Genome Analysis and Geographical Distribution of a novel Emaravirus in Karaka

Lee Owen Rabbidge

A thesis submitted to
Auckland University of Technology
in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

2020

School of Science

Primary supervisor: Dr. Colleen Higgins

Secondary Supervisors: Dr. Robin MacDiarmid, Dr. Arnaud Blouin

Table of Contents

Acknowledgements

This thesis would not have been possible without the love and support of many, many people. During this project I developed my scientific and technical skills, but more importantly it allowed me to meet, collaborate with, and enjoy the company of people who have left an incredible impact on my life. My heartfelt thanks to the following:

My primary supervisor - Dr. Colleen Higgins. You took me under your wing when I had practically no idea about molecular biology and with your gentle guidance I have been able to achieve something that, once upon a time, I would never have imagined possible.

My secondary supervisors - Dr. Robin MacDiarmid and Dr. Arnaud Blouin. You both gave me an incredible project and helped me get up to the level where I could complete it.

My colleagues at Auckland University of Technology - Chris, Priya and Gardette for showing me the ropes on my first plant virology project and helping me ever since. Toni, Shweta, Liz, Zoe, Tony, Jack and Kendall for being the best lab group meeting (and pub) buddies.

My colleagues at Plant and Food Research – Karmun, Roshni, Tianyi, Waqas, Jay, Liz and Rebecca. Thanks for always being there to help and making me feel welcome.

My colleagues at the Plant Health and Environment Laboratory - Lisa, Catia, Subuhi, Joe, Stella, Lia, Zoila, Filomena, Michelle and Deepika from the virology team and everyone else for helping me to further develop my skills and allowing me to apply my knowledge.

My family and friends - Belinda and John (my mum and dad), my brothers Michael, Shaun, Andrew and Nick and their partners and kids. You have all always been an unending source of support and love. My aunties, uncles and cousins – especially Fionna and Graham for sorting out samples for me and putting my sampling instructions to the test! My friends Ben, James, Matt, Chris, Nik and Sam for helping me keep my balance and sanity. Olly, Jeremy and Sanjana for always being there and willing to lend an ear.

Finally (and most importantly): my eternal gratitude to Suzy. Thank you for putting up with me and generally being the best human being I know.

List of Figures

List of Tables

Abbreviations

AcCRaV	Actinidia chlorotic ringspot-associated virus
BLAST	Basic Local Alignment Search Tool
BLMaV	Blackberry leaf mottle-associated virus
BUNV	Bunyamwera virus
CDD	Conserved Domain Database
Contig	Contiguous sequence
cRNA	Complementary RNA
CTAB	Cetyl trimethylammonium bromide
ddH_2O	Double-distilled water
DMB	Double membrane-bound body
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide trisphosphate
dsDNA	Double stranded DNA
DUGV	Dugbe virus
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
EMARaV	European Mountain Ash Ringspot-associated virus
ER	Endoplasmic reticulum
FMV	Fig mosaic virus
GPS	Global Positioning System
GPP	Glycoprotein precursor

HPWMoV	High Plains wheat mosaic virus
ICTV	International Committee on Taxonomy of Viruses
JYMaV	Jujube yellow mottle-associated virus
KOPV	Karaka Okahu purepure virus (tentative name)
LAMP	Loop-mediated isothermal amplification
LNYV	Lettuce necrotic yellows virus
MRStV	Maize red stripe virus (Currently referred to as HPWMoV)
MARC	Mount Albert Research Campus (at The Institute for Plant and Food Research, Auckland)
NAD5	NAD dehydrogenase subunit 5
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
ORF	Open reading frame
P#	Protein #
PCR	Polymerase chain reaction
PFR	The Institute for Plant and Food Research
PiVB	Pistachio virus B
PolyA	Poly adenylated
PPSMV-1	Pigeonpea sterility mosaic virus 1
PPSMV-2	Pigeonpea sterility mosaic virus 2
PUUV	Puumala virus
PVP	Polyvinylpyrrolidone
RACE	Rapid amplification of cDNA ends
RCA	Rolling circle amplification

RdRp	RNA dependent RNA polymerase
RGSV	Rice grassy stunt virus
RNA	Ribonucleic acid
RLBV	Raspberry leaf blotch virus
RRV	Rose rosette virus
RT-PCR	Reverse transcription polymerase chain reaction
RT-qPCR	Quantitative real-time reverse transcription polymerase chain reaction
RVFV	Rift valley fever virus
RYRSaV	Redbud yellow ringspot-associated virus
SGS	Second generation sequencing
siRNA	Small interfering RNA
SMRT	single molecule real-time (sequencing)
SMS	Single molecule sequencing
sRNA	Small RNA
SYNV	Sonchus yellow net virus
TBE Buffer	Tris/Borate/EDTA Buffer
TEM	Transmission electron microscope
TGS	Third generation sequencing
TiRSaV	Ti ringspot-associated virus
TSWV	Tomato spotted wilt virus
UTR	Untranslated region
WBYVV	Wooly burdock yellow vein virus
WMoV	Wheat mosaic virus

Abstract

Chlorotic spot symptoms were observed on a specimen of a tree endemic to New Zealand, karaka (*Corynocarpus laevigatus*), leading to an investigation of the potential viruses associated with the symptoms. RNA-seq analysis identified a new-to-science emaravirus along with two other novel DNA sequences. Full genome sequencing, a survey to identify the distribution of symptomatic karaka trees, and symptom correlation studies were undertaken.

Emaravirus is an established genus of viruses with plant hosts. Each species within the genus has a segmented, negative-sense monocistronic RNA genome comprising five to eight segments, and is vectored by eriophyid mites. Five negative sense segments were identified by high throughput sequencing, which were all confirmed by Sanger sequencing. The sequence of each segment was completed following rapid amplification of cDNA ends (RACE). Sequence comparison with other established emaravirus sequences revealed that this represents a unique virus, with only 31.4% amino acid identity within the replicase of its closest relative European mountain ash ringspot-associated virus.

A diagnostic test was developed and implemented to assess the virus' distribution across New Zealand. The novel virus was primarily found throughout Auckland in both naturally occurring and planted karaka trees, with only a few examples identified beyond this region.

Using the diagnostic test, the presence of virus was confirmed to be highly correlated with the chlorotic spot symptoms. In addition, the virus was found to be only detectable within symptomatic regions of the leaf and not systematic. Systemic movement may be dependent on the suspected eriophyid mite vector, the endemic karaka gall mite.

Evidence is presented for the sequence of a unique emaravirus, that represents the first endemic plant virus reported in New Zealand. The virus is strongly associated with the observed symptoms; therefore, we suggest the name *Karaka Okahu purepure virus*, which describes the chlorotic spots in *te reo Māori*.

Chapter One: General Introduction

1.1 Introduction

Viruses are ubiquitous biological agents that are associated with every type of cellular life and are the most abundant biological entities on earth (Koonin et al., 2006). As of the latest International Committee on Taxonomy of Viruses (ICTV) there were 6,590 different species of viruses divided into 168 families and 55 orders and of these, many species are known to infect plants (ICTV, 2019). Plant viruses are abundant, diverse and have many different modes of transmission and can show a range of symptoms on potentially different plant species. Symptoms and morphology are often not enough to diagnose the specific virus infecting a symptomatic plant and, therefore, it is important to develop modern molecular techniques to be able to detect and study these viruses more effectively.

1.2 Family Fimoviridae

Fimoviridae is a family of plant viruses in the order of *Bunyavirales* with linear, segmented, single-stranded, negative sense RNA genomes (Elbeaino et al., 2018). Negative sense RNA viruses have the defining feature of having a genome that is opposite polarity to host messenger RNA and are replicated through a complementary RNA (cRNA) intermediate (York et al., 2013). Members of the *Fimoviridae* family are known to transmit to plants through eriophyid mite vectors and produce double membrane bound bodies in the cytoplasm that are suspected to be the virion (Elbeaino et al., 2018; Mielke & Muehlbach, 2007). The *Fimoviridae* family currently includes the relatively new *Emaravirus* genus as the only member (Elbeaino et al., 2018). Species and candidates within this genus are referred to henceforth as emaravirids and are shown in Table 1.1.

Table 1.1 - Current confirmed and candidate emaravirus species with the virus abbreviation, the host plant that the virus infects, the number of genome segments present and where the virus was first referenced.

Virus Name	Abbreviation	Host Plant	Recognised Species?	Genomic segments	Reference where first described
Actinidia chlorotic ringspot-associated emaravirus	AcCRaV	*Actinidia* spp.	Yes	5	Zheng et al. (2017)
Actinidia emaravirus 2	AcEV-2	*Actinidia* spp.	Candidate	6	Wang et al. (2020)
Alfalfa ringspot-associated virus	ARaV	*Medicago sativa* L.	Candidate	4	Samarfard et al. (2020)
Aspen mosaic-associated virus	AsMav	*Populus tremula*	Candidate	5	von Bargen et al. (2020)
Blackberry leaf mottle-associated virus	BLMaV	*Rubus* spp.	Candidate	5	Hassan et al. (2017)
European mountain ash ringspot-associated emaravirus	EMARaV	*Sorbus aucuparia*	Yes	6	Mielke and Muehlbach (2007)
Fig mosaic virus	FMV	*Ficus carica*	Yes	6	Elbeaino et al., (2009)
High Plains wheat mosaic emaravirus	HPWMoV	*Zeas mays; Triticum aestivum*	Yes	8	Tatineni et al. (2014)
Jujube yellow mottle-associated virus	JYMaV	*Ziziphys jujuba* Mill.	Candidate	6	Yang et al., (2019)
Lilac chlorotic ringspot-associated virus	LiCRaV	*Syringa* spp.	Candidate	5	Wang et al. (2020)
Palo verde broom virus	PVBV	*Parkinsonia florida*	Candidate	4	Ilyas et al., (2018)
Perilla mosaic virus	PerMV	*Perilla frutescens* var. *crispa*	Candidate	10	Kubota et al. (2020)
Pigeonpea sterility mosaic emaravirus 1	PPSMV1	*Cajanus cajan*	Yes	5	Elbeaino et al. (2014)
Pigeonpea sterility mosaic emaravirus 2	PPSMV2	*Cajanus cajan*	Yes	5	Elbeaino et al. (2015)
Pistacia virus B	PiVB	*Pistacia* spp.	Yes	7	Buzkan et al. (2019)
Raspberry leaf blotch emaravirus	RLBV	*Rubus* spp.	Yes	8	McGavin et al. (2012)
Redbud yellow ringspot-associated emaravirus	RYRSaV	*Cercis* spp.	Yes	5	Di Bello et al. (2016)
Rose rosette emaravirus	RRV	*Rosa* spp.	Yes	7	Laney et al. (2011)
Ti Ringspot-associated virus	TiRSaV	*Cordyline fruitcosa*	Candidate	5	Olmedo-Velarde et al., (2019)

1.3 Genus Emaravirus

The *Emaravirus* genus was recently established in 2012 and has the *European Mountain Ash ringspot-associated virus* (EMARaV) as its type species (Mielke-Ehret & Mühlbach, 2012; Mielke & Muehlbach, 2007). The genus is characterised by each member species having a segmented single-stranded, negative-sense RNA genome, having double-membraned bodies present in the cytoplasm of infected plants, and most being vectored by eriophyid mites (Mielke-Ehret & Mühlbach, 2012). There are currently nine accepted species in the genus and five candidate species as shown in Table 1.1 above (Elbeaino et al., 2018). High Plains wheat mosaic virus (HPWMoV) has been referred to as High Plains virus (HPV), maize red stripe virus (MRStV) and wheat mosaic virus (WMoV) in previous publications but will be referred to as HPWMoV throughout this thesis based on the most current research using this name. At the present time, fig mosaic virus is the only emaravirus reported to be in New Zealand (Veerakone et al., 2015).

1.3.1 Host Range

Each emaravirus appears to infect one plant species, although there are some exceptions such as HPWMoV, which is known to infect many species, and FMV, which has been detected in cyclamen in addition to fig (Elbeaino et al., 2018; Tatineni et al., 2014). Emaravirids are transmitted by eriophyid mites that appear to be specific to their host plants, as shown in Table 1.1 above. This makes it difficult to transmit these viruses to other, well studied, model plants as the mites will not feed upon these other species, even when starved. Kulkarni et al. (2002) showed that *Aceria cajani*, the mite that transmits PPSMV-1 and PPSMV-2 survived for only 13 hours after being removed from its pigeonpea food source. Mechanical inoculation onto model plants such as *Nicotiana benthamiana, N. tabacum* and *N. clevelandii* have been tried with mixed results. FMV has been mechanically transmitted to various *Nicotiana* species and PPSMV could be transmitted with some difficulty to *N. clevelandii* and *N. benthamiana* with a success rate between 10%-40% (Kumar et al., 2003; Serrano et al., 2004). HPWMoV has been transmitted by vascular puncture inoculation from

barley (*Hordeum vulgare* L.) to maize (*Zeas mays* L.), but not with leaf-rub mechanical inoculation (Louie et al., 2006). AcCRaV was shown to transmit to *N. benthamiana* through mechanical inoculation, with virus being detected in 50% of inoculated plants as well as showing chlorotic spot symptoms on systemic leaves (Zheng et al., 2017). It appears that even though the mites may not be able to transmit the virus to other plant species, mechanical inoculation could be possible depending on the virus and plant species, as well as the inoculation technique used.

1.3.2 Symptomatology

Emaravirids display symptoms on each of their host plants that are different to each of the other emaravirus symptoms with some examples shown in Figure 1.1. Most infected host plants have chlorotic ringspots, mottling, yellowing of veins and mosaic patterns occurring separately or together on leaves that is typical of plant virus infections (Di Bello et al., 2016; Elbeaino et al., 2015; Mielke-Ehret & Mühlbach, 2012; Zheng et al., 2017). Emaraviruses can affect the fruit and nut development of plants that produce them, causing economic impact. FMV, BLMaV, RLBV, HPWMoV, PPSMV-1, PPSMV-2 and PiVB all cause reduction in yields of their respective host fruit or nut (Buzkan et al., 2019; Elbeaino et al., 2015; Hassan et al., 2017; Mielke-Ehret & Mühlbach, 2012). FMV results in reduced yield and premature fruit drop, and may even show mosaic on fruit produced by infected trees (Açikgöz & Timur Döken, 2001). These effects were discovered while studying the diseases that each emaravirus causes, or are a partial cause for, whereas the trees that do not have economic importance are less studied. It may be that these economically less important trees produce additional symptoms when infected that are not known yet. RRV causes arguably the most severe symptoms with rapid stem elongation that is followed by breaking of axillary buds, increased thorniness, bright red pigmentation and leaf deformation causing the characteristic “Witches Broom” appearance of rose rosette disease (Laney et al., 2011). The symptoms produced from infection by each emaravirus show typical viral infection symptoms, but also vary among plant species. A new emaravirus would not be easily identified based on symptoms alone and trying to diagnose like this may cause false positives. For some known emaraviruses, such as EMARaV and RRV, the presence of symptoms may be useful to diagnosis of emaravirus

infection due to symptoms produced having a strong correlation to the presence of virus. For other viruses, such as BLMaV where ten viruses are associated with the black-berry yellow vein disease, symptoms may be caused by the presence of multiple viruses and a diagnosis based on symptoms alone would be impossible (Hassan et al., 2017).

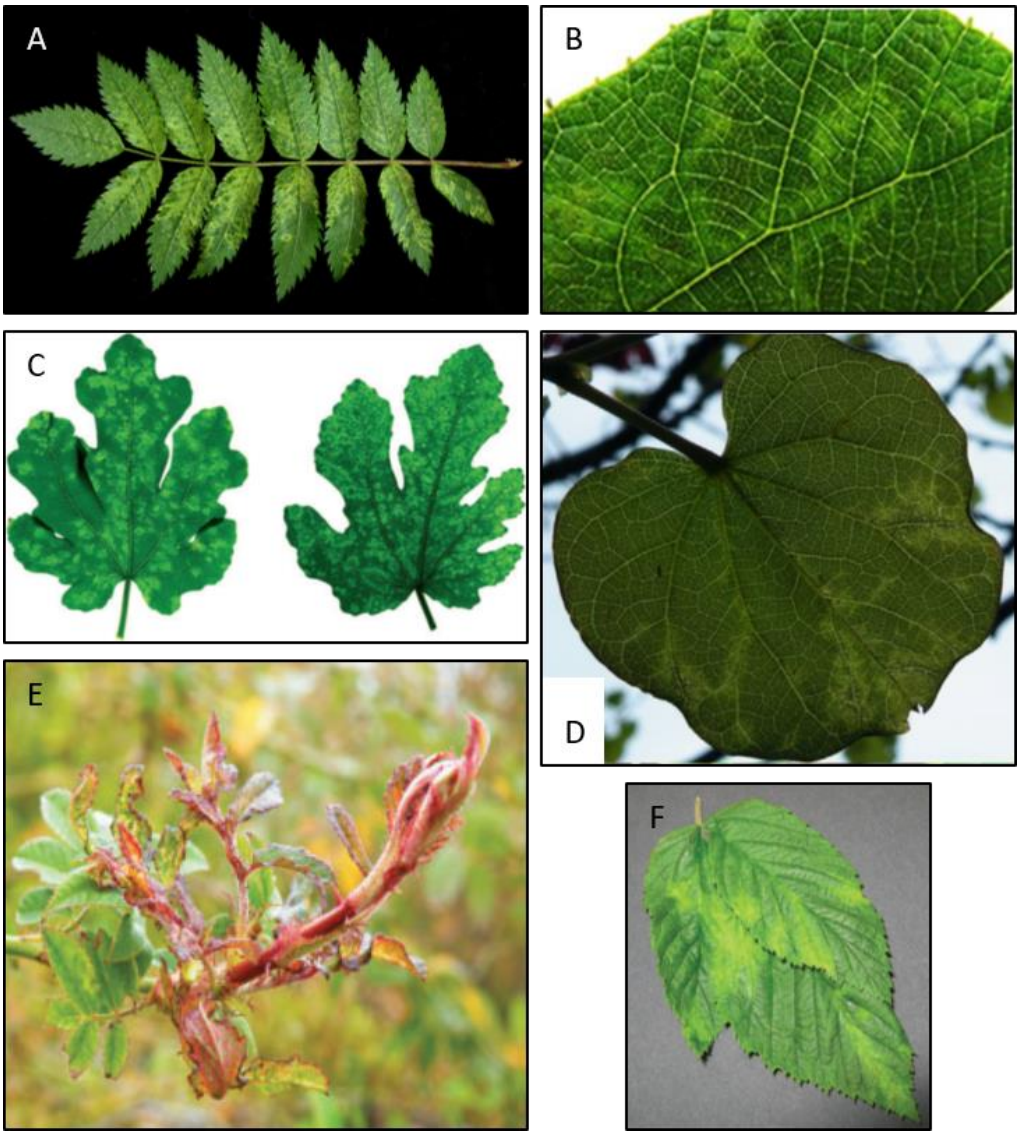


Figure 1.1 - Symptoms of emaravirus infected plants. A) European mountain ash (*Sorbus aucuparia*) leaves infected with EMARaV (Mielke-Ehret & Mühlbach, 2012). B) Kiwifruit (*Actinidia chinensis*) leaf infected with AcCRaV (Zheng et al., 2017). C) Fig (*Ficus carica*) leaves infected with FMV (Elbeaino et al., 2009). D) Redbud (*Cercis canadensis* L.) leaf infected with RYRSaV (Di Bello et al., 2016). E) Rose (*Rosa multiflora*) leaves infected with RRV (Laney et al., 2011). F) Raspberry (*Rubus* spp.) leaves infected with RLBV (McGavin et al., 2012).

1.3.3 Transmission

Emaravirids have been shown to be transmitted through eriophyid mites specific to each virus, or at least is strongly suspected of being transmitted through these mites (Shown in Table 1.1). Information on the mechanism of virus transmission is extremely limited at this time. Generally, eriophyid mites have short stylets that can only pierce the epidermal cells on the under-surface of plant leaves of host plants that the mites browse on. Due to this it follows that for effective virus transmission, these cell types must be able to be infected (Mielke-Ehret & Mühlbach, 2012; Serrano et al., 2004). Virus-like particles were found in parenchyma and subepidermal leaf cells of plants infected with FMV, PPSMV and HPWMoV, which confirms that these cell types can be infected by emaraviruses (Mielke-Ehret & Mühlbach, 2012; Serrano et al., 2004). EMARaV has been found by reverse transcription polymerase chain reaction (RT-PCR) and electron microscopy analyses, but little else is known about *P. pyri* (Mielke-Ehret et al., 2010). Other emaravirids have been shown to, or are suspected to, be transmitted by eriophyid mites, a key characteristic of the genus. FMV, RRV, RLBV, HPWMoV, PPSMV-1 and PPSMV-2 have been shown to be transmitted by identified mite species (Elbeaino et al., 2015; Mielke-Ehret & Mühlbach, 2012). The mite vector for RRV is shown in Figure 1.2. AcCRaV, RYRSaV, WBYVV, PiVB and BLMaV have either been found in unidentified, or undescribed, mites living on leaves of the plant species, or have been suspected of having a mite vector (Bi et al., 2012; Buzkan et al., 2019; Hassan et al., 2017; Zheng et al., 2017). Eriophyid mites are typically spread by wind and it is by this method that the virus they carry might be spread as well (Slykhuis, 1961). While a majority of emaraviruses have been detected in identified and unidentified mites, little is known about the virus interaction with the mites (Bauchan et al., 2019; Mielke-Ehret & Mühlbach, 2012; Tatineni et al., 2014). It is unknown whether the virus is transmitted in a propagative or semi-propagated manner; it may be that each virus could actually be a mite virus that is vectored by the plant "host" (Jones, 2014). Vegetative transmission is typical for plant viruses and several emaravirids have been shown to transmit through grafting symptomatic branches onto healthy branches. PPSMV, FMV and EMARaV were shown to be passed by cuttings, or to be graft transmissible with the virus moving systemically through the plant (Mielke-Ehret & Mühlbach, 2012). No vertical transmission by seed or pollen has been reported for any emaravirus yet except for HPWMoV, which showed a very low seed transmission rate of

0.008% (3 of 38,482 seedlings showed typical mosaic symptoms). High temperatures in the greenhouse that would not occur in nature were suggested to have influenced this (Forster et al., 2001). Virus was detected in the seeds of EMARaV infected *S. aucuparia* and PPSMV-1 infected *C. cajanus*, with the virus present only in the seed coat of plants infected with PPSMV-1, not the cotyledons. The virus was not transmitted to the new seedling when germinated (Divya et al., 2005). Mechanical transmission, as described in section 1.2.1, is possible depending on the virus, plant species, and inoculation method. No transmission by soil has been reported for any emaravirus so far.

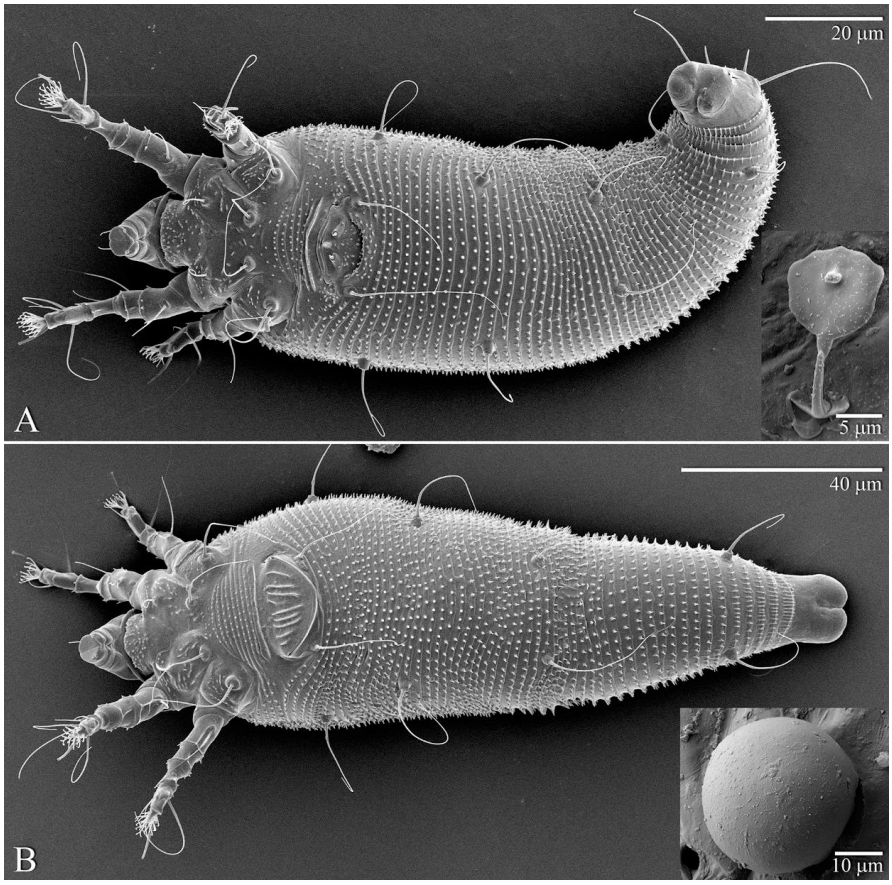


Figure 1.2 – Low-Temperature Scanning Electron Microscopy images of *Phyllocoptes fructiphilus*, the eriophyid mite vector of RRV. A) Male with spermatophore insert. B) Female with egg insert.

1.3.4 Epidemiology

The spread of each emaravirus is not well studied in most species although it is suspected that the eriophyid mite vectors are involved. For EMARaV, the spread of the virus is limited by the range of its suspected mite vector, *Pytoptus pyri* (Mielke & Muehlbach, 2007).. As each emaravirus could be vectored by a mite that is specific to each plant host there have not been other plant hosts found, with the exception of HPWMoV, which infects wheat (*Triticum aestivum*) and maize (*Zea mays*), and has been detected in several other economic and herbaceous species (Alemandri et al., 2017; Mielke-Ehret & Mühlbach, 2012). Little is known about the epidemiology of emaraviruses due to the genus being relatively new and the relationship between virus and mite not being understood.

1.3.5 Virion Morphology

Double membrane-bound bodies (DMBs), as shown in Figure 1.3, are found in the cytoplasm near the endoplasmic reticulum (ER) and Golgi complexes in emaravirus infected plants. DMBs are one of three components required for genus classification and are considered to be the virions (Mielke-Ehret & Mühlbach, 2012). The DMB sizes varying between species and are between 80 – 150 nm as described in the emaravirus review paper by (Elbeaino et al., 2009; Kumar et al., 2003; Laney et al., 2011; Mielke & Muehlbach, 2007; Tatineni et al., 2014). These structures are similar to bodies present in tospovirus infections and were shown to be present in EMARaV, FMB, RRV, PPSMV-1, and HPWMoV infected leaves. Filamentous structures were also detected in PPSMV-1, HPWMoV and RLBV infected leaves with the latter only having filamentous structures present and no sign of any DMBs. Filamentous elements are also present alongside DMBs when viewing naturally infected fig leaves, which may explain why both DMBs and filamentous particles were also observed in HPWMoV (Elbeaino et al., 2009). It is unknown whether these filamentous particles are part of these emaraviruses.

Subsequent studies of RLBV and other emaraviruses appear to have not attempted to detect DMBs. As the DMBs are found near the ER and Golgi complex, it is suggested that particle morphogenesis occurs as described for tospoviruses and takes place at these intracellular host membranes. Of the more recently described emaraviruses, AcCRaV is the only virus that DMBs have been reported for. DMBs between 95 and 110 nm were found to be frequently located near ER membranes within the cytoplasm of leaf tissue (Zheng et al., 2017). Sequence information of emaraviruses seems to be the focus in modern studies and so only a few of the publications describing these new viruses have included images and analyses of DMBs present in infected tissue.

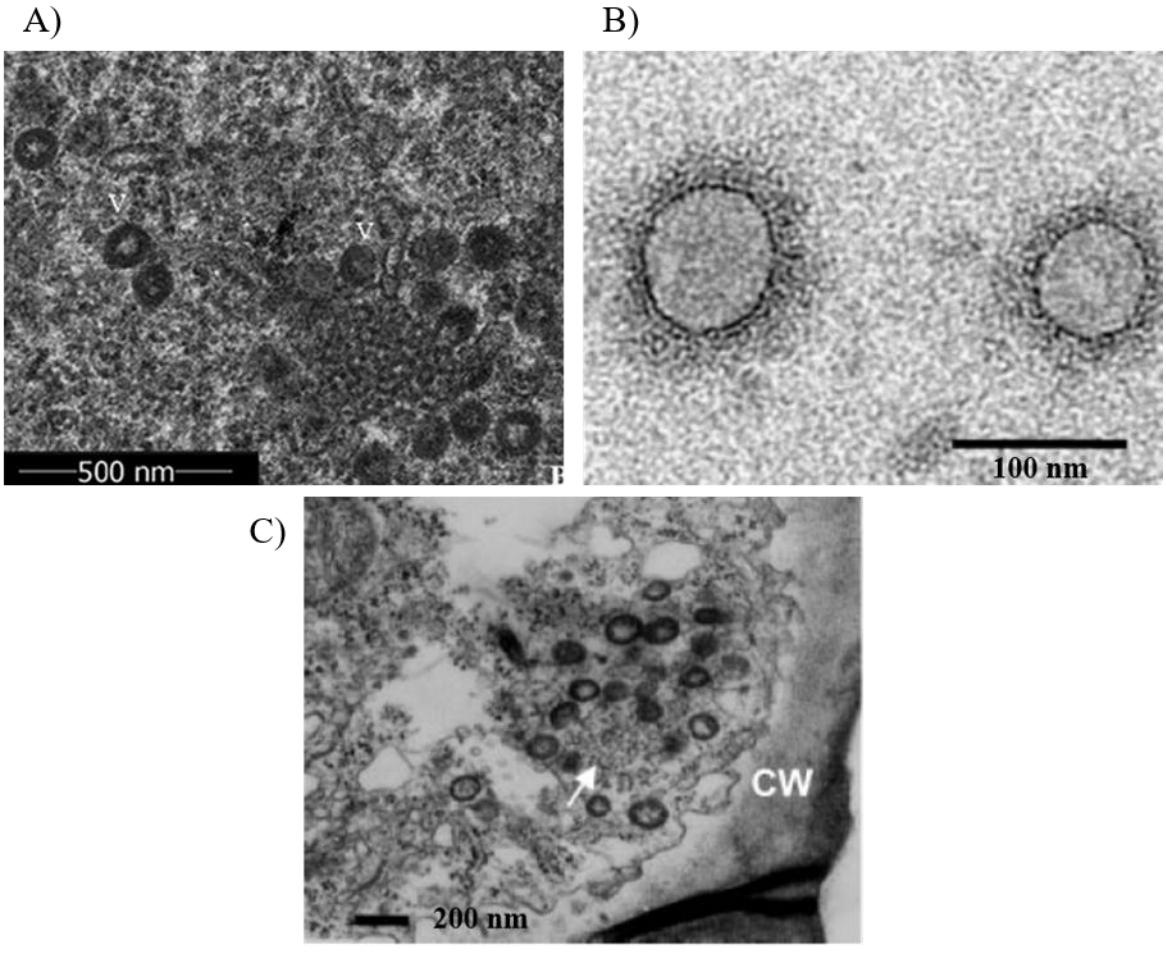


Figure 1.3 – Double membraned bodies (DMBs) present in AcCRaV (top-left) in an ultra-thin leaf slice of kiwifruit leaves (Zheng et al., 2017), EMARaV purified from infected mountain ash leaves (Mielke-Ehret & Mühlbach, 2012), and FMV from a naturally infected fig leaf. The arrow in this image points to filamentous elements that usually accompany DMBs. CW is cell wall (Elbeaino et al., 2009). The bar in each image represents the scale.

1.3.6 Genome Structure and Sequence Analysis

Each emaravirus has between four and ten negative-sense RNA segments that make up the genome, with each segment encoding a single protein. These segments are generally ordered by their length from largest to smallest, with RNA 1 encoding the RNA-dependent, RNA polymerase, RNA 2 encoding the glycoprotein precursor, RNA 3 encoding the nucleocapsid and RNA 4 encoding the movement protein. RNAs 5 to 10 can be present and of various sizes but are listed after the first four segments that encode the proteins of known functions. Figure 1.4 shows the genome representation of AcCRaV, an example of an emaravirus with five genomic RNA segments.

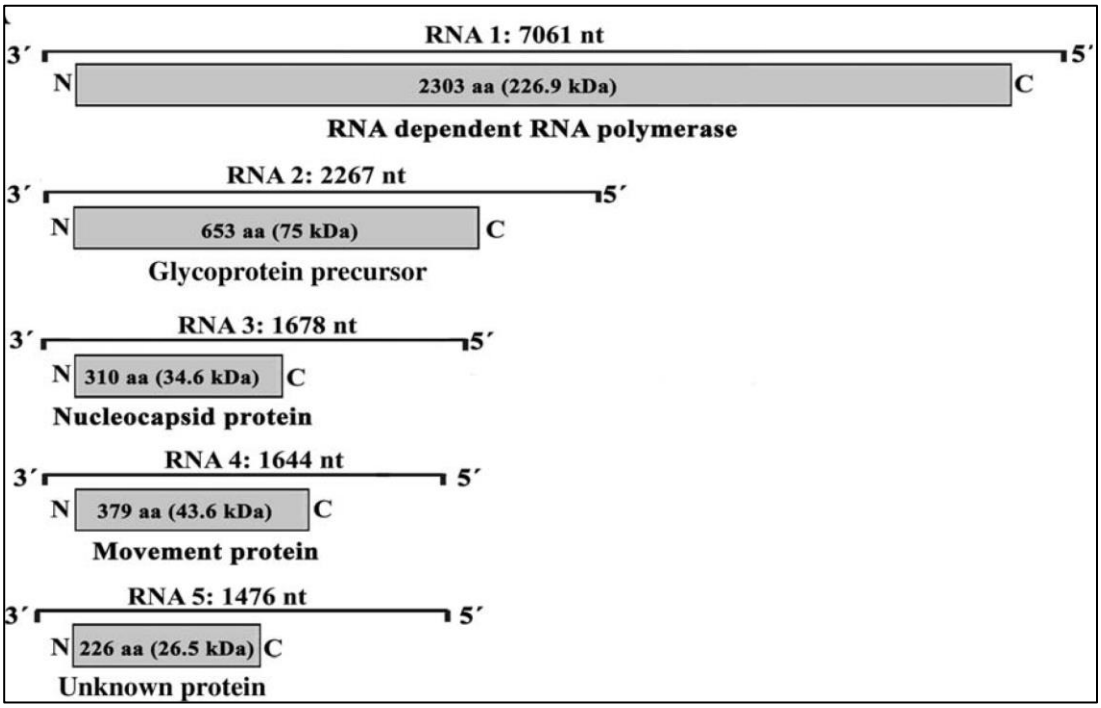


Figure 1.4 – Schematic representation of the genome of AcCRaV, an emaravirus species with five segmented RNA segments (Zheng et al., 2017).

RNA 1 encodes the RNA-dependent, RNA polymerase (RdRp) and is between 7009 and 7062 nucleotides in length. This protein functions as an enzyme that catalyses the replication of RNA from and RNA template and is used in viral transcription (York et al.,

2013). Emaravirus RdRps show similarity to RdRps of bunyaviruses and tenuiviruses, and also have the same conserved motifs of premotif A and A-E, as well as a conserved endonucleolytic centre that may be involved in cap snatching (Mielke-Ehret & Mühlbach, 2012). The five motifs (A to E) in the translated amino acid sequence form the core of the polymerase module of the RdRp active site, shown in Figure 1.5. Motif A (DASKWS) and C (SDD) are 100% conserved across all known emaravirids and are thought to be involved in divalent-metal cation bonding, as shown in RRV (Hassan et al., 2017; Laney et al., 2011). Motif B ($QGNXNX_2S_2$) is also conserved across sequenced emaraviruses, although some amino acid residues are interchangeable, and is thought to bind RNA through the motility promoted by the glycine (Di Bello et al., 2016). Motif D (KK) is thought to have catalytic activity due to being located near motif A when viewing the tertiary structure of the translated protein (Laney et al., 2011). Motif E (EFLST) has been thought to be involved in cap-snatching by RRV, as is highly conserved amongst emaraviruses except in RLBV where the sequence is EFLSS (Di Bello et al., 2016; Laney et al., 2011). In FMV, there are three residues within a 19 aa sequence upstream of motif A that are conserved across bunyavirus RdRps as well as other emaraviruses that has been called Pre-motif A. The 19 aa sequence is highly conserved within emaraviruses (Elbeaino et al., 2009). The N-terminus of AcCRaV RNA 1 contains endonuclease domains likely involved in cap-snatching. This genome expression strategy cleaves cellular capped RNAs to generate capped fragments that are used as primers for viral RNA synthesis, and has been proposed as functioning for members of the *Bunyaviridae* family and *Tenuivirus* genus (Rao et al., 2003; Zheng et al., 2017). These motifs are highly conserved in other emaravirids to varying degrees such as FMV, RYRSAV, AcCRaV and BLMaV (Di Bello et al., 2016; Elbeaino et al., 2009; Hassan et al., 2017; Zheng et al., 2017).

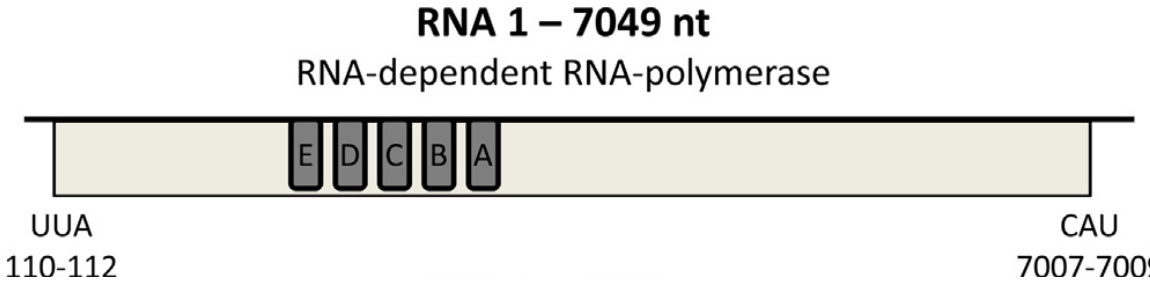


Figure 1.5 – Schematic representation of the genomic segment RNA 1 from RYRSaV showing the approximate locations of motifs A to E (Di Bello et al., 2016).

RNA 2 is between 2135 and 2335 nucleotides in length and encodes a glycoprotein precursor that has the same domain that is conserved in genus *Phlebovirus* (of *Bunyaviridae*) (Hassan et al., 2017; Mielke-Ehret & Mühlbach, 2012). In a viral genome, the glycoprotein forms part of the capsid that protects the virion from host nucleases (Banerjee & Mukhopadhyay, 2016). *In silico* analysis of the BLMaV glycoprotein revealed four transmembrane helices and six glycosylation sites, as well as a tetrapeptide sequence. This latter sequence may indicate the cleavage site to create two functional, mature glycoproteins predicted to form homo-dimers (Hassan et al., 2017; Laney et al., 2011). This tetrapeptide was identical to sequences of EMARaV, FMV and RRV and immediately follows the second cleavage site found in the glycoprotein of EMARaV and FMV, which is also present in bunyaviruses (Hassan et al., 2017). Analysis of the glycoprotein of AcCRaV and RYRSaV also predicted two cleavage sites, although the two sites found for RYRSaV had low probability scores (Di Bello et al., 2016; Zheng et al., 2017). A signal peptide is present in the first 20 amino acids; the glycoprotein appears to be secreted to the endoplasmic reticulum (ER) in RRV, FMV and RYRSaV where the virus likely acquires membranes for its particles (Di Bello et al., 2016; Laney et al., 2011). A schematic representation of RYRSaV is presented in Figure 1.6.

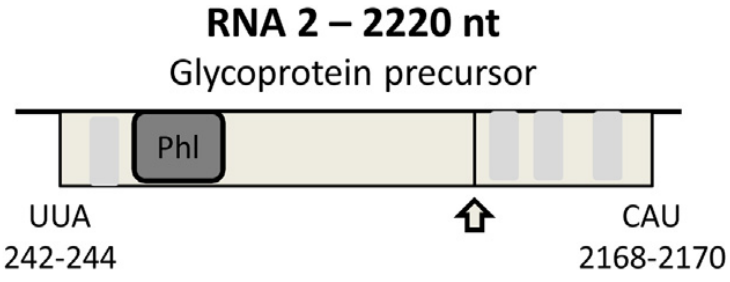


Figure 1.6 - Schematic representation of the genomic segment RNA 2 from RYRSaV showing the approximate locations of Phlebovirus conserved domain (Phl), four putative N-glycosylation sites (light grey) and the predicted cleavage site (arrow indicating approximate position) (Di Bello et al., 2016).

RNA 3 is between 1365 and 1678 nucleotides in length and encodes a nucleocapsid protein. When EMARaV was initially reported, the RNA 3 showed similarity to HPWMoV and PPSMV sequences, which were unclassified at the time (Mielke & Muehlbach, 2007). The initial analysis of EMARaV involved purifying the protein encoded by RNA 3, called P3, to support the assumption that RNA 3 encodes the nucleocapsid. This was based on HPWMoV

having P3 as its nucleocapsid (Mielke & Muehlbach, 2007). Subsequent new emaravirid RNA 3 also share high similarity to each other, with three conserved amino acid regions identified by Elbeaino et al. (2009). The encoded proteins of several species also contain stretches of positively charged amino acids that are likely involved in RNA binding and may form homo-dimers, as would be expected of a nucleocapsid (Di Bello et al., 2016; Laney et al., 2011; Zheng et al., 2017). There are three conserved amino acid stretches conserved across all emaraviruses, and in AcCRaV they are NVLSFNK, NRLA and GYEF (Zheng et al., 2017). In BLMaV, RYRSaV, HPWMoV and RLBV there is a substitution of arginine to lysine in the NRLA conserved region (Hassan et al., 2017). In FMV, NVVS(F/Y)N is present rather than NVLSFNK while the other two regions are present (Elbeaino et al., 2009). These regions are reported to be possible nucleocapsid motifs in the Emaravirus genus (Hassan et al., 2017). A schematic representation of RYRSaV is presented in Figure 1.7.

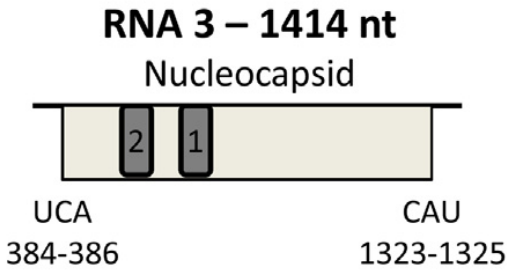


Figure 1.7 - Schematic representation of the genomic segment RNA 3 that encodes the nucleocapsid from RYRSaV. The approximate locations of two of the three conserved emaravirus domains are shown (Di Bello et al., 2016).

RNA 4 is between 1348 and 1675 nucleotides long and, while it has not been proven, evidence points to RNA 4 encoding the putative movement protein for emaraviruses. Initially, when EMARaV was first described, there were no regions in the RNA 4 sequence that indicated that it encoded a movement protein (Mielke & Muehlbach, 2007). Subsequent emaravirid sequences revealed that RNA 4 could encode the movement protein. The protein encoded by RRV RNA 4 contains an ATPase and a DnaK motif, both of which were shown to be involved in cell-to-cell movement in *Closterovirus* species (Laney et al., 2011). Evidence suggests that the protein encoded by RLBV RNA 4 that RNA 4 is also involved in movement, with a signal peptide at its N-terminus. Fusion proteins with green fluorescent protein were

localised to the plasma membrane and plasmodesmata in transgenic *N. benthamiana*, while also co-localising with the tobacco mosaic virus movement protein 30K in the plasmodesmata indicating that the RNA 4 protein was involved in movement (McGavin et al., 2012; Mielke-Ehret & Mühlbach, 2012). Recently, two more RNA segments have been discovered from EMARaV with the encoded protein from one of these segments s showing amino acid sequence identity between 15.9% and 29.5% with other emaravirid RNA 4 encoded movement proteins (von Bargen et al., 2019). The RNA 4 originally described shows similarity to RNA 5 in RYRSaV indicating RNA 4 was initially misread (Di Bello et al., 2016).

RNAs 5 to 10 in emaravirids are highly varied, if present, and encode proteins with unknown functions. There is similarity between certain segments in some species but not to the extent of the first four RNA segments. Initial analysis of some species revealed several segments and were later found to have additional RNA segments present. Three additional segments found in RRV, bringing the total up to seven segments, after discovery of new emaraviruses with more than four segments warranted further investigation. These additional segments also share homology with other emaravirus genomic RNA 5 to 10 segments (Di Bello et al., 2015). Another recent addition was in the type species, EMARaV, where two additional segments were discovered in an isolate extracted from different mountain ash species, one being the putative movement protein that was conspicuously absent, as described above (von Bargen et al., 2019). In RYRSaV, *in silico* analysis of RNA 5 identified two O-glycosylation sites that potentially interact with the protein encoded by RNA 2, and may be involved in the secretory pathway indicating it could function in assembly of the virion (Di Bello et al., 2016). The authors also comment that the five genomic segments discovered for RYRSaV may not represent the entire genome due to additional segments having been discovered in other emaraviruses. Adding to this, there may be variation in the number of genomic RNA segments present between different isolates of the same emaravirid and it may be that RNA 1 to 4 are the required segments for replication and transmission, and the extra RNA segments provide pathogenicity (Di Bello et al., 2016; Stewart, 2016). Recently, RNA 7 and RNA 8 of HPWMoV were shown to involved in RNA silencing suppression (Gupta et al., 2018; Gupta et al., 2019). Additional study is required of different isolates of already described emaraviruses to reveal the extent of such variations.

The 5’ and 3’ termini of each RNA among emaravirids are almost identical between species and show reverse complementarity. For example, EMARaV has the sequence 5’-AGUAGUGUUCUCC … GGAGUUCACUACU-3’ at its termini (Mielke & Muehlbach, 2007). Sequences similar to this were also found at RNA termini of the *Hantavirus* and *Orthobunyavirus* genera, in the *Bunyaviridae* family. In contrast, no similarity was found between the sequences of emaraviruses and that of tospoviruses and tenuiviruses (Mielke-Ehret & Mühlbach, 2012). The sequences in the RNA termini show reverse complementarity and are between 19 to 23 nt long. These complementary ends are believed to be function in a regulatory capacity during encapsidation, replication and transcription. They are thought to form a “panhandle” structures that functions as a promoter for transcription and replication, as seen in orthobunyaviruses and hantaviruses (Elliott, 2014; Mielke-Ehret & Mühlbach, 2012). Because the 5’ and 3’ termini are highly conserved, if not identical, between emaraviruses, it is important to verify the sequences of 3’ and 5’ ends of new species to provide additional evidence that the virus belongs to the emaravirus genus.

1.3.7 Genetic diversity and evolution

Genetic diversity between emaravirids is substantial. The most conserved segment of the emaravirus RNA genome appears to be RNA 1 encoding the RdRp with FMV and PPSMV-2 showed the greatest identity overall at 72.1% at the amino acid level, as shown in Table 1.2. RdRps and glycoproteins appear to be more conserved, probably due to their conserved function, while other proteins are not as conserved, possibly due to the different requirements for infecting their respective hosts (Appendix C).

When compared to other RNA plant viruses there is a close relationship between emaraviruses. The phylogenetic tree in Figure 1.8 shows the relationship between emaraviruses for the amino acid sequences of RNA 1 and 3 (Wang et al., 2020) . The emaraviruses form a monophyletic group for all proteins, indicating they have common ancestry. The exception to this is for P4 where the EMARaV sequence does not group with other emaraviruses. This can be explained due to the tree being constructed with the original

amino acid sequence for EMARaV P4, where if the tree was constructed with the putative movement from the EMARaV (*S. intermedia*) isolate, the sequence is with the subclade with AcCRaV and RYRSaV (von Bargen et al., 2019).

Table 1.2 – Percentage similarity of amino acid sequences from translated sequences of RNA 1 for each emaravirus.

	RNA 1 (P1) RdRp												
	AcCRaV	BLMaV	EMARaV	FMV	HPWMoV	JYMaV	Palo ver.	PiVB	PPSMV1	PPSMV2	RLBV	RRV	RYRSaV
BLMaV	46.4												
EMARaV	54.3	49.0											
FMV	48.8	66.4	48.6										
HPWMoV	33.5	31.9	33.0	32.6									
JYMaV	34.6	33.9	33.9	34.2	41.1								
Palo verde	32.0	32.0	33.5	33.2	50.0	43.0							
PiVB	47.1	67.6	49.1	69.6	32.4	34.6	33.2						
PPSMV1	47.7	52.5	47.9	51.8	31.7	33.4	33.4	53.5					
PPSMV2	47.2	66.6	49.4	72.1	33.0	34.4	33.5	69.9	53.2				
RLBV	34.6	33.9	34.4	33.6	41.9	70.8	42.1	34.0	33.5	33.5			
RRV	47.1	67.7	49.4	68.5	33.1	34.2	33.4	70.5	53.2	68.1	33.4		
RYRSaV	64.5	46.1	53.9	46.8	33.1	34.2	34.5	48.1	47.3	46.7	34.7	47.0	
TiRSaV	35.3	34.4	34.5	33.7	41.9	44.0	41.7	34.0	33.2	34.5	44.7	33.4	33.9

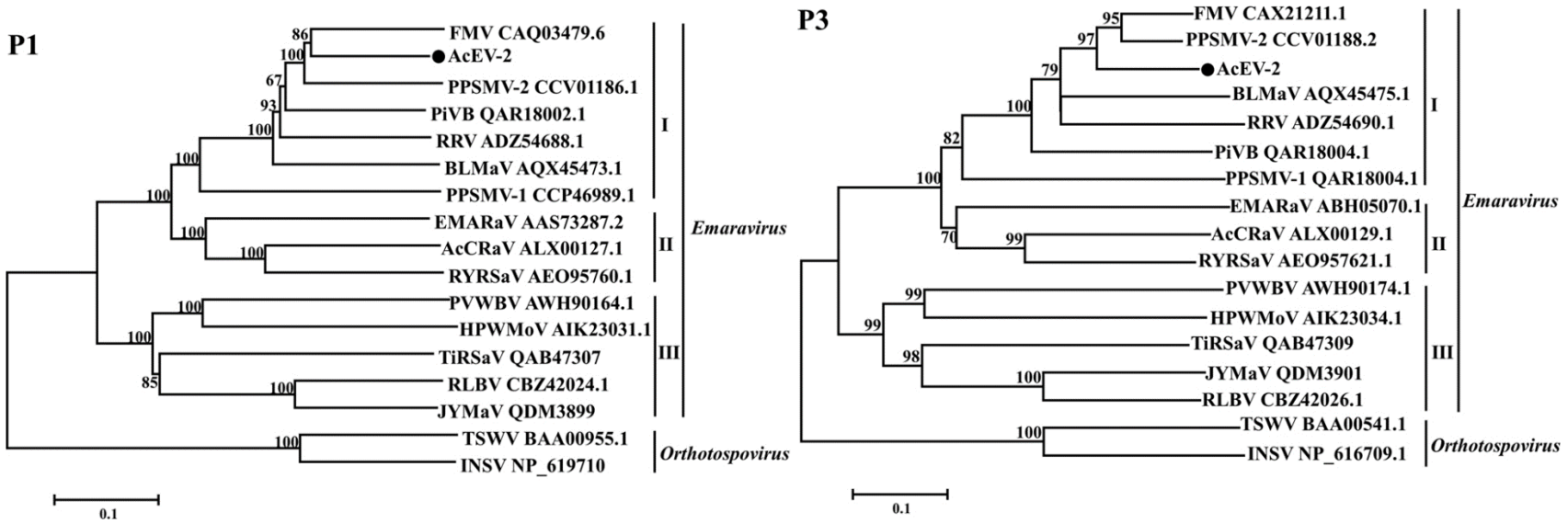


Figure 1.8 - Phylogenetic tree of emaravirus amino acid sequences for proteins 1 and 3 (Wang et al., 2020). Representative members of the genus *Orthotospovirus* are included as the outgroup. TSWV: tomato spotted wilt virus, INSV: Impatiens necrotic spot orthotospovirus.

While there appears to be a high level of genetic diversity within the same species of emaravirids it is less studied than the diversity between different species, this type of analysis is beginning to be reported. HPWMoV was found to have a 12.5% sequence divergence between two RNA 3 consensus sequences extracted from wheat, whereas no heterogeneity was observed in the other seven genomic segments (Tatineni et al., 2014). In another study, five isolates of HPWMoV from two states in the United States of America, Ohio and Kansas, were sequenced and compared to the type isolate from Nebraska. Differences between isolates were found with a sequence identity between the Nebraska and an Ohio isolate being as low as 77.6% for the nucleotide of the nucleocapsids. Two distinct sequences for RNA 3, encoding the nucleocapsid protein, were reported and they both were present in isolates from each state. Some isolates sequenced failed to identify each of the eight genomic RNA segments that should be present in HPWMoV, suggesting that some isolates may not have needed the segments present to be pathogenic, although they may have been absent due to an error in the sequencing process (Stewart, 2016). RYRSaV population structure was studied with 18 different isolates; little divergence was observed with nucleotide sequences being between 98 to 99% identical when analysing selected open reading frames (ORFs) (Di Bello et al., 2016). BLMaV was also shown to have low diversity when compared to other emaraviruses (Hassan & Tzanetakis, 2019). Additional population studies are needed to determine the extent of genetic diversity within each emaravirid. Recently, a study of the population structure of RRV from 16 different states in the USA revealed polymorphisms primarily in the 3' and 5' untranslated region, with some being present in the coding region of the genome (Katsiani et al., 2020). A phylogenetic analysis showed isolates from different geographical locations had few differences suggesting that virus evolution was not driven by location. This suggests that the virus is under positive selection to retain its biological functions, or that regular movement of the virus negates any virus evolution due to location (Katsiani et al., 2020).

Phylogenetic studies indicate common ancestry for emaraviruses (Figure 1.8), and place each species into one of three separate clades (I, II and III). The differing number of segments between emaraviruses and the virus-host relationship of each species indicate that the genus may be monophyletic in origin. EMARaV, RLBV and RRV are all rosaceous-

infecting emaraviruses that occupy different nodes, as seen in Figure 1.8, and not as closely related to each other (Di Bello et al., 2016; Wang et al., 2020; Zheng et al., 2017). PPSMV-1 and PPSMV-2, two distinct viruses that infect the same host, are also in separate clades (Elbeaino et al., 2015). The mite vector phylogeny adds another layer of complexity to the question of a common ancestor with different genera of mites transmitting different species. Eriophyid mites are considered host-specific; however, it is suggested that the mite vectors actually do move and transmit between hosts and this is how speciation of the mites and emaraviruses occurs (Di Bello et al., 2016; Elbeaino et al., 2018). New emaraviruses, or new segments of existing emaraviruses, will need to be discovered to add more information regarding the phylogeny and to narrow down a common ancestor.

