nature microbiology

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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*]

[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

Supplementary discussion.

Choice of viral contigs for analysis. The DNA sequenced for this study was extracted from the "viral size fraction," specifically 0.02-0.2 μm. This size fraction contains not only viruses, but also tiny cells that pass the 0.2 filter, cell debris, free DNA or that associated with colloids, gene transfer agents, etc. For practical reasons, the extraction from frozen filters does not include steps to remove non-viral DNA, such as DNase or density gradients. To be sure our analysis applied to the viruses themselves and not any non-viral DNA, we chose only those contigs that we could be fairly sure were viral. Hence, of the total of 99,722 non-redundant contigs >5kb, we used the 19,907 that were classified bioinformatically as viral with substantial confidence (VirSorter and VirFinder, see Methods). The other ~80K contigs no doubt include many viral ones, but we do not know at this time which they are.

Sensitivity analysis and alternative detection and quantification approach. A potential criticism is that the viral persistence and temporal patterns we reported could theoretically, in part, result from detecting moderately short, highly conserved genomic regions in different viruses. To avoid such artifacts and provide a more conservative determination we performed a sensitivity analysis in which we varied the minimum threshold of abundance to consider a contig as present (SI figure 1). Alternatively, we also used an existing published tool for detecting and quantifying viruses by read recruitment, Fast Virome Explorer (FVE[1]), to examine the contig temporal distributions. FVE is more conservative (and less sensitive) than our other approach because it requires that reads have more coverage, and somewhat even coverage, of a contig to be counted. We ran FVE with default settings, briefly: >10% of the contig covered, more than 10 reads, and masking high coverage regions.

FVE-based analyses reiterated what we reported with our original approach. When analyzed by FVE, each metagenomic sample recovered 80-99% of all the contigs (SI figure 2), overlapping with what we reported using our original methods (95-97%?), with lower numbers as expected from the more demanding criteria in FVE. The FVE analysis still shows no significant gains of contig populations over the 5 years (SI figure 2a,b), and the FVE rarefaction curve still contrasts similarly to previously reported global spatial patterns (Figure 1a, main text). Finally, the seasonal pattern that we found with our original approach looked the same when calculated from FVE results (SI figure 2c). So, while FVE necessarily found fewer instances of our contigs, due to its more demanding detection criteria, it still yielded the same overall patterns and conclusions.

Read recruitment vs cross assembly. One interesting consideration is the conundrum of contrasting expectations from read recruitment vs. cross-assembly statistics. On the one hand, the contig populations were found to be persistent on the basis of read recruitment (at 98% identity). But on the other hand, cross-assembly statistics show that our reported >5kb contigs made from individual months can be cross-assembled (merged by >95% similarity over at least 1000 bp) from individual monthly contigs only rarely, specifically each final contig was cross-assembled from an average of only 2.4 monthly contigs (Methods, Extended data 5). If the same contig population is always present, one might expect that it would assemble each month (hence many months would merge into final contigs), but this did not happen. Our data support the hypothesis that this is due to extensive microvariation (documented in Fig 3), and/or different abundances of microvariant genotypes, interfering with the creation of >5kb contigs on most months. The difficulty in assembling metagenomic mixtures of closely related strains is well documented[2,3] . To test our hypothesis, we looked at recovery of contigs of different lengths, > 4kbp, > 3kbp, > 2kbp, > 1kbp and >300 bp, as they map onto the final cross-assembled contigs we reported. Our results show that as the length of the contig shortens, more of them map to the cross-assembly, i.e., are present in the underlying sequencing effort (SI figure3, left panels). Using a specific example from within our data, our most abundant contig was recovered in > 5 kb form from only one month (it was 5.5 kb then), but when we looked for shorter contigs that mapped to it, we found > 3 kb contig versions occurred in two months, > 2kb versions occurred in 16 months, and > 1 kb contig versions occurred in 50 months. We conclude that while the underlying population was present persistently, microvariation usually prevented long enough assemblies to pass our 5 kb criterion. However, that microvariation did not interfere with recruitment of individual reads at the 98% identity level that we used to show presence of each population.

