FVE-novel: Recovering Draft Genomes of Novel Viruses and Phages in Metagenomic Data

Saima Sultana Tithi (saima5@vt.edu)
Virginia Polytechnic Institute and State University https://orcid.org/0000-0001-5087-0125
Frank O. Aylward
Virginia Polytechnic Institute and State University
Roderick V. Jensen
Virginia Polytechnic Institute and State University
Liqing Zhang
Virginia Polytechnic Institute and State University

Research

Keywords: Metagenomics, Viral metagenomics, Virus, Phage, Viral genome assembly, Assessment of virus assembly

Posted Date: March 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17154/v1

Abstract

Background Despite the recent surge of viral metagenomic studies, recovering complete virus/phage genomes from metagenomic data is still extremely difficult and most viral contigs generated from de novo assembly programs are highly fragmented, posing serious challenges to downstream analysis and inference.

Results Here we develop FVE-novel, a computational pipeline for reconstructing complete or near-complete viral draft genomes from metagenomic data. FVE-novel deploys FastViromeExplorer to efficiently map metagenomic reads to viral reference genomes or contigs, performs de novo assembly of the mapped reads to generate scaffolds, and extends the scaffolds via iterative assembly to produce final viral scaffolds. We applied FVE-novel to an ocean metagenomic sample and obtained 268 viral scaffolds that potentially come from novel viruses. Through manual examination and validation of the ten longest scaffolds, we successfully recovered four complete viral genomes, two are novel as they cannot be found in the existing databases and the other two are related to known phages.

Conclusions The hybrid reference-based and de novo assembly approach used by FVE-novel represents a powerful new approach for exploring viral diversity in metagenomic data. FVE-novel is freely available at https://github.com/saima-tithi/FVE-novel .

Introduction

Viruses/phages (or bacteriophages whose hosts are bacteria or archaea, hereafter used interchangeably) are the most abundant biological entities on Earth and our recognition of their ubiquitous presence and enormous diversity has been made possible only recently by culture-independent high-throughput sequencing [1-8] . Viruses/phages are now increasingly realized as a major player of the global ecosystem [9, 10] and a key biotic component influencing climate change [11, 12]. However, despite the significance, our understanding is greatly limited by the vast unknown of their sequence space, more than 99% of the viruses/phages are uncharacterized for their genome sequences and considered part of the “microbial dark matter” [9, 13-17].

With the fast development of biotechnology, recovering unknown/novel viral genomes is now made possible by the increasing availability of metagenomic data. The most commonly used approach is to conduct *de novo* assembly of metagenomic reads followed by sequence comparison to determine whether or not the contigs assembled are unknown (aka uncultivated virus genome, UViG). However, using *de novo* assembly alone is highly insufficient for recovering complete or near-complete unknown genomes (i.e., contigs are ≥90% of expected genome sizes, [18]). Indeed, IMG/VR 2.0 [19], the largest database for uncultivated virus/phage genomes, houses 739,759 uncultivated genome fragments, of which only ~3% represent complete or near-complete genomes. A recent large-scale marine metagenomic study showed that of a set of 27,346 viral contigs generated by de novo assembly, <0.2% are complete

[20]. Another study found that *de novo* assembly of Illumina short reads generated from marine environmental samples could not fully recover any of the thousand full-length phage genomes revealed by single-molecule nanopore sequencing of the same samples [21]. Taken together, these observations show that the traditional *de novo* short read assembly methods alone are seriously inadequate for building the complete genomes of unknown viruses/phages from metagenomic data.

To address the limitation, we develop FVE-novel, a computational pipeline that combines short reads mapping, de novo assembly, and iterative local assembly together to reconstruct complete or near-complete viral genomes. FVE-novel leverages FastViromeExplorer (FVE) [22], a program we developed recently that uses a pseudoalignment approach to quickly align reads to a database of viral sequences to identify the presence and relative abundance of known viral groups in a metagenomic sample. FVE-novel calls FastViromeExplorer to quickly identify viral reads from metagenomic data. It then uses SPAdes [23, 24] to assemble the viral reads to generate viral scaffolds. The mapping step allows the removal of nonviral reads, making the assembly faster than the commonly used approaches that usually assemble all the reads in the sample. Next FVE-novel grows each scaffold using iterative local assembly to generate final viral scaffolds. Those scaffolds are then compared to the input references and the average nucleotide identity (ANI) is used to determine whether the viral scaffolds are potentially novel. FVE-novel reports potential novel viral scaffolds together with per base depth of coverage, which can be used to examine the quality of the generated scaffolds and whether they are chimeric.

Methods

Algorithm overview

The inputs of FVE-novel are metagenomic short reads and a reference database of genomes/contigs. As reference databases often contain genomes/contigs that are highly similar to one another, FVE-novel includes a preprocessing step that bins or clusters the genomes/contigs in the database based on similarity into groups. This step is optional if the input reference database is already binned or does not has any redundancy. The FVE-novel pipeline contains three main steps, (1) the read mapping and seed scaffold generation step, (2) the scaffold extension step where the seed scaffolds are extended through iterative local assembly, and (3) generating ANI and coverage statistics for all extended scaffolds. Figure 1 shows a comprehensive outline of all the steps.

Preprocessing the reference database

The reference database is preprocessed and clustered into bins based on sequence similarity. This step is optional and may be unnecessary depending on the reference database users choose. FVE-novel was tested using 24,411 viral contigs collected from the Global Ocean Virome (GOV) study [9]. The 24,411

viral contigs are grouped into 15,280 bins, with 15,222 bins corresponding to epipelagic and mesopelagic viral populations and 58 bins corresponding to bathypelagic viral populations.

Step 1. Read mapping and seed scaffold generation

In the first step, FVE-novel takes single-end or paired-end reads and a reference database as input and invokes FastViromeExplorer for read mapping/alignment. FastViromeExplorer outputs the viruses present in the sample and their abundances, and the reads mapped to the viruses. As metagenomic data contains reads from various organisms, using FastViromeExplorer, all the reads coming from bacteria, archaea, and other hosts are quickly filtered out and only viral reads are retained for the next assembly step. In this way, assembling genomes other than viruses can be avoided, making assembly much faster and more efficient.

After getting the mapping result from FastViromeExplorer, all the mapped reads are binned based on the binning of reference genomes. Here all the reads mapped to the reference contigs/genomes from the same bin are binned together. Based on the type of input reads, either SPAdes (version 3.10.1, for single-end reads) or metaSPAdes (option: --meta, for paired-end reads) is used with default settings and with the assembly mode (option: --only-assembler). Here, SPAdes breaks the reads into fixed-length k-mers, builds a de bruijn graph using the overlap of k-mers, and traverses the graph to produce longer sequences or genome fragments.

After the initial assembly step, all the generated scaffolds are examined and only the ones longer than 2 kb are kept and used as “seed scaffolds” or input scaffolds for the next extension step. The 2 kb cutoff is implemented so that the pipeline does not attempt to extend every short segment that can be time consuming yet may contribute little to the overall scaffolds. Nonetheless, users can adjust this parameter depending on their research needs.

Step 2. Extending seed scaffolds using iterative assembly

In the second step or extension step of FVE-novel, for each seed scaffold, start and end edges are extracted using BEDTools [25]. Read length * 1.5 is used as default edge length. Then, for each seed scaffold, all the reads are mapped to the two edges of that scaffold using Salmon [26]. It is found that assembling the original scaffold along with the overhanging reads mapped to the edges of that scaffold can effectively extend the scaffold in one or both ends. SPAdes is used for assembly using the original scaffold as “--trusted-contigs” and the reads mapped to the edges of the scaffold as input reads. The

mapping-and-assembly step is run iteratively for each scaffold until the scaffold stops growing. All the extended scaffolds are then clustered using CD-HIT [27] with 95% global ANI and only the longest one in the cluster is kept as the final extended scaffolds.

Step 3: Analysis of new contigs and comparison to reference sequences

The third and final step involves generating summary statistics for the scaffolds including ANI, percentage of aligned nucleotides, and per base depth of coverage. As each scaffold is assembled from the reads mapped to a set of reference genomes belonging to a bin, the reference genomes in the bin are used for calculating ANIs of the corresponding scaffold using MUMmer's "dnadiff" program [28]. The output of FVE-novel contains ANI and the percentage of aligned nucleotides between the extended scaffold and each of its reference genomes. All the reads are mapped back to each scaffold using Bowtie2 [29] and depth of coverage is calculated using SAMtools [30]. Summary statistics can be used to evaluate the quality and the novelty of the scaffolds. For example, a low ANI with all the reference genomes indicates that the scaffold might be novel or unknown. If the depth of coverage is fairly uniform along the scaffold, it is less likely to be chimeric.

Benchmarking on real data

To demonstrate and evaluate the performance of FVE-novel, we downloaded 12 metagenomic samples from NCBI SRA (SRX2912986, SRX2912968, SRX2912972, SRX2912964, SRX2912992, SRX2912996, SRX2912975, SRX2912979, SRX2912983, SRX2912998, SRX2913002, and SRX2912985), corresponding to 12 time points at the same location/station in an ocean metagenomic study [31] (see description in Suppl. Table 1), and applied FVE-novel to the sample SRX2912986 and used the GOV database [9] as reference. Other samples were used to evaluate the scaffolds generated by FVE-novel. A Linux based cluster with 64 CPUs and 128 GB RAM was used to generate all the results.

Results

Length distribution of the FVE-novel scaffolds

We applied the FVE-novel pipeline to the ocean virome sample [31] (HOE Legacy II diel viral-size metagenome: Station 70). This sample contains 18,471,506 paired-end reads with read length 151 bps. Using the GOV database containing 24,411 contigs as reference and the reads from the station 70 as input, FVE-novel generated 268 scaffolds. Figure 2 shows the length distribution of the scaffolds, ranging from 2,026 bps to 193,112 bps, with a median length of 4,561 bps. There are 66 scaffolds longer than 10 kb.