1.4 Virus Diagnostic Tests

Virus infections can be diagnosed several different ways including using serological assays, polymerase chain reaction (PCR), reverse transcriptase polymerase chain reaction (RT-PCR), quantitative PCR (qPCR) and electron microscopy (EM). Since viruses cannot be cultivated ad hoc, serological techniques involving using antisera proteins have been used to detect viruses in plants (Martinelli et al., 2015). Enzyme-linked immunosorbent assay (ELISA), the most widely used serological technique, allows rapid detection of viruses and is less labour intensive when processing many samples. ELISA and other serological techniques have been used with emaraviruses with limited success. Mielke-Ehret and Mühlbach (2012) recommended to not use serological techniques with EMARaV against the glycoprotein and nucleocapsid proteins, due to a weak interactions and possible background signals. PPSMV-1 and HPWMoV were reliably detected by nucleocapsid specific antibodies; however, PPSMV-1 detection was weaker when used on some isolates. More recently, Shahmirzaie et al. (2019) tested three techniques: double-antibody sandwich ELISA (DAS-ELISA), dot-immuno binding and western blot against a recombinant nucleocapsid protein of FMV. Each test showed lower sensitivity then the current RT-PCR diagnostic test for FMV.

Reverse-transcription polymerase chain reaction (RT-PCR) can be used as a sensitive, although time-consuming, assay to detect RNA plant viruses (Martinelli et al., 2015). Oligonucleotides, or primers, are designed using sequence data obtained from previous studies through GenBank® Nucleotide Sequence Search found at the NCBI website (NCBI, Bethesda, MD, USA). Sequence data can also be obtained through technologies utilising next-generation sequencing (NGS), such as Illumina sequencing. The primers are designed to amplify a region within the viral genome during the RT-PCR and has been used with good success with emaraviruses. All of the emaravirids reported to date have been detected in host plants using RT-PCR (Di Bello et al., 2016; Mielke-Ehret & Mühlbach, 2012; Zheng et al., 2017). Laney et al. (2011) identified 84 RRV in all symptomatic cultivated and wild *Rosa multiflora* samples using primer pairs developed from initial sequencing data. There were also 30 asymptomatic samples that were tested as negative controls with the same primers from which a product was not amplified, showing the possibility to discriminate virus-infected from virus-uninfected plants. A disadvantage of RT-PCR is that it is a laborious process. RNA needs to be extracted from plant tissue that is free from any inhibitory molecules, such as phenolics or polysaccharides, to get efficient and reproducible results (Martinelli et al., 2015). This is a crucial step in the process that can severely affect results down-stream in the process. The RNA then needs to be used in a RT-PCR and products visualised using gel electrophoresis, both taking up a significant amount of time. This process becomes more unwieldy when there is a need to test large numbers of samples, although is still faster than ELISA, which requires production of an antibody that may not be entirely specific to the target.

1.5 Sequencing Viral Genomes

Obtaining the genome sequence for a virus allows for a better understanding of the function of each translated protein as well as giving more information to accurately detect it with diagnostic techniques. There are various ways to sequence DNA and RNA, although RNA usually must first be made into complementary DNA (cDNA) via a reverse transcriptase reaction. Sanger sequencing, developed in 1977, is a chain-termination method that requires a single-stranded DNA template. While many new sequencing technologies have been developed, Sanger sequencing continues to be used when the number of samples is relatively

small or to confirm sequence generated by newer next generation sequencing (NGS) methods (Baudhuin et al., 2015). With the turn of the century came new technological developments in sequencing in the form of NGS, revolutionising molecular biology (Schuster, 2008). NGS combined new technologies and bioinformatics software with rapid development of computer processing ability to be able to rapidly amplify, sequence and analyse large amounts of DNA. These NGS technologies share three major advantages over Sanger sequencing. Firstly, they rely on a cell-free preparation of NGS libraries instead of bacterial cloning. Secondly, these technologies produce thousands-to-millions of sequencing reactions in parallel compared to the hundreds that Sanger sequencing produces. Thirdly, the sequencing output is performed in parallel with the sequencing reactions instead of using electrophoresis (van Dijk, 2014). Sanger sequencing is considered first generation sequencing and second-generation methods (SGS) include Illumina, Ion Torrent, SOLiD sequencing methods (Barba et al., 2013; Schuster, 2008). These methods generate short sequence reads that must be assembled to make longer contiguous sequences (contigs). Such data contain millions of short sequence reads that can be assembled *de novo*, meaning "from new", where the reads are matched to each other to create a contiguous sequence (a contig). The reads can also be assembled by reference assembly by mapping the reads to a known reference sequence, such as a previously reported genome (Barba et al., 2013). Of the SGS technologies, Illumina is the most widely used due to the high throughput capabilities and relatively low cost (van Dijk, 2014). Illumina, which uses bridge PCR as the amplification method and reversible terminators as the sequencing chemistry, produces small fragments of DNA that are between 100-300 bp long that are sequenced at a rate of 25 Mbp per hour with an accuracy of 99.9% (Barba et al., 2013; Huang et al., 2019).

Third generation sequencing (TGS) represents a new set of technologies that feature single molecule sequencing (SMS) that does not require PCR, resulting in higher throughput, faster turnaround times and longer read lengths, which make it easier to detect haplotypes and single nucleotide polymorphisms (Schadt et al., 2010). The technology suffers from high error rate, however, and are under active development. Two TGS technologies being developed are nanopore sequencing by Oxford Technologies, and single molecule real-time sequencing (SMRT) by Pacific Biosciences (now a part of Illumina). Nanopore sequencing works by using library preparation protocols that attach a motor protein and leader adaptor to one strand of double stranded DNA (dsDNA). The motor protein enzymatically splits the dsDNA molecule

and feeds one strand through a biological nanopore that is built into an electrically resistant artificial membrane with a voltage applied across it. As the strand passes through it causes a disturbance in the current running across the membrane which is specific to each DNA base passing through. These disturbances are electronically recorded and give the sequence to the molecule (Leggett & Clark, 2017). SMRT sequencing also begins with preparing a library from dsDNA and requires a hairpin adaptor to be ligated to each end making the dsDNA circular. A primer and polymerase are then annealed to the circular strand and loaded into the zero-mode waveguide observation chamber. Inside the chamber, the polymerase incorporates fluorescently tagged bases that are recorded by camera in real time and converted to continuous long read data (Ardui et al., 2018). Both technologies are under development and may prove to be beneficial to virus sequencing in the future. Of particular interest is that TGS methods are capable of sequencing RNA directly, which would be useful when sequencing viral RNA genomes to further reduce turnaround times (Kumar et al., 2019; Schadt et al., 2010)

In the field of plant virology, NGS technologies have enabled the discovery of novel viruses at an unprecedented rate and made sequencing of their genomes more accessible. Viruses have been discovered when RNA from symptomatic and asymptomatic plants has been sequenced; is predicted that in the following decades this rate of discovery will rise exponentially (Ho & Tzanetakis, 2014; Kuhn et al., 2019). While NGS has made sequencing a viral genome cheaper and more accessible, it may not provide a complete and accurate sequence. When using total RNA, a low viral titre compared to the background plant RNA can lead to a relatively low number of sequence reads specific to the virus, which may lead to low coverage of desired sequence and may not give a strong consensus across the whole viral genome (Pecman et al., 2017). In particular, a low number of sequence reads at the 5’ and 3’ ends of the viral genome is also common, especially when using *de novo* assembly, that reduces confidence in the sequence of genome termini (Alfson et al., 2014). This results in further work needing to be done to verify the sequences in these regions. There are several methods that can be employed to confirm a viral genome sequence, two of which are mapping small RNA (sRNA) sequences generated by NGS onto the *de novo* assembled genome, or Rapid Amplification of cDNA Ends (RACE).

Some virus infected plants will accumulate a large number of small interfering RNAs (siRNA) generated by RNA interference (RNAi) processes present in the plants, which defend against invasive nucleic acids (Seguin et al., 2014). Sequencing these small RNAs can supplement the longer sequence reads generated using NGS sequencing, and can potentially be assembled reliably into whole viral genomes using *de novo* assembly (Pecman et al., 2017). The small RNAs can also be mapped to a reference genome to provide additional confidence and consensus.

1.6 Karaka (*Corynocarpus laevigatus*)

Corynocarpus laevigatus, called by the common name "Karaka" by Māori, is an endemic coastal tree occurring naturally in coastal and lowland forest of New Zealand's North Island. The Latin name is taken from the features used to describe the plant: *Corynocarpus* meaning 'club seed' and *laevigatus* meaning 'smooth', in reference to the orange fruit it produces, while the common name, Karaka, means 'orange' as an adjective in the Māori language (Costall et al., 2006; Sawyer et al., 2003). Karaka grow to between 15 to 20 metres tall, have thick, elliptic and glossy leaves up to 30 cm long, and produce large, orange fruit drupes that can be up to 40 mm long as (Figure 1.9) (Van Essen & Rapson, 2005). There are five species in the *Corynocarpus* genus in New Zealand, Australia, Vanuatu and New Caledonia, and it is an introduced species in Hawaii (Macaskill et al., 2015). *Corynocarpus laevigatus* is first thought to have been brought to New Zealand by the ancestors of the Moriori, an indigenous people preceding the Māori. It may have originated in the Pacific islands, although the evolutionary history of the tree remains unknown (Atherton et al., 2015; Sawyer et al., 2003).

The fruit, and the kernel contained within, has historical and cultural importance to Māori as an important food source from the time before Europeans arrived in New Zealand. The kernel contains a virulent glucoside toxin, karakin, that can cause spasmodic pains and permanent partial paralysis if ingested. Extensive treatment involving steaming and soaking for days to weeks is needed before the kernel to be safe to eat (Macaskill et al., 2015). The

treated kernel was a valuable food source, sometimes second only to kumara tubers, and eaten during winter when other food was in short supply. For this reason, Karaka trees were planted and cultivated near Māori occupation sites around New Zealand to supplement their diets, which explains how the trees are also found in the South Island of New Zealand (Leach & Stowe, 2005; Sawyer et al., 2003). The kernels were also important culturally and held importance in traditional ceremonies, funerals and formal exchanges between *iwi* (tribes) (Macaskill et al., 2015).

Figure 1.9 - Karaka (*Corynocarpus laevigatus*). A: medium sized mature karaka tree flowering in spring. B: a karaka tree in summer without fruit. C: the glossy green Karaka leaves. D: the ripening karaka fruit in summer. Pictures taken by Lee Rabbidge (2018-2019).

Karaka trees in contemporary New Zealand are widespread and appear to be a mixture of naturally sown and artificially planted trees, being desired as an ornamental tree in public parks and private gardens. The tree is available to purchase retail and is sold readily at nurseries throughout New Zealand. The natural compounds present in karaka present a risk and danger to some urban and agricultural life. The pollen collected by bees is toxic and can cause the death of an entire colony, while the karaka fruit nut is toxic to dogs, as it is to humans, and can cause skeletal muscle lesions and damage to muscle (Brown et al., 2018; Hunt et al., 2018). For this reason, human interference may interfere with the natural and artificial propagation of karaka trees. The tree is an important food supply for birds in New Zealand, however, and endangered bird species must be considered when discussing the tree. The kereru, a native pigeon, eats the fruit and most likely acts as its main, non-human, dispersal agent (Sawyer et al., 2003).

1.6.1 Karaka Gall Mite

The Karaka gall mite (*Aculus corynocarpi*) is another native creature that relies on the karaka tree to survive (Figure 1.10). The eriophyid mite belongs to the lives, feeds and breeds on the underside of the karaka leaf specifically (Martin, 2017). The mite causes damage to the leaves and flower buds of the karaka tree, feeding results in distorted leaves, brown spots on leaves, blackening of flower buds in spring and scarring of stalks.

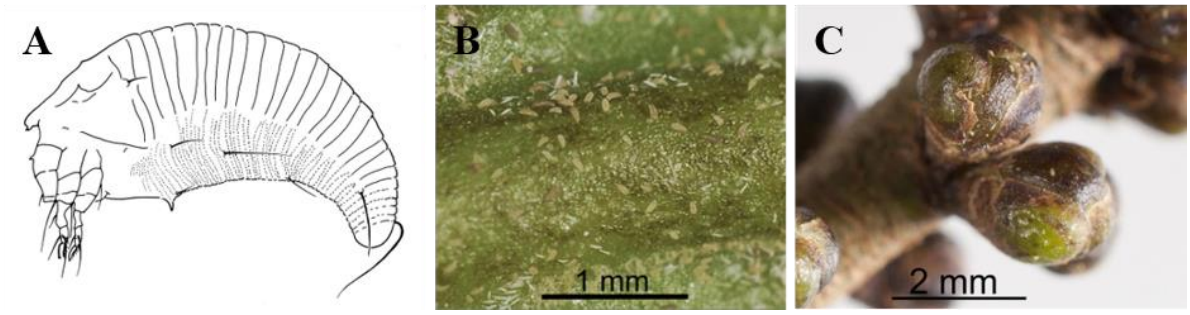


Figure 1.10 – Karaka gall mite (*Aculus corynocarpi*) (Martin, 2017). A: A drawing of the side view of the mite. B: Karaka gall mites on the underside of a karaka leaf. C: karaka gall mites feeding on and damaging a karaka flower bud.

Eriophyid mites are known to transmit plant viruses that can cause plant diseases and death; *Aceria, Phytoptus* and *Phyllocoptes* genera are three genera that has mite species known to be vectors for plant viruses (de Lillo et al., 2018). Further, several species are known to transmit emaraviruses. Most emaraviruses described have an eriophyid mite identified as either a known or suspected vector between plants and generally that mite is specific to the host plant, with HPWMoV being a notable exception (Di Bello et al., 2015; Di Bello et al., 2016; Mielke-Ehret & Mühlbach, 2012; Zheng et al., 2017). Little is known about the mode of virus acquisition and delivery that the eriophyid mites employ (de Lillo et al., 2018).

The effects of the virus infected plants on the biology of each eriophyid mite species is not well understood although there is evidence to believe it is beneficial. For example, *cauliflower mosaic virus* induces elevated emissions of volatile chemicals that attract the aphid vector to the plant to feed and pick up the virus to spread to other plants (Epstein & Hill, 1999; Mauck et al., 2010). Field populations of *Phyllocoptes fructiphilus*, the mite vector of RRV, were found to be up to 17 times denser on RRV symptomatic roses than on symptomless plants (Epstein & Hill, 1999). Negative effects have also been observed with *Aceria tosichella* feeding on Triticum mosaic virus infected plants where the mite reproduction was negatively impacted (de Lillo et al., 2018). The effect may be dependent on the host-mite relationship and the co-evolutionary path of each mite-insect interaction.

Recently, virus-like symptoms were observed on karaka trees where karaka gall mites were present, suggesting the presence of a virus. Analysis of these plants formed the basis of the research described in this thesis.

1.7 Discovery of a potential new emaravirus species

Recently, Dr Robin MacDiarmid from The Institute for Plant and Food Research (PFR) identified virus-like symptoms on several karaka trees at the PFR Mount Albert Research Centre (MARC) in Auckland, New Zealand. The symptoms were distinct chlorotic pale spots on the leaves (Figure 1.11) of varying size and abundance on each leaf.

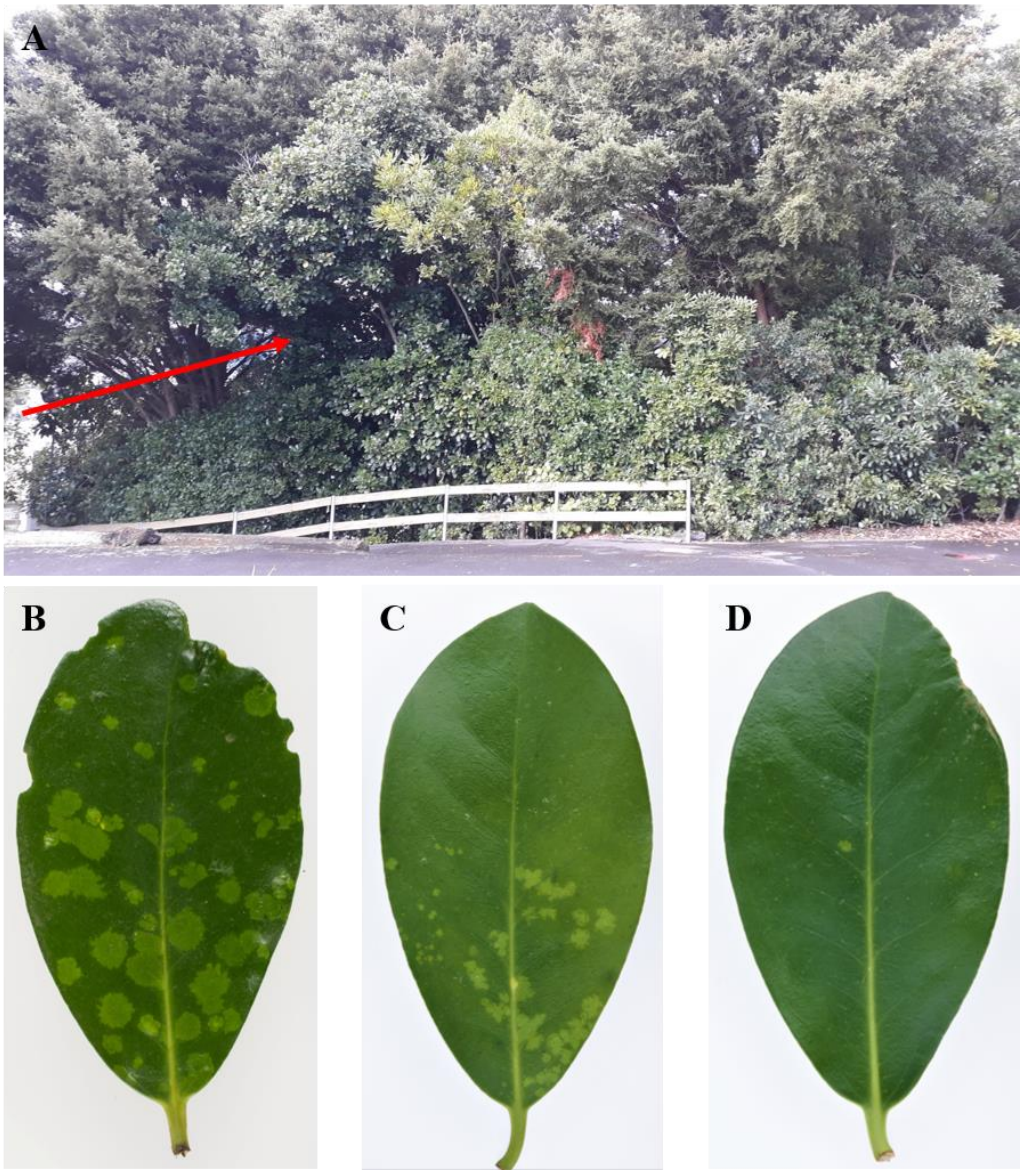


Figure 1.11 – Symptomatic karaka tree and leaves found in Auckland, New Zealand. A: Several karaka trees at PFR MARC that had virus-like symptoms. The red arrow points to the karaka tree in question. B: Karaka leaf with symptoms resembling pale chlorotic circles. C: Karaka leaf showing less severe symptoms. D: Karaka leaf showing one pale circle in the centre of the leaf near the midrib. A and B pictures taken by Dr. Robin MacDiarmid and Dr Arnaud Blouin. C and D pictures taken by Lee Rabbidge.

Symptoms appeared to be unevenly distributed on leaves throughout the symptomatic trees with no symptoms observed on any of the other parts of the tree, including the fruit when the trees were observed in summer. Several smaller trees nearby did not have symptoms even though they were within proximity to the symptomatic trees. Dr Arnaud Bloiun, also from PFR, used Illumina sequencing to determine if a virus was present. He identified an RNA virus with five segments, and two DNA viruses. The two DNA viruses were found to share sequence similarity with badnaviruses while the sequences from the RNA virus were similar to emaravirids (Table 1.3) after using the Basic Local Alignment Search Tool (BLAST) found at the NCBI website (https://www.ncbi.nlm.nih.gov/) (Geer et al., 2009). The largest RNA segment had 34.53% sequence similarity to the EMARaV amino acid sequence encoding the RdRp suggesting that the sequences isolated from the symptomatic karaka tree may belong to a new-to-science virus within the *Emaravirus* genus. Initial testing showed that the badna-like viruses could be detected in asymptomatic and symptomatic regions of the leaves, and the emara-like virus could only be detected in symptomatic regions of the leaves, although the badna-like virus results need to be confirmed. This thesis focuses on the emara-like virus which has tentatively been named, and referred to in this document as, *karaka Okahu purepure virus* (KOPV). The name is in the Māori language (*te reo Māori*) in respect of the treasured status (*taonga*) of the karaka tree in New Zealand. Further discussion on why this name was chosen and why it is tentative will be presented in Chapter Five.

Table 1.3 – Consensus RNA segments found in symptomatic karaka that share homology with *Emaravirus* species.

RNA Segment	Protein Encoded (Based on other emaraviruses)	Nucleotide Length (of initial NGS sequence)	Emaravirus with highest amino acid sequence similarity	Amino acid sequence similarity (%)
1	RNA-dependent, RNA-polymerase (RdRp)	7131	European Mountain Ash ringspot-associated emaravirus (EMARaV)	34.53
2	Glycoprotein precursor	1927	Palo verde broom virus (PVBV)	20.99
3	Nucleocapsid	1771	Ti ringspot-associated emaravirus (TiRSaV)	27.73
4	Movement Protein	1528	Palo verde broom virus (PVBV)	26.04
5	Unknown	1622	No homology with known emaraviruses	N/A

The 3’ and 5’ ends of the new karaka virus appeared to be incomplete. As previously stated, all emaraviruses have complementary sequences between 19 to 23 nt long at the 3’ and 5’ ends that likely form a panhandle structure similar to orthobunyaviruses and hantaviruses (Mielke-Ehret & Mühlbach, 2012). This homology is also conserved between species in the genus and based on comparisons with the other emaraviruses, the new karaka virus appeared to be incomplete. Therefore, the ends of the karaka virus needed verification to complete the genome, followed by an analysis of the completed genome before it could be placed in the genus.

No mention of these symptoms could be been found in literature first describing the tree, or in any transcripts of Māori oral histories. The only virus that has been recorded to infect karaka is cucumber mosaic virus (CMV), which was found in symptomatic karaka leaves in Christchurch (Ashby, 1977; Veerakone et al., 2015). The image in the article by Ashby (1977) was in black and white, and of poor quality, but the symptoms shown appeared to be more like a mosaic and concentric rings. Due to this, it is likely not the same virus that was present in the karaka at the MARC. Further, there was no sequence identity match between the MARC karaka virus and CMV. Likewise, no symptoms could be found when examining digital photos of karaka leaf collections maintained by the Auckland War Memorial Museum.

The virus discovered in karaka trees at the MARC in Auckland, New Zealand is likely a new-to-science virus belonging to the *Emaravirus* genus. To confirm this, the 3’ and 5’ ends of the sequences obtained by NGS needed to be verified. The distribution of the virus within Auckland and New Zealand also needed to be determined to lay the groundwork for future research by showing how widespread the virus is within naturally sown and artificially planted trees. This would be achieved by developing a diagnostic test based on the completed genome to accurately discriminate between virus-infected and virus-free karaka samples obtained by surveying trees throughout Auckland and New Zealand. Finally, the presence of the virus needed to be correlated with symptoms to help determine if the virus is the causal agent of the pale chlorotic circular symptoms.

1.8 Research Aims and Objectives

1.8.1 Aim 1: Completion and analysis of the karaka emaravirus genome

Obtaining the complete sequence for the karaka genome was important for informing the development of a diagnostic assay as well as providing evidence for the virus inclusion in the emaravirus genus. This part of the research is described in Chapter Two. The aim was achieved by the following two objectives.

Objective 1.1: Complete the 3’ and 5’ ends of each RNA of the karaka emaravirus genome

Objective 1.2: Compare the completed RNA sequences with genomes of other emaraviruses

1.8.2 Aim 2: Determine the distribution of the virus

The second research aim was to develop a molecular diagnostic assay to discriminate between virus-infected and virus-free samples. This was then applied to samples gathered from around Auckland and New Zealand to determine whether or not the virus was present. This virus has not been described in New Zealand so far so it is important to have a specific diagnostic assay that can reliably detect the virus and not give any false positives. Determining the distribution of the virus is also important as it will show how far the virus has spread and may give information on when the virus first came to New Zealand. This aspect of the research is present in Chapter Three and was achieved by the following objectives.

Objective 2.1: Develop a PCR-based diagnostic test for the karaka emaravirus

Objective 2.2: Use the diagnostic test developed in objective 2.1 to survey the distribution of the virus within the Auckland region.

1.8.3 Aim 3: Assess the correlation between the symptoms and virus presence

The third research aim is to determine whether or not the virus is correlated with the symptoms present on karaka leaves. This is an important step toward showing that the new virus is the causative agent for the symptoms. This part of the research is presented in Chapter Four and is achieved by the following objective.

Objective 3.1: Use the diagnostic test developed in objective 2.1 to test symptomatic and asymptomatic regions of a karaka leaf.

Chapter Two: Genome Sequence Analysis of a Novel Karaka Emaravirus

2.1 Introduction

The initial Illumina NGS sequencing data obtained from Arnaud Blouin revealed the bulk of the KOPV genome. Sequence reads for the 3' and 5' ends of the genome had low consensus and needed to be confirmed to get the entire sequence of the KOPV genome. RACE is one method that allows determination of the ends of sequences. For a virus such as KOPV, 3' RACE targets both the negative and positive sense RNA segments that may be present in a symptomatic sample, allowing both ends of the viral genome to be determined. A completed genome would be useful for informing a diagnostic test by giving more sequence information, and by relating the novel virus to the current literature surrounding emaraviruses by comparing sequences and finding any conserved motifs.

2.1.1 Rapid Amplification of cDNA Ends (RACE)

RACE is a polymerase chain reaction (PCR) based method that can amplify both 5' and 3' ends of a cDNA strand, generating DNA fragments that can be sequenced using Sanger sequencing. The method was first described in 1990 as a technique to obtain the challenging sequence ends from cDNA copies of mRNA (Frohman, 1990) and has since proven very useful for determining the very ends of viral RNA sequences. This technique requires some prior knowledge of the target RNA sequence to develop gene-specific primers. For eukaryotic mRNA, 3' RACE relies on the presence of the polyA tail at the 3' end which can be targeted by complementary poly-T primers in combination with the gene-specific primers to amplify the ends. These post-transcriptional features are not present in some RNA viral genomes; however, this problem was overcome with a method developed to use RACE with viral genomes by ligation of adapters to each end that could be targeted by specific PCR primers (Hsieh et al., 1997; Li et al., 2005). Many contemporary methods are now available to use RACE on viruses with negative-sense genomes including commercial kits for 3' and 5' RACE. The process for 3' RACE is outlined in Figure 2.1.

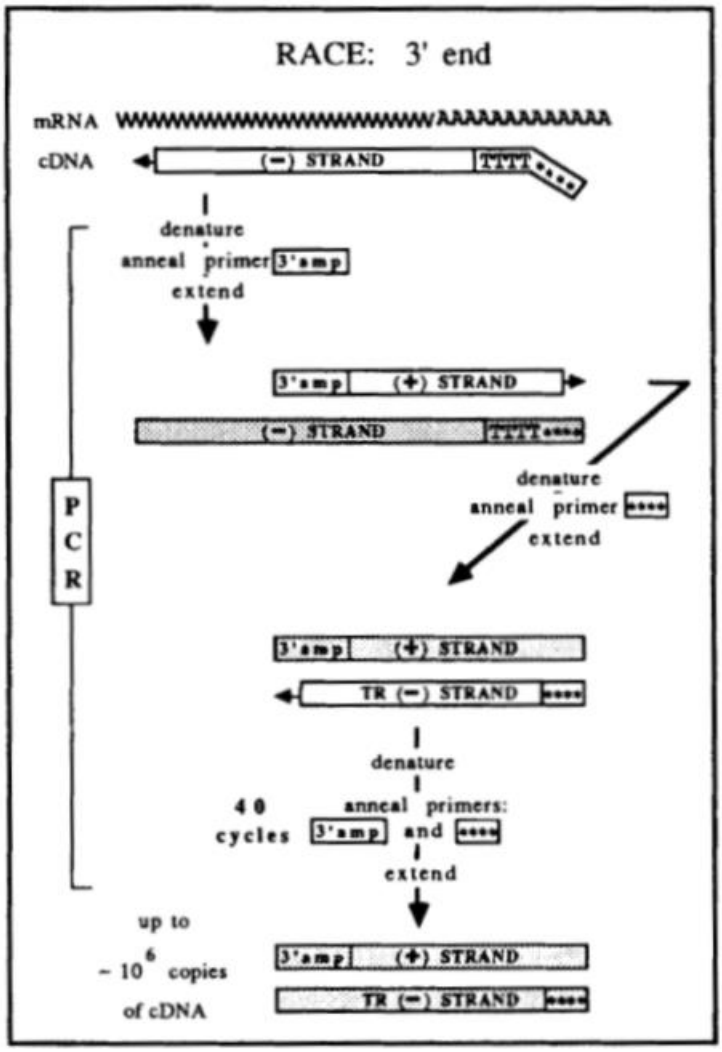


Figure 2.1 – An outline of the 3' RACE process (Frohman, 1990)

Negative-sense RNA viruses, such as emaraviruses, have the defining feature of being the opposite polarity to RNA that can be transcribed and require a cRNA copy of their genome before transcription and translation can begin (York et al., 2013). This means that there will be both negative and positive polarity genome sequences in a virus infected plant as the virus uses the plant cell to replicate. Taking advantage of this, a RACE method can be designed to target the 3' end of both the original negative sense and the cRNA positive-sense strands. Total RNA extracted from a virus infected plant sample can be polyadenylated *in vitro*; during this process both the negative and positive sense viral genomes will be polyadenylated at their 3' ends. This means that both ends of the viral genome can be targeted for 3' RACE from one

polyadenylation reaction using separate gene specific primers designed to amplify in the direction of the RNAs' 3' ends. This method relies on the antisense of a virus, the positive-sense strand in this case, being present in sufficient amounts compared to the genome in a sample (Li et al., 2005). Testing should be carried out to ensure that the positive-sense strand is present in high enough quantities.

Previous scientific communications describing new emaravirids used RACE to verify the 5' and 3' ends of the viral genome where NGS was used to obtain the initial sequence. Examples include AcCRaV, BLMaV and RYRSaV (Di Bello et al., 2016; Hassan et al., 2017; Zheng et al., 2017). Two candidate species, PiVB and *palo verde broom virus*, have reported genomes that were not verified using RACE, only the initial NGS *de novo* assemblies (Buzkan et al., 2019; Ilyas et al., 2018). It is possible that these latter genomes may not be complete.

2.1.2 Aims

This Chapter describes the work to achieve the aim of completion and analysis of the karaka emaravirus genome. The first objective was to complete the KOPV genome that was obtained previously through Illumina NGS data. The number of reads that had been obtained for each end of each of the five segments of the viral genome was low; therefore, the confidence in this sequence was low. Further, initial analysis showed that the segments did not match the conserved emaravirus ends that had been previously reported. A RACE strategy was implemented where 3' RACE was performed on the negative and positive genomes for all five genomic segments hypothesised to be present in a symptomatic sample. Once sequencing data was obtained from these ends, they would be combined with the initial NGS sequencing data to create a complete genome.

The second objective was to analyse the completed genome and compare it to the previously reported emaravirids with the goal of confirming that KOPV is a member of emaravirus genus.

2.2 Methods and Materials

2.2.1 Plant Material

Plant material was taken from the plant PFR0001 at Plant and Food Research, Mount Albert Research Campus, Auckland, New Zealand, the same tree where the RNA was taken to do the initial NGS sequence analysis by Dr Arnaud Blouin.

2.2.2 RNA Extraction

Prior to RNA extraction, leaves were washed with 70% ethanol to remove any contamination. Total RNA was extracted from fresh leaf samples, or leaf samples stored at 4°C, using an adapted version of the modified cetyltrimethylammonium bromide (CTAB) method (White et al., 2008). The extraction buffer consisted of 2% cetyltrimethylammonium bromide, 2% soluble polyvinylpyrrolidone (PVP) K-40, 25 mM ethylenediaminetetraacetic acid (EDTA), 100 mM Tris-HCL (pH 8.0), 2 M NaCl, 0.5 g/: spermidine and 3% β-mercaptopethanol. The buffer, excluding the 3% β-mercaptopethanol, was combined and autoclaved. The 3% β-mercaptopethanol was added to the pre-heated buffer at 65°C before the extraction process began. A portion of the leaf sample was cut out of the leaf using sterilised razor blades and weighed to obtain 250 mg of leaf material (+/- 10%). The leaf material was put into a plastic pouch and ground using a drill press with a modified grinding tip. For each sample, 1.2 mL of pre-heated extraction buffer (at 65°C) with 3% β-mercaptopethanol was pipetted into the pouch and mixed. The mixture was then squeezed out of the pouch and into a 2 mL plastic microtube for each sample then, once all of the samples in the batch had been mixed, the tubes were vortexed for 15 seconds and placed in a 65°C heating block. Each tube was kept in the heating block for 30 minutes and vortexed for 15 seconds at five-minute intervals. The tubes were then centrifuged at maximum speed (16,000 x g) for 10 minutes and the supernatant transferred to a new 2 mL plastic microtube. An equal volume of chloroform:isoamyl Alcohol (C:I, 24:1)) was added to each tube, vortexed for 15 seconds then

centrifuged at maximum speed (16,000 x g) at 4°C for 15 minutes. The aqueous phase was transferred to a new 2 mL plastic microtube and another equal volume of C:I was added to this. The tubes were then vortexed for 15 seconds and centrifuged at maximum speed (16,000 x g) at 4°C for 15 minutes. The aqueous phase was again transferred to a fresh 2 mL plastic microtube and a one third volume of 8 M lithium chloride (LiCl) was added, mixed by pipetting and stored at 4°C overnight to let the RNA precipitate.

The following day, each tube was centrifuged at maximum speed (16,000 x g) for 60 minutes at 4°C to pellet the precipitated RNA. The supernatant was removed after this and the pellet was washed in 800 µL of 70% ethanol then centrifuged at maximum speed (16,000 x g) for two minutes. The ethanol was removed carefully with a pipette and the tubes were left open with a paper towel loosely covering them to air dry for 10 minutes. The dried pellet was re-suspended in 50 µL of sterile ddH_2O and immediately placed on ice.

The concentration of each sample was then determined using a Nanodrop 2000 (ThermoFisher Scientific, USA). Ideal RNA concentrations were generally between 150 and 400 ng/µL with a 260/280 absorbance of between 1.9 and 2.0. Some samples were outside this range but were tested using RT-PCR any way. If those samples gave an unexpected PCR result, they were re-extracted and re-tested using the redo method specified below. This method was scaled down to use as little as 150 mg of starting leaf material that produced lower concentrations of RNA while still having amplifiable RNA when doing a reverse-transcription polymerase chain reaction procedure. RNA was stored at -80°C until used three days later.

2.2.3 Initial NGS data

The initial Illumina NGS deep sequencing data received from Dr Arnaud Blouin had already been assembled *de novo* into five genomic segments constituting the KOPV genome. Initial analysis of these segments revealed that each contained a large ORF in the negative sense orientation. RNA 1 had an ORF that encoded a protein similar to the RdRps of other

emaraviruses. RNA 2 and 3 ORFs also showed similarity to ORFs from published emaraviruses. RNA 4 was of similar size to the emaravirus movement genomic segments and contained an ORF of similar size to the ORFs present in other emaravirus segments. RNA 5 had an ORF that encoded a protein with no similarity to any known viral protein. sRNA sequence data from the original analysis was also obtained from Dr Blouin and consisted of 26,546,769 sequences between 18 and 44 nt in length.

2.2.4 3’ RACE

3’ RACE was used to determine the very ends of each of the five identified KOPV genome segments. Figure 2.2 shows a flow chart overview of the experimental process for both negative and positive sense segment for each KOPV RNA. Total RNA from a symptomatic karaka leaf was used, expected to contain each viral RNA segment, in both positive and negative polarity. Following polyadenylation, gene specific primers for each positive and negative genome segment were designed to pair with a primer to target the PolyA tail to amplify the 3’ regions of each genome segment.

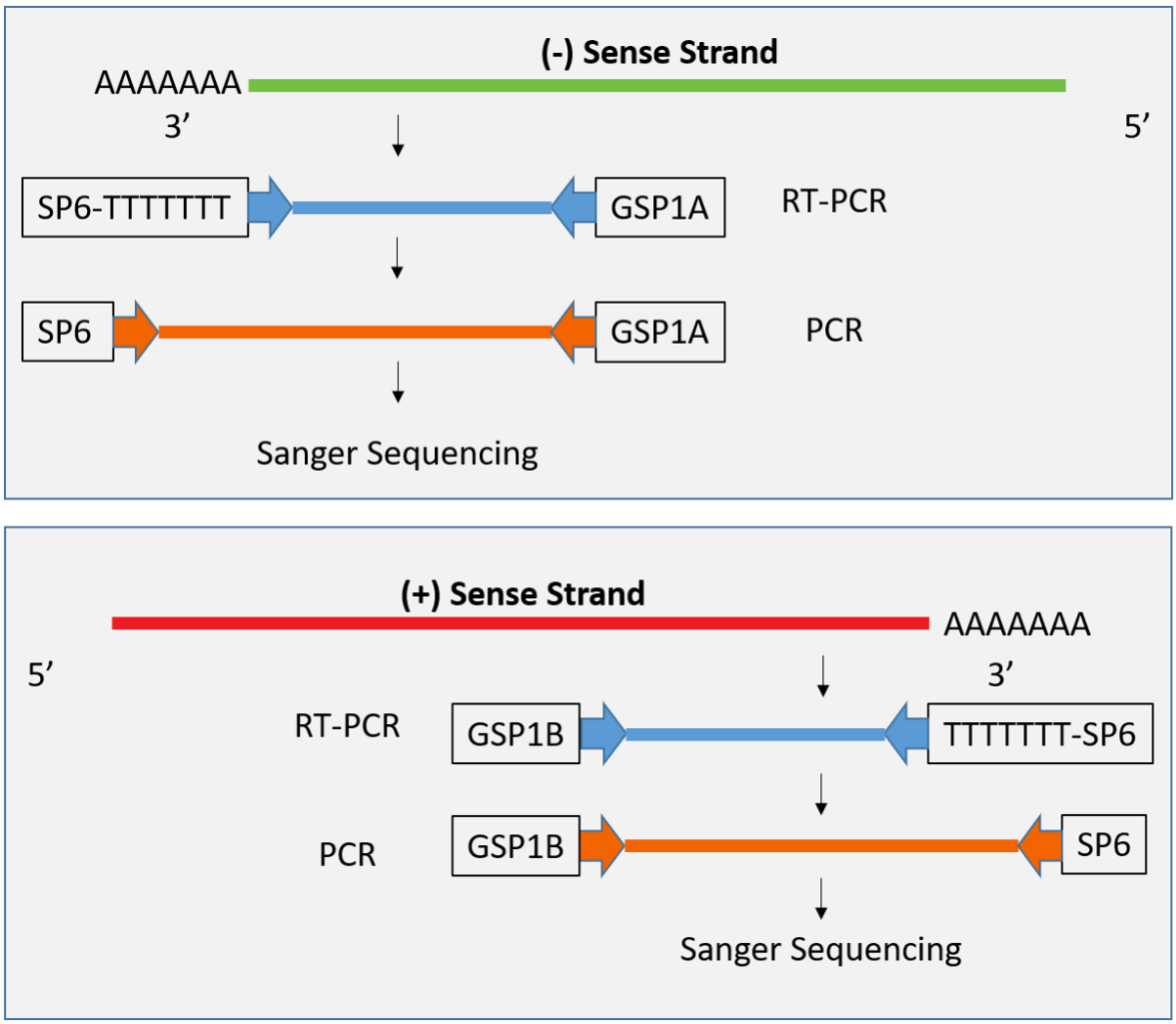


Figure 2.2 – Overview of the strategy using 3’ RACE using KOPV RNA 1 as an example. A polyA tail is added to the 3’ ends of both the negative (-) and positive (+) sense strands of each genomic RNA. For the negative sense strand, GSP1A and SP6-VdT primers are used in a one-step RT-PCR reaction to amplify the 3’ end, including the polyA tail. The GSP1A primer is then used with a SP6 primer in a PCR reaction to amplify the sequence to get enough DNA to be sequenced using Sanger sequencing. The same process is repeated with the positive sense strand.

2.2.4.1 3’ RACE Primer Design

3’ RACE primer pairs were designed that could amplify the 3’ ends of both positive and negative sense viral genomic RNA. All primers used in this chapter are listed below in Table 2.1. These primer pairs needed to bind to a strand specific region (the gene specific primer), and to the PolyA tail that was to be added to each 3’ end. Gene specific primers were designed using the initial Illumina data to target a sequence 240 – 830 bp from the 3’ end. The

priming direction of the primer sequence was toward the 3' end of the negative sense (all primers ending with "A") or positive sense (all primers ending with "B") strand targeted. Primers targeting the PolyA tail to be added to each 3' end were adapted from Mackenzie et al. (1998). PV1/SP6 was adapted to bind to the PolyA tail and attach the SP6 sequence to the PCR product in the first round of RT-PCR (Figure 2.2). The SP6 sequence was then targeted by the SP6 primer in the second PCR to amplify enough product for sequencing. All primers were checked for self-dimers, secondary structures and optimal melting temperatures using OligoAnalyzer by Integrated DNA Technologies™ (https://sg.idtdna.com/calc/analyzer) and were designed to have a melting temperature between 50°C and 52°C. The primers were then checked for potential homology with other untargeted sequences that may be present using the BLAST tool on the NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Table 2.1 – Primer names and sequences for primers used for 3'RACE. The expected product sizes are also given. The targeted RNA segment is indicated in the primers name. Primers ending with A targeted the negative sense strand and primers ending with B targeted the positive sense strand. Primers targeting RNA3 had to be redesigned, as indicated.

Primer Name	Nucleotide sequence 5' → 3'	Expected product size (bp)
Gene Specific Primers		
3'RACE RNA1A	TAC CAT GTA TGA GAT TCT GAA CC	~315
3'RACE RNA1B	CTC ATC TTC AAA GAT ATC AAT GGG	~341
3'RACE RNA2A	CAC CCA TTA TGC TGT TAG ATG	~300
3'RACE RNA2B	AGT TGA CTC AGA GGT ATT CC	~370
3'RACE RNA3A	CAT ACT GGT GAG TTC CTC C	~320
3'RACE RNA3B	CAT CAA TTC ACG AGT TTG GG	~830
3'RACE RNA4A	CCA TAT AAG TCA CTA CAT CTC AC	~290
3'RACE RNA4B	TCA TTG TAT TTT ATT ATG CTT GGT TG	~430
3'RACE RNA5A	TCC TTC TTG AGC TTA GTC AT	~300
3'RACE RNA5B	GAA GAA GAG ACA ATA GAG TAG AGA	~240
Gene Specific Primers – RNA3 redesign		
3'RACE RNA3A-2	CAT ACT GGT GAG TTC CTC C	~380
3'RACE RNA3B-2	CAT CAA TTC ACG AGT TTG GG	~300
PolyA Tail Primers		
PV1/SP6	GAT TTA GGT GAC ACT ATA GTTTTTTTTTTTTTTTTV	
SP6	CAT ACG ATT TAG GTG ACA CTA TAG	

2.2.4.2 Polyadenylation (PolyA Tailing)

Total RNA was extracted from a symptomatic region on a karaka leaf and the concentration of RNA was determined as described in section 2.2.2. RNA was combined with sterile ddH_2O t ogive a total of 4.888 µg of RNA in 15 µL of ddH_2O. Duplicate reactions were set up with 2 µL of 10X *E. coli* Poly(A) Polymerase Reaction Buffer, 2 µL of ATP (10 mM), and 1 µL of *E. coli* Poly(A) Polymerase (New England Biolabs) to make up a 20 µL reaction solution that was incubated for 30 minutes at 37°C. Aliquots of 4 µL were taken from the reaction solution at 10, 15, 20, 25, and 30 minutes and transferred to a 1.5 mL tube on ice to stop the reaction. A negative control reaction included water in place of RNA. One of the duplicate reactions taken at each time point was purified using the Quick PCR Clean Up Kit (Life Technologies). Concentrations of the PolyA RNA was obtained by using a Nanodrop 2000 (ThermoFisher Scientific, USA) and used in the following steps.

2.2.4.3 RT-PCR

One-step RT-PCR was used with each gene specific primer and the PV1/SP6 primer on PolyA tailed RNA for a total of 10 reactions. One-step RT-PCR experiments were done using the InVitrogen Superscript™ RT-PCR with Platinum™ *Taq* DNA polymerase kit (Thermofisher, Massachusetts, USA). One-step reactions were amplified using either an Eppendorf Mastercycler® Gradient, an Eppendorf Mastercycler, or a BioRad T100 thermocycler. A PCR product was amplified for each strand, in each polarity, by adding 1 µL of PolyA RNA standardised to 75 ng/µL in sterile ddH_2O into a reaction with 0.4 µL of 10 µM gene specific primer (Table 2.1) and PV1/SP6 primer. This was mixed with 5 µL of 2X Reaction mix, 0.2 µL of Superscript™ III RT/Platinum *Taq* Mix and 3 µL of ddH_2O. The RT-PCR conditions were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 2 minutes at 98°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 98°C, annealing for 30 seconds at 55°C, extension for 50 seconds at 72°C, followed by a final extension for 7 minutes at 72°C. The amplified product was held at 15°C after the RT-PCR had completed and then run immediately on a gel or stored

at -20°C until needed. Larger volume reactions used the same proportions of reagents as described here.

2.2.4.4 PCR

A PCR was done with each gene specific primer and the SP6 primer on the product obtained from the previous one-step RT-PCR to gain enough product for sequencing. The PCR used 1 µL of the RT-PCR product synthesised as described in Section 2.2.4.3 added to 3 µL of ddH_2O, 5 µL of CloneAmp HiFi PCR Premix (Takara Bio USA, Inc.), 0.5 µL of 10 mM gene specific primer (the same primer that was used to create the RT-PCR product) and 0.5 µL of 10 mM SP6 primer. The PCR conditions were 1 cycle of initial denaturation for 30 seconds at 98°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 98°C, annealing for 30 seconds at 52°C, extension for 50 seconds at 72°C, followed by a final extension for 7 minutes at 72°C. The amplified product was held at 15°C after the PCR had completed and then run immediately on a gel or stored at -20°C until needed.

2.2.4.5 Agarose Gel Electrophoresis

To visualise results, all PCR and RT-PCR products were electrophoresed through a 2% agarose/1x Tris/Borate/EDTA (TBE) buffer gel containing Red Safe nucleic acid staining solution at a concentration of 10 µg/mL. The gel was run for 30 minutes at 80 V using a Mini-sub® Cell GT Gel electrophoresis chamber (Bio-Rad, Auckland, New Zealand). A 1 kb plus DNA marker was used with 25 ng at 5 µL volume (Thermofisher Scientific, USA). The gels were viewed and photographed under UV light using a BioRad Gel Documentation system.

2.2.4.6 Sequencing

PCR products visualised using agarose gel electrophoresis were purified and sequenced to obtain the 3' ends of each KOPV RNA segment. A30 µL PCR was electrophoresed as described in Section 2.2.4.5 for each gene specific and SP6 primer pair combination. Bands of the correct size were excised using a sterile scalpel blade and purified using the Sigma GenElute Gel Extraction Kit (Merck KgaA, Germany) as per the manufacturer's instructions. The concentration of the purified product was checked using a Nanodrop 2000 (Thermofisher Scientific, USA). The purified products were sequenced by Macrogen (South Korea).

2.2.5 Assembling the Complete Genome

Sequencing data from Macrogen was analysed using Geneious 6.1.8 software (https://www.geneious.com). Sequences were trimmed of low-quality reads, and the forward (gene specific) and reverse (SP6) sequences aligned to create a single consensus sequence for each end of each viral RNA segments. This sequence was then aligned pairwise to its respective RNA segment from the initial NGS sequence data using the Geneious Alignment tool with default parameters (a cost matrix of 65% similarity, a gap open penalty of 12, a gap extension penalty of 3 and alignment type of Global alignment with free gaps). The resulting alignment gave a sequence that had a complete 3' end with a PolyA sequence attached. A sequence was extracted from this alignment that included all nucleotides up to the start of the PolyA tail for both the negative and positive sense primer sets, and these were aligned with the original NGS data using the Geneious Alignment tool as above. The completed genomic sequence for each segment was extracted from the respective RNA segment alignments and this represented a tentatively completed genome.

To confirm the tentatively completed genome was correct, sRNA from the initial NGS data was mapped to the completed genome sequence for each segment by reference assembly

using Geneious v6.1.8. The sRNA data contained 26,546,769 sequences between 18 to 44 nt. The default parameters for the assembly were “Medium-Low Sensitivity / Fast”, “Fine tuning of Iterate up to 5 times”, and “Trim sequences” at the default values. The consensus sequence was extracted from the assembly and compared to the tentatively completed genome obtained from the RACE alignment above for each segment. Any discrepancies between the two sequences were analysed and the genome sequence amended based on the number of reads in the original NGS sequence and the number reads in the sRNA assembly giving the final genome. A simplified view of this workflow is shown in Figure 2.3.

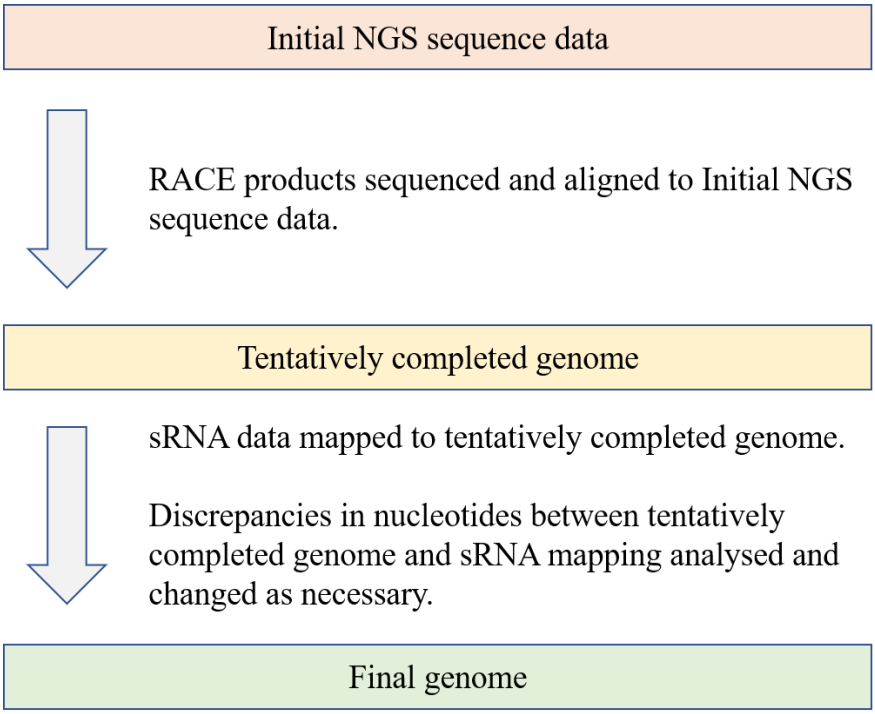


Figure 2.3 – The workflow used to determine the final genome using the initial NGS data.

2.2.6 Sequence Analysis of Complete Genome

Geneious v6.1.8 was used to calculate the total nucleotide, 3’ UTR and 5’ UTR lengths while the Open Reading Frame (ORF) Finder tool was used to find ORFs that could be translated into viral proteins. Completed sequences for each segmented were submitted to a BLASTn search to find similarity to other nucleotide sequences in the National Center for

Biotechnology Information (NCBI) database (www.ncbi.nlm.nih.gov). The Open Reading Frame (ORF) Finder tool in Geneious v6.1.8 was used to find the longest ORFs that could be translated into viral proteins; identified ORFS were extracted and translated into amino acid sequences using the Geneious Translate tool (https://www.bioinformatics.org/sms/prot_mw.html). Translated sequences were named P1, P2, P3, P4 and P5 based on their respective RNA segment number. The translated sequences were submitted to a BLASTp search to find similarity to other amino acid sequences in the NCBI non-redundant protein database. A conserved domain search was also performed on each translated amino acid sequence using the Conserved Domain Database search on NCBI (Marchler-Bauer et al., 2015). Previously reported motifs for each RNA nucleotide sequence and translated protein amino acid sequence were searched for in Geneious v6.1.8.

The predicted amino acid sequences translated from the longest ORF of each RNA were analysed with the following tools. The Protein Molecular Weight tool (https://www.bioinformatics.org/sms/prot_mw.html) was used to estimate the molecular weight of each translated sequence. N-Glycosylation sites were found using NetNGlyc 1.0 and O-Glycosylation sites were found using NetOGlyc 4.0 (Gupta et al., 2004; Steentoft et al., 2013). Signal peptides were predicted using SignalP 5.0 (Almagro Armenteros et al., 2019). Tmpred was used to predict any transmembrane regions present (Hofmann, 1993) while TargetP 2.0 was used to predict the presence of N-terminal pre-sequences that may suggest involvement in a secretory pathway (Emanuelsson et al., 2007). PROMALS was used to identify the 30K superfamily in KOPV by aligning it with the movement protein of EMARaV, FMV, RRV, RYRSaV, BLMaV, AcCRaV and the palo verde broom virus (Pei & Grishin, 2007). COILS was used to predict the presence of any coiled-coils present (Lupas et al., 1991).

2.2.6.1 Comparison with other emaraviruses

Genome data for all known emaravirids was obtained from NCBI and accessions used in each analysis were noted. Only the first reported isolate, where possible, for each virus was used. Geneious v6.1.8 was used to calculate the total nucleotide length, 3’ and 5’ UTRs, and

ORF length for each species. Each ORF for each species was translated into an amino acid sequence and pairwise identity percentages were calculated by aligning the KOPV amino acid sequence against each emaravirus orthologue. Multiple sequence alignments using MUSCLE in the MEGA7 software platform (https://www.megasoftware.net/) were also carried out each for P1-P5 to compare all emaraviruses protein sequences. These data are reported in Appendix C. Note that any emaraviruses reported in 2020 were not used in any analysis as they would have appeared after the completion of the study.

2.2.6.2 Phylogenetic analysis

A phylogenetic tree for RNA 1 was constructed by aligning the RNA1 nucleotide sequence from KOPV with the RNA1 sequences of all available emaraviruses in MEGA7 using ClustalW. The tree was created using the default neighbour-joining method parameters (1000 bootstrap replications, Poisson model, uniform rates, and the same pattern among lineages (homogeneous)).

Phylogenetic trees for the amino acid sequences of P1, P2, P3 and P4 were constructed by aligning all emaravirus protein specific sequences with the KOPV sequence. Amino acid sequences for each protein were imported into MEGA7 and aligned using ClustalW with the default parameters (Gonnet protein weight matrix, a gap opening penalty of 10, a gap extension penalty of 0.2, gap separation of 4 and a delay divergent cut off of 30%). The alignment was then used to create a phylogenetic tree using the default neighbour-joining parameters. The relevant proteins from lettuce necrotic yellows virus (LNYV), a negative sense cytorhabdovirus, were used as the outgroup for each tree. Relevant proteins from other negative sense viruses, namely Bunyamwera virus (BUNV), an animal infecting orthobunyavirus, and tomato spotted wilt virus (TSWV), a tospovirus, were used to compare the emaraviruses to these closely related genera. NCBI accession numbers of sequences used are included in the names of all sequences in the phylogenetic trees.

