frontiers in Microbiology

ORIGINAL RESEARCH
published: 23 January 2017
doi: 10.3389/fmicb.2017.00013

Virome Assembly and Annotation: A Surprise in the Namib Desert

Uljana Hesse[1,2], Peter van Heusden[2], Bronwyn M. Kirby[1], Israel Olonade[1], Leonardo J. van Zyl[1] and Marla Trindade[1]*

[1] *Institute for Microbial Biotechnology and Metagenomics, University of the Western Cape, Bellville, South Africa,* [2] *South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa*

Sequencing, assembly, and annotation of environmental virome samples is challenging. Methodological biases and differences in species abundance result in fragmentary read coverage; sequence reconstruction is further complicated by the mosaic nature of viral genomes. In this paper, we focus on biocomputational aspects of virome analysis, emphasizing latent pitfalls in sequence annotation. Using simulated viromes that mimic environmental data challenges we assessed the performance of five assemblers (CLC-Workbench, IDBA-UD, SPAdes, RayMeta, ABySS). Individual analyses of relevant scaffold length fractions revealed shortcomings of some programs in reconstruction of viral genomes with excessive read coverage (IDBA-UD, RayMeta), and in accurate assembly of scaffolds ≥50 kb (SPAdes, RayMeta, ABySS). The CLC-Workbench assembler performed best in terms of genome recovery (including highly covered genomes) and correct reconstruction of large scaffolds; and was used to assemble a virome from a copper rich site in the Namib Desert. We found that scaffold network analysis and cluster-specific read reassembly improved reconstruction of sequences with excessive read coverage, and that strict data filtering for non-viral sequences prior to downstream analyses was essential. In this study we describe novel viral genomes identified in the Namib Desert copper site virome. Taxonomic affiliations of diverse proteins in the dataset and phylogenetic analyses of circovirus-like proteins indicated links to the marine habitat. Considering additional evidence from this dataset we hypothesize that viruses may have been carried from the Atlantic Ocean into the Namib Desert by fog and wind, highlighting the impact of the extended environment on an investigated niche in metagenome studies.

Keywords: Namib Desert, virome, circovirus, assembly, annotation, simulated/virtual metagenomes

OPEN ACCESS

Edited by:
Akio Adachi,
Tokushima University, Japan

Reviewed by:
Hano Maree,
Agricultural Research Council, South Africa
Hilary G. Morrison,
Marine Biological Laboratory, USA
Patrick William Laffy,
Australian Institute of Marine Science, Australia

****Correspondence:***
Uljana Hesse
uljana@sanbi.ac.za

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *26 September 2016*
Accepted: *03 January 2017*
Published: *23 January 2017*

Citation:
Hesse U, van Heusden P, Kirby BM, Olonade I, van Zyl LJ and Trindade M (2017) Virome Assembly and Annotation: A Surprise in the Namib Desert. Front. Microbiol. 8:13. doi: 10.3389/fmicb.2017.00013

INTRODUCTION

Viruses are important players in their respective ecological niches. Not only do they drive host adaptation through horizontal gene transfer, they also regulate microbial abundance and diversity through predation (reviewed in Kimura et al., 2008). The latter effect is so extensive that they are now recognized as a major driving force in global biogeochemical nutrient cycles (Fuhrman, 1999). The advances in next generation sequencing technologies have boosted studies on viral diversity and functionality through metagenome and virome analysis. An increasing number of studies investigate virome data obtained from purified viral particles, following a workflow similar to the one established for marine viromes (Solonenko and Sullivan, 2013). In short, it involves

filtering and concentration of viral particles from the environmental sample, DNAse, and RNAse treatments to reduce genomic contamination from non-viral organisms, degradation of the protein coats, purification of the DNA, and sequencing of the viral genomes. To date, few studies have used this approach to investigate viromes derived from soils (reviewed in Zablocki et al., 2016). This medium is challenging, as it is rich in enzyme-inhibiting compounds that interfere with library construction (Fenner and Freeman, 2011; Carvalhais et al., 2012; Tveit et al., 2014). To minimize the effects of such inhibitors, additional purification steps, and whole genome amplification (WGA) are sometimes unavoidable (e.g., Zablocki et al., 2014; Adriaenssens et al., 2016). Yet, such methods can introduce biases, such as DNA fragmentation and selective amplification of DNA (Kim and Bae, 2011; Yuan et al., 2012; van Dijk et al., 2014), which are then reflected in the read datasets.

Assembly of reads into longer sequences facilitates identification of coding genes, as well as taxonomic and functional annotation. However, correct reassembly of virome data is challenging. Highly variable read coverage between and across the diverse viral genomes may lead to missing sequence information and fragmented genome recovery (García-López et al., 2015). The mosaic structure of viral genomes caused by recurrent horizontal exchange of genetic material (reviewed in Hatfull, 2008) further complicates the matter. Several metagenome assemblers that account for varying sequencing depths in the data have been developed (e.g., Peng et al., 2011; Bankevich et al., 2012; Boisvert et al., 2012). Their suitability for virome assembly has been assessed in recent studies (de Cárcer et al., 2014; Vázquez-Castellanos et al., 2014; García-López et al., 2015; Laffy et al., 2016). This was effectively done using simulated data: virtual viromes consisting of known viral genomes at varying abundance levels serve to generate synthetic reads, which are reassembled into contigs and scaffolds. The reassembled sequences can then be mapped back to the original viral genomes, providing information on assembly precision. Based on these studies, the programs IDBA-UD, CLC, RayMeta, and SPAdes produced the most accurate assemblies. However, these assemblers have not been tested for their ability to correctly reconstruct fragmented DNA and genomic sequences with excessive read coverage, as may occur during DNA purification procedures and WGA.

Currently, virome annotation is often conducted using only the viral subset of the RefSeq database (e.g., Minot et al., 2013; Carlos et al., 2014; Weynberg et al., 2014; Zablocki et al., 2014; Hannigan et al., 2015; Mohiuddin and Schellhorn, 2015). This can be achieved through direct BLAST analyses or using the convenient and informative online annotation server MetaVir (Roux et al., 2011). Other programs for virome annotation include VMGAP (Lorenzi et al., 2011), VIROME (Wommack et al., 2012), and MG-Rast (Meyer et al., 2008); however, they have been found less effective in identification of viral strains and taxa as compared to MetaVir or direct BLAST analysis (Tangherlini et al., 2016). One major drawback of this approach is that it does not discriminate between viral and non-viral DNA: any contaminating non-viral sequence that shows significant similarity to a viral gene will be annotated as viral. MetaVir does not filter for non-viral sequences, but relies on the user to perform such analyses. Diverse methods (e.g., filtering for 16S rRNA and tRNA) and tools (e.g., Schmieder and Edwards, 2011; Roux et al., 2015) for prescreening virome data have been applied. The recently published virome annotation program HoloVir (Laffy et al., 2016), addresses this problem by prescreening the sequences against the entire SILVA database and an extensive database with viral and cellular marker genes. However, comparative analyses assessing the efficiency of these procedures are still outstanding.

Despite severe environmental constraints, deserts host a multitude of animals, plants, and microorganisms, and are estimated to store almost one third of the global terrestrial carbon (Pointing and Belnap, 2012; Makhalanyane, 2015). Virus research in hot desert soils using classical techniques (Prigent et al., 2005; Prestel et al., 2008, 2012) and metagenomic approaches (Fierer et al., 2007; Adriaenssens et al., 2015; Vikram et al., 2015; Adriaenssens et al., 2016), revealed a wide diversity of viral taxa with a prevalence of tailed viruses (Caudovirales), and described a range of novel viral proteins. Yet, the extent of extracellular viral diversity and ecological functions of viruses in these arid ecosystems remain severely understudied. The Namib Desert is described as one of the oldest deserts in the world, deemed semi-arid for an estimated 55–80 million years (Goudie, 2010). Large areas are currently considered hyper-arid with scant, erratic and low precipitation inputs. Surface temperatures are subject to extreme seasonal fluctuation, reaching up to 70°C for short periods of time. High levels of UV radiation, physical disturbance and low nutrient status further accentuate the severe conditions (Nicholson, 2011; Eckardt et al., 2013b). The surface geomorphology and mineralogy in the Namib Desert also varies widely, with reported surface outcrops of uranium and copper (Eckardt et al., 2013b). These create environmental pockets where microorganisms and their respective viral pathogens may have adapted to poly-extreme conditions. Studies have shown that ore microbes attach to ore particles, exhibiting a general tolerance to high concentrations of metallic and other ions (Dopson et al., 2003; Rawlings, 2005). The dynamic, poly-extreme conditions of desert ore sites make them particularly attractive environments for studies on viral ecology and evolution. Viral communities associated with desert ores have not been assessed before, and their adaptation to such an environment could be exploited in numerous ways for biotechnological application.

In this paper we investigate viral diversity at a copper deposit in the Namib Desert and describe novel viral genomes. Furthermore, we assess effectiveness and accuracy of five assemblers using simulated datasets that mimic environmental data challenges (DNA fractionation and excessive coverage of selected genomes), and address presence of non-viral DNA in the copper soil virome and other datasets.

MATERIALS AND METHODS

Simulated Virome Data Analyses

Initial assembly tests of the Namib Desert copper site data indicated excessive read coverage for small circovirus-like genomes (∼2 kb) and fragmental representation of larger viral

genomes. To investigate which assembly program excels in accuracy and genome recovery given such data, we generated three virtual viromes (Sim1, Sim2, and Sim3). Sim1 consisted of 572 complete viral genomes at relative abundances as published by Vázquez-Castellanos et al. (2014). Sim2 (testing fractionation effects) consisted of non-overlapping 2 kb sequences from these viral genomes, and Sim3 (testing excessive read coverage for small circovirus-like genomes) consisted of the above 572 complete viral genomes at original abundance levels plus five published circoviral genomes with exceptionally high abundance levels. The simulated read datasets were produced using MetaSim (Richter et al., 2008). Since a 250 nt error model was not available, we generated datasets of read length 80 nt (errormodel-80bp.mconf; MetaSim homepage) and read length 300 nt (errormodel-300nt.mconf; http://seqanswers.com/forums/showthread.php?t=44676). Using the empirical error model and an insert size of 450 nt (10% std), we produced 4.5 Mio read pairs with Illumina-specific errors (error rate 0.9 and 1% for 80 and 300 nt, respectively) for each of the six datasets, respectively.

The six virome datasets were assembled using five assemblers. IDBA-UD (Peng et al., 2012) and SPAdes (Bankevich et al., 2012) conduct assembly iterating through several kmers. With these assemblers, default settings were used with the 80 nt read datasets. With the 300 nt datasets additional kmers (IDBA-UD: kmer 40, 80, 120; SPAdes: default, 79, 119) were used. Assemblies with the CLC-Genomic-Workbench (version 7.5, http://www.clcbio.com) were conducted using automated kmer length (resulting in kmer 23), and specifying 95% read coverage and 95% nucleotide identity for read mapping. ABySS (Simpson et al., 2009) and RayMeta (Boisvert et al., 2012) were run with default assembly parameters and a kmer length of 23.