Comparison between the FVE-novel scaffolds and their reference sequences

As the scaffolds are originally assembled from the reads that are mapped to a set of contigs or genomes, these contigs or genomes can be considered the “reference contigs/genomes” of the scaffolds. Among the 268 scaffolds, 59 are longer than their corresponding reference sequences. Figure 3 shows the lengths of the 59 scaffolds with respect to those of their reference genomes, indicating that FVE-novel can generate considerably longer scaffolds than the original references. Moreover, some scaffolds have high ANIs (e.g., >95%) to their references, suggesting that parts of these scaffolds are present in the reference database and FVE-novel successfully extended those partial references. Some scaffolds have low ANIs to their references, indicating that these scaffolds are potentially novel.

Comparison of the FVE-novel scaffolds against several databases

To examine the quality of the scaffolds generated by FVE-novel, we analyzed the longest ten scaffolds in detail. The ten scaffolds range from 72,532 bps to 193,112 bps, with average depths of coverage ranging from 48.5x to 338x.

As the scaffolds are assembled using the GOV database as the input reference database, it is expected that the scaffolds have sequence similarity to the GOV database sequences. Table 1 shows the BLAST results against the GOV sequences, with average ANIs (%) ranging from 89.21 to 97.78 and alignment percentage from 6.25-99.98. Figure 4, generated using Artemis [32], shows how the scaffolds align to their corresponding references in the GOV database and the sequence similarity, further supporting that FVE-novel successfully extended the reference genomes.

To see how the FVE-novel scaffolds compared to the 483 scaffolds reported previously in the original study [31] from where we collected our read sample, we used BLASTN [33] to compare FVE-novel generated scaffolds with the 483 assembled scaffolds. Table 1 shows the best hit result. Nine FVE-novel scaffolds have very high ANIs to their best hits (95.25 to 99.73%). This is not surprising as the same read sample was used to generate both the FVE-novel scaffolds and also some of the 483 scaffolds. Interestingly, scaffold *S6* does not have any hits in the 483 scaffolds, suggesting that FVE-novel was able to reconstruct a viral scaffold that was not recovered by the original study [31]. Scaffold 6 has a best hit to the GOV sequence with 97.78% ANI and 93.09% alignment length. Notably, all nine scaffolds generated by FVE-novel are longer than the scaffolds from the original study.

We also used BLASTN to compare the ten scaffolds against the nr nucleotide database and found that four scaffolds (*S0, S1, S2*, and *S6*) have no significant hits and the other six scaffolds (*S3, S4, S5, S7, S8*, and *S9*) have similarity to *Prochlorococcus phage P-SSM4* (ANIs ranging from 86.46 to 92.68%) and *Prochlorococcus phage P-HM2* (ANI 84.89%). This result is consistent with the original study [31] where parts of the Prochlorococcus phage genomes were also recovered from the data.

Comparison within the FVE-novel scaffolds

The observation that some of the ten scaffolds have similar BLAST hits suggests that their sequences might be similar. Therefore we also compared the scaffolds against each other (Figure 5). Four groups of

scaffolds are observed: group 1 containing *S0, S1*, and *S6*, group 2 containing only *S2* with no significant similarity to any other nine scaffolds, group 3 containing *S3, S4, S5, S7*, and *S9*, and group 4 containing only *S8*. As FVE-novel removed all scaffolds except one with identity greater than 95%, these scaffolds have similarity less than 95% and could be different viral species or divergent strains of the same virus. Note that 95% identity has been commonly used in the literature as the cutoff score for clustering/binning viral contigs into viral operational taxonomic units (vOTUs) [9, 34, 35]. Group 3 scaffolds (*S3, S4, S5, S7*, and *S9*) all have similarity to *Prochlorococcus phage P-SSM4* (Table 1), thus likely represent different strains of *Prochlorococcus phage P-SSM4*. In the following, we analyzed the four groups of scaffolds in detail.

Group 1 scaffolds (S0, S1, S6) *S0* is the longest scaffold in group 1 and has a length 193,112 bps. It is generated from the station 70 sample, thus we checked if this scaffold is also present in the other 11 viral metagenomic samples. Figure 6(a) shows the depth of coverage of *S0* in all 12 samples. The average depth of coverage of *S0* in station 6, 14, 18, 22, 28, 32, 37, 52, 56, 61, 67, and 70 samples are 95.26, 28.39, 32.24, 94.78, 52.08, 73.75, 17.71, 20.18, 82.55, 148.61, 85.54, and 93.02 respectively, with station 37 having the lowest coverage (17.71) and station 61 (148.61) the highest coverage. The pairwise Pearson correlation coefficient of per base coverage of *S0* between any two samples ranges from 0.77 to 0.94, and therefore even though the abundances of *S0* differ among stations, the per base read coverages along the scaffold are highly correlated among samples.

Figure 6(a) also shows that coverage dropped greatly after 150 kb for all samples, which means fewer reads got mapped to this region. We further analyzed this region for sequence composition and did not find any homopolymer or short repeats, and thus the lower mapping rate is not caused by repeat sequences. Another possible cause of coverage drop is that this part of the scaffold was the misassembly from a less abundant strain of the same virus. Here, we explored this possibility by reassembling the scaffold. Specifically, we used the “Map to Reference” algorithm implemented in Geneious 11.0.4 [36] to examine whether there are multiple viral strains and if there are, whether a complete assembly of the dominant strain can be generated. All of the metagenome reads in the station 70 sample were aligned to scaffold *S0* using the “Low sensitivity/Fastest” settings allowing for 10% mismatches. The alignment showed a large number of Single Nucleotide Polymorphisms (SNPs), revealing the existence of two or more strains for this phage as well as distinct regions of high and low coverages consistent with the suspected chimeric assembly of these strains. To recover the dominant strain, the consensus sequence from the alignment was segmented into contigs with high coverage > 40X. These contigs were binned into lists of contigs with similar coverage for further assembly. Next, contigs in each bin were iteratively extended using Geneious by mapping reads to the contig ends with high stringency. Specifically, all of the paired-end reads that were previously mapped to the contigs were aligned to these high coverage contigs using “Map to Reference” with stringent “Custom Sensitivity” settings allowing no more than 1% “Mismatches per Read” and 1% “Gaps per Read” and requiring that both of the paired-end reads map to the new consensus sequence. This process was iterated for each contig using Geneious’ “Fine Tuning” settings up to “100 times”. This process was continued until the extended contigs merged together, maintaining approximately uniform coverage, and could no longer be

extended or closed into a circular genome sequence. Finally, we recovered a 153 kb scaffold from scaffold S0 with uniform coverage across all 12 samples (Figure 6(b)). Comparison of this 153 kb scaffold with *S0* reveals that the middle 90 kb of *S0* (starting at 60 kb and ending at 150 kb) are exactly the same as the 153 kb strain assembled by Geneious (Suppl. Figure 1[1] [2]). But the first 60 kb of *S0* has around 80% similarity to the 153 kb strain, so this part of *S0* could be from a different strain and the last 40 kb of *S0* (staring at 150 kb and ending at 193 kb) is the result of assembly artifact (Suppl. Figure 1). Comparison of the 153 kb strain with *S1* (155 kb) shows that FVE-novel recovered this dominant strain successfully in *S1* (Suppl. Figure 2). As scaffold *S1* is a correct representation of the dominant strain of this novel virus, from figure 7, we can see that *S1* has a uniform coverage across all 12 samples. In scaffold *S6*, our tool captured the 80 kbp of the 153 kbp dominant strain (Suppl. Figure 3) and *S6* also has a uniform depth of coverage across all 12 samples (Suppl. Figure 4). In order to find out if this virus has multiple strains present in station 70 sample, we sub-sampled the 153 kb strain into 10 pieces of 20 kb long and applied TenSQR [37], a viral quasispecies reconstruction tool, to each piece. For each piece, TenSQR reported two or three strains, consistent with our observation that multiple strains of this virus are present in the station 70 sample (Suppl. Table 2).

Group 2 scaffold (S2) Scaffold *S2* is present in all samples with varying abundances but highly correlated depth of coverage among the samples (Figure 8(a)). With the same aforementioned procedure, we generated a 151 kb scaffold that most likely represents a complete novel virus genome recovered from *S2*. Figure 8(b) shows that this 151 kb scaffold has a uniform depth of coverage across all 12 samples. Comparison of this scaffold to *S2* shows that our tool successfully recovered most of this virus except a 15 kb long portion of it (Suppl. Figure 5). We also checked if multiple strains of this virus are present in the station 70 sample using TenSQR. Results show that multiple strains of this virus are also present in this sample (Suppl. Table 3).

Group 3 scaffolds (S3, S4, S5, S7, S9) As scaffolds *S3, S4, S5, S7*, and *S9* have similarity to *Prochlorococcus phage P-SSM4*, we chose to analyze in detail the longest scaffold *S3* with length 132,604 bp. Figure 9(a) shows consistent depth of coverage of *S3* across all 12 samples, but coverage dropped dramatically in all samples from 50 kb to 65 kb. We then aligned all the reads from station 70 sample to scaffold *S3* using Geneious and observed the presence of multiple strains. Using the same procedure as above, we were able to recover a 177 kb scaffold representing the dominant strain of *Prochlorococcus phage P-SSM4* present in station 70 sample, with a uniform coverage across all 12 samples (Figure 9(b)). Comparison of this 177 kb scaffold to *S3* shows that *S3* matches the dominant strain with about 100% similarity for all except the 15 kb segment (the 50-65 kb part in *S3*) where the similarity dropped to 80% (Suppl. Figure 6). This result implies that our tool successfully captured about 117 kb of the dominant strain except from a piece of length 15 kb where the assembly switched to a different strain with lower coverage. Consistently, analysis with TenSQR suggests the presence of three to seven strains in the sample (Suppl. Table 4). Interestingly, alignment of the 177 kp scaffold to *Prochlorococcus phage P-SSM4* (length 178,249 bp) shows similar as well as more divergent regions (ANI 82.9%) (Suppl. Figure 7).

