Diverse circovirus-like genome architectures revealed by environmental metagenomics

Karyna Rosario,1 Siobain Duffy2 and Mya Breitbart1

1College of Marine Science, University of South Florida, St Petersburg, FL 33701, USA
2School of Environmental and Biological Sciences, Rutgers, New Brunswick, New Jersey, USA

Single-stranded DNA (ssDNA) viruses with circular genomes are the smallest viruses known to infect eukaryotes. The present study identified 10 novel genomes similar to ssDNA circoviruses through data-mining of public viral metagenomes. The metagenomic libraries included samples from reclaimed water and three different marine environments (Chesapeake Bay, British Columbia coastal waters and Sargasso Sea). All the genomes have similarities to the replication (Rep) protein of circoviruses; however, only half have genomic features consistent with known circoviruses. Some of the genomes exhibit a mixture of genomic features associated with different families of ssDNA viruses (i.e. circoviruses, geminiviruses and parvoviruses). Unique genome architectures and phylogenetic analysis of the Rep protein suggest that these viruses belong to novel genera and/or families. Investigating the complex community of ssDNA viruses in the environment can lead to the discovery of divergent species and help elucidate evolutionary links between ssDNA viruses.

Single-stranded DNA (ssDNA) viruses that replicate through a rolling circle mechanism are the smallest viruses (<3 kb) known to infect plants and animals. Circular ssDNA viruses are mechanistically predisposed to recombination (Lefeuvre et al., 2009) and experience high nucleotide substitution rates (Duffy et al., 2008; van der Walt et al., 2008). These characteristics may contribute to the emergence of ssDNA viruses as serious pathogens (Hou & Gilbertson, 1996; Lefeuvre et al., 2009; Ma et al., 2007; Varsani et al., 2008; Zhou et al., 1997). Circular ssDNA eukaryotic viruses include plant pathogens from the families Nanoviridae and Geminiviridae and animal viruses from the genera Circovirus, Gyrivirus and Anellovirus (de Villiers & zur Hausen, 2009; Gronenborn, 2004; Gutiérrez, 1999; Mankertz et al., 1997; Navidad et al., 2008; Noteborn & Koch, 1995; Okamoto & Mayumi, 2001). Although circular ssDNA viruses include pathogens of agricultural, veterinary and clinical concern (Almeida et al., 2009; Chae, 2005; de Villiers & zur Hausen, 2009; Moffat, 1999; Rishi, 2009; Todd, 2000), little is known about their prevalence or diversity in natural environments.

To date, most of the research regarding eukaryotic circular ssDNA viruses has focused on recombinant expression or specific detection methods, such as PCR, to further characterize known, economically important pathogens or detect related viruses in known hosts (e.g. Almeida et al., 2009; Banda et al., 2007; Johne et al., 2004, 2006; Kakkola et al., 2007; Mankertz et al., 2000; Ninomiya et al., 2009; Seal et al., 2006; Zhou et al., 2005). However, the incorporation of multiple displacement amplification (MDA) into metagenomic analyses provides a method to detect ssDNA viruses in environmental viral communities. MDA can be used to enrich for ssDNA circular genomes as it can selectively amplify circular ssDNA genomes by two or three orders of magnitude in a mixed community (Haible et al., 2006; Kim et al., 2008). A number of recent viral metagenomic studies utilizing an MDA step have identified novel sequences related to circular ssDNA eukaryotic viruses in different samples, including rice paddy soil (Kim et al., 2008), reclaimed water (Rosario et al., 2009) and marine animal tissues (Ng et al., 2009a, b; Vega Thurber et al., 2008). Therefore, small circular ssDNA viruses may be more widespread in the environment than previously recognized.