Assembly quality was assessed by (1) aligning all contigs and scaffolds larger than 500 nt from the simulated datasets against their respective genomes using MUMmer v-3 (Kurtz et al., 2004), (2) mapping all reads back to the assembled contigs using Bowtie2 v-2.2.6 (Langmead and Salzberg, 2012), selecting end-to-end matches (–no-1 mm-upfront switched on) and reporting multiple alignments (-k 3); and (3) running ALE (Clark et al., 2013) on all assembled datasets to compare the performance of the different assemblers. Genomes were considered "recovered" if they were reassembled into one sequence.

Virus Isolation, Concentration, DNA Extraction, and Sequencing

This study concentrates on a specific environmental niche: isolated heaps of copper ore material found in the vicinity of former copper mines in the Namib Desert. Sample processing was conducted similar to previously published methodologies (Adriaenssens et al., 2015). Two 25 L carboys were each half filled with rocks and soil from the sample site (4 × 2.5 m; 23°33′35.27″S; 15°16′50.63″E; **Figure S1**), then completely filled with distilled water and shaken for 20 min. The resulting slurry was left to settle for an hour and the supernatant was decanted into a clean carboy. The drums with sediment were again filled with distilled water and shaken, left to settle for an hour, and the supernatant decanted into the same collection carboy. The supernatant was filtered through a Millipore 1 μm Polygard-CR filter (CR0101006) using a Millipore peristaltic pump (XX80EL230) and the filtrate collected in a clean carboy. The 1 μm filtrate was filtered through a 0.22 μm Opticap® XL10 Durapore® filter (KVGLA10HH1) and the filtrate collected in a clean carboy. The 0.22 μm filtrate was subjected to tangential flow filtration (TFF) using a Millipore TFF cartridge holder (XX42PS001) and 30 kDa cut-off filter (CDUF001TT). The ±50L of 0.22 μm filtrate was concentrated to ±140 ml (360 × concentrated) at the Gobabeb research center, and the viral fraction stored at 4°C for 2 weeks prior to DNA extraction in the laboratory at the University of the Western Cape. Virus and virus-like particles were collected for DNA extraction by centrifuging 50 ml of the concentrate at 25.000 × g for 6 h using a Beckman Avanti J-26 XPI centrifuge in a JA20 rotor. The pellet was resuspended in 200 μl TE buffer. This viral suspension was treated with DNaseI (EN0521) and RNaseA (EN0531) (Fermentas—final concentration of 0.1 μg/ml) at 37°C for 1 h. We tested for presence of bacterial DNA by amplifying the 16S rRNA gene using the primers E9F and U1510R (Hansen et al., 1998; Baker et al., 2003) as follows: 1 μl of genomic DNA was mixed with 2.5 μl of each primer (10 mM), 2.5 μl of 2 μM dNTPs, 2.5 μl of 10X DreamTaq buffer (ThermoFisher Scientific, MA, USA), 1 μl BSA 10 mg/ml, 0.125 μl DreamTaq polymerase (ThermoFisher Scientific, MA, USA) and milliQ water to a total volume of 25 μl. PCR was conducted under the following thermal regime: 95°C for 5 min, 95°C for 30 s, 52°C for 30 s, 72°C for 85 s (30 cycles), and 10 min at 72°C. The sample was deemed free of bacterial DNA when 16S rDNA could be amplified only from the positive controls, but not from the sample or the negative controls. Then, the virus particles were treated with Proteinase K (Fermentas—final concentration 1 μg/ml) at 55°C for 2 h. Thirty microliters of 20% SDS were added and the sample incubated at 37°C for 1 h. The DNA was extracted with three phenol:chloroform:isoamylalcohol (25:24:1) extractions, followed by two extractions using chloroform:isoamylalcohol (24:1). Phase separation was achieved by centrifugation in an Eppendorf 5810R centrifuge at 5000 RPM for 10 min. Precipitation of the DNA was performed through addition of 1/10 volume of 3 M NaOAc (pH 5.2) and 2 volumes 95% ethanol, with overnight incubation at 4°C. Precipitated DNA was recovered by centrifugation at 13,000 RPM for 10 min and the resulting pellet was resuspended in 30 μl of TE buffer. DNA was purified using the QIAamp DNA stool mini kit (cat. no. 51504) using half of an Inhibit EX tablet (provided with this kit) per purification. Then, 10 ng of the extracted, purified DNA was used to perform WGA (GE Healthcare GenomiPhi HY DNA amplification kit cat. no. 25-6600-20) using the manufacturer's recommendations. The resulting DNA was purified using the Qiagen Gel Extraction kit (Qiaex II, cat. no. 20021). Throughout the extraction and purification process the sample was assessed for the presence of polymerase inhibitors via PCR. Using the above primers, a 16S rRNA gene PCR reaction of genomic *E. coli* DNA was spiked with ∼1 ng extracted metagenomic DNA and the level of amplification was compared to an unspiked control reaction containing only genomic DNA. The sequence library

was prepared with the Illumina Nextera XT library prep kit with minor modifications. The amount of input DNA was decreased to 0.8 ng and 1U Phusion polymerase (Thermo Scientific, cat no. F-530S) was included in the tagmentation reaction. The amplified DNA was sequenced with a MiSeq Reagent V2 kit (2 × 250 nt reads) on the Illumina MiSeq Sequencing platform and included a 20% PhiX V3 spike as per manufacturer's instructions (Preparation guide, Part #15031942, May 2012 revision).

Transmission Electron Microscopy of Viruses and Virus Like Particles

Virus particle suspensions were prepared according to the ammonium acetate method as described by Ackermann (2009). Three microliters taken from the 140 ml concentrate were pipetted onto carbon coated 200 mesh copper grids and stained with 2% aqueous uranyl acetate for 30 s. The samples were viewed using a LEO 912 Omega TEM (Zeiss, Oberkochen, Germany) at 120 kV. Images were collected using a ProScan CCD camera.

Copper Site Data Assembly

The raw reads were filtered for duplicates, sequences matching the Nextera XT adapters, and transposase sequences, and were then quality trimmed using CLC (quality limit 0.05, ambiguous limit 3, adapter trimming, minimum read length 50). PhiX reads not removed by the Illumina MiSeq reporter software (version 3) or through duplicate removal were filtered by mapping all reads to PhiX-174 using RAMICS (Wright and Travers, 2014). Similarly, reads matching the human genome (Hg19; http://tinyurl.com/jay436s) were filtered using consecutively Bowtie2 and RAMICS. This filtered sequence data has been submitted to the European Nucleotide Archive (ENA) under accession number PRJEB13486. The assembly workflow is visualized in **Figure S2**. Reads were assembled using CLC with stringent assembly settings (automatic word size equaling 22, mismatch cost 2, insertion cost 3, deletion cost 3, min contig length 200, with length and similarity fraction for read mapping equaling 0.95%, both), to generate a primary assembly. Our analyses of the simulated datasets had shown that CLC reconstructs excessively covered genomes into multiple copies and subsequences of nearly 100% sequence identity. We identified such incomplete assemblies through network analysis: all sequences of the primary assembly were aligned against each other using MUMmer and sequence names of identical sequence pairs were clustered using network analysis (in-house R-script). With this approach we found one cluster of 286 contigs. Read mapping using RAMICS showed that these contigs combined 3.6 M reads. We also mapped the remaining reads against another 11 contigs that had an average read coverage above 5000 and were represented by single copies. We then separately assembled all unmapped reads (sub-assembly 1) and the reads that had mapped to the contig cluster (sub-assembly 2). The sub-assembly 1 was then filtered once more for human sequences using BLASTn (BLAST+ suite: Camacho et al., 2009) against the human genome. The final assembled dataset was limited to sequences ≥300 nt and contained 20,097 contigs and scaffolds from subassembly 1, one contig from subassembly 2, and 8 of the highly covered contigs from the primary assembly.

Copper Site Data Annotation

The annotation workflow for the copper site protein dataset is shown in **Figure S3**. First, all contigs and scaffolds ≥300 nt were submitted to MetaVir (Roux et al., 2011) for gene prediction and taxonomic classification. MetaVir annotates a protein as "viral" if it matches a protein from the viral subset of the RefSeq protein database with a bit-score above 50. All protein sequences predicted by MetaVir were also compared against the non-redundant protein database (NCBI-nr) and the Refseq protein database (RefSeq-P) from NCBI using BLASTp (BLAST+). The BLAST+ algorithms were executed to provide a tabular output with information on the subject title and subject taxonomy with a preliminary *e*-value threshold of 0.1e-3, and the top hit was used for protein annotation. We then compared the three protein annotations (MetaVir, NCBI-nr, RefSeq-P) and selected the highest scoring annotation for "score-based taxonomic kingdom assignment." The "final taxonomic kingdom assignment" was "Viruses" if (a) the MetaVir annotation had the highest score (in this case, the respective protein annotations were used to construct a "viral keyword" dictionary); (b) the NCBI-nr and/or RefSeq-P annotations had the highest score, taxonomically classified the protein to "Viruses," and the *e*-value was below 1.0e-04 (min bit-score 45); (c) the NCBI-nr and/or RefSeq-P annotations had the highest score, the subject title contained a "viral keyword," and the *e*-value was below 1.0e-04 (min bit-score 45); or (d) the Pfam annotation indicated viral taxonomy. The "final taxonomic kingdom assignment" was "putative Viruses" if the NCBI-nr and/or RefSeq-P annotations had the highest score, indicated viral taxonomy based on taxonomic classification or contained a "viral keyword," and the *e*-value was above 1.0e-04. Alternatively, proteins were classified as "archaea," "bacteria," or "eukaryota" (min *e*-value 1.0e-04) based on NCBI-nr and/or RefSeq-P annotations or remained unclassified. The results were extensively verified through manual annotation (Vm flag). To determine the lowest common ancestry (LCA) affiliation for the viral proteins (V/Vm), we repeated BLASTp against the RefSeq viral protein database and analyzed the results using MEGAN5 (Huson et al., 2007) with min score 45, max expected 1.0e-4, top percent 10, min support 1, and LCA percent 100. The taxonomic IDs provided by MEGAN5 were submitted to phyloT (http://phylot.biobyte.de/index.html) and the resulting tree was visualized using iTOL (Letunic and Bork, 2007). In addition, all contigs and scaffolds ≥300 nt were compared against the non-redundant nucleotide (NCBI-nt) and the RefSeq genomes (RefSeq-G) databases from NCBI using BLASTn (BLAST+). Protein annotations of the "final taxonomic kingdom assignments" were collated by contig and then compared with the corresponding BLASTn annotations to obtain indications on contig taxonomy. For genome comparisons, conducted with *Thalassotalea loyana* phage BA3 and contig_13, we used Easyfig v2.1 (Sullivan et al., 2011).

Screening Copper Site Data for Cellular Contamination

The SILVA_128_LSUParc and the SILVA_128_SSUParc fasta files (rRNA genes for large and small subunits, SILVA version 128)

were obtained from http://ftp.arb-silva.de/release_128/Exports. Both, the 13.9 M raw reads, as well as the 6.9 M quality filtered reads were mapped to the SILVA database using BLASTn (BLAST+) using a min bit-score threshold of 80. The extensive dataset of viral and cellular marker proteins (Laffy et al., 2016), was downloaded from XXX, and all proteins predicted on the final copper site assembly were mapped using BLASTp (BLAST+) with a maximum *e*-value threshold of 1.0e-10.