Group 4 scaffold (S8) The depth of coverage of the scaffold *S8* is quite uniform across all 12 samples along the region except the beginning 6 kb and the last 10 kb where coverage is higher than the remaining region, which can be caused by a different strain (Figure 10(a)). According to the blast result, *S8* has similarity to *Prochlorococcus phage P-HM2*. We then used Geneious to recover the dominant strain from *S8* and obtained a 183 kb scaffold. Figure 10(b) shows that the 183 kb scaffold has uniform coverage across all 12 samples. Comparison to *S8* shows that in *S8* we recovered about 60 kb of this dominant strain (Suppl. Figure 8). Analysis with TenSQR suggests the presence of two or three strains in this sample. Alignment of the 183 kb scaffold to *Prochlorococcus phage P-HM2* (length 183,806 bp) also shows many similar regions and some dissimilar regions (ANI 86.56%) (Suppl. Figure 9).

Function annotation of the four complete scaffolds

We annotated genes from the four completed scaffolds, (a) the 153 kb dominant strain of a novel virus recovered from *S0*, (b) the 151 kb dominant strain of a novel virus recovered from *S2*, (c) the 177 kb dominant strain of *Prochlorococcus phage P-SSM4* recovered from *S3*, and (d) the 183 kb dominant strain of *Prochlorococcus phage P-HM2* recovered from *S8*. The four scaffolds have 195, 160, 220, and 231 proteins respectively predicted by Prodigal with metagenomics mode (i.e., Prodigal option: -p meta) [38]. We then annotated those proteins using eggNOG-mapper with the virus database and HMMER option [39]. For all the four scaffolds, eggNOG-mapper annotated some viral structural proteins such as capsid, baseplate, virion, and so on, indicating that they are potential virus genomes. Figure 11 shows the protein annotation of the two novel genomes recovered from *S0* and *S2*. Figure 12 shows the protein annotation of the two Prochlorococcus phage strains recovered from *S3* and *S8* respectively. As both phages are Prochlorococcus phage strains, as expected, the arrangement of viral structural proteins in these phage genomes are similar to that of Prochlorococcus.

Discussion

In this paper, we present FVE-novel, a new computational pipeline for recovering novel viral scaffolds based on reference-based mapping and iterative assembly. By applying our tool to an ocean metagenome sample, we assembled 268 viral scaffolds. Some of these viral scaffolds are quite long, which can be potential near-complete viral genomes. Manual curation and validation of the ten longest scaffolds led to successful recovery of four complete viral genomes. Among these four viral genomes, two are novel genomes as they were not found in the existing databases, one represents strain of *Prochlorococcus phage P-SSM4* and another one represents strain of *Prochlorococcus phage P-HM2*. We also noted substantial microdiversity in the phage genomes present in the metagenomic data, which should be considered further in future work given it is a potential complication to the recovery of full-length viral genomes.

Out of the 18,471,506 paired-end reads from station 70 sample, 1,240,164 reads (~6.7%) were mapped to the 268 scaffolds with 8.6% mapped to the 153 kb novel virus recovered from *S0*, 4.2% mapped to the 151 kb novel virus recovered from *S2*, 35.3% mapped to the 177 kb dominant strain of *Prochlorococcus*

phage P-SSM4, and 13.9% mapped to the 183 kb strain of *Prochlorococcus phage P-HM2*. Thus, the four new viruses contribute to more than 62% of the mapped reads, indicating that they are very abundant in station 70 sample.

Assembly of abundant marine viruses revealed a high incidence of microdiversity in naturally-occurring viral populations in the environment. Genomic chimeras often have anomalous coverage since they represent distinct genotypes that have been incorrectly merged, and most metagenome assemblers use the coverage information to break contigs or scaffolds into smaller pieces to avoid mis-assembly [24, 40-42]. Given that the recovery of complete viral genomes from metagenome assemblies is rare, it is plausible that microdiversity may be a prevalent cause of assembly fragmentation. Moreover, given that genomic microdiversity is difficult to identify and may not always lead to contig breakage by assembler, it is possible that a substantial number of viral contigs or scaffolds present in publicly-available repositories are in fact chimeras of multiple viral strains. The prevalence of this is difficult to ascertain given the paucity of high-quality complete viral genomic references for benchmarking. Future studies should therefore prioritize the rigorous identification of strain-level microdiversity in viral populations to identify the nature and extent of this genome-wide microdiversity in nature.

Conclusion And Usage Notes

Overall, FVE-novel implements a novel strategy to recover viral genomes and will serve as a powerful tool for future studies that will continue to enhance existing viral databases. The three steps in the FVE-novel pipeline (Figure 1) are implemented as separate modules so that users can use each module separately if needed. For example, for any metagenomic data, users can run the entire pipeline, i.e., all three steps, to obtain viral scaffolds. Alternatively, if a user has a read sample and some viral sequences generated from those reads (e.g., using a different assembler), the user can directly run the second seed extension step to extend the input viral sequences (all the modules are freely available at https://github.com/saima-tithi/FVE-novel).

For each scaffold generated, FVE-novel reports the ANI and percentage of aligned nucleotide between the scaffold and its reference. FVE-novel also reports per base depth of coverage along the scaffold, which can be used similarly to what we have shown above for examining the scaffold quality and identifying “suspicious regions” of misassembly. As our detailed analyses show that having multiple strains of the same viruses can create great challenges to assembly programs, it is paramount that users examine the report on depth of coverage for the generated scaffolds to ensure the quality of the putative novel viral genomic sequences.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

FVE-novel is freely available to download at https://github.com/saima-tithi/FVE-novel. The 12 ocean microbiome samples used in our study is publicly available in NCBI SRA under the accession numbers SRX2912986, SRX2912968, SRX2912972, SRX2912964, SRX2912992, SRX2912996, SRX2912975, SRX2912979, SRX2912983, SRX2912998, SRX2913002, and SRX2912985 and these samples were originally prepared and analyzed by Aylward et al. [31].

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no funding for this work.

Authors' contributions

LZ and SST conceived and designed the experiments. SST performed the experiments, implemented the software, wrote the manuscript, prepared figures and tables. All authors contributed to the analysis of data and review of the manuscript before submission for publication. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. DeLong, E.F., *The microbial ocean from genomes to biomes.* Nature, 2009. **459**(7244): p. 200-206.
2. Hanson, C.A., et al., *Beyond biogeographic patterns: processes shaping the microbial landscape.* Nature Reviews Microbiology, 2012. **10**(7): p. 497-506.
3. Roux, S., et al., *Viral dark matter and virus–host interactions resolved from publicly available microbial genomes.* elife, 2015. **4**: p. e08490.
4. Hug, L.A., *Sizing up the uncultured microbial majority.* MSystems, 2018. **3**(5): p. e00185-18.
5. Pace, N.R., *A molecular view of microbial diversity and the biosphere.* Science, 1997. **276**(5313): p. 734-740.
6. Rappé, M.S. and S.J. Giovannoni, *The uncultured microbial majority.* Annual Reviews in Microbiology, 2003. **57**(1): p. 369-394.
7. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: the unseen majority.* Proceedings of the National Academy of Sciences, 1998. **95**(12): p. 6578-6583.
8. Zou, S., et al., *Research on the human virome: where are we and what is next*. 2016, Springer.
9. Roux, S., et al., *Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses.* Nature, 2016. **537**(7622): p. 689.
10. Suttle, C.A., *Marine viruses—major players in the global ecosystem.* Nature Reviews Microbiology, 2007. **5**(10): p. 801-812.
11. Danovaro, R., et al., *Marine viruses and global climate change.* FEMS microbiology reviews, 2011. **35**(6): p. 993-1034.
12. Puxty, R.J., et al., *Viruses inhibit CO2 fixation in the most abundant phototrophs on earth.* Current Biology, 2016. **26**(12): p. 1585-1589.
13. Brum, J.R., et al., *Illuminating structural proteins in viral "dark matter" with metaproteomics.* Proceedings of the National Academy of Sciences, 2016. **113**(9): p. 2436-2441.
14. Reyes, A., et al., *Going viral: next-generation sequencing applied to phage populations in the human gut.* Nature Reviews Microbiology, 2012. **10**(9): p. 607-617.
15. Villarreal, L.P. and G. Witzany, *Viruses are essential agents within the roots and stem of the tree of life.* Journal of Theoretical Biology, 2010. **262**(4): p. 698-710.
16. Mizuno, C.M., et al., *Expanding the marine virosphere using metagenomics.* PLoS genetics, 2013. **9**(12).
17. Brum, J.R. and M.B. Sullivan, *Rising to the challenge: accelerated pace of discovery transforms marine virology.* Nature Reviews Microbiology, 2015. **13**(3): p. 147-159.
18. Roux, S., et al., *Minimum information about an uncultivated virus genome (MIUViG).* Nature biotechnology, 2019. **37**(1): p. 29-37.
19. Paez-Espino, D., et al., *IMG/VR v. 2.0: an integrated data management and analysis system for cultivated and environmental viral genomes.* Nucleic acids research, 2019. **47**(D1): p. D678-D686.
20. Coutinho, F.H., et al., *Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans.* Nature communications, 2017. **8**(1): p. 1-12.

