22–1 μm size fractions. Occurrence for prokaryotes can be in either small and free-living or large and particle-attached size fractions. Data are plotted with a slight vertical jitter to help with over-plotted points. One chloroplast OTU was omitted (fraction dates observed = 0.95) because no ASV exceeded the minimum number of sequences required to be considered (that is, the OTU was very microdiverse). Inset graphs show relationship between ASV and relative abundance of each OTU, showing the patterns in the large graphs are not due to OTU abundance.

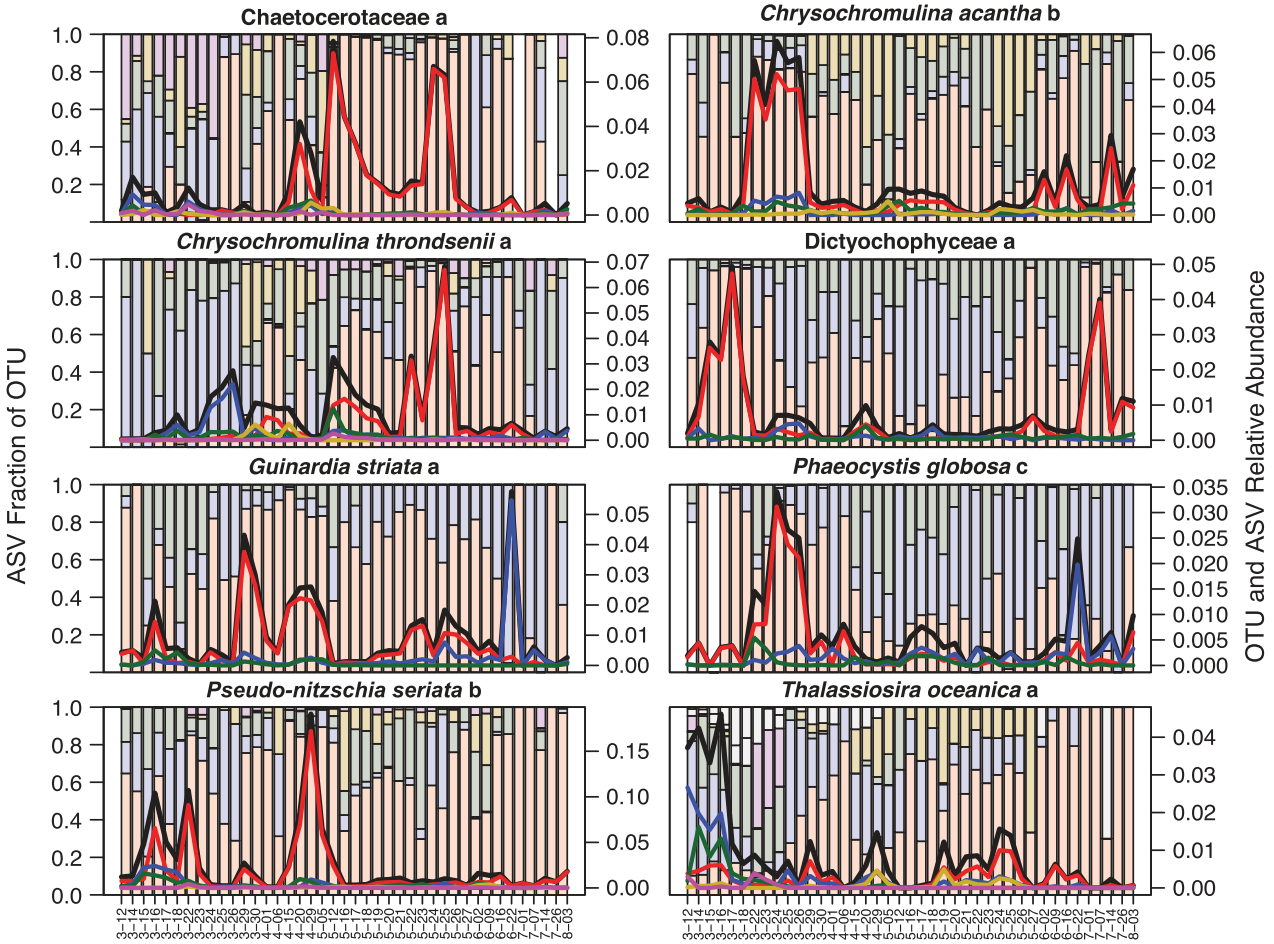

Figure 2 Dynamics of phytoplankton OTUs, which had >2 ASVs that became more than 50% of the sequences of a given OTU for at least one date (in which OTU abundance was >0.1%). The red, blue, green, purple or gold segments of bargraphs represent the proportion that each ASV made up of an OTU on a given day. That is, if a 'red' ASV is made up of 25% of the sequences of the OTU on a given day, the red bar is 25%. The black line represents the relative abundance over time of each OTU as a proportion of the whole community; the colored lines (which correspond to the colors of the bar segments) are the estimated relative abundance of each ASV as a proportion of the whole community of chloroplast sequences (that is, ASV proportion of OTU *OTU relative abundance of all chloroplast sequences). Only the top five most abundant ASVs estimated abundances, that is, lines, are shown for each OTU. All ASVs proportions making up >1% on average of an OTU, but not in the top five, are shown as individual gray bar segments (for example, at the top left of the '*Thalassiosira oceanica* a' panel). The letter following the taxon name corresponds to the average abundance of the taxa relative to other OTUs taxa with that name (a, most abundant; b, second most abundant and so on).

(Figures 2 and 3, exceptions include the diatom *Thalassiosira* in Figure 2), and in some cases different ASVs peaked in different months (for example, diatom *Guinardia* and haptophyte *Phaeocystis*, Figure 2), indicating ecological differences.

Overall dynamics of T4-like-myovirus ASVs

The relative abundance of T4-like-myovirus ASVs (ASV proportion of OTU× abundance of parent OTU) was highly dynamic from day-to-day during the initial decline of the phytoplankton bloom, with 9 different viral taxa that became most abundant over 18 nearly consecutive days of sampling (Figure 4a). Over the full 5 months, 40 sampling point, time series (Figure 4b), 17 different ASVs became most abundant. However, among the abundant T4-like-myovirus OTUs (>1% on average), the ASVs, where observed, tended to vary relatively little within OTUs, thus the dynamics of OTUs largely dictated the observed dynamics for these abundant T4-like-myoviruses (Figure 4c). The high variability of T4-like-myovirus taxa, especially during the initial bloom decline, is comparable to that observed in the large or particle-attached prokaryotic community, but greater than the small and free-living prokaryotic community where a single SAR11 OTU was most abundant on 30 of 39 sampling dates (Needham and Fuhrman, 2016).