Phylogenetic Analyses

For phylogenetic analyses of putative circovirus-like replication-associated (pRepAs) and capsid (pCAPs) proteins, we used MAFFT (Katoh and Standley, 2013) to align the selected protein sequences from our dataset with all matching sequences from the NCBI-nr and RefSeq-P databases (max *e*-value 10). If those proteins originated from metagenomic studies, we used MAFFT to generate preliminary alignments and neighbor-joining trees with all circovirus-like pRepAs and pCAPs from the respective studies (downloaded from UniProt), retaining those protein sequences that showed informative sequence similarities to our proteins. The final alignments were generated using MAFFT, the final phylogenetic trees were generated using the RAxML BlackBox web server (Stamatakis et al., 2008) with default settings and Maximum Likelihood search for best scoring tree after bootstrapping (100) turned on. Figtree was used for tree visualization (Rambaut, 2008).

Viral Host Identification Assay

The copper site virome assembly indicated presence of a *T. loyana* BA3-like virus in the environmental sample. To verify this, we conducted PCR with the primers TH5For (5′-AGGCGC TAACCTGTGGTCAC-3′) and TH5Rev (5′-CGTTCATGTGTG GCGCTACA-3′), expected to amplify a 5 kb region from the corresponding contig_13. We prepared 50 μl reactions using 200 ng of template DNA, 0.5 μM of the primers, 1X Phusion buffer, 200 μM dNTPs, and 0.02 U/μl Phusion DNA polymerase from Thermo Scientific; and used the following cycling conditions: 1 cycle of 98°C for 3 min; 34 cycles of 98°C for 10 s, 58°C for 30 s, and 72°C for 1 min; followed by 72°C for 10 min and refrigeration at 4°C.

Potential hosts for the *T. loyana* BA3-like virus were investigated by challenging known species of *Thalassomonas/Thalassotalea* with the environmental virus fraction and by PCR screening. *T. loyana* (ID: 280483), *Thalassomonas viridans* XOM25 (ID: 137584), *Thalassotalea agariperforans* (ID: 864068), *Thalassotalea ganghwensis* (ID: 221989), *Thalassotalea agarivorans* (ID: 349064), *Thalassomonas actiniarum* (ID: 485447), and *Thalassomonas haliotis* (ID: 485448) were tested. Each strain was assayed for virus infection using the standard soft agar overlay technique. Briefly, overnight *Thalassomonas/Thalassotalea* cultures were used to inoculate 50 ml of fresh marine broth (1% v/v) and grown until O.D. 600 reached 0.4. Growth temperatures ranged from 20 to 35°C depending on the strain (Yi et al., 2004; Jean et al., 2006; Thompson et al., 2006; Hosoya et al., 2009). Then, 200 μl of the resulting culture were added to 1.5 ml tubes containing 100 μl of serially diluted environmental TFF virus fraction and incubated at room temperature for 10 min. The mixture was added to 3 ml broth containing 0.5% bacteriological agar and spread evenly on agar plates. The plates were checked for the presence of plaques every 24 h. To detect possibly low levels of viral infection, total DNA from virus challenged bacterial cells was used to amplify a region of the large terminase subunit (50 μl reactions using 200 ng of template DNA, 0.5 μM of the primers TerLF (5′-TGGGAA AACCTAACAGATGCC-3′) and TerLR (5′-ATGCAAGCCCAT TTGCTGAAG-3′), 1X Phusion buffer, 200 μM dNTPs, and 0.02 U/μl Phusion DNA polymerase from Thermo Scientific). The following conditions were applied: 1 cycle of 98°C for 3 min; 34 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 1 min; followed by 72°C for 10 min and refrigeration at 4°C.

RESULTS

Assembly of Simulated Viromes

We tested five assembly programs for their accuracy in reconstruction of viromes unaffected by sample processing (SIM1), and virome datasets with sample processing biases such as severe genome fragmentation (SIM2) and excessive read coverage for selected small circular virus genomes (SIM3). At constant read numbers, increased read length (300 vs. 80 nt) improved assembly accuracy and genome recovery, but did not change the performance of the assemblers relative to each other. Below we show the results for the 300 nt datasets, the results for the 80 nt datasets are provided in **Table S1**.

Sim1: This dataset consisted of 4.5 M read pairs from 572 complete viral genomes representing a published simulated virome (Vázquez-Castellanos et al., 2014). The results for the five assemblers are summarized in **Table 1**. The highest number of recovered genomes (195) was achieved using CLC. This assembler also had the highest N50 (∼27 kb). This was associated with top performance in assembly of large scaffolds (99 ≥ 50 kb), 90% of which were reconstructed correctly. Comparable results were obtained with IDBA-UD, which showed exceptional assembly precision for most scaffold length categories and successfully reconstructed 147 genomes. In comparison, SPAdes correctly reconstructed a larger number of scaffolds <50 kb, resulting in the longest total correctly assembled sequence length (31.5 Mb). However, for scaffolds ≥50 kb assembly precision was considerably reduced (16% misassembled sequences). This trend was even more accentuated with RayMeta, which reconstructed only 44 correct scaffolds ≥50 kb. ABySS, which had performed comparatively well with the 80 nt dataset (**Table S1**), was outperformed by the other assemblers in most measured parameters.

Sim2: With this dataset we tested the different assemblers for their accuracy in reconstructing short distinct fragments. We fragmented the genome sequences of the 572 viruses into 2 kb pieces, discarded every second sequence and generated 4.5 M read pairs from the remaining 11,743 sequence fragments. Allowing for a maximal over-length of 5%, any scaffolds ≥2.1 kb were considered misassembled. The results for this dataset are summarized in **Table 2**. With this dataset, ABySS performed best in terms of assembly precision: 90% of the scaffolds were equal to or smaller than the expected maximum length of

TABLE 1 | Assembly results for the simulated virome dataset SIM1 for all contigs larger than 500 nt.

	CLC	IDBA-UD	RayMeta	SPAdes	ABySS
Read pairs mapped concordantly to assembly (%)	97	74	91	97	50
>1 times to same scaffold (%)	0.1	0.0	8.8	0.5	5.5
>1 times to different scaffolds (%)	8.1	4.2	6.1	6.0	1.3
Number of scaffolds (% misassembled)	7545 (3)	9454 (1)	8333 (1)	10461 (1)	8179 (2)
500–999 nt	3400 (1)	4336 (1)	4016 (0)	5288 (1)	3974 (2)
1–1.999 kb	1946 (2)	2655 (1)	2127 (0)	2811 (0)	2190 (2)
2–4.999 kb	1207 (4)	1556 (1)	1285 (0)	1416 (1)	1336 (1)
5–9.999 kb	426 (5)	458 (1)	451 (1)	411 (4)	374 (1)
10–19.999 kb	227 (5)	168 (2)	224 (4)	212 (9)	161 (4)
20–49.999 kb	240 (7)	189 (3)	169 (10)	229 (7)	110 (6)
50–99.999 kb	68 (10)	62 (5)	47 (30)	66 (14)	25 (60)
100–199.999 kb	27 (7)	26 (19)	12 (17)	24 (21)	9 (44)
≥200 kb	4 (25)	4 (0)	2 (50)	4 (25)	0 (na)
Max scaffold length:	277,192	257,672	249,422	262,666	178,000
N50:	27,249	15,340	11,876	17,270	5723
Scaffolds correctly mapped to genomes	7360	9389	8266	10,338	8022
Scaffolds correctly mapped to genomes (%)	98	99	99	99	98
Number of misassembled scaffolds	185	65	67	123	157
Total correctly assembled length (Mb)	30.1	31.0	24.8	31.5	19.0
Genomes hit	558	557	531	554	490
Genomes recovered	195	147	124	181	45
ALE score [0,1]	0.2	0.6	0.3	0.3	1.0

TABLE 2 | Assembly results for the simulated virome dataset SIM2 for all contigs larger than 500 nt.

	CLC	IDBA-UD	RayMeta	SPAdes	ABySS
Read pairs mapped concordantly to assembly (%)	91	62	67	95	43
>1 times to same scaffold (%)	0.1	0.0	0.3	42.0	0.8
>1 times to different scaffolds (%)	8.3	3.8	6.6	4.8	0.8
Number of scaffolds (% misassembled)	11,991 (38)	19,346 (46)	14,196 (51)	12,414 (47)	11,279 (18)
500–999 nt	4268 (20)	5593 (30)	4389 (21)	5826 (25)	5093 (9)
1000–1499 nt	1684 (30)	2513 (43)	1881 (37)	2040 (37)	1914 (14)
1500–2099 nt	3373 (17)	5955 (16)	3050 (22)	1475 (42)	3217 (8)
≥2100 nt:	2666 (100)	5285 (100)	4876 (100)	3073 (100)	1055 (100)
Max scaffold length:	226,834	47,930	24,446	365,072	134,592
N50:	4023	2001	2454	14,846	2001
Scaffolds correctly mapped to genomes	4697	9136	5254	4524	5586
Scaffolds correctly mapped to genomes (%)	39	47	37	36	50
Number of misassembled scaffolds	7294	10,210	8942	7890	5693
correctly assembled length (Mb)	6.3	13.3	6.9	4.6	6.9
Genomes hit	4592	8267	4995	4030	5120
Genomes recovered	1809	4420	1881	451	1517
ALE score [0,1]	0.4	0.8	0.7	0.3	1.0

2 kb, and only 10% of those were misassembled. The other assemblers produced high numbers of scaffolds ≥2.1 kb and misassembly rates for the investigated smaller scaffold fractions were generally high (16–43%). Of those four assemblers, CLC performed best in terms of precision (lowest misassembly rates for all length categories), while IDBA-UD prevailed in

terms of numbers, resulting in the highest number of correctly reconstructed scaffolds (9136) and the longest total correctly assembled sequence length (13.3 Mb).

Sim3: Preliminary analyses of the copper virome sequencing data indicated severe excessive read coverage for selected circovirus-like genomes in the dataset. Therefore, Sim3 consisted of 4.5 M read pairs generated from the above genomes with previous relative read coverage and five additional circovirus genomes with exceptionally high relative read coverage. The results are summarized in **Table 3**. CLC again produced the highest number of correctly assembled large scaffolds (only 1 of the 47 scaffolds ≥50 kb was misassembled). It also recovered the highest number of genomes (148), including the five highly covered circoviruses. We observed that these highly covered circovirus genomes had been reconstructed by CLC into multiple copies and subsequences of nearly 100% sequence identity. The next best assembler in terms of precision was IDBA-UD, which outperformed CLC in correct assembly of smaller scaffolds (misassembly rate for scaffolds <50 kb only 0.7%). It correctly reconstructed 40 scaffolds ≥50 kb (10% misassembly rate), and recovered 111 genomes. Noticeably, with this dataset IDBA-UD used only 31% of the reads, and did not reassemble the highly covered circoviral sequences. As with SIM1, SPAdes, and RayMeta showed exceptional average assembly precision (average misassembly rates 0.9 and 0.6%, respectively), which, however, masked high misassembly rates for the few large scaffolds ≥50 kb (15 and 21%, respectively). Although SPAdes produced the longest correct total sequence length (∼26 Mb), this was mainly due to the correct assembly of high numbers of smaller scaffolds. Besides CLC, SPAdes was the only other assembler that recovered the circoviral genomes. ABySS was again outperformed by all other assemblers for most measured parameters.