This study identified circular ssDNA viral genomes in the environment through data-mining of public viral metagenomes generated with an MDA step. Specifically, ssDNA viruses were found in reclaimed water (RW) (Rosario et al., 2009) and three different marine environments, including the Chesapeake Bay (CB) (MOVE858 shotgun dataset), British Columbia coastal waters (BBC) (Angly et al., 2006) and the Sargasso Sea (SAR) (Angly et al., 2006). These datasets were chosen because viral concentrates from these environments were available for confirmation of metagenomic sequence assemblies through PCR. For metagenomic sequence assemblies, the reclaimed water dataset was retrieved from the NCBI short read archive (accession no. SRA008294.7), the MOVE858 dataset was collected from the community cyberinfrastructure for advanced

A supplementary table of primer sequences and the PCR conditions that were used are available with the online version of this paper.
marine microbial ecology research and analysis (CAMERA) website (http://camera.calit2.net/index.php), while the BBC (SEED accession no. JF959077–JF959086) and SAR (SEED accession no. JF959084) metagenomes were obtained from the SEED platform (http://www.theseed.org/DinsdaleSupplementalMaterial; Dinsdale et al., 2008). Sequences from each dataset were assembled using the SeqMan program (DNASTAR) with a criteria of >95% identity over at least 35 nt. To identify eukaryotic circular ssDNA viruses, contigs larger than 1000 nt were compared against GenBank using BLASTX (E-value ≤0.001) and results were summarized using the metagenome analyser (MEGAN) software (Huson et al., 2007).

All of the datasets contained hundreds of sequences that assembled into contigs with similarities to viruses in the genus Circovirus. Therefore, this study focused on circovirus-like sequences. It is important to note that contigs similar to nanoviruses were detected in the SAR, CB and RW datasets and geminivirus-like sequences were present in the CB and RW datasets. No anellovirus- or gyrovirus-like sequences were detected in any of the datasets. Contigs containing complete circular genomes with similarities to circoviruses were verified by PCR (see Supplementary Table S1, available in JGV Online) and are described further below. Any contigs that contained only partial genomes or that could not be confirmed by PCR (possibly due to chimeras) were excluded from the analysis. Since only a small portion of the circovirus-like contigs met these criteria, this study is a conservative estimate of the novelty and diversity of circoviruses in the environment. In addition, there was a high degree of sequence variability between clones consistent with the “quasispecies” diversity observed in ssDNA viruses (Ng et al., 2009a).

Ten circovirus-like genomes, ranging in size from 1739 to 2819 nt, were reconstructed from reclaimed water and marine metagenomic sequences (Fig. 1; GenBank accession nos JF959077–JF959086). All the genomes contained a major open reading frame (ORF) with similarities to the viral replication protein (Rep) of circoviruses. According to the Pfam database classification (Finn et al., 2008; http://Pfam.sanger.ac.uk/search), all of the putative Reps contained a domain similar to the viral Rep family (PF02407), but some Reps also had similarities to an RNA helicase domain (PF00910) (Table 1). Pfam analysis revealed that some known circoviruses only have the viral Rep domain (e.g. canary, finch, columbid and raven circoviruses) while others have both the viral Rep and the RNA helicase domains (e.g. bovine, porcine, beak and feather disease, and duck circoviruses). Overall, amino acid identities to known Rep proteins were less than 40% for all the environmental circovirus-like genomes, suggesting that these are novel viruses (Table 1).

A phylogenetic analysis was performed on the viral Reps to evaluate the relationship between the novel circovirus-like genomes and known viruses. For this purpose, deduced Rep sequences from environmental circovirus-like genomes were aligned against members of the viral Rep family (PF02407) in the Pfam database. This protein family also includes non-viral replication-associated proteins from a plasmid (Bifidobacterium pseudocatenulatum pM4) and protists (Giardia intestinalis and Entamoeba histolytica) (Gibbs et al., 2006). The CB_B (GenBank accession no. JF959083) Rep was not included in the alignment as this sequence was too divergent and missed several conserved amino acids identified through the Hidden Markov model logo for the viral Rep family (Schuster-Böckler et al., 2004). Although the SAR_A (GenBank accession no. JF959084) Rep had several conserved amino acids present in the viral Rep family, this sequence was also excluded due to its small size (180 aa). Amino acid sequence alignments indicate that the novel circular genomes are more related to circoviruses than to nanoviruses (Fig. 2). None of the genomes clustered with known circovirus, plasmid or protist Rep proteins, suggesting that these viruses belong to novel families and/or genera of ssDNA viruses.