SAR11 dynamics via 16S and ITS sequencing

The SAR11 clade had a large amount of sub-OTU microdiversity, with 28 16S ASVs (7 per OTU) and 322 ITS ASVs (12.4 per ITS OTU). Overall, the total SAR11 relative abundance tended to dictate the relative abundances of the SAR11 16S and ITS ASVs (ASV proportion of OTU× abundance of

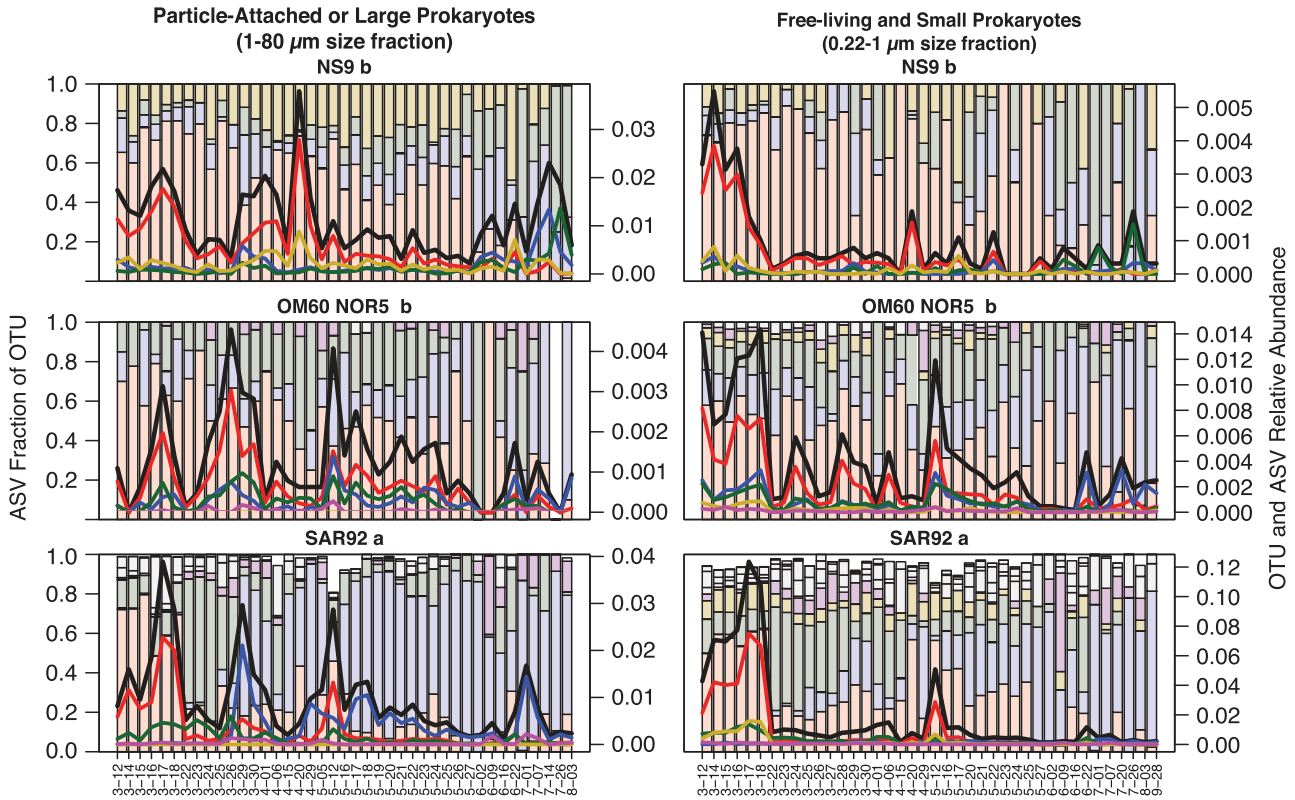

Figure 3 Dynamics of bacterial OTUs which had >2 ASVs that became more than 50% of the sequences for a given OTU for at least one date in both size fractions (not necessarily on the same day), in which OTU abundance was >0.1%. As in Figure 2, the black line represents the total abundance over time of each OTU; the colored lines (corresponding to bar segments) are the estimated relative abundance of each ASV (that is, ASV proportion of OTU *OTU relative abundance). Only the top five most abundant ASVs estimated abundances, that is, lines, are shown for each OTU. All ASVs proportions making up >1% on average of an OTU, but not in the top five, are shown as individual gray bar segments. The letter following the taxon name corresponds to the average abundance of the taxon relative to other taxa with that name (a, most abundant; b, second most abundant and so on).

parent SAR11 OTU × total SAR11 relative abundance via 16S) (Figures 5a and b; Supplementary Figure 12). However, the estimated abundances of the ASVs did vary from one another (Table 1; Figures 5a and b), with the variation being much more considerable for the ITS data, including a change in the most abundant ASV during the March phytoplankton bloom decline, which was not observed in the 16S data (Figures 5a and b). In both 16S and ITS, the dominant ASV decreased as a proportion of total SAR11 toward the summer, whereas a variety of ASVs, from the dominant OTU and others tended to increase (Figures 5a and b; Supplementary Figure 12). On the other hand, SAR11 S4 peaked in April (Supplementary Figure 12), briefly becoming the most abundant SAR11 ITS ASV then decreasing rapidly, coincident with an overall decrease in SAR11 abundance.

Correlations among viruses and cellular communities

Overall, the best predictor of the T4-like-virus community composition (as assessed via the Mantel statistic) was the composition of the large or particle-attached prokaryotic community, followed by (successively) the small and free-living prokaryotic community, SAR11 ITS via ASVs community, and physical and chemical (environmental) conditions (Supplementary Table 2). Consideration of ASVs, rather than OTUs, often only marginally improved Mantel test Rho values (strength of community correlation), except in the case of SAR11 taxa, where the correlations of viruses to SAR11 ITS ASVs was much stronger than to SAR11 16S OTUs, 16S ASVs or ITS OTUs (0.49 vs 0.27 for SAR11 ITS ASVs to virus ASVs vs SAR11 ITS OTUs to myoviral OTUs) (Supplementary Table 2).