21. Beaulaurier, J., et al., *Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities.* bioRxiv, 2019: p. 619684.
22. Tithi, S.S., et al., *FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data.* PeerJ, 2018. **6**: p. e4227.
23. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* Journal of computational biology, 2012. **19**(5): p. 455-477.
24. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome research, 2017: p. gr-213959.
25. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
26. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nature methods, 2017. **14**(4): p. 417.
27. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-3152.
28. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome biology, 2004. **5**(2): p. R12.
29. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.
30. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
31. Aylward, F.O., et al., *Diel cycling and long-term persistence of viruses in the ocean's euphotic zone.* Proceedings of the National Academy of Sciences, 2017. **114**(43): p. 11446-11451.
32. Carver, T., et al., *Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.* Bioinformatics, 2011. **28**(4): p. 464-469.
33. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.
34. Paez-Espino, D., et al., *Uncovering Earth's virome.* Nature, 2016. **536**(7617): p. 425.
35. Gregory, A.C., et al., *Marine DNA viral macro-and microdiversity from pole to pole.* Cell, 2019. **177**(5): p. 1109-1123.
36. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-1649.
37. Ahn, S., Z. Ke, and H. Vikalo, *Viral quasispecies reconstruction via tensor factorization with successive read removal.* Bioinformatics, 2018. **34**(13): p. i23-i31.
38. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**(1): p. 119.
39. Huerta-Cepas, J., et al., *Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper.* Molecular biology and evolution, 2017. **34**(8): p. 2115-2122.

40. Namiki, T., et al., *MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.* Nucleic acids research, 2012. **40**(20): p. e155-e155.
41. García-López, R., J.F. Vázquez-Castellanos, and A. Moya, *Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations.* Frontiers in bioengineering and biotechnology, 2015. **3**: p. 141.
42. Vázquez-Castellanos, J.F., et al., *Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut.* BMC genomics, 2014. **15**(1): p. 37.

Table

Please see the supplementary files section to access the table.

Figures

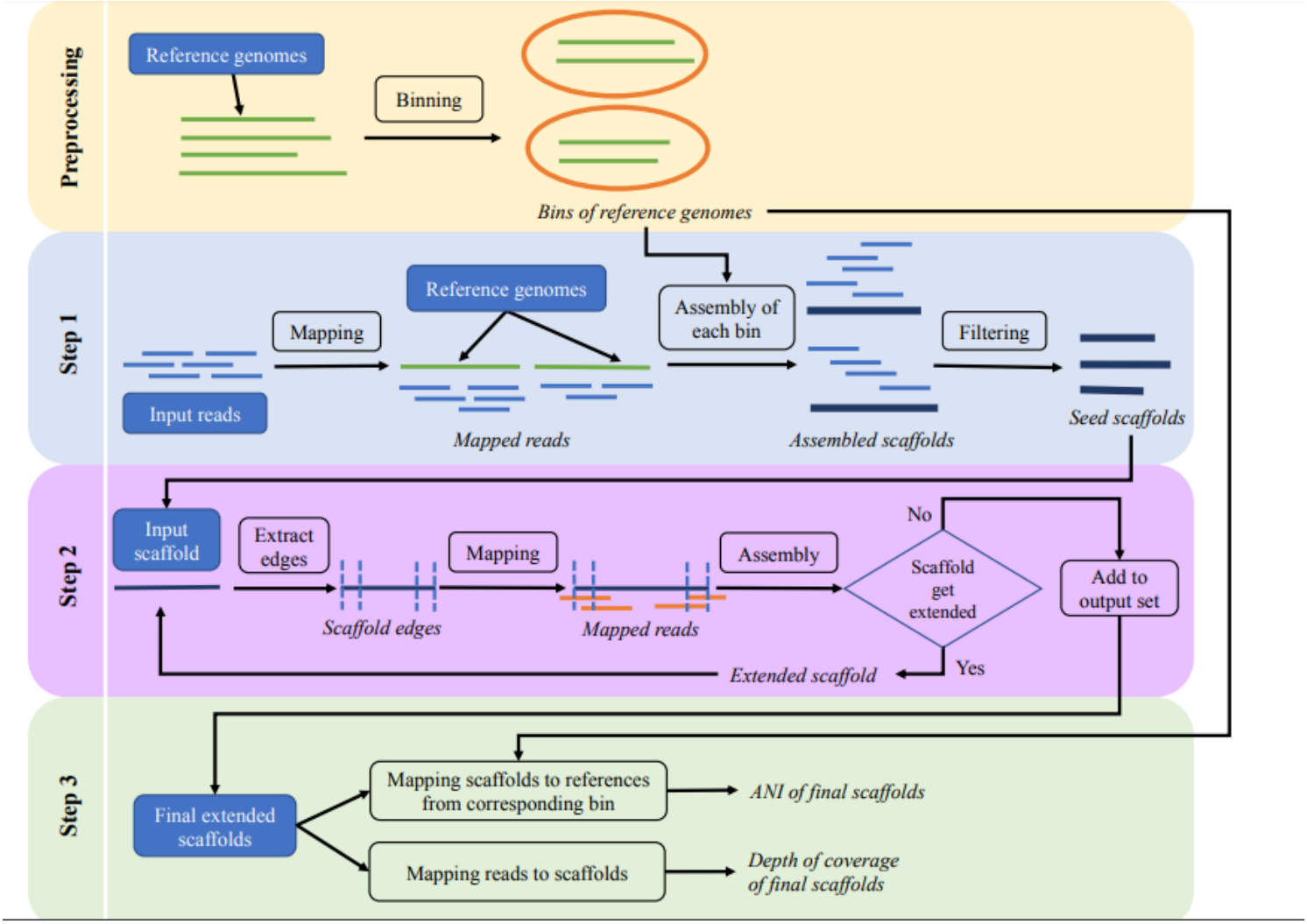

Figure 1

Overview of the FVE-novel pipeline, where the inputs are single-end or paired-end reads and reference database and the output are a set of final extended scaffolds along with ANI and depth of coverage of the output scaffolds.

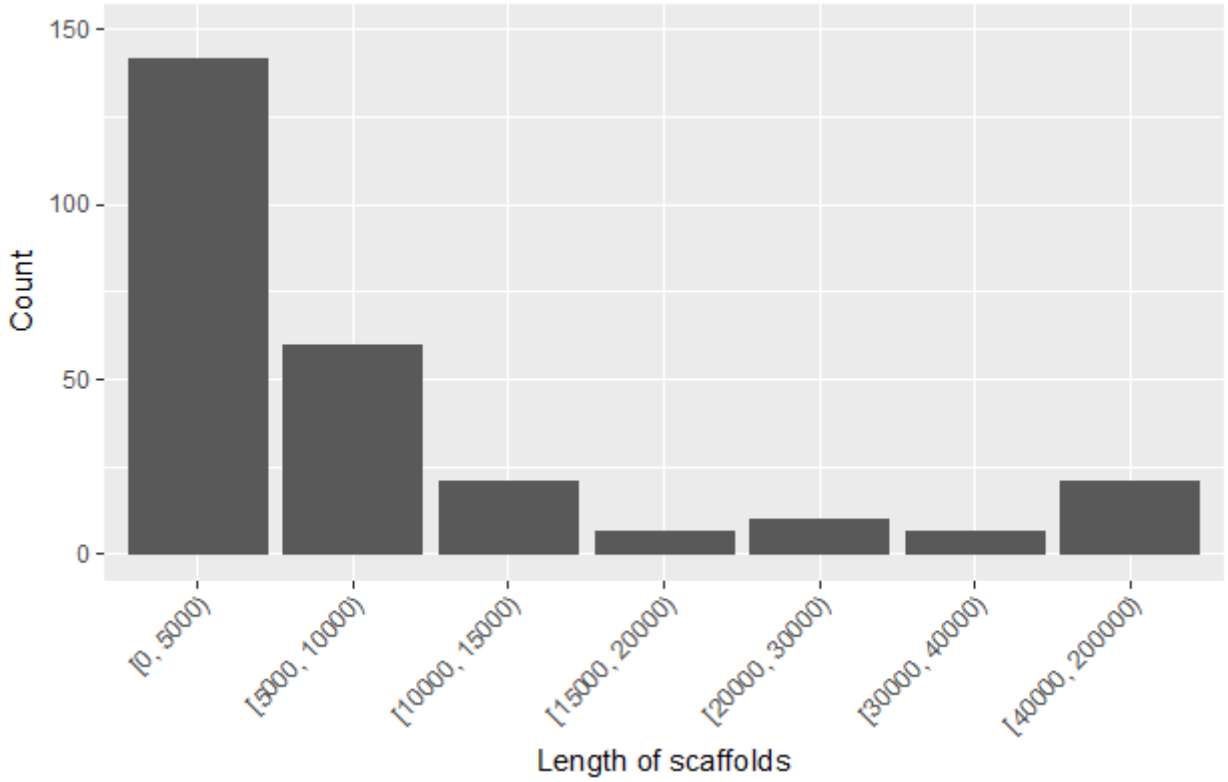

Figure 2

Length distribution of the 268 scaffolds generated by FVE-novel for the ocean metagenome sample.

