nature microbiology

LETTERS

https://doi.org/10.1038/s41564-019-0628-x

Long-term stability and Red Queen-like strain dynamics in marine viruses

J. Cesar Ignacio-Espinoza[1], Nathan A. Ahlgren[1,2] and Jed A. Fuhrman[1*]

Viruses that infect microorganisms dominate marine microbial communities numerically, with impacts ranging from host evolution to global biogeochemical cycles[1,2]. However, virus community dynamics, necessary for conceptual and mechanistic model development, remains difficult to assess. Here, we describe the long-term stability of a viral community by analysing the metagenomes of near-surface 0.02–0.2 µm samples from the San Pedro Ocean Time-series[3] that were sampled monthly over 5 years. Of 19,907 assembled viral contigs (>5 kb, mean 15 kb), 97% were found in each sample (by >98% ID metagenomic read recruitment) to have relative abundances that ranged over seven orders of magnitude, with limited temporal reordering of rank abundances along with little change in richness. Seasonal variations in viral community composition were superimposed on the overall stability; maximum community similarity occurred at 12-month intervals. Despite the stability of viral genotypic clusters that had 98% sequence identity, viral sequences showed transient variations in single-nucleotide polymorphisms (SNPs) and constant turnover of minor population variants, each rising and falling over a few months, reminiscent of Red Queen dynamics[4]. The rise and fall of variants within populations, interpreted through the perspective of known virus–host interactions[5], is consistent with the hypothesis that fluctuating selection acts on a microdiverse cloud of strains, and this succession is associated with ever-shifting virus–host defences and counterdefences. This results in long-term virus–host coexistence that is facilitated by perpetually changing minor variants.

Naturally occurring microbial and viral communities are characterized by high levels of genomic, genetic, metabolic and phenotypic diversity[1,6]. In the oceans, viruses are numerically dominant, and the vast majority of these viruses infect prokaryotes and protists[1,2,6]. Viruses are central to marine food webs and represent one of the largest reservoirs of genetic novelty[2,6]. The mechanisms that generate and maintain biological diversity are fundamental topics of ecological inquiry; in complex communities, species–species interactions are thought to play a central role[5]. Virus–host interactions are one such mechanism, and marine viruses are thought to have a considerable influence on microbial diversity and food-web processes[2]. Although recent research from ocean transects has revealed much about global-scale distributions and the metagenomic diversity of marine viruses[1], time-series studies are needed to understand the stability, resilience, seasonality and long-term changes of such communities. Viral dynamics can also inform models of host dynamics.

To investigate the long-term dynamics of marine viruses, we examined the double-stranded-DNA viral community by sequencing total genomic DNA from the 0.02–0.2 µm viral size fraction, which was collected monthly between June 2009 and September 2014 ($n=53$) at the San Pedro Ocean Time-series (SPOT) off the coast of California[3,7]. Metagenome assembly (cross-assembled from individual sample assemblies) generated 99,722 (median = 12.5 kb) non-redundant contigs larger than 5 kb, of which 19,907 were classified bioinformatically as viral with high confidence (using VirSorter and VirFinder). We studied the 19,907 putative viral contigs to investigate their long-term dynamics. The abundances of each viral contig were estimated by competitive recruitment, at 98% sequence identity, of all metagenomic reads to all contigs, normalized to sequencing depth and contig length[1]. Although some contigs represent different portions of one genome, the 15 kb average contig size comprises a substantial portion of a typical viral genome, and our interpretations are unaffected by multiple contigs that represent one genome. We refer to these contigs operationally as populations (see Methods), in accordance with usage in recent work in marine viral ecology[8,9] and recognizing that they are a consensus of very close relatives.

Three main outcomes emerged: (1) the taxonomic composition (presence) hardly changed throughout the 5 years of sampling. Viral communities at SPOT were remarkably stable, with no notable gains or losses in populations. Any given metagenome (by read recruitment at 98% identity) contained up to 97% of the cumulative viral contig richness (Fig. 1a), and even increasing the threshold for presence by 10- to 100-fold hardly changed the pattern (Supplementary Discussion). Furthermore, 95% of contigs were always detected, and 89% of contigs were always detected even after increasing the number of hits required to record presence by 10-fold. This pattern of persistence contrasts sharply with the considerable amount of spatial variation over the global ocean[1]. Similar multi-year stability has been reported for the North Pacific Subtropical Gyre, which the authors pointed out is oceanographically very stable over time[10], much more so than SPOT[3]. Many of the detected phages putatively infect the most abundant marine bacteria, cyanobacteria and *Pelagibacter*, with some rare phages that are nearly identical to sequenced isolates[8,11] (Extended Data Fig. 1, Supplementary Tables). (2) Relative abundances were generally stable, that is, there was limited rank reordering ($z=-8.9$; $P<0.00001$). A bird's-eye view of community composition revealed striking stability, that is, most viral contigs had similar ranks on all of the sampled dates (Fig. 1b). Cumulative abundance of each contig varied proportionally to its median abundance, showing low variation over time, particularly for abundant contigs (Fig. 1c). Note that Fig. 1b,c uses log scales, and individual contig abundances vary many-fold (Extended Data Fig. 2; linear graphs better visualize smaller variations), but our point is that the abundances of contigs generally stay within a similar percentile range. Rank reordering that did occur was dominated by medium- and low-rank contigs (Fig.1c–e), such that some of the

[1]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA. [2]Present address: Department of Biology, Clark University, Worcester, MA, USA. *e-mail: fuhrman@usc.edu

