ORIGINAL ARTICLE

Short-term observations of marine bacterial and viral communities: patterns, connections and resilience

David M Needham, Cheryl-Emiliane T Chow, Jacob A Cram, Rohan Sachdeva, Alma Parada and Jed A Fuhrman
University of Southern California, Department of Biological Sciences, Los Angeles, CA, USA

Observation of short-term temporal variation in bacterial and viral communities is important for understanding patterns of aquatic microbial diversity. We collected surface seawater once daily for 38 consecutive days with seven more samples interspersed over 40 more days at one location ~2 km from Santa Catalina Island, California. Bacterial communities were analyzed by automated ribosomal intergenic spacer analysis (ARISA) and viral communities were analyzed by terminal restriction fragment length polymorphism (TRFLP) of the conserved T4-like myoviral gene encoding the major capsid protein (g23). Common bacterial and viral taxa were consistently dominant, and relatively few displayed dramatic increases/decreases or 'boom/bust' patterns that might be expected from dynamic predator-prey interactions. Association network analysis showed most significant covariations (associations) occurred among bacterial taxa or among viral taxa and there were several modular (highly-interconnected) associations ($P \leqslant 0.005$). Associations observed between bacteria and viruses ($P \leqslant 0.005$) occurred with a median time lag of 2 days. Regression of all pairwise Bray-Curtis similarities between samples indicated a rate of bacterial community change that slows from 2.1%–0.18% per day over a week to 2 months; the rate stays around 0.4% per day for viruses. Our interpretation is that, over the scale of days, individual bacterial and viral OTUs can be dynamic and patterned; resulting in statistical associations regarded as potential ecological interactions. However, over the scale of weeks, average bacterial community variation is slower, suggesting that there is strong community-level ecological resilience, that is, a tendency to converge towards a 'mean' microbial community set by longer-term controlling factors.

Introduction

Marine bacterial and viral communities are diverse, active, interconnected and critical components to ecosystem production, nutrient recycling and biogeochemical pumps (Brussaard *et al.*, 2008; Fuhrman, 2009). Time-series studies aiming to observe the temporal dynamics in marine microbial communities typically do so at a monthly scale and indicate that bacterial and viral communities are seasonally dynamic and annually repeating (for example, Brown *et al.*, 2005; Fuhrman *et al.*, 2006; Carlson *et al.*, 2009; Eiler *et al.*, 2009; Campbell *et al.*, 2011; Eiler *et al.*, 2011; Gilbert *et al.*, 2011; Parsons *et al.*, 2011; Treusch *et al.*, 2011; Chow and Fuhrman, 2012; Giovannoni and Vergin, 2012). Extended, daily sampling might reveal ecological community dynamics—both subtle and dramatic— in bacteria and virus communities missed by

monthly sampling, and would help contextualize the longer-term monthly and intra-annual investigations. The few published reports so far using molecular fingerprinting approaches for these shorter time-scales have suggested that bacterial and virus community composition is relatively stable over days to weeks (Riemann and Middelboe, 2002; Hewson *et al.*, 2006; Steele, 2010).

Turnover times of bacteria and virus communities in the surface ocean is generally in the range of 3–5 days (Noble and Fuhrman, 2000) and can vary slightly over a diel cycle (Fuhrman *et al.*, 1985; Shiah, 1999; Clokie *et al.*, 2006). Over the course of days, viral production can be steady (Winget and Wommack, 2009) or dynamic: a study with extended daily sampling of a phytoplankton bloom showed very close coupling between an increase in bacterial and viral production following the bloom demise (Matteson *et al.*, 2011). This coupling had a 1-day time lag such that increased virus production preceded an increase in bacterial production; both corresponded to a dramatic environmental nutrient drawdown event.

In addition to short-term observations of production and abundance of marine bacteria and viruses,

molecular community analyses suggest that the rate at which the community composition changes is variable and likely depends on location and season. Using various community composition metrics such as Bray-Curtis similarity or Sørensen's index, bacterial community similarity is estimated to be between 80–92% between adjacent days (Acinas et al., 1997; Hewson et al., 2006). A study of virus community composition showed undetectable change over 48 h (Riemann and Middelboe, 2002). A week-long study of the bacterial community from the same coastal Catalina Island location of this study indicates that samples collected within a week of one another are very similar, that rare taxa are more variable than those that are more common and that bacterial taxa could increase their proportion in a fashion resembling a growth curve despite a stable 'background' community (Steele, 2010).

The role that viruses have in structuring microbial community composition has been investigated since the seminal observations that marine viruses outnumber and infect bacterial hosts (Bergh et al., 1989; Proctor and Fuhrman, 1990). An early and persisting hypothesis of how this influence might manifest is often referred to as the 'Kill-the-Winner' (KtW) hypothesis, which was initially a qualitative description of changes over time (for example, Fuhrman and Suttle, 1993) whereby abundant taxa are more likely to be infected than more rare hosts do primarily to potential encounter rates. This description would lead to the demise of the dominant taxa and open a niche for another to fill. Later, the hypothesis was formally quantified by Thingstad and Lignell, (1997) and Thingstad, (2000), and shown to include steady-state solutions under simplifying assumptions, including that faster growing bacteria are more susceptible to viruses. If a non-steady-state boom-and-bust dynamic occurs, it may have important implications on ecosystem robustness or resilience, that is, the ecosystem's ability to sustain its functions and processes at a relatively constant rate.

We aim here to observe short-term dynamics of marine bacterial and virus communities through daily sampling and molecular fingerprinting analyses. However, because all viruses do not share universally conserved genes, PCR studies of viral communities are more limited taxonomically than cellular life. We target the T4-like-myovirus group from which there have been many isolated representatives from the ocean infecting both cyanobacteria (for example, Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993; for a review see Clokie et al., 2010) and heterotrophic bacteria (for example, Wichels et al., 1998). Additionally, myoviruses make up a ubiquitous and significant percentage (about 9–42%) of sequence matches in aquatic viral metagenomes (Breitbart et al., 2002; Bench et al., 2007; Williamson et al., 2008). Ecologically, marine cyano-myoviruses have been shown to have broader host range compared with cyano-podoviruses and cyano-siphoviruses (Sullivan et al., 2003). We studied the T4-like-myovirus group via amplification of major capsid gene, g23, which has been shown to serve as a reasonable proxy for variation in globally ubiquitous myovirus genomes (Filée et al., 2005; Comeau and Krisch, 2008). Here we study the dynamics of this group in conjunction with the co-occurring bacterial community at one location in order to (a) investigate short-term dynamics of distinct bacteria and T4-like myovirus operational taxonomic units (OTUs) (b) detect statistical associations between distinct bacteria and viruses to understand how viruses may influence the bacterial community composition and determine potential phage-host relationships (c) quantify the rates at which the bacterial and myovirus community composition changes as a whole over days to months and (d) contextualize long-term observations of microbial community composition.

Materials and methods

Sampling
Surface water (10 l) was collected from the top 1 m depth at one geographic location (N 33° 27′ 11″, W 118° 29′ 2″) about 2 km from USC Wrigley Marine Science Center on Santa Catalina Island, CA, USA, by bucket from a small boat between 1100 and 1700 hours. The sampling location has a total depth of 85 m and is open to the San Pedro Channel. Samples were collected for 38 consecutive days (June 13, 2010—July 20, 2010) with seven additional samples collected over the following 40 days. Samples were processed within 2–3 h. Seawater was prefiltered through 80 μm nylon mesh and then filtered sequentially through 47 mm Type A/E glass fiber filter (~ 1.0 μm pore), 47 mm 0.22 μm Durapore (Polyvinylidene fluoride, PVDF, Millipore, Billerica, MA, USA), and 25 mm 0.02 μm Anotop (Aluminum-oxide, Whatman) filters; all stored at -80 °C.