2.3 Results

2.3.1 3’ RACE and Genome Completion

2.3.1.1 Polyadenylation (PolyA Tailing)

The 5’ and 3’ ends of the genome obtained from the NGS data needed to be confirmed by RACE. A PolyA tail was added to the 3’ end of both negative and positive sense strands of the KOPV present in total RNA. The PolyA sequence would be targeted by a specific primer combined with a gene specific primer to amplify the 3’ region of each strand (Figure 2.2).

Table 2.2 shows the results of the PolyA tail reaction. At the end of polyadenylation, samples R001 and R002 had a higher concentration of RNA than what was originally put into the reaction. Curiously, the negative control, sample R003, appeared to have RNA present following polyadenylation. For these samples, the A260/A280 ratios were higher than the ideal 2.0, indicating possible interference from *E. coli* polymerase and other artefacts present in the reaction. Sample R002 was purified using the Quick PCR Clean Up Kit to create sample R200P the RNA concentration of this samples was lowered substantially along with the A260/A280 absorbance, indicating most of these contaminants were removed by this process. Note that the Quick PCR Clean Up Kit is used for DNA products obtained from PCR reactions, not RNA, and there may have been loss of RNA due to degradation during the process.

Since the polyadenylated RNA (R001 and R002) appeared to have contaminants that might interfere with PCR, and the purified polyadenylated RNA (R200P) was low in concentration, it was uncertain if either of these would result in amplified products using one-step RT-PCR. Therefore, both types of samples were tested.

Table 2.2 – Concentration and A260/A280 absorbance values for PolyA tailed RNA. R002 was subsequently purified using the Quick PCR Clean Up Kit (Life Technologies) and labelled R002P.

Sample	Starting RNA in reaction (ng)	Concentration of RNA pre PolyA Tail reaction (ng/µL)	Concentration of RNA post PolyA Tail reaction (ng/µL)	A260/280 Absorbance post PolyA Tail reaction
R001	4888	244	424.252	3.29
R002	4888	244	470.680	3.23
R002P	4888	244	38.035	1.63
R003	0	0	347.148	4.61

2.3.1.2 One-step RT-PCR – Initial cDNA ends amplification

Using the polyadenylated RNA templates described in section 2.3.1.1, one-step RT-PCR was used to amplify a product from the 3' end of each negative and positive sense RNA strand. Using samples R001 (crude), R002P (purified) and R003 (No RNA), amplification of products from the negative and positive sense strands of RNA1 were initially tested. Amplification from the genome negative sense strand was carried out using the primers (3' RACE RNA1A and PV1/SP6 (Table 2.1) while amplification from the positive sense strand was carried out using the 3'RACE RNA1B and PV1/SP6 primers.

Figure 2.4 shows that a product of the expected size was amplified for both primer pairs using both the crude R001 sample (lane 1) and the purified sample R002P (lane 2). A product of approximately 315 bp was amplified from the negative sense RNA1 strand using the RNA1A/PV1SP6 primer pair, while a product of approximately 341 bp was amplified from the positive sense RNA1 strand using the RNA1B/PV1SP6 primer pair. No product was amplified for R003, which was expected as there was no RNA added before the PolyA tailing reaction. This showed that polyadenylation of the viral RNA had occurred and could be used to amplify from both the negative and positive sense viral RNA genomes. While it also showed that purification of the template following polyadenylation is unnecessary, both types of template were used to amplify from the remaining RNA strands.

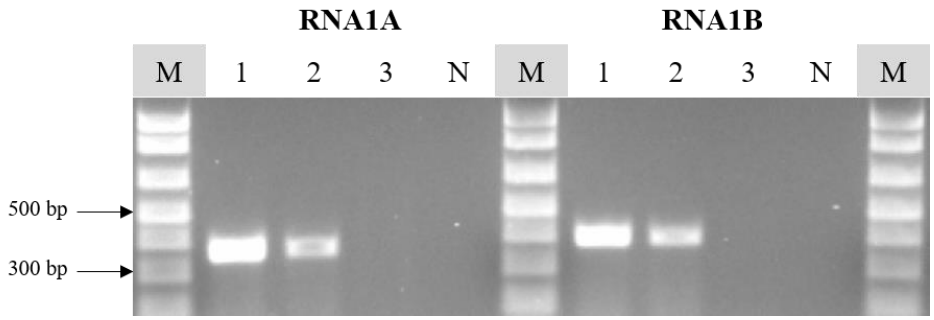


Figure 2.4 – Analysis of a one-step RT-PCR using RNA1A and PV1/SP6, and RNA1B and PV1/SP6 primer pairs. Lane 1 is sample R001 – crude polyA tailed RNA. Lane 2 is sample R002P – purified polyA tailed RNA. Lane 3 is R003 – no RNA in the polyA tailed reaction. Lane M: 1 kb Plus DNA ladder. Lane N: No template control.

Amplification for each positive and negative sense RNA segment gave rise to the expected products except for the positive sense RNA3 strand (RNA3B, lane 1) and negative sense RNA5 strand (RNA5A, lane 1). For sample R002P, products were amplified from both negative and positive strands for all RNAs 2-5 (lane 2), although the RNA3B, RNA5A and RNA5B primers resulted in weak bands. Where products were amplified, the product was the expected size (Table 2.1), except for RNA3B where the band was around 200 bp lower than the expected ~830 bp. These amplified products from both R001 and R002P were used as templates for the next round of PCR.

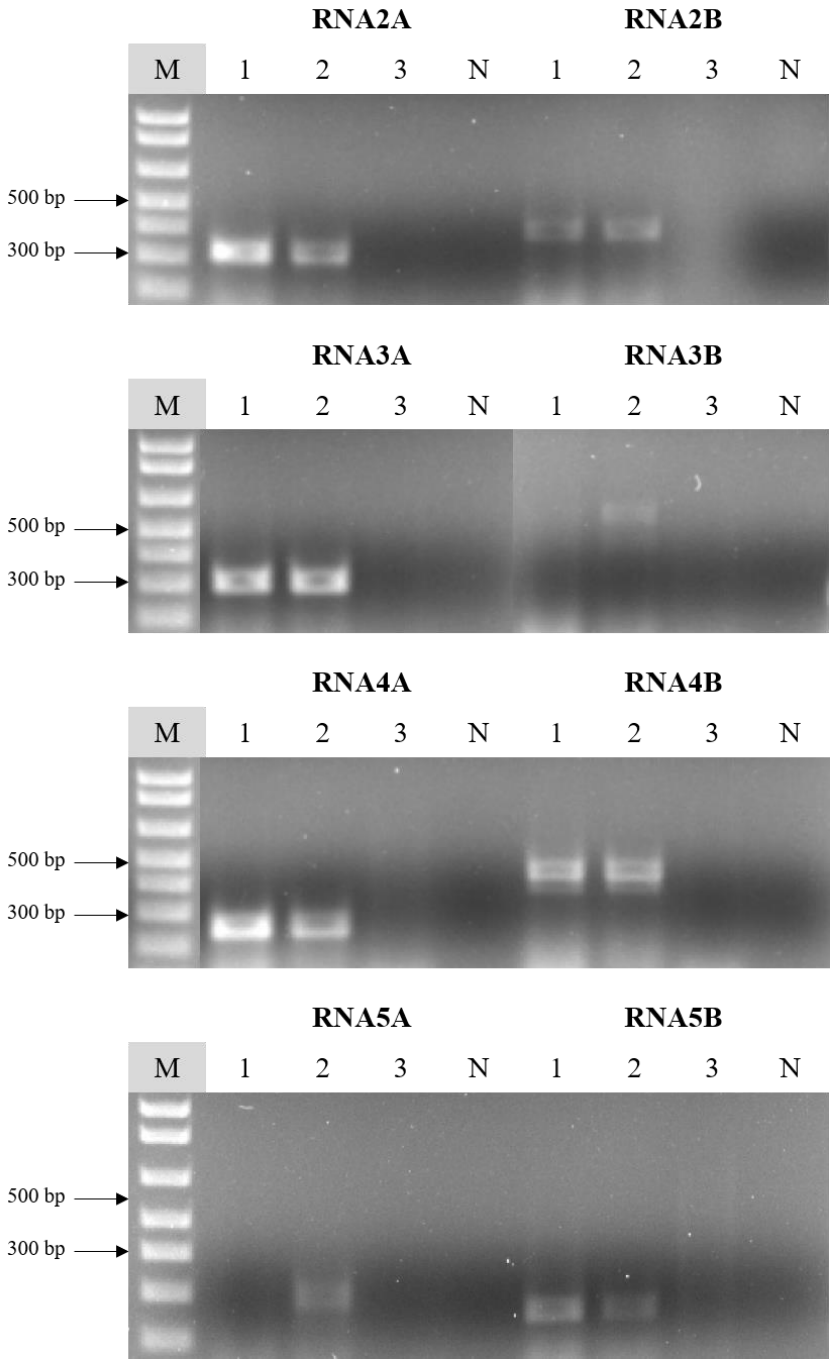


Figure 2.5 – Analysis of one-step RT-PCR using gene specific primers for RNA2, RNA3, RNA4 and RNA5 each paired with the PV1/SP6 primer. Lane 1 is sample R001 – crude polyA tailed RNA from. Lane 2 is sample R002P – purified polyA tailed RNA. Lane 3 is sample R003 – no RNA in the polyA tailed reaction. Lane M: 1 kb Plus DNA ladder. Lane NTC: No template control.

2.3.1.3 PCR amplification for sequencing

Following the polyadenylation and RT-PCR steps described in sections 2.2.4.2 and 2.2.4.3, a PCR was carried out to increase the amount of specific product for sequencing. The RT-PCR products amplified from samples R001 and R002P for RNAs 1-5 were used as templates in a PCR with SP6 primer and the gene specific primer for each end of each viral RNA. The exception was RNA1, where only the R002P template was used. These PCRs were expected to amplify a product that contained the 3' end of each negative and positive strand for each RNA along with the added PolyA tail.

Amplification from the R002P RT-PCR product gave rise to the expected products (Table 2.1) for all RNAs, and for both positive and negative genome strands (Figure 2.4, Lanes 2A and Lanes 2B). The reactions that contained the R001 RT-PCR product as template gave rise to the expected products for positive and negative strands of RNAs 2 and 4 (Lanes 1A and 2A for each RNA), the negative strand for RNA3 (lane 1A), and the positive strand for RNA5 (Lane 1B). Previous RT-PCR using the R001 template did not give rise to a product for the positive stand of RNA3 and the negative stand for RNA 5 (Figure 2.5), although the templates were included here to ensure there truly was no product, rather than a very low level of amplification during the RT-PCR. Figure 2.4, lanes 1B for RNA3, and 1A for RNA5 show no product using the SP6 and gene specific primers, indicating no amplification in the previous RT-PCR step. All bands amplified from the R002P template (Lanes 1B and 2B) were excised, purified, and sequenced by Sanger sequencing.

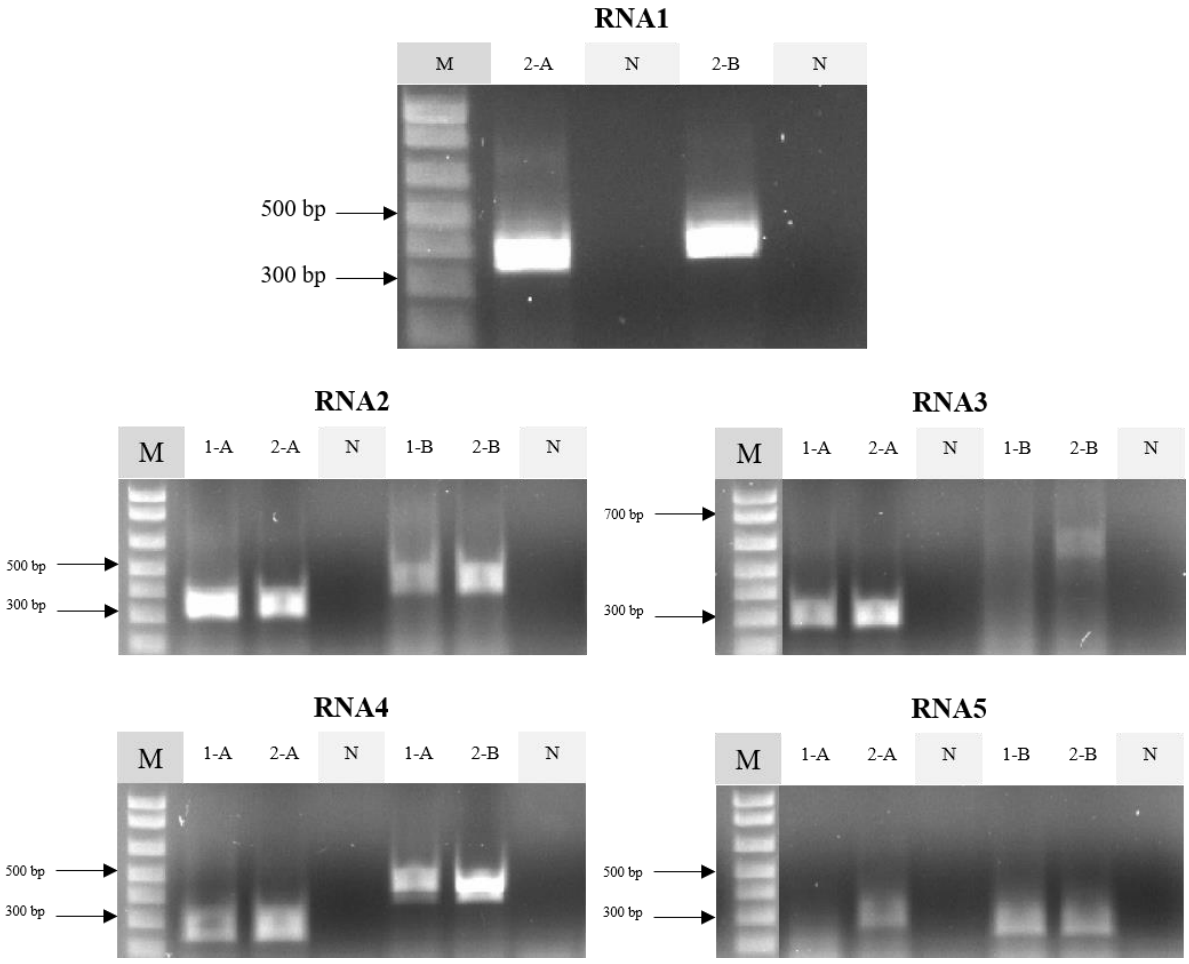


Figure 2.6 –PCR products generated by PCR with SP6 and gene specific primers. The RNA strand used as template is indicated above each photograph. Lanes 1A and 1B contained the R001 sample as template while the 2A and 2B lanes had R002P as template. Lanes 1A and 2A amplified from the negative strand, while 1B and 2B for the positive strand. Lane N: No template control. Gene specific primers are listed in Table 2.1.

The sequence returned for all PCR products was of high quality (discussed in section 2.3.1.4), except for the RNA3A and RNA3B products. Sequence data is discussed in section 2.3.1.4. The gene specific primers for RNA3 were redesigned and used in an RT-PCR with the polyadenylated R001, R200P and R003 samples. The expected product of 380 bp was observed in a 2% agarose gel (Figure 2.5A) for the negative strand template (RNA3A-2). For the positive strand of RNA3 (RNA3B-2), the RNA3B-2 and PV1/SP6 primers gave a product was that was not of the expected size. This band was around 125 to 150 bp long, confirmed with the 4% gel, which was around 200 bp smaller than the expected size of 300 bp. This was a similar size difference between the expected product sizes observed for the original primers

designed for RNA3 (Figure 2.5). To understand this disparity, the RNA3 products amplified using the re-designed gene specific primers were purified and sent for sequencing.

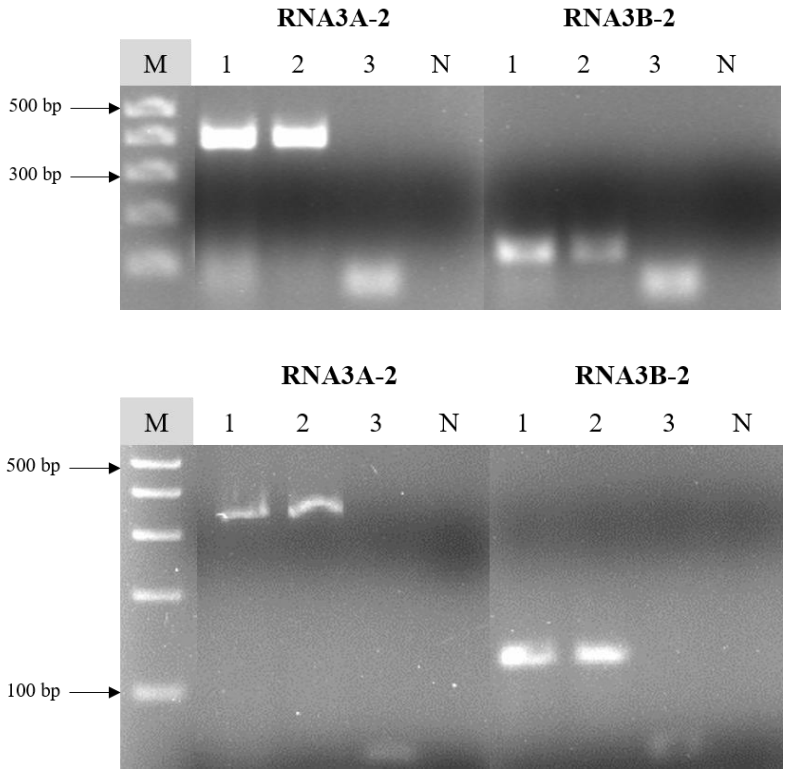


Figure 2.7 – Amplification of RT-PCR products from the negative (RNA3A-2) and negative (RNA2B-2) strands of RNA3 using re-designed gene specific primers. RT-PCR products were electrophoresed through a 2% agarose/1XTBE gel (A), and a 4% agarose gel (B) following amplification of samples R001 (lane 1), R002P (lane 2) and the negative control R003 (lane 3). Lane M: 1 kb Plus DNA ladder. Lane NTC: No template control.

2.3.1.4 Genome Completion

Completion of the genome was done by aligning the sequences obtained through 3' RACE with the original NGS sequences. Sequence data was returned from Macrogen and trimmed of low-quality reads, processed and aligned to give a consensus sequence for the 3' end of both negative and positive sense strands for RNA1 to RNA5. The positive sense sequence was reverse complemented and aligned with the negative sense sequence for each strand to match the negative sense polarity of the strand. The consensus sequence was extracted and for each strand and this constituted the tentative complete genome.

To verify the sequence of each segment, the 3’ and 5’ ends of each segment were aligned to check that they had the characteristic conserved sequences at each of the genome segments observed for other emaraviruses. A 50 bp sequence was extracted from the 3’ and the reverse complement of the 5’ ends of each segment and were aligned using ClustalW in Geneious (Figure 2.8). Around 7 bp was expected show reverse complementarity between the 5’ and 3’ ends of all viral RNA segment.

Figure 2.8 shows the 12 nucleotide sequence GTAGTG(A/T)(G/T)(T/C)TCC is conserved between the viral 3’ ends and the reverse complement of the 5’ ends. This conforms generally with the current literature on emaraviruses, which states that up to 13 nucleotides were reverse complementary on each end, and that seven last nucleotides were always reverse complementary. For KOPV, it appears that the conserved sequence is 12 nucleotides long, with six (GTAGTG) that are perfectly complementary. RNA 1 appears to have an additional guanine at each end, while RNA 3 appears to have an additional adenine.

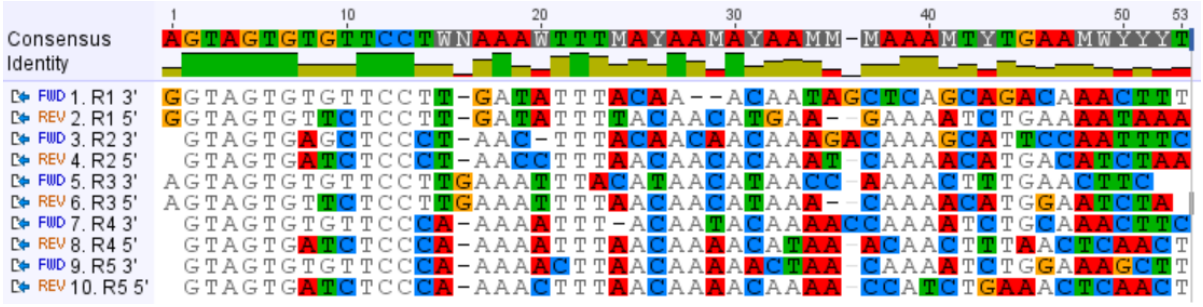


Figure 2.8 – ClustalW alignment of 50 bp from the 3’ and 5’ ends of each RNA segment of the tentatively completed genome. The 5’ ends are the reverse complement sequence of the negative segment genome. Each individual segment was expected to have a 5’ end that show reverse complementarity of up to at least seven nucleotides to the 3’ end of each genome segment. R1-R5 refers to RNA 1-RNA5 of the KOPV genome.

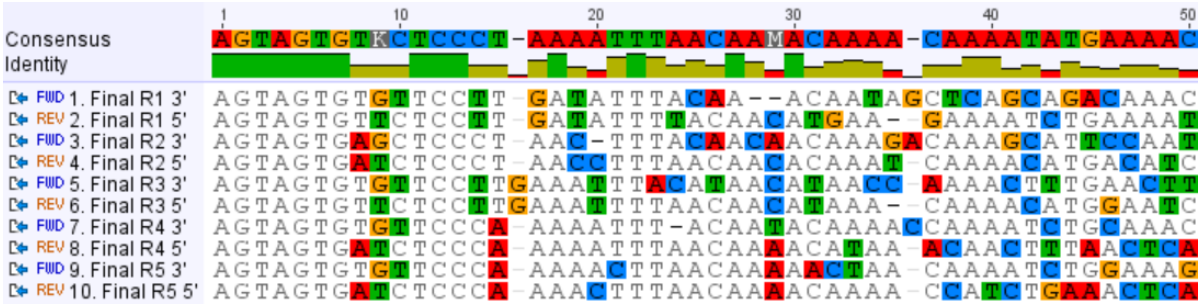


Figure 2.9 – ClustalW alignment of 50 bp from the 3' and 5' end of each RNA segment of the final completed genome. These sequences included data from the Map to Reference assembly of the initial NGS sRNA data mapped to the tentative final genome. The 3' and reverse complemented 5' end of each segment has the same seven nucleotides from the ends (AGTAGTG) and are also the same at position 11 – 13 (TCC). The nucleotides at positions 8 – 10 are variable in each segment.

The sRNA assembly also revealed discrepancies between the tentative and final completed genomes. The final genome was curated and adjusted based on the data shown in Table 2.3 below. The NGS coverage, number of reads and quality were compared to the sRNA assembly coverage, as well as the likelihood of variability within the species, and the genome was adjusted according to which nucleotide was most likely. These changes did not affect the size of the genome or the ORFs, although some of the changed nucleotides had an effect of changing the translated amino acid. Discrepancies were detected for RNA 1, 2 and 5 and no discrepancies were detected for RNA 3 and 4. The NGS original nucleotide was from the original NGS assembly and used in the tentatively completed genome, and the NGS coverage quality was the value given to the coverage quality in Geneious based on depth of coverage and probability of error. The sRNA consensus nucleotide is where the sRNA map to reference assembly differed to the tentatively completed genome. The final genome was adjusted based on the number of reads for a certain nucleotide from both NGS and sRNA assemblies and considering the NGS coverage quality. There may be variability within the KOPV species which could be reflected by these discrepancies; however, to obtain an unambiguous completed genome, the nucleotides with the best likelihood were chosen.

Table 2.3 – Discrepancies between the tentative and final completed genomes for RNA 1, 2 and 5. No discrepancies were present for RNA 3 and 4. The genome was adjusted based on the coverage in the original NGS and sRNA assemblies.

Segment	Position in Final Sequence	NGS Original Nucleotide	NGS Coverage Quality	sRNA Consensus Nucleotide	Number of reads from NGS and sRNA assemblies.	Genome adjusted to:
RNA1	903	C	1554	T	NGS: 49xC, 5xT sRNA: >500 reads of T	T
	1113-1114	CA	1262 1311	TT	NGS: 37xCA, 1xAA sRNA: 3xTT	CA
	1671	A	1382	R	NGS: 38xA sRNA: 6xA, 3xG, 3xT	A
	1686	T	716	C	NGS: 8xC, 27T sRNA: >500 reads of C	C
	2630	T	1528	K	NGS: 48xT, 3xG sRNA: 14xT, 20xG	T
	7061	G	647	A	NGS: 21xG, 3xA sRNA: >100 reads of A	A
RNA2	71	R	1788	A	NGS: 29xA, 20xG sRNA: 15xA	A
RNA5	1065	A	321	G	NGS: 17xA, 7xG sRNA: >200 reads of G	G

After these adjustments, the KOPV genome was complete. Table 2.4 shows the difference in nucleotide length between the original NGS genome, the tentatively completed genome and the final complete genome for the RNA segment, ORF and UTR. For RNA 1, RACE and sRNA mapping resulted in an additional four nucleotides added to the 3’ end and six nucleotides added to the 5’ end compared to the original NGS data. For RNA 2, an additional eight nucleotides were added to both 3’ and 5’ ends. For RNA 3, 22 nucleotides were removed from the 3’ end and 270 nucleotides were removed from the 5’ end. For RNA 4, seven nucleotides were added to both 3’ and 5’ ends. For RNA 5, three nucleotides were added to the 3’ end and eight nucleotides were added to the 5’ end. These comparisons highlight the need to use extra tools to complete genomes assembled from NGS data.

Table 2.4 – Comparison of sequence data for each KOPV RNA segment from original NGS and sRNA sequencing, and 3'RACE. ORF = Open Reading Frame, UTR = Untranslated Region.

	Original NGS Genome			Tentative Complete Genome			Final Complete Genome		
RNA Segment	Segment Length (nt)	ORF Length (nt)	UTR Length (nt)	Segment Length (nt)	ORF Length (nt)	UTR Length (nt)	Segment Length (nt)	ORF Length (nt)	UTR Length (nt)
RNA 1	7131	6963	163	7141	6963	173	7141	6963	173
RNA 2	1927	1800	127	1941	1800	141	1943	1800	143
RNA 3	1771	924	847	1479	924	555	1479	924	555
RNA 4	1504	954	550	1516	954	562	1518	954	564
RNA 5	1571	810	761	1574	810	764	1576	810	766

Table 2.5 shows the characteristics of the completed genome and Figure 2.10 shows a cartoon schematic of the genome. The full nucleotide sequences of each RNA segment and full amino acid sequences for proteins encoded are in Appendix A and Appendix B, respectively. Comparisons with the genomes of other emaraviruses is described in Section 2.3.3.

Table 2.5 – The KOPV completed genome characteristic showing each RNA segment length, untranslated region (UTR) length, open reading frame (ORF) length, protein encoded, protein length and protein molecular weight in kilo-Daltons.

RNA segment	Segment Length (nt)	3' UTR Length (nt)	5' UTR Length (nt)	ORF Length (nt)	ORF encodes protein	Protein Length (aa)	Molecular Weight (kDa)
RNA 1	7141	63	115	6963	RNA-dependent RNA polymerase (RdRp)	2320	271.13
RNA 2	1943	58	85	1800	Glycoprotein	599	68.23
RNA 3	1479	41	514	924	Nucleocapsid	307	34.49
RNA 4	1518	102	62	954	Movement	317	36.20
RNA 5	1576	86	680	810	Unknown Function	269	30.30

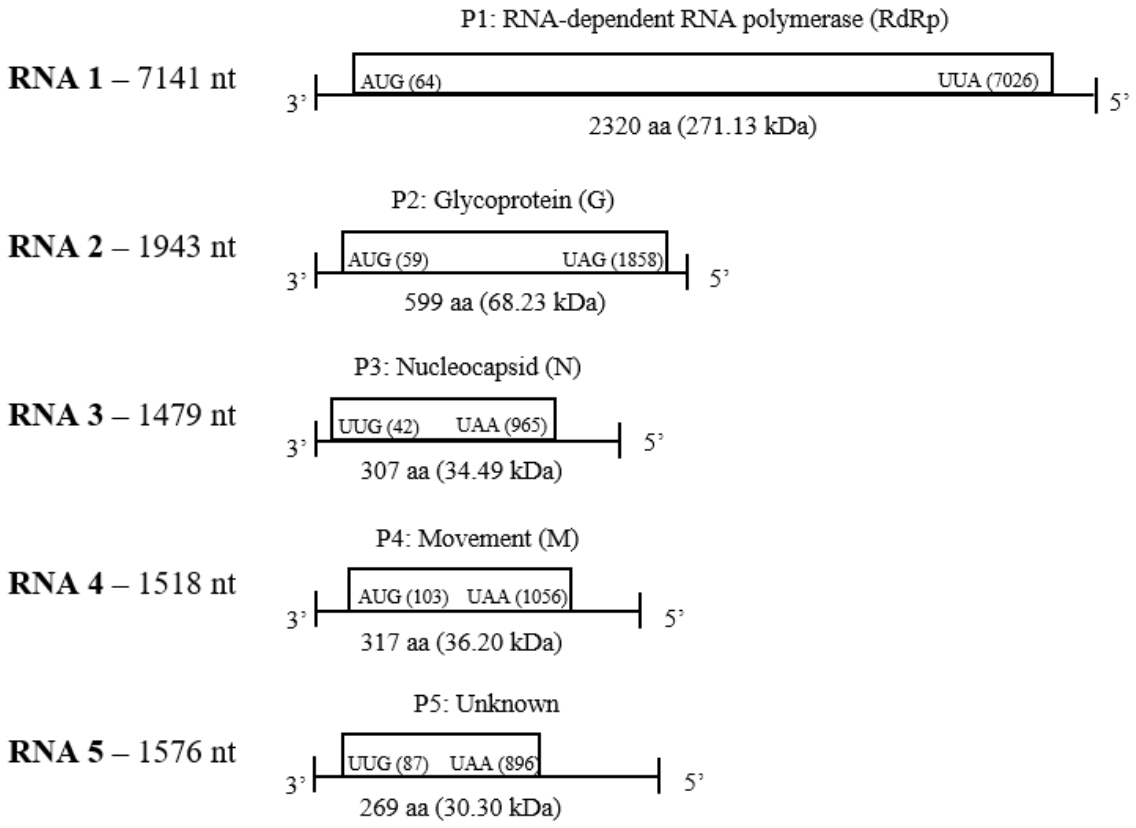


Figure 2.10 – Genomic structure of the five negative-sense RNA segments identified for the KOPV genome. Each RNA segment encodes a single protein. Note: these segments are not to scale.

2.3.2 Sequence Analysis of Complete KOPV Genome

The completed KOPV genome appears to have five negative-sense RNA segments that each encode a putative protein. The entire nucleotide sequence for each segment was submitted to a BLASTn search and no significant similarity to any other sequence in the NCBI nucleotide database was returned. The translated sequences (P1, P2, P3, P4 and P5) from the ORF of each RNA segment (RNA1-5, respectively) were then submitted to a BLASTp search with the top five results for each protein shown in Table 2.6. The proteins P1, P2, P3 and P4 appear to have similarity to the RdRp, glycoprotein, nucleocapsid and movement proteins, respectively, of other emaravirids. P1 had the highest scoring and closest identity match, which was 34.53% identity to the EMARaV RdRp. P1 also showed identity to RdRps belonging to species within the *Orthobunyavirus* genus of between 24 to 26% (data not shown). P2 appears to be the envelope glycoprotein for KOPV since the most similar sequences were glycoproteins from

other emaraviruses such as Palo verde broom virus and EMARaV. P3 is most likely to be nucleocapsid protein as its closest match was this protein from TiRSaV and fig mosaic emaravirus. The only viral sequence to show similarity to P4 was the movement protein from *Palo verde broom virus*, suggesting this is the function of P4. It appears that no available sequence is similar to P5, thus its function remains undetermined.

Sequence identity of the KOPV protein sequences showed that they are like other emaravirid and provides evidence that it belongs in the genus. The overall identity between the KOPV sequences and those of other emaraviruses was not high, indicating that KOPV is a distinct emaravirus. Emaravirus protein sequences also have conserved traits within them that, if present in the proteins from KOPV, will provide more evidence that the virus should be included in the genus.

Table 2.6 – Top five BLASTp results for each KOPV protein from NCBI using the non-redundant protein database (blast.ncbi.nlm.nih.gov/Blast.cgi). P1, P2, P3 and P4 all have a percent identity to other emaravirus species. P5 has no percent identity to any other emaravirus species.

Karaka emaravirus Protein	Description	Max Score	Total Score	Query Cover	E Value	Per. Ident.	Accession
P1	RNA-dependent RNA polymerase [European mountain ash ringspot-associated emaravirus]	1054	1054	85%	0	34.53%	VFU05375.1
	RNA-dependent RNA-polymerase [Redbud yellow ringspot-associated emaravirus]	1044	1044	85%	0	33.15%	YP_009508083.1
	RNA-dependent RNA polymerase [Actinidia chlorotic ringspot-associated virus]	1042	1042	86%	0	33.70%	YP_009507925.1
	RNA-dependent RNA polymerase [European mountain ash ringspot-associated emaravirus]	1032	1032	85%	0	34.00%	YP_003104764.1
	RNA dependent RNA polymerase [Pigeonpea sterility mosaic emaravirus 2]	1009	1009	82%	0	34.44%	CEJ20912.1
P2	envelope glycoprotein [Palo verde broom virus]	78.2	78.2	77%	3.00E-11	20.98%	AWH90170.1
	envelope glycoprotein [Palo verde broom virus]	77.8	77.8	91%	4.00E-11	21.40%	AWH90172.1
	envelope glycoprotein [Palo verde broom virus]	77.8	77.8	91%	5.00E-11	21.40%	AWH90168.1
	Glycoprotein precursor [European mountain ash ringspot-associated emaravirus]	76.6	76.6	69%	1.00E-10	22.32%	VFU05376.1
	glycoprotein precursor p2 [Pistacia emaravirus]	75.9	75.9	75%	2.00E-10	21.44%	QAR18003.1
P3	putative nucleocapsid protein [Ti ringspot-associated emaravirus]	72.4	72.4	74%	9.00E-11	27.73%	QAB47309.1
	putative nucleocapsid protein [Fig mosaic emaravirus]	72.4	72.4	79%	1.00E-10	27.56%	YP_009237270.1
	nucleocapsid protein [Fig mosaic emaravirus]	70.9	70.9	79%	5.00E-10	27.56%	AWS21342.1
	nucleocapsid protein [Fig mosaic emaravirus]	70.5	70.5	79%	6.00E-10	27.73%	BAM13808.1
	nucleocapsid protein [Fig mosaic emaravirus]	70.1	70.1	79%	8.00E-10	27.56%	AQR59327.1
P4	movement protein [Palo verde broom virus]	99.4	99.4	83%	7.00E-20	26.04%	AWH90178.1
	1,25-dihydroxyvitamin D(3) 24-hydroxylase, mitochondrial-like [Saccoglossus kowalevskii]	42	42	19%	2.4	32.79%	XP_006816050.1
	dTDP-glucose 4,6-dehydratase [Candidatus Methylopumilus universalis]	40	40	19%	7.7	37.10%	WP_139929806.1
P5	AsmA family protein [Sphingomonas sp.]	40.8	40.8	33%	4.1	25.77%	PZU11974.1

2.3.2.1 P1 – RNA-Dependent RNA-Polymerase

The KOPV RNA 1 is 7141 nucleotides in length with an ORF found between AUG_{64-66}-$UUA_{7024-7026}$ depicted in Figure 2.11. This ORF encodes a putative 271.13 kDa protein (P1) of 2320 amino acids. P1 has a conserved domain between $aa_{653-1395}$ with an E-value of 1.22e-34 that belongs to the bunyavirus RNA-dependent RNA polymerase super family (Accession cl20265) providing strong evidence to support P1 as the RdRp (Marchler-Bauer et al., 2015). P1 is the only KOPV protein that returns a result using CDD search tool. P1 also has the expected conserved emaravirus amino acid motifs. Motif A is $DASKWS_{1128-1133}$ in the P1 aa sequence and is highly conserved across all emaravirus RdRps (Di Bello et al., 2016; Elbeaino et al., 2009; Mielke & Muehlbach, 2007). Motif B is at $QGNLNRLSS_{1213-1221}$ in KOPV, which matches the conserved sequence QGNXNXXSS – where X represents any residue – reported previously (Di Bello et al., 2016). Motif C is also highly conserved across all emaraviruses and is at $SDD_{1254-1256}$ in P1, while Motif D is reported to be KK and lies between Motif C and Motif E on the amino acid sequence (Di Bello et al., 2016; Zheng et al., 2017). In KOPV there appear to be two Motifs D, one at $KK_{1266-1267}$ and the other downstream at $KK_{1301-1302}$. Motif E is also highly conserved and is present at $EFLST_{1311-1315}$ in KOPV. P1 shares the N-terminus endonuclease domain as reported for emaraviruses and is located at $RHD_{106-108}X_{35}PD_{145-146}X_{12}EVK_{158-160}$. P1 also shares the conserved premotif A sequence at $KDQRTYNDREIYTGNKEAR_{1050-1068}$ which has closest similarity to the premotif A sequences in JYMaV, RLBV and HPWMoV (Di Bello et al., 2016; McGavin et al., 2012; Tatineni et al., 2014; Yang et al., 2019).

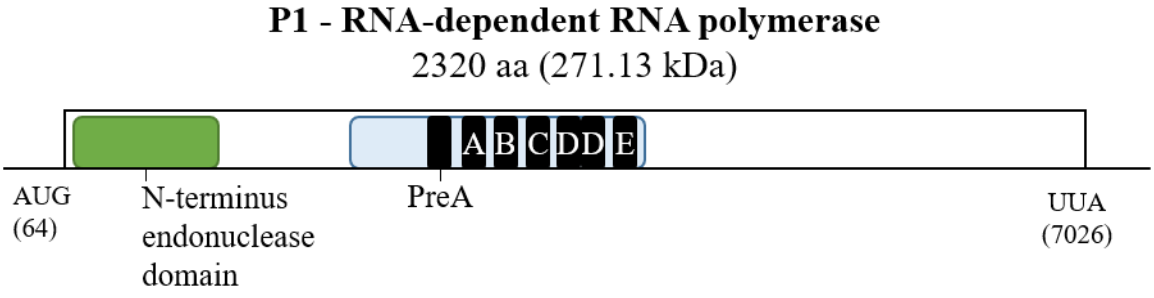


Figure 2.11 – A schematic representing P1 of KOPV depicting the relative locations of conserved emaravirus motifs. The blue box containing the motifs represents bunyavirus RNA-dependent RNA polymerase super family (Accession cl20265). This schematic is not to scale.

2.3.2.2 P2 – Glycoprotein

RNA 2 is 1943 nt long and contains a large ORF between $AUG_{59\text{-}61}$ to $UAG_{1856\text{-}1858}$ that is predicted to encode a glycoprotein (P2) of 599 aa with a molecular weight of 68.23 kDa. P2 does not appear to contain any N-glycosylation sites when analysed using NetNGlyc 1.0, or any O-glycosylation sites when analysed using NetOGlyc 4.0, which is unexpected. Analysis of previously reported emaraviruses using the same tool revealed that emaraviruses usually contain between four and six N-glycosylation sites (Elbeaino et al., 2009; Hassan et al., 2017; Laney et al., 2011). Analysing P2 with SignalP 5.0 revealed a potential signal peptide with a cleavage site between $TTS_{17\text{-}19}$ and $KY_{20\text{-}21}$. The likelihood that the site is a signal peptide was 0.9692 and the probability that it is a cleavage site was 0.3522 giving weak support for this region being a cleavage site. Analysis using TargetP 1.1 also indicated a signal peptide was present in the first 20 aa that belonged to a secretory pathway. This signal peptide may be involved in targeting P2 to the endoplasmic reticulum, based on similar analysis of RRV (Laney et al., 2011). Tmpred was used and predicted that five transmembrane helices are present in P2, more than the two to four helices present in other emaraviruses (Di Bello et al., 2016; Laney et al., 2011; Mielke & Muehlbach, 2007). The tetrapeptide sequence (ADDN) that has been reported in the glycoprotein of other emaraviruses predicted to cleave the protein into two smaller proteins, is not present in P2 (Di Bello et al., 2016). Overall, there is not enough evidence present to conclude that P2 is the glycoprotein for KOPV.

2.3.2.3 P3 – Nucleocapsid

RNA 3 is 1479 nt long and contains an ORF between $UUG_{42\text{-}44}$ and $UAA_{963\text{-}965}$ that is predicted to encode a nucleocapsid protein (P3) of 307 aa with a molecular weight of 34.49 kDa. P3 contains three conserved amino acid stretches; regions A, B and C (Figure 2.12), these are $NKFVMSSNR_{119\text{-}127}$, $NRLA_{173\text{-}176}$ and $GVEN_{194\text{-}197}$, respectively. Region A of KOPV is the most dissimilar to the other emaraviruses whereas region B is a similar sequence as all other emaraviruses. For region C, the KOPV sequence is the only one to have a valine (V)

residue in the second position. The similarity of these domain suggests that KOPV P3 is the nucleocapsid protein.

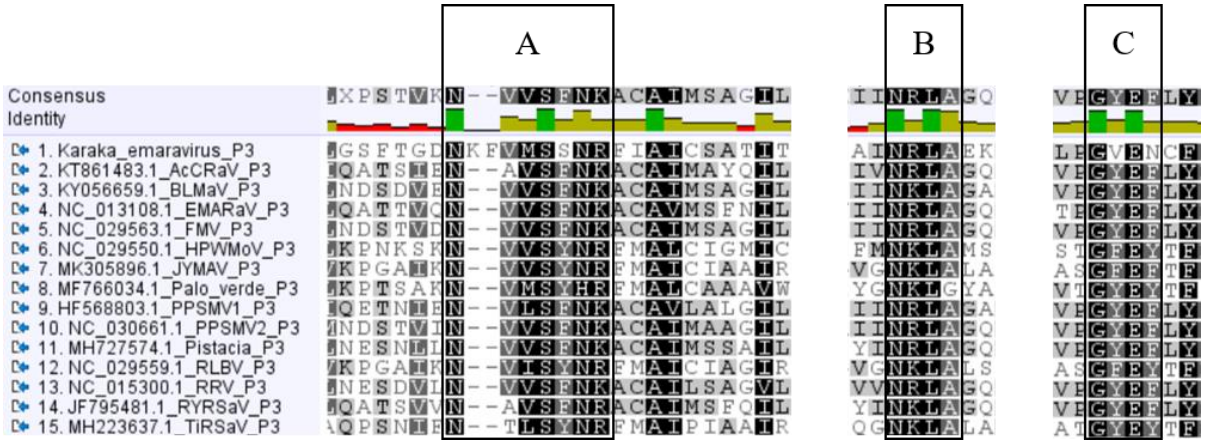


Figure 2.12 – ClustalW alignment of RNA 3 translated amino acid sequences from known emaravirus species. In KOPV, region A is between aa 119-127, region B is between aa 173-176 and region C is between aa 194-197.

2.3.2.4 P4 – Movement Protein

RNA 4 is 1518 nt long with an ORF between $AUG_{103\text{-}105}$ and $UAA_{1054\text{-}1056}$ that is predicted to encode a movement protein (P4) of 317 aa with a molecular weight of 36.20 kDa. A signal peptide was detected using SignalP 5.0 between aa positions 1 and 18, similar to AcCRaV (Zheng et al., 2017). P4 does not appear to have any conserved domains. A PROMALS alignment was done with several other emaraviruses to compare the previously reported 30K superfamily domain of this protein, shown in Figure 2.13 (Di Bello et al., 2016; Zheng et al., 2017). The KOPV P4 sequence has a large gap between consensus aa 160 and 178 that marks it as different to the other emaraviruses. There are regions that appear conserved in other emaravirids, such as between aa 187 and 195, and between aa 209 and 217 that KOPV does not share. It does appear to have weak similarity to the core structure of other emaravirus 30K superfamily domains. The PVBV P4 also does not share high similarity with these apparently conserved sequences, which may explain why it was the only sequence to match with KOPV when using BLASTp. It has not been identified as the movement protein. The only evidence that P4 is the movement protein in KOPV is its relationship to PVBV P4 and as such there is not sufficient evidence to conclude that KOPV P4 is the movement protein.

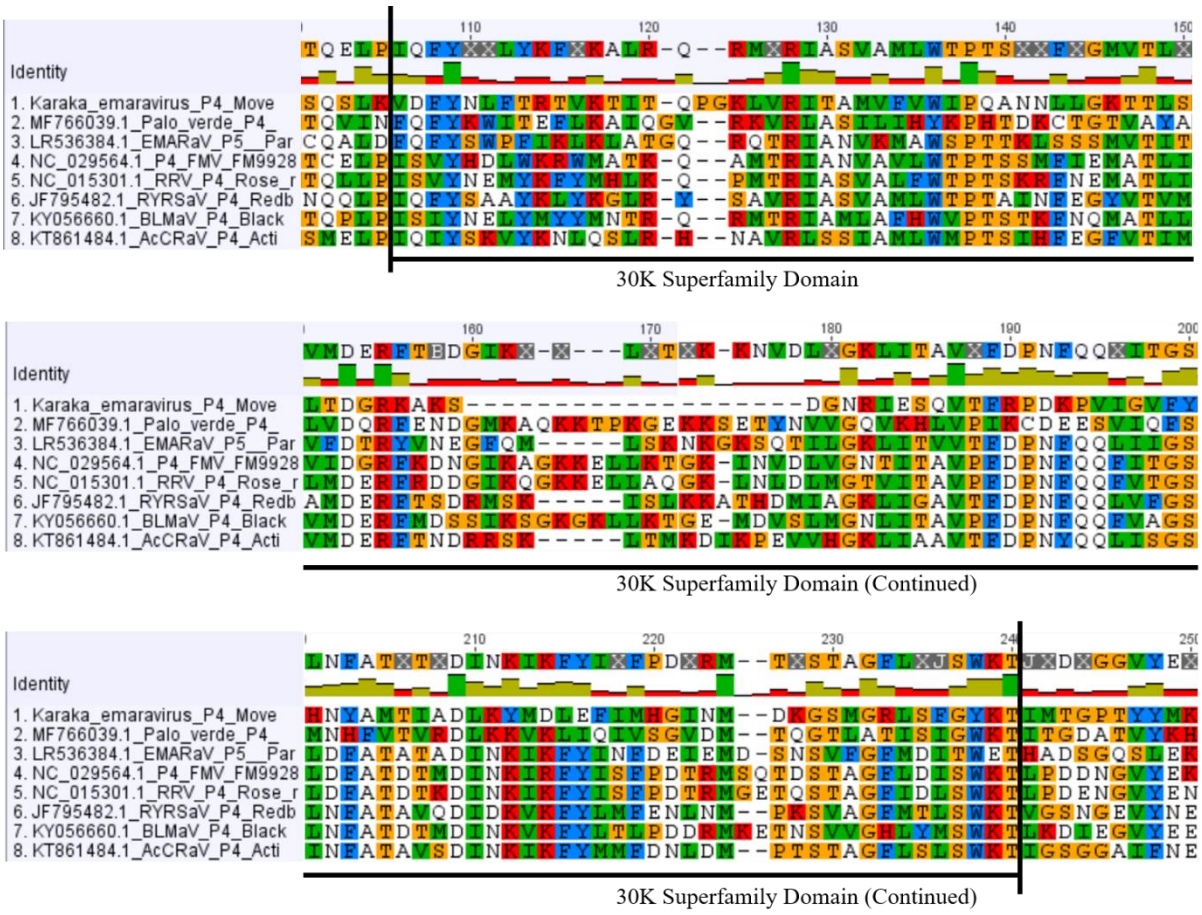


Figure 2.13 – PROMALS alignment of movement protein (P4) sequences with the previously reported 30K superfamily domain highlighted. This figure shows the sequences between aa 100 and 250 of the consensus sequences. The sequences are: KOPV P4, MF766039.1 *palo verde broom virus* P4, LR536384.1 *European Mountain Ash ringspot-associated emaravirus* P5 (which is the movement protein), NC_029564.1 *fig mosaic emaravirus* P4, NC_015301.1 *rose rosette emaravirus* P4, JF795482.1 *redbud yellow ringspot-associated emaravirus* P4, KY056660.1 *blackberry leaf mottle-associated emaravirus* P4 and KT861484.1 *Actinidia chlorotic ringspot-associated emaravirus* P4. The highlighted region between the black vertical lines (consensus aa 106 and 241), is the 30K superfamily domain reported by Zheng et al. (2017).

2.3.2.5 P5 – Protein with Unknown Function

RNA 5 is 1576 nt long with an ORF between $UUG_{87\text{-}89}$ and $UAA_{894\text{-}896}$ that encodes a protein with unknown function (P5) of 269 aa with a molecular weight of 30.30 kDa. The amino acid sequence appears to share no similarity with any other viral protein nor any conserved domains. The amino acid sequence has no coiled-coil domains, no signal peptide for a secretory pathway, no signal peptide for any cleavage sites, and weak likelihood of transmembrane helices. The protein does contain four N-glycosylation sites and potentially ten

O-glycosylation sites when using NetNGlyc 1.0 and NetOGlyc 4.0, respectively, suggesting that the protein could function as a glycoprotein.

2.3.3 Comparison with other emaravirus species

2.3.3.1 Genome Structure and Nucleotide Length Comparison

The genome of structure of KOPV is similar to other emaraviruses, as shown in Table 2.7 and Table 2.8. KOPV has five genomic RNA segments compared to between five and eight for other emaravirids. The nucleotide length of KOPV RNA segments 1, 3 and 4 are similar to other emaraviruses. The nucleotide length of RNA strand 2, and the ORF that encodes the protein, is substantially shorter than the other emaraviruses which may be why no N-glycosylation sites were found in the analysis in section 2.3.2.2. RNA 5 has a nucleotide length and ORF length similar to some of the other emaraviruses, such as RRV RNA 6 and RYRSaV RNA 5.

Table 2.7 – Comparison table of nucleotide and amino acid lengths for emaravirus species RNA 1 to 4

RNA	Virus	Virus ID - Accession	Genome	3' UTR	ORF	ORF	5' UTR
			Length (nt)	Length (nt)	Length (nt)	Length (aa)	Length (nt)
RNA 1	Karaka Emaravirus	-	7141	63	6963	2320	115
(RdRp)	AcCRaV	KT861481	7061	38	6912	2303	111
	BLMaV	KY056657	7050	42	6915	2304	93
	EMARaV	NC_013105	7040	44	6882	2293	114
	FMV	NC_029562	7039	39	6894	2297	106
	HPWMoV	NC_029570	6981	68	6819	2272	94
	JYMaV	MK305894	7143	47	6984	2327	112
	Palo verde	MF766024	7015	60	6849	2282	106
	PiVB	MH727572	7027	52	6900	2299	75
	PPSMV1	HF568801	7022	49	6885	2294	88
	PPSMV2	NC_030660	7009	45	6885	2294	79
	RLBV	NC_029567	7062	48	6888	2295	126
	RRV	NC_015298	7026	62	6876	2291	88
	RYRSaV	JF795479	7049	40	6900	2299	109
	TiRSaV	MH223635	7217	56	7023	2340	138
RNA 2	Karaka Emaravirus	-	1943	58	1800	599	85
(Glycoprotein)	AcCRaV	KT861482	2267	251	1962	653	54
	BLMaV	KY056658	2271	47	1959	652	265
	EMARaV	NC_013106	2335	58	1941	646	336
	FMV	NC_029565	2252	52	1926	641	274
	HPWMoV	NC_029549	2211	79	2004	667	128
	JYMaV	MK305895	2233	21	2022	673	190
	Palo verde	MF766029	2098	40	1929	642	129
	PiVB	MH727573	2245	56	1923	640	266
	PPSMV1	HF568802.3	2223	41	1947	648	235
	PPSMV2	NC_030662	2229	47	1950	649	232
	RLBV	NC_029558	2135	49	1953	650	133
	RRV	NC_015299	2245	50	1938	645	257
	RYRSaV	JF795480	2220	50	1929	642	241
	TiRSaV	MH223636	2399	57	2154	717	188
RNA 3	Karaka Emaravirus	-	1479	41	924	307	514
(Nucelocapsid)	AcCRaV	KT861483	1678	645	933	310	100
	BLMaV	KY056659	1510	97	951	316	462
	EMARaV	NC_013108	1559	119	945	314	495
	FMV	NC_029563	1491	99	948	315	444
	HPWMoV	NC_029550	1441	219	870	289	352
	JYMaV	MK305896	1259	70	885	294	304
	Palo verde	MF766034	1356	85	879	292	392
	PiVB	MH727547	1525	101	918	305	506
	PPSMV1	HF568803	1442	102	927	308	413
	PPSMV2	NC_030661	1335	295	939	312	101
	RLBV	NC_029559	1365	53	879	292	433
	RRV	NC_015300	1544	99	951	316	494
	RYRSaV	JF795481	1414	89	942	313	383
	TiRSaV	MH223637	1106	128	861	286	117
RNA 4	Karaka Emaravirus	-	1518	102	954	317	462
(Movement)	AcCRaV	KT861484	1664	426	1140	379	98
	BLMaV	KY056660	1504	29	1143	380	332
	EMARaV	**LR536379***	1629	109	1122	373	398
	FMV	NC_029564	1472	81	1086	361	305
	HPWMoV	NC_029551	1682	112	1095	364	475
	JYMaV	MK305897	1547	101	1125	374	321
	Palo verde	MF766039	1481	70	1101	366	310
	PiVB	MH727575	1550	81	1086	361	383
	PPSMV1	HF568804	1563	77	1086	361	400
	PPSMV2	NC_029560	1833	306	1422	473	105
	RLBV	NC_029560	1675	81	1122	373	472
	RRV	NC_015301	1541	83	1086	361	372
	RYRSaV	JF795482	1513	99	1131	376	283
	TiRSaV	MH223638	1342	81	999	332	262

Table 2.8 – Comparison of KOPV RNA 5 against other emaravirus genomic strands that have unknown functions.