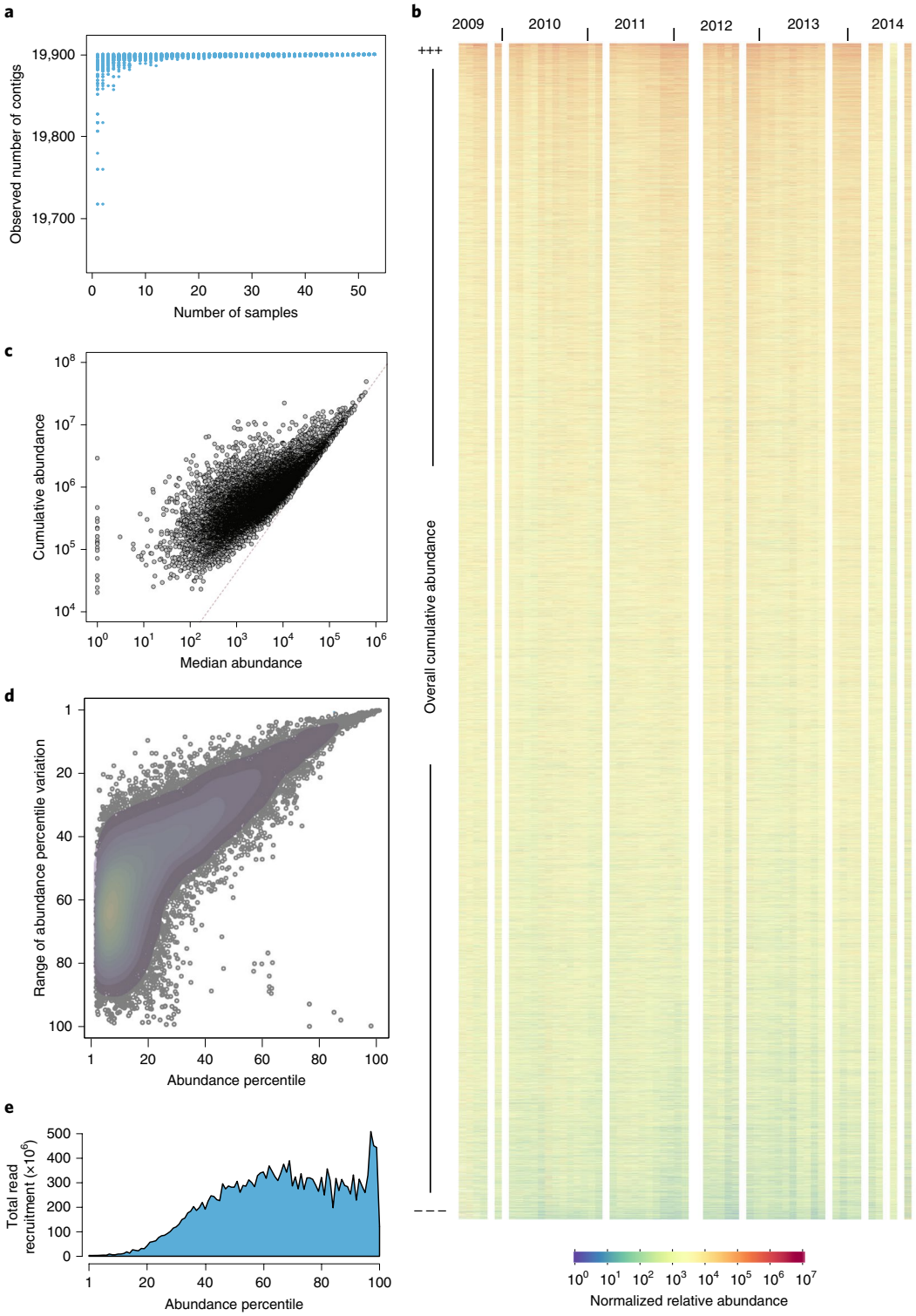


Fig. 1 | Persistence and stability of community patterns. a, An accumulation curve showing the presence of viral contigs over 53 monthly surface ocean viral (0.02–0.2 μm) metagenomes, showing that almost all contigs were present almost all of the time; a similar plot in which the threshold for presence is ramped up 10,000-fold is provided in the Supplementary Discussion. **b**, A heatmap showing relative abundance, determined by competitive metagenomic read recruitment, of the 19,907 viral contigs (note 7-decade log scale), one contig per row, during monthly sampling. Contigs are ordered by average abundance over all months, with the highest at the top. The white columns represent months with missing data (all data are provided in the Supplementary Tables). Note that the colours remain similar through time for the large majority of contigs. Versions of this heatmap with data plotted linearly are provided in Extended Data Fig 2. **c**, Median versus cumulative abundance of contigs across all of the samples, indicating that most contigs retained similar ranks across the 5 years. **d**, Viral contig population abundances (as in **b**) were converted to percentiles and plotted as mean versus the range for all of the samples. Note how the more abundant contigs (top right) changed very little in percentile. **e**, Collective-abundance histogram of contigs within each individual percentile (not a cumulative sum across) plotted against the same percentile axis scale as in **d**. Note the substantial contribution of the middle ranks to the total abundance, even though each is orders of magnitude less abundant than the higher ranked contigs as shown in **b**.

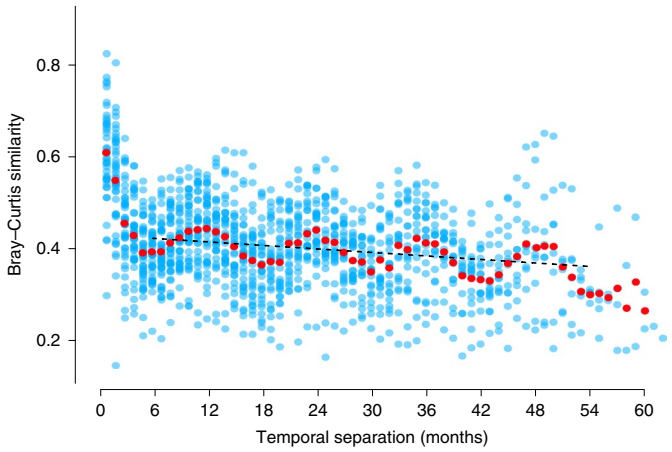


Fig. 2 | Recurrent seasonality superimposed on a stable average viral community. Normalized abundances of viral contigs (n = 19,907) were calculated from competitive per-sample read recruitment. The Bray–Curtis community similarity index (blue dots) was calculated among all of the possible sample pairs (n = 1,378 combinations) and plotted as a function of the number of months separating their sampling. The red dots correspond to the mean of all of the observations at a given separation. Stability is shown by the slightly declining average similarity over time, after accounting for the strongest similarity of samples that were separated by one or a few months and excluding long gaps with few data points (the dashed line is a linear regression of the data spanning 6 to 54 months gaps). Seasonality is shown by sine-wave-like similarity peaks at intervals of 12, 24, 36, 48 and 60 months (same calendar months in different years) in contrast to dips in opposite months (at intervals of 6, 18, 30, 42 and 54 months). Note that the spread of the data shows that any individual sample can deviate considerably from the mean.

viruses had widely varying abundances (Fig. 1b–d, Extended Data Fig. 2, Supplementary Tables); these could include viruses responding to stochastic queues, or responding to more predictable, but episodic, events such as phytoplankton blooms. (3) Moderate- and low-abundance members of the community produced a seasonal pattern that was superimposed over a general stability. The Bray–Curtis similarities of the virus community for all of the pairwise combinations plotted against the temporal gap between samples showed clear seasonal patterns, with peaks of maximum average similarity at intervals of around 12, 24, 36 and 48 months, representing the same season, and local minimum average similarity at intervals of 6, 18, 30 and 42 months, representing opposite seasons (Fig. 2). This pattern for the entire viral community is similar to data that we previously reported for free-living bacteria[3] and T4-like phages[12], indicated by marker genes. Although the sinusoid-like pattern is striking, the overall long-term average similarity of the entire community was steady, declining only slightly over time (Fig. 2, dashed line). If there had been substantial migration, extinction or changes in relative proportions, the average similarity would decline considerably with increasing time lags; however, similarity declined only slightly. Even though any given pair of samples was about 40% similar, we interpret the steady average similarity over medium-to-long time lags as suggesting that the entire viral community fluctuated mathematically around the same average community throughout the entire study period. However, we recognize that data from few, if any, sampling dates showed close to the average composition. Dividing the community into different abundance ranges showed reduced seasonality in the most abundant contigs, with most of the seasonality in the middle and tail of the rank-abundance distribution (Extended Data Fig. 3).

We observed what seems to be a largely resilient, and also regulated (such as seasonal), viral community comprising thousands of members. However, such long-term coexistence of viruses and their putative hosts is challenged by some theoretical expectations[13] and laboratory-based model-systems[14,15], in which limited improvement of infection mechanisms[5] generally leads to viral extinction. Although we know that viruses must coexist with hosts—and models[16,17] can predict stable coexistence under certain conditions[18,19]—the underlying mechanisms for coexistence in a large semi-open system such as our time-series site are not immediately clear. Furthermore, our long-term stability contrasts with the transient nature of viral populations over 12 months in an Irish lake[20], but it is reminiscent of shorter-term observations of stability in enclosed human-controlled aquatic environments[21].

We investigated the mechanisms that might explain this multi-year stable coexistence by leveraging our metagenomic dataset to track temporal changes in underlying diversity. We called single-nucleotide polymorphisms (SNPs)[22] from all contigs and focused on those with more than 10× coverage (n = 3.2 million SNPs distributed within 4,002 contigs), recovered by at least 4 reads and present at a minimum frequency of 1% (ref. [22]). First, we observed that contig intrapopulation genetic diversity (quantified as SNP density) increased as a function of overall population abundance (Fig. 3a). Second, smaller contigs tended to have greater intra-contig variability (Fig. 3b)—consistent with the idea that population heterogeneity hinders long metagenomic assemblies of common and abundant types—of which high coverage and associated microvariation break the De Brujn assembly paths[23]; our observations suggest that this is common.

Although we know that environmental populations have much more genetic heterogeneity than laboratory clonal strains[8,11,24], few studies have tracked natural microvariation over years[22,25]. Our dataset reveals that a large majority of contig populations maintained a dynamic cloud[8] of genetic microdiversity throughout our 5-year study (Fig. 3c,d, Extended Data Fig 4). Thus, in contrast to reports of long-term maintenance of nearly clonal freshwater microbial populations[24], stable clonality was not observed. This observed microvariation could correspond to multiple closely related viral populations exploiting multiple niches or to a single coherent population with access to a common variable gene pool. Unfortunately, with metagenomics (read lengths of ~150 bp) we cannot link most single-nucleotide variants in a given genome, but we can determine SNP frequencies within populations. Regardless of the underlying genetic structure, we sought to compare different population variants (albeit from averages) at different time points. We did this by recovering the natural variation from all of the reads in a sample that mapped to a contig within 98% identity (well within the reported operational definition of marine-viral-species-like units[8,9]) and compared it with the variation from all of the other samples (see Methods). We observed that contig population similarity (the percentage of shared SNPs between members of the same contig population on different dates) decreases sharply as a function of the time lag between dates, irrespective of the reference point (Fig. 3c,d), with time scales of months. Therefore, although each viral contig population stayed within its own cloud of approximately 98% sequence identity[8,11], fluctuating about a fairly steady average abundance over 5 years, there was an ever-changing intrapopulation microdiversity. This pattern strongly resembles Red Queen dynamics, which are rapid changes of genotypes within a population from ecological and evolutionary mechanisms, and may include fluctuating Red Queen, in which fluctuating selection drives genotypic frequency oscillations[4].