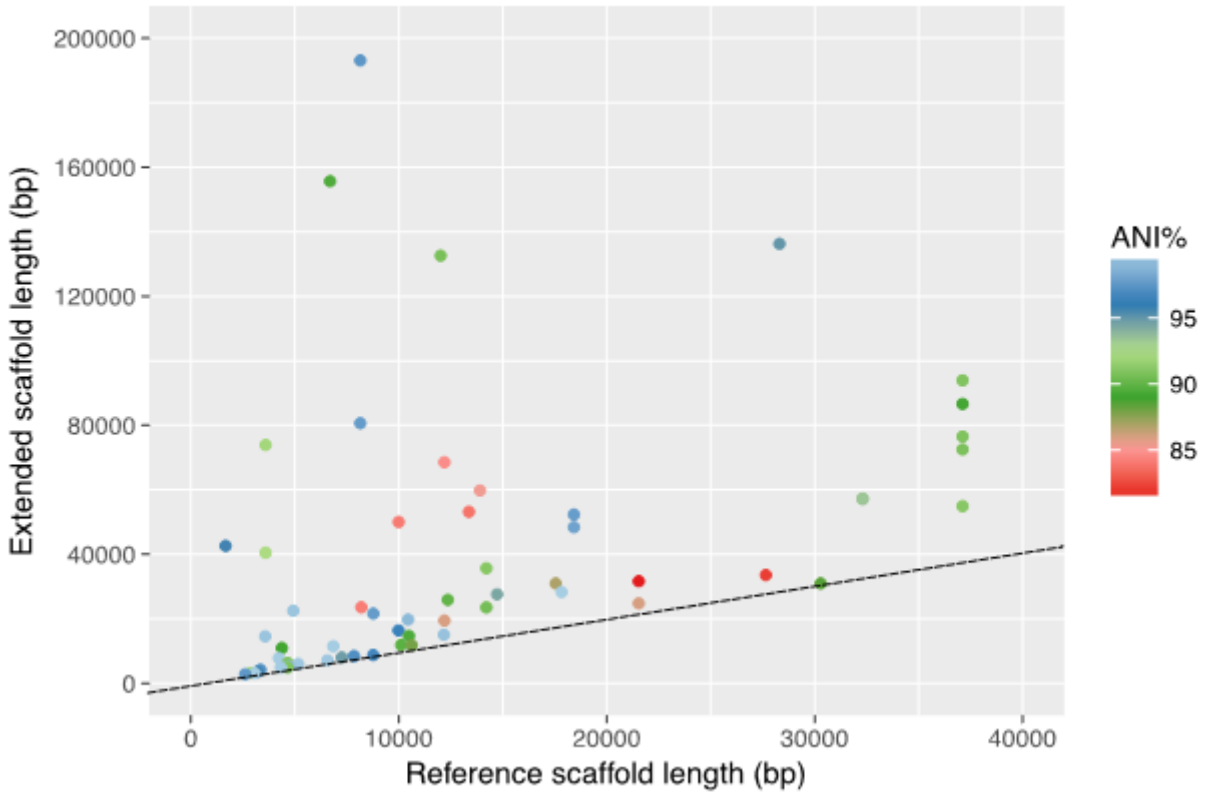

Figure 3

The length comparison of the 59 scaffolds against their corresponding references (GOV database). These 59 scaffolds are a subset of the total 268 scaffolds which are extended by FVE-novel tool. The dotted line represents 1:1 ratio between the x and y axis. The color of the dots represents the ANI between the scaffolds produced by FVE-novel and their references.

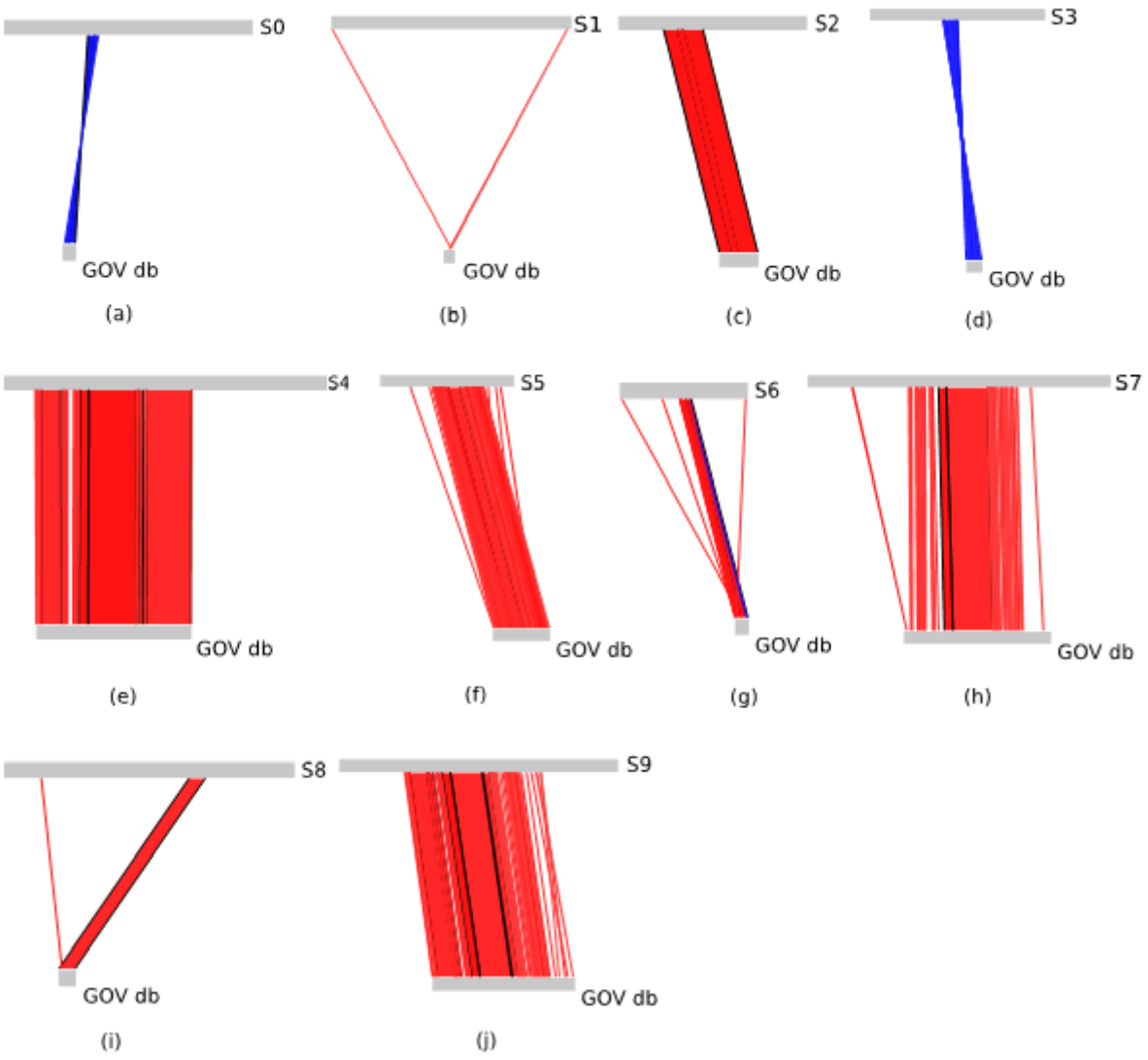

Figure 4

Alignment of the ten scaffolds to their corresponding references in GOV database. Here the red and blue lines represent the forward and reverse alignment respectively.

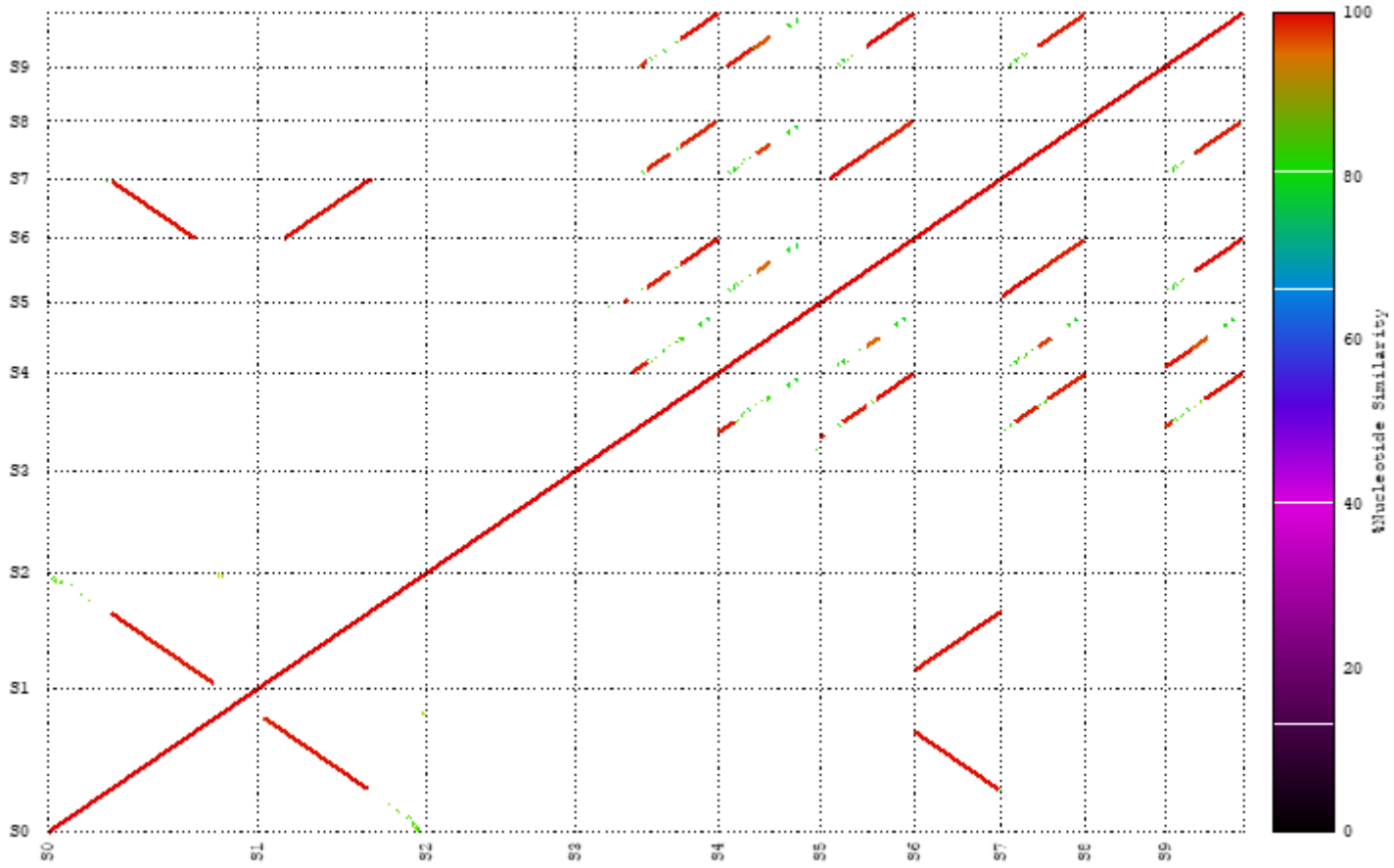

Figure 5

Percentage of similarity between each pair of the longest ten scaffolds of the 268 scaffolds generated by applying FVE-novel to the ocean metagenome sample.

