Australian Government
Department of Defence
Defence Science and
Technology Organisation

Characterisation and Next-generation Sequencing Analysis of Unknown Arboviruses

Jane McAllister, Penny Gauci and Ania Gubala

Human Protection and Performance Division
Defence Science and Technology Organisation

DSTO-TR-2744

ABSTRACT

In the last few years there has been an increased incidence of neurological and encephalitic syndromes in Australia that have not been linked to a known pathogen. Viruses from the families *Rhabdoviridae* and *Bunyaviridae* are postulated to play a role. This report describes the characterisation of two unknown bunyaviruses, designated Maprik (MPKV) and Buffalo Creek (BCV), and two unknown rhabdoviruses, designated Harrison Dam (HARDV) and Holmes Jungle (HOJV), using techniques such as PCR-select subtraction and next-generation sequencing. Preliminary analysis of the four sequenced viruses has shown that they are all novel, and phylogenetic studies will now be performed to further classify these organisms. In addition, the report details the success in using a novel method to prepare MPKV and BCV RNA for next-generation sequencing. The use of this method allows for an increased level of efficiency when sequencing unknown viruses and has now been incorporated into our standard viral preparation protocols.

RELEASE LIMITATION
Approved for public release

Published by

Human Protection and Performance Division
DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia

Telephone: (03) 9626 7000
Fax: (03) 9626 7999

AR-015-333
September 2012

APPROVED FOR PUBLIC RELEASE

Characterisation & Next-generation Sequencing Analysis of Unknown Arboviruses

Executive Summary

The term arbovirus is used to describe viruses transmitted via blood-feeding arthropods, such as mosquitoes, biting midges and ticks. Two families of arboviruses, the *Rhabdoviridae* and *Bunyaviridae*, are suspected to play some role in the incidence of encephalitis and neurological disease occurring in northern regions of Australia, for which there is no known cause. Significant numbers of Australian Defence Force (ADF) personnel are posted to these areas, and military exercises are frequently conducted. In addition, bunyaviruses such as Rift Valley fever, Crimean Congo haemorrhagic fever and hantavirus, and rhabdoviruses such as rabies virus and Australian bat lyssavirus have been implicated as potential bioterrorism agents. This is due to their infectivity, ability to induce a fatal or seriously incapacitating illness, lack of adequate control measures, and the ease of production of large quantities of virus. Characterisation by sequencing is therefore an important step in identification and detection of these viruses. It also assists in determining their role in disease, which may lead to the implementation of intervention strategies by the ADF.

This report focuses on characterisation of two unknown rhabdoviruses, Holmes Jungle virus (HOJV) and Harrison Dam virus (HARDV), and two unknown bunyaviruses, Buffalo Creek Virus (BCV) and Maprik virus (MPKV). It describes the techniques such PCR select subtraction, Rapid Amplification of cDNA Ends (RACE), and next-generation high throughput sequencing, which are used to obtain the full genetic sequence of these viruses. While PCR select subtraction is a gold standard method of obtaining unknown viral sequence, it is a relatively complicated, time consuming process. Next-generation sequencing is a new powerful technology which can be utilised to identify and characterise unknown viruses with greater speed and at lower cost. The rapid advancement of new generation sequencing techniques allows for highly specific acquisition of gigabases of sequence information in just a few days.

This report also describes a new method to prepare the virus samples for next-generation sequencing analysis, which our results show is equally as effective, yet considerably faster, simpler and cheaper than the current standard preparation methods. Preliminary analysis of the four sequenced viruses has also shown that they are all novel viruses, and we are currently in the process of preparing collaborative manuscripts for publication in international scientific journals with the Berrimah Veterinary Laboratory and the CSIRO Australian Animal Health Laboratory describing the novel findings.

Authors

Jane McAllister
Human Protection and Performance Division, DSTO

Jane McAllister joined HPPD in 2003 with a B. Med Lab Sci and Honours in Immunology. Prior to DSTO, Jane worked at the University of Canberra, identifying potential antigens for vaccines against respiratory pathogens, and assessing their effect on the immune response. Her initial work at DSTO focused on the construction and immunological analysis of DNA vaccines against biological agents, in particular Burkholderia pseudomallei. In 2009 she moved into the field of arbovirology under the guidance of Dr Ania Gubala, focusing on characterisation and identification of viruses from the families Rhabdoviridae and Bunyaviridae.

Penny Gauci
Human Protection and Performance Division, DSTO

Penny Gauci graduated from the University of Melbourne with a B Sc in 1992. After working as a technical officer at RMIT University, she joined DSTO in 1998. She initially worked in the AMBRI ion channel switch (ICS) biosensor program before joining the DNA vaccine program in 2001, producing vaccine candidates for western equine encephalitis virus. She has worked on the Virology program since late 2009 and has been involved in the identification and characterisation of unknown rhabodoviruses and bunyaviruses

Dr Ania Gubala
Human Protection and Performance Division, DSTO

Ania Gubala began working for DSTO's Human Protection and Performance Division in 2002 following the completion of a Bachelor of Medical and Pharmaceutical Biotechnology (Honours) at the University of South Australia. Ania's role was in the area of rapid diagnostics of biowarfare agents, on the development of real-time PCR detection assays for Vibrio cholerae, a water-borne bacterium responsible for severe enteric disease. From 2005 to 2009 Ania was on secondment to the CSIRO Australian Animal Health Laboratory in Geelong where she undertook PhD studies in virology. Since Ania's return to DSTO in 2009 she has been developing a new area of research within the division in the field of arbovirology, focusing on viruses from the families Rhabdoviridae and Bunyaviridae. Ania has authored, co-authored and reviewed numerous scientific publications and book chapters on rhabdoviruses and cholera.