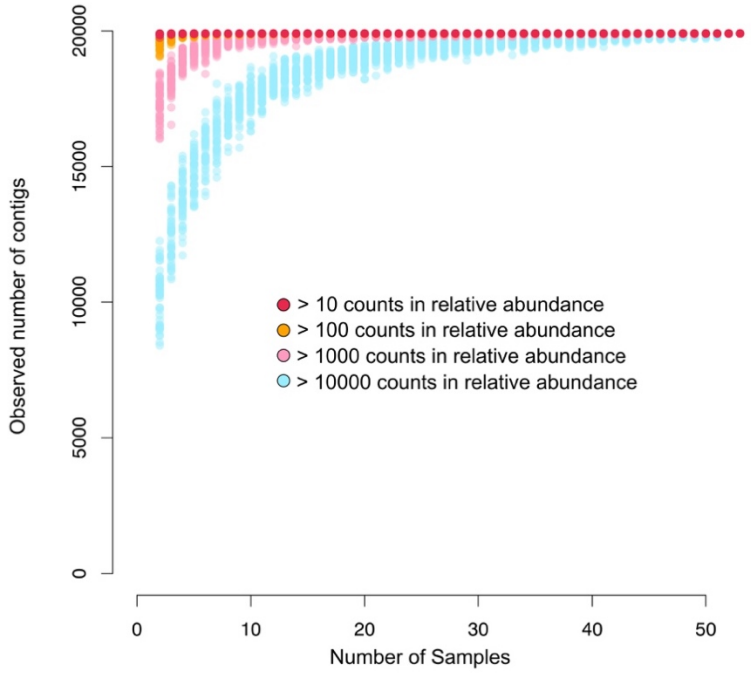
Additionally, when we mapped the temporal gap between two sampling times with a contig (> 300bp, > 1, 2, 3, 4 and 5kb) present that maps to our final merged contig set (SI figure 3, right panels), we observed a decline in

number of mapped contigs with temporal distance; sampling dates closer in time have more similar versions of contigs. This is consistent with our SNP-based observation that nearby dates are most likely to share microvariant composition (Fig 3). Interestingly, SDF2 shows cross-assembly local peaks at yearly intervals, consistent with seasonal re-occurrences.