RNA	Virus	Virus ID - Accession	RNA #	Genome	5' UTR	ORF	ORF	3' UTR
				Length (nt)	Length (nt)	Length (nt)	Length (aa)	Length (nt)
RNA 5 - 8	Karaka Emaravirus	-	RNA 5	1576	86	810	269	680
Unknown Function	AcCRaV	KT861485	RNA 5	1476	75	702	233	699
	BLMaV	KY056661	RNA 5	1224	68	681	226	475
	EMARaV	NC_013107	RNA 4	1348	145	699	232	504
		LR536380	RNA 6	1362	147	690	229	525
	FMV	NC_029566	RNA 5	1752	57	1509	502	186
		NC_029568	RNA 6	1212	68	567	188	577
	HPWMoV	NC_029552	RNA 5	1715	120	1437	478	158
		NC_029553	RNA 6	1752	118	1479	492	155
		NC_029554	RNA 7	1434	117	918	305	399
		NC_029555	RNA 8	1339	93	531	176	715
	JYMaV	MK305898	RNA 5	1267	83	837	278	347
		MK305899	RNA 6	980	89	585	194	306
	Palo verde	-	-	-	-	-	-	-
	PiVB	MH727576	RNA 5	1716	87	1449	482	180
		MH727578	RNA 6	1340	68	723	240	549
		MH727579	RNA 7	1709	14	1515	504	180
	PPSMV1	HF945448.2	RNA 5	1801	84	1422	473	295
		KX363891	RNA 6	1194	29	756	251	409
	PPSMV2	NC_030658	RNA 5	1833	105	1422	473	306
		NC_030659	RNA 6	1194	29	756	251	409
	RLBV	NC_029561	RNA 5	1718	67	1431	476	220
		NC_029571	RNA 6	1095	93	567	188	435
		NC_029572	RNA 7	1089	77	579	192	433
		NC_029573	RNA 8	1273	86	702	233	485
	RRV	NC_034979	RNA 5	1665	61	1404	467	200
		NC_034980	RNA 6	1402	68	702	233	632
		NC_034981	RNA 7	1649	61	1398	465	190
	RYRSaV	KU904300	RNA 5	1272	86	672	223	514
	TiRSaV	MH223639	RNA 5	1323	69	516	171	738

2.3.3.2 Protein Comparison

KOPV shares amino acid identity with other emaravirus species to varying degrees. To determine the relationships between each emaravirus species and candidate species, each encoded protein sequence was compared to KOPV using a MUSCLE alignment. Amino acid comparisons between KOPV and the other emaraviruses are listed in Table 2.9. Generally, KOPV predicted proteins have similar amino acid pairwise identities to known emaravirus proteins with no species having a greater identity match overall. For the RdRp ORF encoded by RNA 1, pairwise identity was between 28.9% and 31.4% with the greatest identity match between KOPV and EMARaV. For the glycoprotein, encoded by RNA 2, the pairwise identity

was between 17.6% and 21.6% with the highest being with HPWMoV. For the nucleocapsid, encoded by RNA 3, the pairwise identity was between 19.3% and 25.7% with the highest being with JYMaV. For the movement protein, encoded by RNA 4, the pairwise identity was between 15.7% and 23.2% with the highest being with PVBV. RNA 5 was compared with all proteins with unknown functions from other emaraviruses and had the highest pairwise identity to PPSMV2 P6 at 17.6%. Four other proteins shared a relatively high percentage similarity to KOPV P5: JYMaV P5 at 17.0 %, RYRSaV P5 at 17.0%, HPWMoV P7 at 16.7% and PPSMV1 P6 at 16.0%. These proteins also share a similar amino acid length suggesting they may be orthologues although with a lower percentage identity match there is not enough evidence to conclude this. It is difficult to compare P5 with any other emaravirus amino acid sequence due to the low similarity KOPV has with these other species sequences, as seen in Table 2.9.

The pairwise identities are low between KOPV and other emaraviruses when comparing what has been observed between other emaravirus species (See Appendix C). For example, the EMARaV P1 has an average of 44.0% pairwise identity with other emaraviruses with the highest being with RYRSaV at 53.9%. This is also seen for P2, P3 and P5 (the EMARaV movement protein). P2 has an average of 31.9% pairwise identity with the highest being with RYRSaV at 41.6%, P3 having an average of 30.3% with the highest again with RYRSaV at 43.3%, and P5 having an average of 25.9% with the highest being with FMV at 32.6%. These percentages are similar with other emaravirus pairwise identities to the rest of the genus and are higher than KOPV percentages with the genus. This data suggests that KOPV could be more distantly related to other emaraviruses. Amino acid pairwise similarities between each emaravirus species and candidate for P1, P2, P3 and P4 can be found in Appendix C.

Table 2.9 – Amino acid sequence comparisons between KOPV and other emaravirus sequences. The comparison was done using translated ORFs for all RNAs of each species with MUSCLE and the % pairwise identity is listed for each comparison. Accession numbers refer to sequences located on the NCBI database. *LR536379 is P5 from the recently reported update to the EMARaV genome as it is shares greater homology to other emaravirus movement proteins (von Bargen et al., 2019). The RNA 5 column lists the protein and amino acid length of each emaravirus protein that KOPV P5 is compared to.

	Karaka Emaravirus											
	RNA 1		**RNA 2**		**RNA 3**		**RNA 4**		**RNA 5**			
	RdRp		**GP**		**N**		**M**		**Unknown**			
Species	**Accession**	**%**	**Accession**	**%**	**Accession**	**%**	**Accession**	**%**	**Accession**	**Protein**	**aa**	**%**
Karaka V	-	-	-	-	-	-	-	-	-	P5	269	-
AcCRaV	KT861481	30.3	KT861482	19.6	KT861483	19.6	KT861484	17.3	KT861485	P5	233	14.0
BLMaV	KY056657	30.1	KY056658	19.6	KY056659	21.9	KY056660	15.7	KY056661	P5	226	14.8
EMARaV	NC_013105	31.4	NC_013106	19.6	NC_013108	21.3	**LR536379***	16.0	NC_013107	P4	232	13.4
									LR536380	P6	108	15.3
FMV	NC_029562	30.6	NC_029565	19.2	NC_029563	24.0	NC_029564	16.4	NC_029566	P5	502	14.5
									NC_029568	P6	188	15.3
HPWMoV	NC_029570	28.9	NC_029549	21.6	NC_029550	23.7	NC_029551	22.0	NC_029552	P5	478	12.2
									NC_029553	P6	492	11.9
									NC_029554	P7	305	16.7
									NC_029555	P8	176	15.2
JYMaV	MK305894	29.0	MK305895	19.0	MK305896	25.7	MK305897	18.0	MK305898	P5	278	17.0
									MK305899	P6	194	13.8
Palo verde	MF766024	31.0	MF766029	19.2	MF766034	22.5	MF766039	23.2	-	-	-	-
PiVB	MH727572	30.0	MH727573	19.0	MH727547	19.3	MH727575	16.0	MH727576	P5	482	13.3
									MH727578	P6	240	15.7
									MH727579	P7	504	12.2
PPSMV1	HF568801	30.4	HF568802.3	17.6	HF568803	20.0	HF568804	18.3	HF945448.2	P5	473	15.0
									KX363891	P6	251	16.0
PPSMV2	NC_030660	31.0	NC_030662	18.3	NC_030661	23.0	NC_029560	17.6	NC_030658	P5	473	13.0
									NC_030659	P6	251	17.6
RLBV	NC_029567	29.9	NC_029558	18.9	NC_029559	21.8	NC_029560	21.8	NC_029561	P5	476	11.9
									NC_029571	P6	188	15.0
									NC_029572	P7	192	14.7
									NC_029573	P8	233	14.8
RRV	NC_015298	30.6	NC_015299	19.5	NC_015300	20.3	NC_015301	17.8	NC_034979	P5	467	15.0
									NC_034980	P6	233	14.4
									NC_034981	P7	465	13.0
RYRSaV	JF795479	30.5	JF795480	17.7	JF795481	21.3	JF795482	18.8	KU904300	P5	223	17.0
TiRSaV	MH223635	30.5	MH223636	18.2	MH223637	23.2	MH223638	22.5	MH223639	P5	171	13.3

2.3.3.3 Conserved End Sequence Comparison

A feature of emaraviruses is that the RNA termini are conserved across the genus. To determine if the RNA ends of KOPV are similar to that of other emaraviruses, 50 nt from the 3’ and 5’ ends of RNA1 from each emaravirus sequence were compared. Figure 2.11 shows that the first seven nucleotides are fully conserved among all emaravirus species and candidates and the nucleotides at positions 10 to 15 are also highly conserved. KOPV differs from the other species with a T instead of a C at position 10 at the 3’ end, and a T instead of a C at position 14 at both 3’ and 5’ ends. HPWMoV and the palo verde broom virus also differ by having an A instead of a T at position 15 in both 3’ and 5’ ends. The PiVB 5’ end and both 3’ and 5’ ends of the palo verde broom virus do not share this conserved region and it appears that there is sequence missing from these strands. This is likely due to these two viruses not having confirmation of their 3’ and 5’ ends using RACE or any other method (Buzkan et al., 2019; Ilyas et al., 2018). This homology with other emaravirus species supports KOPV belonging to the emaravirus genus.

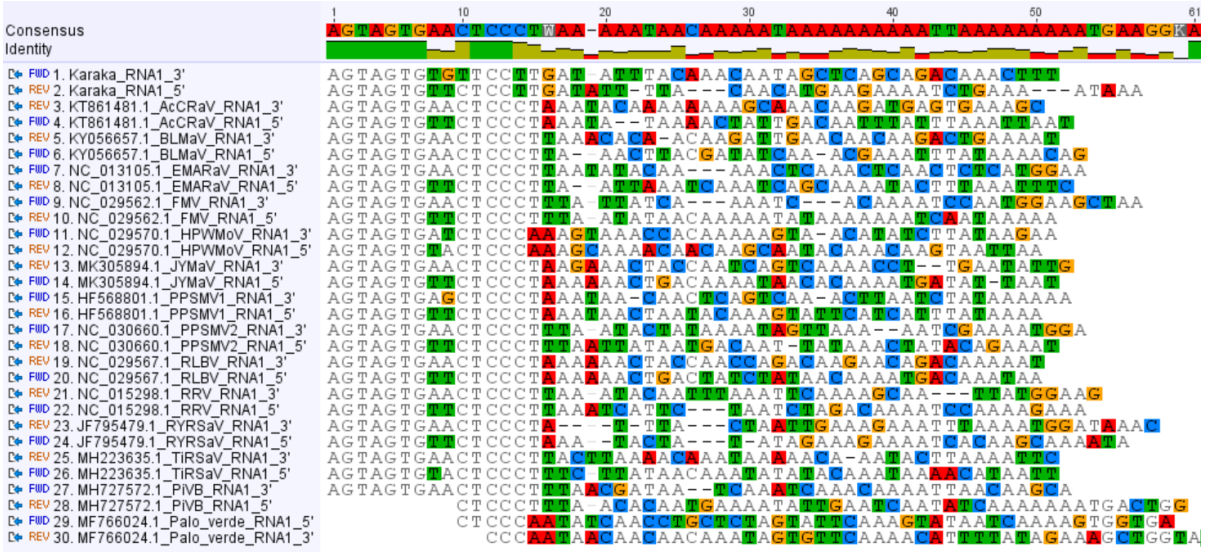


Figure 2.14 – Alignment of 50 bp from the 3’ and 5’ ends of RNA1 of each emaravirus species and candidate species including KOPV (labelled as Karaka_RNA1_3’ and Karaka_RNA1_5’ in this figure).

.

2.3.3.4 Phylogenetic Analysis

Phylogenetic analysis was carried out to determine the relationships between KOPV and the other emaravirids. Initially, a neighbour joining tree was constructed for RNA 1, shown in Figure 2.15. The relationships between the emaraviruses is similar to the relationship shown in previous research on AcCRaV, which utilised amino acid alignments rather than nucleotide alignments (Zheng et al., 2017). All the emaravirus sequences group together, including KOPV. The emaraviruses group into two clades; interestingly, HPWMoV is not grouped with RLBV and appears more distantly related than previously reported (Wang et al., 2020). The grouping of HPWMoV with EMARaV etc is not well supported (57%), suggesting its position in the tree is still unresolved. KOPV shares a common ancestor with Palo verde broom virus, TiRSaV, JYMaV and RLBV. The lineage leading to KOPV appears to have emerged earlier from this ancestor than for the other viruses in this clade.

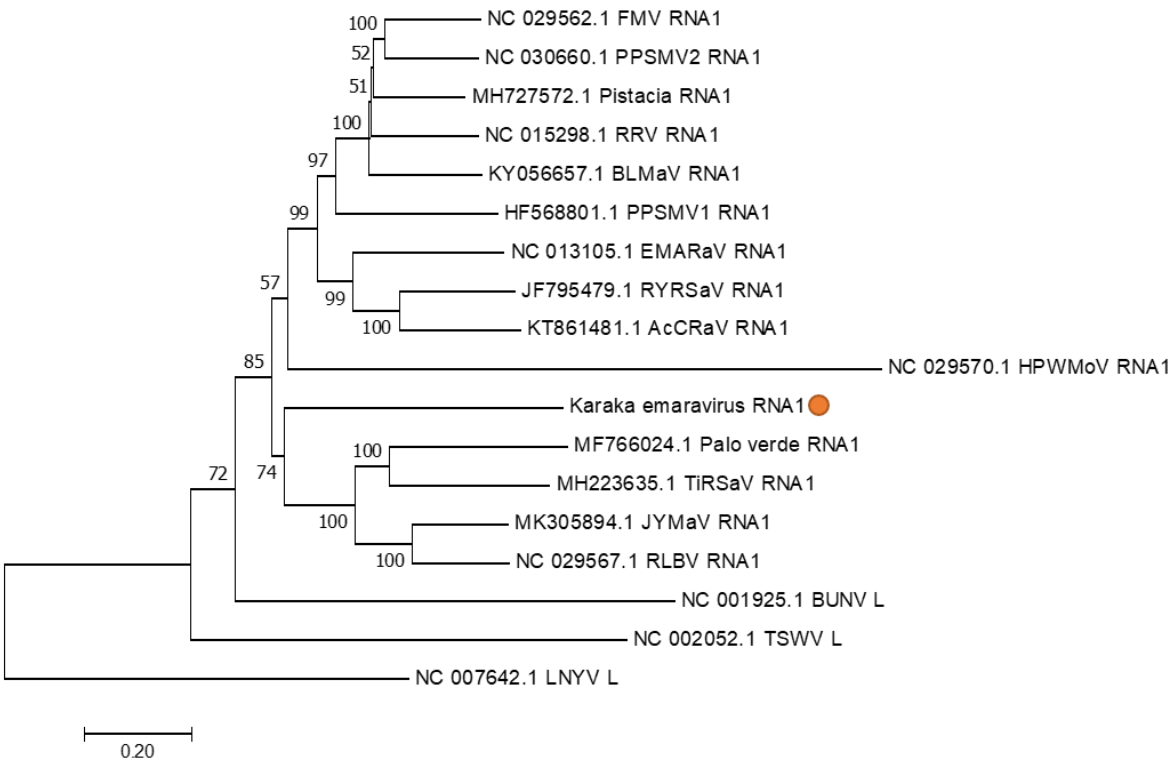


Figure 2.15 – Neighbour joining tree for the nucleotide sequences of RNA 1 from all available species and candidates of the Emaravirus genus. Accession numbers refer to respective records in the NCBI database. BUNV is Bunyamwera virus, an orthobunyavirus, and TSWV is tomato spotted wilt virus, a tospovirus. Both relative amino acid sequences from BUNV and TSWV were included as they belong to closely related genera to emaraviruses. LNYV is lettuce necrotic yellows virus, a cytorhabdovirus, and is used as the outgroup.

To be able to relate to previous studies, which used amino acid sequence alignments for their trees, and to be consistent with previous analysis in this research, subsequent phylogenetic analyses were created using the amino acid sequences. Neighbour joining trees for the amino acid sequences of P1, P2, P3 and P4 from KOPV and other emaravirids are shown in Figure 2.16 (P1 and P2) and Figure 2.17 (P3 and P4).

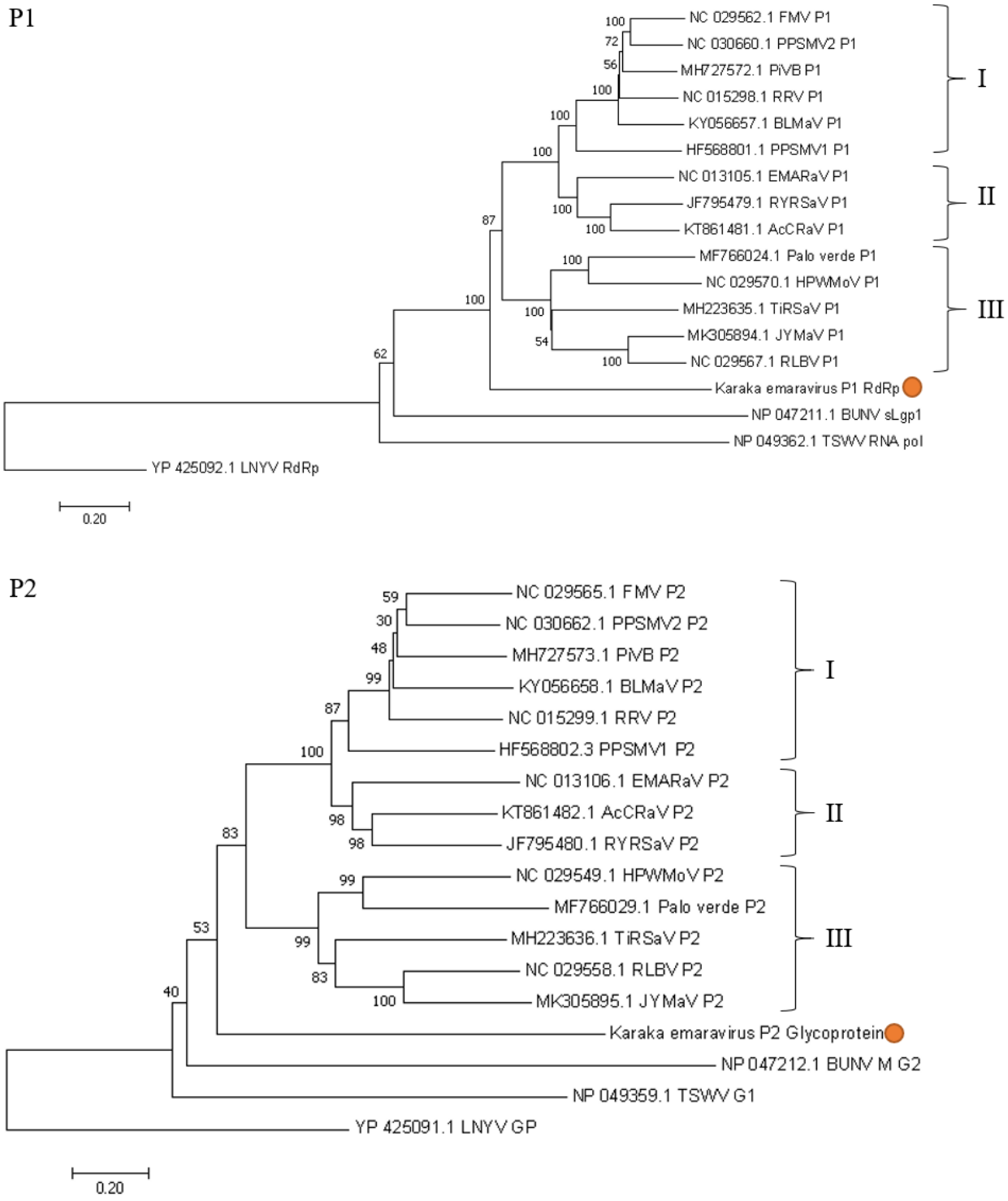


Figure 2.16 – Phylogenetic trees of the amino acid sequences of all species and candidates of the Emaravirus genus. P1 is the RNA-dependent RNA polymerase (RdRp) and P2 is the glycoprotein. Accession numbers refer to respective records in the NCBI database. BUNV is Bunyamwera virus, an orthobunyavirus, and TSWV is tomato spotted wilt virus, a tospovirus. Both relative amino acid sequences from BUNV and TSWV were included as they belong to closely related genera to emaraviruses. LNYV is lettuce necrotic yellows virus, a cytorhabdovirus, and is used as the outgroup. Groups I, II and III represent clades previously reported by Wang et al. (2020).

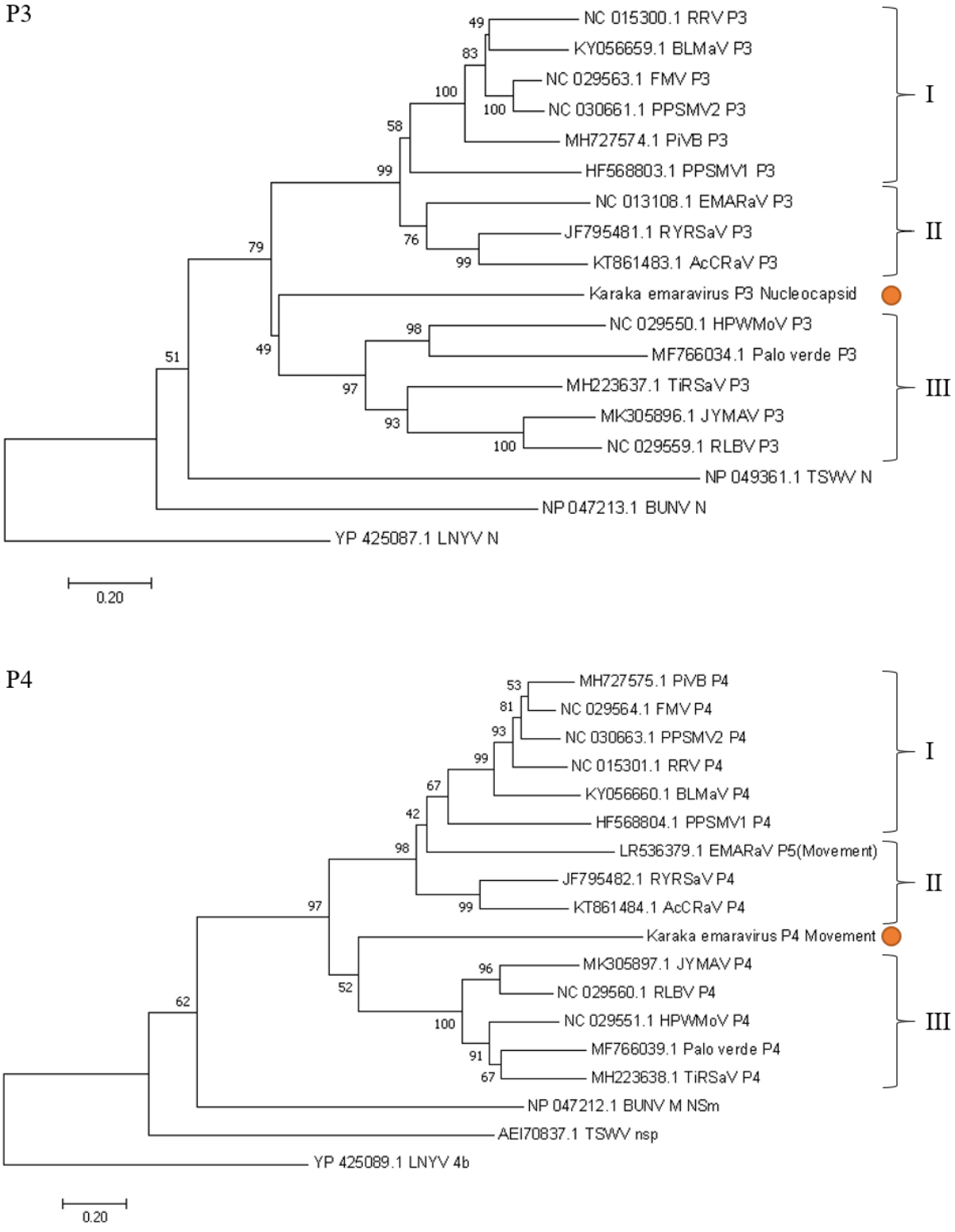


Figure 2.17 – Phylogenetic trees of the amino acid sequences of all species and candidates of the Emaravirus genus. P3 is the nucleocapsid and P4 is the movement protein. Accession numbers refer to respective records in the NCBI database. BUNV is Bunyamwera virus, an orthobunyavirus, and TSWV is tomato spotted wilt virus, a tospovirus. Both relative amino acid sequences from BUNV and TSWV were included as they belong to closely related genera to emaraviruses. LNYV is lettuce necrotic yellows virus, a cytorhabdovirus, and is used as the outgroup. Groups I, II and III represent clades previously reported by Wang et al. (2020).

In all phylogenetic trees KOPV forms its own clade outside of the previously reported clades (I, II and III) by Wang et al. (2020). This is especially evident for P1, the most conserved amino acid sequence in emaraviruses, where KOPV P1 belongs to a different lineage than all other emaraviruses, but KOPV P1 inclusion with all emaraviruses is well supported with a bootstrap value of 100%. KOPV P2 belongs to a different lineage than all other emaraviruses, but KOPV P2 inclusion with all emaraviruses not as well supported as P1 with a bootstrap value of 57%. In P3 and P4, it appears that KOPV is more closely related to viruses in clade; however, their inclusion in this clade is not well supported (with a bootstrap value of 49% and 52%, respectively). In all trees, the KOPV sequence is more closely related to the emaraviruses than it is to the orthobunyavirus and tospovirus representatives. These results reflect what has been previously reported for the JYMaV and AcCRaV phylogenetic analyses, and in the latest emaravirus to be described, AcEV-2 (Wang et al., 2020; Yang et al., 2019; Zheng et al., 2017). These phylogenetic trees show that KOPV is related to other emaravirids when compared to its relationship with the orthobunyavirus and tospovirus.

There does not appear to be any correlation between the emaravirus clades and the virus' host plants and mite vectors. Table 2.10 shows each virus listed according to their clade (Wang et al., 2020) with their host plant family and genus, whether the host is a monocot or dicot, and the mite vector. Viruses infecting plants belonging to both Fabaceae and Rosaceae are present in all three clades whereas viruses infecting monocot hosts are only present in clade C, together with dicot infecting viruses. All mite vectors belong to the Eriophyidae family, but there is no correlation with mite species.

Table 2.10 – Known emaravirus species and candidates grouped according to the clades (I, II and III) described by Wang et al. (2020). Host family, genus and whether the host is a dicot or monocot is listed. The mite vector species and family are listed (if known). All vectors come from the Eriophyidae family. *Not confirmed.

Clade	Virus Species	Host Family	Host Genus	Monocot/ Dicot	Vector	Vector Family
I	AeEV-2	Actinidiaceae	*Actinidia*	Dicot	-	-
	FMV	Moraceae	*Ficus*	Dicot	*Aceria ficus*	Eriophyidae
	PPSMV2	Fabaceae	*Cajanus*	Dicot	*Aceria cajani*	Eriophyidae
	PiVB	Anacardiaceae	*Pistacia*	Dicot	-	-
	RRV	Rosaceae	*Rosa*	Dicot	*Phyllocoptes fructiphilus*	Eriophyidae
	BLMaV	Rosaceae	*Rubus*	Dicot	-	-
	PPSMV1	Fabaceae	*Cajanus*	Dicot	*Aceria cajani*	Eriophyidae
II	EMARaV	Rosaceae	*Sorbus*	Dicot	*Phytoptus pyri**	Eriophyidae
	RYRSaV	Fabaceae	*Cercis*	Dicot	-	-
	AcCRaV	Actinidiaceae	*Actinidia*	Dicot	-	-
III	Palo Verde	Fabaceae	*Cercidium*	Dicot	-	-
	HPWMoV	Poaceae Poaceae	*Zea* *Triticum*	Monocot Monocot	*Aceria tosichella*	Eriophyidae
	TiRSaV	Asparagaceae	*Cordyline*	Monocot	-	-
	JYMaV	Rhamnaceae	*Ziziphus*	Dicot	-	-
	RLBV	Rosaceae	*Rubus*	Dicot	*Phyllocoptes gracilis*	Eriophyidae
-	KOPV	Corynocarpaceae	*Corynocarpus*	Dicot	*Aculus corynocarpi*	Eriophyidae

2.5 Discussion

The aims of the work described here was to complete the karaka emaravirus genome, characterise it and assess its relatedness with other emaraviruses. The completed KOPV genome sequence shares the fundamental prerequisites of membership to the emaravirus genus. It has at least five segmented, negative-sense RNA segments of which one to four appear to encode an RdRp, a glycoprotein, a nucleocapsid and a movement protein, respectively, that share motifs and features of other reported emaraviruses. The termini of each RNA segment share homology with each other segment in the KOPV genome, and with other emaravirids. It is, however, more genetically distinct than other emaraviruses are within the genus, with relatively low pairwise identity homology and a distant relationship when analysing the sequences in a phylogenetic tree. Regardless, the evidence of this research suggests that KOPV belongs in the emaravirus genus.

2.5.1 Completing the genome

RACE and sRNA mapping were used to complete the initial genome obtained using NGS. For RNA 3, the RACE experiment revealed that this viral genomic segment appeared to have fewer nucleotides than the initial segment generated *de novo* from the NGS data. From the original *de novo* assembly, RNA 3 was thought to be 1771 nt long; however, after RACE it was confirmed that a section of the original sequence was of poor quality and most likely not part of the RNA 3 segment. This was discovered when the product amplified from the one-step RT-PCR for the 5’ end of RNA 3; that is the 3’RACE of the positive-sense strand. The product size was predicted to be approximately 840 bp yet the product appeared to be around 650 bp long. Due to the poor quality of the sequencing data, the RNA3A and RNA3B primers were redesigned and when used in an RT-PCR, the primer pairs showed the same discrepancy between the expected and actual product size of RNA3B. Sequencing of this second product showed the RNA segment was indeed smaller than the original assembly, and mapping of the sRNAs confirmed this; the tentative RNA 3 segment length was 1479 nt long, corresponding

to RNA3 from other emaraviruses (Di Bello et al., 2016; Mielke-Ehret & Mühlbach, 2012; Zheng et al., 2017).

The genome was completed by creating a tentatively completed genome using 3' RACE and then mapping small RNAs to the tentative genome to obtain the final genome. It was necessary to use both techniques as the initial tentative genome appeared to be incomplete, even after the RACE. In the tentatively completed genome, the final nucleotide for the end of each segment was ambiguous, likely due to the PV1/SP6 polyT tail primer having the degenerate nucleotide 'V' at the very end. This was done to ensure that the PV1 sequence bound to the interface between the genome and the added polyA tail and did not bind randomly to any region along the polyA tail. It was not known how long the polyA tail would be for each segment, and indeed the polyA lengths of the polyA tailed RNA would be of variable length based on the experimental design. The ambiguous nucleotides at the end of each segment could be explained by the PV1/SP6 primer's degenerate nucleotide. Primer solutions with a degenerate nucleotide contain three sequences, each one with a different nucleotide other than thymine at the V position (Linhart & Shamir, 2005). To avoid this problem in the future, a primer without the degenerate V at the end could be tried, although this would result in the primer binding anywhere randomly along the PolyA tail, producing sequences with different lengths. This could be troublesome for sequencing as the product would be a mixture of different length cDNA. To avoid this, the sequencing could be done only in one direction – from the gene specific primer towards the 3' end. This would result in less certainty of the sequence, however, as a consensus between both forward and reverse sequences would be absent. This method could potentially resolve the issue of the end nucleotide being ambiguous and might remove the need for mapping sRNA to obtain the genomic ends of each segment.

In other emaravirus studies where RACE was used, the process used a nested gene specific primer in the second PCR to amplify enough product for sequencing. For instance, in the GeneRacer kit (Invitrogen, USA) used by Zheng et al. (2017), the kit specifies using a nested GSP to obtain enough product for sequencing and to reduce background noise. This did not seem to be necessary for the experiment outlined in this research, as enough product was

produced that could be sequenced; however, it may be useful in the future to eliminate the variable nucleotide at the end of each sequence.

The sRNA data from the initial NGS data was mapped to the tentative genome obtained from 3' RACE to confirm the final genome sequence for all five segments. This was necessary as the consensus sequence for the ends of each segment were of low quality and, when comparing the ends to other emaravirids, appeared to be incomplete or incorrect. It has been previously reported that 13 nt from the 3' ends are complementary to the 13 nt from the 5' ends within each species and this feature is highly conserved across all known emaravirids to date (Di Bello et al., 2016; Zheng et al., 2017). An initial alignment of KOPV ends showed that the final nucleotide was not identical. Combining this alignment with sRNA mapping to each tentative genomic segment, it was apparent from the consensus that the 3' and 5' ends from the RACE were incorrect. The sequence obtained from the mapping assembly had a strong consensus of the final nucleotides from each RNA segment. These 3' and 5' ends matched to the first seven nucleotides exactly when aligned with other emaravirids, and exactly with the nucleotides from 11-13 (Shown in Figure 2.14). This evidence shows that mapping the sRNA data to each segment was necessary to verify the ends for KOPV. It is important to note that the 5' termini of PiVB, and 3' and 5' termini of palo verde RNA 1 sequences differ from the rest of the emaravirus sequences; the PiVB 3' terminus is 10 nucleotides shorter while the palo verde 5' and 3' termini are 10 and 12 nucleotides shorter, respectively. Both viruses were reported to be found using NGS, but there is no indication that RACE or any other technique was used to confirm the genomic ends (Buzkan et al., 2019; Ilyas et al., 2018). This contrasts with other recent emaravirus reports such as AcCRaV, RYRSaV, TiRSV and JYMaV that have included RACE as a part of the process to determine the full sequence (Di Bello et al., 2016; Olmedo-Velarde et al., 2019; Yang et al., 2019; Zheng et al., 2017). RACE, at the very least, appears to be necessary to obtain a completed emaravirus genome.

The sRNA mapping also revealed nucleotide differences between the tentative and final genome that were not at the 3' and 5' ends. These nucleotide differences were rare with the bulk being in RNA 1, the largest contiguous sequence generated from the NGS data. RNA 1 had seven differences, RNA 2 had one difference and RNA 5 had one difference. The final

genome was curated based on these differences and either the original NGS nucleotide was chosen, or the sRNA consensus nucleotide was chosen based on the number of reads each assembly contributed to the consensus sequence. The reads of each NGS sequence was compared to the reads of the sRNA sequence and the most frequent nucleotide was chosen. In most cases, there was a clear candidate based on the frequency of reads. For example, on RNA 1, at the position of 903 in the final sequence, the original NGS data had 49 reads of C and 5 reads of T, with the consensus being C. The sRNA data mapped had over 500 reads of T. In this case T was chosen for the final sequence as it occurred more frequently than C when comparing both NGS and sRNA data. This was the method used for the rest of the discrepancies.

Although these nucleotides were chosen, there is a possibility that both were correct and that there is variability within KOPV population. It is well established that other RNA viruses exist as quasi species and it would not be surprising if emaraviruses did also (Domingo et al., 2012; Eigen, 1993). Variability within other emaravirids has been observed when comparing different isolates (Di Bello et al., 2016; Hassan & Tzanetakis, 2019; Stewart, 2016), but no studies to date have reported variation within a single host plant. The RNA used here to gather the initial NGS data (including the sRNA data) came from a single host plant leaf, thus any variation within the sequence would possibly have been present within that one host, suggesting a quasispecies model of population structure. It is possible that this variation may have arisen during the initial RT-PCR steps. It is known that reverse transcriptase can introduce errors (Potapov et al., 2018). Further, the Platinum Taq used in the RT-PCR is not a proof-reading polymerase. The error rate for *Taq* polymerase is 2.28×10^{-5} errors/bp/doubling (Retrieved from https://www.thermofisher.com/nz/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/pcr-fidelity-calculator.html). It is expected that no further errors would have been introduced during the second round of PCR since a proof-reading polymerase was used. This suggests for a 350bp PCR product amplified using 35 cycles, 28% of the products generated would be expected to have an error. Therefore, having multiple reads of each nucleotide was important for completing the genome with confidence. To determine if the variation in the sequence was due to technical artefacts, or reflected true biological variability, a proof-reading polymerase should be used for all PCRs. To assess the extent of the biological variability within the Karaka emaravirus, RNA could be gathered from

different symptomatic leaves on the same tree, the genome sequenced and compared with each RNA segment described in this research. Further experiments could involve comparing genome sequences from geographically distinct symptomatic leaves to understand the extent of the variability within the species. Information gathered from this could further inform the process in designing a diagnostic test to detect the virus.

Mapping the sRNA to the tentative genome obtained through RACE gave the final genome and the question was raised if the RACE experiment was necessary and if the sRNA data could have just been mapped to the initial NGS consensus sequences. The sRNA sequences were mapped to the initial NGS genomic segments and showed that the sRNA consensus added on average an additional nucleotide to the end of each segment than what was expected, based on other emaravirus genomes. It appears that without the additional segment information obtained by RACE, the sRNA mapping was unreliable and so, in this case, both methods were required to be used. The sRNA data was crucial in confirming KOPV genomic ends for each segment; without it, re-design of the PV1/SP6 primer and sequencing the subsequent PCR product in only the gene specific primer to polyA tail direction, as mentioned above, would have needed to be done. An alternative could be using a commercial kit, as specified in several emaravirus reports, such as for AcCRaV, as they appear to be accurate when sequencing the products generated by these kits (Zheng et al., 2017). The main difference between these commercial kits and the process described here is the addition of a nested gene specific primer in the second PCR step to reduce noise and increase accuracy. Another reason why the sRNA assemblies should not be used on their own was that the sRNA that was derived from viral RNA may not give a strong, even consensus along the reference segments due to “hotspots” where sRNA is more abundant along certain areas of the viral genome (Mitter et al., 2013; Zheng et al., 2017).

2.5.2 Sequence Analysis and Comparison with other Emaravirus species.

The completed negative-sense genome found in karaka appeared to belong to the emaravirus genus. NCBI BLAST searches returned similarities to other emaraviruses for the amino acid sequences of translated ORFs for RNAs 1 to 4. When using BLAST with the nucleotide sequences, no matches were found, therefore subsequent analysis used the amino acid sequences when comparing KOPV to the other emaravirids. Amino acid pairwise identities between KOPV and other emaraviruses were low when comparing the other emaraviruses between themselves. The ORF encoded by KOPV RNA 1 had the highest sequence identity to other emaraviruses with the closest being to EMARaV. This was to be expected as RNA 1 encodes the RdRp, which requires high conservation to maintain functionality, as described for FMV (Elbeaino et al., 2009). The ORFs for RNA 2, 3 and 4 had relatively low pairwise identity when compared with the identities between other emaraviruses. RNA 5 appears to encode a protein that has no similarity to any other known virus, which is not uncommon for other emaraviruses (Di Bello et al., 2016; Zheng et al., 2017). Considering that the demarcation criteria for a new emaravirids is stated as greater than 25% sequence similarity, KOPV is a strong candidate as a new emaravirus species (Elbeaino et al., 2018).

Features common to the other emaravirids are present in KOPV. The genome is a segmented, negative-sense RNA genome that has five RNA segments which encode a single ORF each. The segments have 3’ and 5’ termini that are complementary to each other to 13 nt within the single segment within each species, with the first seven nucleotides being exactly complementary between other emaraviruses (Mielke-Ehret & Mühlbach, 2012). The individual proteins encoded by the ORFs of each segment also share features and conserved motifs with other emaraviruses, strongly suggesting that the virus belongs in the genus. The ORFs encoded by RNA 1 has premotif A and motifs A-E present, as well as the endonuclease domain required in emaraviruses and other bunyavirus RdRps (Elbeaino et al., 2009; Mielke & Muehlbach, 2007). The ORF encoded by RNA 3 also has the three expected motifs present in other emaraviruses and the pairwise identity homology is consistent when compared to other emaraviruses.

An important difference between KOPV and other emaraviruses is in P2, the glycoprotein precursor encoded by RNA 2, where KOPV is missing the predicted N-glycosylation sites as well as one cleavage site. These N-glycosylation sites are present in all other glycoprotein precursors and there are usually between four and six sites (Di Bello et al., 2016; Elbeaino et al., 2009; Hassan et al., 2017). These other proteins also have two cleavage sites that are predicted to cleave the precursor protein into the peptides that make up the final glycoprotein form (Elbeaino et al., 2009). KOPV is predicted to have other expected features of P2; however, including five transmembrane helices, one cleavage site and one signal peptide predicted to be involved in the secretory pathway. The fact that some of the expected features are present indicates that the final sequence obtained for this genomic RNA segment may not be accurate. The initial *de novo* assembly from the NGS data had a low number of reads and a weak consensus with RNA 2. This was also shown with the sRNA data with only 85,135 of 26,546,769 reads mapped to the tentative genome, the lowest number of reads of all the RNA segments. This may have influenced the quality of the sequence and could explain the differences between KOPV RNA 2 sequences and other emaravirus RNA 2 sequences, and may also explain why the nucleotide length of the RNA and amino acid length of the protein are substantially shorter. Considering that the sequence was missing potentially vital components required for functionality, it could be that the sequence is incorrect. Repeating the NGS process on another symptomatic karaka RNA sample to get another genome to compare with this final genome could reveal if the sequence is correct.

Another important difference between KOPV and other emaraviruses is in P4, the movement protein encoded by the ORF in RNA 4. In other emaraviruses the amino acid sequence for P4 contains a conserved domain from the emaravirus P4 superfamily (pfam16505) involved in movement (Di Bello et al., 2016; Zheng et al., 2017). In contrast, the KOPV P4 amino acid sequence appears to have no such domain, which may explain why the P4 of KOPV only shows similarity with palo verde broom virus Interestingly, the palo verde broom virus P4 sequence (NCBI accession MF766043) does contain the emaravirus p4 superfamily domain. As with P2, the absence of a conserved domain may indicate an error in the sequencing and further NGS analysis would be ideal to determine if that is the case. If the sequence is correct, it could mean that the movement protein may have a diminished functional

capacity and that the virus is restricted from moving through the host plant, instead relying on the mite vector to move the virus. This hypothesis is tested and explored in Chapter 4.

The unknown protein encoded by the ORF in KOPV RNA 5 appears to have no similarity to any other virus protein, and therefore its function is unknown. The closest potential orthologue among other emaraviruses is the amino acid sequence of PPSMV2 P6 where KOPV has a 17.6% pairwise identity match. Another potential orthologue is HPWMoV P7, which is of a similar length and has a pairwise identity of 16.7%. There is limited information about the additional proteins encoded for by each emaravirus. For HPWMoV, there is evidence that P7 and P8 both encode proteins involved in RNAi silencing suppression (Gupta et al., 2018; Gupta et al., 2019). RNAi silencing is a plant defence mechanism against foreign nucleic acids and silencing suppression is the viral counter measure against this defence (Kormelink et al., 2011; Pumplin & Voinnet, 2013). The KOPV P5 may be involved in silencing suppression based on these studies, although the evidence is weak. A study to emulate the research by Gupta et al. (2018) by using the green fluorescent protein reporter gene in combination with the KOPV P5 nucleotide sequence in an agrobacteria infiltration assay could be used to determine if the protein has silencing suppression functions. Overall, the sequence analysis supports KOPV belonging to the emaravirus genus.

2.5.3 Phylogeny

Although KOPV appears to sit outside the three clades reported by (Wang et al., 2020), the phylogenetic tree for the proteins encoded by RNA 1, 2, 3 and 4 all show that it shares a common ancestor with the other emaraviruses. RNA 1 and P1 provide the most reliable evidence for the relatedness between KPV and other emaraviruses. These analyses suggest that KOPV has had an independent evolutionary path to the other emaraviruses. Phylogenetic analysis of the other proteins suggested the position of KOPV was not well supported. This may be due to the amino acid sequences of P3 and P4 within the emaravirus genus being more diverse than the amino acid sequences for P1, and to P2 not showing all the hallmarks of a predicted emaravirus glycoprotein. Rapid discovery of new emaraviruses could “fill the gap”

between the existing emaraviruses and KOPV, further revealing the place that it holds within the genus.

The KOPV genome is complete, and evidence suggests that it belongs in the emaravirus genus.

Chapter Three: Creating a Diagnostic Test to Detect the Karaka Emaravirus and Determining the Distribution in Auckland and New Zealand.

3.1 Introduction

As the novel karaka emaravirus is new to science, to determine the distribution and abundance of the virus in New Zealand work needed to be done. To do this, a reliable, sensitive, cost effective diagnostic test needed to be developed that could be applied to many symptomatic and asymptomatic samples gathered from around New Zealand.

3.1.1 Diagnostic Test

For this project, RT-PCR was chosen as the diagnostic test for several reasons. Firstly, primers had already been developed by Dr Arnaud Blouin to detect each individual RNA segment present in the novel karaka emaravirus genome but had not been tested. This provided a starting point for the project testing the primers on RNA extracted from symptomatic karaka leaves. If the primers amplified a product then they could be used in a diagnostic test, or else they would need to -be redesigned. RT-PCR has been shown to be the main diagnostic tool to detect emaraviruses within the literature, for example when detecting RYRSaV and AcCRaV (Di Bello et al., 2016; Zheng et al., 2017). Sensitivity, reliability, and the accuracy to which the test discriminated between virus-infected and virus-uninfected plants were important factors in deciding to use RT-PCR. These factors are important due to not knowing whether the pale circle symptoms were being caused by the virus or some other factor. Once the diagnostic test was developed, many samples were to be collected and tested to determine if, and to what extent, the virus is present in symptomatic and asymptomatic samples. Extracting RNA and running an RT-PCR followed by gel electrophoresis is laborious with many samples, but this is a worthy trade-off to obtain the sensitivity and accuracy required.

3.1.2 Surveying Trees

Surveying is important when studying plant viruses as it will give information about the spread and possible point of origin of the virus. There are many ways to survey virus symptoms present in trees in an urban landscape depending on what information is to be determined. One way is to focus on the distribution, to determine whether a virus is present in an area, another is to focus on abundance, which is to determine how many symptomatic plants are present in an area. The aim of this project focused on the distribution of symptomatic karaka. Distribution can be determined by creating a map of an area and dividing it into discrete grids, then surveying each grid area to detect presence of the virus. It can also be done by gathering data of virus reports in an area (such as a city or country) and plotting those locations on a map to visualise. Symptomatic and asymptomatic trees can be distinguished by using different markers or colours on the map to create a visual of where the virus may be present. Software such as the QGIS (QGIS, 2019) can be used to create a map and to plot the survey points. An example of this is shown below in Figure 3.1, where GPS coordinates of free-living grapevine in California have been plotted on a map with different colours depicting different infection states (Cieniewicz et al., 2018). The map clearly shows where two viruses are present within the range of the survey.

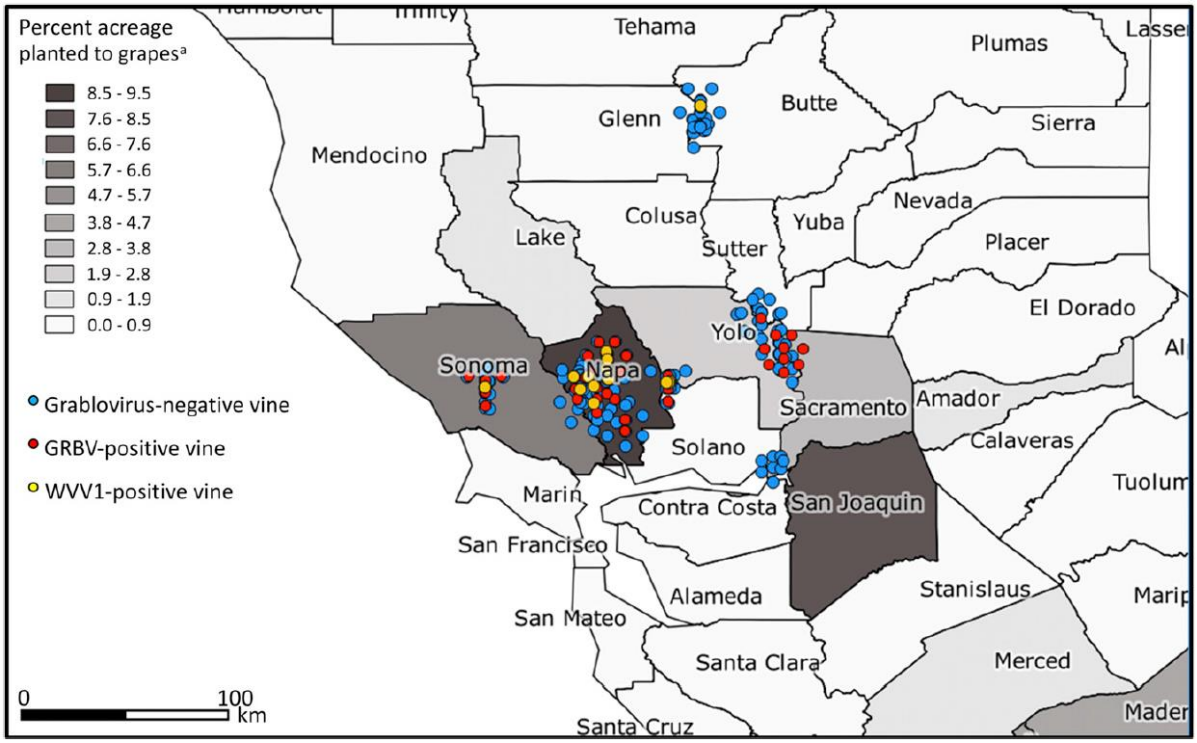


Figure 3.1 - Map of California showing free-living grapevines and grabloviruses that have been detected (Cieniewicz et al., 2018). GRBV is grapevine red blotch virus, WVV1 is wild Vitis virus 1. This is an example of how virus survey data can be represented using QGIS (QGIS, 2019).

Surveying on foot alone results in a limited yield of data. Sampling instructions that could be distributed to other scientists and the members of the public asking them to gather samples is a potential way to get more samples. The locations of surveyed trees would show a potential bias towards trees that had easier access for humans although surveying would still help with gaining an initial understanding of how the virus is distributed. Sampling instructions that allowed people with no scientific background to identify potential symptomatic plants would be beneficial. Instructions would also need to ensure that samples that were sent from around the country could be sent to the laboratory and be able to detect the virus, if present, reliably.

3.1.3 Aims

To complete the aim of determining the distribution of the virus, the objectives of this chapter were twofold. The first was to develop a diagnostic test for KOPV that could accurately detect the virus within karaka plant tissue using RT-PCR. The diagnostic test also needed to discriminate between virus infected and virus uninfected tissue. Once the diagnostic test was developed, it was used on a small number of samples collected from around Auckland to check that it could detect virus in other trees. The RNA extracted from these samples were tested using the diagnostic test.

The second objective was to determine how widely distributed KOPV is within Auckland, and to develop a strategy for the future to determine the distribution within New Zealand. This was done to identify a possible emergence point for KOPV in Auckland to surrounding areas. Trees were surveyed for symptoms from around central Auckland and the upper North Island and samples were gathered for testing using the diagnostic test. To set the groundwork for future survey work being done by people from around New Zealand, a sampling protocol was developed to ensure sample fidelity when being posted to Auckland. Monthly sampling of the PFR0001 tree at Plant and Food Research was also done to determine if there was a best time to sample throughout the year.

3.2 Methods and Materials

3.2.1 Plant Materials

Symptomatic and asymptomatic leaf material was required to obtain RNA for use in developing the diagnostic test. Initial development of the diagnostic test used plant materials gathered onsite from tree PFR0001 at Plant and Food Research, Mount Albert Research Campus, Auckland, New Zealand. Further plant material from survey and sampling, or plant material supplied from around the country was used.

RNA needed to be extracted from symptomatic and asymptomatic leaves that were free from PCR inhibitors and could amplify a product in an RT-PCR reaction. Prior to RNA extraction, the leaves were washed with 70% ethanol and photographs were taken of samples to show what portion of the leaf was excised from the leaf sample. If a sample was symptomatic, a symptomatic section of leaf was always taken. A photograph of the leaf was taken before tissue was excised from each leaf using a sterile razor blade. The leaf was then photographed afterwards to show the area that was taken, as the example shown in Figure 3.2 below.

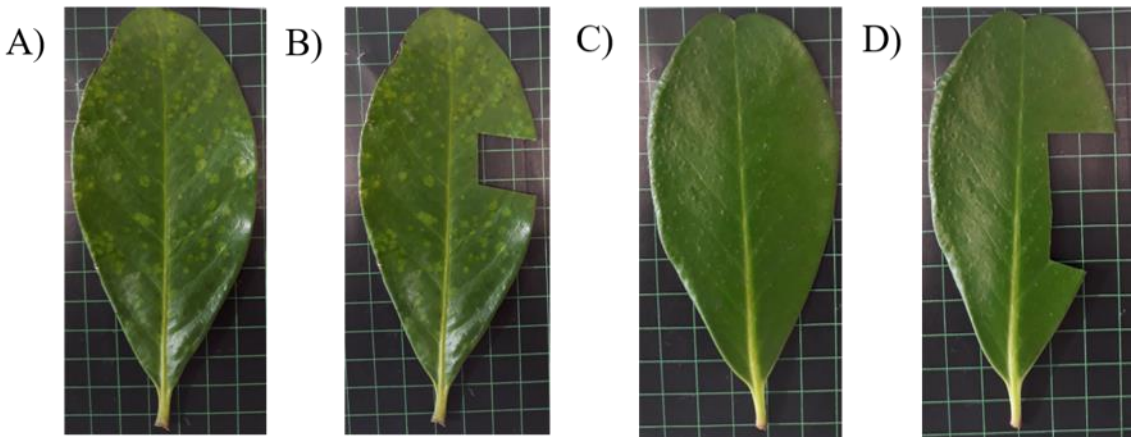


Figure 3.2 - Karaka leaf tissue shown before (A and C) and after excision (B and D) of material used in an RNA extraction. Note that the area of leaf tissue taken is different between B) and D); however, the weight of tissue taken was similar.

3.2.2 RNA Extraction Methods

Two RNA extraction methods were compared against each other to determine what yield, concentration and quality of RNA could be extracted from Karaka samples.

3.2.2.1 RNA Extraction –CTAB

See section 2.2.2 for a description of the method used for RNA extraction using CTAB buffer and LiCl precipitation.

3.2.2.2 RNA Extraction – Spectrum™ Plant Total RNA kit

To determine the best way to extract RNA, another method of RNA extraction was used to compare to the CTAB process. Total RNA was extracted from fresh leaf samples, or leaf samples stored at 4°C, using the Spectrum™ Plant Total RNA extraction kit (Sigma Aldrich, Missouri, USA). The manufacturer's instructions were used with starting leaf material weighing approximately 150 mg. The concentration of each sample was then determined using a Nanodrop 2000 (ThermoFisher Scientific, USA). Ideal RNA concentrations were generally between 150 and 400 ng/uL with a A_{260}/A_{280} absorbance ratio of between 1.9 and 2.0.

3.2.2.3 RNA Concentration Standardisation

The extraction methods were to be compared by determining the yields of each method then standardising them to approximately the same RNA concentration. RNA was standardised to 75 ng/µL in 15 µL volumes before being used in any following test, unless specified in a specific method. The RNA concentration from a nanodrop machine was used in the following equation to get the amount of RNA needed: 75.0 ng/µL x 15 µL / (RNA Concentration). This volume of RNA was pipetted into a 1.7 mL plastic microtube and ddH_2O

was added to make up the rest of the 15 µL to be tested using RT-PCR. If the initial RNA concentration was below 75 ng/µL, then the RNA was not standardised, and the initial RNA concentration was used in the following RT-PCR tests.