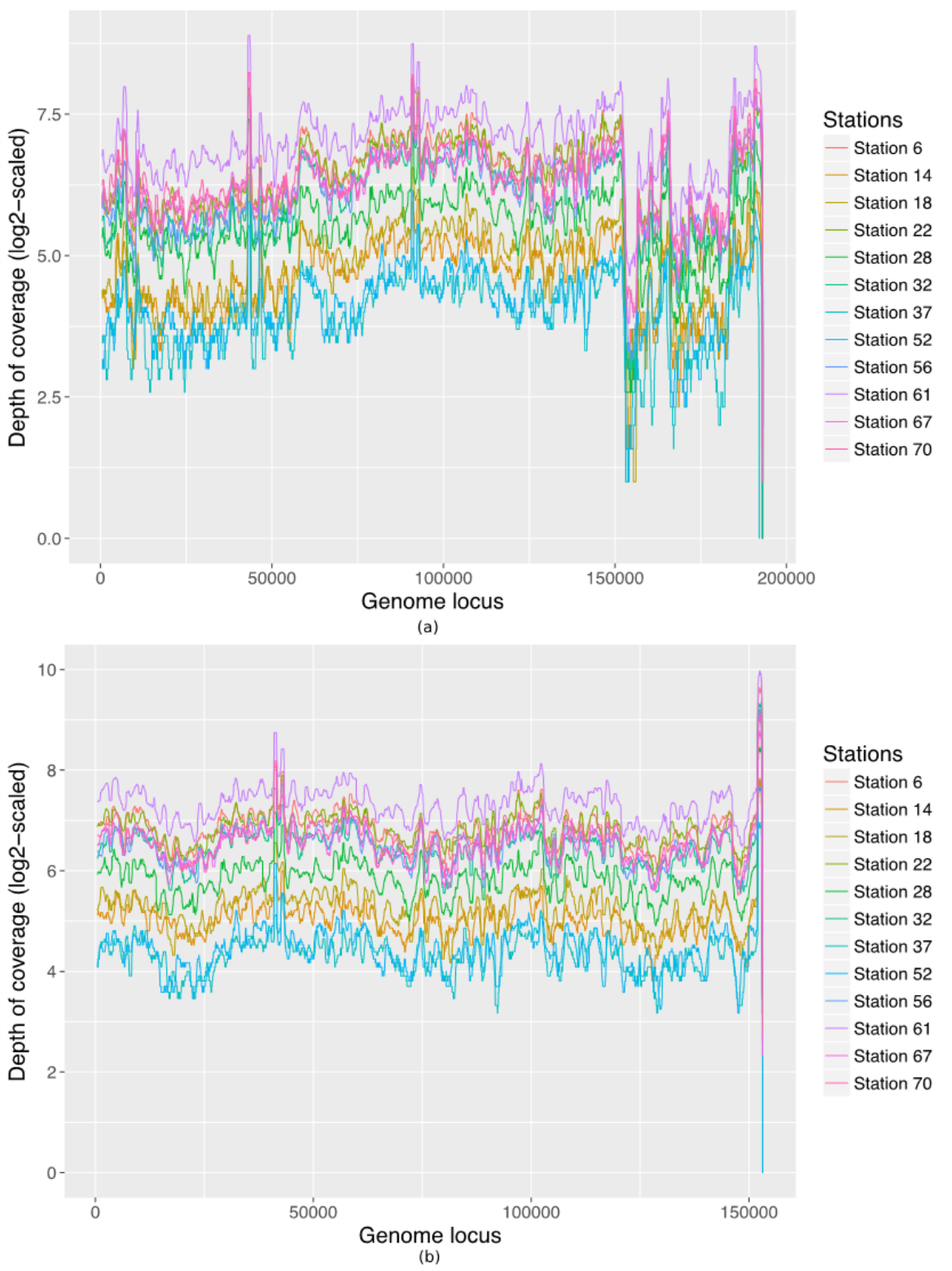

Figure 6

The log2-scaled depth of coverage of (a) S0 (193 kb) and (b) 153 kb scaffold representing the dominant strain of the novel virus (recovered from S0) across 12 ocean samples (showing median coverage of per 1000 bp window with step size 1).

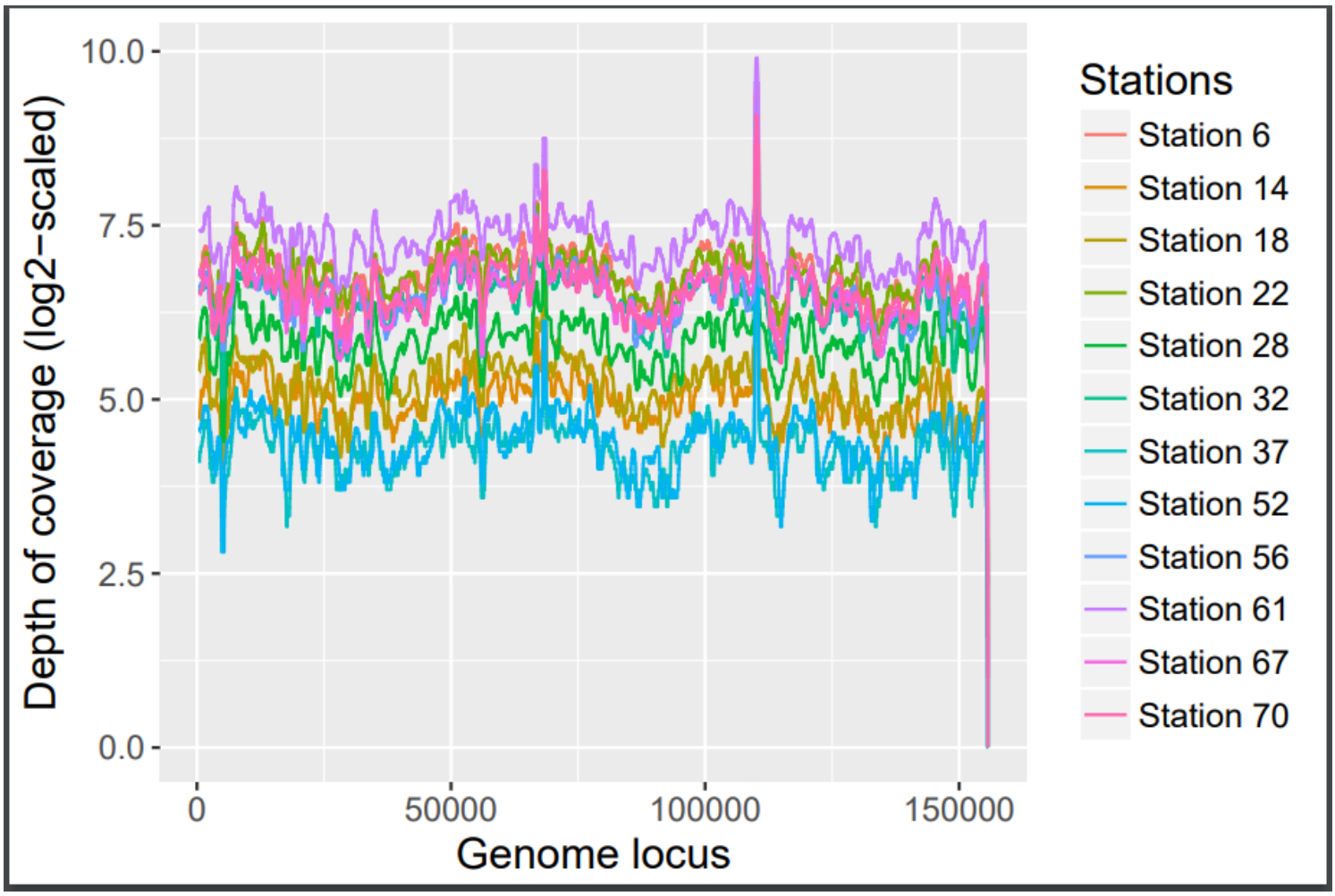

Figure 7

The log2-scaled depth of coverage of S1 (155 kb) across 12 ocean samples (showing median coverage of per 1000 bp window with step size 1).