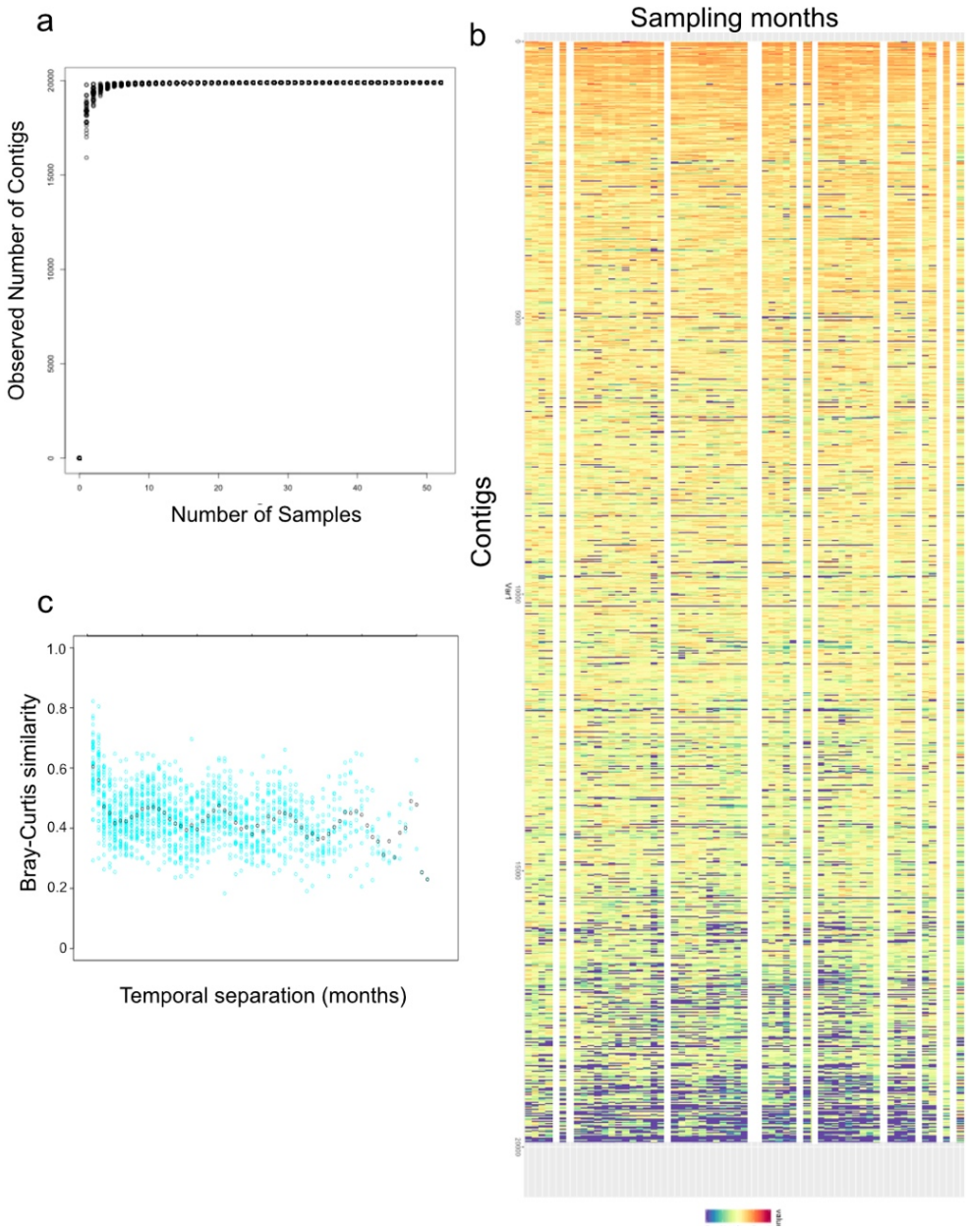
Cultured references vs assembled contigs. To address the possibility that our SNP profile patterns (similarity peaks lasting months) might somehow be an artifact of the metagenomic contig assembly and merging process (e.g. because closely related contigs from adjacent months were often merged), we also evaluated SNP profile patterns using pure culture virus reference genomes from viruses that had been isolated off California and which had close relatives at our site. These yielded the same kinds of SNP profile patterns (Extended data 6).

Relating to previous models. Our results generally support the "Royal Family" model of phage succession[4], whereby dominant phage are replaced with their close relatives as a result of arms-race-like dynamics, though we may be seeing fluctuating selection (recurrence possible) in addition to arms race dynamics (in which new dominants are always different), and we extend the concept to the non-dominant phage as well. Considering continual changes and thus no steady state, our results do not support a steady state kill-the-winner model, though elements of such models that consider invasions by new strains[5] could be relevant.