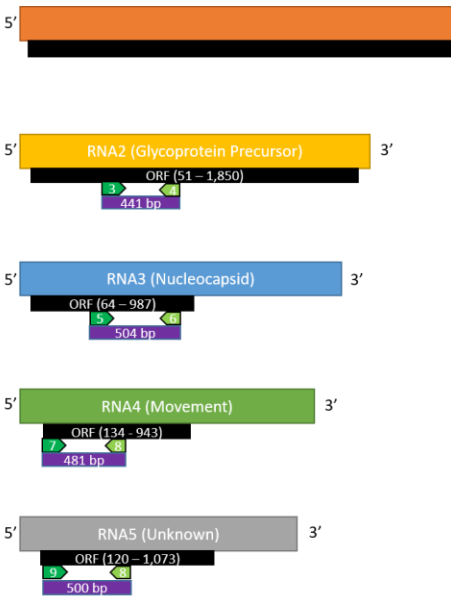
3.2.3 Reverse-Transcription Polymerase Chain Reaction (RT-PCR)

3.2.3.1 Karaka Emaravirus Individual Segment Detection of RNA 1-5

Individual segment detection was carried out using one-step RT-PCR. A PCR product was amplified by adding 1 µL of RNA standardised to 75 ng/µL into a reaction with 0.4 µL each of 10 µM forward and reverse primer specific to each segment as shown in Table 3.1 below. These primers were developed by Dr Arnaud Blouin (unpublished). This was mixed with 5 µL of 2X Reaction mix, 0.2 µL of Superscript™ III RT/Platinum *Taq* Mix and 3 µL of ddH_2O. The RT-PCR conditions were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 2 minutes at 94°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 94°C, annealing for 30 seconds at 61°C, extension for 30 seconds at 68°C, followed by a final extension for 3 minutes at 68°C. The amplified product was held at 15°C after the RT-PCR had completed and then run immediately on an agarose gel or stored at -20°C until needed. Any larger volume reactions that were required used the same proportions of reagents as described here. Primer bind sites are depicted in Figure 3.3 below.

Table 3.1 – Primers used to detect karaka virus (KOPV) RNAs and karaka NAD5 transcripts. Primers targeted KOPV individual segment primers for RT-PCR detection. These primers were designed by Dr Arnaud Blouin at Plant and Food Research and are unpublished. NAD5 primers targeted the NADH dehydrogenase subunit 5 sequence in grapevine (*Vitis vinifer*a). Note that the expected product size is approximate as the primers were designed for *V. vinifera*, not karaka (*Corynocarpus laevigatus*). Generic primers were designed to target a conserved sequence in any emaravirus species. W represents A or T, R represents A or G, H represents A or C or T, Y represents C or T, I represent any base. *RNA4 and RNA5 primers were initially mislabelled, the RNA4 primers target the RNA5 sequence and the RNA5 primers target the RNA4 sequence from the initial NGS data obtained by Arnaud Blouin.

Target Segment	Primer Name	Nucleotide sequence 5' → 3'	Expected product size (bp)	Uniplex (U) or Multiplex (M)?	Reference
KOPV Primers					
RNA 1	KOPV RNA1 4669F	AGA AAT CAA CAG CTC ATC CAG AG	500	U, M	Arnaud Blouin (Unpublished)
	KOPV RNA1 5168R	ATC AAC AGA GCT CAG ATG ACC TG			
RNA 2	KOPV RNA2 482F	ACT AAG AGT TTG TCA TAT TTG TAA TCT ACC	441	U, M	Arnaud Blouin (Unpublished)
	KOPV RNA2 922R	AGT TGA ACA AGA GCA CAT GAC CAC			
RNA 3	KOPV RNA3 390F	AAC TGG GTT CCT TTA CTG GTG AC	504	U, M	Arnaud Blouin (Unpublished)
	KOPV RNA3 863R	CTG AGA GCA TGA AGT CAA CAG CA			
RNA 4	KOPV RNA4 169F *	GCA GAT TGC CTT ATT GCT TAT CAC C	481	U, M	Arnaud Blouin (Unpublished)
	KOPV RNA4 649R *	TCA AGA TTG TGA TAT GAG GAT GGC A			
RNA 5	KOPV RNA5 110F *	ATG ATT CTG TTA GGT CTG GTT GCT	500	U, M	Arnaud Blouin (Unpublished)
	KOPV RNA5 609R *	AAT GAA AGC CTA CCC ATA CTT CCC			
NAD5 Primers					
NAD5	Vv-NAD5-F	GCT TCT TGG GGC TTC TTG TTC G	220	U	(Chooi et al., 2013)
	Vv-NAD5-R	TCC CTC CCA TCC CAG GAA TAA TTG			
Generic Primers					
RNA 1	G13-1AF	GAT GCA TCD AAA TGG TCW GC	360	U	Elbeaino et al (2013)
	G13-1CR	ATC ATC WGA RTG HAC CAT			
RNA 1	G16-F&R-7	TCT TGT GGT GAT CCA TGI ARR CCY TTA TTW CC	302	U	Olmedo-Velarde et al (2016)
	G16-F&R-8	CCG CGC AGA TAA TCT TAT ARA IGA YAA RYT RGA AT			



Code	Primer	Sequence
1	KTPV-RNA1 F (4669F)	AGAAATCAACAGCTCATCCAGAG
2	KTPV-RNA1 R (5168R)	ATCAACAGAGCTCAGATGACCTG
3	KTPV-RNA2 F (482F)	ACTAAGAGTTTGTCATATTTGTAATCTACC
4	KTPV-RNA2 R (922R)	AGTTGAACAAGAGCACATGACCAC
5	KTPV-RNA3 F (395F)	AACTGGGTTCCTTTACTGGTGAC
6	KTPV-RNA3 R (898R)	CTGAGAGCATGAAGTCAACAGCA
7	KTPV-RNA4 F (129F)	GCAGATTGCCTTATTGCTTATCACC
8	KTPV-RNA4 R (609R)	TCAAGATTGTGATATGAGGATGGCA
9	KTPV-RNA5 F (132F)	ATGATTCTGTTAGGTCTGGTTGCT
10	KTPV-RNA5 R (631R)	AATGAAAGCCTACCCATACTTCCC

Figure 3.3 – Binding sites for KOPV RNA 1 - 5. Each primer binds to a section on the open reading frame (ORF) of each RNA. Note: RNA1 is not to scale with the rest of the RNA segments.

3.2.3.2 Karaka emaravirus RNA 1 -5 Multiplex Detection

Multiplex detection reactions were carried out using one-step RT-PCR. RNA template standardised to 75 ng/µL at 1 µL was added to a mix containing 0.38 µL of each RNA 1-5 primer listed in Table 3.1, 5 µL of 2X Reaction Mix and 0.2 µL of Superscript™ III RT/Platinum *Taq* Mix. The RT-PCR conditions were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 2 minutes at 94°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 94°C, annealing for 30 seconds at 61°C, extension for 30 seconds at 68°C, followed by a final extension for 3 minutes at 68°C. The amplified product was held at 15°C after the RT-PCR had completed and then run immediately on an agarose gel or stored at -20°C until needed. Any larger volume reactions required used the same proportions of reagents as described here.

3.2.3.3 Karaka Host Plant NAD5 Transcript Detection

Primers developed to detect the karaka host's NAD5 transcripts were used as a quality control to determine if the RNA was amplifiable. The NAD5 transcript detection using RT-PCR was initially done as a two-step reaction using the *Vitis vinifera* primers in Table 3.1 to test if they were suitable for use with karaka RNA. NAD5 is a constitutive gene regularly used as a housekeeping gene in molecular biological experiments. The NAD5 primers used in the chapter were designed by Karmun Chooi (The Institute for Plant and Food Research, MARC) for use with *V. vinifera* (grapevine). For this test 1 µL of RNA standardised to 75 ng/µL was mixed with 0.5 µL of 10 µM dNTPs, 0.2 µL of 10 µM reverse primer (Vv-NAD5-R) and 4.8 µL of ddH_2O. The mixture was incubated for 5 minutes at 65°C then transferred onto ice for 1 minute. It was then added to a mixture of 2 µL of 5X Reaction Mix, 0.5 µL of 0.01 µM DTT, 0.5 µL of 2 mM RNaseOUT and 0.5 µL of 10 µM Superscript™ III RT (ThermoFisher Scientific, USA). The cDNA was synthesised for 1 hour at 55°C, followed by 15 minutes at 70°C and then held at 15°C until the next step. For amplification, 1 µL of the cDNA synthesised in the first step was added to 3 µL of ddH_2O, 5 µL of CloneAmp HiFi PCR Premix (Takara Bio USA, Inc.), 0.5 µL of 10 mM Vv-NAD5-F primer and 0.5 µL of 10 mM Vv-NAD5-R primer. The PCR conditions were 1 cycle of initial denaturation for 2 minutes at 98°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 94°C, annealing for 10 seconds at 60°C, extension for 10 seconds at 72°C, followed by a final extension for 5 minutes at 72°C. The amplified product was held at 15°C after the PCR had completed and then run immediately on an agarose gel or stored at -20°C until needed.

The RT-PCR for detecting NAD5 mRNA was optimised for one-step RT-PCR using the *V. vinifera* primers in Table 3.1. The RNA template was standardised to 75 ng/µL and 1 µL was mixed with 0.4 µL of 10 µM Vv-NAD5-F, 0.4 µL of 10 µM Vv-NAD5-R, 5 µL of 2X Reaction Mix and 0.2 µL of Superscript™ III RT/Platinum *Taq* Mix (ThermoFisher Scientific, USA). The RT-PCR conditions were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 2 minutes at 98°C, 35 cycles of PCR amplification with denaturation for 10 seconds at 94°C, annealing for 10 seconds at 60°C, extension for 10 seconds at 72°C, followed by a final extension for 5 minutes at 72°C. The

amplified product was held at 15°C after the RT-PCR had completed and then run immediately on an agarose gel or stored at -20°C until needed.

It is important to note that the one-step RT-PCR for NAD5 only amplified using the Eppendorf Mastercycler® Gradient machine.

3.2.3.4 Agarose Gel Electrophoresis

RT-PCR products were electrophoresed through an agarose gel as described in section 2.2.4.5.

3.2.3.5 Sequencing of PCR Product.

RT-PCR products were sequenced to ensure that the tests were targeting the correct sequence in the viral genome and the karaka NAD5 transcript. Following a 30 µL RT-PCR using appropriate primers, 5 µL of the products from the RNA 1-5 individual segment detection and NAD5 transcript detection methods were electrophoresed on an agarose gel (refer to 2.2.4.5) to confirm amplification. The remaining 25 µL of RT-PCR product was then purified by removing primer dimer and other contaminants using the Invitrogen PureLink PCR Purification Kit (ThermoFisher Scientific, USA) as per the manufacturer's instructions. The concentration of the purified product was checked using a Nanodrop machine. The purified products were sequenced by Sanger sequencing by Macrogen Inc., South Korea). The relevant forward and reverse primers for each product were used as sequencing primers. Sequence data was returned by email and analysed using the Geneious 6.1.8 software (https://www.geneious.com). Sequences were trimmed to remove low quality data and the forward and reverse reactions were aligned to create a consensus sequence for each product. These sequences were translated to amino acid sequences using Geneious and were use as a query sequence in BLASTP searches on the NCBI website (https://blast.ncbi.nlm.nih.gov/) for similarity to other sequences. The karaka RNA 1-5 sequences were aligned to the genome

sequence of the karaka emaravirus to ensure that the respective sequence amplified was the same as the original karaka emaravirus sequence.

3.2.4 Comparison of RNA Extraction Methods

Given that many leaf samples were to be extracted, a comparison was carried out between the CTAB method and the method based on the Spectrum Total RNA kit. Parameters considered were sensitivity of virus detection, throughput, yield, and cost-effectiveness.

3.2.4.1 RNA Extraction

Four fresh leaf samples were taken from Karaka and Kawakawa (*Piper excelsum*) leaves on the Plant and Food Mt. Albert Research Campus. The first leaf was symptomatic and taken from Tree 1, the second leaf was an asymptomatic leaf taken from the same tree. The third leaf was taken from a tree that had no observable symptoms. The fourth leaf was taken from a Kawakawa plant with no observable symptoms to be used as a negative control. One 200 mg sample was taken from each leaf and used in a CTAB extraction, and then another sample of the same size was taken and used in a Spectrum extraction. The RNA concentration was calculated and standardised using the methods above.

3.2.4.2 RNA Dilution

The RNA from both extraction methods was standardised to 75 ng/µL and then diluted in a ten-fold dilution series to concentrations of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5}. These diluted RNA samples were then tested using the individual segment primers RNA1, RNA2, RNA3, RNA4 and RNA5 from Table 3.1 above.

3.2.4.3 RT-PCR

The diluted RNA was tested using the individual primer segments using RT-PCR as described above and using the same conditions in section 3.2.3.1.

3.2.5 Testing Generic Primers

Generic primers developed to detect all emaravirids were tested on karaka RNA to see if they amplified the expected product. Primers developed by Elbeaino et al (2013) and Olmedo-Velarde et al (2016), as shown in Table 3.1 were used. Firstly, each primer pair was aligned to their respective RNA segment using the completed genome data described in Chapter 2. The alignment was carried out using ClustalW in Geneious 6.1.8. Both primer sets were aligned to the KOPV RNA 1 sequence, the RdRp, as described by the authors.

The primers developed by Elbeaino et al (2013) were tested using one-step RT-PCR using the method outlined in section 3.2.3.1. The RT-PCR conditions as described in the literature were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 5 minutes at 94°C, 35 cycles of PCR amplification with denaturation for 30 seconds at 94°C, annealing for 50 seconds at 48°C, extension for 40 seconds at 72°C, followed by a final extension for 7 minutes at 72°C. The expected product size for a positive reaction is around 360 bp.

The primers developed by Olmedo Velarde et al (2016) were tested using one-step RT-PCR. The RT-PCR conditions were adapted from the literature to work in a one-step reaction. The conditions were 1 cycle of reverse transcription for cDNA synthesis for 30 minutes at 50°C, 1 cycle of initial denaturation for 2 minutes at 97°C, 35 cycles of PCR amplification with denaturation for 20 seconds at 95°C, annealing for 30 seconds at 45°C, extension for 30 seconds at 72°C, followed by a final extension for 5 minutes at 72°C. The expected product size for a positive reaction is around 304 bp.

Both RT-PCR reactions were held at 15°C when completed and transferred to a 4°C refrigerator until they were electrophoresed as described in section 2.2.4.5.

3.2.6 Auckland Tree Survey and Leaf Sampling

To determine the distribution of KOPV around Auckland, symptomatic and asymptomatic karaka trees were surveyed initially where they could be found on foot, or by travelling around in a car and recognising the trees by sight. Karaka trees were photographed close-up showing the detail of symptomatic leaves (or asymptomatic leaves if no symptoms were present). A second photograph was taken further away that showed the entire tree and in the context of its surroundings so that it could be identified later. The photographs were taken using the "GPS Camera" application on an android phone and recorded the approximate GPS co-ordinates of the trees based on the phone's GPS data (GPS Camera for Android, Google Play Store). Later, surveying around Auckland was done by creating a map of Auckland with a 2 km^2 grid overlay in QGIS (QGIS, 2019). This approach focused on distribution rather than abundance of karaka in Auckland. GPS co-ordinates of trees already surveyed were uploaded and layered onto the map and grid areas where no trees had been surveyed in central Auckland were identified. Surveying efforts then concentrated on those areas to get a more even distribution of Karaka in central Auckland.

Auckland samples were gathered from symptomatic and asymptomatic karaka trees as they were being surveyed. From symptomatic trees, a symptomatic leaf and an asymptomatic leaf were sampled. From asymptomatic trees, only an asymptomatic leaf was sampled. Each sample was placed in a ziplocked bag and stored at ambient temperature until the end of the day where it was transferred to a 4°C refrigerator. As samples were gathered from symptomatic and asymptomatic trees, as well as symptomatic and asymptomatic leaves being sampled from symptomatic trees, a symptom labelling system was created. Samples were labelled according to whether the tree was symptomatic (+/) or asymptomatic (-/), and then whether the leaf sampled was symptomatic (/+) or asymptomatic (/-). For example, a symptomatic tree was

sampled, and a symptomatic leaf was taken then labelled as +/+, whereas an asymptomatic leaf from the same tree was labelled as +/-. An asymptomatic tree with an asymptomatic leaf sample taken was labelled as -/-.

3.2.7 Testing Sampled Trees

The diagnostic test developed as part of this work was applied to the samples collected in central Auckland. RNA was extracted from plant material using the CTAB method outlined above, and the RT-PCR diagnostic test was performed using the conditions above. A subset of 40 +/- and 40 -/- samples were tested using the VvNAD5 primers as described section 3.2.3.3 to ensure that the RNA extracted was amplifiable.

3.2.8 Determining Optimal Storage Conditions for Leaf Samples

To determine the best way to transport samples from around the country, five sets of samples consisting of a +/+, a +/- and a -/- leaf were taken from the PFR0001 tree. Each set was stored according to different storage conditions for four weeks. The first set was stored in a paper envelope at room temperature. The second set was stored in a paper envelope at room temperature for one week, then transferred to a ziplocked bag and stored at 4°C for the remaining weeks. The third set was stored in a ziplocked bag at 4°C. The fourth set was stored in a ziplocked bag at room temperature for one week, then transferred in the same bag to 4°C for the remainder weeks. The fifth set was stored in a ziplocked bag at room temperature. Room temperature was the ambient temperature in the PFR MARC level 6 laboratory and ranged from between 20°C to 25°C. A sub-sample of each sample was taken each week and the diagnostic test developed as part of this work was applied to each sub-sample. Sampling instructions were developed as a result of this (see Appendix D). To understand the distribution of KOPV around New Zealand, samples from the wider New Zealand areas were surveyed with the help of friends and colleagues using the sampling instructions developed from the

weekly sampling. These samples were not posted, but were hand delivered from their respective locations to the PFR MARC.

3.2.9 Determining Optimal Sampling Time

To determine the best time to sample Karaka to detect the virus, samples were gathered from the PFR0001 tree every month, stored at -80°C and tested using the KOPV diagnostic test. Four branches with the highest density of symptomatic leaves on the tree were labelled as A, B, C and D, and samples were collected approximately monthly over seven months from these branches according to Table 3.2 below.

Table 3.2 - Samples collected from the PFR0001 tree. Samples were collected from branches labelled A, B, C and D approximately monthly from July 2018 to February 2019. Branch A and branch C had both symptomatic and asymptomatic samples taken. Branch B and branch D had only symptomatic samples taken.

	Samples Collected from:			
Date	Branch A	Branch B	Branch C	Branch D
25/07/2018	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
22/08/2018	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
03/10/2018	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
01/11/2018	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
11/12/2018	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
07/02/2019	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic
27/02/2019	Symptomatic Asymptomatic	Symptomatic	Symptomatic Asymptomatic	Symptomatic

Samples were prepared by dividing the leaves into quarters and a 150 mg sub-sample of symptomatic tissue was excised from each section. Some of the sections did not have symptoms so 150 mg of asymptomatic tissue was taken instead. Other sections did not have 150 mg of symptomatic tissue, so as much symptomatic tissue was taken and then asymptomatic tissue from that region was used to make up the remainder. RNA was then extracted from each sub-sample using the CTAB extraction method, standardised to 75 ng/µL and tested using the KOPV diagnostic test to detect presence of the virus. Positive amplifications were identified and sets where a section was consistently amplified across all seven months were selected to be diluted. These RNA extractions were diluted in a 10-fold dilution series up to 1:100,000 and tested using the KOPV diagnostic test to determine the limit of detection for the RT-PCR.

3.3 Results

3.3.1 Comparison of RNA extraction methods

KOPV is new to science and there is a need for a diagnostic test to be established that can accurately detect the virus in Karaka leaves. This project focused on using RT-PCR to detect the virus within Karaka leaves. An important part of this test is being able to extract RNA from symptomatic and asymptomatic plants that will be amplifiable using virus specific primers if the virus is present in the sample. To this end, a comparison was made between the commercial Spectrum kit and the CTAB extraction method. The key focus of this comparison was to determine if each of the KOPV RNA segment specific primers from Table 3.1 (primers and primer bind sites are shown above in Figure 3.3) could be amplified in an RT-PCR reaction. Kawakawa was also extracted to provide additional support for the RNA extractions as a control for specificity. RNA was extracted using Spectrum and CTAB methods from four leaf samples from tree PFR0001.

Symptomatic leaves from symptomatic trees (+/+) showed amplification of a PCR product of the expected size for each RNA in Figure 3.4. Asymptomatic leaves from asymptomatic trees (-/-) gave an unexpected product for RNAs 3 and 4 using RNA extracted with the Spectrum kit. The CTAB method amplified unexpected products for RNA 3 and RNA 4 in asymptomatic leaves (+/- and -/-) and for RNA 4 in the Kawakawa sample. All NTC lanes do not have amplification of a product.

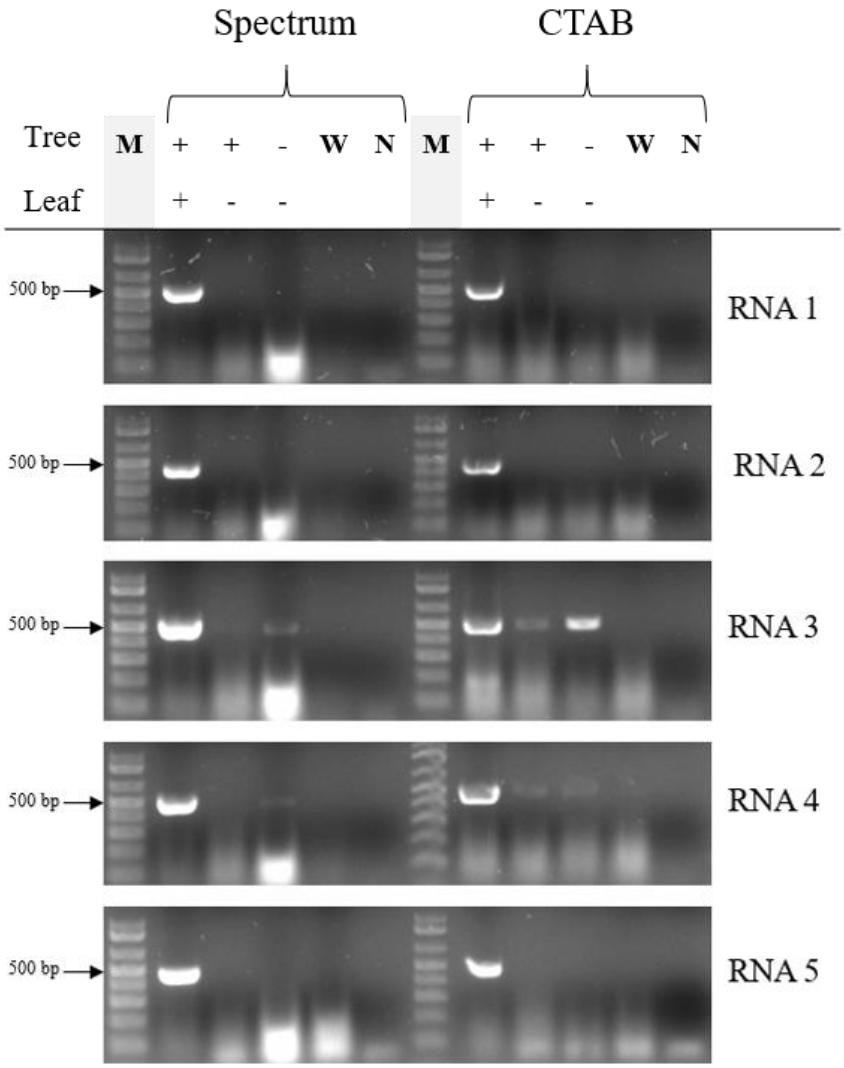


Figure 3.4 – Analysis of a one-step RT-PCR on 2% agarose gels using each of the RNA segment specific primer pairs from Table 3.1 with an annealing temperature of 61°C. RNA was extracted from Karaka and Kawakawa using both Spectrum and CTAB methods as labelled. + is symptomatic, - is asymptomatic and W is Kawakawa. Lane M: 1 kb Plus DNA ladder. Lane NTC: No template control

Both CTAB and Spectrum RNA extraction methods can extract RNA that is amplifiable using each virus segment specific primer. The unexpected products for RNA 3 and RNA 4 were likely due to contamination at the RNA extraction step. Later extractions showed that amplified products in +/- and -/- samples were absent more often than it was present, and if they were present, likely represented a contamination during the extraction.

3.3.2 CTAB vs. Spectrum Sensitivity

The RNA extracted using both the CTAB and Spectrum methods allow amplification of expected products using virus segment-specific primer pairs. The aim of the following experiment was test which method is more sensitive when detecting the virus. RNA extracted from each method was standardised to 75 ng/µL and diluted 10-fold in a dilution series to 1:10,000 and tested using the RNA specific primer pairs.

In Figure 3.5, viral RNA extracted using the Spectrum method can be detected when 75 ng/µL total RNA was diluted up to 10^{-2}, with RNA 3 and RNA 4 primers being able to detect their targets at a10^{-4} dilution. The CTAB method also resulted in amplification of the expected PCR product at a 10^{-2} dilution for every primer pair. While this is not strictly a quantitative assay, the CTAB method appeared to result in less product for each RNA. The RNA extracted with the Spectrum method can be detected at more diluted concentrations for every RNA when compared to RNA extracted using the CTAB method.

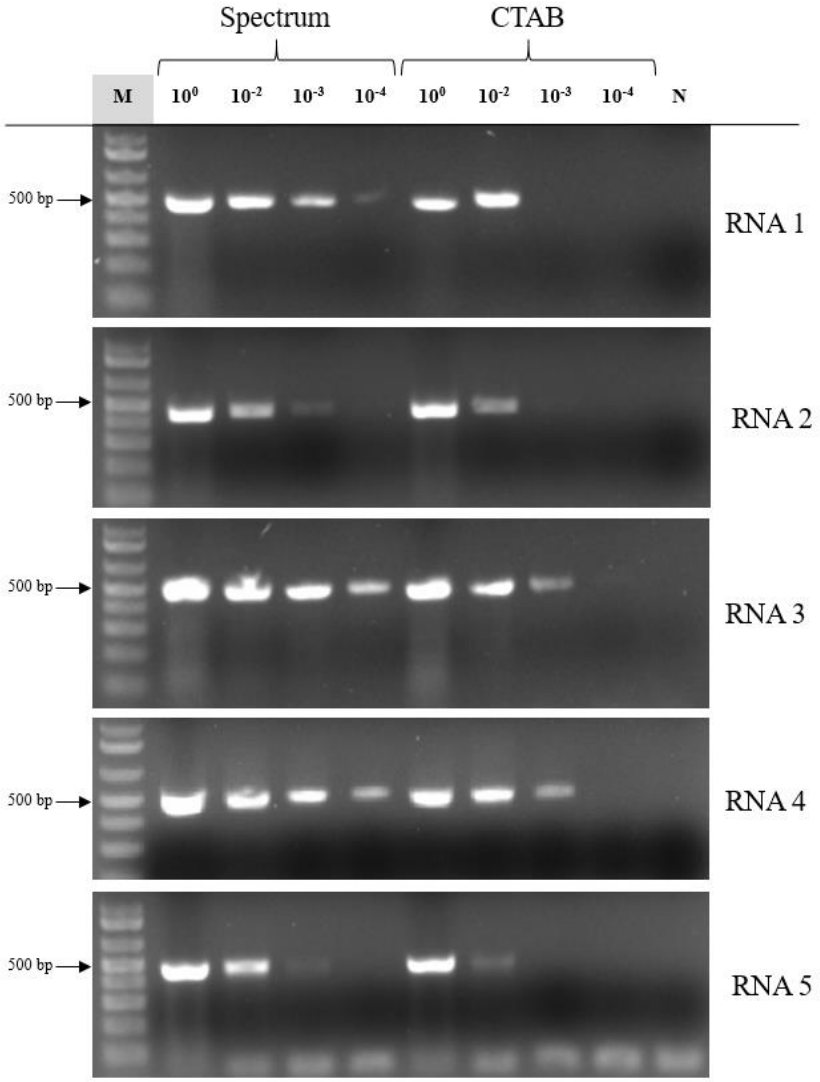


Figure 3.5 – Comparison of the sensitivity of each RNA extraction method for detecting the Karaka emaravirus RNAs. Total RNA was extracted from a symptomatic karaka leaf using both Spectrum and CTAB methods as labelled. RNA was diluted in a 10-fold dilution series up to 1:100,000 (10^{-4}). Lane M: 1 kb Plus DNA ladder. Lane N: No template control.

Both Spectrum and CTAB methods can detect virus at total RNA concentrations of 75 x 10^{-2} ng/µL. Spectrum can detect RNA with greater sensitivity than the CTAB method, but for a diagnostic test, amplifying a viral product at 75 x 10^{-2} ng/µL was deemed to be adequate. CTAB was chosen as the RNA extraction method for the diagnostic test after considering cost factors when extracting large numbers of samples.

3.3.3 Amplification and Analysis of the NAD5 Internal Control

Testing +/-, -/- and Kawakawa samples show that no virus product is amplified. To ensure that this result is truly virus negative and not due to the RNA not being amplifiable, amplification of the internal control sequence NAD5 was carried out. Amplification of this sequence would confirm that the RNA was amplifiable, and that the virus negative result was correct. The NAD5 primers had been designed for *Vitis vinifera* and were tested on Karaka. The forward primer, Vv-NAD5-F, was designed at an exon-exon junction in *V. vinifera* and the reverse primer, Vv-NAD5-F was designed completely within an exon. The design means that the primers will amplify only mRNA and not genomic DNA, and can be used to test if RNA extracted from a plant is amplifiable (Chooi et al., 2013; Menzel et al., 2002). It was unknown if these primers would work with Karaka and Kawakawa, so they were tested using the RNA extractions used in the Spectrum vs. CTAB experiment.

3.3.3.1 NAD5 Two-step RT-PCR

The RNA amplification was initially carried out using in a two-step RT-PCR reaction with the Vv-NAD-5 primers to test if the primers would amplify. A negative RT control was used where the reverse-transcription enzyme was replaced with water to ensure that only RNA was being amplified, not genomic DNA. Once again, Kawakawa was used to provide an additional control, and to determine if the Vv-NAD5 primer pairs can amplify this species too.

Figure 3.6 shows that the NAD5 primers amplified a product for symptomatic and asymptomatic RNA from Karaka, as well as RNA from Kawakawa. All +RT samples were positive while all -RT samples were negative, indicating that the primers only target RNA in Karaka and Kawakawa. This shows that two-step RT-PCR amplification of NAD5 using the grapevine primers could be used to confirm the virus negative status of Karaka.

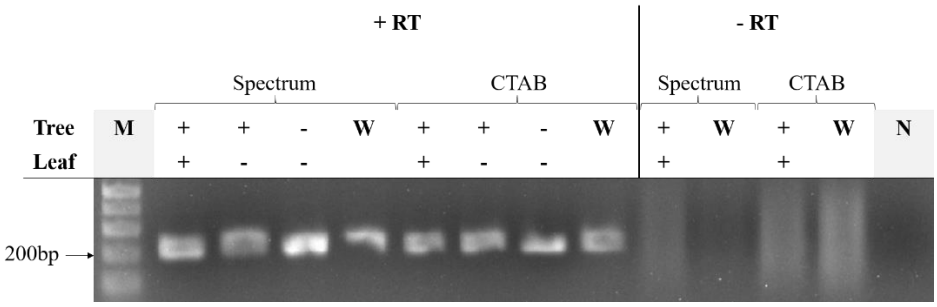


Figure 3.6 – Testing of *V. vinifera* NAD5 primers in a two-step RT-PCR of Karaka and Kawaka RNA. + is symptomatic, - is asymptomatic and W is kawakawa. +RT is where the reverse transcriptase was added, -RT was where water was added instead of the reverse transcriptase. The method of RNA extraction is indicated as Spectrum or CTAB. Lane M: 1 kb Plus DNA ladder. Lane N: No template control.

3.3.3.2 NAD5 One-step RT-PCR

While the two-step NAD5 RT-PCR test showed that the grapevine primers could amplify a product from Karaka and Kawakawa, a two-step RT-PCR takes time and, with many samples, can be a burden. A one-step test is potentially more time efficient. A one-step reaction was tested and optimised to amplify the NAD5 region in Karaka and Kawakawa. A -RT reaction was not done for the one-step RT-PCR as the commercial kit used had the reverse transcriptase and *Taq* polymerase combined into one solution.

Figure 3.7 shows that NAD5 was amplified from all samples to give the expected sized products present in the plant mRNA. Curiously, this assay only gave these results when using the Eppendorf Mastercycler® Gradient machine. No amplification was observed on any other thermocycler using the same conditions, RNA and reagents.

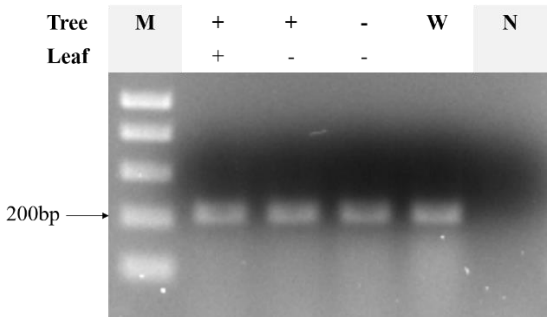


Figure 3.7 – Testing of *V. vinifera* NAD5 primers in a one-step RT-PCR of Karaka and Kawakawa RNA. + is symptomatic, - is asymptomatic and W is kawakawa. Lane M: 1 kb Plus DNA ladder. Lane N: No template control.

3.3.3.3 Sequencing of NAD5 Products

The RT-PCR products amplified from the and NAD5 primers were sequenced to ensure that the primers were targeting the correct region in the karaka mRNA transcripts. Sequence data was sent to Macrogen for sequencing, returned by email, processed, and analysed using BLASTP (https://blast.ncbi.nlm.nih.gov/). The NAD5 product size was estimated as 220 bp based on the size of the Vitis vinifera NAD5 product amplified using the Vv-NAD5-F and Vv-NAD5-R primers.

The translated amino acid sequence of the NAD5 primer pairs had the closest similarity to the NAD5 sequence from *V. vinifera* with 100% identity match. The NAD5 sequence also matched many other species' NAD5 sequence with 100% identity match. The NAD5 primer pairs appear to be amplifying the correct region and can be used as an internal control going forward.

3.3.4 Sequencing of RNA 1-5 PCR products

The RT-PCR products amplified from the RNA 1-5 primers were sequenced to ensure that the primers were targeting the correct region in the viral genome. Sequence data was sent to Macrogen for sequencing, returned by email, processed, and analysed using BLASTP

(https://blast.ncbi.nlm.nih.gov/). KOPV was aligned to the completed genome obtained from Chapter 2. The sequence length of each primer pair for nucleotide and amino acid is shown in Table 3.3 below. For KOPV pairs (RNA 1-5), the expected product size is based on the primer binding sites on the completed genome.

Table 3.3 - Primer pairs that were used in sequencing KOPV RT-PCR products

Primer Pair	Expected Product Size (nt)
KOPV RNA1 4669F & KOPV RNA1 5168R	500
KOPV RNA2 482F & KOPV RNA2 922R	441
KOPV RNA3 393F & KOPV RNA3 863R	504
KOPV RNA4 169F & KOPV RNA4 649R	481
KOPV RNA5 110F & KOPV RNA5 609R	500

The KOPV RNA 1-5 primer pairs all aligned with 100% identity to the respective segment of the completed KOPV genome, showing that the primer pairs amplify the expected product. BLASTP analysis shows that portion of RNA 1 and RNA 3 sequences have similarity to other emaraviruses, which adds evidence to this virus belonging to the emaravirus genus. RNA 2, RNA 4 and RNA 5 did not show similarity, although as seen in Chapter 2, the completed sequence of each RNA does show similarity to other emaraviruses.

3.3.5 Detection of Karaka emaravirus by uniplex and multiplex RT-PCR

A diagnostic test relying on detection of one of the viral RNA segments may result in a false negative reaction if that RNA is present at low level. While the data shown in Figure 3.5 is not quantitative, it shows that the different RNAs are detectable down to different dilutions suggesting that the RNAs are not necessarily present in equal amounts. RNAs 3 and

5 are detectable at higher dilutions than the other RNAs. A test that could identify all or any one of the five RNAs in one reaction would be more robust. Combining all five RNA specific primers into a single RT-PCR was tested in comparison to each primer pair separately (Figure 3.8). RNA was extracted from symptomatic and asymptomatic samples taken from central Auckland from symptomatic and asymptomatic trees. In Figure 3.8, + is symptomatic and - is asymptomatic. Bands were expected in samples that were + for tree and + for leaf. RNA was tested in an RT-PCR using the individual RNA segment specific primers and using a multiplex test, where all the individual primers were used in a single reaction.

RT-PCR products were expected in all +/+ samples, with no amplification expected in +/- and -/- samples. Each RNA primer pair gave the expected results except for RNA 3 and RNA 4 where amplification was observed in a +/- sample. This was likely due to contamination as the positive band did not appear in the multiplex reaction. Further, very faint amplification of RNA 3 and RNA 5 was observed in the positive control sample (P). Multiplexing all the five primer pairs gave the expected product in the appropriate samples, with no products observed in +/- or -/- samples. The positive control sample gave rise to a very faint product; subsequent experiments used a different positive control that gave rise to a strong product (data not shown). Issues with the P sample shown here may have been due to RNA degradation starting to occur. Note that the internal control targeting NAD5 transcripts was not being used at this stage in the experimental process.

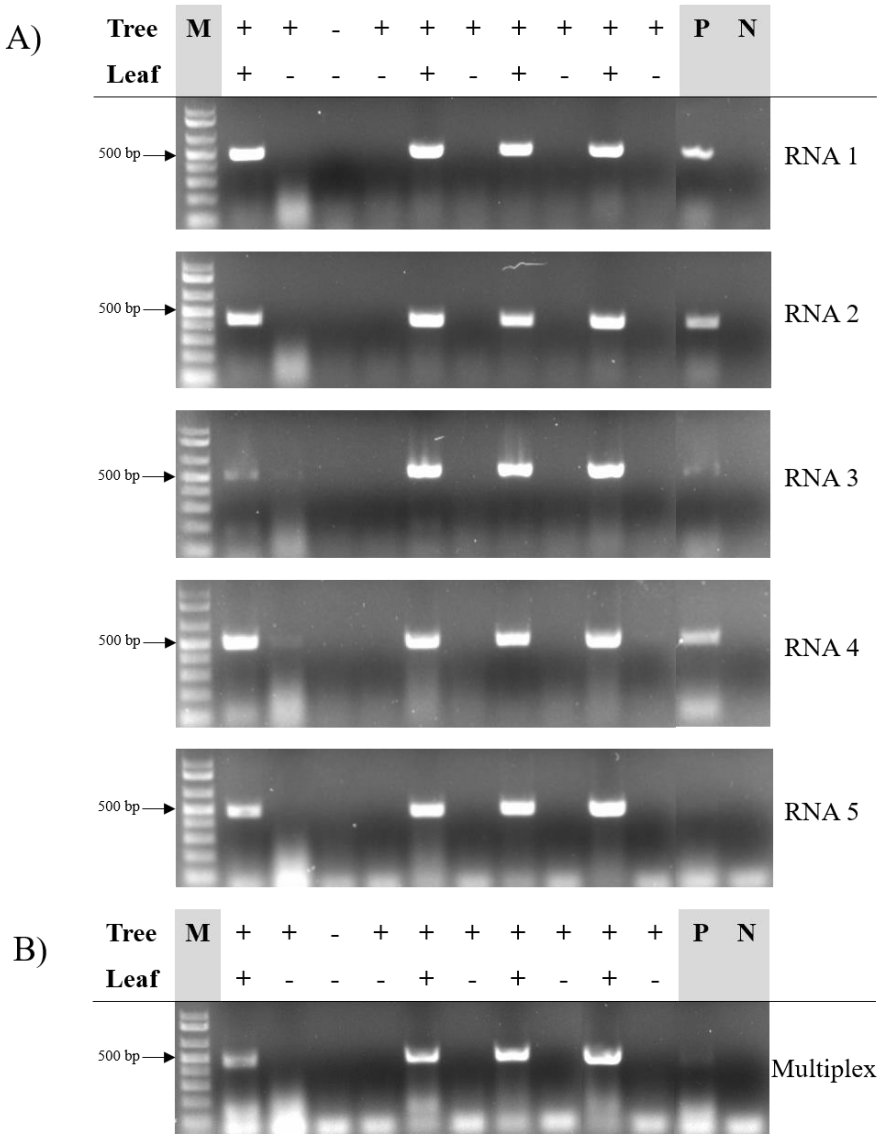


Figure 3.8 – RT-PCR products resulting from combining all five RNA specific primer pairs into a single reaction, compared with each separately. Lane M: 1 kb Plus DNA ladder. Lane P: Positive control (L19b-N). Lane N: No template control

3.3.6 Testing Generic Primers

Generic primers have been described to detect all known emaraviruses. There are two primer pairs that have been described to date, the first by Elbeaino et al. (2013), and the second by Olmedo-Velarde et al. (2016), both described in Table 3.1. These primers are designed to bind to conserved regions in RNA 1 of emaravirus that had been described to their respective dates, and theoretically can amplify any species belonging to the genus. These primers were

tested with KOPV to determine that they are useful for diagnosing this virus, and to compare to the primers used in this study. These primers were first aligned to the KOPV genome using Geneious to determine if they would bind and to obtain the expected product size, and then tested using one-step RT-PCR and virus-positive karaka RNA.

Figure 3.9 shows the pairwise alignment of the generic primers to the RNA 1 sequences of all known emaraviruses, including KOPV. Both generic primer sets did not perfectly align to KOPV RNA 1 genome or the rest of the emaravirids. The primers developed by Elbeaino et al (2013) were expected to bind as they showed some similarity to the relevant KOPV sequence. The Olmedo-Velarde (2016) primers, however, showed low similarity to the KOPV relevant sequence and were not expected to amplify a product in an RT-PCR.

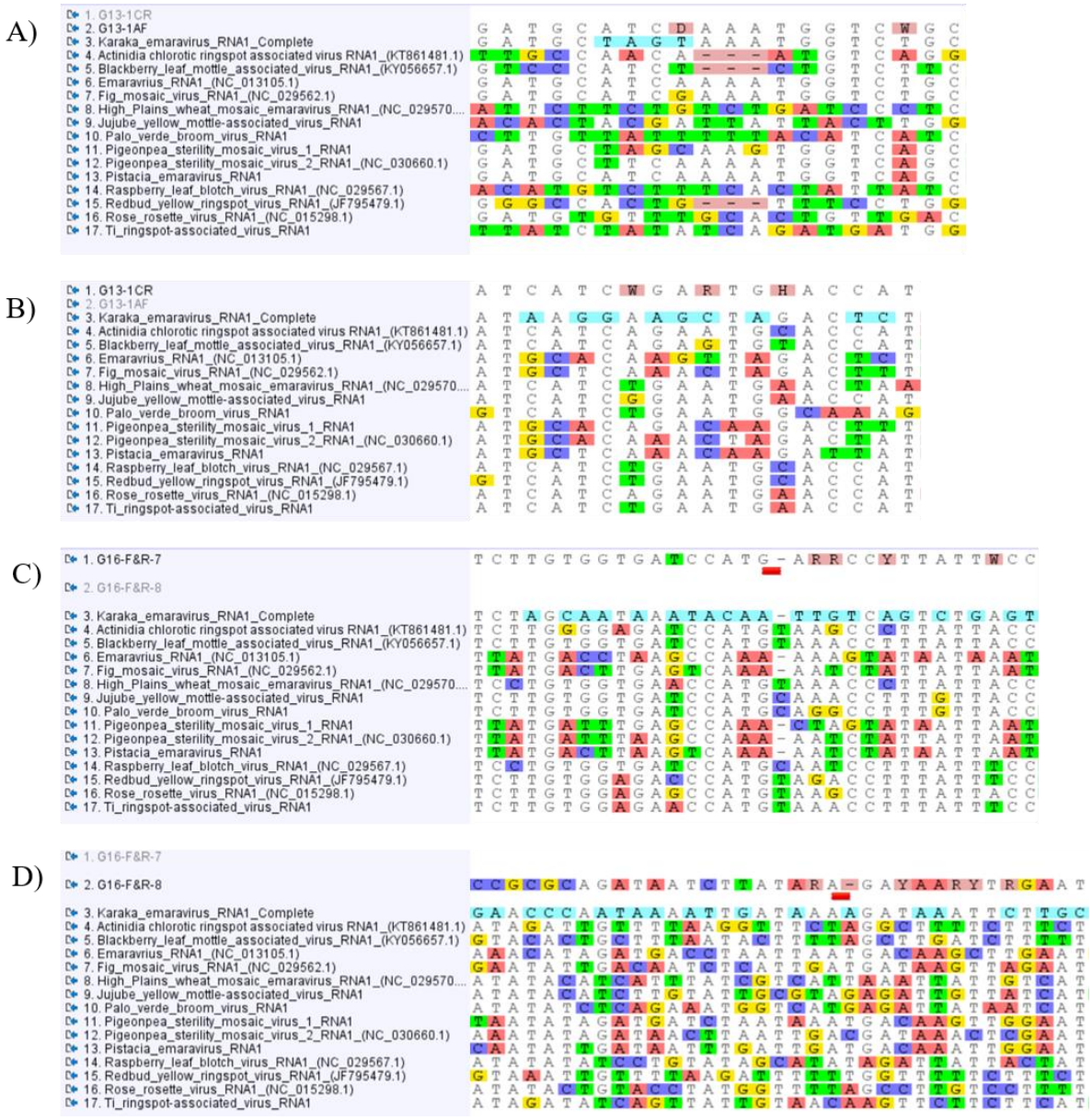


Figure 3.9 – Alignment of published generic primers to RNA 1 of all known emaraviruses, including KOPV. The primers developed by Elbeaino et al (2013) are shown in A) the forward primer, G13-1AF, and B) the reverse primer G13-1CR. The primers developed by Olmedo-Velarde et al (2016) are shown in C) the forward primer G16-F&R-7, and D) the reverse primer G16-F&R-8.

Figure 3.10 shows the comparison of the generic primers and the karaka emaravirus specific primers for diagnosing KOPV. Samples known to be positive and negative for KOPV were used for this comparison and two sets of +/+ RNA were used to check if the age of the

RNA would affect the results. Lane 1 was RNA extracted from a symptomatic leaf 3 months before the RT-PCR and Lane 2 was extracted 9 months before the RT-PCR.

The KOPV specific multiplex RT-PCR gave the expected products. The 2013 primers gave the expected products for both +/+ samples and the positive control. The faint result in lane 2 was possibly due to there being low levels of RNA 1 in this sample; this RNA was older than that in lane 1 and may have been degraded. This did not appear to be an issue when using the KOPV specific primers. Further, the 2013 primers gave a non-specific 700 bp product in Kawakawa. The 2016 primers did not appear to amplify any products. Note, the positive control was omitted from this RT-PCR.

Considering that the KOPV diagnostic test and the 2013 generic primers amplified the expected products, it can be concluded that KOPV is present within the symptomatic leaf RNA. The 2013 generic primers could be used instead of the KOPV diagnostic test, although it may not be as sensitive. Therefore, the KOPV multiplex diagnostic test was used in the subsequent sections of this Chapter.

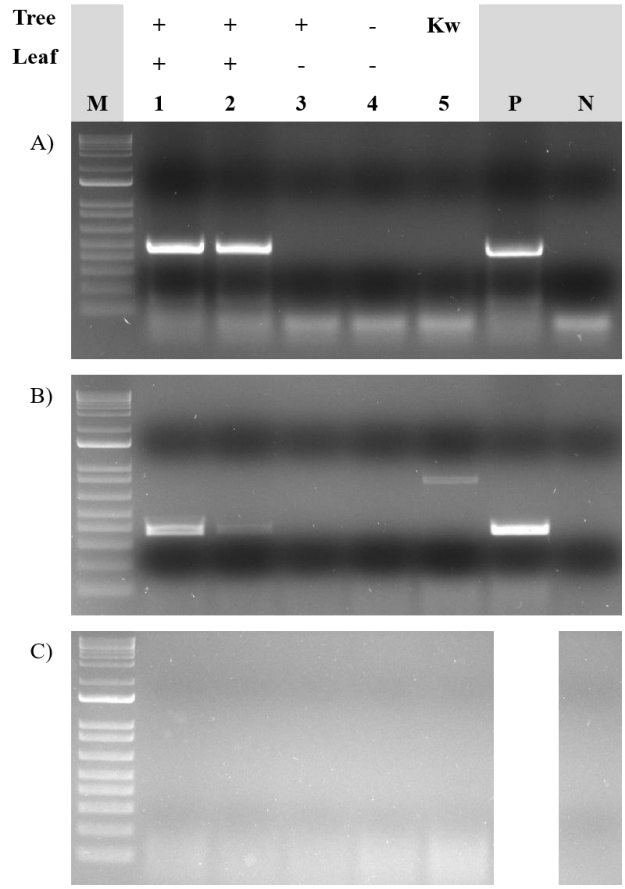


Figure 3.10 – Comparison of RT-PCR using generic emaravirus and KOPV specific primers. A) The karaka multiplex specific diagnostic test. B) the generic primers of Elbeaino et al (2013); G13-1AF and G13-1CR. C) the generic primers of Olmedo-Velarde et al (2016); G16-F&R-7 and G16-F&R-8. RNA was extracted from karaka and kawakawa using the CTAB method and standardised to 75 ng/µL. + is symptomatic, - is asymptomatic and Kw is kawakawa. Lane M: 1 kb Plus DNA ladder. Lane P is the positive control (L19b-N). Lane N: No template control. Primers described on Table 3.1.

3.3.7 Survey and Sampling

Using the multiplex diagnostic test, the distribution of KOPV around Auckland was investigated. This was done to determine how widely the virus was distributed around Auckland and to set the groundwork and data collection for future studies that might gather samples from around the country. Symptomatic and asymptomatic trees were surveyed around central Auckland and samples were taken.

Table 3.4 shows that the number of symptomatic and asymptomatic trees surveyed in New Zealand, with 47.6% of trees surveyed being symptomatic. Figure 3.11 shows that the samples were also relatively evenly distributed around central Auckland. Some grids are empty where karaka trees could not be found, or where surveying did not take place. There appears to be no area that has a large concentration of symptomatic trees compared to asymptomatic trees, and vice versa. Figure 3.12 shows the limited number of samples surveyed outside Central Auckland. The samples gathered in Kerikeri (Figure 3.12 – Location 1) were all asymptomatic, whereas the samples gathered from Waipu (Figure 3.12 – Location 2) had a mixture of symptomatic and asymptomatic trees as seen in central Auckland. This view also shows two symptomatic outliers in Auckland. One from the western beach at Murawai (Figure 3.12 – Location 3) and one from Waiheke Island (Figure 3.12 – Location 4). The Waipu, Murawai and Waiheke symptomatic trees show that the virus can be found outside of central Auckland. The symptomatic trees in Central Auckland appear to be relatively evenly distributed alongside asymptomatic trees and have the same symptoms as those found on the PFR0001 tree. While the presence of symptoms suggests the presence of virus, it was necessary to confirm this using the multiplex diagnostic test.

Table 3.4 - Symptomatic and asymptomatic Karaka trees surveyed in Central Auckland and in other North Island areas. *High percentage due to low number of samples observed.

	Symptomatic	Asymptomatic	% Symptomatic
Central Auckland	112	138	47.7 %
Other North Island Areas	3	5	37.5 %*
Total	115	143	47.6 %

Figure 3.11 – Distribution of the karaka emaravirus in Central Auckland. Red dots represent symptomatic trees, orange dots represent asymptomatic trees and light-blue dots represent trees with uncertain symptoms. This map is overlaid with a 2 km^2 grid.

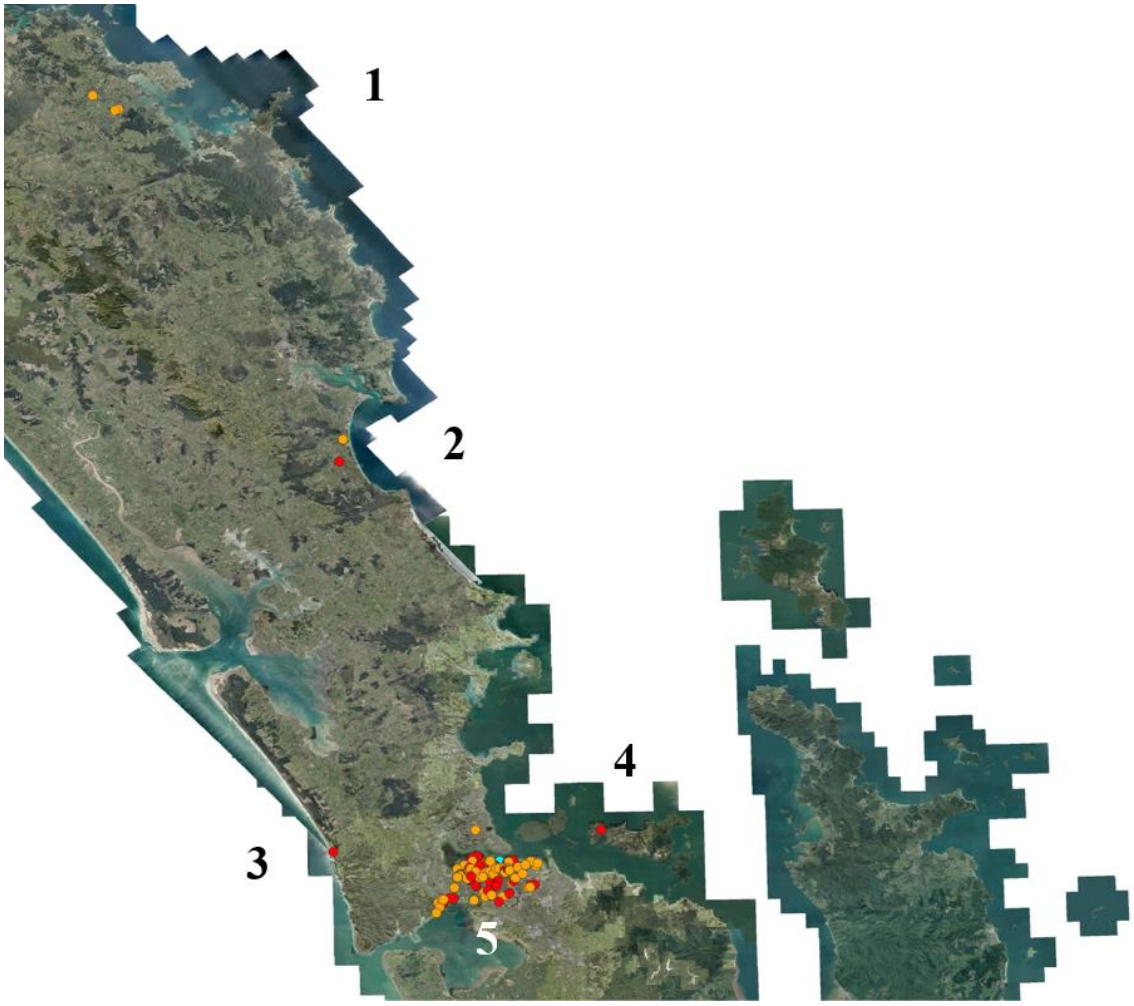


Figure 3.12 – Distribution of Karaka emaravirus surveyed in Central Auckland and other North Island Locations. Red dots represent symptomatic trees, orange dots represent asymptomatic trees and light-blue dots represent trees with uncertain symptoms. Location 1, 2, 3, 4 and 5 are Kerikeri, Waipu, Murawai Beach, Waiheke Island and Central Auckland

3.3.8 Application of the Diagnostic Test on Sampled Trees

The symptomatic trees show the same symptoms of the virus that were observed at the MARC at PFR on tree PFR0001. The next step was to apply the KOPV diagnostic test to determine whether the virus was present within the symptomatic samples gathered. A key part of this test was also to determine if the virus could be found in asymptomatic samples gathered from trees with symptoms and trees without. A total of 258 trees were sampled, with 115 symptomatic, 115 asymptomatic from these symptomatic trees and 143 asymptomatic from

asymptomatic trees. If a leaf was symptomatic, the section taken for RNA extraction always included some symptoms.