The observed continual changes in virus genotypes are not consistent with dispersal (by ocean current advection) of variants as the sole cause, because individual SNP profiles persist for months (Fig. 3c,d), whereas the water in the San Pedro Basin has a physical

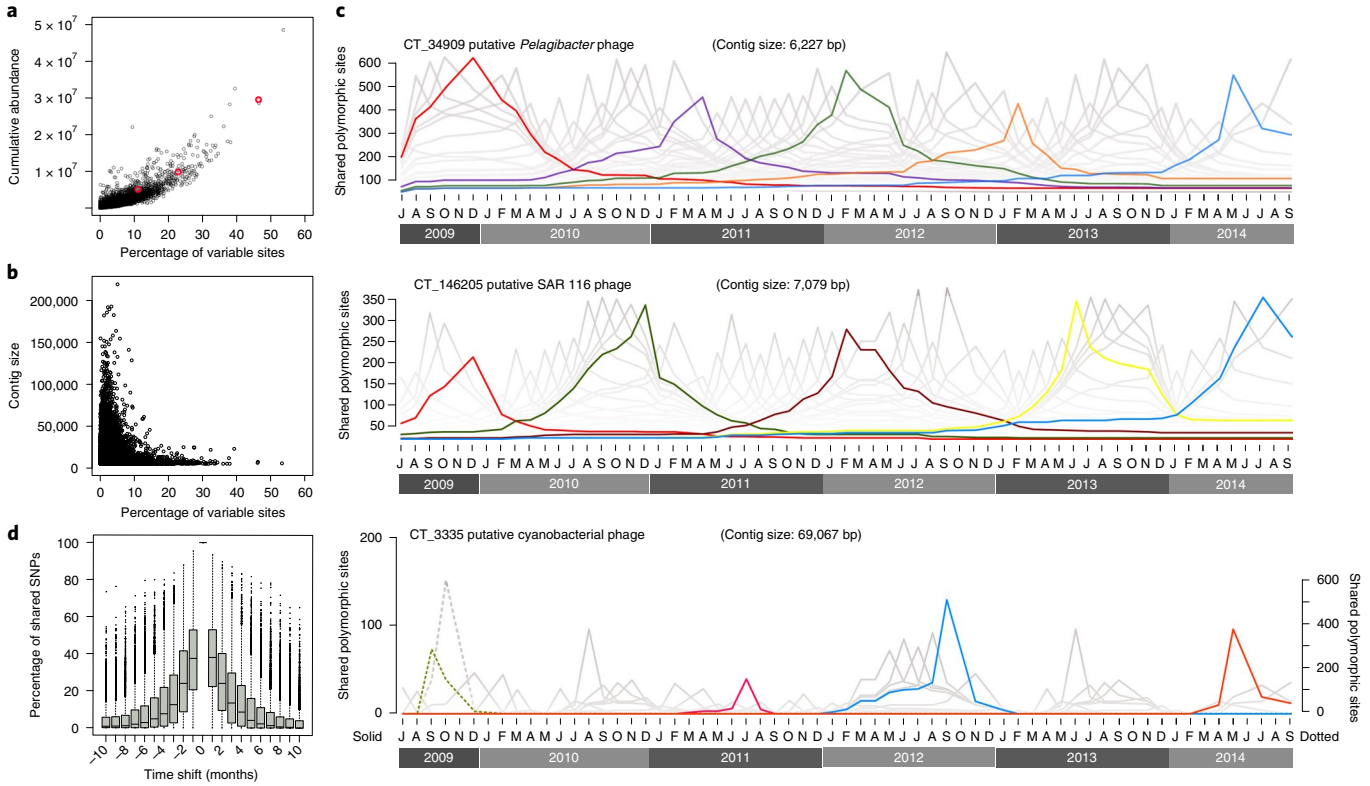


Fig. 3 | Comparisons of SNP profiles show a constant turnover of intrapopulation variants. a,**b**, SNP analysis shows that the average amount of variability per base in a given contig is positively correlated with the abundance of the contig (**a**) and only high within short contigs (**b**). **c**, Dynamics of SNP profiles show Red Queen-like succession of intrapopulation variants of three example contigs over time, with each combination of SNPs lasting only months. The graphs show the number of shared individual SNPs between a given date and the reference date defined by the peak of each line (the peak height represents the number of SNPs in that contig); for example, the red line on the top graph shows comparisons using the SNP profile from December 2009 as the reference for comparison to all of the other dates and the orange line uses a February 2013 reference. Presence of a SNP is defined as >1% in the population. To aid visualization, colours were added to the lines of randomly selected reference dates for emphasis. These three demonstration viral contigs are marked in **a** with red circles, and they occupy the 3rd, 135th and 2,198th ranks, respectively (from top to bottom, that is). They also range from highly abundant with a high SNP density to moderately rare with lower SNP density. Note that there are two scales for the bottom right panel; the left *y* axis applies to the solid lines and the right *y* axis applies to the dotted lines. **d**, A summary of similar calculations and similar results for all of the suitable contigs (those with adequate coverage, $n = 4{,}002$ contigs); the values were normalized to percentage values to be shown on the same scale. The boxes show the median and the interquartile range, and the whiskers show the minimum and maximum values marked at 2× the interquartile range from the median. Data are all-versus-all population profiles at each month compared with previous (negative numbers) and subsequent (positive numbers) months. Extended Data Fig. 6 shows that very similar patterns are obtained when contigs are replaced by reference genomes of cultured isolates with very close relatives in our dataset, suggesting that the patterns are not assembly-related artefacts.

residence time of 2–3 weeks[26]; invoking mixing or advection of viruses with these continual changes also raises questions about the mechanisms of variation in the source locations. Although the accumulation of SNPs might potentially arise from genetic drift, the consistent rapid loss of each consecutive SNP profile resembles a series of selective purges[24]. We suggest that the most parsimonious explanation for these genetic changes is coevolutionary interactions of viruses and hosts, reflected by changes (including oscillations) in genotypes or allele frequencies on ecological to evolutionary timescales, for example, fluctuating selection dynamics[4]. Specifically, we posit that the polymorphic variants that we observe are either directly responsible or linked to variants with infection-related phenotypic differences, such as allowing phage evasion of restriction enzymes and CRISPR-like defence systems, or altering host range[27]. As our SNP profiles are a population statistical composite and do not necessarily distinguish between multiple strains within a given SNP profile, recurrence over time of particular mutations in a given strain would contribute to the similarity between months but could be masked by other changing strains. Although a comprehensive analysis of the 3.2 million SNPs is far beyond the scope of this report, we performed a preliminary analysis of the locations of SNPs to see whether they occurred relatively uniformly across genomes or in distinct patches, and in particular kinds of genes. As the majority of marine viral genes are unannotated, we examined the relatively well-known T4-like myoviral contigs, and we found no clear SNP patterns (both uniform and patchy distributions, structural and non-structural genes), and very different distributions even among the few that we examined (Extended Data Fig. 4, Supplementary Tables Supplementary Information). The vast majority (96%; χ^2 $P < 0.01$) of observed polymorphic variants encode no changes in amino acids (within bioinformatically determined coding regions), suggesting that their purifying selection occurs mostly at the nucleotide level. This suggests that the changes were mostly involved in protection against restriction enzymes or other nucleotide-level host defences, but it is possible that some neutral changes are linked with successful non-synonymous changes elsewhere on the genome. We posit that these variants are constantly purged (local extinction or falling below detection limit) from the population when enough of the host

population develops resistance to infection, by new mutations or ascension of pre-existing resistant types. New successful viral variants will then arise from rarity (that is, the bank model[28]), mutation or advection (migration). The rapidity of changes is probably too fast to be accounted for by de novo mutations alone, so the ascension of rare types (persistent or imported from other regions) may be particularly important. Thus, the community can change constantly at the strain level while remaining relatively stable near the species level. A similar inference was made previously for two controlled aquatic systems (aquaculture pond and saltern, both designed for relative stability) by Rodriguez-Brito et al.[21] on the basis of a much smaller and shorter-time-span metagenomic dataset.

Patterns of community assembly and persistence are central ecological questions. We uncovered long-term quasi-steady coexistence of thousands of viral contig populations, and intrapopulation genotypic changes revealed patterns that are consistent with coevolution-driven dynamics. Despite its importance and vast theoretical work[29], direct field observations of coevolutionary outcomes are rare or limited to model systems[5,30]. Our observations offer genomic insights into the long-term dynamics and coexistence of naturally occurring virus–host pairs.

Methods

Sample collection and DNA extraction. Seawater was collected monthly in a Niskin bottle at a depth of 5 m as part of the SPOT (https://dornsife.usc.edu/spot/); 0.5–1 l was filtered through a 0.22 µm Sterivex cartridge (Millipore, using a Durapore filter) then onto a 25 mm 0.02 µm Anotop (Whatman) filter assembled in tandem using a peristaltic pump. DNA (operationally viral, 0.02–0.22 µm) was extracted from the Anotop filter membranes using the Epicentre Total DNA kit as described by Steward and Culley[31].

Library preparation and sequencing. Aliquots of DNA were sent for library preparation and sequencing at the DOE Joint Genome Institute as part of the Community Science Program on a grant to N.A. (proposal ID, 2799). All of the libraries were prepared according to the manufacturer's instructions (Swift 1S Plus or Nextera XT; details are provided under proposal 2799 at the JGI Genome Portal) from a targeted DNA quantity of 1 ng per sample, and included as many PCR cycles as were necessary to obtain 200 pM of DNA for sequencing, with a maximum of 20 cycles.

Bioinformatic analyses. *Quality control of reads and assembly.* Initial sample-by-sample assembly was performed at the DOE Joint Genome Institute. In brief, trimmed, screened paired-end Illumina reads (see documentation for bbtools filtered reads) were read corrected using bfc[32] (v.r181) with the options '−1 -s 10g -k 21 -t 10'. Reads with no mate pair were removed. The resulting reads were then assembled using SPAdes[33] assembler (SPAdes v.3.11.1) using a range of *k*-mers with the following options: '-m 2000 --only-assembler -k 33,55,77,99,127 --meta -t 32' (memory limit: 2,000 GB; *k*-mer sizes: 33, 55, 77, 99 and 127; number of threads: 32; using the metagenomic flag). The original raw data, quality controlled data and original assemblies were deposited at the DOE JGI Genome portal under the proposal ID 2799.