Copper Site Data Assembly

The copper dataset consisted of 13.8 M raw sequencing reads (6.9 M read pairs). Despite high quality, a large number of reads was lost during preassembly filtering: 4.3 M reads were duplicates, 1.3 M reads matched PhiX-174, and 1.1 M reads matched the human genome. The remaining 6.9 M high quality reads were used for assembly with CLC, which performed best in recovery of viral genomes. The results are summarized in **Table 4**. The primary assembly of the reads produced 38,365 contigs longer than 200 nt, and the largest contig was 35.5 kb. As discovered in our SIM3 analyses, CLC sometimes reconstructs highly covered sequences into several copies and subsequences. We therefore aligned the assembly to itself and investigated it for replicated sequences using network analyses. This lead to the identification of a contig cluster of 286 nearly identical sequences, assembled from a total of 3.6 M reads. Another 11 contigs that had an average read coverage >5000 were present as single copies. In total, these sequences had combined 67% (4.6 M reads) of the dataset. After removing these reads, we proceeded with the separate assembly of the 2.3 M unmapped reads (sub-assembly 1),

TABLE 3 | Assembly results for the simulated virome dataset SIM3 for all contigs larger than 500 nt.

	CLC	IDBA-UD	RayMeta	SPAdes	ABySS
Read pairs mapped concordantly to assembly (%)	89	31	40	92	18
>1 times to same scaffold (%)	0.2	0.0	29.1	1.1	1.1
>1 times to different scaffolds (%)	2.8	3.1	3.5	1.4	0.9
Number of scaffolds (% misassembled)	9068 (2)	9543 (1)	8430 (1)	12,231 (1)	7248 (1)
500–999 nt	4433 (1)	4537 (1)	4571 (0)	6657 (0)	4067 (1)
1–1.999 kb	2449 (2)	2694 (1)	2295 (0)	3280 (1)	1873 (1)
2–4.999 kb	1352 (3)	1548 (1)	1041 (1)	1573 (1)	867 (1)
5–9.999 kb	456 (3)	436 (1)	266 (2)	374 (4)	226 (1)
10–19.999 kb	166 (5)	154 (2)	117 (4)	152 (6)	130 (4)
20–49.999 kb	165 (5)	132 (6)	111 (8)	155 (7)	75 (8)
50–99.999 kb	35 (3)	29 (10)	23 (22)	28 (14)	7 (43)
100–199.999 kb	11 (0)	12 (8)	6 (17)	11 (18)	3 (67)
≥200 kb	1 (0)	1 (0)	0 (na)	1 (0)	0 (na)
Max scaffold length:	244,834	244,953	163,001	245,076	141,829
N50:	8073	5958	5262	4759	3766
Scaffolds correctly mapped to genomes	8900	9471	8380	12,126	7158
Scaffolds correctly mapped to genomes (%)	98	99	99	99	99
Number of misassembled scaffolds	168	72	50	105	90
Total correctly assembled length (Mb)	25.0	24.2	18.1	26.1	13.6
Genomes hit	565	552	514	559	479
Genomes recovered	148	111	83	126	23
ALE score [0,1]	0.2	0.9	0.8	0.2	1.0

TABLE 4 | Assembly results for the Namib Desert copper-site dataset for all contigs larger than 200 nt.

	Reads used	Contigs	Assembled reads	% reads assembled	N50	Max
Primary assembly	6,877,654	38,365	5,356,821	77.9	913	35,461
Sub-assembly 1	2,241,214	35,966	1,562,711	69.7	961	35,670
Sub-assembly 2	3,629,636	135	3,385,364	93.3	–	1835

and the 3.6 M reads that had mapped to the contig cluster (sub-assembly 2). Despite the significantly lower read number, sub-assembly 1 produced a similar number of contigs, a comparable contig length distribution, and a marginally higher N50 than the primary assembly. Sub-assembly 2 produced 135 contigs, which really represented only two different sequences: a potentially circular DNA sequence of 1308 nt and a linear contig consisting of repeats. The remaining 133 sequences were shorter versions of these two contigs.

Copper Site Data Annotation

A total of 20,106 scaffolds ≥300 nt were submitted to MetaVir for gene prediction and annotation. The program predicted 33,254 ORFs, and assigned 4027 ORFs from 3382 scaffolds to viral taxonomic groups based on BLAST matches to the viral subset of the RefSeq database. Our subsequent protein annotations using the NCBI-nr and RefSeq-P databases and manual verifications could only confirm viral origin for 1274 proteins. Of these, 577 had Pfam domain annotations, many of which indicated viral origin (e.g., terminase, phage-integrase, phage portal, phage capsid). The remaining 2753 proteins (68%) had better matches to archaeal, bacterial, or eukaryotic proteins in the NCBI-nr and/or RefSeq-P databases, and Pfam annotations were not virus-related. Only 305 of these proteins were found on contigs that encoded at least one confirmed viral protein, i.e., may potentially represent viral genes. For 51% of these proteins the score from MetaVir was less than half of the top score. To avoid false viral annotation, we did not include these proteins into our final viral protein dataset. Furthermore, we identified eight proteins that had Pfam annotations indicating viral origin, and 392 proteins that had matched viral proteins in the RefSeq-P and/or NCBI-nr databases (max 1.0E-4, min bit-score 45), but had not been detected by MetaVir. Our final dataset of predicted viral proteins therefore consisted of 1674 sequences (**Table S2**).

To determine the LCA affiliation for these proteins, we repeated BLASTp against the viral protein RefSeq database and analyzed the results using MEGAN5. Taxonomic affiliations were obtained for 1288 proteins and the results indicated a diverse spectrum of viral species in the dataset (**Figure 1**). The majority of the proteins (748) were assigned to Caudovirales, including siphoviridae (259), podiviridae (213), and myroviridae (159) species, supporting our TEM results (**Figure S4**). Most species were represented by a single protein. However, within the unclassified podoviridae, 30 proteins were affiliated with *Pelagibacter* phage HTVC010P, another 30 proteins with *Thalassomonas* phage BA3, 12 proteins with *Pseudoalteromonas* phage RIO-1 and nine proteins with *Cellulophaga* phage phi14:2. Among the siphoviridae, higher protein numbers were found for *Azospirillium* phage Cd (6), *Geobacillus* phage GBK2 (9), and for diverse *Lactococcus* phages. Of the myoviridae species *Aeromonas* phage vB_AsaM-56 had the highest protein count (11). Among the unclassified dsDNA viruses, archaeal viruses prevailed by far: 197 proteins were assigned to haloviruses, with *Halovirus* HRTV-4 representing the most common top hit. Two thirds of the quality filtered reads from the copper site dataset (4.1 M) assembled into just 35 contigs that encoded proteins with sequence similarity to proteins from small circular viruses. This included the 1308 nt potentially circular contig from sub-assembly 2, which combined 3.6 M reads. It encoded one geminivirus-like capsid protein (Pfam: PF00844) and two other ORFs that could not be annotated. Since the replication-associated protein could not be identified, it remains unclear whether this particular contig represents a complete viral genome. Most of the other circovirus-like contigs encoded proteins with similarity to sequences from recent virome studies on seawater samples from Tampa Bay, Florida, USA (McDaniel et al., 2014), mollusks from the Avon Heathcote estuary in New Zealand (Dayaram et al., 2015), and sewage-associated circular viruses (Kraberger et al., 2015).

Phylogenetic Analyses on Circovirus-Like Proteins

Manual annotation of the contigs with circoviral-like proteins allowed identification of nine putative replication-associated proteins (pRepAs). All but the pRepA from contig_176 contained the viral replication (PF02407) and the RNA-helicase (PF00910) domains (**Figure S5**). Our phylogenetic analyses (**Figure 2**) showed that these proteins differed from described animal and bird circoviral RepAs (clade-1), forming three distinct, well-supported clades. Most of the pRepAs discovered in this study formed a monophyletic sub-clade within clade-2, co-segregating with a sub-clade containing the pRepA from contig_351 as well as one pRepA from a marine water study (Labonté(Labonté and Suttle, 2013)), a mollusk (Dayaram et al., 2015) and a bat feces virome study (Ge et al., 2011), respectively. In all proteins of clade-2, the Rep motif I was unrecognizable, while the motifs II (H.Q) and III (Y..KD/E) were conserved, resembling those described for animal and bird circoviral RepAs. Clade-3 contained the pRepA from contig_176, six RepAs from diverse marine invertebrates and three RepAs from uncultured marine viruses. In these sequences, all three RepA motifs (I: FTINN, II: HLQG, III: YCKKD) were conserved, again resembling those described for animal circoviral RepAs. For the pRepA of contig_2pa, the Rep motifs I (FLTF), II (HLH), and III (YCMKD) resembled those of plant geminiviruses. However, it did not contain the recently described geminivirus replication (GRS)

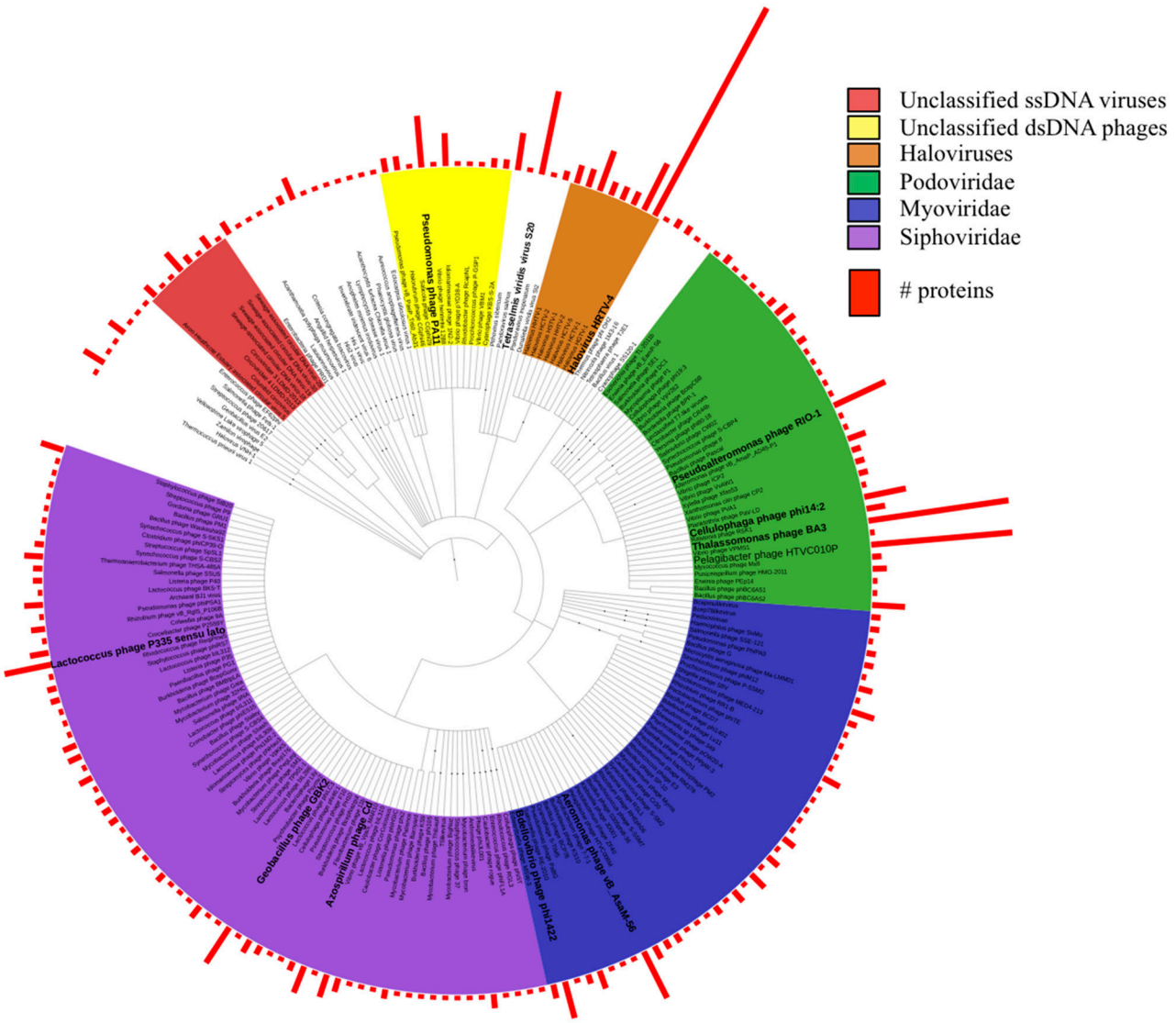

FIGURE 1 | Taxonomic tree of all viral protein sequences discovered in this study. Proteins were assigned taxonomic IDs using the lowest common ancestor method. Bars depict numbers of proteins assigned to the corresponding taxa.

domain. Accordingly, phylogenetic analyses placed this protein, as well as the three pRepAs from the Sewage-associated circular DNA viruses 18, 14, and 28 (Kraberger et al., 2015) into clade-4 which harbored three plant viral RepAs (chloris striate mosaic virus, mulberry mosaic dwarf associated virus and Mulberry crinkle-associated virus).