Although all the genomes have similarities to the Rep protein of circoviruses, only half of them contained other genomic features consistent with known circoviruses. Circoviruses are characterized by a small (~2063 nt), circular ssDNA genome that contains two major ORFs, encoding the Rep and capsid proteins (Cap), in an ambisense organization (Todd et al., 2005). The only genomes with this organization are RW_A, RW_B, RW_C, SAR_A and BBC_A (Type I; Fig. 1). These genomes contain two major ORFs, a putative Rep and an unknown ORF (no significant homologues in GenBank, E-value <0.001) that are divergently organized (Fig. 1). Known circoviruses have an origin of replication located upstream of both major ORFs and this region contains the nonanucleotide motif at the apex of a stem–loop (Mankertz, 2008; Todd, 2000). RW_B, RW_C and SAR_A contain a conserved nonanucleotide motif (TAGTATTAC) at the apex of a potential stem–loop. Genomes RW_A and BBC_A contain a similar motif (Table 1). Although RW_C and BBC_A genomes contained a genome organization similar to circoviruses, the intergenic region containing the potential stem–loop is located downstream from major ORFs (Fig. 1).

In contrast with the genomes discussed above, genomes RW_D, RW_E, SAR_B, CB_A and CB_B (Types II and III; Fig. 1) do not have genomic features consistent with circoviruses. The CB_B genome (Type II) contained broad organizational similarities to circoviruses, but the Rep is split into three overlapping ORFs and no potential stem–loops were identified near the conserved nonanucleotide sequence (Fig. 1). The ORF upstream (CB_BRep1) had hits to the viral Rep in the Pfam database, the middle ORF (CB_BRep2) did not have any significant matches, while the partial ORF downstream (CB_BRep3) had hits to the RNA helicase family (E-value=0.0053; Table 1). This suggests that the Rep domains of CB_B are in an order consistent with circovirus Reps. Furthermore CB_BRep3 had significant hits to the Rep protein of porcine
circoviruses (PCV) in GenBank (Table 1). PCV Reps, which contain the viral Rep–RNA helicase architecture, are differentially transcribed to produce two proteins essential for virus replication, Rep and Rep’ (Cheung, 2003, 2004; Mankertz, 2008). PCV transcriptional splicing results in the removal of an intron and the expression of the Rep’ C terminus in a different reading frame (Mankertz & Hillenbrand, 2001). It is possible that CB_B has a similar
**Table 1.** Circovirus-like features identified in the environmental genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Nonanucleotide sequence</th>
<th>Rep family domains identified in Pfam (E-value)</th>
<th>Best hit*</th>
<th>Arginine/lysine-rich region†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Top hit (E-value)</td>
<td>Amino acid</td>
</tr>
<tr>
<td>RW_A</td>
<td>CAGTATTAC</td>
<td>Viral Rep (1 × 10⁻¹⁸) RNA helicase (7.2 × 10⁻⁴)</td>
<td>Canary circovirus (1.3 × 10⁻¹²)</td>
<td>36</td>
</tr>
<tr>
<td>RW_B</td>
<td>TAGTATTAC</td>
<td>Viral Rep (4.3 × 10⁻²⁶)</td>
<td>Porcine circovirus 1 (1.3 × 10⁻¹⁰)</td>
<td>33</td>
</tr>
<tr>
<td>RW_C</td>
<td>TAGTATTAC</td>
<td>Viral Rep (1.4 × 10⁻⁰⁶)</td>
<td>Columbid circovirus (6.5 × 10⁻²⁰)</td>
<td>38</td>
</tr>
<tr>
<td>RW_D</td>
<td>AAGTATTAC</td>
<td>Viral Rep (5.2 × 10⁻¹⁵)</td>
<td>Porcine circovirus 2 (4.8 × 10⁻¹⁰)</td>
<td>32</td>
</tr>
<tr>
<td>RW_E</td>
<td>AAGTATTAC</td>
<td>Viral Rep (6.5 × 10⁻⁰⁵) RNA helicase (7.9 × 10⁻⁰³)</td>
<td>Porcine circovirus 2 (1.3 × 10⁻¹⁰)</td>
<td>31</td>
</tr>
<tr>
<td>CB_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (5.2 × 10⁻¹⁵)</td>
<td>Gull circovirus (1.3 × 10⁻⁰²)</td>
<td>32</td>
</tr>
<tr>
<td>CB_B</td>
<td>TAGTATTAC</td>
<td>ORF1: Viral Rep (8.6 × 10⁻⁰⁵) ORF3: RNA helicase (5.3 × 10⁻⁰³)</td>
<td><em>Gossypium darwinii</em> symptomless virus (5.5 × 10⁻¹⁰)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Porcine circovirus 1 (5.2 × 10⁻¹⁰)</td>
<td>44</td>
</tr>
<tr>
<td>SAR_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (6.2 × 10⁻¹⁶)</td>
<td>Gull circovirus (1.1 × 10⁻¹⁴)</td>
<td>56</td>
</tr>
<tr>
<td>SAR_B</td>
<td>TAGTATTAC</td>
<td>Viral Rep (3.7 × 10⁻¹⁵)</td>
<td>Starling circovirus (2.3 × 10⁻¹¹)</td>
<td>40</td>
</tr>
<tr>
<td>BBC_A</td>
<td>TAGTATTAC</td>
<td>Viral Rep (2.5 × 10⁻¹⁰)</td>
<td>Finch circovirus (1.3 × 10⁻¹⁶)</td>
<td>28</td>
</tr>
</tbody>
</table>