Contents

1. INTRODUCTION 1

2. MATERIALS AND METHODS 2
2.1 Virus propagation and RNA Extraction 2
2.2 PCR-select cDNA subtraction of HOJV and BCV 3
2.3 Preparation of double stranded cDNA (ds cDNA) for next-generation sequencing of MPKV and HARDV 3
2.4 Next-generation sequencing and analysis of MPKV and HARDV 3
2.5 Confirmatory PCR amplification of HARDV, MPKV, BCV and HOJV genomes 4
2.6 Capillary sequencing of PCR products 4
2.7 Rapid amplification of cDNA ends (RACE) 4

3. RESULTS AND DISCUSSION 6
3.1 Characterisation of HOJV and BCV via PCR-select cDNA subtraction 6
3.1.1 HOJV sequence analysis 7
3.1.2 BCV sequence analysis 7
3.1.3 RACE of HOJV and BCV 8
3.2 Subtractive hybridisation vs crude extract purification 9

4. CONCLUSION 13

5. REFERENCES 14

This page is intentionally blank

1. Introduction

The term arbovirus is used to describe viruses transmitted via blood-feeding arthropods, such as mosquitoes, biting midges and ticks. In the last few decades, the incidence of emerging arbovirus infection has been on the rise, due to increased urbanisation in tropical and subtropical areas, globalisation, and as a consequence of environmental and ecological change [1]. Of particular interest is the increased incidence in Australia of encephalitic and neurological disease for which the causative agent is unknown or poorly understood, but likely to be linked to arthropod transmission [2]. Two viral families, the *Rhabdoviridae* and *Bunyaviridae*, are suspected to play a role in the incidence of unidentified encephalitic symptoms, particularly in Northern regions of Australia, an area where military exercises are frequently conducted, and where significant numbers of Australian Defence Force (ADF) personnel are posted.

In addition to a potential increase of endemic exposure, bunyaviruses such as Rift Valley fever, Crimean Congo haemorrhagic fever and hantavirus, and rhabdoviruses such as rabies virus and Australian bat lyssavirus have been implicated as potential bioterrorism agents. This is due to their infectivity, transmissibility, ability to induce a fatal or seriously incapacitating illness, the lack of adequate control measures, and the ease of production of large quantities of virus [3, 4]. Characterisation by sequencing is therefore an important step in detection and identification of these viruses. It also assists in determining their role in disease, which may lead to the implementation of intervention strategies by the ADF.

Viruses of the family *Bunyaviridae* contain enveloped, spherical virions, approximately 100 nm in diameter, and comprise the largest family of RNA viruses, containing more than 350 named isolates [5]. Rhabdoviruses by comparison have a bullet-shaped morphology and range in size from 100-400 nm in length and 45-100 nm in width. The virion envelope contains transmembrane glycoproteins that form 'spikes' which project from the surface. One of the more well known genera of the *Rhabdoviridae* is the *Lyssavirus* whose members include the deadly rabies virus and Australian bat lyssavirus. Rhabdovirus genomes comprise a negative sense, single stranded RNA molecule, ranging in size from approximately 8.9-15 kb. The prototype rhabdovirus contains at least 5 open reading frames, encoding five structural proteins as follows (in order 5'-3'): nucleoprotein (N), phosphoprotein (P); matrix protein (M), glycoprotein (G) and polymerase (L) [6]. In contrast, the prototype bunyavirus contains three negative sense, single stranded RNA segments, which encode four structural proteins. The two external virion glycoproteins (Gn and Gc) are encoded by the M (medium sized) segment, whereas the two internal virion proteins, a nucleocapsid protein (N) and the viral RNA polymerase, are encoded on the small (S) and large (L) segment respectively [7].

In recent times a number of powerful technologies have emerged that can be utilised to identify and characterize unknown viruses with greater speed and at lower cost. The rapid advance of new generation sequencing techniques has allowed for highly specific acquisition of gigabases of sequence information in just a few days [8]. Such technologies include the 454 Life Sciences pyro-sequencing based instrument (Roche) and the Illumina genetic analyser. Where original Sanger methods of genetic analysis would generate 100

kilobases of sequence data, 454 and Illumina sequencing are capable of generating up to 400 megabases and 20 gigabases respectively [9].

This report describes the techniques that were developed for *de novo* whole genome sequencing of rhabdoviruses and bunyaviruses. It focuses on two unknown rhabdoviruses, Holmes Jungle virus (HOJV) and Harrison Dam virus (HARDV), and two unknown bunyaviruses, Buffalo Creek Virus (BCV) and Maprik virus (MPKV), that were isolated in northern Australia and Papua New Guinea during routine arbovirus surveillance (see Table 1). The report also describes the development of an alternative approach to the traditional PCR-select cDNA subtraction method for identification of unknown viruses, which is considerably more efficient and cost effective.

Table 1: Unassigned rhabdoviruses and bunyaviruses from Australia and Papua New Guinea [10-12]

Virus name	Isolated from	Isolate #	Year of isolation	Location	Family	Neutralising antibody detected
Holmes Jungle	*Culex annulirostris* (mosquito)	DPP1163	1987	Darwin, NT	*Rhabdoviridae*	Cattle, buffalo, humans
Harrison Dam	*Culex annulirostris* (mosquito)	CSIRO75	1975	Beatrice Hill, NT	*Rhabdoviridae*	unknown
Buffalo Creek	*Anopheles meraukensis* (mosquito)	DPP186	1982	Darwin, NT	*Bunyaviridae*	Cattle, pig, human
Maprik	*Aedes funereus* (mosquito)	MK7532	1966	Maprik, New Guinea	*Bunyaviridae*	unknown

2. Materials and Methods

2.1 Virus propagation and RNA Extraction

BCV (isolate DPP186), HOJV (isolate DPP1163), HARDV (isolate CS75) and MPKV (isolate MK7532), were obtained from the Berrimah Veterinary Laboratory, Darwin. Viruses were propagated in BHK-BSR cells (a subclone of the baby hamster kidney BHK-21 cell line) grown in basal medium Eagle (Gibco) supplemented with 10mM HEPES, 2mM L-glutamine, 137µM streptomycin, 80 U/ml penicillin and 5% (growth media, GM) or 2.5% (maintenance media, MM) fetal calf serum (Gibco) at 37˚C. Three to four days post infection, at the first signs of the cytopathic effect (CPE), the infected cell culture supernatant was collected, centrifuged at 1600 xg for 10 min to pellet cell debris, and the viral pellet was obtained by ultracentrifugation at 70 000 x g for 1 hr in a Beckman 70Ti rotor. The pellet was resuspended in Buffer RLT (Qiagen) containing β-mercaptoethanol and total RNA extracted from the crude virus pellet using the RNeasy Mini Kit (Qiagen)

and quantified using the NanoDrop 2000 (Thermo Scientific). For MPKV and HARDV, total RNA extraction involving the addition of RNase One (Promega) to the crude pellet was performed with the aim of eliminating host cell RNA prior to using the RNeasy Kit.