Beyond the whole-community level, pairwise correlations of ASVs revealed that many T4-like-myoviruses that were abundant during the initial bloom decline were positively correlated over the full-time series with a small, but diverse set of potential hosts, including SAR92, Flavobacteria (genera: *Polaribacter* and *Formosa*), and a Verrucomicrobium taxon (genus: *Persicirhabdus*) (Spearman correlation >0.85, $P<0.001$; Figure 6a). Many strong positive pairwise associations to phage were also observed among T4-like ASVs that were abundant late in spring or summer, including *Synechoccocus* and *Prochlorococcus.* Many associations between viruses, prokaryotes and phytoplankton (Spearman's correlation >|0.85|) were consistent whether or

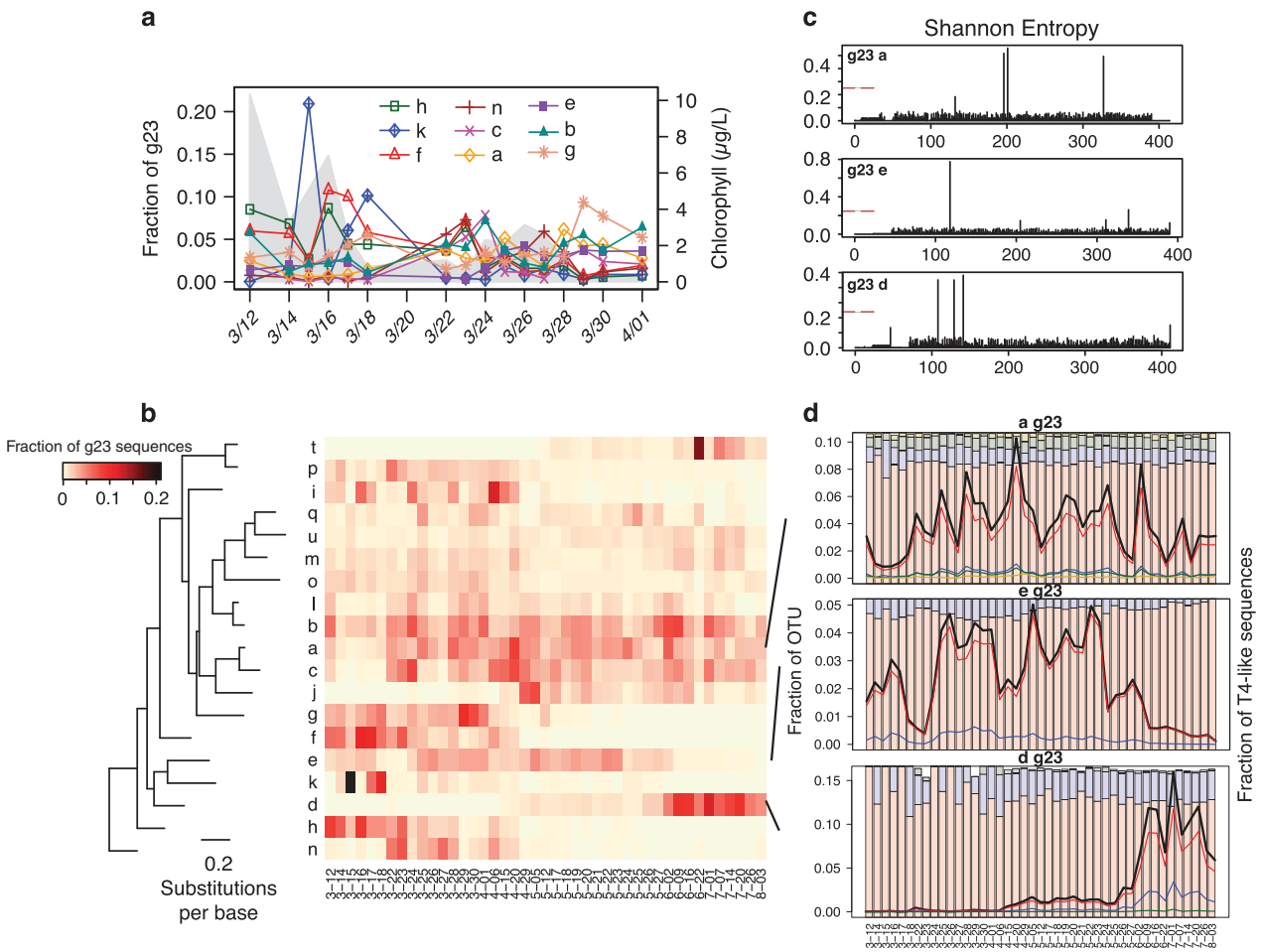

Figure 4 Dynamics and underlying diversity of T4-like-myovirus ASVs from San Pedro Ocean Time-series March–August 2011. (**a**) Dynamics of all T4-like-virus ASVs that were most abundant for at least 1 day during 12 March–1 April, along with chlorophyll concentration as the gray background (values along the secondary *y*-axis) and (**b**) for all ASVs that were >1% on average over the full-time series. (**c**) Underlying Shannon entropy and (**d**) dynamics are shown for three g23 OTUs (>1% on average) that showed significant changes in abundance of ASVs as proportion of the parent OTU during the time series (Mann–Kendall test, <0.005). For **d**, as in Figures 2 and 3, the black line represents the total abundance over time of each OTU; the colored lines (corresponding to bar segments) are the estimated relative abundance of each ASV (that is, ASV proportion of OTU *OTU relative abundance). g23 OTUs that had more than one ASV become most abundant are shown in Supplementary Figure 10. T4-like-myovirus OTUs did not have high sequence similarities to cultured representatives, so were assigned generic names (a, most abundant on average; b, second most abundant and so on).

not ASVs were considered, but the networks were much more sparse for the OTU networks (Figure 6b), with notable absences of some positive correlations such as T4-like phage to *Synechococcus* and Flavobacteria (*Formosa*), and *Chrysochromulina* to UCYN-A. This correlation to the UCYN-A ASV (an important symbiotic diazotroph (Thompson *et al.*, 2012)) was unexpected because we previously reported OTU level and slightly lower correlations to three taxa: *Mesopedinella,* Prasino-clade-7, and also the known host, *Braarudosphaera bigelowii*, not *Chrysochromulina* (Needham and Fuhrman, 2016).

Overall, correlations between myoviruses and SAR11 were not as high as the best correlations between myoviruses and individual, non-SAR11 prokaryotic taxa. However, there were many significant correlations ($P<0.001$, $r>0.7$) between SAR11 ITS ASVs and myovirus ASVs and the average positive pairwise correlation between SAR11 taxa and myovirus taxa was higher with consideration of ASVs rather than OTUs (0.61 vs 0.58, Welch's *t*-test, $P<0.001$) and the total number of correlations was much higher (Figure 7c). Like the other virus-to-bacteria correlations, most of these highly correlated taxa peaked either in March or summer (Figure 7; Supplementary Figure 12).