*Best hit in GenBank using Rep as the query.
†The arginine/lysine-rich region is present (P) or not present (NP) at the N terminus of the major unknown ORF.

Replication mechanism, needing two differentially transcribed proteins.

Similar to known circoviruses, the type III genomes (RW_D, RW_E, SAR_B and CB_A) were also characterized by two major ORFs, a putative Rep and an unknown ORF. However, these ORFs are non-overlapping and organized in the same orientation (Fig. 1). These genomes contained the conserved nonanucleotide (CB_A and SAR_B) or a similar sequence (RW_D and RW_E) (Table 1) located within a potential stem-loop. These genome organizations have not been observed in circoviruses or other eukaryotic circular ssDNA viruses. On the other hand, these genome organizations are characteristic of linear ssDNA (4–6 kb) viruses from the family Parvoviridae. Some parvovirus genomes possess two major ORFs, encoding the Rep and Cap proteins, with the same polarity (Tattersall et al., 2005). In addition, other parvoviruses have the Rep and Cap proteins encoded on complementary strands and the Rep is split into three minor ORFs (Tattersall et al., 2005). Based on similarities to the Rep of circoviruses but genome organizations that have only been observed in linear ssDNA viruses, the type II and III circular genomes most likely represent members of novel virus families.

Extrapolating from similar genome organizations in known ssDNA viruses, the unknown ORFs in the circovirus-like genomes may encode a Cap. Caps of many eukaryotic ssDNA viruses contain a region rich in basic amino acids at the N terminus (Niagro et al., 1998). The RW_A, RW_C and CB_A genomes had a lysine/arginine-rich region at the N terminus of the major unknown ORF which is characteristic of known circovirus Caps (Table 1). A search in Pfam revealed similarities to circovirus capsid proteins (PF02443) for the RW_A and RW_C genomes; however, the matches were not well supported (E-values=0.09 and 0.004, respectively). The major unknown ORF from RW_B had weak similarities (E-value=0.006) to geminivirus coat proteins (PF00844). This is interesting since the putative Rep from this genome had significant matches to the Rep from circoviruses (PF02447). Although geminiviruses and circoviruses have similar genome organizations (Niagro et al., 1998), they belong to different families and infect different hosts. None of the unknown ORFs in the other circovirus-like genomes had similarities to known Pfam protein families. Although the functions of these unknown ORFs are still undefined, this study places these sequences into circovirus-like genomes, thus assigning a portion of the large percentage of unknown sequences in metagenomic surveys (Angly et al., 2006) to genomes. This process may help identify divergent structural viral genes in the future.