Of the 14 putative putative circoviral-like capsids (pCAPs) discovered in this study, nine shared sufficient sequence similarity to each other and to known circoviral-like putative capsid proteins to generate a phylogenetic tree (**Figure 3**, **Figure S6**). Clade-1 contained the pCAPs from this study, five pCAPS from a sewage virome study (Kraberger et al., 2015), six pCAPs from a seawater virome study (Tampa Bay, FL, USA; McDaniel et al., 2014), two pCAPs from the Avon Heathcote Estuary mollusk virome study (Dayaram et al., 2015) and one putative capsid from a fiddler crab virus. Within clade-1, the sewage virus pCAPs and the seawater virus pCAPs formed separate sub-clades, but branch support was low. Clade-2 contained selected plant satellite viruses (Maize white line mosaic Satellite virus and three Satellite tobacco necrosis viruses). None of the investigated sequences showed similarity to any of the known animal or bird circoviral capsid proteins.

Novel Small Circovirus-Like Genomes

Here, we describe three novel circovirus-like genomes identified in this study (**Table 5**). The pRepAs and the pCAPs of these genomes are included in the above analyses and show that contig_21 and contig_71 are potentially closely related viral species as both the pRepAs and the pCAPs co-segregate on the respective phylogenetic trees. Contig_176 represents a more

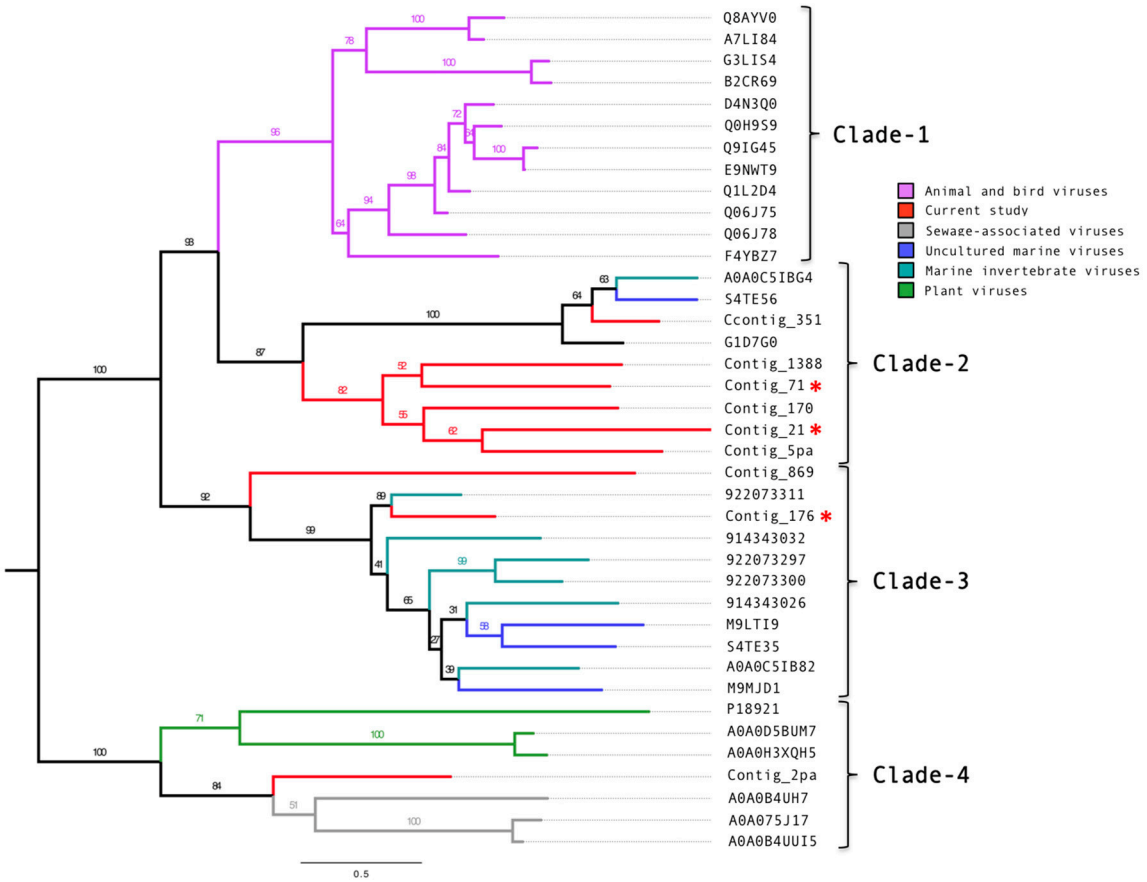

FIGURE 2 | Phylogenetic tree of the circovirus-like replication-associated proteins. Alignments were generated using MAFFT, the phylogenetic tree was produced using the RAxML BlackBox, the tree was visualized using FigTree. The asterisk serves to identify the replication and capsid proteins identified in this study.

distantly related viral species, with a pRepA that shows similarity to invertebrate circoviral pRepAs.

Novel *Thalassotalea loyana* Phage Ba3-Like Viral Genome

The largest contig of the assembly (Contig_13) represented a potentially complete 35.461 nt linear viral genome with a uniform high average read coverage [1221]. It encoded 47 ORFs, with ORFs 1–39 located on one, and ORFs 40–47 on the complementary DNA strand (**Table S3**). Gene annotations using BLAST and Pfam predicted two putative early genes (ORF42, ORF40), four putative late genes (ORF2, ORF8, ORF17, ORF18), one structural gene (ORF12), and one gene that may potentially be involved in host lysis (ORF21). A 5 kb region of the BA3-like virus could be amplified from the environmental fraction confirming presence of the virus in the sample. Furthermore, end sequence analysis of the amplified product showed 100% sequence identity to Contig_13, providing confidence in the assembly. However, host infection assays with seven *Thalassomonas/Thalassotalea* species using the environmental virus fraction failed to produce plaques, and PCR analyses of virus challenged bacterial cells did not amplify the target region.

In total, 22 ORFs had top matches to *T. loyana* phage BA3, including the ORF with the highest score (used for taxonomic assignment by MetaVir). Sequence comparisons between Contig_13 and BA3 indicated high levels of conservation for gene order, gene orientation, and protein sequences (average percent identity between the 22 ORFs was 64%) for the corresponding 5′ regions of the two sequences (**Figure 4**). However, most of the early genes of BA3 are located in the 3′ region of the genome and had no homologs on contig_13. Further investigations are necessary to identify the host of the Contig_13 virus.

Non-viral Sequences in the Copper Site Dataset

Contig annotation indicated that at least 5.1 M reads of the quality filtered dataset originated from viral DNA. However, we also found 1.1 M human reads, 50% of which assembled into contigs of up to 8 kb. Another 0.6 M quality filtered reads assembled into sequences that were annotated as archaeal, bacterial or eukaryotic. This degree of contamination was not

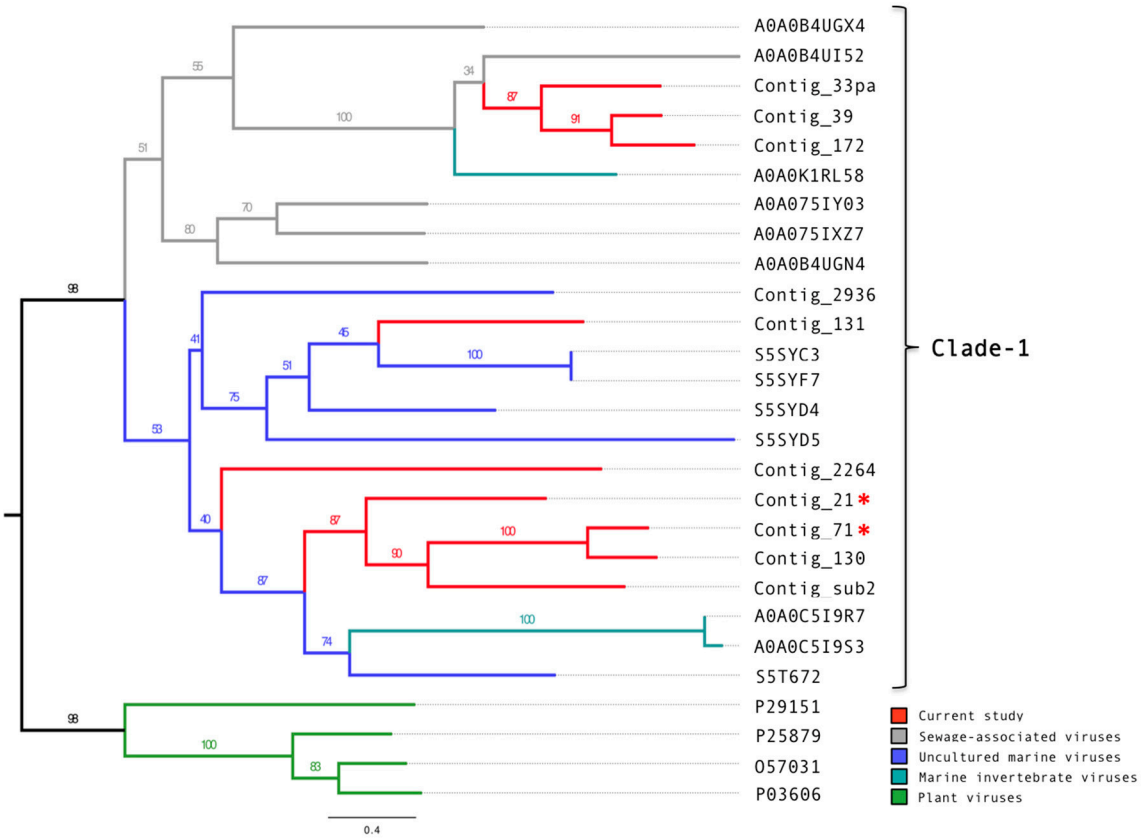

FIGURE 3 | Phylogenetic tree of the circovirus-like capsid proteins. Alignments were generated using MAFFT, the phylogenetic tree was produced using the RAxML BlackBox, the tree was visualized using FigTree. The asterisk serves to identify the replication and capsid proteins identified in this study.

