Infection, Genetics and Evolution 31 (2015) 73–86

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond

Simona Kraberger [a], Gerardo R. Argüello-Astorga [b], Laurence G. Greenfield [a], Craig Galilee [a], Donald Law [c], Darren P. Martin [d], Arvind Varsani [a,e,f,g,*]

[a] School of Biological Sciences, University of Canterbury, Christchurch 8140, New Zealand
[b] División de Biología Molecular, Instituto Potosino de Investigación Científica y Tecnológica, Camino a la Presa San José 2055, 78216 San Luis Potosí, S.L.P., Mexico
[c] The Laboratories, Christchurch City Council, Christchurch, New Zealand
[d] Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa
[e] Electron Microscope Unit, Division of Medical Biochemistry, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, 7701 Cape Town, South Africa
[f] Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, FL 32611, USA
[g] Biomolecular Interaction Centre, University of Canterbury, Christchurch 8140, New Zealand

ARTICLE INFO

Article history:
Received 4 November 2014
Received in revised form 23 December 2014
Accepted 2 January 2015
Available online 9 January 2015

Keywords:
Circular ssDNA viruses
Treated sewage
Gemycircularvirus
Next-generation sequencing
Viral metagenomics

ABSTRACT

Our knowledge of circular replication-associated protein encoding single-stranded (CRESS) DNA virus diversity has increased dramatically in recent years, largely due to advances in high-throughput sequencing technologies. These viruses are apparently major virome components in most terrestrial and aquatic environments and it is therefore of interest to determine their diversity at the interfaces between these environments. Treated sewage water is a particularly interesting interface between terrestrial and aquatic viromes in that it is directly pumped into waterways and is likely to contain virus populations that have been strongly impacted by humans. We used a combination of high-throughput sequencing, full genome PCR amplification, cloning and Sanger sequencing to investigate the diversity of CRESS DNA viruses present in a sewage oxidation pond. Using this approach, we recovered 50 putatively complete novel CRESS viral genomes (it remains possible that some are components of multipartite viral genomes) and 11 putatively sub-genome-length circular DNA molecules which may be either defective genomes or components of multipartite genomes. Thirteen of the genomes have bidirectional genome organisations and share similar conserved replication-associated protein (Rep) motifs to those of the gemycircularviruses: a group that in turn is most closely related to the geminiviruses. The remaining 37 viral genomes share very low degrees of Rep similarity to those of all other known CRESS DNA viruses. This number of highly divergent CRESS DNA virus genomes within a single sewage treatment pond further reinforces the notion that there likely exist hundreds of completely unknown genus/family level CRESS DNA virus groupings.

1. Introduction

Sewage is a biological sink for a wide variety of infectious agents including viruses. Consisting largely of human excrement, sewage harbours a rich diversity of viruses. Besides viruses that infect the multitude of environmental microbes which degrade faeces, sewage also contains viruses infecting humans, their gut flora, and their food (Blinkova et al., 2009; Cantalupo et al., 2011; Metcalf et al., 1995; Ng et al., 2012; Parsley et al., 2010; Rosario et al., 2009c; Symonds et al., 2009; Tamaki et al., 2012). Secondary stages of sewage treatment involve aeration in 'oxidation ponds' which might also introduce viruses that infect birds, algae, fungi, insects and aerobic bacteria into the treated sewage water that is ultimately discharged into the natural environment.

To date virus research on sewage systems has predominantly focused on viruses of clinical importance to humans, such as poliovirus and other enteroviruses (Blomqvist et al., 2004; Hewitt et al., 2011; Katayama et al., 2008; Lodder and de Roda Husman, 2005; Symonds et al., 2009; Vaidya et al., 2002). Only a handful of studies have taken an unbiased metagenomics-based approach to study viral diversity associated with either raw sewage (Cantalupo et al., 2011; Ng et al., 2012; Tamaki et al., 2012) or reclaimed water (Rosario et al., 2009b). These revealed that sewage contains genetic

* Corresponding author at: School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand. Tel.: +64 3 366 7001x4667; fax: +64 3 364 2590.
E-mail address: arvind.varsani@canterbury.ac.nz (A. Varsani).

http://dx.doi.org/10.1016/j.meegid.2015.01.001
1567-1348/

material derived from a wide variety of vertebrate, invertebrate and plant-infecting viruses. A significant proportion of this genetic material encodes proteins that share low, but nevertheless significant, degrees of similarity to those encoded by circular single-stranded DNA (ssDNA) viruses in the families *Geminiviridae*, *Circoviridae* and *Nanoviridae*. One study identified as much as 30% of viral-related sequences assembled from a metagenomic analysis of raw sewage shared significant similarities to proteins of these three ssDNA families (Ng et al., 2012). If this genetic material is indeed viral-derived, much of it represents completely novel ssDNA virus groups. For example, three complete ssDNA genomes that were recently isolated from raw sewage (named Nepavirus, Nimivirus and Baminivirus) encode replication-associated proteins (Reps) that are most closely related to, but clearly distinct from, Rep proteins expressed by plant-infecting geminiviruses and each likely represents an undescribed genus/family level taxonomic grouping (Ng et al., 2012). Similarly, divergent ssDNA viruses have been found in faecal samples from humans (Castrignano et al., 2013), caribou (Ng et al., 2014), sheep (Sikorski et al., 2013b), New Zealand fur seals (Sikorski et al., 2013a), pigs , foxes (Bodewes et al., 2013), bats (Castrignano et al., 2013; Ge et al., 2011, 2012; Li et al., 2010b), rodents (Phan et al., 2011), mustelids (Smits et al., 2013; van den Brand et al., 2012) and birds (Phan et al., 2013; Reuter et al., 2014; Sikorski et al., 2013b).