Five of the novel circovirus-like genomes identified in this study originated from reclaimed water, the end product of wastewater treatment. The discovery of circovirus-like genomes in reclaimed water suggests that these viruses can be disseminated through the discharge of treated wastewater (Rosario et al., 2009). Although circovirus-like genomes have also been identified in stool from children (Victoria et al., 2009), the source of these viruses in reclaimed water remains to be determined.
The five marine circovirus-like genomes originated from an estuarine environment (CB), coastal waters (BBC) and the open ocean (SAR). To our knowledge, the only known marine virus with similarities to circoviruses is the Chaetoceros salsugineum nuclear inclusion virus (CsNIV; Nagasaki et al., 2005). The inferred Rep protein sequence of CsNIV had weak hits to bird circoviruses. However, CsNIV is not similar to known circoviruses in terms of size and genome organization (Nagasaki et al., 2005; Park et al., 2009). In contrast with CsNIV, the circovirus-like SAR, BBC and CB genomes described in this study have stronger hits to circovirus Rep proteins, have similar genome sizes to known circoviruses and some have similar genome organization. Therefore, the novel genomes found in this study...

Fig. 2. Condensed maximum-parsimony phylogenetic tree of deduced Rep amino acid sequences showing the relationship between reclaimed water (RW), Chesapeake Bay (CB), Sargasso Sea (SAR) and British Columbia coastal waters (BBC) genomes and members of the viral Rep protein family (PF02407). The tree includes a putative Rep sequence from a circovirus-like genome from soil. Alignments were performed using the CLUSTALW algorithm (Thompson et al., 1994) and BLOSUM62 as the similarity matrix in BioEdit version 7.0.9.0 (Hall, 1999). All sequences were trimmed to match a lysine residue at position 15 of the porcine circovirus 2 translated rep sequence (NC_005148) and the arginine residues at positions 276–277 of the same sequence. Alignments were performed over at least 200 aa and inspected manually. Phylogenetic trees were constructed in MEGA 4 (Tamura et al., 2007) using the close-neighbour-interchange algorithm (CNI=3) with random addition of sequences (10 replicates). All alignment gaps were treated as missing data. The tree was manually rooted between the nanovirus and circovirus clades. One thousand bootstrap resamplings were performed to assess statistical support (only bootstrap values >70 are shown). Samples isolated for this study are printed in a larger font.
study are the first circo-like viruses identified in the marine environment.

The data presented here demonstrate the diversity of circovirus-like genome architectures in the environment. Intriguingly, some of the genomes revealed a mixture of genomic features associated with different families of ssDNA viruses. The abundance of circovirus-like sequences in environmental metagenomic studies and the presence of unique genome sequences and architectures suggest that there is a complex community of ssDNA viruses in the environment. Investigating these communities may lead to the discovery of divergent species that could illuminate the evolutionary links between ssDNA viruses. Future studies need to continue to explore the diversity of circular ssDNA viruses and investigate the ecology of these novel viruses.

Acknowledgements

This research was funded by a grant to M.B. from the National Science Foundation (MCB-0701984). K. R. was funded by a Bridge to the Doctorate Fellowship from the National Science Foundation and the Saint Petersburg Progress Endowed Fellowship. The authors would like to acknowledge Eric Wommack and Danielle Winget for providing viral concentrates from the Chesapeake Bay.

References