Discussion

We hypothesized that ecological interpretation of *in situ* dynamics of microbial population interactions would benefit from highly resolving sequence analytical methods as it is well known that ecologically distinctive microbial species are found within typical 97 or even 99% similarity clusters by 16S rRNA (Rocap *et al.*, 2002; Eren *et al.*, 2013; Tikhonov

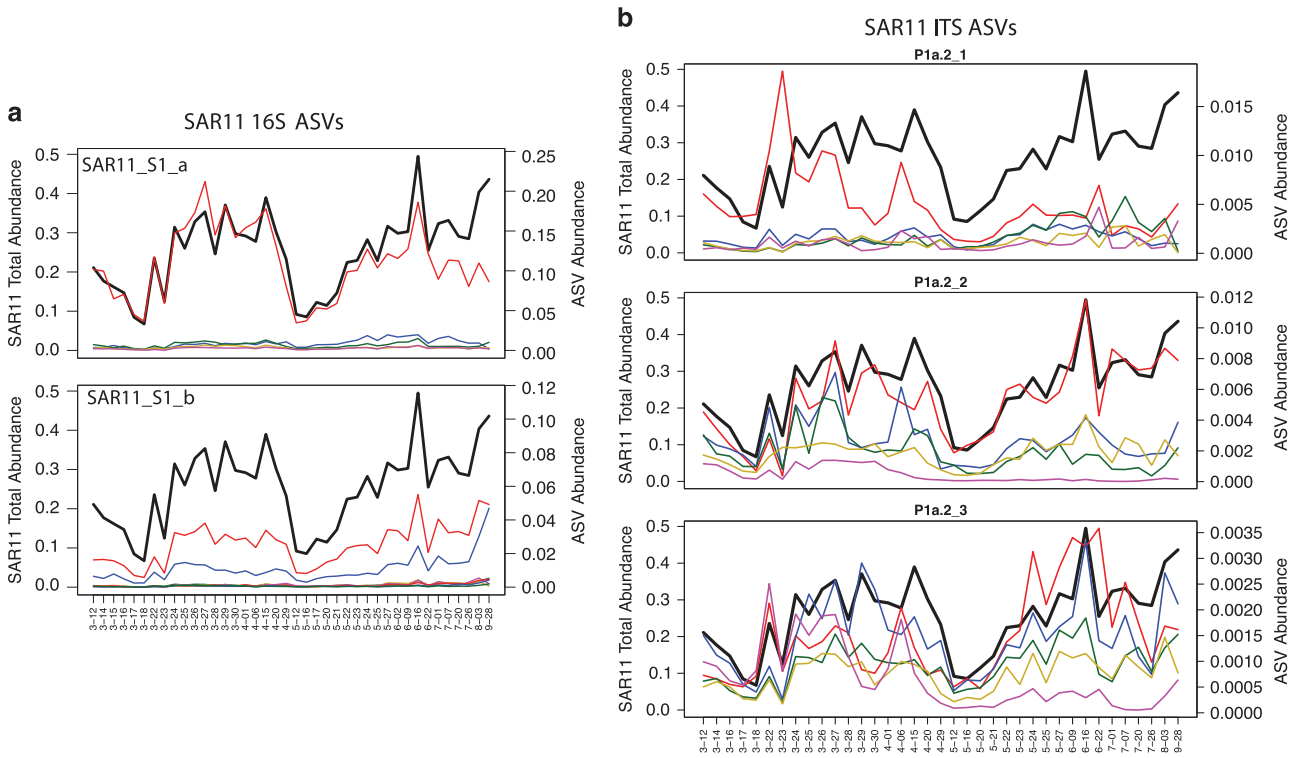

Figure 5 Dynamics of SAR11 (**a**) 16S and (**b**) ITS OTUs, and ASVs March–September 2011. The black line represents the overall relative abundance of SAR11 (all SAR11 OTUs cumulatively). Lines represent the ASV relative abundances (cumulative SAR11 16S relative abundance *ASV relative abundance) over time. Only the top five most abundant ASVs are shown. Note the *y*-axis on the left of each panel corresponds to the SAR11 total abundance and the *y*-axis on the right corresponds to ASV relative abundance. SAR11 16S taxonomy is from the SILVA taxonomy and the ITS taxonomy comes from Brown *et al.* (2012). The letter (16S) and number (ITS) following the last underscore in the taxon name correspond to the average abundance of the taxon relative to other OTUs with the same taxon name.

et al., 2015) and that such ecological interactions, such as between host and viruses, can be strain specific. By applying such methods, where we resolved finely as single-base variations in the ITS region, we found what appeared to be ecologically relevant patterns of phytoplankton, bacterial, archaeal and viral taxa, but the extent of microdiversity varied from taxon-to-taxon, with the more common taxa tending to have more microdiversity. In many cases, the patterns corresponded to distinctive temporal patterns and an improvement in detection of correlations between viruses and potential hosts, consistent with narrow host ranges of viruses. This was particularly true for the SAR11 clade, where ITS-based ASVs correlated best to viruses (Figure 7d; Supplementary Table 2). Further, our high-resolution analysis revealed a stronger Mantel correlation between myovirus community composition and particle attached or large prokaryotes (<1 μm) compared to free-living small prokaryotes. This is consistent with the hypothesis that particle attached or large bacteria are a major source of free viruses (Riemann and Gros sart, 2008; Bettarel *et al.*, 2015).

Extent of microdiversity

Our study used three genes, the 16S rRNA gene of bacteria, archaea and phytoplankton chloroplasts, g23 major capsid protein genes of T4-like-myoviruses viruses and ITS of SAR11 to examine the extent of microdiversity throughout the microbial community during and after a marine phytoplankton bloom, enabling assessment of the extent and ecological relevance of the microdiversity. About 60% of sequence cluster OTUs (99% sequence similarity clusters) had significant sequence variation within them (ASVs); within about 59% of those OTUs were ASVs with distinctive time-based pattern of abundance that suggest the sequence variation is ecologically relevant. We consider these taxa to be ecologically distinctive units, not neutral variants, because population sizes of bacterial lineages in the ocean are so large that we do not expect founder effects or genetic drift to have any effect at these time and spatial scales (Luo *et al.*, 2014). Accordingly, changes in the relative abundance of variants are considered to be the result of selection.