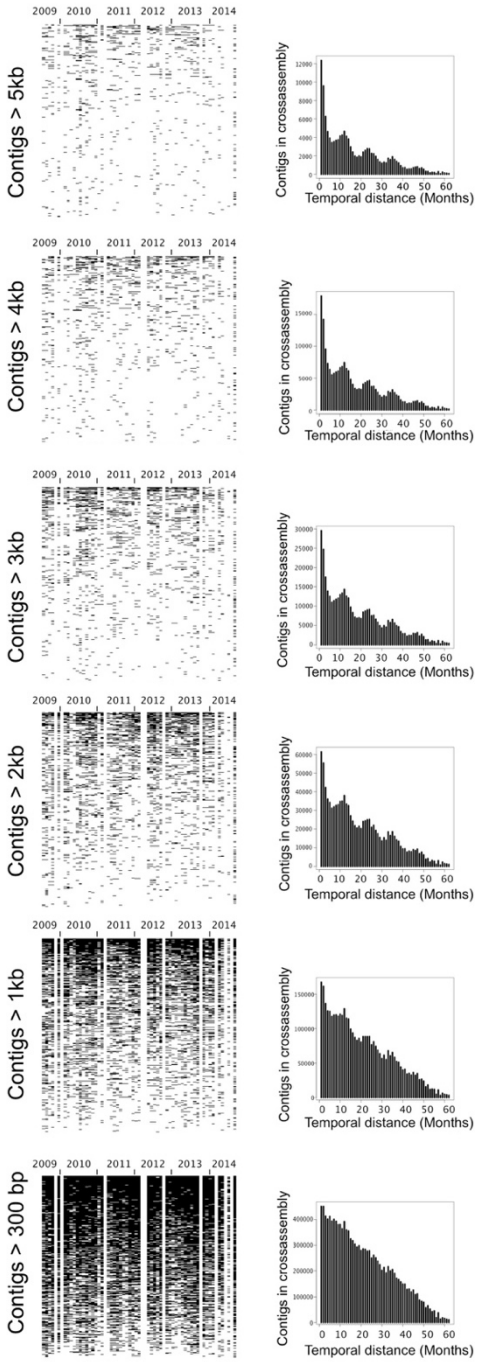
1. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer: A pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* (2018). doi:10.7717/peerj.4227
2. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* (2017). doi:10.7717/peerj.3817
3. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* (2017). doi:10.1038/ncomms15892
4. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nature Microbiology* (2018). doi:10.1038/s41564-018-0166-y
5. Thingstad, T. F., Pree, B., Giske, J. & Våge, S. What difference does it make if viruses are strain-, rather than speciesspecific? *Front. Microbiol.* (2015). doi:10.3389/fmicb.2015.00320



SI figure 1. Sensitivity analysis of persistence. Accumulation curve showing the presence of viral contigs over 53 monthly surface ocean viral (0.02-0.2 μm) metagenomes. All curves were calculated as in figure 1a, and different colors denote different cutoffs in abundance for a contig to be considered 'present'. Note that 10 to100-fold increases in the threshold for presence give extremely similar plots to that shown in Figure 1a. Note also that relative abundances are adjusted for contig length and sampling depth (see methods), and in practice the lowest abundance of any "present" contig in Figure 1a was actually 12 reads.



SI figure 2. Fast Virome Explorer reveals similar patterns to our own approach. Accumulation curve showing the presence of the majority of viral contigs over 53 monthly surface ocean viral (0.02-0.2 μm) metagenomes **b)** Heat map 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, highest at the top. White columns represent months with missing data. Note colors remain similar across time for the large majority of contigs. **c)** The Bray-Curtis community similarity index (blue dots) was calculated among all possible sample pairs (n= 1378 combinations) and plotted as a function of the number of months separating their sampling. Red dots correspond to the mean of all observations at a given separation.



SI figure 3S. Mapping of individual assemblies into cross-assembly. All panels are organized in similar manner, each row (n = 19907) represents a contig in the final cross-assembly. Cells (or lines due to resolution limitations) are colored black if a contig from the original sample-by-sample assembly (columns) match the corresponding contig in the final cross-assembly. As more contigs are considered (from top to bottom) due to lower minimum length cut-offs, the sequence space in the final cross-assembly gets more covered. Right panels: Each histogram represents the frequency of the temporal distances, in months, between all cross assembled contigs when the different size cutoffs are applied.