Several of these novel ssDNA viruses share similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (SsHADV-1), a virus species found in benthic river sediments (Kraberger et al., 2013) and was first isolated from the fungus, *S. sclerotiorum* (Yu et al., 2010). The Rep of SsHADV-1 contains geminivirus-like Rep motifs (Dayaram et al., 2012; Nash et al., 2011). Viral genomes related to SsHADV-1 have also been recovered from animal faecal samples (Sikorski et al., 2013b; van den Brand et al., 2012), dragonflies, mosquitoes (Ng et al., 2011b; Rosario et al., 2012a), bovine and human serum (Lamberto et al., 2014) and plant material (Dayaram et al., 2012; Du et al., 2014) and, as a result, a new genus known as gemycircularviruses has been proposed for this group of Geminivirus-like, possibly fungal-infecting, viruses. Novel ssDNA viruses which have Reps that share similarity to circoviruses and cyclovirus proteins have also been identified in faecal samples (Ge et al., 2011; Li et al., 2010a,b).

The genomes of monopartite eukaryote-infecting circular ssDNA viruses (such as those in the families *Circoviridae* and *Geminiviridae*) typically have at least two open reading frames (ORF): one encoding a Rep and the other encoding a coat protein (CP). Multipartite ssDNA virus genomes, such as those found in members of the *Nanoviridae*, are comprised of up to eight individual genome components, with each component encoding a single protein; of those components at least one encodes a Rep and another a CP. The CPs of viruses are often highly diverse, possibly because they are involved in interactions with host and/or vector species cell surface receptors. In many instances the putative Rep encoded by divergent environmental circular Rep encoding single-stranded (CRESS) DNA viruses is the only protein that has any detectable homology to other known ssDNA virus Rep proteins. This similarity together with the presence of evolutionarily conserved signature motifs within the Reps encoded by these molecules that have proven functionality during the replication of well characterised CRESS DNA viruses (Rosario et al., 2012b) is the strongest evidence that these environmental ssDNA molecules are in fact either virus genomes or components thereof.

Sampling sewage provides a convenient and non-invasive approach to studying viral diversity within human impacted environments. As mentioned previously, aeration of treated sewage occurs in oxidation ponds where clarified sewage is circulated for approximately two weeks before being discharged into the ocean or rivers. This open-air stage allows algal growth and UV exposure, both of which reduce coliform populations (Abdel-Raouf et al., 2012; Sinton et al., 2002). Despite a number of viral studies on raw sewage, only one study (Rosario et al., 2009b) has looked at viral diversity in treated sewage prior to discharge. To address this lack of knowledge, we used a viral metagenomic sequence-informed approach to determine the diversity of CRESS DNA viruses within treated sewage oxidation ponds. We find a rich diversity of CRESS DNA viruses associated within this environment and describe the recovery and characterisation of 50 novel CRESS DNA virus genomes and 11 sub-genome-length circular DNA molecules that may be either defective genomes or individual genome components of viruses belonging to divergent groups of multipartite CRESS DNA viruses.

2. Materials and methods

2.1. Sample collection and viral DNA isolation

A sewage oxidation pond (final 'open air' stage of treatment) water sample was collected at the Christchurch Wastewater Treatment Plant, Christchurch, New Zealand in September, 2012. 50 ml of the sample was successively passed through 0.45 μm and 0.2 μm syringe filters (Sartorius Stedim Biotech, Germany). The filtrate was precipitated using 15% PEG at 4 °C overnight and pelleted by centrifugation at 10,000*g* for 10 min. The resulting pellet was resuspended in 1 ml of SM buffer [0.1 M NaCl, 50 mM Tris–HCl (pH 7.4)]. Nucleic acid was extracted from 200 μl of this re-suspension using the High Pure Viral Nucleic Acid Kit (Roche Diagnostics, USA). The isolated viral nucleic acids were enriched using TempliPhi™ (GE Healthcare, USA).

2.2. Next-generation sequencing-informed recovery of complete viral genomes

Enriched viral DNA was sequenced at the Beijing Genomics Institute (Hong Kong) using an Illumina HiSeq 2000 sequencer. The paired end reads were assembled using ABySS 1.3.5 (Simpson et al., 2009) with a k-mer setting of 64. Contigs >1000 nt were analysed by BLASTx (Altschul et al., 1990) for detectable homology to known viral proteins.

Abutting primers were designed (Sup. Table 1) to recover complete circular genomes for contigs found to have credibly high degrees of similarity to known viruses (BLASTx E-scores $<10^{-7}$; (Altschul et al., 1990). The circular genomes were recovered using polymerase chain reaction with KAPA HiFi Hotstart DNA polymerase (Kapa Biosystems, USA) with the specific abutting primers using the following thermocycler program: 94 °C for 3 min, 25 cycles of 98 °C (20 s), 55 °C (30 s), 72 °C (3 min) and a final extension of 72 °C for 3 min. The PCR amplicons were gel purified and ligated into pJET1.2 plasmid (Thermo Fisher Scientific, USA) and sequenced at Macrogen Inc. (Korea) by Sanger sequencing using primer walking.

Sanger sequencing reads were assembled using DNA Baser (Heracle BioSoft S.R.L. Romania). Putative CP and Rep ORFs were identified and preliminary genome analysis was carried using BLASTx (Altschul et al., 1990) against the non-redundant database in GenBank. Pairwise similarity comparisons (1 – p-distance, with pairwise deletion of gaps) of Rep amino acid sequences predicted to be expressed by circular CRESS DNA viruses obtained in this study along with those available in GenBank (as of 14th Sept 2014), were determined using SDT v1.2 (Muhire et al., 2014).

2.3. Phylogenetic analyses

The Rep amino acid sequences potentially encoded by the newly determined CRESS virus genomes together with those