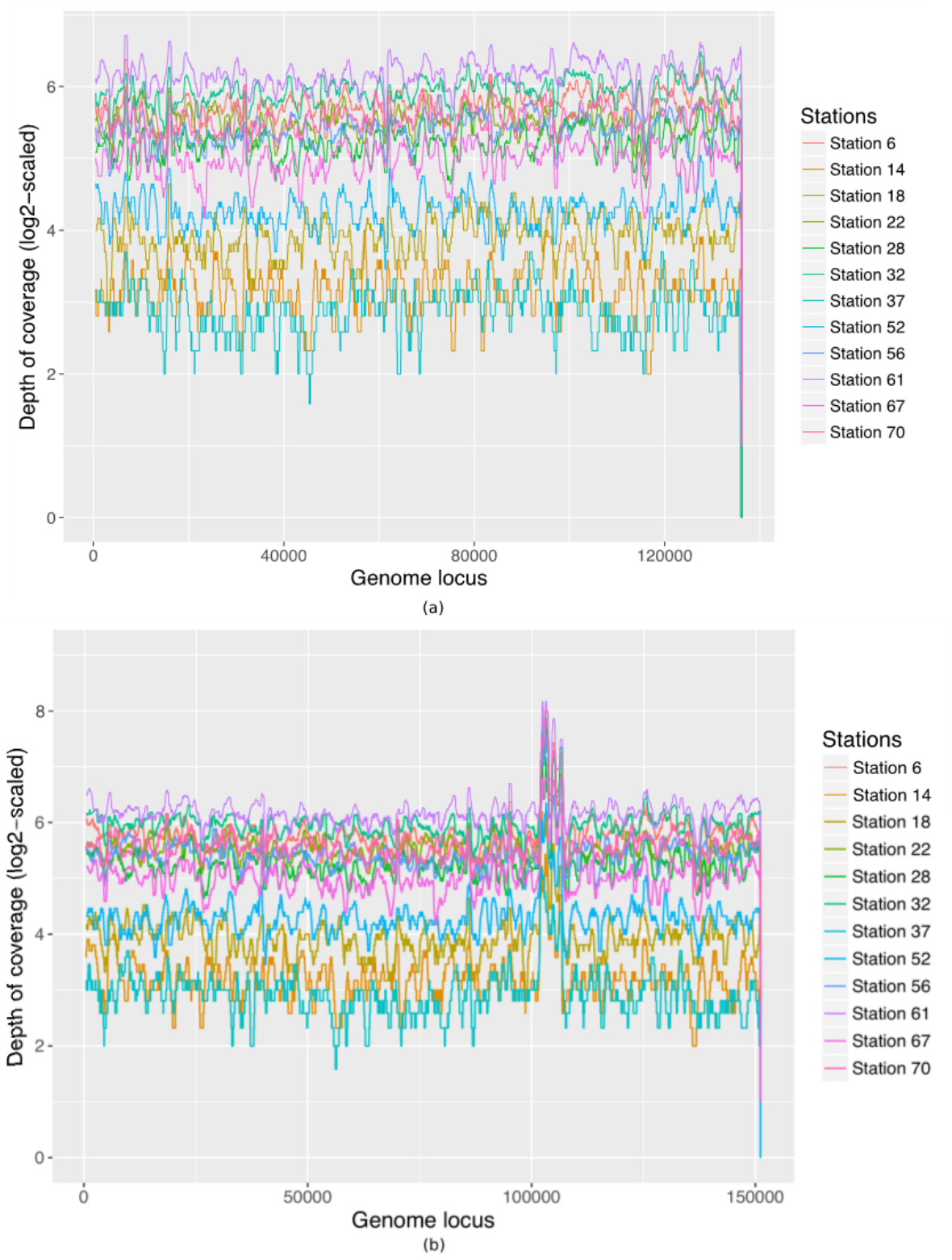

Figure 8

The log2-scaled depth of coverage of (a) S2 (136 kb) and (b) 151 kb scaffold representing the extended and complete version of S2 across 12 ocean samples (showing median coverage of per 1000 bp window with step size 1).

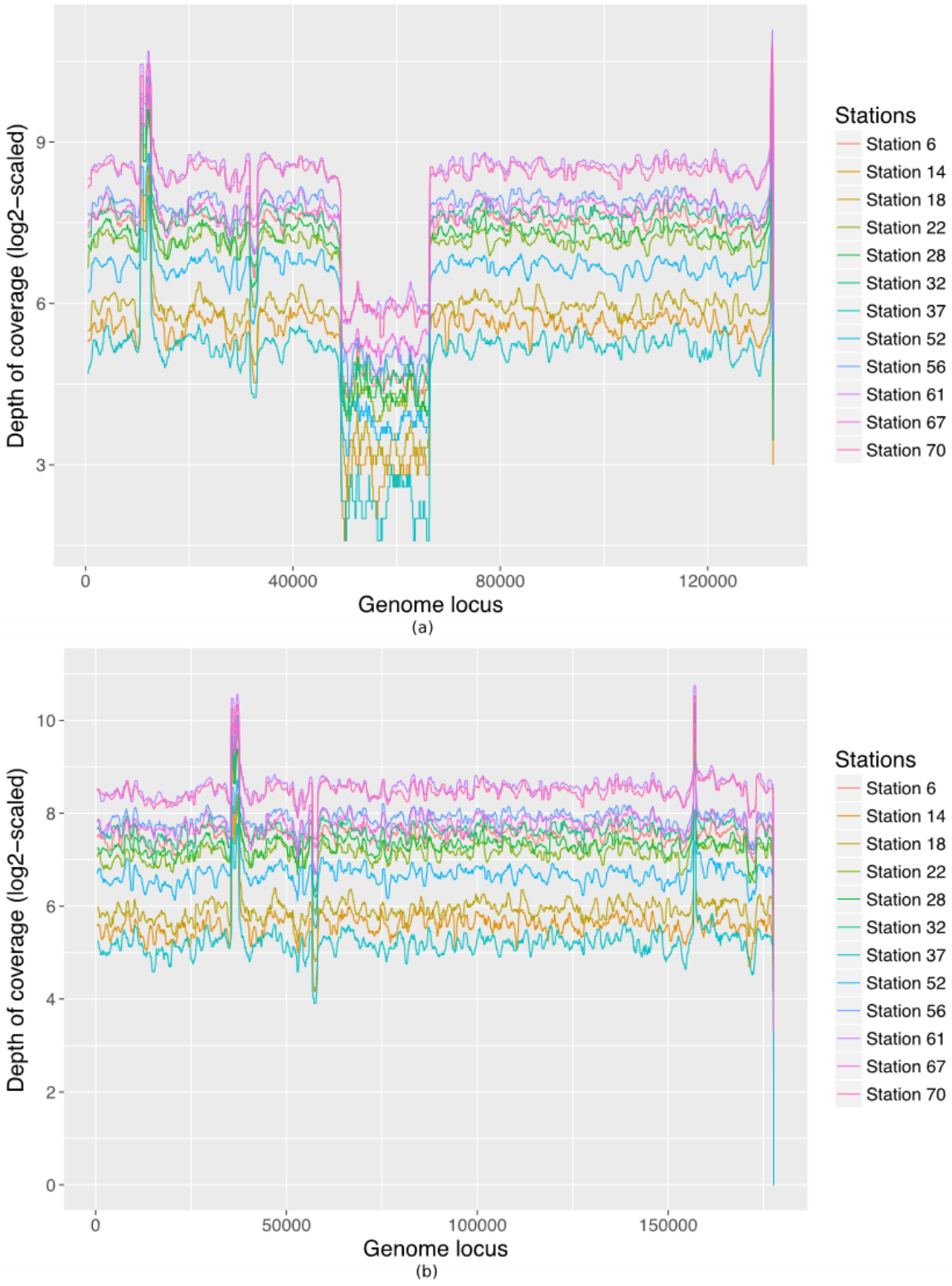

Figure 9

The log2-scaled depth of coverage of (a) S3 (133 kb) and (b) 177 kb scaffold representing the dominant strain of Prochlorococcus phage P-SSM4 (recovered from pieces of S3) across 12 ocean samples (showing median coverage of per 1000 bp window with step size 1).