2.2 PCR-select cDNA subtraction of HOJV and BCV

The sequencing of HOJV and BCV was performed using the traditional approach using PCR-select cDNA subtraction and capillary sequencing. The subtractive hybridization reactions were performed by staff at the Berrimah Veterinary Laboratory using the PCR-Select cDNA Subtraction kit (Clontech) according to manufacturer's instructions. Tibrogargan virus and Akabane virus were used as the driver in the reactions for HOJV and BCV, respectively. The genome fragments generated were cloned into pCR-Blunt II TOPO and sequenced via traditional capillary sequencing as described previously [13]. DSTO received the generated raw sequence data from Berrimah Veterinary Laboratory, and subsequently constructed contigs to which PCR primers were designed for sequence confirmation and to fill the gaps between contigs.

2.3 Preparation of double stranded cDNA (ds cDNA) for next-generation sequencing of MPKV and HARDV

RNA was extracted from crude viral MPKV and HARDV pellets and was concentrated by adding a 1:10 volume of 3M sodium acetate and 2.5 x volumes of 100% ethanol. A precipitated pellet was obtained by centrifugation for 30 min at maximum speed in a microfuge, and the pellet was washed in 500µl of 70% ethanol. The pellet was air-dried and resuspended in 1µl DEPC-treated water and used as template in ds cDNA synthesis reactions using the Superscript ds cDNA sythesis kit (Invitrogen) according to manufacturer's instructions. The final concentration of ds cDNA was determined using the NanoDrop 2000 (Thermo Scientific).

2.4 Next-generation sequencing and analysis of MPKV and HARDV

High throughput sequencing was performed by Monash University using the second-generation Illumina Genetic Analyser. Primary assembly of raw data and generation of consensus sequences were performed using the programs Velvet 1.1.04, Geneious Pro 5.4 and Artemis (Sanger) by Dr Dieter Bulach of CSIRO Livestock Industries, Geelong. Routine sequence management and the design of PCR primers was performed using the programs SeqMan Pro v. 8.0.2 (Lasergene v. 8 DNASTAR), CloneManager v. 9 (Sci Ed Central) and Sequencher 4.9.

2.5 Confirmatory PCR amplification of HARDV, MPKV, BCV and HOJV genomes

A volume of 100 µl of virus-infected cell culture supernatant was added to 600 µl Buffer RLT (Qiagen) containing β-mercaptoethanol and RNA was extracted using the Qiagen RNeasy Mini Kit according to manufacturer's instructions. The extracted total RNA was subsequently used as template for ss cDNA synthesis using the Invitrogen Superscript III Kit according to manufacturer's instructions. The cDNA was diluted 1:10 in sterile water and 1 µl (equivalent to 1-10 ng) was used as template in PCR reactions. PCR primer pairs were designed using the sequences obtained from the Illumina Genetic Analyser or from PCR-select cDNA subtraction, with each amplifying 500-800 nt. regions spanning the genome (excluding the termini). Primers were synthesized by Geneworks, Australia. Each PCR reaction contained 2.5 µl of 10x Advantage 2 PCR buffer, 0.2 mM dNTPs, 0.5 µl Advantage 2 DNA Polymerase (all from Clontech), 0.2 µM each primer, 100 ng cDNA template, and sterile water to a final volume of 25 µl. PCR products were subsequently sequenced to generate a contiguous, consensus sequence for the entire viral genome, excluding the genome termini (which were sequenced using the RACE protocol, described below).

2.6 Capillary sequencing of PCR products

Twenty to forty ng of amplified PCR product was used as the template in a sequencing reaction containing 1 x sequencing buffer (Applied Biosystems), 3.2 pmol of the appropriate forward or reverse primer, 0.5 µl of BigDye Terminator (Applied Biosystems) and sterile water to a final volume of 20 µl. The reaction was performed under standard sequencing thermal cycling conditions on an Eppendorf Mastercycler pro (Eppendorf). The resulting amplified product was purified using the BigDye XTerminator Purification Kit (Applied Biosystems) according to manufacturer's instructions and analysed on a Genetic Analyzer 3130xl (Applied Biosystems).

2.7 Rapid amplification of cDNA ends (RACE)

To determine the sequence of the 5'and 3'genome termini of the unknown viruses, a modified protocol for the rapid amplification of cDNA ends (RACE) was used [6]. Total genomic RNA was isolated from 100ul of infected cell culture supernatant using the RNeasy Mini kit (Qiagen) , and eluted in 30ul water.

For 5' RACE, cDNA was synthesized from total genomic RNA using the SuperScript III Reverse Transcriptase system (Invitrogen) and a virus specific synthesis primer (5RACE synthesis primer) located upstream of the 5' terminus. Briefly, 31 µl of the viral RNA was mixed with 1 x FS buffer, 0.25 µM "5RACE synthesis primer" and 1 mM dNTPs in a 50 µl reaction and incubated at 65 ˚C for 5 min before placing on ice for 1 min. One hundred units of Superscript III Reverse Transcriptase and 40 U RNaseOut were added and incubated at 50 ˚C for 1 hr followed by 70 ˚C for 15 min before placing the reaction on ice. The RNA strand bound to the cDNA was subsequently destroyed by treatment with 5 U

RNaseH at 37 ˚C for 20 minutes to yield a single stranded cDNA molecule to which a RACE adaptor could be ligated. The reaction was purified using the QIAquick PCR Cleanup kit (Qiagen) and eluted in 30 µl water. A 5'-phosphorylated 3'-end cordecypin-blocked anchor adaptor (DT88) was ligated to the cDNA using T4 RNA ligase, as previously described [6, 14, 15]. Briefly, 12 µl cDNA was mixed with 0.4 µM DT88, 1 x Ligase buffer, and 10 U T4 RNA ligase in a 15 µl reaction and incubated overnight at 4 ˚C. The adaptor-ligated cDNA was diluted 1:10 and subjected to PCR amplification with the Advantage 2 PCR kit (Clontech) using a virus-specific primer upstream of the 5 RACE synthesis primer (5RACE1) and the adaptor-specific primer (DT89). This was followed by a hemi-nested PCR using a 10 fold dilution of the primary PCR product as template and a virus specific primer located upstream of the 5RACE1 primer (5RACE2), and the DT89 primer. For the rhabdoviruses HOJV and HARDV, a modified rhabdovirus-specific DT89 primer (DT89-rhabdo) containing three additional nucleotides homologous to the conserved rhabdovirus genome termini was used as previously described [6].