The extent of microdiversity varied across taxa and we found the highest microdiversity was associated with the most temporally ubiquitous parent taxon, and less associated with taxonomy (for 16S OTUs) or abundance. Low microdiversity was found both in prevalent taxa and rare taxa. Thus it appears from the 99% OTUs that the more the ephemeral the taxon, the more likely we would observe them to have little microdiversity, even

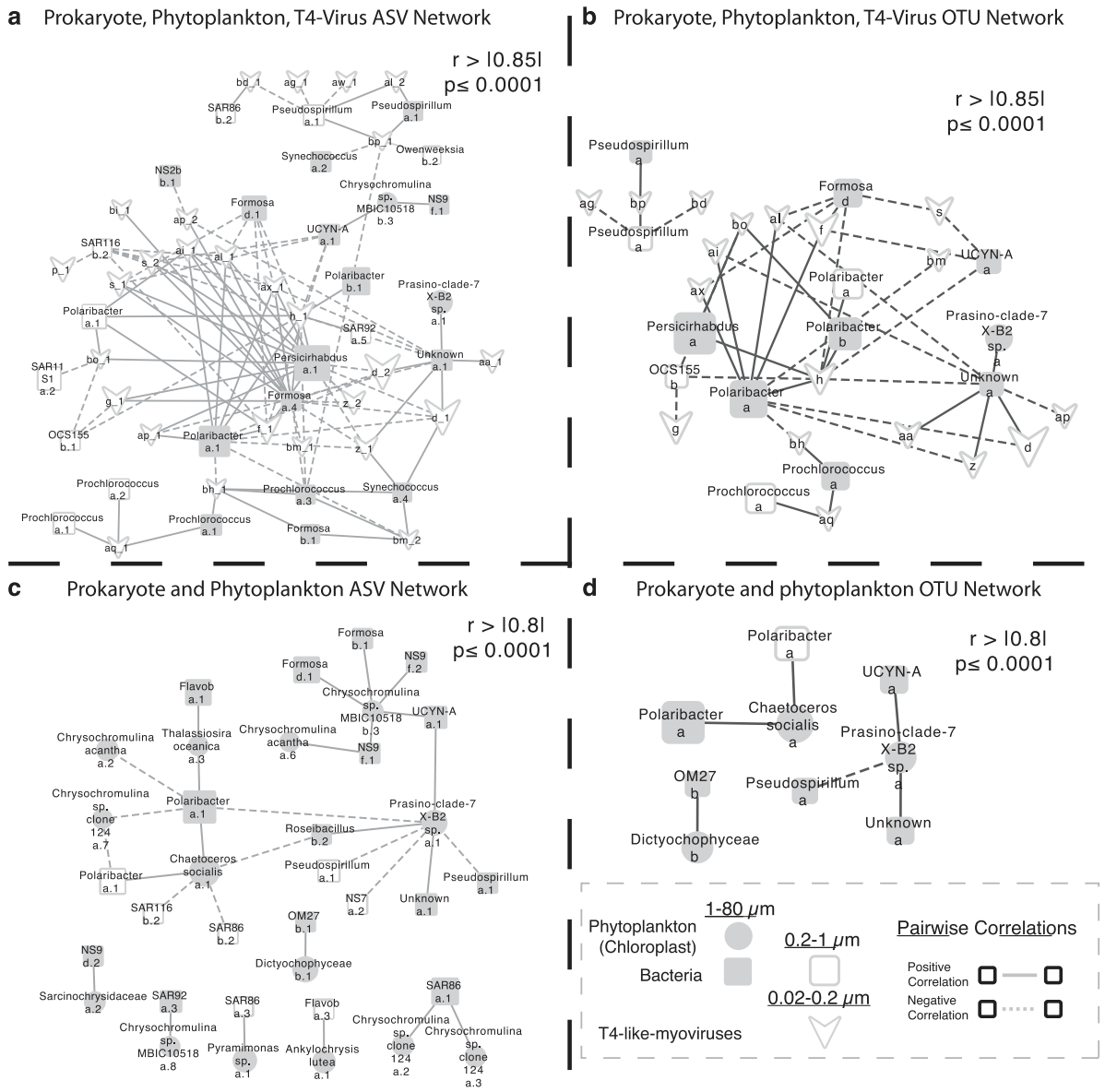

Figure 6 Pairwise correlation networks show correlations (Spearman >0.85, $P<0.001$) between prokaryotes and phytoplankton or T4-like-myoviruses for (**a**) ASVs and (**b**) OTUs over the full-time series. Slightly lower correlations are shown between (**c**) phytoplankton and prokaryotes ASVs (Spearman >0.8, $P<0.001$) and (**d**) OTUs over the full-time series. In each network, gray-filled circles represent phytoplankton, gray filled-squares represent bacteria from the 1-80 μm size fraction, unfilled squares represent bacteria from 0.2-1 μm size fraction, and 'V' symbols represents T4-like-viruses. Thick outline surrounding nodes indicates the node that represents the dynamics of a taxon within the large or particle-attached fraction. The size of the symbols represents the average relative abundance of each taxon. Solid and dashed lines represent positive and negative, respectively, correlations between the connected nodes. As in Figures 2 and 3, the letters following taxon name correspond to the average abundance of the taxon relative to other OTUs with the same name. The number following these letters (**a** and **c**) corresponds to the average abundance of the ASV within a given OTU. A full color version of this figure is available at the *The ISME Journal* online.

when accounting for abundance of the taxon (by sub-sampling). Without exhaustive sequencing effort, rare variants will always be missed, especially considering our conservative approach to calling ASVs (requiring at least 50 sequences across all samples); thus the taxa we observed as single ASVs could have rare variants.

As we need to focus our analysis on OTUs with enough sequence reads to discriminate between significant base changes vs errors, we have fewer intermediate and low abundance taxa for comparison. In any case, it appears that there is not a strong linear relationship between microdiversity and prevalence. The data suggest that it is necessary to be prevalent (temporally common) in order to have high microdiversity, but other factors could limit the microdiversity even for the most prevalent taxa.