Cross-assembly and genome de-replication. The 53 original sample-by-sample-based assemblies were merged using minimus2 (ref. [34]) with the following settings: '-D OVERLAP = 1000 –D MINID= 95', requiring 95% identity over 1,000 bp. This enabled us to bridge the regions that did not assemble in SPAdes, such as through microdiversity (which breaks such assemblies), and it also merges long regions that exceed 95% identity. The latter step means that any contigs that were merged yield consensus sequences, each representing a population of >95% identical sequences, a percentage identity that has been reported to include members of species-like units in marine bacteriophages[8,9]. We recognize that merging contigs this way could potentially lose some information on spatial or temporal patterns of individual variants (although our SNP analysis that compared individual bases from all of the original reads avoided this problem, because the contigs were just used for initial mapping; see below). The individual assemblies (each of the 53 dates performed independently) totalled 224,216 contigs that were larger than 5 kb with an N50 of 10 kb, and after our merging step there were 99,722 contigs with an N50 of 12.5 kb. The fact that 224,000 contigs in 53 samples were reduced by cross-assembly to 99,000 contigs means that on average each final contig was created from 2.4 original contigs. This relatively small amount of merging is itself an indication of similarity in composition among all of the samples. We further evaluate the cross-assembly below.

Bioinformatic evaluation of the cross-assembly. We used an overlap-based cross-assembly of individual assemblies owing to the impracticalities of de novo assembling 5.1 billion reads. Furthermore, adding up the microvariation present at each month into a single cross-assembly would break the assemblies in regions of high coverage and high variation as discussed previously[23,35]. However, cross-assembly could generate artefacts such as chimaeras. We evaluated the cross-assembly using several methods (Extended Data Fig. 5, Supplementary Discussion). First, we investigated how much of the cross-assembly, in practice, was merged in regions with 95%, 96%, 97%, 98%, 99% and 100% identity. We observed that 92% of the alignments within the merged regions were at least 98% identical, and only a small percentage were 95–96% or 96–97% (Extended Data Fig. 5). Second, whereas cross-assembly required at least 1,000 bp overlap, the final assemblies had a large distribution of merged alignments that spanned into tens of kb, the large majority of which were 5–10 kb in length (Extended Data Fig. 5). When these merged lengths were normalized to the percentage of the contig covered, 86% of the alignments covered at least 90% of the contigs (Extended Data Fig. 5), meaning most of the merging was between almost completely overlapping (including nested) contigs, rather than bridging between long contigs with short overlaps. Finally, we computed the number of contigs that were merged per new contig (Extended Data Fig. 5), and the vast majority (73%) of cross-assembled contigs came from merging 2 or 3 contigs. Overall these analyses reveal a low risk of substantial numbers of chimeric assemblies in our dataset, and indicate that this merging step was, in practice, primarily a 'dereplication' that created consensus sequences in a large majority of cases between completely overlapping 98% identical regions.

Bioinformatic identification of viral contigs. The metagenomes used in the present work were operationally extracted from viral-size fractions (0.02–0.2 µm), yet we then used a very conservative approach to identify viral genomes (as this size of fraction can include cellular fragments with DNA, tiny cells, non-viral free DNA and gene transfer agents). We used two different computational methods that discover viral sequences from metagenomic assemblies—(1) VirSorter[36] and (2) VirFinder[37]. VirSorter uses signature marker genes, known viral genes and known genomic characteristics (such as strand bias and gene density) to assign contigs to different categories, each with different degrees of confidence. VirFinder is a reference-independent method that identifies viral contigs on the basis of *k*-mer frequency distributions; this machine-learning approach uses logistic regression models to predict the likelihood that a sequence of is viral. For our work, we used categories one and two from VirSorter (that is, contigs with known viral marker genes and contigs with genes annotated as viral) and from category three (genomic characteristics shared by viral genomes) only if they had been identified with VirFinder with a score larger than 0.98 and $P < 0.01$. For both programs, the databases used for training the algorithms correspond to the defaults. For VirFinder, the database was the nucleotide viral sequences on NCBI from before 1 January 2014 (ref. [37]). For VirSorter the database was predicted proteins from viral sequences on NCBI in January 2014 in addition to viral sequences from curated metagenomes[36]. We added the original method of identification as viral for all contigs as a column in the Supplementary Tables. In total, original cross-assembly yielded 99,722 contigs that were larger than 5 kb; we identified 19,907 of these as viral, and only the latter is evaluated and discussed in this manuscript.

Gene calling and annotation. Open reading frames (ORFs) were predicted using Prodigal[38] from all contigs larger than 5 kb (cross-assembled final set, 99,722 contigs). These ORFs were functionally annotated by top BLASTP hits against the non-redundant protein dataset of GenBank (accessed 28 August 2018). A summarized annotation of all of the putative viral contigs is provided in the Supplementary Data. The full annotation is available at NCBI under the BioProject PRJNA550983.

Taxonomy assignment. Taxonomy of each viral contig has two aspects (Supplementary Tables)—(1) we can potentially identify the virus itself by matching its sequence to known viruses using BLASTN, and (2) we may be able to match viruses to potential hosts using VirHostMatcher (d2* distance); each approach is described in brief below.

(1) Top BLASTN: a nucleotide blast database was constructed using all of the viral contigs in RefSeq (accessed 8 August 2018), viral genomes from the viral proteomic tree server[39], and further expanded including fosmids[40] and assemblies from global metagenomic projects[1,6] (see the 'In-house genome database' section). We considered significant hits if the alignment was longer than 1,000 bp. All of the hits, percentage identities and alignment lengths are provided in the Supplementary Tables under the 'Taxonomy' tab in the 'I. Virus ID by Best blastN hit' section.

(2) VirHostMatcher: this method measures the similarity of virus and host *k*-mer word patterns by the d2* distance[41] on the basis of word frequency usage between viral contigs and cellular genomes. For the potential hosts, we combined two databases. The first was a subset of the microbial genome database[42] ($n = 6,318$) selecting only marine bacteria, archaea and the relatively few marine eukaryotes whose genomes have been fully sequenced. The second was our own curated collection of metagenomically assembled genomes and single-cell genomes from several different collections, selected because they have significant hits in cellular metagenomes from our study location. Their taxonomy was evaluated

by phylogenetic placement of marker genes using the Genome Taxonomy Database Tool kit (GTDB-Tk)[43]. These databases can be found at https://doi.org/10.6084/m9.figshare.8968316.v1. All distances are reported in Supplementary Tables under the 'Taxonomy' tab, in the 'II. Likely Host by d2* distance (VirHostMatcher)' section.

Contig abundance calculations. Competitive read mapping was performed using Bowtie2 (ref. [44]); only reads that mapped with a quality score larger than 1 were retained; importantly, Bowtie2 was run non-deterministically. Only completely aligned reads were considered with a minimum identity of 98%, fairly consistent with previous methods and within what is broadly considered to be a virus species/population[8,9]. Abundance of viral contigs was then calculated (see equation below) as the number of reads recruited per contig (H) and normalized to sequencing depth (N) and contig length (L). Read abundances were then rescaled by dividing by the smallest number across all samples; final normalized relative abundances span 7 orders of magnitude; these values are provided in Supplementary Tables. This approach is identical to that used in other metagenomic projects[1,10,40,45], albeit using different rescaling methods. Although we do not consider a minimum read count per sample per contig, our recruitments, in practice, had a minimum of 12 reads for the lowest recovered contig; furthermore, 4,002 contigs had a coverage of at least 10× across 90% of the full length, and these were the contigs that were used for determining SNPs. We also used the FastViromeExplorer[46] pipeline, because it further considers evenness across the contig, and its results are provided in the Supplementary Discussion.

$$\text{Abundance} = \frac{(H \times N^{-1} \times L^{-1})}{\min(H \times N^{-1} \times L^{-1})}$$

Bray–Curtis similarity. Bray–Curtis similarity (1 − Bray–Curtis dissimilarity), which compared proportions (determined from recruitment) of all contigs, was calculated using vegan[47] for all of the possible pairwise combinations among all of the sampling points. Relative abundances were calculated as described above, and all of the community members were used. Each sampling date was assigned a sequential value from 1 (June 2009) to 64 (September 2014); the time lag represents the difference between these sequential numbers. All of the individual data points are shown in Fig. 2, and means were calculated for all time lags.

Sensitivity analyses. We performed a sensitivity analysis to investigate the relative contributions of abundant versus rare members of the community to seasonality. The overall abundance table was subsetted multiple times to include populations that would account for 5%, 10%, 25%, 50% and 75% of the data both moving from the most abundant to the least abundant and vice versa. Bray–Curtis similarity was calculated as described above for each independent subset.