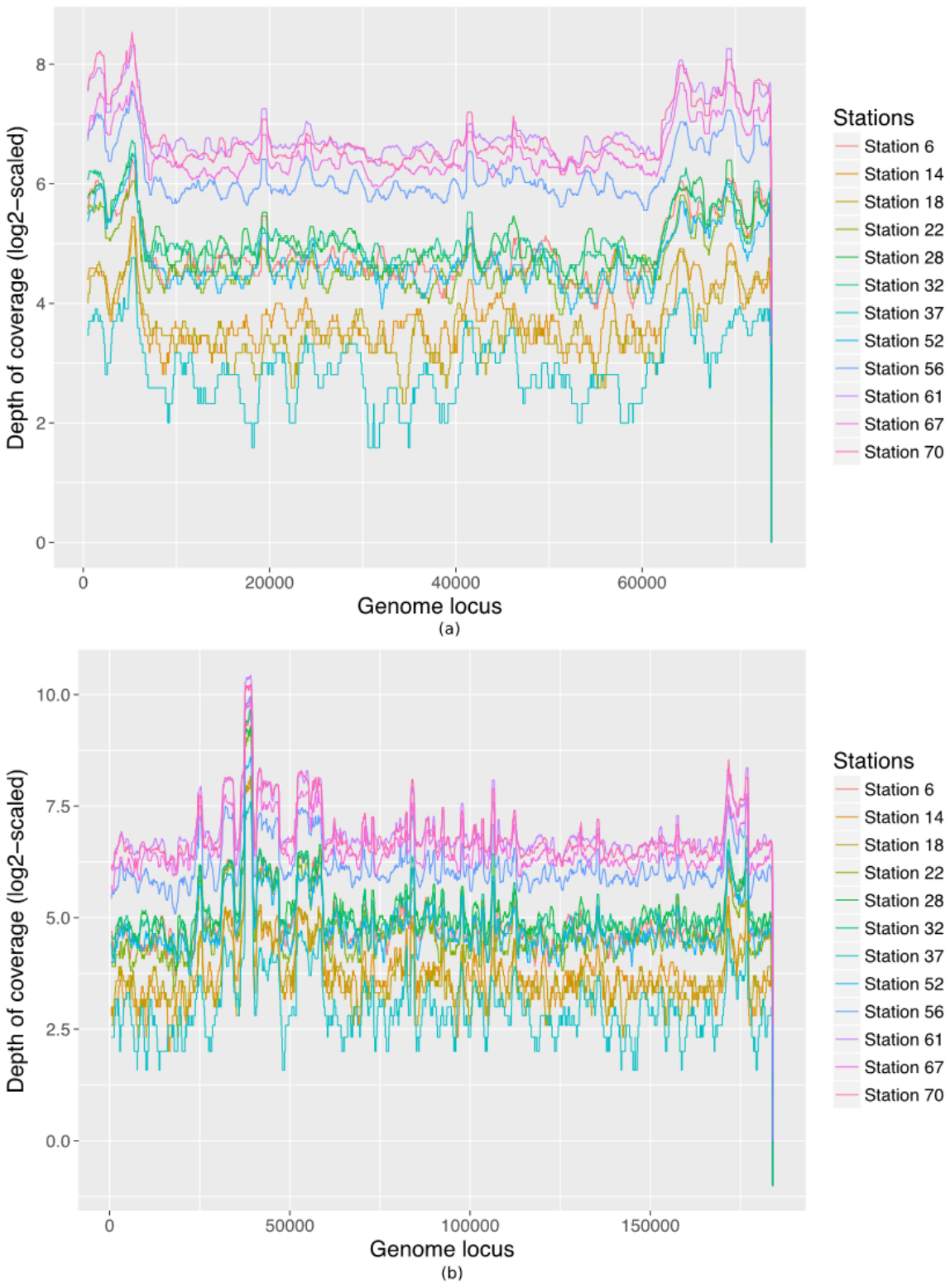

Figure 10

The log2-scaled depth of coverage of (a) S8 (73 kb) and (b) 183 kb scaffold representing the dominant strain of Prochlorococcus phage P-HM2 (recovered from pieces of S8) across 12 ocean samples (showing median coverage of per 1000 bp window with step size 1).

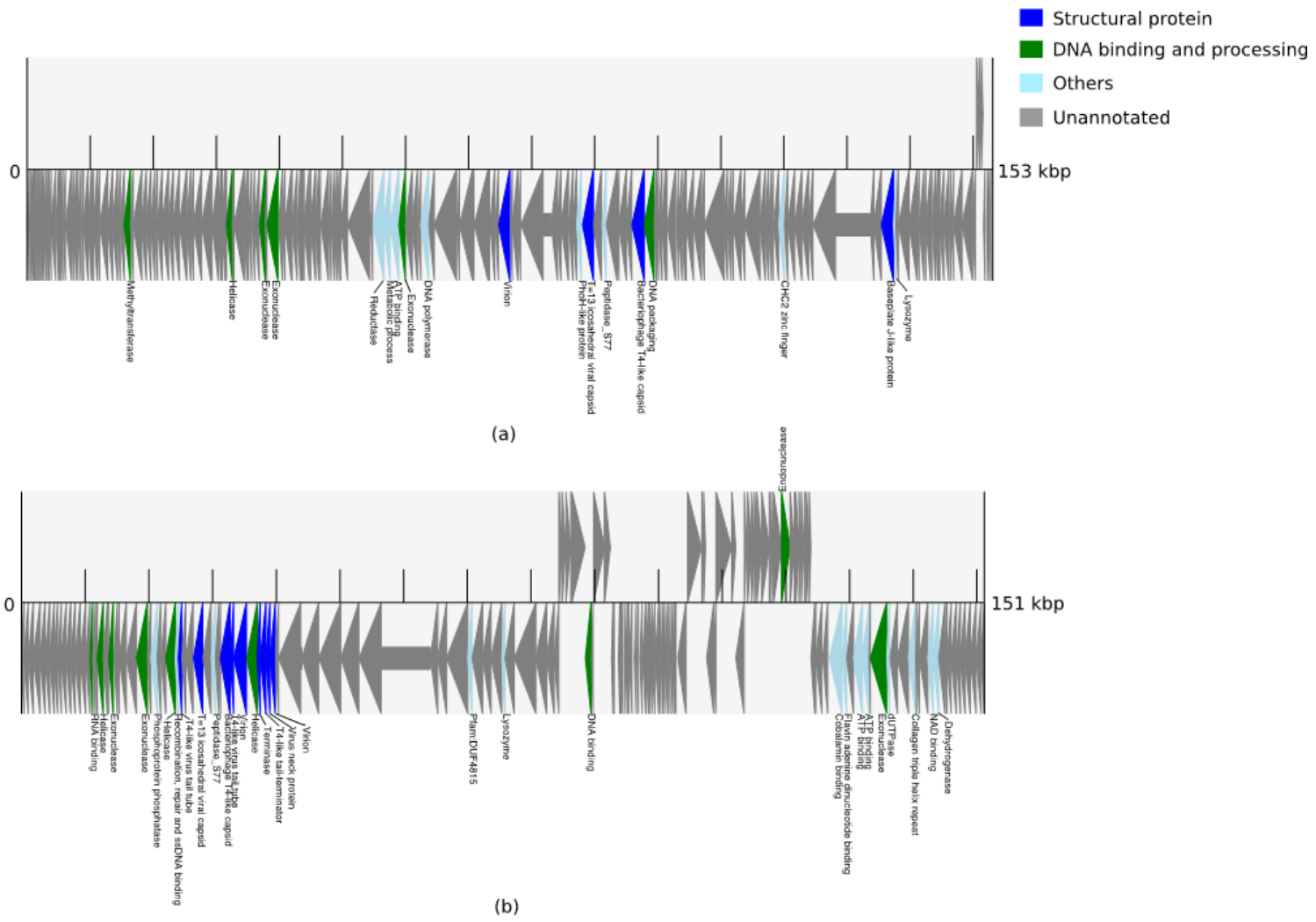

Figure 11

Protein annotation of the two novel scaffolds (a) 153 kb scaffold of dominant strain of a novel virus, (b) 151 kb scaffold of dominant strain of a novel virus.

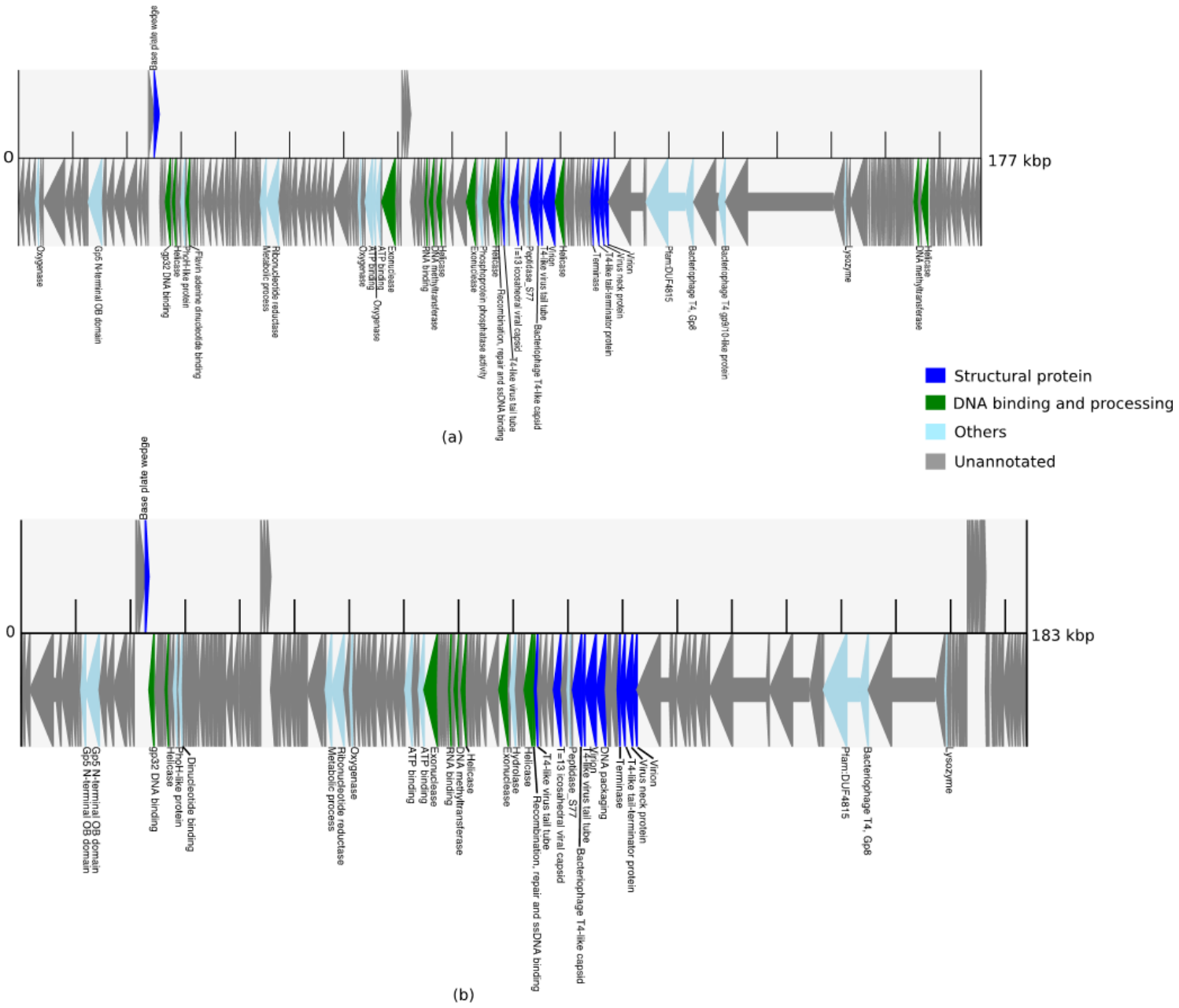

Figure 12

Protein annotation of the two strains of Prochlorococcus phage (a) 177 kb scaffold of dominant strain of Prochlorococcus phage P-SSM4 and (b) 183 kb scaffold of dominant strain of Prochlorococcus phage P-HM2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterialfinal.docx
- Table1.docx