SAR11 microdiversity was very high, especially when assayed by the ITS sequencing assay, with 322 ASVs occurring among the dominant OTUs ($n=26$) that we investigated for microdiversity. This amount of marine microbial diversity is not surprising, as a

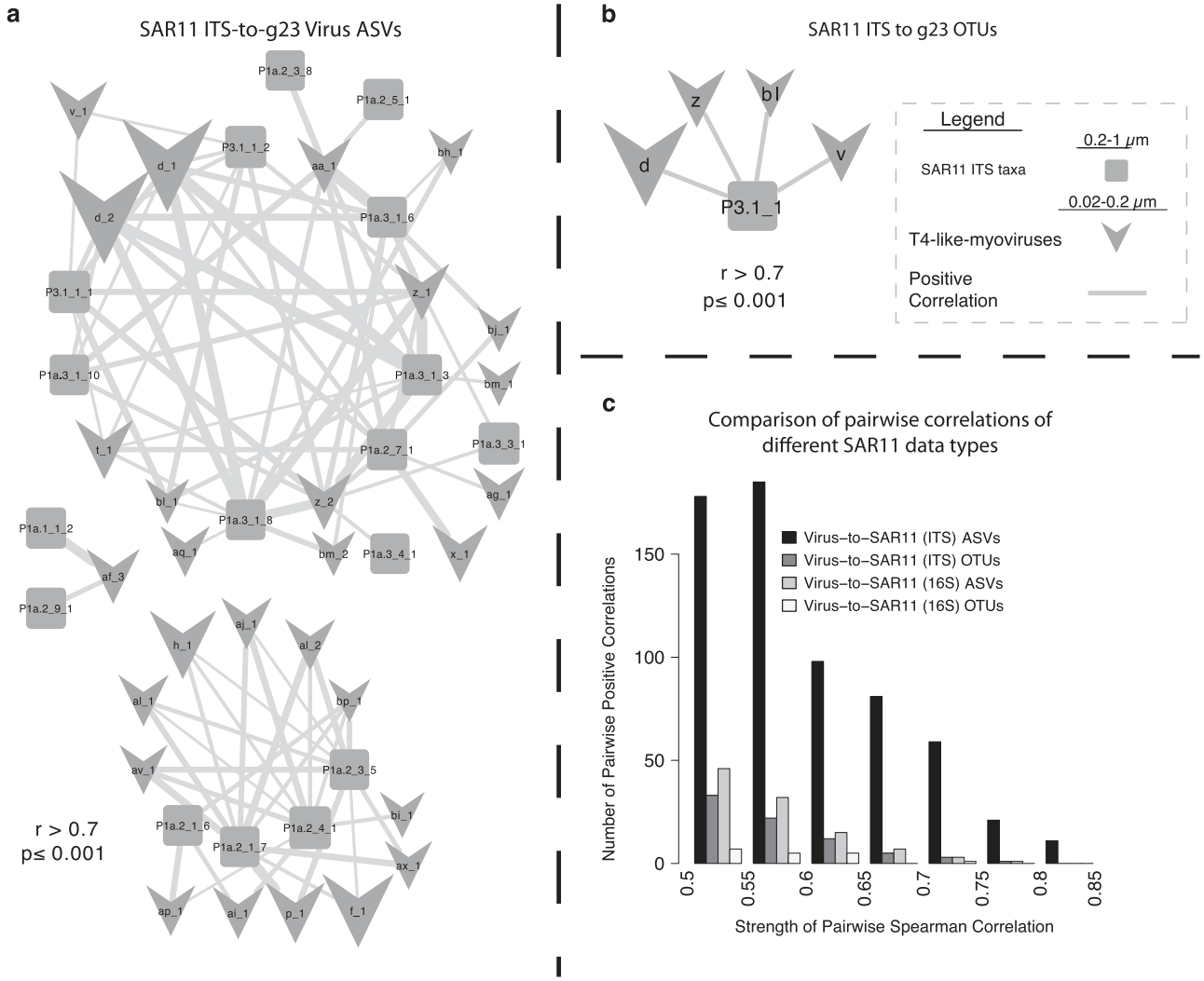

Figure 7 Correlation network between T4-like-myovirus ASV and (**a**) SAR11 ITS ASVs and (**b**) SAR11 ITS OTUs showing many more and higher correlations at the ASV level. (**c**) The total number of significant positive correlations ($P<0.001$, $Q<0.001$) between SAR11 (ITS and 16S) OTUs and ASVs to virus OTUs and ASVs. All SAR11 data from 0.22 to 1 μm size fraction. As in Figure 5, ITS taxonomy comes from Brown *et al*. (2012) and the number following the first underscore in the taxon names, corresponds to the average abundance of the taxon relative to other OTUs with the same taxon name. The value following these numbers (**a**), corresponds to the average abundance of the ASV within a given OTU.

previous study of 139 cloned SAR11 ITS sequences reported that no two were identical (Brown and Fuhrman, 2005). More recently, Kashtan *et al*. (2014) concluded *Prochlorococcus* is made of hundreds of co-occurring distinct lineages, each millions of years old, based on analysis of 100 s of single *Prochlorococcus* cells, which showed extensive genomic variations that were also reflected in small variations in ITS sequences. The genomic changes were often associated with transporters and cell surface biosynthesis or modification. These cell surface variations could result in the success or failure of attachment, and thus infection, by viruses, and may alter grazing susceptibility. By analogy, we expect that the SAR11 ITS ASVs may be from organisms with differing surface properties, and this may partly explain better correlations to viruses at the ITS ASV level.

Compared to the bacterial subtypes, the T4-like-myovirus subtypes were less likely to vary significantly over time as a proportion of their parent OTUs. For the g23 OTUs where g23 ASVs did not vary significantly as a proportion of their parent OTU, it is likely that these taxa infect the same hosts during this period of study. On the other hand, g23 ASVs that vary significantly as a proportion of the parent OTU may infect different taxa within lineages or with varying levels of efficiency (Holmfeldt *et al*., 2014; Howard-Varona *et al*., 2016).

Ecological implications

The extent of microdiversity within the various lineages allows inference of ecological and evolutionary forces that play roles in shaping the function, lifestyle and diversity of lineages, which has

important ecological implications. Our data show that the highest microdiversity consistently occurred among the most prevalent taxa, even though other prevalent taxa might have lower or no observed microdiversity. In contrast, the rarely occurring taxa generally had lesser or no observed microdiversity.

The common taxa, such as SAR11, SAR116 and Actinobacteria, maintain their moderately high abundance as primarily free-living plankton year round and therefore experience few, if any, bottlenecks. The sets of conditions that these lineages prefer are apparently nearly ubiquitous, but the contrast in the amounts of diversity within these common taxa (Actinobacteria being markedly lower) may reflect the influence of viral defense or the result of individual SAR11 variant taxa being highly specialized for very specific but stable niches. For a given population, one mode of defense to viruses is the presence of many subtypes (as we observed in SAR11 and SAR116) such that strain-specific viral infection of a single sub-type will not result in decimation of the whole population, as has been suggested for *Prochlorococcus* (Avrani *et al.*, 2011; Kashtan *et al.*, 2014). These slight changes would probably have competitive trade-offs, and the lineages with the most competitive strains have been modeled to become most abundant under this scenario (Thingstad *et al.*, 2014). In contrast, some common lineages have fewer co-existing close relatives (for example, OCS 155 Actinobacteria). One explanation for how taxa may have relatively little microdiversity is a recent genetic sweep, whereby an adaptive trait evolved within a strain and other similar strains have been outcompeted (Acinas *et al.*, 2004; Polz *et al.*, 2006). Whether or not these taxa with low microdiversity utilize a different mode of defense that enables them to maintain their abundance or if their population-level diversity is just not evident in the 16S sequences remains to be seen (note that many SAR11 subtypes required ITS-level resolution to resolve). Functionally, the high diversity in the SAR11 clade and SAR116 clade may correspond to flexibility in terms of usable substrates and/or preferred substrate concentration ranges within the lineage. Thus, the ecosystem functions performed by these highly microdiverse taxa may have greater resilience to changes or disturbances (Roger *et al.*, 2012; Sjöqvist and Kremp, 2016).