For 3' RACE, the DT88 adaptor was ligated directly to total RNA prior to cDNA synthesis as described previously [15]. Briefly, the 20 µl ligation reaction consisting of 13 µl RNA, 0.9 µM DT88, 1 x ligase buffer, 20 U T4 RNA ligase and 40 U RNaseOut was incubated overnight at 4 ˚C. Synthesis of cDNA was performed directly on an aliquot of the ligation reaction. Eight microlitres of the adaptor-ligated RNA was added to a reaction containing 0.4 M DT89 Primer, 1 mM dNTPs, 1 x FS buffer in a final volume of 50 µl and incubated at 65 ˚C for 5 minutes before placing on ice for 1 minute. One hundred units of Superscript III Reverse Transcriptase and 40 U RNaseOut was added and the mixture incubated at 50 ˚C for 1 hr. This was followed by the mixture being incubated at 70 ˚C for 15 min and placed on ice. The sample was diluted 1:10 and subjected to PCR amplification using a virus-specific primer, 3RACE1 and the DT89 primer. This was followed by a hemi-nested PCR using a virus-specific primer located downstream of the 3RACE1 primer (3RACE2), and the DT89 (for bunyaviruses) or DT89-rhabdo (for rhabdoviruses) primer.

Both 5' and 3' RACE PCR amplifications were performed on an Eppendorf Mastercycler Pro (Eppendorf) using standard thermal cycling conditions (94 ˚C for 2 min, followed by 35 cycles of 94 ˚C for 30 sec, 50 ˚C for 30 sec, 72 ˚C for 1 min) and primers at a final concentration of 2µM. A final extension at 72 ˚C for 7 minutes concluded the thermal cycling. A 'touchdown' PCR cycle was utilised when the desired PCR product could not be achieved through the standard cycling conditions. The touchdown cycle parameters were 94 ˚C for 2 min, followed by 16 cycles consisting of 94 ˚C for 30 sec followed by an annealing temperature that decreased by 0.5 ˚C with each cycle starting from, 56 ˚C to 48 ˚C for 30 sec, 72 ˚C for 1 min. This was followed by 20 cycles of 94 ˚C for 30 sec, 48 ˚C for 30 sec, 72 ˚C for 1 min and concluding with a 7 min extension at 72 ˚C. The resulting PCR products were analysed on a 2% agarose gel and bands of appropriate size were excised and purified using the QIAquick Gel Extraction Kit (Qiagen) and cloned into pCR2.1-TOPO (Invitrogen). The cloned fragments were subsequently sequenced using the 3130*xl* Genetic analyser (Applied Biosystems) and the consensus sequence was determined. The final sequences were analysed using Sequencher analysis software v4.9.

This method was used for both the rhabdoviruses and the bunyaviruses. As bunyaviruses contain 3 genome segments, certain modifications are needed to the basic method. For the 5'end, it is necessary to perform three separate cDNA synthesis reactions using virus-

specific primers for each segment. The method for the 3' end remains the same since it is non-specific.

3. Results and Discussion

3.1 Characterisation of HOJV and BCV via PCR-select cDNA subtraction

Berrimah Veterinary Laboratories are the key centre for the surveillance of arbovirus activity in the Northern Territory. An unknown rhabdovirus, designated Holmes Jungle virus, and an unknown bunyavirus, designated Buffalo Creek virus, were detected during routine surveillance via virus isolation and serological testing. Twenty cattle at the Coastal Plains Research station tested seropositive to the virus, and in 1993 neutralising antibody was also detected in a human serum sample collected at the Royal Darwin Hospital [10]. Holmes Jungle virus was isolated in 1987 from mosquitoes collected at Palm Creek near Darwin. Subsequently neutralising antibody has been detected in cattle, buffalo and humans, however the association with disease remains unknown [11].

PCR-select subtraction is a gold standard method for *de novo* virus sequencing as there is no need for virus purification which can be complicated and time consuming [16-21]. This method enriches for DNA sequences that are unique to the sample of interest through a series of hybridisations and PCRs, generating a sample that is highly enriched in virus-specific sequence [22]. The length of time from harvesting virus to generating samples that are ready for sequencing takes about two weeks, which is a considerable advancement on the technologies of the past.

This method was performed to obtain new sequence data for both HOJV and BCV. Using an RsaI restriction enzyme, restriction libraries of the viral genome were created to form overlapping fragments, which were cloned, sequenced and screened for virus-specific sequences via BLASTX similarity searches of NCBI databases of available bunyavirus and rhabdovirus sequences.

The total amount of sequence data generated for HOJV via this method approximated 6680 nt, representing 50% of the entire genome. This was in the form of 9 distinct contigs, which were separated by regions of unknown sequence. Similarly, sequence data for BCV approximated 6600 nt, or 60% of the entire genome and was used to form 10 distinct contigs that spanned each of the three RNA segments of the virus (L fragment, 6 contigs; M fragment, 2 contigs; S fragment, 2 contigs),

The sequence data obtained via subtraction was used to design 11 primer pairs for HOJV and 13 primer pairs for BCV, to enable amplification of the entire genomes (with the exception of the ends which were sequenced using the RACE method, described below),and obtain sequence data in the areas of unknown sequence. The primers amplified overlapping regions of between 500 nt and 800 nt. The amplified PCR products were sequenced by capillary sequencing, resulting in a single consensus sequence for the

genome of HOJV, and three consensus sequences representing each of the BCV genome fragments.

3.1.1 HOJV sequence analysis

The sequencing of the HOJV genome has been completed and consists of 13,168 nt. In addition to the 5 structural proteins classic to all rhabdoviruses, HOJV contains an additional four ORFs encoding putative proteins of unknown function: three between the P and M genes and one overlapping the G gene (see Figure 1). The genome organisation is similar to that of Wongabel rhabdovirus (WONV). A distinguishing feature between the two viruses is the absence of a small ORF overlapping the N gene in HOJV (designated U4 in WONV). The predicted proteins of HOJV and WONV show a high degree of amino acid similarity (ranging from 76 to 95% sequence identity).

Phylogenetic analysis demonstrates that HOJV has a close relationship with WONV. N and G protein phylogenetics place HOJV within the proposed 'Hart Park group' of viruses, consisting of WONV, and two additional unassigned viruses, Ngaingan virus (NGAV) and Flanders virus (FLAV), which have been associated with insects, birds, macropods and mammals [11]. Antigenic comparisons between the two viruses are currently underway at the Berrimah Veterinary Laboratory and a collaborative scientific manuscript describing the findings is in preparation.