Ancillary parameters
Seawater temperature was measured daily from the surface by a hand-held YSI 63 (Yellow Springs, OH, USA). Bacteria and viruses were enumerated using SYBR green epifluorescence microscropy (Noble and Fuhrman, 1998; Patel et al., 2007). Chlorophyll-a estimates were downloaded from NOAA CoastWatch browser using Science Quality 8-day composites from the MODIS sensor on-board NASA's Aqua spacecraft.

Microbial community analyses
Bacterial and myoviral communities were analyzed by observation of the variation of the length of the intergenic spacer in bacteria and of the major capsid protein gene of T4-like myoviruses (g23-terminal restriction fragment length polymorphism (TRFLP)). Bacterial DNA was extracted from a crushed

Durapore by phenol/chloroform chemical extraction (Fuhrman *et al.*, 1988) and viral DNA was extracted from the Anotop with Masterpure Epicentre kit (Steward and Culley, 2010) and stored at $-80\,°C$. Upon thawing, bacterial and viral DNA in the extract was quantified with PicoGreen (Invitrogen, Carlsbad, CA, USA) and diluted to $2\,ng\,\mu l^{-1}$ and $2.5\,ng\,\mu l^{-1}$, respectively, for analysis by automated ribosomal intergenic spacer analysis (ARISA) and TRFLP. Fifty microliter ARISA PCR reactions contained $2\,ng$ of bacterial DNA extract, $1\times$ Thermopol PCR Buffer (New England Biolabs, NEB, Ipswich, MA, USA), $10\,\mu M$ dNTPs (Promega, Madison, WI, USA), $20\,ng\,\mu l^{-1}$ BSA (Sigma-Aldrich, Catalog Number: A7030, St Louis, MO, USA), $5\,U$ Thermopol Taq polymerase (NEB). Primer concentrations, sequences and thermocycling for ARISA proceeded as described elsewhere (Brown *et al.,* 2005) except we used a 3-minute initial denaturation and 31 cycles of amplification. g23-TRFLP is described elsewhere (Chow and Fuhrman, 2012); only modifications are mentioned here. We used $2.5\,ng$ of viral DNA extract, $40\,ng\,\mu l^{-1}$ BSA and cycling conditions of $95\,°C$ for $3\,min$, 41 cycles of $95\,°C$ for $30\,s$, $55\,°C$ for $45\,s$ and $72\,°C$ for $45\,s$ with a final extension at $72\,°C$ for $5\,min$. HincII restriction digestion was carried out as in Chow and Fuhrman, (2012). ARISA and digested TRFLP PCR products were cleaned and concentrated using Zymo DNA Clean and Concentrate -5 kit and then diluted to 5 and $7.5\,ng\,\mu l^{-1}$, respectively. Purified and diluted ARISA and TRFLP PCR products ($1\,\mu l$) were loaded in duplicate onto an ABI 377XL sequencer in non-adjacent lanes as in Chow and Fuhrman, (2012). Only $5'$ TRFs were analyzed in this study.

ABI 377 peak analysis
Analysis of electropherograms was carried out similarly to previous studies (Steele *et al.*, 2011) with modifications (Chow and Fuhrman, 2012). Bacterial ARISA peak relative abundances from one PCR (machine duplicates) were averaged. These averaged OTUs were placed into bins created from a database built from >10 years of dynamically binned data from the nearby San Pedro Ocean Time-Series (SPOT). Dynamic binning of the SPOT OTUs was performed as in Ruan *et al.*, (2006a). The bin bounds were the lesser of either (a) the distance between adjacent upper (larger) bound of observed SPOT peaks or (b) the upper bound of a bin minus the maximum bin size. Maximum bin widths for fragments 390–450, 450–650, 650–900 and 900–1400 bp were 1, 2, 3 and 5 bp, respectively. If a peak did not fall into a predetermined bin from the SPOT database, a bin was created. OTUs separated by less than 0.2 bp were manually merged. Myovirus TRFs between 100 and 500 bp were dynamically binned with a 1 bp maximum bin size as described elsewhere (Ruan *et al.*, 2006a; Chow and Fuhrman, 2012). After binning, OTUs (peaks) with relative

abundances (that is, peak area divided by cumulative area of all peaks in that sample) less than 0.05% of the total area were removed and the data were renormalized.

Bacterial OTU identification
Putative identification of bacterial OTUs was determined by comparison with various taxonomic and ITS-length sources listed below in order of priority: (1) 16S rRNA gene of clone libraries (16S-ITS-23S) constructed from surface waters at the SPOT, (2) ARISA fragment sizes were determined from marine cyanobacteria by *in-silico* PCR of available genomes, (3) clone libraries (16S-ITS-23S) constructed from environmental ARISA from various global ocean locations: (a) Pacific Ocean, (b) Atlantic Ocean and (c) Indian Ocean, (4) Global-Ocean Survey open ocean rRNA-containing scaffolds (Yooseph *et al.*, 2007) and (5) whole-genomes from marine isolates were queried by *in-silico* PCR using a custom Python script and EMBOSS primersearch (Rice *et al.,* 2000) with a 14% mismatch. The primary source of identification, as noted, were surface water clones from SPOT, which includes 184 clones from four dates: October 2000, April 2001, August 2001 and December 2001. Putative IDs were determined by searching within the taxonomy/ITS-length sources in order of priority; thus, if a putative ID was determined from the SPOT clones, the identification process was halted. If there were multiple IDs associated with one fragment length within the same level of priority, the most common ID was used. If there were two equally common IDs, then the length is identified as both. If there is no ID associated with the fragment length, then it is identified only as bacteria. 16S sequences were classified via the Greengenes (McDonald *et al.*, 2012), Ribosomal Database Project (RDP) (Cole *et al.*, 2009) and SILVA 108 truncated SSU database (Pruesse *et al.,* 2007) using BLAST (word size 7). Full taxonomy and associated 16s sequence (where applicable) of reported bacterial OTUs are available in Supplementary Table S1. The top hit (by e-value with a minimum alignment length of 200 bp and minimum percent identity of 97%) was classified with consistent hierarchical taxonomy by Greengenes and the Silva identifier for the most discriminating taxonomy given before 'uncultured' or 'unidentified.' SAR11 clade designations came from Ribosomal Database Project release 10. To identify bacteria from ARISA fragment lengths, we prefer empirical measurements of cloned ARISA fragment lengths rather than simply counting base pair from sequencing, because electrophoretic migration rates vary with factors like GC content. 95 clones were amplified using ARISA-PCR and run on the same instrument (ABI 377) as environmental samples to determine empirical (machine-estimated) ARISA lengths. For fragments $<800\,bp$ the empirical length corresponds well with sequence length; however,

for longer fragments (to 1200 bp) the empirically-determined lengths tended to underestimate sequence-length. Therefore, for clones or *in-silico* PCR where the ARISA product size was not explicitly determined, an adjusted length was used to make results directly comparable. Specifically, by plotting the variation in empirical-length and sequence length vs sequence length of the 95 ARISA amplified clones, we determined that the best fit to these data was by application of linear regressions of regions 400–800 bp and 800–1200 bp. To calculate what the apparent ARISA length would have been for the clones we did not run on the instrument, lengths 400–800 and 800–1200 bp were estimated by the equations $y = 2.428 - 0.00227x$ and $y = 29.46 - 0.03683x$, respectively, where y is the calculated apparent length and x is ARISA sequence length. The cloning and ARISA PCR for these steps can be found in Brown *et al.*, (2005).