The rare taxa with relatively less microdiversity include organisms such as 'copiotrophs,' that are successful in short-lived, nutrient-rich conditions such as phytoplankton blooms. This may be the result of clonal expansion from a small seed population and/or gene exchange when growing in close proximity to each other, for example, when growing on or near nutrient sources, such as live or decaying phytoplankton. The success of these lineages is probably dictated by exploitation of relatively narrow niches or sets of environmental conditions.

Although space limitations prevent interpretation of the many particular microbial associations we observed (Figure 7), we note that these results can point to focused hypotheses about numerous possible microbe–microbe interactions such as viral infection, parasitism, nutrient exchange or specific symbioses.

There are limited data on what percent identity constitutes a single virus population, but a recent report regarding one marine cyanobacterial host suggested >95% average nucleotide identity overall (Deng *et al.*, 2014). Unfortunately, there are no high sequence similarity matches (not even at >80% amino-acid identity) among the T4-like-myovirus taxa that we investigated, compared to cultivated phages, including the only known T4-like-myovirus infecting a member of the SAR11 cluster. Furthermore, the extent of cross-infectivity is unclear, and myoviruses in particular are reported to have a relatively broad host range (Sullivan *et al.*, 2003). However among these g23 sequences from viruses with unknown hosts, we note that unlike the monthly scale bacteria–virus and bacteria–protist networks previously reported from San Pedro Ocean Time-series that had only 2–17% negative associations (Chow *et al.*, 2014), these higher resolution (time and phylogeny) networks show a much larger fraction (33–44%) of negative associations. This may be because at monthly scales the viruses primarily follow host abundances, but daily scales and high resolution better permit observation of predator–prey cycles.

We conclude that using highly resolving methods improves observation of ecological dynamics, and more so for common taxa than ephemeral taxa. Thus, for ecological studies, it is better to be 'too' taxonomically resolving and maintain the ability to *post hoc* merge taxa with ecological similar coherent patterns (Preheim *et al.*, 2013) than it is to not be able to resolve taxa with different ecologically distinct niches. However, this merging may be difficult because two taxa that are correlated over one time period, or set of conditions, may not be necessarily correlated over another set of conditions (Fuhrman *et al.*, 2015). Rather than being too resolving, marker genes analyses probably often obscure important variation, regardless of the resolution used (Kashtan *et al.*, 2014). And higher resolution also allows more accurate detection of dispersal. Genomic characterizations should, theoretically, be an improvement over marker genes, but given the difficulty in assembling close relatives from community sequencing data and the depth required to genomically track hundreds of types (Brown, 2015), it may be some time before we are able to observe the dynamics of many closely related host and virus populations via shotgun metagenomics. Traditionally, sequence clusters have been used to approximate microbial 'species'; however, our results show these classical OTU-based 'species' represent anything from highly microdiverse clusters to clusters with no apparent significant sequence variation. These factors likely contribute to the ephemerality, resilience, function and biodiversity of microbial lineages.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank the USC Wrigley Institute of Environmental Science, especially Roberta Marinelli, Sean Conner and Captain Gordon Boivin, and the crew of the Miss Christie for sampling opportunities and laboratory space. We thank Mark V Brown for his providing of SAR11 database used for classification of SAR11 ITS sequences. We thank Jennifer Chang, Cheryl Chow, Alle Lie, Sean McCallister and Elizabeth Teel for sampling assistance. We thank Dave Caron, Frank Corsetti, John Heidelberg, Eric Webb, Nathan Ahlgren, Lyria Berdjeb, Jacob Cram, Laura Gómez Consarnau, Erin Fichot, Alma Parada, Ella Sieradzki, Yi-Chun Yeh and Yuanqi Wang for insightful discussion and feedback on the manuscript. This work was supported by NSF Grants 1031743 and 1136818, grant GBMF3779 from the Gordon and Betty Moore Foundation Marine Microbiology Initiative and a National Science Foundation Graduate Student Research Fellowship to DMN.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Amin SA, Hmelo LR, van Tol HM, Durham BP, Carlson LT, Heal KR *et al.* (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**: 98–101.

Amin SA, Parker MS, Armbrust EV. (2012). Interactions between diatoms and bacteria. *Microbiol Mol Biol Rev* **76**: 667–684.

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. (2011). Genomic island variability facilitates Prochlorococcus–virus coexistence. *Nature* **474**: 604–608.

Bettarel Y, Motegi C, Weinbauer MG, Mari X. (2015). Colonization and release processes of viruses and prokaryotes on artificial marine macroaggregates. *FEMS Microbiol Lett* **363**: 1–8.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D *et al.* (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99**: 14250–14255.

Brown CT. (2015). Strain recovery from metagenomes. *Nat Biotechnol* **33**: 1041–1043.

Brown MV, Fuhrman JA. (2005). Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquat Microb Ecol* **41**: 15–23.

Brown MV, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T *et al.* (2012). Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* **8**: 595.

Brown MV, Schwalbach MS, Hewson I, Fuhrman JA. (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**: 1466–1479.

Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. (2016). DADA2 : high resolution sample inference from amplicon data. *Nat Methods* **13**: 581–583.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Chow C-ET, Fuhrman JA. (2012). Seasonality and monthly dynamics of marine myovirus communities. *Environ Microbiol* **14**: 2171–2183.

Chow C-ET, Kim DY, Sachdeva R, Caron DA, Fuhrman JA. (2014). Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J* **8**: 816–829.

Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P *et al.* (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513**: 242–245.

Ducklow HW, Steinberg DK, Buesseler KO. (2001). Upper ocean carbon export and the biological pump. *Oceanography* **14**: 50–58.

Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996–998.