SNPs. SNPs were identified using standard tools on a per-sampling-date basis, that is, each sampling date was treated completely independently for the purposes of calling SNPs. In brief, reads were mapped to the viral contigs using Bowtie2 (ref. [44]) as above (independently per sampling date, mapped to our cross-assembly), and the resulting alignment files were converted to a BAM format and sorted using samtools[48] using the options 'view -S -b | sort'. Variation among reads per site (reads were not compared with the underlying contig sequence) was calculated using samtools[48] and bcftools[49] as: 'mpileup -g -f | bcftools call –ploidy 1 -mv'. Variants were considered to be real if they passed the following filter 'filter -e '%QUAL<20 || DP<10'; this filter removes low quality reads and SNPs that occur in less than 10 reads. Finally, owing to the high coverage of some contigs, variants were only considered to be bona fide and used for downstream analysis if they had a frequency of >1%. Similar methods and thresholds have been used previously[9,22].

Intrapopulation variation, SNPs profiles and temporal succession. SNPs were called on a sample-by-sample basis as described above. To compare SNPs in a contig between dates, we used only those in contigs with a minimum of 10× coverage[22] across at least 90% of the contig length on all dates; a total of 4,002 contigs (~20% of the viral contigs) met these criteria. If a polymorphism was shared between two dates, it was counted, then the sum of all of the shared polymorphic sites was determined. Intrapopulation variation was calculated as this sum of all of the polymorphic sites across all sampled dates, divided by the length of the contig, and is expressed in Fig. 3a and elsewhere, where appropriate, as SNP density. Analyses to estimate the date to date variation shown in Fig. 3 were performed using pairwise comparisons of all of the SNPs profiles (extracted as described above) from all 4,002 contigs. Importantly, no consensus sequence was generated, and alignments were used at 98% only to determine the placement of a read in a particular contig. The variation associated to each contig at each time point was then compared to all of the other time points. The specific scripts used to generate Fig. 3c are now stored in figshare under the following link: https://doi.org/10.6084/m9.figshare.8872796.v1.

In-house genome database. Annotation and identification of viral contigs was aided by our in-house aquatic environment viral isolate genomic database; details and accessions numbers are part of the Supplementary Tables, under tab 'InHouseGenomeDatabase'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data are available at the JGI Genome portal under the proposal ID 2799 (to N.A. and J.A.F.). All data needed to evaluate the conclusions in the paper are provided in the paper or the Supplementary Information. Final cross-assembled sequences are deposited at NCBI under the BioProject ID PRJNA550983.

Code availability

Custom code is available at https://doi.org/10.6084/m9.figshare.8872796.v1.

Received: 29 April 2019; Accepted: 4 November 2019;
Published online: 09 December 2019

References

1. Brum, J. R. et al. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
2. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).
3. Cram, J. A. et al. Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J.* **9**, 563–580 (2015).
4. Brockhurst, M. A. et al. Running with the Red Queen: the role of biotic conflicts in evolution. *Proc. R. Soc. B* **281**, 20141382 (2014).
5. Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275–278 (2010).
6. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
7. Chow, C. E. T. et al. Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *ISME J.* **7**, 2259–2273 (2013).
8. Deng, L. et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**, 242–245 (2014).
9. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from Pole to Pole. *Cell* **177**, 1109–1123 (2019).
10. Aylward, F. O. et al. Diel cycling and long-term persistence of viruses in the ocean's euphotic zone. *Proc. Natl Acad. Sci. USA* **114**, 11446–11451 (2017).
11. Gregory, A. C. et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genom.* **17**, 930 (2016).
12. Chow, C. E. T. & Fuhrman, J. A. Seasonality and monthly dynamics of marine myovirus communities. *Environ. Microbiol.* **14**, 2171–2183 (2012).
13. Lenski, R. E. Coevolution of bacteria and phage: are there endless cycles of bacterial defenses and phage counterdefenses?. *J. Theor. Biol.* **108**, 319–325 (1984).
14. Hall, A. R., Scanlan, P. D., Morgan, A. D. & Buckling, A. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol. Lett.* **14**, 635–642 (2011).
15. Van Houte, S. et al. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
16. Weitz, J. S., Hartman, H. & Levin, S. A. Coevolutionary arms races between bacteria and bacteriophage. *Proc. Natl Acad. Sci. USA* **102**, 9535–9540 (2005).
17. Thingstad, T. F., Pree, B., Giske, J. & Våge, S. What difference does it make if viruses are strain-, rather than species-specific? *Front. Microbiol.* **6**, 320 (2015).
18. Martiny, J. B. H., Riemann, L., Marston, M. F. & Middelboe, M. Antagonistic coevolution of marine planktonic viruses and their hosts. *Ann. Rev. Mar. Sci.* **6**, 393–414 (2013).
19. Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.* **59**, 3393–3399 (1993).
20. Arkhipova, K. et al. Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *ISME J.* **12**, 199–211 (2018).
21. Rodriguez-Brito, B. et al. Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**, 739–751 (2010).
22. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
23. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
24. Bendall, M. L. et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1602 (2016).
25. Minot, S. et al. Rapid evolution of the human gut virome. *Proc. Natl Acad. Sci. USA* **110**, 12450–12455 (2013).
26. Hickey, B. M. Circulation over the Santa Monica-San Pedro Basin and Shelf. *Prog. Oceanogr.* **30**, 37–115 (1992).

27. Marston, M. F. et al. Rapid diversification of coevolving marine *Synechococcus* and a virus. *Proc. Natl Acad. Sci. USA* **109**, 4544–4549 (2012).
28. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trend. Microbiol.* **13**, 278–284 (2005).
29. Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569–577 (2002).
30. Betts, A., Gray, C., Zelek, M., MacLean, R. C. & King, K. C. High parasite diversity accelerates host adaptation and diversification. *Science* **360**, 907–911 (2018).
31. Steward, G. F. & Culley, A. I. in *Manual of Aquatic Viral Ecology* (eds. Wilhelm, S. W. et al.) Ch. 16 (2010).
32. Li, H. BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**, 2885–2887 (2015).
33. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
34. Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinform.* **8**, 64 (2007).
35. Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B. & Fuhrman, J. A. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat. Commun.* **10**, 1169 (2019).
36. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
37. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
38. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
39. Nishimura, Y. et al. Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere* **2**, e00359-16 (2017).
40. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
41. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free *d2** oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).
42. Uchiyama, I., Mihara, M., Nishide, H. & Chiba, H. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic Acids Res.* **43**, D270–D276 (2015).
43. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
44. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
45. Zhao, Y. et al. Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013).
46. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**, e4227 (2018).
47. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
48. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
49. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

Acknowledgements

We thank the directors and staff of the USC Wrigley Institute for Environmental Studies for supporting the SPOT, T. Gundersen and Fuhrman laboratory members, and the crew of the R/V *Yellowfin* for their help during our monthly sampling. We also thank J. McNichol for curating the local genome database and D. Needham, E. Sieradzki and S. Hou for providing metagenomically assembled genomes. This work was supported by NSF grant no. 1737409, the NIH (1R01GM120624-01A1), the Gordon and Betty Moore Foundation Marine Microbiology Initiative grant no. 3779 to J.A.F., and the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES; grant ID 549943) to J.A.F.

Author contributions

J.C.I.-E., N.A. and J.A.F. designed the study and wrote the manuscript. N.A. initiated the project and performed the DNA extractions. J.C.I.-E. performed the bioinformatic analyses related to the identification, annotation, read mapping, variant calling and population genetics of the viral contigs.

Competing interests

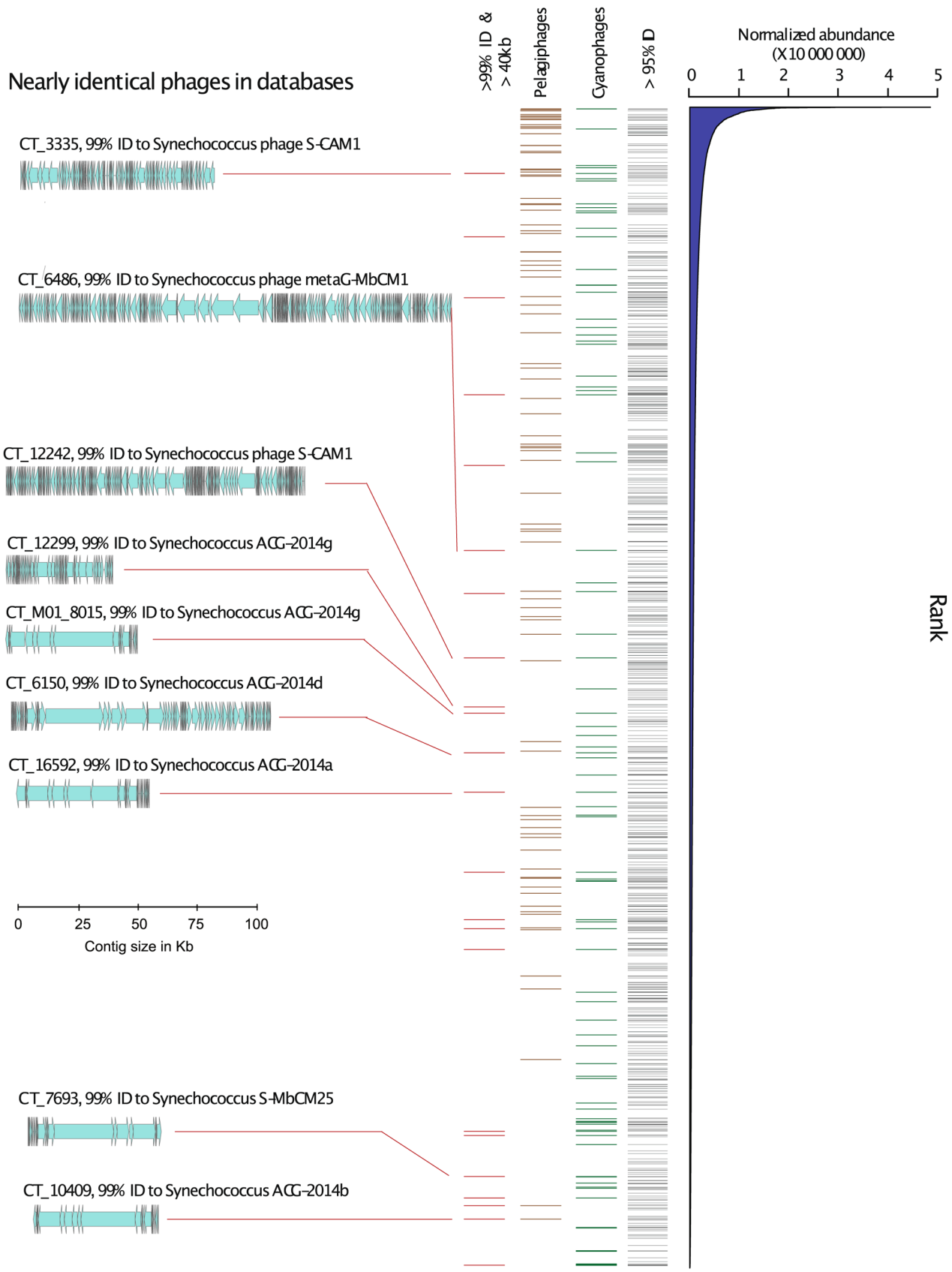
The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-019-0628-x.