An example gel depicting the analysis is shown in Figure 3.13, with bands at around 500 bp showing a virus-positive result. A positive result was expected in all lanes that had RNA from a symptomatic tree that had symptoms on the leaf (+/+). Table 3.5 below shows that 88.8% of +/+ samples tested positive using the diagnostic test suggesting that the virus is present within them. Only 6.1% percent of +/- and 2.8% of -/- samples tested positive, suggesting that the virus is not found commonly outside of symptomatic leaves. The +/+ samples that tested negative (potentially a false negative) as well as +/- and -/- samples that tested positive (potentially a false positive) were viewed as suspect and re-tested using the re-test method. Also, to be sure that a negative was truly a negative result, a subset of the +/- and -/- RNA was tested using NAD5 primers to ensure that it was amplifiable.

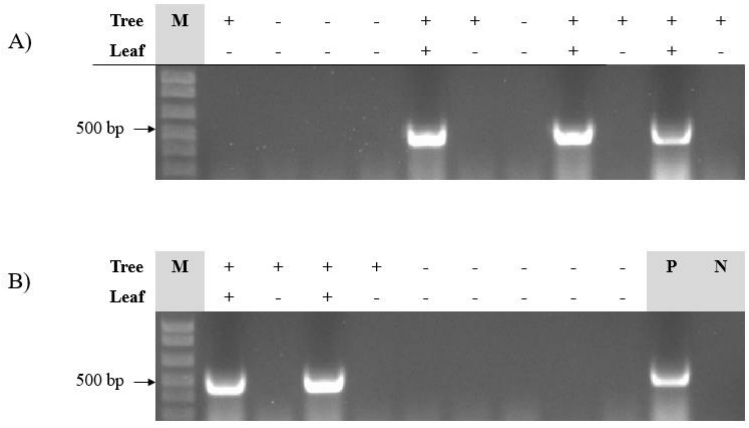


Figure 3.13 – Example gel using the diagnostic test on symptomatic and asymptomatic samples collected from around Auckland. + is symptomatic and - is asymptomatic. Bands were expected in samples that were + for tree and + for leaf. Lane M: 1 kb Plus DNA ladder. Lane P: Positive control (L19b-N). Lane N: No template control.

Table 3.5 - KOPV diagnostic test results for sampled trees. Leaves were sampled from symptomatic and asymptomatic trees. Virus positive results are where the expected product successfully amplified using the RT-PCR diagnostic test developed in section 3.3.5.

Sample type	Symptom code	Samples collected and tested	Virus Positive	Virus Positive (%)
Symptomatic tree and leaf	+/+	115	101	88.8%
Symptomatic tree with asymptomatic leaf	+/-	115	7	6.1%
Asymptomatic tree and leaf	-/-	143	4	2.8%

A small percentage of samples returned unexpected results from the diagnostic test: a symptomatic sample returning a negative result, for example, or an asymptomatic sample returning a positive result. These samples were viewed with suspicion as they did not produce the expected results. To ensure that these samples were not due to contamination, or an error, in the RNA extraction or RT-PCR diagnostic test, they were re-tested by extracting new RNA from the leaves and running the RT-PCR diagnostic test again. Table 3.6 shows that, after re-testing samples that had unexpected results, five of the fourteen +/+ samples that initially showed no amplification, returned a positive result. All seven of the +/- samples that showed an initial virus positive result returned an expected negative result. Only 1 of the 4 -/- samples that showed an initial virus positive result returned an expected negative result. One of the +/+ and three of the two of the -/- samples with unexpected results were not re-tested due to time constraints. A revised table combining the initial results from Table 3.5 and the re-tested results from Table 3.6 is shown on Table 3.7.

Table 3.6 - Results for re-testing sampled trees with unexpected results. Leaves were sampled from symptomatic and asymptomatic trees. Virus positive results are where product was amplified using the multiplex RT-PCR diagnostic test.

Sample type	Symptom code	Samples with unexpected results	Samples re-tested	Virus Positive
Symptomatic tree and leaf	+/+	14	13	5
Symptomatic tree with asymptomatic leaf	+/-	7	7	0
Asymptomatic tree and leaf	-/-	4	2	1

Table 3.7 shows that 92.2% of +/+ samples returned a positive result after re-testing unexpected results. None of the +/- samples amplified a virus product in the diagnostic test and only 2.1% of the -/- amplified a virus product. The re-testing was necessary to help eliminate any false results; however, there were still nine +/+ samples that did not amplify, and three -/- that did amplify even after re-testing. This suggests that there might be an anomaly that is outside of the initial hypothesis that virus is, and only is, found in symptomatic regions of karaka leaves.

Table 3.7 - Results combining the initial results from Table 3.5 and the re-tested results from Table 3.6. Leaves were sampled from symptomatic and asymptomatic trees. Virus positive results are where product was amplified using the multiplex RT-PCR diagnostic test.

Sample type	Symptom code	Samples tested and re-tested	Virus Positive	Virus Positive (%)
Symptomatic tree and leaf	+/+	115	106	92.2%
Symptomatic tree with asymptomatic leaf	+/-	115	0	0.0%
Asymptomatic tree and leaf	-/-	143	3	2.1%

3.3.8.1 NAD5

To determine whether the RNA samples that were negative for the virus were PCR competent for use with the RT-PCR diagnostic test, the one-step NAD5 diagnostic test was applied to a four subsets of virus negative samples. This was done to ensure that RNA that was used from +/- and -/- samples was amplifiable and diagnostic results were true negatives. Table 3.8 shows that at least 22 of the 24 samples of each subset amplified the NAD5 product and Figure 3.14 shows an example RT-PCR analysis of +/- and -/- subsets. A total percentage of 96.9% of samples gave rise to the expected NAD5 product. The samples that did not amplify had A_{260}/A_{280} of 1.02 to 1.86 – far greater than the expected for good quality RNA. For these samples it was most likely due to a poor A260/280 reading. This shows that the RNA is amplifiable and that the negative results from the diagnostic tests were indeed negative because they had no virus present.

Table 3.8 - NAD5 amplification of a selection of virus negative samples.

Subset	Symptoms	RNA Samples Tested	Samples Positive for NAD5	Samples Positive for NAD5 (%)
+/- (1)	Asymptomatic leaves from a tree with symptoms.	24	22	91.6%
+/- (2)	Asymptomatic leaves from a tree with symptoms.	24	23	95.8%
-/- (1)	Asymptomatic leaves from a tree without symptoms.	24	24	100%
-/- (2)	Asymptomatic leaves from a tree without symptoms.	24	23	95.8%
Total		98	95	96.9%

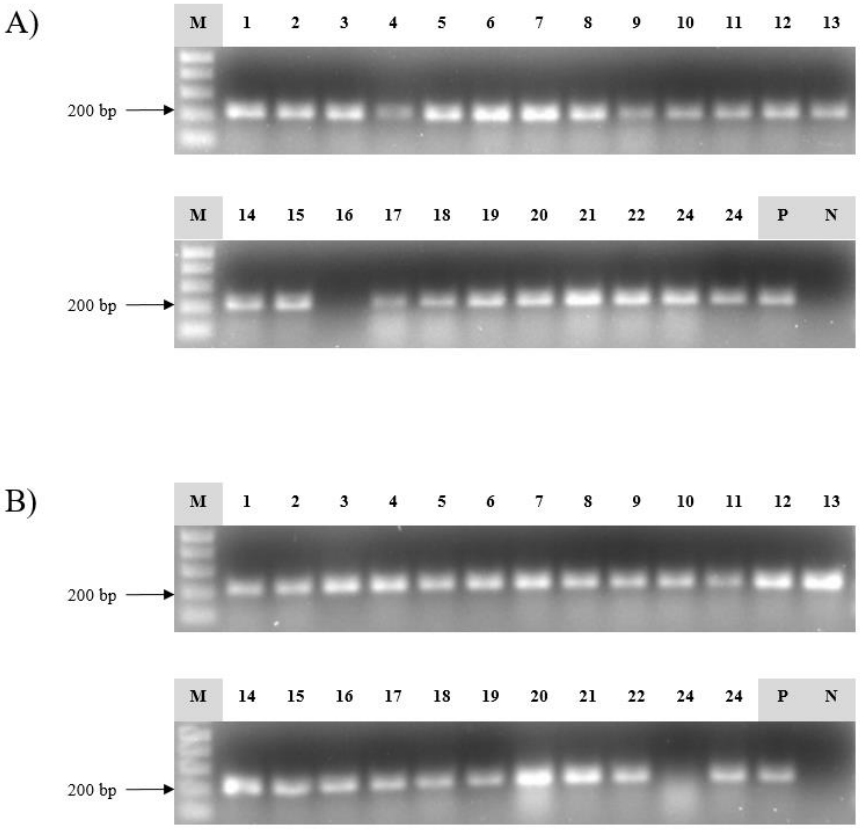


Figure 3.14 - Example gels of NAD5 testing of +/- and -/- samples. Lane M: 1 kb Plus DNA ladder. Lane P: Positive control L21b-N. Lane N: No template control. A) Subset +/- (1) samples extracted in October and November. B) Subset -/- (2) samples extracted in October and November. Note that in B), samples 14 to 24 appear to be different sizes due to the gel setting incorrectly.

3.3.9 Determining Optimal Storage Conditions for Leaf Samples

Few samples were obtained from outside of Auckland. This was because the survey method relied on on-foot and first-person observations of symptomatic and asymptomatic trees. To further understand the distribution of symptomatic and virus infected Karaka in New Zealand, a strategy was developed to enlist the aid of fellow New Zealander's to help survey and sample the trees. This was done by developing a sampling instructions flyer that could be distributed in hard-copy or digitally (Appendix D). The first step was to determine what would be the best way to get samples to Auckland that could be tested with a reliably positive result from symptomatic samples. Sending the samples via the New Zealand postal system was identified as being the most reliable option, and different treatments were designed to simulate a sample in transit via post. A selection of +/+, +/- and -/- samples were taken from tree PFR0001 at the MARC. Leaves from these samples at week 0 were stored in various packaging and temperatures, and were tested weekly for the presence of the virus for four weeks.

Table 3.9 and Figure 3.15 each show the results for each storage treatment for each week. Treatment C (samples stored in a ziplock bag at 4°C for four weeks) and treatment D (samples stored in a ziplock at room temperature for one week, then in a ziplock bag at 4°C for three weeks) both had positive results for the entire four weeks. Treatment B (leaf samples stored in an envelope at room temperature for one week, then in a ziplock bag at 4°C for three weeks) showed a faint positive at week 0 and was continued to be used in the experiment. Week 3 samples were not used as there was contamination in the RNA extraction causing all results to appear positive. The extractions were not re-done due to limited leaf material and not having enough to extract RNA at week 4. These results show that there are several ways of storing the tissue with and without refrigeration, that will allow amplification of the virus if it is present, even after four weeks. While both treatment C and D allowed amplification of a virus product consistently over the four-week period, it is impractical to refrigerate a sample when sending through the postal system. Therefore, treatment D was chosen as the preferred method of sending samples due to the one week of room temperature (allowing postage time) followed by refrigeration (following receipt of the sample) being effective.

Table 3.9 - Table showing results of treatments A through E. Room temperature of storage type was between 20°C and 25°C. +/+ is a symptomatic leaf from a symptomatic tree, +/- is an asymptomatic leaf from a symptomatic tree and a -/- is an asymptomatic leaf from an asymptomatic tree. P is a positive RT-PCR result using the diagnostic test, N is a negative result. Week 3 is omitted as there was contamination from the RNA extraction and not enough leaf material to extract the RNA again.

Treatment	Storage Type	Symptoms	Week 0	Week 1	Week 2	Week 3	Week 4
A	Paper envelope at room temperature for four weeks	+/+	P	N	P	-	N
		+/-	N	N	N	-	N
		-/-	N	N	N	-	N
B	Paper envelope at room temperature for one week, Ziplock bag at 4°C for three weeks	+/+	P	N	P	-	P
		+/-	N	N	N	-	N
		-/-	N	N	N	-	N
C	Ziplock bag at 4°C for four weeks	+/+	P	P	P	-	P
		+/-	N	N	N	-	N
		-/-	N	N	N	-	N
D	Ziplock at room temperature for one week, Ziplock bag at 4°C for three weeks	+/+	P	P	P	-	P
		+/-	N	N	N	-	N
		-/-	N	N	N	-	N
E	Ziplock bag at room temperature for four weeks	+/+	P	P	P	-	N
		+/-	N	N	N	-	N
		-/-	N	N	N	-	N
		+/-	N	N	N	-	N

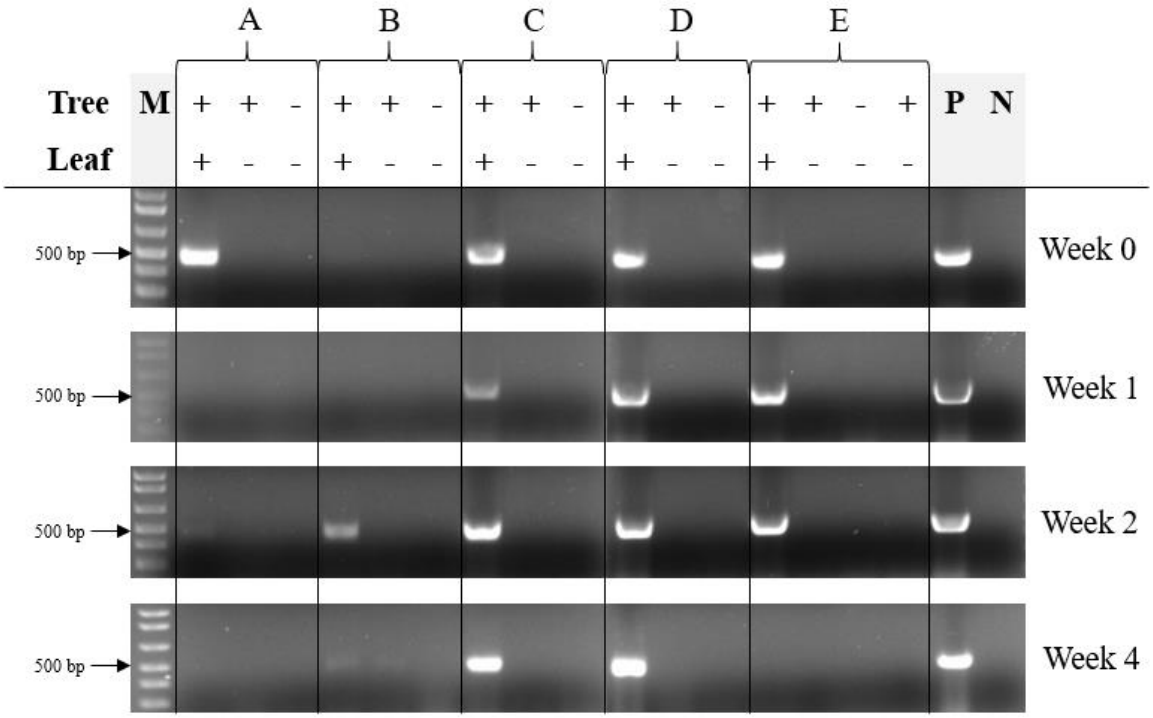


Figure 3.15 – RT-PCR results from different storage conditions. A) Leaf samples stored in an envelope at room temperature for four weeks. B) Leaf samples stored in an envelope at room temperature for one week, then stored in a ziplock bag at 4°C for three weeks. C) Leaf samples stored in a ziplock bag at 4°C for four weeks. D) Leaf samples stored in a ziplock bag at room temperature for one week, then stored in a ziplock bag at 4°C for three weeks. E) Leaf samples store in a ziplock bag at room temperature for four weeks. All RNA is from karaka and the symptoms of KOPV are recorded as + for symptomatic and – for asymptomatic. Lane M: 1 kb Plus DNA ladder. Lane P: Positive control L19b-N. Lane N: No template control.

3.3.10 Determining Optimal Sampling Time

To determine whether there was an optimal sampling period throughout the year to detect KOPV, symptomatic and asymptomatic samples were taken from the PFR0001 tree over seven months. The symptomatic regions of each sample were extracted, diluted and tested to determine whether the sensitivity of the KOPV diagnostic test changed depending on the month it was sampled from. Symptomatic and asymptomatic samples were taken from four different branches of the same tree at near-monthly intervals for seven months. Each leaf was divided into four sections, as in Figure 3.16 below, and RNA was extracted from each sample, then tested using the diagnostic test.

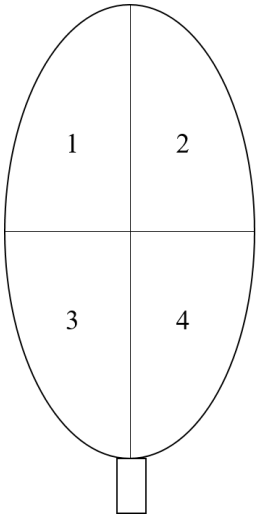


Figure 3.16 - Diagram showing how each leaf was divided. RNA was extracted from each section.

Four sections, two from leaf A and two from leaf C, were selected to be diluted to test the viral titre. These sections all showed amplification through all seven time periods in Figure 3.17. These samples were Leaf A section 2, Leaf A section 3, Leaf C section 1, and Leaf C section 2. Leaf A section 3 had a faint band at month 3. Leaf C section 1 had no amplification at month 4, and leaf C section 2 had no amplification at month 6. These samples were still used as they were the leaf sections that most consistently amplified the expected product over all seven time periods. The selected samples were diluted in a 10-fold dilution series up to 75 ng/μL x 10^{-5} and the karaka diagnostic test was applied.

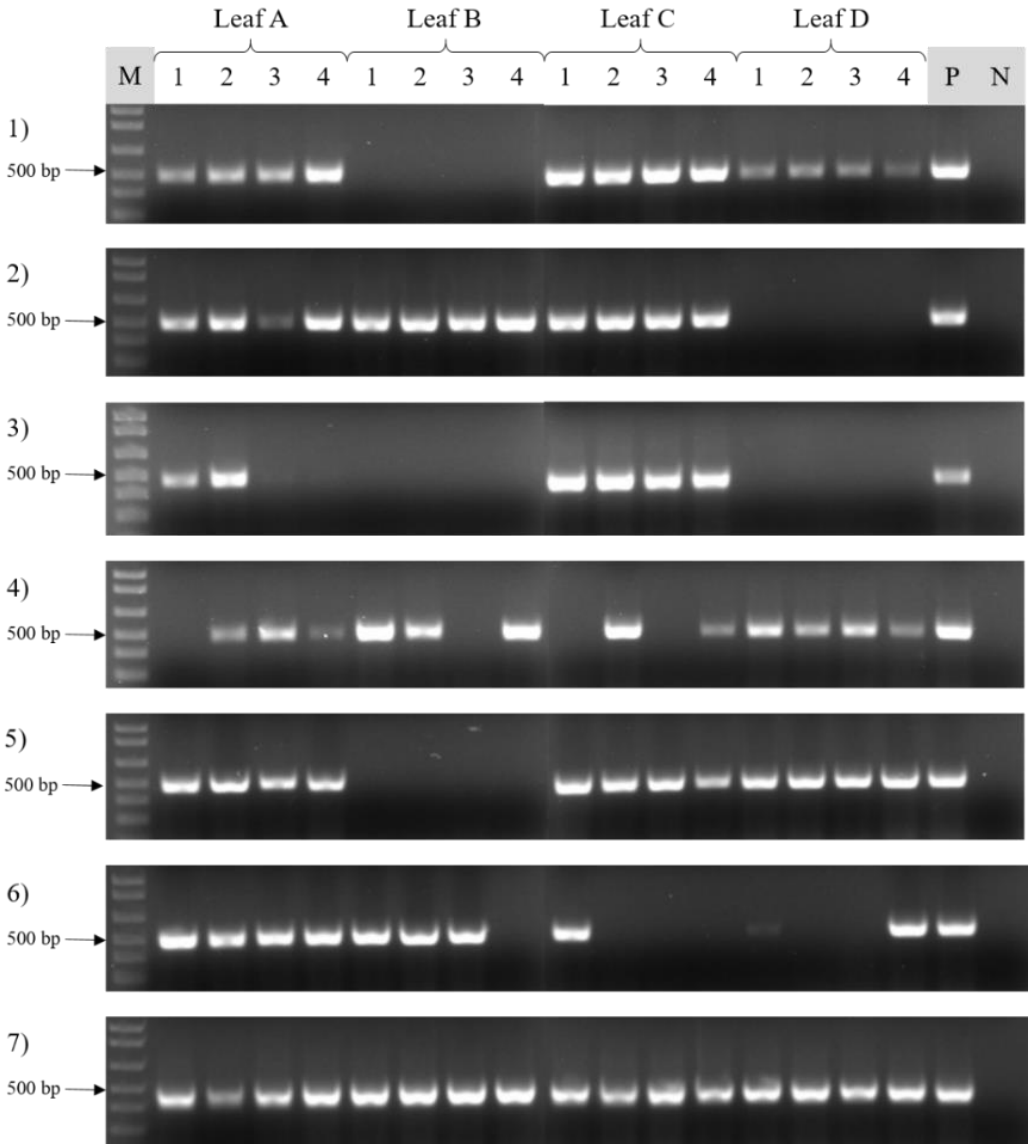


Figure 3.17 – RT-PCR analysis of standardized RNA extracted samples gathered between July 2018 and February 2019. Four symptomatic leaves were gathered each month. The leaves (A, B, C and D) were gathered from the same branch each for each time period and divided into sections as shown in Figure 3.16. Leaves were gathered on 1) 25/07/2018, 2) 22/08/2018, 3) 03/10/2018, 4) 01/11/2018, 5) 11/12/2018, 6) 07/02/2019, and 7) 27/02/2019. Some of the sections did not have symptomatic regions. RNA was extracted from these regions anyway.

Table 3.10 shows the most samples amplified a product at 10^{-2} dilutions. Leaf C section 2 at 1/11/2018 showed amplification up to 10^{-4}, and a few other samples showed amplification at 10^{-3}. There appears to be no month or sample section that amplifies at a higher dilution. Further, there appears to be no difference to when a karaka sample should be taken to be able to detect virus in symptomatic leaves. Thus, sampling can occur at any time of the year.

Table 3.10 - Table showing results of a dilution series of RNA extracted from Leaf A, section 2 and 3, and Leaf C, section 1 and 2, at different time periods. RNA was standardised to 75 ng/µL (the 10^{0} column) then diluted in a ten-fold dilution series to 75 x 10^{-5} (the 10^{-5}) column. Diluted RNA was tested using the multiplex diagnostic test. Positive results were a strong band at the 500 bp mark, a faint result was a faint band 500 bp, and a negative result was no band present in the lane.

Leaf	Section	Sampling Date	10^{-0}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
A	2	25/07/2018	Positive	Positive	Negative	Negative	Negative
A	3	25/07/2018	Positive	Positive	Negative	Negative	Negative
C	1	25/07/2018	Positive	Positive	Faint	Negative	Negative
C	2	25/07/2018	Positive	Positive	Negative	Negative	Negative
A	2	22/08/2018	Positive	Negative	Negative	Negative	Negative
A	3	22/08/2018	Positive	Positive	Negative	Negative	Negative
C	1	22/08/2018	Positive	Negative	Negative	Negative	Negative
C	2	22/08/2018	Positive	Positive	Faint	Negative	Negative
A	2	3/10/2018	Positive	Positive	Negative	Negative	Negative
A	3	3/10/2018	Positive	Positive	Negative	Negative	Negative
C	1	3/10/2018	Positive	Positive	Negative	Negative	Negative
C	2	3/10/2018	Positive	Positive	Negative	Negative	Negative
A	2	1/11/2018	Positive	Positive	Negative	Negative	Negative
A	3	1/11/2018	Positive	Positive	Negative	Negative	Negative
C	1	1/11/2018	Negative	Positive	Negative	Negative	Negative
C	2	1/11/2018	Positive	Positive	Positive	Faint	Negative
A	2	11/12/2018	Positive	Positive	Negative	Negative	Negative
A	3	11/12/2018	Positive	Faint	Negative	Negative	Negative
C	1	11/12/2018	Positive	Positive	Negative	Negative	Negative
C	2	11/12/2018	Positive	Positive	Negative	Negative	Negative
A	2	7/02/2019	Positive	Positive	Negative	Negative	Negative
A	3	7/02/2019	Positive	Positive	Negative	Negative	Negative
C	1	7/02/2019	Positive	Faint	Negative	Negative	Negative
C	2	7/02/2019	Negative	Negative	Negative	Negative	Negative
A	2	27/02/2019	Positive	Faint	Negative	Negative	Negative
A	3	27/02/2019	Positive	Positive	Negative	Negative	Negative
C	1	27/02/2019	Positive	Positive	Faint	Negative	Negative
C	2	27/02/2019	Positive	Positive	Faint	Negative	Negative

3.4 Discussion

An accurate and sensitive diagnostic test was created to detect KOPV within symptomatic plants. This test was applied to 373 karaka leaf samples predominately from around Auckland, and the virus was detected in 92.2% of symptomatic samples tested. Creating an accurate diagnostic assay that can be applied to many samples was an important first step to detecting the virus presence in Auckland and New Zealand. In nearly all previously described emaraviruses, primer pairs were designed for RT-PCR that were specific to the virus (Di Bello et al., 2016; Hassan et al., 2017; Mielke-Ehret & Mühlbach, 2012). Zheng et al. (2017) used the generic primers designed by Elbeaino et al. (2013) as well as primers designed to be specific to AcCRaV. No other paper took the approach to use a multiplex test to detect all emaravirus segments in one reaction. In most cases, primers were designed to only detect the RdRp or nucleocapsid protein coding RNAs. The test developed here was used on samples to determine the distribution of KOPV within Auckland.

3.4.1 CTAB vs. Spectrum extraction methods

The CTAB RNA extraction method was chosen over the Spectrum kit as, while the Spectrum kit had greater sensitivity, the CTAB method was less costly, which was a major factor when extracting many samples. There were unexpected products amplified in RNA 3 and RNA 4 for both CTAB and Spectrum that showed that both methods extracted similar quality RNA although the CTAB method had more unexpected products. This may have been to contamination resulting from initial inexperience in extracting nucleic acids from plant tissue. In later experiments these unexpected products appeared less frequently due to increased familiarity and skill in extracting RNA, with unexpected results being rare and not present in most cases when re-tested. Regardless, between CTAB and Spectrum there was no extraction method that gave fewer unexpected products. The Spectrum kit was more sensitive than the CTAB method and could amplify virus product in solutions at least an order of magnitude more dilute for every RNA segment-specific primer pairs. Both methods could amplify a virus product when RNA was diluted to 75×10^{-2} ng/µL using each segment-specific

primer pair. This was deemed an acceptable limit for use in a diagnostic capability. Other articles describing emaravirus specific RT-PCRs all appeared to use commercial RNA extraction kits that are regularly used in their labs, although some do not specify how nucleic acid was extracted at all. In future experiments, it will be worth considering using the Spectrum kit when a greater sensitivity is needed.

The CTAB RNA extraction method, while being cheaper, is a laborious process and a more rapid approach would save time, allowing more samples to be extracted in a given time frame. This is a trade-off that was acceptable with developing the diagnostic test for this project; however, it is worth noting that the Spectrum kit would be chosen if possible as it provides greater sensitivity and faster RNA extraction. Protocols for commercial RNA extraction kits have fewer steps in general, which may reduce the potential for contamination. In this project entire sets of experiments had to be repeated due to contamination in the RNA extraction step as many samples were processed in a single day that increased the chance for error. While repeating these experiments had a lower financial cost than the Spectrum kit, the time spent did prove costly. A typical CTAB extraction took two days to complete and preparing the standardised aliquot, completing the RT-PCR diagnostic test and running the products through gel electrophoresis took substantial time as well. Contamination in samples was not detected until this final step. Then it was a matter of troubleshooting the entire process to see where the error had occurred. In most cases, it was the RNA extraction step, which meant a complete repeat of the whole process starting at extraction of RNA from leaf material. If the Spectrum kit was used it may have resulted in less time spent on these repeats in two ways: firstly, fewer steps and a simpler extraction process may have reduced the chance for errors and contamination, secondly, if a repeat was necessary, the time taken to do this would be drastically less than the CTAB extraction method.

There are other extraction methods that may have been considered but were not explored due to the unavailability of equipment or were shown to be ineffective. The KingFisher Purification System (ThermoFisher Scientific, Massachusetts, USA) allows total nucleic acid to be extracted quickly by automation. This would have significantly reduced the time to extract RNA in this project; however, the cost of the machine and the reagents required

are prohibitive. Another way could be using an even cruder extraction method than CTAB. Thomson and Dietzgen (1995) developed a crude extraction technique to rapidly release virus particles from plant material without using organic solvents or ethanol precipitation. This method was tested unsuccessfully on lettuce necrotic yellows virus (LNYV) infected *Nicotiana glutinosa* where all results were negative, even though the virus was shown to be present previously (Ajithkumar, 2018). The reason for the false negatives may have been due to the inability of the extraction method to remove PCR inhibitors from the final RNA solution. Additional time taken to optimise the process may have resulted in success. Another method has been described for potyviruses where crude RNA can be extracted in as little as 30 minutes to be used with a reverse transcription-recombinase polymerase amplification assay (Silva et al., 2018). In future KOPV experiments, these methods could be attempted and combined with a common commercial PCR inhibitor removal kit, such as the Zymo™ OneStep PCR Inhibitor Removal Kit, as a rapid assay to detect whether the virus is present.

3.4.2 NAD5 Transcript Detection as an Internal Control

Primers that were originally designed to amplify NAD5 transcripts in grapevine (*V. vinifera*) were adapted for use to test the quality of RNA extracted from Karaka. Internal controls are helpful when testing RNA extracted from plants using PCR methods to ensure the extracted RNA is amplifiable and that a negative result is truly a negative result (Menzel et al., 2002). This was especially important when attempting to detect KOPV in karaka, as the correlation between virus presence and symptoms was not well understood. As the NAD5 primers used in this research were designed to bind to grapevine mRNA, testing on karaka RNA had to be done before the primers were employed as part of the diagnostic test. While initial tests using one-step RT-PCR failed, a two-step reaction was able to amplify the expected product. For diagnostics, one-step RT-PCR is often preferred as it takes substantially less time. Therefore, time was spent troubleshooting the lack of amplification from the one-step RT-PCR. After adjusting the annealing temperature and repeating the one-step reaction several times it was noted that the reaction only amplified products on one of the three thermocycler machines, the Eppendorf Mastercycler® Gradient. Further investigation revealed that the temperature ramp time for this machine was lower than the other two machines. This allowed

the reaction to take place at the correct temperatures for longer and thus amplified the product. Thus, this machine was used in all subsequent NAD5 one-step RT-PCRs. In the future, conditions of the one-step RT-PCR could be altered to allow the reaction to run on any machine. Increasing the denaturation, annealing, extension, and final extension times could result in successful amplification.

Ideally, amplification of NAD5 should be tested for each sample to mitigate the chance of false negatives being reported. In this present research, only a subset of the samples tested with the diagnostic assay were tested with the NAD5 primers, and so it is possible that some of the negative results reported were false negatives. This is possible as, although the subsets tested were mostly positive for NAD5, there was a small percentage that failed. The absorbance A_{260}/A_{280} value for these samples was lower than the value of 2.0 suggesting the RNA was not pure and contaminants affecting the PCR could have been present. The only way to be sure would be to repeat the RNA extraction and NAD5 RT-PCRs. Future research could attempt to create a duplex reaction with both virus diagnostic primers and NAD5 primers present. Running the diagnostic test in a duplex reaction with NAD5 primers amplifying different sized products would save in time and costs. In the current literature focused on emaraviruses, it does not appear that there are any assays that simultaneously detect specific emaraviruses and internal control transcripts. Furthermore, use of internal controls has not been reported when new emaraviruses have been described. This may be because determining if extracted nucleic acid is competent is a common practice and assumed to be happening as a part of routine research.

3.4.3 Individual vs. Multiplex RT-PCR reactions

The multiplex reaction using all RNA segment-specific primer sets in one reaction amplified virus specific products of similar size that were observable when using gel electrophoresis. In the comparison between single and multiplex reactions, shown in Figure 3.8, RNA 2 and RNA 5 in one of the symptomatic samples failed to amplify a product. If these segments were solely used to determine whether the virus was present, then these samples

would have been falsely labelled as negative. The multiplex reaction removes this outcome as all segments are detected if present. Interestingly, in the segment-specific reactions for RNA 3 and RNA 4, unexpected products were present in asymptomatic leaves that were not present in the multiplex reaction. This could be due to contamination when setting up the RT-PCR reaction, resulting in a false positive. However, if these virus segments are truly present in these samples, then this could be due to the multiplex reaction having the opposite effect as intended – reducing the sensitivity of the assay. Although the sensitivity of detecting the individual segments was determined in this present research, the sensitivity of the multiplex reaction was not measured against these results. It may be useful in the future to test the sensitivity of the multiplex reaction as a diagnostic assay.

The multiplex reaction is cumbersome and may not require all segment-specific primer sets to be included for it to be sensitive enough to be used as a diagnostic test. The large concentration of primer in the reaction causes formation of large amount of primer dimer that, while not impacting the readability of the results, could potentially be interfering in the reaction. Previously described emaraviruses typically used only one set of primers targeting one of the RNA segments in any detection tests. RNA 1 or RNA 3 were typically used and were sensitive enough to determine if the virus was present for RYRSaV, AcCRaV, BLMaV and JYMaV (Di Bello et al., 2016; Hassan et al., 2017; Yang et al., 2019; Zheng et al., 2017). These segments appeared to be targeted due to the sequence information available. RNA 1 is typically more conserved with regions being highly conserved within and between emaravirids, hence why it was chosen as the target for emaravirus generic primers (Elbeaino et al., 2013). However, in HPWMoV isolates sequenced from different states in the USA show that similarity of RNA 1 within a species can be as low as 83.4% with large geographical separation (Stewart, 2016). As only one KOPV genome has been sequenced to date, it is difficult to compare to the intraspecies diversity of other emaraviruses. However, it may explain why some segments do not amplify in some samples. More complete genome sequences are needed to understand the intraspecies diversity of KOPV. This would help in determining the effectiveness of each individual segment-specific primer and would help in determining which would be the minimum required for a diagnostic test.

Multiplex reactions in PCR in general can be problematic and should be planned and executed carefully. In usual cases, primer concentrations should be adjusted to allow an even amplification of all targeted sequences (Sint et al., 2012). This was not necessary when creating the KOPV diagnostic test. Balancing product sizes was not required as the products from each segment were roughly the same size and only one was needed to be amplified to be declared a positive result.

3.4.4 Emaravirus Degenerate Primers

The emaravirus specific generic primers developed by Elbeaino et al. (2013) detected KOPV, and the generic primers developed by Olmedo-Velarde et al. (2016) did not. The 2013 generic primers target the motif A and C regions, which are present and conserved across all known emaravirids (Elbeaino et al., 2013; Yang et al., 2019). This further supports KOPV belonging to the *Emaravirus* genus, as KOPV contains these motifs and can be amplified by the 2013 generic primers. The 2016 generic primers did not amplify a product for any of the symptomatic plants, which may be due to multiple factors. The sequences of the primer pairs are quite different compared to KOPV RNA 1 genomic target, meaning that the primers probably could not bind to the target sequence resulting in failed amplification. Conditions of the RT-PCR, such as annealing temperature and annealing time could be adjusted and tested. A gradient RT-PCR with different annealing temperatures to determine the optimal binding conditions for the primers could be tried. Another reason could be that the assay was adapted to a one-step RT-PCR, where Olmedo-Velarde et al. (2016) had used a two-step RT-PCR. Future studies could trial a two-step reaction with the 2016 generic primers and KOPV RNA to determine if a product could be amplified. It could be that the sequence identity of KOPV is too dissimilar to the 2016 primers and, now with the completed KOPV genome, the primers could be adjusted to include the new sequence. Indeed, the original research presented for the 2016 generic primers only tested six of the fourteen currently known emaravirids, so there is room to develop primers that could amplify each virus (Olmedo-Velarde et al., 2016).

The emaravirus specific generic primers described by Elbeaino et al. (2013) have been tested on many of the known emaravirids and could be used in a diagnostic test for KOPV.

Zheng et al. (2017) used these generic primers targeting RNA 1 in addition to specific primers targeting RNA 3 of AcCRaV in separate reactions, to discriminate against tomato zonate spot virus infected kiwifruit (*Actinidia* spp.). The generic primers have also been used in research on RRV, EMARaV, PPSMV-1 and PiVB (Buzkan et al., 2019; Di Bello et al., 2015; Elbeaino et al., 2014; Robbach et al., 2015). As the generic primers amplify a product in symptomatic karaka and not in asymptomatic leaves, they could be used as the diagnostic test. Comparisons would need to be done to the KOPV diagnostic test described in this research regarding sensitivity, accuracy, and reliability with many different samples to ensure robustness. The products amplified would also need to be sequenced to confirm that the target was KOPV and not a separate emaravirus infecting karaka.

3.4.5 Karaka Emaravirus Diagnostic Test

The multiplex one-step RT-PCR using all the segment-specific primers with RNA extracted from karaka using the CTAB method was chosen to be the diagnostic test to detect the emaravirus presence. This was chosen as it reduced the chance for false negatives where some segments in a sample did not amplify when other strands did amplify. This test can be refined in the future to possibly include fewer primer sets targeting fewer segments. In the current research into emaraviruses, there are no instances where detection assays have used a multiplex approach. Only single segments have been targeted, usually RNA 1, the RdRp, and RNA 3, the nucleocapsid (Di Bello et al., 2015; Elbeaino et al., 2013; Yang et al., 2019; Zheng et al., 2017). Testing the segment-specific primers for RNA 1 and RNA 3 for KOPV on the already extracted RNA could result in a diagnostic test that will save time and cost of materials.

Alternative PCR technologies could be used to develop a diagnostic test to detect the emaravirus in karaka, such as quantitative real-time RT-PCR (RT-qPCR) or loop-mediated isothermal amplification PCR (LAMP). RT-qPCR uses fluorescent probes that allow light to be emitted when the target strand is synthesised in the reaction. It is often the method of choice for contemporary diagnostic applications requiring sensitive and fast detection, as well as quantification, of target organisms (Khan et al., 2015). Developing a RT-qPCR diagnostic test

for KOPV would reduce the time taken to get results when running many samples at once. It would also be beneficial to any future studies aiming at understanding how concentrations of viral RNA can vary between samples as qPCR can be used to quantify starting DNA and RNA. The ability to multiplex the virus specific reaction with primers targeting an internal control transcript can also be helpful in reducing time taken as well as verifying the extracted RNA integrity (Khan et al., 2015). LAMP, first described in the year 2000, is a technology that has been used extensively in virus diagnostics by being able to amplify a target sequence at a static temperature (Notomi et al., 2000). LAMP amplifies DNA with high sensitivity under isothermal conditions (typically between 60°C and 65°C) using a combination of 4-6 sets of specific primers and a DNA polymerase that extends strands concomitantly with dissociation of strands (Ishikawa et al., 2015). The method has been adapted for RNA by adding a reverse transcription component, making it extremely useful to detect viruses in the field without access to a laboratory. LAMP assays have been developed to target RRV and FMV with accurate results. The isothermal assay developed for RRV was able to accurately and specifically detect virus in multiple tissue types from multiple isolates, with a detection limit of 1 fg/µL. The entire process was rapid, taking less than 25 minutes including extraction, all done with less sophisticated equipment than found in a laboratory (Babu et al., 2017). Similar results were reported for the isothermal assay developed for FMV, where the assay could detect differences between geographically distinct isolates in around 15 minutes. Results were comparable to RT-PCR tests on isolated total RNA (Ishikawa et al., 2015). The drawback to LAMP is the complexity of designing an assay that combines 4 to 6 primer sets. Laboratories that already have RT-PCR and RT-qPCR facilities can already run specific diagnostic tests that they are familiar with, so there is little need to develop a LAMP test. Now that KOPV genome is complete, future research could focus on developing an isothermal test to take advantage of the rapid screening of large quantities of samples in the field.

3.4.6 Survey and Sampling

Karaka is distributed widely within Auckland and both symptomatic and asymptomatic trees appear to be evenly distributed. One aim of this chapter was to determine if there was an origin point of the virus; if there is one, it cannot be identified based on the survey presented in this research. One of the main problems with this study is that it was not known if the trees surveyed were naturally sown or planted from nurseries. This makes it difficult to narrow down an origin point as the virus could have naturally spread through a nursey, followed by an artificial spread by human activity to the rest of Auckland and potentially wider New Zealand. Future research could investigate ways of differentiating naturally sown trees from artificially planted trees to determine the natural spread. This may be difficult to achieve as the planted trees may have influenced the natural spread of KOPV, assuming the virus is endemic to New Zealand. Survey and sampling focused on distribution over abundance and so areas that could potentially have a denser abundance of symptomatic trees were never revealed. Future studies focusing on determining the abundance of symptomatic karaka in Auckland could give insight as to where the origin point of the virus is. Measuring how many symptomatic trees were in multiple 2 km^2 areas and measuring the distance between symptomatic trees would help. The distribution presented here is indicative for Auckland and cannot be applied to the greater area of New Zealand. It could be that the virus originated outside of Auckland.

The survey in this present research was a basic first step into determining the distribution of symptomatic karaka in New Zealand and it was found that the symptoms were widespread around Auckland. A limitation of this survey was that it was done predominately on foot by one person using only first-person observations. Private property and sacred places (the *urupa* at Orakei Beach, for example) were avoided even though there were signs that there was symptomatic karaka present. The number of symptomatic leaves present on trees varied greatly between each karaka tree. Some trees had only a few symptomatic leaves whereas others had most of the foliage covered with symptoms. Karaka trees in public spaces were searched thoroughly for symptoms; however, there are potentially some of these that were missed and resulted in the tree being falsely recorded as asymptomatic. A major problem encountered were tall trees that could not be completely checked without a ladder or cherry picker, for example. This could be done in the future although the time taken might be prohibitive when attempting to cover a large territory.

3.4.7 Testing Samples

A total of 373 samples were collected from the surveyed locations around Auckland and the upper North Island. There were 115 samples collected from symptomatic leaves, 115 samples collected from asymptomatic leaves from symptomatic trees, and 143 samples collected from asymptomatic leaves from asymptomatic trees. After repeating the diagnostic test on samples with unexpected results, 106 of the 115 (92.2%) symptomatic samples tested positive for KOPV. Contrasting this with zero of 115 (0%) asymptomatic leaves from symptomatic trees testing positive, and three of 143 (2.1%) asymptomatic leaves from asymptomatic trees testing positive, shows that the virus presence appears to be highly associated with symptom presence. Other recently reported emaraviruses have a similarly high rate of virus positive results from symptomatic leaves. Di Bello et al. (2016) reported that RYRSaV was detected in 100% of redbud (*Cercis* spp.) with symptoms and Laney et al. (2011) also reported that RRV was detected in 84 of 84 (100%) rose (*Rosa* spp.) plants, suggesting an absolute association for both viruses. Yang et al. (2019) reported that JYMaV was detected in 37 of 45 (82.2%) symptomatic jujube (*Ziziphus jujube*) trees tested in which the authors state there is a close association between virus and symptoms. The JYMaV also had only 1 in 33 (3%) asymptomatic trees test positive These results resemble those obtained for KOPV. The association of virus and symptoms with AcCRaV in kiwifruit at 38 of 146 (26.0%), and BLMaV in blackberry plants (*Rubus* spp.) at 225 of 524 (43%), were lower than this due to the symptomatic samples that were tested having other species of virus that may have caused the symptoms (Hassan et al., 2017; Zheng et al., 2017). There have been two other viruses (badnaviruses) detected in the karaka trees where symptoms were present and further work needs to be done to determine if they have an association with the symptoms as well. Regardless, at this stage the symptoms present on karaka leaves have been shown to be associated with KOPV. Further analysis of this is presented in Chapter Four.

A small percentage of asymptomatic leaves from asymptomatic trees returned an unexpected positive result where virus was present without symptoms. As the diagnostic test

process was repeated for these samples, starting with the RNA extraction, it is unlikely the result is due to contamination. These samples may have been virus infected but had not had the chance to display symptoms yet, or that the leaves had symptoms that were too small to observe. This raised questions on how the virus moves within and between karaka hosts and if it could be detected in other parts of the plants such as the phloem. In the symptomatic leaves tested, only the symptomatic region was taken and tested, but a question was raised as to whether the virus could be detected in the asymptomatic regions of the symptomatic plants. This is addressed in this present research in 0 Correlating Virus and Symptoms.

3.4.8 Determining Virus Detection in Different Storage Conditions for Leaf Samples

Due to the limitations of the survey in this research and the need for a large quantity of survey data and samples, sampling instructions were developed to employ the help of interested citizens from around New Zealand. Citizen science is equated with a partnership between scientific research and public participation that has involved millions of people helping with thousands of scientific projects around the world (Bonney et al., 2016). A sampling instruction flyer (Appendix D) to distribute was created to raise awareness and help people without scientific backgrounds to sample and send to the lab. The flyer was not sent out due to a long wait on confirmation from some stakeholders in the project; however, the sampling instructions with information obtained from testing samples in potential mailing conditions in the lab. The best way to send a sample was to put it in a zip-locked bag and mail it, and when the sample arrived to refrigerate it at 4°C. Symptomatic samples stored this way were able to be successfully tested for virus presence. The critical factor seemed to be preserving the moisture in the leaf as samples stored in paper envelopes dried out too fast and virus was not able to be detected. A damp tissue included in the zip-locked bag might preserve the sample longer at room temperatures, but this likely unnecessary.

3.4.9 Determining Optimal Sampling Time and Viral Titre

Samples were collected from a single tree at PFR over a period of seven months and were intended to be used to determine the viral titre at these time intervals. This proved difficult to do as leaves gathered from each branch of the tree had varying symptoms, some with more than others, making it almost impossible to measure the viral titre with conventional RT-PCR methods. RT-qPCR could be explored for this experiment in the future as it could be used to relatively quantify the viral RNA present in each sample. There appears to be little research into viral titre studies for emaraviruses as most of the RT-qPCR applications focus on detection. Katsiani et al. (2018) developed an RT-qPCR assay to successfully determine relative viral titre quantities of *Little cherry virus 1* in *Prunus* spp. throughout the year finding that early spring was the best time to sample for this virus. This research could be used as a framework in KOPV research to find when the viral titre is highest. The samples collected from the karaka tree were used in a dilution series to attempt to gain information on the best time for sampling, in which all time periods seemed to be equal. A year-long study would be essential to capture every season unlike this present research, which only captured winter, spring and summer. Unfortunately, this experiment used up all the symptomatic plant material sampled, and future work would need to repeat this lengthy sampling process.

The movement of the virus within the karaka tree is something that needs to be considered in future work as taking leaves from the same branch each month may have interfered with the propagation of the virus. Initial branches chosen had a high density of symptomatic leaves and it was noticed each time the branch was sampled that no new symptoms were appearing on the leaves. The karaka gall mite is active throughout the spring and summer months and therefore new symptoms were expected through the mite spreading and feeding on new leaves (Martin, 2017). It is also possible that the karaka tree serves as a passive vector for the emaravirus, which could potentially be a mite-based virus, rather than a plant-based virus. Species of *Cripaviruses* infecting aphids employ this transmission method with the virus infecting and replicating within the insect when it feeds upon the plant (Jones, 2014). If this is the case with KOPV, then removing symptomatic material would be removing the vector and could possibly result in lower chance of mites carrying the virus to other leaves.

This would imply that the virus does not move through the plant on its own which could be possible and will be discussed in the next chapter.

Chapter Four: Correlating Virus and Symptoms

4.1 Introduction

Chapter Two revealed that the negative-sense RNA KOPV likely belongs to the emaravirus genus and Chapter Three revealed that the pale circle symptoms on karaka leaves are associated with the virus. It has been shown in previous literature, such as research presenting RYRSaV, RRV and JYMaV, that high rates of detection of the virus in symptomatic leaf material suggests a close association between virus and symptoms (Di Bello et al., 2016; Laney et al., 2011; Yang et al., 2019). In this Chapter, research is presented using Koch's postulates to further correlate the symptoms with KOPV by testing the asymptomatic and symptomatic sections of a single leaf to detect the virus presence. In Chapter Two, it was also noted that the movement protein could potentially have an impaired function that would limit the systemic movement of the virus through the host. In this chapter, the movement of the virus is tested by attempting to detect presence of the virus within the phloem.

4.1.1 Koch's Postulates

Koch's postulates are four conditions that must be fulfilled to determine whether a biological agent is the cause of a disease. Fulfilling some of these conditions can also add evidence that an agent is associated, or correlated, with the disease. Koch's postulates, initially described in 1884, were considered inapplicable to viruses due to these biological agents being unculturable (Bos, 1981; Inglis, 2007). Increase in knowledge around viruses and advancement of technology resulted in an adapted set of postulates developed specifically for plant viruses in 1981, shown in Table 4.1 (Bos, 1981). This new adapted set of postulates has been adopted and used since, and has even been further adapted to use with viroids, the smallest biological agent able to cause infection in plants (Di Serio et al., 2018). Fulfilling all four of these postulates allows causality of the symptoms caused by the virus to be determined.

Table 4.1 – Koch's Postulates adapted for viruses as described by (Bos, 1981)

1	The virus must be concomitant with the disease
2	It must be isolated from the diseased plant: • separated from contaminating pathogens, • multiplied in a propagation host, • isolated from plant sap and purified physico-chemically, and • identified for its intrinsic properties
3	When inoculated into a healthy host plant, it must reproduce the disease
4	The same virus must be demonstrated to occur in and must be re-isolated from the experimental host.

4.1.2 Virus Movement

Virus movement through a host plant is referred to as systemic movement. It is where the virus moves from cell-to-cell via the plasmodesmata, which are membrane-lined channels that penetrate the cell wall of each plant cell, allowing nutrient and signal molecules to travel throughout the plant (Hipper et al., 2013). Some emaraviruses have been shown to be able to move systemically through a host plant, spreading to other areas from where the initial infection is established. In RLBV, the P4 protein encoded by RNA 4 was shown to be involved in movement through experimental and bioinformatic methods (Yu et al., 2013). *In silico* analysis revealed the presence of a conserved domain that shared homology with the 30K superfamily of other viruses, which has been shown to be involved in movement (Mushegian & Elena, 2015). Since then, a domain called emaravirus P4 superfamily (pfam16505) has been established as a conserved domain and is present in the P4 of all known emaravirids to date. It is not present in KOPV (as discussed in Chapter Two). The lack of the conserved domain may suggest that the systemic movement of KOPV through the karaka tree has been impaired, or that the RNA has not been correctly identified.

4.1.3 Aims

In the previous chapter it was established that there is a high correlation between symptoms on a karaka leaf and KOPV. For this chapter, to complete the aim of assessing the correlation between the symptoms and virus presence, the objectives were twofold. The first was to further determine the correlation between symptoms and virus by testing symptomatic and asymptomatic regions on an entire symptomatic leaf. It was hypothesised that the virus would only be detected in the symptomatic regions while the asymptomatic regions would be virus free. The second objective was to begin an investigation into the systemic movement of KOPV. This was done by testing asymptomatic petiole parts of heavily symptomatic leaves to determine if the virus had entered the phloem and was potentially moving through the plant.

4.2 Method and Materials

4.2.1 Correlating Virus and Symptoms – Testing whole leaf

An entire symptomatic karaka leaf was divided into sections that contained symptomatic and asymptomatic regions and RNA was extracted from each of these sections to determine whether the virus could be detected in both symptomatic and asymptomatic regions of the same leaf.

A total of ten symptomatic leaves were selected from the PFR0001 tree that had discernible symptomatic and asymptomatic regions, and were of similar size. These leaves were labelled as whole leaf 1 – 10. The leaves were washed with 70% ethanol and equally divided up into sections that were between 200 and 300 mg and RNA was extracted from each section using the method outlined in section 3.2.2.1. An example of a leaf divided into eight sections is shown in Figure 4.1, although some larger leaves were divided into ten sections. The diagnostic test developed in Chapter 3 was used to detect the virus in each RNA sample from each section. Care was taken to avoid taking symptomatic regions that were close to the midrib and petiole.

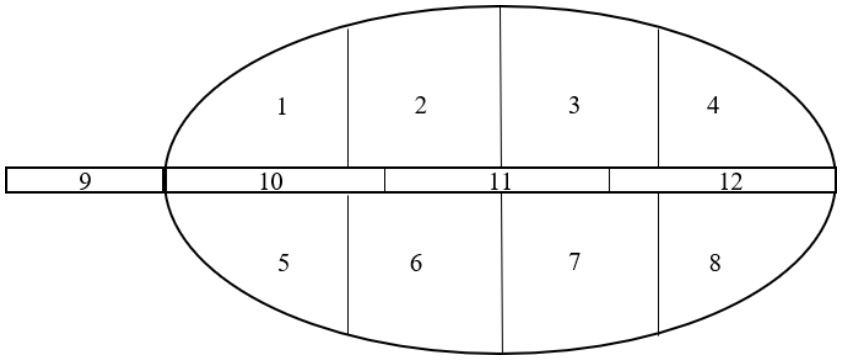


Figure 4.1 – Example of the division of a leaf to test correlation of symptoms including the midrib and petiole.

4.2.2 Testing Petiole for Systemic Virus Movement

To determine if KOPV could move systemically through a leaf and into the petiole, a symptomatic leaf that had symptoms present on, or close to, the midrib was divided into sections that had no other leaf material attached to it including the petiole. Five additional leaves labelled as petiole leaf 1 – 5 with extensive symptoms were selected. The leaf was then divided into sections as specified in section 4.2.1 and tested using the diagnostic test outlined in Chapter 3. Refer to relevant sections in Figure 4.1.

4.3 Results

4.3.1 Correlating Virus and Symptoms – Testing entire leaf

While KOPV has been found in most symptomatic samples tested from around Auckland, these samples had only their symptomatic regions tested. To further correlate the virus with the symptoms, a whole leaf was extracted by sectioning it into symptomatic and asymptomatic parts. The diagnostic test described in Chapter 3 was then used to detect the virus from RNA extracted from each section. The hypothesis was that the virus would only be amplified from symptomatic regions of a leaf. Ten leaves were chosen that had symptoms that did not completely cover the leaf so that there could be a separation of symptomatic and asymptomatic sections. An example of the leaf sectioning is shown in Figure 4.2 as it was done with leaf 3 and an example gel of the results of the diagnostic test for this leaf is shown in Figure 4.3.

Figure 4.2 – An example showing how whole leaf 3 was sectioned. The bottom leaf is the same as the top leaf with the outline of each section overlaid.