Eren AM, Borisy GG, Huse SM, Mark Welch JL. (2014). Oligotyping analysis of the human oral microbiome. *Proc Natl Acad Sci USA* **111**: E2875–E2884.

Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG *et al.* (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* **4**: 1111–1119.

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9**: 968–979.

Filée J, Tetart F, Suttle CA, Krisch HM. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci USA* **102**: 12471–12476.

Fuhrman JA, Caron DA. (2016). Heterotrophic Planktonic Microbes: Virus, Bacteria, Archaea, and Protozoa. In: Manual of Environmental Microbiology, Yates M, Nakatsu C, Miller R, Pillai S (eds), ASM Press: Washington, DC, USA, pp 4.2.2 –1.4.2.2 –34.

Fuhrman JA, Cram JA, Needham DM. (2015). Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**: 133–146.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* **103**: 13104–13109.

García-Martínez J, Rodríguez-Valera F. (2000). Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* **9**: 935–948.

Goldsmith DB, Brum JR, Hopkins M, Carlson CA, Breitbart M. (2015). Water column stratification

structures viral community composition in the Sargasso Sea. *Aquat Microb Ecol* **76**: 85–94.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Holmfeldt K, Howard-Varona C, Solonenko N, Sullivan MB. (2014). Contrasting genomic patterns and infec tion strategies of two co-existing bacteroidetes podovirus genera. *Environ Microbiol* **16**: 2501–2513.

Howard-Varona C, Roux S, Dore H, Solonenko NE, Holmfeldt K, Markillie LM *et al.* (2016). Regulation of infection efficiency in a globally abundant marine bacteriodetes virus. *ISME J* **11**: 284–295.

Ignacio-Espinoza JC, Sullivan MB. (2012). Phylogenomics of T4 cyanophages : lateral gene transfer in the 'core' and origins of host genes. *Environ Microbiol* **14**: 2113–2126.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al.* (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science* **344**: 416–420.

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

Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ *et al.* (2015). Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J* **9**: 2386–2399.

Lu H-P, Yeh Y-C, Sastri AR, Shiah F-K, Gong G-C, Hsieh C. (2016). Evaluating community–environment relationships along fine to broad taxonomic resolutions reveals evolutionary forces underlying community assembly. *ISME J* **10**: 2867–2878.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. (2014). Comparing effective population sizes of dominant marine alphaproteobacteria lineages. *Environ Microbiol Rep* **6**: 167–172.

Marston MF, Amrich CG. (2009). Recombination and microdiversity in coastal marine cyanophages. *Environ Microbiol* **11**: 2893–2903.

Marston MF, Pierciey FJ, Shepard A, Gearin G, Qi J, Yandava C *et al.* (2012). Rapid diversification of coevolving marine synechococcus and a virus. *Proc Natl Acad Sci USA* **109**: 4544–4549.

Morris RM, Rappé MS, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Needham DM, Chow C-ET, Cram JA, Sachdeva R, Parada AE, Fuhrman JA. (2013). Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J* **7**: 1274–1285.

Needham DM, Fuhrman JA. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* **1**: 16005.

Newton RJ, Mclellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM *et al.* (2015). Sewage reflects the microbiomes of human populations. *MBio* **6**: 1–9.

Nolan JM, Petrov V, Bertrand C, Krisch HM, Karam JD. (2006). Genetic diversity among five T4-like bacteriophages. *Virol J* **3**: 30.

Pagarete A, Chow C-ET, Johannessen T, Fuhrman JA, Thingstad TF, Sandaa RA. (2013). Strong seasonality and interannual recurrence in marine myovirus communities. *Appl Environ Microbiol* **79**: 6253–6259.

Parada AE, Needham DM, Fuhrman JA. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environ Microbiol* **18**: 1403–1414.

Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD. (2010). Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol J* **7**: 292.

Polz MF, Hunt DE, Preheim SP, Weinreich DM. (2006). Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philos Trans R Soc B* **361**: 2009–2021.

Preheim SP, Perrott AR, Martin-Platero AM, Gupta A, Alm EJ. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**: 6593–6603.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria.

Reveillaud J, Maignien L, Murat Eren A, Huber JA, Apprill A, Sogin ML *et al.* (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J* **8**: 1198–1209.

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

Riemann L, Grossart HP. (2008). Elevated lytic phage production as a consequence of particle colonization by a marine flavobacterium (Cellulophaga sp.). *Microb Ecol* **56**: 505–512.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of prochlorococcus and synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180–1191.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M *et al.* (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739–751.

Roger F, Godhe A, Gamfeldt L. (2012). Genetic diversity and ecosystem functioning in the face of multiple stressors. *PLoS One* **7**: e45007.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson SJ, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Sjöqvist CO, Kremp A. (2016). Genetic diversity affects ecological performance and stress response of marine diatom populations. *ISME J* **10**: 2755–2766.

Storey JD, Tibshirani R. (2003). Statistical significance for genome wide studies. *Proc Natl Acad Sci USA* **100**: 9440–9445.

Sullivan MB, Waterbury JB, Chisholm SW. (2003). Cyanophages infecting the oceanic cyanobacterium prochlorococcus. *Nature* **424**: 1047–1052.

Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM *et al.* (2012). Substrate-

controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**: 608–611.

Teeling H, Fuchs BM, Bennke CM, Krüger K, Chafee M, Kappelmann L *et al.* (2016). Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *Elife* **5**: 1–31.

Tesson SVM, Montresor M, Procaccini G, Kooistra WHCF. (2014). Temporal changes in population structure of a marine planktonic diatom. *PLoS One* **9**: 1–23.

Thingstad TF, Våge S, Storesund JE, Sandaa R-A, Giske J. (2014). A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc Natl Acad Sci USA* **111**: 7813–7818.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D *et al.* (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**: 1546–1550.

Tikhonov M, Leach RW, Wingreen NS. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* **9**: 68–80.

Tully BJ, Nelson WC, Heidelberg JF. (2011). Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* **14**: 254–267.

Turlapati SA, Minocha R, Long S, Ramsdell J, Minocha SC. (2015). Oligotyping reveals stronger relationship of organic soil bacterial community structure with N-amendments and soil chemistry in comparison to that of mineral soil at Harvard Forest, MA, USA. *Front Microbiol* **6**: 1–16.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**: 77–80.

Xia LC, Ai D, Cram JA, Fuhrman JA, Sun F. (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* **29**: 230–237.

Xia LC, Steele JA, Cram JA, Cardon ZG, Simmons SL, Vallino JJ *et al.* (2011). Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol* **5**(Suppl 2): S15.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013). Abundant SAR11 viruses in the ocean. *Nature* **494**: 357–360.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)