TABLE 5 | Annotation table for novel circoviral-like genomes discovered in this study.

	Annotation	Coordinates			
Contig_21	ORF1	128–937 (−1)	PF00844	44% id to hypothetical protein from Circoviridae 3 LDMD-2013	putative gemini coat protein
	ORF2	1127–2130 (+2)	PF02407, PF00910	38% id to replication-associated protein of the Avon-Heathcote Estuary associated circular virus 5	putative viral replication protein
	ORI	1078–1115	GACC**AGTATTAC** flanked by a perfect 13 nt palindromic repeat		
Contig_71	ORF1	109–1410 (+1)	PF02407, PF00910	40% id to the replication-associated protein of the Avon-Heathcote Estuary associated circular virus 5	putative viral replication protein
	ORF2	1572–2360 (−3)	PF00844	35% id to a hypothetical protein from Circoviridae 3 LDMD-2013	putative gemini coat protein
	ORI	42–83	CT**AGTATTAT** flanked by an imperfect 15 nt palindromic repeat		
Contig_176	ORF1	50–955 (+2)	PF02407	59% id to the putative replication initiation protein of the *Gammarus* sp. amphipod associated circular virus	putative viral replication protein
	ORF2	1187–1951 (−3)	–	–	–
	ORI	1–33	CCT**ATTATTAC** flanked by a perfect 11 nt palindromic repeat		

The nonanucleotide sequence where ssDNA synthesis is initiated is shown in bold.

evident when screening the reads against the complete SILVA database: 0.06% of the reads from the raw dataset and 0.1% of the reads from the filtered dataset mapped to rRNA genes at min bit-score 80, indicating negligible amounts of cellular contamination as defined by Roux et al. (2013). Furthermore, at max 1.0E-10, only 577 of the 33296 predicted proteins (1.6%) mapped to cellular protein sequences from the extensive marker gene database published by Laffy et al. (2016).

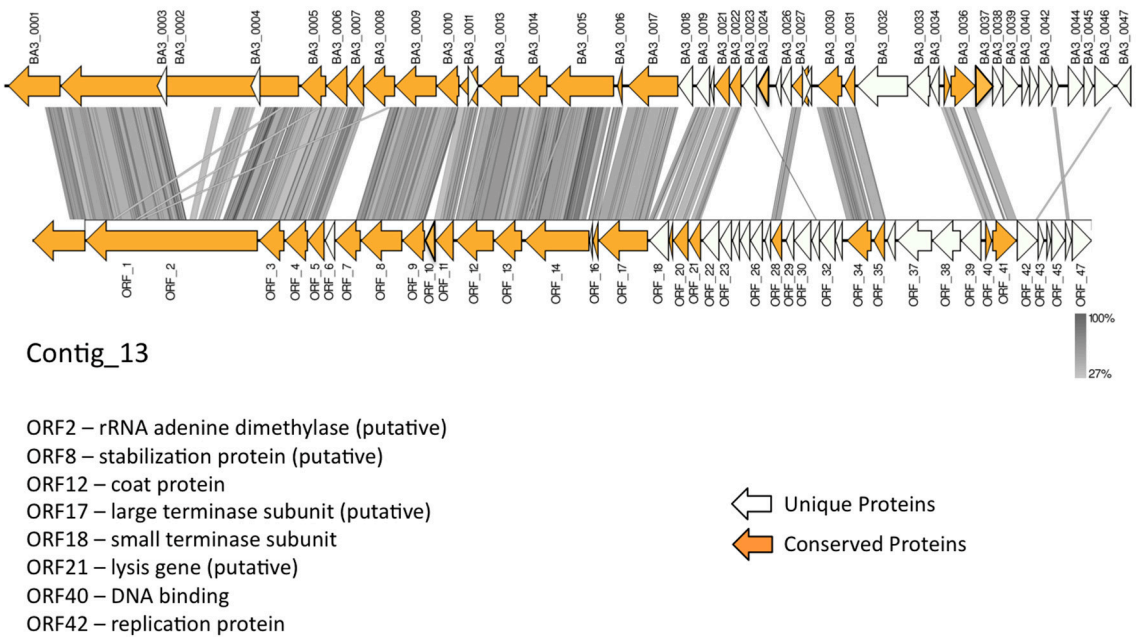

FIGURE 4 | Comparison of the published *Thalassotalea loyana* phage BA3 genome and contig_13, representing a potentially complete BA3 phage–like genome.

We used this ancillary environmental data to identify new prophage genes. We had obtained substantial genome coverage for a *Limnobacter* strain: 1429 contigs (cumulative length 2 Mb) matched the *Limnobacter sp.* MED105 genome at 85% identity over 95% of their length; and 3694 proteins had top hits to this bacterium. Twelve proteins had BLAST and Pfam annotations indicating viral origin (e.g., toprim proteins, phage integrase).

DISCUSSION

Initial analyses of the copper dataset indicated fragmented representation of the original genomes and excessive read coverage for a number of small viral genomes. Previous studies with simulated datasets had not addressed these particular data issues. To identify a suitable assembly program, we simulated datasets that mimicked the copper-site data challenges, and compared performance of five assemblers that had previously shown best results in metagenome reconstruction (de Cárcer et al., 2014; Vázquez-Castellanos et al., 2014; García-López et al., 2015; Laffy et al., 2016). Of particular interest was the assembly precision for different size sequences, which are expected in metagenomes. In the above citations, this is assessed using averaged values for the whole dataset (average contig numbers, N50, %accuracy, %chimeric sequences). However, large scaffolds that are particularly interesting for downstream analyses usually represent a very small fraction of an assembly, and the degree of their misassembly is not discernable based on those values. We therefore increased the resolution of assembly accuracy assessment by analyzing misassembly rates for distinct scaffold size fractions. We also investigated the ability of each assembler to recover genomes as single scaffolds, since multi-scaffold representation of genomes necessitates additional assembly steps that may generate chimeras. Based on these parameters, the CLC Genomics Workbench emerged as the most suitable assembler: it consistently recovered the highest number of genomes, accurately reconstructed large scaffolds, and outperformed most other assemblers in recognition of fragmented genome data (SIM2 virome). Shorter sequences were often misassembled by CLC with the 80 nt datasets, but these could be identified by long stretches of Ns, and filtering sequences with N% $\geq$ 20% alleviated the problem. With the 300 nt datasets, CLC assembly accuracies for shorter sequences were comparable to those achieved by the other assemblers. IDBA-UD generally excelled in assembly precision, but failed to assemble the 80 nt SIM2 dataset. SPAdes, RayMeta and ABySS had high average assembly accuracy, but often misassembled sequences $\geq$20 kb (80 nt) and $\geq$50 kb (300 nt). Only CLC and SPAdes recovered the circoviral genomes with excessive read coverage (SIM3 virome), which were entirely missing in the assemblies generated by IDBA-UD, RayMeta and ABySS. Interestingly, with the SIM3 dataset, these three assemblers used only a fraction of the reads for sequence reconstruction, which may indicate stringent settings for high-frequency kmers in these programs. Since it is likely

that virome composition (relatedness of species, variation in genome sizes, genome coverages) affect assembler performance, kmer assessment of environmental datasets may prove a useful tool for choosing a suitable program. With SIM3 we observed that CLC reassembled the highly covered circoviral genomes into multiple copies and subsequences of nearly 100% sequence identity. To minimize such artificial duplicates in the final copper site assembly, we investigated the primary assembly for duplicates using network analysis. We identified one cluster of 286 contigs. Separate reassembly of the corresponding reads produced 135 contigs, which represented only two distinct sequences based on manual verification. One may argue that such duplicates may represent different viral strains, and in fact, the approach could be applied to identify viral strain variants. With the copper site dataset, however, it would be difficult to distinguish computational and laboratorial artifacts from true viral strain variants. Filtering for reads from excessively covered sequences did not improve assembly of the remaining reads, which contrasts with previous reports.

A number of papers document contamination of currently published virome datasets with non-viral sequences. In 2011, Schmieder and Edwards analyzed 202 published microbial and viral metagenomes and found indications for human DNA contamination in 145 datasets. Roux et al. (2013) reported on substantial cellular DNA contaminations found in a quarter of 67 published viromes. According to the authors, all investigated viromes derived from complex matrices (e.g., feaces, gut, coral, insect, and animal samples) contained DNA encoding ribosomal genes, suggesting that purification of viral particles from such environments may be challenging. Soil samples definitely fall into this category. Enzyme-inhibiting compounds not only affect library construction (Carvalhais et al., 2012; Tveit et al., 2014); they can also protect extracellular DNA (eDNA) from degradation, as demonstrated for humic acids (Crecchio and Stotzky, 1998), organic matter (Nguyen et al., 2010), clay (Mao et al., 2014), and extracellular polymeric substances produced in microbial biofilms (Grande et al., 2011, 2014; Rosario et al., 2015). With soil samples, additional purification steps and WGA are sometimes unavoidable to overcome inhibitors and avoid failure of sequencing reactions. Yet, these procedures can introduce biases such as selective purification of DNA from specific organisms, DNA fractionation and selective sequence amplification (Kim and Bae, 2011; Yuan et al., 2012; van Dijk et al., 2014) that will affect representation of the viral and non-viral DNA. Lower target/contaminant DNA ratios result in higher numbers of reads from contaminants in the read dataset, as demonstrated by Lusk (2014), who found that frequencies of contaminating human reads in an *E. coli* study increased with decreasing sample DNA concentrations. In this study, we followed all standard laboratorial practices for virome analyses, including sterile handling of material, microscopic examination of the sample for bacterial cells, TEM to verify presence of viral particles, as well as DNAse and RNAse treatment of the viral particles prior to protein coat digestion (verified through 16S rRNA PCR with positive and negative controls). We therefore hypothesize that trace amounts of human and environmental DNA escaped DNAse treatment and were preferentially amplified by Phi29, leading to the observed contaminations. Phi29 amplification is also the likely reason for the high read duplication rates and the very selective amplification of circovirus-like genomes: previous analyses have shown that this enzyme predominantly amplifies small circular DNA (Kim and Bae, 2011). Therefore, dedicated optimization studies on sample preparation and sequencing methodologies for soil virome analysis (as conducted for human skin and gut viromes by Hannigan et al., 2015 and Džunková et al., 2015, respectively) would help to address these problems.