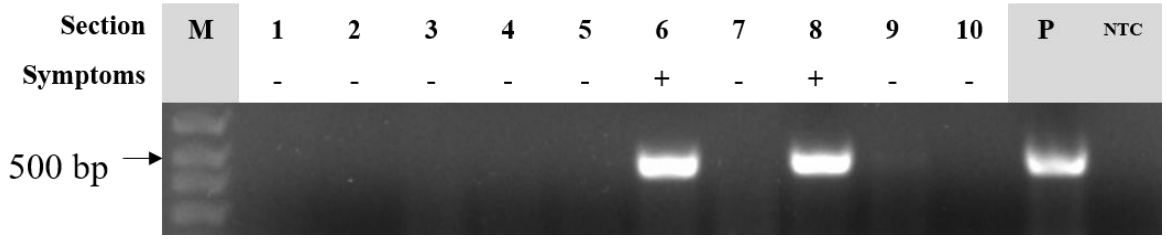


Figure 4.3 – KOPV amplification from each section of whole leaf 3. Lane M: 1 kb Plus DNA ladder. Lane P is the positive control (L19b-N). Lane NTC: No template control.

Table 4.2 below shows the diagnostic test results for each of the ten leaves. All sections of leaves 3 and 5, both symptomatic and asymptomatic, appeared to contain the virus – most likely due to contamination. Leaf 10 had an unexpected positive result in one asymptomatic section and a negative result in one symptomatic section. All the other leaves amplified virus products in symptomatic sections only. In the midrib sections, leaves 3, 4 and 6 returned a weak positive, and leaf 8 reported a strong positive. In the petiole sections, leaf 9 and 10 returned a positive result.

Symptomatic sections all amplified a virus product except for one section on leaf 10. This suggests a strong correlation between the virus and symptoms. Some petiole and midrib sections amplified a product where there appeared to be no symptoms. This may mean that the virus is moving out of the leaf and into the phloem, or it could just mean that there may be symptoms close to the midrib that were picked up when sectioning the leaf. To show that the virus may be moving into the phloem, the petiole would need to be sub-sampled separate from the midrib and tested.

Table 4.2 – Results of the KOPV diagnostic test run on RNA extracted from asymptomatic and symptomatic sections of a karaka leaf. (W) indicates a weak band on the gel was observed. The percentage asymptomatic, symptomatic, midrib and petiole positive results are listed at the bottom of the table.

	Asymptomatic Sections		Symptomatic Sections		Midrib Section		Petiole Section	
Whole Leaf	No. of Sections	Positive (%)	No. of Sections	Positive (%)	No. of Sections	Positive (%)	No. of Sections	Positive (%)
1	6	0	4	100	1	0	1	0
2	3	100	5	100	1	100	1	100
3	6	0	2	100	1	100 (W)	1	0
4	2	0	6	100	1	100 (W)	1	0
5	4	100	4	100	1	100	1	100
6	5	0	3	100	1	100 (W)	1	0
7	5	0	3	100	1	0	1	0
8	3	0	5	100	1	100	1	0
9	4	0	4	100	3	0	1	100
10	2	50	6	83.3	1	0	1	100
Total	40	20.0%	42	97.6%	12	50.0%	10	40.0%

4.3.2 Testing Petiole for Systemic Virus Movement

The virus has been shown to be strongly associated with the pale circle symptoms present and was rarely found outside of the symptoms. In the previous experiment, virus was amplified in the midrib and petiole sections where symptoms were absent, raising the question of whether the virus is moving systemically through the plant even though it was not detected in asymptomatic leaf sections. Five additional leaves were chosen (labelled as petiole leaf 1 to 5) with extensive symptoms that were also present on the midrib but were not observable on the petiole. An example of the leaf sectioning is shown in Figure 4.4, where the symptomatic petiole of leaf 5 was sectioned prior to RNA extraction and Figure 4.5 shows a gel with the results of the diagnostic test applied to each section. All sections were extracted and tested to add further evidence of correlation between the virus and symptoms.

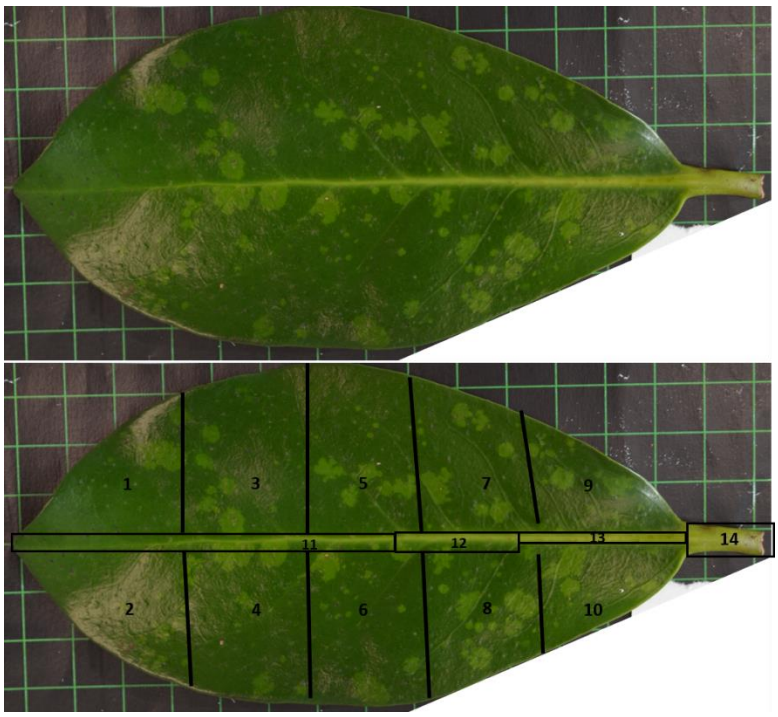


Figure 4.4 – An example showing how petiole leaf 5 was sectioned. The bottom leaf is the same as the top leaf with the outline of each section overlaid. Note that the petiole is taken separate from the rest of the leaf and appears to have no symptoms.

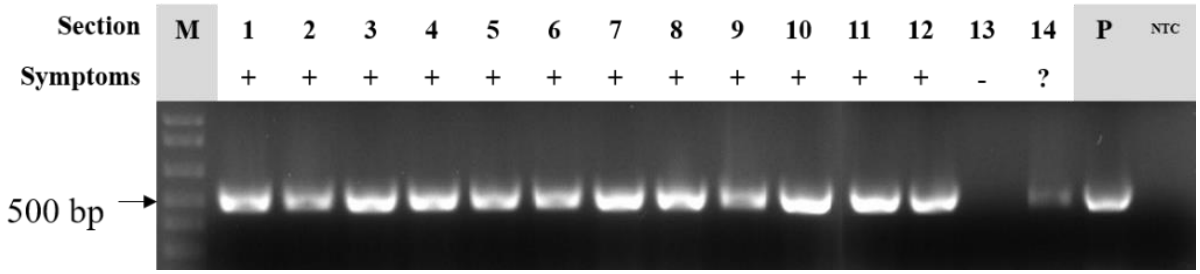


Figure 4.5 – An example 2% agarose gel showing the results of the diagnostic test for sections of leaf 5. Lane M: 1 kb Plus DNA ladder. Lane P is the positive control (L19b-N). Lane NTC: No template control.

Table 4.3 shows the diagnostic test results for each of the five leaves in this petiole experiment. Once again, all five leaves show strong correlation between the presence of the virus and symptoms with all symptomatic regions amplifying a virus product. Due to the selection of leaves focusing on obtaining heavily symptomatic leaves with symptoms on the midrib, there were few asymptomatic regions. Most importantly, all petiole sections returned a positive result.

Table 4.3 – KOPV detection in RNA extracted from asymptomatic and symptomatic sections of petiole leaves 1-5. Symptomatic, midrib and petiole positive results are listed at the bottom of the table as a percentage.

Petiole Leaf	Asymptomatic Sections (including midrib)		Symptomatic Sections (including midrib)		Petiole Section	
	No. of Sections	Positive (%)	No. of Sections	Positive (%)	No. of Sections	Positive
1	7	0	4	100	1	Yes
2	0	0	11	100	1	Yes
3	0	0	7	100	1	Yes
4	0	0	7	100	1	Yes
5	1	0	11	100	1	Yes
Total	8	0	40	100	5	100 %

The virus can be detected in the petiole where no symptoms are observed indicating that the virus can potentially move systemically through the leaf if symptoms are present on the midrib. Contrasting this with the results from the whole leaf testing in 4.3.1 where the petiole did not return a virus positive result most of the time, it may mean that the virus can move systemically through the karaka tree only if symptoms are present on the midrib. An alternative explanation may be that not all of the mites were removed during the leaf cleaning process and what was detected was a mite carrying the virus.

4.4 Discussion

Symptomatic and asymptomatic sections of ten Karaka leaves were tested for the presence of KOPV using the diagnostic test described in Chapter Three. This was done to determine whether the virus could be found in the asymptomatic sections as well as the symptomatic sections to satisfy Koch's first postulate – the virus being concomitant with the disease. The results showed that the virus is strongly correlated with the symptoms and that the virus was rarely found in the asymptomatic sections. Interestingly, in some of the leaves there was a positive result returned for samples containing the petiole section, even though there were no symptoms present on this part. This raised questions about whether the virus can systemically move throughout Karaka. The same experiment was repeated on five heavily symptomatic leaves with the difference that the petiole of each leaf was taken as a separate section, with no leaf or midrib material attached to it. These leaves had symptoms on the midrib and when tested it was found that the petiole returned a positive result on all five leaves. This suggested that the virus may be able to move systemically through the phloem of the karaka leaf.

4.4.1 Correlation of virus and symptoms

KOPV appears to be strongly correlated with symptoms. As shown in Chapter Three section 3.4.7, previous studies of emaraviruses have shown strong correlation between virus and symptoms by being able to detect the virus in symptomatic leaves. This chapter describes taking this a step further by attempting to detect the virus in both asymptomatic and symptomatic regions of the same leaf in order to discriminate between virus infected and virus free leaf tissue. From this, it appears that the virus is strongly associated with the symptoms. Indeed, the virus was rarely found in areas without symptoms with only one of 40 asymptomatic samples testing positive that could not be ruled out by contamination in the diagnostic test process. This may have been due to several reasons, including contamination during the extraction process, or that the virus was present without symptoms or that a symptom was present but too small to be visible. Conversely, the virus was detected in 32 of

the 33 symptomatic samples, which demonstrates a strong correlation between symptoms and presence of the virus. Further work needs to be done to determine whether the two badnaviruses detected in the initial NGS work are not contributing to the symptoms. Developing a diagnostic test to detect the two badnaviruses and using it with symptomatic and asymptomatic leaves is required as a first step. Badnaviruses can be integrated into the host's genome, therefore this should also be investigated in the case of Karaka. Rolling circle amplification (RCA) is a method that can discriminate between episomal and integrated badnavirus genomes by only targeting the circular episomal dsRNA, eliminating any false positives that an integrated genome would pose (Bhat et al., 2016).

Koch's postulates have only been partially fulfilled so it cannot be said that the virus is causing the symptoms, it is only correlated with the symptoms at this stage. Based on the modified Koch's postulates for plant viruses by Bos (1981), only postulate one has been fulfilled. Postulate one states the virus must be concomitant with the disease, which this research has shown by correlating the symptoms and the virus using the diagnostic test. Postulate two states that the virus must be separated from contaminating pathogens. While the KOPV sequence has been separated from other sequences within an infected Karaka sample, the virion itself has yet to be purified. Therefore, postulate two as not been achieved. Further research would also need to be done to fulfil the remaining postulates. Postulate two also requires that the virus is multiplied in a propagation host, isolated from plant sap, purified, and identified it for its intrinsic properties. To do this, KOPV must be mechanically transmitted to an indicator plant such as *N. benthamiana*, which may prove difficult for an emaravirid. In early research it was difficult to mechanically transmit an emaravirus to commonly used indicator plants, only being successful with low transmission rates in a few species, such as PPSMV-1 and PPSMV-2, and more recently with RYRSaV (Di Bello et al., 2016; Mielke-Ehret & Mühlbach, 2012). More recent research has reported AcCRaV could be mechanically inoculated into *N. benthamiana* with seven out of fourteen plants testing positive for the virus, showing that it is possible (Zheng et al., 2017). If KOPV can be purified, not only could it be mechanically inoculated into a propagation host, it could also be inoculated into a healthy Karaka plant to cause disease and satisfy Koch's third postulate. Isolation of KOPV from this mechanically infected plant would fulfil Koch's fourth postulate. This could be achieved by creating cDNA clones of the virus and inoculating them into *N. benthamiana* using agro-

infiltration, where cDNA of virus sequence is inserted into plasmids, transformed into *Agrobacterium tumefaciens* that is used to insert the plasmids into the model plant host. In previous emaravirus research this was thought to be unviable since the virus has a multipartite and negative-sense genome (Di Bello et al., 2016). Recently, progress has been made with infectious clones of negative-sense RNA viruses with development of an infectious clone for RRV (Pang et al., 2019; Wang et al., 2015). Using the protocol developed for RRV, an infectious clone for KOPV could be created and inserted into a host model plant and attempts to rescue or detect the virus through the RT-PCR diagnostic test could be done. Rescue of the virus from a systemic leaf grown after successful inoculation could be difficult, however, due to the potentially restricted systemic movement by the virus.

The midrib and petiole sections sometimes amplified a weak product where there were no visible symptoms. Once again this might have been due to contamination due to some small, symptomatic regions from the leaf regions being picked up when sub sampling the midrib and petiole. It may also have been due to KOPV being present in the vector and carrying over unnoticed due to their microscopic size This is unlikely as the leaves were washed with 70% EtOH before being subsampled. Another possibility is that the virus may have been present but causing no symptoms, or the virus was present and was in the process of causing symptoms, but they had not manifested yet. The positive result could also have been caused by contamination. Indeed, the whole leaf extraction of leaves two and five in section 4.3.1 all returned a positive result even though some parts were clearly symptomatic, likely due to contamination in the RNA extraction step. This could also be true with the asymptomatic midrib samples. To avoid this in the future, a negative control could be used in the form of a buffer only extraction where no plant leaf material is added to the extraction tube and the process is carried out with buffer only. If the buffer control is positive, it would indicate that contamination has occurred in the RNA extraction process.

4.4.2 Systemic Movement of Virus

KOPV can be detected in the petiole if the symptoms are present on the midrib but is not detectable if there are no symptoms on the midrib, suggesting that the virus may be able to move through the phloem, but not enter it from the leaf naturally. Graft transmission experiments demonstrated that EMARaV, FMV, PPSMV-1 and AcCRaV virus particles can move systemically through host plants (Mielke-Ehret & Mühlbach, 2012; Zheng et al., 2017). This would suggest that as KOPV is an emaravirus, it would be likely move through the karaka tree, although it appears that the virus can be detected in the petiole only if there are symptoms present on the midrib. This may be due to the compromised sequence of KOPV P4 (the putative movement protein). As discussed previously (Chapter Two), the KOPV P4 does not contain the emaravirus P4 30k domain superfamily conserved domain, whereas all other emaraviruses do, suggesting that KOPV P4 may not function as a movement protein. As systemic virus movement relies on the viral movement protein modifying the plasmodesmata, the compromised P4 may not be able to do this, and the virus may only be able to enter the vasculature if it is directly on the midrib. Vascular transport of viruses differs between entry and exit, to and from, the phloem. Virus entry can occur in all vein classes of source leaves whereas virus exit is limited only to major veins of sink tissues, which suggests that there are likely different mechanisms for the virus entering and exiting the phloem (Hipper et al., 2013). With KOPV, it may be that the virus is able to enter the phloem in the midrib tissues more easily but is unable to exit again, essentially limiting it to, and from, the phloem. Further experiments are needed to confirm this with the first being to establish infection through graft transmission. If the virus can be transmitted this way, then the KOPV movement protein would likely be working as any other emaravirus movement protein. If not, then further experiments involving movement could be attempted such as the rescue of the movement protein. In these experiments, the movement protein of a virus is mutated so that it cannot modify the plasmodesmata to move the virus systemically through a model plant. The mutated virus is then co-infected into a model plant with a virus that has been demonstrated to move systemically. Some days later, the mutated virus is isolated from systemic leaves. Rescue of the movement of RLBV and other negative-sense RNA viruses has been successfully demonstrated, and the technique could be adapted to rescue the movement of KOPV using RLBV. (Yu et al., 2013; Zhou et al., 2019). If successful, then it may show that KOPV can

move systemically provided it has a functioning movement protein. A method for mechanical transmission of KOPV into a model plant would need to be developed first. An alternative mechanism of virus movement might rely on the mite vector. As the mite moves around the leaf while browsing it may be moving the virus from one location to another. The leaves chosen for the petiole experiment were heavily symptomatic and mites browsing on these leaves may have carried and deposited the virus all throughout the leaf, with no movement of the virus inside the leaf occurring due to molecular processes. The mites could even have browsed on the petiole depositing virus particles that were picked up by the RT-PCR diagnostic test.

An alternative hypothesis is that KOPV is a mite virus that uses the karaka tree as a passive vector. The assumption that KOPV is a plant virus is based on its homology to other emaraviruses. Little is known about how each emaravirid interacts with its eriophyid mite host other than through transmission. It may be that the karaka tree is acting as a passive vector for KOPV, which infects mites. Plant vectors facilitating horizontal virus transmission provides a way for viruses that infect mites to enhance their dispersal and survival (Jones, 2018). Two viruses belonging to the *Cripavirus* family, *Aphid lethal paralysis virus* and *Rhopalosiphum padi virus* are both known to infect their mite hosts and are transmitted by the mite feeding on the passive vector plant (Jones, 2018). Similarly, with KOPV, the karaka gall mite may become infected as it feeds on symptomatic karaka tree leaves. The mite would then migrate along the leaf, depositing the virus as it browsed on the leaf, allowing the virus to move locally and infect a new part of the leaf. This would explain the seemingly random appearance of symptoms, lack of symptoms on nearby leaves as the mites may not browse too far, and the lack of systemic movement. Indeed, it may be that the movement protein of KOPV has lost its conserved movement functions due to this interaction between the mite and virus. This hypothesis, while interesting, will rely on much work toward proving it and it will be difficult. Firstly, once again, whether KOPV is graft transmissible needs to be determined to understand if and how the virus moves within the karaka tree. Testing of roots and other parts of the tree can then be done to determine presence of the virus. Failure to establish graft transmission would provide evidence for this hypothesis. Once this has been determined subsequent research could be focused on mite transmission studies, determining firstly whether the karaka gall mite can transmit the virus and then secondly if the virus can replicate within the mite in a propagative manner. Research on these two transmission methods will then inform the

direction that additional research needs to go to further understand the virus-vector-host relationship.

Chapter Five: General Discussion

Emaravirus is a relatively new genus of plant viruses that includes species that have monocistronic, segmented, negative-sense RNA genomes, have double-membraned bodies as the virus particle, and are vectored by eriophyid mites (Mielke-Ehret & Mühlbach, 2012) The research described in this thesis has demonstrated that the novel negative-sense RNA virus found in a Karaka tree in Auckland, New Zealand, tentatively named *Karaka Okahu purepure virus* (KOPV) has all of the attributes required to belong to the emaravirus genus. The virus was found to be widely distributed within Auckland and the pale chlorotic symptoms were found to be highly correlated with the virus.

5.1 Completion and analysis of the genome

The first aim of this research was to complete and analyse the genome of KOPV to determine if it belonged in the emaravirus genus by completing two research objectives. The first objective was to verify the 3' and 5' ends of the genome and the second was to analyse and compare the completed genome to current emaravirids.

Completion of the KOPV genome was accomplished using a combination of 3' RACE and sRNA mapping to the initial Illumina sequenced reference genome, allowing the complete genome to be obtained. The nucleotide sequence of the 3' and 5' ends of the completed genome had the expected conservation found in other emaravirids. This approach has been used to obtain the genomes of other emaraviruses. Illumina sequencing, coupled with bioinformatic analysis, has been used to discover and/or sequence several emaraviruses, such as AcCRaV, RRV, HPWMoV, PPSMV-1 and PPSMV-2 (Zheng et al., 2017). In other studies of emaraviruses, such as AcCRaV, RRV and RYRSaV, RACE was used to verify the ends of the genome (Di Bello et al., 2016; Laney et al., 2011; Zheng et al., 2017). As previously discussed in Chapter Two, two emaravirus candidate species, PVBV and PiVB had not been confirmed with RACE and the 3' and 5' termini of these sequences appear to be incomplete (Buzkan et al., 2019; Ilyas et al., 2018). Therefore, RACE seems to be essential to verifying the genome. sRNA mapping to the original Illumina genome was also crucial for verifying the KOPV genome. The original primer design for the RACE experiment returned an ambiguous result

for very last nucleotide at the 3' and 5' termini. The sRNA mapping confirmed that last nucleotide, which was also verified by mapping the sequences to other emaraviruses. Although the sRNA mapping was useful in this situation, it may not be necessary for genome completion with better designed RACE primers, as discussed in Chapter Two. Alternatively, sRNA mapping could be used instead of RACE to verify the ends, but results obtained (not presented in this research) were not expected. sRNA mapping to the original sequences obtained from Illumina did not complete the ends of each RNA sequence to give the correct final nucleotides. The combination of both RACE and sRNA was necessary to achieve the completed ends. Further testing with the map to reference tool in Geneious could, however, lead to a method that allows verification of the ends of each RNA genomic segment without using the time-consuming RACE process.

Comparison of the KOPV genome sequence with genomes of other emaraviruses support KOPV as an emaravirus. KOPV has a segmented, negative-sense RNA genome comprised of five segments all encoding a single ORF. The final 13 nucleotides of the 3' and 5' termini are complementary and are suspected to hybridise to each other to form panhandles, as is common with other emaraviruses and negative-sense RNA viruses (Kormelink et al., 2011; Zheng et al., 2017). The termini sequences are also similar to that observed in other emaravirus genomes, giving further support to KOPV being an emaravirus. While the overall sequence identity is low between KOPV and the rest of the emaravirids, indeed having the lowest identity between all emaraviruses, the presence of conserved motifs and domains in translated ORF amino acid sequences strongly indicate that P1, P2 and P3 of KOPV are the RdRp, glycoprotein and nucleocapsid, respectively. KOPV P4 is missing the emaravirus movement protein superfamily domain, but it still shares similarity with PVBV which has this superfamily domain. The PVBV P4 shares similarity with other emaraviruses that provides weak evidence that KOPV P4 may be the movement protein (Ilyas et al., 2018). Current known emaraviruses have between four and eight segments in their genome. KOPV has five, but there may be more. Initial reports of EMARaV listed four segments where now there are six confirmed segments, while RRV initially had four segments reported and now seven segments are attributed to the virus (Di Bello et al., 2015; Mielke-Ehret & Mühlbach, 2012; von Bargen et al., 2019). This could mean that there are more segments in KOPV that are yet to be discovered. In other emaravirus research, primer pairs designed around the conserved termini

of each segment have been used to detect all segments except for RNA 1, which is too large to be amplified (Di Bello et al., 2016; Zheng et al., 2017). Similar primer pairs could be designed for KOPV to answer whether the virus has more genome segments. Additional segments found may reveal a sequence which contains the movement protein superfamily domain, as was the case for EMARaV (von Bargen et al., 2019). If this is the case, then the RNA 4 of KOPV described in this research may encode a protein with unknown function.

Based on species demarcation criteria, KOPV belongs in the emaravirus genus but is a separate species from other emaraviruses. The species demarcation criteria for emaraviruses is based upon the amino acid products of the RNAs encoding the RdRp, glycoprotein and nucleocapsid. If these sequences differ by more than 25% the virus belongs to a different species (Elbeaino et al., 2018). KOPV differs substantially more than this, with the closest RdRp sequence similarity being 31.4% with EMARaV. The closest glycoprotein precursor similarity is with HPWMoV at 21.6% and the closest nucleocapsid similarity is with JYMaV at 25.7%. The phylogenetic analysis showed that KOPV was outside of the three previously established clades described by Zheng et al. (2017) although it still shared a common ancestor with the emaraviruses that is distinct from closely related tospovirus and tenuivirus genera. KOPV has had a different evolutionary history from other emaraviruses, further supporting it as a distinct species. With the discovery of more emaraviruses the relationships between viruses in this genus will become clearer, and closer relatives of KOV may be identified.

Overall, KOPV has been shown to have several of the key defining characteristics of emaraviruses, such as the segmented, negative-sense RNA genome encoding amino acid sequences that resemble other emaraviruses but are different enough for it to be considered a separate species. Further, the genome has the expected conserved termini on each RNA segment. The karaka gall mite, an eriophyid mite, is the candidate mite vector for KOPV. Further research is needed (discussed in Section 5.5), but there is strong evidence that KOPV belongs in the emaravirus genus.

5.2 Diagnostics and Distribution

The second aim of this research was to determine the distribution of KOPV in Auckland by completing two objectives. The first objective was to create a diagnostic test that would accurately discriminate between virus-infected and virus-free leaf tissue. This was achieved with development of the KOPV diagnostic assay that could detected virus in 92.2% of symptomatic leaves (Table 3.7). While the virus could not be detected in 100% of symptomatic samples, a 92.2% detection rate does suggest a high correlation between the virus and symptoms as discussed in Chapter Three. The diagnostic test detected virus in 0% of asymptomatic samples taken from symptomatic trees, and 2.1% of samples taken from, what was observed to be, completely asymptomatic trees. The second objective of Aim 2 was to use the diagnostic test to survey the distribution of KOPV in Auckland. This was achieved with limitations due to the nature of the survey. Conclusions about the data were also limited due to not knowing whether the trees were naturally sown or planted from a nursery. If it was the latter, then the virus spread around Auckland could be artificial and any conclusions about the distribution would need to factor this in. Survey efforts could utilise data collected by Atherton et al. (2015) detailing the locations of karaka throughout New Zealand. The work they have done could be used to determine if there are any symptoms on plants that have not been accessed in this present research. The authors of that study have also suggested using specific genetic markers to determine the dispersal of karaka, which may be useful in determining whether the trees were naturally sown or through man-made means.

5.3 Correlation of Virus and Symptoms

The third aim of this research was to determine how closely the virus is correlated with the pale circle symptoms present on some karaka. This was achieved by completing objective 3.1, using the diagnostic test developed in objective 2.1 to test symptomatic and asymptomatic regions of a karaka leaf. Results from Chapter Four show that the symptoms are highly correlated with virus presence (Table 4.2) with 97% of symptomatic sections and 3% of asymptomatic sections amplifying a virus product. As discussed in that chapter, these results

indicate a strong association between KOPV and pale circle symptoms. This work also revealed another interesting factor: the virus can be detected if the symptoms are present on the midrib of the leaf. This, coupled with the fact that the movement protein appears to be impaired based on it not having the conserved emaravirus movement domain, may mean that the virus has impaired systemic movement. As discussed in Chapter Four, the mite may play an important role in not only the virus spread from tree to tree, but also local movement of the virus within a leaf. Mite transmission studies will be invaluable in determining the movement of the virus.

A major issue that has not been addressed in this research is the significance of the presence of two badnaviruses, which may be influencing the generation of symptoms in infected karaka. These two viruses were initially detected in symptomatic and asymptomatic material (not described in this research) and may be endogenous within the karaka genome, a feature that is common for badnaviruses (Bhat et al., 2016). Endogenous badnaviruses could be influencing symptom generation as the karaka tree is exposed to abiotic or other stressors (Bhat et al., 2016; Iskra-Caruana et al., 2014). This would need to be investigated further. Indeed, it may be a combination of KOPV and the badnaviruses that are causing the symptoms and both the RNA and DNA viruses may need to be present. Virus synergy has been reported in grapevine, with presence of both vitiviruses and grapevine leafroll viruses correlating with stem-pitting symptoms, and the former viruses being present in higher titres where the latter viruses are present (Rowhani et al., 2018). Blackberry yellow vein disease is another case where virus co-infection appears to be causing symptoms in *Rubus*. To date, at least ten viruses have been associated with the disease, with BLMaV, an emaravirus species candidate, as one of these (Hassan et al., 2017; Susaimuthu et al., 2008). The interaction of these viruses has been hypothesised to have led to the decline of blackberries, indicating that symptoms are the result of these interactions. Similarly, KOPV and the two badnaviruses may together be responsible for causing the pale circle symptoms present on some karaka trees. Further investigation into these badnaviruses is a high priority.

5.4 Karaka Okahu purepure virus

The novel emaravirus found in karaka has tentatively been named *Karaka Okahu purepure emaravirus* (KOPV). The name is in the Māori language to reflect the importance of the host plant to Māori. Many existing karaka groves in New Zealand have cultural significance as a historical food source and are therefore *taonga*, meaning "treasured", creating an obligation for conservation of the trees in these groves (Costall et al., 2006). Okahu in the name refers to the location of Okahu Bay in Auckland where there is an abundance of naturally sown karaka trees that exhibit KOPV symptoms. *Purepure* in Māori means spotted, which refers to the symptoms and the emaravirus suffix adheres to the current naming convention of emaraviruses (Elbeaino et al., 2018). Virus naming conventions typically use English words to describe the virus, for example *European Mountain Ash ringspot-associated emaravirus* describes the host and symptoms associated with the virus (Mielke & Muehlbach, 2007). As KOPV may be the first fully sequenced virus to be described in a native plant and due to the virus possibly being native to New Zealand, a Māori name was deemed to be most appropriate. There have been other viruses reported that have genus names using the regional languages from where the virus was found. Varsani and Krupovic (2017) describe several genera of viruses that have non-English names, such as *Gemykibivirus* genus, in which *kibi* means circular in Amharic, and the *Gemygorvirus*, in which *gor* means round in Hindi. These are genera names that are in non-English, however, and not the species name. Regardless, the *karaka Okahu purepure emaravirus* name is tentative at this stage due to, at the time of this research, discussions of the virus name were still being undertaken between key stakeholders. The name may be subject to change in the future.

5.5 Future Research

This present research has provided the foundation for future studies involving KOPV and there is much work to do to gain a better understanding of the virus. As discussed in Chapters Three and Chapter Four, the next focus of this research should be to focus on transmission of the virus. This involves studying the karaka gall mite suspected vector and

developing procedures to use the mites to transmit the virus from infected to uninfected trees. Graft transmission of the virus into uninfected karaka trees and a mechanical transmission procedure that can efficiently and reliably mechanically transmit KOPV into a model plant, such as *N. benthamiana*, should also be a high focus. With procedures that allow transmission of the virus, further research can be undertaken to learn more about the virus and its interactions between host and vector. One of these interactions, as discussed in Chapter Four, is the systemic movement of KOPV through the host plant.

Electron microscopy can also be used to detect the double-membrane bound bodies that are likely to be the KOPV virion. In previously described emaraviruses, such as AcCRaV and EMARaV, ultrathin slices have been taken from symptomatic leaf tissue and examined under a transmission electron microscope (TEM) where these DMBs were detected (Ebrahim‐Nesbat & Izadpanah, 1992; Olmedo-Velarde et al., 2019; Zheng et al., 2017). Doing this for KOPV to detect potential DMBs would give further evidence that it is an emaravirus. However, with the recent advances in next generation sequencing and bioinformatics, detecting the DMBs may not be necessary. RYRSaV has been classified as an emaravirus species and it appears to not have had any electron microscopy work done to support this (Di Bello et al., 2016). Therefore, while it would add evidence, TEM analysis of KOPV is not required for it to be considered an emaravirus.

This present research has raised questions about KOPV, such as where it came from and is it endemic to New Zealand. The phylogenetic analysis described in Chapter Two indicated that KOPV shares a common ancestor with the rest of the emaraviruses but appears more genetically distinct. A molecular clock study could be done to determine the evolutionary age of the virus, which has been done before for many plant RNA viruses (Guterres et al., 2017; Holmes, 2003). This, coupled with work by Atherton et al. (2015) on determining the evolutionary history of the karaka tree, could give great insight into the origins of the virus. If KOPV is endemic to New Zealand it likely has co-evolved with karaka. If this is the case, then herbarium samples and photos need to be searched for signs of the symptoms. This has been done to a limited degree in preparation for this research (data not shown) by searching the Auckland War Memorial Museum's photo collections of Karaka. Symptoms could not be

found, although that may be due to the age of the samples but also due to collectors selecting "unblemished" samples to keep. Furthermore, there appears to be no mention of the symptoms in any literature. Research is limited around Karaka; therefore, it may be prudent to ask the *tangata whenua* (the local Māori people) around areas where karaka has been cultivated if these symptoms are known to them. If KOPV is endemic to New Zealand it will be the first fully sequenced plant virus native to the country and therefore important to study to further understand the virus and its interactions in this country.

Appendices

Appendix A. Completed nucleotide genome (Complementary DNA sequences)

>Karaka_emaravirus_RNA1_Final
AGTAGTGTGTTCCTTGATATTTACAAACAATAGCTCAGCAGACAAACTTTAAAGAAAACAAGTATGAATTCAGAGAAGCT
TAATGTTTCTGCTAAGAGCAAGATTGGAGAAATAGTTGATAATTATGTCAGGAATAAAGAAATCTATAAGGATGATGAGG
TCATCAAGGCATTCTCAATGATAATAGGAGCTCATAGCTCTTACCGAGAACCAAAGTCAATTGGAAAAGTGCCAGAGAAG
ATCTACAATGTCATCAAGGATGCTTTAGATGAGGAGAATGAGACATACATAAACTTGGTTCAGAATCTCATACATGGTAC
TCCAAATGAGAATAGAATTGATATTGCTGTGACTGTTGTTAAGATAACAGAGATGAGGAGGCATGATTACTTGATCAAGT
ATCTATCTGAATCATTGGAGTCTGCTGGCATAAGGGTACTTGATACTGAGTTCAATATGAGTGATCACTACAACATCCAG
AATAAATTGACACCTGACCTGTTGGTATCCATGAATGATGTTGAATATGTCATTGAAGTTAAAATGAGAGATCCTAGCCA
TTTGGATATGGACTTCTTCTATAAGAAATATAAGATTGAAGACTTGAATTTGGTATTAGTTGTCAACTTTTCTAGGGGTT
TATTGTACTTCCATGGAGATAACAAAGATTATTTCCAAGACAGAATCAGATTAACTCAAGAGCAATCAGAACATCTAGAG
TCAATGATTAGAGAGACAATAAATATAAGGAATAACTATGCCAAATATGAGCTTTTTAATCAGTTGATGAATAATGAACT
CTTTGAAGTTGAATCAGATATGATAAGTGGTTTTCGAGATGCATGCCAGGATCTAGAACCATTCAAGGAAATAAAATCTT
TGTATGGTGATTATTATGCTGATATAGAGGAAGCAATTAGCAATTTTAACTTGACAAACCAATCTTGTTTAGATATTGTT
GAGGGTGCTGAAGATAAGCTTCAACAATATATAATCACCAATTTGTTGGATGATATAAAATCTGAGTATATATCCTTGAA
GGAGAATAATGATTATAGTCATTTTCCGGATATCAATATAAACATTAGAGATTTGATAGATAATGAAAATATCATTAAGT
ATGATATAACATCAAAATACTCTCCATCTTTATATGTTCCATTTGTGCATGTCAAACAAATAAACACTAGTAGAATAACA
CTCTACAATCAGTTATTTGAAAATTTAACTATCAAAAACACAGATGATTACTCATTATCAGCTTTTAATATGATCAAAAA
TATATTGACCAATGATAAGCTGGGAGTTCTACTTGGACATGAGAGAGTGGCTGATGAAAGCATTGAGATATCCAACTCTG
CAACTCACACTGTGCCCAATCCATTTGTTAAAAGTAAATTATTAACCAAGGAGTCTATGCAATCTTATGATCTCAATATC
TATGTTAATAACACATTTTCCTACACTGATAAGAATATAGCTGATAGTAAATCTAAGATATGTGGTTTCAATAGAAAAAA
ATACAATCTTGGTACAGATTTCAAGAAAAGCTCATTATTACTGTCAGAGAAGGTTGATGATATAAAATTTGTTGAGAATT
ATCTTAAAGGGTTTTCAGCATATTATACAAGTGAATCTATATTGGAAGATATGAGAACTTATCATTCATCATTTAATGAT
TTCACAGATCTTGATGTGGACCTGATAGATGAAGTAGCCTACTATTCATATTACACTTCAAATCTGTATAGTTCAATGAT
AGCTGTCTCCCAAATAAATAATAAGAAATTCAGATTGATACAATGTTATGATCCAAATACAATATGTATAATGTTGCCCA
ATGGTGACATAAAAGATAACAAACCAATAAGATTCTTCACCATTGTTATCACTGAAGATGAAGATGATATAATATTGAAT
AAAATAATGGGGAATTACCATTCGCATGAGGTTTCTGGCAAGAAGTTTATAATACTGAGTAAAGTTCTAACCCTTGATAT
AAACAGGATGAAATTACTGGCTAACTCATTTGTTAAATATACAATATTAACATCATATTATAAGACATTCATAGGTTATG
ATACCAAGAGGATATTGACTAATTTCAATAGCTTGTTGTACACCAACATGGTCACACTTCAAACACTAGATGTAACCGAA
AATTTCAAAAACATAATAATGGTGTGCTATTCTGATTATGGGAACCCAATAAAATTGATAAAAGATAAATTCTTGCCAAG
ACTGAAGAATTACACATGTATGCAGCTTTGTGACAAGATGATCAATGCTATAGGATTATCACACAGTCAGAGAAATATGA
TAATAAAAAATCAAAATAAGGCTGAGATCAGCACAGATGGCAGAGAGATACTTCACACTGGCTTTAGAGGTGTGGATGAT
ATGAAGTTACCTATATCAGGATTGTTGACAAATAACCCCAGGGAGATATTGCAAGAAGCCTATATAGTATTTTATCTTGG
AAATAAGACTCTGCATGGTTCACCACAAGAATTAATATCACTATACCATGTCCCATACTTATTTGAGACTGAGTATAGTA
ATTACTTGAGTGAGAATAAGACAGTCATACAGGAAATCACTGGTGATTACAAGTATGGTTTTTCATATGAAAGCATGAAA
TTGACAACAATATGTGCATATTCAAATCAGATGTCTGATAAGAAAAAAATAAGGGCTATGATAGCAAAAGATCTTGATTT
GGACTCACCACTCTTATCAAAACCTCAGTTTTCATCAACTAAGAGCATGGTTACAAATGACATGAAGATTGATGCACCAG
TACTTAATGAGCATAGCACATTTATTGATTTTAAGAATTGGCTTCAGCACACCAAATTTGATGATGTAGACAAGTTTGTA
TCATCATCCAATGTAATGATTAGGGTACTGAATTCAAAAAGAGGTGTTTCTAATGAAATAGCACAAGAAATATTTGGTAT
GGCCAGTGAGATTGATAGATCACCATTCCCATTTTTAAAGAAAGAGATATTACCTAATGGTAAATATTACATAACTGCAA
AACCAACAAATTACTGTAGGTATTCATTGACTGACACTAAAGACTCAGAAAATAACAAAGTGTTTGATGAGATGCTTCAA
TTGACAAAGGATGATTCATCATTGTCAGTTAGGAATGCACTAGTTTCACTCCTAGACAATGATAATGAGAACAATTTTAG
AGTATTCAATAAAGATCAGAGAACATATAATGATAGGGAGATATATACTGGAAATAAGGAAGCTAGACTCTGCCTCTATG
CTATAGAAAAGTTGTTTCTAAGTCACAACAAGTTGTTACCAGAAGAAGCTATCACAATATCTGGGGACATGAAACATAAG
AGAATGTATGACCAGAGATGCTCAATAATGAAAGAAGGCAGATACCAATCAGCATATGGTAGAACTTTGATGTCTGTATC
TTCTGATGCTAGTAAATGGTCTGCAAGGGATTCTTATATTAAGTTCCTAATATGCATTGCTTATAATCCATATTTATCAC
CAGAAGAAAAATGGTTTTATATTTATTGTTTCAGCAGGTATTATTACAAAAATATAGTATTGACTGAGAAAGTATTACAT
AGTATTTATGAGTTAAGTAATAAGGACAGAAATGGTAGATTCGAAGAGATGACTGACAACATGGAGAAGAATTATTTTAG
AGTCAGGTCAAATTGGCTTCAAGGCAATTTAAATAGATTATCATCGTTTGTACATTATTGCTCATCATTCATGGTAAAAA
AAACTTTGGATGTCCTTAATTCAAAATATGGTGATCTAAACCTAATGAGATTCATGGTTCATTCTGATGATAGTACATAT
GATTTTTCTATGTCAATAAAAAAAGAGAACAAGAGAGATCTATCAATGAGAACTGATTTTGGTAGATACATAATATCTAT

AATAAAAGTTTGTGAAAGGAGGCATAGCATAATTTTGAATGAAAAAAAGACATATATCTCAACATTTTACAAAGAATTTT
TGTCAACACTCATAATAGGAAATGTGCTATCTTACTTCTATTTAGCTGATTTGATACCTATAGCATCAGATCTATCCTAC
AACAGCCCTTTGGAGGATACAGCATCATTAAGCAGCTTTATACAGAATGCATATGCACACTCAGCACCATATGGTATAAT
CAAAACTGCAATAAATCTTATAAACTACACCTCATTAAAAACATACAATTTAAATAACACATCAAATAAATCACCCATAA
ATCATTTTTTCACAAAATCCATATACAAAGAAAATTATTCCATGGTTCCTATATCAATATTACCAATATACAGGTTTGAT
GTGAAGCTTGCTGGGATTATACCATATCACTGTGCTGATGCATACTATATACTAAAGAATATATTAGAAATTGTTGGTGT
TGATGATGGAAAAACAATAGAAGATCAAATCACAAAAGAAACCTTGACCAATTATATCTCTAAGGCTGATAAAACAGTCC
TAAATTACATAAAGGCATGCATGTTTTCATATGACTCAACAATGTTCAAAATAGATGATGAAGACCCTTACAACACAAAA
GATCTAGCAATAAATACAATTGTCAGTCTGAGTAAGCCAAGGTCTGGAAAGAGAAAAGATCCCAAGTATAAATCCTATCA
GGAATTTAAGCATAGGAGAACAGAAATTTTGAAGAAATCAACAGCTCATCCAGAGTGGTTATTAACTCGACCTAGAGATA
TGGATGATTCTAGAGATAGTATCTTAGCTAATTATTTAAAACCCACATTTATAGAGTCTTTATCCTTTTCAACACCAGCA
ATGGAATATGGTTCTAGAGTGATCCATTCAAACAGGGAGATGTATAGATTAAATATAAAAGGTTACACAAGTGAGAAGTT
ATATAATATAAAGGATATATACAATGAGATTGTTAATCTATCACACAACATTGACTTAGACCCAGTCAAATTATACTCAT
ACCTATTTACATTTTTATTCTCTGACAAGGAGATATCGTATGCCATACAAACAATAATGACTAAGACACTAAAATCAACA
CAACAGAAAGTACATATGAGTCTTAAAATAACTACACCAGCATCCATATTTAGAAGAGATAAGGGAGCATTGTCAGTTAC
AAGTATACTAAATGAGATTTTCACATATACAGGTCATCTGAGCTCTGTTGATTCAAAATATAGAAGCATAATTGATTACT
CAAAAAACCTTTTGACTAGTTTCAACCTTGATGGTGTTAAATTATATAGAAGCTATCTTGATATAGATGATAATTACTTA
GAATATGCATCATTTATGAAGACAAAGACAGATTTTGATGATGTTATTTCCATACTAGATATTACCGATGATGAAGAAGT
GAGGTTGGCTGTTCACGAGATAAGAGTCATATACAAGTCTATACTTATTAGATATTTCTCTGATATGATATTGGCTGATA
CTGATAAGGACTTTAAATTTGATAATTACCCAACACCACAGTCCATACTATCAACCATAAGCAAATACTCAACTAAAGAT
GTCACCACATCAAAAGTTTATATGGCAGAAAAGAGATCGAGTAGGAAAGACGATTATTGGTTATCCAGATTAGGTTATTA
TCAAGATGAGAACATGTATTATAAATATATTCTCAGCAATAAGTATAATCTTGATCATTCAGACCAGATGATAATAAGGG
AGTCTAATTATGAGGATATAAATACATTTACTTCTTTGATGAATTTCATAATAAAAGGTGAAATGATGGATAATTATGAC
AAACTTAAGTACAATGGTATGTCATATGATGACTTTTTAAGGAAATTGCATAATAGTGGAAGTTTGAACCACAAAATACT
TCTAAATAAACTTGGTTATATGGATCAAAGGTCACTGTCAAGGTATATAAATATGGATCAAAGGATAAAAAATTACTGGT
TGATATCTCATGGACAATCAAGAGAAGATCCAAACTTATCACATGTTGTCTATATGTTCAGAACTACATTTTTATCAGTT
CAAACTGCCACTGAGAGAGGCAAAATACTGTTCAAGATAGAATTATTCCACCAAGAAGCTTATCCAGATGATGATGCTGT
TTTCAGATTGATAACAAAGTTGAGTAAAGACAATAGATCAACAATAATGAACAATAGTGTATATTCCATAAAACAAGGTT
TAAACATATATATGGATTCTCAGAAGAATTTGACCACCAGAATCAAAGATACATTTGGTAGGATGAACAGCTCAACATTG
ACTTATGTTCCTATTAGGTACAATGATATAAGGTTAACTGATATAAAACTGACAGACGAGAACAGACACATCTATCACAT
TGATATATATAGCCCTCTGGGAAATAGATATTTAGGTTTTGAATTCAAAGACAACTTTAGTTACAACAGTGAATCACTAC
TACATATATGTCTGGAGAATCTTGAAAAGCACCAAATTTTATCAAACATTATAATTGATAATAGATTGTATAATGAAGAT
AATATGTTTAGCATAATAAAGAGATTGACTAAAGACAATTTAATGGCATTTGTTTTAAGGATAAAAAATGTTGGTAGGCC
AAAGATAAGTAACAAACTAGTTAAAAATATAAAGACTGCTAAAAATAGAATGTACATGAAGTCTTTTCTAGTTAAGTATG
AACCAGAAATATTGGAAACACTCTATGAGAGATACTCAAGATATGGTTCCATGTTTTACAATTTTTATATGATGCTATTG
GAGGATTCTGATCAACTGGATATTGTTGATCATAATGTTATAGATATCATGAATTCAGTTCAGTTAGATAGAAGCTCATC
TTCAAAGATATCAATGGGTAATACAAAACCATTTGATTCTGTATTCTATAGTTATTTAGTAACATCTATAAATCCAGATG
ACATCATCTCTATAATAGAATCAATACTGATGTACATAATAAAATCTGCAGTCTTTGGGACAGATTTCTATATAGACCTA
GTTGATAATTTTCCTGATTCTGATTCTGATAATGATGATGATGATGACTTCAGTTATGGTGATTAATTCATAGTATTCAG
CTATAAAACTACAGTTAGCGTATCTTAGATCTATATTCACTTGGTTTATTATTTATTTTCAGATTTTCTTCATGTTGTAA
AATATCAAGGAGAACACTACT

>Karaka_emaravirus_RNA2_Final
AGTAGTGAGCTCCCTAACTTTACAACAACAAAGACAAAGCATTCCAATTTCTAAAAATATGGTTAATAAGATTTACCTAG
CTTGCGCAATCATTGCCATTGTAGTCACAACATCAAAGTATAGTACAACCTCAGGTCCATGCAGTTGTAACATCAAGTTG
AATCATATACACAATGATTATGTTCTTGTGAAACCTAAAAATATGTGTCCAATTGTACTTCCAAACAATCATATAGAGAC
AGGGAATTACATAACTTCATCTAACAGCATAATGGGTGTTTTATGCAATGATGATTTACTTACATTTGAGAAACCAAAGA
TTATATCCAAATGCAAATACTTTGATGATCATACCTGTTCACTGTATAACTCATTCAAAGTATTCTTCAACCATGTTGTG
TCCTACATAATAGTATATCTTGTCCATGAGCATATAGTCACAGCTCTAATCCTGATGTATGCTATATTTATAACAATAAC
CAAGAATAAACTAAGAGTTTGTCATATTTGTAATCTACCATATATTCTGTATCACAACTGTAATACAGCATATAAGCTGA
ATGGTAATAAATTCTTCATGTTAGCTGCCTTGATATTGCATATATTTTCTTACAAAGTCAGTTCAGTTCAGATGCACAGT
GTATTGAAAACTGGTGATGGTTCATTGGTAACATTTGATGATCATGTTGGTTTAATACAACAATATGATATGGATGGTAG
CAAGCTCCAGTTTCAGCTGATATCAACAACTCTAGTTTATGAGAGTAGCCATTTGCATGATTTATATAAGATTGTCGATG
AACCATTTATTGAGAAACAATGTTCAGTTTGTGAGCCAAAAATTGAGGATTGTGTGAAATCTCTTGGTGGTAATGCTGAT
TATGCTTATTCAAAGTTAGTTGATAAGTGGTCATGTGCTCTTGTTCAACTGAAAGTTTGCTTTATATGCTCAAATAATGT
TGTGAAGATAGGATCAGTATTCCAGTTAACTAAACCAAAGATAGTTGTAGATTATCTTGAAACTATTGATGGAATAAAGA
CTAATACTTTAAATCTTGATAAGAGAAACTCACAATATTTTGAAACCATGGTTGTGTTAGTCACTGAAAAAGATAGCTAT
GAGGGGAACATTTGCATGAAACCAGATTTATCTTGCTATGGTAAACATATTAAGAGATCTAGTGATACAGATGTTGTTAT
ATATAAAAAAGTAGAAATTATTGATATGCCAGGACAATCATATACATTGAAGACTTGTGGTGTTTTAGAATCAGATCCAA
AGAATTATCTCAGAATTGTTGATGGATTTTTATCTGAAACTAAGTTTGTTAAGAACATAAATTTAGGATCAATAACATTT
AATATAAATAAGAATCTGATACCAGACAAAGATATTTGCAACTCAAAAGCAAATCTAATTTTTAGAGTTGATGGTTGCTA
TAGTTGTCATCAAGGTTATAAATTAATTATCAGTCATGATTGTATGGATTGCTGCAAGAATATAGTCAAGATAAATGGTG
TTGATGAGATGATCACAACTTACAGTAAAGCTGAGGTGACATTGAGAAGATATTCTGATAACAAGCAAATCACAGTTGAC
TCAGAGGTATTCCAGTTATCTGAGCCTGAGGATTATCATGAATCAGTTATAATTGGAACTCATTATCATTATCCAATGAA
ATCACTGATACCACAAGATATAGAATTCTCATTTCCATTTATTAGGAGAATAATAACAATTGTGGTATTATCTTTTCTAA

CAATAATTGTGATATATTCAATTTTGAATAGTTGTAAACTACTACAATCATACACTCCAAATGAAGTATTATTGAGAAAT
AAGAGCAAGAATCTGTAGACATGCATTGTAGCTGATCTTCACCATATTTCTATTAGATGTCATGTTTTGATTTGTGTTGT
TAAAGGTTAGGGAGATCACTACT

>Karaka_emaravirus_RNA3_Final
AGTAGTGTGTTCCTTGAAATTTACATAACATAACCAAAACTTTGAACTTCTTACTAATTCAGTTTAAATCTTCCAAACTT
GCGTCAAACATGTCTAGCACAAAATTCACTATTAAGTTCACTGGTGGAGCATCTAAGGATGTAGATATCAAAGAGATCAA
AGATATCTCCTCAATTAGATTCAAGCCAAATGAGGCAGCAAATGATAGTTATGACATTCTAAATTTCTGCAAGAGCTTCT
ATAATTTTGAGGACTGTATTTCGGTTTGCACATCTAGCCTTGAAATAAAGAAAGAAATCAAGGACAAGCAAATTGTCAAG
AGACTAATAAAAACAGCAAGTGGAGGAACTCACCAGTATGTATTAGTTTACAAACTGGGTTCCTTTACTGGTGACAACAA
GTTTGTTATGTCCTCAAATAGGTTCATTGCTATTTGTTCTGCTACAATCACAATTGGCCTATTTGGGACAATTTTTGATT
GGGAATCAAGAAGCAATAAACCAGTGGGTGAGACTACAAAGAAAGCTTATAATCAAGTCACAACTGGTGATGCAATAAAT
AGGTTAGCAGAGAAGATTGGAATATCAAAACAGAGCAGACTTTATTGGTTGTATTTGCCTGGTGTTGAGAATTGTTTTGA
AATCTTTCCTGAGGAAGTTACAACTATATGCTGTTGGAGGGTTCTTCACAAAGAAAAATCTGGCTTATCAGGTGATGCTG
TTAGTTCTATCAAGGCATTATCTAACAAATTTGTTAAGAGAGGTATAAGCATAGGTGATCTAAATCAGGGAAAAATAATG
GAATACTACAAAGAGTTAGAAAGGTGTGTCCATACTAGTGCTTCTCATCAACATGCTGTTGACTTCATGCTCTCAGTCAC
CAGTATGTTTAATAAGCATGCAGAAGATTACATAGATGCTAAAGATTTGAAGAGTACTAAGGCATCATCTTCTCGTATGC
CTTAAGAGCTGTATTTTAATATAATTATATATAGTCAAGTATTATTATCTATCATTACATATAATCATATTTAATCCAGT
ATAATCATATTGAGCTTAGCATAATTATTAATTATTATATATAATTCAGTATAATTAAATTGCATCCTGGTTAATTATGC
ATAATCTAAATCAATCATTGGTTATTTAAATCCATTATGCATAATTCTATATAATCACATATTGAAAAATGCAATTGTTA
TGAATTCAGTATCATTACCATATATTATTAGGAATCTAATTTCCCATAATTTCATTATAGTTAATTATGGTTGATTAGTA
TTGGTATTATCAATAAAATATAATTATGAAAAAACAAAAAAATCAAAAATTGAAAAAAATCCAAAAAAAATCTAAAAAACA
TATAAAAAAACTAAAAAAATATATGGCATCAATTCACGAGTTTGGGCAAATCTTTCACTGTATTAATATTAGATTCCATG
TTTTGTTTATGTTGTTAAAATTTCAAGGAGAACACTACT

