Virus Research 219 (2016) 51–57

Contents lists available at ScienceDirect

Virus Research

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

Comparison of Illumina *de novo* assembled and Sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus *Sclerotinia sclerotiorum*

Mahmoud E. Khalifa[a,f], Arvind Varsani[b,c,d], Austen R.D. Ganley[e], Michael N. Pearson[a,*]

[a] *School of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand*
[b] *Biomolecular Interaction Centre and School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand*
[c] *Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, 7701 Cape Town, South Africa*
[d] *Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, Florida, FL 32611, USA*
[e] *Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand*
[f] *Faculty of Sciences, Damietta University, Damietta, Egypt*

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 21 October 2015
Accepted 1 November 2015
Available online 12 November 2015

Keywords:
Mycovirus
Virus detection
De novo viral genome assembly
Next generation sequencing
Illumina sequencing

ABSTRACT

The advent of ‘next generation sequencing’ (NGS) technologies has led to the discovery of many novel mycoviruses, the majority of which are sufficiently different from previously sequenced viruses that there is no appropriate reference sequence on which to base the sequence assembly. Although many new genome sequences are generated by NGS, confirmation of the sequence by Sanger sequencing is still essential for formal classification by the International Committee for the Taxonomy of Viruses (ICTV), although this is currently under review. To empirically test the validity of *de novo* assembled mycovirus genomes from dsRNA extracts, we compared the results from Illumina sequencing with those from random cloning plus targeted PCR coupled with Sanger sequencing for viruses from five *Sclerotinia sclerotiorum* isolates. Through Sanger sequencing we detected nine viral genomes while through Illumina sequencing we detected the same nine viruses plus one additional virus from the same samples. Critically, the Illumina derived sequences share >99.3 % identity to those obtained by cloning and Sanger sequencing. Although, there is scope for errors in *de novo* assembled viral genomes, our results demonstrate that by maximising the proportion of viral sequence in the data and using sufficiently rigorous quality controls, it is possible to generate *de novo* genome sequences of comparable accuracy from Illumina sequencing to those obtained by Sanger sequencing.

1. Introduction

The introduction of the Roche GS-FLX 454 sequencer in 2004, heralded the start of the ‘next generation sequencing’ (NGS) revolution. Subsequently, other commercial NGS technologies have emerged, making DNA sequencing more affordable (Boria et al., 2013; Voelkerding et al., 2009) and substantially increasing throughput compared to Sanger sequencing (Chan et al., 2013). These NGS technologies have enabled an enormous acceleration in genomic research, including virus research where they have been applied to genome sequencing, analysing polymorphisms within a population, viral ecology, the detection of known viruses in ecosystems, and viral transcriptomics (Radford et al., 2012). The ability of NGS platforms to generate extensive viral sequence data is transforming the field of virology and enabling the detection and identification of many previously unknown viruses and viroids (Barba et al., 2014; Roossinck, 2015; Roossinck et al., 2015; Massart et al., 2014). For example, almost complete genome sequences of known and novel mycoviruses were discovered in RNA viral metagenomic studies of grapevine (Al Rwahnih et al., 2011; Coetzee et al., 2010). These methods can provide sequence for complete (or near complete) genomes, not only by scaffold-based sequence assembly but also by *de novo* assembly.

Complete sequencing of viral genomes using conventional sequencing methods typically requires cloning and sequencing of many DNA/cDNA fragments, a process that is time, effort and resource consuming and often requires designing sequence-specific primers to fill in sequence gaps. The method most often used to sequence plant and fungal RNA viral genomes is *via* a dsRNA approach that requires sufficiently high RNA abundance to

* Corresponding author.
E-mail address: m.pearson@auckland.ac.nz (M.N. Pearson).

http://dx.doi.org/10.1016/j.virusres.2015.11.001
0168-1702/© 2015 Elsevier B.V. All rights reserved.

be visualised on gels, with the consequence that viruses with low titres are not sampled. There is little doubt that NGS approaches have vastly improved our ability to detect the presence of virus sequences, including many that were previously unknown. However, there is continuing debate about the validity and reliability of draft genome sequences derived solely from *de novo* assembly of short sequence fragments generated from current NGS platforms, especially for formal classification of highly divergent novel viral genomes/sequences. The acceptance of genome sequences based only on *de novo* assembled sequences for taxon assignment is currently being reviewed by the International Committee for the Taxonomy of Viruses (ICTV). Nonetheless, many viruses, such as persistent viruses belonging to the *Chrysoviridae*, *Endornaviridae, Partitiviridae* and *Totivirdiae* in uncultivated plants (Roossinck et al., 2015), are yet to be assigned to appropriate taxons. The extent to which NGS technologies are being adopted in virology makes it imperative to understand how reliable the genome sequences generated by these technologies are and under what circumstances *de novo* assembled sequences are acceptable without confirmation by conventional sequencing and/or supporting biological evidence. To empirically test the validity of *de novo* assembled viral genomes, we compared complete mycoviral genome sequences derived from Illumina sequencing with those from random cloning plus targeted PCR coupled with Sanger sequencing, from five isolates of the plant pathogenic fungus *Sclerotinia sclerotiorum*.