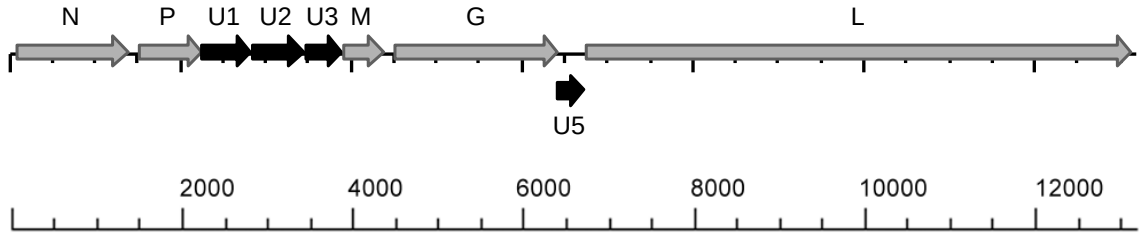


Figure 1: Genetic Map of the 13912-nt HOJV genome. Shaded arrows represent ORFs for the five typical rhabdovirus proteins N, P, M, G and L shown in the direction of mRNA transcription. Black arrows indicate ORFs encoding putative proteins of unknown function; U1, U2 and U3 are encoded by independent genes, whereas U5 overlaps the G gene.

3.1.2 BCV sequence analysis

BLAST searches suggest that BCV belongs to the genus *Orthobunyavirus*. These preliminary analyses suggest the closest relationship is with the *California encephalitis virus* (a group of viruses involved in cases of human disease in the USA) and *Bunyamwera virus* groups, however, phylogenetic analyses are required to confirm these observations. An external paper providing more detailed characterisation of BCV is currently being prepared by Dr Richard Weir of Berrimah Veterinary Laboratories.

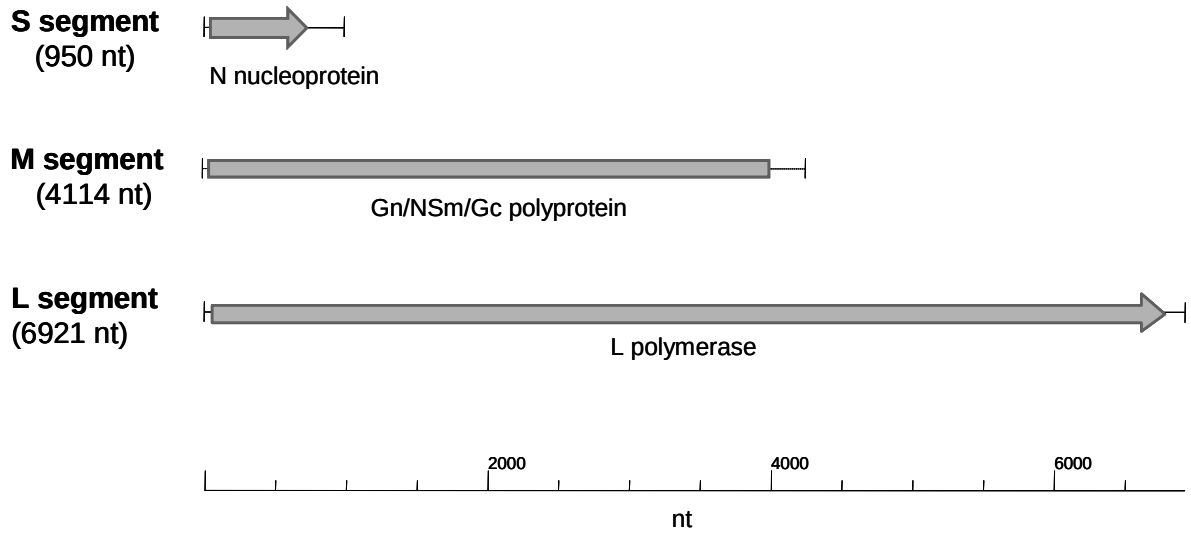


Figure 2: Genome map of BCV. Sequencing of the 5′ (genome-sense) region of the M segment is not yet completed. The sizes of the three genome segments (S, 950 nt; M, ~4500 nt anticipated once finished; L, 6921 nt) are similar to those of the prototype orthobunyavirus (S, 1000 nt; M 4500 nt, L 6900 nt) [7]). Each of the segments contains a single open reading frame which codes for a single protein, with the exception of the M segment, which typically in orthobunyaviruses codes for a polypeptide that is cleaved by host cell proteases into three distinct proteins. These cleavage locations cannot be effectively predicted by bioinformatics and thus detailed protein studies are required.

3.1.3 RACE of HOJV and BCV

A feature of (-) ssRNA viruses is the complementarity of the 3′and 5′ genome termini [7]. This feature provides confidence that the genome has been successfully sequenced to the ends following the RACE procedure. Figure 3 shows the final 20 nucleotides at each end of the HOJV and BCV genomes following the RACE procedure. The RACE procedure confirmed that the ends of HOJV, which were previously obtained by BVL, were correct and did not require further sequencing. Sequencing of the 3′ genome terminus of HOJV revealed conservation of the three terminal nucleotides common to rhabdoviruses, that signifies the ends of the genome. We have also confirmed that 12 of the 16 terminal nucleotides in the leader and trailer sequences showed perfect complementarity, which is also a common feature in rhabdoviruses.

The genome termini of BCV, however, were not sequenced by BVL. Following the RACE procedure, we obtained a further 79, 54 and 99 nucleotides at the 3′ termini of the L, M and S segments respectively. Similarly, 125, 458 and 33 additional nucleotides were obtained at the 5′ termini for the L, M, and S segments respectively.

Bunyaviruses often contain 8 to 11 nt at their genome segment termini which are highly conserved between members of the same genus [7]. Viruses within the genus *Orthobunyavirus* are reported to contain 11 conserved terminal nucleotides. In addition, there is often high conservation of up to 20 terminal nt within each genome segment, which can be used to assess if the terminal sequences have indeed been obtained. Based on

these criteria, we have confidently determined the terminal genome sequences for all three segments with the exception of the 5′ terminus of the M segment (Fig 3). Because there was a large section (~1000 nt) of missing sequence data at the 5′ M segment terminus, we have yet to complete this segment. It is suspected that 1000 nt is beyond the capability of the RNA polymerase in the RACE protocol, hence further sequencing is required to complete this region. Consequently, a new synthesis primer and RACE primers have been designed to the newly generated sequence, and the procedure will be repeated until the sequence is completed.