Statistical analyses

Local similarity analysis. To observe correlations between OTUs and environmental parameters over the first 38 days, we used local similarity analysis (Ruan *et al.*, 2006b). Only OTUs observed at a frequency greater than 10 times were analyzed. A maximum time-lag of five steps (days) was allowed. *P*- and *Q*-values were determined by 1000 random permutation tests. Missing values were linearly interpolated. The local similarity analysis output data were input into visualization software, Cytoscape (Shannon *et al.*, 2003), to visualize association networks of microbial taxa. Only correlations with *P*-value and *Q*-values less than or equal to 0.005 and 0.05 were examined, respectively. Modules were detected by AllegroMCODE with the following settings: Degree cutoff (2), Node Score Cutoff (0.4), K-Core (2), Max. Depth (100) and the 'haircut' option was selected (Bader and Hogue, 2003).

Community similarity metrics. Group average Bray-Curtis similarities of microbial communities were calculated in Primer E- 6 as unweighted entities. Time-dependent community similarity 'decay' values were calculated by plotting similarity values between all pairwise samples separated by 1–77 days and fitting linear and logarithmic regressions in Sigma-Plot Version 11.0 (San Jose, CA, USA).

Results

Environmental parameters
Over the first 40 days of the time-series, surface temperature varied between 16.4 and 19.7 C and peaked on day 14 (Figure 1). Bacterial and viral counts varied less than fourfold reaching maxima on day 15 and 13 (Figure 1), respectively, and were positively correlated from days 14–25 ($r = 0.797$, $P < 0.001$;
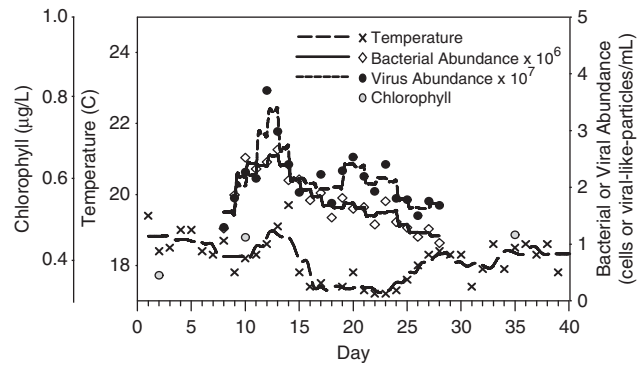
Figure 1 Environmental parameters over the first 38 days of the time-series. Moving averages for environmental parameters were calculated by smoothing with 0.1 sampling proportion and polynomial degree 1, roughly corresponding to 3-day averages.

$LS = 0.43$, $P < 0.05$). Eight-day averages for satellite chlorophyll varied between 0.3 and 0.45 $\mu g\,l^{-1}$.

Bacterial and T4-like-myovirus community overview
Over the 78 day time-series 286 bacterial and 153 T4-like myoviral OTUs were detected at least twice with an average richness of 123.7 ± 14.6 and 60.3 ± 14.2 OTUs each day. The majority (79.5 and 80% of amplified DNA) of the total bacterial and T4-like myovirus communities consisted of 'common' (detected $> 90\%$ of days) OTUs (Figures 2a and c). However, the majority of the bacteria and T4-like myoviral OTUs observed were rare (detected $< 25\%$ of days) and totaled 16.8% and 11.6% of each community (abundance), respectively (Figures 2b and d).

Individual bacterial/viral OTU fingerprint dynamics
As observed by ARISA, the relative abundances of the top 10 (highest average relative abundance) individual bacterial taxa showed different patterns, including gentle variation around stable mean abundance, 'boom-bust' dynamics and monotonic increases/decreases (Figure 3a). For example, the most abundant bacterial OTU (by average contribution over the 78 days), a putative *Synechococcus* (Bacterial OTU 1056), increased from its lowest detected relative abundance (2%) on day 1–20% by day 10, and then oscillated about an average of 18.2% for the remainder of the time-series. The second most abundant OTU, a member of the SAR11 clade (Bacterial OTU 666.4), was at its highest relative abundance on day 1 and oscillated about a mean relative abundance of 5% after day 7. In contrast to both, the third most abundant OTU, an Actinobacterium (Bacterial OTU 435; Acidimicrobiales, clade OCS155) three times showed steady, greater than sixfold increases in relative abundance over the course of about 10 days followed by reductions of equal proportion over 4 days (Figure 3a). Other common bacteria displayed variation about their mean, but did not show obvious net increases or decreases.
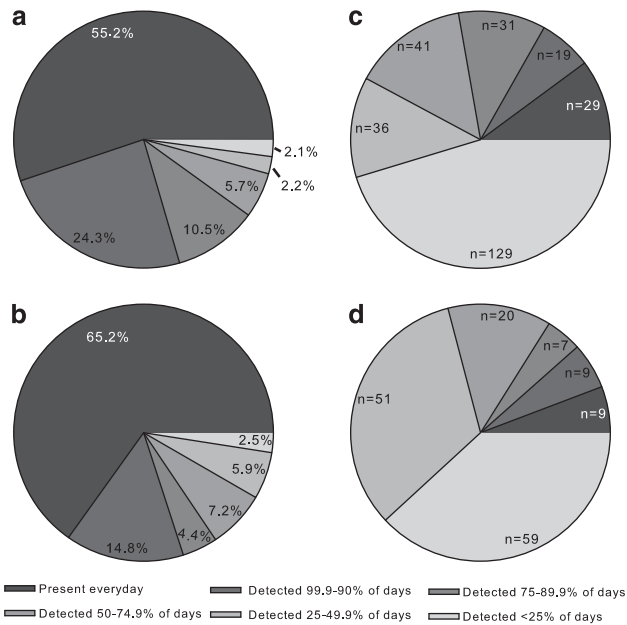
Figure 2 Contribution of the most common to the most rare taxa to total community composition and total OTU count. Percent of the total summed (**a**) bacterial and (**b**) myoviral community (peak area/total area) made up of OTUs which were detected on the indicated percentage of days of the 78 day time-series. Number of (**c**) bacterial and (**d**) myoviral OTUs, which were detected greater than twice for the indicated percentage of days of the 78-day time-series.

By g23-TRFLP, the four most abundant T4-like myovirus OTUs were at a maximum abundance on day 1 and showed steady decreases in relative abundance throughout the time-series (Figure 3b) of ~0.01% per day. Notably, viral OTU 296 showed no upward or downward trend overall, but twice showed steady increases followed by steady decreases of >1% per day. In contrast, other dominant T4-like myoviruses maintained stable abundances throughout the time-series (Figure 3b).