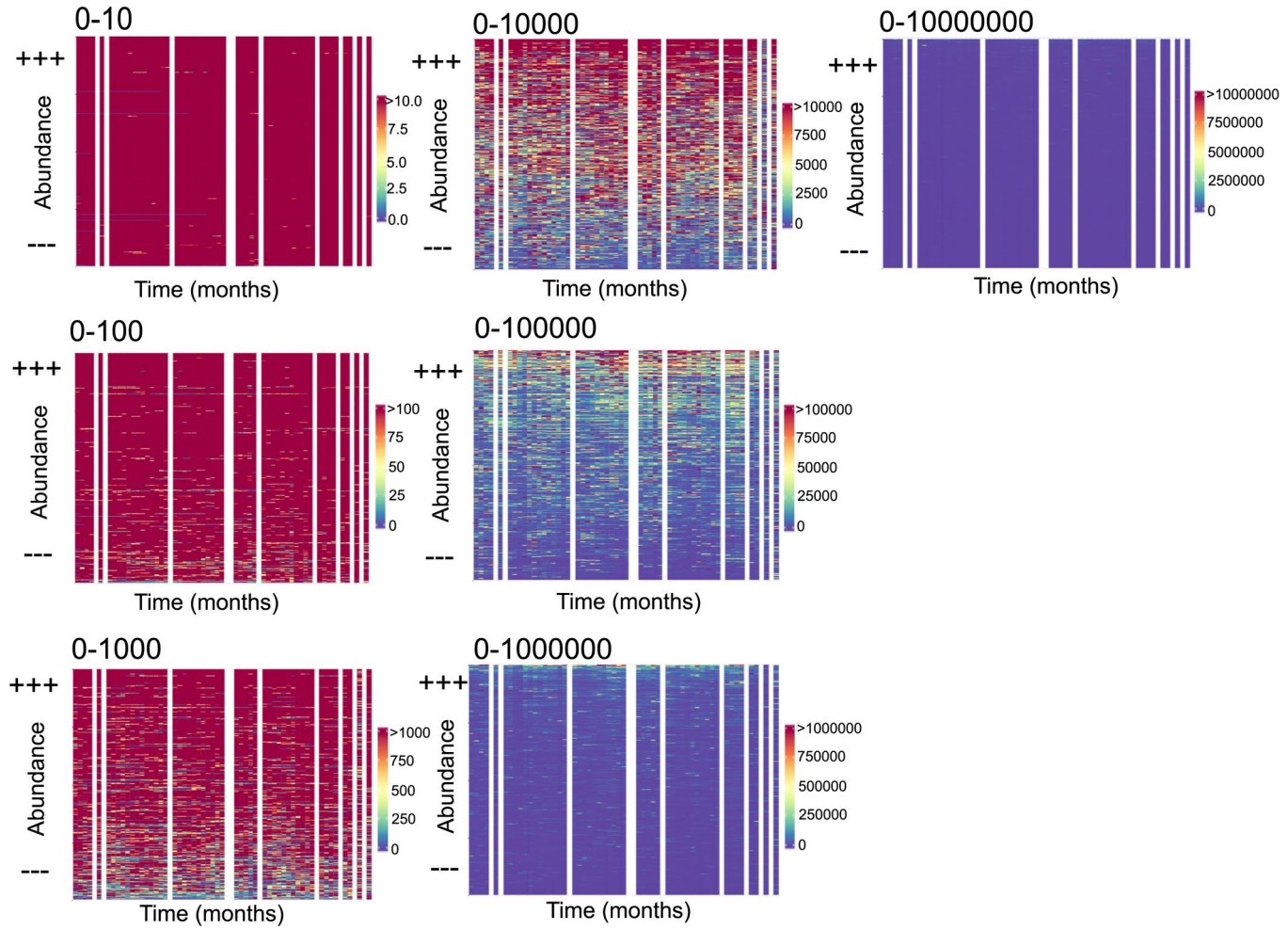
Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0628-x.

Correspondence and requests for materials should be addressed to J.A.F.