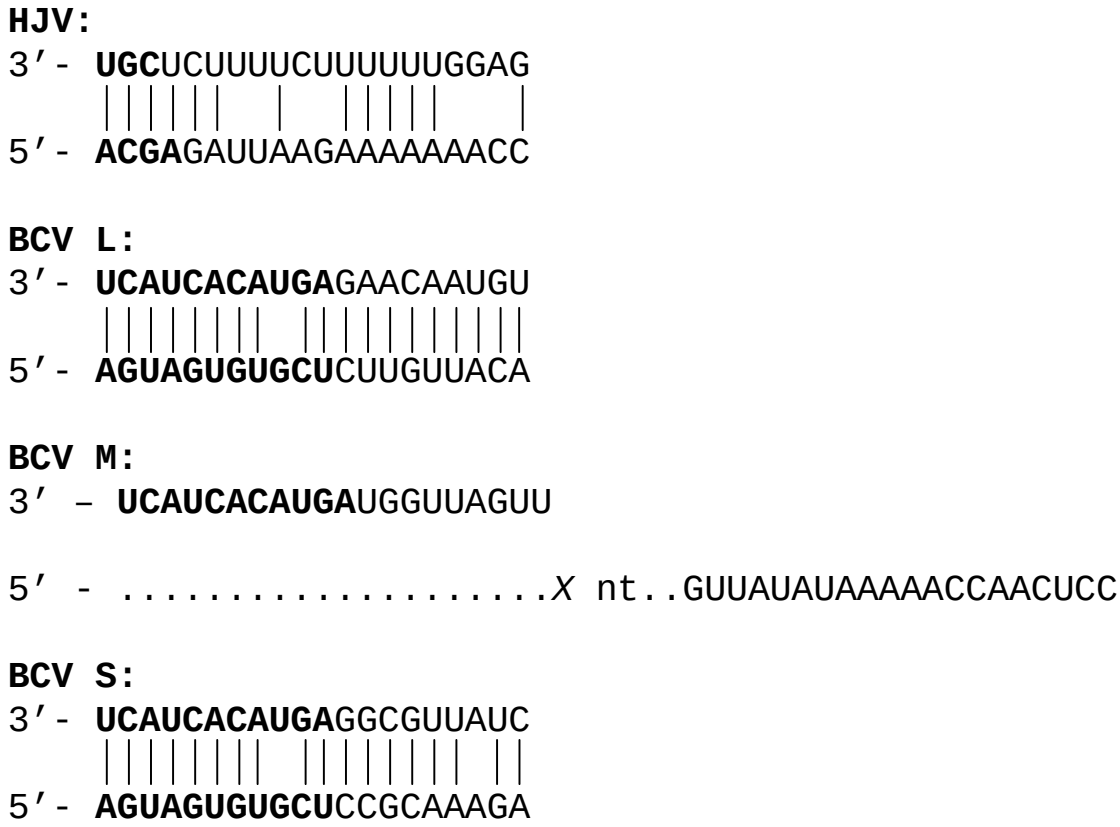


*Figure 3: End sequence determined for each of the ends analysed (genome sense). The complementary bases are indicated by vertical lines whilst the conserved termini sequence is indicated in **bold**.*

3.2 Subtractive hybridisation vs crude extract purification

The process of PCR-select subtraction described above is considered one of the most effective ways to identify unknown viral sequence. Although this technique is extremely successful for identification of novel viruses, it requires significant preparation time, is time consuming and expensive. High throughput sequencing using platforms such as the Illumina Genetic Analyser provides a rapid, cost-effective and more accurate means of generating sequence data [23].

The Illumina sequencing process involves preparation of fragmented ds cDNA libraries using a reverse transcriptase method that utilises extracted viral RNA ligated to adaptors via its 5′ and 3′ ends. The Illumina technology utilises hybridisation and amplification on a multi-lane glass flowcell to generate hundreds of thousands of DNA clusters per lane. After amplification, each cluster is analysed using reversible terminator dye chemistry [24].

We utilised this next-generation sequencing technology to sequence a further 2 unknown arboviruses: Maprik virus (MPKV), a bunyavirus belonging to the unclassified Mapputta serogroup [25], and Harrison Dam virus (HARDV), an uncharacterised rhabdovirus (Table 1). We attempted a proof-of-concept approach to sample preparation for next-generation sequencing with the aim of eliminating the need to perform PCR-select subtraction prior to sequencing, thereby significantly reducing preparation time.

BSR cells were infected with the two viruses, and on day 4 post infection significant and widespread cytopathic effect was observed. The viruses were harvested by ultracentrifugation and RNA was extracted and pooled. Half of the viral RNA harvested for each virus was subjected to an additional RNAse One ribonuclease treatment prior to RNA extraction, with the aim of removing contaminating host cell RNA, under the assumption that viral genomic RNA would be protected by the nucleoprotein which inherently protects viral RNA from host cell RNAses. Quantification of the RNA on the Nanodrop, however, demonstrated that this procedure was unsuccessful because the RNAse One appeared to also destroy the viral RNA. Consequently, the RNA preparations without RNAse One treatment were converted to ds cDNA, fragmented into libraries and analysed on the Illumina instrument.

Following next generation sequencing of HARDV, a near complete genome sequence (with the exception of the genome terminal regions) was obtained, with greater than 20-50 times coverage (see Figure 4). A similarly high level of genome coverage was also obtained for all three genome segments of MPKV (see Figure 5). While there were areas of limited sequence coverage (gaps that were present were sequenced using PCR as described for HOJV and BCV above), for the vast majority of the genome a minimum of 5, 10 and 20 fold coverage was obtained for the L, M and S segments respectively.

This proof-of-concept approach demonstrates the power of next generation sequencing for analysis of crudely prepared viral RNA that contains a significant proportion of ‘contaminating’ genetic material from the host cell. We have developed a method that effectively omits the necessity to enrich viral RNA for sequencing using methods such as the PCR-select subtraction. This is a significant advance on currently available methods for *de novo* virus sequencing. Currently, we are sequencing other viruses using this method with a similarly high level of success. We are also in the process of developing a method to enable sequencing of viruses directly from frozen stocks without the need for propagation in cells. If successful, this will not only further simplify the viral purification process, but will also allow for sequencing of viruses difficult to cultivate and grow in cell culture. Ultimately, it is hoped that these steps will lead to the development of a method for next generation virus sequencing directly from clinical samples such as blood.