Observations from association network analysis
We found tens to hundreds of significant correlations ($P \leqslant 0.005$, $Q < 0.05$) between taxa depending on the correlation strength considered (Figure 4). The majority of statistically significant positive correlations between OTUs were intra-bacterial or intra-viral associations with no time-lag (95.2%, 82.7%, 78.1% and 69.4% for minimum LS values of 0.5, 0.45, 0.4 and 0.35, respectively) rather than T4-like-myoviral-to-bacterial or intra-group with time-lag (Figures 4a and b). In contrast to the majority intra-group correlations that had no time-lag, the bacteria-to-virus correlations had median time lags of 1.5, 2, 2 and 2 for LS of 0.5, 0.45, 0.4 and 0.35, respectively (Figure 4c).

Highly interconnected groupings of T4-like-myoviruses (Figure 5, Module 1) and bacteria (Figure 6) had many connections outside of their core members. For example, two members of the virus module

correlated with a member of the SAR86 clade (OTU 402.4). In turn, this SAR86 OTU was positively associated with four bacterial OTUs and negatively associated with one T4-like-myovirus (384.3). The bacterial module (Figure 6) was largely composed of positively associated members of the SAR11 clade; one SAR11 OTU was positively associated with the most abundant T4-like-myovirus (OTU 382.3) with a 3-day delay. In addition to those associations shown, there were other significant correlations between OTUs, which were only associated for a period of time shorter than the full 38 days or at correlations of less strength. For example, Actino-bacterium (OTU 435; Acidimicrobiales, clade OCS155) and T4-like myovirus 296 were correlated (LS = 0.249, $P = 0.001$) from days 10 to 31 with a 1-day time-lag (not shown, see Figure 3 for relative abundances).

*Microbial community similarity*
Bacterial and T4-like myoviral Bray-Curtis similarity was greater between communities with a closer temporal proximity. Adjacent days were, on average, 74.7% ± 1.7 (± s.e.m.) and 81.2% ± 1.5 similar, respectively (Figure 7). Notably, true duplicates, collected on day 20 were 84% and 86% similar by Bray-Curtis similarity for the ARISA and g23 TRFLP while technical replicates (same PCR reaction, measured more than once) during the time of study averaged 84% ± 1.1 ($n = 7$) for ARISA and 84% ± 1.3($n = 14$) for g23-TRFLP. Communities of 30 and 60 days apart were 62.9% ± 3.3 and 55.4% ± 2 similar for bacteria and 59.9% ± 2.5 and 53% ± 2.5 similar for T4-like myoviruses. Pairwise comparison of all bacterial and T4-like myoviral communities, plotted against the time lag between them, resulted in a linear decrease in un-weighted similarity of 0.18% (r = 0.296, $P < 0.001$) and 0.4% (r = 0.720, $P < 0.001$) per day, respectively (Figure 7). The linear rate of bacterial community differentiation over the first week (2.1%/day) was more rapid than the change over the remainder of the time-series. Thus, the fit of a logarithmic function may be more appropriate and has a higher correlation coefficient for bacteria (0.357, $P < 0.001$).