As follows from above, virome data must be inspected for contaminating DNA, in particular when the samples were obtained from complex environmental matrices. This does not appear to be common practice: 19 of 25 investigated virome studies published since 2014 appear to rely solely on 16S rDNA PCR to detect bacterial contamination and do not mention biocomputational inspection of their data prior to submitting it to MetaVir for annotation. The MetaVir server does not filter for non-viral sequences, but annotates the submitted data using the viral subset of the RefSeq database. In case of contamination, all non-viral proteins showing significant similarity to proteins in the database will be misannotated as viral. This may not only lead to vast overestimation of bacterial genes in viral sequences identified through virome studies (as found for antibiotic resistance genes by Enault et al., 2017), but also result in error propagation in databases. With our dataset, screening reads against rRNA genes and proteins against the most extensive cellular marker gene database published to date (Laffy et al., 2016) did not indicate substantial contamination of the dataset with non-viral DNA. Yet, large scaffolds of non-viral DNA (up to 18 kb) with high nucleotide identity ($\geq$95%) over their entire length ($\geq$95% coverage) were recovered. Two scenarios can explain this result: (a) genomic regions encoding marker genes were absent, i.e., contamination originated from fragments of free DNA that had escaped DNAse treatment and were selectively amplified; or (b) minute amounts of cellular DNA was present, but only genomic regions not encoding marker genes were selectively amplified. The studies by Schmieder and Edwards (2011), Roux et al. (2013), and Lusk (2014) suggest that such type of contaminations may be quiet common in metagenome datasets. These findings indicate that the virome community would benefit from benchmarking studies using simulated and environmental datasets to establish optimal screening and filtering procedures. In fact, when annotated correctly, data from non-viral DNA can provide interesting ancillary information about the sampled environment, as found in this study. Considering the presence of non-viral DNA in the copper site data, we chose a stringent approach to annotate sequences as viral and used the complete NCBI-nr and RefSeq-P databases rather than the viral RefSeq database for annotation (Hurwitz et al., 2016). While this method results in underestimation of viral sequences, it provides higher confidence in the assignment of viral annotations.

Our sequence data from the Namib Desert copper-site indicated a diverse range of dsDNA viruses from all three families of the Caudovirales. Viruses from ubiquitous bacterial species included *Geobacillus*, *Bacillus*, and *Aeromonas* phages, confirming previous findings for the Namib Desert

(Adriaenssens et al., 2015). A number of the detected viruses may be associated with plants and their rhizobial symbionts, since several proteins showed sequence similarity to diverse *Lactococcus* phages and *Azospirillium* phage Cd. A high number of proteins was affiliated with Haloviruses that infect archaea, which is also in agreement with previous findings for this environment (Adriaenssens et al., 2016). Furthermore, 35 contigs encoded proteins that showed sequence similarity to selected circovirus-like and geminivirus proteins. Three novel circovirus-like genomes could be identified. Only few published proteins shared sequence similarity with the putative replication-associated (pRepAs) and the putative capsid (pCAPs) proteins from these contigs. This is not surprising, considering that these small viruses evolve exceptionally fast (reviewed in Krupovic, 2013). Interestingly, most of these published proteins originated from studies of aqueous environments. So, all analyzed pCAPs from this study co-segregated with viral proteins from sewage (Kraberger et al., 2015), marine water (McDaniel et al., 2014), and estuarian/marine invertebrate virome studies (Dayaram et al., 2015; Rosario et al., 2015). These proteins appeared to be distantly related to selected plant viruses and shared no sequence similarity with known circoviral pCAPs from animals or birds. The pRepAs from Contig_176 and Contig_869 co-segregated with pRepAs of circular ssDNA viruses from diverse marine invertebrates and several uncultured marine viruses. The pRepAs from another six contigs formed a well-supported clade with just three published proteins from one marine (Labonté and Suttle, 2013), one mollusk (Dayaram et al., 2015) and one bat feces (Ge et al., 2011) virome study, respectively. Interestingly, the bat sequence YN-BtCV-1 (phylogenetically unplaced in the above study) originates from a roosting site in Yunnan populated by diverse *Myotis* species. Several members (including *M. pilosus*, which is found in this area) are known to eat fish (Stadelmann et al., 2004). It is possible, that the sequences from the mollusk virome and the uncultured marine virus also represent fish circovirus-like species, which would explain the highly-supported proximity of clade-2 to clade-1 (animal/bird circoviral pRepAs) relative to clade-3 (putative invertebrate circoviral pRepAs) and clade-4 (putative geminivirus-like RepAs). However, considering that our circovirus annotations could be biased due to the limited amount of sequence information on circoviral genomes from other habitats, these interpretations remain to be validated.

Yet, other results of this study also link the investigated virome to the marine environment. Contig_13, which represents a potentially complete 35.5 kb viral genome, shares high sequence similarity with *T. loyana* phage BA3, which infects the coral pathogen *T. loyana* (Efrony et al., 2009). Although we could not confirm that *Thalassomonas/Thalassotalea* species are hosts for the identified virus, the extensive degree of conservation between the two genomes may indicate common ancestry. Another 51 protein sequences were taxonomically affiliated with Pelagibacter phage HTVC010P, Cellulophaga phage phi14:2, and Pseudoalteromonas phage RIO-1, all of which infect marine bacteria. Incidentally, these four species all belong to the podoviridae family, which represents a major fraction in marine viromes, but appears to be less common in viromes from non-marine environments including the Namib Desert (Prestel et al., 2008; Adriaenssens et al., 2015). Moreover, a large number of contigs (totaling 2.6 Mb) mapped to the genome of *Limnobacter sp.* MED105, a common marine bacterium. Last but not least, a previous study on viromes from Namib Desert salt pans mentions several contigs that clustered with viral sequences of aquatic origin (Adriaenssens et al., 2016).

Considering that samples analyzed in this study were derived from a soil sample in the Namib Desert, these results are, at first glance, surprising. However, this desert is characterized by seasonal fog from the Atlantic Ocean that reportedly reaches as far as 100 km inland (http://tinyurl.com/zwuky7v). The sample was collected within this distance (∼30 km to the south-east of the Gobabeb research station), where "high-fog" is prevalent (Eckardt et al., 2013a). Sampling was conducted at the end of April 2013, just after the typical "high-fog" months in this location (September-March). It is therefore possible that microorganisms and viral particles carried by fog and wind from the Atlantic Ocean represent a sampling fraction of this study. Seasonal variations of this phenomenon could also explain why few viral particles were found on following sampling occasions (April 2014 and 2015, respectively). Investigations on the marine virome diversity in proximate coastal regions may help to verify our hypothesis.

AUTHOR CONTRIBUTIONS

Lv and MT designed and conceptualized the project and collected the samples, Lv and BK conducted DNA isolation and NGS, UH conceptualized and conducted the biocomputational data analyses and wrote the manuscript, Pv assisted in biocomputational data analyses, IO conducted the Thalassomonas infection studies. All authors read and approved the manuscript.

FUNDING

The project was supported by the NRF South Africa through the DST/NRF SARChI programme (UID87326).

ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Mary Seely from the Gobabeb Research Station for ongoing support in sampling.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00013/full#supplementary-material

Table S1 | Assembly results for the simulated virome datasets SIM1, SIM2, and SIM3 for all contigs larger than 200 nt.

Table S2 | Annotation table for all proteins discovered in this study comparing MetaVir annotations with those derived from the nr and RefSeq databases from NCBI, considering Pfam.

Table S3 | Annotation table for the Thalassomonas phage-like genome described in this study (contig_13).

Figure S1 | Satellite image of the sampling site (Cu-site) showing its position in relation to nearby copper mines (Hope and Gorob). Inset depicts an enlargement of the site showing the trench dug alongside the road with heaps of sampled copper laden material, as well as images of the sampled material (including material with a green copper patina).

Figure S2 | Assembly workflow.

Figure S3 | Protein annotation workflow.

Figure S4 | TEM images of a selection of phage morphologies identified in material from the sampling site. Siphovirus and myovirus morphologies were observed in particular abundance.

Figure S5 | Multiple sequence alignment of circoviral replication associated proteins (MSA conducted using MAFFT).

Figure S6 | Multiple sequence alignment of circoviral capsid-like proteins (MSA conducted using MAFFT).

REFERENCES

Ackermann, H.-W. (2009). "Basic phage electron microscopy," in *Bacteriophages: Methods and Protocols,* eds M. R. J. Clokie and A. M. Kropinski (New York, NY: Humana Press), 113–126.

Adriaenssens, E. M., Van Zyl, L. J., Cowan, D. A., and Trindade, M. I. (2016). Metaviromics of Namib desert salt pans: a novel lineage of haloarchaeal salterproviruses and a rich source of ssDNA viruses. *Viruses* 8:14. doi: 10.3390/v8010014

Adriaenssens, E. M., Van Zyl, L., De Maayer, P., Rubagotti, E., Rybicki, E., Tuffin, M., et al. (2015). Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* 17, 480–495. doi: 10.1111/1462-2920.12528

Baker, G. C., Smith, J. J., and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555. doi: 10.1016/j.mimet.2003.08.009

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S. et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). RayMeta: scalable *de novo* metagenome assembly and profiling. *Genome Biol.* 13:R122. doi: 10.1186/gb-2012-13-12-r122

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Carlos, C., Castro, D. B. A., and Ottoboni, L. M. M. (2014). Comparative metagenomic analysis of coral microbial communities using a reference-independent approach. *PLoS ONE* 9:e111626. doi: 10.1371/journal.pone.0111626

Carvalhais, L. C., Dennis, P. G., Tyson, G. W., and Schenk, P. M. (2012). Application of metatranscriptomics to soil environments. *J. Microbiol. Methods* 91, 246–251. doi: 10.1016/j.mimet.2012.08.011

Clark, S. C., Egan, R., Frazier, P. I., and Wang, Z. (2013). ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* 29, 435–443. doi: 10.1093/bioinformatics/bts723

Crecchio, C., and Stotzky, G. (1998). Binding of DNA on humic acids: effect on transformation of *Bacillus subtilis* and resistance to DNase. *Soil Biol. Biochem.* 30, 1061–1067. doi: 10.1016/S0038-0717(97)00248-4

Dayaram, A., Goldstien, S., Argüello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J. S., et al. (2015). Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect. Genet. Evol.* 31, 284–295. doi: 10.1016/j.meegid.2015.02.010

de Cárcer, D. A., Angly, F. E., and Alcamí, A. (2014). Evaluation of viral genome assembly and diversity estimation in deep metagenomes. *BMC Genomics* 15:989. doi: 10.1186/1471-2164-15-989

Dopson, M., Baker-Austin, C., Koppineedi, P. R., and Bond, P. L. (2003). Growth in sulphidic mineral environments: metal resistance mechanisms in acidophilic micro-organisms. *Microbiology* 149, 1959–1970. doi: 10.1099/mic.0.26296-0

Džunková, M., D'Auria, G., and Moya, A. (2015). Direct sequencing of human gut virome fractions obtained by flow cytometry. *Front. Microbiol.* 6:955. doi: 10.3389/fmicb.2015.00955

Eckardt, F. D., Livingstone, I., Seely, M., and von Holdt, J. (2013b). The surface geology and geomorphology around Gobabeb, Namib Desert, Namibia. *Geogr. Ann. Ser. A* 95, 271–284. doi: 10.1111/geoa.12028

Eckardt, F. D., Soderberg, K., Coop, L. J., Muller, A. A., Vickery, K. J., Grandin, R. D., et al. (2013a). The nature of moisture at Gobabeb, in the central Namib Desert. *J. Arid Environ.* 93, 7–19. doi: 10.1016/j.jaridenv.2012.01.011

Efrony, R., Atad, I., and Rosenberg, E. (2009). Phage therapy of coral white plague disease: properties of phage BA3. *Curr. Microbiol.* 58, 139–145. doi: 10.1007/s00284-008-9290-x

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B., and Petit, M. A. (2017). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analysis. *ISME J.* 11, 237–247. doi: 10.1038/ismej.2016.90

Fenner, N., and Freeman, C. (2011). Drought-induced carbon loss in peatlands. *Nat. Geosci.* 4, 895–900. doi: 10.1038/ngeo1323

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., et al. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* 73, 7059–7066. doi: 10.1128/AEM.00358-07

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548. doi: 10.1038/21119

García-López, R., Vázquez-Castellanos, J. F., and Moya, A. (2015). Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front. Bioeng. Biotechnol.* 3:141. doi: 10.3389/fbioe.2015.00141

Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., et al. (2011). Genetic diversity of novel circular ssDNA viruses in bats in China. *J. Gen. Virol.* 92, 2646–2653. doi: 10.1099/vir.0.034108-0

Goudie, A. (2010). "Namib Sand Sea: large dunes in an ancient desert," in *Geomorphological Landscapes of the World*, ed P. Migon (New York, NY: Springer), 163–169.