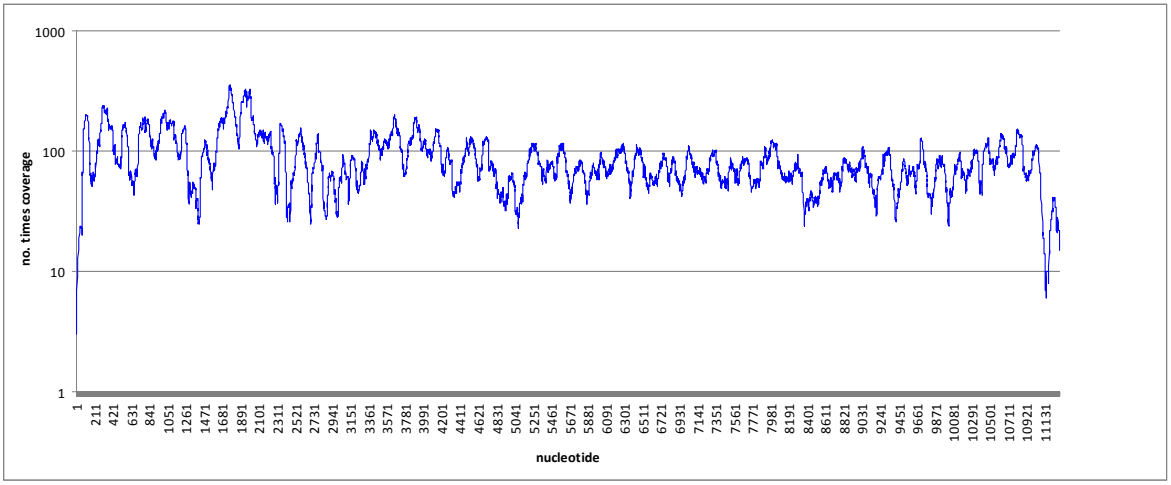
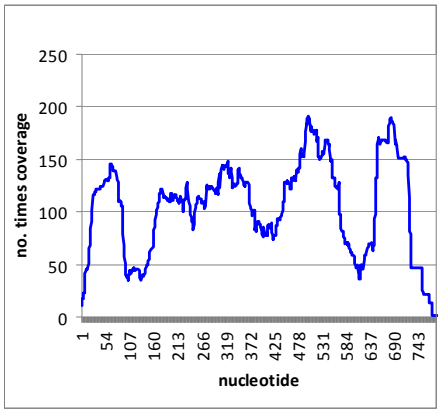
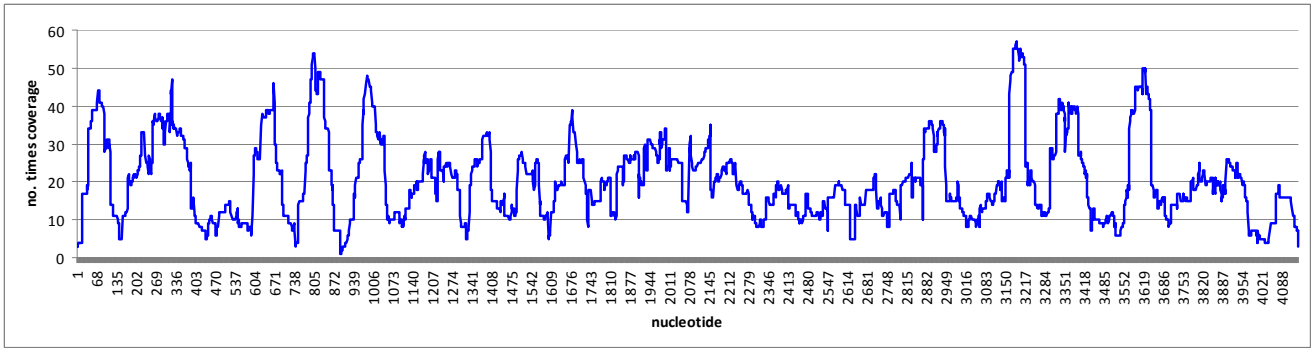


Figure 4: Coverage plot for Harrison Dam rhabdovirus following next-generation sequencing. Greater than 20-50 times coverage was obtained for the entire genome length, with the exception of the ends which will be sequenced via RACE.

a)



b)



c)

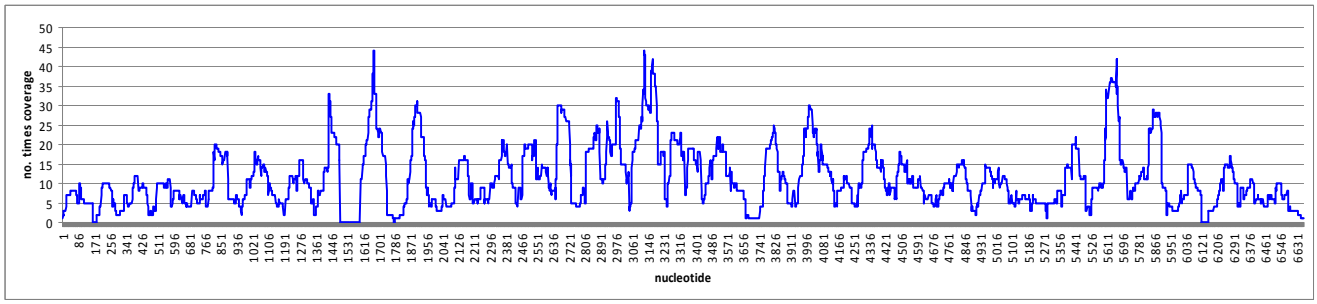


Figure 5: Coverage map for the three segments of Maprik bunyavirus following next-generation sequencing. Graphs indicate no. of times sequence coverage for a) S segment, b) M segment and c) L segment. A minimum of 5 times coverage was observed for most of the L segment genome, a minimum of 10 times coverage was obtained for the majority of the M segment, and greater than 20 times coverage was obtained for the S segment, with the exception of the ends. All regions of low coverage were confirmed using PCR, and the end sequences will be analysed using RACE.

4. Conclusion

In summary, we have described a new method for genomic sequencing of novel viruses which is a considerable improvement on current methods for *de novo* virus sequencing. We have shown that crude viral RNA preparations can be used to prepare ds cDNA for direct analysis by high throughput sequencing, and that this method produces quality sequence data that is comparable to that generated using the subtractive hybridisation method but is considerably less time consuming, simpler and cheaper. This provides an increased level of efficiency when sequencing unknown viruses, and we have now incorporated this technique as our standard method of virus preparation for next-generation sequencing. Preliminary analysis of the four sequenced viruses has shown that they are all new to science, and we are currently in the process of preparing manuscripts for publication in international scientific journals with the Berrimah Veterinary Laboratory and the CSIRO Australian Animal Health Laboratory describing the novel findings.