2. Materials and methods

2.1. Source of the viral nucleic acids

Twenty two New Zealand *S. sclerotiorum* isolates were screened for the presence of high molecular weight dsRNAs using CF11 chromatography as described by Valverde et al. (1990). The dsRNAs were processed by random primed RT-PCR, cloned into pGEM-T easy vector (Promega, USA), transformed into *Escherichia coli* DH5α (Invitrogen, USA), and Sanger sequenced. The terminal sequences were determined by ligating adapter T4L (5′-PO_4-CCCGTCGTTTGCTGGCTCTTT-NH_2-3′) to the 3′ end of the dsRNAs using T4 RNA ligase (Promega, USA), as described by the manufacturer. The genome sequences obtained by this approach were previously published by Khalifa and Pearson (2013, 2014a,b,c).

2.2. Illumina sequencing

From the original dsRNA samples used for Sanger sequencing (see Section 2.1), dsDNA fragments of random lengths were generated, essentially as described in Khalifa and Pearson (2013, 2014a), using a common PCR primer but with a different 4-nt tag (barcode) for each separate sample (fungal isolate) at the 5′-proximal end (Roossinck et al., 2010) and with a PCR extension step of 30 s instead of 2 min. Fragments longer than 100 bp were purified using an Agencourt AMPure XP PCR purification kit (Beckman Coulter, USA), as described by the manufacturer. Purified fragments were quantified using a Qubit 2.0 Fluorometer (Invitrogen, USA) and equimolar concentrations of DNA from each sample were pooled and adjusted to a final concentration of 100 ng/μl. Illumina sequencing was performed by Macrogen Inc. (Seoul, South Korea) using an Illumina HiSeq2000 (Illumina, USA) platform with ~100 nt paired end reads.

2.3. Bioinformatic analysis

Illumina sequence reads were imported into Geneious 5.6.5 (Drummond et al., 2011) and assigned to individual fungal isolates based on their barcodes, and the primer and barcode sequences trimmed. Reads with quality scores of less than Q20, as determined by The Galaxy Project server (Goecks et al., 2010), were filtered out and the remaining reads were trimmed to remove unreliable 5′ sequence based on the FastQC report (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The filtered reads were assembled using the Geneious 5.6.5 *de novo* assembly tool set to medium sensitivity and default parameters. Consensus sequences of all contigs generated from the assembly of reads for each dataset were imported into the Galaxy Project and contigs shorter than 200 nt were discarded. Viral-like contigs were identified by BLASTX (Altschul et al., 1990) analysis against the non-redundant (nr) database of NCBI using Blast2GO software (Conesa et al., 2005) with an *E*-value cut-off of 1×10^{-6}.

2.4. Confirmation of de novo assembled contigs generated from Illumina sequencing by PCR and Sanger sequencing

Consensus sequences of assembled contigs and genomes that matched viral sequences identified by BLASTX (Altschul et al., 1990) searches were categorized by the closest hit and used to design primers for their subsequent detection by RT-PCR (Supplementary Table S1). First strand cDNA synthesis, PCR and Sanger sequencing were conducted as described in Khalifa and Pearson (2013, 2014a).

Supplementry material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2015.11.001.

3. Results

3.1. DsRNA analyses and Sanger sequencing

DsRNAs from nine of the twenty two *S. sclerotiorum* isolates were selected for further analysis (Supplementary Fig. S1) of which five were chosen for Sanger sequencing of the individual bands (Table 1). The Sanger sequencing resulted in the identification of nine distinct viral genomes, several of which were found in more than one isolate, giving a total of fifteen viral sequence detections (Table 1). We have previously published the genomes (based on a minimum of three fold coverage) of *S. sclerotiorum* mitovirus 3 (SsMV3) and *S. sclerotiorum* mitovirus 4 (SsMV4) (Khalifa and Pearson 2013), as well as *S. sclerotiorum* mitovirus 5 (SsMV5), *S. sclerotiorum* mitovirus 6 (SsMV6) and *S. sclerotiorum* mitovirus 7 (SsMV7) (Khalifa and Pearson, 2014a). *S. sclerotiorum* mitovirus 2 (SsMV2) was described by Xie and Ghabrial (2012). The three other viruses detected are: (i) a virus closely related to *S. sclerotiorum* negative sense RNA virus 1 (SsNsRV-1), the only negative-sense mycovirus described so far (Liu et al., 2014); (ii) a novel hypovirus, *S. sclerotiorum* hypovirus 2 (SsHV2) (Khalifa and Pearson, 2014b); and (iii) a novel endornavirus, *S. sclerotiorum* endornavirus 1 (SsEV1) (Khalifa and Pearson, 2014c).

Supplementry material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2015.11.001.

3.2. Illumina sequencing, de novo assembly, and virus identification

To determine how Illumina sequencing of virus genomes compares to the traditional Sanger approach, we performed random RT-PCR on dsRNAs from nine *S. sclerotiorum* isolates (including the five that were Sanger sequenced). This resulted in abundant cDNA ranging from 150 to 500 nt. Illumina paired-end sequencing of the cDNAs from all nine isolates generated 50,224,769 reads of 101 nt, of which ~48% could be assigned to individual fungal isolates based on their barcodes (Table 2). However, the reads per isolate were quite variable, ranging from 143,843 (isolate 17,019) to >12 × 10^6 (isolate 11,691). Following trimming and removal of poor quality reads, the number of usable reads (Table 2) ranged from 476 (isolate 17,019) to >8 × 10^6 (isolate 11,691). Quality filtered reads from each