Grande, R., Di Giulio, M., Bessa, L. J., Di Campli, E., Baffoni, M., Guarnieri, S., et al. (2011). Extracellular DNA in *Helicobacter pylori* biofilm: a backstairs rumour. *J. Appl. Microbiol.* 110, 490–498. doi: 10.1111/j.1365-2672.2010.04911.x

Grande, R., Nistico, L., Sambanthamoorthy, K., Longwell, M., Iannitelli, A., Cellini, L., et al. (2014). Temporal expression of agrB, cidA, and alsS in the early development of *Staphylococcus aureus* UAMS-1 biofilm formation and structural role of extracellular DNA and carbohydrates. *Pathog. Dis.* 70, 414–422. doi: 10.1111/2049-632X.12158

Hannigan, G. D., Meisel, J. S., Tyldsley, A. S., Zheng, Q., Hodkinson, B. P., SanMiguel, A. J., et al. (2015). The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* 6:e01578-15. doi: 10.1128/mBio.01578-15

Hansen, M. C., Tolker-Nelson, T., Givskov, M., and Molin, S. (1998). Biased 16S rDNA PCR amplification caused by interference from DNA flanking template region. *FEMS Microbiol. Ecol.* 15, 25–36. doi: 10.1111/j.1574-6941.1998.tb00500.x

Hatfull, G. F. (2008). Bacteriophage genomics. *Curr. Opin. Microbiol.* 11, 447–453. doi: 10.1016/j.mib.2008.09.004

Hosoya, S., Adachi, K., and Kasai, H. (2009). *Thalassomonas actiniarum* sp. nov. and *Thalassomonas haliotis* sp. nov., isolated from marine animals. *Int. J. Syst. Evol. Microbiol.* 59, 686–690. doi: 10.1099/ijs.0.000539-0

Hurwitz, B. L., U'Ren, L. M., and Youens-Clark, K. (2016). Computational prospecting the great viral unknown. *FEMS Microbiol. Lett.* 363, 1–12. doi: 10.1093/femsle/fnw077

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Jean, W. D., Shieh, W. Y., and Liu, T. Y. (2006). *Thalassomonas agarivorans* sp. nov., a marine agarolytic bacterium isolated from shallow coastal water of An-Ping Harbour, Taiwan, and emended description of the genus Thalassomonas. *Int. J. Syst. Evol. Microbiol.* 56, 1245–1250. doi: 10.1099/ijs.0.64130-0

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kim, K. H., and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77, 7663–7668. doi: 10.1128/AEM.00289-11

Kimura, M., Jia, Z., Nakayama, N., and Asakawa, S. (2008). Ecology of viruses in soils: past, present and future perspectives. *Soil Sci. Plant Nutr.* 54, 1–32. doi: 10.1111/j.1747-0765.2007.00197.x

Kraberger, S., Argüello-Astorga, G. R., Greenfield, L. G., Galilee, C., Law, D., Martin, D. P., et al. (2015). Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infect. Genet. Evol.* 31, 73–86. doi: 10.1016/j.meegid.2015.01.001

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12. doi: 10.1186/gb-2004-5-2-r12

Krupovic, M. (2013). Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr. Opin. Virol.* 3, 578–586. doi: 10.1016/j.coviro.2013.06.010

Labonté J. M., and Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

Laffy, P. W., Wood-Charlson, E. M., Turaev, D., Weynberg, K. D., Botté, E. S., van Oppen, M. J. H., et al. (2016). HoloVir: a workflow for investigating the diversity and function of viruses in invertebrate holobionts. *Front. Microbiol.* 7:822. doi: 10.3389/fmicb.2016.00822

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128. doi: 10.1093/bioinformatics/btl529

Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., et al. (2011). The Viral MetaGenome Annotation Pipeline (VMGAP): an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand. Genomic. Sci.* 4, 418–429. doi: 10.4056/sigs.1694706

Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE* 9:e110808. doi: 10.1371/journal.pone.0110808

Makhalanyane, T. P. (2015). Microbial ecology of hot desert edaphic systems. *FEMS Microbiol. Rev.* 39, 203–221. doi: 10.1093/femsre/fuu011

Mao, D. Q., Luo, Y., Mathieu, J., Wang, Q., Feng, L., Mu, Q. H., et al. (2014). Persistence of extracellular DNA in river sediment facilitates antibiotic resistance gene propagation. *Environ. Sci. Technol.* 48, 71–78. doi: 10.1021/es404280v

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

McDaniel, L. D., Rosario, K., Breitbart, M., and Paul, J. H. (2014). Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environ. Microbiol.* 16, 570–585. doi: 10.1111/1462-2920.12184

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12450–12455. doi: 10.1073/pnas.1300833110

Mohiuddin, M., and Schellhorn, H. E. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front. Microbiol.* 6:960. doi: 10.3389/fmicb.2015.00960

Nguyen, T. H., Chen, K. L., and Elimelech, M. (2010). Adsorption kinetics and reversibility of linear plasmid DNA on silica surfaces: influence of alkaline earth and transition metal ions. *Biomacromolecules* 11, 1225–1230. doi: 10.1021/bm901427n

Nicholson, S. (2011). *Dryland Climatology*. New York, NY: Cambridge University Press.

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* 27, 94–101. doi: 10.1093/bioinformatics/btr216

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Pointing, S. B., and Belnap, J. (2012). Microbial colonization and controls in dryland systems. *Nat. Rev. Microbiol.* 10, 551–562. doi: 10.1038/nrmicro2831

Prestel, E., Regeard, C., Andrews, J., Oger, P., and DuBow, M. S. (2012). A novel bacteriophage morphotype with a ribbon-like structure at the tail extremity. *Res. J. Microbiol.* 7, 75–81. doi: 10.3923/jm.2012.75.81

Prestel, E., Salamitou, S., and DuBow, M. S. (2008). An examination of the bacteriophages and bacteria of the Namib Desert. *J. Microbiol.* 46, 364–372. doi: 10.1007/s12275-008-0007-4

Prigent, M., Leroy, M., Confalonieri, F., Dutertre, M., and DuBow, M. S. (2005). A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles* 9, 289–296. doi: 10.1007/s00792-005-0444-5

Rambaut, A. (2008). *FigTree v1.1.1: Tree Figure Drawing Tool*. Available online at: http://tree.bio.ed.ac.uk/software/figtree/

Rawlings, D. E. (2005). Characteristics and adaptability of iron-and sulphur-oxidizing microorganisms used for the recovery of metals from minerals and their concentrates. *Microb. Cell Fact.* 4:13. doi: 10.1186/1475-2859-4-13

Richter, D. C., Ott, F., Auch, A. F., Schmid, R., and Huson, D. H. (2008). MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* 3:e3373. doi: 10.1371/journal.pone.0003373

Rosario, K., Schenck, R. O., Harbeitner, R. C., Lawler, S. N., and Breitbart, M. (2015). Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Front. Microbiol.* 6:696. doi: 10.3389/fmicb.2015.00696

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. doi: 10.7717/peerj.985

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., et al. (2011). MetaVir: a web server dedicated to virome analysis. *Bioinformatics* 27, 3074–3075. doi: 10.1093/bioinformatics/btr519

Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* 3:130160. doi: 10.1098/rsob.130160

Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123. doi: 10.1101/gr.089532.108

Solonenko, S. A., and Sullivan, M. B. (2013). Preparation of metagenomic libraries from naturally occurring marine viruses. *Methods Enzymol.* 531, 143–165. doi: 10.1016/B978-0-12-407863-5.00008-3

Stadelmann, B., Herrera, L. G., Arroyo-Cabrales, J., Flores-Martinez, J. J., May, B. P., and Ruedi, M. (2004). Molecular systematics of the fishing bat Myotis (Pizonyx) vivesi. *J. Mammal.* 85, 133–139. doi: 10.1644/1545-1542(2004)085<0133:MSOTFB>2.0.CO;2

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* 57, 758–771. doi: 10.1080/10635150802429642

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Tangherlini, M., Dell'Anno, A., Allen, L. Z., Riccioni, G., and Corinaldesi, C. (2016). Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Sci. Rep.* 6:28428. doi: 10.1038/srep28428

Thompson, F. L., Barash, Y., Sawabe, T., Sharon, G., Swings, J., and Rosenberg, E. (2006). *Thalassomonas loyana* sp. nov., a causative agent of the white

plague-like disease of corals on the Eilat coral reef. *Int. J. Syst. Evol. Microbiol.* 56, 365–368. doi: 10.1099/ijs.0.63800-0

Tveit, A. T., Urich, T., and Svenning, M. M. (2014). Metatranscriptomic analysis of artic peat soil microbiota. *Appl. Environ. Microbiol.* 80, 5761–5772. doi: 10.1128/AEM.01030-14

van Dijk, E. L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322, 12–20. doi: 10.1016/j.yexcr.2014.01.008

Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M., and Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15:37. doi: 10.1186/1471-2164-15-37

Vikram, S., Guerrero, L. D., Makhalanyane, T. P., Le, P. T., Seely, M., and Cowan, D. A. (2015). Metagenomic analysis provides insights into functional capacity in a hyperarid desert soil niche community. *Environ. Microbiol.* 18, 1875–1888. doi: 10.1111/1462-2920.13088

Weynberg, K. D., Wood-Charlson, E. M., Suttle, C. A., and van Oppen, M. J. (2014). Generating viral metagenomes from the coral holobiont. *Front. Microbiol.* 5:206. doi: 10.3389/fmicb.2014.00206

Wright, I. A., and Travers, S. A. (2014). RAMICS: trainable, high-speed and biologically relevant alignment of high-throughput sequencing reads to coding DNA. *Nucl Acids Res.* 42:e106. doi: 10.1093/nar/gku473

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., et al. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic. Sci.* 6, 427–439. doi: 10.4056/sigs.2945050

Yi, H., Bae, K. S., and Chun, J. (2004). *Thalassomonas ganghwensis* sp. nov., isolated from tidal flat sediment. *Int. J. Syst. Evol. Microbiol.* 54, 377–380. doi: 10.1099/ijs.0.02748-0

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS ONE* 7:e33865. doi: 10.1371/journal.pone.0033865

Zablocki, O., Adriaenssens, E. M., and Cowan, D. (2016). Diversity and ecology of viruses in hyperarid desert soils. *Appl. Environ. Microbiol.* 82, 770–777. doi: 10.1128/AEM.02651-15

Zablocki, O., van Zyl, L., Adriaenssens, E. M., Rubagotti, E., Tuffin, M., Cary, S. C., et al. (2014). High-level diversity of tailed Phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. *Appl. Environ. Microbiol.* 80, 6888–6897. doi: 10.1128/AEM.01525-14

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.