5. References

1. Mahy, B.W.J., *Emerging and reemerging virus diseases of vertebrates* in *Desk Encyclopedia of Human and Medical Virology*, B.W.J. Mahy and M.H.V. Regenmortel, Editors. 2008, Academic Press: San Diego. p. 93-97.
2. Huppotz, C., et al., *Encephalitis in Australia, 1979-2006: Trends and Aetiologies* CDI, 2009. **33**(2): p. 192-197.
3. Geisbert, T.W. and P.B. Jahrling, *Exotic emerging viral diseases: Progress and challenges.* Nature Medicine, 2004. **10**(12 SUPPL.): p. S110-S121.
4. Sidwell, R.W. and D.F. Smee, *Viruses of the Bunya- and Togaviridae families: Potential as bioterrorism agents and means of control.* Antiviral Research, 2003. **57**(1-2): p. 101-111.
5. Elliot, R.M., *Bunyaviruses and Climate Change.* Clin Microbiol Infect, 2009. **15**(2009): p. 510-517.
6. Gubala, A., et al., *Ngaingan virus, a macropod-associated rhabdovirus, contains a second glycoprotein gene and seven novel open reading frames.* Virology, 2010. **399**(1): p. 98-108.
7. Fauquet, C.M., et al., eds. *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses*. 8th Edition ed. 2005, Elsevier Academic Press: London.
8. MacLean, D., J.D.G. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nature Reviews Microbiology, 2009. **7**(4): p. 287-296.
9. Bexfield, N. and P. Kellam, *Metagenomics and the molecular identification of novel viruses.* Veterinary Journal, 2010. **190**(2): p. 191-198.
10. Weir, R.P., *Classification and Identification of Viruses isolated from Mosquitoes in the Northern Territory, 1982 - 1992, Using a Range of Techniques*. 2002, University of Sydney.
11. Bourhy, H., et al., *Animal Rhabdoviruses*, in *Encyclopedia of Virology*, B.W.J. Mahy and M.H.V. van Regenmortel, Editors. 2008, Elsevier Academic Press. p. 111-21.
12. CDC. *Divison of Vector-borne Diseases Arbovirus Catalog*. 2010 [cited 2011 May 26]; Available from: http://wwwn.cdc.gov/arbocat/catalog-listing.asp?VirusID=284&SI=1.
13. Gubala, A.J., et al., *Genomic characterisation of Wongabel virus reveals novel genes within the Rhabdoviridae.* Virology, 2008. **376**(1): p. 13-23.
14. Tillett, D., B.P. Burns, and B.A. Neilan, *Optimized rapid amplification of cDNA ends (RACE) for mapping bacterial mRNA transcripts.* Biotechniques, 2000. **28**(3): p. 448, 450, 452-3, 456.
15. Cowled, C., et al., *Genetic and epidemiological characterization of Stretch Lagoon orbivirus, a novel orbivirus isolated from Culex and Aedes mosquitoes in northern Australia.* Journal of General Virology, 2009. **90**(6): p. 1433-1439.

16. Cowled, C., et al., *Genetic and epidemiological characterization of Middle Point orbivirus, a novel virus isolated from sentinel cattle in northern Australia.* Journal of General Virology, 2007. **88**(12): p. 3413-3422.
17. Bowden, T.R., et al., *Molecular characterization of Menangle virus, a novel paramyxovirus which infects pigs, fruit bats, humans.* Virology, 2001. **283**(2): p. 358-373.
18. Jack, P.J.M., et al., *The complete genome sequence of J virus reveals a unique genome structure in the family paramyxoviridae.* Journal of virology, 2005. **79**(16): p. 10690-10700.
19. Crabtree, M.B., et al., *Genetic and phenotypic characterization of the newly described insect flavivirus, Kamiti River virus.* Archives of virology, 2003. **148**(6): p. 1095-1118.
20. Lambeth, L.S., et al., *Complete genome sequence of Nariva virus, a rodent paramyxovirus.* Arch Virol, 2009. **154**(2): p. 199-207.
21. Tang, P. and C. Chiu, *Metagenomics for the discovery of novel human viruses.* Future Microbiology, 2010. **5**(2): p. 177-189.
22. Illumina, I. *Genomic sequencing*. 2011 [cited 2011 May 25]; Available from: http://www.illumina.com/applications/detail/sequencing/dna_sequencing.ilmn.
23. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in genetics, 2008. **24**(3): p. 133-141.
24. Newton, S.E., et al., *The Mapputta Group of Arboviruses: Ultrastructural and Molecular Studies which place the Group in the Bunyavirus Genus of the Family Bunyaviridae.* Aust J Exp Biol Med Sci, 1983. **61**(2): p. 201-217.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Characterisation and Next-generation Sequencing Analysis of Unknown Arboviruses

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Jane McAllister, Penny Gauci and Ania Gubala

5. CORPORATE AUTHOR

DSTO
506 Lorimer St
Fishermans Bend Victoria 3207 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2744	AR-015-333	Technical Report	September 2012

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2011/1098620/1	07/258	JHC	15	24

DSTO Publications Repository

http://dspace.dsto.defence.gov.au/dspace/

14. RELEASE AUTHORITY

Chief, Human Protection and Performance Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS

Arbovirus, next-generation sequencing, rhabdovirus, bunyavirus

19. ABSTRACT

In the last few years there has been an increased incidence of neurological and encephalitic syndromes in Australia that have not been linked to a known pathogen. Viruses from the families *Rhabdoviridae* and *Bunyaviridae* are postulated to play a role. This report describes the characterisation of two unknown bunyaviruses, designated Maprik (MPKV) and Buffalo Creek (BCV), and two unknown rhabdoviruses, designated Harrison Dam (HARDV) and Holmes Jungle (HOJV), using techniques such as PCR-select subtraction and next-generation sequencing. Preliminary analysis of the four sequenced viruses has shown that they are all novel, and phylogenetic studies will now be performed to further classify these organisms. In addition, the report details the success in using a novel method to prepare MPKV and BCV RNA for next-generation sequencing. The use of this method allows for an increased level of efficiency when sequencing unknown viruses, and has now been incorporated into our standard viral preparation protocols.