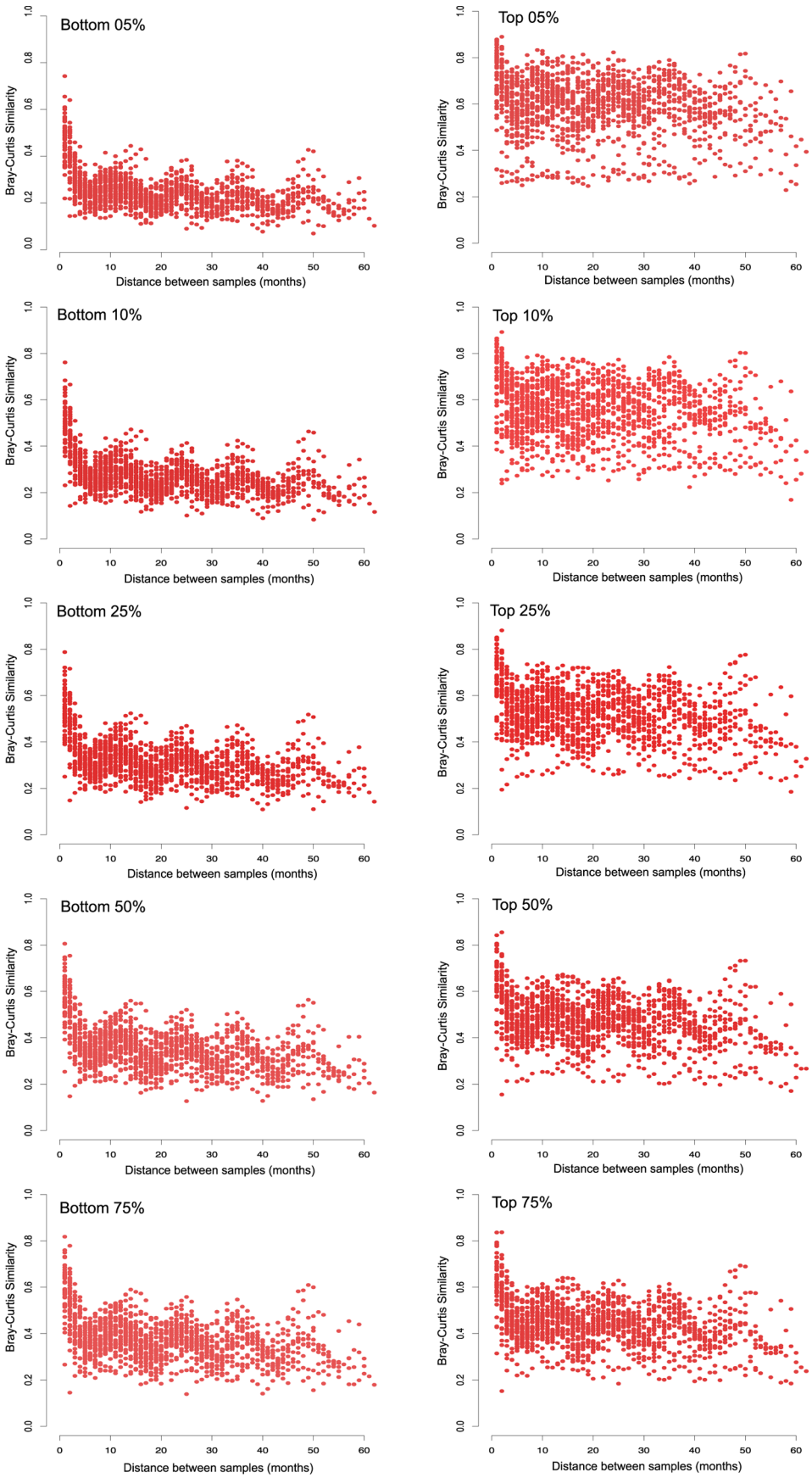


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Rank abundance and placement of identifiable viral populations. Rank abundance and placement of identifiable viral populations. From right to left: Right-most column is a rank abundance plot, where the most abundant viral contigs are at the top and the least abundant at the bottom. Cumulative abundances across all time points are shown, there are the product of normalization as stated in the methods. Middle columns show (i) In black, the positions of contigs with high (> 95% ID over 5kb) identity to previously identified virus sequences in databases, n = 746 (Details in Supplementary Tables). Due to constraints on visible line thickness, they appear to represent a majority of contigs, but note they are only 746 out of 19,907 total contigs (See details on Supplementary Tables); (ii) Green lines show viral contigs identified as cyanophage n = 73 (Supplementary Tables). (iii) Maroon lines show viral contigs identified as *Pelagibacter* phages n = 68 (Supplementary Tables). (iv) Red lines show viral contigs that are nearly identical, > 99% ID over at least 40 kb at the nucleotide level to previously described viruses (Supplementary Tables). Left-most panel shows genomic diagrams from selected contigs and their identified hit in databases (>99% identity), all drawn to the same scale. Remarkably, all these illustrations represent *Synechococcus* phages isolated off the coast of California.

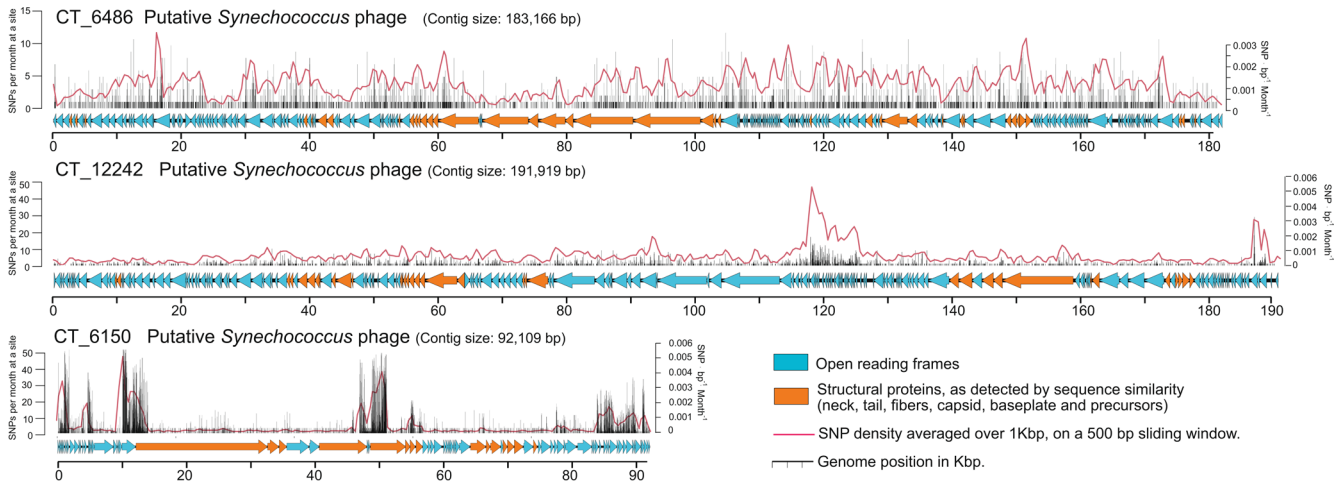


Extended Data Fig. 2 | Relative abundance of viral contigs on a linear scale. Heat Maps otherwise similar to Fig. 1b (which has a 7-decade log scale), showing relative abundance of the 19,907 viral contigs, one contig per row, during monthly sampling. Contigs are ordered by average abundance (over all months), highest at the top. White columns represent months with missing data (all data are in Supplementary Tables). Each panel has a different range, where abundances at or exceeding the maximum value appear as red. This display better allows visualization of temporal changes in contig abundance within orders of magnitude, compared to Fig. 1b where colors change little within each order of magnitude. Different subpanels are needed to visualize the ranges of all the contigs. "Zooming" the image optimizes the ability to visualize details.

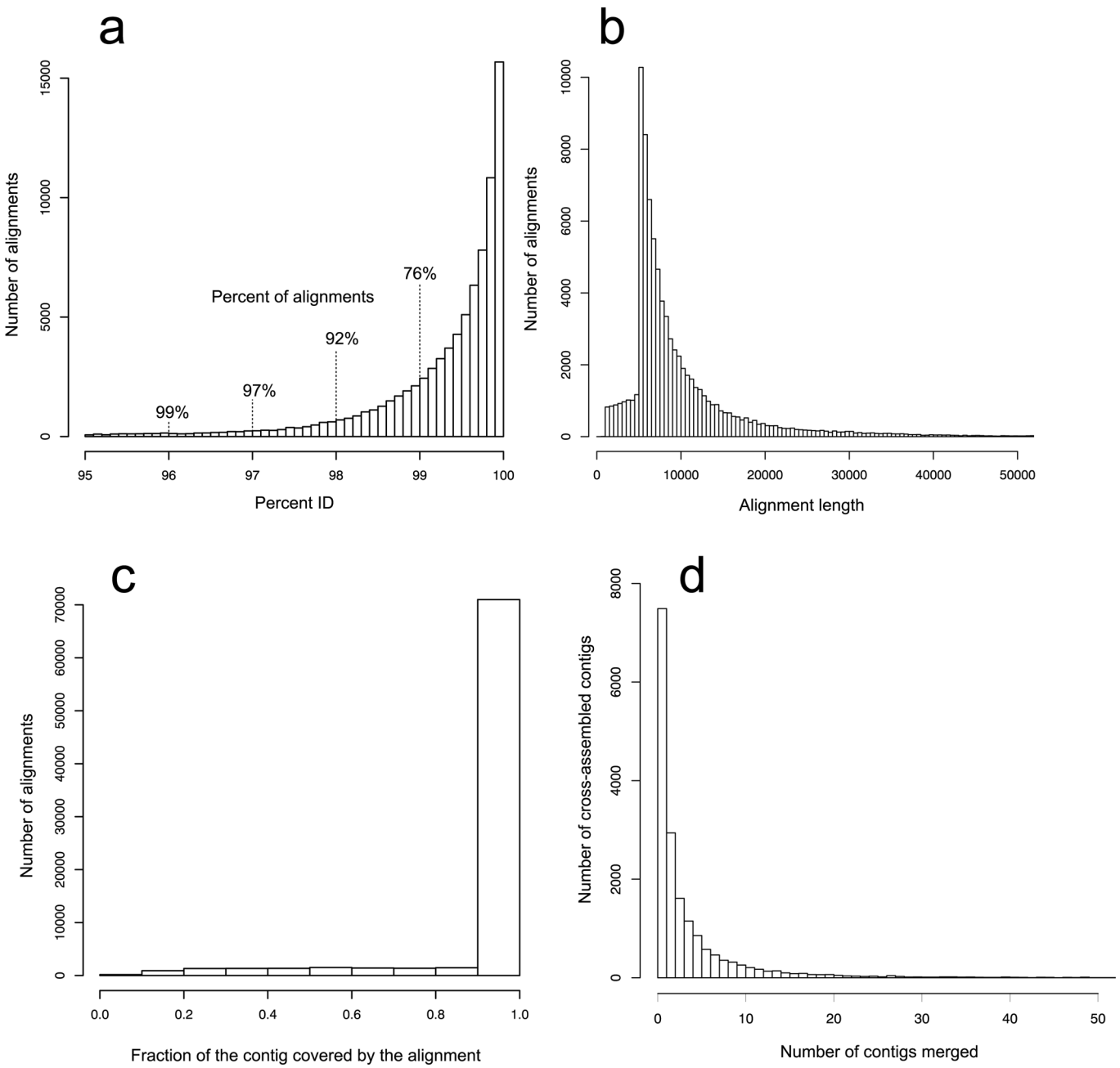


Extended Data Fig. 3 | See next page for caption.