Discussion

The results from this study suggest that, over days-to-weeks, microbial communities exhibit characteristics of subtle-to-dramatic short-term variation and resilience. Individual bacterial and T4-like myoviral OTUs of the surface ocean vary in their relative abundance in connected ways over the extended daily time-series. Over weeks to months the total community changed, on average, relatively little. This suggests that underlying the ephemeral, connected variation of individual taxa, there is a relatively stable core, resilient community about which the variation
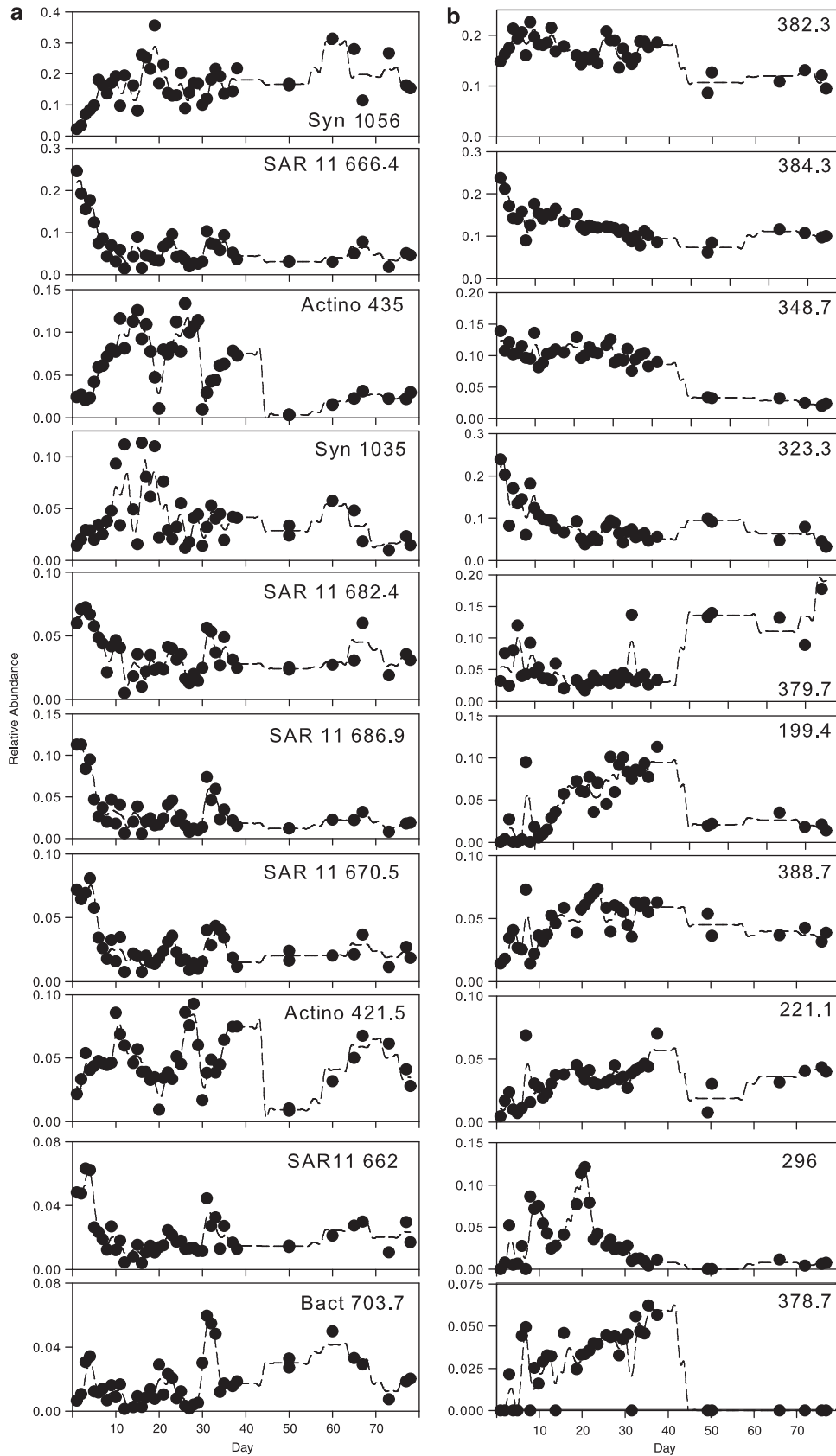
Figure 3 Normalized relative abundances of distinct (**a**) bacterial and (**b**) myoviral OTUs over the 78-day time series. The plots are organized from top to bottom from most abundant (on average) to 10th most abundant.
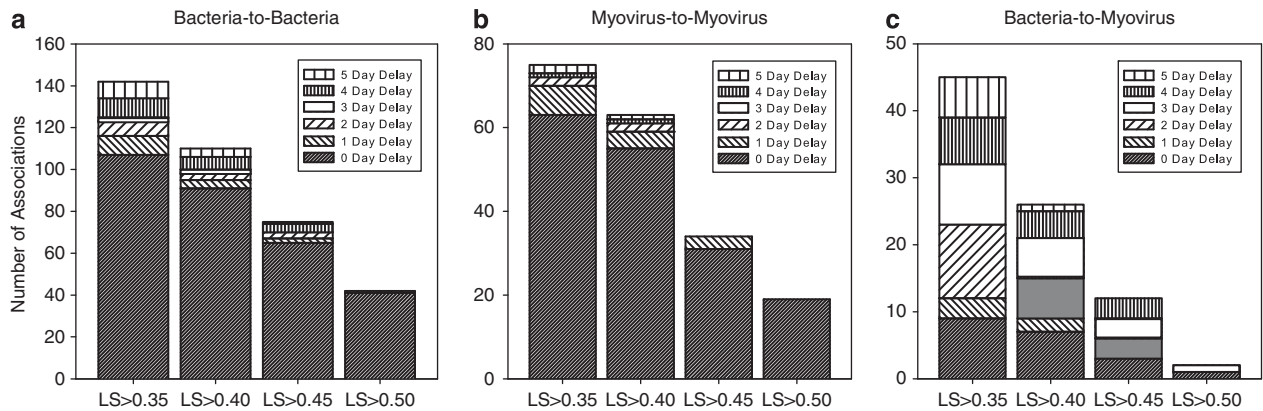
Figure 4 Statistical positive associations between bacteria and T4-like myoviruses as determined by local similarity analysis. Only data for the first 38 days were used in this analysis and only those associations over the entire 38 days are displayed. (**a–c**) Stacked bar graphs indicate the total number of positive associations ($P \leqslant 0.005$, $q < 0.05$) between (**a**) Bacteria-to-bacteria, (**b**) myovirus-to-myovirus and (**c**) bacteria-to-myovirus. Each column is broken down into segments indicating the number of associations corresponding to delays of 0–5 days. Statistical associations are organized according to their corresponding local similarity scores (LS), a measure of correlation strength similar to Pearson's r, and corresponding time-delay of the optimal association as determined by local similarity analysis. Like Pearson's correlation coefficient, higher LS indicate a stronger association and is thus more discriminating. Each column is cumulative: for example, correlations > 0.5 are also included within LS > 0.4 column.
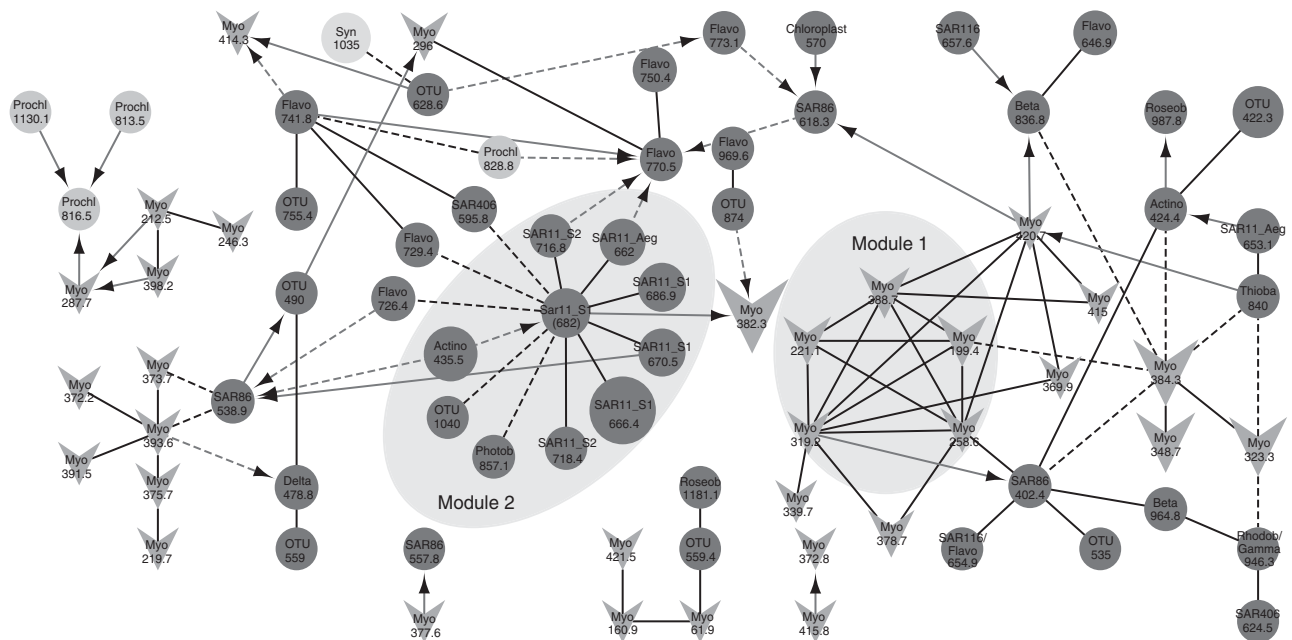
Figure 5 Network visualization of all statistical associations to OTUs associated with T4-like Myoviruses ($|LS| > 0.45$, $P \leqslant 0.005$, $q < 0.05$) over the full 38 days are shown. The network also displays correlations one-step beyond a connection to a T4-like Myovirus (LS > 0.45, $P \leqslant 0.005$, $q < 0.05$), but correlations (lines) between these taxa are not displayed for clarity. Circles indicate bacteria taxa; myoviruses by gray V-shapes. Non-cyanobacterial bacterial OTUs are dark gray, *Prochlorococcus* OTUs light gray and a *Synechoccocus* OTU light gray. Abbreviated, putative identities are followed by the associated fragment size. The size of each node corresponds to the OTU's average relative abundance over the time-series. Solid and dashed lines indicate positive and negative associations between taxa, respectively. The darkness of the line indicates the length of delay of correlations: black lines are non-lagged, whereas lighter shades are delayed correlations. In delayed associations, arrows point at the OTU that was delayed, or trailed, in the association. Shaded gray areas indicate OTUs that made up highly-interconnected regions (or modules) of the network. A full colour version of this figure is available at the *International Society for Microbial Ecology* journal online.

occurs. A resilient community likely results in a similar retention of function. The discussion first focuses on variation and later on stability.

Spatial and error considerations
Sampling from one geographic location, likely resulted in sampling of multiple regional water masses, which mixed and advected past our sampling location. Although most of this discussion is framed within the temporal context of the study, space and time are confounding variables. The observation that variation between communities from day-to-day can be as great as that observed a month apart may be from sampling different water masses. However, when averaging the points
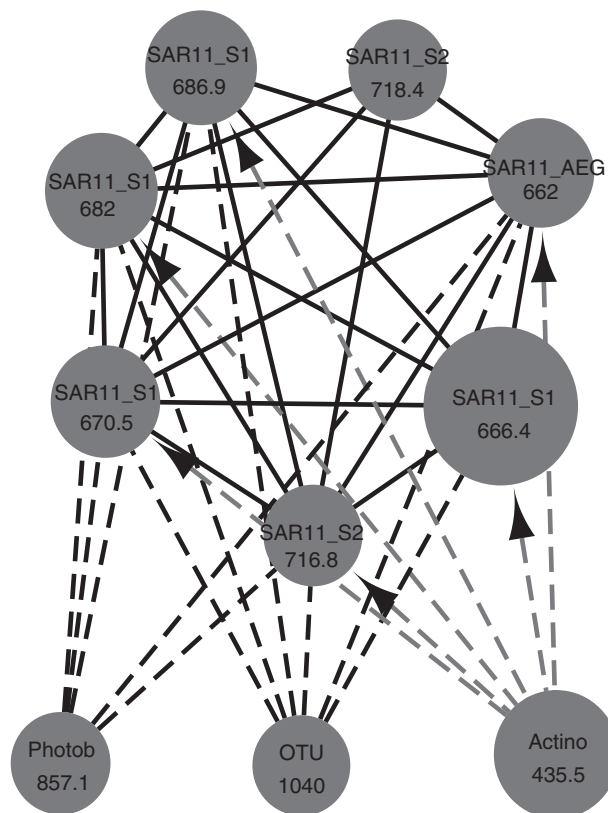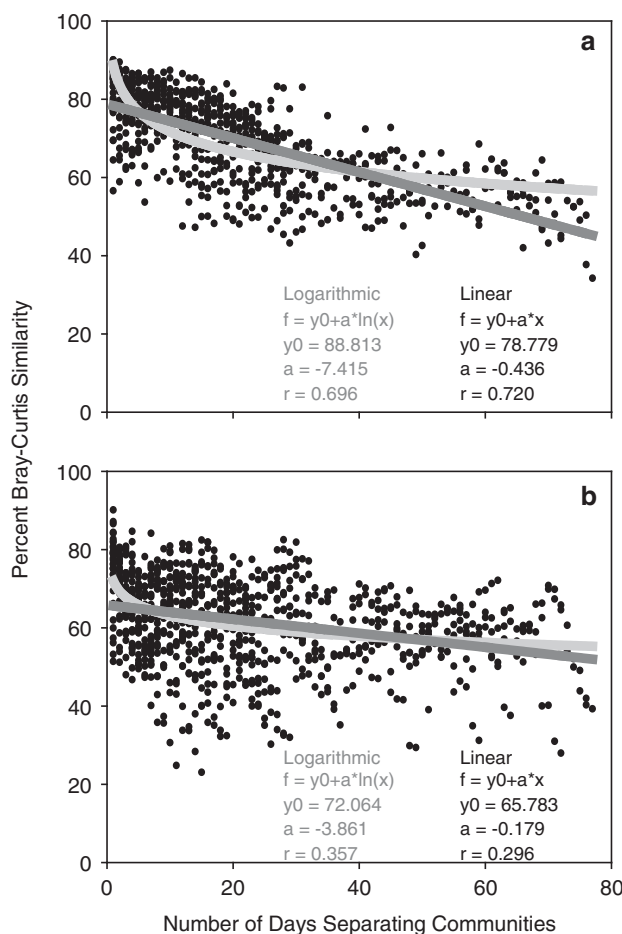
Figure 6 Network visualization of the largest highly-interconnected nodes exceeding the threshold of modularity (see Methods). All positive and negative correlations ($|LS| > 0.45$, $P \leqslant 0.005$, q < 0.05) are shown between members of the module. The network shows seven SAR11 OTUs that are positively correlated with no time-lag and three OTUs, which are negatively correlated to the SAR11 OTUs, one, an Actinobacterial OTU with a 1-day time-lag.

Figure 7 Community similarity time-dependent decay. Community composition was assessed by group average Bray–Curtis similarity for all possible pairwise comparisons between days for (**a**) T4-like myovirus and (**b**) bacteria. Both a linear and a logarithmic function are shown ($P < 0.001$).

together there is a robust pattern and the slow variation observed at the community level over the short-term is similar to that which has shown community coherence over short spatial scales (Hewson *et al.*, 2006; Steele, 2010). Further, co-occurrence patterns may be expected to be consistent over space or time. That we find robust associations over the time-series suggests that this may be the case.

Due to the fact that this study is based on relatively small variation between samples, it is important to consider the implication of errors in both the ARISA and g23-TRFLP methods. As noted, both technical (same PCR reaction, run more than once, in duplicate) and true replicates average about 85% by un-weighted Bray–Curtis similarity. Therefore, most differences between true replicates stem from accumulation of slight errors in the fragment analysis and binning procedures. Our ARISA reproducibility is close to those reported previously (Hewson *et al.*, 2006). Calculated most easily via Bray–Curtis *Dissimilarity* (1-similarity), our results suggests that about 60% of the apparent day-to-day change in bacterial community similarity (average

~25% Bray–Curtis Dissimilarity) is due to noise in the measurements (average 15% Bray–Curtis Dissimilarity between replicates), and for viruses about 84% of the apparent day-to-day change is due to noise (16% and 19% Bray–Curtis Dissimilarity for replicates and adjacent days, respectively). Regardless, this source of error should not influence the rates of change over time, determined by the slopes of regressions between all samples with respect to number of days separating them.

Short-term microbial variation
As assessed by community fingerprints, the dominant 10 bacterial and T4-like myoviral OTUs often varied in observable patterns over the course of the time-series, including subtle and dramatic oscillation, monotonic increases and decreases (Figure 2) over the 78-day time-series. Relatively dramatic oscillations were observed in particular taxa, for example, an Actinobacterium (OTU 435; Acidimicrobiales, clade OCS155); Actinobacteria, in general, are heterotrophic and prolific producers of

secondary compounds (Ward and Bora, 2006). Such rapid fluctuation (either temporally or spatially) may be an important to ecosystems as the functional redundancy within the bacterial community is not that well understood.

Likewise, particular T4-like myovirus OTUs were found to exhibit dramatic oscillation, for example T4-like myovirus OTU 296. This T4-like myovirus OTU twice increased in relative abundance by about 2% per day over 5 days. We do not make an attempt to extrapolate to the absolute number of viruses produced by such variation, but future investigations should attempt to estimate the ecosystem implications, for example, the number of bacterial cells lysed. A better understanding of the spatial extent of such oscillations, T4-like myovirus community abundance and decay rates would be useful to determine such production values. Our results illustrate that estimation of the productivity of individual viral OTUs over the short-term may be informative.

Short-term microbial connectivity and modularity
The short-term variation of individual OTUs produced statistical association between taxa, including highly interconnected groups of bacteria and potential host-phage interactions. Local similarity analysis, the method we employed to detect covariation of variables and/or OTUs, can identify correlations with time-delays, like those observed between predators and prey. Among the many significant positive and negative associations between many OTUs (Figures 5 and 6), the bacterial module of inter-connected bacteria (Figure 5) identified several OTUs from the SAR11 clade that were correlated with no time-lag. Notably, some of the SAR11 OTUs from Module 1 are >98% similar at 16S level (Brown *et al.*, 2005), and it is possible they respond similarly to abiotic and biotic forcing at the daily time-scale, but interestingly many of these same OTUs were not correlated when previously examined monthly at SPOT over 3 years at the Chlorophyll maximum depth (Fuhrman and Steele, 2008; Steele *et al.*, 2011). Synchronous oscillations of these OTUs, even when only distantly related, could be due to a variety of non-exclusive causes, for example, (a) similar positive or negative response to environmental parameters such as varying inputs of particular substrates, (b) tight ecological coupling due to auxotrophy or cross-feeding, and (c) top-down pressures by selective protists or viruses. Many OTUs were negatively correlated to the bacterial module which might indicate, for example, inter-specific competitions for organic substrates or nutrient regime shifts that favor bacteria not associated with Module 1.

Similarly, intra-connected groups of viruses were associated with zero time-delay (Figure 5b). The ecological connectedness of individual viral OTUs could be due to periods, in which a lineage of bacteria became particularly susceptible to lytic infection by multiple phage types (Comeau *et al.,* 2006; Holmfeldt *et al.,* 2007). We hypothesize that these intra-correlated groups of viruses may include close relatives and infect ecologically similar hosts. Inter-correlated groups of viruses might instead result from environmentally influenced viral production, for example, from pseudo-lysogens (for example, Wilson *et al.,* 1996). In this study, interconnected viruses were loosely associated with two distinct, putative SAR86 OTUs (Figure 6), suggesting that this group (and likely many others) may experience pseudo-lysogeny in the environment.

Virus–bacteria associations. Most of the bacteria-to-myovirus associations observed in this study had a time lag. Time-lagged associations, in which peaks in abundance of prey and predators are temporally offset are described classically by the Lotka-Volterra equations or the non-steady-state KtW hypothesis. These predator-prey-like dynamics are often depicted with an arbitrary time-scale along the x-axis (Wommack and Colwell, 2000); the results of this study suggest that a reasonable time-scale for bacteria–virus dynamics can be short (days). However, most of the bacteria–virus correlations we observed were not indicative of a boom-bust, predator-prey-like dynamic. The correlations we found were mostly among taxa that varied in their relative abundance but remained relatively abundant. For example, T4-like-myovirus OTU 382 was significantly correlated to SAR11 (682.4) with a 3-day time-delay despite both remaining relatively abundant throughout the time-series. These results are in fact consistent with a modestly-changing system where the steady-state KtW dynamics (Thingstad 2000) may apply to a significant extent, even though the system is not truly in steady state.

Given that the growth rates of specific bacterial taxa influence the dynamics of their relationships with viruses (Middelboe, 2000), it should be expected that different hosts would have dissimilar temporal dynamics with their viruses in the environment. Our median time-lag between bacteria and virus correlations of 2 days was shorter than that recently proposed by Parsons *et al.,* (2011) that used a range of average ocean values for marine bacterial growth rates and virus decay rates. This implied that many host-phage relationships may be more closely coupled than suggested by bulk rates and due to bacteria which grow faster than the average community growth rate.

Short-term stability within microbial communities. Characteristics of community stability was equally notable. Generally, most common bacteria and T4-like-myovirus OTUs were persistent; bacterial and T4-like myoviral OTUs that were detected in at least 90% of the samples made up 80% of the cumulative communities (Figure 1). Thus, as previous research has shown (Hewson *et al.*, 2006; Steele, 2010), a microbial community from one location and

one time provides a fairly good estimate of the community of surrounding days at that location.

Although about 60% and 84% of the apparent day-to-day change in the bacterial and T4-like-myovirus community comes from analytical noise, the average day-to-day change in the bacterial and T4-like-myovirus communities was about 10% and 4% greater than between replicates, respectively, with some days that exhibited large changes and others that showed no significant change (Figure 7). The average rate of change of the bacterial community over 1 week was about 2% per day (calculated from the linear slope of days separated by 1–7 days, Figure 7) which, if continued for a month, would lead to substantially dissimilar communities of about 25% similarity (that is, subtraction of the rate 2% per day for 30 days, considering the initial similarity between replicates of 85%). However, we observed that this modest rate of change over a week was 12-fold higher than the rate of change within the bacterial community when averaged over 2 months (the linear slope of the regression of all combinations of days separated by 1–60 days is 0.18% per day). We interpret this decelerating rate of change in the bacterial community to suggest that, although there were significant fluctuations over days-to-weeks, ecological processes forced the communities towards a relatively steady and even predicable average composition for a given time period of weeks to months. This is consistent with the predictability of bacterial communities at the nearby SPOT station reported previously (Fuhrman et al., 2006) but from that study we would have had no expectation of the more rapid variation over the first week observed here. The virus community change is relatively more steady, 0.48%–0.37% per day over a week and 2 months, respectively. However, the rate of change we observe in the T4-like-myovirus community (0.4% per day) is about twice that which would be expected from the T4-like myovirus community similarity at 6 month intervals over 3 years based on data in Chow and Fuhrman, (2012). These two data sets were taken from two locations roughly 10 km apart and SPOT is likely more open-ocean-like. Taken together with the observed dynamics of individual taxa, the bacterial (and to a smaller extent the myovirus) communities decelerating rate of change is evidence that the community retained a resilient background composition with significant, ephemeral variation over the short-term, with a stable core bacterial community likely providing ecosystem resilience that was regulated by outside forces. A useful analogy might be the relatively 'noisy' or chaotic variations in weather, as compared with longer term and much more predictable seasonal variations in climate.

Recognizing sensitivity of methods. Higher taxonomic or genomic resolution would likely provide more information. For example, studies investigating the source of resistance and infectivity in picocyanobacteria and cyanophage reported that successful infection could be controlled by single genomic mutations: this level of interaction between host and phage helps to explain the ubiquity and high abundance of picocyanobacterial populations in the environment (Avrani et al., 2011; Marston et al., 2012). Further, isolated marine bacteria can vary in their susceptibility to viruses at a taxonomic resolution finer than that observed by ARISA (Holmfeldt et al., 2007) and metagenomic approaches have illustrated that a greater degree of temporal variation occurs at a finer-taxonomic resolution for both bacterial and viral communities (Rodriguez-Brito et al., 2010). Even with these issues and possible future directions, our study clearly demonstrated that, at daily resolution, variation occurred within the bacterial and virus communities using DNA fingerprinting techniques.

Summary, implications, and the KtW hypothesis. Given that marine bacterial communities turn over on the scale of days, are subject to top-down controls from protists and viruses, horizontally transfer genes, and can be functionally redundant, it might be expected that over days-to-months distinct bacterial taxa might fluctuate wildly or exhibit unpredictable, boom-and-bust cycles. From samples collected over three summer months including 38 consecutive days, we found that it was uncommon for bacterial and viral taxa to exhibit boom and bust-type dynamics. In contrast, at our taxonomic resolution, most dominant taxa retained their dominance over the period of study. Further, most common and abundant taxa persisted at stable abundances. Thus, the steady-state solutions in Thingstad's formal KtW hypothesis may apply to a significant extent. At the community level, the relative rate of change over days-to-weeks is greater than that over weeks-to-months showing that there was variation about a mean community over days that is smoothed over time, especially within the bacterial community. This suggested that there are barriers to fundamental change in the bacterial community composition over the short term, and/or strong external or community forcings promoting particular bacterial and viral types. Our results also indicated that bacterial and T4-like myoviral communities were highly interconnected over days-to-months potentially suggesting a steady, core community with relationships that sustained ecosystem function over the short-term. In particular, the observation of modularity within marine microbial communities over the short-term warrants further study to understand their implication on community structure and to understand how microbial communities respond to and influence ecosystem perturbations.

Acknowledgements

References

Acinas SG, Rodríguez-Valera F, Pedrós-Alió C. (1997). Spatial and temporal variation in marine bacterioplankton diversity as shown by RFLP fingerprinting of PCR amplified 16S rDNA. *FEMS Microbiol Ecol* **24**: 27–40.

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

Bader GD, Hogue CWV. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform* **27**: 1–27.

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K *et al.* (2007). Metagenomic characterization of Chesapeake Bay virioplankton. *Appl and Environ Microbiol* **73**: 7629–7641.

Bergh O, Borsheim KY, Bratbak G, Heldal M. (1989). High Abundance of viruses found in aquatic environments. *Nature* **340**: 467–468.

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 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 Microb* **7**: 1466–1479.

Brussaard CPD, Wilhelm SW, Thingstad TF, Weinbauer MG, Bratbak G, Heldal M *et al.* (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J* **2**: 575–578.

Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. (2011). Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci USA* **108**: 12776–12781.

Carlson CA, Morris RM, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. (2009). Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3**: 283–295.

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

Clokie MRJ, Millard AD, Mann NH. (2010). T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. *Virol J* **7**: 1–19.

Clokie MRJ, Millard AD, Mehta JY, Mann NH. (2006). Virus isolation studies suggest short-term variations in abundance in natural cyanophage populations of the Indian Ocean. *J Bio Mar Biol Assoc UK* **86**: 499.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Comeau AM, Chan AM, Suttle CA. (2006). Genetic richness of vibriophages isolated in a coastal environment. *Environ Microbiol* **8**: 1164–1176.

Comeau AM, Krisch HM. (2008). The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol Biol Evol* **25**: 1321–1332.

Eiler A, Hayakawa DH, Church MJ, Karl DM, Rappé MS. (2009). Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. *Environ Microbiol* **11**: 2291–2300.

Eiler A, Hayakawa DH, Rappé MS. (2011). Non-random assembly of bacterioplankton communities in the subtropical north pacific ocean. *Front Microbiol* **2**: 1–12.

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. (2009). Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Fuhrman JA, Comeau DE, Hagstrom A, Chan AM. (1988). Extraction from natural Planktonic Microorganisms of DNA Suitable for Molecular Biological Studies. *Appl Environ Microbiol* **54**: 1426–1429.

Fuhrman JA, Eppley RW, Hagstrom A, Azam F. (1985). Diel variations in bacterioplankton, phytoplankton, and related parameters in the Southern California Bight. *Mar Ecol Prog Ser* **27**: 9–20.

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.

Fuhrman JA, Steele JA. (2008). Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microbl Ecol* **53**: 69–81.

Fuhrman JA, Suttle CA. (1993). Viruses in marine planktonic systems. *Oceanography* **6**: 51–63.

Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B *et al.* (2011). Defining seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.

Giovannoni SJ, Vergin KL. (2012). Seasonality in ocean microbial communities. *Science* **335**: 671–676.

Hewson I, Steele JA, Capone DG, Fuhrman JA. (2006). Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Mar Ecol Prog Ser* **311**: 67–77.

Holmfeldt K, Middelboe M, Nybroe O, Riemann L. (2007). Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. *Appl Environ Microbiol* **73**: 6730–6739.

Marston MF, Pierciey FJJ, 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.

Matteson AR, Loar SN, Pickmere S, Debruyn JM, Ellwood MJ, Boyd PW *et al.* (2011). Production of viruses during a spring phytoplankton bloom in the South Pacific Ocean near of New Zealand. *FEMS Microb Ecol* **79**: 709–719.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Middelboe M. (2000). Bacterial growth rate and marine virus-host dynamics. *Microb Ecol* **40**: 114–124.

Noble RT, Fuhrman JA. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat Microb Ecol* **14**: 113–118.

Noble RT, Fuhrman JA. (2000). Rapid virus production and removal as measured with fluorescently labeled viruses as tracers. *Appl Environ Microbiol* **66**: 3790–3797.

Parsons R, Breitbart M, Lomas MW, Carlson CA. (2011). Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* **6**: 273–284.

Patel A, Noble RT, Steele JA, Schwalbach MS, Hewson I, Fuhrman JA. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols* **2**: 269–276.

Proctor LM, Fuhrman JA. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60–62.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Rice Peter, Longden Ian, Bleasby A. (2000). The european molecular biology open software suite emboss: the european molecular biology open software suite. *Trends Genet* **16**: 276–277.

Riemann L, Middelboe M. (2002). Stability of bacterial and viral community compositions in Danish coastal waters as depicted by DNA fingerprinting techniques. *Aquat Microb Ecol* **27**: 219–232.

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.

Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. (2006b). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22**: 2532–2538.

Ruan Q, Steele JA, Schwalbach MS, Fuhrman JA, Sun F. (2006a). A dynamic programming algorithm for binning microbial community profiles. *Bioinformatics* **22**: 1508–1514.

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.

Shiah F-K. (1999). Diel cycles of heterotrophic bacterioplankton abundance and production in the ocean surface waters. *Aquat Microb Ecol* **17**: 239–246.

Steele JA. (2010). Marine Bacterioplankton Biogeography over short-to-medium spatio-temporal scales. PhD Dissertation, University of Southern California, UMI Number 3403649.

Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* **5**: 1414–1425.

Steward GF, Culley AI. (2010). *Extraction and Purification of Nucleic Acids from Viruses* In:American Society of Limnology and Oceanography: Manual of Aquatic Virology, Waco, TX, USA, pp 154–165.

Sullivan MB, Waterbury JohnB, Watson StanleyW, Valois FredericaW, Franks DG, Chisholm SW. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature* **424**: 1047–1052.

Suttle CA, Chan AM. (1993). Marine cyanophages infecting oceanic and coastal strains of Synechococcus: abundance, morphology, cross-infectivity and growth characteristics. *Mar Ecol Prog Ser* **92**: 99–109.

Thingstad TF. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.

Thingstad TF, Lignell R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* **13**: 19–27.

Treusch AH, Demir-Hilton E, Vergin KL, Worden AZ, Carlson CA, Donatz MG et al. (2011). Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J* **6**: 481–492.

Ward AC, Bora N. (2006). Diversity and biogeography of marine actinobacteria. *Curr Opin Microbiol* **9**: 279–286.

Waterbury JB, Valois FW. (1993). Resistance to co-occurring phages enables marine synechococcus communities to coexist with cyanophages abundant in seawater. *Appl Environ Microbiol* **59**: 3393–3399.

Wichels A, Biel SS, Gelderblom HR, Brinkhoff T, Muyzer G, Schütt C. (1998). Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* **64**: 4128–4133.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI et al. (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS One* **3**: e1456.

Wilson WH, Carr NG, Mann HN. (1996). The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium Synechococcus sp. WH7803. *J Phycol* **32**: 506–516.

Wilson WH, Joint IR, Carr NG, Mann NH. (1993). Isolation and molecular characterization of five marine cyanophages propagated on Synechococcus sp. strain WH7803. *Appl Environ Microbiol* **59**: 3736–3743.

Winget DM, Wommack KE. (2009). Diel and daily fluctuations in virioplankton production in coastal ecosystems. *Environ Microbiol* **11**: 2904–2914.

Wommack KE, Colwell RR. (2000). Virioplankton: viruses in aquatic ecosystems. *Microb Molec Biol Rev* **64**: 69–114.

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 biology* **5**: e16.

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