>Karaka_emaravirus_RNA4_Final
AGTAGTGTGTTCCCAAAAATTTACAATACAAAACCAAAATCTGCAAACTTCTCTAAAAGTTGGTTATTTCTCTAAGTATC
TTGCACATTTCGATTAGTCACCATGTTTCCCATCATGATTCTGTTAGGTCTGGTTGCTACAACATCAGCTAGTTCCACTC
ATGATGTTTCAACCTGGGACAGTGAAGAGGCAATCTCAGATGCACTTGAGACTGTCACTGTTGATGATATGTCAATGACT
AAGCTCAAGAAGGATTTTGAGATAACCAACAAGAGTCAAAGCTTGAAAGTTGATTTTTATAATCTTTTCACAAGAACTGT
TAAGACAATCACTCAGCCAGGAAAATTGGTAAGAATTACTGCCATGGTTTTTGTCTGGATTCCACAAGCGAACAATCTGC
TTGGGAAAACTACATTATCGCTCACTGATGGAAGGAAAGCTAAAAGTGATGGCAATAGAATTGAATCACAAGTAACTTTC
AGACCTGACAAGCCTGTCATTGGAGTTTTCTATCATAATTATGCTATGACCATTGCTGATTTGAAGTATATGGATCTAGA
ATTCATCATGCATGGCATAAACATGGATAAGGGAAGTATGGGTAGGCTTTCATTTGGTTACAAAACTATAATGACTGGCC
CTACATACTACATGAAGAAAGATCCAGAAGTATTTTATCTTCCAATTGAGCAACTTCCTGAGCTTTCTGTTGAAACTCCA
ACAGACATTTATGTTTCATTTATTGAAAAGATGAAGAAGAAAAAGAGTCAAGAAGTTGAGTACTTTTCTAATCTACAGAA
GTTTATTGATCTGACAAAGAGACCTATAAATGCTGCAACAGCTAATCAACAGGCTCTCATCGATGATTTAAAGAGAGATC
TTCATGTATTGAATAAATACCAAGATGATGCTCAGAAATTCAAGCTCCAGAAAGAAAAGGTTTCTGGATTAGAGAGTGCT
ATTGCTGACACAGAGAAAAGAATCCATAACTTGATGTCAGCTCAAGAACAAATAGAGGAGCCAATGAATGTTATCAATGT
TGACAGAGAAATTTAAGATAATCTCTGCTTAATAATTGGAAATGAAATGAATGAACAAAATAACAAAAAAATAAAAAAAA
TTAAAAAATCAACAAAAAATATAATGATAAATCCAAAAAAAATGAAAAAATCAAAAAATAAAAAAACATCAAAAAACTAAG
AAAAATCCAAAAAAAATCATATATGAGAAAAAAACAAAAATGAAATTTGCTTTGTTTTTATCTCTCTTCCCTATAGTCTC
TTCTTGATTTATTATAATGAAGAAGAGACAATAGAGTAGAGAGATAAAAACAAAGTAAATCAAATTATTGTAGTTGCATG
ATAATTTGATTCAATAACCAGTTCAGAAATCCTTGAACTGTGTAAATGGGTAAATTTCGACAGACAACTAATATGATGTA
AGCTAGGTGAGATCAAATACCATTCTGAGTTGAGTTAAAGTTGTTTATGTTTTGTTAAATTTTTGGGAGATCACTACT

>Karaka_emaravirus_RNA5_Final
AGTAGTGTGTTCCCAAAAACTTAACAAAAACTAACAAAATCTGGAAAGCTTAAATAACATTTGGTAAAATTATCTCAATC
AGCAGATTGCCTTATTGCTTATCACCTATTAATATATTTAGTGGTGAACCAAACAATTCAATCATGAATAGCATGATTTC
AAGAATTGAGGTAGATTGTGACAAAGGCAAGGTGAGATGTAGTGACTTATATGGTAATCACTTTAGCATCGAGATCAGAG
GTGAGAACTTGAAGAGCTTGGAAGATCTTTATGTTAAGCTGATTTCTGGTATCAAGAGCTATGATACTAATGAAAGAATC
AGGATTGATGTCATAAAGAAGTCAAAGCTGTATAACTTCTTTGCTACATGGTCAGAACCAAATGCTAGTTCTGTGGCATT
CTTGGATTCTATGCTTAATAATAAATATAAGCATATAACACTGAATTATTCTTTCCAGGAATCATTTGTTTGCAAACTTA
CTGCAATTTATCTGATTCCATTGATCATAATCAAGACAATTTTCTGTGACATTGATTTGCCATCCTCATATCACAATCTT
GATAACTATGTGTCTGTTTTCAAGAAAGTTCCAGAAGCAGAAGTTGCTCAAGTGCCATTGTCACCAGCATCACAATGGGG
TATGGTCAGGACATTTACAAATAGTATAAAAAGTCTTGGCAAAACATCAAGCAACAAATCTAAAGATAGGTCAACTACGA
AGACTATTGATGATAGTAAGTCTATATCAAGCTCAATCAATATTCCATTTGTTGGTCCTGTTTCTGGCTCTTATAGTTCA
GGACACAAAGAGATTACAGAGACAAAGAATATTGAGCTTGATGTTGATACTAAGAGTCTTGAGGTTTATCCTGATATTTA
TCCAAAAATCAATTAAATATATGTACTACTATTTCTATTCATTATTTCTATTTTATTACATTATAACTTATAAATTATAA
ATATCCAAATGTAAATTGGTTTTATATACAATCTAGATCTTAAAAAAATCTAGATAAAATTAATCACGATATTGAAACAT
CATCTGGTAATTTGCTCTCAATCAGAATTATTAATCCTCAAATTTTATTTGTCCATCATCATTAGCTATTATAATGAATC
CTGAATTATCATAGTTATTGGGAAATCATATTCAATCATCTACAATTATTATTATTCATTGTATTTTATTATGCTTGGTT
GATTAATAATCACTGTTATGTCAATTGATTATTTTCATTTATTATTATTCATTATTTTCTATCACAAATCATTCTGATTA
ATTATCTAATTGTTTCCATAATTGTTGGTTGACCTCAGTCATTCTCCACTATTAGATTCTATTATGATTCATCCTCATTA
ATTTGGATTTGATTGCCATAATCACTATTCATTATCATTTGCTATACTTAGCTATGCTTTATAGTATTTGGAAATTGTTG

TCAATTAGATGATCTGCTGCACATTAGTAATTAAATGTCTGTGAATTATGGACATTAGTTGCTACTGATTAAATACCACA
TTATAAGTTGAGTTTCAGATGGTTTTGTTTTGTTAAAGTTTTGGGAGATCACTACT

Appendix B. Completed amino acid sequences for RNA 1-5

>Karaka_emaravirus_P1_RdRp
MNSEKLNVSAKSKIGEIVDNYVRNKEIYKDDEVIKAFSMIIGAHSSYREPKSIGKVPEKIYNVIKDALDEENETYINLVQ
NLIHGTPNENRIDIAVTVVKITEMRRHDYLIKYLSESLESAGIRVLDTEFNMSDHYNIQNKLTPDLLVSMNDVEYVIEVK
MRDPSHLDMDFFYKKYKIEDLNLVLVVNFSRGLLYFHGDNKDYFQDRIRLTQEQSEHLESMIRETININRNNYAKYELFNQ
LMNNELFEVESDMISGFRDACQDLEPFKEIKSLYGDYYADIEEAISNFNLTNQSCLDIVEGAEDKLQQYIITNLLDDIKS
EYISLKENNDYSHFPDININIRDLIDNENIIKYDITSKYSPSLYVPFVHVKQINTSRITLYNQLFENLTIKNTDDYSLSA
FNMIKNILTNDKLGVLLGHERVADESIEISNSATHTVPNPFVKSKLLTKESMQSYDLNIYVNNTFSYTDKNIADSKSKIC
GFNRKKYNLGTDFKKSSLLLSEKVDDIKFVENYLKGFSAYYTSESILEDMRTYHSSFNDFTDLDVDLIDEVAYYSYYTSN
LYSSMIAVSQINNKKFRLIQCYDPNTICIMLPNGDIKDNKPIRFFTIVITEDEDDIILNKIMGNYHSHEVSGKKFIILSK
VLTLDINRMKLLANSFVKYTILTSYYKTFIGYDTKRILTNFNSLLYTNMVTLQTLDVTENFKNIIMVCYSDYGNPIKLIK
DKFLPRLKNYTCMQLCDKMINAIGLSHSQRNMIIKNQNKAEISTDGREILHTGFRGVDDMKLPISGLLTNNPREILQEAY
IVFYLGNKTLHGSPQELISLYHVPYLFETEYSNYLSENKTVIQEITGDYKYGFSYESMKLTTICAYSNQMSDKKKIRAMI
AKDLDLDSPLLSKPQFSSTKSMVTNDMKIDAPVLNEHSTFIDFKNWLQHTKFDDVDKFVSSSNVMIRVLNSKRGVSNEIA
QEIFGMASEIDRSPFPFLKKEILPNGKYYITAKPTNYCRYSLTDTKDSENNKVFDEMLQLTKDDSSLSVRNALVSLLDND
NENNFRVFNKDQRTYNDREIYTGNKEARLCLYAIEKLFLSHNKLLPEEAITISGDMKHKRMYDQRCSIMKEGRYQSAYGR
TLMSVSSDASKWSARDSYIKFLICIAYNPYLSPEEKWFYIYCFSRYYYKNIVLTEKVLHSIYELSNKDRNGRFEEMTDNM
EKNYFRVRSNWLQGNLNRLSSFVHYCSSFMVKKTLDVLNSKYGDLNLMRFMVHSDDSTYDFSMSIKKENKRDLSMRTDFG
RYIISIIKVCERRHSIILNEKKTYISTFYKEFLSTLIIGNVLSYFYLADLIPIASDLSYNSPLEDTASLSSFIQNAYAHS
APYGIIKTAINLINYTSLKTYNLNNTSNKSPINHFFTKSIYKENYSMVPISILPIYRFDVKLAGIIPYHCADAYYILKNI
LEIVGVDDGKTIEDQITKETLTNYISKADKTVLNYIKACMFSYDSTMFKIDDEDPYNTKDLAINTIVSLSKPRSGKRKDP
KYKSYQEFKHRRTEILKKSTAHPEWLLTRPRDMDDSRDSILANYLKPTFIESLSFSTPAMEYGSRVIHSNREMYRLNIKG
YTSEKLYNIKDIYNEIVNLSHNIDLDPVKLYSYLFTFLFSDKEISYAIQTIMTKTLKSTQQKVHMSLKITTPASIFRRDK
GALSVTSILNEIFTYTGHLSSVDSKYRSIIDYSKNLLTSFNLDGVKLYRSYLDIDDNYLEYASFMKTKTDFDDVISILDI
TDDEEVRLAVHEIRVIYKSILIRYFSDMILADTDKDFKFDNYPTPQSILSTISKYSTKDVTTSKVYMAEKRSSRKDDYWL
SRLGYYQDENMYYKYILSNKYNLDHSDQMIIRESNYEDINTFTSLMNFIIKGEMMDNYDKLKYNGMSYDDFLRKLHNSGS
LNHKILLNKLGYMDQRSLSRYINMDQRIKNYWLISHGQSREDPNLSHVVYMFRTTFLSVQTATERGKILFKIELFHQEAY
PDDDAVFRLITKLSKDNRSTIMNNSVYSIKQGLNIYMDSQKNLTTRIKDTFGRMNSSTLTYVPIRYNDIRLTDIKLTDEN
RHIYHIDIYSPLGNRYLGFEFKDNFSYNSESLLHICLENLEKHQILSNIIIDNRLYNEDNMFSIIKRLTKDNLMAFVLRI
KNVGRPKISNKLVKNIKTAKNRMYMKSFLVKYEPEILETLYERYSRYGSMFYNFYMMLLEDSDQLDIVDHNVIDIMNSVQ
LDRSSSSKISMGNTKPFDSVFYSYLVTSINPDDIISIIESILMYIIKSAVFGTDFYIDLVDNFPDSDSDNDDDDDFSYGD

>Karaka_emaravirus_P2_Glycoprotein
MVNKIYLACAIIAIVVTTSKYSTTSGPCSCNIKLNHIHNDYVLVKPKNMCPIVLPNNHIETGNYITSSNSIMGVLCNDDL
LTFEKPKIISKCKYFDDHTCSLYNSFKVFFNHVVSYIIVYLVHEHIVTALILMYAIFITITKNKLRVCHICNLPYILYHN
CNTAYKLNGNKFFMLAALILHIFSYKVSSVQMHSVLKTGDGSLVTFDDHVGLIQQYDMDGSKLQFQLISTTLVYESSHLH
DLYKIVDEPFIEKQCSVCEPKIEDCVKSLGGNADYAYSKLVDKWSCALVQLKVCFICSNNVVKIGSVFQLTKPKIVVDYL
ETIDGIKTNTLNLDKRNSQYFETMVVLVTEKDSYEGNICMKPDLSCYGKHIKRSSDTDVVIYKKVEIIDMPGQSYTLKTC
GVLESDPKNYLRIVDGFLSETKFVKNINLGSITFNINKNLIPDKDICNSKANLIFRVDGCYSCHQGYKLIISHDCMDCCK
NIVKINGVDEMITTYSKAEVTLRRYSDNKQITVDSEVFQLSEPEDYHESVIIGTHYHYPMKSLIPQDIEFSFPFIRRIIT
IVVLSFLTIIVIYSILNSCKLLQSYTPNEVLLRNKSKNL

>Karaka_emaravirus_P3_Nucleocapsid
MNFLLIQFKSSKLASNMSSTKFTIKFTGGASKDVDIKEIKDISSIRFKPNEAANDSYDILNFCKSFYNFEDCISVCTSSL
EIKKEIKDKQIVKRLIKTASGGTHQYVLVYKLGSFTGDNKFVMSSNRFIAICSATITIGLFGTIFDWESRSNKPVGETTK
KAYNQVTTGDAINRLAEKIGISKQSRLYWLYLPGVENCFEIFPEEVTTICCWRVLHKEKSGLSGDAVSSIKALSNKFVKR
GISIGDLNQGKIMEYYKELERCVHTSASHQHAVDFMLSVTSMFNKHAEDYIDAKDLKSTKASSSRMP

>Karaka_emaravirus_P4_Movement
MFPIMILLGLVATTSASSTHDVSTWDSEEAISDALETVTVDDMSMTKLKKDFEITNKSQSLKVDFYNLFTRTVKTITQPG
KLVRITAMVFVWIPQANNLLGKTTLSLTDGRKAKSDGNRIESQVTFRPDKPVIGVFYHNYAMTIADLKYMDLEFIMHGIN
MDKGSMGRLSFGYKTIMTGPTYYMKKDPEVFYLPIEQLPELSVETPTDIYVSFIEKMKKKKSQEVEYFSNLQKFIDLTKR
PINAATANQQALIDDLKRDLHVLNKYQDDAQKFKLQKEKVSGLESAIADTEKRIHNLMSAQEQIEEPMNVINVDREI

>Karaka_emaravirus_P5_Unknown_Protein
MPYCLSPINIFSGEPNNSIMNSMISRIEVDCDKGKVRCSDLYGNHFSIEIRGENLKSLEDLYVKLISGIKSYDTNERIRI
DVIKKSKLYNFFATWSEPNASSVAFLDSMLNNKYKHITLNYSFQESFVCKLTAIYLIPLIIIKTIFCDIDLPSSYHNLDN
YVSVFKKVPEAEVAQVPLSPASQWGMVRTFTNSIKSLGKTSSNKSKDRSTTKTIDDSKSISSSINIPFVGPVSGSYSSGH
KEITETKNIELDVDTKSLEVYPDIYPKIN

Appendix C. Emaravirus interspecies amino acid pairwise identity comparison (P1 to P4)

	RNA 1 (P1) RdRp													
	KOPV	AcCRaV	BLMaV	EMARaV	FMV	HPWMoV	JYMaV	Palo ver.	PiVB	PPSMV1	PPSMV2	RLBV	RRV	RYRSaV
AcCRaV	30.3													
BLMaV	30.1	46.4												
EMARaV	31.4	54.3	49.0											
FMV	30.6	48.8	66.4	48.6										
HPWMoV	28.9	33.5	31.9	33.0	32.6									
JYMaV	29.0	34.6	33.9	33.9	34.2	41.1								
Palo verde	31.0	32.0	32.0	33.5	33.2	50.0	43.0							
PiVB	30.0	47.1	67.6	49.1	69.6	32.4	34.6	33.2						
PPSMV1	30.4	47.7	52.5	47.9	51.8	31.7	33.4	33.4	53.5					
PPSMV2	31.0	47.2	66.6	49.4	72.1	33.0	34.4	33.5	69.9	53.2				
RLBV	29.9	34.6	33.9	34.4	33.6	41.9	70.8	42.1	34.0	33.5	33.5			
RRV	30.6	47.1	67.7	49.4	68.5	33.1	34.2	33.4	70.5	53.2	68.1	33.4		
RYRSaV	30.5	64.5	46.1	53.9	46.8	33.1	34.2	34.5	48.1	47.3	46.7	34.7	47.0	
TiRSaV	30.5	35.3	34.4	34.5	33.7	41.9	44.0	41.7	34.0	33.2	34.5	44.7	33.4	33.9

	RNA 2 (P2) Glycoprotein													
	KOPV	AcCRaV	BLMaV	EMARaV	FMV	HPWMoV	JYMaV	Palo ver.	PiVB	PPSMV1	PPSMV2	RLBV	RRV	RYRSaV
AcCRaV	19.6													
BLMaV	19.6	36.2												
EMARaV	19.6	41.1	36.2											
FMV	19.2	40.7	53.7	37.5										
HPWMoV	21.6	26.9	24.4	24.1	25.6									
JYMaV	19.1	26.8	24.0	25.0	25.6	35.0								
Palo verde	19.2	26.0	23.3	24.7	22.2	38.9	32.2							
PiVB	19.0	39.3	51.5	35.9	53.3	24.9	23.9	23.5						
PPSMV1	17.6	40.0	40.8	39.3	43.8	25.2	26.1	23.4	43.7					
PPSMV2	18.3	39.8	54.1	36.0	57.3	25.0	24.0	22.9	56.7	43.2				
RLBV	18.9	24.3	25.3	25.3	24.1	35.4	50.8	33.1	24.4	27.5	24.1			
RRV	19.5	38.4	51.2	36.4	49.6	27.7	25.4	23.4	52.5	43.3	51.9	24.8		
RYRSaV	17.7	47.1	36.9	41.6	38.5	25.6	25.1	25.3	38.5	39.4	39.7	24.4	39.1	
TiRSaV	18.2	22.8	22.8	23.8	22.2	32.8	34.3	29.0	24.0	23.9	23.5	35.5	24.3	24.4

	RNA 3 (P3) Nucleocapsid													
	KOPV	AcCRaV	BLMaV	EMARaV	FMV	HPWMoV	JYMaV	Palo ver.	PiVB	PPSMV1	PPSMV2	RLBV	RRV	RYRSaV
AcCRaV	19.6													
BLMaV	21.9	37.1												
EMARaV	21.3	39.9	33.3											
FMV	24.1	37.7	61.1	39.7										
HPWMoV	23.7	17.2	17.9	19.1	19.5									
JYMaV	25.7	19.2	23.0	23.2	23.0	34.1								
Palo verde	22.5	19.6	18.6	17.7	20.6	33.8	29.0							
PiVB	19.3	34.2	55.4	39.9	59.3	20.0	24.3	19.9						
PPSMV1	20.0	35.9	42.5	32.3	41.6	22.8	25.2	22.2	43.0					
PPSMV2	23.1	38.3	62.0	37.2	80.3	21.5	22.0	18.8	61.6	41.9				
RLBV	21.8	18.2	21.6	22.9	22.8	31.4	65.3	28.5	21.2	25.1	22.1			
RRV	20.3	35.0	58.3	32.4	60.1	19.0	23.0	19.8	54.5	42.1	61.5	23.7		
RYRSaV	21.3	53.3	39.0	43.3	38.2	22.0	21.6	21.5	36.7	35.6	36.2	20.4	37.1	
TiRSaV	23.2	22.8	19.6	21.8	20.5	30.4	38.3	29.4	22.3	25.3	20.7	37.8	21.0	25.8

	RNA 4 (P4) Movement													
	KOPV	AcCRaV	BLMaV	EMARaV	FMV	HPWMoV	JYMaV	Palo ver.	PiVB	PPSMV1	PPSMV2	RLBV	RRV	RYRSaV
AcCRaV	17.3													
BLMaV	15.7	31.9												
EMARaV	16.0	31.3	30.2											
FMV	16.4	31.3	55.1	32.6										
HPWMoV	22.0	28.8	18.5	19.8	19.7									
JYMaV	18.1	21.4	19.2	20.7	18.2	42.6								
Palo verde	23.2	18.8	20.7	20.2	20.0	50.8	44.3							
PiVB	16.0	31.2	57.6	31.4	72.9	20.7	21.7	21.1						
PPSMV1	18.3	29.3	38.6	31.1	38.8	19.7	19.8	18.7	39.3					
PPSMV2	17.6	31.0	52.0	32.1	75.1	22.8	22.4	20.2	66.8	40.8				
RLBV	21.8	21.4	20.6	19.4	20.2	44.9	63.3	42.8	22.1	23.3	19.6			
RRV	17.8	29.9	52.1	28.1	59.0	21.9	21.6	21.3	57.9	38.4	61.6	21.7		
RYRSaV	18.8	51.7	34.7	31.9	30.5	20.0	21.6	18.4	34.7	28.1	31.6	21.1	31.5	
TiRSaV	22.5	17.7	17.9	17.4	16.7	50.6	45.1	49.3	16.7	19.0	17.7	47.2	18.6	19.3

Appendix D. Sampling Instructions Flyer

Karaka Leaf Samples Needed!

Samples Needed! Karaka leaves are needed for a Masters project studying a virus that is new to science and could potentially be endemic to New Zealand. **We need your help!** Each sample collected will be tested for the presence of the virus and will provide greater confidence for correlating presence of virus with symptomatic leaves. Instructions on how to sample are on the reverse side.

Karaka, a native New Zealand tree, is distributed nationwide as seen in the pictures above. **It has been identified with a disease that is caused by a plant virus that results in symptoms that appear as pale circles on the leaves** (symptomatic pictures below). Karaka trees are naturally found in coastal areas, but have been planted in reserves and parks around Auckland and New Zealand.

How to identify a Karaka Tree*:

- Karaka is a leafy canopy tree with erect or spreading branches. It grows to heights up to 15 m and has a stout trunk up to 1 m in diameter
- The thick, leathery leaves are glossy, dark green above and paler beneath, 50–200 mm long, and 30–70 mm wide.
- In winter and spring (August to November), karaka produces clusters of tiny flowers, greenish-cream to off-white or pale yellow
- The fruit ripens in summer and autumn (January to April) and is an oval drupe 25–46 mm long, with pale yellow to orange flesh
- Not to be confused with a griselinia, Moreton Bay Fig or Magnolia (which has a brown underleaf).

*From Wikipedia (https://en.wikipedia.org/wiki/Karaka_(tree)

Karaka Leaf Sampling Instructions

- Find a Karaka tree (regardless of symptoms)
- Take two pictures. One that shows the entire tree and any features around it. The other, a close up of the leaf that has been sampled. (Example top right).
- Note the GPS coordinates (this can be done with a phone using the "GPS Camera" app available on Android or iPhone.) The app adds an overlay to photos with GPS data of where the photo was taken from. An example of the coordinates are on the right. More instructions can be found on the app.
- **Important**: If the tree has symptomatic leaves, take a symptomatic AND a symptomless leaf from the same branch. If the tree only has symptomless trees, take one symptomless leaf as a sample.
- Break the stem that attaches the leaf to the branch (by either snapping or cutting).
- Place the leaves in separate ziplock bags along with a tag stating the date, general location and sample number from that location, GPS coordinates of the sample and whether the tree and leaf has symptoms. + for symptomatic, - for symptomless. (Examples on right).
- Send the zip-locked sample in an envelope by post to:

  School of Science
  C/- Dr Colleen Higgins
  Auckland University of Technology (AUT)
  Private Bag 92006
  Auckland 1142,
  New Zealand
- Send photos to Lee.Rabbidge@plantandfood.co.nz

ID:	Location Name + Number		
Date:	30/11/2018		
Species:	Karaka		
Location:	Auckland CBD		
GPS:	-36.849959,	174.766673	
Symptoms - Tree:	-	Leaf:	-

My name is Lee Rabbidge and I am currently working on a Masters thesis project studying viruses within the native Karaka tree.
I am interested in understanding how these virus symptoms are distributed in NZ and plan to test each sample collected for the presence of the virus using various molecular tools.
Each sample tested will contribute to a greater confidence when correlating the virus with the symptomatic leaves.

Important
Please respect sacred spaces, e.g. the urupa (cemetery) at Okahu Bay, Auckland by not collecting samples.
Please also respect private property by not collecting samples.

If you have any questions, please email me at:
Lee.Rabbidge@plantandfood.co.nz

Sample 1

ID:
Date:
Species:
Location:
GPS:
Symptoms - Tree: Leaf:

Sample 2

ID:
Date:
Species:
Location:
GPS:
Symptoms - Tree: Leaf:

References

Açikgöz, S., & Timur Döken, M. (2001). *The Determination of Sampling Time for DSRNA Isolation of the Agent of Fig Mosaic Disaese Prevalent in Agean Region-Turkey.* Paper presented at the II International Symposium on Fig 605.

Ajithkumar, P. (2018). *Diagnosis and Genome Analysis of Lettuce Necrotic Yellows Virus Subgroups.* (Master of Science), Auckland University of Technology, Auckland, New Zealand.

Alemandri, V., Mattio, M. F., Rodriguez, S. M., & Truol, G. (2017). Geographical distribution and first molecular detection of an Emaravirus, High Plains wheat mosaic virus, in Argentina. *European Journal of Plant Pathology*, 1-8. doi:10.1007/s10658-017-1207-8

Alfson, K. J., Beadles, M. W., & Griffiths, A. (2014). A new approach to determining whole viral genomic sequences including termini using a single deep sequencing run. *Journal of Virological Methods, 208*, 1-5. doi:10.1016/j.jviromet.2014.07.023

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., . . . Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology, 37*(4), 420-423. doi:10.1038/s41587-019-0036-z

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Research, 46*(5), 2159-2168. doi:10.1093/nar/gky066

Ashby, J. (1977). Infection of karaka (Corynocarpus laevigatus JR & G. Forst.) by cucumber mosaic virus. *New Zealand Journal of Agricultural Research, 20*(4), 533-534.

Atherton, R., Lockhart, P., McLenachan, P., de Lange, P., Wagstaff, S., & Shepherd, L. (2015). A molecular investigation into the origin and relationships of karaka/kōpi (Corynocarpus laevigatus) in New Zealand. *Journal of the Royal Society of New Zealand, 45*(4), 212-220.

Babu, B., Washburn, B. K., Ertek, T. S., Miller, S. H., Riddle, C. B., Knox, G. W., . . . Paret, M. L. (2017). A field based detection method for Rose rosette virus using isothermal

probe-based Reverse transcription-recombinase polymerase amplification assay. *Journal of Virological Methods, 247*, 81-90. doi:10.1016/j.jviromet.2017.05.019

Banerjee, N., & Mukhopadhyay, S. (2016). Viral glycoproteins: biological role and application in diagnosis. *VirusDisease, 27*(1), 1-11. doi:10.1007/s13337-015-0293-5

Barba, M., Czosnek, H., & Hadidi, A. (2013). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses, 6*(1), 106-136. doi:10.3390/v6010106

Bauchan, G. B., Otero-Colina, G., Hammond, J., Jordan, R., & Ochoa, R. (2019) Rose rosette disease: It all started with a small mite. *Vol. 1232. Acta Horticulturae* (pp. 227-232).

Baudhuin, L. M., Lagerstedt, S. A., Klee, E. W., Fadra, N., Oglesbee, D., & Ferber, M. J. (2015). Confirming variants in next-generation sequencing panel testing by sanger sequencing. *Journal of Molecular Diagnostics, 17*(4), 456-461. doi:10.1016/j.jmoldx.2015.03.004

Bhat, A. I., Hohn, T., & Selvarajan, R. (2016). Badnaviruses: The current global scenario. *Viruses, 8*(6). doi:10.3390/v8060177

Bi, Y., Tugume, A. K., & Valkonen, J. P. T. (2012). Small-RNA deep sequencing reveals Arctium tomentosum as a natural host of Alstroemeria virus X and a new putative emaravirus. *PLoS ONE, 7*(8). doi:10.1371/journal.pone.0042758

Bonney, R., Phillips, T. B., Ballard, H. L., & Enck, J. W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science, 25*(1), 2-16.

Bos, L. (1981). Hundred years of Koch's Postulates and the history of etiology in plant virus research. *Netherlands Journal of Plant Pathology, 87*(3), 91-110. doi:10.1007/BF01976645

Brown, P., Newstrom-Lloyd, L. E., Foster, B. J., Badger, P. H., & McLean, J. A. (2018). Winter 2016 honey bee colony losses in New Zealand. *Journal of Apicultural Research, 57*(2), 278-291.

Buzkan, N., Chiumenti, M., Massart, S., Sarpkaya, K., Karadağ, S., & Minafra, A. (2019). A new emaravirus discovered in Pistacia from Turkey. *Virus Research, 263*, 159-163. doi:10.1016/j.virusres.2019.01.012

Chooi, K. M., Cohen, D., & Pearson, M. N. (2013). Generic and sequence-variant specific molecular assays for the detection of the highly variable Grapevine leafroll-associated virus 3. *Journal of Virological Methods, 189*(1), 20-29. doi:10.1016/j.jviromet.2012.12.018

Cieniewicz, E., Thompson, J. R., McLane, H., Perry, K. L., Dangl, G. S., Corbett, Q., . . . O'Connell, J. (2018). Prevalence and Genetic Diversity of Grabloviruses in Free-Living Vitis spp. *Plant Disease, 102*(11), 2308-2316. doi:10.1094/PDIS-03-18-0496-RE

Costall, J. A., Carter, R. J., Shimada, Y., Anthony, D., & Rapson, G. L. (2006). The endemic tree Corynocarpus laevigatus (karaka) as a weedy invader in forest remnants of southern North Island, New Zealand. *New Zealand Journal of Botany, 44*(1), 5-22. doi:10.1080/0028825X.2006.9513002

de Lillo, E., Pozzebon, A., Valenzano, D., & Duso, C. (2018). An intimate relationship between eriophyoid mites and their host plants – A review. *Frontiers in Plant Science, 9*. doi:10.3389/fpls.2018.01786

Di Bello, P. L., Ho, T., & Tzanetakis, I. E. (2015). The evolution of emaraviruses is becoming more complex: seven segments identified in the causal agent of Rose rosette disease. *Virus Research, 210*, 241-244. doi:10.1016/j.virusres.2015.08.009

Di Bello, P. L., Laney, A. G., Druciarek, T., Ho, T., Gergerich, R. C., Keller, K. E., . . . Tzanetakis, I. E. (2016). A novel emaravirus is associated with redbud yellow ringspot disease. *Virus Research, 222*, 41-47. doi:10.1016/j.virusres.2016.05.027

Di Serio, F., Ambrós, S., Sano, T., Flores, R., & Navarro, B. (2018). Viroid diseases in pome and stone fruit trees and Koch's postulates: A critical assessment. *Viruses, 10*(11). doi:10.3390/v10110612

Divya, P., Kumar, L., Rangaswamy, K., & Muniyappa, V. (2005). Detection of pigonpea sterility mosaic virus in floral parts and seeds. *Indian J. Virol, 16*, 36.

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews, 76*(2), 159-216. doi:10.1128/MMBR.05023-11

Ebrahim‐Nesbat, F., & Izadpanah, K. (1992). Viruslike particles associated with ringfleck mosaic of mountain ash and a mosaic disease of raspberry in the Bavarian Forest. *Forest Pathology, 22*(1), 1-10. doi:10.1111/j.1439-0329.1992.tb01426.x

Eigen, M. (1993). Viral quasispecies. *Scientific American, 269*(1), 42-49.

Elbeaino, Digiaro, Uppala, & Sudini. (2014). Deep sequencing of pigeonpea sterility mosaic virus discloses five RNA segments related to emaraviruses. *Virus Research, 188*, 27-31.

Elbeaino, T., Digiaro, M., & Martelli, G. P. (2009). Complete nucleotide sequence of four RNA segments of fig mosaic virus. *Archives of Virology, 154*(11), 1719-1727. doi:10.1007/s00705-009-0509-3

Elbeaino, T., Digiaro, M., Mielke-Ehret, N., Muehlbach, H. P., Martelli, G. P., & Ictv Report, C. (2018). ICTV Virus Taxonomy Profile: Fimoviridae. *The Journal of General Virology, 99*(11), 1478-1479. doi:10.1099/jgv.0.001143

Elbeaino, T., Digiaro, M., Uppala, M., & Sudini, H. (2015). Deep sequencing of dsRNAs recovered from mosaic-diseased pigeonpea reveals the presence of a novel emaravirus: pigeonpea sterility mosaic virus 2. *Archives of Virology, 160*(8), 2019-2029. doi:10.1007/s00705-015-2479-y

Elbeaino, T., Marais, A., Faure, C., Trioano, E., Candresse, T., & Parrella, G. (2018). High-Throughput Sequencing Reveals Cyclamen persicum Mill. as a Natural Host for Fig Mosaic Virus. *Viruses, 10*(12), 684.

Elbeaino, T., Whitfield, A., Sharma, M., & Digiaro, M. (2013). Emaravirus-specific degenerate PCR primers allowed the identification of partial RNA-dependent RNA polymerase sequences of Maize red stripe virus and Pigeonpea sterility mosaic virus. *Journal of Virological Methods, 188*(1-2), 37-40. doi:10.1016/j.jviromet.2012.11.037

Elliott, R. M. (2014). Orthobunyaviruses: recent genetic and structural insights. *Nature Reviews Microbiology, 12*(10), 673-685. doi:10.1038/nrmicro3332

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols, 2*(4), 953-971. doi:10.1038/nprot.2007.131

Epstein, A., & Hill, J. (1999). Status of rose rosette disease as a biological control for multiflora rose. *Plant Disease, 83*(2), 92-101.

Forster, R., Seifers, D., Strausbaugh, C., Jensen, S., Ball, E., & Harvey, T. (2001). Seed transmission of the High Plains virus in sweet corn. *Plant Disease, 85*(7), 696-699. doi:10.1094/PDIS.2001.85.7.696

Frohman, M. A. (1990). RACE: Rapid Amplification of cDNA Ends. *PCR Protocols: A Guide to Methods and Applications, 28*.

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., . . . Bryant, S. H. (2009). The NCBI biosystems database. *Nucleic Acids Research, 38*(suppl_1), D492-D496.

Gupta, A. K., Hein, G. L., Graybosch, R. A., & Tatineni, S. (2018). Octapartite negative-sense RNA genome of High Plains wheat mosaic virus encodes two suppressors of RNA silencing. *Virology, 518,* 152-162. doi:10.1016/j.virol.2018.02.013

Gupta, A. K., Hein, G. L., & Tatineni, S. (2019). P7 and P8 proteins of High Plains wheat mosaic virus, a negative-strand RNA virus, employ distinct mechanisms of RNA silencing suppression. *Virology*. doi:https://doi.org/10.1016/j.virol.2019.06.011

Gupta, R., Jung, E., & Brunak, S. (2004). Prediction of N-glycosylation sites in human proteins.

Guterres, A., de Oliveira, R. C., Fernandes, J., de Lemos, E. R. S., & Schrago, C. G. (2017). New bunya-like viruses: Highlighting their relations. *Infection, Genetics and Evolution, 49,* 164-173. doi:10.1016/j.meegid.2017.01.019

Hassan, M., Di Bello, P. L., Keller, K. E., Martin, R. R., Sabanadzovic, S., & Tzanetakis, I. E. (2017). A new, widespread emaravirus discovered in blackberry. *Virus Research, 235,* 1-5.

Hassan, M., & Tzanetakis, I. E. (2019). Population structure, evolution and detection of blackberry leaf mottle-associated virus, an emerging emaravirus. *Plant Pathology*. doi:10.1111/ppa.12992

Hipper, C., Brault, V., Ziegler-Graff, V., & Revers, F. (2013). Viral and cellular factors involved in phloem transport of plant viruses. *Frontiers in Plant Science, 4*(MAY). doi:10.3389/fpls.2013.00154

Ho, T., & Tzanetakis, I. E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. *Virology, 471-473,* 54-60. doi:10.1016/j.virol.2014.09.019

Hofmann, K. (1993). TMbase-A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler, 374,* 166.

Holmes, E. C. (2003). Molecular clocks and the puzzle of RNA virus origins. *Journal of Virology, 77*(7), 3893-3897. doi:10.1128/JVI.77.7.3893-3897.2003

Hsieh, S. Y., Yang, P. Y., Chen, H. C., & Liaw, Y. F. (1997). Cloning and characterization of the extreme 5′ -terminal sequences of the RNA genomes of GB virus C/hepatitis

G virus. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(7), 3206-3210. doi:10.1073/pnas.94.7.3206

Huang, B., Jennison, A., Whiley, D., McMahon, J., Hewitson, G., Graham, R., . . . Warrilow, D. (2019). Illumina sequencing of clinical samples for virus detection in a public health laboratory. *Scientific reports, 9*(1), 1-8.

Hunt, H., Cave, N., Gartrell, B., Cogger, N., Petersen, J., & Roe, W. (2018). An epidemiological investigation of an idiopathic myopathy in hunting dogs in New Zealand. *New Zealand Veterinary Journal, 66*(4), 199-204.

ICTV. (2019). *International Committee on Taxonomy Virus (ICTV): 2019 Release [Internet]*. Retrieved from Washington DC, United States of America: https://talk.ictvonline.org/taxonomy/

Ilyas, M., Avelar, S., Schuch, U. K., & Brown, J. K. (2018). First report of an emaravirus associated with witches' broom disease and eriophyid mite infestations of the blue palo verde tree in Arizona. *Plant Disease, 102*(9), 1863. doi:10.1094/PDIS-01-18-0124-PDN

Inglis, T. J. (2007). Principia aetiologica: taking causality beyond Koch's postulates. *Journal of Medical Microbiology, 56*(11), 1419-1422.

Ishikawa, K., Maejima, K., Netsu, O., Fukuoka, M., Nijo, T., Hashimoto, M., . . . Namba, S. (2015). Rapid detection of fig mosaic virus using reverse transcription loop-mediated isothermal amplification. *Journal of General Plant Pathology, 81*(5), 382-389. doi:10.1007/s10327-015-0603-1

Iskra-Caruana, M.-l., Duroy, P.-O., Chabannes, M., & Muller, E. (2014). The common evolutionary history of badnaviruses and banana. *Infection, Genetics and Evolution, 21*, 83-89. doi:10.1016/j.meegid.2013.10.013

Jones, R. A. C. (2014). Trends in plant virus epidemiology: Opportunities from new or improved technologies. *Virus Research, 186*, 3-19. doi:10.1016/j.virusres.2013.11.003

Jones, R. A. C. (2018) Plant and Insect Viruses in Managed and Natural Environments: Novel and Neglected Transmission Pathways. *Vol. 101. Advances in Virus Research* (pp. 149-187).

Katsiani, A., Stainton, D., Lamour, K., & Tzanetakis, I. E. (2020). The population structure of Rose rosette virus in the USA. *Journal of General Virology, 101*(6), 676-684. doi:https://doi.org/10.1099/jgv.0.001418

Katsiani, A. T., Pappi, P., Olmos, A., Efthimiou, K. E., Maliogka, V. I., & Katis, N. I. (2018). Development of a real-time RT-PCR for the universal detection of lchv1 and study of the seasonal fluctuation of the viral titer in sweet cherry cultivars. *Plant Disease, 102*(5), 899-904. doi:10.1094/PDIS-01-17-0107-RE

Khan, S., Mackay, J., Liefting, L., & Ward, L. (2015). Development of a duplex one-step RT-qPCR assay for the simultaneous detection of Apple scar skin viroid and plant RNA internal control. *Journal of Virological Methods, 221*, 100-105. doi:10.1016/j.jviromet.2015.04.032

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biology Direct, 1*. doi:10.1186/1745-6150-1-29

Kormelink, R., Garcia, M. L., Goodin, M., Sasaya, T., & Haenni, A. L. (2011). Negative-strand RNA viruses: The plant-infecting counterparts. *Virus Research, 162*(1-2), 184-202. doi:10.1016/j.virusres.2011.09.028

Kubota, K., Usugi, T., Tomitaka, Y., Shimomoto, Y., Takeuchi, S., Kadono, F., . . . Tsuda, S. (2020). Perilla mosaic virus is a highly divergent emaravirus transmitted by Shevtchenkella sp. (acari: Eriophyidae). *Phytopathology, 110*(7), 1352-1361. doi:10.1094/PHYTO-01-20-0013-R

Kuhn, J. H., Wolf, Y. I., Krupovic, M., Zhang, Y. Z., Maes, P., Dolja, V. V., & Koonin, E. V. (2019). Classify viruses — the gain is worth the pain. *Nature, 566*(7744), 318-320. doi:10.1038/d41586-019-00599-8

Kulkarni, N., Kumar, P. L., Muniyappa, V., Jones, A. T., & Reddy, a. V. (2002). Transmission of Pigeon pea sterility mosaic virus by the eriophyid mite, Aceria cajani (Acari: Arthropoda). *Plant Disease, 86*(12), 1297-1302. doi:10.1094/PDIS.2002.86.12.1297

Kumar, K. R., Cowley, M. J., & Davis, R. L. (2019). Next-Generation Sequencing and Emerging Technologies. *Seminars in Thrombosis and Hemostasis, 45*(7), 661-673. doi:10.1055/s-0039-1688446

Kumar, P. L., Jones, A. T., & Reddy, D. (2003). A novel mite-transmitted virus with a divided RNA genome closely associated with pigeonpea sterility mosaic disease. *Phytopathology, 93*(1), 71-81. doi:10.1094/PHYTO.2003.93.1.71

Laney, A. G., Keller, K. E., Martin, R. R., & Tzanetakis, I. E. (2011). A discovery 70 years in the making: Characterization of the Rose rosette virus. *Journal of General Virology, 92*(7), 1727-1732. doi:10.1099/vir.0.031146-0

Leach, H., & Stowe, C. (2005). Oceanic arboriculture at the margins - The case of the Karaka (Corynocarpus laevigatus) in Aotearoa. *Journal of the Polynesian Society, 114*(1), 7-27.

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of Experimental Botany, 68*(20), 5419-5429. doi:10.1093/jxb/erx289

Li, Z., Yu, M., Zhang, H., Wang, H. Y., & Wang, L. F. (2005). Improved rapid amplification of cDNA ends (RACE) for mapping both the 5′ and 3′ terminal sequences of paramyxovirus genomes. *Journal of Virological Methods, 130*(1-2), 154-156. doi:10.1016/j.jviromet.2005.06.022

Linhart, C., & Shamir, R. (2005). The degenerate primer design problem: Theory and applications. *Journal of Computational Biology, 12*(4), 431-456. doi:10.1089/cmb.2005.12.431

Louie, R., Seifers, D. L., & Bradfute, O. E. (2006). Isolation, transmission and purification of the High Plains virus. *Journal of virological methods, 135*(2), 214-222. doi:10.1016/j.jviromet.2006.03.023

Lupas, A., Van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, 1162-1164.

Macaskill, J. J., Manley-Harris, M., & Field, R. J. (2015). Quantification of nitropropanoyl glucosides in karaka nuts before and after treatment. *Food Chemistry, 175*, 543-548. doi:10.1016/j.foodchem.2014.12.022

Mackenzie, A. M., Nolan, M., Wei, K. J., Clements, M. A., Gowanlock, D., Wallace, B. J., & Gibbs, A. J. (1998). Ceratobium mosaic potyvirus: Another virus from orchids. *Archives of Virology, 143*(5), 903-914. doi:10.1007/s007050050341

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., . . . Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research, 43*(D1), D222-D226. doi:10.1093/nar/gku1221

Martin, N. A. (2017). Karaka gall mite - *Aculus corynocarpi. Interesting Insects and other Invertebrates. New Zealand Arthropod Factsheet Series Number 87.*

Martinelli, F., Scalenghe, R., Davino, S., Panno, S., Scuderi, G., Ruisi, P., . . . Dandekar, A. M. (2015). Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development, 35*(1), 1-25. doi:10.1007/s13593-014-0246-1

Mauck, K. E., De Moraes, C. M., & Mescher, M. C. (2010). Deceptive chemical signals induced by a plant virus attract insect vectors to inferior hosts. *Proceedings of the National Academy of Sciences of the United States of America, 107*(8), 3600-3605. doi:10.1073/pnas.0907191107

McGavin, W. J., Mitchell, C., Cock, P. J., Wright, K. M., & MacFarlane, S. A. (2012). Raspberry leaf blotch virus, a putative new member of the genus Emaravirus, encodes a novel genomic RNA. *Journal of General Virology, 93*(2), 430-437. doi:10.1099/vir.0.037937-0

Menzel, W., Jelkmann, W., & Maiss, E. (2002). Detection of four apple viruses by multiplex RT-PCR assays with coamplification of plant mRNA as internal control. *Journal of Virological Methods, 99*(1-2), 81-92. doi:10.1016/S0166-0934(01)00381-0

Mielke-Ehret, N., & Mühlbach, H. P. (2012). Emaravirus: Anovel genus of multipartite, negative strand RNA plant viruses. *Viruses, 4*(9), 1515-1536. doi:10.3390/v4091515

Mielke-Ehret, N., Thoma, J., Schlatermund, N., & Mühlbach, H. P. (2010). Detection of European mountain ash ringspot-associated virus-specific RNA and protein P3 in the pear leaf blister mite Phytoptus pyri (Eriophyidae). *Archives of Virology, 155*(6), 987-991. doi:10.1007/s00705-010-0667-3

Mielke, N., & Muehlbach, H.-P. (2007). A novel, multipartite, negative-strand RNA virus is associated with the ringspot disease of European mountain ash (Sorbus aucuparia L.). *Journal of General Virology, 88*(4), 1337-1346. doi:10.1099/vir.0.82715-0

Mitter, N., Koundal, V., Williams, S., & Pappu, H. (2013). Differential Expression of Tomato Spotted Wilt Virus-Derived Viral Small RNAs in Infected Commercial and Experimental Host Plants. *PLoS ONE, 8*(10). doi:10.1371/journal.pone.0076276

Mushegian, A. R., & Elena, S. F. (2015). Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology, 476*, 304-315. doi:10.1016/j.virol.2014.12.012

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N., & Hase, T. (2000). Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research, 28*(12).

Olmedo-Velarde, A., Ochoa-Corona, F., & Elbeaino, T. (2016). Toward broad detection of emaraviruses: Endpoint RT-PCR. *Phytopathology, 106*, S4.

Olmedo-Velarde, A., Park, A. C., Sugano, J., Uchida, J. Y., Kawate, M., Borth, W. B., . . . Melzer, M. J. (2019). Characterization of Ti Ringspot-Associated Virus, a Novel Emaravirus Associated with an Emerging Ringspot Disease of Cordyline fruticosa. *Plant Disease*, PDIS-09-18-1513-RE.

Pang, M., Gayral, M., Lyle, K., Shires, M. K., Ong, K., Byrne, D., & Verchot, J. (2019). Infectious DNA clone technology and inoculation strategy for Rose Rosette Virus that includes all seven segments of the negative-strand RNA genome. *bioRxiv*, 712000.

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., & Ravnikar, M. (2017). Next generation sequencing for detection and discovery of plant viruses and viroids: Comparison of two approaches. *Frontiers in Microbiology, 8*(OCT). doi:10.3389/fmicb.2017.01998

Pei, J., & Grishin, N. V. (2007). PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics, 23*(7), 802-808.

Potapov, V., Fu, X., Dai, N., Corrêa, I. R., Jr., Tanner, N. A., & Ong, J. L. (2018). Base modifications affecting RNA polymerase and reverse transcriptase fidelity. *Nucleic Acids Research, 46*(11), 5753-5763. doi:10.1093/nar/gky341

Pumplin, N., & Voinnet, O. (2013). RNA silencing suppression by plant pathogens: Defence, counter-defence and counter-counter-defence. *Nature Reviews Microbiology, 11*(11), 745-760. doi:10.1038/nrmicro3120

QGIS. (2019). QGIS Geographic Information System. Retrieved from http://qgis.org

Rao, P., Yuan, W., & Krug, R. M. (2003). Crucial role of CA cleavage sites in the cap‐snatching mechanism for initiating viral mRNA synthesis. *The EMBO Journal, 22*(5), 1188-1198. doi:10.1093/emboj/cdg109

Robbach, J., Dieckmann, H. L., Büttner, T., Mühlbach, H. P., Von Bargen, S., & Büttner, C. (2015). Genetic variability and phylogeny of European mountain ash ringspot-associated virus RNA3 and RNA4. *Forests, 6*(11), 4072-4087. doi:10.3390/f6114072

Rowhani, A., Daubert, S., Arnold, K., Al Rwahnih, M., Klaassen, V., Golino, D., & Uyemoto, J. K. (2018). Synergy between grapevine vitiviruses and grapevine leafroll viruses. *European Journal of Plant Pathology, 151*(4), 919-925. doi:10.1007/s10658-018-1426-7

Samarfard, S., McTaggart, A. R., Sharman, M., Bejerman, N. E., & Dietzgen, R. G. (2020). Viromes of ten alfalfa plants in Australia reveal diverse known viruses and a novel RNA virus. *Pathogens, 9*(3). doi:10.3390/pathogens9030214

Sawyer, J. W. D., McFadgen, B., & Hughes, P. M. (2003). *Karaka (Corynocarpus laevigatus JR et G. Forst.) in Wellington Conservancy (excluding Chatham Islands)*: Citeseer.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics, 19*(R2), R227-R240. doi:10.1093/hmg/ddq416

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods, 5*(1), 16-18. doi:10.1038/nmeth1156

Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R. R., Kasschau, K., Dolja, V. V., . . . Pooggin, M. M. (2014). De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PLoS ONE, 9*(2). doi:10.1371/journal.pone.0088513

Serrano, L., Ramon, J., Segarra, J., Medina, V., Achón, M. A., López, M., & Juárez, M. (2004) New approach in the identification of the causal agent of fig mosaic disease. *Vol. 657. Acta Horticulturae* (pp. 559-566).

Shahmirzaie, M., Safarnejad, M. R., Rakhshandehroo, F., Safarpour, H., Rabbani, H., Zamanizadeh, H. R., & Elbeaino, T. (2019). Production of a polyclonal antiserum against recombinant nucleocapsid protein and its application for the detection of fig mosaic virus. *Journal of Virological Methods, 265*, 22-25. doi:10.1016/j.jviromet.2018.12.016

Silva, G., Oyekanmi, J., Nkere, C. K., Bömer, M., Kumar, P. L., & Seal, S. E. (2018). Rapid detection of potyviruses from crude plant extracts. *Analytical Biochemistry, 546*, 17-22. doi:10.1016/j.ab.2018.01.019

Sint, D., Raso, L., & Traugott, M. (2012). Advances in multiplex PCR: Balancing primer efficiencies and improving detection success. *Methods in Ecology and Evolution, 3*(5), 898-905. doi:10.1111/j.2041-210X.2012.00215.x

Slykhuis, J. T. (1961). Eriophyid mites in relation to the spread of grass viruses in Ontario. *Canadian Journal of Plant Science, 41*(2), 304-308.

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester‐Christensen, M. B., Katrine, T., . . . Pedersen, N. B. (2013). Precision mapping of the human O‐GalNAc glycoproteome through SimpleCell technology. *The EMBO Journal, 32*(10), 1478-1488.

Stewart, L. R. (2016). Sequence diversity of wheat mosaic virus isolates. *Virus Research, 213*, 299-303. doi:10.1016/j.virusres.2015.11.013

Susaimuthu, J., Tzanetakis, I. E., Gergerich, R. C., Kim, K. S., & Martin, R. R. (2008). Viral interactions lead to decline of blackberry plants. *Plant Disease, 92*(9), 1288-1292. doi:10.1094/PDIS-92-9-1288

Tatineni, S., McMechan, A. J., Wosula, E. N., Wegulo, S. N., Graybosch, R. A., French, R., & Hein, G. L. (2014). An eriophyid mite-transmitted plant virus contains eight genomic RNA segments with unusual heterogeneity in the nucleocapsid protein. *Journal of Virology, 88*(20), 11834-11845. doi:10.1128/JVI.01901-14

Thomson, D., & Dietzgen, R. G. (1995). Detection of DNA and RNA plant viruses by PCR and RT-PCR using a rapid virus release protocol without tissue homogenization. *Journal of Virological Methods, 54*(2-3), 85-95. doi:10.1016/0166-0934(95)00022-M

van Dijk, E. L. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*(9), 418-426. doi:10.1016/j.tig.2014.07.001

Van Essen, R., & Rapson, G. (2005). Fruit size of karaka (Corynocarpus laevigatus) in relation to potential selection by Maori. *New Zealand Botanical Society Newsletter, 81*, 13-16.

Varsani, A., & Krupovic, M. (2017). Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. *Virus Evolution, 3*(1), vew037.

Veerakone, S., Tang, J., Ward, L., Liefting, L., Perez-Egusquiza, Z., Lebas, B., . . . Guy, P. (2015). A review of the plant virus, viroid, liberibacter and phytoplasma records for New Zealand. *Australasian Plant Pathology, 44*(5), 463-514.

von Bargen, S., Al Kubrusli, R., Gaskin, T., Fürl, S., Hüttner, F., Blystad, D. R., . . . Büttner, C. (2020). Characterisation of a novel Emaravirus identified in mosaic-diseased Eurasian aspen (Populus tremula). *Annals of Applied Biology*. doi:10.1111/aab.12576

von Bargen, S., Dieckmann, H. L., Candresse, T., Mühlbach, H. P., Roßbach, J., & Büttner, C. (2019). Determination of the complete genome sequence of European mountain ash ringspot-associated emaravirus from Sorbus intermedia reveals two additional genome segments. *Archives of Virology*. doi:10.1007/s00705-019-04275-0

Wang, Q., Ma, X., Qian, S., Zhou, X., Sun, K., Chen, X., . . . Li, Z. (2015). Rescue of a Plant Negative-Strand RNA Virus from Cloned cDNA: Insights into Enveloped Plant Virus Movement and Morphogenesis. *PLoS Pathogens, 11*(10). doi:10.1371/journal.ppat.1005223

Wang, Y., Zhai, L., Wen, S., Yang, Z., Wang, G., & Hong, N. (2020). Molecular characterization of a novel emaravrius infecting Actinidia spp. in China. *Virus Research, 275*. doi:10.1016/j.virusres.2019.197736

Wang, Y. Q., Song, Y., Cao, M. J., Cheng, Q., Wu, J. X., & Hu, T. (2020). Identification of a novel emaravirus infecting lilac through next-generation sequencing. *Journal of Integrative Agriculture, 19*(8), 2064-2071. doi:10.1016/S2095-3119(19)62806-6

White, E. J., Venter, M., Hiten, N. F., & Burger, J. T. (2008). Modified cetyltrimethylammonium bromide method improves robustness and versatility: The benchmark for plant RNA extraction. *Biotechnology Journal, 3*(11), 1424-1428. doi:10.1002/biot.200800207

Yang, C., Zhang, S., Han, T., Fu, J., Di Serio, F., & Cao, M. (2019). Identification and characterization of a novel emaravirus associated with jujube (Ziziphus jujuba Mill.) yellow mottle disease. *Frontiers in Microbiology, 10*.

York, A., Hengrung, N., Vreede, F. T., Huiskonen, J. T., & Fodor, E. (2013). Isolation and characterization of the positive-sense replicative intermediate of a negative-strand RNA virus. *Proceedings of the National Academy of Sciences of the United States of America, 110*(45), E4238-E4245. doi:10.1073/pnas.1315068110

Yu, C., Karlin, D. G., Lu, Y., Wright, K., Chen, J., & MacFarlane, S. (2013). Experimental and bioinformatic evidence that raspberry leaf blotch emaravirus P4 is a movement protein of the 30K superfamily. *Journal of General Virology, 94*(PART9), 2117-2128. doi:10.1099/vir.0.053256-0

Zheng, Y., Navarro, B., Wang, G., Wang, Y., Yang, Z., Xu, W., . . . Hong, N. (2017). Actinidia chlorotic ringspot‐associated virus: a novel emaravirus infecting kiwifruit plants. *Molecular Plant Pathology, 18*(4), 569-581. doi:10.1111/mpp.12421

Zhou, X., Lin, W., Sun, K., Wang, S., Zhou, X., Jackson, A. O., & Lia, Z. (2019). Specificity of plant rhabdovirus cell-to-cell movement. *Journal of Virology, 93*(15). doi:10.1128/JVI.00296-19