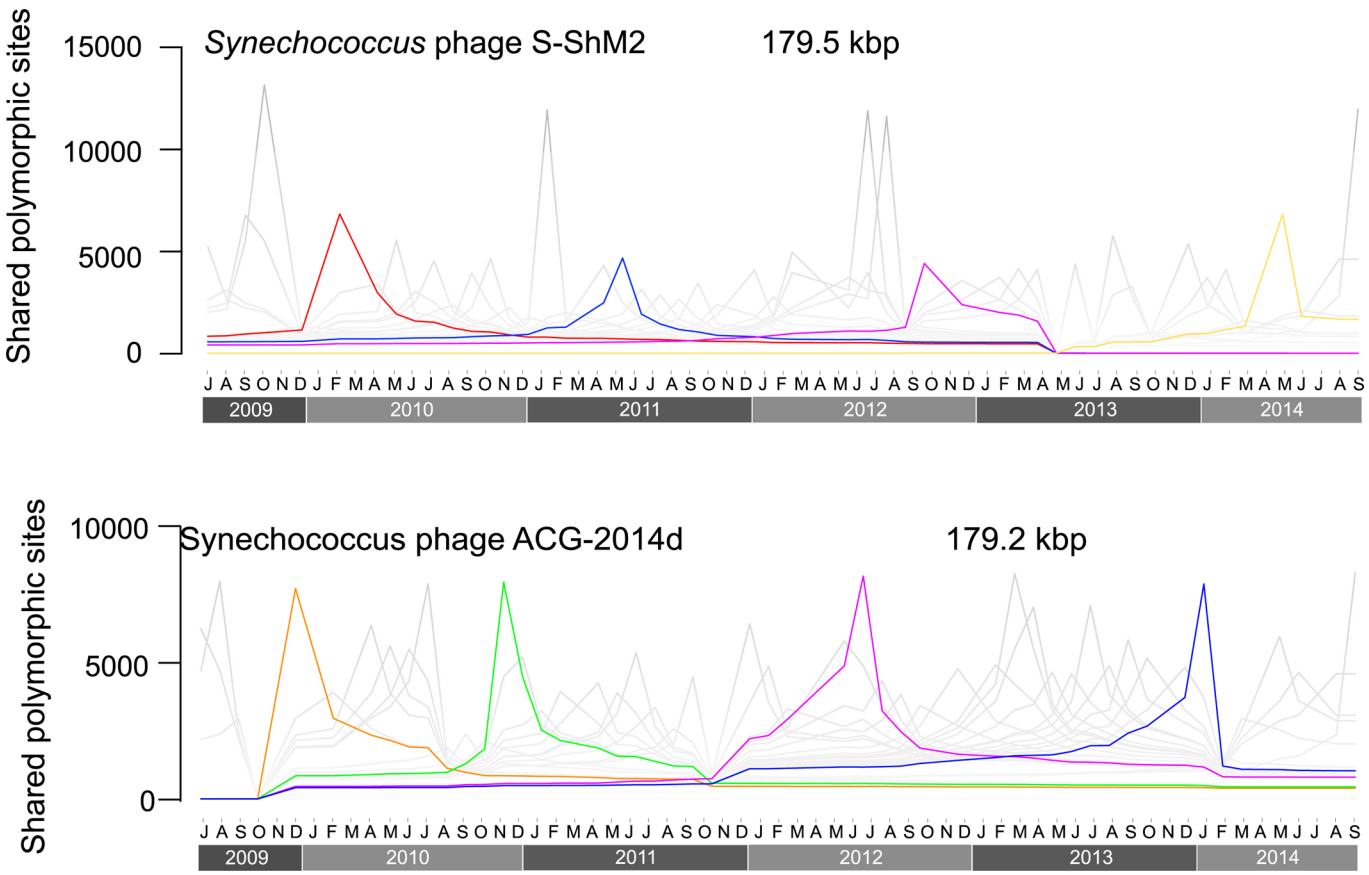
Extended Data Fig. 3 | Seasonality is driven by medium range abundance contigs. Similar to Fig. 2, these depict the Bray-Curtis distance vs time lag between samples, but here divided into different fractions of the rank abundance curve. The left column of graphs starts from the bottom (rare) part of the curve, with an increasing fraction of the contigs included in graphs displayed from top to bottom. The right column of graphs starts with the top most abundant contigs, with an increasing fraction of the curve included in graphs from top to bottom. Note on the right that as more members are included from the long rare tail of the rank-abundance curve, consistent seasonality increases and average similarity decreases. Generally, the rarer contigs show stronger seasonality than the most abundant ones. Because Bray Curtis similarity is proportionately more affected by the more abundant organisms in general, these indicate that middle-high percentiles (top 50th-75th) may dominate the collective community seasonality. Bottom 5% n = 4587 viral contigs, bottom 10 % n = 7092, bottom 25 % n = 12031, bottom 50% n = 16741, bottom 75% n=19093. Top 5% n = 61, Top 10% n=168, Top 25% n= 814, top 50% n = 3166, top 75% n = 7876.



Extended Data Fig. 4 | Distribution of polymorphic sites along three selected genomes. The three longest T4-like genome fragments were chosen (ranked 9433rd, 10382nd and 13113rd) because T4-likes are the most extensively studied group and have the best annotations. Each diagram shows all predicted open reading frames and their sense direction as depicted by an arrow. The y-axis on the left side (for thin black bars) shows the number of months that each location exhibited a SNP. The position along the x axis corresponds to their position along the genome. Y axis on the right (red line) shows the average number of SNPs per basepair on a 500 bp moving average. Note that only about ~5% of the sites are polymorphic although this is hard to visualize. For details, please refer to annotations and the per gene density of 20 representative (including the ones shown here) T4-like viruses is included in Supplementary Tables under the tab "SNPsAmongT4LikeViruses" Although not selected for this reason, these three show strikingly different patterns in SNP distributions, from relatively uniform (top panel) to very patchy with a few hotspots (bottom panel). Note y axes are scaled for each panel, and the top and middle ones have similar SNP densities to each other over most of their lengths.



Extended Data Fig. 5 | *Post hoc* evaluation of our cross-assembly strategy. Cross assembly merged contig sequences from different months when overall identity of overlapping regions greater than 1000 bp in length exceeded 95%. We evaluated how often these merged overlaps occurred at different percent identities to assess how much variation was combined, and also examined other useful statistics. a) Distribution of percent identity of all alignments used to merge contigs during our cross-assembly step, dotted lines represent the percent of alignments covered to the right of the line. Note that 92% of merges had >98% sequence identity b). Distribution of lengths of all alignments used to merge contigs during our cross-assembly step. Note most merged regions were 5,000–10,000 bp in length. c) Distribution of the fractions of the contig used during our merging step (that is length of the alignment divided by the contig length). Note that the vast majority of merges occurred over almost the entire lengths (90–100%) of the contigs d). Distribution of the number of contigs that were merged into a single contig during cross- assembly. Note that the vast majority of merged contigs came from 2 or three individual contigs. All panels taken together show that while merging occurred, the vast majority (86%) was between almost completely overlapping (including nested) and >98% identical sequence contigs, rather than bridging between long contigs with short overlaps.



Extended Data Fig. 6 | Read recruitment to fully sequenced isolates reveals identical patterns of succession. Dynamics of SNP profiles calculated from reads recruited (within 98% ID) to two reference genomes. These profiles were generated as those shown in Fig. 3 and as described in the methods. NOTE: The absolute number of polymorphic sites is bigger since a full genome is being considered.

nature research

Corresponding author(s): Jed A. Fuhrman

Last updated by author(s): Nov 1, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina software at sequencing facility.
Data analysis	We used only the following tools, which are all publicly available: bbtools; SPAdes version: 3.11.1; minimus2; Virsorter; Virfinder; Prodigal v2.6.2; BLAST v2.2.3; bowtie2 v2.2.6; VirHostMatcher-Net; Samtools v1.2; Fast-Virome Explorer; Their specific use and parameters are noted in the methods section. Custom code available at https://doi.org/10.6084/m9.figshare.8872796.v1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The manuscript includes the statement: "All sequencing data is available at the JGI Genome portal under the proposal ID 2799 (to NA and JF). All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials."
The Supplemental GenBank-like file includes all contig sequences
The Supplemental Excel spreadsheet includes raw contig length, %GC, abundance, identity information, and abundance in all samples.
Final cross-assembled sequences are deposited at NCBI under the BioProject ID PRJNA550983.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Community DNA from operationally identified virus-size fraction was collected monthly for 5 years without experimental manipulation.
Research sample	Seawater virus assemblages (as defined from being retained in between 0.02 - 0.2 um filters) were collected monthly. No experimental manipulation was done.
Sampling strategy	Seawater virus assemblages (as defined from being retained in between 0.02 - 0.2 um filters) were collected monthly. No experimental manipulation was done.
Data collection	Viromes were collected monthly; Seawater was filtered through a 0.2 um filter and then to 0.02 um filter, the latter was then used for DNA extraction. DNA amplification and sequencing was done at JGI as described in the methods section.
Timing and spatial scale	Viromes were collected monthly; Seawater was filtered through a 0.2 um filter and then to 0.02 um filter, the latter was then used for DNA extraction.
Data exclusions	Not all months are represented due to weather conditions, equipment failure, and failed DNA extractions; missing dates are noted in the figures as blank columns.
Reproducibility	Not relevant for an environmental time series.
Randomization	N/A. We described an ecological time series. Continuous sampling for ~20 years.
Blinding	NA

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Field conditions varied monthly, this project is the result of 53 cruises. For all environmental data associated please see: https://dornsife.usc.edu/spot/datasets-summary/ . We also cited publications by Cram et al., and Chow et al., that overlapped in time with this study, and reported and used these environmental conditions for interpretation.
Location	San Pedro Ocean Time series is located at 33°33'N and 118°24'W.
Access and import/export	NA
Disturbance	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging